

Strategies to Improve Quantitative Proteomics:  
Implications of Dimethyl Labelling and Novel Peptide Detection

by

Joseph Michael Boutilier

Submitted in partial fulfilment of the requirements  
for the degree of Master of Science

at

Dalhousie University  
Halifax, Nova Scotia  
March 2012

© Copyright by Joseph Michael Boutilier, 2012

DALHOUSIE UNIVERSITY  
DEPARTMENT OF CHEMISTRY

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled “Strategies to Improve Quantitative Proteomics: Implications of Dimethyl Labelling and Novel Peptide Detection” by Joseph Michael Boutilier in partial fulfilment of the requirements for the degree of Master of Science.

Dated: March 21, 2012

Supervisor: \_\_\_\_\_

Readers: \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

DALHOUSIE UNIVERSITY

DATE: March 21, 2012

AUTHOR: Joseph Michael Boutilier

TITLE: Strategies to Improve Quantitative Proteomics: Implications of Dimethyl  
Labelling and Novel Peptide Detection

DEPARTMENT OR SCHOOL: Department of Chemistry

DEGREE: MSc CONVOCATION: May YEAR: 2012

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions. I understand that my thesis will be electronically available to the public.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

The author attests that permission has been obtained for the use of any copyrighted material appearing in the thesis (other than the brief excerpts requiring only proper acknowledgement in scholarly writing), and that all such use is clearly acknowledged.

---

Signature of Author

# Table of Contents

<b>List of Tables</b> .....	<b>vii</b>
<b>List of Figures</b> .....	<b>viii</b>
<b>Abstract</b> .....	<b>xi</b>
<b>List of Abbreviations</b> .....	<b>xii</b>
<b>Acknowledgements</b> .....	<b>xiii</b>
<b>Chapter 1 Introduction</b> .....	<b>1</b>
1.1 Overview.....	1
1.2 Proteomics.....	2
1.3 Objectives of Proteomics.....	3
1.4 MS-based Proteomics.....	4
1.5 Biomarker Discovery.....	8
1.6 Motivation For Thesis Research.....	9
1.7 Outline of Thesis.....	11
<b>Chapter 2 Isotopic Labelling Techniques</b> .....	<b>15</b>
2.1 Chemically Based Methods.....	15
2.1.1 Isotope-Coded Affinity Tags (ICATs).....	16
2.1.2 Isobaric Tags for Relative and Absolute Quantitation (iTRAQ).....	20
2.1.3 Stable-Isotope Dimethyl Labelling.....	24
2.2 Biological Based Methods.....	26
2.2.1 Metabolic Methods.....	26
2.2.2 Enzymatic Methods.....	27
2.3 Absolute Quantitation by Spiked Synthetic Peptide Standards.....	28
2.4 Label-Free Methods.....	29
2.5 Summary.....	30
<b>Chapter 3 Methods and Data Analysis</b> .....	<b>31</b>
3.1 Materials and Methods.....	31
3.1.1 Reagents and Standards.....	31
3.1.2 Yeast Extraction and Preparation.....	32
3.1.3 Tryptic Digestion.....	32

	3.1.4	Isotopic Labelling.....	33
	3.1.5	Liquid Chromatography (LC).....	33
	3.1.6	LC-MS/MS.....	34
	3.1.7	SEQUEST Parameters.....	35
	3.2	Data Preprocessing.....	35
<b>Chapter 4</b>		<b>Chromatographic Behaviour of Peptides Following Dimethylation with H<sub>2</sub>/D<sub>2</sub>-Formaldehyde: Implications for Comparative Proteomics.....</b>	<b>40</b>
	4.1	Introduction.....	40
	4.2	Methods.....	43
	4.2.1	Data Selection and Preprocessing.....	43
	4.2.2	Signal Processing.....	45
	4.2.3	Computational Calculations of Compound Polarity.....	46
	4.3	Results.....	46
	4.3.1	Retention Time Differences.....	47
	4.3.2	Resolution.....	49
	4.3.3	Quantitative Ratios.....	51
	4.4	Discussion.....	52
	4.5	Concluding Remarks.....	58
<b>Chapter 5</b>		<b>Peptide Pair Detection Algorithm.....</b>	<b>59</b>
	5.1	Strategy for Peptide Detection Based on MS Data.....	59
	5.2	Methods to Predict Isotopic Patterns of Tagged Peptides.....	62
	5.2.1	Estimating Elemental Composition.....	62
	5.2.2	Calculating Isotopic Ratios.....	65
	5.3	Algorithms for Peptide Detection.....	69
	5.3.1	Autocorrelation Function.....	70
	5.3.2	Parallel Isotopic Tag Screening.....	73
	5.3.3	Mapping of the Classification Matrix, $\Theta$ .....	85
	5.4	Performance of PITS Algorithm.....	87
	5.5	Summary.....	94
<b>Chapter 6</b>		<b>Conclusions.....</b>	<b>96</b>
	6.1	Conclusions.....	96

6.2	Future Work.....	98
	<b>References.....</b>	<b>101</b>
	Appendix A – Procedure to Convert RAW File to mzXML File.....	112
	Appendix B – base64cvt.m MatLab <sup>®</sup> Code.....	113
	Appendix C – rawexportfull.m MatLab <sup>®</sup> Code.....	114
	Appendix D – Complete List of 71 BSA Peptide Pairs Used in Chapter 4.....	116
	Appendix E – Complete List of 535 Peptide Pairs Found by PITS.....	118

## List of Tables

<b>Table 2.1</b> Combination of isotopes for iTRAQ reagents to maintain a combined molecular weight of 145.....	21
<b>Table 5.1</b> Subwindow ranges for the PITS algorithm.....	75
<b>Table 5.2</b> Summary of the 535 detected peptide pairs from the PITS algorithm.....	89

## List of Figures

<b>Figure 1.1</b> Visual representation of the two common MS-based quantitative proteomics approaches.....	5
<b>Figure 1.2</b> Visual representation of the proposed approach for the research in this thesis.....	12
<b>Figure 2.1</b> Structure of the original ICAT reagent.....	16
<b>Figure 2.2</b> Structure of the cICAT reagent.....	19
<b>Figure 2.3</b> Structure of the iTRAQ reagent.....	21
<b>Figure 2.4</b> Reductive amination reaction for CH <sub>2</sub> O and CD <sub>2</sub> O.....	25
<b>Figure 3.1</b> A selection of an mzXML file from the BSA data set with the Header (A), Sub-header (MS scan (B), ZOOM scan (C) and MS/MS scan (D)) and Body (E) sections.....	37
<b>Figure 3.2</b> The procedure for the first <i>m/z</i> data value (A) which is put into a data matrix (B) and plotted as a mass spectrum (C).....	39
<b>Figure 4.1</b> Two-dimensional representation of a mass chromatogram for one BSA replicate (A), an example of a peptide pair (B), a three dimensional surface plot (C) of the response indicated on (A), and the calculated XIC (D) for the response, where the dashed lines are the estimated median retention times for the peptide pair....	45
<b>Figure 4.2</b> Calculated mean retention time differences as a function of apparent peptide mass for the 71 unique BSA peptide pairs.....	48
<b>Figure 4.3</b> Four selected XIC replicates for the indicated peptide pairs at positions “a” through “d” in Figure 4.2.....	49
<b>Figure 4.4</b> Calculated mean resolution as a function of apparent mass for the 71 unique BSA peptide pairs.....	50
<b>Figure 4.5</b> (A) Strategy for calculation of ratios based on a single time region (R <sub>1</sub> ) and (B) based on two time regions (R <sub>2</sub> ); (C) R <sub>1</sub> ratios calculated for 71 peptides with inset histogram; (D) R <sub>2</sub> ratios with inset histogram; (E) Ratio differences plotted against resolution.....	53
<b>Figure 4.6</b> Four replicate XICs for the peptide pair identified as QTALVELLK.....	54
<b>Figure 4.7</b> Dipole moments of labelled surrogates (deuterated version).....	56



<b>Figure 5.1</b> Common isotopic patterns for singly charge peptides with one (A) and two tags (B), doubly charge peptides with one (C), two (D) and three tags (E), and triply charged peptides with one (F), two (G) and three tags (H).....	60
<b>Figure 5.2</b> The amino acid distribution (B) based on the BSA sequence (A) and the elemental distribution in each amino acid (C).....	63
<b>Figure 5.3</b> The calculated isotopic ratios (A) and the isotopic pattern (B) for a BSA peptide detected at 600 $m/z$ with an elemental composition of $C_{53}H_{83}N_{14}O_{16}S_1$ .....	68
<b>Figure 5.4</b> Selected mass spectral regions (black) compared to calculated isotopic profiles (red) for BSA peptides.....	69
<b>Figure 5.5</b> The logarithmic heat map of a BSA sample (A) with an example of a correlation spot for a pair of differently labelled peptides (B) and the mass spectrum of a peptide pair (C) indicated in (B).....	71
<b>Figure 5.6</b> Two selected BSA correlation spots (A & E) with the corresponding mass spectra (B, C & F) at the indicated locations “w” and “x” on (A) and “y” and “z” on (E). (D) shows a region of interfering correlation spots.....	72
<b>Figure 5.7</b> Visual representation of the Parallel Isotopic Tag Screening (PITS) algorithm.....	74
<b>Figure 5.8</b> A two-dimensional representation of the mass chromatogram for isotopically labelled yeast (A), one particular mass spectrum (B) indicated on (A) and an expanded mass spectrum (C) of section $i$ showing $r_{j1}$ and $r_{j2}$ respectively for each combination.....	76
<b>Figure 5.9</b> The observed mass spectral response (black) for a particular peptide pair compared to the predicted mass spectral response (red) calculated for both $k$ segments with the respective $f_{j1}$ and $f_{j2}$ .....	78
<b>Figure 5.10</b> The overall calculation procedure for the PITS algorithm at one particular mass spectral window $i$ for both $k$ segments.....	81
<b>Figure 5.11</b> The isotopic pattern for a doubly charge yeast peptide with two labels at 400 $m/z$ (A) and 1700 $m/z$ (B).....	82
<b>Figure 5.12</b> A selected mass spectrum (D) from the yeast mass chromatogram with three particular peptide pairs highlighted.....	84
<b>Figure 5.13</b> A two-dimensional representation of the classification matrix for the yeast mass chromatogram (A) and (B) is a expanded region of the classifications indicated on (A).....	85

<b>Figure 5.14</b> A two dimensional representation of the modified classification matrix for the yeast mass chromatogram (A) and (B) is a expanded region of the modified classifications indicated on (A).....	86
<b>Figure 5.15</b> A two dimensional representation of the classification matrix for the four replicate BSA experiments.....	89
<b>Figure 5.16</b> An example of a peptide pair close to the baseline where the isotopic pattern was not clear. The gray dots correspond to the mass of the light and heavy components for the peptide indicated on the top of the plot.....	91
<b>Figure 5.17</b> The observed mass spectral response (black) and the predicted mass spectral response (red) calculated for two peptide pairs not found by PITS algorithm.....	93

## **Abstract**

In quantitative proteomics, many of the LC-MS based approaches employ stable isotopic labelling to provide relative quantitation of the proteome in different cell states. In a typical approach, peptides are first detected and identified by tandem MS scans prior to quantifying proteins. This provides the researcher with a large amount of data that are not useful for quantitation. It is desirable to improve the throughput of current approaches to make proteomics a more routine experiment with an enhanced capacity to detect differentially expressed proteins. This thesis reports the developments towards this goal, including an assessment of the viability of stable dimethyl labelling for comparative proteomic measurements and the evaluation of a dynamic algorithm called Parallel Isotopic Tag Screening (PITS) for the detection of isotopically labelled peptides for quantitative proteomics without the use of tandem MS scans.

## List of Abbreviations Used

$\Delta$ CN	Delta correlation value
2D-PAGE	Two dimensional polyacrylamide gel electrophoresis
ABC	Ammonium bicarbonate
AQUA	Absolute quantitation of proteins
BSA	Bovine serum albumin
cICAT	Cleavable isotope-coded affinity tags
CID	Collision induced dissociation
Da	Dalton
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
IAA	Iodoacetamide
ICAT	Isotope-coded affinity tags
iTRAQ	Isobaric tags for relative and absolute quantitation
LC	Liquid chromatography
LC-MS	Liquid chromatography-mass spectrometry
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
<i>m/z</i>	Mass-to-charge
MRM	Multiple reaction monitoring
mRNA	Messenger RNA
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
mzXML	Extensible markup language
NHS	N-hydroxysuccinimide
PAI	Protein abundance index
PITS	Parallel isotopic tag screening
PTM	Post-translational modification
RNA	Ribonucleic acid
RPLC	Reversed phase liquid chromatography
$R_{Sp}$	Ranked preliminary score
SCX	Strong cation exchange
SDS	Sodium dodecyl sulfate
SILAC	Stable isotope labelling by amino acid in cell culture
SPC	Seattle Proteome Center
TEAB	Triethylammonium bicarbonate
TFA	Trifluoroacetic acid
$X_{corr}$	Cross-correlation value
XIC	Extracted ion chromatogram

## Acknowledgements

Last but not least, I need to write the acknowledgements for my thesis. I have been at Dalhousie for six and a half years while completing my BSc and MSc and there are a number of people I would really like to thank.

First, my family, the people who have made me the person I am today. I am so grateful to my amazing mother and father who have always been there for me. Always providing support when I needed it, never questioning me about my plans for the future. I love them both very much and I am very proud to be their son. Thank you for financially supporting me over the years. To my nanny, the most wonderful woman I know, thank you for never letting me get into debt. Thank you for our weekly chats we have while I am walking to work, they make my day. Family is one of the most important things to me and I am so grateful to have them in my life.

Second, Club Wentzell, over the years there has been multiple people who have helped me out in some type of way. Thanks to Robert Flight for taking me as his student in Summer 2007, it meant a lot to me. Thanks to Siyuan Hou, Rukhsi Jabeen, Hunter Warden, and Bjorn Wielens for their help with MatLab programming advice, helping with my Seminar, discussions, and ultimately providing a great working environment over the years.

Third, Peter Wentzell, thanks for a great six and a half years under your supervision. Thank you for introducing me to chemometrics, it has changed how I view chemistry and has made me be very excited about research in the future. You have provided me with the knowledge and experience I need to be successful in the “real world”.

Finally, Emma, the person who makes me laugh, smile and always puts me into a better mood. Thank you for your encouragement, support, hugs and consistently telling me everything is going to be OK. As new adventures and experiences start to unfold there is no other person I want by my side enjoying them with me. I love you with all my heart.

# Chapter 1

## Introduction

### 1.1 Overview

Over the past decade, research in proteomics has emerged as a major field of study in analytical chemistry, driven by the demands of systems biology and health research. Proteomics has the potential to provide a massive amount of biological information because proteins are involved in virtually all biological activities in various capacities, and elucidating these roles can add to the understanding of biological systems. Presently, a number of international research journals are published exclusively to cover proteomic studies, including *Proteomics*, *Molecular Cellular Proteomics*, and *Journal of Proteomic Research*. Mass spectrometry (MS) has become a crucial component in proteomics because of its speed, selectivity, and sensitivity in the detection and identification of proteins [1-2]. Given the complexity of the proteome, a considerable amount of sample preparation, including sample purification, digestion, and several stages of separation, is typically required before proteins can be indentified through MS analysis. Unfortunately, despite its potential, practical results from proteomics have been slow to emerge due to the complexity of the biological systems under study [3]. There is a need to improve the current proteomic tools and strategies. A major motivation for this is the research being done in the field of protein biomarker discovery [3-4]. The need is apparent when taking into account the throughput of modern proteomic experiments, shorter analysis time compared to existing strategies is essential to making proteomics a routine experiment, especially in the situation of protein biomarker discovery.

This introductory chapter will focus on the objectives and challenges of quantitative proteomics in the context of understanding biological systems. An overview of current proteomic methods is presented, with a particular attention to MS-based methods.

## **1.2 Proteomics**

Proteomics is the study of the proteome [5], introduced by Anderson and Anderson in 1998. The term ‘proteome’ comes from PROTEins being expressed by the geOME and was coined by Marc Wilkins in 1994 [6-7]. Wilkins defined proteomics as “the study of proteins, how they’re modified, when and where they’re expressed, how they’re involved in metabolic pathways and how they interact with one another”. Proteomics has expanded very quickly. The first publication related to proteomics appeared in 1995 [8], and in 2011 there were approximately 3,600 publications related to proteomics. The proteome contains a broad variety of unique proteins ranging in physical and chemical properties, present over a wide concentration range [9]. A typical proteome can contain millions of unique proteins. For example, in humans, taking into account a large number of possible post-translational modifications (PTMs) [10], the number of unique proteins has been estimated to be in the millions [11].

In 2001, the human genome project was completed and sequenced approximately 39,000 genes [12]. The principle objective of human functional genomics is now to assign function to all of these genes. Gene function is derived from the protein product it encodes. Proteins define the functional output of the cell, and therefore would be able to provide relevant information, particularly when considering that their presence takes into

account any specific biological/environmental factors in the cell (i.e. stress, drug administration, disease). A strong gene expression, resulting in abundant mRNA transcribed from the DNA, does not necessarily mean that the respective protein translated from the mRNA is also abundant or indeed active in the cell [13]. Substantial modifications can also be introduced during or after translation, (i.e. glycosylation, phosphorylation) leading to several protein products from a single gene. For example, this number is typically one or two in bacteria, two or three in yeast, and three to six in humans [14]. Ultimately studying the proteins rather than the genes is more promising to achieve pertinent information.

### **1.3 Objectives of Proteomics**

The ultimate objective of proteomics is the rapid identification and quantitation of all of the proteins expressed by a cell or tissue, an objective that has yet to be achieved for any species. This is due to the complexity of the proteome, although improvements in instrumentation (e.g. separation techniques and mass spectrometry) allow for more proteins to be identified by mass spectrometry each year [15-16]. There are many diverse areas of proteomics research, some of these focus on understanding the different properties of proteins, protein structure [17-19], protein-protein interaction [20-22], and protein modification [10, 23, 24], in many cases however it is necessary to quantify the amount of proteins present in different cell states and this is the focus of this research. One major goal of studying these properties is to provide appropriate information to medical doctors about human diseases (e.g. various types of cancer), with a purpose of developing refined protocols for disease diagnosis and treatment.



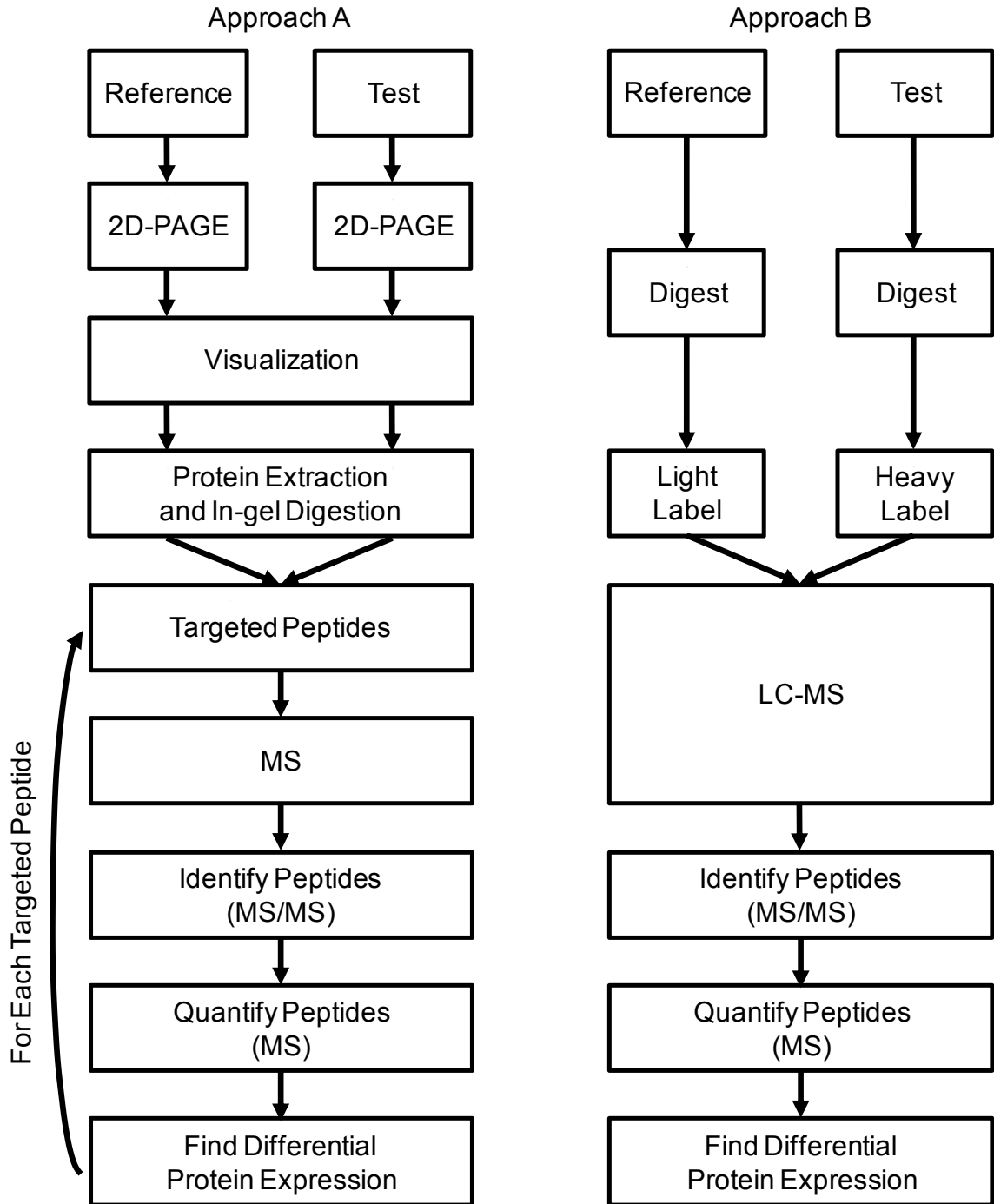
Many different technologies have been and are still being developed to collect the information about the properties of proteins. Two characteristics of these proteomic technologies are immediately apparent: first, there is no single technology platform that can satisfy all of the desired proteomic measurements, and second, there is no mature, ‘true’ proteomic technology as yet. The primary focus of this thesis is quantitative proteomics and the ability to identify protein biomarkers. There are a number of prominent MS-based proteomic approaches widely used but in this thesis only two will be focused on, one involving two dimensional polyacrylamide gel electrophoresis (2D-PAGE) [25-27] and the other using reversed phase liquid chromatography (RPLC) [28-29]. Section 1.4 will discuss these two approaches.

#### **1.4 MS-based Proteomics**

In the context of proteomic analysis, the approach can be described in terms of the experimental techniques and instruments used to extract information from the proteomic experiment. In quantitative proteomics, the proteome of two cell states under different conditions are analyzed. In general, these states can be referred to as the reference (e.g. healthy) and test (e.g. diseased) states of the cells of an organism. The objective of the experiment is to quantify the relative change in protein expression in the two different cell states. The two MS-based approaches commonly used in quantitative proteomics to achieve this objective involve similar experimental techniques and instruments at different positions in the experiment. In this work, Approach A is the 2D-PAGE based approach and Approach B is the LC-MS based approach. An enzymatic digestion is completed in both approaches prior to (Approach B) or after (Approach A) a separation of proteins (Approach A) or peptides (Approach B). Comparison of peptides or proteins

is done via staining of gels (Approach A) or isotopic labelling techniques (Approach B).

The final step in both approaches is to use mass spectrometry for quantitation and



**Figure 1.1** Visual representation of the two common MS-based quantitative proteomics approaches.

identification of proteins. Figure 1.1 shows a visual representation of the two common quantitative proteomics approaches.

Approach A employs 2D-PAGE to separate proteins from the two different cell states (reference versus test). Proteins are separated on the basis of charge in the first dimension and molecular mass in the second. The separation of charge is done through isoelectric focusing [25-27]. Isoelectric focusing allows proteins to be separated and focused on an immobilized pH gradient gel strip by their isoelectric point. The separation by molecular mass in the second dimension is achieved by attaching negative sodium dodecyl sulfate (SDS) molecules to proteins which allows them to migrate dependent on molecular mass [30-31]. The SDS molecules are negatively charged, resulting in proteins having approximately the same mass-to-charge ( $m/z$ ) ratios [32]. This is because the length of an unfolded protein is approximately proportional to its molecular weight. The completed gels for both cell states are then stained (e.g. Coomassie staining [33]), and visually compared. Proteins which are only present in one of the gels are cut out and individually digested in-gel to be analyzed by the mass spectrometer. In this work, the proteins which are cut out of the gels are called the targeted proteins.

In Approach B, proteins from the two cell states are enzymatically digested prior to separation by RPLC. To distinguish the resulting peptides from each state, peptides are isotopically labeled. In this technique, peptides from different cell states are labeled using compounds with nearly identical chemical properties but different stable isotope compositions, resulting in different masses that can be detected by the mass spectrometer. A detailed discussion of the most common isotopic labelling techniques is presented in Chapter 2. Once the labelling is complete, the labelled samples from the two cell states

are combined into one sample. The peptides in the combined sample are separated by liquid chromatography and analyzed by the mass spectrometer (LC-MS).

In the framework of quantitative proteomics, the mass spectral analysis for both approaches is identical, except that in Approach A analysis is done separately on each protein cut out of a gel and in Approach B analysis is done on the one combined labeled sample (Figure 1.1). Only those peptides identified to be associated with a protein are quantified. Identification is done by peptide sequencing using tandem mass spectrometry (MS/MS) [34]. Peptide ions detected in the mass spectrometer are collected and then fragmented, normally by collision induced dissociation (CID). The amino acid sequence of a peptide can be determined by using a search engine (e.g. SEQUEST [35]) which matches experimentally collected fragmentation spectra to computationally generated fragmentation spectra of potential peptide matches from proteomics databases (e.g. Swiss-Prot [36]), created using genomic information. The identified peptides are then quantified by comparing relative ion abundances. For Approach B, the quantitative analysis is typically done by determining a ratio of the ion abundances in the two cell states at a single time point. A more in-depth discussion on how quantitation is done is presented in Chapter 2. The goal of quantifying all of the identified peptides is to find peptides (corresponding to a protein) that have a differential expression level in the two cell states of interest. This quantitative information provides insight on finding potential biomarkers, discussed in the next section.

In recent years, quantitative proteomics has evolved from the 2D-PAGE approach to the LC-MS based approaches. Several of the main reasons are: (1) 2D-PAGE is subject to bias in the selection of targeted proteins on the gel, where the detection of the

proteins largely relies on relatively insensitive staining methods; (2) 2D-PAGE followed by in-gel digestion and mass spectral analysis of individual proteins is labour-intensive and difficult to automate, resulting in low throughput; and (3) the LC approach is more comprehensive, allowing more proteins from both cell states to be analyzed by the mass spectrometer, providing as much mass spectral information as possible.

## **1.5 Biomarker Discovery**

As previously stated, a major goal of proteomics is the discovery of protein biomarkers that are characteristic of particular human diseases for early diagnosis and improved treatment strategies. A biomarker, in this work, is a protein that can be used as an indicator of a particular disease state of an organism. One of the first protein biomarkers used in disease diagnosis was the prostate-specific antigen for the detection of prostate cancer [37]. Unfortunately many single protein biomarkers have proven to be unreliable and multiple biomarkers are maybe needed [38].

The biomarker discovery workflow has two stages, identification of candidate biomarkers and validation of those biomarkers. In identifying candidate biomarkers, the goal is to generate a list of prospective candidate protein biomarkers by profiling the proteome of the different cell states. The candidate biomarkers are acquired by carrying out an exhaustive profiling of the healthy and diseased proteomes for a relatively small number of samples, probing for differences in protein expression that may be suggestive of the disease state [3]. Once the list of candidate biomarkers is compiled, the validation stage begins. The candidate biomarkers are subject to a detailed and comprehensive analysis over a broader population and possibly over a number of different conditions

such as age, gender and disease stage. This allows the candidate biomarkers' sensitivity and selectivity to be assessed in different variations of the population. This stage uses targeted mass spectral strategies that only focus on the candidate biomarkers, ignoring the rest of the proteome. Typically multiple reaction monitoring (MRM) is used to target particular peptide fragments from the targeted protein to quantify candidate biomarkers in a large number of samples [39-40]. Statistical analysis is then performed on the quantitative data to determine if any of the candidate biomarkers are reliable and can be used for early diagnosis [41].

Currently, the main challenge of biomarker discovery is the identifying candidate biomarkers stage [9, 42]. This process is extremely slow, due to the number of separations required per sample and the time spent identifying peptides in the mass spectrometer. This limited throughput constrains the number of samples that can be analyzed in a given time period, consequently reducing the reliability of candidates selected by this procedure. Presently, a number of efficient protocols for protein separation prior to mass spectral analysis can be found in literature [15-16]. However, as separation methods become more efficient there is a need to develop methods that can provide high-throughput mass spectral proteome analysis of complex samples and decrease the duty cycle of the mass spectrometer. The development of such a method is the focus of the work presented in this thesis.

## **1.6 Motivation For Thesis Research**

Much of the analytical research effort in proteomics has been focused on expanding proteome coverage, with an emphasis on identifying the largest number of

proteins in a particular proteome and to a lesser extent, quantifying proteins expressed in biological systems. However, unlike transcriptomics and metabolomics, the availability of clinically useful proteomics data is much more limited because of the qualitative nature of most data and the lack of studies with a sufficient number of samples. A prerequisite of clinical data is that it be drawn from a sufficiently large sample size to ensure reliable inference within a highly variable population. The number of clinical samples that can be analyzed is currently limited by the low throughput of proteomics procedures. It is also important that such data quantify the relative abundance of proteins rather than simply reporting their detection. While it is not critical that such methods provide comprehensive coverage of the proteome to be useful, it is of course desirable that as many proteins as possible are detected. Finally it is important to note the *identification* of peptides is not necessary for the initial screening of potential biomarkers, but can be done after such candidates have been selected.

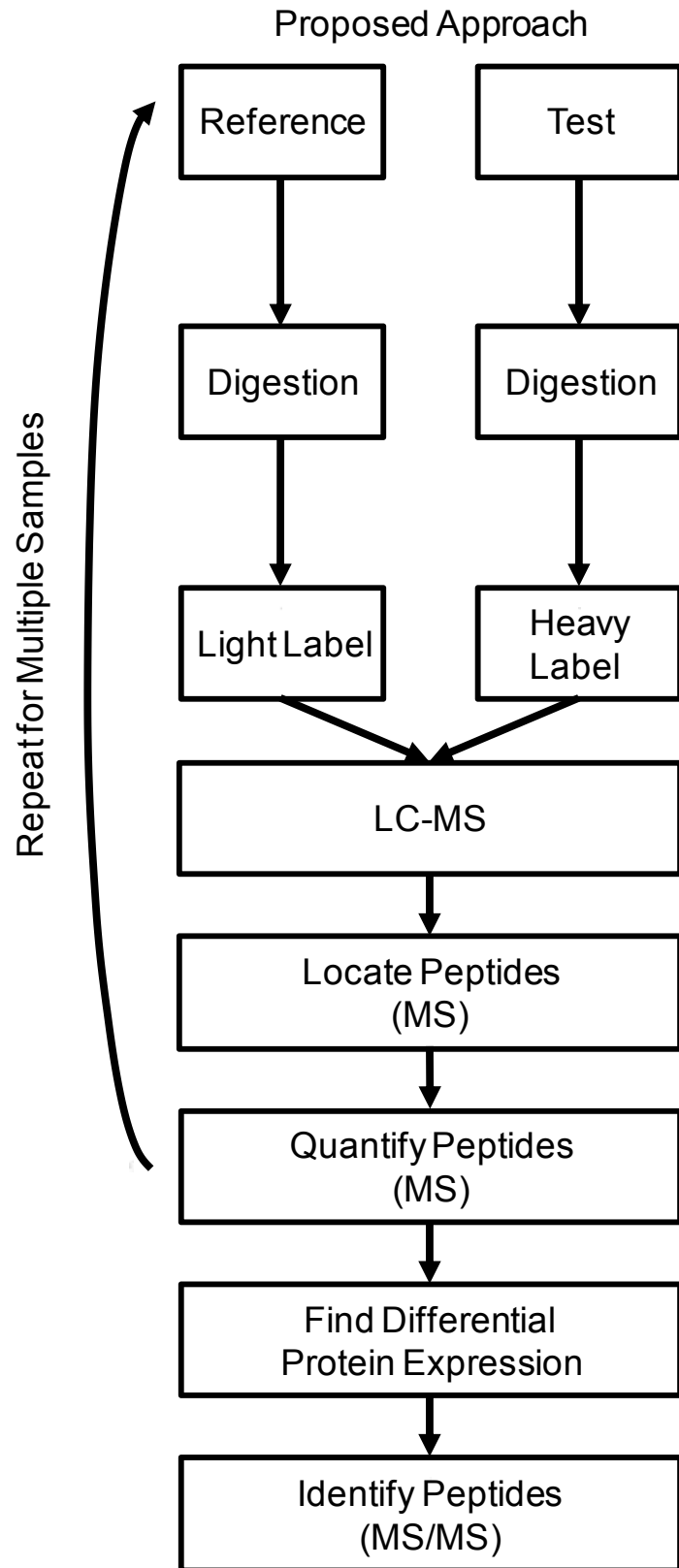
The majority of quantitative LC-MS approaches employ stable isotope labeling to create exact mass tags that can be detected by the mass spectrometer to provide relative quantitation of two (or more) proteins samples. However, to quantify proteins, the peptides must first be identified by tandem MS scans, which increases the analysis time, decreases the duty cycle of the mass spectrometer and provides a large amount of data that is not useful for quantitation. These restrictions have limited the utility of conventional proteomics for routine biological applications, since the analysis of a simple fractionated sample can take several days. In traditional approaches (Figure 1.1), the mass spectrometer compiles data from three different mass scans: (1) a full mass scan used for quantitation, (2) a high resolution zoom scan to provide accurate masses used for

the tandem mass scan and (3) a tandem mass scan used for identification of peptides. In the quantitative experiments presented in this thesis, 49% of the scans are tandem mass scans and only 17% are full mass scans. The goal of quantitative proteomics is to quantify the peptides but only 17% of the scans from the experiment are actually used for quantitation. If peptides could be located and quantified without the need of tandem MS scans or zoom scans, a number of improvements could be made: (1) the number of MS scans could be increased, (2) the chromatographic separation time could be reduced, (3) throughput could be improved, (4) data quality could be improved and (5) more peptides could be detected. The focus of this thesis is the development of a method that can locate and quantify peptides from quantitative proteomic experiments without the use of tandem MS scans. The proposed approach is shown in Figure 1.2. The experimental techniques are identical to Approach B in Figure 1.1 but the mass spectrometer is used differently. In the proposed approach, only MS scans are performed at first. The peptides are found and quantified using an algorithm based on the data obtained from the MS scans. Only those peptides that have differences in protein expression are then identified using MRM scans.

## **1.7 Outline of Thesis**

Since the principal focus of this work is quantitative methods of proteomics, Chapter 2 reviews the history and current practice of labelling and quantitation. Different approaches to tagging peptides are described with an emphasis on the chemically based methods. Chapter 3 describes the experimental aspects of the studies carried out in this work, which included analysis of bovine serum albumin (BSA) and the yeast proteome. Details of the preparation, digestion, labeling and MS analysis are presented.





**Figure 1.2** Visual representation of the proposed approach for the research in this thesis.

Additionally, to analyze the mass spectral data, it was necessary to import it into MatLab<sup>®</sup> and the process for doing this is described in this chapter.

The initial goal of this thesis was to develop an algorithm to locate and quantify peptides indicative for biomarker discovery. Unfortunately, the quantitation aspect was not completed because preliminary research showed that there could be a problem with the stable-isotope dimethyl labelling technique, the labelling technique chosen for the research in this thesis. The reasons for this are described in Chapter 2. As previously stated, protein quantitation is achieved by comparing relative ion abundances. However, due to the composition of the stable isotopes there is a chance that differentially labeled peptides could separate in the LC analysis. This poses complications for quantitative analysis which typically assumes coeluting analytes and determines the ratio at a single time point. A number of studies have investigated the extent of coelution in several systems, with mixed conclusions. In Chapter 4, a comprehensive study was carried out to investigate the effect of dimethyl labelling with H<sub>2</sub>/D<sub>2</sub> formaldehyde on the retention characteristics of differentially labeled peptides. It is demonstrated here that the widely used dimethyl labelling technique does not adversely affect the quantitation in single point methods because the isotopic effect is inversely related to the polarity of the label.

Once it was determined that deuterium based dimethylation should not be discounted as an isotopic labelling strategy, the development of an algorithm to locate peptides was undertaken. Since the approaches described in this thesis make no assumption of the availability MS/MS data, the detection of peptides has to be based solely on MS data, which necessitates the exploitation of unique characteristics to distinguish peptides from other species that may be present. The features used in this

work are the isotopic patterns arising in the mass spectral domain from isotopically labelled peptide pairs and a sustained chromatographic presence. Owing to the nature of the labelling and the charges that exist on the peptides, there are very distinctive patterns observed for the different cell states of interest. In Chapter 5, a detailed discussion on the development of an algorithm to locate peptides produced from quantitative proteomics is presented. Finally, describing the successful algorithm called Parallel Isotopic Tag Screening (PITS) and the performance of the PITS algorithm compared to traditional methods for detecting peptides. It was found that the new algorithm located more peptides overall than conventional database searches.

## **Chapter 2**

### **Isotopic Labelling Techniques**

As stated in Chapter 1, most quantitative proteomic methods employ stable isotope labelling to create an internal standard in the form of an exact mass tag that has the ability to be distinguished by the mass spectrometer and simultaneously provide information for quantitation. Mass tags can be introduced into proteins or peptides chemically [43], metabolically [44], enzymatically [45] or by spiked synthetic peptide standards [46]. In contrast, protein quantitation can also be performed using label-free quantitation techniques [47]. All of these methods have been documented in the literature as achieving significant results and are in use today. Each method has its own advantages and disadvantages and the use of each method is dependent on the preferences of the researcher, instrument capability, cost and the nature of investigation. This chapter focuses on stable isotope labelling methods that are chemically introduced to proteins, but will also briefly cover biological introduced stable isotope methods and label-free methods.

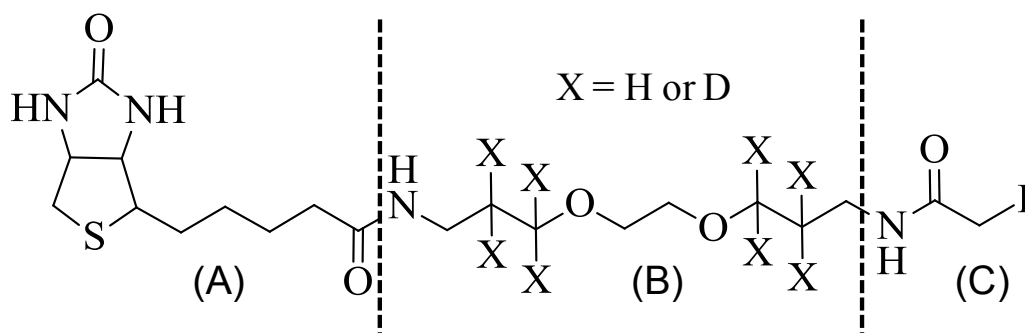
#### **2.1 Chemically Based Methods**

Some amino acids have side chains with reactive functional groups that can be chemically modified to integrate an isotope-coded mass tag. In practice, side chains of lysine and cysteine are primarily used for this purpose. This section focuses on the most common chemically introduced stable isotope mass tags that are in use today, with a particular emphasis on isotope-coded affinity tags (ICAT) [43], isobaric tags for relative and absolute quantitation (iTRAQ) [48] and stable-isotope dimethyl labelling [49]. These

tags are all designed to create internal standards that are isotopically labelled versions of the molecule that is to be quantified. This is a key characteristic because isotopically labelled internal standards will have similar extraction recovery, ionization response in mass spectrometry, and chromatographic retention time. Given that the mass spectrometer can distinguish the mass difference between the light- and heavy-labelled forms of peptides, quantitation is attained by comparing their respective signal intensities.

### 2.1.1 Isotope-Coded Affinity Tags (ICATs)

The ICAT method was first described by Gygi et al. in 1999 [43] and is the first in a series of similar commercial cysteine synthetic tags for protein quantitation. Referring to Figure 2.1, the ICAT is a cysteine specific tag that consists of three components: (1) a biotin moiety (A) that allows for specific isolation of labelled peptides by means of biotine-avidin affinity chromatography, (2) a linker (B) incorporating stable isotopes that has two forms, the light form containing eight hydrogen atoms and the heavy form containing eight deuterium atoms, and (3) a cysteine-reactive iodoacetamide group (C) that allows the specific attachment of the label to the thiol group of the cysteine side chain. This label provides the ability to do comparative analysis between two different cell states.



**Figure 2.1** Structure of the original ICAT reagent.

The procedure for the ICAT method consists of the following steps. First, the side chains of cysteinyl residues in a protein sample from one cell state are labelled with the isotopically light form of the ICAT tag ( $X = H$ ) and cysteinyl residues from the second cell state are labelled with the heavy form of the ICAT tag ( $X = D$ ). As described by Gygi (1999), this is done by breaking the disulfide bonds in the denatured protein samples from each cell state using a reducing agent and then adding the ICAT reagents (light or heavy). Second, the two samples are combined and digested with trypsin to generate peptide fragments, some of which are tagged. Third, the combined sample of peptides is passed through a monomeric avidin column to remove unlabelled peptides. Finally, the isolated cysteine-containing peptides are separated and analyzed by LC-MS/MS. The peptides are quantified by measuring the relative signal intensities in the MS mode for pairs of peptide ions that differ in mass corresponding to the labelling of the isotopically light or heavy forms of ICAT reagent. The labelling produces a mass difference of 8 Da for every cysteine in a peptide. The detected mass difference changes as the number of tags and charge state vary. The proteins are then identified through peptide sequencing in the fragmentation spectra.

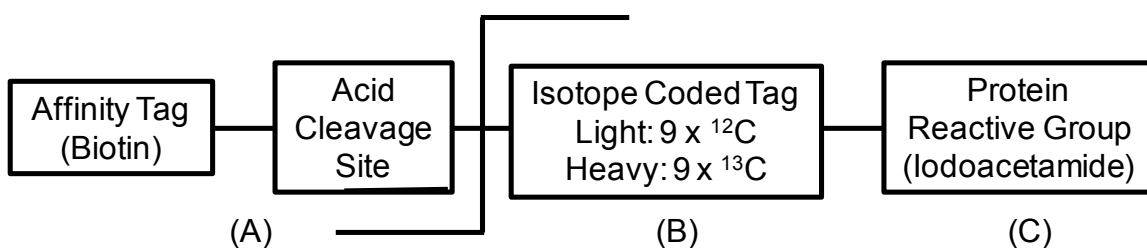
A number of applications of ICAT have been documented in the literature [50-53]. Gygi et al. [50] also showed that they could detect and quantify proteins of low abundance in complex mixtures. This study combined the use of ICAT reagents and three-dimensional chromatography (cation exchange, biotin affinity, and reversed-phase) of the peptides generated by enzymatic digestion of the tagged proteins in the yeast proteome. A number of groups were able to use ICAT reagents and LC-MS/MS for the detection and quantitation of differentially expressed proteins from two different cell

states for example in *pseudomonas aeruginosa* [51], human myeloid leukemia (HL-60) cells [52], and LNCaP human prostate cancer cells [53], with improved results over previous reports in literature.

A number of limitations of ICAT have been reported in literature. First, LC separation can occur between the light and heavy forms of the ICAT labelled peptides [54, 55]. Extreme cases show that baseline separation of the two forms of the ICAT reagents can occur. This separation considerably affects the accuracy of the quantitation because point measurements of signal intensity ratios are not reliable. Second, the ICAT tag is quite large and, as a result, the labelled peptide produces many fragments in the fragmentation spectrum related to the tag rather than the peptide, making identification more difficult [56]. Third, because cysteine is a rare amino acid and its relative abundance is quite low [57], only one or two peptides are typically labelled and quantified per digested protein. In some cases, proteins contain no cysteine, which would prevent any quantitative information from being obtained. Finally, ICAT is unable to identify post-translationally modified peptides since the majority of the PTMs are removed during affinity chromatography. Despite these limitations, ICAT can be a useful method for broad (bodily fluid) or targeted (cysteine containing protein) analyses. To overcome the third and fourth limitations, another amino acid specific labelling method is required and will be discussed in Section 2.1.2. The first and second limitations have been overcome by an enhanced version of the original ICAT label described by Hansen et al. in 2003 [58] and referred to as the cleavable isotope-coded affinity tag (cICAT). This version employs isotope labels that use  $^{12}\text{C}$  and  $^{13}\text{C}$  as the isotope pair (light and heavy), since these substitutions do not lead to chromatographic

separation of differentially labelled peptides as is the case for the D<sub>0</sub> / D<sub>8</sub> ICAT reagents [54, 55].

The cICAT method has a mass tag similar to the original but with some modifications as shown in Figure 2.2. The structure contains a biotin affinity tag group, similar to the original ICAT reagent, connected to a site that can be cleaved under acidic conditions (A). This is then connected to a linker group that is labelled with nine <sup>12</sup>C's or nine <sup>13</sup>C's (B) and which is connected to a cysteine-reactive iodoacetamide group the same as in the original ICAT (C). For proprietary reasons, the structure of the acid cleavage site of the cICAT label has not been described. The benefits of cICAT structure are: (1) the acid-cleavable site can be removed, resulting in a smaller moiety being attached to the peptide (indicated by the vertical solid line in Figure 2.2) and improving the quality of the fragmentation spectra, and (2) cICAT employs <sup>12</sup>C and <sup>13</sup>C as its light and heavy labels rather than <sup>1</sup>H and <sup>2</sup>H, thereby allowing the peptides to co-elute by chromatography, resulting in more reliable quantitation.



**Figure 2.2** Structure of the cICAT reagent.

The procedure for the cICAT method is similar to the one for the original ICAT tag with a few differences. After avidin affinity chromatography, the acid-cleavable site is then removed from the labelled peptide under acidic conditions. Again, for proprietary reasons, the chemistry associated with the removal of the cleavable site is unknown. The



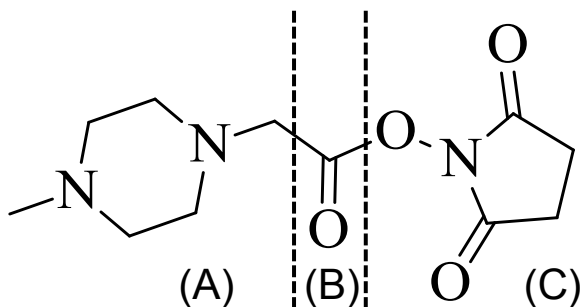
acid cleaved labelled peptides are separated and analyzed by LC-MS/MS in the same way as in the original ICAT method, except that the mass difference is 9 Da's for every cysteine in a peptide. A number of examples showing successful application of the cICAT method can be found in literature [59-62].

It should also be noted that there are a number of other cysteine-specific tagging methods [63-67]. These include visible isotope-coded affinity tags [63, 64], solid-phase ICAT [65], acid-labile isotope-coded extractions [66], and element-coded affinity tags [67], to name a few. All of these methods attempt to improve on some of the limitations of the original ICAT, but are based on the same principles and will not be discussed in detail here. However, due to its advantages over the original ICAT, cICAT is the most widely used cysteine-specific isotopic labelling method to date.

### **2.1.2 Isobaric Tags for Relative and Absolute Quantitation (iTRAQ)**

First described by Ross in 2004 [48], the iTRAQ method employs a multiplexed set of reagents for quantitative protein analysis that consists of a set of isobaric (same weight) mass labels that attach to the N termini and lysine side chains of peptides in a digested mixture. With this approach, the protein from up to four sources (as opposed to two) can be distinguished in a single mixture. All four samples are labelled with different combinations of isotopes of carbon, nitrogen, and oxygen such that all labelled peptides are isobaric, but have characteristic 'reporter ions' following CID that can be used to identify and quantify differentially expressed proteins in the four cell states (samples). Shown in Figure 2.3, the iTRAQ reagent consists of three components: (1) a reporter group (structure of N-methylpiperazine) (A), (2) a mass balance group (carbonyl) (B) and

(3) an amine specific peptide reactive group (N-hydroxysuccinimide (NHS) ester) (C). The overall mass of the reporter (A) and balance (B) groups of the molecule remain constant using differential isotopes ( $^{13}\text{C}$ ,  $^{15}\text{N}$ ,  $^{18}\text{O}$  indicated in Table 2.1), and the use of these isotopes removes any chromatographic separation that can be observed with hydrogen and deuterium isotopic labelling. As illustrated in Table 2.1, the reporter group has a mass range of 114 to 117 Da and the balance group ranges from 28 to 31 Da, giving combined mass of 145 Da for each of the four reagents. iTRAQ provides the ability to do comparative analysis among four different cell states.



**Figure 2.3** Structure of the iTRAQ reagent.

**Table 2.1** Combination of isotopes for iTRAQ reagents to maintain a combined molecular weight of 145.

A (MW = 113)			B (MW = 28)		
Da	Carbon and Nitrogen isotopes		Da	Carbon and Oxygen isotopes	
114	$^{13}\text{C}$		31	$^{13}\text{C}$	$^{18}\text{O}$
115	$^{13}\text{C}_2$		30		$^{18}\text{O}$
116	$^{13}\text{C}_2$	$^{15}\text{N}$	29	$^{13}\text{C}$	
117	$^{13}\text{C}_3$	$^{15}\text{N}$	28		

To carry out the iTRAQ method, each protein sample of a different cell state is reduced, alkylated and then digested with trypsin. Equal aliquots of the isotopic reagents are

added to each cell state, labelling each peptide in a particular cell state with a specific reporter group and balance group combination (Table 2.1). The NHS ester in the iTRAQ reagent reacts with the N-terminus or lysine to form an amide linkage to the peptide. The four different peptide samples are then combined into one sample, which is separated and analyzed by LC-MS/MS. In the MS mode, a single unresolved precursor ion is detected for each peptide. Each MS ion detected represents an identical peptide labelled with the four iTRAQ reagents. Following CID, the four reporter group ions appear in the fragmentation spectra as distinct masses of 114.1, 115.1, 116.1 and 117.1 Da. The sequence information is also determined from the same fragmentation spectrum. This process is carried out for each precursor ion detected, where the relative concentrations of the peptides are obtained from the relative intensities of the corresponding reporter ions at 114.1 to 117.1 Da.

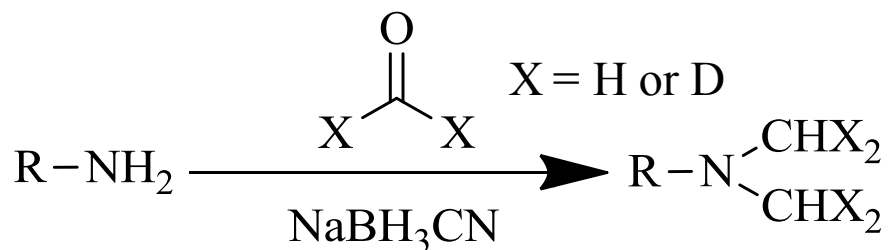
A number of applications for iTRAQ have been documented in literature [68-70]. Zieske [68] showed that the iTRAQ method was successful for a number of biological applications in drug induced-protein expression, discovery and elucidation of disease markers, and protein-protein interactions. This was all done in multi-protein complexes and no protein sequence information was lost from samples involving PTM. Keshamouni et al. [69] were able to identify differentially expressed proteins during epithelial-mesenchymal transition of epithelial cells in both normal embryonic development and certain pathological contexts. Glückmann et al. [70] used iTRAQ methods to prevalidate potential biomarkers in early molecular process of hepatocarcinogenicity, which is the production of cancer cells in the liver. They obtained results similar to their previous work, thus validating the potential biomarkers for the

system of interest. The iTRAQ method also detected potential biomarkers that were not observed in Glückmann's previous work, showing that iTRAQ is a very useful method for biomarker prevalidation.

The most significant drawback of the iTRAQ method is that fragmentation spectra must be acquired for quantitation of the proteins in the complex mixture. This means more analysis time is required than for other isotopic methods like ICAT and cICAT, which only need MS scans for quantitation. The authors of the original iTRAQ paper comment on this drawback but believe that the "ability to identify more proteins with increased confidence and greater peptide coverage outweighs this disadvantage" [48]. Whether to use iTRAQ or another technique that doesn't require tandem MS is one of the major debates in quantitative proteomics. There are three other drawbacks to the iTRAQ method. First, iTRAQ is not ideally suited for some mass spectrometers, since the reporter group ions appear in the range of 114-117 Da; Q-traps have a low mass cut-off during tandem MS. However, pulsed-Q dissociation (PQD) [71, 72], was introduced to overcome this drawback. PQD fragmentation mode allows for low  $m/z$  ions to be detected in linear ion traps. A second drawback is that higher collision energies are required to efficiently liberate the reporter ions from lysine residues, which results in a loss of sequence information [73]. This drawback of the fragmentation spectra cannot be overcome. Third, using  $^{13}\text{C}$ ,  $^{15}\text{N}$ , and  $^{18}\text{O}$  in the iTRAQ reagents makes the method quite expensive. However, the last three drawbacks can be overcome by using another amine-specific isotopic labelling method like the one described in the next section.

### 2.1.3 Stable-Isotope Dimethyl Labelling

In 2003, Hsu et al. [49] developed stable-isotope dimethyl labelling method using the well-known reductive amination mechanism. This method is similar to iTRAQ with respect to the labelling location of the tag (N-terminus and lysine), but it is different from iTRAQ in that much simpler tags are employed and isotopic labelling is performed with formaldehyde ( $\text{CH}_2\text{O}$ ) and deuterated formaldehyde ( $\text{CD}_2\text{O}$ ). The procedure for this method is quite simple compared to the other methods. The proteins in the two different cell states are denatured, reduced, alkylated and then digested with trypsin. The peptides from each cell state are separately labelled via reductive amination (shown in Figure 2.4) with either  $\text{CH}_2\text{O}$  or  $\text{CD}_2\text{O}$ . After the addition of formaldehyde, sodium cyanoborohydride is added to each mixture to complete the reductive amination reaction. After the reactions are complete, the labelled samples are mixed and analyzed simultaneously by LC-MS. The mass difference of the dimethyl labels is used to compare the peptide abundance in the different samples. In dimethyl labelling, all primary amines (the N terminus and the side chain of lysine residues) in the peptide mixture are converted (through reaction with  $\text{CH}_2\text{O}$  or  $\text{CD}_2\text{O}$ ) into dimethylamines which serve as the labels. The labelling produces a mass difference of 4 Da per labelling site for every tryptic peptide. The detected mass difference changes as the number of tags and charge state vary. The proteins are then identified through peptide sequencing in the fragmentation spectra.



**Figure 2.4** Reductive amination reaction for CH<sub>2</sub>O and CD<sub>2</sub>O.

A number of applications of this method have been reported in literature producing encouraging results [74-77]. Melanson et al. [74] illustrated that stable-isotope dimethyl labelling leads to significant enhancement of the a<sub>1</sub> ion in fragmentation spectra of all peptides studied. The a<sub>1</sub> ion was used as the precursor ion for MRM and removed tedious method development and optimization for each peptide studied. Stable isotope dimethyl labelling has been used in several studies to identify and quantify proteins in mouse macrophage-like cells [75], and an E-cadherin-deficient human carcinoma cell line and its transfectants expressing E-cadherin [76]. Stable isotope dimethyl labelling was also used to determine proteome differences between different parts of the rod outer segment of photoreceptor cells in bovine retina [77]. Differences identified with isotope labelling were in agreement with those identified by western blotting, but many more proteins could be quantified with dimethyl labelling.

There are a number of drawbacks to the stable isotope dimethyl labelling method. First, sodium cyanoborohydride is a toxic compound that can release hydrogen cyanide gas if exposed to strong acid, and formaldehyde is a suspected carcinogen. This can easily be overcome by doing the reaction in the fume hood. Second, as with the ICAT method, due to the composition of the stable isotopes, there is a chance that differentially labelled peptides could separate in the LC analysis. This poses complications for

quantitative analysis which typically assumes co-eluting analytes and determines the ratio at a single time point. Hsu et al. [49] do comment about this problem and say “we have investigated this isotopic effect and found that inaccuracy in quantitation caused by such an isotopic effect is negligible for the samples that we have analyzed”. However, a more comprehensive study on the effect of isotope labelling on retention characteristics is needed and is addressed in Chapter 4.

## **2.2 Biological Based Methods**

### **2.2.1 Metabolic Methods**

Labelling of different metabolic precursors in the cell during growth and division is the earliest possible time to introduce a stable isotope label into proteins. Through cell growth, the whole unlabeled protein populations are replaced with stable isotope-labelled versions before the quantitative proteomics experiment is started. The cells’ states of interest are grown in light and heavy media, producing the mass difference that is distinguished by MS. This was first described by Langen et al. in 1998 [44], where  $^{15}\text{N}$ - and  $^{13}\text{C}$ -labeled precursors were used for the quantitative comparison of proteins with 2D gels. In 1999, Oda et al. [78] used  $^{15}\text{N}$ -labeled precursors to quantify yeast phosphopeptides and in 2001 Conrads et al. [79] used the same method to quantify a mammalian cell line. However in 2002, Ong et al. developed the today’s most frequently used method called stable isotope labelling by amino acids in cell culture (SILAC) [80]. The initial SILAC employed D0-leucine and D3-leucine as its light and heavy media [80]. However, separation was observed between the SILAC D0 and D3 grown peptides, which prompted the development of the  $^{12}\text{C}_6$ -arginine and  $^{13}\text{C}_6$ -arginine media [81]. These media ensure that all tryptic digested peptides have at least one labelled amino acid

resulting in a mass that is different from its unlabeled counterpart. The key advantage of all metabolically labelling methods is that labelled intact cells can be combined before digestion, fractionation and purification, which removes quantitation errors and any experimental variance associated with separate preparation steps. However, the time and cost for creating and maintaining these methods is often substantial compared to the information provided from the experiment.

### **2.2.2 Enzymatic Methods**

Stable isotopes essential for quantitation can also be introduced into the peptides by the protease cleaving the protein to a peptide mixture. This was first described in 1983 by Rose et al. to assist *de novo* sequencing of peptides by MS [45], but recently has been applied to quantitative proteomics [82-87]. Enzymatic labelling can be executed either during proteolytic digestion with  $^{18}\text{O}$  labelled water [82-84] or after proteolysis in a second incubation step with the protease in a small volume of  $\text{H}_2^{18}\text{O}$  water [85, 86] or deactivating the protease through a reduction and alkylation step [87]. When trypsin or Glu-C is used for digestion, two oxygen atoms are introduced, resulting in a 4 Da mass difference, which is ample for differentiation of peptide isotopic profiles [82-84]. However, if Lys-N and other enzymes are used, they introduce only one oxygen atom, resulting in a 2 Da mass difference, which is insufficient to identify peptide isotopic profiles [88]. Major disadvantages are that labelling is rarely fully complete and that the rate of labelling is different for each peptide, which complicates data analysis [89, 90]. The overall advantage is that, because peptides are labelled enzymatically, common side reactions in chemically introduced labelling are avoided.



### **2.3 Absolute Quantitation by Spiked Synthetic Peptide Standards**

Originally described in 1983 by Desiderio et al. [46], this is one of the first methods of quantitative mass spectrometry where stable isotope-labelled synthetic peptides are used as internal standards that are spiked into samples at known quantities. This method today is commonly termed as AQUA (absolute quantitation of proteins) [91, 92]. AQUA-like methods are the only isotopic labelling methods that can achieve absolute quantitation by comparing the mass spectrometric signal of the known quantity of synthetic peptides to peptides in the sample. However, even though protein quantitation is absolute, AQUA has a number of disadvantages. One drawback is that one has to estimate the amount of the labelled standard that should be added to a sample. This amount may be different for all proteins of interest as their expression levels may differ greatly within a sample. Another limitation is that there are likely other peptides present in the sample that match the mass of the synthetic peptide. As a result, the protein quantitation determined by AQUA may not reflect the true expression level of the protein in the cell. Both of these issues can be improved by multiple reaction monitoring [93, 94] in which the (triple quadrupole) mass spectrometer monitors both the intact peptide mass and one or more specific fragment ions of that peptide over the course of an LC-MS experiment. The main advantage of this method is for studies aimed at, for example, measuring the levels of particular peptide modifications or the analysis and validation of potential biomarkers in clinical samples.

## 2.4 Label-Free Methods

Presently, two generally used but essentially different label-free quantitation strategies have been described: (1) measuring and comparing the mass spectrometric signal intensity of peptide precursor ions belonging to a particular protein [47, 95-99] and (2) counting and comparing the number fragment spectra identifying peptides of a given protein [100-102]. In the first approach, for each peptide, an ion chromatogram is extracted from an LC-MS/MS run and the mass spectral peak areas are integrated over the chromatographic peak. This is typically done by creating extracted ion chromatogram (XIC) for the  $m/z$  values determined for each peptide. The intensity value for each peptide in one experiment can then be compared to the respective signals in one or more other experiments to yield relative quantitative information. However, to achieve better quantitative precision (more mass spectra), proteome coverage is reduced (fewer fragmentation spectra), and vice versa. The relatively new spectral counting approach is based on the assumption that, if a particular protein is more abundant in a sample, more fragmentation spectra are collected for peptides of that protein. Therefore, relative quantitation can be obtained by comparing the number of peptides identified for a protein between a pair of experiments. However, in the original spectral counting it is assumed that the linearity of response is the same for every protein. Actually, the spectral count value is different for every peptide because the chromatographic behaviour (retention time, peak width) changes for every peptide [103]. Because the relationship between amount of protein and number of fragmentation spectra is not valid, in 2002 Rappsilber et al. [104], developed a new estimator for protein quantitation called protein abundance index (PAI). This index is the number of observed peptides divided by the

number of all possible tryptic peptides from a particular protein. In 2005, Ishihama et al. [105] optimized the PAI into an exponentially modified PAI that showed a better relationship to known peptide amounts. The obvious advantages of label-free methods are: (1) the steps of introducing the isotopic label into proteins or peptides can be omitted and (2) there is no limit to the number of experiments that can be compared. Overall, label-free methods look promising, but certainly they are the least accurate among the mass spectrometric quantitation methods when considering the whole approach of a quantitative proteomics experiment.

## **2.5 Summary**

The field of quantitative proteomics has grown in recent years and one reason for this is stable isotopic labelling. This chapter has reviewed the most commonly used quantitative methods, including chemical and biological labelling, and label free methods for protein quantitation. As previous noted, the use of each method is dependent on the preferences of the researcher, instrument capability, cost and the nature of investigation. Ultimately a researcher would prefer to use a method that can do absolute quantitation, but there are no fully validated methods to date. The next obvious choice would be biologically-based methods, but the costs of these methods are extremely high. For these reasons the best choice for the present work was to use a chemically-based method. While the methods described in the remainder of this thesis could be adapted to any of the labelling techniques, stable-isotope dimethyl labelling was chosen for its simplicity, low cost and relatively small retention time differences for isotopically labelled peptide pairs (which will be discussed in Chapter 4).

## **Chapter 3**

### **Methods and Data Analysis**

In this thesis, two quantitative proteomic experiments were performed, one on bovine serum albumin (BSA) and the other on *S. cerevisiae* (yeast), using the procedure described in Section 3.1. These data sets will be designated as BSA and YEAST in this work. The BSA data set consisted of four replicate experiments of the same BSA sample and was carried out by Mark Wall in the Chemistry Department at Dalhousie University in December 2009. The BSA sample was split into five sub-samples, digested separately and combined to make four replicates as described in Section 3.1. The YEAST data set consisted of one fraction of yeast from the cation exchange separation and was carried out by the author.

In both cases, the raw data files were obtained from the mass spectrometer and preprocessed as described in Section 3.2 using MatLab<sup>®</sup> (R2009b, MathWorks, Natick, MA) functions written by the author and software from the Seattle Proteome Center (SPC).

### **3.1 Materials and Methods**

#### **3.1.1 Reagents and Standards**

BSA, bovine trypsin (catalogue T8802), trifluoroacetic acid (TFA), triethylammonium bicarbonate (TEAB), ammonium bicarbonate (ABC), formaldehyde, D<sub>2</sub>-formaldehyde, sodium cyanoborohydride and formic acid were obtained from Sigma (Oakville, Canada). Dithiothreitol (DTT) and iodoacetamide (IAA) were purchased from

Bio-Rad (Hercules, CA). Milli-Q grade water was purified to  $18.2 \text{ M}\Omega \text{ cm}^{-1}$ . Solvents were of HPLC grade and were from Fisher Scientific (Ottawa, Canada).

### **3.1.2 Yeast Extraction and Preparation**

*S. cerevisiae* was grown and protein concentration was determined as described in Wall et al. [106]. A 1 mg yeast pellet was used for digestion, isotopic labelling and quantitative analysis.

### **3.1.3 Tryptic Digestion**

BSA, prepared in 250 mM Tris-HCl (pH 8) was reduced through addition of DTT (in 250 mM Tris) to a final concentration of 5 mM, with incubation at 55 °C for 20 minutes. Then, IAA (250 mM Tris) was added to a final concentration of 12.5 mM, with incubation at room temperature in the dark for 20 minutes. The final BSA concentration was 0.5 g/L. The BSA sample was divided into five  $\times$  100  $\mu\text{L}$  aliquots (50  $\mu\text{g}$  protein) to which 390  $\mu\text{L}$  water was added, along with 10  $\mu\text{L}$  trypsin (0.1 g/L in water) per vial. Digestion proceeded overnight (16 hrs) at 37 °C. The digests were terminated through addition of 50  $\mu\text{L}$  of 10% TFA per vial.

The 1 mg yeast pellet was reconstituted in 100  $\mu\text{L}$  of 8 M urea and Tris-buffer. 530  $\mu\text{L}$  ABC was then added to sample to decrease the concentration of urea to be less than 1.5 M. DTT was added to a final concentration of 2 mM, incubated at 55 °C for 20 minutes, followed by 69  $\mu\text{L}$  of 200 mM IAA with incubation at room temperature in the dark for 20 minutes. To digest 100  $\mu\text{L}$  of the reduced yeast sample (final protein concentration 5 g/L) 10  $\mu\text{L}$  of 0.5 g/L trypsin was added to the sample. Digestion was terminated by adding of 60  $\mu\text{L}$  of 10% TFA, following an overnight digestion at 37 °C.

### 3.1.4 Isotopic Labelling

Digested BSA and yeast peptides were subject to sample cleanup *via* reversed phase HPLC as described by Wall et al. [106]. The strategy employs a C<sub>18</sub> column with a 0.1% TFA, water/acetonitrile gradient to separate non-protein components (Tris, DTT, IAA) while capturing peptides as a single fraction. The peptide fractions (50 µg per vial) were evaporated to dryness in a SpeedVac ahead of isotopic labelling.

Peptide dimethylation with D<sub>0</sub> (CH<sub>2</sub>O) and D<sub>2</sub> (CD<sub>2</sub>O) formaldehyde was performed as previously described [49]. Briefly, dried peptide samples (50 µg) were reconstituted in 100 µL of 100 mM TEAB (pH 8.5) to which 3.6 µL of 20% D<sub>0</sub> or D<sub>2</sub> formaldehyde was added. Each sample was incubated at room temperature for 5 minutes, followed by addition of 4.2 µL of 6 M sodium cyanoborohydride, with incubation for 2 hours at room temperature. Each sample was then subjected to RP-HPLC sample cleanup [106], dried, and then frozen (-20°C) until LC-MS/MS analysis (for BSA) or strong cation exchange (for yeast).

### 3.1.5 Liquid Chromatography (LC)

Before cation exchange, light and heavy labelled yeast peptide fractions were reconstituted (0.1% TFA in water), and combined in a 1:1 ratio. Cation exchange of the combined yeast sample was executed on a self-packed PolySULFOETHYL A™ strong cation exchange (SCX) column (5 µm beads, 1000 Å pore size, 100 x 1 mm i.d.) from The Nest Group, Inc. (Southboro MA). Separation was done using a linear gradient between solvent A (0.1% TFA in water) and solvent B (0.1% TFA in water with 500 mM NaCl), beginning at 100% A and increasing to 40% B over 40 minutes, then ramping to

60% B over 10 minutes. A total of ten fractions (every 5 minutes) were collected over the gradient. Fractions were dried, reconstituted with 0.1% TFA in water and desalted by reversed phase [106] and dried again before LC-MS/MS analysis.

### **3.1.6 LC-MS/MS**

Prior to LC-MS/MS, the heavy and light labelled BSA peptide fractions were reconstituted (0.1% TFA, water with 5% acetonitrile), and combined in a 1:1 ratio. For BSA, four replicates were prepared, combining the light/heavy labelled fractions (1+2, 2+3, 3+4, and 4+5). A total of 1 pmol BSA was injected per analysis. For yeast, the first SCX fraction was reconstituted with 0.1% TFA in water with 5% acetonitrile, and a total of 1 µg of yeast was injected for LC-MS/MS analysis.

A ThermoFisher LTQ linear ion trap mass spectrometer (Waltham, MA) equipped with a nanospray ionization source coupled to an Agilent 1200 nanoflow HPLC system (Palo Alto, CA) was used to analyze the protein digests. Separation was on a self-packed C<sub>12</sub> reversed phase column (30 cm x 75 µm i.d., 3 µm Jupiter beads from Phenomenex, Torrance, CA) flowing at 0.25 µL min<sup>-1</sup>. The gradient consisted of a linear increase from 5% to 30% acetonitrile over 100 minutes, followed by an instantaneous increase to 80% acetonitrile to regenerate the column. The nanospray ionization voltage was set at 2.5 kV and the transfer capillary temperature was set to 225 °C. The MS scan range was 400-1700 *m/z*. The ion trap had the maximum fill time set at 100 ms, and the automatic gain control was set to allow up to 1 x 10<sup>5</sup> ions to enter the trap for MS and 1 x 10<sup>4</sup> ions for MS/MS. Data acquisition used dynamic exclusion, collecting one MS scan followed by tandem MS scans of the top five ions, with an exclusion duration time of 30 seconds.

### 3.1.7 SEQUEST Parameters

Peptide identification for BSA was performed using the SEQUEST algorithm within the Thermo Xcalibur Bioworks (v. 3.3) software package. Peptide filters were set to achieve a false positive rate of 1% or less when the reversed BSA database was included in the search.  $X_{\text{corr}}$  (cross-correlation value) versus charge state was set to 1.50 (+1), 2.00 (+2) and 3.25 (+3) for searches. Peptide probability was set to 0.01 with  $\Delta\text{CN}$  (delta correlation value)  $\geq 0.1$  and  $R_{\text{Sp}}$  (ranked preliminary score)  $\leq 4$ . Two searches were completed: the first employed fully tryptic peptides with up to 2 missed cleavages; the second employed fully non-tryptic peptides (i.e. non-specific cleavage at any amino acid), again permitting up to 2 missed cleavages (at lysine or arginine). The data set was searched for both fully tryptic as well as non-tryptic peptides against BSA, sequence-reversed BSA, trypsin and a collection of 10 commonly observed protein contaminants.

## 3.2 Data Preprocessing

A significant amount of data preprocessing was carried out to manipulate the data into MatLab<sup>®</sup>. The proprietary RAW file that is obtained from the Thermo LTQ mass spectrometer cannot be directly imported into MatLab<sup>®</sup>. The most straightforward way is to convert the RAW file format into an open mzXML (extensible markup language) format [107] that is compatible with MatLab<sup>®</sup>. The mzXML format is a protocol for storage and exchange of mass spectrometry data developed at the SPC Institute for Systems Biology and provides a standard container for MS and MS/MS proteomic data. RAW files are converted into mzXML files using the command line program called ReAdW which was downloaded from the SPC website [108]. A detailed procedure of how the ReAdW software was used to convert RAW files into mzXML files is described



in Appendix A. The mzXML files can be observed using a standard text file reader like gVim [109] (Figure 3.1).

Each mzXML file consists of three different types of sections; the header, sub-header and body. The header provides the details about the experiment: number of scans (scanCount), start time (startTime), end time (endTime), file name (fileName) and file type (fileType). It also provides information on the mass spectrometer: manufacturer (msManufacturer), model (msModel), ionization (msIonisation), mass analyzer (msMassAnalyzer), detector (msDetector) and acquisition software (see Figure 3.1A). The sub-header is the section that has the information corresponding to each scan that was compiled in the experiment: scan number (scan num), MS level (msLevel), peak count (peaksCount), scan type (scanType), retention time (retentionTime), low  $m/z$  (lowMz) and high  $m/z$  (highMz) (see Figure 3.1B-D). The sub-header for a MS/MS scan also provides the collision energy (collisionEnergy), activation method (activationMethod) and the parent ion for the MS/MS scan (highlighted in Figure 3.1D). There is a sub-header for each scan compiled and it provides important information to distinguish what type of scan it is, MS (Figure 3.1B) or ZOOM (Figure 3.1C) or MS/MS (Figure 3.1D). The variables used to distinguish between each scan is the msLevel (MS = "1", ZOOM = "1", MS/MS = "2") and scanType (MS = "Full", ZOOM = "Z", MS/MS = "Full"). The body is the section where the  $m/z$  and intensity data is stored (see Figure 3.1E). However, the data are stored in a Base64 encoding scheme that represents binary data in an ASCII string format. Base64 encoding schemes are commonly used when there is a need to encode binary data that needs to be stored and transferred over media. This is to ensure that the data remain intact without modification during transport.

<b>Header – A</b>		
<pre>&lt;?xml version="1.0" encoding="ISO-8859-1"?&gt; &lt;mzXML xmlns="http://sashimi.sourceforge.net/schema_revision/mzXML_3.1" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://sashimi.sourceforge.net/schema_revision/mzXML_3.1 http://sashimi.sourceforge.net/schema_revision/mzXML_3.1/mzXML_idx_3.1.xsd" &gt; &lt;msRun scanCount="8204" startTime="PT0.5285S" endTime="PT5401.21S" &gt; &lt;parentFile fileName="Aon-1-2.raw" fileType="RAWData" fileSha1="ad72bffc9f55f3bbdefe6d55037f33bde5303a83" /&gt; &lt;msInstrument&gt; &lt;msManufacturer category="msManufacturer" value="Thermo Scientific" /&gt; &lt;msModel category="msModel" value="LTQ" /&gt; &lt;msIonisation category="msIonisation" value="NSI" /&gt; &lt;msMassAnalyzer category="msMassAnalyzer" value="ITMS" /&gt; &lt;msDetector category="msDetector" value="unknown" /&gt; &lt;software type="acquisition" name="Xcalibur" version="2.2" /&gt; &lt;/msInstrument&gt; &lt;dataProcessing&gt; &lt;software type="conversion" name="ReAdW" version="4.3.1(build Sep 9 2009 12:30:29)" /&gt; &lt;/dataProcessing&gt;</pre>		
<b>Sub-Header</b>		
<b>MS Scan – B</b>	<b>ZOOM Scan – C</b>	<b>MS/MS Scan – D</b>
<pre>&lt;scan num="5" msLevel="1" peaksCount="15600" polarity="+" scanType="Full" filterLine="ITMS + p NSI Full ms [400.00-1700.00]" retentionTime="PT4.5938S" lowMz="400.083" highMz="1700" basePeakMz="1352" basePeakIntensity="3459.79" totIonCurrent="1.13515e+007" &gt; &lt;peaks precision="32" byteOrder="network" contentType="m/z-int" compressionType="none" compressedLen="0" &gt;</pre>	<pre>&lt;scan num="6" msLevel="1" peaksCount="500" polarity="+" scanType="Z" filterLine="ITMS + p NSI d Z ms [1253.00-1263.00]" retentionTime="PT4.9434S" lowMz="1253.02" highMz="1263" basePeakMz="1255.38" basePeakIntensity="17.6573" totIonCurrent="1617.35" &gt; &lt;peaks precision="32" byteOrder="network" contentType="m/z-int" compressionType="none" compressedLen="0" &gt;</pre>	<pre>&lt;scan num="7" msLevel="2" peaksCount="1" polarity="+" scanType="Full" filterLine="ITMS + c NSI d Full ms2 1257.80@cid35.00 [335.00-2000.00]" retentionTime="PT6.1991S" lowMz="1143.4" highMz="1143.4" basePeakMz="1143.4" basePeakIntensity="1.55171" totIonCurrent="1.55171" collisionEnergy="35" &gt; &lt;precursorMz precursorIntensity="2104.21" precursorCharge="5" activationMethod="CID" &gt;1257.8 &lt;/precursorMz&gt; &lt;peaks precision="32" byteOrder="network" contentType="m/z-int" compressionType="none" compressedLen="0" &gt;</pre>
<b>Body – E</b>		
<pre>Q8gKq0L8tj9DyBVVQq2oqkPIIABCK/KtQ8gqq0HsxUNDyDVVQxv2RUIPIQABDq+39Q8hKq0N8vxVDyFVVQm9I8EPIYABAE JTTQ8hqqwAAAAABDyHVVAAAAAEPiGABB6lShQ8iKq0LFTZ5DyJVvQsZkaUPloABBrQbMQ8iqq0FwlcNDyLVVQpjTakPIw ABCK12EQ8jKq0JInwNDyNVVQT6FtkPI4AA+4dDCQ8jqqwAAAAABDyPvVAAAAAEPJAAAAAQA8kKqzuV7o1DyRvVQ P9d10PIIABCmfBMQ8kqq0IyrOtDyTVVQOgc90PjqABBqhQ5Q8IKq0LCJspDyVVVQv2iUUPJYABCjYKeQ8lqq0FPUUpDyXV VAAAAAEPJgAAAAAQA8mKqwAAAABDyZVVAAAAAEPJoAA7+b/oQ8mq0GAZYFDybVVQsroB0PJwABCekW0Q8nKq 0GQ5ppDyDyVvVQvp+HEPJ4ABDJiUeQ8nqq0NxAOdDyFVvQ00+A0PKAABDMbS7Q8oKq0QG6NNDyHVVQ9cdPEPKIABDIzWf Q8oqq0HBavVDyJvVAAAAAEPKQAAAAAQA8pKq0IaqJpDylVvQIYggkPKYABDA3asQ8pqq0MemSIDynVVQj47N0PKgA BAg9GxQ8qKq0K9mDtDypVVQzJe40PKoABDGL/G8qqq0KXjrZDyrVVQbSM+kPKwABDarbpQ8rKq0OkeINDytVVQtB/5EP K4ABBHW7pQ8rqwAAAABDyVvVQf3feEPLAABDetKsQ8sKq00TnrJDyxVVQwSjx0PLIABC/fgbQ8sqq0IsQmdDyZVVQpZ/8 UPLQABCOpDQ8tkq0OG9UIDy1VVQ4vrMkPLYABCsneSQ8tqq0Hng8JDy3VVQ0n5K0PLgABDxOBbQ8uKq0N5S3hDy5Vv QqLuskPloABDENT2Q8uqq0LZ9CpDy7VvVQgHoU0PLwABCqDMDQ8vKq0KpMipDy9VVQwzz4EPL4ABDH9clQ8vqq0NFGv xDy/VVQ6RsvUPMAABD52cSQ8wKq0PIP8xDzBVVQ8jFG0PMIABDggBuQ8wqq0MtWyxDzDVVQ4P56PEPMQABC8MngQ ...</pre>		

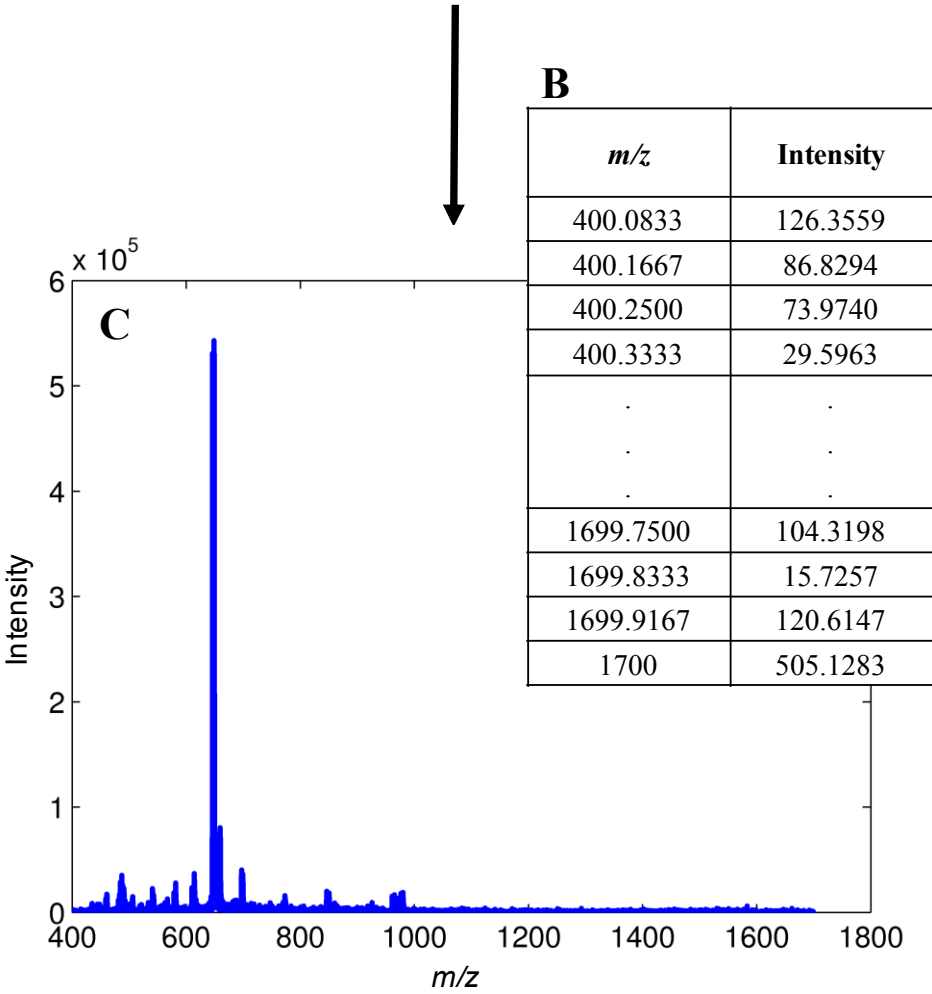
**Figure 3.1** A selection of an mzXML file from the BSA data set with the Header (A), Sub-header (MS scan (B), ZOOM scan (C) and MS/MS scan (D)) and Body (E) sections.

Base64 is used commonly in a number of applications including email via multipurpose internet mail extensions and storing complex data in XML in this case. The Base64 encoded data in each body is converted using the `base64cvt.m` MatLab<sup>®</sup> function (Appendix B). This converts the Base64 encoded data into single precision data values that can be used in MatLab<sup>®</sup>. The procedure for this conversion is shown in Figure 3.2. In brief, the conversion of each single precision data value is done by taking the Base64 body characters and converting them into a Base64 index [110], translate into binary code, transform into 8 bits, convert into decimal index, reorder the bytes, and finally convert into a single precision data value [111] (Figure 3.2A). These single precision data values are assembled into a data matrix (Figure 3.2B) which composes a mass spectrum (Figure 3.2C).

The information from the whole `mzXML` file is exported into MatLab<sup>®</sup> using the function called `rawexportfull.m` (Appendix C) written by the author. This function requires an `mzXML` file and will output a structured array that contains the processed data that is used in Chapter 4 and 5.

**A**

<b>Body character</b>	Q	8	g	K	q	0	L	8
<b>Packet reversal</b>	K	g	8	Q	8	L	0	q
<b>Base64 index</b>	10	32	60	16	60	11	52	42
<b>Binary code</b>	010100	000001	001111	000010	001111	110100	001011	010101
<b>Convert to 8 bits</b>	01010000	00010011	11000010	00111111	01000010	11010101	01010101	
<b>Decimal index</b>	10	200	67	252	66	171		
<b>Byte reordering</b>	67	200	10	171	66	252		
<b>Convert to 4 bytes</b>	67	200	10	171	66	252		
<b>Byte reordering</b>	171	10	200	67	159	180		
<b>Binary code</b>	11010101	01010000	00010011	11000010	...			
<b>Single precision data value</b>	400.0833					...		



**Figure 3.2** The procedure for the first *m/z* data value (A) which is put into a data matrix (B) and plotted as a mass spectrum (C).

## Chapter 4

# Chromatographic Behaviour of Peptides Following Dimethylation with H<sub>2</sub>/D<sub>2</sub>-Formaldehyde: Implications for Comparative Proteomics

### 4.1 Introduction

In mass spectrometry, the use of stable isotopes is a preferred approach to quantitative proteomics. Various protein labelling strategies incorporate isotopes through metabolic processes [80-81], enzymatic digestion [82-87], or reaction with chemical tags containing <sup>2</sup>H, <sup>13</sup>C, <sup>15</sup>N, or <sup>18</sup>O [43, 48, 58]. In an LC-MS experiment, the chromatographic profiles of the isotope analogues should ideally exhibit perfect overlap so that the relative MS ion abundance of the eluting peptide pair indicates the concentration ratio of peptides. Unfortunately, the chromatographic separation of isotopically labelled peptide pairs is commonly observed, especially for deuterated compounds. In this case, the heavy isotope generally elutes in advance of its lighter counterpart [112]. This phenomenon has influenced current practices of quantitative proteomics.

Numerous studies have investigated the extent of isotopic separation in proteomics. As seen with the original ICAT, Gygi et al. acknowledged a 1-2 s separation between D<sub>0</sub> and D<sub>8</sub>-labelled ICAT peptides [43], which could have implications for quantitation. In contrast, it has been observed that labels incorporating isotopes other than deuterium generally lead to a much smaller separation of isotopic pairs. This is not surprising given that there is a doubling of mass for D vs H, while the change for <sup>13</sup>C over <sup>12</sup>C, for example, is only about 8%. A general consequence of this observation is that

labels employing isotopes other than deuterium, though more costly, have become the preferred choice in proteomics studies [113]. To illustrate, while the original SILAC reagent employed D<sub>3</sub>-leucine [80], chromatographic separation of the resulting peptides (on the order of half a peak width) prompted development of a <sup>13</sup>C<sub>6</sub>-arginine SILAC reagent [81]. Likewise, an updated cleavable ICAT reagent eliminated deuterium in favour of <sup>13</sup>C [58]. The iTRAQ reagent also avoids deuterium in favour of <sup>13</sup>C, <sup>15</sup>N, and <sup>18</sup>O [48]. With all such commercial reagents, the isotopic pairs have been shown to co-elute during RPLC, ensuring correlation of the MS intensity of respective isotopic pairs and providing more reliable quantitation based on point measurements.

To address quantitative issues arising from isotopic separation of deuterated peptide pairs in the original ICAT method, Gygi et al. integrated the peptide peak areas over their respective elution profiles [43]. Given that peak integration can potentially overcome the issue of chromatographic separation of isotopes, numerous researchers have weighed the benefits of deuterated labelling reagents over more costly <sup>13</sup>C isotopically labelled compounds. Although peak integration is a potential solution to the problem of isotopic separation, it increases the complexity of automated data analysis. Moreover, comparison of peak areas for chromatographic signals separated in time implies that the sensitivity of the MS detector remains constant. Given the stochastic nature of the electrospray process, such an assumption may not be valid. Thus, a preferred solution to balance cost and quantitative reliability would be to identify deuterated labels that minimize isotopic separation. Such labels have been identified in the literature [49], though they have not come into widespread use, perhaps owing to a

lack of commercial availability, or a perception that all deuterated labels lead to isotopic separation.

Considering alternatives to commercial isotopic reagents, one of the simplest and increasingly popular methods is the use of D<sub>0</sub> or D<sub>2</sub> formaldehyde [75-77], which dimethylates primary amines (lysine residues and N-termini of peptides). Through peptide dimethylation, the resulting 4 Da mass difference per label is preferred to the 2 Da difference obtained with <sup>13</sup>C formaldehyde. Expanding on previous observations for deuterated analogues, one would assume a chromatographic time shift following reaction with D<sub>0</sub>/D<sub>2</sub> formaldehyde. Perhaps surprisingly, in an early study proposing this reagent for quantitative proteomics, Hsu et al. noted a negligible chromatographic time shift [49]. It should be noted however that such an observation was made from a single BSA peptide, and may not be representative of a complex series of labelled peptides.

While deuterium isotopic effects are more pronounced than those of <sup>13</sup>C, the incorporation of deuterium alone does not dictate chromatographic separation. Zhang et al. [55] observed that the chromatographic time shift increased as a function of the number of deuterium atoms in the labelled peptide (for structurally similar labels). This gave rise to a concept of “specific resolution”, referring essentially to the time shift afforded per deuterium isotope. These authors also noted that the relative time shift was larger for smaller peptides with the same label. Boersema et al. [114] speculated that the partial separation of dimethyl-labelled peptides was due to the higher hydrophilicity of the C-D bond over that of the C-H bond. Such an explanation would imply a differential separation dependent on the relative contribution of the C-D bond to the retention of a given peptide. Furthermore, the location of deuterium in a given isotopic compound

would also be important in influencing the relative retention of isotopic pairs on a reversed phase chromatographic support. This argument was presented by Zhang et al. [55], who noted that the incorporation of a deuterium in a more polar (charged) region of the molecule led to smaller observed differences in the chromatographic retention of H/D labelled peptides.

In this work, a comprehensive study was carried out to investigate the effect of dimethyl labeling with H<sub>2</sub>/D<sub>2</sub> formaldehyde on the retention characteristics of differentially labeled peptides. The objectives of the work reported here were to (1) determine whether there is a significant difference in retention (measured in terms of time as well as peak resolution) across a complex mixture of peptides, and (2) assess the consequences of any observed separation on proteome quantitation. Consistent with previous observations using deuterium, a statistically significant separation of dimethylated peptides (D ahead of H) was observed. However, this separation is inconsequential in terms of peptide quantitation based on point measurements (i.e. without peak integration). These results are explained with consideration of the relative high polarity of the dimethylated peptides during chromatographic separation.

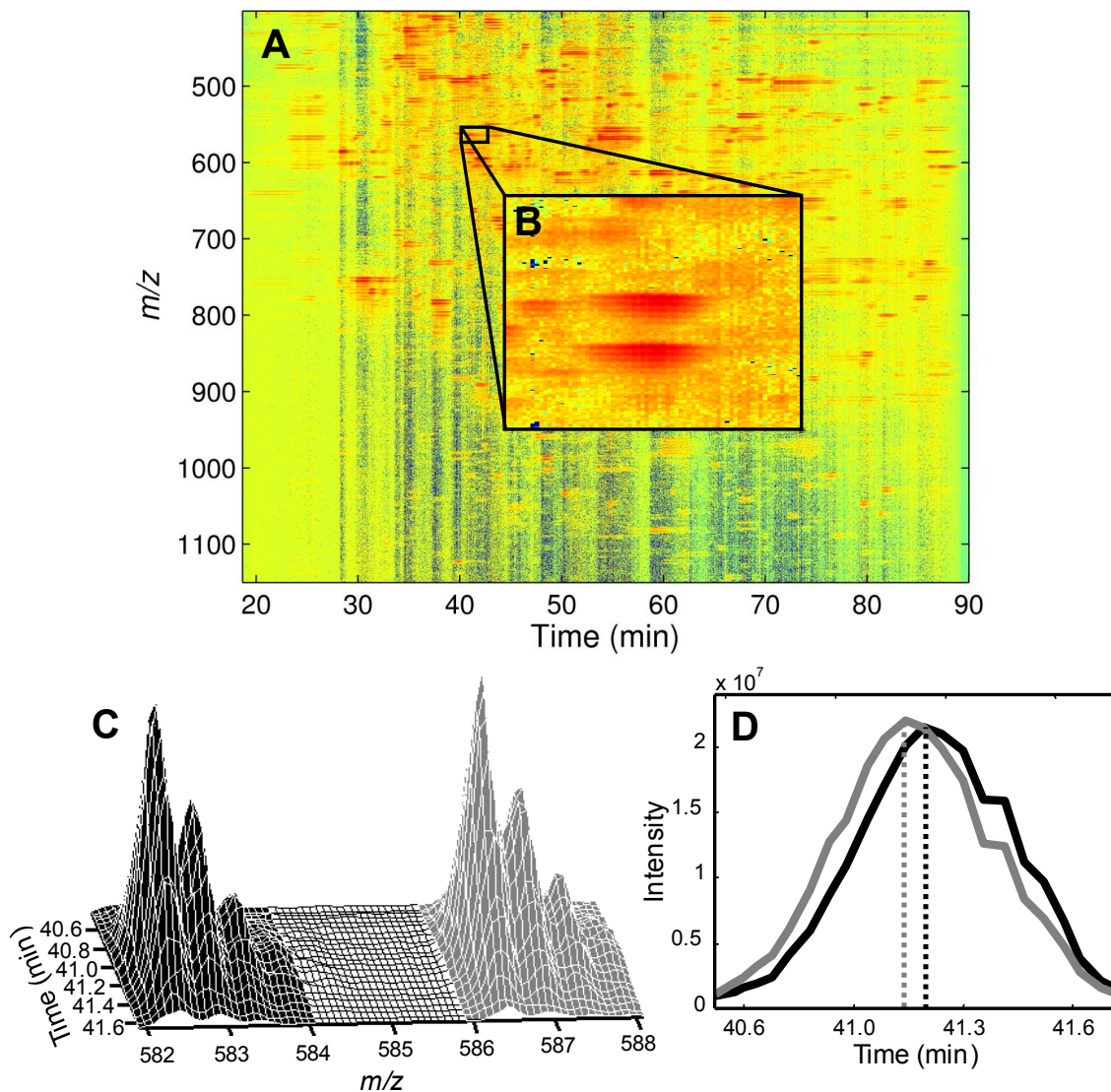
## **4.2 Methods**

### **4.2.1 Data Selection and Preprocessing**

A multiconcensus SEQUEST search of the database resulted in a total of 155 unique BSA peptides identified as either the heavy or the light tag in one or more of the four replicate runs. For each of the 155 identified peptides, extracted ion chromatograms of the respective isotopic profiles for heavy and light pairs were obtained for all of the



replicates in which the peptide was identified. For example, given a mass chromatogram (Figure 4.1A) for one replicate, the XICs (Figure 4.1D) for peptide pairs are created by integrating the light (black shading) and heavy (gray shading) component regions as indicated on Figure 4.1C. The regions correspond to the mass and retention time ranges required to construct the XIC for the particular peptide pair. These regions are different for each peptide pair combination (charge and number of labels). For peptides that were not initially identified in all four replicates, a manual search was carried out, resulting in additional positive hits when the presence of a peptide overlooked by SEQUEST could be verified. The number of peptides used for this study was subsequently reduced to 71 by applying the following criteria: (1) both peptide peaks (heavy and light) needed to be confirmed in all four replicates, (2) the isotopic patterns needed to be clearly distinguishable from baseline noise and free from obvious mass interferences, and (3) the signal, as represented by the XIC, needed to be of an adequate quality. The first two criteria were sufficient to eliminate most of the discarded peptides, and the last was subjectively assessed based on the likelihood of the XIC yielding reliable statistical parameters (median, width, etc.). From the initial set of 155 peptides pairs, a total of 64 identified peptides were rejected by the first criterion, 15 by the second criterion, and 5 by the third criterion. The resulting 71 reliable unique BSA peptides were used in this work.



**Figure 4.1** Two-dimensional representation of a mass chromatogram for one BSA replicate (A), an example of a peptide pair (B), a three dimensional surface plot (C) of the response indicated on (A), and the calculated XIC (D) for the response, where the dashed lines are the estimated median retention times for the peptide pair.

#### 4.2.2 Signal Processing

After preprocessing, statistical parameters including time difference, chromatographic resolution and quantitative ratio were calculated using the median retention time and peak widths at half height for each XIC for every replicate of each peptide pair. The median of the elution profiles, representing the time encompassing half

of the peak area, was chosen as the most objective representation of retention time, especially when peak maxima were not clearly defined. A summary table that includes the sequence, masses of the light and heavy molecular ions, charge, number of labels, time difference and resolution for all 71 unique peptide pairs is available in Appendix D.

### **4.2.3 Computational Calculations of Compound Polarity**

All calculations were performed by Stephen Christensen with Spartan (Wavefunction Inc., Irvine CA) '08 computational software. The DFT functional B3LYP with 6-31G\* basis set was used for the dipole moment calculations for the labelled species presented in Figure 4.7.

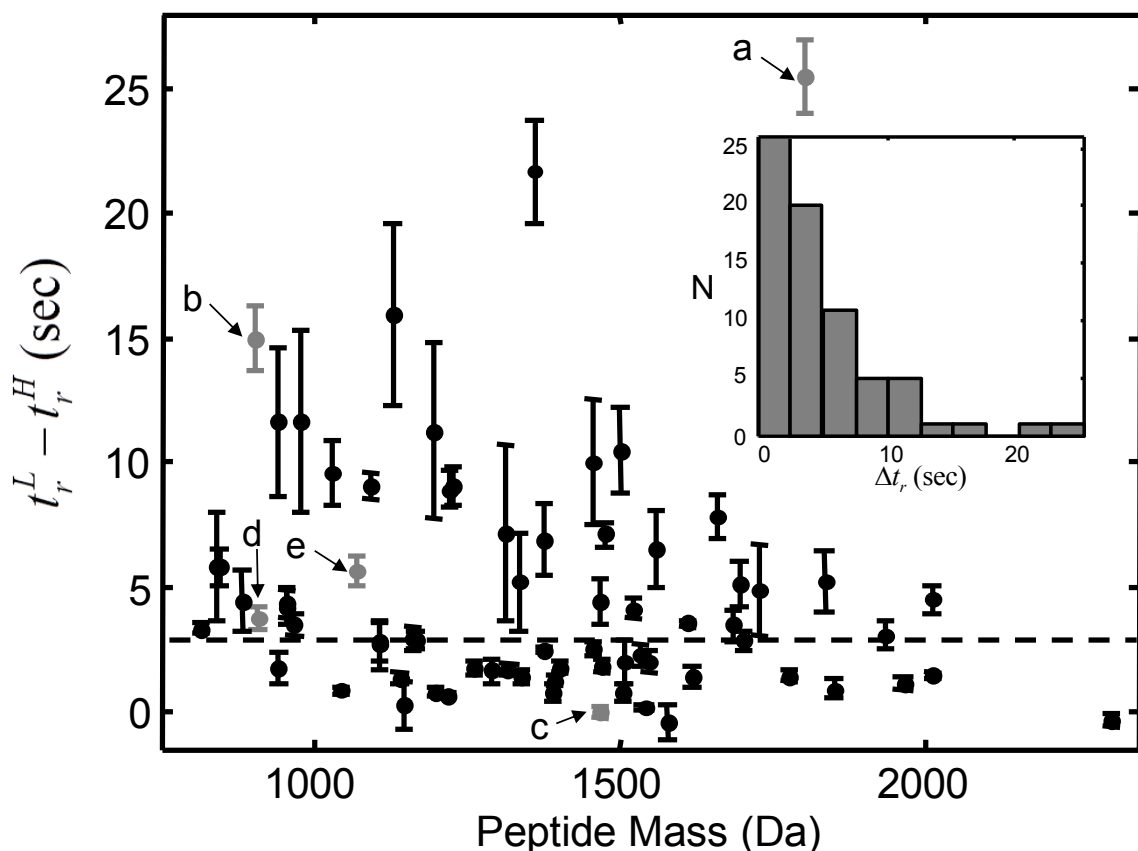
## **4.3 Results**

A total of 71 unique labelled BSA peptide pairs were identified in all four replicate samples of labelled BSA using the criteria previously noted. In cases of multiple charge states, only the most abundant was included in this list. The 71 peptides included 54 doubly charged ions (14 with one label, 37 with two labels and 3 with three labels), 15 triply charged ions (2 with one label, 9 with two labels and 4 with three labels) and 2 quadruply charged ions (two labels). Out of the 71 peptide pairs, 38 were fully tryptic peptides (28 fully cleaved and 10 with one missed cleavage), 31 were semi-tryptic peptides (29 full cleaved and 2 with one missed cleavage) and 2 were non-tryptic peptides. The peptides account for a BSA sequence coverage of 67% with masses ranging from 813 to 2303 Da with a median mass of 1375 Da. Retention times varied between 20.9 and 83.5 min with a median value of 49.1 min.

### 4.3.1 Retention Time Differences

Figure 4.2 summarizes the time differences calculated from each of the 71 heavy/light peptide pairs, plotted as a function of the apparent mass of the light peptide in the pair. The time difference is calculated as the average retention time difference (light,  $t_r^L$ , minus heavy,  $t_r^H$ ) of each peptide pair across four replicate runs. Here, the retention time is defined as the median time calculated for each XIC, as shown in Figure 4.1D. Because the sampling interval was not uniform on the time axis, a weighted calculation was used to determine the median from the chromatographic profile. The limits for the median calculation were 10% of the maximum peak intensity. In Figures 4.2, 4.4, 4.5C and 4.5D the error bars represent one standard deviation of the mean ( $\pm s/\sqrt{4}$ ).

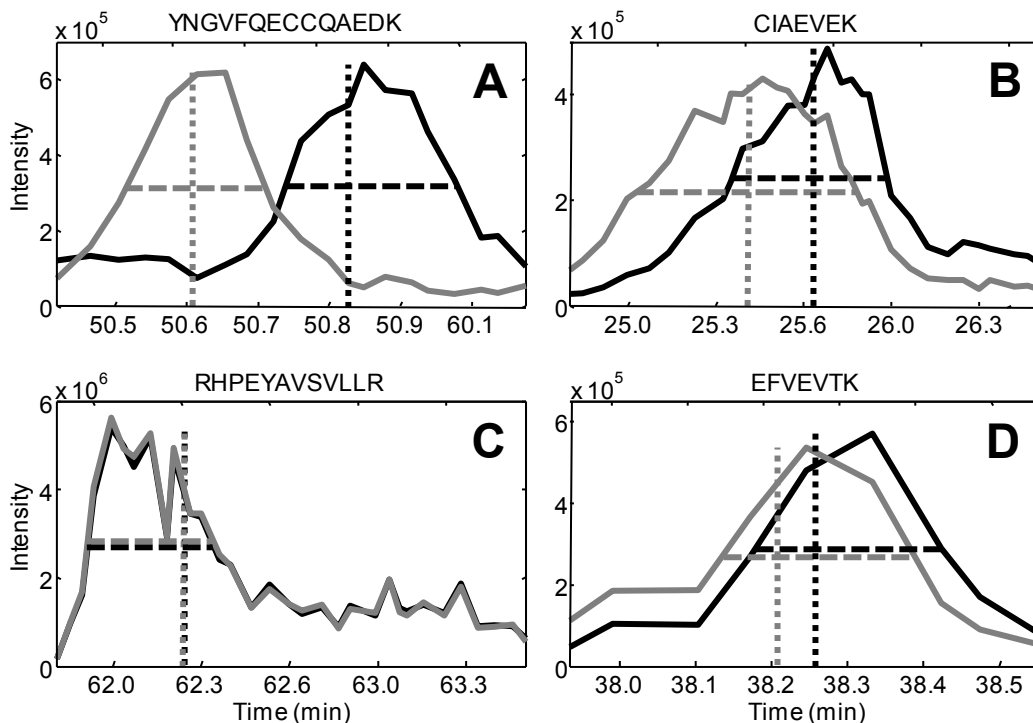
Figure 4.2 shows that the retention time differences of the various peptides are highly variable, ranging from -0.40 to 25.45 seconds. The majority of the retention time differences (69 out of 71) are in the positive region and correspond to elution of the heavy-labelled peptide ahead of the light. This positive bias is expected based on previous studies [54, 55, 115]. No meaningful correlation was observed ( $r^2 = 0.154$ ) between retention time difference and the apparent peptide mass. Similarly, the time difference also did not correlate with the retention time of the peptide pairs, their observed mass, charge state, number of labels (related to the number of missed cleavages) or cleavage type (fully or non-specific).



**Figure 4.2** Calculated mean retention time differences as a function of apparent peptide mass for the 71 unique BSA peptide pairs. A histogram of the time differences is inset. Error bars indicate one standard deviation of the mean. Representative points shown in other figures are labelled “a”-“e”. The dashed line corresponds to the median retention time difference.

As seen in the inset histogram of Figure 4.2, the retention time differences have a tailed distribution with most of the differences between 0 and 8 seconds and a median value of 3.36 seconds. Four peptides had an average time difference greater than 15 seconds. The selected XICs for two of these peptide pairs (indicated by “a” and “b” in Figure 4.2) are provided in the subplots of Figure 4.3 (A and B). As seen from these XIC profiles, these peptides display a significant degree of chromatographic separation, with the heavy component (gray) eluting ahead of the light component (black). The other two peptide pairs with a time difference greater than 15 seconds were characterized by features similar to the ones shown in Figure 4.3A and 4.3B. Also shown in Figure 4.3 are

the selected XICs from a peptide pair with a time difference near zero (Figure 4.3C, indicated by “c” in Figure 4.2) and a pair with a time difference near the median value (Figure 4.3D, indicated by “d” in Figure 4.2).

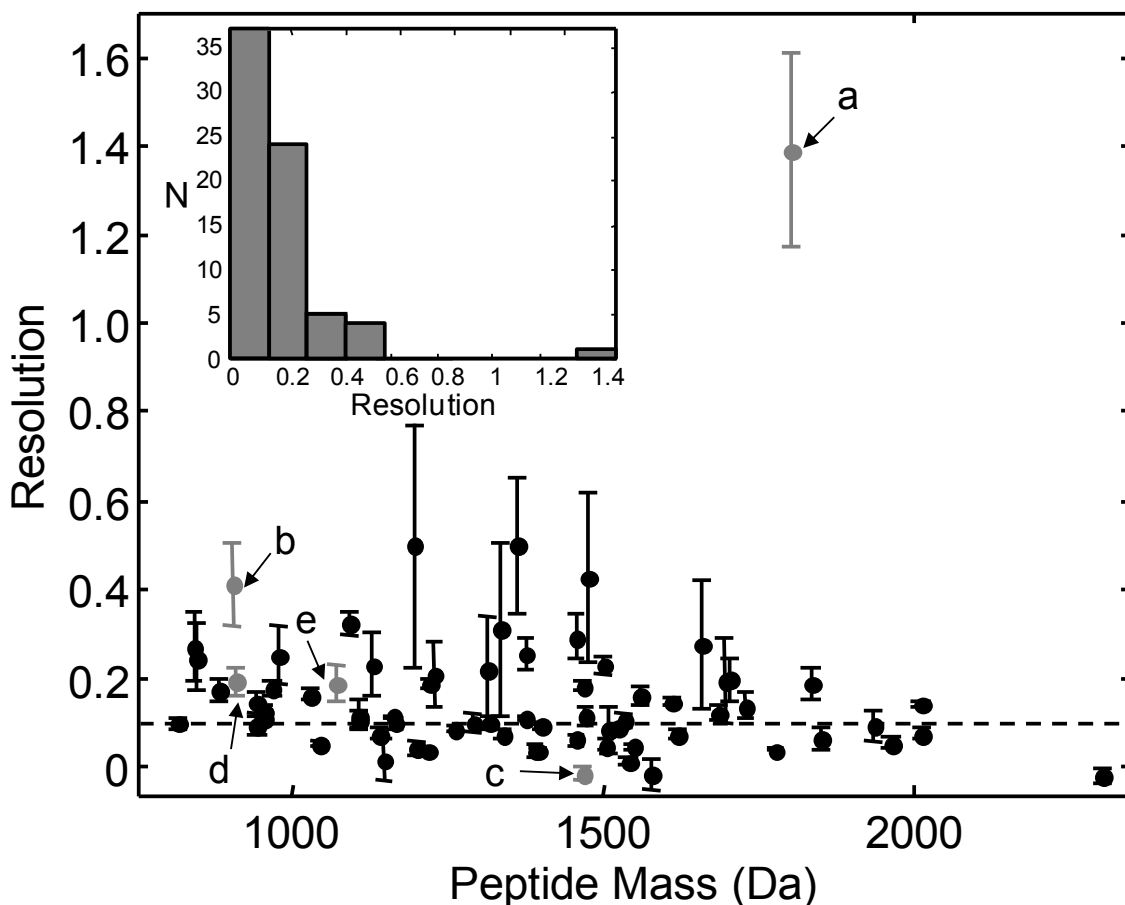


**Figure 4.3** Four selected XIC replicates for the indicated peptide pairs at positions “a” through “d” in Figure 4.2. The vertical lines correspond to the median retention times and horizontal lines represent the peak widths at half height for the light (black) and heavy (gray) components. The confirmed amino acid sequence of each peptide pair is provided in the figure.

### 4.3.2 Resolution

The absolute time differences between heavy/light labelled peptide pairs highlight the retention differences experienced by hydrogen vs deuterium labelled peptides. Chromatographic resolution provides a quantitative measure of the degree of separation resulting from this time difference, and is presented in Figure 4.4. Again, peak resolution is plotted as a function of the apparent peptide mass of the light peptide in the pair. The resolution for one peptide pair is defined as the average retention time difference

$(t_r^L - t_r^H)$  divided by the average peak width at half maximum for the light and heavy components. Thus resolution can be negative if the light component elutes before the heavy component.



**Figure 4.4** Calculated mean resolution as a function of apparent mass for the 71 unique BSA peptide pairs. A histogram of the resolutions is provided in the inset. Error bars indicate one standard deviation of the mean. The dashed line corresponds to the median resolution.

A peak width was calculated from the maximum of the XIC for both components to the points of earliest intersection at half height as shown in Figure 4.3. Despite potential variations in the peak widths of the eluting peptides, the resolution correlates strongly with the time difference ( $r^2 = 0.98$ ). The resolution between peaks ranges from -0.022 to 1.39. The peptides pairs labelled “a” through “e” in Figure 4.2 are similarly

labelled in Figure 4.4. The peptide pair with the largest time difference also has the largest resolution. With the exception of the peptide pair labelled “a” in the figure, the remaining peptides have resolution below 0.6. As shown in the inset of Figure 4.4, the majority of peptide pairs (67 out of 71) have resolutions between 0 and 0.6, with a median value of 0.114.

### 4.3.3 Quantitative Ratios

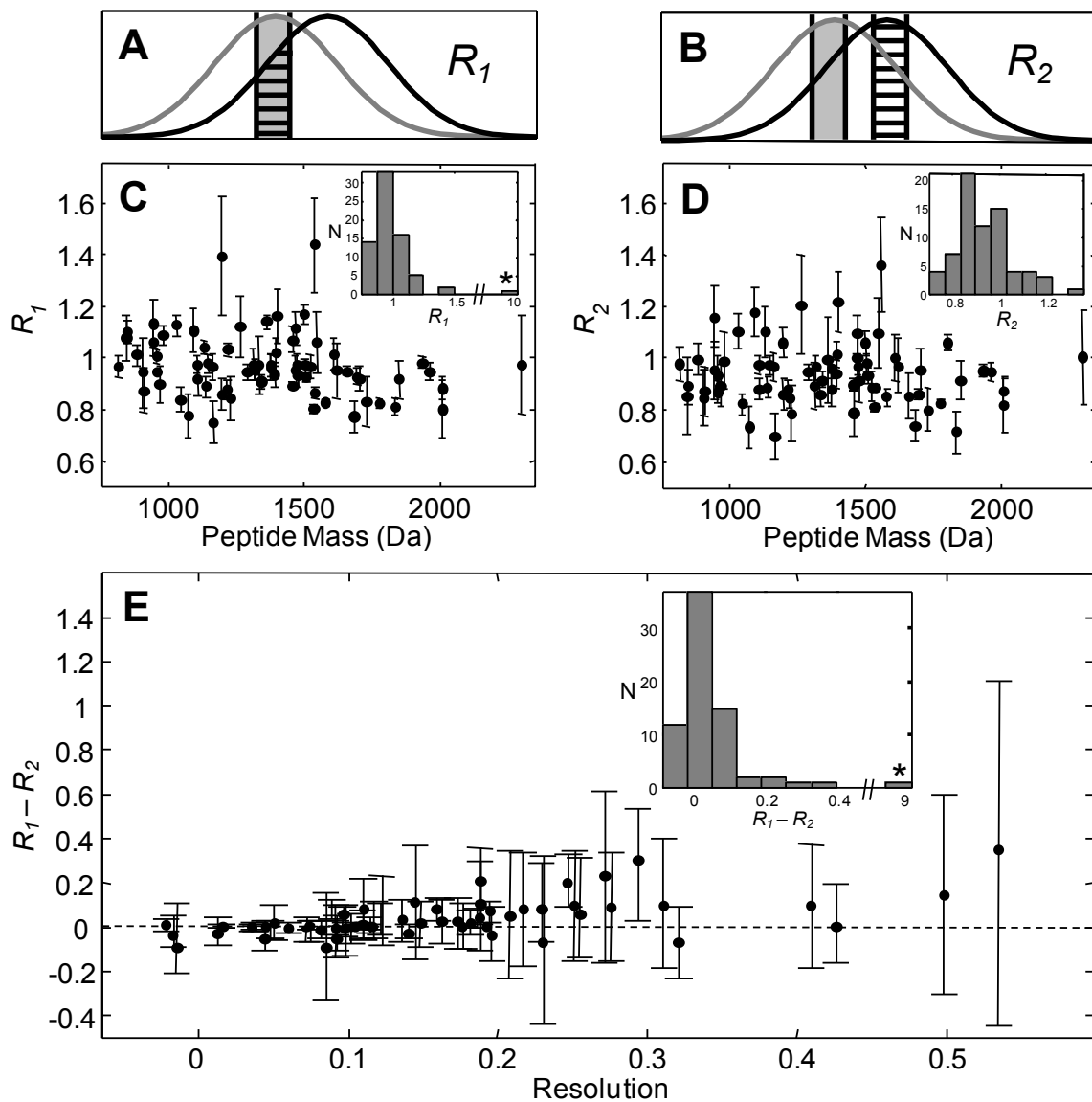
For isotope labelled peptides that do not co-elute, it has been recommended to integrate each peptide signal over their respective elution profiles. The effects of chromatographic separation on the quantitation of peptides following formaldehyde dimethylation are investigated here. Based on experimental design, the theoretical ratio of the heavy to light tagged peptides would be unity. Two approaches were used to calculate the relative ratio of heavy/light peptides, as illustrated in Figures 4.5A and B. The first approach (shown in Figure 4.5A) calculates the intensity ratio ( $R_1$ ) of the two peaks by summing the intensities of the five time channels of the XICs centered on the time channel corresponding to the maximum observed signal for the heavy component. The second approach (Figure 4.5B) calculates the intensity ratio ( $R_2$ ) in the same manner, but centers the intensity measurements at the respective peak maxima for the heavy and light components. (Multiple time channels were included to improve signal averaging, but the window, *ca* 15 s, was sufficiently small to approximate point measurements.) Given the observed separation of isotopes, one would assume that the second ratio calculation would lead to a more accurate determination of the relative peptide quantity. Each ratio corresponds to the area over the indicated region associated with the heavy component (gray shading) divided by the area associated with the light component



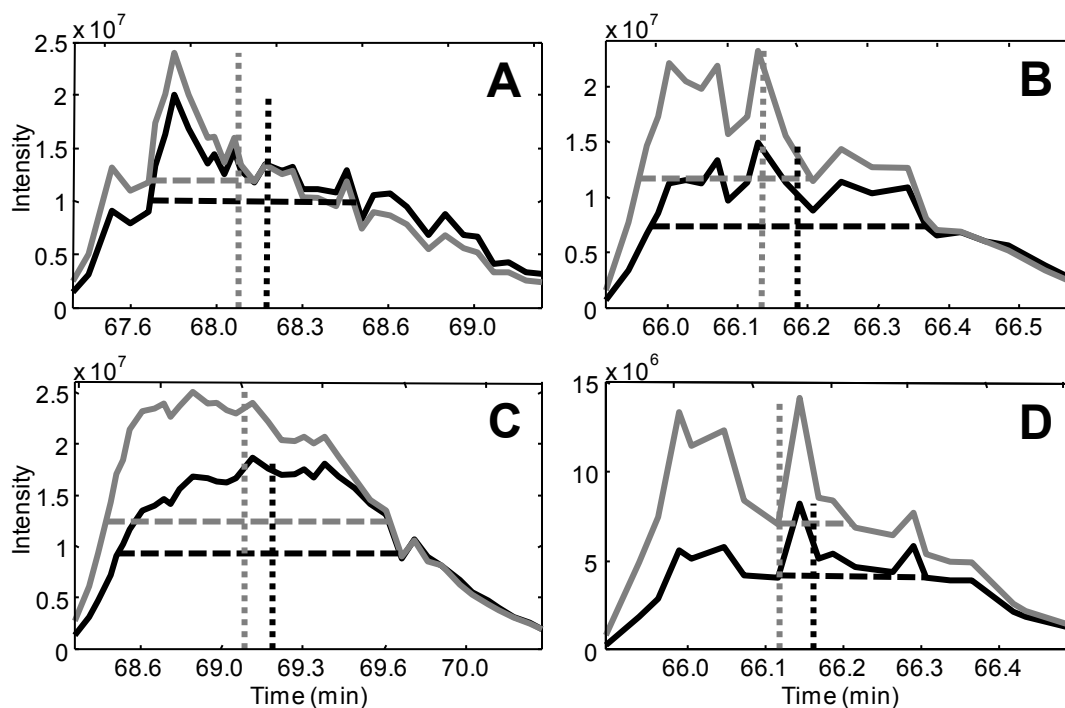
(horizontal lines). The averages for the four replicate ratios for each peptide pair are shown in Figures 4.5C ( $R_1$ ) and 4.5D ( $R_2$ ) as a function of apparent peptide mass. The median values for  $R_1$  and  $R_2$  were 0.95 and 0.92 (the extreme was excluded in this calculation). To examine the change in the ratios, a plot of the difference between  $R_1$  and  $R_2$  for every peptide pair as a function of resolution is also shown in Figure 4.5E. The ratio for the extreme case is not shown in the main plot for Figures 4.5C and 4.5E but is indicated by the asterisk in the inset histograms. In Figure 4.5E, the error bars represent the 95% confidence interval and the dashed line corresponds to the zero ratio difference.

#### **4.4 Discussion**

As previously stated, Hsu et al. [49] indicated that there was no isotopic effect observed when labelling BSA with the dimethyl method. However, this conclusion was established by examining only one particular BSA peptide, QTALVELLK. This peptide was also observed in this work (labelled “e” in Figure 4.2 and 4.4) and the XICs for all replicates of this peptide are shown in Figure 4.6. The figure shows very similar elution patterns for the differentially labelled peptides across all replicates, consistent with the observations of Hsu et al. for this peptide pair. However, the ratio of signal intensities (heavy to light) is consistently higher on the leading edge of the profile, resulting in a relatively small but reproducible difference in median retention times. While a small retention time difference is observed for this particular peptide, the magnitude of the isotopic effect varies substantially depending on the peptide pair, as shown in Figures 4.2 and 4.4. Thus, one cannot generalize co-elution of deuterated peptide pairs from this one peptide.



**Figure 4.5** (A) Strategy for calculation of ratios based on a single time region ( $R_1$ ) and (B) based on two time regions ( $R_2$ ); (C)  $R_1$  ratios calculated for 71 peptides with inset histogram; (D)  $R_2$  ratios with inset histogram; (E) Ratio differences plotted against resolution.



**Figure 4.6** Four replicate XICs for the peptide pair identified as QTALVELLK. The vertical lines correspond to the median retention times and horizontal lines represent the peak widths at half height for the light (black) and heavy (gray) components.

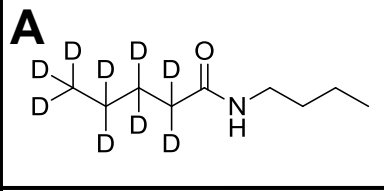
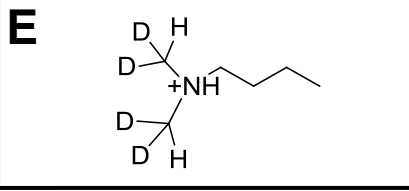
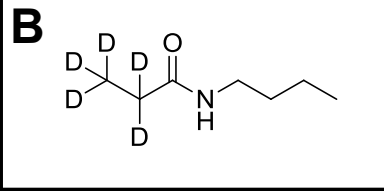
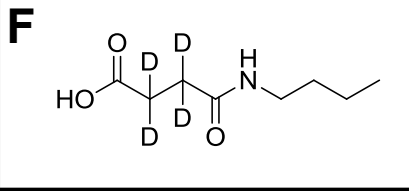
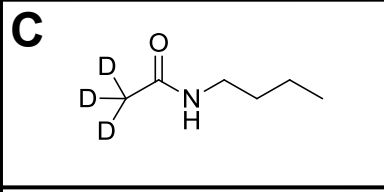
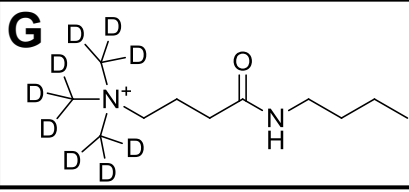
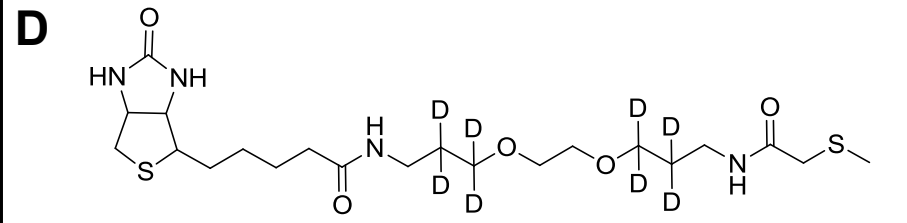
In the context of studies of other isotopic labels, Zhang et al. [54] reported that when BSA was labelled with  $D_0/D_8$  ICAT reagents, 4 of 19 peptides (20%) had a resolution greater than 0.5, and no peptides had a resolution below 0.1. It was also reported that, for peptides labelled with  $^{12}C$  and  $^{13}C$  succinate reagents, no peptides exceeded a resolution of 0.01 (i.e. they effectively co-elute). In the present work, only 3% (2 out of 71) of peptides had a resolution greater than 0.5, and 42% (30 out of 71) of peptides had a resolution below 0.1. No peptides had a resolution below 0.01, although the error bars suggest that such precise measurements are not meaningful. These results suggest that dimethyl labelling consistently produces resolution values smaller than those for deuterated ICAT reagents, but larger than those for the  $^{13}C$  succinate reagents. Zhang et al. [54, 55] reported that there was a negative correlation between the resolution and the peptide mass, wherein a greater isotopic separation was observed for smaller peptides.

This trend was not observed in the current study, owing to the narrower range of peptide masses observed. Low mass peptides (<800 Da) were not observed given the selected mass range of the detector (400 – 1700  $m/z$ ). Such a mass range is consistent with most proteomic investigations.

Zhang et al. [55] have also provided a qualitative rationale for the relative isotopic effects of deuterated labels based on the polarity of the label. Deuterated compounds, in general, are expected to elute earlier because their interactions with the non-polar stationary phase are not as strong as the protonated compounds [112]. It has been shown that chromatographic separation of isotopes is related to the number of deuterium atoms for structurally similar molecules [55]. However, this effect will only be observed if the deuterium atoms are able to interact with the stationary phase. Thus, Zhang et al. argued that the polarity of the label will influence the chromatographic separation, with less polar compounds showing greater separation of isotope pairs. This argument was consistent with a group of deuterated compounds used to for protein labelling, though a quantitative investigation has not been attempted.

To place dimethyl labelling in the context of a quantitative assessment of polarity, computations were carried out in this work to obtain estimates of the polarity of the labels. To achieve this, model compounds were assessed as shown in Figure 4.7, representing the product after reaction of the label with 1-butylamine or, in the case of ICAT, methylthiol. Thus, these compounds mimic the environment of the tag following reaction with a lysine or cysteine group. Figure 4.7 lists the computed dipole moment (described in Section 4.2.3). As suggested by Zhang et al., the fragments with the highest polarity correlate with those that have the smallest isotopic effect. In this group, the

dimethyl label (Figure 4.7E) falls between the ICAT (Figure 4.7D) and succinic anhydride (Figure 4.7F) labels, which is consistent with the experimental results observed for the resolutions obtained. Thus our observations confirm the arguments previously put forward [55].

Structure	Dipole Moment	Structure	Dipole Moment
<b>A</b> 	3.316	<b>E</b> 	5.942
<b>B</b> 	3.402	<b>F</b> 	8.095
<b>C</b> 	3.515	<b>G</b> 	15.423
<b>D</b> 			5.762

**Figure 4.7** Calculated dipole moments (Debye) of labelled surrogates (deuterated version) for: (A) pentanoic acid 2,5-dioxopyrrolidin-1-yl ester, (B) propionic acid 2,5-dioxopyrrolidin-1-y ester, (C) acetic acid 2,5-dioxopyrrolidin-1-y ester, (D) ICAT, (E) dimethyl labelling, (F) succinic anhydride, (G) [3-(2,5-dioxopyrrolidin-1-yloxycarbonyl)-propyl]-trimethylammoniumchloride.

The primary purpose of this work was to examine the extent of retention time shifts in dimethyl labelling and their impact on quantitation. While a time shift was observed between heavy and light labelled peptides, it was very small except for one

anomalous peptide pair. To evaluate the effect on quantitation, intensity ratios were calculated based on time synchronous measurements ( $R_1$ ) and measurements made at chromatographic maxima ( $R_2$ ), with the latter expected to be more reliable. The majority of the differences of the ratios ( $R_1-R_2$ ) are in the positive region (Figure 4.5E), indicating that  $R_1$  is larger than  $R_2$  for most of the peptide pairs, as expected. This positive bias is also indicated by the median ratios reported. While the difference in ratios is significant ( $P = 2 \times 10^{-4}$  by a paired t-test), none of the individual peptides showed a significant difference in the ratio calculation (i.e. the time shift was unimportant) except for the single anomalous case (Figure 4.3A). Because proteomics studies (e.g. biomarker studies) often focus on extreme values, the presence of such anomalies should be a consideration, but retention time shifts are expected to be inconsequential for the large majority of dimethylated peptides.

It has been suggested in literature that  $^{12}\text{C}/^{13}\text{C}$  based methods generate more reliable quantitative results due to smaller isotopic separation over H/D based methods [58]. Although deuterium labelled compounds produce some separation of isotopic peaks, in many cases, as exhibited here, this shift is inconsequential for purposes of quantitation. The majority of  $^{12}\text{C}/^{13}\text{C}$  labelling methods are more expensive and the labelling agent is commonly larger. It has also been reported that larger labelling reagents may interfere with MS/MS sequencing [56]. Therefore, it is felt on the basis of the present study that deuterium based dimethylation should not be discounted as a peptide labelling strategy.

## 4.5 Concluding Remarks

Isotopic effects on retention time were observed when stable isotope dimethyl labelling is employed for comparative proteomic experiments. This effect is small for most peptide pairs observed in the normal operational range, but is peptide dependent and was found to be more substantial in a few cases. If there was a significant separation of differentially labelled peptides, the deuterium labelled component of the peptide pair eluted first. No reproducible trends were observed when retention time difference or resolution was plotted as a function of apparent mass, retention time, or other relevant variables. It was determined that the isotopic effect is smaller than the effect shown when labelling with deuterated ICAT reagents, but larger when labelling with  $^{12}\text{C}/^{13}\text{C}$  succinate reagents. Through computational studies, it was demonstrated that these results are consistent with arguments that the magnitude isotopic effect is inversely related to the polarity of the label. Except for one anomalous peptide, the effect of time shifts resulting from dimethyl labelling on single point quantitation were found to be negligible when compared to measurement uncertainty. It is therefore concluded that it is possible to employ dimethyl labelling to quantitative proteomics without the need for peak integration.

## **Chapter 5**

### **Peptide Pair Detection Algorithm**

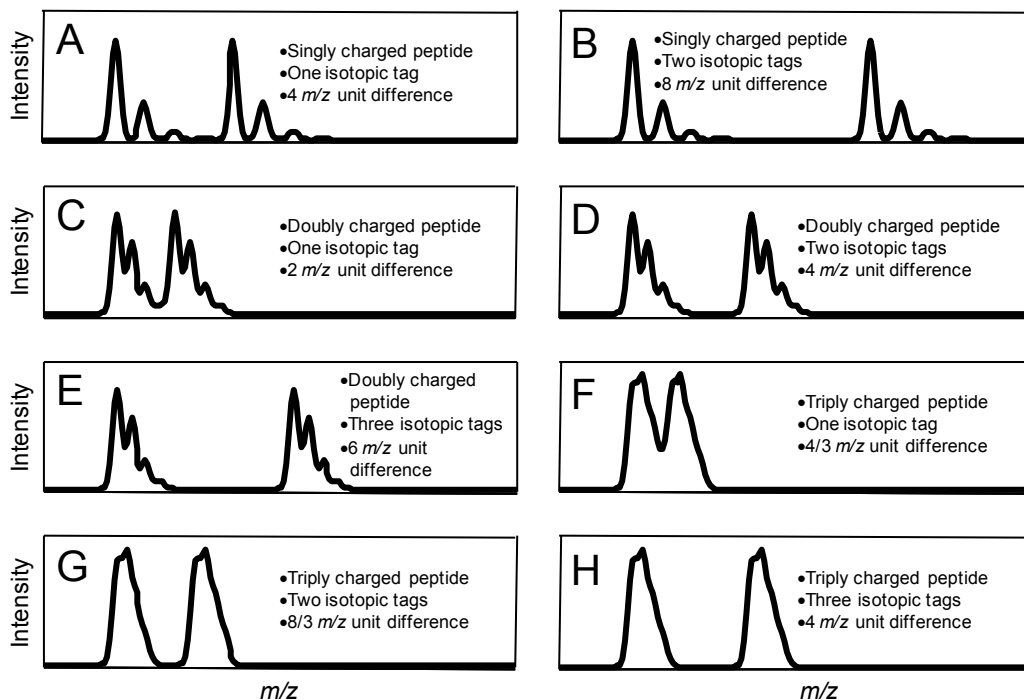
In Chapter 1, the limitations of conventional approaches to quantifying isotopically labelled peptides were introduced. The challenge was presented as being able to detect peptides without the need for MS/MS scans. In this chapter, strategies for meeting this challenge are presented. A number of approaches were investigated, some of which will be only briefly described, followed by a complete description of the method that was actually used for this research. Following this, the application of this method to the BSA and YEAST data sets is described.

#### **5.1 Strategy for Peptide Detection Based on MS Data**

Since the approaches described in this thesis make no assumption of the availability of MS/MS data, the detection of peptides has to be based solely on MS data, which necessitates the exploitation of unique characteristics to distinguish peptides from other species that may be present. The features used in this work are the patterns arising in the mass spectral domain from isotopically labelled peptide pairs and a sustained chromatographic presence. Owing to the nature of the labelling and the charges that exist on the peptides, there are very distinctive patterns observed for the two cell states of interest. As discussed in Chapter 2, the nature of stable-isotope dimethyl labelling can result in a variety of different labels and charges applied to a particular peptide pair. Generally it is assumed that peptides can be singly, doubly, triply or quadruply charged and may have 1, 2, 3 or 4 isotopic labels. However, not all of these combinations are commonly observed. In this work, the approach was to search for isotopic patterns of the



most common cases. The combinations and the typical isotopic patterns that were examined in this work are shown in Figure 5.1. These were expected to be the most common combination of the charge and tags.



**Figure 5.1** Common isotopic patterns for singly charge peptides with one (A) and two tags (B), doubly charge peptides with one (C), two (D) and three tags (E), and triply charged peptides with one (F), two (G) and three tags (H).

The presence of quadruply charged peptides is possible and can be distinguished in the MS/MS data; however, it is virtually impossible to distinguish the isotopic pattern in the MS data. This is a limitation of the mass resolution of the LTQ mass spectrometer. It can't resolve individual isotopic peaks that are separated by less than  $\frac{1}{4}$  mass units and thus produces unclear isotopic patterns. In addition, isotopic patterns for peptides with four isotopic tags were not examined in this work since they contain two missed cleavages of lysine or arginine and are therefore rarely observed.

The inclusion of singly charged peptides with 1 or 2 isotopic tags as targets in this work is a significant departure from traditional MS/MS-based approaches. These approaches do not carry out MS/MS scans on singly charged ions, effectively ignoring their identification. This is done for several reasons. First, the quality of the MS/MS spectra for singly charged peptides tends to be poor because the fragmentation pattern is worse than that of more highly charged peptides, where the liberation of the b- and y-ions is more feasible. Second, compiling an MS/MS scan for each singly charged peptide would diminish the number of MS/MS scans for more highly charged ions without producing an equal amount of useful information because the peptides wouldn't be identified using MS/MS data. Finally in practice it is found that many singly charged peptides have relatively short sequences and therefore are generally less useful in protein identification than the multiply charged ions. Because MS/MS data are not assumed or used in the present work, the first two reasons for ignoring singly charged ions are not relevant. Although it may still be true that singly charged peptides are less useful for identification, they may still have value for confirmation, or for detecting differential responses from proteins that do not yield multiply charged peptides. Nevertheless, singly charged peptides with three labels were not examined here because the mass unit difference of 12 is quite large compared to the other isotopic patterns, giving more likelihood of false positives.

The model representations in Figure 5.1 have essentially four parameters relating to their appearance: (1) the mass separation of the isotope peaks for each peptide, (2) the mass separation of the peptide pairs, (3) the relative heights of isotopic peaks within each peptide, and (4) the width of the isotope peaks. The first two parameters are determined

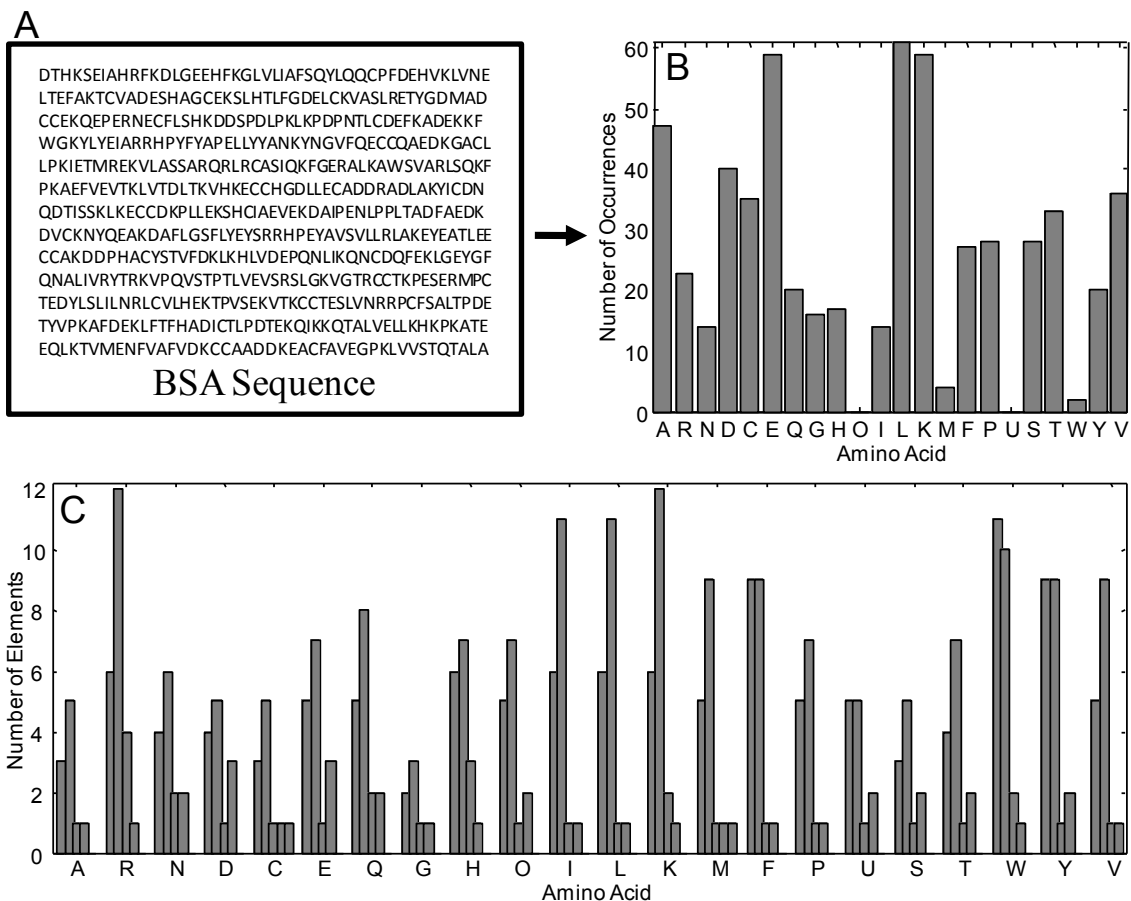
by charge state and number of labels. The last parameter, the peak width, is a function of the mass spectrometer and is assumed to be fixed in this work. The value employed was 0.15  $m/z$  and was estimated by nonlinear fitting of a representative singly charged cluster. The relative height of the isotopic peaks (the third parameter) depends on the peptide elemental composition. These were estimated as described in Section 5.2.

## **5.2 Methods to Predict Isotopic Patterns of Tagged Peptides**

Based on labelling and charge, it is relatively straightforward to predict isotopic patterns that will be observed. The complicating factor is the relative peak heights of the isotopic peaks for each peptide, which depends on the elemental composition of the peptide. Without identification of the peptides, this composition is an unknown quantity. However, we can make the assumption that the average molecular formula of a protein or proteome can be used to estimate the elemental composition of the observed mass of interest. Once the elemental composition is estimated, the isotopic ratios can be calculated using the simple multinomial distribution. The procedures to estimate the elemental composition and calculate the isotopic ratios are described in Sections 5.2.1 and 5.2.2.

### **5.2.1 Estimating Elemental Composition**

As previously stated, a reasonable estimate of the elemental composition can be made using the average molecular formula based on the amino acid distribution calculated from the sequence of a protein or proteome. For example, based on the protein or proteome sequence (Figure 5.2A), BSA in this case, the amino acid distribution is calculated (Figure 5.2B) by finding the total number of each amino acid in the sequence.



**Figure 5.2** The amino acid distribution (B) based on the BSA sequence (A) and the elemental distribution in each amino acid (C). The order of elements for (C) is carbon, hydrogen, oxygen, nitrogen and sulphur.

In generalized terms, the average amino acid composition can be expressed as:

$$C_v H_w N_x O_y S_z \quad (5.1)$$

where  $v$ ,  $w$ ,  $x$ ,  $y$ , and  $z$  are the number of atoms of each element present in the hypothetical amino acid. Knowing the frequency and the number of carbon, hydrogen, nitrogen, oxygen and sulphur atoms in each amino acid (Figure 5.2C), the quantities  $v$ ,  $w$ ,  $x$ ,  $y$  and  $z$  can be calculated. Equation 5.2 shows this calculation for carbon.

$$v = \frac{\sum f_i n_{ci}}{\sum f_i} \quad (5.2)$$

Here the summation is over all 22 amino acids, where  $f_i$  is frequency of amino acid  $i$  and  $n_{ci}$  is the number of carbon atoms in amino acid  $i$ . Similar equations can be written for the other coefficients. For example, for BSA, the formula for the average amino acid is  $C_{5.03}H_{7.91}N_{1.33}O_{1.53}S_{0.0669}$ . Once the molecular formula and the molecular weight for the average amino acid are calculated, the elemental composition for a given peptide is estimated based on the observed molecular weight by Equation 5.3. This is done by taking the ratio of the observed molecular weight to that of the average amino acid and multiplying it by the coefficients of the average amino acid. The values for the average amino acid for BSA are inserted into Equation 5.3.

$$C_{v_e} H_{w_e} N_{x_e} O_{y_e} S_{z_e} = C_{5.03\omega} H_{7.91\omega} N_{1.33\omega} O_{1.53\omega} S_{0.0669\omega} \quad (5.3)$$

$$\text{where } \omega = \left( \frac{MW_o Z}{113.56} \right) \quad (5.4)$$

In this equation,  $MW_o$  is the observed molecular weight,  $Z$  is the charge and  $v_e$ ,  $w_e$ ,  $x_e$ ,  $y_e$  and  $z_e$  are the coefficients for the estimated elemental composition of the peptide. For example, a doubly charged BSA peptide observed at 600  $m/z$  has an elemental composition of  $C_{53}H_{83}N_{14}O_{16}S_1$  calculated from Equation 5.3. This same procedure can also be applied to a proteome.

Obviously Equation 5.3 will have some limitations, particularly for small peptides, because the statistical realizations of the proportions of atoms will deviate from the average. However, the impact of this on isotopic ratios should be relatively small as long as the peptide mass is reasonably large. In the absence of an exact solution this was deemed the optimal approach to use. The next section describes how to implement the elemental compositions to calculate the isotopic ratios given an observed peptide mass.

## 5.2.2 Calculating Isotopic Ratios

Once the elemental composition is estimated from the observed molecular weight, the isotopic ratios for the M+1 ion ( $R_{M+1}$ ), M+2 ion ( $R_{M+2}$ ) and M+3 ion ( $R_{M+3}$ ) relative to the molecular ion are calculated. In this work, the molecular ion (M) corresponds to  $[M+nH]^{n+}$ , the M+1 ion corresponds to  $[M+1+nH]^{n+}$ , the M+2 ion corresponds to  $[M+2+nH]^{n+}$  and the M+3 ion corresponds to  $[M+3+nH]^{n+}$ . Where n represents the charge of the peptide. The ratios for these ions are calculated by taking the sum of the all probabilities of all isotopic compositions leading to the formation of an M+l ion divided by the probability of forming the molecular ion.

$$R_{M+l} = \frac{h_{M+l}}{h_M} = \frac{P(M+l)}{P(M)} = \frac{\sum P(X_k^l)}{P(X_0)} \quad (5.5)$$

Here  $h_{M+l}$  is the height of the peak for the M+l ion,  $h_M$  is the height of the peak for the molecular ion,  $P(M+l)$  is the overall probability of forming a M+l ion,  $P(M)$  is the probability of forming the M ion,  $P(X_k^l)$  is the probability of forming the M+l ion by isotopic combination  $k$ , and  $P(X_0)$  (=  $P(M)$ ) is the probability of forming the lowest isotope combination (i.e. the molecular ion). The quantities  $P(X_k^l)$  and  $P(X_0)$  are calculated based on the multinomial distribution. For  $P(X_k^l)$  this is given by:

$$P(X_k^l) = \frac{n_1! n_2! n_3! \dots}{n_{11}! n_{12}! \dots n_{21}! n_{22}! \dots} p_{11}^{n_{11}} p_{12}^{n_{12}} \dots p_{21}^{n_{21}} p_{22}^{n_{22}} \dots \quad (5.6)$$

where  $n_i$  is the number of atoms of element  $i$  in ion X,  $n_{ij}$  is the number of atoms of isotope  $j$  of element  $i$  in ion X with mass M+l and  $p_{ij}$  is the fractional abundance of

isotope  $j$  of element  $i$ . For  $P(X_0)$ , a simplified version of  $P(X_k^\ell)$  is to calculate the probability of the molecular ion in which  $n_{ij} = 0$  for  $j > 1$ .

$$P(X_0) = \frac{n_1!n_2!n_3!\dots}{n_{11}!n_{21}!\dots} p_{11}^{n_{11}} p_{21}^{n_{21}} \dots \quad (5.7)$$

Note that these calculations do not adjust the molecular formula to account for the presence of labels or other modifications, but these variations are expected to be of negligible consequence, especially given that an average formula is already used in the calculation.

For example, the isotopic ratios for the  $M+\ell$  ions ( $R_{M+1}$ ,  $R_{M+2}$ ,  $R_{M+3}$ ) for a doubly charged BSA peptide observed at 600  $m/z$  with an elemental composition of  $C_{53}H_{83}N_{14}O_{16}S_1$  are calculated from Equation 5.5. The probability of forming the lowest isotope combination (i.e. the molecular ion) is calculated by Equation 5.7.

$$\begin{aligned} P(X_0) &= P(^{12}C_{53} \ ^1H_{83} \ ^{14}N_{14} \ ^{16}O_{16} \ ^{32}S_1) \\ &= \frac{53!83!14!16!1!}{53!83!14!16!1!} (0.9889)^{53} (0.99985)^{83} (0.9964)^{14} (0.9976)^{16} (0.9500)^1 \\ &= 0.4751 \end{aligned}$$

There are multiple  $M+1$  ions that can be formed. These include ions that contain only  $^{13}C$ ,  $^2H$ ,  $^{15}N$ ,  $^{17}O$  or  $^{33}S$ . For the  $^{13}C$  case,  $P(X_1^1)$ , is calculated from Equation 5.6.

$$\begin{aligned} P(X_1^1) &= P(^{12}C_{52} \ ^{13}C_1 \ ^1H_{83} \ ^{14}N_{14} \ ^{16}O_{16} \ ^{32}S_1) \\ &= \frac{53!83!14!16!1!}{52!1!83!14!16!1!} (0.9889)^{52} (0.0111)^1 (0.99985)^{83} (0.9964)^{14} (0.9976)^{16} (0.9500)^1 \\ &= 0.2826 \end{aligned}$$

Likewise the probabilities can be calculated for the other four possibilities and finally  $R_{M+1}$  is calculated.

$$R_{M+1} = \frac{P(M+1)}{P(M)} = \frac{(0.2826 + 0.0059 + 0.0240 + 0.0030 + 0.0038)}{0.4751} = \frac{0.3194}{0.4751} = 0.672$$

$R_{M+2}$  and  $R_{M+3}$  can be calculated in the same way with different combinations, however the results for  $P(X_k^\ell)$  are not shown since the number of combinations increases to 17 for the X+2 ion and 45 for the X+3 ion.

$$R_{M+2} = \frac{\sum P(M+2)}{P(M)} = \frac{0.1421}{0.4751} = 0.299$$

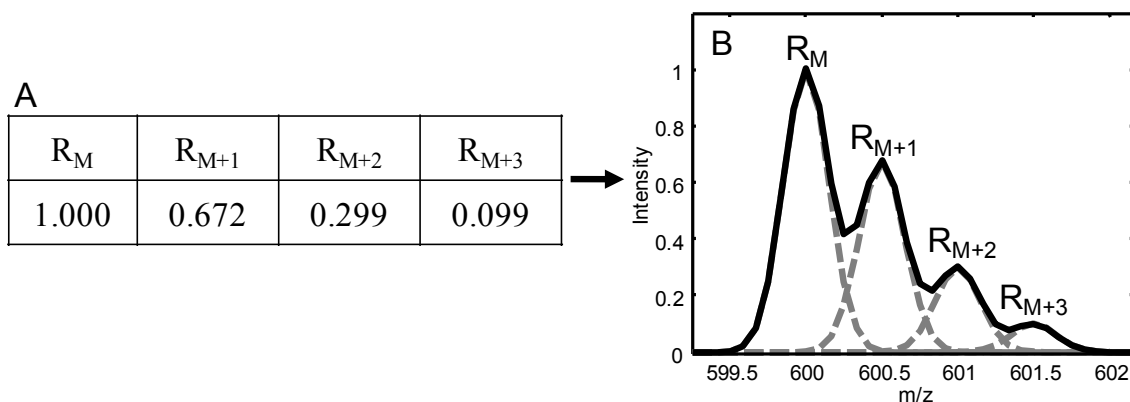
$$R_{M+3} = \frac{\sum P(M+3)}{P(M)} = \frac{0.0047}{0.4751} = 0.099$$

Once these ratios are calculated they are included in the generation of a mixed gaussian model for the isotopic pattern with the ratios for the M+l ions. The isotopic pattern is calculated using Equation 5.8 which is based on a gaussian distribution.

$$\begin{aligned} f(x) = & R_M e^{-(x-X_M)^2/2\sigma^2} \\ & + R_{M+1} e^{-(x-X_M-\Delta x)^2/2\sigma^2} \\ & + R_{M+2} e^{-(x-X_M-2\Delta x)^2/2\sigma^2} \\ & + R_{M+3} e^{-(x-X_M-3\Delta x)^2/2\sigma^2} \end{aligned} \quad (5.8)$$

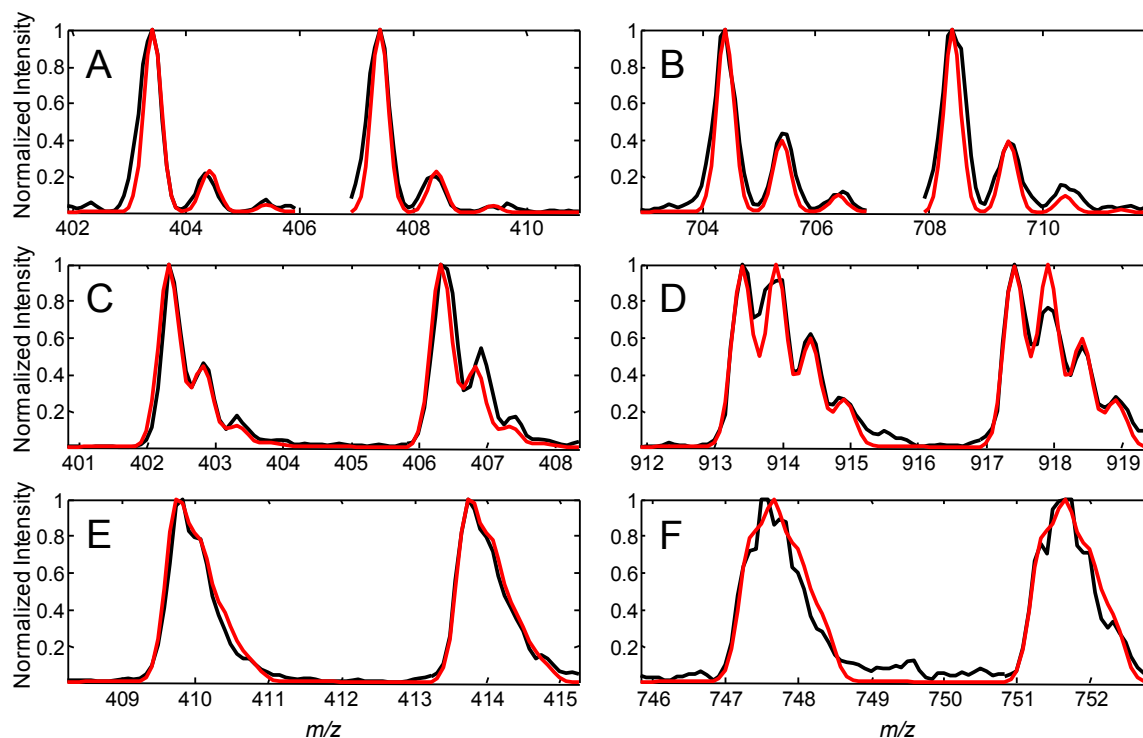
Here,  $x$  represents the mass-to-charge ratio on the x-axis,  $X_M$  is the position of the M ion in the pattern,  $\sigma^2$  is the variance of the distribution which is set to 0.15 and  $\Delta x$  is the isotopic separation of the peptide in question (in this case,  $\Delta x = 0.5$  for doubly charged). The calculated isotopic ratios (Figure 5.3A) and the isotopic pattern (Figure 5.3B – black) for the doubly charged BSA peptide observed at 600  $m/z$  with an elemental composition of  $C_{53}H_{83}N_{14}O_{16}S_1$  are shown in Figure 5.3. The gray dashed lines in Figure 5.3B represent the individual gaussian distributions that make up the total isotopic pattern (black).





**Figure 5.3** The calculated isotopic ratios (A) and the isotopic pattern (B) for a BSA peptide detected at 600  $m/z$  with an elemental composition of  $C_{53}H_{83}N_{14}O_{16}S_1$ .

To illustrate the accuracy of the calculated isotopic patterns, Figure 5.4 was created to compare experimental mass spectra (black) of different BSA peptide pair combinations (charge and number of isotopic tags) to the calculated isotopic patterns (red). Each experimental mass spectrum was normalized to the maximum intensity in the selected mass-to-charge region. The first row corresponds to a singly charged BSA peptide with one isotopic tag at a low (Figure 5.4A) and high (Figure 5.4B) mass-to-charge ratio, the second row corresponds to a doubly charged BSA peptide with two isotopic tags at a low (Figure 5.4C) and high (Figure 5.4D) mass-to-charge and the third row corresponds to a triply charged BSA peptide with three isotopic tags at a low (Figure 5.4E) and high (Figure 5.4F) mass-to-charge. In addition to changes in the spacing of isotopic peaks with different charge states, there is a distinct change in the ion ratios between low and high masses. Taking these isotopic pattern differences into account, the calculated patterns model the experimental mass spectra very well for all different types of peptide combinations and mass-to-charge ratios. This procedure to estimate the elemental composition and calculate the isotopic ratios was used for the remainder of this work.



**Figure 5.4** Selected mass spectral regions (black) compared to calculated isotopic profiles (red) for a singly charged BSA peptides with one tag (low (A) and high (B) mass), doubly charged peptides with two tags (low (C) and high (D) mass) and triply charged peptides with three tags (low (E) and high (F) mass).

The next section will outline the development of an algorithm to detect isotopically labelled peptide pairs using the calculated isotopic patterns described in this section.

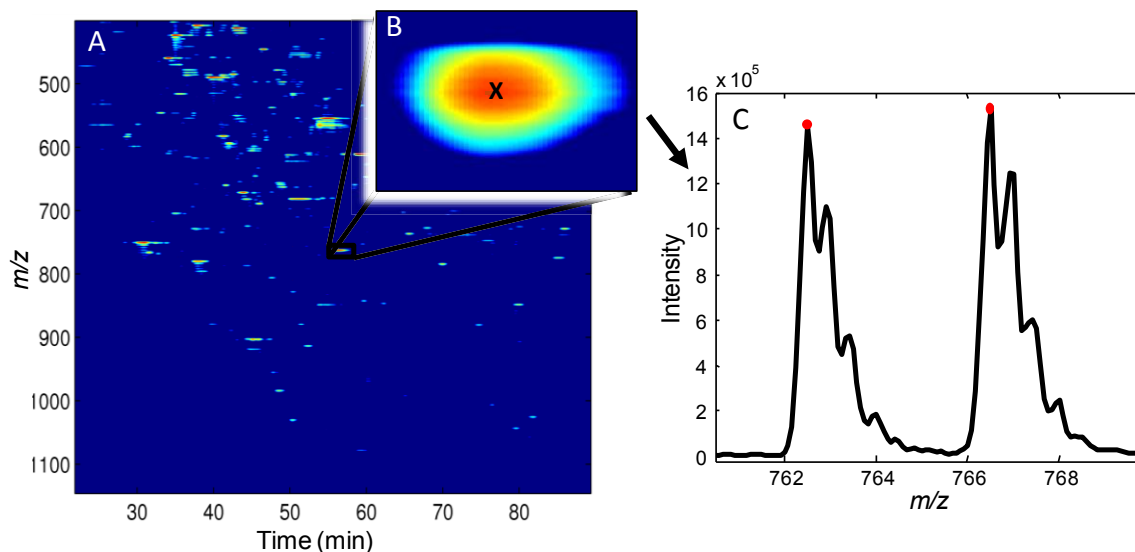
### 5.3 Algorithms for Peptide Detection

Once the isotopic patterns were established, it was necessary to find a method to detect where they arise in the mass chromatogram. A number of strategies were employed for this, some unsuccessfully. One of these will be described briefly, followed by a more extensive description of the algorithm used successfully in this work.

### 5.3.1 Autocorrelation Function

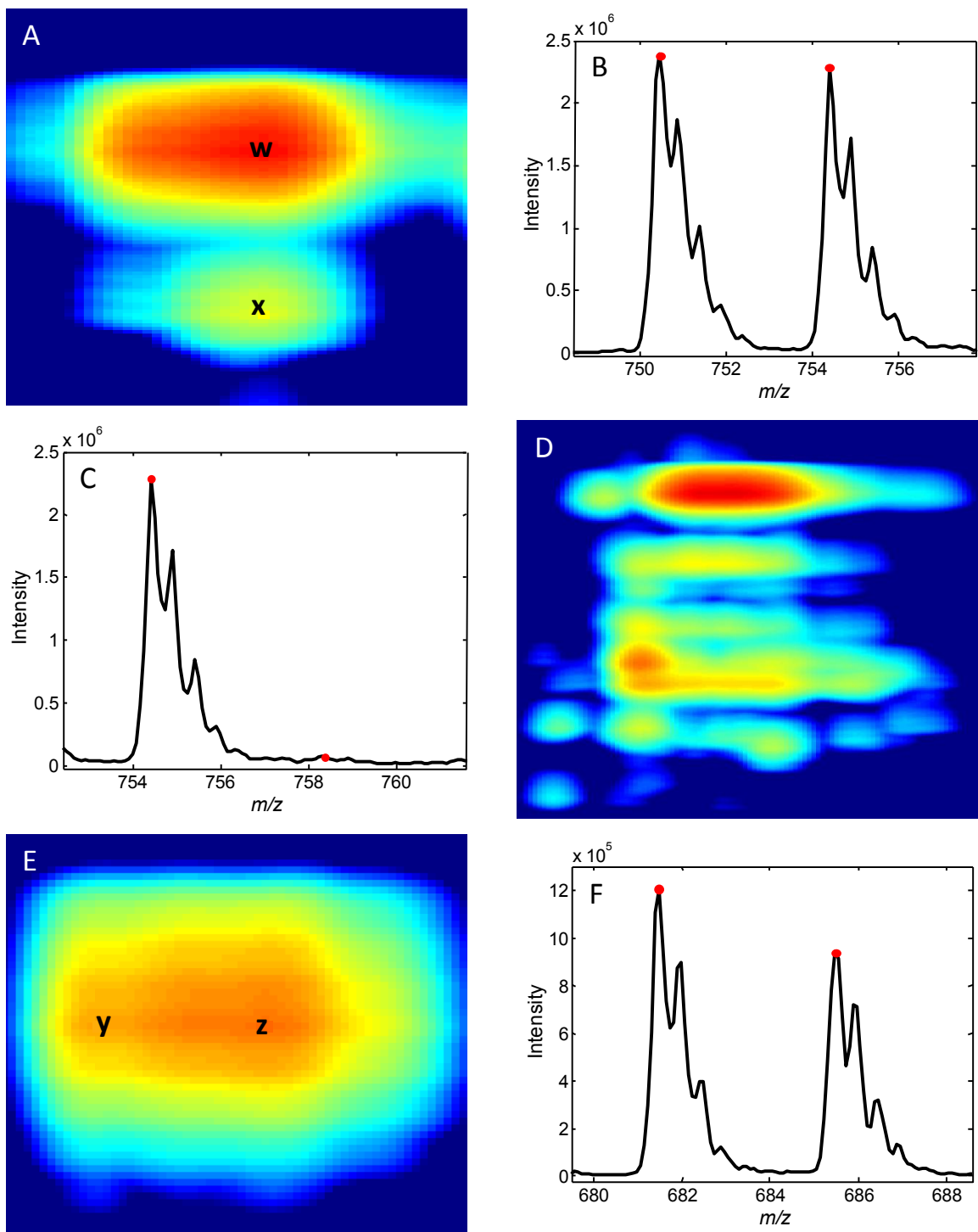
The first strategy that was employed to detect isotopically labelled peptide pairs through specific patterns in a mass chromatogram was a novel two dimensional weighted autocorrelation function. The weighting for the function in the mass domain was the predicted isotopic patterns presented in Section 5.2.2, while in the time domain an estimated Gaussian distribution based on the peptide chromatography was used. As the autocorrelation function is passed over sections of the mass chromatogram for BSA (Figure 4.1A) a correlation value is calculated for each mass-to-charge channel and scan number. This correlation value corresponds to how well the function models the extracted section from the mass chromatogram. As the function and the extracted section from the mass chromatogram become similar the correlation increases and the correlation decreases when the two become different. Each isotopically labelled peptide pair creates a correlation spot where the maximum correlation corresponds to the location of the peptide pair. The correlation values can be visualized using a logarithmic heat map (Figure 5.5A) where a correlation spot on the map (Figure 5.5B) corresponds to a labelled peptide pair (Figure 5.5C). In the heat map, the colourbar gradually goes from blue (lower correlation) to red (higher correlation).

This strategy was successful in locating labelled peptide pairs as shown in Figure 5.6A and B. Figure 5.6B represents a mass spectrum of a found peptide pair for the correlation spot indicated on Figure 5.6A at “w”. However, after an extensive investigation on the reliability of each correlation spot (Figure 5.5B) a number of problems were discovered. First, close to a half of the correlation spots were false-positives due to improper detection of peptide pairs. For example, Figure 5.6C shows a



**Figure 5.5** The logarithmic heat map of a BSA sample (A) with an example of a correlation spot for a pair of differently labelled peptides (B) and the mass spectrum of a peptide pair (C) indicated in (B). The red dots in (C) corresponds to the masses of the light and heavy components in a peptide pair that produce the correlation spot indicated in (B).

mass spectrum of a false-positive which corresponds to the correlation spot indicated on Figure 5.6A at “x”. The majority of the false-positives spots were due to prominent heavy peptide peaks of legitimate pairs (Figure 5.6B) correlating with the baseline (Figure 5.6C) producing lower correlation spots as shown in Figure 5.6A at “x”. These correlations spots are very problematic because the algorithm should have little to no false-positives. Second, when sections of a mass chromatogram are complex (a lot of peptides eluting at once), the correlation spots tend to interfere with one another and it becomes difficult to distinguish the boundaries of individual correlation spots (Figure 5.6D). Third, several spots produce two maximum correlation locations (Figure 5.6E at “y” and “z”) where each location corresponds to the same mass spectral data for a peptide pair (Figure 5.6F). This problem produces redundant and unwanted information when compiling a list of peptide pairs found. This type of spot would allow a peptide to be counted twice and distinguished as two different peptide pairs. These problems are the



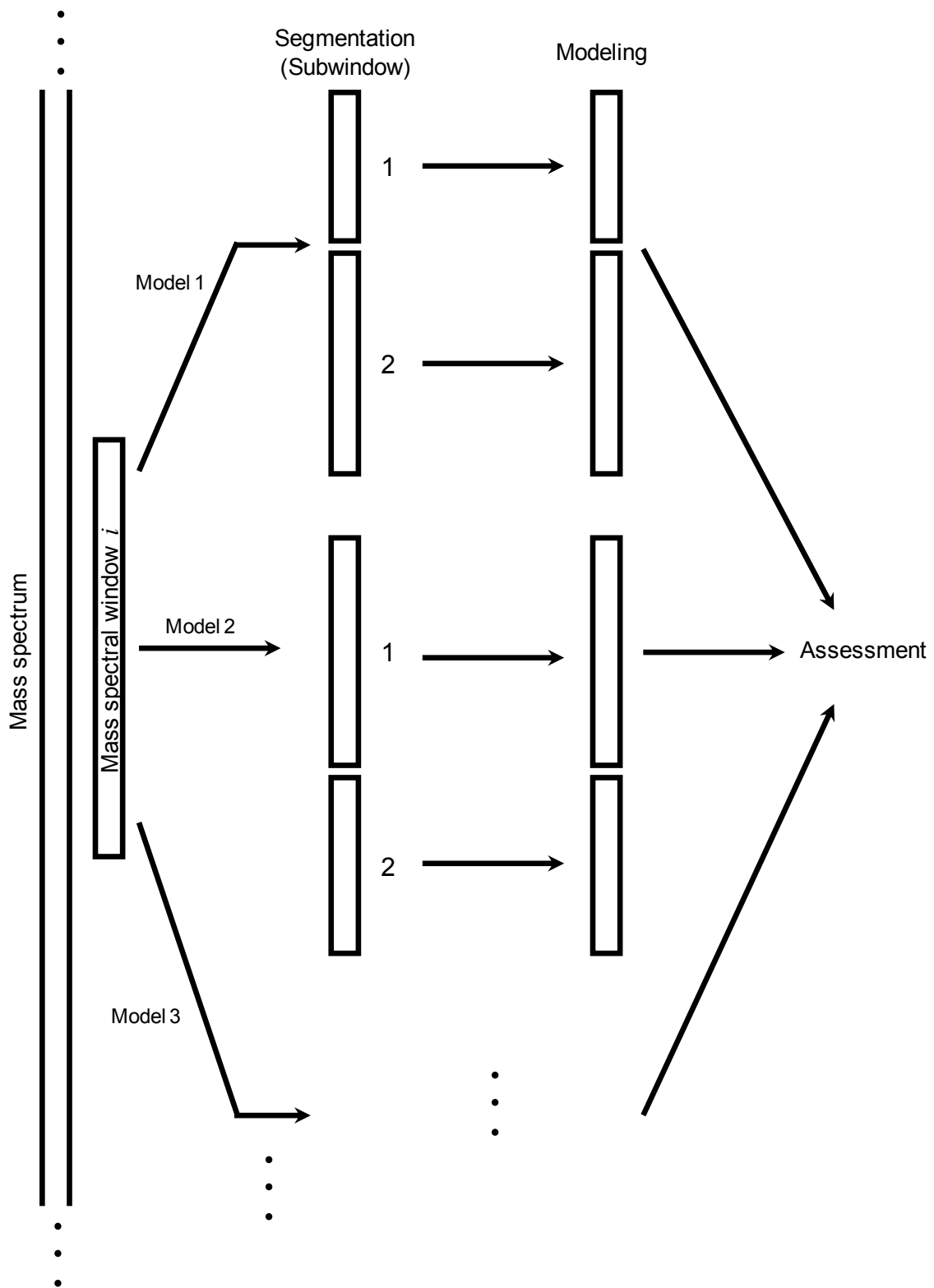
**Figure 5.6** Two selected BSA correlation spots (A & E) with the corresponding mass spectra (B, C & F) at the indicated locations “w” and “x” on (A) and “y” and “z” on (E). (D) shows a region of interfering correlation spots.

reasons why this strategy was not used in this work. This directed the research towards more comprehensive statistical methods that would allow for better control of the parameters used in the algorithm. The next sections will explain the details of the algorithm that was used in this thesis to detect peptide pairs.

### 5.3.2 Parallel Isotopic Tag Screening

The most promising method for the detection of peptide pairs was based on a parallel implementation of classical least squares modeling. In this approach, which is outlined in Figure 5.7, mass spectral data within a designated window of the mass spectrum are fit to calculated isotopic patterns to assess whether or not a peptide pair is present. In the current work, there are eight models employed, one for each of the label/charge combinations commonly observed (Figure 5.1). This approach will be referred to as the Parallel Isotopic Tag Screening (PITS) method.

The PITS method begins by segmenting an MS window  $i$  into two subwindows best suited to encompass the light ( $k = 1$ ) and heavy ( $k = 2$ ) clusters for each model. In each case, the windows chosen differ according to the number of labels, the charge state, and whether the light or heavy ion cluster is being modeled. The range of each subwindow within the mass spectral window is shown in Table 5.1 for the eight models used. These are represented both in terms of the index within the window and relative to the molecular ion,  $M$ , of the light component of the pair. Modeling assumes that this ion is located at index 19 of mass spectral window  $i$ , and that  $\Delta(m/z) = 1/12$  on the y axis.



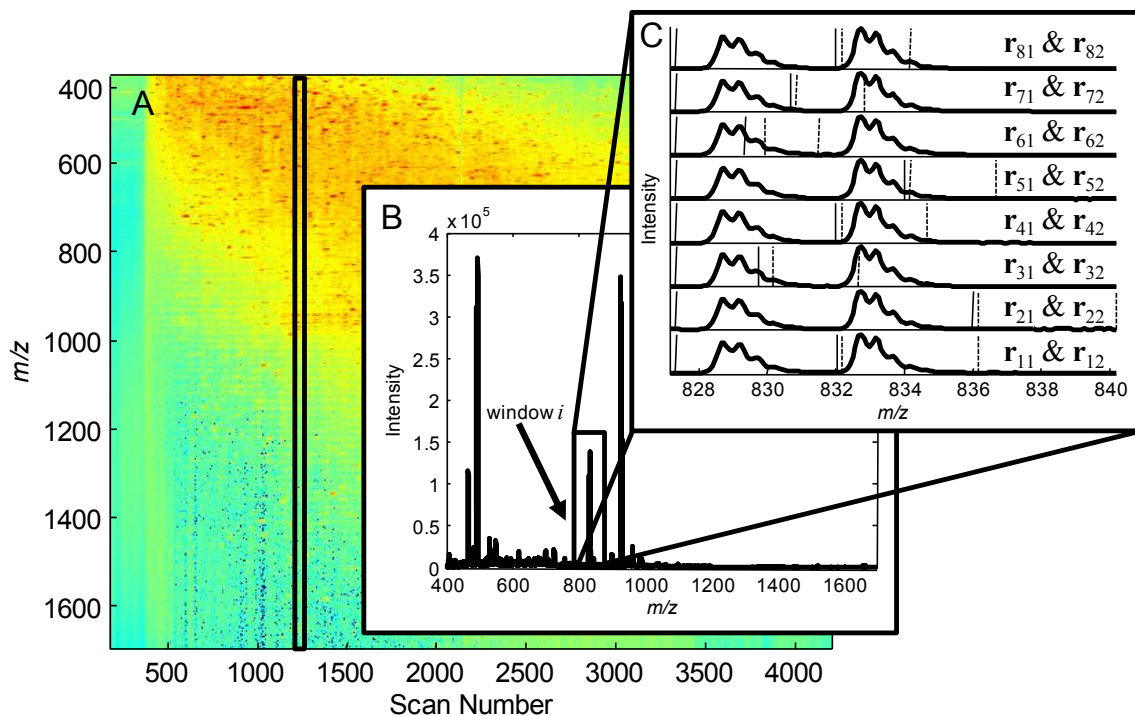
**Figure 5.7** Visual representation of the Parallel Isotopic Tag Screening (PITS) algorithm.

**Table 5.1** Subwindow ranges for the PITS algorithm.

Model	No. of Labels	Charge	Subwindow 1 Range, $r_{j1}$		Subwindow 2 Range $r_{j2}$	
			Index	Mass	Index	Mass
1	1	1	1-61	M - 1.5	61-109	M + 3.5
				M + 3.5		M + 7.5
2	2	1	1-109	M - 1.5	109-157	M + 7.5
				M + 7.5		M + 11.5
3	1	2	1-34	M - 1.5	37-67	M + 1.5
				M + 1.25		M + 4.0
4	2	2	1-61	M - 1.5	61-91	M + 3.5
				M + 3.5		M + 6
5	3	2	1-85	M - 1.5	85-115	M + 5.5
				M + 5.5		M + 8.0
6	1	3	1-29	M - 1.5	34-53	M + 1.25
				M + 0.833		M + 2.833
7	2	3	1-45	M - 1.5	45-69	M + 2.167
				M + 2.16		M + 4.167
8	3	3	1-61	M - 1.5	61-85	M + 3.5
				M+3.5		M + 5.5

Figure 5.8C is a pictorial representation of these subwindows for a selected mass spectral window from the yeast mass chromatogram. For reference, a region of the mass spectrum was chosen that shows a doubly charged, doubly labelled peptide pair. The vertical lines for each model are the subwindow boundaries, with the solid lines representing the first subwindow and the dashed lines representing the second. Note that the size of the subwindows varies both with the model and between the light and heavy labels. Subwindow sizes were chosen to optimally consider aspects such as the inclusion of baseline regions, overlap of the patterns and the number of isotopic peaks included.





**Figure 5.8** A two-dimensional representation of the mass chromatogram for isotopically labeled yeast (A), one particular mass spectrum (B) indicated on (A) and an expanded mass spectrum (C) of section  $i$  showing  $r_{j1}$  and  $r_{j2}$  respectively for each combination.

For each of the eight isotopic patterns tested, each subwindow is modeled independently using the isotopic profiles calculated as described earlier. The calculated profile depends on the charge, the peptide mass and the assumed position of the molecular ion within the subwindow. Also included in the fit is a baseline parameter to account for offsets in the mass spectrum. Mathematically, the model can be represented as:

$$\begin{bmatrix} \mathbf{r}_{ijk} \end{bmatrix} = \begin{bmatrix} b_{ijk} & c_{ijk} \end{bmatrix} \begin{bmatrix} \mathbf{1} \\ \mathbf{f}_{ijk} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_{ijk} \end{bmatrix} \quad (5.9)$$

Here  $\mathbf{r}_{ijk}$  is the  $1 \times n_{jk}$  row vector for mass spectral measurements corresponding to subwindow  $k$  ( $k = 1$  or  $2$ ) of model  $j$  ( $j = 1 \dots 8$ ) for the spectral window  $i$ , where  $n_{jk}$  is the size of subwindow  $k$  for model  $j$  (see Table 5.1). The scalar quantity  $b_{ijk}$  represents the

baseline offset and  $c_{ijk}$  represents the magnitude of the contribution from  $\mathbf{f}_{ijk}$  to estimate  $\mathbf{r}_{ijk}$ . This magnitude is arbitrarily scaled to the  $1 \times n_{jk}$  vector  $\mathbf{f}_{ijk}$ , which is the calculated isotopic pattern for subwindow  $k$  of model  $j$ , corresponding to the masses in window  $i$ . For convenience, the  $\mathbf{f}_{ijk}$  vectors are normalized to a maximum height of unity and should be essentially the same for both subwindows of a given model except for a shift on the mass axis and possibly a very small modification of isotopic ratios. The boldface “ $\mathbf{1}$ ” in Equation 5.9 represents a  $1 \times n_{jk}$  vector of ones, corresponding to the basis function for the baseline. The  $1 \times n_{jk}$  vector  $\mathbf{e}_{ijk}$  is the vector of residuals from the model. Represented in matrix notation, we can rewrite Equation 5.9 as

$$\mathbf{r}_{ijk} = \mathbf{a}_{ijk} \mathbf{F}_{ijk} + \mathbf{e}_{ijk} \quad (5.10)$$

where  $\mathbf{r}_{ijk}$  is  $1 \times n_{jk}$ ,  $\mathbf{a}_{ijk}$  is a  $1 \times 2$  vector of model parameters, and  $\mathbf{F}_{ijk}$  is the  $2 \times n_{jk}$  matrix of basis functions. The least squares solution for the fitted model parameters is

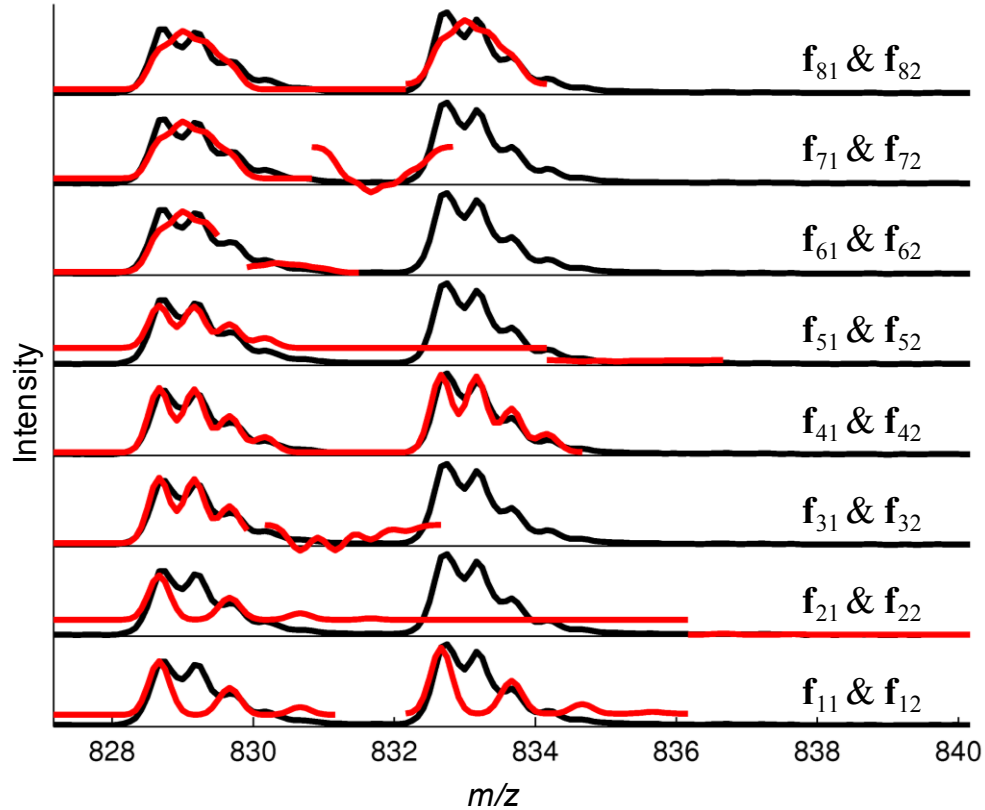
$$\hat{\mathbf{a}}_{ijk} = \mathbf{r}_{ijk} \mathbf{F}_{ijk}^T (\mathbf{F}_{ijk} \mathbf{F}_{ijk}^T)^{-1} \quad (5.11)$$

and the calculated mass spectral response in the subwindow is given by

$$\hat{\mathbf{r}}_{jk} = \hat{\mathbf{a}}_{ijk} \mathbf{F}_{ijk} \quad (5.12)$$

In the modeling step of the PITS algorithm, the fitted parameters and calculated mass spectrum are calculated for each subwindow in each model. To further illustrate the modeling step, Figure 5.9 was created to show the mass spectral response for window  $i$  ( $\mathbf{r}_{jk}$ ) in black plotted along with the fitted mass spectral responses ( $\hat{\mathbf{r}}_{jk}$ ) in red for each subwindow and model for the peptide pair in Figure 5.8C. The discontinuity in the red lines corresponds to the  $\hat{\mathbf{r}}_{ij1}$  calculated with  $\mathbf{f}_{ij1}$  for the first section and  $\hat{\mathbf{r}}_{ij2}$  with  $\mathbf{f}_{ij2}$  for the second section (note that the window lengths are not generally the same size). Figure 5.9

shows that the best  $\hat{\mathbf{r}}_{ijk}$  fit is obtained for the  $\mathbf{f}_{ijk}$  with the correct charge and label combination (+2 with two labels), represented by  $\mathbf{f}_{41}$  and  $\mathbf{f}_{42}$  in this case.



**Figure 5.9** The observed mass spectral response (black) for a particular peptide pair compared to the predicted mass spectral response (red) calculated for both  $k$  segments with the respective  $\mathbf{f}_{j1}$  and  $\mathbf{f}_{j2}$ .

The final step in each iteration of the PITS algorithm is the model assessment, which attempts to determine which, if any, of the tested models matches the mass spectral response in window  $i$ . Most of the time, it is expected that none of the models will accurately represent the data in the window. This can occur for a number of reasons, including (1) there is no legitimate peptide pair in the window, (2) a peptide pair is present, but is not aligned with the model at this iteration of the algorithm, (3) a peptide pair is present, but exists in a state not included in the model set, (4) a peptide pair is

present, but is subject to mass spectral interferences and (5) one of the subwindows fits well for one of the models, but the other subwindow does not. This last situation will arise inevitably as the window moves through the mass spectrum and each isotopic cluster traverses the window. To detect a valid model pair, the PITS algorithm calculates a standard error  $(s_e)_{ijk}$ , for each subwindow and model at each iteration  $i$ .

The standard error is calculated by taking the square root of the sum of the squared residuals divided by the length of  $\mathbf{r}_{ijk}$  ( $n_{jk}$ ) corrected for the loss of degrees of freedom. The standard error indicates the standard deviation in the residuals  $(\mathbf{r}_{ijk} - \hat{\mathbf{r}}_{ijk})$  and estimates the measurement uncertainty when the model is valid. When the standard error is small, the fit of  $\hat{\mathbf{r}}_{ijk}$  to  $\mathbf{r}_{ijk}$  is good and when it is large, the fit is poor.

$$(s_e)_{ijk} = \sqrt{\frac{(\mathbf{r}_{ijk} - \hat{\mathbf{r}}_{ijk})(\mathbf{r}_{ijk} - \hat{\mathbf{r}}_{ijk})^T}{n_{jk} - 2}} \quad (5.13)$$

The standard error is used to assess whether the model fits are reliable. To be acceptable, both subwindows of a model must show good fits. For this purpose, the relative standard is calculated using the observed maximum within each window.

$$Q_{ij} = \max\left(\frac{(s_e)_{ij1}}{\max(r_{ij1})}, \frac{(s_e)_{ij2}}{\max(r_{ij2})}\right) \quad (5.14)$$

Here,  $Q_{ij}$  is the quality measure for the fit of model  $j$  at window  $i$ . If the value of  $Q_{ij}$  falls below some threshold, the model is deemed to be acceptable, otherwise it is not. In this work the threshold was set to 0.10 (10% minimum relative standard error).

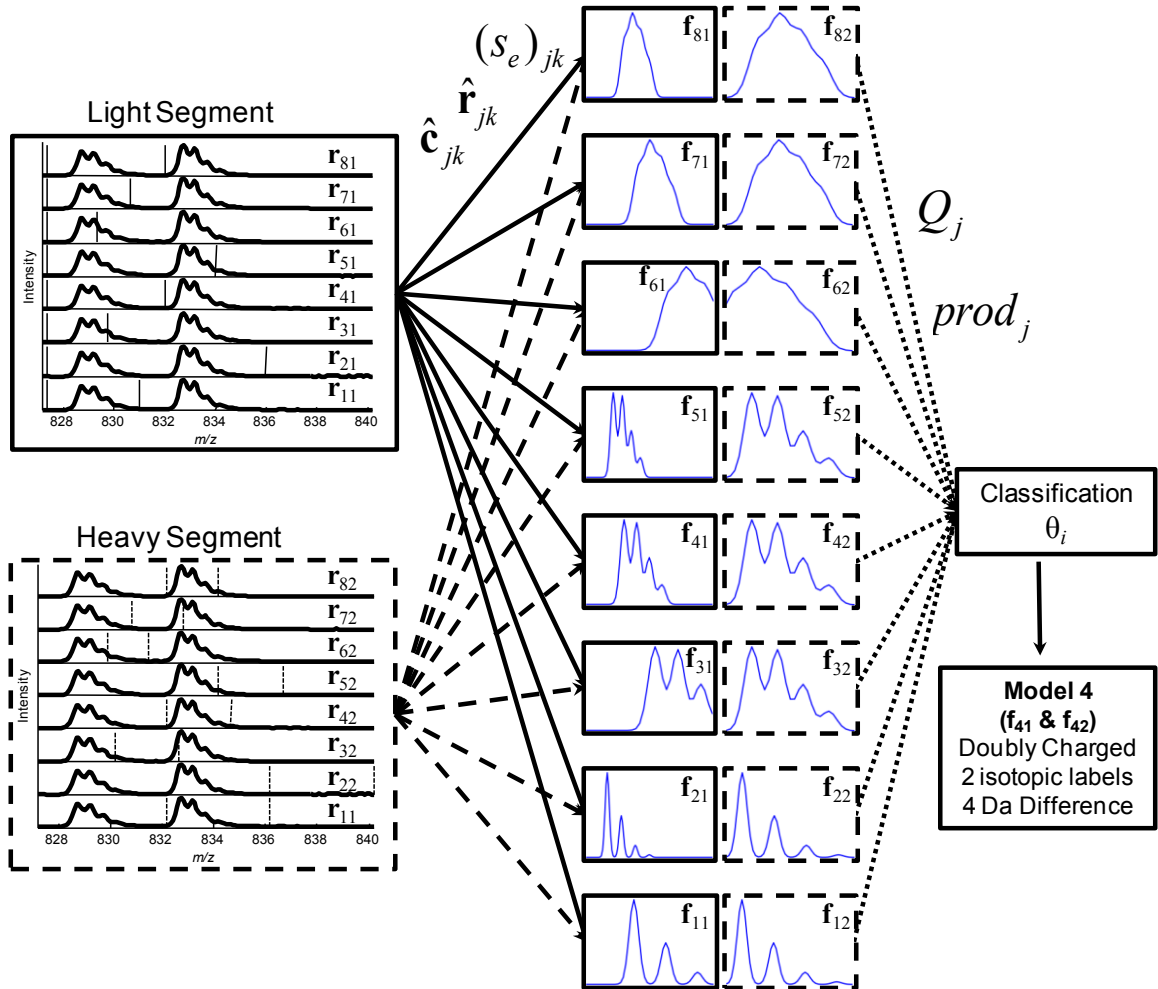
Another parameter calculated for each model will be referred to as the product value, and is simply defined as the product of the two fitted contributions.

$$prod_{ij} = \hat{c}_{ij1}\hat{c}_{ij2} \quad (5.15)$$

This scalar quantity is meant to reflect the intensity of the measured peptide pair. A product was incorporated rather than a sum because it also reflects the covariance of the peptide pair. However, it is important to note that if the quality parameter,  $Q_{ij}$ , is above the threshold for a given model,  $prod_{ij}$  was set to zero.

The last step in the model assessment is to create a classification for mass spectral window  $i$ , which will be designated as  $\theta_i$ . If none of the models are considered acceptable,  $\theta_i$  is set to zero. If only one model is considered acceptable,  $\theta_i$  is set to that number ( $j$ ). It is possible; however, that more than one model will give an acceptable fit. For example, in Figure 5.9, both model 4 ( $\mathbf{f}_{41}$  &  $\mathbf{f}_{42}$ ) and 8 ( $\mathbf{f}_{81}$  &  $\mathbf{f}_{82}$ ) give reasonable residuals, although model 4 is clearly the better fit. In such circumstance,  $\theta_i$  is assigned to the model with the smaller  $Q$  value (better fit).

Figure 5.10 presents an overview of one iteration of the PITS algorithm. This figure shows the process necessary to establish the information needed to locate and recognize the type of peptide pair. Each solid line in the figure represents the calculations completed for the light subwindow ( $k = 1$ ) of a peptide pair, which includes  $\hat{c}_{ij1}$ ,  $\hat{\mathbf{r}}_{ij1}$  and  $(s_e)_{ij1}$  respectively for each peptide model. The dashed lines represent the calculations completed for the heavy subwindow ( $k = 2$ ) of a peptide pair ( $\hat{c}_{ij2}$ ,  $\hat{\mathbf{r}}_{ij2}$ ,  $(s_e)_{ij2}$ ) for each model. Once these calculations are finished  $Q_{ij}$  and  $prod_{ij}$  for each model are calculated (dotted line). From the  $Q_{ij}$  values the algorithm decides on which model, if any, is appropriate. For example, in this figure, the peptide pair is doubly charged with two isotopic labels and a four mass unit difference.

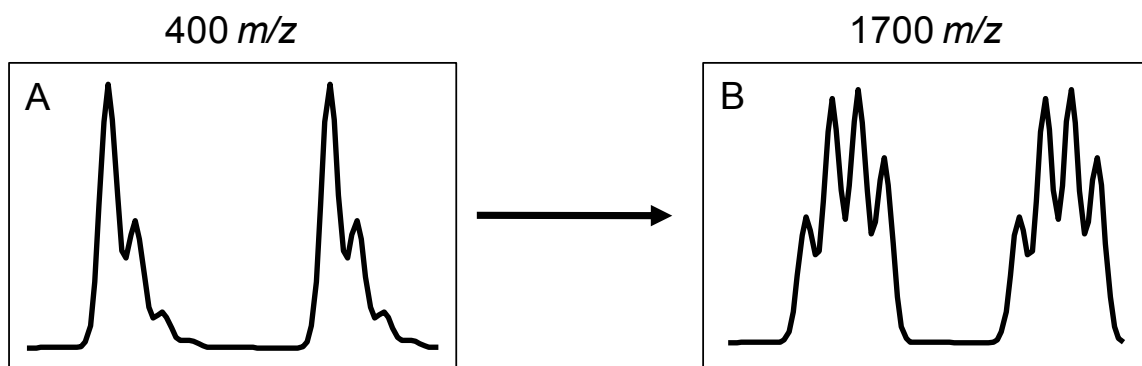


**Figure 5.10** The overall calculation procedure for the PITS algorithm at one particular mass spectral window  $i$  for both  $k$  segments.

The preceding discussion applies to the parallel calculations carried out for one iteration of the PITS algorithm, i.e. for one mass spectral window. This window moves across each channel of the mass spectrum,  $i$ , for every time point,  $j$ , in the mass chromatogram, resulting in a classification,  $\theta_{ij}$  for every point in the mass chromatogram (excluding a small number that are inaccessible at the beginning and end). The resulting classification matrix,  $\Theta$ , is a map of the location of all detected peptide pairs in the mass chromatogram and will be discussed in Section 5.3.3. It is apparent that this is a computationally intensive approach. A typical computation time for a BSA mass

chromatogram of dimensions 15000 x 1502 was approximately 12 hours. Although this is an extended calculation, it is still relatively short compared to the experiment itself and not much different from extensive searches of MS/MS data. Moreover, steps have not yet been taken to fully optimize the algorithm or exploit parallel computing strategies.

An important concept that should be noted is the dynamics of the PITS algorithm, in particular the calculated isotopic patterns described in Section 5.2.2. As the PITS algorithm is applied to each window over a mass spectrum, the calculated ratios in the isotopic profile are adjusted every ten mass units. This interval was chosen because simulations found that the isotopic ratios only exhibit noticeable changes with a change of about ten mass units and the efficiency of the algorithm could be improved by incorporating these values into a lookup table rather than calculating them on the fly. Figure 5.11 shows the calculated isotopic profile for a doubly charged yeast peptide with two labels used for the PITS algorithm at the start (400 $m/z$  – Figure 5.11A) and at the end (1700 $m/z$  – Figure 5.11B).



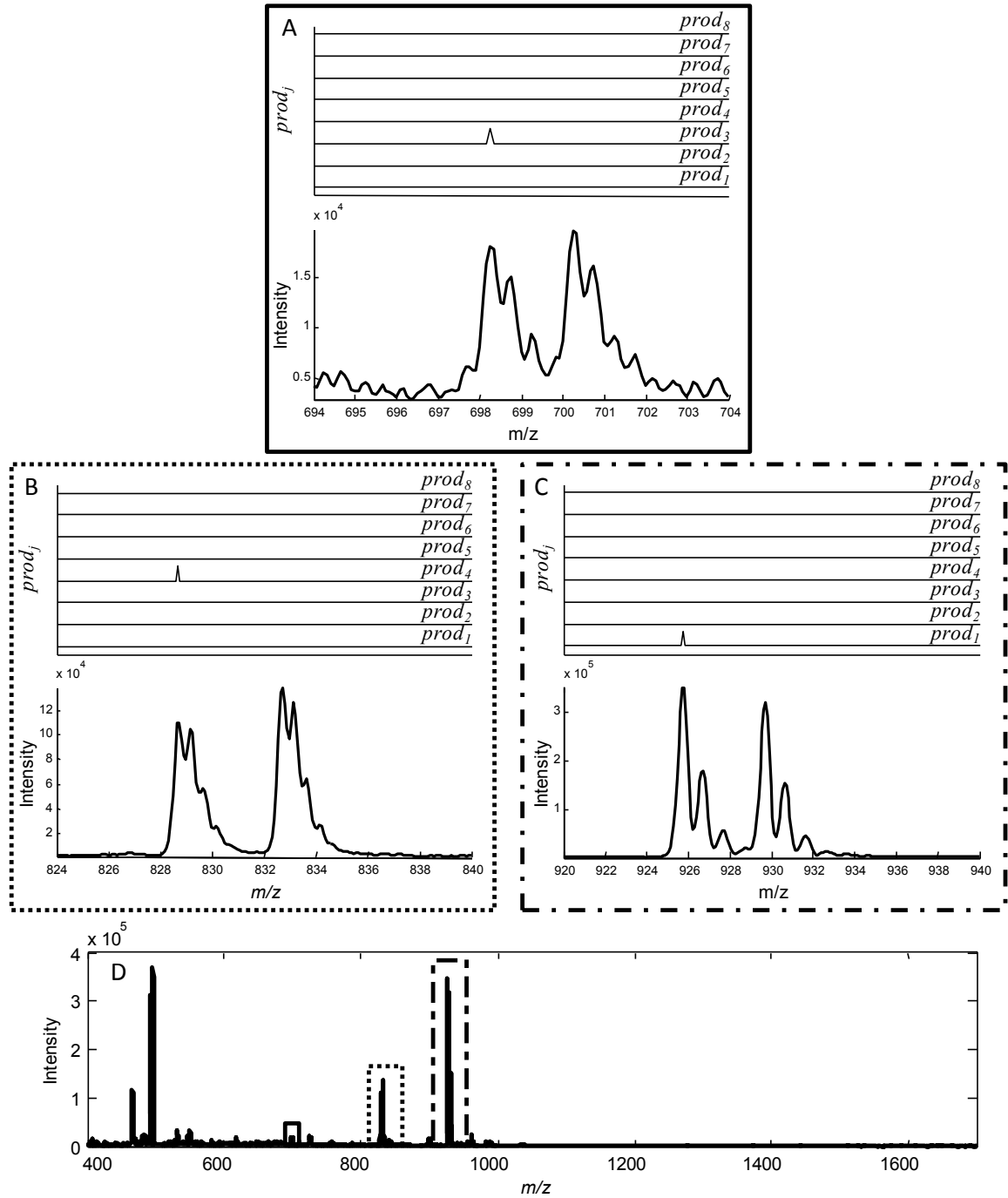
**Figure 5.11** The isotopic pattern for a doubly charge yeast peptide with two labels at 400  $m/z$  (A) and 1700  $m/z$  (B).

Figure 5.12 illustrates the performance of the PITS algorithm, when applied to the mass spectrum from Figure 5.8B. A number of labelled peptide pairs are located. Figure

5.12 highlights three particular peptide pairs of interest. The upper section of Figures 5.12A-C is a plot of  $prod_j$  values corresponding to mass-to-charge range as in the lower section. A  $prod_j$  value will be greater than zero when the PITS algorithm detects a peptide pair with one of the models. Figure 5.12B shows the detection of a doubly charged peptide with a four  $m/z$  unit difference, which was the most common labelled peptide pair located. This was expected because tryptic digestion produces mainly doubly labelled peptides and mass spectrometry most commonly forms doubly charged peptides. Figure 5.12A shows a doubly charged peptide with a two  $m/z$  unit difference located near the baseline. This was very encouraging because peptide pairs near the baseline are hard to locate when employing traditional quantitative methods. Figure 5.12C shows a singly charged peptide with a four  $m/z$  unit difference. As already stated in Section 5.1.1, these peptide pairs are important to locate because quantitation and MS/MS scans are never done on singly charged peptides using traditional methods. This is a benefit of using the PITS algorithm, which allows singly charge peptides that might never be found with traditional methods to be included. The changes in the isotopic ratios among the peptides, which are dynamically modeled by the PITS algorithm, should also be noted.

It is apparent from the results in this section that the PITS algorithm developed has the ability to recognize differentially labelled peptide pairs through specific isotopic patterns without identifying the peptides through MS/MS scans.

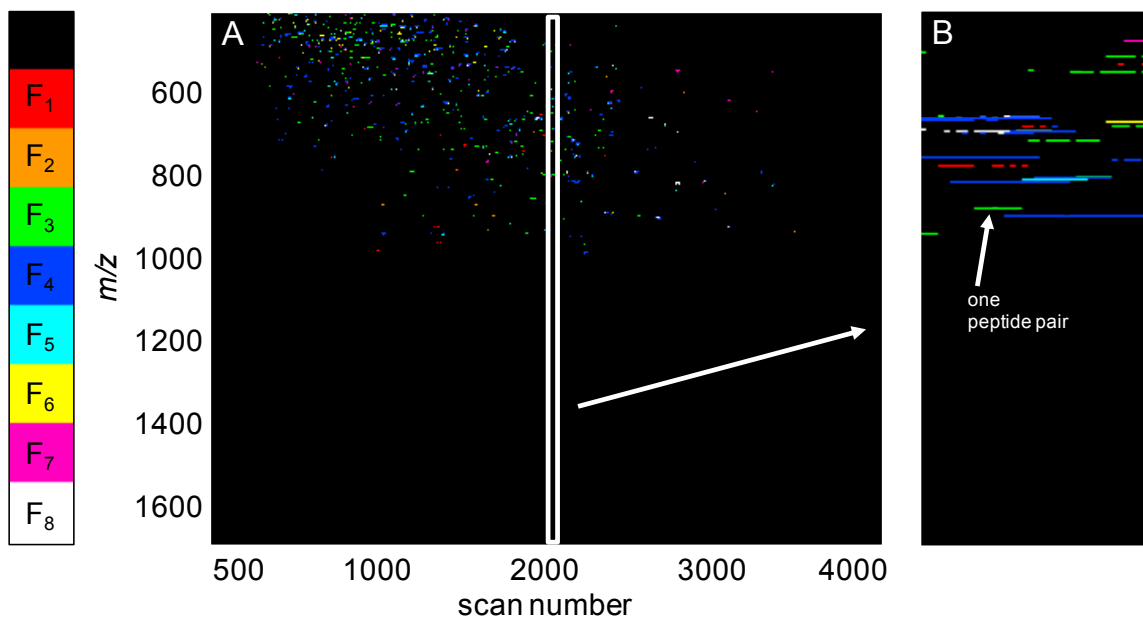




**Figure 5.12** A selected mass spectrum (D) from the yeast mass chromatogram with three particular peptide pairs highlighted. For (A) to (C) the lower section corresponds to the expanded mass spectral regions indicated on (D) and the upper section shows the  $prod_j$  values for each combination respectively over the same mass spectral range.

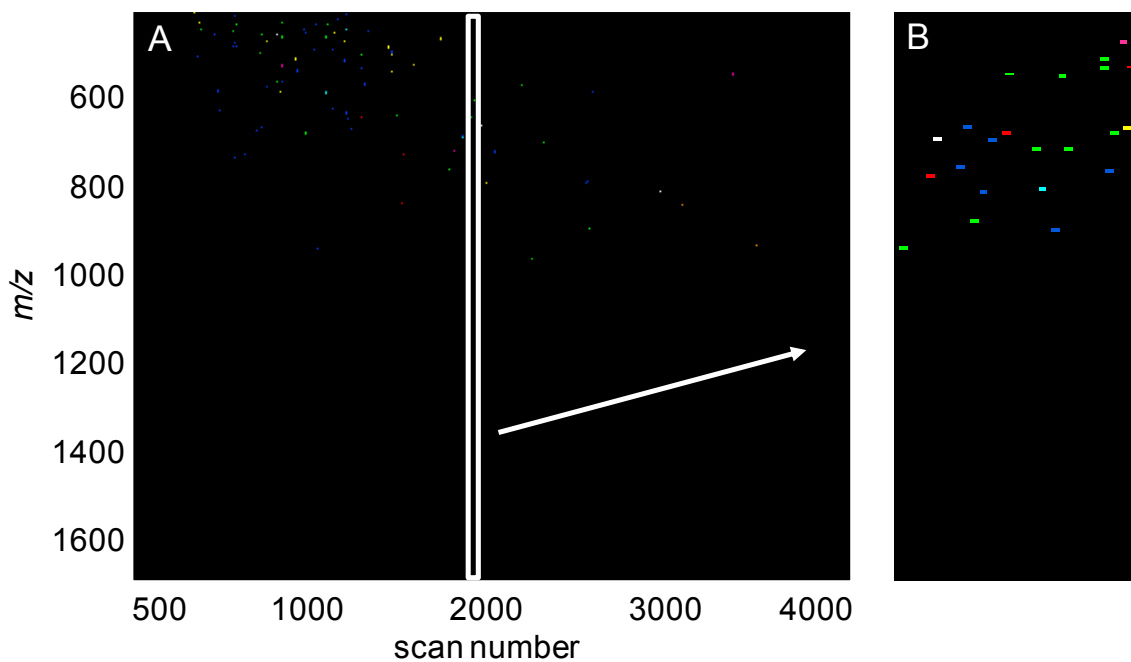
### 5.3.3 Mapping of the Classification Matrix, $\Theta$

As previously stated in Section 5.3.2, when the PITS algorithm moves across a mass chromatogram a classification matrix,  $\Theta$ , is produced. This classification matrix provides the location (mass channel and scan number) for all detected peptide pairs in the mass chromatogram. To visualize the classification matrix a unique colour map was constructed that consisted of 9 different colours to represent the eight isotopic patterns modeled and one for when the PITS algorithm did not detect a peptide pair. For example, the classification matrix for the yeast mass chromatogram is shown in Figure 5.13A. In addition, Figure 5.13B shows an expanded section for the region indicated on Figure 5.13A.



**Figure 5.13** A two-dimensional representation of the classification matrix for the yeast mass chromatogram (A) and (B) is a expanded region of the classifications indicated on (A). See inset colour bar for corresponding colours that represent each PITS model. Black corresponds to when no peptide pair was detected.

The classification matrix is very complex as shown in Figure 5.13B. This is because for every time point that a peptide pair was detected is marked with a point. One detected peptide pair corresponds to a continuous line of points as indicated in Figure 5.13B. To reduce the complexity of the classification matrix, the  $Q_{ij}$  values (fit parameter) were taken into account for each detected peptide pair. Only the time points that had the lowest  $Q_{ij}$  value for a detected peptide pair were kept. This was done by using a smoothing method (binning approach) that moved across the classification matrix to identify points belonging to a single peptide pair and replace them with a single point corresponding to the minimal  $Q_{ij}$  value. The resulting modified matrix is shown in Figure 5.14.

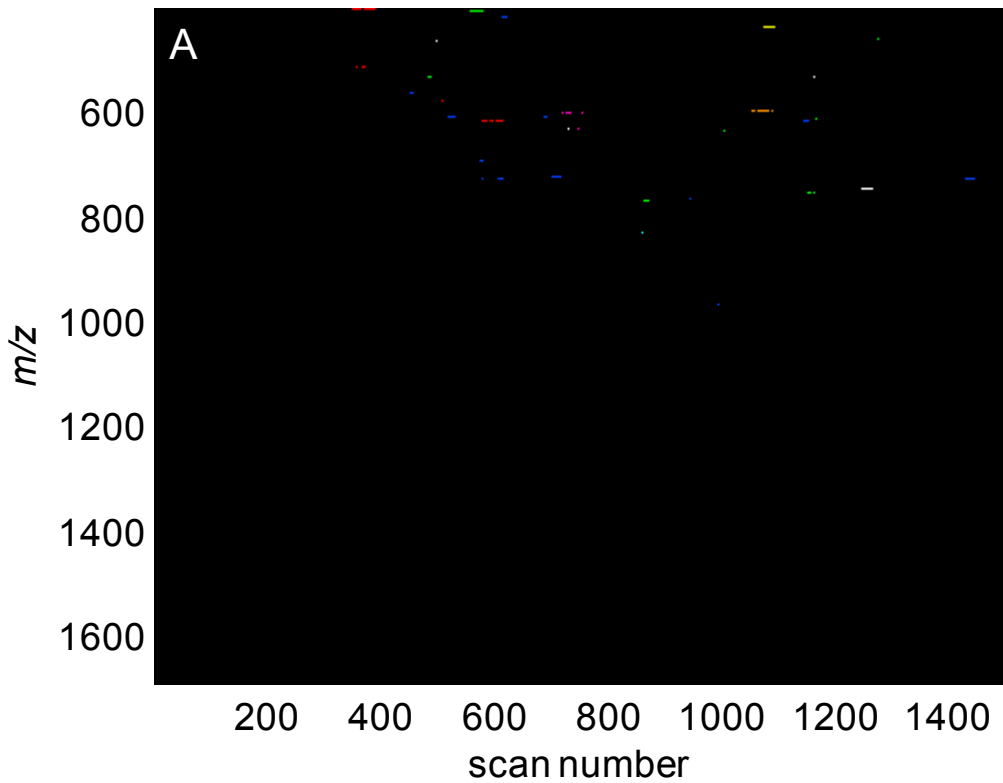


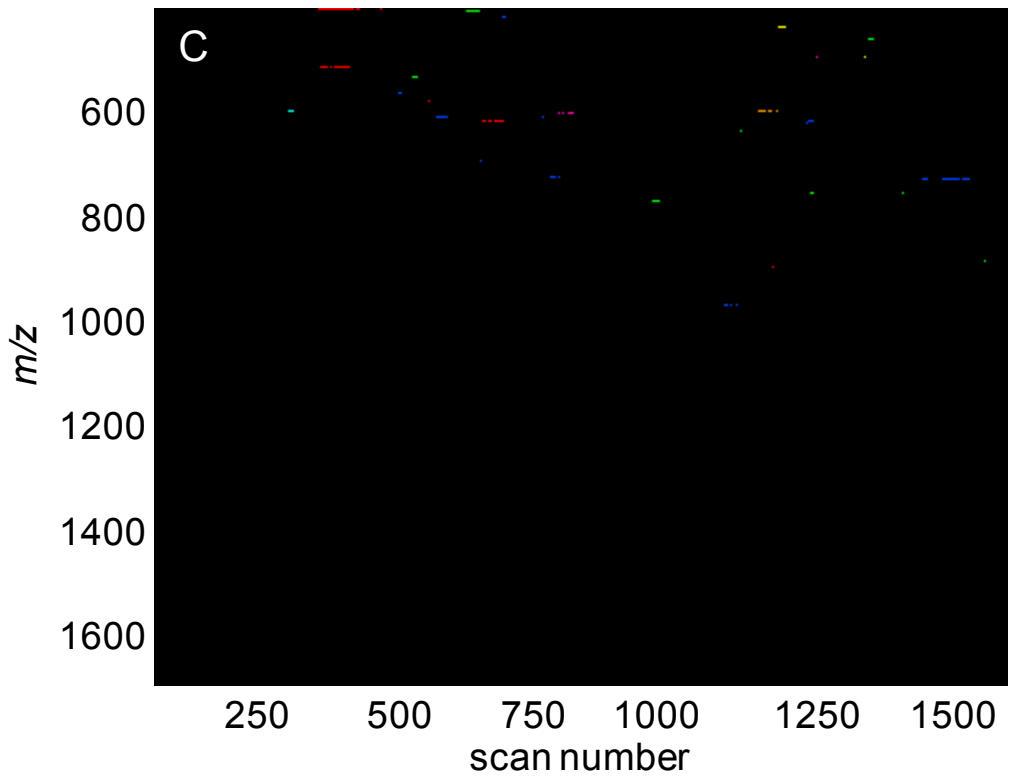
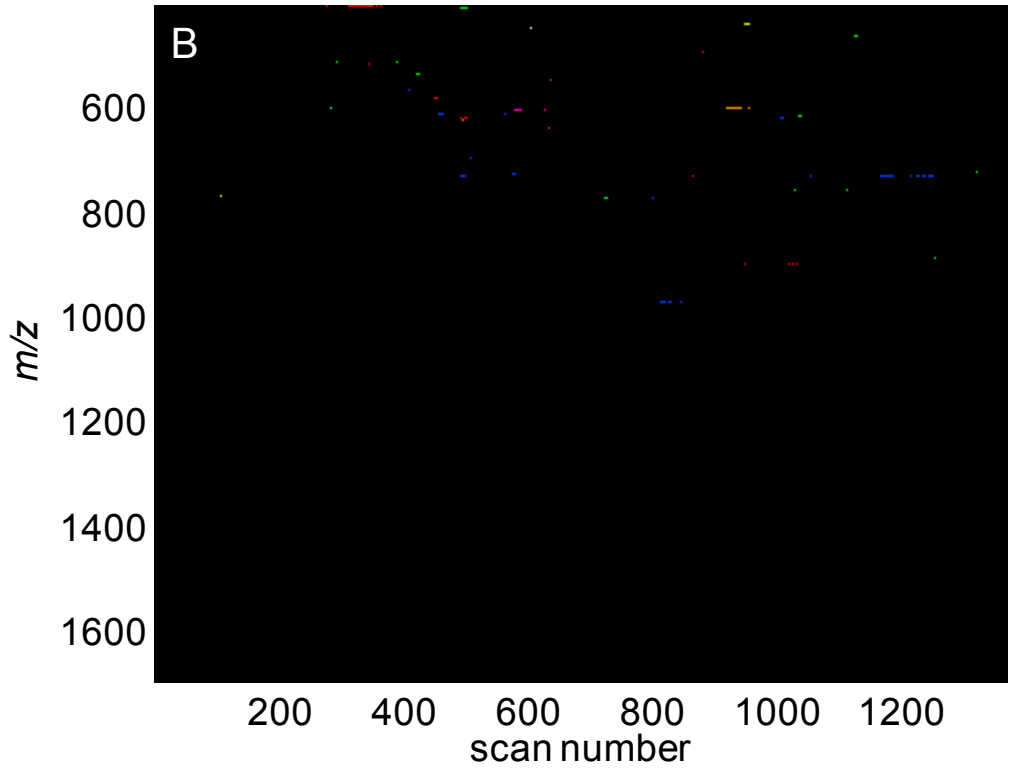
**Figure 5.14** A two dimensional representation of the modified classification matrix for the yeast mass chromatogram (A) and (B) is a expanded region of the modified classifications indicated on (A).

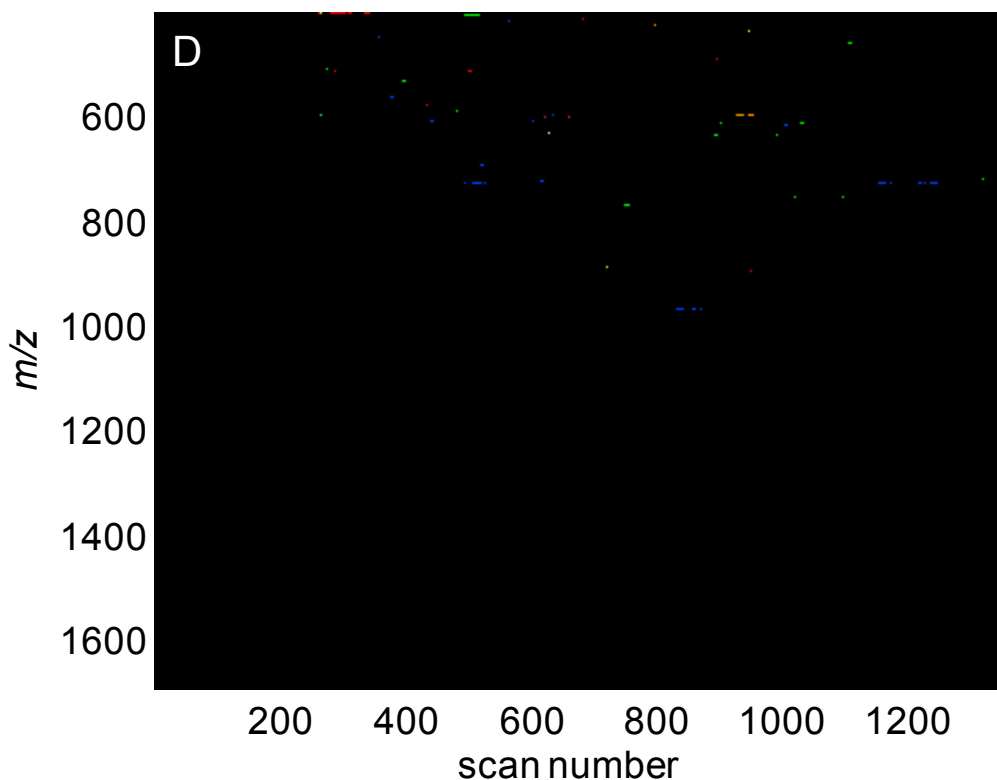
Using the modified classification matrix shown in Figure 5.14 a total list of detected peptides pairs found using the PITS algorithm can be compiled.

## 5.4 Performance of PITS Algorithm

Section 5.3 described how the PITS algorithm works, but before it is really used in practice the algorithm has to be validated. The yeast proteome is much too complex to do this, so the BSA dataset, which consisted of four replicate runs as described in Section 3.1, was employed. The resulting classification matrices for the four replicates are shown in Figure 5.15. This section will describe the performance of the PITS algorithm in detecting peptide pairs in BSA (Figure 5.15) compared to the traditional SEQUEST protocol.







**Figure 5.15** A two dimensional representation of the classification matrix for the four replicate BSA experiments.

Using the classification matrices from Figure 5.15 a complete list of detected peptide pairs was compiled by combining the four replicate lists into one. This resulted in a list of 535 detected unique peptide pairs in BSA. A summary of this list is presented in Table 5.2 and the complete list is in Appendix E.

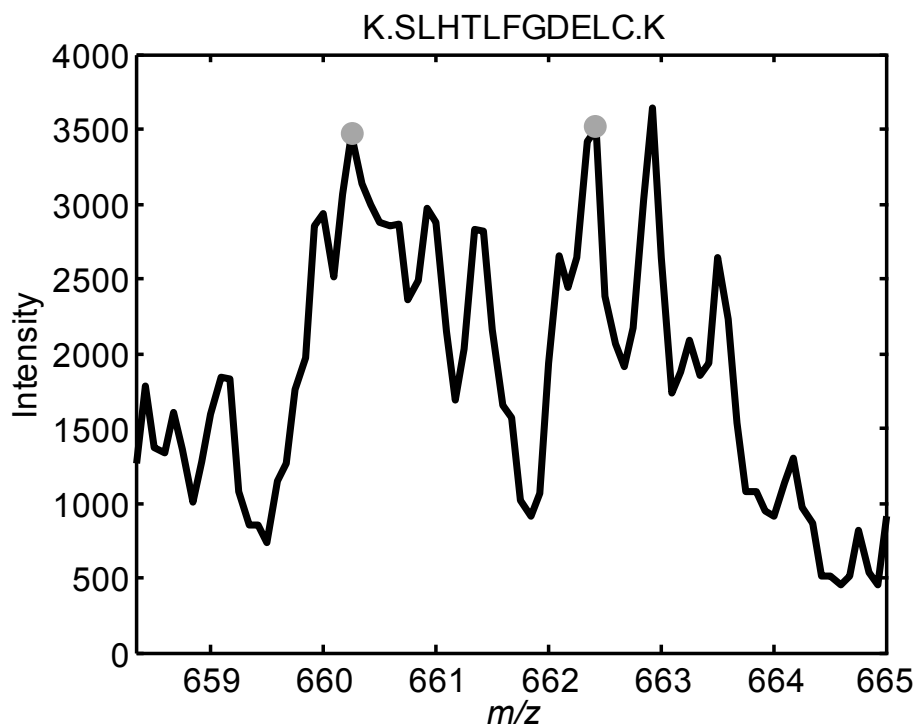
**Table 5.2** Summary of the 535 detected peptide pairs from the PITS algorithm.

Model	No. of Labels	Charge	No. of Peptide Pairs Found
1	1	1	106
2	2	1	11
3	1	2	132
4	2	2	159
5	3	2	13
6	1	3	20
7	2	3	64
8	3	3	30

The 535 detected peptide pairs included 117 singly charge peptides, 304 doubly charged peptides, and 114 triply charge peptides. There were 258 peptides with one label, 234 peptides with two labels, and 43 with three labels. A total of 170 were found only in one file, 101 were found in two files, 79 were found in three files, and 185 were found in all four files. Out of the 535 detected peptide pairs, 43 were found to be associated with the same peptide but detected at different charges. This was based on comparing the apparent peptide mass between the peptide pairs with a tolerance of 0.3 mass units. A total of 11 were found as singly and doubly charged with one label, 6 as singly and doubly charged with two labels, 5 as doubly and triply charged with one label, 19 as doubly and triply charged with two labels, and 2 as doubly and triply charged with three labels. In other words, it can be said this list corresponds to 535 peaks found, which represents 492 estimated peptides.

To validate the PITS algorithm, the 535 detected peptide pairs were compared to the list produced from a SEQUEST search of the replicate files. As described in Section 3.1.7 a multiconcensus SEQUEST search of the BSA database resulted in a total of 155 unique BSA peptides, being identified as either the heavy or the light label, in one or more of the four replicate runs. The SEQUEST search was run allowing PTMs to occur in BSA, but generally PTMs do not arise in BSA. Therefore, 9 of the 155 unique BSA peptides were removed from the list. Another problem that occurred was that SEQUEST incorrectly identified peptide pairs and, when manually searched, the majority of the peptide pairs had mass spectral interference because the pairs were close to baseline, see Figure 5.16, for example. This was problematic for the PITS algorithm because the isotopic patterns couldn't be distinguished and it was unclear if the mass spectral data

was actually for a peptide pair. This problem produced 35 incorrectly identified peptides and these were also removed from the list. Finally, two peptides that SEQUEST found had four labels, and in this work there was no PITS model to fit this particular peptide pair combination. After this preprocessing, a total of 105 unique peptide pairs were retained. Out of the 105 unique peptide pairs, 22 were found at multiple charges. This increased the list from 105 peptide pairs to 127 peptide peaks found. Out of the 127 peaks, there were 4 peaks corresponding to peptides that were quadruply charged and these were removed from the list. After removing the quadruply charged peptides the list had a total of 123 peaks corresponding to 105 unique BSA peptides pairs.



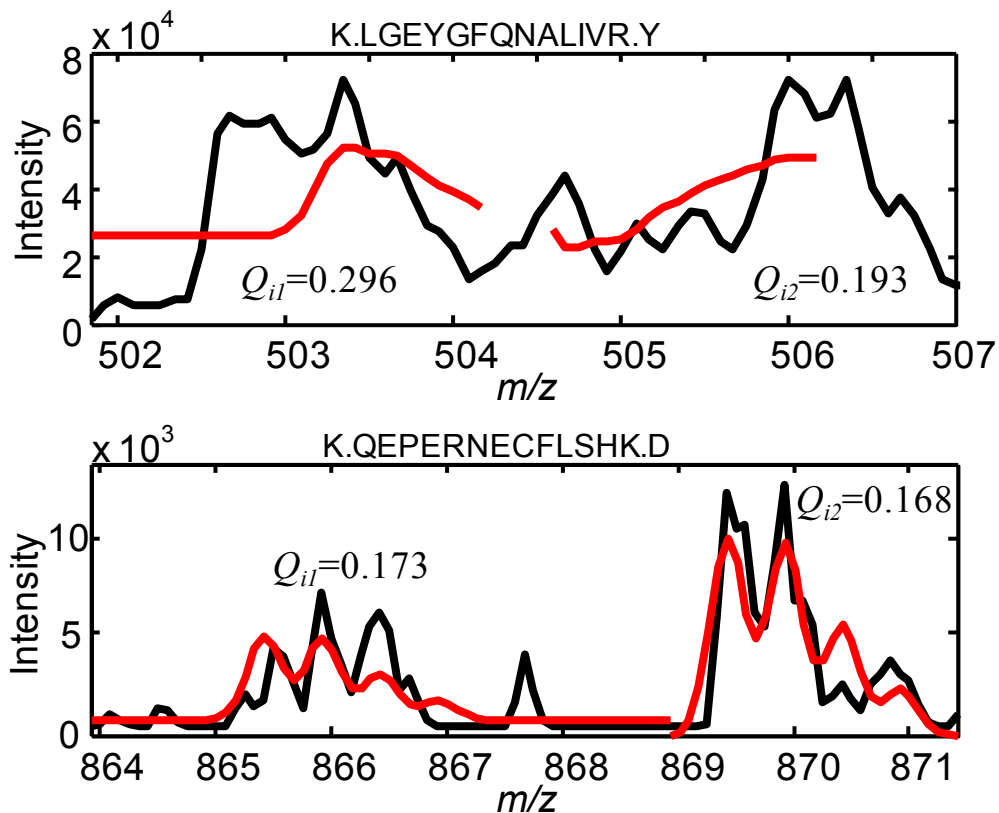
**Figure 5.16** An example of a peptide pair close to the baseline where the isotopic pattern was not clear. The gray dots correspond to the mass of the light and heavy components for the peptide indicated on the top of the plot.

Ultimately, when comparing the two lists to each other, the PITS list was reduced to 418 peaks (535 minus 117) because the SEQUEST list does not contain singly charged



peptides. This is done to verify that the PITS algorithm found all of the 123 SEQUEST peaks. This process compared four values between the two lists: (1) the detected mass of the light component of the peptide pair, (2) the charge of the peptide pair, (3) the number of labels in the peptide pair and (4) the scan number where the peptide pair was detected. The tolerance for the mass difference was 0.3 units and if there was more than one peptide from the PITS list that had a tolerance of less than 0.3, then the peptide that gave the smallest difference was said to be the corresponding peptide. It was found that 93% (114 of 123) SEQUEST peaks were found when compared to the PITS list. A total of 9 SEQUEST peaks were not found by the PITS algorithm. By manual inspection of these peptide pairs, it was concluded that the peaks were not found because of the quality of the mass spectral data. Figure 5.17 shows two examples of two different SEQUEST found peptide pairs that were not found by the PITS algorithm and their corresponding  $Q_{ij}$  values. It is clear the quality of the mass spectral data is inadequate, the isotopic patterns are not clear and the  $Q_{ij}$  values from the PITS algorithm are greater than the threshold of 0.10. The PITS algorithm would never be able to locate these peptide pairs unless the  $Q_{ij}$  threshold was increased, but this would introduce more low quality mass spectral peptide pairs unsuitable for quantitative measurement and more false positives.

The anticipated question that comes from these results is why is the number of peaks in the PITS algorithm list much larger compared to the SEQUEST list. At present the identity of the other peaks that the PITS algorithm finds is unclear, as are the reasons they were not found by SEQUEST. One obvious reason for the lack of SEQUEST results could be that no MS/MS scan was done for these detected peaks. However, a manual search was done and, for the majority of the peaks found by PITS (excluding the singly



**Figure 5.17** The observed mass spectral response (black) and the predicted mass spectral response (red) calculated for two peptide pairs not found by PITS algorithm. The corresponding sequence and  $Q_{ij}$  values for each peptide pair are inset.

charged peptides for which no MS/MS scan is done) at least one MS/MS scan was completed at the appropriate mass near the time point that the PITS algorithm found the peptide pair. The other obvious reason would be the presence of non-BSA peptide contaminants in the samples, but when the SEQUEST search was done, it was also searched against the sequence-reversed BSA data base and a data base with the 10 most common protein contaminants with no significant identifications. Following this a manual inspection of the majority of the unidentified PITS found peptide pairs was carried out and it was confirmed that all of the identified peptide pairs fit a model well and the ratio of the two components was close to unity. It is likely then that the MS/MS data was of insufficient quality to identify the ions as BSA peptides. This is possible

because the MS/MS scans were not always performed at the maximum of the chromatography peak due to the limited duty cycle of the instrument. Beyond this, it is difficult to perform a complete validation of the peaks identified by the PITS algorithm because of the lack of a method that can identify all peptide pairs. However, it is felt that all of the peaks identified were valid peptide pairs.

## **5.5 Summary**

The premise of this work was that the identification of peptides in proteomics by mass spectrometry is limited by the duty cycle of the instrument required for MS/MS analysis of the ions. Based on this, the PITS algorithm was applied to the replicate BSA samples and it was demonstrated that even operating with MS/MS scans more peaks were found with a single MS scan using the PITS algorithm than were identified through SEQUEST using MS/MS data. The majority (114 of 123) of the SEQUEST peaks were found by the PITS algorithm, in addition 421 more peaks were found. Of these, 117 were singly charged peptides which may have limited utility in identification of peptides and the rest were 304 unidentified peaks. It is expected that this will improve further when dedicated mass spectral scanning is used without MS/MS analysis. However, MS/MS (SEQUEST searching) is still necessary to positively identify peptides. It is expected that this high throughput method will be advantageous in proteomics studies especially where differential expression of proteins is sought. Further work needs to be done on the PITS algorithm, however using more complex mixtures like the yeast data set described in this thesis. To date, preliminary studies have been carried out on the yeast data set, where 768 peptide pairs have been found but no extensive searches of the proteome have been

done. These included 46 singly charged peptides (28 with one label and 18 with two labels), 576 doubly charged peptides (196 with one label, 315 with two labels and 65 with three labels), and 136 triply charged peptides (80 with one label, 35 with two labels and 21 with three labels).

# Chapter 6

## Conclusions

### 6.1 Conclusions

The work presented in this thesis has described the development of an algorithm to locate isotopically labelled peptide pairs to provide an approach for high-throughput proteomic analysis. It is essential to improve the throughput of modern proteomic experiments to make quantitative proteomics more routine, particularly for protein biomarker discovery. In Chapter 1, the restrictions of traditional approaches to quantifying isotopically labelled peptides were introduced. The challenge was presented as being able to locate peptides without the need for MS/MS scans. The elimination of MS/MS scans can lead to a number of potential enhancements to proteomic analysis, including an increase in the number of MS scans, a reduction in separation time, improve sample throughput, and the quantitation of more peptides. It was established that quantitative proteomics has evolved from the 2D-PAGE approach to the LC-MS based approaches which use isotopic labelling techniques, to distinguish peptides from different cell states.

In Chapter 2, a detailed description of the most common isotopic labelling techniques was presented, which included techniques that can be introduced to a peptide or protein chemically, metabolically, enzymatically or by spiked synthetic peptide standards. Label-free methods were also discussed in Chapter 2. It was concluded that a chemically-based labelling technique was the best choice for this thesis. Stable-isotope dimethyl labelling was chosen for this work because of its simplicity, low cost and

overall performance. Chapter 3 provided the experimental details for the quantitative proteomic experiments performed in this thesis and details of data conversion.

In Chapter 4, a complete study on the effect of isotope labelling on retention characteristics for the widely employed strategy of dimethyl labelling with H<sub>2</sub>/D<sub>2</sub> formaldehyde was performed. Isotopic effects on retention time were observed when stable dimethyl labelling was employed for quantitative proteomic experiments. The effect was small for the majority of peptide pairs investigated but can be substantial for other cases. If there was a separation of the labelled peptides, the deuterium labelled peptide eluted first. It was determined that the isotopic effect is larger than the effect shown when labelling with <sup>13</sup>C based reagents, but smaller when labelling with deuterated ICAT reagents. The results from the computational studies were found to be consistent with speculations that the magnitude of the isotopic effect was inversely related to the polarity of the label. Excluding the one anomalous case observed, it was shown that the isotopic effect on quantitation by single point methods was negligible when compared to the measurement uncertainty. Therefore, stable-isotope dimethyl labelling was deemed an appropriate labelling technique for quantitative proteomics and should not be dismissed as suggested in the literature.

The development of an algorithm for the detection of dimethylated peptides was described in Chapter 5. A number of strategies were employed, but the most promising method for the detection of peptide pairs was the PITS method. The PITS algorithm is based on a model that uses predicted isotopic patterns arising from labelled peptide pairs and comparing these to experimental mass spectral patterns to locate and classify appropriate peptide pairs in a mass chromatogram. PITS is a dynamic algorithm that

adapts the model to changes in the mass-to-charge ratio and performs parallel calculations for every iteration. It was demonstrated that the PITS algorithm is successful in locating peptides pairs in the simple BSA protein and the extensively more complex yeast proteome. When compared to traditional quantitative methods using replicate BSA samples, it was demonstrated that the PITS algorithm found more peaks (using MS scans only) than were identified by SEQUEST operating with MS/MS scans. Although, steps have not yet been taken to fully optimize the algorithm, it was shown that the PITS algorithm has the potential to be an effective tool in systems biology and improve throughput in proteomic experiments. In contrast to traditional methods, the emphasis of this proposed approach is peptide pair discovery followed by identification, rather than identification followed by discovery. It is hoped that this proposed approach will provide a new view towards quantitative proteomics and perhaps become the method of choice for proteomic analysis in the years to come.

## **6.2 Future Work**

With the development of the PITS algorithm described in this work and its application to an appropriate isotopic labelling technique (stable-isotope dimethyl labelling), there are a number of challenges ahead for the work presented in this thesis. Besides the obvious challenge of quantitation, described later in this section, a number of features of the PITS algorithm need to be optimized. First, the isotopic patterns need to be extended to an appropriate mass spectral range to include more than just the first four isotopic peaks as presented here. This will improve the selectivity and accuracy of the PITS algorithm, allowing the algorithm to fit more molecular ions and provided an improved quality fit parameter, particularly for higher mass ions and higher charge states.

This will also allow a second improvement, which is the conversion of the mathematical model from two sub-window fits to a single window. This would make the model contiguous and more robust than the current model. Third, a new method is required to cluster results from the PITS algorithm so that single points identified can be merged into a single chromatographic peak for a given peptide. In this work, a smoothing method was used to identify points belonging to the same peptide detection event and replace them with a single point. This method was unreliable and a majority of the time, manual inspection was required. Finally, to detect replicate peptides in other experiments, alignment of the chromatograms for replicate samples is needed. One possible method for this would be the correlation optimized warping algorithm [116]. Once the alignment is complete it should be possible to compile a coordinated list of detected peptide peaks and determine whether or not they were really found in multiple replicate samples.

Perhaps most importantly, a dependable strategy is needed to perform relative quantitation. This was one of the initial goals for this thesis, but the study on retention characteristics of isotopically labelled peptides was essential to define the strategy for quantitation. The expected strategy would be to find a ratio of the majority of the area for both XICs of the light and heavy labelled peptide components. However, this is not an easy task because the XICs for the components are not always clear and some type of weighted average of the ratios across the chromatographic peak will be needed to accomplish this. Additionally, there is also another weakness of the PITS algorithm that has not yet been discussed. The PITS algorithm requires both labelled components of a peptide pair to be detectable which, in the context of protein biomarker discovery, may not always be the case. There are at least two strategies to deal with this. The first



approach would be to use mixed labelling experiments, in which the test and reference are present in both samples in different ratios (e.g. 10:1 and 1:10). This would provide all the pertinent peptide information to locate every peptide when the labelling experiment is done at a 1:1 ratio. The second strategy would entail combination experiments involving ref/ref, test/test and test/ref samples. The first two experiments would be used to detect the peptides, which are then sought in the third experiment. Taking these two strategies into account, it is apparent that the PITS algorithm is a useful strategy to improve quantitative proteomics.

## References

- [1] Aebersold R, Mann M; Mass Spectrometry-based Proteomics. *Nature*. **422:198-207**, (2003).
- [2] Mann M, Hendrickson RC, Pandey A; Analysis of Proteins and Proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70:437-473**, (2001).
- [3] Rifai N, Gillette MA, Carr SA; Protein Biomarker Discovery and Validation: the Long and Uncertain Path to Clinical Utility. *Nat. Biotechnol.* **24:971-983**, (2006).
- [4] Vlahous A, Fountoulakis M; Proteomic Approaches in the Search for Disease Biomarkers. *J. Chromatogr. B.* **814:11-19**, (2005).
- [5] Anderson NL, Anderson NG; Proteome and Proteomics: New Technologies, New Concepts, and New Words. *Electrophoresis.* **19:1853-1851**, (1998).
- [6] Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, Williams KL; Progress with Proteome Projects: Why All Proteins Expressed by a Genome Should be Identified and How to do it. *Biotechnol. Genet. Eng. Rev.* **13:19-50**, (1996).
- [7] Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humphery-Smith I, Williams KL, Hochstrasser D; From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Amino Acid Analysis. *Nat. Biotechnol.* **14:61-65**, (1996).
- [8] Kahn P; Molecular Biology – From Genome to Proteome – Looking at a Cells Proteins. *Science.* **270:369-370**, (1995).
- [9] Anderson NL, Anderson NG; The Human Plasmas Proteome – History, Character, and Diagnostic Prospects. *Mol. Cell. Proteomics.* **1:845-867**, (2002).
- [10] Mann M, Jensen ON; Proteomic Analysis of Post-translational Modifications. *Nat. Biotechnol.* **21:255-261**, (2003).
- [11] Jensen ON; Modification-specific Proteomics: Characterization of Post-translational Modifications by Mass Spectrometry. *Curr. Opin. Chem. Biol.* **8:33-41**, (2004).
- [12] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al.; The Sequence of the Human Genome. *Science.* **291:1304-1351**, (2001).

- [13] Anderson L, Seihamer J; A Comparison of Selected mRNA and Protein Abundances in Human Liver. *Electrophoresis*. **18:533-537**, (1997).
- [14] Wilkins MR, Sanchez JC, Williams KL, Hochstrasser DF; Current Challenges and Future Applications for Protein Maps and Post-translational Vector Maps in Proteome Projects. *Electrophoresis*. **17:830-838**, (1996).
- [15] Washburn MP, Wolters D, Yates; Large-scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology. *Nat. Biotechnol.* **19:242-247**, (2001).
- [16] Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP; Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-scale Protein Analysis: The Yeast Proteome. *J. Proteome Res.* **2:43-50**, (2003).
- [17] Chou KC; Progress in Protein Structural Class Prediction and its Impact to Bioinformatics and Proteomics. *Curr. Protein Peptide Sci.* **6:423-436**, (2005).
- [18] Liu HL, Hsu JP; Recent Developments in Structural Proteomics for Protein Structure Determination. *Proteomics*. **5:2056-2068**, (2005).
- [19] Naylor S, Kumar R; Emerging Role of Mass Spectrometry in Structural and Functional Proteomics. *Adv. Protein Chem.* **65:217-248**, (2003).
- [20] Downard KM; Ions of the Interactome: The Role of MS in the Study of Protein Interactions in Proteomics and Structural Biology. *Proteomics*. **6:5374-5384**, (2006).
- [21] Schwikowski B, Uetz P, Fields S; A Network of Protein-protein Interactions in Yeast. *Nat. Biotechnol.* **18:1257-1261**, (2000).
- [22] von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P; Comparative Assessment of Large-scale Data Sets of Protein-protein Interactions. *Nature*. **417:399-403**, (2002).
- [23] Meng FY, Forbes AJ, Miller LM, Kelleher NL; Detection and Localization of Protein Modifications by High Resolution Tandem Mass Spectrometry. *Mass Spectrom. Rev.* **24:126-137**, (2005).
- [24] Temporini C, Callerli E, Massolini G, Caccialanza G; Integrated Analytical Strategies for the Study of Phosphorylation and Glycosylation in Proteins. *Mass Spectrom. Rev.* **27:207-236**, (2008).

- [25] O'Farrell PH; High-resolution Two-dimensional Electrophoresis of Proteins. *J. Biol. Chem.* **250:4007-4021**, (1975).
- [26] Gorg A, Obermaier C, Boguth G, Harder A, Scheibe B, Wildgruber R, Weiss W; The Current State of Two-dimensional Electrophoresis with Immobilized pH Gradients. *Electrophoresis.* **21:1037-1053**, (2000).
- [27] Gorg A, Weiss W, Dunn MJ; Current Two-dimensional Electrophoresis Technology for Proteomics. *Proteomics.* **4:3665-3685**, (2004).
- [28] Opitzek GJ, Lewis KC, Jorgenson JW, Anderegg RJ; Comprehensive On-line LC/LC/MS of Proteins. *Anal. Chem.* **69:1518-1524**, (1997).
- [29] Wolters DA, Washburn MP, Yates JR; An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. *Anal. Chem.* **73: 5683-5690**, (2001).
- [30] Laemmli UK; Cleavage of Structural Proteins During Assembly of Head of Bacteriophage-T4. *Nature.* **227:680-685**, (1970).
- [31] de Godoy LMF, Olsen JV, Cox J, Nielsen ML, Hubner NC, Froehlich F, Walther TC, Mann M; Comprehensive Mass-spectrometry-based Proteome Quantification of Haploid Versus Diploid Yeast. *Nature.* **455:1251:U60**, (2008).
- [32] Shapiro AL, Vinuela E, Maizel JV; Molecular Weight Estimation of Polypeptide Chains by Electrophoresis in SDS-Polyacrylamide Gels. *Biochem. Biophys. Res. Commun.* **28:815-820**, (1967).
- [33] Neuhoff V, Arold N, Taube D, Ehthardt W; Improved Staining of Proteins in Polyacrylamide Gels Including Isoelectric-focusing Gels with Clear Background at Nanogram Sensitivity Using Coomassie Brilliant Blue G-250 and R-250. *Electrophoresis.* **9:255-262**, (1988).
- [34] Hunt DF, Yates JR, Shabanowitz J, Winston S, Hauer CR; Protein Sequencing by Tandem Mass Spectrometry. *Proc. Natl. Acad. Sci.* **83:6233-6237**, (1986).
- [35] Eng JK, McCormack AL, Yates JR; An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **5:976-989**, (1994).
- [36] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M; The SWISS-PROT Protein Knowledgebase and its Supplement TrEMBL in 2003. *Nucleic Acids Res.* **31:367-370**, (2003).

- [37] Stamey TA, Yang N, Hay AR, McNeal JE, Freiha FS, Redwine E; Prostate-specific Antigen as a Serum Marker for Adenocarcinoma of the Prostate. *New England Journal of Medicine*. **317:909-916**, (1987).
- [38] Polanski M, Anderson NL; A List of Candidate Cancer Biomarkers for Targeted Proteomics. *Biomark. Insights*. **2:1-48**, (2006).
- [39] Kitteringham NR, Jenkins RE, Lane CS, Elliott VL, Park BK; Multiple Reaction Monitoring for Quantitative Biomarker Analysis in Proteomics and Metabolomics. *J. Chromatogr. B*. **877:1229-1239**, (2009).
- [40] Kim K, Kim SJ, Yu HG, Yu J, Park KS, Jang I, Kim Y; Verification of Biomarkers for Diabetic Retinopathy by Multiple Reaction Monitoring. *J. Proteome Res*. **9:689-699**, (2010).
- [41] Pencina MJ, D'Agostino (Sr) RBD, D' Agostino (Jr) RBD, Vasan RS; Evaluating the Added Predictive Ability of a New Marker: From Area Under the ROC Curve to Reclassification and Beyond. *Statist. Med*. **27:157-172**, (2008).
- [42] Gutman S, Kessler LG; The US Food and Drug Administration Perspective on Cancer Biomarker Development. *Nat. Rev. Cancer*. **6:565-571**, (2006).
- [43] Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R; Quantitative Analysis of Complex Protein Mixtures Using Isotope-coded Affinity Tags. *Nat. Biotechnol*. **17:994-999**, (1999).
- [44] Langen H, Fountoulakis M, Evers S, Wipf B, Berndt P; In From Genome to Proteome. *3<sup>rd</sup> Siena 2D Electrophoresis Meeting*. Wiley-VCH, Weinheim, Germany, Siena, (1998).
- [45] Rose K, Simona MG, Offord RE, Prior CP, Otto B, Thatcher DR; A New Mass-spectrometric C-terminal Sequencing Technique Finds a Similarity Between  $\gamma$ -interferon and  $\alpha_2$ -interferon and Identifies a Proteolytically Clipped  $\gamma$ -interferon that Retains Full Antiviral Activity. *Biochem. J*. **215:273-277**, (1983).
- [46] Desiderio DM, Kai M; Preparation of Stable Isotope-Incorporated Peptide Internal Standards for Field Desorption Mass-spectrometry Quantification of Peptides in Biologic Tissue. *Biol. Mass. Spectrom*. **10:471-479**, (1983).
- [47] Bondarenko PV, Chelius D, Shaler TA; Identification and Relative Quantitation of Protein Mixtures by Enzymatic Digestion Followed by Capillary Reversed Phase Liquid Chromatography Tandem Mass Spectrometry. *Anal. Chem*. **74:4741-4749**, (2002).

- [48] Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ; Multiplexed Protein Quantitation in *Saccharomyces cerevisiae* Using Amine-reactive Isobaric Tagging Reagents. *Mol. Cell. Proteomics*. **3:1154–1169**, (2004).
- [49] Hsu JL, Huang SY, Chow NH, Chen SH; Stable-isotope Dimethyl Labeling for Quantitative Proteomics. *Anal. Chem.* **75:6843-6852**, (2003).
- [50] Gygi SP, Rist B, Griffin TJ, Eng J, Aebersold R; Proteome Analysis of Low-abundance Proteins Using Multidimensional Chromatography and Isotope-coded affinity Tags. *J. Proteome Res.* **1:47-54**, (2002).
- [51] Guina T, Wu M, Miller SI, Purvine SO, Yi EC, Eng J, Goodlett DR, Aebersold R, Ernst RK, Lee KA; Proteomic Analysis of *Pseudomonas Aeruginosa* grown under magnesium limitation. *J. Am. Soc. Mass. Spectrom.* **14:742-752**, (2003).
- [52] Han DK, Eng J, Zhou H, Aebersold R; Quantitative Profiling of Differentiation-induced Microsomal Proteins Using Isotope-coded Affinity Tags and Mass Spectrometry. *Nat. Biotechnol.* **19:946-951**, (2001).
- [53] Meehan KL, Sadar MD; Quantitative Profiling of LNCaP Prostate Cancer Cells Using Isotope-coded Affinity Tags and Mass Spectrometry. *Proteomics*. **4:1116-1134**, (2004).
- [54] Zhang R, Regnier FE; Minimizing Resolution of Isotopically Coded Peptides in Comparative Proteomics. *J. Proteome Res.* **1:139-147**, (2002).
- [55] Zhang R, Sioma CS, Thompson RA, Xiong L, Regnier FE; Controlling Deuterium Isotope Effect in Comparative Proteomics. *Anal. Chem.* **74:3662-3669**, (2002).
- [56] Borisov OV, Goshe MB, Conrads TP, Rakov VS, Veenstra TD, Smith RD; Low Energy Collision Induced Dissociation Fragmentation Analysis of Cysteinyl Modified Peptides. *Anal. Chem.* **74:2284-2292**, (2002).
- [57] Gilis D, Massar S, Cerf NJ, Rooman M; Optimality of the Genetic Code With Respect to Protein Stability and Amino Acid Frequencies. *Genome Biology*, **2**, (2001), arXiv:physics/0102044v1.
- [58] Hansen KC, Schmitt-Ulms G, Chalkley RJ, Hirsch J, Baldwin MA, Burlingame AL; Mass Spectrometric Analysis of Protein Mixtures at Low Levels Using Cleavable <sup>13</sup>C-Isotope-coded Affinity Tag and Multidimensional Chromatography. *Mol. Cell. Proteomics*. **2:299-314**, (2003).

- [59] Wu WW, Wang G, Baek SJ, Shen RF; Comparative Study of Three Proteomic Quantitative Methods, DIGE, cICAT, and iTRAQ Using 2D Gel- or LC-MALDI TOF/TOF. *J. Proteome Res.* **3:651-658**, (2006).
- [60] Li J, Steen H, Gygi SP; Protein Profiling with Cleavable Isotope-coded Affinity Tag (cICAT) Reagents – The Yeast Salinity Stress Response. *Mol. Cell. Proteomics.* **2:1198-1204**, (2003).
- [61] Sethuraman M, McComb ME, Heibeck T, Costello CE, Cohen RA; Isotope-coded Affinity Tag Approach to Identify and Quantify Oxidant-sensitive Protein Thiols. *Mol. Cell. Proteomics.* **3:273-278**, (2004).
- [62] Molloy MP, Donohoe S, Brzezinski EE, Kilby GW, Stevenson TI, Baker JD, Goodlett DR, Gage DA; Large-scale Evaluation of Quantitative Reproducibility and Proteome Coverage Using Acid Cleavable Isotope Coded Affinity Tag Mass Spectrometry for Proteomic Profiling. *Proteomics.* **5:1204-1208**, (2005).
- [63] Bottari P, Aebersold R, Turecek F, Gelb MH; Design and Synthesis of Visible Isotope-coded Affinity Tags for the Absolute Quantification of Specific Proteins in Complex Mixtures. *Bioconjugate Chem.* **15:380-388**, (2004).
- [64] Lu Y, Bottari P, Turecek F, Aebersold R, Gelb MH; Absolute Quantification of Specific Proteins in Complex Mixtures Using Visible Isotope-coded Affinity Tags. *Anal. Chem.* **76:4104-4111**, (2004).
- [65] Zhou H, Ranish JA, Watts JD, Aebersold R; Quantitative Proteome Analysis by Solid-phase Isotope Tagging and Mass Spectrometry. *Nat. Biotechnol.* **19:512-515**, (2002).
- [66] Qui Y, Sousa EA, Hewick RM, Wang JH; Acid Labile Isotope-coded Extractants: A Class of Reagents for Quantitative Mass Spectrometric Analysis of Complex Protein Mixtures. *Anal. Chem.* **74:4969-4979**, (2002).
- [67] Whetstone PA, Butlin NG, Corneillie TM, Meares CF; Element-coded Affinity Tags for Peptides and Proteins. *Bioconjugate Chem.* **15:3-6**, (2004).
- [68] Zieske LR; A Perspective on the Use of iTRAQ Reagent Technology for Protein Complex and Profiling Studies. *J. Exp. Bot.* **57:1501-1508**, (2006).
- [69] Keshamouni VG, Michailidis G, Grasso CS, Anthwal S, Strahler JR, Walker A, Arenberg DA, Reddy RC, Akulapalli S, Thannickal VJ, Standiford TJ, Andrews PC, Omenn GS; Differential Protein Expression Profiling by iTRAQ-2DLC-MS/MS of Lung Cancer Cells Undergoing Epithelial-mesenchymal Transition Reveals a Migratory/Invasive Phenotype. *J. Proteome Res.* **5:1143-1154**, (2006).

- [70] Glückmann M, Fella K, Waidelich D, Merkel D, Krufft V, Kramer PJ, Walter Y, Hellmann J, Karas M, Kröger M; Prevalidation of Potential Protein Biomarkers in Toxicology Using iTRAQ Reagent Technology. *Proteomics*. **7:1564-1574**, (2007).
- [71] Guo T, Gan CS, Zhang H, Zhu Y, Kon OL, Sze SK; Hybridization of Pulsed-Q Dissociation and Collision-Activated Dissociation in Linear Ion Trap Mass Spectrometer for iTRAQ Quantitation. *J. Proteome Res*. **7:4831-4840**, (2008).
- [72] Yang F, Wu S, Stenoiien DL, Zhao R, Monroe ME, Gritsenko MA, Purvine SO, Polpitiya AD, Paša-Tolić L, Camp DG, Smith RD; Combined Pulsed-Q Dissociation and Electron Transfer Dissociation for Identification and Quantification of iTRAQ-Labeled Phosphopeptides. *Anal. Chem*. **81:4137-4143**, (2009).
- [73] Wiese S, Reidegeld KA, Meyer HE, Warscheid B; Protein Labeling by iTRAQ: A New Tool for Quantitative Mass Spectrometry in Proteome Research. *Proteomics*. **7:340-350**, (2007).
- [74] Melanson JE, Chisholm KA, Pinto DM; Targeted Comparative Proteomics by Liquid Chromatography/Matrix Assisted Laser Desorption/Ionization Triple-quadrupole Mass Spectrometry. *Rapid Commun. Mass. Spectrom*. **20:904-910**, (2006).
- [75] Rogers LD, Foster LJ; The Dynamic Phagosomal Proteome and the Contribution of the Endoplasmic Reticulum. *Proc. Natl. Acad. Sci*. **104:18520-18525**, (2007).
- [76] Ji C, Li L, Gebre M, Pasdar M, Li L; Identification and Quantification of Differentially Expressed Proteins in E-Cadherin Deficient SCC9 Cells and SCC9 Transfectants Expressing E-Cadherin by Dimethyl Isotope Labeling LC-MALDI MS and MS/MS. *J. Proteome Res*. **4:1419-1426**, (2005).
- [77] Kwok MCM, Holopainen JM, Molday LL, Foster LJ, Molday RS; Proteomics of Photoreceptor Outer Segments Identifies a Subset of SNARE and Rab Proteins Implicated in Membrane Vesicle Trafficking and Fusion. *Mol. Cell. Proteomics*. **7:1053-1066**, (2008).
- [78] Oda Y, Huang K, Cross FR, Cowburn D, Chait BT; Accurate Quantitation of Protein Expression and Site-specific Phosphorylation. *Proc. Natl. Acad. Sci*. **96:6591-6596**, (1999).
- [79] Conrads TP, Alving K, Veenstra TD, Belov ME, Anderson GA, Anderson DJ, Lipton MS, Paša-Tolić L, Udseth HR, Chrisler WB, Thrall BD, Smith RD; Quantitative Analysis of Bacterial and Mammalian Proteomes Using a



- Combination of Cysteine Affinity Tags and  $^{15}\text{N}$ -Metabolic Labeling. *Anal. Chem.* **73:2132-2139**, (2001).
- [80] Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M; Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell. Proteomics.* **1:376-386**, (2002).
- [81] Ong SE, Kratchmarova I, Mann M; Properties of  $^{13}\text{C}$ -Substituted Arginine in Stable Isotope Labeling by Amino Acid in Cell Culture (SILAC). *J. Proteome Res.* **2:173-181**, (2003).
- [82] Yao X, Freas A, Ramirez J, Demirev PA, Fenselau C; Proteolytic  $^{18}\text{O}$  Labeling for Comparative Proteomics: Model Studies with Two Serotypes of Adenovirus. *Anal. Chem.* **73:2836-2842**, (2001).
- [83] Reynolds K, Yao X, Fenselau C; Proteolytic  $^{18}\text{O}$  Labeling for Comparative Proteomics: Evaluation of Endoprotease Glu-C as the Catalytic Agent. *J. Proteome Res.* **1:27-33**, (2002).
- [84] Mirgorodskaya OA, Kozmin YP, Titov MI, Körner R, Sönksen CP, Roepstorff P; Quantitation of Peptides and Proteins by Matrix-assisted Laser Desorption/Ionization Mass Spectrometry Using  $^{18}\text{O}$ -labeled Internal Standards. *Rapid Commun. Mass Spectrom.* **14:1226-1232**, (2000).
- [85] Yao X, Afonso C, Fenselau C; Dissection of Proteolytic  $^{18}\text{O}$  Labeling: Endoprotease-catalyzed  $^{16}\text{O}$ -to- $^{18}\text{O}$  Exchange of Truncated Peptide Substrates. *J. Proteome Res.* **2:147-152**, (2003).
- [86] Bantscheff M, Dimpelfeld B, Kuster B; Femtomol Sensitivity Post-digest  $^{18}\text{O}$  Labeling for Relative Quantification of Differential Protein Complex Composition. *Rapid Commun. Mass Spectrom.* **18:869-876**, (2004).
- [87] Staes A, Demol H, VanDamme J, Martens L, Vandekerckhove J, Gevaert K; Global Differential Non-gel Proteomics by Quantitative and Stable Labeling of Tryptic Peptides with Oxygen-18. *J. Proteome Res.* **3:786-791**, (2004).
- [88] Rao KCS, Palamalai V, Dunlevy JR, Miyagi M; Peptidyl-Lys Metalloendopeptidase-catalyzed  $^{18}\text{O}$  Labeling for Comparative Proteomics. *Mol. Cell. Proteomics.* **4:1550-1557**, (2005).
- [89] Johnson KL, Muddiman DC; A Method for Calculating  $^{16}\text{O}$   $^{18}\text{O}$  Peptide Ion Ratios for the Relative Quantification of Proteomes. *J. Am. Soc. Mass Spectrom.* **15:437-445**, (2004).

- [90] Ramos-Fernández A, López-Ferrer D, Vázquez J; Improved Method for Differential Expression Proteomics Using Trypsin catalyzed 18O Labeling with a Correction for Labeling Efficiency. *Mol. Cell. Proteomics*. **6:1274-1286**, (2007).
- [91] Stemmann O, Zou H, Gerber SA, Gygi SP, Kirschner MW; Dual Inhibition of Sister Chromatid Separation at Metaphase. *Cell*. **107:715-726**, (2001).
- [92] Gerber SA, Rush J, Stemmann O, Kirschner MW, Gygi SP; Absolute Quantification of Proteins and Phosphoproteins from Cell Lysates by Tandem MS. *Proc. Natl. Acad. Sci. U S A*. **100:6940-6945**, (2003).
- [93] Hopfgartner G, Varesio E, Tschäppät V, Grivet C, Bourgoigne E, Leuthold LA; Triple Quadrupole Linear Ion Trap Mass Spectrometer for the Analysis of Small Molecules and Macromolecules. *J. Mass Spectrom*. **39:845-855**, (2004).
- [94] Kirkpatrick DS, Gerber SA, Gygi SP; The Absolute Quantification Strategy: A General Procedure for the Quantification of Proteins and Post-translational Modifications. *Methods*. **35:265-273**, (2005).
- [95] Chelius D, Bondarenko PV; Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry. *J. Proteome Res*. **1:317-323**, (2002).
- [96] Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M, Becker CH; Quantification of Proteins and Metabolites by Mass Spectrometry without Isotopic Labeling or Spiked Standards. *Anal. Chem*. **75:4818-4826**, (2003).
- [97] Wiener MC, Sachs JR, Deyanova EG, Yates NA; Differential Mass Spectrometry: A Label-Free LC-MS Method for Finding Significant Differences in Complex Peptide and Protein Mixtures. *Anal. Chem*. **76:6085-6096**, (2004).
- [98] Higgs RE, Knierman MD, Gelfanova V, Butler JP, Hale JE; Comprehensive Label-Free Method for the Relative Quantification of Proteins from Biological Samples. *J. Proteome Res*. **4:1442-1450**, (2005).
- [99] Wang G, Wu WW, Zeng W, Chou CL, Shen RF; Label-Free Protein Quantification Using LC coupled Ion Trap or FT Mass Spectrometry: Reproducibility, Linearity, and Application with Complex Proteomes. *J. Proteome Res*. **5:1241-1223**, (2006).
- [100] Washburn MP, Wolters D, Yates JR III; Large-scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology. *Nat. Biotechnol*. **19:242-247**, (2001).

- [101] Liu H, Sadygov RG, Yates JR III; A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Anal. Chem.* **76:4193-4201**, (2004).
- [102] Gilchrist A, Au CE, Hiding J, Bell AW, Fernandex-Rodriguez J, Lesimple S, Nagaya H, Roy L, Gosline SJC, Hallett M, Paiement J, Kearney RE, Nilsson T, Bergeron JM; Quantitative Proteomics Analysis of the Secretory Pathway. *Cell.* **127:1265-1281**, (2006).
- [103] Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG; Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics. *Mol. Cell. Proteomics.* **4:1487-1502**, (2005).
- [104] Rappsilber J, Ryder U, Lamond AI, Mann M; Large scale Proteomic Analysis of the Human Spliceosome. *Genome Res.* **12:1231-1245**, (2002).
- [105] Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M; Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. *Mol. Cell. Proteomics.* **4:1265-1272**, (2005).
- [106] Wall MJ, Crowell AMJ, Simms GA, Carey GH, Liu F, Doucette AA; Implications of Partial Tryptic Digestion in Organic-Aqueous Solvent Systems for Bottom-Up Proteome Analysis. *Anal. Chim. Acta.* **703:194-203**, (2011).
- [107] Pedrioli PGA, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, McComb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu W, Aebersold R; A Common Open Representation of Mass Spectrometry Data and Its Application to Proteomics Research. *Nat. Biotechnol.* **22:1459-1466**, (2004).
- [108] **Software: ReAdW – Seattle Proteome Center**  
[<http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW>]
- [109] **gVim – Text file reader** [<http://www.vim.org/download.php>]
- [110] **The Base16, Base32, and Base64 Data Encodings**  
[<http://tools.ietf.org/html/rfc3548>]
- [111] IEEE Standard for Floating-Point Arithmetic. *VDE VERLAG Conference Proceedings*. DOI: [10.1109/IEEESTD.2008.4610935](https://doi.org/10.1109/IEEESTD.2008.4610935), (2008).

- [112] Filer CN; Isotopic Fractionation of Organic Compounds in Chromatography. *J. Labelled Cpd. Radiopharm.* **42:169-197**, (1999).
- [113] Ong SE, Mann M; Mass Spectrometry-based Proteomics Turns Quantitative. *Nat. Chem. Biol.* **1:252-262**, (2005).
- [114] Boersema PJ, Raijmakers R, Lemeer S, Mohammed S, Heck AJR; Multiplex Peptide Stable Isotope Dimethyl Labeling for Quantitative Proteomics. *Nat. Protoc.* **4:484-494**, (2009).
- [115] Zhang R, Sioma CS, Wang S, Regnier FE; Fractionation of Isotopically Labeled Peptides in Quantitative Proteomics. *Anal. Chem.* **73:5142-5149**, (2001).
- [116] Vest Nielsen NP, Carstensen JM, Smedsgaard J; Aligning of Single and Multiple Wavelength Chromatographic Profiles for Chemometric Data Analysis Using Correlation Optimised Warping. *J. Chromatogr. A.* **805:17-35**, (1998).

## Appendix A

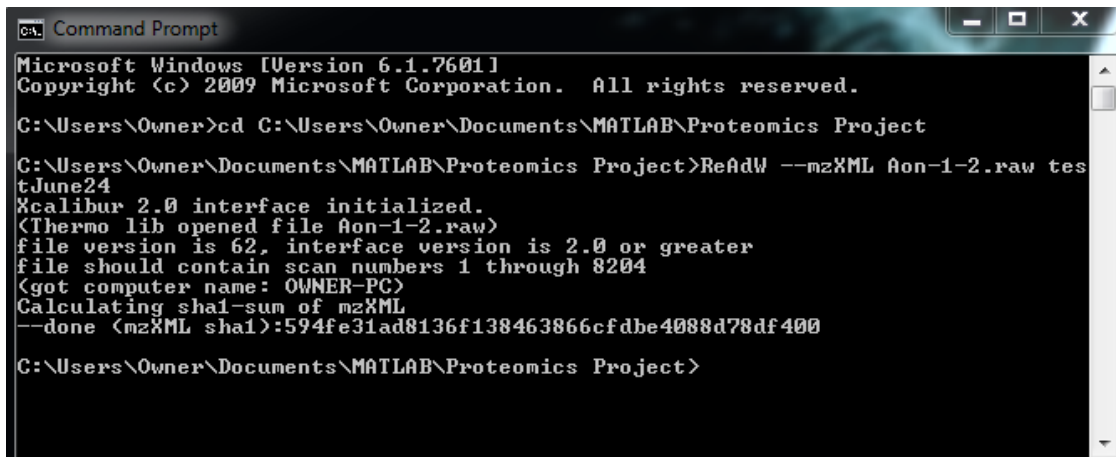
### Procedure to Convert RAW File to mzXML File

#### Installing ReAdW

- (1) Go to the Seattle Proteome Center (SPC) wikipedia website for the ReAdW software at <http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW>
- (2) Go to the Sashimi Source Forge release website using the link from the SPC website.
- (3) Download ReAdW-4.3.1.zip from the Sashimi website.
- (4) Download TPP\_Setup\_v4\_4\_VUVUZELA\_rev\_1.exe from the Sashimi website.
- (5) Go to Active State website at <http://www.activestate.com/activeperl/downloads>
- (6) Download ActivePerl-5.12.4.1205-MSWin32-x64-294981 from ActivePerl website.
- (7) Install ActivePerl-5.12.4.1205-MSWin32-x64-294981
- (8) Install TPP\_Setup\_v4\_4\_VUVUZELA\_rev\_1.exe
- (9) Transfer all files from ReAdW-4.3.1.zip into the folder C:\inetpub\tpp-bin

#### Running ReAdW

- (1) Open Command Prompt on the computer.
- (2) Find the directory that the RAW file is in that is to be converted.
- (3) Convert RAW file using ReAdW in the command prompt using the input:  
ReAdW --mzXML\_filename.raw\_savename( \_ = space)



```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\Owner>cd C:\Users\Owner\Documents\MATLAB\Proteomics Project

C:\Users\Owner\Documents\MATLAB\Proteomics Project>ReAdW --mzXML Aon-1-2.raw testJune24
Xcalibur 2.0 interface initialized.
(Thermo lib opened file Aon-1-2.raw)
file version is 62, interface version is 2.0 or greater
file should contain scan numbers 1 through 8204
(got computer name: OWNER-PC)
Calculating sha1-sum of mzXML
--done (mzXML sha1):594fe31ad8136f138463866cfdbe4088d78df400

C:\Users\Owner\Documents\MATLAB\Proteomics Project>
```

## Appendix B

### base64cvt.m MatLab<sup>®</sup> Code

```
function byteout=base64cvt(charlst)

b64codes=ones(122,1)*-1;
b64set=[65:90 97:122 48:57 43 47]';
b64vals=[0:63]';
b64codes(b64set)=b64vals;

[pos]=find(charlst ~= '');
charlst = charlst(pos);
nchar=length(charlst);
nsets=floor(nchar/4);
nleft=nchar-nsets*4;
bytelst=[];

for i=1:nsets
    indx=(i-1)*4+1;
    charset=double(charlst(indx:indx+3));
    hexset=uint8(b64codes(charset));
    tmp1=uint8(bitshift(hexset(1),2)+bitshift(hexset(2),-4));
    tmp2=uint8(bitshift(hexset(2),4)+bitshift(hexset(3),-2));
    tmp3=uint8(bitshift(hexset(3),6)+hexset(4));
    bytelst=[bytelst uint8([tmp1 tmp2 tmp3])];
end

if nleft==2
    charset=double(charlst(nchar-1:nchar));
    hexset=uint8(b64codes(charset));
    tmp1=uint8(bitshift(hexset(1),2)+bitshift(hexset(2),-4));
    bytelst=[bytelst uint8(tmp1)];
elseif nleft==3
    charset=double(charlst(nchar-2:nchar));
    hexset=uint8(b64codes(charset));
    tmp1=uint8(bitshift(hexset(1),2)+bitshift(hexset(2),-4));
    tmp2=uint8(bitshift(hexset(2),4)+bitshift(hexset(3),-2));
    bytelst=[bytelst uint8([tmp1 tmp2])];
end
byteout=bytelst;
```

## Appendix C

### rawexportfull.m MatLab<sup>®</sup> Code

```
function [dat]=rawexportfull(mzXML,savename)
% mzXML and savename needs to be string variables
% mzXML is the filename of the file being exported
% savename is the filename that you want to save the data as

fid = fopen(mzXML,'r+'); % read in mzXML data file

main = textscan(fid,'%c',170000); main=main{1}';
[a,stop,c,d,tokens] = regexp(main,'<?xmlversion="(.*?)"</dataProcessing>');
header = tokens{1}{1,1};
main=main(stop+1:end);

[a, b, c, d, tokens] = regexp(header,'fileName="(.*?)"');
dat.fileName=tokens{1}{1,1};
[a, b, c, d, tokens] = regexp(header,'scanCount="(.*?)"');
dat.scanCount=str2num(tokens{1}{1,1});

ScanDone = 0;
FulCount=1;

while ~feof(fid) % end-of-file indicator
    keepgoing=false;
    [a, b, c, match] = regexp(main,'<scannum="(.*?)"</peaks>');
    if isempty(match)
        keepgoing = true;
    end
    while keepgoing
        current = main;
        TEMP = textscan(fid,'%c',20000);
        TEMP=TEMP{1}';
        current = [main,TEMP];
        [a, b, c, match] = regexp(current,'<scannum="(.*?)"</peaks>');
        if length(match) ~= 0
            keepgoing = false;
        end
        if isempty(match) && isempty(TEMP)
            keepgoing = false;
        end
        main = current;
    end
    [a, stop, c, match] = regexp(main,'<scannum="(.*?)"</peaks>');
```

```

nscan=length(match);
if nscan ~= 0
    for iscan = 1:nscan;

        [a, b, c, d, tokens] = regexp(match{1,iscan},'msLevel="(.*?)"');
        msLevel=str2num(tokens{1}{1,1});

        [a, b, c, d, tokens] = regexp(match{1,iscan},'scanType="(.*?)"');
        scanType=tokens{1}{1,1};

        msLevelCheck=msLevel==1;
        scanTypeCheck=strncmp(scanType,'Full',4);
        if msLevelCheck && scanTypeCheck == true

            [a, b, c, d, tokens] =
regexp(match{1,iscan},'compressedLen="0">(.*?)</peaks>');
            rawdata=tokens{1}{1,1};
            base64 = base64cvt(rawdata);
            [row,col]=size(base64);
            indx=[1:4:col-1];
            for i = 1:length(indx)
                start2=indx(i); stop2=indx(i)+3;
                DATA(1,i)=typecast(fliplr(base64(start2:stop2)),'single');
            end

            [a, b, c, d, tokens] = regexp(match{1,iscan},'scannum="(.*?)"');
            curscan=str2num(tokens{1}{1,1});
            dat.scanNum(1,FulCount)=curscan;

            [a, b, c, d, tokens] = regexp(match{1,iscan},'retentionTime="PT(.*?)S"');
            dat.rTime(1,FulCount)=str2num(tokens{1}{1,1});

            dat.intensity(:,FulCount)=DATA([2:2:length(indx)]);
            FulCount=FulCount+1;
            if FulCount == 2
                dat.massCharge(:,1)=DATA([1:2:length(indx)-1]);
            end
        end
    end
    ScanDone=ScanDone+nscan
    main=main(stop(1,nscan)+1:end);
end
end
fclose(fid)
save(savename,'dat')

```



## Appendix D

### Complete List of 71 BSA Peptide Pairs Used in Chapter 4

Peptide Sequence	Mass of Light	Mass of Heavy	Charge State	# of Labels	Time Diff (sec)	Resolution
A.EFVEVTK.L	454.33	458.25	2	2	3.736	0.191
C.CAADDKEACFAVEGPK.L	618.00	622.00	3	3	0.930	0.060
C.FSALTPDETYVVPK.A	762.41	766.41	2	2	4.142	0.089
D.FAEDKDVCK.N	598.33	604.33	2	3	11.26	0.535
F.HADICTLPDTEK.Q	728.41	732.33	2	2	10.01	0.293
F.LGSFLYEYSR.R	631.83	633.83	2	1	1.796	0.085
F.SALTPDETYVVPK.A	688.91	692.83	2	2	2.489	0.109
F.SQYLQQCPFDEH.V	790.41	792.33	2	1	-0.415	-0.018
F.SQYLQQCPFDEHVK.L	612.33	615.00	3	2	5.255	0.188
F.TFHADICTLPDTEK.Q	568.66	571.33	3	2	2.858	0.195
F.YAPELLYANK.Y	700.91	704.91	2	2	1.800	0.092
G.VFQECCQAEDK.G	735.33	739.33	2	2	4.418	0.182
H.ACYSTVFDK.L	573.83	577.75	2	2	0.276	0.015
H.ADICTLPDTEK.Q	659.83	663.83	2	2	1.698	0.098
H.CIAEVEK.D	452.75	456.75	2	2	14.97	0.409
K.AEFVEVTK.L	489.83	493.75	2	2	11.71	0.251
K.CCAADDKEACFAVEGPK.L	671.33	675.33	3	3	4.523	0.140
K.CCTESLVNR.R	583.83	585.83	2	1	2.890	0.096
K.DAIPENLPLTAD.F	697.41	699.41	2	1	1.282	0.036
K.DAIPENLPLTADFAEDK.D	671.41	674.00	3	2	1.485	0.074
K.DDPHACYSTVF.D	670.33	672.33	2	1	1.426	0.072
K.DDPHACYSTVFDK.L	537.58	540.25	3	2	3.585	0.148
K.DDSPDLPK.L	471.75	475.75	2	2	11.65	0.145
K.DLGEEHFK.G	515.75	519.75	2	2	9.588	0.162
K.EACFAVEGPK.L	582.33	586.33	2	2	2.954	0.112
K.ECCDKP LLEK.S	688.33	694.33	2	3	6.934	0.254
K.ECCHGDLLECADDR.A	593.25	594.58	3	1	1.450	0.038
K.ECCHGDLLECADDRADLAK.Y	576.75	578.75	4	2	-0.336	-0.025
K.EYEATLECCA.K.D	779.83	783.83	2	2	6.524	0.157
K.GACLLPK.I	407.75	411.75	2	2	3.367	0.090
K.HLVDEPQNLIK.Q	681.41	685.41	2	2	21.67	0.494
K.KQTALVELLK.H	613.91	619.91	2	3	9.096	0.201
K.LFTFHADICTLPDTEK.Q	655.33	658.00	3	2	1.190	0.055
K.LGEYGFQNAL.I	570.33	572.33	2	1	1.367	0.079

Peptide Sequence	Mass of Light	Mass of Heavy	Charge State	# of Labels	Time Diff (sec)	Resolution
K.LGEYGFQNALIVR.Y	754.41	756.41	2	1	2.021	0.081
K.LKPDNTLCDEF.K	752.91	756.91	2	2	0.801	0.042
K.LKPDNTLCDEFK.A	554.33	558.33	3	3	7.831	0.271
K.LVNELTEFAK.T	610.33	614.33	2	2	8.935	0.184
K.LVTDLTK.V	423.33	427.25	2	2	5.814	0.243
K.QEPERNECFLSHK.D	433.25	435.25	4	2	4.894	0.138
K.QTALVELLK.H	535.83	539.83	2	2	5.703	0.183
K.SHCIAEVEK.D	564.83	568.83	2	2	15.91	0.234
K.SLHTLFGDELCK.V	492.58	495.25	3	2	7.121	0.422
K.TVM*ENFVAFVDK.C	736.41	740.41	2	2	1.825	0.119
K.TVMENFVAFVDK.C	728.41	732.41	2	2	2.534	0.060
K.VPQVSTPTLVEVSR.S	770.50	772.50	2	1	0.176	0.013
K.YICDNQDTISSK.L	750.41	754.33	2	2	10.48	0.229
K.YLYEIAR.R	478.33	480.33	2	1	4.344	0.123
K.YNGVFQECCQAEDK.G	902.41	906.41	2	2	25.53	1.390
L.TPDETYVPK.A	553.33	557.33	2	2	2.811	0.106
L.VNELTEFAK.T	553.83	557.83	2	2	2.682	0.117
N.FVAFVDK.C	441.33	445.25	2	2	4.435	0.173
N.TLCDEFK.A	484.75	488.75	2	2	3.519	0.177
P.CFSALTPDETYVPK.A	842.50	846.41	2	2	3.490	0.121
Q.TALVELLK.H	471.83	475.83	2	2	1.748	0.092
Q.VSTPTLVEVSR.S	608.41	610.33	2	1	0.667	0.034
Q.YLQQCPFDEHVK.L	540.66	543.25	3	2	1.398	0.075
R.ETYGDMADCCEK.Q	767.83	771.83	2	2	2.283	0.102
R.FKDLGEEHFK.G	445.25	449.25	3	3	5.204	0.310
R.HPEYAVSVLLR.L	656.41	658.41	2	1	7.181	0.216
R.KVPQVSTPTLVEVSR.S	566.00	568.66	3	2	5.113	0.194
R.LCVLHEK.T	477.83	481.75	2	2	4.223	0.110
R.NECFLSHK.D	545.83	549.75	2	2	9.071	0.321
R.RHPEYAVSVL.L	599.83	601.83	2	1	0.803	0.040
R.RHPEYAVSVLLR.L	490.00	491.33	3	1	0.002	-0.019
R.RPCFSALTPDETYVPK.A	646.33	649.00	3	2	3.091	0.096
S.ALTPDETYVPK.A	645.41	649.33	2	2	1.660	0.098
T.ALVELLK.H	421.33	425.33	2	2	5.853	0.271
Y.FYAPPELLYYAN.K	696.33	698.33	2	1	0.799	0.035
Y.FYAPPELLYYANK.Y	774.41	778.41	2	2	2.000	0.044
Y.GFQNALIVR.Y	523.33	525.33	2	1	0.860	0.050

## Appendix E

### Complete list of 535 peptide pairs found by PITS

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
402.4167	2	2	532	474	554	464
407.3333	2	1	571	495	603	505
414.75	2	1	808	683	874	706
421.4375	2	2	1040	927	1124	932
429.3958	1	1	876	778	951	803
438	3	1	1084	949	1179	953
440.75	2	1	1235	1106	1316	1095
441.3542	1	1	571	495	603	508
441.4167	2	2	815	697	882	723
442.3333	1	1	927	829	1005	841
449.9583	2	2	495	437	499	417
452.3333	2	1	578	504	611	513
452.9167	2	2	372	371	346	314
454.375	2	2	589	523	631	543
454.6875	3	2	828	676	869	703
455	3	2	864	724	932	754
458.3958	2	2	625	543	666	568
459.3542	2	1	1003	882	1088	897
459.6667	3	2	733	596	785	639
460.4167	2	2	1018	891	1104	905
460.75	2	1	1276	1127	1346	1107
471.25	2	1	808	681	874	706
472.8542	2	1	511	454	523	433
475.875	2	1	375	355	346	319
477.9167	2	2	571	495	607	508
478.4167	2	1	772	660	831	686
480.2708	2	1	842	720	913	746
481.3542	1	1	449	443	427	422
484.8542	2	2	671	561	713	591
485.9167	3	2	582	498	622	513
486.3333	1	1	564	520	586	525
489.8958	2	2	620	537	661	568
496.75	2	1	1068	949	1158	946
496.9167	2	2	752	614	808	645
504.7708	2	1	1090	967	1174	961

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
508.6667	3	2	913	773	994	803
510.9167	2	2	302	284	261	268
514	3	1	868	726	942	756
514.4167	1	1	372	345	366	289
515.75	2	1	808	683	874	706
515.8333	2	2	544	484	581	474
518.9167	2	2	1058	933	1155	934
521.75	2	1	897	771	968	803
522.2708	2	1	1321	1161	1389	1144
524.3959	2	1	701	587	744	620
530.2708	2	1	1199	1071	1277	1064
534.9792	2	2	1232	1095	1318	1084
536.4376	2	2	1185	1054	1259	1053
537.6876	3	2	761	634	815	669
538.375	2	1	501	443	509	422
542.9167	2	2	1177	1031	1250	1027
543.2292	2	1	1276	1127	1344	1107
543.9584	2	3	496	437	499	418
545.8959	2	2	507	454	515	433
546.8333	2	1	726	598	777	636
553.9167	2	2	820	697	879	718
554.4167	1	1	571	495	603	506
555.375	1	1	1099	991	1183	991
562.3333	3	1	646	525	691	568
563.9167	2	2	452	411	460	378
568.6667	3	2	853	712	921	744
570.2708	2	1	1090	967	1175	961
573.4167	2	1	490	431	497	406
573.9167	2	2	745	630	799	657
576.4167	2	1	719	605	769	639
582.4167	2	2	649	544	693	575
584.3541	2	1	498	425	488	400
598	3	1	722	593	769	632
598.4792	1	2	1063	921	1140	936
599.8333	2	1	888	745	953	779
602	3	2	705	577	750	617
603.4167	1	1	371	323	368	296
608.4167	2	1	776	649	832	680
611.7292	3	2	863	720	933	749

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
612.3959	3	2	798	664	858	691
613.9376	2	2	1159	1018	1240	1015
614	2	3	1075	921	1140	932
617.4375	2	2	1148	1010	1232	1009
631.9167	2	1	1084	965	1172	967
645.4167	2	2	723	591	771	632
656.3124	3	2	754	647	843	677
656.4376	2	1	1084	954	1177	954
659.8959	2	2	654	537	698	573
662.75	3	3	787	639	840	671
665.4167	2	2	606	521	645	543
670.3124	2	1	838	722	904	748
671.4167	3	3	695	561	737	604
680.4376	2	2	876	734	952	762
681.9375	2	2	864	721	933	752
688.4376	2	3	521	460	538	445
688.9167	2	2	733	596	785	639
696.25	2	1	1367	1181	1426	1167
697.2708	2	1	999	860	1089	872
700.9167	2	2	1063	937	1156	937
706.9167	2	2	953	825	1044	840
716.4167	2	2	1060	932	1152	932
727.4167	1	1	1120	1001	1205	1001
735.375	2	2	409	374	402	327
750.9167	2	2	494	427	501	402
752.4792	1	2	299	342	204	318
756.7709	2	1	744	612	798	647
758.4167	2	2	438	381	445	337
762.4792	2	2	913	773	989	803
770.4167	2	1	868	726	941	756
779.9167	2	2	589	504	638	527
784.3334	1	1	1153	1030	1237	1024
805.9167	2	2	761	633	815	667
814.5	1	2	588	524	625	543
830.9792	2	3	888	743	967	779
842.4167	2	2	972	831	1050	844
846.5	2	2	1018	879	1107	893
857.4167	2	2	744	618	797	650
879.8125	2	1	827	698	893	722

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
902.4167	2	2	706	577	751	618
903.3959	2	2	749	624	792	655
913.4167	2	2	844	708	913	739
983.9167	2	2	778	647	843	679
401.8333	2	1	NaN	911	1116	922
424.25	2	1	830	674	NaN	702
445.4167	3	3	741	NaN	792	643
459.3611	3	2	NaN	764	979	796
459.75	2	2	564	NaN	603	509
483.3889	3	1	1145	977	1232	NaN
485.3611	2	2	NaN	626	793	654
497.2778	2	1	1052	933	NaN	936
503.9167	2	1	236	296	NaN	278
515.4167	2	1	666	NaN	709	585
515.4167	1	1	837	672	NaN	700
533.3333	2	1	482	426	NaN	400
540.75	3	2	747	622	801	NaN
541.3611	2	1	1001	878	NaN	893
541.9167	2	1	1301	1144	NaN	1124
546.4167	1	1	422	394	NaN	361
551.25	2	1	1075	958	NaN	953
553.4167	2	2	448	408	458	NaN
565.3889	1	1	569	NaN	596	518
570.6945	3	2	NaN	816	1039	831
570.75	2	1	NaN	982	1192	980
577.3889	3	2	549	NaN	573	481
579.25	1	1	506	454	NaN	436
582.9167	2	2	NaN	580	739	617
591.5	2	2	1337	1165	NaN	1152
592.6667	3	1	746	617	799	NaN
593.3333	3	1	NaN	548	708	577
596.3611	1	1	295	371	NaN	336
598.4167	2	3	NaN	284	259	268
600.3611	3	2	859	717	927	NaN
603.75	2	1	NaN	1263	1529	1257
605.4167	1	1	538	480	NaN	468
609.4445	3	2	844	708	NaN	739
615.3333	1	1	569	513	590	NaN
617.4167	1	1	584	498	622	NaN

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
621.8333	2	1	1026	896	NaN	907
622.4167	1	1	833	677	890	NaN
627.4445	3	3	1029	878	1121	NaN
654.3333	1	1	NaN	432	454	406
663.0555	3	2	NaN	712	922	744
671.75	3	3	NaN	577	765	624
693.9167	2	2	578	NaN	615	521
710.3333	1	1	819	715	NaN	739
727.4167	2	2	727	595	NaN	635
734.3611	2	2	646	NaN	688	564
740.3333	3	2	NaN	874	1107	891
746.4167	2	2	NaN	1008	1238	1008
756.3611	1	1	841	NaN	908	744
768.3333	1	1	793	NaN	852	701
769.4167	2	2	NaN	867	1091	882
775.3611	2	2	NaN	341	337	300
786.9167	2	2	NaN	577	751	619
817.4167	1	1	423	NaN	397	364
848.5	2	2	NaN	726	949	756
880.4167	1	1	NaN	1106	1315	1091
887.3333	1	1	1047	NaN	1134	927
889.3333	2	1	NaN	547	705	581
969	2	2	NaN	818	1066	834
1025.583	1	1	NaN	1192	1430	1179
1110	2	2	NaN	875	1108	890
409.75	2	2	NaN	922	1139	936
412.5	2	2	717	577	763	NaN
412.25	2	2	678	NaN	720	NaN
448.7083	3	2	511	449	NaN	NaN
448.9167	3	2	NaN	NaN	849	NaN
478.4167	2	2	NaN	210	NaN	191
478.4167	2	2	NaN	NaN	639	NaN
569.8611	2	2	203	298	NaN	279
569.9167	2	2	NaN	771	NaN	NaN
616	3	2	NaN	914	NaN	925
616	3	2	NaN	NaN	1325	NaN
646.5833	3	2	NaN	865	NaN	882
646.5833	3	2	1378	NaN	NaN	NaN
658.4167	1	1	NaN	846	1032	NaN

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
658.4167	1	1	591	NaN	628	NaN
661.4167	3	2	NaN	795	NaN	819
699.4584	2	2	NaN	1143	NaN	1122
738.8889	2	2	504	435	NaN	414
738.9167	2	2	NaN	NaN	NaN	1087
749.3333	1	1	NaN	1106	1315	1091
749.3611	1	1	NaN	1161	1390	1144
749.3334	1	1	NaN	NaN	1481	NaN
767.8333	2	2	NaN	433	507	409
1006.417	2	2	NaN	1054	1278	1047
405.75	2	1	808	NaN	874	NaN
412.1667	3	3	NaN	544	NaN	578
416.125	3	3	644	523	NaN	NaN
430.4167	2	2	NaN	273	NaN	266
430.6667	3	2	723	NaN	772	NaN
439.75	2	1	NaN	1070	NaN	1065
449	3	2	833	677	NaN	NaN
458.75	2	2	659	NaN	701	NaN
463.9583	2	3	449	NaN	453	NaN
470.875	2	2	NaN	838	1042	NaN
479.4583	2	2	478	NaN	479	NaN
486.75	3	2	NaN	NaN	763	620
491.4167	1	1	NaN	882	NaN	897
492.3333	2	1	NaN	1010	NaN	1007
493	3	2	NaN	NaN	1374	1086
493.8333	2	1	NaN	210	NaN	196
497.4167	2	2	NaN	712	NaN	740
498.0833	3	2	NaN	1008	1238	NaN
500.6667	3	2	NaN	NaN	423	337
506.8333	2	2	520	NaN	538	NaN
509.3333	1	1	NaN	312	NaN	323
516.4167	1	1	NaN	598	NaN	630
516.6667	3	2	1423	NaN	1481	NaN
529.875	2	1	NaN	625	NaN	650
531.4167	2	1	NaN	294	NaN	276
532.2916	2	1	NaN	965	NaN	961
561.8333	2	1	NaN	961	NaN	956
562.9167	2	2	NaN	869	NaN	886
566.3333	3	2	940	NaN	NaN	813



Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
566.875	2	2	NaN	468	NaN	452
572.875	2	2	NaN	278	NaN	264
573.8333	2	1	NaN	1101	NaN	1087
590.4167	1	1	NaN	1066	NaN	1060
605.4167	2	2	712	NaN	NaN	633
607.4167	1	1	514	213	NaN	NaN
612.7084	3	2	NaN	715	NaN	746
613.4167	1	1	NaN	965	NaN	961
614.4167	1	1	NaN	776	NaN	803
622.4167	2	2	421	NaN	NaN	335
622.75	2	1	NaN	979	NaN	977
637	3	1	NaN	815	NaN	833
639.2916	2	1	NaN	1201	NaN	1197
639.7916	2	1	NaN	890	NaN	905
641.0833	3	3	746	621	NaN	NaN
648.4167	3	2	864	720	NaN	NaN
651.6667	3	2	NaN	810	1030	NaN
653.4167	1	1	569	NaN	603	NaN
654.3333	3	3	689	NaN	NaN	601
661.9167	2	2	NaN	378	NaN	334
668.9167	2	2	NaN	252	NaN	253
670.75	3	3	NaN	626	NaN	660
674.4167	1	1	NaN	288	NaN	266
677.9167	2	2	NaN	1078	NaN	1069
680.2916	2	1	NaN	875	NaN	891
687.4167	2	3	660	NaN	702	NaN
704.9167	2	2	560	NaN	595	NaN
733.375	1	1	NaN	861	NaN	878
734.4584	2	1	NaN	922	NaN	941
752.9167	2	2	969	NaN	NaN	840
762.3334	1	1	NaN	854	NaN	870
774.4167	1	1	NaN	NaN	580	481
809.7917	3	3	NaN	1175	1405	NaN
813.4584	1	1	NaN	NaN	607	506
853.4167	3	3	1306	NaN	1344	NaN
862.4167	1	1	NaN	1258	NaN	1260
955.4167	1	1	NaN	649	NaN	675
966.9584	2	2	NaN	965	NaN	961
1019.417	1	1	NaN	1091	NaN	1080

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
1078.5	1	1	970	NaN	1049	NaN
1090.5	1	2	NaN	454	NaN	435
415.875	2	2	577	NaN	608	NaN
415.75	2	2	NaN	NaN	797	NaN
444.9167	2	2	NaN	189	NaN	196
444.9167	2	2	NaN	296	NaN	279
444.9167	2	2	NaN	NaN	NaN	138
529.25	2	1	NaN	NaN	1522	NaN
529.4167	2	1	NaN	NaN	NaN	269
540.2916	2	1	NaN	298	NaN	277
553.5	3	3	1436	NaN	NaN	NaN
553.5	3	3	NaN	857	NaN	NaN
584.75	3	1	1444	NaN	NaN	NaN
584.75	3	1	NaN	938	NaN	NaN
613.6667	2	1	NaN	1039	NaN	1032
613.75	2	1	NaN	NaN	NaN	903
682.9584	2	1	NaN	1218	NaN	1216
839.375	1	1	NaN	NaN	712	591
839.4167	1	1	NaN	NaN	NaN	519
401.8333	2	2	292	NaN	NaN	NaN
408.3333	1	1	NaN	308	NaN	NaN
408.3333	2	2	NaN	570	NaN	NaN
409.25	2	1	NaN	NaN	395	NaN
409.75	3	3	1063	NaN	NaN	NaN
410.3333	2	1	NaN	NaN	515	NaN
411.8333	2	2	NaN	754	NaN	NaN
415.6667	3	3	758	NaN	NaN	NaN
425.25	2	1	NaN	1127	NaN	NaN
437.3333	1	1	NaN	NaN	NaN	314
439.4167	1	1	NaN	640	NaN	NaN
439.4167	2	2	NaN	NaN	808	NaN
443.5833	3	1	NaN	517	NaN	NaN
445.5	2	2	NaN	610	NaN	NaN
450.3333	2	1	NaN	403	NaN	NaN
452.9167	2	1	NaN	NaN	401	NaN
458.0833	3	1	NaN	NaN	523	NaN
460.3333	2	1	NaN	NaN	NaN	302
462.8333	2	2	NaN	NaN	NaN	495
464.0833	2	2	NaN	NaN	679	NaN

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
468.3333	2	1	NaN	513	NaN	NaN
469.75	2	1	763	NaN	NaN	NaN
483.4167	2	1	NaN	NaN	NaN	1008
484.8333	2	1	NaN	NaN	NaN	777
485.4167	1	2	1059	NaN	NaN	NaN
485.4167	2	1	NaN	NaN	NaN	770
486.4167	3	2	651	NaN	NaN	NaN
488.9167	2	2	NaN	435	NaN	NaN
501.5	3	2	1321	NaN	NaN	NaN
502.3333	1	1	565	NaN	NaN	NaN
506.8333	2	1	NaN	NaN	NaN	705
507.3333	1	1	NaN	NaN	NaN	423
510.8333	2	3	NaN	NaN	1177	NaN
511.6667	3	2	NaN	NaN	NaN	1010
514.8333	2	2	622	NaN	NaN	NaN
516.25	2	1	NaN	721	NaN	NaN
520.25	3	2	NaN	NaN	632	NaN
520.4167	2	1	NaN	1241	NaN	NaN
525.3333	3	1	NaN	NaN	NaN	500
532	3	1	738	NaN	NaN	NaN
535.4167	2	2	NaN	737	NaN	NaN
538	3	2	NaN	NaN	860	NaN
538.8333	2	1	NaN	NaN	NaN	495
543.75	2	1	NaN	1150	NaN	NaN
548.3333	2	1	NaN	863	NaN	NaN
552.9167	2	2	NaN	NaN	1375	NaN
557.3333	1	1	NaN	790	NaN	NaN
557.6667	3	1	NaN	NaN	1150	NaN
559	3	3	NaN	NaN	1148	NaN
562.9167	3	1	NaN	570	NaN	NaN
576.9167	2	1	NaN	NaN	NaN	648
577.1667	3	3	NaN	NaN	1140	NaN
577.6667	3	2	NaN	NaN	NaN	492
578.3333	1	1	NaN	NaN	NaN	314
580.4167	1	1	NaN	279	NaN	NaN
580.4167	2	1	NaN	1171	NaN	NaN
583.25	2	1	NaN	1056	NaN	NaN
583.3333	2	2	NaN	NaN	NaN	439
583.4167	3	2	NaN	673	NaN	NaN

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
584.4167	2	2	NaN	628	NaN	NaN
590.3333	3	1	NaN	NaN	1542	NaN
593.1667	3	3	NaN	NaN	1295	NaN
593.4167	1	1	NaN	NaN	380	NaN
598.5833	3	1	647	NaN	NaN	NaN
599.4167	1	1	NaN	798	NaN	NaN
599.4167	2	2	NaN	896	NaN	NaN
600.4167	2	1	1173	NaN	NaN	NaN
604.75	2	1	NaN	901	NaN	NaN
612.1667	2	1	NaN	NaN	1527	NaN
622.3333	2	1	NaN	873	NaN	NaN
624.3333	2	2	NaN	NaN	702	NaN
632.75	3	3	729	NaN	NaN	NaN
635.3333	2	1	NaN	313	NaN	NaN
636.5833	3	2	NaN	634	NaN	NaN
641.4167	1	1	NaN	NaN	NaN	822
641.9167	2	3	NaN	NaN	757	NaN
642.9167	2	3	NaN	NaN	NaN	495
647.4167	1	1	NaN	NaN	NaN	317
651	3	2	NaN	927	NaN	NaN
656.5833	3	2	NaN	695	NaN	NaN
660.3333	2	1	NaN	1083	NaN	NaN
660.75	3	2	NaN	NaN	1408	NaN
664.8333	2	1	NaN	517	NaN	NaN
669.4167	1	1	NaN	999	NaN	NaN
671.8333	2	2	719	NaN	NaN	NaN
672.25	2	1	NaN	931	NaN	NaN
675.3333	2	1	767	NaN	NaN	NaN
676.3333	2	1	NaN	NaN	NaN	347
677.3333	1	1	NaN	NaN	NaN	654
680.6667	3	2	NaN	NaN	1023	NaN
687	2	2	1160	NaN	NaN	NaN
698.4167	2	1	NaN	NaN	NaN	906
701	3	2	NaN	NaN	1373	NaN
709.9167	2	1	NaN	1083	NaN	NaN
713.3333	1	1	941	NaN	NaN	NaN
715.3333	1	1	NaN	NaN	NaN	1066
720.75	2	1	NaN	1322	NaN	NaN
722.75	3	2	NaN	NaN	1464	NaN

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
729.4167	1	1	NaN	NaN	NaN	550
736.9167	2	2	NaN	NaN	NaN	1119
747.3334	3	3	1254	NaN	NaN	NaN
757.4167	1	1	NaN	NaN	NaN	706
763.9167	2	2	554	NaN	NaN	NaN
767.4167	1	1	NaN	NaN	NaN	878
768.8334	3	3	1268	NaN	NaN	NaN
770.4167	2	2	NaN	803	NaN	NaN
774.8334	3	2	NaN	NaN	1532	NaN
774.8334	2	2	NaN	NaN	NaN	493
778.4167	1	1	NaN	NaN	NaN	649
779.3334	2	1	NaN	NaN	825	NaN
786.5	3	2	NaN	NaN	1406	NaN
791.3334	1	2	NaN	NaN	NaN	874
796.4167	1	1	NaN	445	NaN	NaN
798.4167	2	1	1361	NaN	NaN	NaN
808	2	2	NaN	NaN	NaN	941
810.3334	1	1	NaN	NaN	NaN	706
827.3334	1	1	NaN	NaN	NaN	947
828.3334	1	1	NaN	687	NaN	NaN
830.4167	1	1	NaN	NaN	NaN	998
841.4167	1	1	NaN	NaN	NaN	724
845.5834	1	2	536	NaN	NaN	NaN
858.4167	1	1	NaN	NaN	NaN	280
859.4167	1	1	NaN	647	NaN	NaN
868.4167	1	1	NaN	NaN	NaN	883
885	2	1	NaN	1256	NaN	NaN
893.9167	2	2	NaN	642	NaN	NaN
906.3334	1	1	NaN	923	NaN	NaN
906.8334	3	3	1331	NaN	NaN	NaN
954.5	1	2	NaN	496	NaN	NaN
968.4167	1	2	NaN	NaN	NaN	593
978.5	1	2	NaN	NaN	NaN	572
992.4167	1	1	1068	NaN	NaN	NaN
1006.5	2	3	NaN	NaN	NaN	605
1028.333	1	1	NaN	873	NaN	NaN
1038.333	1	1	NaN	NaN	1198	NaN
1128.5	1	2	NaN	NaN	NaN	296
451.4167	1	1	NaN	577	NaN	NaN

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
896.4167	1	1	NaN	NaN	NaN	951
896.4167	1	1	NaN	NaN	NaN	1018
403.4167	1	1	372	323	369	296
403.3333	1	1	NaN	NaN	432	NaN
407.9167	2	2	586	530	625	547
413.25	1	1	372	323	369	296
418.375	1	1	402	413	381	384
418.3333	1	1	NaN	514	NaN	NaN
423.4167	2	2	538	480	560	466
423.7708	2	1	518	462	532	439
429.8333	2	1	253	301	205	282
435.3333	2	1	1050	922	1135	928
435.3333	2	1	NaN	404	NaN	NaN
444.3333	2	1	547	485	575	477
444.25	2	1	1049	922	1135	928
448.75	2	1	1068	950	1157	949
448.8333	2	1	NaN	NaN	NaN	1092
457.3333	3	1	848	703	NaN	NaN
459.4167	2	2	525	460	538	444
471.8958	2	2	335	348	271	301
471.9792	2	2	1180	1039	1254	1033
490.0833	3	1	1112	919	1174	936
492.6667	3	2	1177	1023	1252	1013
492.7083	3	2	1425	1239	1478	1231
510.2708	2	1	1224	1091	1300	1078
510.3333	2	1	271	298	NaN	279
510.3611	2	1	NaN	390	417	353
512	3	2	383	355	378	300
513.3541	2	1	1361	1192	1432	1182
513.25	2	1	NaN	1261	1529	1257
519.3333	2	2	793	663	851	695
519.4167	2	2	480	NaN	NaN	393
521.4167	2	2	682	570	723	604
521.4167	2	2	NaN	279	NaN	265
523.4167	2	1	885	750	958	781
523.2292	2	1	737	620	791	648
523.9167	2	1	939	827	1017	842
523.9167	2	1	NaN	744	NaN	775
523.9167	2	1	NaN	311	305	NaN

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
530.4167	1	1	1357	1192	1429	1179
530.4376	1	1	1063	921	1143	936
532.4167	1	1	372	322	369	296
533.9376	3	3	966	808	1034	824
534.125	3	3	861	NaN	929	NaN
535.9792	2	2	1112	991	1210	991
546.3333	2	2	581	NaN	614	513
546.3333	2	2	NaN	472	563	458
554.3958	3	3	890	743	966	779
554.7292	3	3	861	720	930	750
554.75	3	3	980	823	NaN	837
564.9167	2	2	369	320	368	296
566.0208	3	2	870	728	948	756
566.0416	3	2	1142	NaN	1146	NaN
583.9167	2	1	427	389	429	350
583.8333	2	1	NaN	288	NaN	272
602.3333	3	2	717	585	765	624
609.9167	2	2	523	462	538	444
609.9167	2	2	690	564	733	604
610.4792	2	2	962	836	1049	851
610.9584	2	2	927	816	1011	831
610.9584	2	2	1008	878	1093	895
613	2	3	1200	1065	NaN	1063
613.0833	2	3	1390	NaN	1452	NaN
628.7292	2	1	1119	991	1199	984
628.6667	2	1	NaN	466	NaN	NaN
629.5	1	1	1358	1192	1430	1182
629.4584	1	1	NaN	922	NaN	940
636.3333	1	1	1236	1104	1315	1095
636.75	2	1	1012	882	1100	895
636.75	2	1	NaN	NaN	NaN	985
637.4375	1	1	299	348	206	318
637.3333	1	1	NaN	556	NaN	NaN
646.3333	3	2	984	823	1064	835
646.3124	3	2	1227	1096	1271	1074
655.4167	3	2	1389	1210	1441	1206
655.4167	3	2	NaN	1068	1323	1063
671.3125	3	2	1268	1054	1278	1047
681.5	2	2	829	676	870	703

Mass of Light	Charge State	# of Labels	Scan Numbers Where PITS Found the Peptide Pair			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
704.4167	1	1	538	481	559	469
717.5	1	1	486	448	492	426
717.4167	1	1	NaN	225	NaN	198
724.4167	2	2	703	579	749	619
724.4167	2	2	943	NaN	NaN	NaN
726.3333	1	1	369	322	376	296
726.3333	1	1	NaN	741	NaN	NaN
728.4792	2	2	1433	1242	1493	1236
728.4167	2	2	582	498	NaN	513
728.4722	2	2	NaN	1171	1441	1159
728.5	2	2	NaN	1057	NaN	NaN
728.9584	2	2	1309	1141	1379	1119
728.9722	2	2	1383	1199	NaN	1192
728.9584	2	2	NaN	1307	NaN	1304
728.9167	2	2	NaN	NaN	NaN	572
736.4375	2	2	1075	949	1164	944
736.4167	2	2	NaN	NaN	NaN	1156
736.4167	2	2	NaN	NaN	NaN	1240
738.4376	2	2	1184	1019	1251	1013
738.4167	2	2	1429	1240	1481	1234
750.4167	2	2	468	408	467	372
750.4167	2	2	NaN	456	NaN	439
754.4584	2	1	1221	1061	1288	1055
754.5209	2	1	1316	1148	1367	1127
754.9584	2	1	1154	1015	1236	1013
754.9792	2	1	1268	1099	1335	1085
774.4792	2	2	1423	1223	1483	1226
774.5	2	2	NaN	1163	NaN	1148
848.0834	3	3	1327	1152	1363	1142
848.0834	3	3	NaN	1052	NaN	NaN
902.9167	2	2	717	582	764	623
902.9167	2	2	NaN	628	813	NaN