

Electronic Thesis and Dissertation Repository

3-29-2023 12:30 PM

De novo sequencing of multiple tandem mass spectra of peptide containing SILAC labeling

Fang Han, *The University of Western Ontario*

Supervisor: Zhang, Kaizhong, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Computer Science

© Fang Han 2023

Follow this and additional works at: <https://ir.lib.uwo.ca/etd>



Part of the [Bioinformatics Commons](#)

Recommended Citation

Han, Fang, "De novo sequencing of multiple tandem mass spectra of peptide containing SILAC labeling" (2023). *Electronic Thesis and Dissertation Repository*. 9283.

<https://ir.lib.uwo.ca/etd/9283>

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.

Abstract

The systematic studies of proteins has gradually become fundamental in the research related to molecular biology. Shotgun proteomics use bottom-up proteomics techniques in identifying proteins contained in complex mixtures using a combination of high performance liquid chromatography coupled with mass spectrometry technology. Current mass spectrometers equipped with high sensitivity and accuracy can produce thousands of tandem mass spectrometry (MS/MS) spectra in a single run. The large amount of data collected in a single LC-MS/MS run requires effective computational approaches to automate the process of spectra interpretation. De novo peptide sequencing from tandem mass spectrometry (MS/MS) has emerged as an important technology for peptide sequencing in proteomics. However, the low identification rate of the acquired mass spectral limits the efficiency of computational approaches. To increase the accuracy and practicality of de novo sequencing, some previous algorithms used multiple spectra to identify the peptide sequence. In this thesis, we focus on de novo sequencing of multiple tandem mass spectra of peptide containing SILAC labeling. Compared with previous approach, our research develop de novo sequencing algorithms based on different idea of how to use multiple spectra. SILAC technology uses medium containing different kinds of isotope-labeled essential amino acids, usually Arginine(R) and Lysine(K), to label newly synthesized proteins with stable isotopes during cell growth. Multiple MS/MS spectra for the same peptide sequence are produced by spectrometer after the SILAC samples are processed by LC-MS/MS shotgun proteomics. Based on the factors such as the type of isotope

labeling, retention time, precursor ion mass, multiple spectra with different type of SILAC modifications for the same peptide in the sample can be used to identify the peptide sequence.

In this study, not only are we aiming to identify the peptide sequence with specific SILAC modifications, but we are also pinpointing locations of SILAC modifications from multiple SILAC labeled MS/MS spectra. We propose two de novo sequencing algorithms to compute the peptide sequence. The first algorithm is based on total number of SILAC modifications which is suitable for the situation that the total number of modifications for the peptide can be determined. The other algorithm is based on the combinations of SILAC modifications of Arginine(R) and Lysine(K) for general cases.

With two dynamic programming algorithms to identify peptide sequence and locating its SILAC modifications, the potential candidates are computed with similarity scores and then refinement algorithms are applied. Finally, a confidence score is designed to measure all of the candidate sequence.

To verify the performance of our algorithm, we compare the experimental results. We also compare the output candidates between our approach and PEAKS de novo.

Keywords: Mass spectrometry, Computational proteomics, De novo sequencing, SILAC, Multiple MS/MS spectra

Summary for lay audience

The systematic studies of proteins have exponentially increased in importance in the field of molecular biology. Understanding the sequence of amino acids of each protein can help us infer on its structure and therefore its role in normal cell development and in diseased tissues. Currently, the identification of proteins within complex mixtures can be done using a combination of techniques, including chromatography and mass spectrometry. The latter can break down and label numerous short amino acid sequences that require the appropriate computational approach to interpret such large body of data, which is also the rate limiting step in improving the efficiency of this method. Here, we developed two de novo sequencing algorithms of multiple Stable isotope labeling by amino acids in cell culture (SILAC) labeled tandem mass spectra by incorporating isotope-labeled amino acids into newly synthesized proteins. At once, we can identify the protein sequence and locate SILAC modifications, both of which are validated using currently available algorithms.

Acknowledgments

I would like to express sincere thanks to my supervisor Dr. Kaizhong Zhang who offered me the opportunity to work on this project. His great enthusiasm and unusual patience trained me both in critical research attitude and strict scientific method; without his meticulous supervision and sound guidance, I could not make this research through to its completion. Also importantly, I'm grateful to his financial support; otherwise, it might have been impossible to pursue my studies.

I sincerely appreciate my fiancée, Qin Dong, for her encouragements and understanding. I owe many thanks to my colleagues for their numerous encouragements and help whenever various difficulties seemed overwhelming. I would also like to express my sincere thanks to the people and staff around for their friendship to make my stay pleasant in the department.

Last but not least, I am deeply indebted to my parents for their love and understanding; you always deserve more than I could give you.

Contents

Abstract	i
Summary for lay audience	iii
Acknowledgments	iv
List of Figures	ix
List of Tables	xvii
List of Abbreviations	xix
1 Introduction	1
1.1 Background	1
1.1.1 Mass spectrometry based proteomics	1
1.1.2 Post-translational modification	3
1.1.3 SILAC	4
1.2 Research rationale and experimental objectives	5
1.3 Chapter outlines	6
2 Fundamentals	8

2.1	Biochemistry basics	8
2.1.1	Amino acid, peptide and protein	8
2.1.2	Post-translational modifications	13
2.2	Mass spectrometry	14
2.2.1	Instruments and configuration	14
2.2.1.1	Ionization source	14
2.2.1.2	Mass analyzer	17
2.2.1.3	Ion detector	20
2.2.1.4	Mass spectrum	21
2.2.2	Tandem mass spectrometry	22
2.2.2.1	Shotgun proteomics	23
2.2.2.2	Fragmentation and ion type	24
2.3	Interpreting MS/MS spectra	27
2.3.1	Database search	28
2.3.2	Peptide de novo sequencing	30
2.3.3	Multiple MS/MS spectra identification	34
2.4	Stable isotope labeling and SILAC	38
2.4.1	Stable isotope labeling	38
2.4.2	SILAC	39
3	De novo sequencing of multiple SILAC labeled MS/MS spectra	42
3.1	Introduction	42
3.2	Problem Definition	44

3.2.1	Notations	44
3.2.2	Ion fragments and ion mass	44
3.2.3	Single MS/MS spectra sequence matching and de novo sequencing	46
3.2.4	De novo sequencing to multiple MS/MS spectra of peptide con- taining SILAC labeling	47
3.3	Algorithms for de novo sequencing of multiple MS/MS spectra of pep- tide containing SILAC labeling	49
3.3.1	Algorithm based on total number of modifications	50
3.3.2	Algorithm based on modification pairs	54
3.4	Program design	59
3.4.1	Dynamic programming	61
3.4.2	Significant value	66
3.4.3	Traceback	67
3.4.4	Candidate refinement and confidence score	69
3.4.4.1	Subsequence array	71
3.4.4.2	Shuffling the candidate sequence	72
3.4.4.3	Confidence score	73
3.5	Complexity analysis	74
4	Experimental result and analysis	75
4.1	Group spectra results compare with spectrum results	75
4.1.1	Test design	75
4.1.2	Experiment 1	79

4.1.3	Experiment 2	86
4.2	De novo sequencing results comparison with PEAKS de novo sequencing result	92
4.2.1	Experiment data, software and algorithm settings	92
4.2.2	Experiments	94
4.2.2.1	Experiment 1	96
4.2.2.2	Experiment 2	98
4.2.2.3	Experiment 3	103
4.2.2.4	Experiment result summary	106
5	Conclusion and future work	110
5.1	Conclusion	110
5.2	Future work	112
	Bibliography	115
	Appendices	128
	Curriculum Vitae	135

List of Figures

2.1	Structure of alpha-amino acid	9
2.2	Structure of alpha-amino acid residue	9
2.3	Chemical Structure of the 20 standard amino acids [1]	10
2.4	Peptide bond formation: Condensation reaction [2]	11
2.5	4-level structure of protein [3]	12
2.6	Basic MALDI principles of time-of-flight (TOF) mass spectrometry analysis	15
2.7	Matrix-assisted laser desorption/ionization (MALDI) [4]	16
2.8	Electrospray ionization (ESI) [5]	16
2.9	Drastic difference exists in the shape of the same signal after measuring under different Resolving Power [6].	18
2.10	Ion detectors:(a) Faraday cup (b) Electron multiplier with discrete dyn- odes (c) Electron Multiplier with one continuous dynode [7].	21
2.11	A mass spectra and the isotopic distribution [8]	22
2.12	General diagram of tandem mass spectrometry	23
2.13	Workflow of LC-MS/MS based shotgun proteomics [9].	24
2.14	Cleavage of Six basic types of fragment ions [10]	25

2.15	Six basic types of fragment ions [10]	26
2.16	Structure of amino-acylium ion, amino-immonium ion and immonium ion [11]	27
2.17	Example of a spectrum of a successfully identified peptide LTKVHKE [12]	28
2.18	Workflow of database search approach [12].	29
2.19	Workflow of de novo sequencing [12].	31
2.20	Schematic of graph spectra model [12]	32
2.21	De novo sequencing problem definitions from Bin Ma and Kaizhong Zhang approaches. [13]	34
2.22	Relationships and ions selected in spectra merging [14]	37
2.23	Ion types considered in CID/HCD spectra [14]	37
2.24	Ion types considered in ECD/ETD spectra [14]	38
2.25	Flowchart of triple SILAC coupled with LC-MS/MS for the comparative analysis of three distinct cell populations [15]	40
2.26	Example of SILAC MS spectra [16]	41
3.1	Logic structure of dynamic programming rows when $n = 3$	62
3.2	Dynamic programming matrix DP	64
3.3	Subsequence array	71
3.4	Subsequences of mass m	73
4.1	Example of group spectra identification results and its single spectrum identification results	78
4.2	FDR curve for Experiment 1 (PEAKS DB)	80

4.3	Groundtruth dataset of experiment 1 ordered by peptide sequence alphabetical order. (From PEAKS DB)	80
4.4	FDR curve for Experiment 2 (PEAKS DB)	87
4.5	Groundtruth dataset of experiment 2 ordered by peptide sequence alphabetical order. (From PEAKS DB)	87
4.6	Experiment 1: Identification result of spectrum 1 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.	97
4.7	Experiment 1: Spectrum 1 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.	98
4.8	Experiment 1: Ion table of spectrum 1 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.	99
4.9	Experiment 1: Identification result of spectrum 2 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.	99
4.10	Experiment 1: Spectrum 2 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.	100

4.11	Experiment 1: Ion table of spectrum 2 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.	100
4.12	Experiment 1: identification result of spectrum 1 only	101
4.13	Experiment 1: identification result of spectrum 2 only	101
4.14	Experiment 1: identification result of spectra 1 and 2	102
4.15	Experiment 1: ion table of spectra 1 and 2 matching the candidate peptide sequence	102
4.16	Experiment 2: Identification result of spectrum 1 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.	103
4.17	Experiment 2: Identification result of spectrum 2 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.	103
4.18	Experiment 2: Identification result of spectrum 3 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.	104

4.19	Experiment 2: Identification result of spectrum 4 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.	104
4.20	Experiment 2: Identification result of all 4 spectra	105
4.21	Experiment 2: Identification result of spectra 1 and 4	105
4.22	Experiment 3: Ion table of spectrum 1 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.	106
4.23	Experiment 3: Ion table of spectrum 2 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.	107
4.24	Experiment 3: Ion table of spectrum 3 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.	107
4.25	Experiment 3: Identification result of all 3 spectra	108
4.26	Experiment 3: ion table of spectra matching the candidate peptide sequence	108

.1	Experiment 2: Spectrum 1 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.	128
.2	Experiment 2: Ion table of spectrum 1 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.	129
.3	Experiment 2: Spectrum 2 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.	129
.4	Experiment 2: Ion table of spectrum 2 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.	130
.5	Experiment 2: Spectrum 3 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.	130
.6	Experiment 2: Ion table of spectrum 3 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.	131

.7	Experiment 2: Spectrum 4 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.	131
.8	Experiment 2: Ion table of spectrum 4 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.	132
.9	Experiment 2: ion table of spectra combined matching the candidate peptide sequence	132
.10	Experiment 3: Spectrum 1 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.	133
.11	Experiment 3: Identification result of spectrum 1 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.	133
.12	Experiment 3: Spectrum 2 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.	133
.13	Experiment 3: Identification result of spectrum 2 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.	134

.14 Experiment 3: Spectrum 3 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red. 134

.15 Experiment 3: Identification result of spectrum 3 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence. 134

List of Tables

2.1	Residue mass and composition of the 20 standard amino acids	11
2.2	Post-Translational Modifications [17]	13
2.3	Comparison of the typical performance characteristics of several types of commonly used mass analyzers [12]	20
3.1	pre_K and pre_R values when $k + l \leq 3$	62
3.2	SILAC combinations corresponding to the index of dynamic program- ming when $k + l = 3$	63
4.1	Total number of group and spectra of Experiment 1	81
4.2	Total number of groups identified real sequence of Experiment 1	81
4.3	Total number of spectra identified real sequence of Experiment 1	81
4.4	Spectra identified real sequence but group can not identify statistics of Experiment 1	82
4.5	Group spectra identified real sequence statistics of Experiment 1	82
4.6	Score comparison statistics of Experiment 1	83
4.7	Recall of Experiment 1	84
4.8	Precision of Experiment 1	84

4.9	Brief information summary for comparison of top one highest score candidate chosen and top five candidate chosen of Experiment 1 . . .	85
4.10	Total number of groups and spectra of Experiment 2	86
4.11	Total number of group identified real sequence of Experiment 2 . . .	88
4.12	Total number of spectra identified real sequence of Experiment 2 . . .	88
4.13	Spectra identified real sequence but group can not identify statistics of Experiment 2	89
4.14	Group spectra identified real sequence statistics of Experiment 2 . . .	89
4.15	Score comparison statistics of Experiment 2	90
4.16	Recall of Experiment 2	90
4.17	Precision of Experiment 2	91
4.18	Brief information summary for comparison of top one highest score candidate chosen and top five candidate chosen of Experiment 2 . . .	92

Abbreviation

MS/MS Tandem mass spectrometry

LC Liquid chromatography

LC-MS/MS Liquid chromatography with tandem mass spectrometry

MALDI Matrix-assisted laser desorption ionization

ESI Electrospray ionization

CID Collision induced dissociation

HCD Higher-energy collisional dissociation

ETD Electron transfer dissociation

PTM Post translational modification

SILAC Stable isotope labeling with amino acids

Chapter 1

Introduction

1.1 Background

1.1.1 Mass spectrometry based proteomics

Proteins are crucial entities of all biological processes. The systematic studies of proteins has gradually become fundamental in the research related to molecular biology. The primary objective is to grasp a comprehensive understanding of the protein interactions that govern aberrant cellular processes, which in turn contribute to disease occurrence. Efficient and reliable analysis of proteomic data invaluablely contribute to current clinical detection and treatment of diseases and underlines basic health science research. Such analysis depend on the accurate identification and quantification of proteins expressed under known and manipulable conditions. A protein molecule has four level structure. Structural stratification of proteins begins with its primary amino acid sequence. The unique sequence of a given amino acid strand determines

its higher level structure, folding, and function. Accurate sequencing analysis can unveil its interactions with other proteins and/or macromolecules and infer on its overall function within the cell. The consensus strategy to decode a protein sequence is to fragment a protein into smaller polypeptide chains and determine its sequence from the twenty different amino acids found in eukaryotes. Hence, peptide sequencing has become a standard of practice and the rate limiting step in current proteomics. In the past two decades, tandem mass spectrometry (MS/MS) has emerged as a major technology for peptide sequencing because of its high throughput data which can be generated in a short amount of time and its exceptional sensitivity [18–20]. In a typical liquid chromatography with tandem mass spectrometry (LC-MS/MS) experiment, protein sample is digested into peptides with proteolytic enzymes (e.g. trypsin) to break protein into relatively short peptide sequences. Then, the peptide mixtures are separated by using liquid chromatography (LC) and subsequently ionized by using matrix-assisted laser desorption ionization (MALDI) [21] or electrospray ionization (ESI) [22]. After ionization, the charged peptides are measured in the first mass analyzer and then fragmented and followed by a measurement in the second mass analyzer. The output of MS/MS spectra is a diagram called a tandem mass spectra (MS/MS spectra) [23–26]. A MS/MS spectra usually contains two kinds of information called the mass-to-charge ratio (m/z) values of fragment ions and the corresponding intensities. Therefore, MS/MS acts as a charged sieve from which we can only obtain information on selected, short, charged fragments. Yet, the main hurdle in protein sequencing using MS/MS remains to be piecing together the information gathered from these short peptides to deduce the whole protein sequence.

Current mass spectrometers equipped with high sensitivity and accuracy can produce thousands of MS/MS spectra in a single run. The large amount of data collected in a MS/MS experiment requires effective computational approaches to automate the process of spectra interpretation [12]. Database search and de novo sequencing are two mainstream computational methods for this purpose. Both approaches design scoring functions to determine the best matched peptide for each MS/MS spectra. However, the main difference between these two approaches is that the database search approach matches the peptide sequence to an existing protein sequence database, whereas de novo sequencing computes the sequence from the spectra directly [27]. De novo sequencing is required for the identification of novel or modified peptides that do not exist in a protein database, such as peptides from un-sequenced organisms, biotech products, and peptide toxins [28]. There are several software packages for the purpose of de novo sequencing, including Lutefisk [29], PEAKS [30], PepNovo [31], pNovo [32], EigenMS [33], Sherenga [34], NovoHMM [35], MSNovo [36], AUDENS [37], SEQUIT [38], PPM-Chain [39].

1.1.2 Post-translational modification

Although the identification of peptide sequence is necessary to infer its native conformation, it is insufficient for determining its regulation such as post-translational modifications (PTM). Post-translational modifications (PTM) refers to the chemical modification of a protein after translation. There are a few hundreds of known

PTMs with the most common being phosphorylation, glycosylation, methylation, acetylation and acylation [40,41]. Many of these modifications affect the activity and specificity of proteins, and may even play a role in stabilizing a protein's structure and in regulating enzymatic activity [9]. In many cases, the proteins are modified at several sites by a number of added functionalities. Unarguably, PTMs are essential for cellular homeostasis by regulating protein activity in response to various stimuli, and thus aberrant PTMs are inevitable identified as contributors to various diseases. PTMs can be determined by MS/MS spectrometry since PTMs change the mass of peptide.

1.1.3 SILAC

Stable isotope labeling by amino acids in cell culture (SILAC) based analysis is a modern method for protein identification and quantitation. SILAC is a simple and powerful approach in mass spectrometry-based quantitative proteomics [42]. In a typical SILAC experiment, cells are grown in a defined medium complemented with certain essential amino acids, usually Arg and Lys, containing normal amino acids (the light culture) or amino acids labeled with stable isotope (the heavy culture). After full incorporation of the labeled amino acid, the light and heavy cultures are subjected to different perturbations, harvested, and combined essential amino acids, usually Arginine(R) and Lysine(K), containing naturally occurring atoms (the light culture) or a specific number of their stable isotope counterparts (the medium and

heavy culture). Peptides resulting from enzymatic digestion of the combined cell lysates are then detected by mass spectrometry [43].

1.2 Research rationale and experimental objectives

Although de novo sequencing can generate rather trustworthy preliminary protein sequences that are not found in existing databases, it has several limitations that remains to be addressed. First, the models constructed usually do not perfectly reflect MS/MS spectra. Second, less information is extracted than could be from a MS/MS spectra. Third, traditional peptide identification methods only use one MS/MS spectra to infer the peptide sequence [44]. To increase the accuracy and practicality of de novo sequencing, some previous approaches focus on the use of alternative MS/MS spectra and multiple spectra of the same peptide. In these approaches, data and information are combined from spectra generated by different fragmentation methods. For example, in the CID fragmentation, b-ions and y-ions are the primary results from cleavage of the peptide bond, while ETD fragmentation mainly generates c-ions and z-ions. Thus, CID and ETD make of the complementary techniques to lead a wider coverage of a peptide sequence with fragment ions [45].

Isotope labeling project (for example, SILAC based) contains same protein molecules with different types of isotope labeling in its sample. Multiple MS/MS spectra can be generated for the same peptide sequence after the samples are processed by LC-MS/MS shotgun proteomics. Based on the factors such as the type of isotope labeling, retention time, precursor ion mass, multiple spectra with different types of SILAC

modifications for the same peptide can be determined and used to identify the peptide sequence.

The work in this thesis focuses on developing novel algorithms for peptide identification from multiple MS/MS spectra of peptide containing SILAC labeling and scoring function for candidate measurement.

1.3 Chapter outlines

The thesis is organized in the following chapters:

In chapter 1, we present a brief introduction to the background, research rationale and objectives of the research.

In chapter 2, we provide the necessary fundamentals for mass spectrometry based proteomics research, which include the biochemical basic knowledge, and the mass spectrometry technology. We also give a general summary on the current computational methods for mass spectra identification and the basic knowledge about PTMs, stable isotoped labeling and SILAC.

In chapter 3, we formulate the multiple SILAC labeled MS/MS spectra de novo sequencing problem mathematically and propose a dynamic programming algorithm for the problem. A scoring function is also presented which is designed to measure the candidates.

In chapter 4, we performed experiments from the real SILAC project to test our algorithm, comparing the output with PEAKS de novo and analysis the results.

In chapter 5, we conclude the major content in the research and discuss the future

work.

Chapter 2

Fundamentals

2.1 Biochemistry basics

2.1.1 Amino acid, peptide and protein

The structure of an amino acid contains of the (α) carbon atom bonded with an amine group, a carboxylic acid group, a hydrogen and a side chain (R group) specific to each amino acid. Amino acids can be represented by formula $H_2NCHRCOOH$ (Figure 2.1). An amino acid residue is formed when an amino acid loses one hydrogen atom (H) from the amine group (N-terminal) and one hydroxyl group (OH) from the carboxylic group (C-terminal) (Figure 2.2). There are 20 common alpha-amino acids, each with their characteristic side chains (R groups). Figure 2.3 shows their chemical structures. Table 2.1 shows mass and composition information regarding all the 20 common amino acid residues. The chemical reaction process that two amino acids joined together called condensation [46]. Two amino acids are joined together through

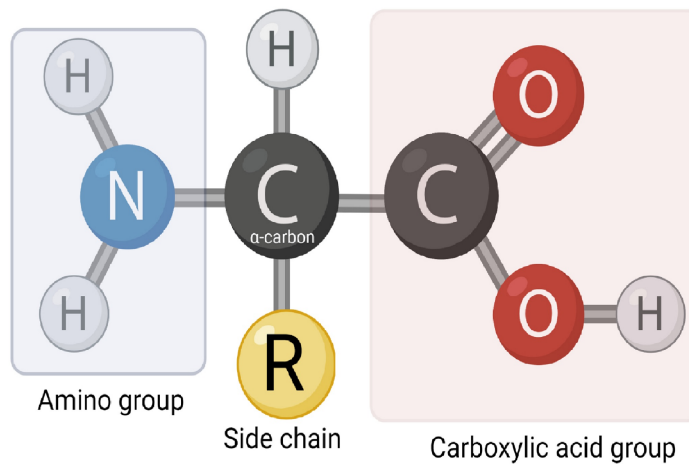


Figure 2.1: Structure of alpha-amino acid

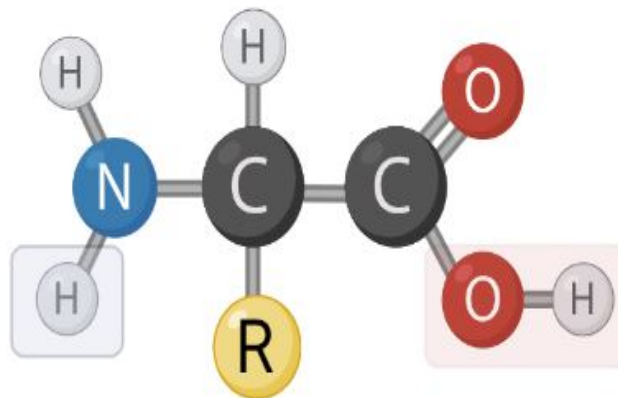


Figure 2.2: Structure of alpha-amino acid residue

the formation of a peptide bond between the carboxylic group of one amino acid and the amino group of the other amino acid. This condensation reaction is characterized by the loss of a hydrogen (H) and a hydroxyl (OH) from the other - or the net loss of a water molecule. Figure 2.4 shows the condensation reaction. Depending on the length of the amino acid chain, it assembles to become either a peptide (shorter) or a protein

Name	Formula	Abbreviations	Name	Formula	Abbreviations
Glycine		Gly G	Cysteine		Cys C
Alanine		Ala A	Methionine		Met M
Valine		Val V	Lysine		Lys K
Leucine		Leu L	Arginine		Arg R
Isoleucine		Ile I	Histidine		His H
Phenylalanine		Phe F	Tryptophan		Trp W
Proline		Pro P	Aspartic Acid		Asp D
Serine		Ser S	Glutamic Acid		Glu E
Threonine		Thr T	Asparagine		Asn N
Tyrosine		Tyr Y	Glutamine		Gln Q

Figure 2.3: Chemical Structure of the 20 standard amino acids [1]

Name	3-letter symbol	1-letter symbol	Monoisotopic mass	Average mass	Composition
Alanine	Ala	A	71.03711	71.08	C_3H_5NO
Arginine	Arg	R	156.10111	156.2	$C_6H_{12}N_4O$
Asparagine	Asn	N	114.04293	114.1	$C_4H_6N_2O_2$
Aspartic Acid	Asp	D	114.02694	115.1	$C_4H_5NO_3$
Cysteine	Cys	C	103.00919	103.1	C_3H_5NOS
Glutamic Acid	Glu	E	129.04259	129.1	$C_5H_7NO_3$
Glutamine	Gln	Q	128.05858	128.1	$C_5H_8N_2O_2$
Glycine	Gly	G	57.02146	57.05	C_2H_3NO
Histidine	His	H	137.05891	137.1	$C_6H_7N_3O$
Isoleucine	Ile	I	113.08406	113.2	$C_6H_{11}NO$
Leucine	Leu	L	113.08406	113.2	$C_6H_{11}NO$
Lysine	Lys	K	128.09496	128.2	$C_6H_{12}N_2O$
Methionine	Met	M	131.04049	131.2	C_5H_9NOS
Phenylalanine	Phe	F	147.06841	147.2	C_9H_9NO
Proline	Pro	P	97.05276	97.12	C_5H_7NO
Serline	Ser	S	87.03203	87.08	$C_3H_5NO_2$
Threonine	Thr	T	101.04768	101.1	$C_4H_7NO_2$
Tryptophan	Trp	W	186.07931	186.2	$C_{11}H_{10}N_2O$
Tyrosine	Tyr	Y	163.06333	163.2	$C_9H_9NO_2$
Valine	Val	V	99.06841	99.13	C_5H_9NO

Table 2.1: Residue mass and composition of the 20 standard amino acids

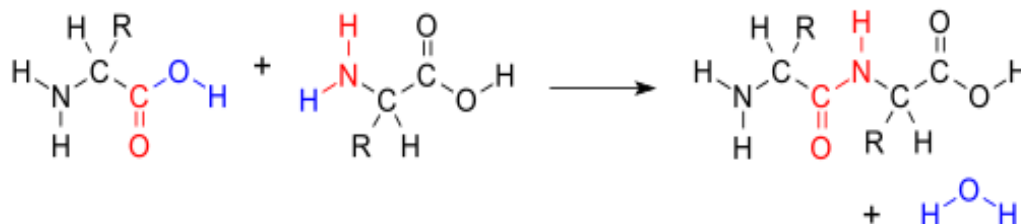


Figure 2.4: Peptide bond formation: Condensation reaction [2]

(longer). Proteins have more defined structure and function, and is stratified into four level structures. The primary structure is the amino acid chain. Secondary Structure refers to the folding of the chain that gives the protein its 3D shape. Different parts of polypeptide bonds by weak hydrogen bonds. There are two major peptide chain backbone conformations: the alpha helix and the beta sheet. Around 60% of the

length of the average polypeptide chain consists of segments of alpha helices or beta sheets. Both structures are based on the hydrogen bonding between peptide bond carbonyl O atoms on one amino acid residue, and the amino group hydrogen on a difference amino acid residue. Alpha helices are defined by their periodicity driven by the tilted axis of hydrogen bonds of 3.6 residues per turn. Beta sheets can create large surfaces by hydrogen linking two adjacent beta strands that may be parallel or anti parallel. The overall three dimensional shape of a protein molecule is the tertiary structure. The molecule bends and twists to achieve maximum stability or lowest energy state. The quaternary structure represents proteins interact with each other and refers to as protein subunits. Figure 2.5 shows 4-level structure of protein.

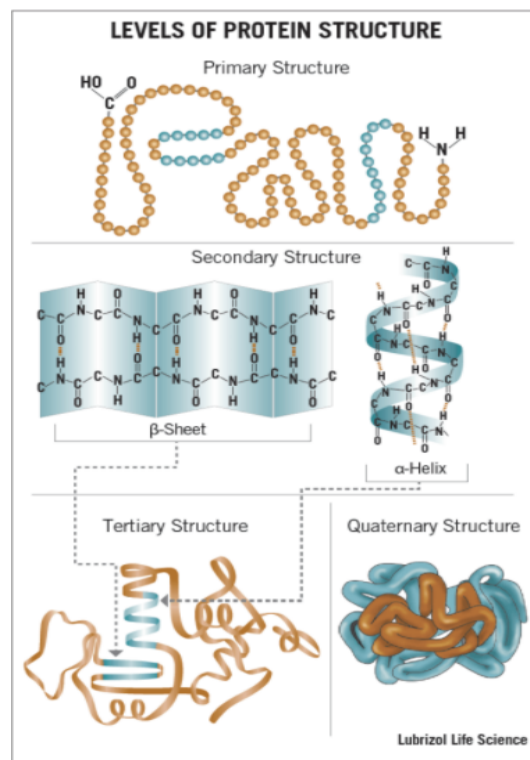


Figure 2.5: 4-level structure of protein [3]

2.1.2 Post-translational modifications

Post-translational modifications (PTMs) are chemical modifications which are key mechanisms to increase proteomic diversity. There are hundreds of post-translational modifications including phosphorylation, glycosylation, ubiquitination, nitrosylation, methylation, acetylation and lipidation which maintain normal cell biology [47], altogether regulating the physical and chemical properties, stability, and activity of a protein by altering its structure [48–50] (Table 2.2).

Entry No.	Mass(Δ m)	Residue	Modification Name
1	-48.003372	M@C-term	Homoserine lactone
2	-29.992805	M@C-term	Homoserine
3	-18.010565	C@N-term	Dehydration
4	-18.010565	E@N-term	Pyroglutamic Acid from E
5	-17.026548	Q@N-term	Pyroglutamic Acid from Q
6	-0.984016	[X]@C-term	Amidation
7	0.984016	R,N,Q	Deamidation
8	14.01565	E,[X]@C-term	Methylation
9	15.994915	W,H,C,M	Oxidation or Hydroxylation
10	21.981943	D,[X]@C-term	Sodium adduct
11	27.994915	T,[X]@N-term	Formylation
12	31.989828	M	Dihydroxy(Di-oxidation)
13	42.010567	C,[X]@N-term	Acetylation
14	43.005814	K,[X]@N-term	Carbamylation
15	44.026215	C	Ethanolation
16	45.98772	C	Beta-methylthiolation
17	57.021465	C	Carbamidomethyl
18	58.005465	C	Iodoacetic acid derivative
19	71.03712	C	Acrylamide adduct
20	79.95682	Y,T,S	O-Sulfonation
21	79.96633	D,Y,H,T,S	Phosphorylation
22	99.06841	C	N-sopropylcarboxamidomethyl
23	105.057846	C	S-pyridylethylation
24	162.0528	Y,[X]@N-term	Hexoses
25	203.0794	S,T,N	N-Acetylhexosamine
26	210.1937	K,[X]@N-term	Myristoylation
27	226.07759	K,[X]@N-term	Biotinylation

Table 2.2: Post-Translational Modifications [17]

2.2 Mass spectrometry

2.2.1 Instruments and configuration

The three significant functions of a mass spectrometer and the associated components are

1. The Ion Source: Produce ionized sample which is usually to cations by loss of an electron.
2. The Mass Analyzer: The ions are sorted and separated according to their mass over charge value.
3. The Detector: The separated ions are detected and formed spectra which is comprised of a series of peaks.

Figure 2.6 shows the a time-of-flight (TOF) mass spectrometer. The basis of TOF mass analysis is that charged ions from the ionization source are accelerated into an electric field free flight region. The larger ion will gain the less energy during acceleration and as a result it will travel slower than smaller ions.

2.2.1.1 Ionization source

Ionization is the process that converts a neutral molecule into a charged molecule by gaining or losing electrons. The electrically charged molecule is called an ion. Molecules are ionized in the ionizer by technique like impacting compounds with electron beam, which renders the formation of charged molecules. MALDI (Matrix-Assisted Laser Desorption/Ionization) and ESI (Electrospray Ionization) are two

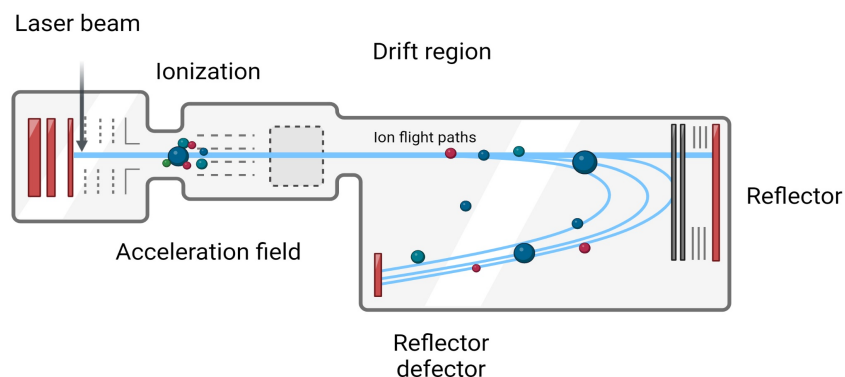


Figure 2.6: Basic MALDI principles of time-of-flight (TOF) mass spectrometry analysis

types of ionizers which are commonly used in mass spectrometry based proteomics study.

Matrix-assisted laser desorption/ionization (MALDI) uses a laser energy absorbing matrix to create ions from large molecules. There are three steps in the process. First, the sample is mixed with a suitable matrix material and applied to a metal plate. Second, a pulsed laser irradiates the matrix-sample spot, triggering desorption and vaporization of the sample. The matrix absorbs the UV laser energy, preventing the sample from being destroyed. Finally, the molecules are ionized by being protonated or deprotonated in the hot plume of ablated gases before they are accelerated into mass spectrometers [51]. Figure 2.7 shows how the ionization happens in MALDI.

Electrospray ionization (ESI) produces ions using an electrospray in which a high voltage is applied to a liquid to create an aerosol. The liquid containing the analytes is dispersed into a fine aerosol before going through a capillary tube carrying high voltage. Then the droplet is heated to aid the solvent evaporation which will sub-

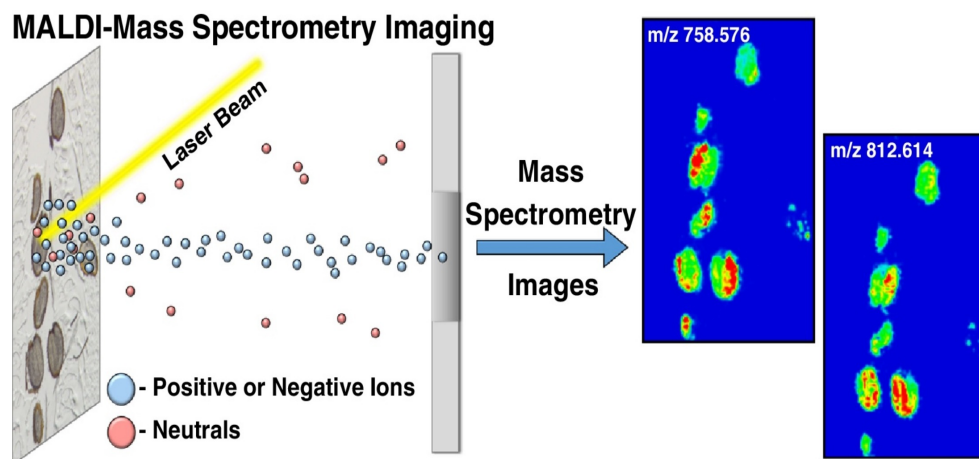


Figure 2.7: Matrix-assisted laser desorption/ionization (MALDI) [4]

sequently makes itself smaller and denser charged [12]. When reaching the Rayleigh limit, the surface tension that holds the droplet together will be surpassed by the electrostatic repulsion of like charges, which will cause the droplet to break into smaller yet stable droplets. The new droplets undergo similar desolvation and fission process and eventually turn into a stream of charged ions [12, 52]. Figure 2.8 shows the principle of Electrospray Ionization (ESI).

In mass spectroscopy, the mass-to-charge ratio (m/z) is the most widely used physi-

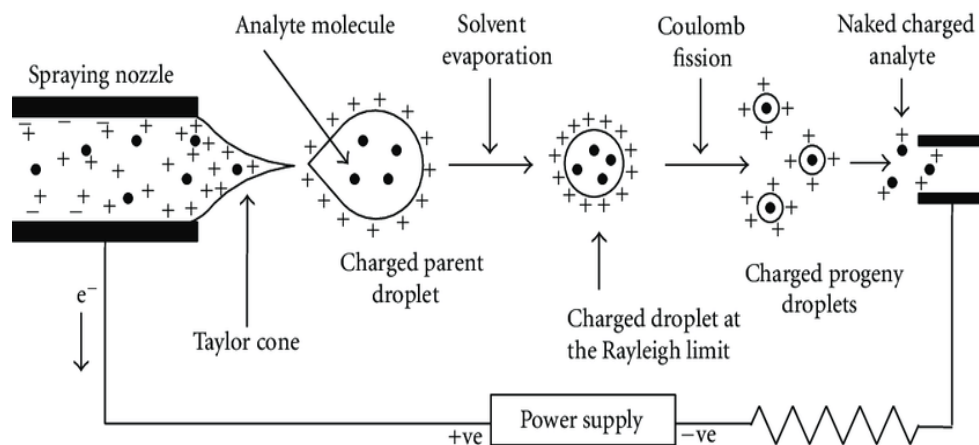


Figure 2.8: Electrospray ionization (ESI) [5]

cal quantity in the electrodynamics of charged particles. It is equal to the mass of the cation divided by its charge. MALDI produces singly charged ions ($z = 1$) and ESI produces singly and multiply charged molecules. ESI has advantages that molecule with large mass can also be measured and profiled in the mass spectrometers when the charge state z is larger than 1 since a relatively large molecule can still fall into the m/z range. Another advantage of ESI is that it produces multi-charge phenomena of the same ions so that instrument can defeat in the low m/z range. However, the existence of multiple charged ions increases the complexity of the spectra. A single type of molecule may produce multiple peaks due to different charge states and the charge state of a peak needs to be determined by other functions in order to convert the m/z value back to the mass value.

The molecules will be transmitted into the electromagnetic field of a mass spectrometer after being ionized. Molecules with different m/z values will demonstrate different motions when passing through the electromagnetic field [12].

2.2.1.2 Mass analyzer

Ions are separated in the mass analyzer according to their mass-to-charge ratios(m/z). Mass range, mass resolution and mass accuracy are three primary parameters of a mass analyzer.

The mass range is the limit of mass-to-charge ratio which measures ions passing through. Ions can be profiled only if their mass-to-charge ratios fall into this range. Dalton(Da) is a unit which equals $1/12$ of the mass of a carbon atom and is approximately the mass of one hydrogen atom. The mass-to-charge ratio range for proteomics

analysis is typically from around 100 Da to a few thousand Da.

Mass resolution evaluates the ability of the mass spectrometer to distinguish two peaks that barely have difference of the mass-to-charge ratio values. Larger resolution indicates a better separation of peak profiles. The capacity of mass resolution is called resolving power defined as $R = M/\Delta M$ where M is the mass of the peak. ΔM can be defined in different ways. In the valley definition, ΔM is the closest spacing between two peaks of equal intensity with the valley between them less than a specified fraction of the peak height. Typical fractions are 5%, 10% or 50%.

For the peak width definition, the value of ΔM is the width of the peak measured at a 50% of the peak height which is also called the Full Width at Half Maximum (FWHM). Figure 2.9 shows achievable measurement under different resolving power where M equals 1999 and ΔM equals to 10, 1 and 0.1. High resolution is considered to be with Resolving Power $R \geq 5000$.

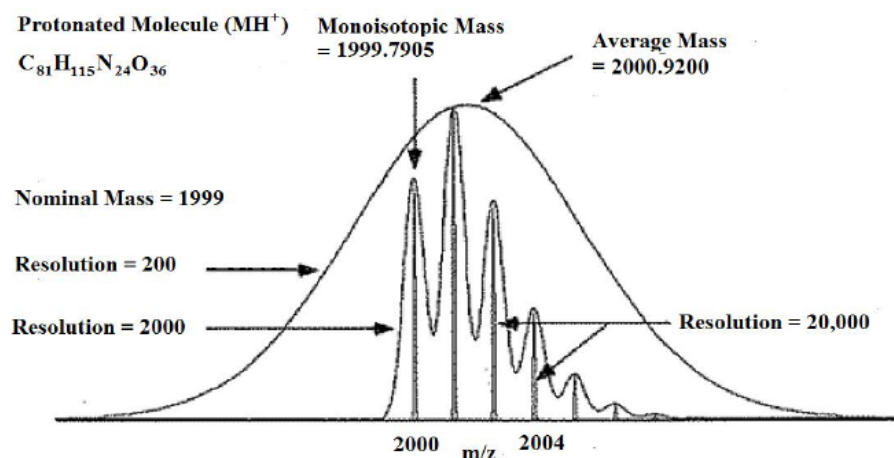


Figure 2.9: Drastic difference exists in the shape of the same signal after measuring under different Resolving Power [6].

Mass accuracy represents how much accuracy of the mass value a mass analyzer

can provide. There are two ways to measure mass accuracy that are millimass unit (mmu) or parts per million (ppm). A mmu is equivalent to 1/1000 of the unified atomic mass unit (u) which is displaced by the unit dalton so that 1 mmu equals to 1 millidalton (mDa). Mass accuracy expressed in ppm is calculated by the formula $MA_{ppm} = (|m_1 - m_2|)/m_2 \times 10^6$ where m_1 is the real mass and m_2 is the mass generated by the mass spectrometer. Millimass unit (mmu) and parts per million (ppm) can convert with each other. Given a ion with mass M , then we can have $MA_{mDa} = (MA_{ppm} \times M)/10^3$

Note that mmu describe the absolute mass difference between theoretical mass and the mass generated by the mass spectrometer. ppm is a relative value that equals to mass difference over mass generated by the mass spectrometer. Therefore, for same mass difference value, its ppm depends on the mass of spectrum. For example, given theoretical mass 100 and observed mass 101, ppm is 10000 which is calculated from formula, but given theoretical mass 10000 and observed mass 10001, ppm is 100. Normally, we prefer using ppm to measure the accuracy of mass analyzer.

There are several types of mass analyzers that have been developed, including the Quadrupole mass analyzers [53], Quadrupole ion trap, QIT [54]; Linear ion trap, LIT/LTQ [55], Time-of-flight (TOF) analyzers [56], Fourier transform ion cyclotron resonance (FTICR) [57] and Orbitrap [58]. Table 2.3 is a summary of the performance characteristics for each mass analyzer.

Mass Analyzer	Resolving Power	Accuracy(ppm)	m/z Range	Scan Rate
Quadrupole	1000	100-1000	50-2000; 200-4000	Moderate
QIT	1000	100-1000	10-4000	Moderate
LTQ	2000	100-500	50-2000; 200-4000	Fast
TOF	10000-20000	10-100	No upper limit	Fast
FT-ICR	100000-750000	<2	50-2000; 200-4000	Slow
Orbitrap	30000-100000	2-5	50-2000; 200-4000	Fast

Table 2.3: Comparison of the typical performance characteristics of several types of commonly used mass analyzers [12]

2.2.1.3 Ion detector

Ions get to ion detector after separated by mass analyzer. A signal is produced by an ion impinging event or the charge induced when an ion passing by and then the signal is recorded by the ion detector. Ion's charge will be neutralized by an electron emitted from the surface onto the ion when it hits the metal surface of the detector. The kinetic movement of electrons forms an electric current which will be recorded by ion detector. Amplification technique is applied to get a signal since the number of ions leaving the mass analyzer at a specific instant is normally small.

Faraday Cup is the simplest ion detector. It is a metal cup that collects all ions leaving the mass filter. The secondary electrons emitted upon an ion impact event can also be captured by Faraday Cup's metal surface. The current flowing away from the Faraday Cup will be recorded as a signal but it is relatively low in sensitivity and slow in response time. Electron Multiplier is the more commonly used detector. An electron multiplier is comprised of a series of dynodes which transfers the kinetic energy of incident ions that in turn generates secondary electrons. An emission of electrons is caused by an ion striking the first dynode surface and then attracted to the next dynode with a higher potential and more secondary electrons are emitted. Finally,

a cascade of electrons is converted into a voltage signal which is proportional to the number of impinging ions. There is another type of electron multiplier that contains one continuous dynode instead of several discrete dynodes. Compared with Faraday Cup, electron multiplier has higher amplification factor and the faster response time.

Figure 2.10 shows these three ion detectors.

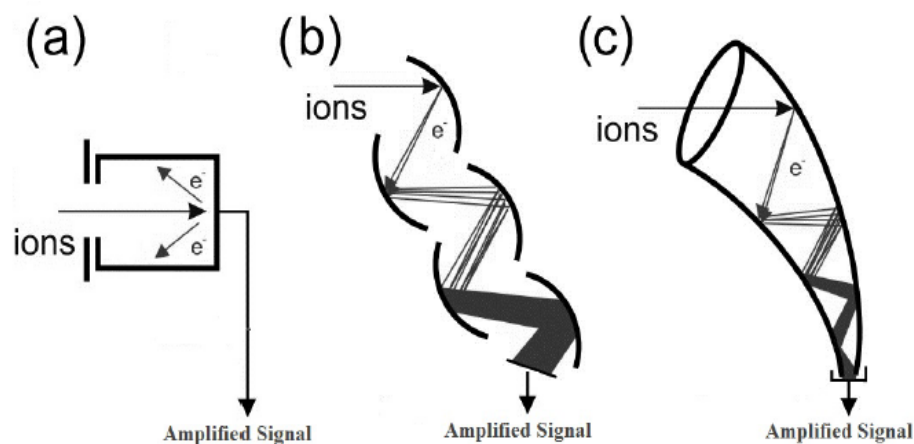


Figure 2.10: Ion detectors:(a) Faraday cup (b) Electron multiplier with discrete dynodes (c) Electron Multiplier with one continuous dynode [7].

2.2.1.4 Mass spectrum

The data generated by ion detector are analyzed by mass spectrometers which are connected to computers with softwares. The detected ions are organized by their individual m/z values and relative abundance and then recorded. In spectra, ions with the same m/z form a peak. The intensity of a peak represents the number of ions observed by the detector which is related to the abundance of the corresponding ions in the original sample.

Figure 2.11 shows an example of a mass spectra. The inner image shows the isotopic

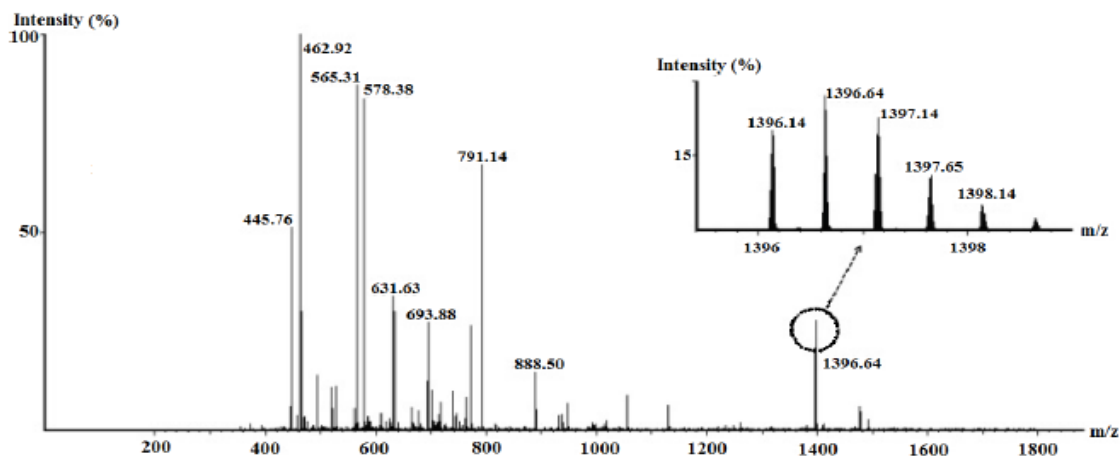


Figure 2.11: A mass spectra and the isotopic distribution [8]

distribution when zooming in a peak with $m/z = 1396.64$. Based on the monoisotopic peak distribution, the charge state of the ion can be determined [59,60]. The adjusted monoisotopic m/z value can be used to accurately interpret the mass spectra [61,62]. In particular, each peak spans a width on the m/z direction. Before any data analysis is done, peaks will undergo a process of centroiding that each peak is assigned with a single m/z value. This value represents the centroid of the peak shape.

2.2.2 Tandem mass spectrometry

In proteomics, the tandem mass spectrometry analysis can provide much more information about a peptide than the mass spectrometry analysis. Thus, most of today's proteomics analyses has to heavily rely on the tandem mass spectrometry (MS/MS) techniques. A tandem mass spectrometer has two mass analyzers or two sequential analyses in the same analyzer. The first analyzer selects ions at a certain m/z window which is usually a very small window of a few Daltons wide so that only copies of

the same ion are selected. This is called the precursor ion or the parent ion. Then the ion is fragmented into fragment ions by some dissociation methods. Finally, the fragments are then separated based on their individual m/z ratios in another mass analyzer and form a MS/MS spectrum.

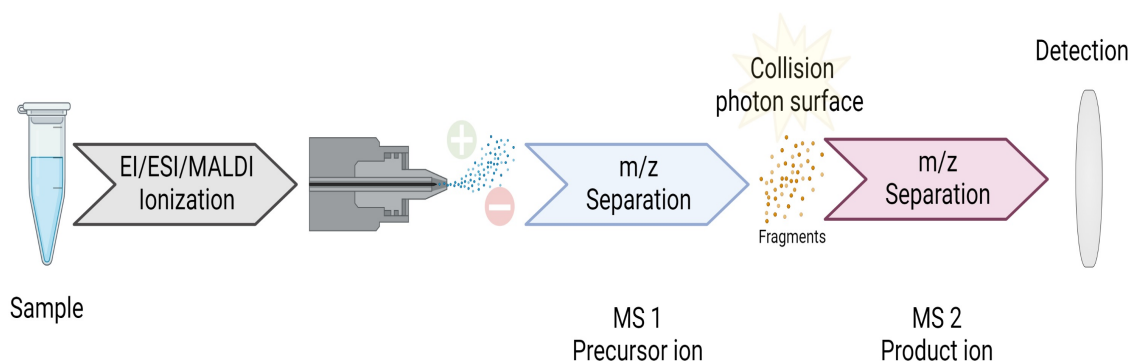


Figure 2.12: General diagram of tandem mass spectrometry

2.2.2.1 Shotgun proteomics

Shotgun proteomics refers to the use of bottom-up proteomic techniques in identifying proteins contained in complex mixtures by combination of high performance liquid chromatography (LC) coupled with mass spectrometry technology [63,64](Figure 2.13). Protein mixture is first undergone the enzymatic digestion broken down into peptides. Trypsin is the most commonly used enzyme. Peptides are then separated by liquid chromatography (LC). Different peptides elute at different times, it's noted that the separation may not be perfect. At the third step, the mass spectrometer scans the peptide ions at a particular time to obtain a MS scan profile. Each peak from the MS spectra is supposed to represent a unique peptide. At the last step,

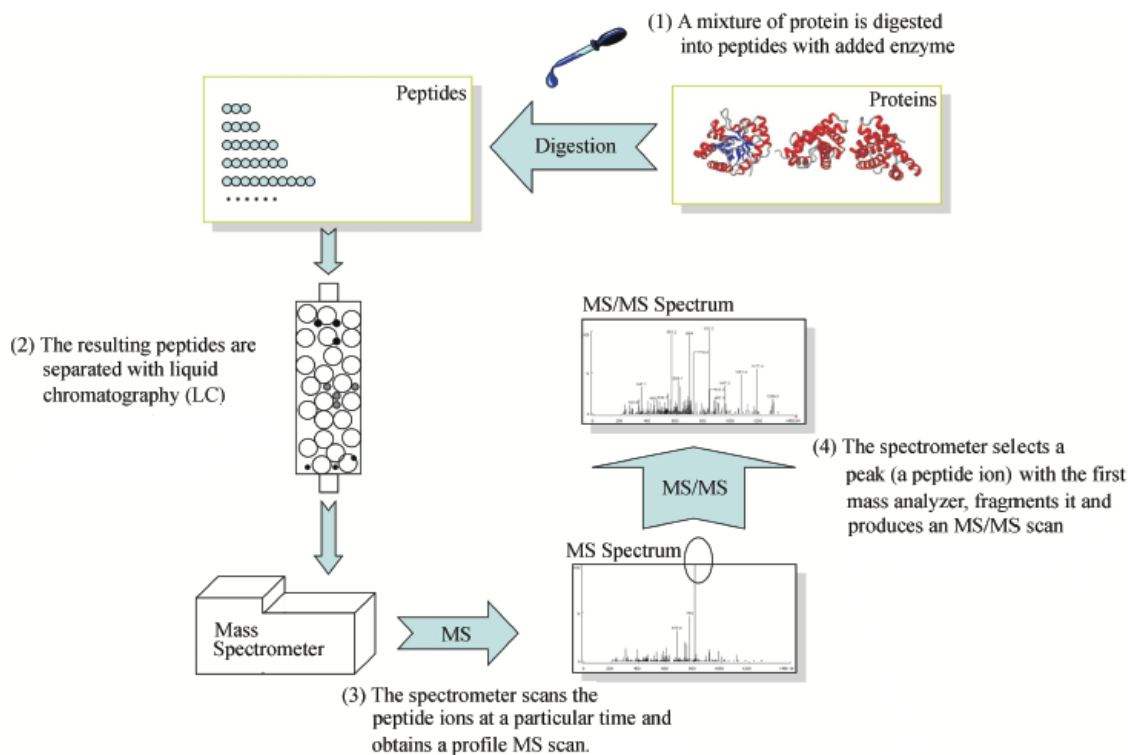


Figure 2.13: Workflow of LC-MS/MS based shotgun proteomics [9].

spectrometer selects a peptide ion generated from the first mass analyzer to get it fragmented and a MS/MS spectrum will be produced from the second mass analyzer. Peptides can be characterized from the MS spectra and MS/MS spectra by a variety of computational approaches. Therefore, the protein sequence can be determined by aligning peptide sequences to existing sequence database or generating a new, unique protein sequence [65–67].

2.2.2.2 Fragmentation and ion type

Peptides precursors are fragmented before they are transferred to the second mass analyzer. Theoretically, one specific peak of the MS/MS spectra stands for one specific fraction of the peptide. There are six basic types of the fragmented ions which can

be classified according to where the cleavage occurs. Fragment ions retained with the amino group (the N-terminus) are a-ion, b-ion and c-ion. Fragment ions retained with the carboxylic acid group (the C-terminus) are x-ion, y-ion and z-ion. As Figure 2.14 and Figure 2.15 shows, the subscript following each label indicates the sequential number of amino acid residue contained in the fragmented ion.

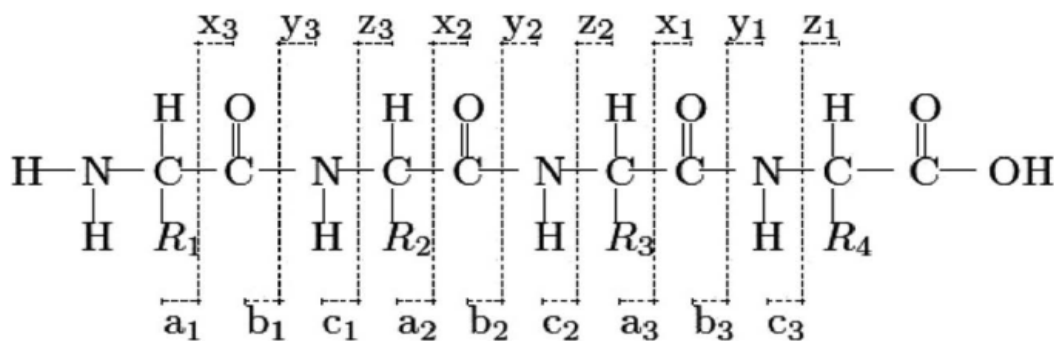


Figure 2.14: Cleavage of Six basic types of fragment ions [10]

Besides six basic types of fragment ions, there are many other types of fragments. If two backbone cleavages happen at the same time, it will generate internal fragment ions. The so-called amino-acylium ion can be formed by combination of *b* type and *y* type cleavage to produce the illustrated structure. It's called an amino-immonium ion if the internal cleavage ions are formed by combination of *a* type and *y* type cleavage. Immonium ion is an internal fragment with only a single side chain formed by a combination of *a* type and *y* type cleavage (Figure 2.16). The side chains of the precursor may also be fragmented to generate several types of satellite ions, including the d-ions, v-ions and w-ions.

Fragment ions observed in an MS/MS spectra depend on many factors including the

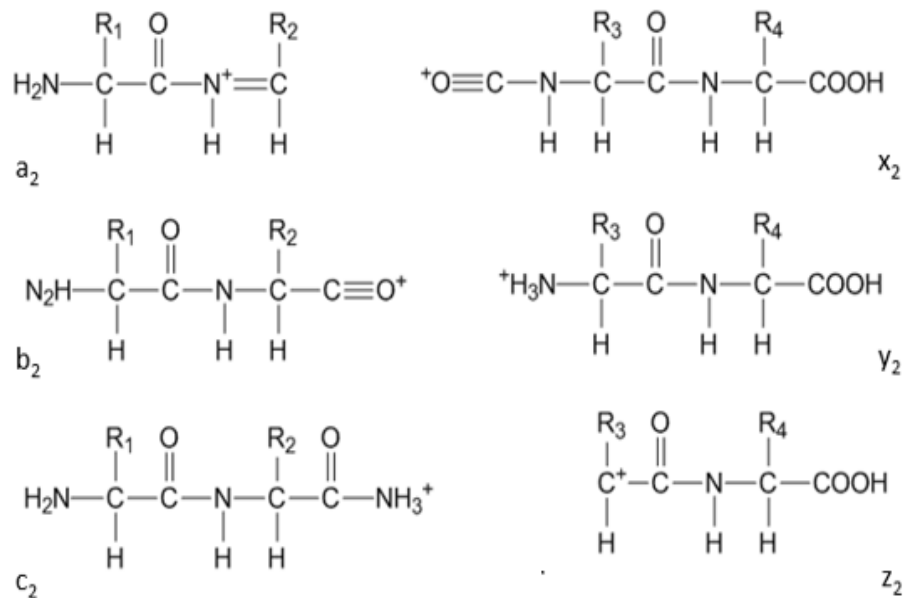


Figure 2.15: Six basic types of fragment ions [10]

primary sequence of peptide, the amount of internal energy, how the energy was introduced and charge state. Different fragmentation techniques for tandem mass spectrometry tend to generate different types of ions. There are three commonly used fragmentation methods in the current mass spectrometry based shotgun proteomics which are the Collision-Induced Dissociation (CID) [68], the Higher-energy Collisional Dissociation (HCD) [69], and the Electron-Transfer Dissociation (ETD) [70]. Collision-induced dissociation (CID) is also known as collisionally activated dissociation (CAD). The precursor ions are usually accelerated by applying an electrical potential to increase the ion kinetic energy and collided with neutral molecules like helium, nitrogen or argon. Some of the kinetic energy is converted into internal energy that induces the breakage of the peptide backbone during the collision process. While y-ion and b-ion are the most frequently observed ion types of the MS/MS spectra generated from CID fragmentation. Higher-energy Collisional dissociation (HCD) is an

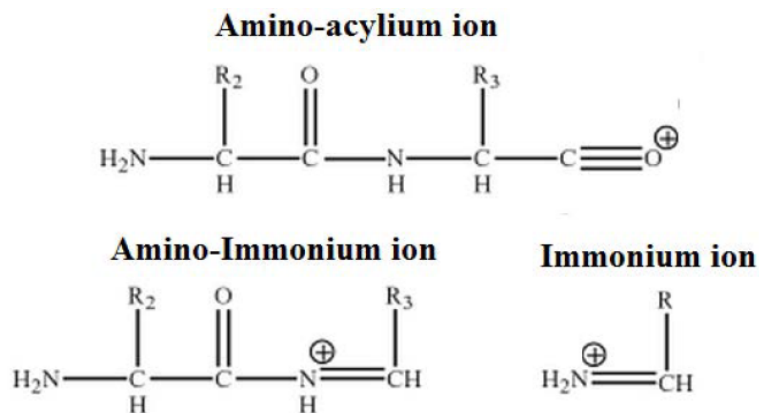


Figure 2.16: Structure of amino-acylium ion, amino-immonium ion and immonium ion [11]

improved CID technique specific to the orbitrap mass spectrometer that the MS/MS spectra generated by HCD always have more accurate m/z values. Compared with CID, HCD can generate not only y -ion and b -ion but also more a -ions and other smaller fragments from which more info regarding the protein structure and composition of the molecule can be acquired. The Electron-Transfer Dissociation (ETD), similar to the Electron-Capture Dissociation, induces peptide fragmentation of large, multiply-charged cations by transferring electrons to them. In the MS/MS spectra generated from ETD fragmentation, c -ion and z -ion are the most frequently observed ion types.

2.3 Interpreting MS/MS spectra

Protein identification is the most relevant application of mass spectrometry in biological proteomic studies. Structural information embedded within the target mass spectra must be efficiently and accurately interpreted to yield a trustworthy protein

sequence (Figure 2.17). Intensive research has been pursued in efforts to develop and optimize computational tools for mass spec data interpretation. Namely, database search and peptide de novo sequencing are the two main computational approaches for mass spectra analyses.

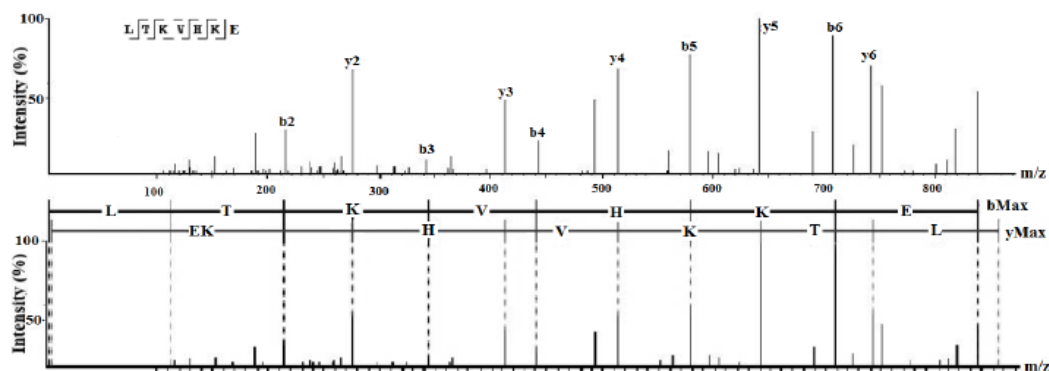


Figure 2.17: Example of a spectrum of a successfully identified peptide LTKVHKE [12]

2.3.1 Database search

In the database search approach, the assumption is that a database contains all the target proteins. The computational task is to select the spectrum of correct proteins from the database. Generally, approaches for database search are very similar. Figure 2.18 shows the workflow of database search approach, it includes the following steps.

1. Proteins are theoretically digested to the peptides. If the peptide sequences satisfy the precursor mass value of the input MS/MS spectra, it is retrieved as the potential candidates.

2. Theoretical spectra are predicted from the selected peptides according to certain fragmentation rules.
3. Experimental MS/MS spectra are compared with all the predicted spectra based on an appropriate scoring function.
4. Peptide candidates are then ranked according to the scoring function and the top ranked candidate will be output.
5. Proteins in the sample can be identified with peptide candidates.

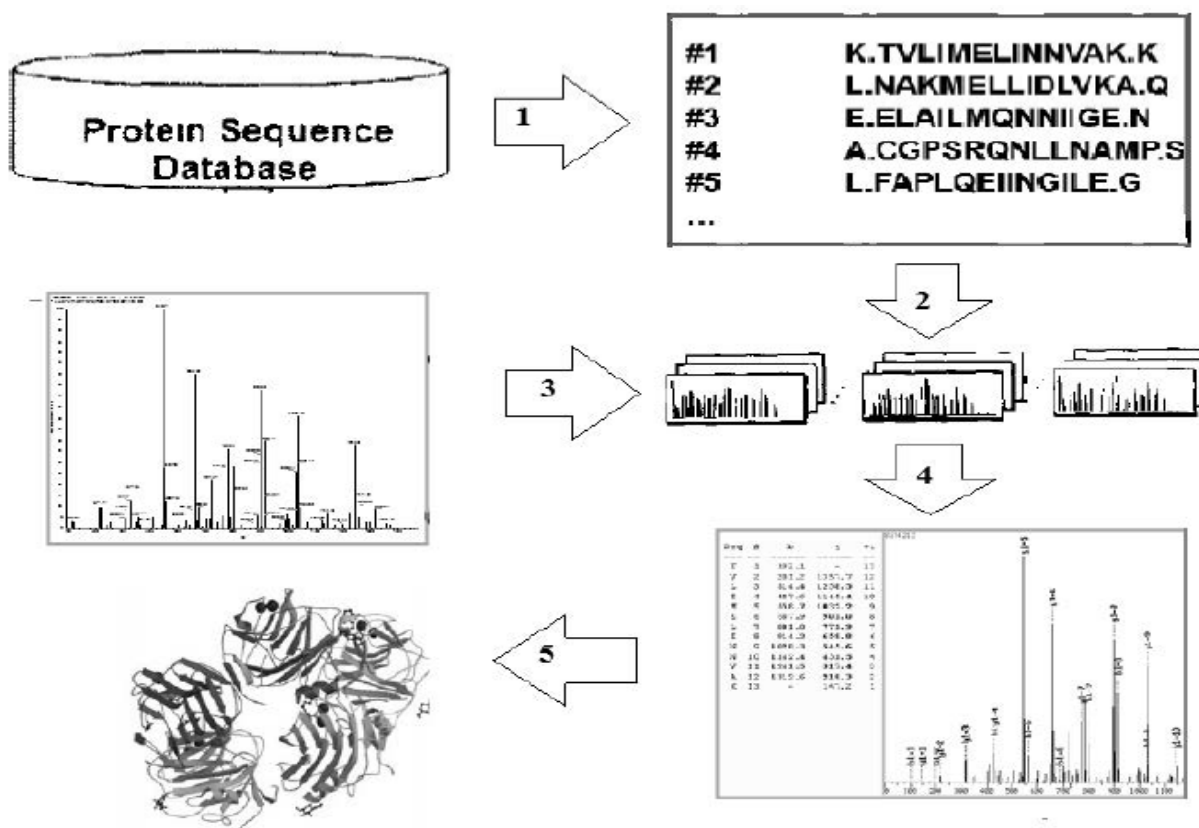


Figure 2.18: Workflow of database search approach [12].

Peptide-spectra Match (PSM) score is calculated to construct the best matched peptide sequence. The score describes the quality of a match between the candidate

peptide and a given spectrum. Thus, it is important to design an effective scoring function to achieve the peptide identification accuracy. A good scoring function should consider many factors like the intensities of peaks, the number of matched peaks and the mass errors amongst other contributing factors. Fragment ion type is a significant factor which should be considered during the scoring procedure. This is because certain types of mass spectrometer usually produce peaks with higher intensity for certain ion types.

2.3.2 Peptide de novo sequencing

De novo sequencing is the process of constructing peptide sequences directly from tandem mass spectra without the assistance of a sequence database (Figure 2.19). This method owns the obvious advantage to be considered for the situation when a target protein or peptide is not included in the sequence database. De novo sequencing process searches for the optimal combination of amino acids to match the input MS/MS spectra. Initially, the analysis has largely been performed manually yet high throughput data analysis is another choice. Database search approach can be as simple as enumerating every possible peptide sequence in the database with proper mass values. However, it will take exponential time for de novo sequencing enumerating every possible amino acid combination for a given molecular mass value. Thus, it is also important for de novo sequencing to design more efficient algorithm component in order to seek the optimal solution. In comparison with database searching, de novo sequencing is harder since that it requires much higher quality data to derive the

complete sequence and suffers low accuracy of the reported result when the complete sequence can not obtain. In recent years, with more advances in both algorithms and instrument analysis, de novo sequencing has been able to identify sequences with higher accuracy and better coverage.

There are still some limitations and problems for de novo sequencing. The first one is that the spectra is typically noisy. Pre-processing and denoising approaches are needed to solve the problem [71], such as the pre-processing in PEAKS and the work done by Sridhara et al [72]. In addition, considering the property of de novo sequencing, the ions used to construct ion ledgers should have the same ion type so that ion type separations are helpful in the subsequent peptide sequencing. The second problem is missing data from MS/MS spectra. Though researchers have tried to get as much useful information as possible from the constructed models, usually they are still not able to successfully reflect the actual MS/MS spectra, and there is still a need to generate more accurate information from the outset [51].

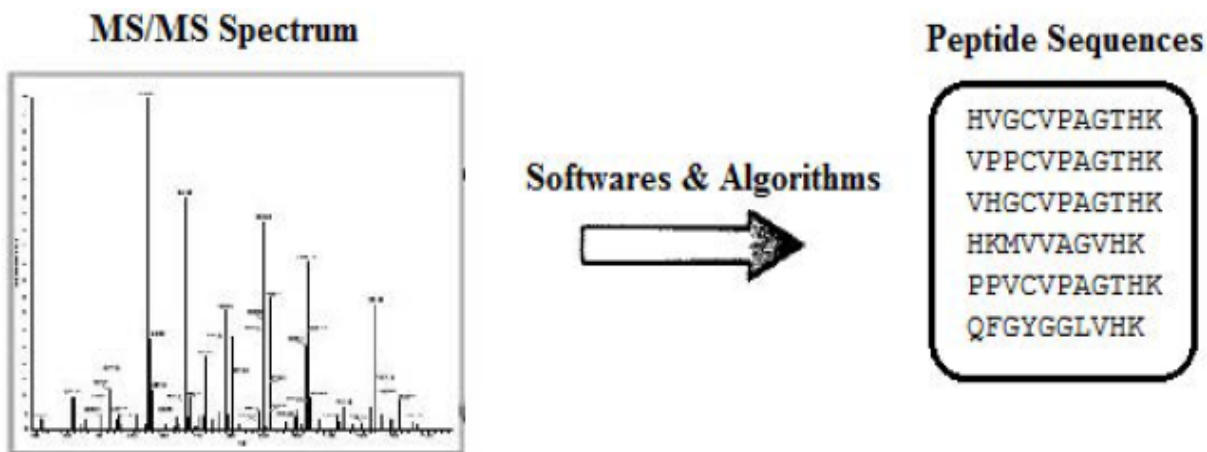


Figure 2.19: Workflow of de novo sequencing [12].

To address this issue, there is a mathematic model can convert spectra into its cor-

responding spectra graph and the solution is to find an optimal path on the graph. Figure 2.20 shows the general procedure of de novo sequencing derived by the spectra graph model. The target MS/MS spectrum is converted to a spectrum graph in which each edge corresponds to an occurrence that the m/z difference of two peaks in the spectra equals to the mass value of a certain amino acid residue. A dynamic programming algorithm is designed to find the path which represents the best solution.

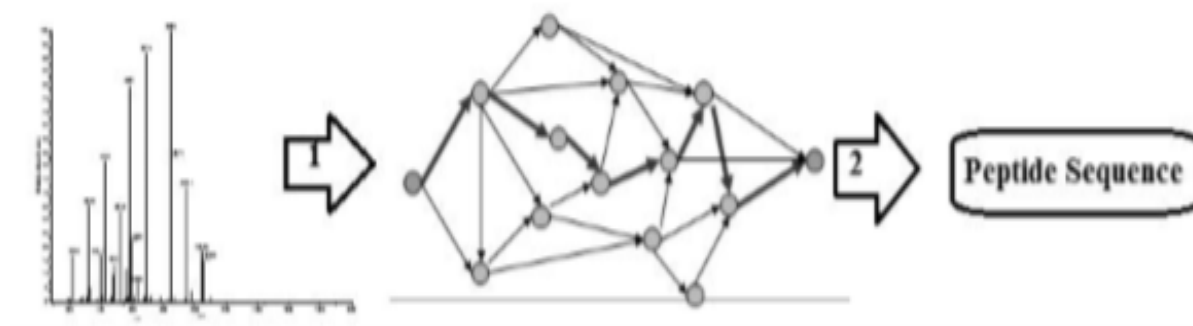


Figure 2.20: Schematic of graph spectra model [12]

Bin Ma and Kaizhong Zhang proposed an effective algorithm for peptide de novo sequencing from MS/MS spectra [13]. Their algorithm try to solve some major difficulties of the de novo sequencing problem. First, each fragmentation may produce a pair of ions. Therefore, both ends of the spectrum must be considered at the same time, in order to evaluate the score contributed by the pair of ions caused by the same fragmentation. Secondly, the types of the peaks are unknown, and a peak might be matched by zero, one, or two different types of ions. In the case a peak is matched by two ions, the height of the peak can only be counted once. An algorithm that counts the height twice has the tendency to match the highest peaks more than once,

instead of matching more peaks. Therefore, the algorithm should know whether a peak has already been matched, before it can evaluate the match between the peak and a new ion. Third, to construct optimal sequence, the algorithm should consider both of ion ladders at same time. Therefore, both of y ion ladders and b ion ladders for the candidate sequence should be construct during the computation.

Because of the difficulties, a straightforward dynamic programming approach, which attempts to construct the optimal peptide from one terminus to the other, does not work. A more sophisticated dynamic programming algorithm is given for the de novo sequencing problem. The algorithm gradually construct optimal pairs of prefixes and suffixes in a carefully designated pathway, until the prefix and the suffix become sufficient long to form the optimal solution.

Let $M = ||P|| + 20$. Let $A = a_1a_2\dots a_k$ be a string of amino acids. If A is a prefix (at the N-terminus) of $P = a_1a_2\dots a_n$, then the mass of the b-ion produced by the fragmentation between a_i and a_{i+1} is $||a_1a_2\dots a_i||_b$. The mass of the y-ion caused by the same fragmentation is $M - ||a_1a_2\dots a_i||_b$. Denote $S_N(A)$ as the set of masses of all the ions caused by the fragmentation between a_i and a_{i+1} .

Similarly, suppose that $A' = a'_k\dots a'_2a'_1$ is a suffix (at the C-terminus) of a peptide $P' = a'_n\dots a'_2a'_1$ and $M = ||P|| + 20$. Denote $S_C(A)$ as the set of masses of all the ions caused by the fragmentation between a'_{i+1} and a_i .

Let peptide P be the optimal solution. For any string A , A' and an amino acid a , such that $P = AaA'$, the following fact is obvious as formula shows(Figure 2.21).

This suggests us to reduce the de novo sequencing problem to the problem of finding

$$\begin{aligned}\mathcal{S}_N(A) &= \bigcup_{i=1}^k [B(\|a_1 a_2 \dots a_i\|_b) \cup Y(M - \|a_1 a_2 \dots a_i\|_b)]. \\ \mathcal{S}_C(A') &= \bigcup_{i=1}^k [Y(\|a'_i \dots a'_2 a'_1\|_y) \cup B(M - \|a'_i \dots a'_2 a'_1\|_y)] \\ \mathcal{S}(P) &= \mathcal{S}_N(A) \cup \mathcal{S}_C(A')\end{aligned}$$

Figure 2.21: De novo sequencing problem definitions from Bin Ma and Kaizhong Zhang approaches. [13]

the appropriate prefix A and suffix A' . There are $l(P)$ different ways to divide P into the form AaA' , where $l(P)$ denotes the length of P .

2.3.3 Multiple MS/MS spectra identification

De novo peptide sequencing methods have challenges in correctly identifying peptide sequences. However, only one MS/MS spectrum is used to conduct peptide sequencing by traditional de novo peptide sequencing methods. The main shortcoming of these approaches is the limited fragmentation information extracted from the only one type of experimental spectra [44]. Therefore, identification accuracy is relatively low. Apart from using more high quality data for sequencing, there is another way to improve performance that is to combine multiple spectra from the same peptide but from different technologies to conduct de novo sequencing [73]. For example, combining the CID spectra and ETD spectra of the same peptide during the de novo procedure, the b-ion and c-ion, as well the y-ion and z-ions can be matched each other and fill out the missing gaps that may occur in each individual fragmentation

mode [74–76]. Combination of multiple spectra greatly increases the possibility of deriving a complete peptide sequence and reduces the misinterpretation of other peaks as ion ladder peaks.

Some major developments of utilizing multiple spectra for de novo peptide sequencing are discussed and compared. Although the current trend is to use a pair types of spectra, some of the methods have the potential or are already able to utilize three or more types of spectra. The first documented method designed for multiple spectra sequencing is the one developed by Savitski et al. [77]. This method utilized spectra of the same peptide from collision activated dissociation (CAD) and electron-capture dissociation (ECD) together. Their algorithm first used ions that had supporting ions in the other spectra to create a backbone of the sequence and then used complementary ion pairs and other ions from the two spectra to extend the sequence until a full sequence was obtained [44]. There are other studies which use multiple spectra. Altelaar et al. [78] utilized a special enzyme to break the peptide backbone at the N-terminal side of lysine residues, and combined CID and ETD spectra to conduct peptide sequencing. Guthals et al. [7] used MS/MS triplets (CID/HCD/ETD) from overlapping peptides produced by different enzymes to infer peptide or even protein sequences. Shen et al. [73] compared the peptide sequencing performance of different methods when using various types of spectra solely or together. Recently, Jeong et al. [79] proposed a universal de novo sequencing tool based on spectra graph model and even claimed it could be used for various types of spectra.

Yan Yan, Anthony J. Kusalik, and Fang-Xiang Wu [14] proposed an algorithm of de novo peptide sequencing for multiple tandem mass spectra. Their spectra merging

method is valuable for combining information from multiple tandem mass spectra. Their basic idea of spectra merging is to select signal ions from each spectrum to form a new spectrum, denoted as S_m , which contains more useful information. Two kinds of relationships between ions are considered: amino acid mass difference and complementarity.

For amino acid difference, with consideration of regular amino acid masses and loss of H_2O and NH_3 , all 2-tags (two amino acids long) in each spectrum are first produced. Here, a length-2 tag consists of three ions (any two consecutive ion pairs having mass difference close to an amino acid mass or an amino acid mass minus some small molecular) from an experimental spectrum. The three consecutive ions can be denoted as $u, v, t \in$ spectrum S . Without loss of generality, we say the masses of u, v, t are in an increasing order. Then, ion v is named as a “middle ion” and selected for the merged spectrum. The middle ion, which has two other supporting ions from the spectrum, is more likely to be a signal ion rather than noise. For the complementary relationship, all complementary ion pairs in each spectrum are selected.

Denoted u^+, v^+, t^+ as the values of these ions in charge state 1. Additionally, we denote A as the set of 20 amino acids, and $a_i \in A$ as a certain amino acid. a_i is also used to represent its residue mass. m_{loss} is defined to be the mass of some small molecules or groups lost from fragment ions, which includes H_2O and NH_3 . Relationships and ions in table are then utilized (Figure 2.22). In this table, $a_i, a_j \in A$; σ can be 0 or m_{loss} (considering the loss of small molecules of fragment ions); θ is a given threshold and m_p is the parent peptide mass. Finally, S_m consists of all middle ions in 2-tags and complementary ion pairs for a given pair of MS/MS spectra.

RELATIONSHIPS AND IONS SELECTED IN SPECTRA MERGING

Relationship	Ions selected
$ (v^+ - u^+) - a_i + \sigma \leq \theta$ and $ (t^+ - v^+) - a_j - \sigma \leq \theta$	v (middle ion)
$ m_p + 2m_H - (v^+ + u^+) \pm \sigma \leq \theta$	v and u if $u, v \in S_c$
$ m_p + 3m_H - (v^+ + u^+) \pm \sigma \leq \theta$	v and u if $u, v \in S_e$

Figure 2.22: Relationships and ions selected in spectra merging [14]

When they implement their algorithm, all ion types of peaks are considered. For example, when de novo sequencing using S_m which is merged by CID spectrum (Figure 2.23) and ETD spectrum (Figure 2.24), the ions listed in tables below are considered in the proposed method based on the availability of spectra and observed frequency of different ions.

ION TYPES CONSIDERED IN CID/HCD SPECTRA

Ion Type	Mass calculation from residues	Mass calculation from other ions
a	$\sum(\text{residue mass}) - 26.9871$	$b_m - m_{CO}$
b	$\sum(\text{residue mass}) + 1.0078$	b_m
x	$\sum(\text{residue mass}) + 44.9977$	$y_m + m_{CO}$
y	$\sum(\text{residue mass}) + 19.0814$	y_m

Figure 2.23: Ion types considered in CID/HCD spectra [14]

ION TYPES CONSIDERED IN ECD/ETD SPECTRA

Ion Type	Mass calculation from residues	Mass calculation from other ions
c	$\sum(\textit{residue mass}) + 18.0344$	$b_m + m_{NH_3}$
$c - 1$	$\sum(\textit{residue mass}) + 17.0265$	$b_m + m_{NH_3} - m_H$
z	$\sum(\textit{residue mass}) + 3.0156$	$y_m - m_{NH_3} + m_H$
$z + 1$	$\sum(\textit{residue mass}) + 4.0156$	$y_m - m_{NH_3} + 2m_H$
w	$\sum(\textit{previous residue mass}) + 73.0290$	$x_{\textit{previous}} + m_{CO}$
b	$\sum(\textit{residue mass}) + 1.0078$	b_m
y	$\sum(\textit{residue mass}) + 19.0814$	y_m

Figure 2.24: Ion types considered in ECD/ETD spectra [14]

2.4 Stable isotope labeling and SILAC

2.4.1 Stable isotope labeling

Stable isotopes have the same number of protons as common elements and consequently share the same physicochemical properties. However, they have different mass due to the difference in number of neutrons. Among biochemically relevant elements, carbon, hydrogen, nitrogen, oxygen and sulfur all have two or more stable isotopes with measurable abundance in nature. The isotopes commonly used are ^{13}C , ^{15}N , 2H (deuterium) and ^{18}O . Isotopologs are metabolites containing stable isotopes and their unlabeled counterparts which have the same chemical formula and structure. Although they behave similarly during chromatographic separation, mass spectrometer can readily differentiate isotopologs by their m/z [15]

Stable isotope labeling involves the use of non-radioactive isotopes which can act as

tracers to model several chemical and biochemical systems. The reactant is "labeled" by replacing specific atoms with its isotopes. The reactant is then allowed to undergo the reaction. Stable isotope labeling can be exploited to study the metabolic pathway of peptides and track the changes of the position and quantity of polypeptides. Stable isotope labeling has become prevailing in the area of comparative proteomics (or differential proteomics) study where two or three samples need to be compared. For example, one sample can be labeled with the naturally occurring isotope abundance (light) element and the other with a stable isotope of low natural abundance (heavy) element. Thereafter, samples can be mixed and analyzed by mass spectrometry [80]. Several different strategies involving stable isotopes label like ICAT, ICPL, IDBEST, iTRAQ, TMT, IPTL and SILAC.

2.4.2 SILAC

SILAC refers to stable isotope labeling with amino acids in cell culture. This method is applied to quantitative proteome research, providing an effective scheme for comprehensive and systematic qualitative and quantitative analysis of complex cell proteome. SILAC technology uses medium containing light, medium or heavy isotope-labeled essential amino acids (mainly lysine and arginine) to label newly synthesized proteins with stable isotopes during cell growth. Two or three cell populations are grown in culture media that are identical except that one of them contains a 'light' and the other two contains a 'medium' or a 'heavy' form of a particular amino acid. As the three

isotopically labeled amino acids are essentially chemically identical, their incorporation does not interfere with normal cell growth, while leading to proteins/peptides that are distinguishable by mass and thus are ideal for mass spectrometric analysis [43]. Figure 2.25 shows the flowchart of triple SILAC coupled with LC-MS/MS for the comparative analysis of three distinct cell populations.

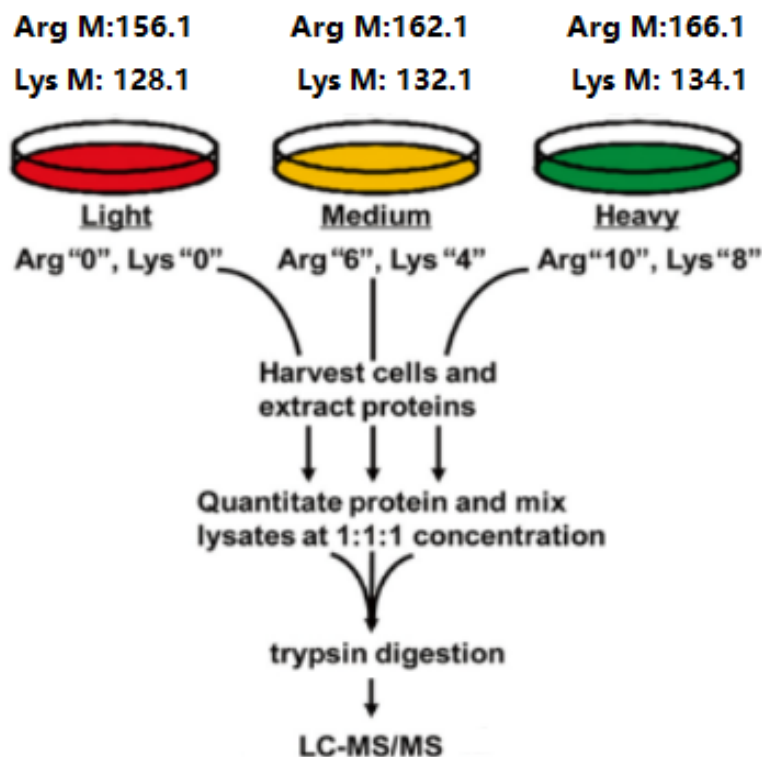


Figure 2.25: Flowchart of triple SILAC coupled with LC-MS/MS for the comparative analysis of three distinct cell populations [15]

Figure 2.26 is an example of SILAC MS spectra with arginine labeled by $^{12}C^{14}N$, $^{13}C^{14}N$ and $^{13}C^{15}N$. The spectra consists of a set of peaks with each individual m/z value and intensity. Each peptide is displayed in three distinct forms which are derived from three types of labeled arginine. The peptide and its differentially la-

beled ones are specified in the spectra based on the principle that the differences of molecular weight between $^{12}\text{C}^{14}\text{N} - ^{13}\text{C}^{14}\text{N}$ and $^{13}\text{C}^{14}\text{N} - ^{13}\text{C}^{14}\text{N}$ are 6Da and 4Da, respectively [16].

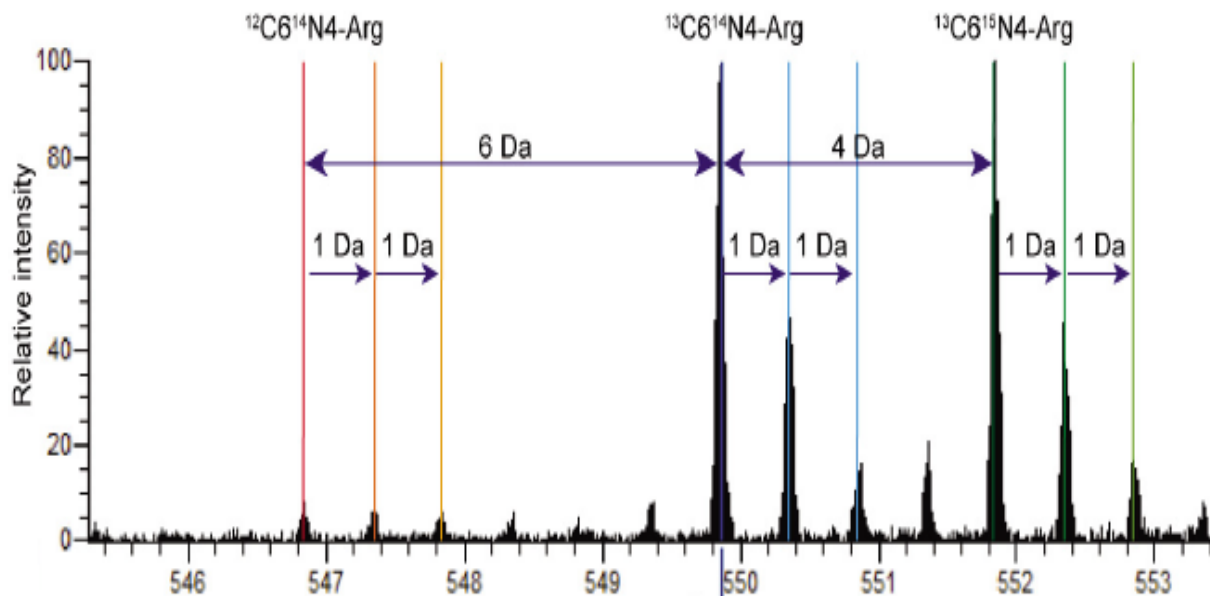


Figure 2.26: Example of SILAC MS spectra [16]

SILAC is widely used in comparative proteomics, protein-protein interaction, protein-DNA interaction and protein-RNA interaction, etc.

Chapter 3

De novo sequencing of multiple SILAC labeled MS/MS spectra

3.1 Introduction

Database search is an effective method to identify peptide sequence from MS/MS spectra. However, it has several shortcomings such as incomplete genome sequencing, inferior gene prediction from the genome, the sequence variations between two individuals of the same species or the target proteins may not be included in the sequence database. De novo sequencing is a computational method to construct peptide sequence but needs further optimization due to the imperfect data. In this case, de novo sequencing become a reasonable choice to identify the peptides from multiple MS/MS spectra this is because more spectra supplements mean more data will be available for identification.

Isotope labeling project (for example, SILAC based) contains same protein molecules

with different types of isotope labeling in its sample. Multiple MS/MS spectra for the same peptide sequence are produced by spectrometer after the sample was processed by LC-MS/MS shotgun proteomics. Based on the factors such as the type of isotope labeling, retention time, precursor ion mass, multiple spectra with different types of SILAC modifications for the same peptide can be used to identify the peptide sequence. Thus, our research focus on de novo sequencing method that aims to construct candidate peptide sequences which can best match spectra and locate the position of SILAC modifications from multiple spectra.

Prior to this research, no de novo sequencing algorithm has been reported for peptide identification from multiple MS/MS spectra of peptide containing SILAC labeling. Previous approach for multiple spectra interpretation does not suitable since the peptide sequences are identified after all information are merged. However, for a specific peak of SILAC based spectrum, to located its corresponding peaks from other spectra and incorporate all information, we have to know the type and the quantity of SILAC modifications it represents which are the information we can get only after the identification. In this case, the methods of previous approaches does not work. In this thesis, we propose algorithms that try to incorporate all information during the computation of identification.

Incorporating all information from spectra sets has advantages. It can strengthen the peak which are matched in those different sets of spectra. Valuable data such as forming ion ladders are much easier be achieved. But noise peaks effects are a more serious problem for masses that don't find strong peak support.

3.2 Problem Definition

3.2.1 Notations

In this approach, we extend the notations from the approach of Ma Bin and Kaizhong Zhang [13].

We use a set of 2-mers to represent an experimental mass spectrum. Let $S = \{(x_i, h_i) | 1 \leq i \leq n\}$ be a set of 2-mers where x_i is the mass and h_i is the intensity of the peak at x_i . Associated with S , there is a precursor ion mass of the peptide in the spectrum S , denoted as $m(S)$. Let Σ be the alphabet that represents the 20 different types of amino acids. Two amino acids Isoleucine(I) and Leucine(L) have exactly same mass values, so we consider them as the same amino acid in our study. When a MS/MS spectrum is produced by spectrometry, we call it raw data. Before being identified, raw data will be pre-processed by standard software packages and methods such as de-convolution. During this pre-processing, multiple charged peaks in spectra will be converted to their singly charged equivalents [81]. Thus, in this research we assume all peaks in spectra are with charge one ($z=1$) and m/z value equals to its mass value(m).

3.2.2 Ion fragments and ion mass

A length n peptide p is an amino acid string $p = \{a_1 a_2 \dots a_n\}$, where $a_i \in \Sigma$. After p is fragmented in a MS/MS spectrometer, each prefix or suffix of the string can form several types of fragment ions. The ion's masses can be computed accurately. Denote $||a||$ as the mass of amino acid residue a . The residue mass of the peptide p

is $\|p\| = \sum_{1 \leq i \leq n} \|a_i\|$ and the actual mass of the peptide p is $\|p\| + \|H_2O\|$.

Denote b_i as the mass of the b-ion (prefix) of p with i amino acids. b_i can be computed with $b_i = 1.01 + \sum_{1 \leq j \leq i} \|a_j\|$. Denote y_i as the mass of the y-ion (suffix) of P with i amino acids. y_i can be computed with $y_i = 19.02 + \sum_{n-i+1 \leq j \leq n} \|a_j\|$. Therefore we have

$$b_{n-i} + y_i = \|p\| + 20.03 \quad 1 \leq i \leq n$$

Let x be the mass value of a b-ion. The corresponding a-ion and c-ion locate at $x - 27.99$ and $x + 17.03$ respectively. A b-ion may lose an ammonia or a water which generates new ion mass $x - 17.03$ and $x - 18.01$ respectively. Denote $B(x)$ as the set of all ion masses corresponding to the b-ion located at x , we have

$$B(x) = \{x, x - 17.03, x - 18.01, x + 17.03, x - 27.99\}$$

For each y-ion with mass x , we have set $Y(x)$ as

$$Y(x) = \{x, x - 17.03, x - 18.01, x + 25.98\}$$

Note that y-ion lost an ammonia and its corresponding z-ion are both $x - 17.03$. For a given peptide p , we use M_p to represent the set of masses of the hypothetical spectrum of peptide p .

$$M_p = \bigcup_{i=1}^{n-1} (B(b_i) \cup Y(y_i))$$

3.2.3 Single MS/MS spectra sequence matching and de novo sequencing

We use $m(p) = \|p\| + \|H_2O\| + 1 = \|p\| + 19$ to denote the actual mass of p when p is measured by a mass spectrometer. For a given experimental spectrum S and a peptide p , let $M_p(S)$ be a subset of S supporting peptide p .

$$M_p(S) = \{(x, h) \in S \mid \exists y \in M_p, \text{ s.t. } |x - y| \leq \delta\}$$

$M_p(S)$ represents that quantity of peaks from M_p matches peaks from experimental spectrum S . When matching two peaks from M_p and S , a small error tolerance δ is accepted. Depending on different spectrometers, the maximum error tolerance is from ± 0.01 Dalton to ± 0.5 Dalton. Note that not only the quantity of peaks, but also the quality of peaks are significant. The more and higher peaks are matched, the more likely the peptide p is correct.

Given a peak (x, h) , $f(h)$ is a function of intensity to weighted h . In addition of intensity, it can involve many factors such as the type of ion that explains h and the mass error between the ion and x . Denote $\sigma(S, M_p)$ as the matching similarity score between p and S . We have:

$$\sigma(S, M_p) = \sum_{(x,h) \in M_p(S)} f(h)$$

We now define the peptide sequencing problem as to find a peptide p with maximum matching similarity score satisfying the condition that the mass of p and the precursor ion mass of S are within the error tolerance δ .

$$P(S) = \arg \max_{\{p \mid |m(p) - m(S)| \leq \delta\}} \sigma(S, M_p)$$

3.2.4 De novo sequencing to multiple MS/MS spectra of peptide containing SILAC labeling

In a SILAC experiment, samples labeled by different Lysine(K) and Arginine(R) modifications are combined and separated by LC and then analyzed by MS and tandem mass spectrometry (MS/MS). Therefore the multiple MS/MS spectra from the same peptide are produced. The mass of Lysine(K), Arginine(R) and peptide will increase based on the specific chemical isotope labeling elements. Ideally, we assume it is fully reacted that all the Lysine(K) and Arginine(R) in the sample have been labeled. Based on the experiment design, a peptide sequence should have only one type of SILAC modification(either light, medium or heavy) no matter the modification is on Lysine(K) or Arginine(R). In this research, we assume that the type of SILAC modification used in the experiment is known so that the algorithm is designed based on the specific conditions.

For Lysine, denote $\|K\|_L$, $\|K\|_M$ and $\|K\|_H$ as the mass of residue labeled by light, medium and heavy modification respectively. For Arginine, denote $\|R\|_L$, $\|R\|_M$ and $\|R\|_H$ as the mass of residue labeled by light, medium and heavy modifications respectively. Denote $\|p\|_L$, $\|p\|_M$, and $\|p\|_H$ as the residue masses of the peptide p with light, medium, and heavy modifications respectively which can be computed with $\|p\|$, $\|K\|_L$, $\|K\|_M$, $\|K\|_H$, $\|R\|_L$, $\|R\|_M$ and $\|R\|_H$. Given a peptide p , we use M_p^L , M_p^M , and M_p^H to represent the set of masses of the hypothetical spectrum of peptide p with light, medium, and heavy modifications for Lysine and Arginine, respectively. We use $m_L(p)$, $m_M(p)$, and $m_H(p)$ to represent the mass of p measured

by spectrometer with light, medium, and heavy modifications, respectively. For a given experimental spectrum S and a peptide p , let $M_p^L(S)$, $M_p^M(S)$ and $M_p^H(S)$ be a subset of S supporting peptide p with light, medium, and heavy modifications, respectively.

$$M_p^L(S) = \{(x, h) \in S \mid \exists y \in M_p^L, \text{ s.t. } |x - y| \leq \delta\}$$

$$M_p^M(S) = \{(x, h) \in S \mid \exists y \in M_p^M, \text{ s.t. } |x - y| \leq \delta\}$$

$$M_p^H(S) = \{(x, h) \in S \mid \exists y \in M_p^H, \text{ s.t. } |x - y| \leq \delta\}$$

$\sigma(S, M_p^L)$, $\sigma(S, M_p^M)$, and $\sigma(S, M_p^H)$ are the matching similarity scores between p and S when Lysine and Arginine in p are subject to light, medium, and heavy modifications.

$$\sigma(S, M_p^L) = \sum_{(x,h) \in M_p^L(S)} f(h)$$

$$\sigma(S, M_p^M) = \sum_{(x,h) \in M_p^M(S)} f(h)$$

$$\sigma(S, M_p^H) = \sum_{(x,h) \in M_p^H(S)} f(h)$$

Given a peptide p and let $S_L = \{S_{L_1}, \dots, S_{L_b}\}$, $S_M = \{S_{M_1}, \dots, S_{M_c}\}$, and $S_H = \{S_{H_1}, \dots, S_{H_d}\}$ be the sets of experimental spectra with light, medium, and heavy modifications respectively. The matching similarity score between p and the set of experimental spectra $\{S_L, S_M, S_H\}$, $\sigma(S_L, S_M, S_H, M_p^L, M_p^M, M_p^H)$, is defined as the summation of the matching similarity scores between p and each individual experimental spectrum.

$$\sigma(S_L, S_M, S_H, M_p^L, M_p^M, M_p^H)$$

$$\begin{aligned} &= \sum_{i=1}^b \sigma(S_{L_i}, M_p^L) + \sum_{i=1}^c \sigma(S_{M_i}, M_p^M) + \sum_{i=1}^d \sigma(S_{H_i}, M_p^H) \\ &= \sum_{i=1}^b \sum_{(x,h) \in M_p^L(S_{L_i})} f(h) + \sum_{i=1}^c \sum_{(x,h) \in M_p^M(S_{M_i})} f(h) + \sum_{i=1}^d \sum_{(x,h) \in M_p^H(S_{H_i})} f(h) \end{aligned}$$

We now define the peptide SILAC sequencing problem as to find a peptide p with maximum matching similarity score satisfying the condition that the mass of p and the precursor ion mass of each spectrum in S_L, S_M, S_H are within the error tolerance.

$$P(S_L, S_M, S_H) = \arg \max_p \left\{ \begin{array}{l} \max \sigma(S_L, S_M, S_H, M_p^L, M_p^M, M_p^H) \\ |m_L(p) - m(S_{L_i})| \leq \delta, \quad 1 \leq i \leq b \\ |m_M(p) - m(S_{M_i})| \leq \delta, \quad 1 \leq i \leq c \\ |m_H(p) - m(S_{H_i})| \leq \delta, \quad 1 \leq i \leq d \end{array} \right.$$

3.3 Algorithms for de novo sequencing of multiple MS/MS spectra of peptide containing SILAC labeling

Given a spectra set $S = \{S_{L_1}, S_{L_2} \dots S_{L_b}, S_{M_1}, S_{M_2} \dots S_{M_c}, S_{H_1}, S_{H_2} \dots S_{H_d}\}$, we first convert each spectrum into an intensity array indexed by mass integer. If for a mass value m , there is no corresponding intensity, then its intensity is set to be zero. Let Σ be an amino acid residue set which contains all amino acid residues and their PTMs. Let Σ_m be an amino-acid residue set which contains modified Lysine(K) and modified Arginine(R).

To incorporate all the information from spectra set S , the most difficult challenge is to determine the corresponding peak p_M in S_M and p_H in S_H of a given peak p_L in S_L . That is because p_M and p_H have shift masses with p_L . For a specific peak of SILAC based spectrum, to located its corresponding fragment ion, we have to know the type and the quantity of SILAC modifications to determine the shift masses. In this case, we need to know the whole sequence but this is the goal of identification. To solve this problem, we propose two algorithms to compute the peptide sequence which are based on total number of modifications and based on SILAC modification pairs

3.3.1 Algorithm based on total number of modifications

If $\|K\|_M - \|K\|_L = \|R\|_M - \|R\|_L$ and $\|K\|_H - \|K\|_L = \|R\|_H - \|R\|_L$, then for any mass position, m , for a peptide with n modified Lysine or Arginine, in a spectrum in S_L , the corresponding mass position in a spectrum in S_M , is $m + n(\|K\|_M - \|K\|_L)$, and in a spectrum in S_H , is $m + n(\|K\|_H - \|K\|_L)$. From $m_L(p) = m(S_{L_i})$, $m_M(p) = m(S_{M_i})$, and $m_H(p) = m(S_{H_i})$, we can determine the number of modified Lysine or Arginine for the peptide we are looking for using the following formula. From

$$\|p\|_L + \|H_2O\| + 1 = m_L(p) = m(S_{L_i})$$

$$\|p\|_M + \|H_2O\| + 1 = m_M(p) = m(S_{M_i})$$

$$\|p\|_H + \|H_2O\| + 1 = m_H(p) = m(S_{H_i})$$

we have

$$n * (||K||_M - ||K||_L) = m(S_{M_i}) - m(S_{L_i}) \quad (3.1)$$

$$n * (||K||_H - ||K||_L) = m(S_{H_i}) - m(S_{L_i}) \quad (3.2)$$

This means that we can determine the total number of modifications for the peptide we are looking for, $n = (m(S_{M_i}) - m(S_{L_i})) / (||K||_M - ||K||_L)$ or $n = (m(S_{H_i}) - m(S_{L_i})) / (||K||_H - ||K||_L)$. With these information, we can design an algorithm based on the number of modifications.

For any mass value m in S_{L_i} and an amino acid a , a similarity score of matching a at location m , as an ending amino acid of a peptide p with mass m and n modifications, $as(n, m, a, S)$, is defined, where n is the number of modified K or R in p and S is the set of spectra. Let $\Delta_M = ||K||_M - ||K||_L$, $\Delta_H = ||K||_H - ||K||_L$. We have

$$\begin{aligned} as(n, m, a, S) &= \sum_{i=1}^b f(S_{L_i}[m - ||a||_L], S_{L_i}[m]) \\ &+ \sum_{i=1}^c f(S_{M_i}[m - ||a||_M + n * \Delta_M], S_{M_i}[m + n * \Delta_M]) \\ &+ \sum_{i=1}^d f(S_{H_i}[m - ||a||_H + n * \Delta_H], S_{H_i}[m + n * \Delta_H]) \end{aligned}$$

For any mass value m in S_{L_i} and a peptide p with matching mass value m including n modifications, the similarity score $ps(n, m, p, S)$ is defined as the summation of the similarity score for each individual amino acid in p . We can also define $ps(n, m, p, S)$ recursively. Assume the peptide sequence p forms $p'a$ consisting of a prefix sequence

p' and an amino acid residue a , we have

$$ps(n, m, p'a, S) = \begin{cases} ps(n, m - \|a\|_L, p', S) + as(n, m, a, S) & \text{if } a \in \Sigma \\ ps(n - 1, m - \|a\|_L, p', S) + as(n, m, a, S) & \text{if } a \in \Sigma_m \end{cases}$$

Note that the approach of Ma Bin and Kaizhong Zhang [13] try to construct two subsequence from both ends(for example, y-ion and b-ion of CID/HCD spectrum) until they become sufficient long to form the optimal solution. To simplify our computation, our algorithm only consider one sequence of the major ion of the spectrum(for example, y-ion sequence of CID/HCD spectrum).

The problem of finding an optimal solution is defined as finding a peptide with maximum similarity score among all the peptide which mass equals to $m(S_{L_i})$ and has $((m(S_{M_i}) - m(S_{L_i}))/\Delta_M)$ modifications, based on equation 3.1 and 3.2.

$$P(S) = \arg \max_{\left\{ p \left| \begin{array}{ll} |m_L(p) - m(S_{L_i})|, & 1 \leq i \leq b \\ |m_M(p) - m(S_{M_i})|, & 1 \leq i \leq c \\ |m_H(p) - m(S_{H_i})|, & 1 \leq i \leq d \end{array} \right. \right\}} ps(((m(S_{M_i}) - m(S_{L_i}))/\Delta_M, m(S_{L_i}), p, S)$$

To calculate $P(S)$, we use dynamic programming approach. Denote $DP[n, m]$ as the maximum similarity score between a peptide with mass m and n modifications with the spectra set S .

$$DP[n, m] = \max_p \{ps(n, m, p, S)\}.$$

Note that $\|a\|_L$, $\|a\|_M$ and $\|a\|_H$ equal to $\|a\|$ when $a \in \Sigma$.

Lemma 3.1 *When $n = 0$,*

$$DP[0, m] = \max_{a \in \Sigma} DP[0, m - \|a\|_L] + as(0, m, a, S)$$

Proof Assume the peptide sequence p with mass m and forms $p'r$ consisting of a prefix sequence p' and an amino acid residue a . $n = 0$ represents that in peptide p , there is no modified K and modified R in peptide sequence. So that the mass of p' is $m - \|a\|_L$ where $a \in \Sigma$ and the similarity score of p' is $DP[0, m - \|a\|_L]$. The similarity score of matching a at location m is $as(0, m, a, S)$. Note that p' must also be the optimal sequence by this recursive method, so that $DP[0, m - \|a\|_L]$ is the maximum similarity score for all the p' with mass $m - \|a\|_L$. Therefore, the similarity score of $p'a$ is $DP[0, m - \|a\|_L] + as(0, m, a, S)$. In order to construct optimal peptide sequence p for m , the amino acid residue a is chosen where a can offer the maximum similarity score $DP[0, m - \|a\|_L] + as(0, m, a, S)$. \square

Lemma 3.2 *When $n > 0$,*

$$DP[n, m] = \max \begin{cases} \max_{a \in \Sigma} DP[n, m - \|a\|_L] + as(n, m, a, S) \\ \max_{a \in \Sigma_m} DP[n - 1, m - \|a\|_L] + as(n, m, a, S) \end{cases}$$

Proof $n > 0$ represents that in peptide p , there are at least 1 modified K or R in peptide sequence. When we choose an amino acid residue a to form $p'a$, both $a \in \Sigma$ and $a \in \Sigma_m$ should be considered. Note that p' must contain all of n modifications when $a \in \Sigma$ so that $p'a$ contains n modifications. In this case, the similarity score of p' is $DP[n, m - \|a\|_L]$ and the similarity score of $p'a$ is $DP[n, m - \|a\|_L] + as(n, m, a, S)$.

When $a \in \sum_m$ which represents a is a modified K or R , p' contains $n-1$ modifications so that $p'a$ contains n modifications. In this case, the similarity score of p' is $DP[n-1, m - \|a\|_L]$ and the similarity score of $p'a$ is $DP[n-1, m - \|a\|_L] + as(n, m, a, S)$. In order to construct optimal peptide sequence p for m , the amino acid residue a is chosen where a can offer the maximum similarity score among all the $a \in \sum$ and $a \in \sum_m$. \square

Based on these lemma, we have algorithm 1.

Algorithm 1: Dynamic programming algorithm based on total number of modifications

Input: A MS/MS spectra set S , the total residue mass $M = m(S_{L_1})$, total number of SILAC modifications $N = (m(S_{M_1}) - m(S_{L_1})) / (\|K\|_M - \|K\|_L)$, an amino acid residue set \sum , a modified SILAC amino acid residue set \sum_m

Output: A peptide sequence p

```

1 Let  $DP[n, m] = 0$  for  $0 \leq n \leq N, 0 \leq m \leq M$ 
2 for  $m := 0$  to  $M$  do
3   |  $DP[0, m] = \max_{a \in \sum} DP[0, m - \|a\|_L] + as(0, m, a, S)$ 
4 end
5 for  $i := 1$  to  $n$  do
6   | for  $m := 0$  to  $M$  do
7     |  $DP[i, m] = \max \left\{ \begin{array}{l} \max_{a \in \sum} DP[n, m - \|a\|_L] + as(n, m, a, S) \\ \max_{a \in \sum_m} DP[n-1, m - \|a\|_L] + as(n, m, a, S) \end{array} \right.$ 
8   | end
9 end
10 Track back from  $DP[N, M]$ 
11 return The peptide obtained from backtracking

```

3.3.2 Algorithm based on modification pairs

If $\|K\|_M - \|K\|_L \neq \|R\|_M - \|R\|_L$ or $\|K\|_H - \|K\|_L \neq \|R\|_H - \|R\|_L$, then for any mass position, m , for a peptide with k modified Lysine and l modified Arginine, in a spectrum in S_L , the corresponding mass position in a spectrum in S_M , is

$m + k(\|K\|_M - \|K\|_L) + l(\|R\|_M - \|R\|_L)$, and in a spectrum in S_H , is $m + k(\|K\|_H - \|K\|_L) + l(\|R\|_H - \|R\|_L)$. From $m_L(p) = m(S_{L_i})$, $m_M(p) = m(S_{M_i})$, and $m_H(p) = m(S_{H_i})$, we can determine the number of modified Lysine and number of modified Arginine for the peptide we are looking for using the following formula. From

$$\|p\|_L + \|H_2O\| + 1 = m_L(p) = m(S_{L_i})$$

$$\|p\|_M + \|H_2O\| + 1 = m_M(p) = m(S_{M_i})$$

$$\|p\|_H + \|H_2O\| + 1 = m_H(p) = m(S_{H_i})$$

we have

$$k * (\|K\|_M - \|K\|_L) + l * (\|R\|_M - \|R\|_L) = m(S_{M_i}) - m(S_{L_i}) \quad (3.3)$$

$$k * (\|K\|_H - \|K\|_L) + l * (\|R\|_H - \|R\|_L) = m(S_{H_i}) - m(S_{L_i}) \quad (3.4)$$

Depend on the solutions of the equations, we can either determine a unique (k, l) pair or a set of pairs of (k, l) for the peptide we are looking for.

Let $\Delta K_M = \|K\|_M - \|K\|_L$, $\Delta K_H = \|K\|_H - \|K\|_L$, $\Delta R_M = \|R\|_M - \|R\|_L$, $\Delta R_H = \|R\|_H - \|R\|_L$. A similarity score of matching a at location m , as an ending amino acid of a peptide p with mass m and (k, l) modifications, $as(k, l, m, a, S)$ is

defined as

$$\begin{aligned}
 as(k, l, m, a, S) &= \sum_{i=1}^b f(S_{L_i}[m - \|a\|_L], S_{L_i}[m]) \\
 &+ \sum_{i=1}^c f(S_{M_i}[m - \|a\|_M + k * \Delta K_M + l * \Delta R_M], S_{M_i}[m + k * \Delta K_M + l * \Delta R_M]) \\
 &+ \sum_{i=1}^d f(S_{H_i}[m - \|a\|_H + k * \Delta K_H + l * \Delta R_H], S_{H_i}[m + k * \Delta K_H + l * \Delta R_H])
 \end{aligned}$$

For any mass value m in S_{L_i} and a peptide p with matching mass value m and (k, l) modifications, the similarity score $ps(k, l, m, p, S)$ is defined as the summation of the similarity score for each individual amino acid in p . We can also define $ps(k, l, m, p, S)$ recursively. Assume the peptide sequence p forms $p'a$ consisting of a prefix sequence p' and an amino acid residue a . We have

$$ps(k, l, m, p'a, S) = \begin{cases} ps(k, l, m - \|a\|_L, p', S) + as(k, l, m, a, S) & \text{if } a \in \Sigma \\ ps(k - 1, l, m - \|K\|_L, p', S) + as(k, l, m, K, S) & \text{if } a = K \\ ps(k, l - 1, m - \|R\|_L, p', S) + as(k, l, m, R, S) & \text{if } a = R \end{cases}$$

The problem of finding an optimal solution is defined as finding a peptide with maximum similarity score from all the peptides which mass equals to $m(S_{L_i})$ and has k modified K and l modified R .

$$P(S) = \arg \max_p \left\{ \begin{array}{l} |m_L(p) = m(S_{L_i})|, \quad 1 \leq i \leq b \\ |m_M(p) = m(S_{M_i})|, \quad 1 \leq i \leq c \\ |m_H(p) = m(S_{H_i})|, \quad 1 \leq i \leq d \end{array} \right\} ps(k, l, m(S_{L_i}), p, S)$$

Denote $DP[k, l, m]$ as the maximum similarity score between a peptide with mass m and k modified K and l modified R . with the spectra set S .

$$DP[k, l, m] = \max_p \{ps(k, l, m, p, S)\}.$$

Lemma 3.3 *When $k = 0$ and $l = 0$,*

$$DP[0, 0, m] = \max_{a \in \Sigma} DP[0, 0, m - ||a||_L] + as(0, 0, m, a, S)$$

Proof When $k = 0$ and $l = 0$, it is the same with lemma 3.1. \square

Lemma 3.4 *When $k > 0$ and $l = 0$,*

$$DP[k, 0, m] = \max \begin{cases} \max_{a \in \Sigma} DP[k, 0, m - ||a||_L] + as(k, 0, m, K, S) \\ DP[k - 1, 0, m - ||K||_L] + as(k, 0, m, K, S) \end{cases}$$

Proof When $k > 0$ and $l = 0$, there are only one kind of modification in peptide.

Therefore, k in this case is equivalent with n in lemma 3.2. The proof is similar with lemma 3.2. \square

Lemma 3.5 *When $k = 0$ and $l > 0$,*

$$DP[0, l, m] = \max \begin{cases} \max_{a \in \Sigma} DP[0, l, m - ||a||_L] + as(0, l, m, R, S) \\ DP[0, l - 1, m - ||R||_L] + as(0, l, m, R, S) \end{cases}$$

Proof When $k = 0$ and $l > 0$, there are only one kind of modification in peptide. Therefore, l in this case is equivalent with n in lemma 3.2. The proof is similar with lemma 3.2. \square

Lemma 3.6 *When $k > 0$ and $l > 0$,*

$$DP[k, l, m] = \max \begin{cases} \max_{a \in \Sigma} DP[k, l, m - \|a\|_L] + as(k, l, m, a, S) \\ DP[k - 1, l, m - \|K\|_L] + as(k, l, m, K, S) \\ DP[k, l - 1, m - \|R\|_L] + as(k, l, m, R, S) \end{cases}$$

Proof When a is a modified K , p' must contains $k - 1$ modified K and l modified R . In this case, the similarity score of p' is $DP[k - 1, l, m - \|K\|_L]$ and the similarity score of $p'a$ is $DP[k - 1, l, m - \|K\|_L] + as(k, l, m, K, S)$. When a is a modified R , p' must contains k modified K and $l - 1$ modified R . In this case, the similarity score of p' is $DP[k, l - 1, m - \|R\|_L]$ and the similarity score of $p'a$ is $DP[k, l - 1, m - \|R\|_L] + as(k, l, m, R, S)$. The proof is similar with lemma 3.2. \square

Let $\Delta_M = \min\{\|K\|_M - \|K\|_L, \|R\|_M - \|R\|_L\}$, based on equation 3.3 and 3.4, a bound of maximum number of modifications should be the mass difference divided Δ_M . So we have:

$$k + l \leq (m(S_{M_i}) - m(S_{L_i})) / \Delta_M.$$

With these information, we can design an algorithm based on the combination on the number of modified Lysine and the number of modified Arginine. Based on these lemma, we have algorithm 2.

Algorithm 2: Dynamic programming algorithm based on modification pairs

Input: A MS/MS spectra set S , the total residue mass $M = m(S_{L_1})$, the number number of modifications $N = (m(S_{M_1}) - m(S_{L_1}))/\Delta_M$, an amino acid residue set Σ

Output: A peptide sequence p

```

1 Let  $DP[k, l, m] = 0$  for  $0 \leq k + l \leq N, 0 \leq m \leq M$ 
2 for  $m := 0$  to  $M$  do
3   |  $DP[0, 0, m] = \max_{a \in \Sigma} DP[0, m - \|a\|_L] + as(0, 0, m, a, S)$ 
4 end
5 for  $i := 1$  to  $N$  do
6   | for  $m := 0$  to  $M$  do
7     |  $DP[i, 0, m] = \max \left\{ \begin{array}{l} \max_{a \in \Sigma} DP[i, 0, m - \|a\|_L] + as(i, 0, m, K, S) \\ DP[i - 1, 0, m - \|K\|_L] + as(i, 0, m, K, S) \end{array} \right.$ 
8   | end
9 end
10 for  $j := 1$  to  $N$  do
11   | for  $m := 0$  to  $M$  do
12     |  $DP[0, j, m] = \max \left\{ \begin{array}{l} \max_{a \in \Sigma} DP[0, j, m - \|a\|_L] + as(0, j, m, R, S) \\ DP[0, j - 1, m - \|R\|_L] + as(0, j, m, R, S) \end{array} \right.$ 
13   | end
14 end
15 for  $n := 2$  to  $N$  do
16   | for  $i := 1$  to  $N - 1$  and  $j := 1$  to  $N - i$  do
17     | for  $m := 0$  to  $M$  do
18       |  $DP[i, j, m] =$ 
19         |  $\max \left\{ \begin{array}{l} \max_{a \in \Sigma} DP[i, j, m - \|a\|_L] + as(i, j, m, a, S) \\ DP[i - 1, j, m - \|K\|_L] + as(i, j, m, K, S) \\ DP[i, j - 1, m - \|R\|_L] + as(i, j, m, R, S) \end{array} \right.$ 
20     | end
21   | end
22 Track back from  $DP[k, l, M]$ 
23 return The peptide obtained from backtracking

```

3.4 Program design

We use real experiment data to test our program. Theoretically, precursor ions of spectra from same set should be identical. But in reality, they may have a small mass difference. Thus, in our program, we use average mass of precursor ions of spectra for

computation. Given a spectra set S , denote $\overline{m(S)}$ as the average mass of precursor ions of S , we have

$$m(S_L) = \overline{m(S'_{S' \in S_L})}$$

$$m(S_M) = \overline{m(S'_{S' \in S_M})}$$

$$m(S_H) = \overline{m(S'_{S' \in S_H})}$$

There are constraints when we select spectra sets. The mass difference between $m(S_L)$ and $m(S_M)$ and the mass difference between $m(S_L)$ and $m(S_H)$ should match the same (k, l) pair. Note that it is possible for some peptide sequence could not be fully reacted in the real experiment and hence it could have less SILAC modifications. Spectrum like this will not be selected in our spectra set.

Besides dynamic programming, some other algorithms are applied to improve the accuracy of algorithm. In the identification stage, all peaks from different spectra should be measured by the same standard and assigned a score to represent its value. Significant value from He and Ma [27] is introduced to measure peaks from different MS/MS spectra. We use min heap to output more possible candidate sequences during traceback stage. Finally, a new algorithm is designed to refine all the candidates and we developed a confidence score function to calculate the rate of a peak used for high score candidate sequences.

3.4.1 Dynamic programming

Algorithm based on total number of modifications can be considered as a special case of algorithm based on modification pairs. Therefore, in our approach, we consider about general case that we design a program based on the combination on the number of modified Lysine and the number of modified Arginine.

To simplify the dynamic programming, given a (k, l) pair, the following function

$$f(k, l) = l + (k + l)(k + l + 1)/2$$

converts a 2-dimension pair (k, l) into 1-dimension point $f(k, l)$. $f(k, l)$ gives a linear increasing order based on the number of modifications. Within the same number of modifications, $n = k + l$, the order for l is from 0 to n . Given an integer i , let $m_i = \lfloor \frac{-1 + \sqrt{1 + 8i}}{2} \rfloor$, the following inverse function computes the corresponding pair (k, l)

$$f^{-1}(i) = (m_i * (m_i + 3)/2 - i; i - m_i * (m_i + 1)/2)$$

Therefore, $DP[k, l, m]$ transformed to $DP[i, m]$. In this program, peptide with small number of SILAC modifications can be computed. Denote q as the total rows needed for computation in the dynamic programming. The value of q depends on the number of SILAC modifications. Therefore, q equals to 3, 6 and 10 corresponding to $k + l$ equals to 1, 2 and 3 in the peptide sequence respectively. Figure 3.1 shows the logic structure of dynamic programming rows. Two integer array pre_K and pre_R are computed by $f^{-1}(i)$ to record the corresponding i which represent $DP[k - 1, l, m]$ and $DP[k, l - 1, m]$, as table 3.1 shows. Note that N/A represent that for the combination

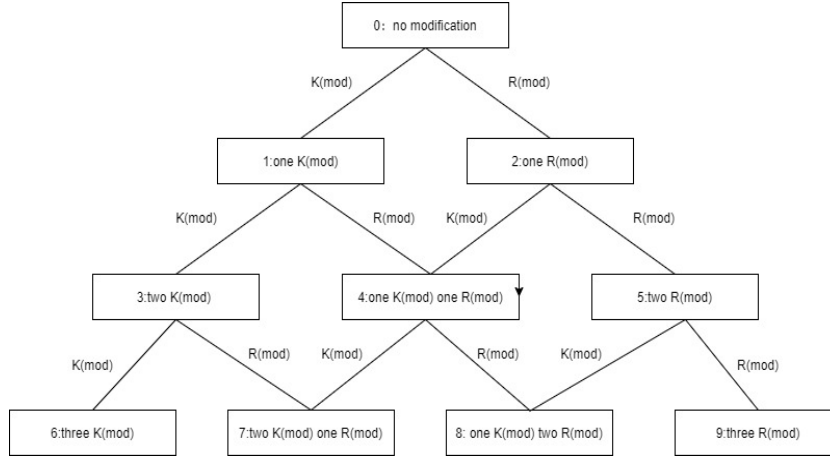


Figure 3.1: Logic structure of dynamic programming rows when $n = 3$
 The relationships between dynamic programming rows i for all combinations of (k, l) pair that $k + l \leq 3$

of (k, l) of $k = 0$ or $l = 0$, $DP[k - 1, l, m]$ or $DP[k, l - 1, m]$ is not applicable. Table 3.2 shows the detail combinations represents by index i when $n = 3$. Note that a missed cleavage is defined as any uncut lysine(K)/arginine(R) peptide bond. In reality, $n = 3$ represent that missed cleavage occurs twice at same peptide which is a quite rare situation.

i	0	1	2	3	4	5	6	7	8	9
$pre_K[i]$	N/A	0	N/A	1	2	N/A	3	4	5	N/A
$pre_R[i]$	N/A	N/A	0	N/A	1	2	N/A	3	4	5

Table 3.1: pre_K and pre_R values when $k + l \leq 3$

We also have the following design for the dynamic programming computation

- The row of dynamic programming matrix is mass value that is transformed to integer by rounding the mass after it multiplied by 100.
- $DP[0, 1902]$ set to 0 since 1902 is mass of Y-ion's C-terminus. This project

Index	SILAC modification combination
i=0	There is no SILAC labeling residue in the sequence.
i=1	There is only one SILAC labeling K in the sequence.
i=2	There is only one SILAC labeling R in the sequence.
i=3	There are two SILAC labeling Ks in the sequence.
i=4	There is one SILAC labeling K and one SILAC labeling R in the sequence.
i=5	There are two SILAC labeling Rs in the sequence.
i=6	There are three SILAC labeling Ks in the sequence.
i=7	There are two SILAC labeling Ks and one SILAC labeling R in the sequence.
i=8	There is one SILAC labeling K and two SILAC labeling Rs in the sequence.
i=9	There are three SILAC labeling Rs in the sequence.

Table 3.2: SILAC combinations corresponding to the index of dynamic programming when $k + l = 3$

is focused on CID and HCD MS/MS spectra which generate more y-ions and b-ions than other ions. Therefore, a reasonable m should equals to $||p|| + 19$ where p is a peptide sequence residue.

- The length of row equals to the precursor ion mass of spectra with light modification. Note that when p is consist of just one amino acid residue(p' is null), a reasonable m should equals to $||r|| + 19$ and $DP[i, m]$ equals to $as(k, l, m, r, S)$.
- To make sure that the peptide sequence of p is reasonable, $DP[i, m]$ will be set to -1 if m can not equals to any mass of peptide sequence. That means, $DP[i, m]$ will be set to -1 if all of its $DP[i, m - m(r)]$ equal to -1 during the computation.
- A integer array *Begin* stores the mass that dynamic programming computation begins for each row i . $DP[i, m]$ will not be calculated if the mass m is less than $Begin[i]$ since there is no reasonable sequence for i and $DP[i, m]$ is set to -1.

Figure 3.2 shows the dynamic programming matrix. As the figure shows, the dark

part of each row represent $DP[i, m]$ are set to -1. For example, when $i = 1$, there is one SILAC labeling K in the sequence. There is no reasonable sequence for i when m is less than $\|K\|_L + 19$.

Based on the design, we propose the Algorithm 3.

The $getScore()$ function of m in Algorithm 3 computes $as()$. We assume that a peak

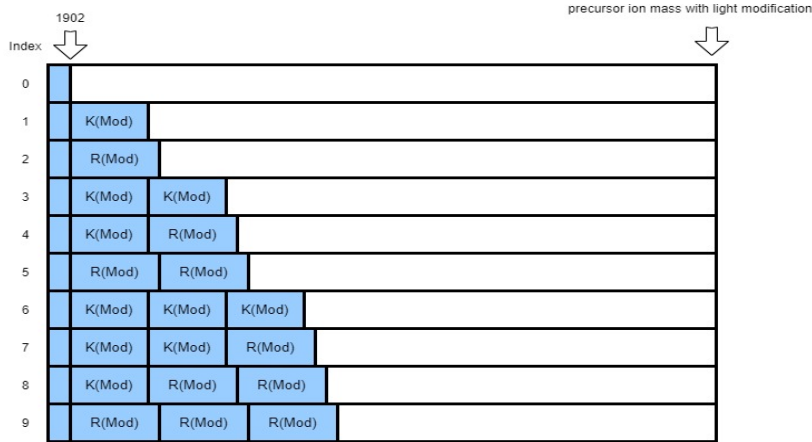


Figure 3.2: Dynamic programming matrix DP

of m is a y-ion peak. Normally, the peaks of y-ion can lose an ammonia or a water and its complement is b-ion and b-ion is losing of an ammonia or a water should be considered. Some other types of ion peak are optional. For example, spectra of multiple charged ($z \geq 3$) precursor ion may contain ion peaks with charge 2 after pre-processing'. The score of peaks are weighted based on their type.

The error tolerance accepted for the spectra. Our algorithm aims to search peaks in the interval between $m - \delta$ and $m + \delta$. The peak with highest score is chosen if there are more than one peak in the interval. Denote c as difference between theoretical m and the real peak mass. The larger c will have more score penalty which is adjusted by the formula

Algorithm 3: Dynamic programming computation in program

Input: A MS/MS spectra set $S = \{S_{L1}, S_{L2} \dots S_{M1}, S_{M2} \dots S_{H1}, S_{H2} \dots\}$ with total N MS/MS spectra with the total precursor ion mass $m(S_L)$, $m(S_M)$, $m(S_H)$, total dynamic programming rows needed q , an amino acid residue set Σ , a modified SILAC amino acid residue set Σ_m , an integer array $Begin$, integer arrays pre_K and pre_R , a positive score b for bonus

Output: candidate sequences

```

1 Let  $DP[0, 1902] = 0$ 
2 for  $i := 0$  to  $q$  do
3   for  $m := 0$  to  $Begin[i] - 1$  do
4     |  $DP[i, m] = -1$ 
5   end
6   for  $m := Begin[i]$  to  $m(S_L)$  do
7     | if  $\forall r \in \Sigma DP[i, m - m(r)] = -1$  and
8       |  $\forall r_m \in \Sigma_m DP[pre_r[i], m - m(r)] = -1$  then
9       |  $DP[i, m] = -1$ 
10      | else
11      |  $DP[i, m] = \max \begin{cases} \max_{r \in \Sigma} (DP[i, m - m(r)] + getScore(i, m, m(r), S)) \\ DP[pre_K[i], m - m(K)] + getScore(i, m, m(K), S) + b \\ \text{if } r = K, pre_K[i] \text{ is applicable} \\ DP[pre_R[i], m - m(R)] + getScore(i, m, m(R), S) + b \\ \text{if } r = R, pre_R[i] \text{ is applicable} \end{cases}$ 
12      | end
13    end
14  end
15 Traceback to get candidates
16 Evaluate the candidates

```

$$score = peakscore * e^{c*(-0.1)}$$

If two peaks formed a specific amino acid, it is more reasonable that the peptide sequence contains this amino acid. Thus, when calculating the $DP[i, m]$, if both mass m and $m - m(r)$ have real y-ion peak, we call it formed ion ladder and give it a bonus score. In order to locate the SILAC modifications, we give extra bonus score if m and $m - m(r)$ can form SILAC labeled K and R .

Algorithm 4: getScore()

Input: A MS/MS spectra set S , the index of dynamic programming rows i , current mass m , residue mass $m(r)$

Output: *score* of current mass

```

1 for  $j := 0$  to  $S.size()$  do
2    $\Delta_m$  is the mass shift based on  $S[j]$  and  $i$ 
3    $ScoreM_{S[j]}$  is the score for current spectrum which equals to summation
     of  $\left\{ \begin{array}{l} PeakScore_{S[j]}(m + \Delta_m) \\ weighted \ PeakScore_{S[j]}(m + \Delta_m - m(H_2O)) \\ weighted \ PeakScore_{S[j]}(m + \Delta_m - m(NH_3)) \\ weighted \ corresponding \ b-ion \ PeakScore_{S[j]}(m_b) \\ weighted \ PeakScore_{S[j]}(m_b - m(H_2O)) \\ weighted \ PeakScore_{S[j]}(m_b - m(NH_3)) \end{array} \right.$ 
4   Add  $ScoreM_{S[j]}$  to the total Score
5 end
6 if Both  $m$  and  $m - m(r)$  have y-ion peaks in  $S$  then
7   Add bonus score to the total Score
8 end

```

3.4.2 Significant value

In this research, Significant value from He and Ma [27] is introduced to measure peaks from different MS/MS spectra with a same standard. Four features of a peak may be associated with the significance of a peak: global rank, local rank, global intensity ratio and local intensity ratio. The global rank is the number of peaks in the spectra with intensity higher than or equal to the current peak. The global intensity ratio is the ratio between the average intensity of the top 3 peaks in the spectra and the intensity of the current peak. If this ratio is lower than 1 then it is set to 1. The local versions of the two features are defined in the same way except that only the peaks within ± 56 Da from the current peak are counted [27].

The peak significance value is defined as the linear combination of the algorithms of these four features.

$$\text{Significant value} = C_{gr} \log \text{global-rank} + C_{gir} \log \text{global-intensity-ratio} + C_{lr} \log \\ \text{local-rank} + C_{lir} \log \text{local-intensity-ratio}$$

The four coefficients $C_{gr}, C_{lr}, C_{gir}, C_{lir}$ are trained with real training data. The best normalized coefficients are $C_{gr} = 0.22, C_{lr} = 0.4, C_{gir} = 0.05, C_{lir} = 0.33$ [82].

From the definition, significant value is positive from 0 to 5. A smaller significant value indicates a stronger peak. The final score of each peak converts its significant value in an opposite way for dynamic programming that a larger score indicates a stronger peak.

3.4.3 Traceback

The traceback determines the actual peptide sequence candidates. We starts with the theoretical candidate sequence which has score $DP[i, m]$. m equals to the precursor ion mass of the spectra. The sequence having $DP[i, m']$ will also be calculated if $DP[i, m']$ is the maximum score in row i , $m \neq m'$ and $|m' - m| \leq \delta$. To obtain the target sequence, in traceback stage, we try to output more potential sequence for refinement. Given a precursor ion mass M , all peptide sequences with $||p|| + 19.02 = M$ are considered to be potential sequences and will be output. For each sequence, the traceback start from mass m and end at 19.02.

However, this procedure produce exponential number of peptide sequences. In our algorithm, a min heap is designed to filter the candidate sequences which have low scores during the computation. Each element in heap contains the following information: the score of candidate, current mass, current subsequence. Current subsequence

is a part of sequence p' ($\|p'\| = M - m'$) which has been completely tracebacked. The current mass m' is the mass of the rest part of sequence which has not been tracebacked. The score of candidate is the key for min heap that equals to score of p' plus $DP[i, m']$. The first candidate sequence has the score $DP[i, m]$, current mass m and current subsequence *null*.

From the dynamic programming, a peptide sequence p form $p'a$ consisting of a prefix sequence p' and a residue a . The first candidate do a one-step traceback and produce several new candidates with the score of candidate equals score of $p'a$, current mass $m - \|a\|$, current subsequence a . All these new candidates are insert into min heap and then candidates remained do one-step traceback again. During the computation, if a candidate sequence has *currentmass* = 1902, it represent that this sequence has complete traceback. Algorithm runs looping this one-step traceback and min heap filtering for candidates until all candidate sequences have *currentmass* = 1902 and their current subsequences now are the complete candidate sequence.

Min heap filtering and one-step traceback have advantages. Min heap can efficiently select high score candidates. One-step traceback ensure that all the candidates take same number of steps in order to avoid that specific high score candidate take too many steps of traceback and produce more candidate than others. Furthermore, we check all candidates for every two steps of traceback. Given a candidate sequence p_s , after two steps of traceback, the sequence is $p_s a_1 a_2$. If there is another candidate sequence that its sequence is $p_s a_2 a_1$, the candidate with lower score will be removed from candidate set and the candidates remained run next round of traceback. This procedure aim to reduce the calculation of traceback. For sequence $p_s a_1 a_2$ and $p_s a_2 a_1$,

the difference of their scores are from the combination of amino acid residue a_1 and a_2 since the score of p_s are the same. If a_1a_2 is better than a_2a_1 , the new candidates traceback from $p_s a_1 a_2$ will be also better than candidates traceback from $p_s a_2 a_1$. Therefore, it is no need to traceback $p_s a_2 a_1$.

When calculating the dynamic programming, a small error tolerance δ is accepted. Denote et as an integer that equals to $\delta * 100$. For the integer mass M , all the candidates from the interval between $M - et$ and $M + et$ are the potential sequences. Thus, the algorithm run the traceback $et * 2 + 1$ times for all acceptable mass in the interval. Assume the heap size is h . There are totally $h * (et * 2 + 1)$ candidate sequences selected.

3.4.4 Candidate refinement and confidence score

In this research, with dynamic programming, it is difficult to apply some complex scoring functions. Therefore, the score function of dynamic programming may not perfectly measure the matching between candidates and spectra. A typical example is that when dynamic programming calculating, it can easily compute one step of ion ladder that dynamic programming consider whether two peaks can form an amino acid residue. In this case, we can only give linear bonus base on the ladder. However, computing continuous ion ladder and giving it more score bonus need extra complex design to restore the length of the ladder.

After traceback, we have a set of candidate sequences. Given a candidate sequence, all of its theoretical ion peaks can be computed. Complex scoring functions can be

much easily applied for matching an exist peptide sequence to spectra. Therefore, new score function and algorithms are designed to refine all the candidates. Our refinement score function is able to consider more factors such as continuous ion ladder, duplicate using of a peak, enzyme effect and peaks with charge 2.

Candidates with highest refined score will be output as the identification results, we name them refined candidates. In this research, a new score is designed to evaluate these candidates. Considering a segment of refined candidate which is supported by peaks in spectra, we replace segment with random subsequence which has same mass with the segment and generate a new candidate. Compared with the refined candidate, if the refined score of the generated candidate dramatically decreases, it represents that the segment is much better matched to the spectra than a random subsequence. It's noted that some slightly change of the candidate sequence may affect the score. If the refined scores almostly do not changed, the random subsequence make the same score contribution with the segment which represent the segment is not 'confident' to the spectra.

For each refined candidate, we try to shuffle every part of its peptide sequence with random subsequences which have same mass to the shuffled part. In this way, certain numbers of the shuffled candidates are generated and then matched to the spectra with the same refinement score function.

A confidence scoring function is developed to evaluate all parts of candidate. The confidence score of the candidate is combined by the confident scores of each part of the candidate subsequence. We have following three steps to calculate confidence scores.

3.4.4.1 Subsequence array

A subsequence array is created which contains all possible subsequences of mass m . Note that this array can be created before dynamic programming identification if the types of SILAC modifications is known. Figure 3.3 shows the structure of the subsequence array.

The first integer array M is index by mass m and stores a pointer of second char array

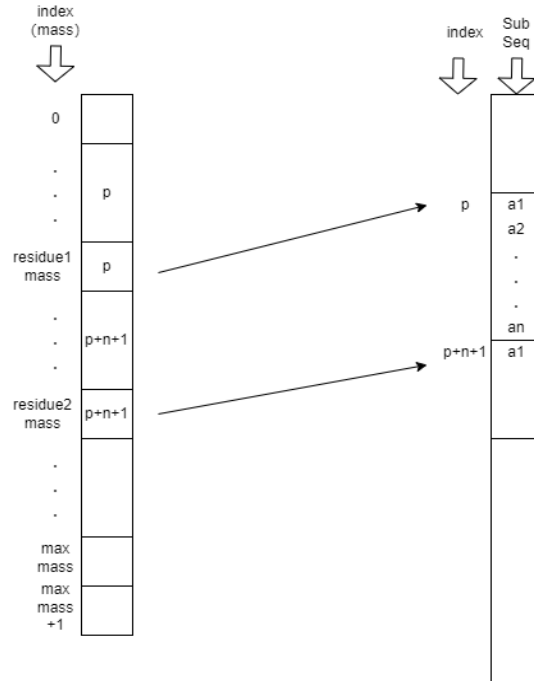


Figure 3.3: Subsequence array

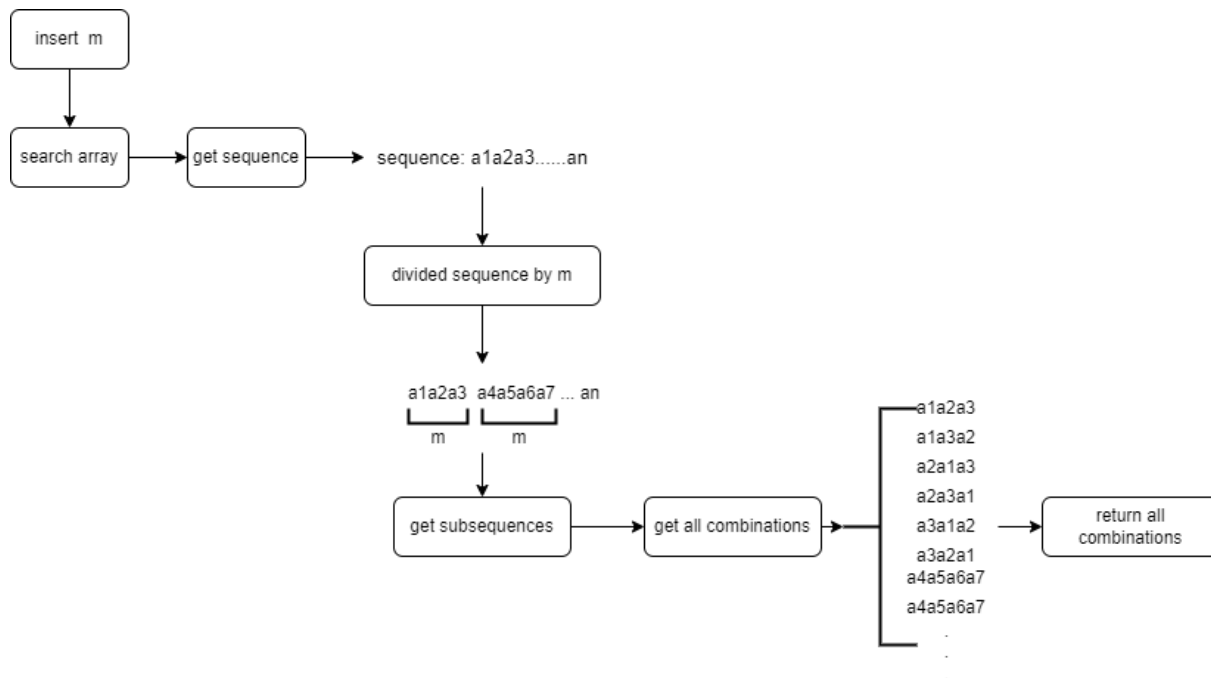
S . To balance the real time cost of array computing and the length of subsequence for shuffling, the maximum m is around 700 Da which is large enough to contain at least three amino acid residue combination. When calculating M and subsequence array S , all of mass of amino acid residue $a(a \in \Sigma$ and SILAC modified K and R) are put

into array in the first place. Then for each m , all of $M[m - ||a||]$ are checked. If all of $M[m - ||a||]$ are equals to their $M[m - ||a|| + 1]$ which represent that this m can not form any subsequence residue. If there is a $M[m - ||a||]$ such that $M[m - ||a|| + 1] - M[m - ||a||]$ is not equals to 0, m can form subsequence of $m - ||a||$ followed by a . We can enumerate all $||a||$ to have all subsequence of m .

Then, subsequences of m are arranged by alphabetical order. If there was permutations, all permutations of subsequences are the same now. We removed some subsequences and keep only one subsequence of same subsequences. Then, the remained subsequences are combined into one string and save into S .

3.4.4.2 Shuffling the candidate sequence

Given a m , we can have all of its corresponding subsequences from subsequence array, as figure 3.4 shows. Given an refined candidate sequence $p = a_1a_2...a_n$, we enumerate all subsequences of $p' = a_i a_{i+1} ... a_k$ that $||a_i a_{i+1} ... a_k|| \leq m$ and $||a_i a_{i+1} ... a_{k+1}|| > m$. These subsequences in refined candidate sequences are shuffled to the random subsequence which can get from subsequence array. New candidates with all possible subsequence combinations are generated and then their refined scores are computed. To filter the new candidate, a threshold is set that the new candidate is abandoned if its refined score is under the threshold (normally, the threshold is a percentage. We use 90% of the refined candidate's refined score e.g.) which are known as not confident candidate. The new candidate set is used to do the statistical analysis.. But the number of new candidate remained for each refined candidate may have difference. In order to make sure all refined candidate have same number of new candidates,


 Figure 3.4: Subsequences of mass m

certain number of refined candidate sequence are add to the set.

3.4.4.3 Confidence score

All of the theoretical peaks of all the candidates in the set are computed. Each peak gets a confidence score which equals to the ratio of the times of it appears to the total number of new candidates in the set. The confidence score of the original candidate equals the ratio of the sequence mass weighted by confidence score of each peaks to the sequence mass. Denote C_{seq} as the confidence score of the refined candidate, p as the theoretical peak of candidate, C_{p_i} as the confidence score of i th peak, m as the mass of the peaks and sequence. We have

$$C_{seq} = (\sum_1^i (m_{p_i} - m_{p_{i-1}}) * C_{p_i}) / m_{seq}$$

3.5 Complexity analysis

Denote M as the mass of precursor ion mass of the spectra, Σ as the set of the all amino acids residues, their PTMs and SILAC modifications(K and R), n as the number of SILAC modifications included in the candidate sequence and S as the set of spectra.

Theorem 3.1 Algorithm computes the candidate solution of SILAC peptide sequence in time bounded by

$$O(M \times n^2 \times |\Sigma| \times |S|)$$

Proof A dynamic programming matrix is designed to solve this problem. One dimension of the matrix is M , the mass of precursor ion mass of the spectra. The other one is the number of SILAC modifications included in the candidate sequence. For n SILAC modifications, there are n^2 combinations needed to be computed. So that we have a $M \times n^2$ table. For each entry of the table, algorithm check all $|S|$ spectra and find peaks of $|\Sigma|$ amino acid residue to compute the one with maximum score. Therefore, the time complexity will be $O(M \times n^2 \times |\Sigma| \times |S|)$.

In our experiment, M is transformed to integer by rounding the mass after it multiplied by 100 so that it is $100M$ in computing. n is equal to 3 so that there are at most 10 combinations needed to be computed. Σ contains amino acid residues, 2 SILAC modifications(K and R) and their PTMs. For each peptide sequence, there are 2 to up to 10 spectra with different SILAC modifications can be found. \square

Chapter 4

Experimental result and analysis

4.1 Group spectra results compare with spectrum results

4.1.1 Test design

The basic idea of our algorithm is that individual spectrum may not contain all information of peptide and a group of spectra of same peptide may contain information complementary with each other. Theoretically, identification from a group of spectra should be better than identification from individual spectrum. To verify our algorithm, we designed experiments to compare the results of individual spectrum de novo identifications and group multiple spectra de novo identifications. In these tests, our algorithm has two modes, group mode and spectra/spectrum mode. Group mode is that algorithm combine all information of multiple spectra to identify the peptide sequence. Spectra/spectrum mode identify peptide from each spectrum individually.

We run our group mode algorithm for group multiple spectra de novo identifications and run spectra/spectrum mode for each individual spectrum and compare the results.

The test is as follows:

- First, we search for a real public SILAC experiment dataset.
- Second, we establish a groundtruth dataset by using a database search tool to identify peptides with 1% FDR. A commercial database search software PEAKS DB is used to establish groundtruth dataset. The sequence reported by PEAKS DB are considered reliable.
- Third, we further filtered those spectra according to our requirements. We only select those groups of spectra given by PEAK DB in which all spectra in the group have identical groundtruth sequence but have different type of SILAC modifications. The selected groundtruth peptide sequences are considered to be the real peptide sequences of group spectra and individual spectrum since the peptide can be more reliable than 1% FDR. These considered real sequences are a standard of our de novo identification results comparison.
- Fourth, we use our two mods of de novo sequencing algorithm to identify group spectra and each of its individual spectrum.
- Last, all identification results are grouped for statistics, comparison and analysis.

The parameters of our algorithm are as follows

- Error tolerance: 0.02 Da
- Maximum SILAC modifications per candidate: 3
- The four coefficients for significant score calculation: 0.22, 0.4, 0,05, 0,33.
- Threshold of peaks: 90%(peaks with last 10% lowest significant scores will not be considered in the algorithm)
- Min heap size: 2000
- Number of candidates with the highest scores for output: 5

For comparison and analysis, we propose a simple scoring function to evaluate each candidate sequence. The full score is the length of the real peptide sequence which is the total number of amino acids contained. Compared with the real peptide sequence, the score of the candidate sequence is the number of the same amino acids contained. Note that we are not only considering the same amino acid but also considering the position of amino acid. Therefore, we transform real sequence and candidate sequences to a list of theoretical mass. We check whether the two lists of mass contains same mass and whether their previous masses are also same in order to make sure the identical masses represent the same amino acid residue. Figure 4.1 is an example of group spectra identification result and its spectra identification result.

For each de novo sequencing result, five candidate sequences are output and we choose the one with highest score as the result of identification for further comparison and analysis. Note that our scoring function is not optimized so that the sequence with

Groundtruth peptide sequence mass:1762.8712 Group ID:14

LM(+15.99)EGPAFNFLDAPAVR(+SILAC)

Group spectra de novo identification result:

YPEGPAFNFLDAPLGR(+SILAC) 12
LM(+15.99)EGPAFNFLDAPAVR(+SILAC) 16
MEEGPAFNFLDAPAVR(+SILAC) 14
EM(+15.99)LGPAFNFLDAPLGR(+SILAC) 11
LM(+15.99)WPAFNFLDAPLGR(+SILAC) 12

Match the groundtruth:true Best score:16

Single spectra de novo identification result - Scan:10301 Index:0 Type:L

PYEGPAFNFLDAPAVR(+SILAC) 14
MEEGPAFNFLDAPAVR(+SILAC) 14
LM(+15.99)QGPAFDFLDAPAVR(+SILAC) 10
SATEGPAFDFLDAPAVR(+SILAC) 8
MEQGPAFDFLDAPAVR(+SILAC) 8

Match the groundtruth:false Best score:14

Single spectra de novo identification result - Scan:10324 Index:1 Type:H

YPEGPAFNFLDALPGR(+SILAC) 11
PYEGPAFGGFLDALPGR(+SILAC) 10
AM(+15.99)GGGGPAFNFLDAPLGR(+SILAC) 11
SMGNGPAFNFLDAPLGR(+SILAC) 11
LM(+15.99)EGPAFNFLDLAPGR(+SILAC) 12

Match the groundtruth:false Best score:12

Figure 4.1: Example of group spectra identification results and its single spectrum identification results

highest score may not be the best sequence for the spectra and ranking of the candidate sequences will change when the scoring function is optimized. Choosing the best candidate sequence as the identification result from five candidate sequences with highest scores should get better results than choosing the candidate sequence with highest score as the identification result. That is because the result select from top five sequence should be no worse than the top one sequence. We also shows the comparison at last of each experiment.

4.1.2 Experiment 1

The public dataset is from ProteomeXchange Consortium and project ID is PXD017465.

The project information are as follows:


Title: Protein exchange rates within mitochondrial protein complexes of differentiated myotubes using pulsed SILAC complexome profiling

Data Processing Protocol: Proteins were identified using mouse reference proteome database UniProtKB with 52538 entries, released in 2/2018. Acetylation (+42.01) at N-terminus and oxidation of methionine (+15.99) were selected as variable modifications and carbamidomethylation (+57.02) as a fixed modification on cysteines. The enzyme specificity was set to trypsin.

SILAC Labeling: For SILAC labeling, the medium was changed into differentiation medium containing $^{13}\text{C}_6,^{15}\text{N}_4$ -L-Arginine and $^{13}\text{C}_6,^{15}\text{N}_2$ -L-Lysine. (Light: normal K, R; Heavy: K(+8.01), R(+10.01))

We choose a part of experiment data for testing. The database search software

PEAKS DB is used to identify peptides with 1% FDR (Figure 4.2). The reference proteome database is mouse UniProtKB with 86947 entries.

False discovery rate (FDR) curve. X axis is the number of peptides being kept. Y axis is the corresponding FDR. 

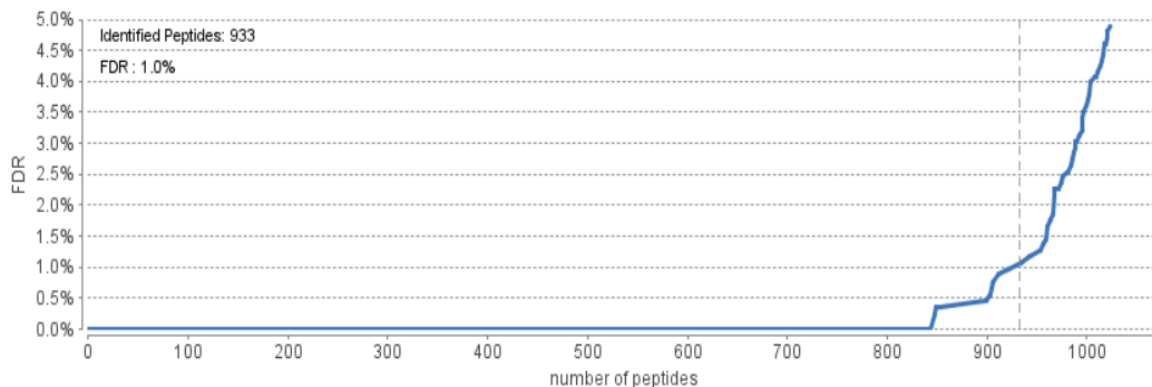


Figure 4.2: FDR curve for Experiment 1 (PEAKS DB)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
1	Peptide	-10lgP	Mass	Len	ppm	m/z	Z	RT	Area	Fraction	Id	Scan	from Chim	Source File	Accession	PTM	AScore	Found By	
2	AADFQLHTHVNDGTEFGGSIYQK	28.4	2534.183	23	1.8	845.7363	3	25.14	5.67E+04	1	3096	7846	No	P18_039_(Q3TX38:Q3TTN3:Q60931:Q5EBC				PEAKS DB	
3	AAELIANSLATAGDGLIELR	52.02	1997.079	20	1.6	999.5485	2	34.35	7.39E+04	1	6486	11086	No	P18_039_(P67778					PEAKS DB
4	AAGTFTEIPASNIR	35.27	1446.747	14	1.2	724.3815	2	25.36	5.01E+05	1	3202	7921	No	P18_039_(Q8BKZ9					PEAKS DB
5	AALAHSEIATTQAASKT	50.68	1669.864	17	1.5	557.6293	3	18.22	2.24E+05	1	400	5262	No	P18_039_(Q65388:Q65392:Q65393:Q65387				PEAKS DB	
6	AALAHSEIATTQAASKT	27.67	1669.864	17	3.1	835.9416	2	18.3	0	1	13033	5272	No	P18_039_(Q65388:Q65392:Q65393:Q65387				PEAKS DB	
7	AAPAAAAAM(+15.99)APPGPR	42.38	1334.676	15	1.3	445.9	3	18.32	0	1	13036	5277	No	P18_039_(Q8BMF4	Oxidation	M9:Oxidat		PEAKS DB	
8	AATFGLILDVSLTHTLTFGK	49.64	2118.136	20	0.4	707.0529	3	36.48	1.99E+05	1	7206	11771	No	P18_039_(P67778					PEAKS DB
9	AAVPSGASTGIYEALRL	52.86	1803.937	18	0.3	902.9758	2	31.12	1.21E+06	1	5303	9969	No	P18_039_(Q5FW97:P17182					PEAKS DB
10	AAYFGIYDTAK	37.66	1218.592	11	0.1	610.3033	2	27.07	5.58E+05	1	3809	8550	No	P18_039_(Q545A2:P51881					PEAKS DB
11	ADLEM(+15.99)QIENLKEELAYLK	40.34	2165.093	18	3.2	722.7071	3	35.56	0	1	13947	11483	No	P18_039_(Q9QWL7	Oxidation	M5:Oxidat		PEAKS DB	
12	ADLEM(+15.99)QIESLKEELAYLK	53.23	2138.082	18	1.1	713.7019	3	35.76	1.32E+06	1	6993	11538	No	P18_039_(Q61781	Oxidation	M5:Oxidat		PEAKS DB	
13	ADLEM(+15.99)QIESLKEELAYLKK	57.97	2266.177	19	0	567.5514	4	33.43	1.81E+06	1	6158	10761	No	P18_039_(Q61781	Oxidation	M5:Oxidat		PEAKS DB	
14	ADLEM(+15.99)QIESLKEELAYLKK	54.46	2266.177	19	1.4	756.4005	3	33.55	2.86E+05	1	6202	10814	No	P18_039_(Q61781	Oxidation	M5:Oxidat		PEAKS DB	
15	ADLEM(+15.99)QIESLKEELAYLR	22.3	2166.088	18	2.7	1084.054	2	35.83	4.33E+05	1	7010	11594	No	P18_039_(Q6IFX2	Oxidation	M5:Oxidat		PEAKS DB	
16	ADLEMQIES(+79.97)LK(+8.01)EELA	23.19	2338.158	19	4.8	780.397	3	36.17	1.35E+06	1	7121	11553	No	P18_039_(Q61781	Phosphory	S9:Phosph		PEAKS DB	
17	ADVDAATLAR	30.6	1001.514	10	-1	501.7637	2	20.06	0	1	13144	5882	No	P18_039_(P31001:Q3V2C6:Q3V1K9				PEAKS DB	

Figure 4.3: Groundtruth dataset of experiment 1 ordered by peptide sequence alphabetical order. (From PEAKS DB)

Figure 4.3 shows the 1% FDR dataset of experiment 1 ordered by peptide sequence alphabetical order (from PEAKS DB). As Table 4.1 shows, there are total 28 group and 67 spectra are selected from dataset.

Type	Groups	Spectra
Total	28	67

Table 4.1: Total number of group and spectra of Experiment 1

In these 28 groups of spectra, there are 61% (17 groups) groups can directly identify the real sequence while 39% (9 groups) groups can not (Table 4.2).

Groups	Total	Real sequence identified	Real sequence not identified
Number of group	28	17	9
Percentage	100%	61%	39%

Table 4.2: Total number of groups identified real sequence of Experiment 1

In these 67 spectra, there are 29.9% (20 spectra) spectra can directly identify the real sequence while 70.1% (47 spectra) spectra can not (Table 4.3).

Spectra	Total	Real Sequence identified	Real sequence not identified
Number of spectra	67	20	47
Percentage	100%	29.9%	70.1%

Table 4.3: Total number of spectra identified real sequence of Experiment 1

There are 3 out of 20 spectra can directly identify the real sequence but their groups result can not identify. In experiment 1, there is no group that all spectra results can identify the real sequence but group result can not, as Table 4.4 shows

For all groups spectra identified real sequence, as Table 4.5 shows

- There are 23.5%(4 out of 17 groups) of groups can directly identify the real sequence. Furthermore, all of their spectra can not identify the real sequence.

	Spectra real sequence identified total	Group not identified spectrum can identify	Group not identified all spectra identified group	Group not identified all spectra identified spectra
Number	20	3	0	0

Table 4.4: Spectra identified real sequence but group can not identify statistics of Experiment 1

- There are 53%(9 out of 17 groups) of groups can directly identify the real sequence but some of their spectra can directly identify the real sequence.

- There are 23.5%(4 out of 17 groups) of groups can identify the real sequence.

However, all of their spectra can directly identify the real sequence.

	Total	Real sequence identified no spectra identified	Real sequence identified group at least one spectra identified	Real sequence identified all spectra identified
Group	17	4	9	4
Percentage	100%	23.5%	53%	23.5%
Spectra	40	12	20	8
Percentage	100%	30%	50%	20%

Table 4.5: Group spectra identified real sequence statistics of Experiment 1

For all groups, we compare the group score with the average spectrum score. Group spectra combined all information of spectrum. Theoretically, group spectra should have better identification result than individual spectrum so that the comparison of group score and the average spectrum score can represent the improvement of identification result. For all 28 groups, as Table 4.6 shows,

- There are 71.4%(20 out of 28 groups) groups have larger score than their average spectrum score.

- There are 17.9%(5 out of 28 groups) groups have same score with their average spectrum score.
- There are 10.7%(3 out of 28 groups) groups score are less than their average spectrum score.

We consider the results that group score is larger than the average spectrum score are positive results. There are total 71.4%(20 out of 28 groups) groups are have positive identification results.

Group score VS average score	Total	Group score larger than average score	Group score equals to average score	Group score less than average score
Number of group	28	20	5	3
Percentage	100%	71.4%	17.9%	10.7%
Positive / Negative	100%	Positive	Neutral	Negative

Table 4.6: Score comparison statistics of Experiment 1

We also compared the recall and precision of group results and spectrum results. Recall is the ratio of the total number of correct amino acids from identification result to the total number of amino acids of real peptide sequences. Precision is the ratio of the total number of correct amino acids from identification results to the total number of amino acids of identification results. Note that group mode spectra identification can be considered as identifying peptide from all spectra in group together and all individual spectrum identification results are the same. Spectra/spectrum mode identify peptide from each spectrum individually. Recall and precision are two ratios to measure results identified from two modes. We compare them to show whether group spectra identification improved the identification result.

There are 953 amino acids total from real peptide sequence. There are 765 amino acids are correct from group spectra. Compared with real peptide sequence, recall is 80.3%. For spectrum results, there are 674 amino acids are correct. Compared with real peptide sequence, recall is 70.7% (Table 4.7).

	Real sequence amino acid total	Group correct amino acid (multiply spectra number)	Spectrum correct amino acid
Number of amino acid	953	765	674
Recall		80.3%	70.7%

Table 4.7: Recall of Experiment 1

There are 973 amino acids total from group identification results. There are 765 amino acids are correct from group spectra. Group result precision is 78.6%. For spectrum results, there are 674 amino acids are correct while there are 996 amino acids total. Precision is 67.7% (Table 4.8).

	Group amino acid total	Group correct amino acid	Spectrum amino acid total	Spectrum correct amino acid
Number of amino acid	973	765	996	674
precision		78.6%		67.7%

Table 4.8: Precision of Experiment 1

In summary, the results of Experiment 1 shows that our algorithm improved the de novo sequencing results. Group spectra identification results are better than spectrum identification results.

In this experiment, we choose the best candidate sequence as the identification result from top five candidate sequences with highest scores. Table 4.9 shows the brief statistics results if we only choose the candidate sequence with highest score as the identification result. Compared with previous results, choosing the best candidate sequence as the identification result from top five candidate sequences with highest scores is better than just choosing the candidate sequence with highest score as the identification result.

	Group real sequence identified	Group real sequence not identified	Spectra real sequence identified	Spectra real sequence not identified
Top 1	9	17	13	54
Top 5	17	9	20	47
	Group identified real sequence no spectra identified	Group identified real sequence at least one spectra identified	Group identified real sequence all spectra identified	
Top 1	3	4	2	
Top 5	4	9	4	
	Group result score larger than average spectra	Group result score equals to average spectra	Group result score less than average spectra	
Top 1	16	7	5	
Top 5	20	5	3	
	Group recall	Spectrum recall	Group precision	Spectrum precision
Top 1	72.2%	64.7%	69.4%	62%
Top 5	80.3%	70.7%	78.6%	67.7%

Table 4.9: Brief information summary for comparison of top one highest score candidate chosen and top five candidate chosen of Experiment 1

4.1.3 Experiment 2

The public dataset is from ProteomeXchange Consortium and project ID is PXD004539.

The project information are as follows:

Title: Proteome of Erythropoietin-induced mouse CFU-E cells

Data Processing Protocol: The search included variable modifications of methionine oxidation, phosphorylation (serine, threonine and tyrosine) and N-terminal acetylation, and fixed modification of carbamidomethyl cysteine. Minimal peptide length was set to six amino acids and a maximum of two missed-cleavages were allowed.


SILAC Labeling: The amino acid labeling information was set up in MaxQuant (Arg 10 and Lys6) to allow the comparison between light and heavy peptides. (Light: normal K, R; Heavy: K(+6.01), R(+10.01))

We choose a part of experiment data for testing. The database search software PEAKS DB is used to identify peptides with 1% FDR (Figure 4.4). The reference proteome database is mouse UniProtKB with 86947 entries.

Figure 4.5 shows the 1% FDR dataset of experiment 2 ordered by peptide sequence alphabetical order (From PEAKS DB). As Table 4.10 shows, there are total 4517 group and 15020 spectra are selected from dataset.

Type	Groups	Spectra
Total	4517	15020

Table 4.10: Total number of groups and spectra of Experiment 2

False discovery rate (FDR) curve. X axis is the number of peptides being kept. Y axis is the corresponding FDR. 

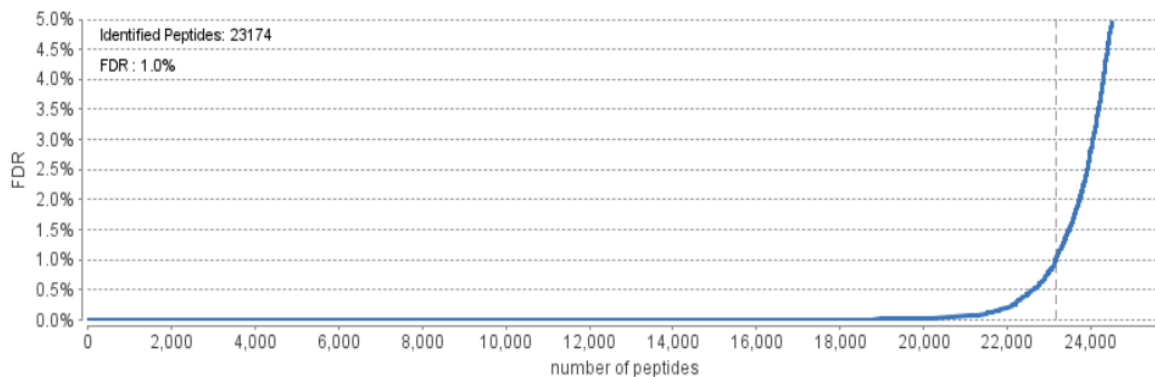


Figure 4.4: FDR curve for Experiment 2 (PEAKS DB)

Peptide	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
A(+42.01)AAAAAGAASGLPGPVAQGLK(+6.01)		-10lgP	Mass	Length	ppm	m/z	Z	RT	Area	Fraction	Id	Scan	from Chim	Source File	Accession	PTM	AScore	Found By
A(+42.01)AAAAAGAASGLPGPVAQGLK(+6.01)		54.77	1795.979	21	4.4	899.0005	2	54.36	6.32E+06	1	191357	23411	No	141211_0	E9QKZ2:Q	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVR		43.53	1795.979	21	5.9	599.6703	3	54.36	3.22E+06	1	191346	23351	No	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVR		58.01	2742.434	25	-2.4	915.1498	3	120.35	6.47E+06	1	220918	48668	No	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVR		51.43	2742.434	25	-2.6	1372.221	2	120.23	2.59E+04	1	220889	48616	No	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVR		22.41	2742.434	25	-2.4	915.1498	3	120.35	6.47E+06	1	220919	48813	No	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVR		14.66	2742.434	25	-2.4	915.1498	3	120.35	6.47E+06	1	220917	48516	No	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVR(+10.01)		53.12	2752.444	25	-3	918.4859	3	120.36	1.15E+07	1	220921	48667	No	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVR(+10.01)		48.37	2752.444	25	-2.9	1377.225	2	120.38	1.40E+05	1	220933	48594	No	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVR(+10.01)		33.98	2752.444	25	-3	918.4859	3	120.36	1.15E+07	1	220922	48811	No	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVR(+10.01)		28.54	2752.444	25	0.3	918.489	3	119.88	0	1	233716	48514	No	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVR(+10.01)K(+6.01)		48.85	2886.549	26	3.2	963.1934	3	103.15	2.76E+06	1	216365	43004	No	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVR(+10.01)K(+6.01)		35.74	2886.549	26	3.2	963.1934	3	103.15	2.76E+06	1	216364	42804	No	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVR(+10.01)K(+6.01)		31.73	2886.549	26	1.2	722.6454	4	103.09	2.82E+05	1	216315	42879	Yes	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVRK		44.31	2870.529	26	1.2	957.8514	3	103.15	1.36E+06	1	216363	42850	No	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAEQQFYLLGNLLSPDNVVRK		24.44	2870.529	26	2.2	718.6411	4	103.15	1.30E+05	1	216359	42905	Yes	141211_0	Q8BKCS	Acetylatio	A1:Acetyla	PEAKS DB
A(+42.01)AAAAATAVGPGAGSAGVAGPGGAGPC(+57.02)ATVSVI		44.27	3049.504	36	4.4	763.3866	4	60.8	5.95E+05	1	195481	26221	No	141211_0	Q9WVG6:1	Acetylatio	A1:Acetyla	PEAKS DB

Figure 4.5: Groundtruth dataset of experiment 2 ordered by peptide sequence alphabetical order. (From PEAKS DB)

In these 4517 groups of spectra, there are 30.2% (1363 groups) groups can directly identify the real peptide sequence while 69.8% (3154 groups) groups can not (Table 4.11). In these 15020 spectra, there are 16.4% (2467 spectra) spectra can directly identify the peptide sequence while 83.6% (12253 spectra) spectra can not (Table 4.12).

Groups	Total	Real sequence identified	Real sequence not identified
Number of group	4517	1363	3154
Percentage	100%	30.2%	69.8%

Table 4.11: Total number of group identified real sequence of Experiment 2

Spectra	Total	Real sequence identified	Real sequence not identified
Number of spectra	15020	2467	12553
Percentage	100%	16.4%	83.6%

Table 4.12: Total number of spectra identified real sequence of Experiment 2

There are 491 out of 2467 spectra can directly identify the real peptide sequence but their groups can not identify. In experiment 2, there are 26 groups that all spectra result can identify the real sequence but group result can not identify which include 54 out of 491 spectra, as Table 4.13 shows

For all groups spectra identified Real sequence, as Table 4.14 shows

- There are 16.6%(226 out of 1363 groups) of groups can directly identify the real sequence. Furthermore, all of their spectra can not identify the real sequence.

	Spectra real sequence identified total	Group not identified spectrum can identify	Group not identified all spectra identified group	Group not identified all spectra identified spectra
Number	2467	491	26	54

Table 4.13: Spectra identified real sequence but group can not identify statistics of Experiment 2

- There are 51.3%(700 out of 1363 groups) of groups can directly identify the real sequence but some of their spectra can directly identify the real sequence.
- There are 32.1%(437 out of 1363 groups) of groups can identify the real sequence. However, all of their spectra can directly identify the real sequence.

	Total	Real sequence identified no spectra identified	Real sequence identified at least one spectra identified	Real sequence identified all spectra identified
Group	1363	226	700	437
Percentage	100%	16.6%	51.3%	32.1%
Spectra	4314	737	2630	947
Percentage	100%	17.1%	60.9%	22%

Table 4.14: Group spectra identified real sequence statistics of Experiment 2

For all groups, we compare the group score with the average spectrum score. For all 4517 groups, as Table 4.15 shows,

- There are 67.9%(3067 out of 4517 groups) groups have larger score than their average spectrum score.
- There are 20.3%(918 out of 4517 groups) groups have same score with their average spectrum score.

- There are 11.8%(532 out of 4517 groups) groups score are less than their average spectrum score.

We consider the results that group score is larger than the average spectrum score are positive results. There are total 67.9%(3067 out of 4517 groups) groups are have positive identification results.

Group score VS average score	Total	Group score larger than average score	Group score equals to average score	Group score less than average score
Number of group	4517	3067	918	532
Percentage	100%	67.9%	20.3%	11.8%
Positive / Negative	100%	Positive	Neutral	Negative

Table 4.15: Score comparison statistics of Experiment 2

We also compared the recall and precision of group result and spectrum results. There are 250636 amino acids total from real sequence. There are 171723 amino acids are correct from group identification. Compared with real sequence, recall is 68.5%. For spectrum results, there are 128961 amino acids are correct. Compared with real sequence, recall is 51.4% (Table 4.16).

	Real sequence amino acid total	Group correct amino acid	Spectrum correct amino acid
Number of amino acid	250636	171723	128961
Recall		68.5%	51.4%

Table 4.16: Recall of Experiment 2

There are 262032 amino acids total from group identification result. There are 171723

amino acids are correct from group spectra. Compared with group result, precision is 65.5%. For spectrum results, there are 128961 amino acids are correct while there are 265123 amino acids total. Precision is 48.6% (Table 4.17).

	Group amino acid total	Group correct amino acid	Spectrum amino acid total	Spectrum correct amino acid
Number of amino acid	171723	262032	265123	128961
precision		65.5%		48.6%

Table 4.17: Precision of Experiment 2

In summary, the results of Experiment 2 shows that our algorithm improved the de novo sequencing results. Group spectra identification results are better than individual spectrum identification results.

In this experiment, we choose the best candidate sequence as the identification result from top five candidate sequences with highest scores. Table 4.18 shows the brief statistics results if we only choose the candidate sequence with highest score as the identification result. Compared with previous results, choosing the best candidate sequence as the identification result from top five candidate sequences with highest scores is better than just choosing the candidate sequence with highest score as the identification result.

	Group real sequence identified	Group real sequence not identified	Spectra real sequence identified	Spectra real sequence not identified
Top 1	822	3695	1576	13444
Top 5	1363	3154	2467	12553
	Group identified real sequence no spectra identified	Group identified real sequence at least one spectra identified	Group identified real sequence all spectra identified	
Top 1	158	437	227	
Top 5	226	700	437	
	Group result score larger than average spectra	Group result score equals to average spectra	Group result score less than average spectra	
Top 1	3059	841	617	
Top 5	3067	918	532	
	Group recall	Spectrum recall	Group precision	Spectrum precision
Top 1	63.7%	46.9%	60.4%	44.1%
Top 5	68.5%	51.5%	65.5%	48.6%

Table 4.18: Brief information summary for comparison of top one highest score candidate chosen and top five candidate chosen of Experiment 2

4.2 De novo sequencing results comparison with PEAKS de novo sequencing result

4.2.1 Experiment data, software and algorithm settings

In this research, a real SILAC MS/MS set from ProteomeXchange Consortium is used to test the performance of the proposed algorithm and its project ID is PXD005149.

Its instrument parameters are as follows

- Ion source: ESI
- Fragmentation mode: high energy CID (y and b ions)

- MS Scan Mode: FT-ICR/Orbitrap
- MS/MS Scan Mode: FT-ICR/Orbitrap

The cells of this SILAC experiment were grown in light (Arg0 and Lys0), medium (Arg6 and Lys4) and heavy (Arg10 and Lys8) media. Therefore, the lysine and arginine of peptide grown in the 'light' group are regular ones. The modifications of SILAC in this MS/MS spectra set are K(+4.03), R(+6.01) for the spectra with 'medium' modifications and K(+8.01), R(+10.01) for the spectra with 'heavy' modifications. In this spectra set, the following PTMs are also included:

- Carbamidomethylation(C): 57.02
- Oxidation (M): 15.99
- Acetylation (N-term): 42.01
- Phosphorylation (S,T,Y): 79.97

The data is pre-processed by PEAKS which is a comprehensive, vendor-neutral, proteomics tool to provide systematic identification and quantification of peptides/proteins in a complex protein mixture using tandem mass spectrometry (LC-MS/MS) [83]. PEAKS is a software with well-trained algorithm for identification and quantification of SILAC. We only compare the results from PEAKS de novo identification with the results from our algorithms. The parameters of our algorithm are as follows

- Error tolerance: 0.02 Da
- Maximum SILAC modifications per candidate: 3

- The four coefficients for significant score calculation: 0.22, 0.4, 0.05, 0.33.
- Threshold of peaks: 90%(peaks with last 10% lowest significant scores will not be considered in the algorithm)
- Min heap size: 2000
- Number of candidates with the highest scores for output: 5

4.2.2 Experiments

To verify the performance of our algorithm, we use the PEAKS software to determine the candidate sequence and compare the output candidates between our approach and PEAKS de novo from real data. We design three experiments:

- Experiment 1 aims to test the performance of function of algorithm. Two low quality spectra were selected and check whether the candidate sequence can be constructed using our method. These two spectra have same type of SILAC modifications.
- Experiment 2 use a four spectra set to test whether candidate sequence can be constructed in both cases of four spectra and two low quality spectra. This spectra set contains two type of SILAC modifications(light (Arg6, Lys4) and heavy (Arg10, Lys8)).
- Experiment 3 use a three spectra set contains three type of SILAC modifications(light (Arg0 and Lys0), medium (Arg6 and Lys4) and heavy (Arg10 and Lys8)).

In these experiments, identification result of spectrum, spectrum alignment to the candidate peptide sequence and ion table of spectrum matching the candidate peptide sequence from PEAKS de novo are shown to compare with identification result of spectra and ion table of spectra matching the candidate peptide sequence from our algorithm.

For PEAKS de novo, we show the following experimental result

Identification result of spectrum (PEAKS de novo): The identification result shows the candidate sequences with their scores from PEAKS de novo. Top 5 candidate sequences with highest scores are listed. Amino acid in candidate sequences are color coded according to their local confidence scores. Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.

Spectrum alignment (PEAKS de novo): The spectrum alignment shows how the fragment ions generated from the top 1 peptide sequence align with the spectrum. N-terminal ions are shown in blue and C-terminal ions are shown in red.

Ion table(PEAKS de novo): Ion table shows the calculated mass of possible fragment ions of top 1 candidate sequence from PEAKS de novo. If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red. A fragment ion is found when there is a matching peak within the mass error tolerance.

For our algorithm, we show the following experimental result

Identification result of spectra: The identification result shows the top 5 candidate sequences with their mass and confident score of sequence from our program.

We also list the mass of y-ions and confident scores of each amino acid.

Ion table: Ion table shows the calculated mass of possible fragment ions, peptide mass and its score of top 1 candidate sequence.

4.2.2.1 Experiment 1

Two spectra were selected from the SILAC MS/MS set which were generated from same peptide. Note that the lower case character represents the amino acid with PTM. The most likely candidate sequence is

K(+SILAC)C(+57.02)Y(+79.97)EM(+15.99)ASHLR(+SILAC)

Figure 4.6, Figure 4.7, Figure 4.8 are identification results, matching graph and ion tables of spectra 1 identified by PEAKS de novo. It can be seen that spectra 1 has few peaks with large mass to support the y ion ladder so that the large y ion mass parts of candidate sequence are constructed by b ion ladder. In this case, compared with the candidate sequence, the candidates output by PEAKS de novo are strongly support parts of the sequence and provide some of high possibility combinations for the rest parts of sequence. Two SILAC modifications from candidate sequence are successfully identified. One of the SILAC modification, the R(+10.01), is accurately located.

Figure 4.9, Figure 4.10, Figure 4.11 are identification results, matching graph and ion tables of spectra 2 identified by PEAKS de novo. As these graph shown, spectra 2 has a few peaks with large mass but it can not form a continuous y ion ladder and there are no peaks to form a reasonable b ion ladder to support the sequence. Compared with the identification result of spectra 1, the candidates output by PEAKS de novo

are very similar to the candidate sequence. However, the candidate sequence can not be identified due to the incomplete ion ladder. Two SILAC modifications from candidate sequence are successfully identified. One of the SILAC modification, the K(+8.01), is accurately located and the other SILAC modification, the R(+10.01), is also located but it has relatively weaker support from spectra data.

Based on the result of PEAKS de novo, the candidate sequence can not be identified from these two spectra of Experiment 1 due to their low quality data. However, the candidate sequence could be identified by their effective data complemented with each other. In our approach, Figure 4.12, Figure 4.13 shows the output candidates which are the results of identification of spectra 1 and spectra 2 respectively. The result is not better than result of PEAKS de novo. Figure 4.14 is the identification result of spectra. The candidate sequence is included in the output candidate sequence which have the highest score. Figure 4.15 is the ion table of spectra 1 and 2 matching the candidate peptide sequence. Compared with ion tables from PEAKS de novo, a relatively long and reasonable ion ladder is constructed from spectra set which highly support the candidate sequence.

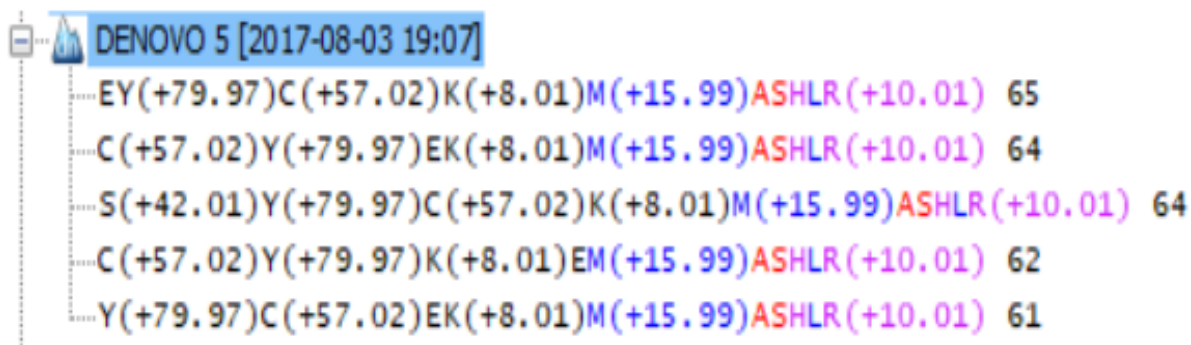


Figure 4.6: Experiment 1: Identification result of spectrum 1 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.

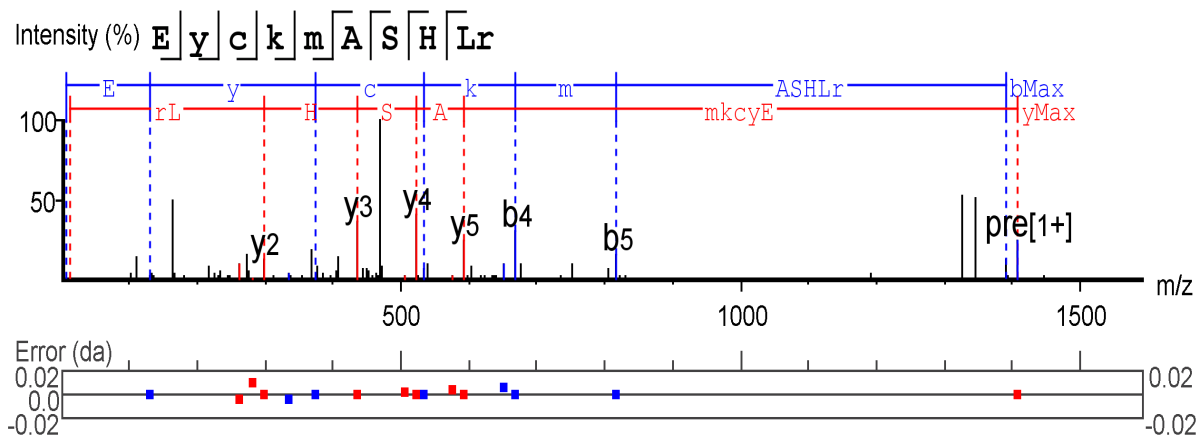


Figure 4.7: Experiment 1: Spectrum 1 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.

4.2.2.2 Experiment 2

In Experiment 2, 4 spectra are selected from the SILAC MS/MS set which are produced from same peptide. The most likely candidate sequence is

TR(+SILAC)TDVEDDDDEY(+79.97)K(+SILAC)

Figure 4.16, Figure 4.17, Figure 4.18, Figure 4.19 show the output candidates identified by PEAKS de novo from four spectra of Experiment 2. As graphs show, all of these four spectra have many identical peaks to form a reasonable y ion ladder which is also represent part of candidate sequence. For some of the spectra, the rest part of candidate sequence can be predicted and constructed by using mathematical algorithm.

Figure 4.20 shows the identification result of our approach from all the four spectra. The candidate sequence is included in the output candidates sequence which has the highest score. From the result of PEAKS de novo, spectra 2 and 3 have relative high quality that the candidate sequence can be identified by single spectrum and

#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	130.05	112.04	113.02	65.53	E					10
2	373.08	355.07	356.05	187.04	Y(+79.97)	1279.54	1261.53	1262.52	640.27	9
3	533.11	515.10	516.08	267.06	C(+57.02)	1036.51	1018.50	1019.49	518.76	8
4	669.22	651.20	652.19	335.11	K(+8.01)	876.48	858.47	859.46	438.74	7
5	816.26	798.24	799.23	408.63	M(+15.99)	740.37	722.36	723.35	370.69	6
6	887.29	869.28	870.27	444.15	A	593.34	575.32	576.31	297.17	5
7	974.32	956.31	957.30	487.66	S	522.30	504.29	505.28	261.66	4
8	1111.38	1093.37	1094.36	556.19	H	435.27	417.26	418.24	218.14	3
9	1224.47	1206.46	1207.44	612.73	L	298.21	280.19	281.18	149.61	2
10					R(+10.01)	185.13	167.12	168.10	93.06	1

Figure 4.8: Experiment 1: Ion table of spectrum 1 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.



Figure 4.9: Experiment 1: Identification result of spectrum 2 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.

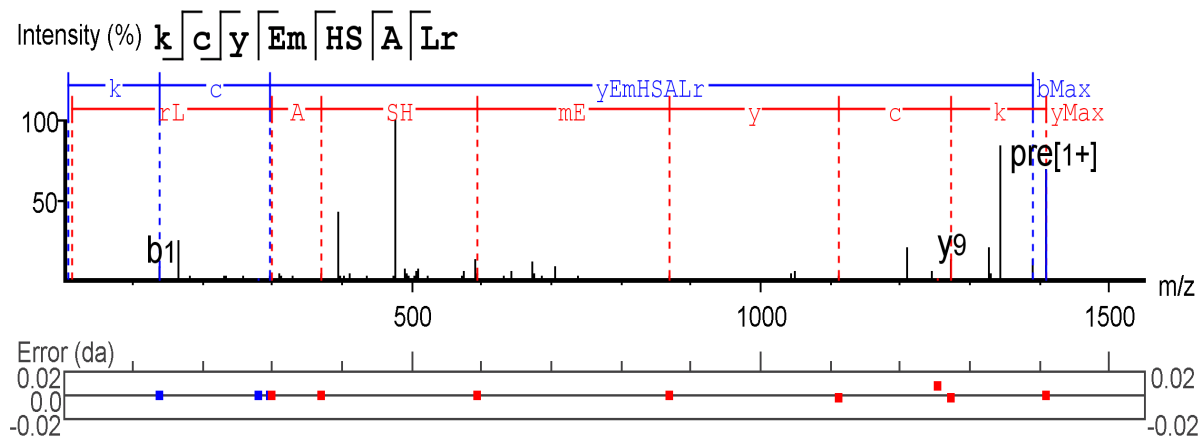


Figure 4.10: Experiment 1: Spectrum 2 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.

#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	137.12	119.11	120.09	69.06	K(+8.01)					10
2	297.15	279.14	280.12	149.07	C(+57.02)	1272.48	1254.46	1255.45	636.74	9
3	540.18	522.17	523.15	270.59	Y(+79.97)	1112.45	1094.44	1095.42	556.72	8
4	669.22	651.21	652.19	335.11	E	869.42	851.41	852.39	435.21	7
5	816.26	798.24	799.23	408.63	M(+15.99)	740.37	722.36	723.35	370.69	6
6	953.31	935.30	936.29	477.16	H	593.34	575.33	576.31	297.17	5
7	1040.35	1022.34	1023.32	520.67	S	456.28	438.27	439.25	228.64	4
8	1111.38	1093.37	1094.36	556.19	A	369.25	351.24	352.22	185.12	3
9	1224.47	1206.46	1207.44	612.73	L	298.21	280.20	281.18	149.61	2
10					R(+10.01)	185.13	167.12	168.10	93.06	1

Figure 4.11: Experiment 1: Ion table of spectrum 2 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.

Candidate Sequence:kyCGEmASHLr Candidate Mass:1409.58 Confident score:0.764											
k	y	C	G	E	m	A	S	H	L	r	
1408.57	1272.47	1029.44	926.43	869.41	740.37	593.34	522.30	435.27	298.21	185.13	
1.00	0.84	0.11	0.13	0.14	0.76	1.00	1.00	1.00	0.98	0.98	
Candidate Sequence:kSSyDmASHLr Candidate Mass:1409.6 Confident score:0.685											
k	S	S	y	D	m	A	S	H	L	r	
1408.59	1272.49	1185.46	1098.43	855.40	740.37	593.34	522.30	435.27	298.21	185.13	
1.00	0.84	0.11	0.10	0.20	0.76	1.00	1.00	1.00	0.98	0.98	
Candidate Sequence:kCtTmmASHLr Candidate Mass:1409.59 Confident score:0.706											
k	C	t	T	m	m	A	S	H	L	r	
1408.58	1272.48	1169.47	988.45	887.40	740.37	593.34	522.30	435.27	298.21	185.13	
1.00	0.84	0.20	0.05	0.21	0.76	1.00	1.00	1.00	0.98	0.98	
Candidate Sequence:kMsSmmASHLr Candidate Mass:1409.58 Confident score:0.697											
k	M	s	S	m	m	A	S	H	L	r	
1408.57	1272.47	1141.43	974.43	887.40	740.37	593.34	522.30	435.27	298.21	185.13	
1.00	0.84	0.01	0.05	0.21	0.76	1.00	1.00	1.00	0.98	0.98	
Candidate Sequence:ktSsPmASHLr Candidate Mass:1409.58 Confident score:0.726											
k	t	S	s	P	m	A	S	H	L	r	
1408.57	1272.47	1091.45	1004.42	837.42	740.37	593.34	522.30	435.27	298.21	185.13	
1.00	0.84	0.19	0.08	0.05	0.76	1.00	1.00	1.00	0.98	0.98	

Figure 4.12: Experiment 1: identification result of spectrum 1 only

Candidate Sequence:kcyEmSHrLA Candidate Mass:1409.58 Confident score:0.596											
k	c	y	E	m	S	H	r	L	A		
1408.57	1272.47	1112.44	869.41	740.37	593.34	506.31	369.25	203.14	90.06		
1.00	1.00	1.00	1.00	0.24	0.14	0.04	0.38	0.28	0.26		
Candidate Sequence:kcyPCGDGArLA Candidate Mass:1409.59 Confident score:0.630											
k	c	y	P	C	G	D	G	A	r	L	A
1408.58	1272.48	1112.45	869.42	772.37	669.36	612.34	497.31	440.29	369.25	203.14	90.06
1.00	1.00	1.00	1.00	0.40	0.20	0.40	0.23	0.29	0.38	0.28	0.26
Candidate Sequence:kcySGDSPGrLA Candidate Mass:1409.6 Confident score:0.589											
k	c	y	S	G	D	S	P	G	r	L	A
1408.59	1272.49	1112.46	869.43	782.40	725.38	610.35	523.32	426.27	369.25	203.14	90.06
1.00	1.00	1.00	1.00	0.19	0.20	0.20	0.20	0.25	0.38	0.28	0.26
Candidate Sequence:kcyPcAGDrAL Candidate Mass:1409.59 Confident score:0.602											
k	c	y	P	c	A	G	D	r	A	L	
1408.58	1272.48	1112.45	869.42	772.37	612.34	541.30	484.28	369.25	203.14	132.10	
1.00	1.00	1.00	1.00	0.40	0.40	0.20	0.20	0.38	0.28	0.01	
Candidate Sequence:kcyGGGEEArLA Candidate Mass:1409.6 Confident score:0.566											
k	c	y	G	G	G	E	E	A	r	L	A
1408.59	1272.49	1112.46	869.43	812.41	755.39	698.37	569.33	440.29	369.25	203.14	90.06
1.00	1.00	1.00	1.00	0.18	0.16	0.16	0.17	0.29	0.38	0.28	0.26

Figure 4.13: Experiment 1: identification result of spectrum 2 only

Candidate Sequence:kcyEmASHLr Candidate Mass:1409.58 Confident score:0.710											
k	c	y	E	m	A	S	H	L	r		
1408.57	1272.47	1112.44	869.41	740.37	593.34	522.30	435.27	298.21	185.13		
1.00	1.00	1.00	1.00	0.80	0.80	0.56	0.41	0.14	0.20		
Candidate Sequence:kcyEmASHrL Candidate Mass:1409.58 Confident score:0.718											
k	c	y	E	m	A	S	H	r	L		
1408.57	1272.47	1112.44	869.41	740.37	593.34	522.30	435.27	298.21	132.10		
1.00	1.00	1.00	1.00	0.80	0.80	0.56	0.41	0.14	0.31		
Candidate Sequence:kcyEmHSrAL Candidate Mass:1409.58 Confident score:0.697											
k	c	y	E	m	H	S	r	A	L		
1408.57	1272.47	1112.44	869.41	740.37	593.34	456.28	369.25	203.14	132.10		
1.00	1.00	1.00	1.00	0.80	0.80	0.16	0.14	0.11	0.31		
Candidate Sequence:kcyPDAGCGrLA Candidate Mass:1409.59 Confident score:0.536											
k	c	y	P	D	A	G	C	G	r	L	A
1408.58	1272.48	1112.45	869.42	772.37	657.34	586.30	529.28	426.27	369.25	203.14	90.06
1.00	1.00	1.00	1.00	0.20	0.20	0.20	0.20	0.19	0.14	0.11	0.03
Candidate Sequence:kCGyEmSHrAL Candidate Mass:1409.58 Confident score:0.631											
k	C	G	y	E	m	S	H	r	A	L	
1408.57	1272.47	1169.46	1112.44	869.41	740.37	593.34	506.31	369.25	203.14	132.10	
1.00	1.00	0.21	1.00	1.00	0.80	0.80	0.05	0.14	0.11	0.31	

Figure 4.14: Experiment 1: identification result of spectra 1 and 2

Sequence mass: 1409.58

#	b	b-H ₂ O	b-NH ₃	b/2	seq	y	y-H ₂ O	y-NH ₃	y/2	#
1	137.11	119.10	120.09	68.56	K(SILAC)					10
2	297.14	279.13	280.12	148.57	C(+57.02)	1272.47	1254.46	1254.45	636.23	9
3	540.17	522.16	523.15	270.08	Y(+79.97)	1112.44	1094.43	1095.42	556.22	8
4	669.21	651.20	652.19	334.61	E	869.41	851.40	852.39	434.70	7
5	816.24	798.23	799.22	408.12	M(+15.99)	740.37	722.36	723.35	370.18	6
6	887.28	869.27	870.26	443.64	A	593.34	575.33	576.32	296.67	5
7	974.31	956.30	957.29	487.15	S	522.30	504.29	505.28	261.15	4
8	1111.37	1093.36	1094.35	555.68	H	435.27	417.26	418.25	217.63	3
9	1224.45	1206.44	1207.43	612.22	L	298.21	280.20	281.19	149.10	2
10					R(SILAC)	185.13	167.12	168.11	92.57	1

Figure 4.15: Experiment 1: ion table of spectra 1 and 2 matching the candidate peptide sequence

candidate sequence can not be constructed from spectra 1 only. Figure 4.21 shows the identification result which only Spectrum 1 and 4 for comparison. In this case, the candidate sequence is also identified.



Figure 4.16: Experiment 2: Identification result of spectrum 1 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.

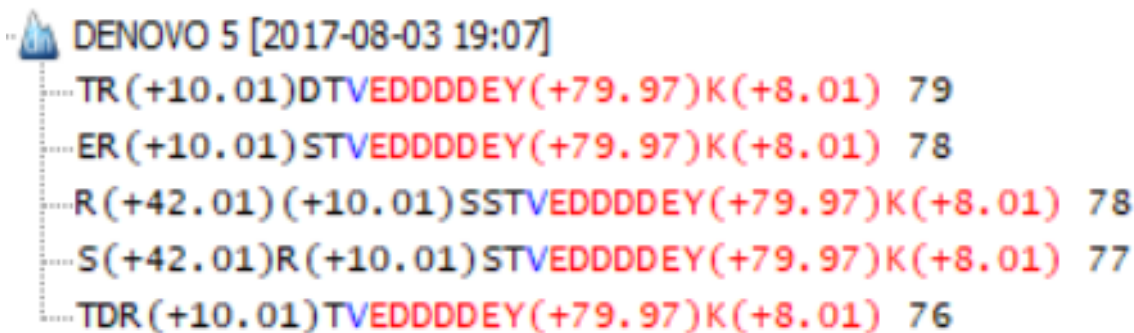


Figure 4.17: Experiment 2: Identification result of spectrum 2 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.

4.2.2.3 Experiment 3

In Experiment 3, 3 spectra are selected from the SILAC MS/MS set which are produced from same peptide. The most likely candidate candidate sequence is

HTDDEM(+15.99)TGY(+79.97)VATR(+SILAC)

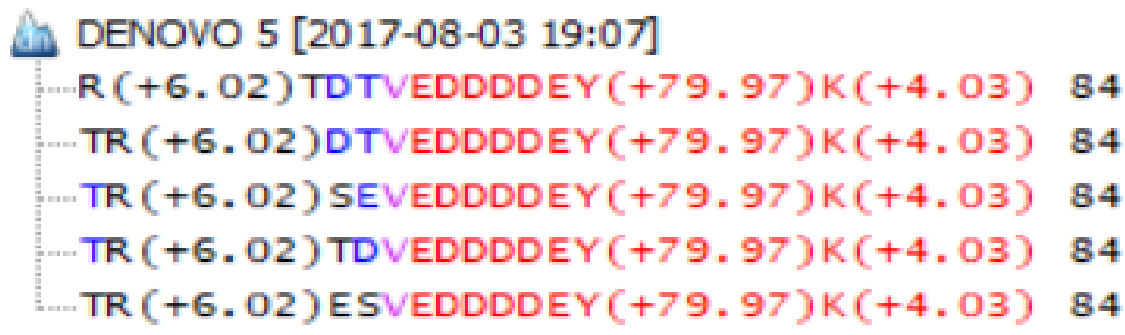


Figure 4.18: Experiment 2: Identification result of spectrum 3 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.

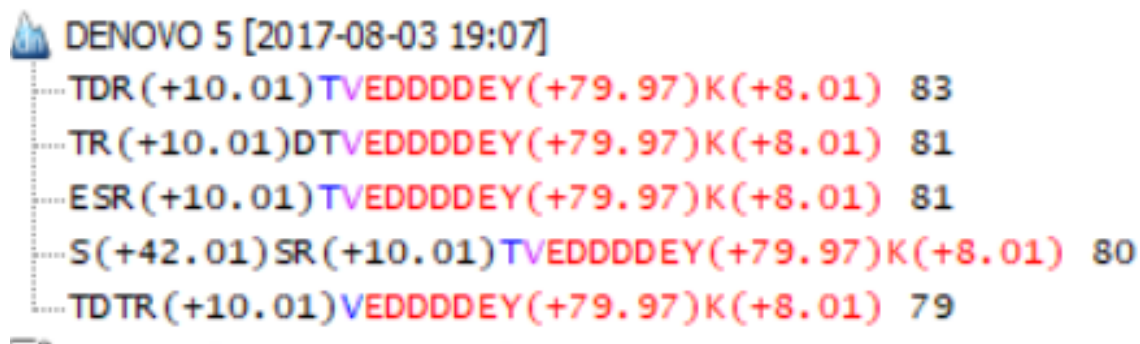


Figure 4.19: Experiment 2: Identification result of spectrum 4 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.

Candidate Sequence:ESrTVEDDDDEyk | Candidate Mass:1691.69 | Confident score:0.816
E S r T V E D D D D E y k
1690.68 1561.64 1474.61 1312.49 1211.44 1112.37 983.33 868.30 753.27 638.24 523.21 394.17 151.14
1.00 0.01 0.25 0.18 0.79 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00

Candidate Sequence:TrDTVEDDDDEyk | Candidate Mass:1691.7 | Confident score:0.818
T r D T V E D D D D E y k
1690.69 1589.64 1427.52 1312.49 1211.44 1112.37 983.33 868.30 753.27 638.24 523.21 394.17 151.14
1.00 0.31 0.20 0.18 0.79 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00

Candidate Sequence:TDrTVEDDDDEyk | Candidate Mass:1691.7 | Confident score:0.820
T D r T V E D D D D E y k
1690.69 1589.64 1474.61 1312.49 1211.44 1112.37 983.33 868.30 753.27 638.24 523.21 394.17 151.14
1.00 0.31 0.25 0.18 0.79 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00

Candidate Sequence:SErTVEDDDDEyk | Candidate Mass:1691.69 | Confident score:0.793
S E r T V E D D D D E y k
1690.68 1603.65 1474.61 1312.49 1211.44 1112.37 983.33 868.30 753.27 638.24 523.21 394.17 151.14
1.00 0.04 0.25 0.18 0.79 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00

Candidate Sequence:TrESVEDDDDEyk | Candidate Mass:1691.69 | Confident score:0.810
T r E S V E D D D D E y k
1690.68 1589.63 1427.51 1298.47 1211.44 1112.37 983.33 868.30 753.27 638.24 523.21 394.17 151.14
1.00 0.31 0.20 0.01 0.79 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00

Figure 4.20: Experiment 2: Identification result of all 4 spectra

Candidate Sequence:TrDTVEDDDDEyk | Candidate Mass:1691.7 | Confident score:0.833
T r D T V E D D D D E y k
1690.69 1589.64 1427.52 1312.49 1211.44 1112.37 983.33 868.30 753.27 638.24 523.21 394.17 151.14
1.00 0.67 0.38 0.59 0.46 0.52 1.00 1.00 1.00 1.00 1.00 1.00 1.00

Candidate Sequence:TDrTVEDDDDEyk | Candidate Mass:1691.7 | Confident score:0.831
T D r T V E D D D D E y k
1690.69 1589.64 1474.61 1312.49 1211.44 1112.37 983.33 868.30 753.27 638.24 523.21 394.17 151.14
1.00 0.67 0.44 0.59 0.46 0.52 1.00 1.00 1.00 1.00 1.00 1.00 1.00

Candidate Sequence:TrESVEDDDDEyk | Candidate Mass:1691.69 | Confident score:0.802
T r E S V E D D D D E y k
1690.68 1589.63 1427.51 1298.47 1211.44 1112.37 983.33 868.30 753.27 638.24 523.21 394.17 151.14
1.00 0.67 0.38 0.01 0.46 0.52 1.00 1.00 1.00 1.00 1.00 1.00 1.00

Candidate Sequence:TrTDVEDDDDEyk | Candidate Mass:1691.7 | Confident score:0.797
T r T D V E D D D D E y k
1690.69 1589.64 1427.52 1326.47 1211.44 1112.37 983.33 868.30 753.27 638.24 523.21 394.17 151.14
1.00 0.67 0.38 0.03 0.46 0.52 1.00 1.00 1.00 1.00 1.00 1.00 1.00

Candidate Sequence:TDrTVEDDDDEyk | Candidate Mass:1691.7 | Confident score:0.781
T D T r V E D D D D E y k
1690.69 1589.64 1474.61 1373.56 1211.44 1112.37 983.33 868.30 753.27 638.24 523.21 394.17 151.14
1.00 0.67 0.44 0.01 0.46 0.52 1.00 1.00 1.00 1.00 1.00 1.00 1.00

Figure 4.21: Experiment 2: Identification result of spectra 1 and 4

Figure 4.22, Figure 4.23, Figure 4.24 show the ion table of the best candidates identified by PEAKS de novo from single spectrum of Experiment 3. Figure 4.25 shows the identification result of our approach from all the four spectra. The candidate sequence is included in the output candidates sequence which has the highest score. From the result of PEAKS de novo, all these 3 spectra have relative low quality that the candidate sequence can not be identified by single spectrum. Figure 4.26 shows the ion table of the candidates.

#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	138.07	120.06	121.04	69.53	H					13
2	239.11	221.10	222.09	120.06	T	1454.55	1436.54	1437.52	727.78	12
3	354.14	336.13	337.11	177.57	D	1353.50	1335.49	1336.48	677.25	11
4	469.17	451.16	452.14	235.08	D	1238.48	1220.47	1221.45	619.74	10
5	598.21	580.20	581.18	299.61	E	1123.44	1105.44	1106.42	562.22	9
6	745.25	727.24	728.22	373.12	M(+15.99)	994.41	976.40	977.38	497.70	8
7	846.29	828.28	829.27	423.65	T	847.37	829.36	830.34	424.19	7
8	1089.32	1071.31	1072.30	545.16	Y(+79.97)	746.32	728.31	729.30	373.66	6
9	1146.35	1128.33	1129.32	573.67	G	503.29	485.28	486.27	252.15	5
10	1245.41	1227.40	1228.39	623.21	V	446.27	428.26	429.25	223.64	4
11	1316.45	1298.44	1299.42	658.73	A	347.20	329.19	330.18	174.10	3
12	1417.50	1399.49	1400.47	709.25	T	276.17	258.16	259.14	138.58	2
13					R	175.12	157.11	158.09	88.06	1

Figure 4.22: Experiment 3: Ion table of spectrum 1 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.

4.2.2.4 Experiment result summary

Based on the confident score, identified sequence of this two experiments are 0.71 and 0.818 respectively which are equal or higher than other candidates. The confident score of each peak in spectra divide into polarization. The peaks' confident score with highly support by spectra are usually larger than 0.75. On the other hand,

#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	138.07	120.06	121.04	69.53	H					13
2	239.11	221.10	222.09	120.06	T	1460.56	1442.56	1443.54	730.79	12
3	354.14	336.13	337.11	177.57	D	1359.52	1341.51	1342.50	680.26	11
4	469.17	451.16	452.14	235.08	D	1244.50	1226.49	1227.47	622.75	10
5	598.21	580.20	581.18	299.61	E	1129.47	1111.46	1112.44	565.23	9
6	745.25	727.24	728.22	373.12	M(+15.99)	1000.43	982.42	983.40	500.71	8
7	846.29	828.28	829.27	423.65	T	853.39	835.38	836.36	427.20	7
8	903.32	885.30	886.29	452.16	G	752.34	734.34	735.32	376.67	6
9	1146.35	1128.33	1129.32	573.67	Y(+79.97)	695.32	677.31	678.29	348.16	5
10	1245.41	1227.40	1228.39	623.21	V	452.29	434.28	435.27	226.65	4
11	1316.45	1298.44	1299.42	658.73	A	353.23	335.21	336.20	177.11	3
12	1417.50	1399.49	1400.47	709.25	T	282.19	264.18	265.16	141.59	2
13					R(+6.02)	181.14	163.13	164.11	91.07	1

Figure 4.23: Experiment 3: Ion table of spectrum 2 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.

#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	138.07	120.06	121.04	69.53	H					13
2	239.11	221.10	222.09	120.06	T	1464.56	1446.55	1447.53	732.78	12
3	354.14	336.13	337.11	177.57	D	1363.51	1345.50	1346.48	682.26	11
4	469.17	451.16	452.14	235.08	D	1248.48	1230.47	1231.46	624.74	10
5	598.21	580.20	581.20	299.61	E	1133.46	1115.45	1116.43	567.23	9
6	745.25	727.24	728.22	373.12	M(+15.99)	1004.41	986.40	987.39	502.71	8
7	846.29	828.28	829.27	423.65	T	857.38	839.37	840.35	429.19	7
8	903.32	885.30	886.29	452.16	G	756.34	738.32	739.30	378.67	6
9	1146.35	1128.33	1129.32	573.67	Y(+79.97)	699.31	681.30	682.28	350.16	5
10	1247.39	1229.38	1230.37	624.20	T	456.28	438.27	439.25	228.64	4
11	1346.46	1328.45	1329.43	673.73	V	355.25	337.22	338.21	178.12	3
12	1417.50	1399.49	1400.47	709.25	A	256.18	238.15	239.14	128.58	2
13					R(+10.01)	185.13	167.10	168.10	93.06	1

Figure 4.24: Experiment 3: Ion table of spectrum 3 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.

Candidate Sequence:HTDDEmTGyVATr | Candidate Mass:1592.63 | Confident score:0.823

H	T	D	D	E	m	T	G	y	V	A	T	r
1591.62	1454.56	1353.51	1238.48	1123.45	994.41	847.38	746.33	689.31	446.28	347.21	276.17	175.12
1.00	1.00	0.98	0.98	0.05	0.09	1.00	1.00	1.00	1.00	1.00	0.80	1.00

Candidate Sequence:HTDEdM TGyVATr | Candidate Mass:1592.63 | Confident score:0.829

H	T	D	E	D	m	T	G	y	V	A	T	r
1591.62	1454.56	1353.51	1238.48	1109.44	994.41	847.38	746.33	689.31	446.28	347.21	276.17	175.12
1.00	1.00	0.98	0.98	0.01	0.09	1.00	1.00	1.00	1.00	1.00	0.80	1.00

Candidate Sequence:HTDDEmTGyVTAr | Candidate Mass:1592.63 | Confident score:0.800

H	T	D	D	E	m	T	G	y	V	T	A	r
1591.62	1454.56	1353.51	1238.48	1123.45	994.41	847.38	746.33	689.31	446.28	347.21	246.16	175.12
1.00	1.00	0.98	0.98	0.05	0.09	1.00	1.00	1.00	1.00	1.00	0.20	1.00

Candidate Sequence:HTDPmmTGyVATr | Candidate Mass:1592.64 | Confident score:0.809

H	T	D	P	m	m	T	G	y	V	A	T	r
1591.63	1454.57	1353.52	1238.49	1141.44	994.41	847.38	746.33	689.31	446.28	347.21	276.17	175.12
1.00	1.00	0.98	0.98	0.01	0.09	1.00	1.00	1.00	1.00	1.00	0.80	1.00

Candidate Sequence:HTDmPmTGyVATr | Candidate Mass:1592.64 | Confident score:0.840

H	T	D	m	P	m	T	G	y	V	A	T	r
1591.63	1454.57	1353.52	1238.49	1091.46	994.41	847.38	746.33	689.31	446.28	347.21	276.17	175.12
1.00	1.00	0.98	0.98	0.01	0.09	1.00	1.00	1.00	1.00	1.00	0.80	1.00

Figure 4.25: Experiment 3: Identification result of all 3 spectra

Sequence mass: 1592.63

#	b	b-H ₂ O	b-NH ₃	b/2	seq	y	y-H ₂ O	y-NH ₃	y/2	#
1	138.07	120.06	121.05	69.04	H					13
2	239.12	221.11	222.10	119.56	T	1454.56	1436.55	1437.54	727.28	12
3	354.15	336.14	337.13	177.07	D	1353.51	1335.5	1336.49	676.76	11
4	469.18	451.17	452.16	234.59	D	1238.48	1220.47	1221.46	619.24	10
5	598.22	580.21	581.2	299.11	E	1123.45	1105.44	1106.43	561.72	9
6	745.25	727.24	728.23	372.63	M(+15.99)	994.41	976.4	977.39	497.2	8
7	846.3	828.29	829.28	423.15	T	847.38	829.37	830.36	423.69	7
8	903.32	885.31	886.3	451.66	G	746.33	728.32	729.31	373.17	6
9	1146.35	1128.34	1129.33	573.17	Y(+79.97)	689.31	671.3	672.29	344.65	5
10	1245.42	1227.41	1228.4	622.71	V	446.28	428.27	429.26	223.14	4
11	1316.46	1298.45	1299.44	658.23	A	347.21	329.2	330.19	173.6	3
12	1417.51	1399.5	1400.49	708.76	T	276.17	258.16	259.15	138.09	2
13					R(SILAC)	175.12	157.11	158.1	87.56	1

Figure 4.26: Experiment 3: ion table of spectra matching the candidate peptide sequence

peaks of less confident parts of sequence are usually below 0.40.

Chapter 5

Conclusion and future work

5.1 Conclusion

De novo peptide sequencing has evolved over several decades and numerous methods have been published. Progresses have been achieved in the applications of mass spectrometry in proteomics studies. However, the low identification rate of the acquired mass spectral data lowers the efficiency of applications of computational approaches. The application of multiple spectra has become popular and major attention has been paid to the use of spectra with different fragmentation methods. Multiple spectra sequencing methods have opened a new door for de novo peptide sequencing and provide a promising way to solve some of the current challenges facing traditional de novo peptide sequencing methods.

The application of stable isotope labeled proteins has been widely adopted to the comparative proteomics. In the area of quantitative study, SILAC (stable isotope labeling with amino acids in cell culture) can provide an effective scheme for comprehensive

and systematic qualitative and quantitative analysis of complex cell proteome. In a SILAC-based isotope labeling application, samples contain same protein molecule with different type of SILAC labeling. There are multiple spectra can be determined based on the factors such as the type of isotope labeling, retention time, precursor ion mass etc. In this approach, we try to combine these multiple SILAC spectra data to increase the identification rate and accuracy. However, theoretical peaks of a candidate sequence are different in these spectra due to different types of SILAC labeling and it becomes a challenge to locate the SILAC modification during the identification stage.

In this thesis, we conducted research regarding the identification of multiple SILAC MS/MS spectra. We propose two algorithms to compute the peptide sequence which are based on total number of SILAC modifications and based on SILAC modification pairs. Multiple possible candidate sequences are determined from traceback with min heap implemented. For pre-processing these algorithm, significant value [27] is used to all the peaks from different MS/MS spectra so that peaks can be measured by a same standard based on their intensity and rank. For post-processing the result, candidates are refined by algorithms that more factors such as continuous ion ladder, duplicate using of a peak and enzyme effect are considered. A confidence score is designed to measure the sequence which equals to the ratio of the sequence mass weighted by confidence score of each peaks to the sequence mass.

This de novo method is verified that it improved the identification results by using group spectra and also shows accurate identification result on our experiment, in which the correct peptide sequence can be identified for each spectra and the SILAC

modifications are accurately located.

5.2 Future work

Recent advances on the identification of multiple SILAC spectra has contributed greatly to proteomic studies, encouraging more researchers to integrate computational approach into biological sciences. Yet, it still demands further revision to better its application. For future work, we will concentrate on several research topics described in the following part.

First, we will continually improve our current work on the de novo sequencing of multiple MS/MS spectra of peptide containing SILAC labeling and its related applications. Additional theoretical efficient algorithms will be tested in order to reduce the time and space complexity.

- In this research, significant value from He and Ma [27] is used to measure peaks from different MS/MS spectra in the same standard. The advantage of significant value in our experiment is that it takes peak distribution into account by balancing the value with peak's rank and relative rank. However, complex design cost more time and space. In this case, the performance of simple and traditional method such as the direct usage of intensive logarithm needs to be tested.
- Additional scoring function from theoretical consideration(e.g. from machine learning) for identification and candidate refinement can be discussed in the

future work.

- Considering the error tolerance, all mass value of peaks are multiplied by 100 and then rounding to integers for the purpose of dynamic programming. Reducing the multiplier to 10 will significantly decrease the time and space cost, and the accuracy of identification will be affected. The balancing of accuracy and the computation time is worth to be tested.

Second, more work will focus on the complex PTM scenarios. In our approach, we assume that there is only one PTM or SILAC modifications on each amino acid. PTMs at same special position such as acetylation at N-term are not computed properly. Rare scenario such as acetylation at N-term with a SILAC modified *K* or *R* is ignored in our method. This reduce the accuracy of identification rate when target sequence is in these scenario, although the computation speed is the preference of our method design. Thus, correctly determining all the possible PTM scenarios will help to improve the identification result of the collected spectra.

Last but not least, multiple spectra for a same peptide with different types of isotope labeling modifications are the theoretical basis of the research. The prevailing advantage of the SILAC technique is that a large amount of multiple MS/MS spectra data can be obtained from a single SILAC experiment, which is the core factor leading us to focus on it. However, SILAC is not the only technique applied stable isotope labeling.

Stable isotope dimethylation is a costeffective, simple, robust, reliable and easy-to-multiplex labeling method which is widely applied to quantitative proteomics using

liquid chromatography mass spectrometry [84]. Multiple spectra for same peptide with different stable isotope dimethyl labeling can be achieved as well as some newly appearing ways such as more isotope labeling modifications of the special location(N-term, for example) instead of specific amino acid, these are all bringing challenges to the identification process. On the other hand, it's also the opportunity to apply our algorithm to these situations, which will be the future research focus.

Bibliography

- [1] Libretexts. <https://chem.libretexts.org/>.

- [2] Condensation reaction from wikipedia. https://en.wikipedia.org/wiki/Condensation_reaction.

- [3] Four level protein structure. <https://lubrizolcdmo.com/technical-briefs/protein-structure/>.

- [4] Drew Sturtevant, Young-Jin Lee, and Kent D Chapman. Matrix assisted laser desorption/ionization-mass spectrometry imaging (maldi-msi) for direct visualization of plant metabolites in situ. *Current Opinion in Biotechnology*, 37:53–60, 2016.

- [5] Shibdas Banerjee and Shyamalava Mazumdar. Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte. *International journal of analytical chemistry*, 2012, 2012.

- [6] Gary Siuzdak. *The expanding role of mass spectrometry in biotechnology*. Mcc Press, 2006.

- [7] J Benedikt, A Hecimovic, D Ellerweg, and A Von Keudell. Quadrupole mass spectrometry of reactive plasmas. *Journal of Physics D: Applied Physics*, 45(40):403001, 2012.
- [8] Lin He. Algorithms for characterizing peptides and glycopeptides with mass spectrometry. 2013.
- [9] Bin Ma. Challenges in computational analysis of mass spectrometry data for proteomics. *Journal of Computer Science and Technology*, 25(1):107–123, 2010.
- [10] Fragmentation. http://www.matrixscience.com/help/fragmentation_help.html.
- [11] Bryan M Ham. *Proteomics of biological systems: protein phosphorylation using mass spectrometry techniques*. John Wiley & Sons, 2011.
- [12] Yi Liu. *Algorithms for Peptide Identification from Mixture Tandem Mass Spectra*. PhD thesis, University of Western Ontario, 2015.
- [13] Bin Ma, Kaizhong Zhang, and Chengzhi Liang. An effective algorithm for peptide de novo sequencing from ms/ms spectra. *Journal of Computer and System Sciences*, 70(3):418–430, 2005.
- [14] Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. A framework of de novo peptide sequencing for multiple tandem mass spectra. *IEEE Transactions on NanoBioscience*, 14(4):478–484, 2015.

- [15] Yasuhiro Hara, Naoko Kawasaki, Ken-ichi Hirano, Yuuki Hashimoto, Jun Adachi, Shio Watanabe, and Takeshi Tomonaga. Quantitative proteomic analysis of cultured skin fibroblast cells derived from patients with triglyceride deposit cardiomyovasculopathy. *Orphanet journal of rare diseases*, 8(1):1–18, 2013.
- [16] Ayumu Saito, Masao Nagasaki, Masaaki Oyama, Hiroko Kozuka-Hata, Kentaro Semba, Sumio Sugano, Tadashi Yamamoto, and Satoru Miyano. Ayums: an algorithm for completely automatic quantitation based on lc-ms/ms proteome data and its application to the analysis of signal transduction. *BMC bioinformatics*, 8(1):1–10, 2007.
- [17] Xi Han, Lin He, Lei Xin, Baozhen Shan, and Bin Ma. Peaksptm: mass spectrometry-based identification of peptides with unspecified modifications. *Journal of proteome research*, 10(7):2930–2936, 2011.
- [18] Richard S Johnson and Klaus Biemann. Computer program (seqpep) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomedical & environmental mass spectrometry*, 18(11):945–957, 1989.
- [19] Leo McHugh and Jonathan W Arthur. Computational methods for protein identification from mass spectrometry data. *PLoS computational biology*, 4(2):e12, 2008.
- [20] Ruedi Aebersold and David R Goodlett. Mass spectrometry in proteomics. *Chemical reviews*, 101(2):269–296, 2001.

- [21] Michael Karas, Doris Bachmann, Ute Bahr, and Franz Hillenkamp. Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International journal of mass spectrometry and ion processes*, 78:53–68, 1987.
- [22] John B Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.
- [23] Vicki H Wysocki, Katheryn A Resing, Qingfen Zhang, and Guilong Cheng. Mass spectrometry of peptides and proteins. *Methods*, 35(3):211–222, 2005.
- [24] Fred W McLafferty, František Tureček, and Frantisek Turecek. *Interpretation of mass spectra*. University science books, 1993.
- [25] James J Pitt. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *The Clinical Biochemist Reviews*, 30(1):19, 2009.
- [26] Alan G Marshall, Christopher L Hendrickson, and George S Jackson. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass spectrometry reviews*, 17(1):1–35, 1998.
- [27] Lin He, Xi Han, and Bin Ma. De novo sequencing with limited number of post-translational modifications per peptide. *Journal of bioinformatics and computational biology*, 11(04):1350007, 2013.
- [28] Swapnil Bhatia, Yong J Kil, Beatrix Ueberheide, Brian Chait, Lemmuel L Tayo, Lourdes J Cruz, Bingwen Lu, John R Yates, and Marshall Bern. Constrained de

- novo sequencing of peptides with application to conotoxins. In *International Conference on Research in Computational Molecular Biology*, pages 16–30. Springer, 2011.
- [29] J Alex Taylor and Richard S Johnson. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Analytical chemistry*, 73(11):2594–2604, 2001.
- [30] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, 17(20):2337–2342, 2003.
- [31] Ari Frank and Pavel Pevzner. Pepnovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry*, 77(4):964–973, 2005.
- [32] Hao Chi, Rui-Xiang Sun, Bing Yang, Chun-Qing Song, Le-Heng Wang, Chao Liu, Yan Fu, Zuo-Fei Yuan, Hai-Peng Wang, Si-Min He, et al. pnovo: de novo peptide sequencing and identification using hcd spectra. *Journal of proteome research*, 9(5):2713–2724, 2010.
- [33] Marshall Bern and David Goldberg. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *Journal of Computational Biology*, 13(2):364–378, 2006.

- [34] Vlado Dančák, Theresa A Addona, Karl R Clauser, James E Vath, and Pavel A Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology*, 6(3-4):327–342, 1999.
- [35] Bernd Fischer, Volker Roth, Franz Roos, Jonas Grossmann, Sacha Baginsky, Peter Widmayer, Wilhelm Gruissem, and Joachim M Buhmann. Novohmm: a hidden markov model for de novo peptide sequencing. *Analytical chemistry*, 77(22):7265–7273, 2005.
- [36] Lijuan Mo, Debojyoti Dutta, Yunhu Wan, and Ting Chen. Msnovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Analytical chemistry*, 79(13):4870–4878, 2007.
- [37] Jonas Grossmann, Franz F Roos, Mark Cieliebak, Zsuzsanna Lipták, Lucas K Mathis, Matthias Müller, Wilhelm Gruissem, and Sacha Baginsky. Audens: a tool for automated peptide de novo sequencing. *Journal of proteome research*, 4(5):1768–1774, 2005.
- [38] Rodion Demine and Peter Walden. Sequit: software for de novo peptide sequencing by matrix-assisted laser desorption/ionization post-source decay mass spectrometry. *Rapid communications in mass spectrometry*, 18(8):907–913, 2004.
- [39] Robert M Day, Andrey Borziak, and Andrey Gorin. Ppm-chain-de novo peptide identification program comparable in performance to sequest. In *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.*, pages 505–508. IEEE, 2004.

- [40] Matthias Mann and Ole N Jensen. Proteomic analysis of post-translational modifications. *Nature biotechnology*, 21(3):255–261, 2003.
- [41] Unimod. <http://www.unimod.org>.
- [42] Shao-En Ong and Matthias Mann. A practical recipe for stable isotope labeling by amino acids in cell culture (silac). *Nature protocols*, 1(6):2650–2660, 2006.
- [43] Wolfgang Schütz, Niklas Hausmann, Karsten Krug, Rüdiger Hampp, and Boris Macek. Extending silac to proteomics of plant cell lines. *The Plant Cell*, 23(5):1701–1705, 2011.
- [44] Yan Yan. *De novo peptide sequencing methods for tandem mass spectra*. PhD thesis, University of Saskatchewan Saskatoon, 2015.
- [45] Roman A Zubarev. Reactions of polypeptide ions with electrons in the gas phase. *Mass spectrometry reviews*, 22(1):57–77, 2003.
- [46] Linus Pauling. The structure of singlet carbene molecules. *Journal of the Chemical Society, Chemical Communications*, (15):688–689, 1980.
- [47] Matthias Mann and Ole N Jensen. Proteomic analysis of post-translational modifications. *Nature biotechnology*, 21(3):255–261, 2003.
- [48] Mark W Duncan, Ruedi Aebersold, and Richard M Caprioli. The pros and cons of peptide-centric proteomics. *Nature biotechnology*, 28(7):659–664, 2010.

- [49] Jacques U Baenziger. A major step on the road to understanding a unique posttranslational modification and its role in a genetic disease. *Cell*, 113(4):421–422, 2003.
- [50] Eric S Witze, William M Old, Katheryn A Resing, and Natalie G Ahn. Mapping protein post-translational modifications with mass spectrometry. *Nature methods*, 4(10):798–806, 2007.
- [51] Michael Karas and Ralf Krüger. Ion formation in maldi: the cluster ionization mechanism. *Chemical reviews*, 103(2):427–440, 2003.
- [52] Boguslaw P Pozniak and Richard B Cole. Current measurements within the electrospray emitter. *Journal of the American Society for Mass Spectrometry*, 18(4):737–748, 2007.
- [53] PH Dawson. Quadrupole mass analyzers: Performance, design and some recent applications. *Mass Spectrometry Reviews*, 5(1):1–37, 1986.
- [54] Karen R Jonscher and John R Yates III. The quadrupole ion trap mass spectrometer—a small solution to a big challenge. *Analytical biochemistry*, 244(1):1–15, 1997.
- [55] James W Hager. A new linear ion trap mass spectrometer. *Rapid Communications in Mass Spectrometry*, 16(6):512–526, 2002.
- [56] BA Mamyrin. Time-of-flight mass spectrometry (concepts, achievements, and prospects). *International Journal of Mass Spectrometry*, 206(3):251–266, 2001.

- [57] Alan G Marshall, Christopher L Hendrickson, and George S Jackson. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass spectrometry reviews*, 17(1):1–35, 1998.
- [58] Qizhi Hu, Robert J Noll, Hongyan Li, Alexander Makarov, Mark Hardman, and R Graham Cooks. The orbitrap: a new mass spectrometer. *Journal of mass spectrometry*, 40(4):430–443, 2005.
- [59] Matthias Mann, Chin Kai Meng, and John B Fenn. Interpreting mass spectra of multiply charged ions. *Analytical Chemistry*, 61(15):1702–1708, 1989.
- [60] Hailing Zhang. *A New Algorithm for Charge State Deconvolution of Electrospray Ionization Mass Spectra*. PhD thesis, Faculty of Graduate Studies, University of Western Ontario, 2005.
- [61] Zhongqi Zhang and Alan G Marshall. A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *Journal of the American Society for Mass Spectrometry*, 9(3):225–233, 1998.
- [62] Zhongqi Zhang, Shenheng Guan, and Alan G Marshall. Enhancement of the effective resolution of mass spectra of high-mass biomolecules by maximum entropy-based deconvolution to eliminate the isotopic natural abundance distribution. *Journal of the American Society for Mass Spectrometry*, 8(6):659–670, 1997.
- [63] N Leigh Anderson and Norman G Anderson. Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*, 19(11):1853–1861, 1998.

- [64] W Hayes McDonald and John R Yates 3rd. Shotgun proteomics: integrating technologies to answer biological questions. *Current opinion in molecular therapeutics*, 5(3):302–309, 2003.
- [65] Michael P Washburn, Dirk Wolters, and John R Yates. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature biotechnology*, 19(3):242–247, 2001.
- [66] Dirk A Wolters, Michael P Washburn, and John R Yates. An automated multi-dimensional protein identification technology for shotgun proteomics. *Analytical chemistry*, 73(23):5683–5690, 2001.
- [67] Marjorie L Fournier, Joshua M Gilmore, Skylar A Martin-Brown, and Michael P Washburn. Multidimensional separations-based shotgun proteomics. *Chemical reviews*, 107(8):3654–3686, 2007.
- [68] R Graham Cooks. Special feature: Historical. collision-induced dissociation: Readings and commentary. *Journal of Mass Spectrometry*, 30(9):1215–1221, 1995.
- [69] Jesper V Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy c-trap dissociation for peptide modification analysis. *Nature methods*, 4(9):709–712, 2007.
- [70] John EP Syka, Joshua J Coon, Melanie J Schroeder, Jeffrey Shabanowitz, and Donald F Hunt. Peptide and protein sequence analysis by electron transfer

- dissociation mass spectrometry. *Proceedings of the National Academy of Sciences*, 101(26):9528–9533, 2004.
- [71] Jingfen Zhang, Simin He, Charles X Ling, Xingjun Cao, Rong Zeng, and Wen Gao. Peakselect: preprocessing tandem mass spectra for better peptide identification. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry*, 22(8):1203–1212, 2008.
- [72] Viswanadham Sridhara, Dina L Bai, An Chi, Jeffrey Shabanowitz, Donald F Hunt, Stephen H Bryant, and Lewis Y Geer. Increasing peptide identifications and decreasing search times for etd spectra by pre-processing and calculation of parent precursor charge. *Proteome Science*, 10(1):1–10, 2012.
- [73] Yufeng Shen, Nikola Tolic, Fang Xie, Rui Zhao, Samuel O Purvine, Athena A Schepmoes, J Moore, Ronald, Gordon A Anderson, and Richard D Smith. Effectiveness of cid, hcd, and etd with ft ms/ms for degradomic-peptidomic analysis: comparison of peptide identification methods. *Journal of proteome research*, 10(9):3929–3943, 2011.
- [74] Andreas Bertsch, Andreas Leinenbach, Anton Pervukhin, Markus Lubeck, Ralf Hartmer, Carsten Baessmann, Yasser Abbas Elnakady, Rolf Müller, Sebastian Böcker, Christian G Huber, et al. De novo peptide sequencing by tandem ms using complementary cid and electron transfer dissociation. *Electrophoresis*, 30(21):3736–3747, 2009.

- [75] Lin He and Bin Ma. Adept: advanced peptide de novo sequencing with a pair of tandem mass spectra. *Journal of bioinformatics and computational biology*, 8(06):981–994, 2010.
- [76] Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. Novopair: De novo peptide sequencing for tandem mass spectra pair. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 150–155. IEEE, 2014.
- [77] Mikhail M Savitski, Michael L Nielsen, Frank Kjeldsen, and Roman A Zubarev. Proteomics-grade de novo sequencing approach. *Journal of proteome research*, 4(6):2348–2354, 2005.
- [78] AF Maarten Altelaar, Danny Navarro, Jos Boekhorst, Bas van Breukelen, Berend Snel, Shabaz Mohammed, and Albert JR Heck. Database independent proteomics analysis of the ostrich and human proteome. *Proceedings of the National Academy of Sciences*, 109(2):407–412, 2012.
- [79] Kyowon Jeong, Sangtae Kim, and Pavel A Pevzner. Uninovo: a universal tool for de novo peptide sequencing. *Bioinformatics*, 29(16):1953–1962, 2013.
- [80] Gerald W Becker. Stable isotopic labeling of proteins for quantitative proteomic applications. *Briefings in Functional Genomics and Proteomics*, 7(5):371–382, 2008.
- [81] A Peter Snyder. *Interpreting protein mass spectra: a comprehensive resource*. American Chemical Society, 2000.

- [82] Xiaowen Liu, Baozhen Shan, Lei Xin, and Bin Ma. Better score function for peptide identification with etd ms/ms spectra. *BMC bioinformatics*, 11(1):1–8, 2010.
- [83] Bsi. <http://www.bioinfor.com>.
- [84] Jue-Liang Hsu and Shu-Hui Chen. Stable isotope dimethyl labelling for quantitative proteomics and beyond. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2079):20150364, 2016.

Appendices

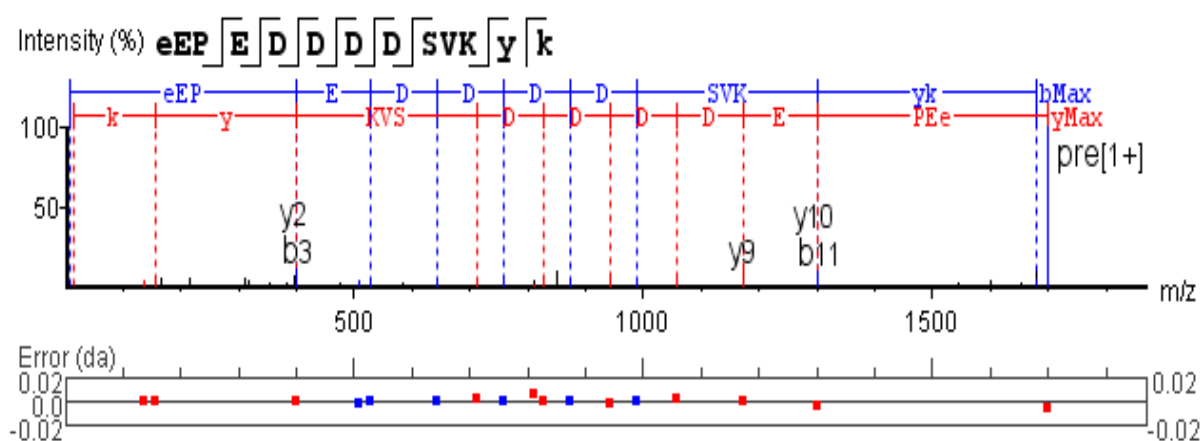


Figure .1: Experiment 2: Spectrum 1 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.

#	b	b-H ₂ O	b-NH ₃	b (2+)	Seq	y	y-H ₂ O	y-NH ₃	y (2+)	#
1	172.06	154.05	155.03	86.53	E(+42.01)					13
2	301.10	283.09	284.08	151.05	E	1527.60	1509.59	1510.57	764.30	12
3	398.16	380.15	381.13	199.58	P	1398.56	1380.54	1381.53	699.78	11
4	527.20	509.19	510.17	264.10	E	1301.51	1283.49	1284.48	651.25	10
5	642.23	624.22	625.20	321.61	D	1172.46	1154.45	1155.43	586.73	9
6	757.25	739.24	740.23	379.13	D	1057.43	1039.42	1040.41	529.22	8
7	872.28	854.27	855.25	436.64	D	942.41	924.40	925.38	471.70	7
8	987.31	969.30	970.28	494.15	D	827.38	809.37	810.35	414.19	6
9	1074.34	1056.33	1057.31	537.67	S	712.35	694.34	695.33	356.68	5
10	1173.41	1155.40	1156.38	587.20	V	625.32	607.31	608.29	313.16	4
11	1301.51	1283.49	1284.48	651.25	K	526.25	508.24	509.22	263.63	3
12	1544.53	1526.52	1527.50	772.77	Y(+79.97)	398.16	380.15	381.13	199.58	2
13					K(+8.01)	155.13	137.12	138.10	78.06	1

Figure .2: Experiment 2: Ion table of spectrum 1 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.

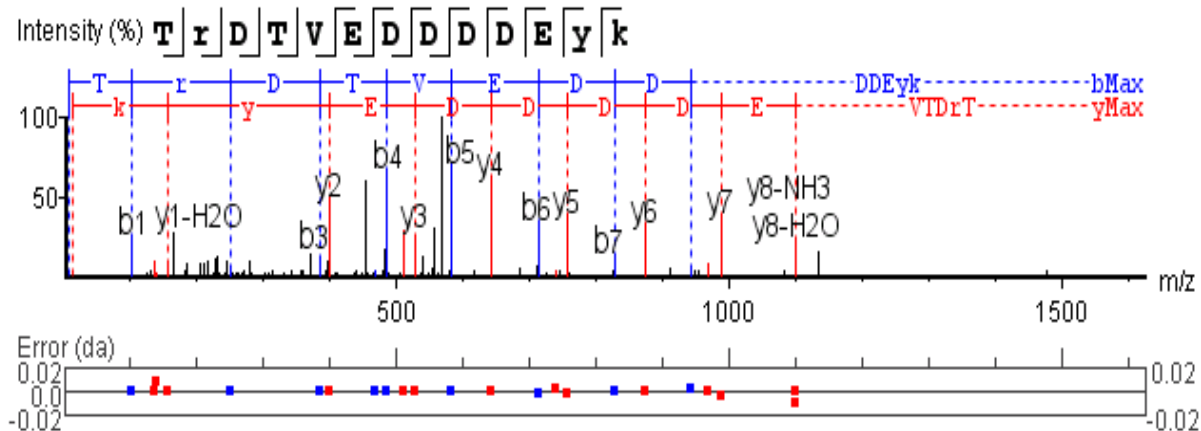


Figure .3: Experiment 2: Spectrum 2 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.

#	b	b-H ₂ O	b-NH ₃	b (2+)	Seq	y	y-H ₂ O	y-NH ₃	y (2+)	#
1	102.05	84.04	85.03	51.53	T					13
2	268.16	250.15	251.14	134.58	R(+10.01)	1597.60	1579.59	1580.58	799.30	12
3	383.19	365.18	366.16	192.10	D	1431.49	1413.48	1414.47	716.25	11
4	484.24	466.23	467.21	242.62	T	1316.47	1298.46	1299.44	658.73	10
5	583.31	565.30	566.28	292.15	V	1215.42	1197.41	1198.39	608.21	9
6	712.35	694.34	695.32	356.68	E	1116.35	1098.34	1099.33	558.67	8
7	827.38	809.37	810.35	414.19	D	987.31	969.29	970.28	494.15	7
8	942.40	924.39	925.38	471.70	D	872.28	854.27	855.25	436.64	6
9	1057.43	1039.42	1040.40	529.22	D	757.26	739.24	740.23	379.13	5
10	1172.46	1154.45	1155.43	586.73	D	642.23	624.22	625.20	321.61	4
11	1301.50	1283.49	1284.47	651.25	E	527.20	509.19	510.17	264.10	3
12	1544.53	1526.52	1527.50	772.77	Y(+79.97)	398.16	380.15	381.13	199.58	2
13					K(+8.01)	155.13	137.12	138.09	78.06	1

Figure .4: Experiment 2: Ion table of spectrum 2 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.

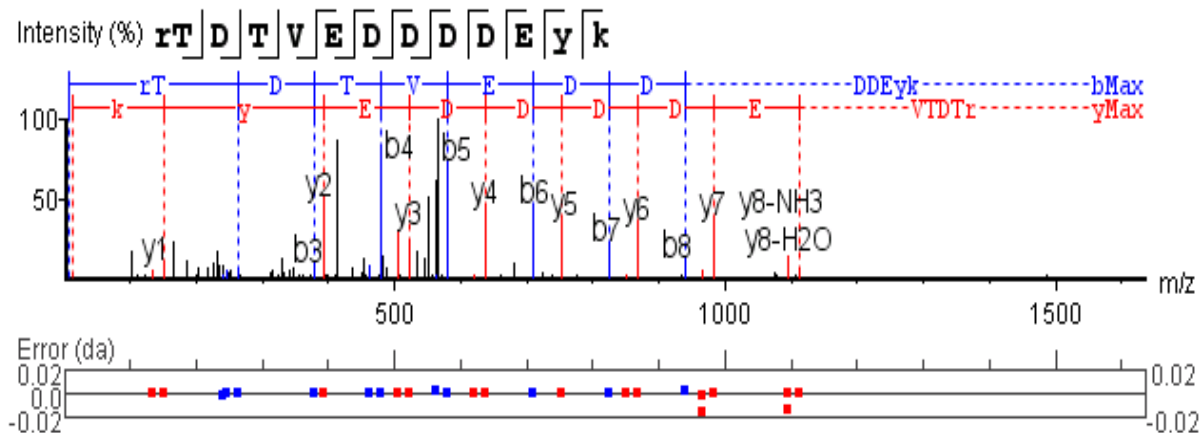


Figure .5: Experiment 2: Spectrum 3 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.

#	b	b-H2O	b-NH3	b (2+)	Seq	y	y-H2O	y-NH3	y (2+)	#
1	163.13	145.12	146.10	82.06	R(+6.02)					13
2	264.18	246.17	247.15	132.59	T	1528.55	1510.54	1511.52	764.78	12
3	379.20	361.19	362.18	190.10	D	1427.50	1409.49	1410.48	714.25	11
4	480.25	462.24	463.22	240.63	T	1312.48	1294.47	1295.45	656.74	10
5	579.32	561.31	562.29	290.16	V	1211.43	1193.42	1194.40	606.21	9
6	708.36	690.35	691.34	354.68	E	1112.36	1094.35	1095.35	556.68	8
7	823.39	805.38	806.36	412.19	D	983.32	965.31	966.31	492.16	7
8	938.41	920.41	921.39	469.71	D	868.29	850.28	851.26	434.65	6
9	1053.44	1035.43	1036.42	527.22	D	753.26	735.25	736.24	377.13	5
10	1168.47	1150.46	1151.44	584.74	D	638.24	620.23	621.21	319.62	4
11	1297.51	1279.50	1280.49	649.26	E	523.21	505.20	506.18	262.11	3
12	1540.54	1522.53	1523.52	770.77	Y(+79.97)	394.17	376.16	377.14	197.58	2
13					K(+4.03)	151.14	133.13	134.11	76.07	1

Figure .6: Experiment 2: Ion table of spectrum 3 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.

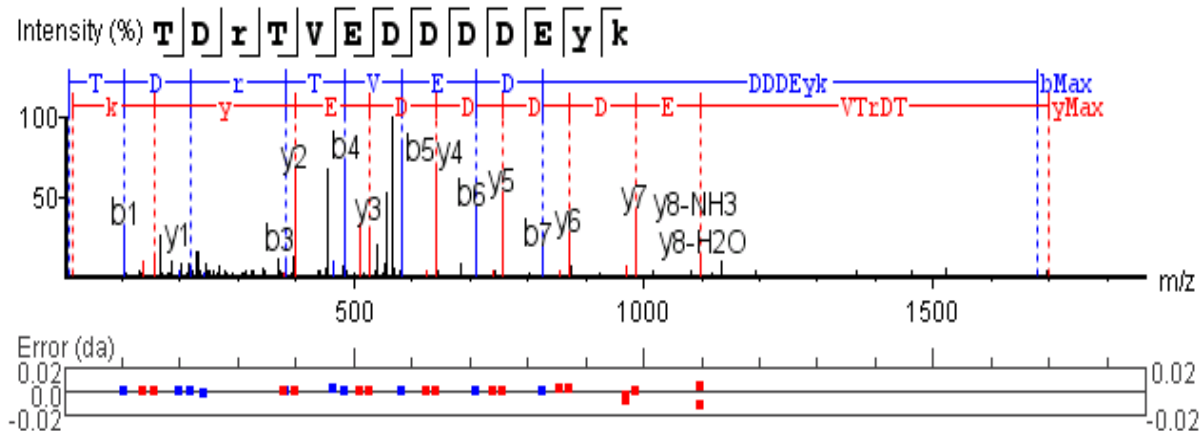


Figure .7: Experiment 2: Spectrum 4 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.

#	b	b-H ₂ O	b-NH ₃	b (2+)	Seq	y	y-H ₂ O	y-NH ₃	y (2+)	#
1	102.05	84.04	85.03	51.53	T					13
2	217.08	199.07	200.06	109.04	D	1597.60	1579.59	1580.58	799.30	12
3	383.19	365.18	366.16	192.10	R(+10.01)	1482.58	1464.56	1465.55	741.79	11
4	484.24	466.23	467.21	242.62	T	1316.47	1298.46	1299.44	658.73	10
5	583.31	565.30	566.28	292.15	V	1215.42	1197.41	1198.39	608.21	9
6	712.35	694.34	695.32	356.68	E	1116.35	1098.33	1099.34	558.67	8
7	827.38	809.37	810.35	414.19	D	987.31	969.30	970.29	494.15	7
8	942.40	924.39	925.38	471.70	D	872.28	854.27	855.25	436.64	6
9	1057.43	1039.42	1040.40	529.22	D	757.25	739.24	740.23	379.13	5
10	1172.46	1154.45	1155.43	586.73	D	642.23	624.22	625.20	321.61	4
11	1301.50	1283.49	1284.47	651.25	E	527.20	509.19	510.17	264.10	3
12	1544.53	1526.52	1527.50	772.77	Y(+79.97)	398.16	380.15	381.13	199.58	2
13					K(+8.01)	155.13	137.12	138.10	78.06	1

Figure .8: Experiment 2: Ion table of spectrum 4 matching the candidate peptide sequence(PEAKS de novo). If fragment ion is found in the spectrum, its mass value is displayed in color. N-terminal ions are shown in blue and C-terminal ions are shown in red.

Sequence mass: 1691.7

#	b	b-H ₂ O	b-NH ₃	b/2	seq	y	y-H ₂ O	y-NH ₃	y/2	#
1	102.06	84.05	85.04	51.03	T					13
2	264.18	246.17	247.16	132.09	R(SILAC)	1589.64	1571.63	1572.62	794.82	12
3	379.21	361.2	362.19	189.6	D	1427.52	1409.51	1410.5	713.76	11
4	480.26	462.25	463.24	240.13	T	1332.49	1294.48	1295.47	656.24	10
5	579.33	561.32	562.31	289.67	V	1211.44	1193.43	1194.42	605.72	9
6	708.37	690.36	691.35	354.18	E	1112.37	1094.36	1095.35	556.18	8
7	823.4	805.39	806.38	411.7	D	983.33	965.32	966.31	491.67	7
8	938.43	920.42	921.41	469.21	D	868.3	850.29	851.28	434.15	6
9	1053.46	1035.45	1036.44	526.73	D	753.27	735.26	736.25	376.64	5
10	1168.49	1150.48	1151.47	584.24	D	638.24	620.23	621.22	319.12	4
11	1297.53	1279.52	1280.51	648.77	E	523.21	505.2	506.19	261.61	3
12	1540.56	1522.55	1523.54	770.28	Y(+79.97)	394.17	376.16	377.15	197.09	2
13					K(SILAC)	151.14	133.13	134.12	75.57	1

Figure .9: Experiment 2: ion table of spectra combined matching the candidate peptide sequence

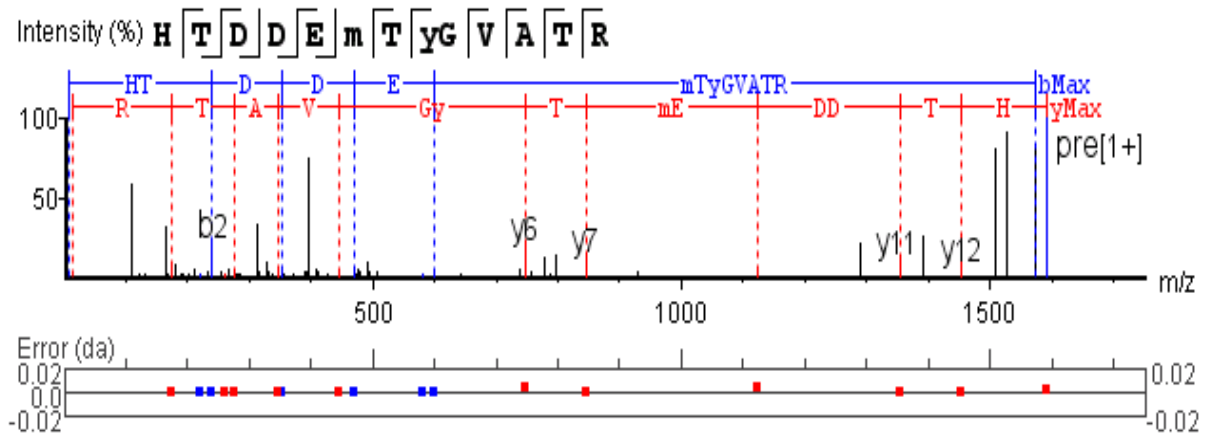


Figure .10: Experiment 3: Spectrum 1 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.

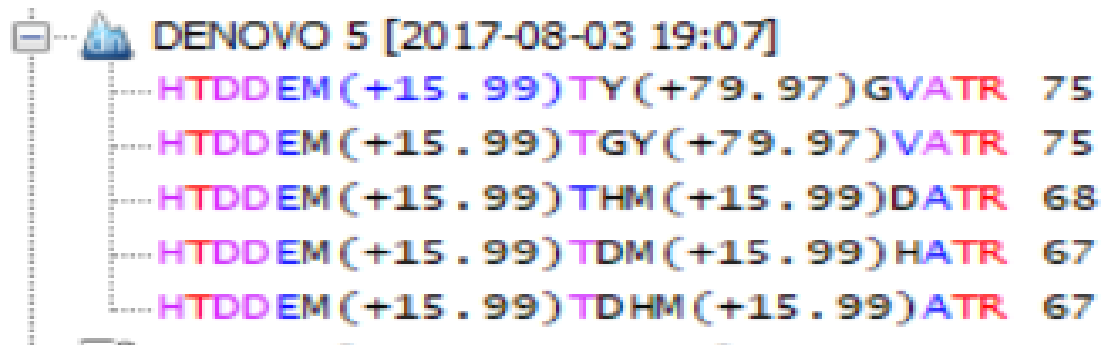


Figure .11: Experiment 3: Identification result of spectrum 1 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.

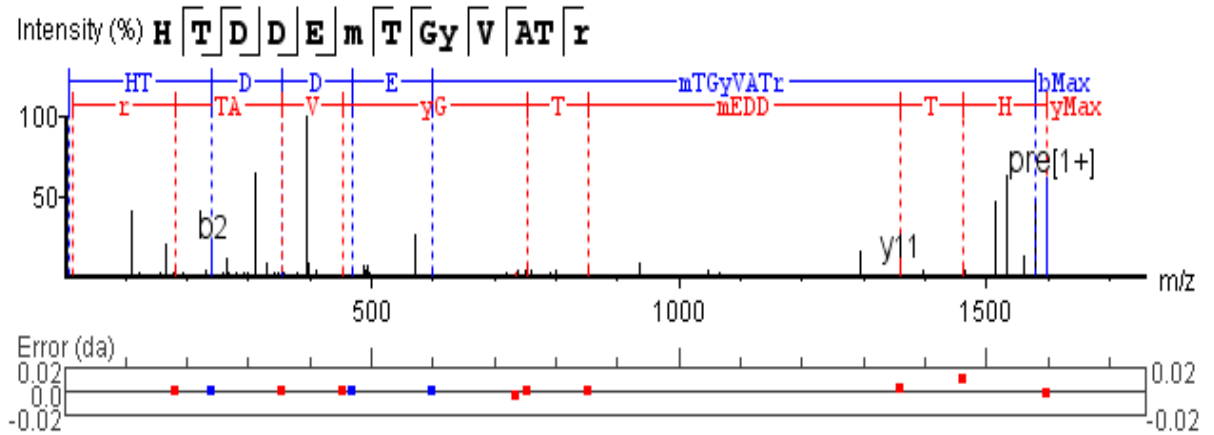


Figure .12: Experiment 3: Spectrum 2 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.

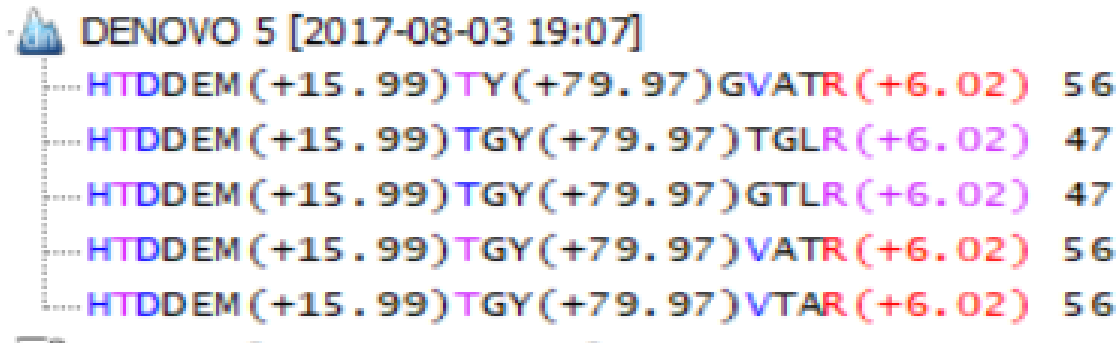


Figure .13: Experiment 3: Identification result of spectrum 2 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.

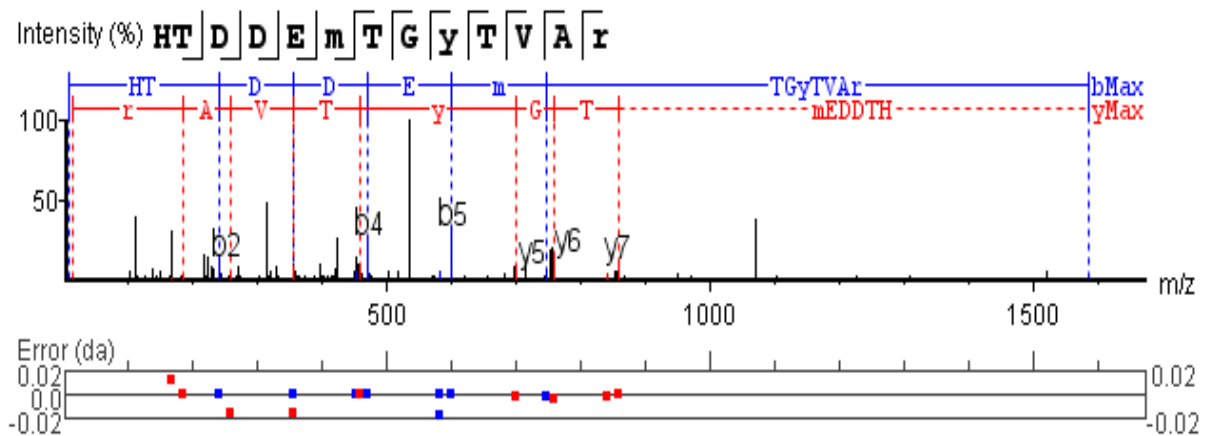


Figure .14: Experiment 3: Spectrum 3 align with candidate peptide sequence(PEAKS de novo). N-terminal ions are shown in blue and C-terminal ions are shown in red.

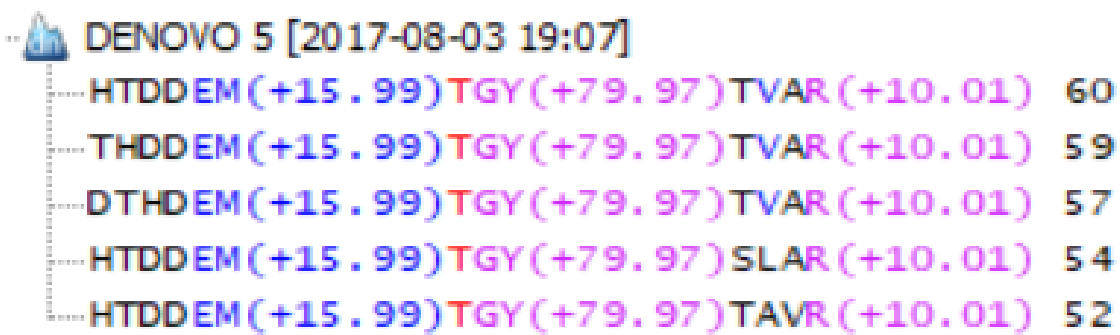


Figure .15: Experiment 3: Identification result of spectrum 3 only (PEAKS de novo). Red represents a very high confidence; purple represents a high confidence; blue represents a medium confidence; black represents a low confidence.

Curriculum Vitae

Name: Fang Han

Post-Secondary Education and Degrees: University of Western Ontario
London, ON
2015 - present PHD candidate

University of Western Ontario
London, ON
2011 - 2014 -2023 MSC

Tianjin University of Technology
Tianjin, China
2006 - 2010 BSc

Honors and Awards: University people's Scholarship
2006-2010

Related Work Experience: Teaching Assistant
The University of Western Ontario
2011 - present

Publications:

[1] Fang Han, Baozhen Shan and Kaizhong Zhang. De novo sequencing of multiple SILAC based tandem mass spectra. *IEEE 21st International Conference on Cognitive Informatics and Cognitive Computing (ICCICC)*, 2022