DEVELOPMENT AND APPLICATION OF NEW MASS SPECTROMETRY-BASED PROTEOMICS TECHNOLOGIES TO POST-TRANSLATIONAL MODIFICATIONS

by

YUE CHEN

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2008

ACKNOWLEDGEMENTS

Though only my name appears on the cover of this dissertation, a great many people have contributed to its production. I owe my gratitude to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

My deepest gratitude is to my advisors, Dr. Bellion and Dr. Zhao. Dr. Bellion’s broad knowledge and constant guidance helped me become a real scientist and serious scholar. I have been extremely blessed with his encouragement and inspiration throughout the past several years. I wouldn’t even have begun this journey of my life without his support. I have also been extremely fortunate to have been working with Dr. Zhao for the past almost seven years. It was at the Zhao lab that I received my formal training in LC/MS. Dr. Zhao continuously stimulated my serious analytical thinking and greatly assisted me with logical scientific writing. I wouldn’t have what I have today without his constant support and belief in me.

I’m also very grateful for having such a wonderful dissertation committee. Dr. Schug and Dr. Heo have given their time and expertise to better my work. I greatly appreciate their continuous support and encouragement. I would also like to thank all the professors who have been teaching me in the past a few years, whose classes have taught me so much, and staff members in the department who have made my graduation such a joyful experience.

I’m deeply indebted to the past and present members in Zhao lab, from whom I have learned so much. I would like to acknowledge Dr. Sungchan Kim, Dr. Robert Sprung, Dr. Zhihong Zhang and Dr. Zhongyi Cheng for their help in biochemsitry analysis and Dr. Yingxin Zhao and Dr. Kai Zhang for their collaborative effort on improving HPLC methodology.

I must acknowledge my wonderful collaborators Dr. Melanie Cobb and Dr. Wei Chen from UT Southwestern, with whom we worked on several projects including developing PTMap

algorithm, and Dr. Wei Gu and Yi Tang from Columbia University, with whom we worked on p53 lysine propionylation and butyrylation projects.

Most importantly, none of these would have been possible without the love and patience from my parents Dehua Chen and Hui Huang, and my wife Yankun Gao. They have been a constant source of support and strength all these years. I would like to express my heart-felt gratitude to my family.

November 6, 2008

ABSTRACT

DEVELOPMENT AND APPLICATION OF NEW MASS SPECTROMETRY-BASED PROTEOMICS TECHNOLOGIES TO POST-TRANSLATIONAL MODIFICATIONS

Yue Chen, Ph.D.

The University of Texas at Arlington, 2008

Supervising Professors: Edward Bellion, Yingming Zhao

Post-translational modification (PTM) represents a major cellular mechanism to diversify the limited number of proteins coded by genome. They provide dynamic regulation of protein functions and the means to fine-tune protein functions in response to the changing cellular environment and physiological conditions. Despite their crucial roles in cellular functions, efficient and sensitive analysis of PTMs remains a daunting analytical challenge. Traditional chemical approaches to characterize PTMs are labor-intensive and time-consuming, often with low sensitivity and specificity. Over the past ten years, mass spectrometry has emerged as an indispensable tool to identify and characterize PTMs. Mass spectrometry-based proteomic study of PTMs has become possible with the development of new analytical instruments, chromatographic technology, new techniques for enriching peptides bearing a PTM, and bioinformatics software. However, the reliability of PTM identifications with current technologies remains challengeable, hindering the process of further understanding the biological functions of PTMs. Therefore, new mass spectrometry-based technologies are much needed to improve the sensitivity and reliability of PTM identification. Towards this end, this document describes the development of novel mass spectrometry-based proteomic approaches for PTMs. It begins by introducing strategies for systematic manual verification of tandem mass spectra for peptide and PTM identification as well as the identification of common types of false positive sequence alignments from representative bioinformatics computer programs. Our results challenge some of the most popular concepts in the analysis of tandem mass spectrometry data and point out the caveats in the common practice of sequence alignment for peptide identification. The application of the strategies and concepts to manual verification led to the successful discovery of two novel protein modifications: lysine propionylation and lysine butyrylation. Our results demonstrated that these two novel modifications are dynamic PTMs that can be regulated by enzymes in a manner similar to lysine acetylation. We further developed a new software tool called PTMap based on the concepts and strategies of manual verification for unrestrictive PTM identification. The software was demonstrated to have high accuracy and efficiency compared to other known sequence alignment tools. The application of the software to the analysis of four selected proteins, human histone H4, HMG2, mouse SGK1 and BSA, led to the identification over seventy novel PTMs and polymorphisms. The data suggest that the complexity of the proteome exceeds far beyond what people have imagined. In another demonstration, the application of the software led to the discovery of four novel chemical modifications introduced mainly during *in vitro* sample handling, providing guidance for better sample preparation. Together, the strategies and tools described in this thesis provide a powerful platform to fully dissect PTMs using mass spectrometry and explore their dynamic implications in biological processes.

TABLE OF CONTENTS

APPENDIX

LIST OF ILLUSTRATIONS

LIST OF TABLES

Table Page

CHAPTER 1

INTRODUCTION

1.1 Proteins And Post-Translational Modifications (PTMs)

1.1.1 Proteins

The protein, as a polymer of 20 types of ribosomal-coded amino acids, is one of the major macromolecules in the cell, along with polynucleotide, lipids and sugar. The primary sequences of proteins are encoded by the genome of an organism, while their functions are regulated through diverse cellular mechanisms including alternative splicing, micro ribonucleic acid (RNA) and post-translational modifications. The size of a protein can range from several thousand Daltons, (*e.g.,* ubiquitin and insulin), to over 500 kDa, (*e.g.*, titin), while the half-life of a protein can extend from a few minutes to over 100 days [1,2].

As fundamental architectural building blocks in the cell, proteins come in various types of sizes and shapes, which are critical for successfully carrying out their functions. Typically, protein structure can be categorized into four distinct levels. First, the primary structure of a protein is its linear sequence of amino acids which is encoded by the DNA sequence. As a consequence of the peptide bond, this linear sequence exhibits directionality with a free amino group on one end and a free carboxyl group on the other, which are termed N-terminal and C-terminal, respectively. Second, the linear polymer of amino acids is spatially arranged into different types of stable structures through non-covalent hydrogen-bonding in a repeated pattern between peptide bonds of different residues that are spatially close to each other. Such structures mainly include α-helix, β-sheet, β-turn and coils. Third, each protein molecule exhibits a three-dimensional arrangement, referred to as its tertiary structure, which is stabilized by hydrophobic interactions and hydrogen bonding between different modules. Tertiary structure holds together the secondary structural elements such as α-helix and β-sheet and can be further stabilized by covalent chemical bonding such as disulfide bonds and isopeptide bonds. Fourth, polypeptide chains containing individual tertiary domains can associate with each other and form a specific structural complex. Typically, such a multimeric protein complex does not function properly unless all the subunits are present and folded into specific structures. An example can be found in the multiple subunits of hemoglobin ($\alpha_2\beta_2$ tetramer): a mutation in one of the polypeptide subunits disrupts the binding between the polypeptide chains as well as the protein function [3].

Proteins carry out diverse types of cellular functions including structure, movement, catalysis, transport, signaling and so on. Although proteins may function alone, proteins in cells are often found to form multi-protein complexes that facilitate highly efficient substrate operation and more stable formation of structures. For example, the transcription pre-initiation complex (PIC) of RNA polymerase II is formed by the Transcription Factor IIE, IIF, and IIH subunits to start the transcription process. Cellular functions of proteins are regulated elegantly at several distinct levels including transcriptional regulation of gene expression activity, post-transcriptional mRNA regulation by mechanisms such as alternative splicing and micro-RNA, post-translational regulation by covalently bonded modifications, and protein degradation or cleavage. These regulation mechanisms often work together to keep a balanced concentration of the active form of the proteins. For example, a protein transcription factor termed p53, named after the protein apparent molecular weight of 53 kDa, is a critical guardian protein that controls the cell cycle and a well-known tumor suppressor. P53 is regularly expressed by the cell but remains inactive through binding to a protein named murine double minute (MDM2) that promotes p53 degradation through ubiquitination of lysine residues in p53 [4]. In case of cell stress such as DNA damage induced by ultraviolet or infrared radiation, p53 can be phosphorylated by specific kinases that target p53, disrupting its binding with MDM2 and further preventing its rapid degradation by the proteasome [5]. Acetyl transferases such as p300 and CREB binding protein

(CBP) further acetylate C-terminal lysines of p53 and activate the DNA-binding domain [6] so that it can activate the transcription of downstream genes such as p21. p21 binding to cyclin-dependent kinase 2 (cdk2) is necessary to prevent the cell from moving to S phase from G1 phase [7]. Therefore, an increased concentration of active p53 protein, regulated through covalent modifications of specific residues and a reduced degradation rate, controls the cell cycle and prevents unlimited cell division.

1.1.1 Covalent post-translational modifications

Post-translational modifications (PTMs) represent a major dynamic mechanism to regulate protein function and cellular activity by forming covalent bonds on the side chains of specific amino acids. In this process, the side chains of the modified amino acids are typically rich in electrons and can attack an electrophilic center in a group transfer reaction catalyzed by specific enzymes. Depending on the types of co-substrates or cofactors required in each modification reaction, PTMs can be grouped into different categories, such as S-adenosyl methionine (SAM)-dependent methylation, coenzyme A (CoA)-dependent acylation, nicotinamide adenine dinucleotide (NAD)-dependent ADP ribosylation, ATP-dependent phosphorylation, and so on. At least 15 out of 20 ribosomally coded amino acids have been found to be post-translationally modified [8], although some of them are limited to certain species such as Asp phosphorylation in bacteria and fungi [9]. Such diverse types of modifications result in a much bigger inventory of proteomes (more than 1 million species) than what is predicted by the genome (around 25,000 genes in human). It was estimated that 5% of the eukaryotic genes code for enzymes that regulate the status of various types of PTMs [10].

ATP-dependent phosphorylation on the side chains of Ser/Thr/Tyr is the most comprehensively studied protein modification. The enzymes catalyzing the phosphoryl group transfer reaction comprise over 500 different proteins in humans and they form a unique group of kinases, termed the kinome. Phosphorylation of Ser/Thr/Tyr converts the hydroxyl side chains (-OH) of the substrate protein that are small and neutral to acidic phosphoryl groups (-

HPO_3^-) that are usually negatively charged and bulky under physiological conditions. Such a dramatic change easily alters some important properties of the protein, such as conformation, isoelectric point and hydrophilicity, which subsequently induces the functional transformation of the target substrate. For example, the MAP kinase p38 is activated by MKK3 by phosphorylation on a TGY motif [11], which results in structural changes within a domain [12], and the active form of p38 is able to phosphorylate downstream substrates such as ATF-2 in the nucleus [13].

Acetyl-CoA-dependent ε-amine acetylation is an important histone modification that has been widely studied. The reaction is typically carried out by histone acetyl transferases (HATs), which activate the electron-rich ε-amine of lysine side chains in the substrate protein to nucleophilically attack the acetyl group on the terminal of acetyl-CoA, a high-energy metabolite intermediate in the cell. The reaction converts the primary amino group into an amide, which neutralizes the positive charge typically carried on the lysine side chain and also induces significant structural change through the addition of a bulky acetyl group. Like phosphorylation, lysine acetylation is reversible through the action of a group of histone deacetylases (HDACs). A recent proteomic survey showed that lysine acetylation is a prevalent event involving a wide range of protein substrates in a variety of cellular processes [14]. The study also showed that over 20% of mitochondrial proteins are lysine-acetylated, suggesting a potentially intimate correlation between lysine acetylation and the energy metabolism processes in mitochondria.

In addition to phosphorylation on Ser, Thr, Tyr and acetylation on Lys, over 200 different types of PTMs have been identified, forming a complex regulatory network that controls diverse cellular physiological processes [15]. Studies have shown that PTM sites are not randomly chosen. Instead, enzymes seem to prefer certain motif structures. For example, N-glycosylation on Asn is preferred in a motif having the consensus sequence Asn-X-Ser/Thr/Cys (X is any amino acid except Pro). Despite the importance of protein modifications, the path of identifying and characterizing PTMs has not been smooth. In the next section some processes traditionally used for PTM identification will be reviewed.

1.1.2 Traditional approach to identify and characterize PTMs

Unveiling the identity of protein PTMs, especially novel modifications, before the introduction of biological mass spectrometry, depended on traditional chemical labeling and analytical technology such as HPLC and electrophoresis. The identification process often required complicated sample preparation and a large quantity of target proteins, making it impossible to systematically identify PTMs or specific interacting proteins on a large scale.

The traditional approach to identify PTMs typically involves chemical hydrolysis of proteins by acids or enzymes. The amino acids bearing the modifications can be isolated by HPLC or electrophoresis due to changes in properties such as pI, hydrophobicity, and molecular weight (MW) induced by the side chain modification. Routine chemical analysis methods can be applied to identify the specific properties of the modifying group such as element composition, MW, amino acid preferences, and characteristic reactions of the functional groups. Based on this information, the structure of the new modification can be deduced. A good example is the first identification of protein phophorylation.[16]

In addition to chemical analysis, radioisotope analysis was widely applied in earlier years to discover new protein modifications. This strategy was first introduced as tracers in the study of intermediates in the biochemistry pathways in 1935[17] and even today it remains a highly sensitive and specific tool. Good examples are the first identification of lysine acetylation in calf thymus histones in 1968[18] and the first identification of lysine acetylation in a non-histone protein, p53, in 1997.[6]

The identification of protein PTMs by the traditional approach is a difficult process due to the complicated sample preparation and lengthy analytical procedure. The maturing mass spectrometry-based proteomic technology brought biochemical analysis into a new era in the late 1980's.

1.2 Biological Mass Spectrometry

Twenty years ago, the development of ElectroSpray Ionization (ESI) and Matrix-Assisted Laser Desorption/Ionization (MALDI) technologies heralded the beginning of biological mass spectrometry. Both approaches solved the long standing technical challenge of the ionization of the large biomolecule and enabled the analysis of proteins and polynucleotides by mass spectrometry. Combined with the latest developments of new mass analyzers, computer hardware and bioinformatics software, biological mass spectrometry makes it possible to study protein expression and modification at the system level and becomes an invaluable tool for modern biochemical studies.

1.2.1 Common ionization source

Two types of ionization methods have been most widely applied in the studies of proteins and polynucleotides: ElectroSpray Ionization and Matrix-Assisted Laser Desorption/Ionization. ESI was developed in 1988 by John Fenn's lab at Yale University [19] .As a so-called soft ionization method compared to traditional Electro Impact (EI) and Chemical Ionization (CI), ESI overcame the difficulty of easy fragmentation upon ionization of the large biomolecule. As illustrated in Figure 1, a protein or peptide of interest is dissolved in volatile solvents, such as water, methanol or acetonitrile. The liquid sample is pushed through a charged capillary with an opening facing towards or orthogonal to the inlet of the mass detector.

The voltage between the capillary tip and the inlet can range from 1.0 to 5.0 kV depending on the composition, flow rate and vacuum. With high voltage, the liquid droplets shoot out of the capillary in a jet stream and the dissolvation process starts while the solvent vaporizes, which is often helped by the weak vacuum generated from the nearby inlet of the mass spectrometer as well as the nebulizer gas such as nitrogen introduced at the capillary tip. At this stage, the liquid droplets form a so-called Taylor cone. As the solvent continues to evaporate, the charges on the surface of the liquid droplets are repulsive to each other and force the liquid droplet to go through series of Coulomb fissions, which continues until the solvent completely evaporates and charges reside on the single molecule. The total number of charges carried by each molecule depends on its chemical properties, most importantly the basicity of the molecule. The more basic groups available, the more charges a molecule may carry after electrospray, and it is common to observe a distribution of different charge states for the same molecule. As the mass spectrum displays mass-to-charge ratios of each charged ion, the mass of the molecule can be determined from simple calculation.

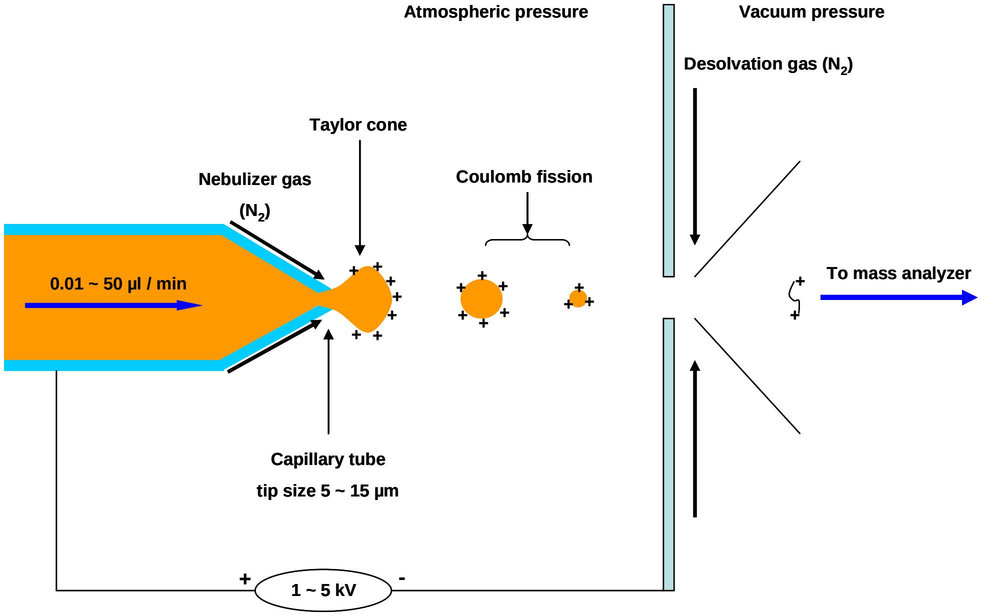


Figure 1 An illustration of electrospray ionization (ESI).

Matrix-assisted laser desorption/ionization was developed in 1985 by Hillenkamp and Karas who were able to ionize peptides up to 2.8 kDa in mass,[20] while in 1987, Tanaka and colleagues achieved a breakthrough with MALDI by using a combination of matrix and laser and were able to ionize the 34.5-kDa protein carboxypeptidase-A.[21] Today, with the development of new matrix molecules and instrument designs, MALDI mass spectrometers can analyze proteins with masses of more than several hundred kilo Daltons. The principle of the MALDI technique is illustrated in Figure 2. Briefly, the biomolecules of interest are first mixed with matrix molecules and deposited on a polished stainless steel surface. While the solvent evaporates, the biomolecules co-crystalize with the matrix on the spot and the metal plate is usually placed in a vacuum chamber. A nitrogen laser of wavelength 337 nm is typically used to fire at the deposited spot. The matrix molecule is designed to specifically absorb the laser energy and upon laser irradiation, the matrix molecules become thermally activated and then ionized. Following this step, the co-crystalized analyte molecules interact with the activated matrix ions and form analyte ions in the gas phase. Unlike ESI, MALDI tends to generate molecule ions bearing limited number of charges, and therefore, MALDI mass spectra are usually much less complex and require less mass deconvolution.

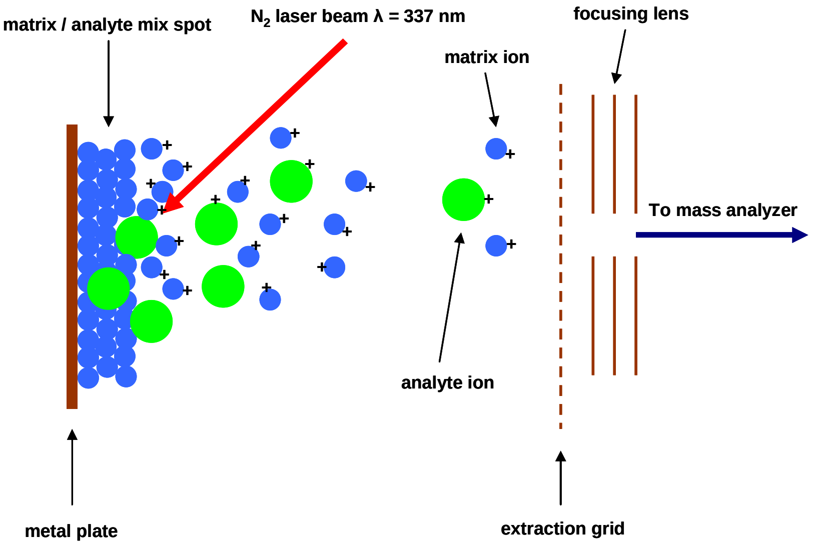


Figure 2 An illustration of matrix-assisted laser desorption/ionization (MALDI).

Although the detailed mechanisms of ESI and MALDI remain to be elucidated, both techniques have been widely applied in the analysis of biological samples. In this thesis, my research projects were mainly carried out using micro-electrospray (micro-ESI) or nanospray technology which, unlike electrospray, typically interfaces with capillary HPLC column and the flow rate is at nanoliter per minute range. At such flow rate, peptide ionization does not require the facility of sheath gas or nebulizer gas. Such an approach typically offers higher sensitivity and higher ionization efficiency with very flow HPLC flow rate [22], and higher chromatographic performance with capillary column.

1.2.2 Common mass analyzers

Following the ionization, the molecules are analyzed by mass analyzers that measure the mass-to-charge ratio of each ion. Currently, there are four types of commonly seen mass analyzers: time-of-flight, quadrupole and quadrupole ion trap, Fourier transform ion cyclotron resonance and traditional magnetic sector. In biochemical analysis, the most widely used mass analyzers are the first three types and they are briefly introduced below.

The time-of-flight mass analyzer, developed in the 1940's by Stephens [23], is based on the simple principle that, given the same kinetic energy, the velocity of ions is dependent on the mass-to-charge ratio of each ion. When these ions travel in the same distance from ionization source to the ion detector, the velocity determines the travel time. Therefore, the ions with higher mass-to-charge ratio travel slower than the ions with lower mass-to-charge ratio. By accurately measuring the travel time of each ion from the starting point (ionization source) to the ion detector, the TOF mass analyzer can accurately determine the mass-to-charge ratio of each ion.

The quadrupole mass analyzer was developed first in 1953 by Paul [24], who later also developed the 3-D quadrupole ion trap. The quadrupole mass analyzer works with four electrodes with oscillating RF electrical fields. The analyzer can select ions with specific m/z passing through the four electrodes by applying a specific DC voltage which destabilizes other ions. The quadrupole ion trap works in a similar fashion except that it traps the ions by applying AC voltage to the two end caps and the ring. The ions are first trapped in the mass analyzer and then sequentially ejected. The RF voltage of the ejection can be used to calculate the mass-to-charge ratio of the ions. A special advantage of quadrupole ion trap is its capability to accumulate or enrich ions for a short period of time, eject the ions that are not relevant and fragment specifically the ions of interest. Ions at low concentrations can therefore be enriched and analyzed. This technique allows the quadrupole ion trap to have high fragmentation efficiency, high sensitivity and high dynamic range for a variety of biological applications. The main research projects in this thesis were carried out using a linear ion trap mass spectrometer from Thermo Fisher, Inc.

Recently, increasing demand for high-resolution mass spectrometers has spurred interest in the Fourier Transform Ion Cyclotron Resonance (FTICR) mass spectrometer[25,26]. This mass analyzer traps ions using either a Penning trap with a magnetic field or a nonmagnetic, spindle-like electrode. The frequency of ion oscillation in the trap is determined by the mass-to-charge ratio of the ions and can be measured by first recording the time-decay image current generated by the ion cycling and then Fourier transforming the signal from the time domain into the frequency domain. The frequency of the ions can be used to calculate the mass-to-charge ratio of each ion. Because the frequency of the ions can be very accurately measured, FTICR's mass resolution and accuracy is very high. This is very important in proteomic analysis for accurate sequence identification and quantification analysis.

1.2.3 HPLC-MS

A number of chromatographic techniques have been applied in conjunction with mass spectrometry, including gas chromatography, liquid chromatography and ion mobility. Among them, high performance liquid chromatography in conjunction with electrospray mass spectrometry has been most widely applied to studying proteins and other biomolecules as shown in Figure 3.

There are several advantages to this instrument configuration. First, HPLC has high separation power and is able to separate complex mixtures chromatographically before they are introduced into mass spectrometer. Highly abundant species in the sample can be separated from species with low abundance. Hence, the approach significantly reduces the sample complexity and increases the dynamic range of the analysis. Second, HPLC offers high chromatographic resolution for each species and is able to elute the analyte in a sharp peak with low peak volume. For a given total amount of analyte in the sample, the concentration is dramatically increased in the eluting peak compared to the whole sample due to the small peak volume. For example, suppose 1 µmol of peptide is dissolved in 10 µL of solvent with a concentration of 0.1 M. Upon HPLC elution, 50% of the peptide can elute in 6 s with a flow rate of 0.1 µL/min and peak volume of 0.01 µL. Therefore, the peptide concentration in the eluting peak is 50 M, five hundred times the peptide concentration in the sample. The mass spectrometer is a concentration-dependent device, meaning that the signal intensity of an ion from a mass spectrometer depends on the concentration of the analyte, instead of the total amount of analyte in the sample. Accordingly, an increase in concentration will allow significant improvement of the sensitivity. Third, HPLC can interface online with an electrospray ionization source, which significantly reduces sample loss compared to offline preparation and allows the detection of analytes of very low concentration.

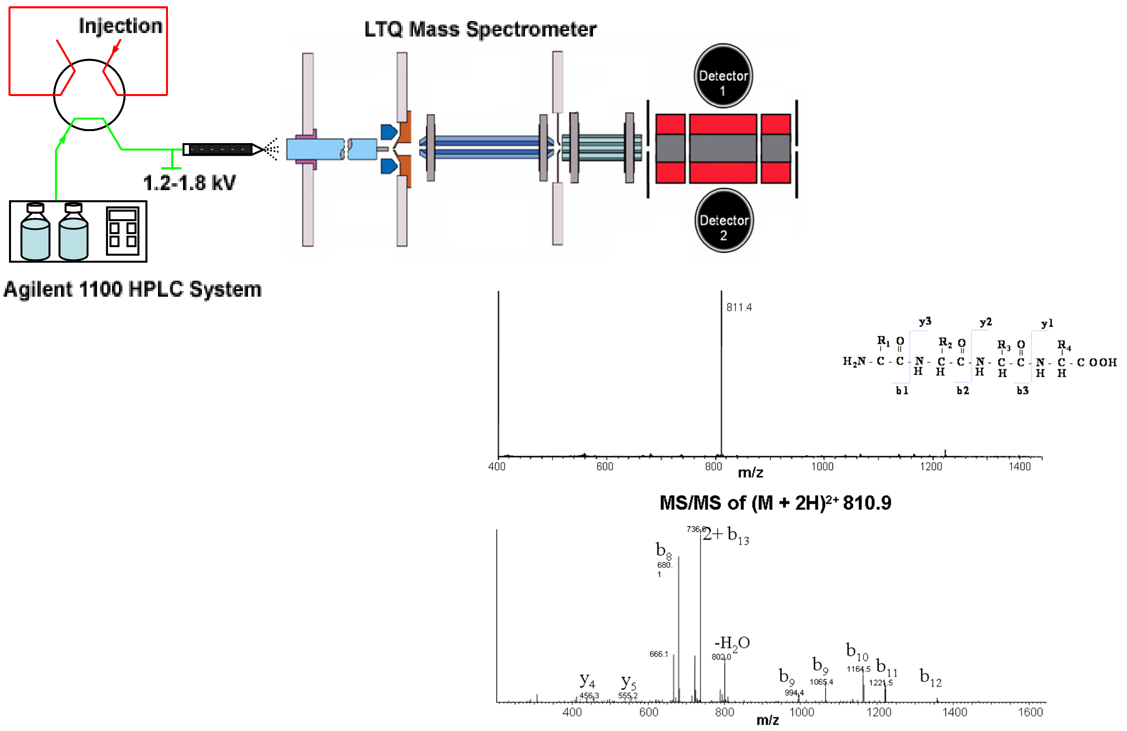


Figure 3 An example of instrument set-up for HPLC-MS/MS analysis.
It shows the Agilent 1100 capillary flow HPLC system interfaced with an LTQ linear ion trap mass spectrometer through a nanospray source. The mass spectrometer generates MS spectrum (middle) and MS/MS spectrum (lower) of bombesin peptide. The spectra for bombesin peptide are shown as examples.

1.2.4 Tandem mass spectrometry for protein identification and mapping PTM sites

Traditionally, peptide mass fingerprinting has been used to identify proteins based on the assumption that the enzymatic digestion profile of a protein is uniquely dependent on the protein's sequence. However, such an approach is seriously challenged in the analysis of complex protein mixture digests, due to the existence of large numbers of peptides with similar masses. Tandem mass spectrometry has been developed to meet these challenges. This approach typically involves multiple stages of mass spectrometry analysis, allowing users to obtain more information regarding the peptide in addition to the mass. For example, a triple quadrupole mass spectrometer has been used in connection with a time-of-flight mass spectrometer. Peptide ions with a specific m/z can be selected by the first quadrupole and then fragmented in the gas phase in the second quadrupole. These fragments are analyzed by the time-of-flight mass spectrometer in tandem. In this way, the user is able to distinguish peptides with similar masses based on different fragmentation patterns, which significantly improves the identification accuracy. This technique is also critical for PTM identification with a mass spectrometer. This is because, in the traditional PMF approach, only the mass of the modified peptide can be detected, with no information regarding the location and type of modification. With tandem mass spectrometry, the user can analyze the fragmentation pattern of the peptide and identify the modification type and site where the mass shift occurs. Tandem mass spectrometry has been widely applied in biochemical analysis and such functions are built into most modern mass spectrometers. In this thesis, all the research projects described were carried out using tandem mass spectrometry technology.

1.3 Proteomics And The Analysis Of PTMs

1.3.1 Definition of proteomics

The term "proteomics" was coined to describe the study of all proteins, similar to "genomics". However, the complexity of proteins is far more than what the genomic sequence predicts, due to cellular regulatory mechanisms such as alternative splicing and post-translational modifications. In addition to various sizes and forms a protein may take in the cell, the expression levels of different proteins vary dramatically and therefore the dynamic range issue becomes one of the most difficult challenges in proteomic analysis.

In the past decade, mass spectrometry has become an indispensable tool to study proteomics due to its unparallel sensitivity and adaptability for high speed, systematic analysis. By interfacing with chromatography such as HPLC or ion mobility, people are able to reduce sample complexity and boost sensitivity of mass spectrometry analysis. The establishment of gas-phase fragmentation theory, accessible genomic sequencing databases, and improvements in computing hardware and software allow mass spectrometry to have high performance in the study of peptides or proteins. Recent advancement in the development of new mass analyzers such as Orbitrap and LTQ-Electron Transfer Dissociation (ETD) opens new fields for quantification and PTM analysis. In this thesis, I will mainly focus on the study of PTMs with mass spectrometry and here I will introduce some general approaches and technical challenges in PTM analysis.

1.3.2 General approach for mass spectrometric analysis of PTMs

Mass spectrometry is especially suitable for studying PTMs because a PTM on the side chain of an amino acid will typically induce a MW change that is easily reflected in the mass spectrum. For example, a tryptic peptide “IHAGYTR” has a monoisotopic mass of M = 816.424. In a mass spectrum from MALDI-TOF mass spectrometer, the singly charged peptide bearing a proton will show a peak at (M+H) = 817.434. If the same peptide also has a PTM of tyrosine phosphorylation, the modified peptide “IHAG$_p$YTR” will have a peak at (M+H) = 897.414 in a MALDI mass spectrum. The increase of about +80 Da is induced by the modification and it is the net increase after adding the mass of H_3PO_4 (98 Da) and removing the mass of H_2O (18 Da). By observing such a characteristic mass shift in the mass spectrometer, people can hypothesize that the peptide “IHAGYTR” may be phosphorylated.

Mass information alone is often not enough for determining a PTM site. In this case, the peptide “IHAGYTR” contains another residue (T) that also can be phosphorylated. The masses of the T-phosphorylated peptide and the Y-phosphorylated peptide are the same and therefore even with a high-resolution mass spectrometer, these two molecular species cannot be distinguished with this approach. To localize a PTM site, people traditionally use Edman degradation, which cleaves amino acids one by one starting from the N-terminal and identifies the free amino acid with HPLC. However, it usually requires a relatively pure sample and a large starting amount (2~5 pmol) of the modified peptide, a requirement that is very difficult to meet in biochemical purification, especially with *in vivo* sample preparation. The application of tandem mass spectrometry (MS/MS) can solve this problem easily. For example, in an LTQ-linear ion trap mass spectrometer, the ions with the same mass as the phosphorylated peptide can be isolated and enriched. These ions are fragmented in the gas phase and then subjected to a second round of mass analysis to measure the masses of all fragments. In collision-induced fragmentation, which we will introduce in more detail later, the backbone of the peptide is fragmented. For different phosphorylated forms, the fragmentation pattern is different. For example, the b ion fragment “IHAGY $^{1+}$” has a mass of 542.27 in the MS/MS spectrum. If the tyrosine is phosphorylated, the fragment “IHAG$_p$Y $^{1+}$“ will have a mass of 622.27. Therefore, in the MS/MS spectrum, people are able to distinguish different modification isoforms with different modification sites.

1.3.3 Challenges and solutions in PTM analysis

Analytical techniques for PTM analysis face the challenge of dynamic range, which mainly comes from two aspects. First, proteins bearing PTMs are often at low stoichiometry in the cell. For example, it may be sufficient to activate a signaling pathway with only 5% of the target protein phosphorylated while 95% remains unmodified. Second, protein expression level varies dramatically in the cell, with some housekeeping genes expressed constantly and some regulatory enzymes expressed at only very low levels. To address the issue of dynamic range in

PTM analysis, various chromatographic techniques have been applied to fractionate the biological sample, reduce sample complexity and enrich the protein of interest. For example, to study the phosphoproteome, instead of analyzing all the proteins in the cell lysate, it would be better to fractionate the proteome with two-dimensional strong cation exchange chromatography - immobilized metal affinity chromatography technology to enrich phosphorylated peptides[27,28].This is based on the principles that first, the phosphorylated peptides bind weakly to SCX and therefore elute earlier than non-phosphorylated peptides; and second, phosphorylated peptides specifically bind to metal ions such as Fe^{3+}, Ga^{3+} or TiO_2 in the IMAC matrix forming chelating complexes, while non-phosphorylated peptides do not bind or bind very weakly. After the two-step separation, non-phosphorylated peptides were largely removed in the sample and phosphorylated peptides were enriched. This approach overcomes the dynamic range problem caused by the low PTM stoichiometry level for each protein. To overcome the dynamic range problem caused by highly abundant protein interference, people have used protein-specific antibodies to enrich the target protein of interest before PTM analysis, or chromatographic fractionations such as size-exclusion chromatography or 2-D electrophoresis to reduce the sample complexity.

1.4 Sequence Alignment In Proteomic Analysis

1.4.1 Peptide fragmentations in tandem mass spectrometry

Tandem mass spectrometry generates an MS/MS spectrum for each peptide fragmentation containing the m/z values of sequence-determined fragments. Since peptides or proteins are linear polymers of amino acids, they tend to fragment in a predictable pattern under the same fragmentation conditions. The types of peptide ions that can be generated in the gas phase are illustrated in the Figure 4, and the most valuable fragmentation pathway for peptide sequencing is along the peptide backbone.

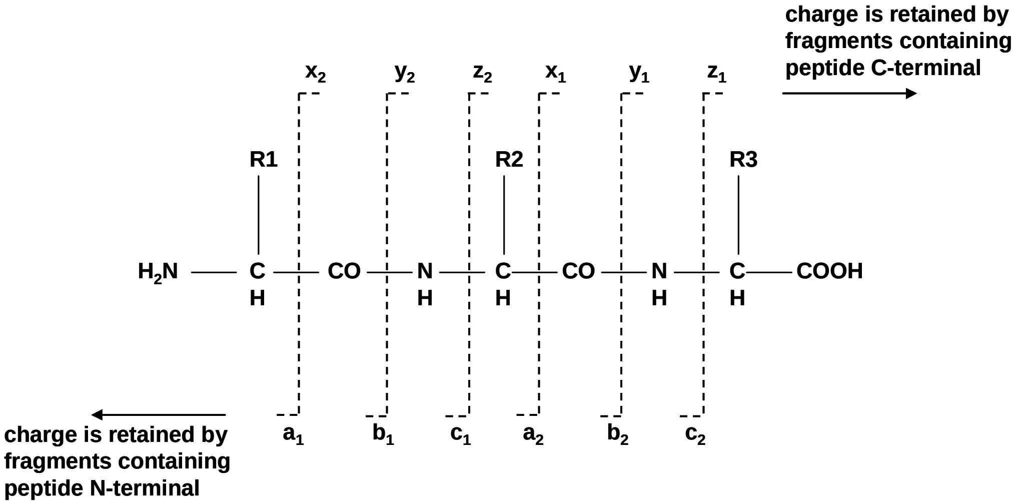

charge is retained by fragments containing peptide C-terminal
x_2
y_2
z_2
x_1
y_1
z_1
R1
R2
R3
H_2N
C
H
CO
N
H
C
H
CO
N
H
C
H
COOH
charge is retained by fragments containing peptide N-terminal
a_1
b_1
c_1
a_2
b_2
c_2

Figure 4 Biemann nomenclatures[29] for peptide fragmentation.
a, b, and c ions represent the corresponding peptide backbone fragment ions containing the peptide N-terminal, and x, y, z ions represent the corresponding peptide backbone fragment ions containing the peptide C-terminal.

Many technologies have been developed to induce peptide or protein fragmentation in the gas phase, such as Collision-Induced Dissociation (CID), InfraRed MultiPhoton Dissociation (IRMPD) and more recently Electron Capture Dissociation (ECD) and Electron Transfer Dissociation (ETD). Among them, collision-induced dissociation has been most widely used because of its simple design and easy-to-interpret spectrum. In a typical CID MS/MS spectrum from an ion trap type of mass spectrometers, a peptide tends to generate b or y ions. However, the intensity of each ion is not equally distributed, creating a problem for interpreting the peptide sequence *de novo* from the spectrum. Several models have been proposed to explain the fragmentation mechanism in peptide CID, and among them, the most widely cited is the mobile-proton model proposed by Wysocki [30]. In this model, the peptide fragmentation pathway is dictated by the balance of available free protons and basic groups that sequester protons, and the fragmentation is induced by the mobile proton transferred to the peptide backbone at different energy levels. As a result of this fragmentation mechanism, the backbone of the peptide is differentially fragmented depending on the sequence and even on different modification groups on the amino acid side chains. Despite of the lack of complete backbone fragmentation, MS/MS spectra still contain rich information regarding the peptide sequences and can be aligned with fragments theoretically predicted by software.

1.4.2 Automatic sequence alignment for peptide and modification identification

The concept of applying computer analysis to peptide sequence alignment was first introduced by Biemann in 1989 [31]. However, it was not widely applied in large-scale peptide or protein analysis until 1994 when Eng in Yates' lab developed Sequest which introduced fragmentation pattern alignment and used a cross-correlation function to evaluate the sequence alignment quality [32]. Today, there are a number of commercial or open-source database search algorithms on the market; the most popular ones include Mascot [33], Sequest [32,34], X!Tandem [35], OMSSA [36], Phenyx [37], and ProteinProspector [38].

The principal approach for these algorithms is similar, involving four steps as illustrated in Figure 5. First, the software processes a protein sequence database and digests the protein sequence into peptide sequences *in silico* based on the user-specified enzymes, maximum number of missing cleavages allowed and any fixed or variable modifications. In this step, the software may also calculate the fragmentation pattern of each peptide. Here, "fixed" or "static" modification usually refers to certain modifications that are considered to happen to every specified amino acid. For example, Cys alkylation by iodoacetamide is a common chemical reaction used prior to enzymatic digestion to alkylate free Cys and prevent the formation of disulfide bonds. Usually, such a reaction can alkylate nearly all free Cys. Therefore, for easy computation, the user can specify the modification as a "fixed" modification so that the software would use the mass of Cys plus 57 Da (57 Da is the net mass increase after alkylation) instead of just the mass of Cys to calculate a peptide's mass or fragment ion mass when a peptide sequence contains a Cys. In contrast to "fixed" modification, "variable" modification means that a modification might or might not occur on a residue, such as phosphorylation on Ser. In this case, the software would need to generate a peptide form with the modification and without, and both are candidate sequences for subsequent spectral alignment. In these cases, the software takes into account the types of modifications and generates all the candidate peptide sequences before sequence alignment. This approach has been termed “restrictive” analysis of PTMs. This is in contrast to “unrestrictive” analysis which will be described in detail later in the chapter.

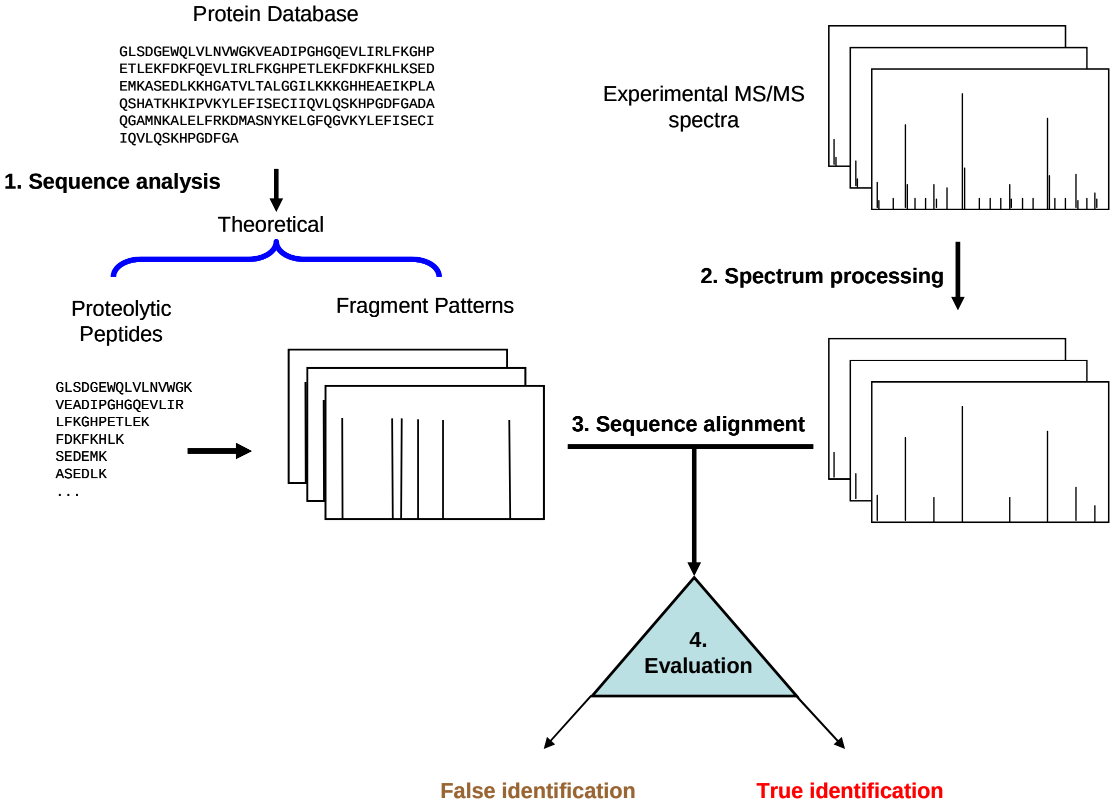


Figure 5 An illustration of the four steps in automatic sequence alignment for peptide identifications.
1. database sequence analysis; 2. experimental spectrum processing; 3. sequence alignment; 4. evaluation of identification results.

Second, the raw data is processed by the software. Since in a typical MS/MS spectrum, the instrument may generate hundreds of peaks with different intensities, different algorithms may choose different ways to convert raw data into a certain manageable data structure for optimized performance. Generally, peak selection should be balanced between sensitivity and performance. If too many peaks are selected, it may include all the true signals but it also includes most noise peaks that will cause false positive identification and also longer computation time. However, if too few peaks are selected, it may cause the loss of sensitivity. The principle is to select the minimum set of representative peaks for the most efficient sequence alignment. Some software will apply a certain signal-to-noise cutoff threshold specified by the user to remove noise signals with low intensity, while others arbitrarily choose the top N most intense peaks in the MS/MS spectrum (N is usually specified by user) for sequence alignment.

Third, the software performs sequence alignment exhaustively between each candidate peptide sequence and the MS/MS spectrum. Since not all the algorithms make their alignment algorithm publicly available, here I will use Sequest as an example [32]. In this step, Sequest first utilizes a simple preliminary scoring function Sp, which is based on the total matched ion intensity, the percentage of matched ions as well as sequence matching continuity to calculate the top 500 peptide candidates. Then, Sequest adapts the pattern recognition function from infrared library searching to further compare the top 500 peptide candidates. It uses a discrete fast-Fourier transform function Cn to correlate the experimental MS/MS spectrum with the theoretical fragmentation patterns and a score is calculated based on the sum of the multiples of all matched fragments. The best match should have the highest correlation value. The advantage of this approach is its fast calculation and easy computer programming, while one of the disadvantages of this approach is its dependence on the simulation of the theoretical fragmentation pattern. Therefore, the application of the algorithm is limited by the people’s knowledge of peptide fragmentation and may introduce bias against the identification of peptides with abnormal fragmentation patterns. For example, the theoretical spectrum generated by the software assumes that the ions from the loss of water and ammonia of the original b or y ions were assigned an intensity value of 1/5 of the original b or y ions, However, in some cases, such as E or Q in N-terminal of the peptide, the water loss or amine loss peak of b ions may be 10 times stronger than the corresponding b ions. Therefore, the similarity between the theoretical spectrum and the experimental spectrum, and the correlation score are compromised. As mentioned before, different software may perform sequence alignment in different ways and the results may vary significantly from one program to another.

Fourth, all the alignment results are analyzed by each algorithm and result output is generated. Here, the software performs a critical task, which is to remove false sequence alignments and select the correct result. We use two algorithms as examples. One is Sequest, which is not based on statistics [32], and the other one is Mascot, which was the first statistical scoring algorithm [33]. In Sequest software results, several scores are typically obtained for each identification, including Cn, Sp and ∆Cn, the difference in Cn between the top and second ranked peptides. People have found that the threshold for positive identification can vary among MS/MS spectra of the same peptides but bearing different charge state. Therefore, systematic evaluation of these scores showed several cutoff thresholds for positive identification, such as Cn >1.5 for a +1 charged peptide, Cn > 2.0 for a +2 charged peptide and Cn > 2.5 for a +3 charged peptide. The results below the thresholds are considered false identifications. In Mascot software, the peptide's original score is not reported after alignment. Instead, the statistical distribution of all the alignments is plotted and each alignment is given a probability, indicating the chance that the alignment is random. Then, an absolute statistical score threshold is calculated by dividing the typical probability threshold such as 0.05 by the total number of alignments performed in that specific search. A sequence alignment with probability better than the threshold is considered a non-random event and the top hit is considered a positive identification. Since the statistical analysis relies only on the dataset generated from the database and does not relate to the charge state of the peptides, the statistical cutoff threshold is typically unrelated to the charge state of the peptides. Following the four steps of data processing, the software will typically assemble the peptide identification list into a protein identification list and output the result in HTML or XML format.

1.5 Technological Challenge Of Molecular Characterization Of Ptm Pathways And Questions Addressed By This Thesis

Despite advances in sequence alignment algorithms, characterization of PTMs using mass spectrometry remains a challenging problem [39,40]. In addition to the previously mentioned dynamic range issue faced by all PTM analyses, there are a number of hurdles in the identification of PTMs with high accuracy and sensitivity from tandem mass spectrometry data. In this section, I will describe these major challenges and introduce the strategies I chose to solve these problems in the following chapters of the thesis.

1.5.1 Statistics-based sequence alignment

Among the most popular sequence alignment programs, statistical analysis of the identification results has become a routine practice. The major advantage of this strategy is its independence from subjective analysis by human interpretation of a mass spectrum, and therefore a unified scoring standard that is statistically significant can be achieved. Such a statistical analysis strategy can be relatively simple in execution and is very effective for removing most of the false-positive sequence alignments in identification of peptides. However, in our daily analysis of mass spectra, we have observed many cases where the false-positive results cannot be distinguished from the positive data. When we looked into the data, it was not too hard to identify the reason.

The statistical significance test is based on the establishment of the Null hypothesis. In sequence alignment results, the analysis of the Null hypothesis suggests that the identification of a peptide is purely random. Therefore, when an identification has a score above a significance threshold, we would assume that the identification is positive, However, in a real situation, the false sequences matching to the spectra are not random events; instead, they are highly sequence-dependent and usually have a high statistical score well above the threshold. Therefore, statistical analysis cannot remove these false positives, while they can be efficiently identified and removed by careful manual verification analysis. We have developed a series of rules for systematic manual verification of tandem mass spectra and our analysis has revealed a series of common false-positives that cannot be removed completely by popular sequence alignment algorithms [41]. This work will be presented in Chapter 2. The application of these manual verification principles resulted in the identification of two novel *in vivo* modifications: lysine propionylation and butyrylation in histone H4, p53, p300 and CBP. This work will be presented in Chapter 3 [42,43].

1.5.2 Evaluation of false positive rate for peptide identifications

A popular approach to evaluate the false positive rate of peptide identification in proteomic analysis is the application of statistical analysis. Different statistical models have been proposed to evaluate the false positive rate. The most widely accepted approaches are the target-decoy database strategy and an empirical Bayesian model [39]. The former strategy assumes that the false positive rate remains the same for forward and reversed (or randomized) protein databases. Therefore, the false positive rate of the peptide identification above a cutoff threshold can be estimated based on the number of false identifications with scores above the threshold divided by the total number of identifications scoring above the threshold. The latter strategy applies an empirical Bayesian model that arbitrarily designates incorrect and correct peptide identification distributions based on large-scale data analysis with specific forms of discriminate scores. Calculation of the false positive rate is based on the probabilities from the distributions and the empirical Bayesian model.

Both strategies have advantages and disadvantages and they seem to complement each other in certain ways. The advantage of target-decoy database searching is its simple execution and independence of statistical simulation, while the disadvantage is that it is incapable of removing false positives that are sequence homologues to the true positives. The advantage of the empirical Bayesian model is its direct statistical simulation of database search results without additional data analysis, while the disadvantage is the dependence on large-scale data collection for the statistical simulation and the parametric assumption on the discriminate score function.

To address these problems of statistics-based strategies for false positive rate evaluation, we have developed manual verification methods to remove false positive sequence alignments based on the principle that a positive peptide identification should adequately explain all the major peaks in the MS/MS spectrum. Our approach eliminates the need for statistical simulation and is able to eliminate false positives that are sequence homologues to the true positives. These strategies are discussed in Chapter 2 to 4 and are incorporated in our new software tool named PTMap, for unrestrictive PTM identification (Chapter 4) [44].

1.5.3 Restrictive vs. unrestrictive sequence alignment

Restrictive sequence alignment has been widely applied in all algorithms to identify PTMs. The advantages of this strategy are that it generates modified peptide sequences based only on user-specified PTMs prior to sequence alignment and it significantly saves computation time and space. However, this strategy is limited to the PTM types specified by users. Unrestrictive sequence alignment, on the other hand, refers to analysis where the modification types are no longer restricted by the user and instead are identified on the fly by calculating the difference between the mass of the precursor ion of an experimental MS/MS spectrum and the mass of a theoretical peptide. The advantage of unrestrictive alignment is its capability to identify all types of modifications, expected or unexpected. However, it suffers from a high false positive rate and ambiguous PTM identification results. We have developed a new software tool called PTMap for accurate unrestrictive sequence alignment [44]. Unlike other unrestrictive sequence alignment software, PTMap implements a series of novel strategies to improve the accuracy, sensitivity and speed of the analysis. This work will be presented in Chapter 4 and Chapter 5.

1.5.3 Combinatorial PTMs

Combinatorial PTMs play critical roles in biology. For example, the “histone code” forms a unique modification pattern on the N-terminal of histone tails, allowing specific gene expression activity [45]. Similar to the “histone code”, “protein language” has been hypothesized that may involve in regulating protein function or cellular pathways [40]. Traditional sequence alignment softwares rely on restrictive PTM analysis and are not capable of identifying co-existing PTMs that are unexpected. We have developed PTMap software that is capable of unrestrictive identification of co-existing PTMs that are either known or unknown with high accuracy and sensitivity. This work will also be presented in Chapters 4 and 5.

CHAPTER 2

IMPROVING ACCURACY OF PROTEIN IDENTIFICATION BY DEVELOPING NEW MANUAL VERIFICATION METHODS

2.1 Summary

Protein identifications are typically carried out by automated searching of a sequence database with tandem mass spectra of peptides. When these spectra contain limited information, automated searches may lead to incorrect peptide identifications. It is therefore necessary to validate the identifications by careful manual inspection of the mass spectra. Not only is this task time-consuming, but the reliability of the validation varies with the experience of the analyst. To address this issue, we developed a systematic approach to evaluating peptide identifications made by automated search algorithms [41]. The method is based on the principle that the candidate peptide sequence should adequately explain the observed fragment ions. To evaluate our method, we studied tandem mass spectra obtained from tryptic digests of *E. coli* and HeLa cells. Candidate peptides were identified with the statistics-based automated search engine Mascot and subjected to the manual validation method. The method found correct peptide identifications that were given low Mascot scores (*e.g.*, 20-25) and incorrect peptide identifications that were given high Mascot scores (*e.g.*, 40-50). The method comprehensively detected false results from searches designed to produce incorrect identifications. Comparison of the tandem mass spectra of synthetic candidate peptides to the spectra obtained from the complex peptide mixtures confirmed the accuracy of the evaluation method.

In this chapter, I will also present studies on the common false positive identifications from automatic sequence alignment software [46]. These false positives are caused by: 1) enzymatic digestion at abnormal sites; 2) misinterpretation of charge states; 3) misinterpretation

of protein modifications; 4) incorrect assignment of protein modification sites; and 5) incorrect use of isotopic peaks. I will present examples, clearly identified as false positives by manual inspection, that nevertheless were assigned high scores by a representative sequence alignment algorithm. A common feature of the false positives is the presence of unmatched peaks in the MS/MS spectra. The studies therefore highlight the importance of using unmatched peaks to remove false positives and offer direction to aid development of better sequence alignment algorithms for protein identification.

2.2 Introduction

Tandem mass spectrometry (MS/MS) is the method of choice for identifying and quantifying proteins, largely due to its unparalleled sensitivity and the speed at which fragment mass fingerprints can be generated [47]. In a typical experiment, a proteolytic digest of interest is subjected to LC/MS/MS analysis to generate MS/MS spectra of individual peptides. The resulting MS/MS data are used in an automated search of a protein sequence database to find the peptide that most closely matches each observed spectrum. During the sequence alignment, the experimentally generated MS/MS spectrum is compared to the theoretical MS/MS spectrum of each peptide in the database and a score, representing the degree of correlation, is calculated for each peptide. Several algorithms have been developed for protein sequence alignment and are currently in widespread use, including SEQUEST [32], PepSea [48], Mascot [33], Sonar [49], ProbID [50], Popitam [51], and Tandem [52]. In addition to the 20 ribosomally encoded amino acids, information about protein modification and isotopic labeling (*e.g.*, ICAT or I-DIRT/SILAC/AACT) can be included into the database search for protein quantification and protein modification sites mapping [34,53-58].

A major problem associated with these automated search algorithms is the appearance of false positive hits caused by random matching between the experimental and theoretical data [49,59-64]. To reduce the number of false positives, different statistical strategies have been developed [62]. Unfortunately, the reliability of these strategies has not been critically evaluated, as could be done, for example, by testing them with highly stringent manual verification, or with MS/MS of synthetic peptides, the gold standard for confirming peptide identification. Accordingly, despite efforts to reduce their incidence, false positives remain a concern in shotgun proteomics. The problem is more serious when non-restrictive sequence alignment is carried out to identify all possible modifications in a substrate protein.

We argue that a true peptide identification should explain all major peaks in the MS/MS spectrum. Based on this rationale, we developed systematic manual verification rules to remove false positives in our work [41]. A common feature of false positives is the presence of unmatched peaks in the MS/MS spectra. During the course of our routine work of manually verifying protein identifications obtained from protein sequence database searches, we have encountered several recurring types of false positives. In this chapter, I will also report five types of false positives that cannot be easily eliminated by statistical methods or evaluated by reversing or scrambling the sequence database, as their sequences share with the true peptide identifications. Our case studies provide insights into false positives of protein identification, highlight the importance of careful inspection of MS/MS spectra to ensure accuracy of peptide identification, and offer direction for development of better methods for removing false positives. Our results also suggest that emphasis should be placed on the unmatched peaks in MS/MS spectra to identify the false positives during protein sequence database searching.

2.3 Results

2.3.1 Methods for evaluation of peptide identification

We have established three rules to evaluate protein identifications made by automated protein sequence database searches as listed in Table 1. The rules are based on our accumulated experience manually analyzing MS/MS spectra and on the principle that a correct result should explain all the major mass spectrometric peaks in the MS/MS spectrum, except peaks resulting from electronic sparks that occasionally occur during data acquisition. The rules were also evaluated with reversed sequence database and cross-species sequences. We consider those peptide identifications that can meet the rules as “Correct” identification. In these rules, isotopically resolved peaks were emphasized because a single peak could come from an electronic spark or chemical noise. Single peaks are less likely to be relevant to peptide fragments. Also, noise peaks are less abundant in the high mass region than in lower mass regions. Therefore, peptide fragments with an m/z ratio higher than the m/z ratio of the doubly charged parent ion must be explained by the peptide sequence. Otherwise, the peptide identification should be considered an incorrect identification. The criteria for low mass peaks (below the doubly charged parent m/z ratio) are less stringent because more significant noise peaks exist in the low mass region.

For singly charged peptides, MS/MS spectra usually have more unpredictable noise peaks with high intensity. Therefore, rules for manual evaluation of multiply charged peptides cannot normally be applied to singly charged ions. For this reason, we developed Rule II (Table 1) for evaluation of the identities of singly charged peptides.

2.3.2 Evaluation of the manual verification rules with cross species database search

To test whether the Rules for manual evaluation described above can exhaustively identify incorrect peptide identifications, we used the Mascot software to search human protein sequences in the NCBI-nr database with MS/MS spectra of tryptic peptides derived from *E. coli* proteins. We reasoned that all human peptide sequences identified using MS/MS spectra of *E. coli* tryptic peptides should be incorrect identifications, unless the peptide sequences are shared between human and *E. coli*.

This cross-species search led to identification of 464 peptides with Mascot scores ranging from 20 to 51 (Figure 6A). Evaluation of peptide identification using our Rules established that all the identifications were false. To further confirm the falseness of these identifications, we synthesized the peptide AQVVPPAR, which had the highest Mascot score 51, which is above the identity threshold given by Mascot for this particular peptide. The MS/MS spectrum of the synthetic peptide showed a different fragmentation pattern than that obtained during the LC/MS/MS analysis (Figure 8), showing that this peptide was incorrectly identified.

Table 1 Manual verification rules for evaluating MS/MS spectra from multiply charged peptides and singly charged peptides.

Rule I: Normal rule for validation of peptide candidates of multiply charged ions

1. Only *y*-, *b*- or *a*-ions or associated peaks arising due to water or amine loss are considered as daughter ions of a parent peptide. At least 5 isotopically resolved, independent fragment peaks must match theoretical peptide fragments.

2. All isotopically resolved peaks with intensities higher than 5% of the maximum intensity and m/*z* ratios larger than that of the doubly charged parent mass must match theoretical peptide fragments.

3. All isotopically resolved peaks with intensities higher than 20% of the maximum intensity and m/*z* values between one-third of the parent m/*z* ratio and the parent m/*z* ratio must match theoretical peptide fragments.

4. The difference in the mass errors of neighboring fragment peaks that are within 200 Da of each other must be lower than 0.4 Da.

Rule II: Validation of peptide candidates of singly charged ions

1. Mascot score should be equal to or above the identity score threshold of the peptide.

2. For peptides ended with arginine or lysine residue, both the *b*-ion and the *y*-ion series should confirm at least 3 consecutive amino acids in the peptide sequence.

3. For C-terminal peptides without C-terminal arginine or lysine residue, either the *b*-ion or the *y*-ion series should confirm at least 3 consecutive amino acids in the peptide sequence.

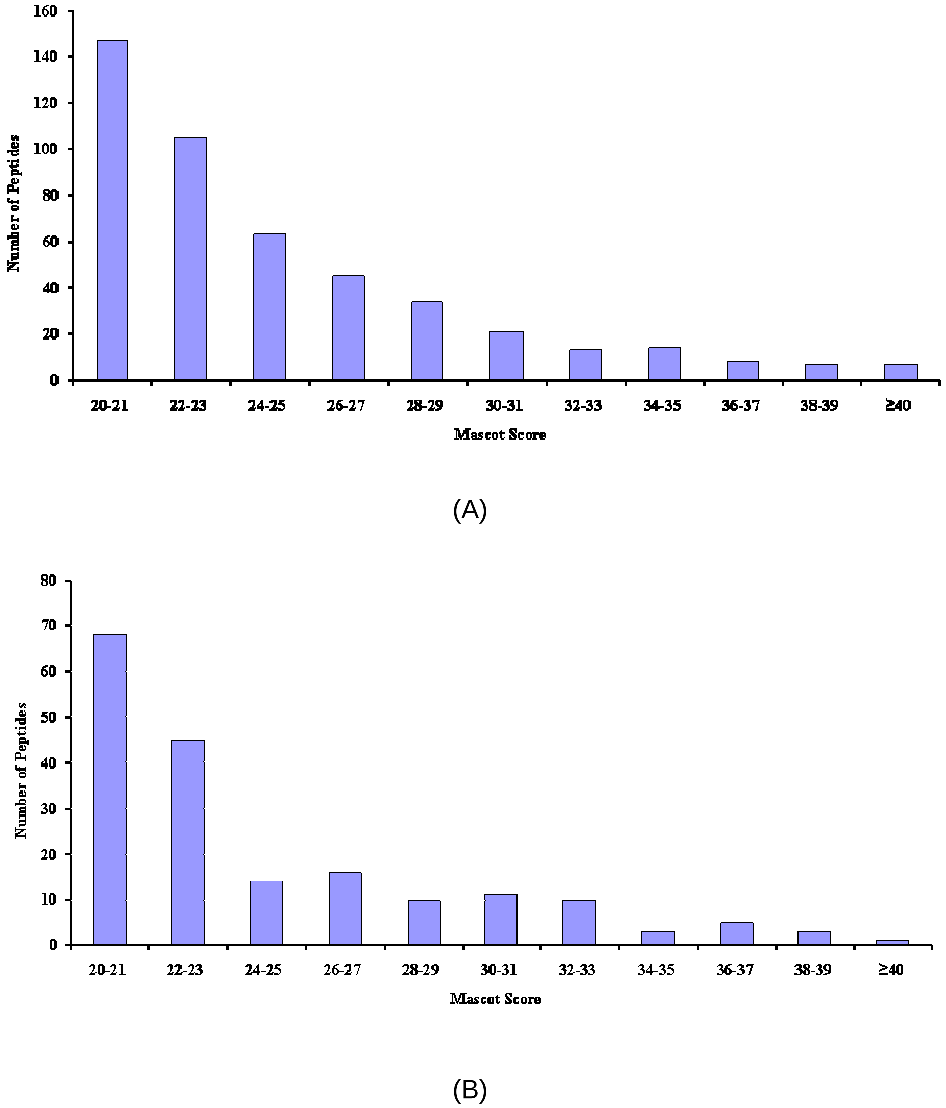
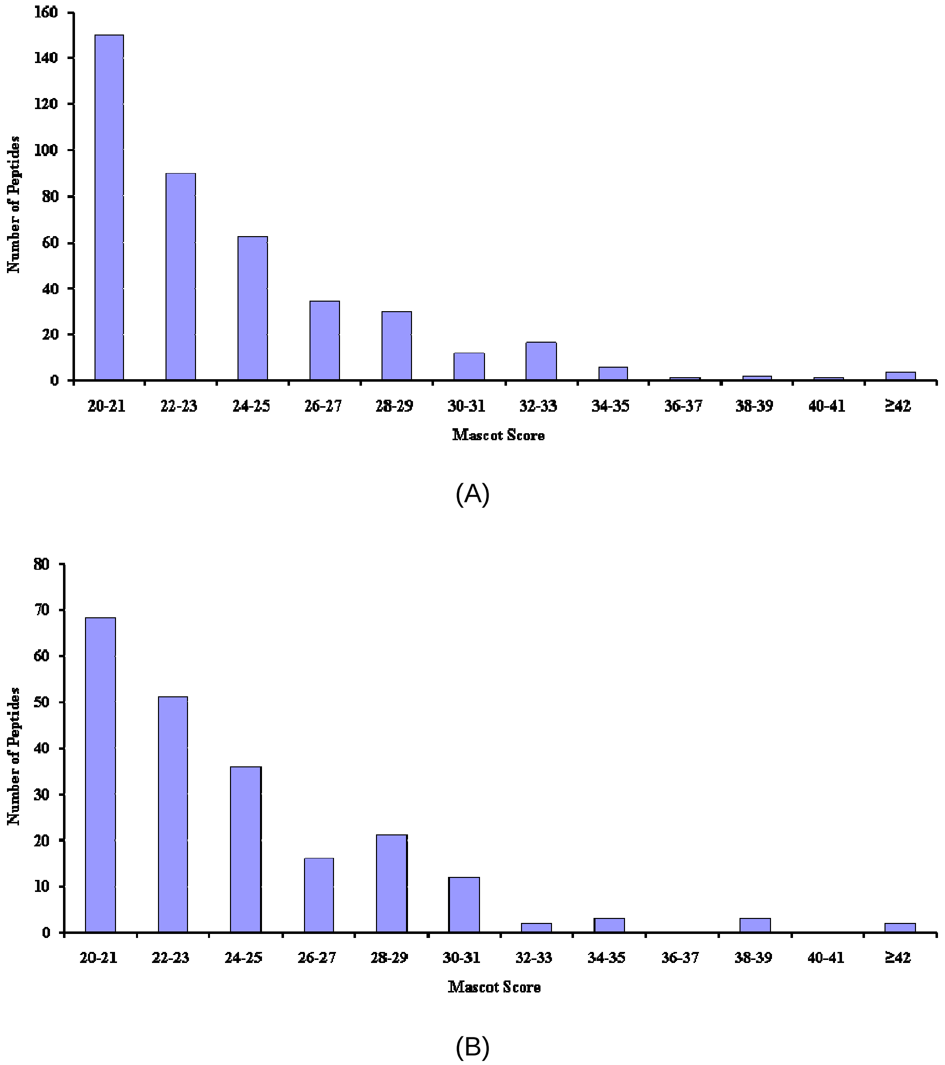


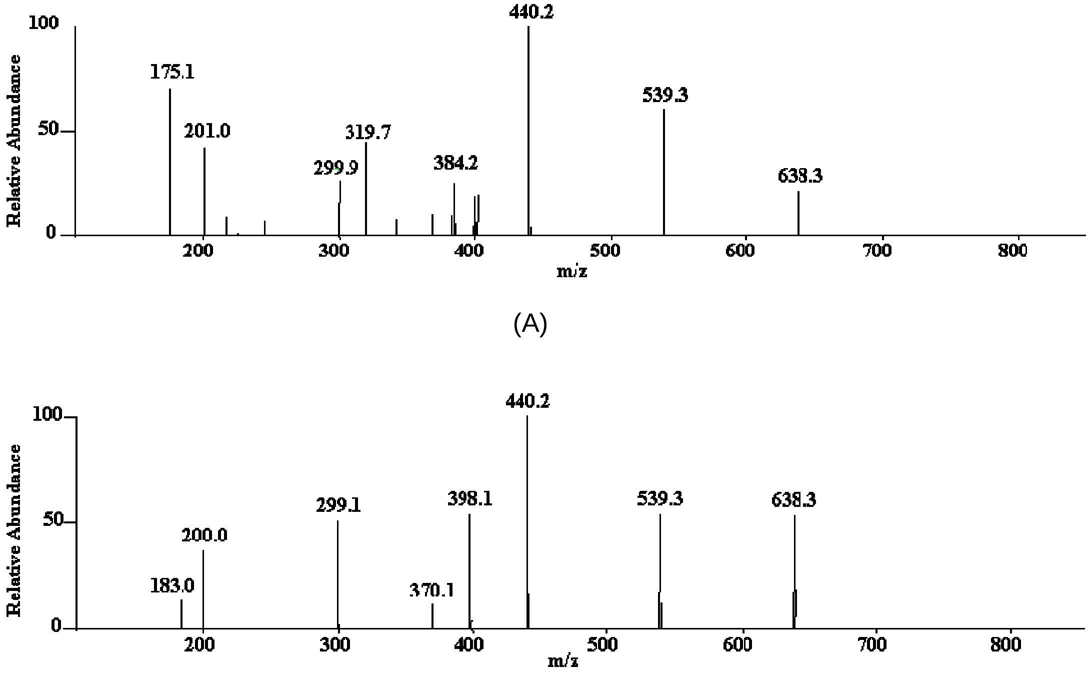
Figure 6 Distribution of Mascot scores for incorrect peptide identifications from cross-species database search. (A) the human sequence database was searched with MS/MS spectra of *E. coli* peptides; (B) the *E. coli* sequence database was searched with MS/MS spectra of peptides derived from HeLa cells.



(A)

(B)

Figure 7 Distribution of Mascot scores for incorrect peptide identifications from reversed protein database searches. (A) the reversed human sequence database was searched with MS/MS spectra of peptides derived from HeLa cells; and (B) the reversed *E. coli* sequence database was searched with MS/MS spectra of *E. coli* peptides.

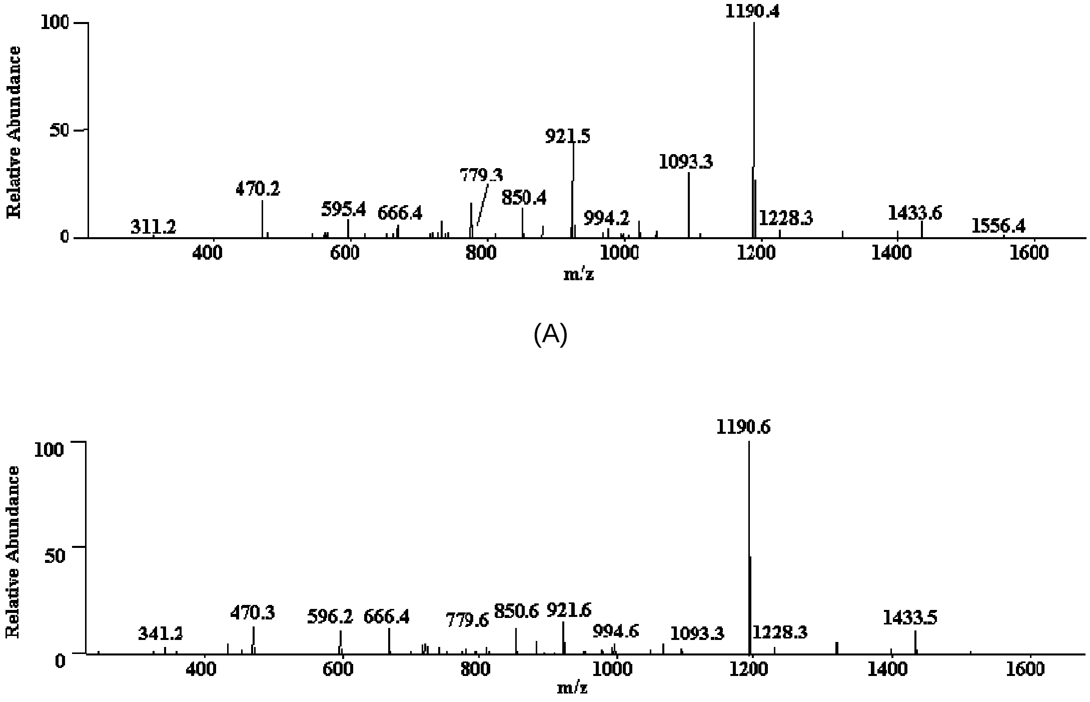


(A)

(B)

Figure 8 Experimental verification of incorrect peptide identification from a cross-species search of the human sequence database. (A) MS/MS spectrum of a tryptic peptide from *E. coli* that resulted in identification of peptide AQVVPPAR from the human sequence database, with MASCOT score 51. (B) MS/MS spectrum of synthetic peptide AQVVPPAR.

In a parallel experiment, we searched the *E. coli* protein sequence database using the MS/MS spectra of tryptic peptides derived from HeLa cell extracts. One hundred eighty-six *E. coli* peptides were identified, with Mascot scores ranging from 20 to 50 (Figure 6B). Evaluation of peptide identification using our approach established that peptide IINEPTAAALAYGLDK from the *E. coli* protein database (Mascot score 50) was the only correct identification. The major peaks of the MS/MS spectrum could be explained by the theoretical mass fingerprint of the peptide. Comparison of the observed MS/MS spectrum with that of a synthetic peptide of the same sequence revealed that the two spectra were almost completely identical, fulfilling Rule II (Figure 9). These results suggest that this sequence is a true peptide identification. A BLAST search using the sequence of this peptide showed that a homologous peptide is present in the 70-kDa human heat shock protein 5, with an isobaric amino acid substitution of leucine for isoleucine. We synthesized a peptide representing identification from this experiment with a Mascot score of 38. Tandem mass spectrometry of the synthetic peptide confirmed that they were incorrectly identified.



(A)

(B)

Figure 9 Experimental verification of correct peptide identification from a cross-species search of the *E. coli* sequence database. (A) MS/MS spectrum of a tryptic peptide from HeLa cells that resulted in identification of peptide IINEPTAAALAYGLDK from the *E. coli* sequence database. (B) MS/MS spectrum of synthetic peptide IINEPTAAALAYGLDK.

2.3.3 Evaluation of the manual verification rules with reversed sequence database search

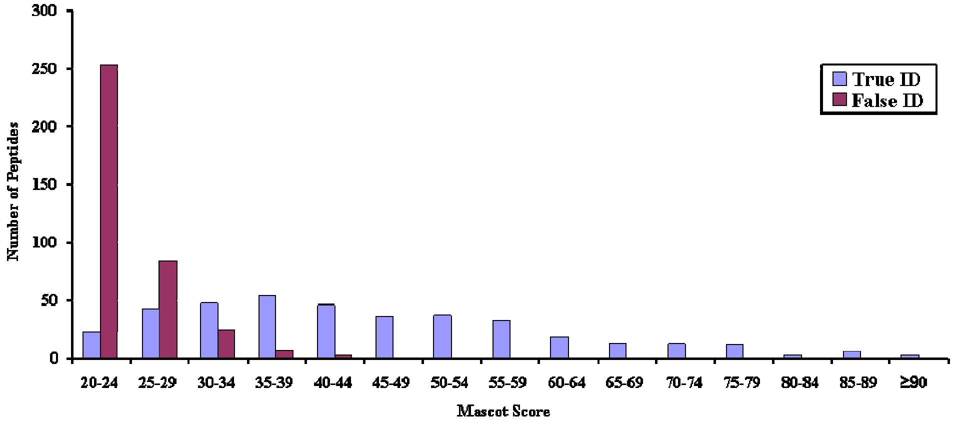
Next, we used the MS/MS spectra of tryptic peptides derived from HeLa cell extracts to search a sequence database composed of all human proteins with their sequences in the reverse order (i.e., from C-terminus to N-terminus). The search led to 408 peptide identifications with MASCOT scores ranging from 20 to 44 (Figure 7A). Similarly, MS/MS spectra obtained from tryptic peptides of *E. coli* proteins were used to search a reversed *E. coli* protein sequence database, resulting in identification of 214 peptide candidates with MASCOT scores ranging from 20 to 43 (Figure 7B). Manual evaluation with our Rules suggested that all these identifications were incorrect.

2.3.4 Application of manual verification rules:

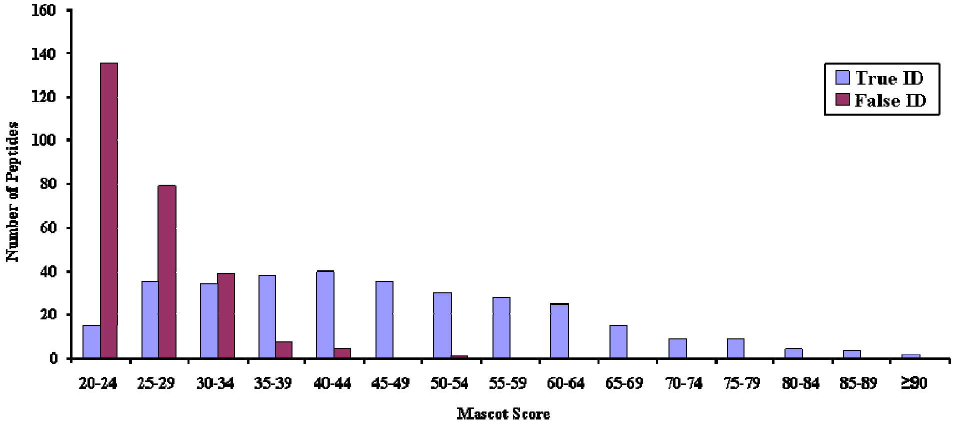
1. Identification of correct peptide identifications with low statistic scores

We used the MS/MS spectra of the tryptic peptides derived from HeLa cells to search the human protein sequence database. The search led to the identification of 745 peptides with Mascot score ranging from 20 to 117 (Figure 10A). Based on our previous experience, few peptide identifications with scores below 20 can be correlated with MS/MS spectra. Therefore, peptides given scores below 20 were not analyzed. Manual evaluation suggested that 376 of the 745 candidates were true identifications while 369 were incorrect (Figure 10A). To test if we had mistakenly considered an incorrect identification to be correct, we synthesized two peptides with Mascot scores of 22 and 24, well below the homolog threshold given by Mascot for these peptides. The MS/MS spectrum of each synthetic peptide contained the same fragment signatures as the spectrum obtained in the HPLC/MS/MS analysis, confirming each of these peptide identifications (Figure 11).

Protein identification and manual evaluation were also carried out for MS/MS spectra of *E. coli* tryptic peptides, searching the *E. coli* sequence database. Five hundred eighty-nine peptide candidates were identified with Mascot scores between 20 and 100. Manual evaluation showed that 322 were true identifications while 267 were incorrect (Figure 10B). MS/MS analysis of two synthetic peptides representing correct identifications with low Mascot scores (20, 27) again confirmed that the identifications were indeed correct.

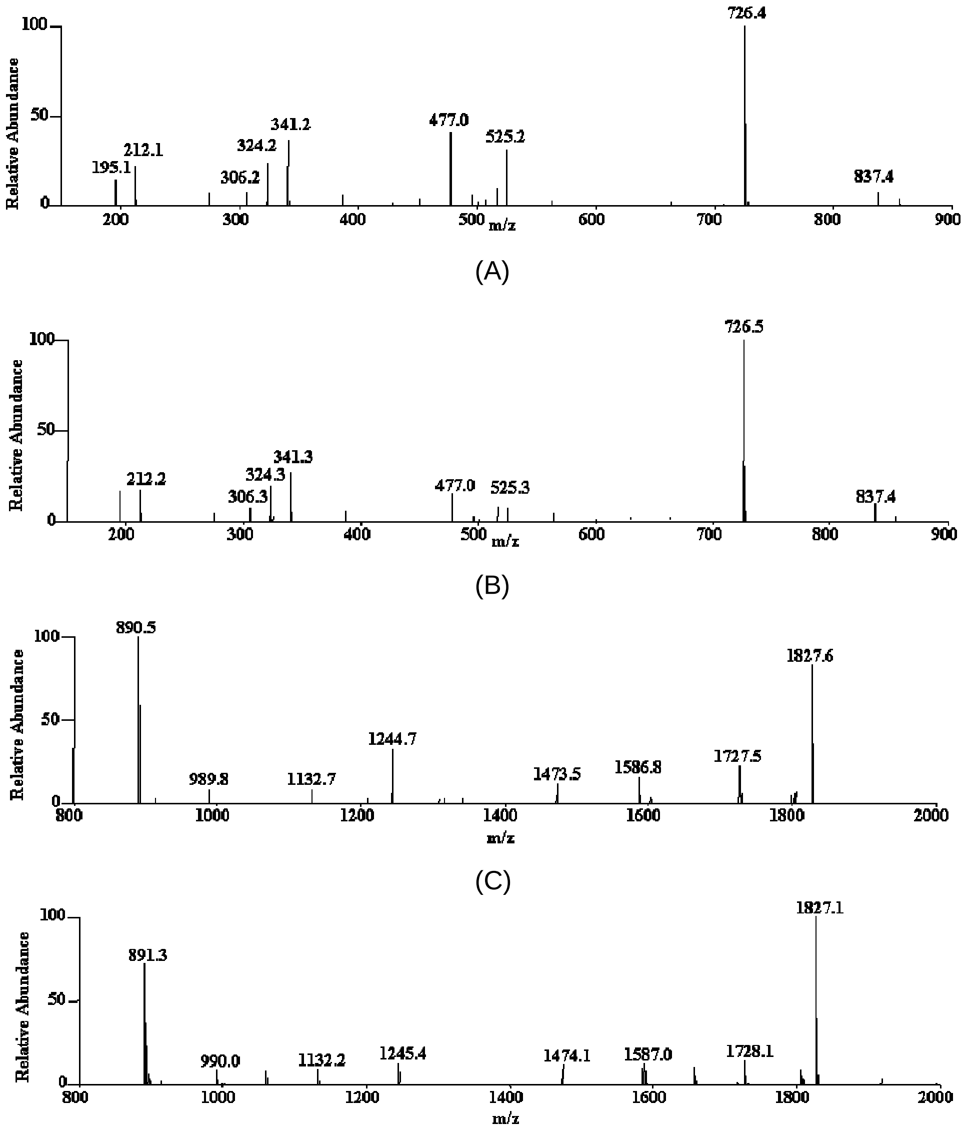


(A)



(B)

Figure 10 Distribution of Mascot scores for true and false peptide identifications of *E. coli* and HeLa cytosolic peptides. MS/MS spectra of *E. coli* and HeLa tryptic peptides were used to search the protein sequence databases of the appropriate species. All peptide identifications with Mascot score of at least 20 were manually verified. The Mascot score distributions of true and false peptide identifications from analysis of (A) HeLa proteins and (B) *E. coli* proteins are shown.



(A)

(B)

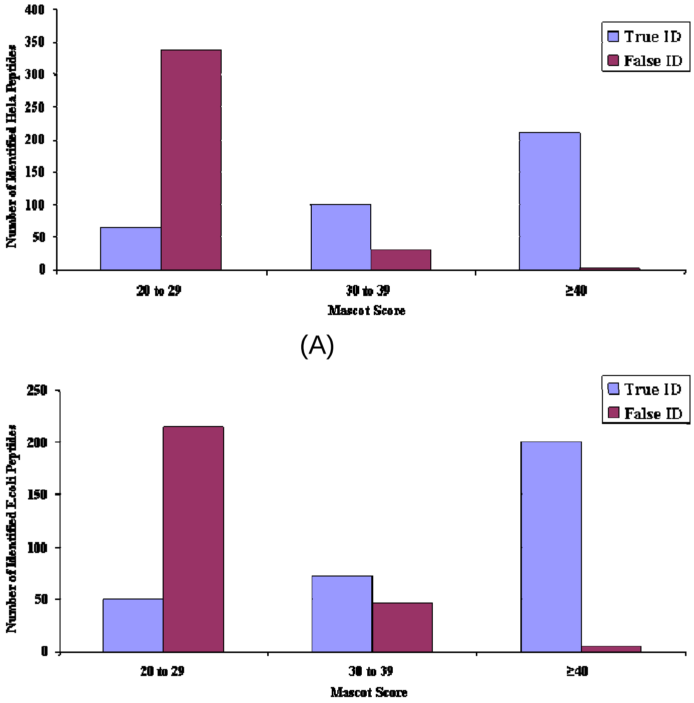
(C)

(D)

Figure 11 Experimental verification of correct peptide identifications with low Mascot scores. Shown are MS/MS spectra of tryptic peptides from HeLa cells that resulted in identification of peptides NPEPELLVR (A) and HSQDLAFLSMLNDIAAVPATAMPFR (C) from the human sequence database, with Mascot scores of 22 and 24, respectively. Also shown are the MS/MS spectra of the corresponding synthetic peptides NPEPELLVR (B) and HSQDLAFLSMLNDIAAVPATAMPFR (D).

Collectively, these experiments demonstrated that (1) correctly identified peptides could have Mascot scores as low as 20 (lower than corresponding homolog score), again indicating that there is no definitive threshold Mascot score for true peptide identifications, and (2) our

Rules were able to evaluate correct peptide identifications with low scores. A Mascot score with a probability of occurring of less than 5% is considered a significant match by default in the Mascot software (http://www.matrixscience.com/help/scoring_help.html). In our analyses, this value corresponded to a Mascot score of 36 for the *E. coli* database search and 42 for the human database search. Given this information, a Mascot score of 40 might be considered a reasonable threshold for correct identification in our experiments. We found that ~99% of human peptide identifications and ~98% of *E. coli* peptide identifications with Mascot scores above 40 would be correct peptide identifications. However, 44% of human peptide identifications and 38% of *E. coli* peptide identifications with scores from 20-39 were also correct identifications (Figure 12, Figure 13), meaning that applying a strict cutoff value of 40 would result in discarding a considerable number of correct identifications.



(A)

(B)

Figure 12 Distribution of Mascot scores for correct and incorrect peptide identifications. The number of true and false peptide identifications in three Mascot score ranges are shown for the nano-HPLC/MS/MS analyses of HeLa peptides (A) and *E. coli* peptides (B).

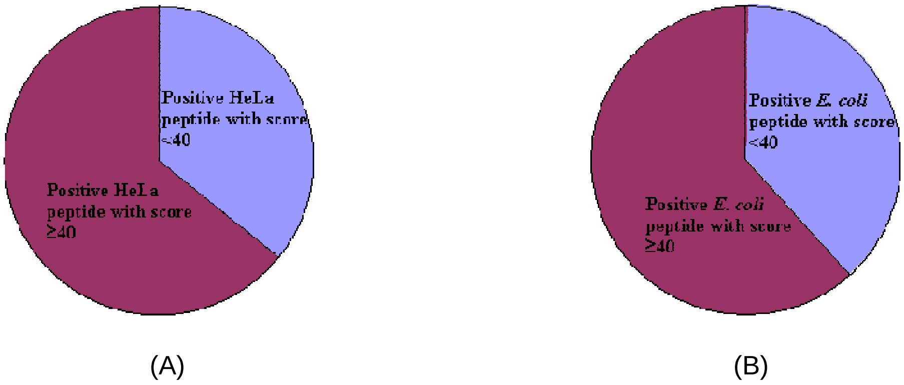


Figure 13 Distributions of positive peptide identifications in the HeLa cell (A) and *E. coli* (B) analyses if a Mascot score of 40 is selected as the threshold for correct identification.

2. Identification and classification of common false positives with high statistical scores

We routinely identify false positives, using manual verification analysis, that were given high statistical scores by a statistics-based protein-sequence alignment algorithm such as Mascot. Based on our accumulated experience, we identified five major categories of commonly observed false positives identified by the Mascot algorithm with high statistical scores.

The first category - enzymatic digestion at unexpected sites. In shotgun proteomics, a protein mixture of interest is usually digested with trypsin. Preparations of trypsin will not only have canonical tryptic activity, cleaving a protein at the C-terminal side of lysine and arginine residues, but will also have weak chymotryptic activity, resulting in cleavage of the peptide bond C-terminal to aromatic or hydrophobic residues such as phenylalanine, tryptophan, tyrosine, leucine and methionine. The chymotryptic activity of trypsin can increase during the course of the incubation, as the enzyme is auto-digested. Accordingly, digestion with trypsin usually generates chymotryptic peptides, the abundance of which depends on trypsin quality, amount of trypsin used, trypsin-to-substrate ratio, and digestion time. When chymotrypsin is not included as a digestion enzyme during a protein sequence database search, the algorithm can assign a high statistical score to a tryptic peptide from the database matched with the MS/MS spectrum of a peptide that arose because of chymotryptic digestion at one or both ends. As an example, protein sequence database searching using the MS/MS spectrum in Figure 14A identified the triply charged tryptic peptide VL^{Ox}MLPTLQNDPPSLETGVQDK with Mascot score 36. Careful inspection of the spectrum discovered three problems with the peptide identification. First, the b series of fragment ions are completely missing, which does not usually happen for a tryptic peptide with a lysine residue at the C-terminus. Second, one of the major ions (at m/z 340) could not be assigned. Third, a triply charged peptide was assigned, even though only one basic amino acid residue (K) is present in the peptide sequence. When chymotryptic digestion was considered, the molecular weight matched a doubly charged peptide, QNDPPSLETGVQDK. The peptide could explain all fragment ions in the spectrum (Figure 14B). Accordingly, the second peptide should be considered the correct identification for the MS/MS spectrum.

In Figure 14, the labels b and y designate the N- and C-terminal fragment ions, respectively, of the peptide produced by breakage at the peptide bond in the mass spectrometer. The label a designates N-terminal fragments produced by breakage at the backbone C-C bond adjacent to the peptide bond. The subscripted number in each label represents the number of N- or C-terminal residues present in the peptide fragment. The label Δ designates b, y or a ions with loss of water, ammonia or both. All unassigned peaks with relative intensity more than 5% of the base peak are labeled with a question mark. The same nomenclature system is used for all the other figures.

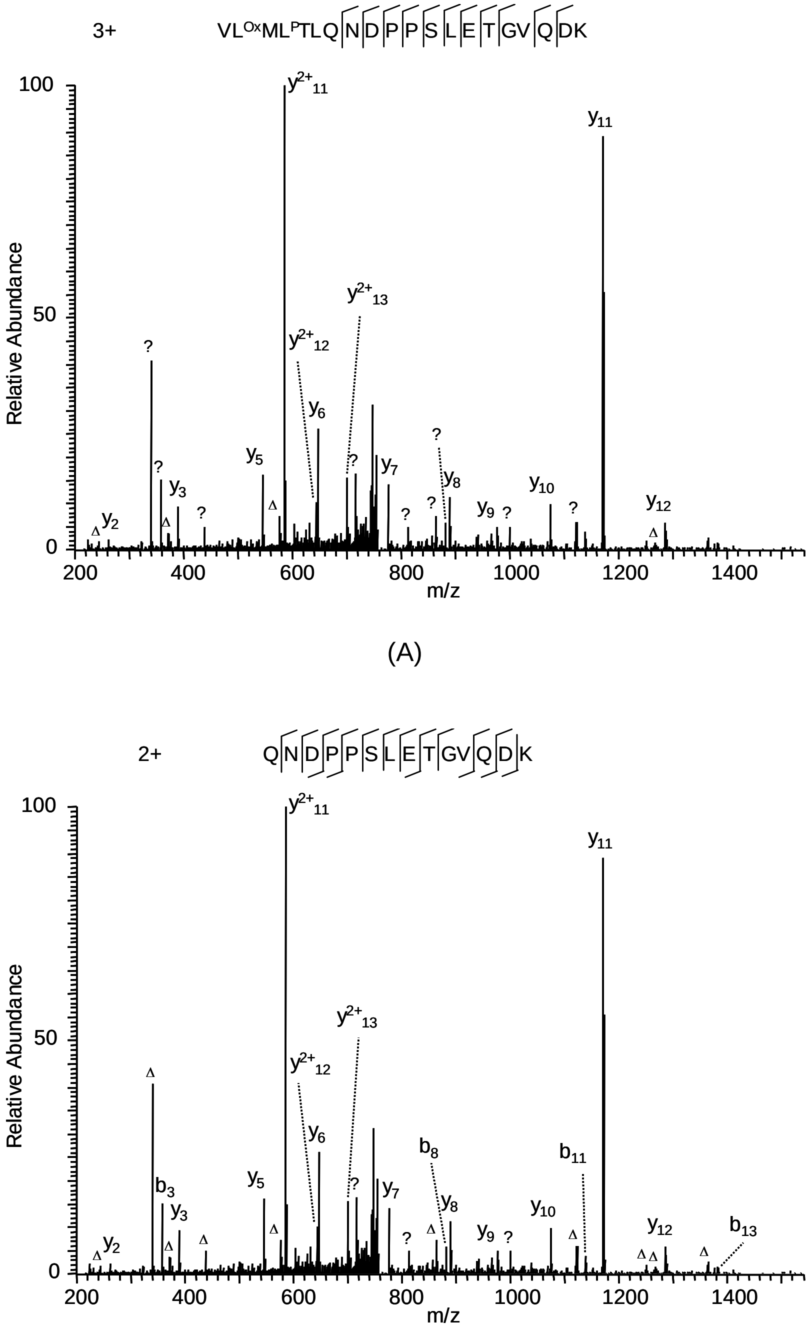
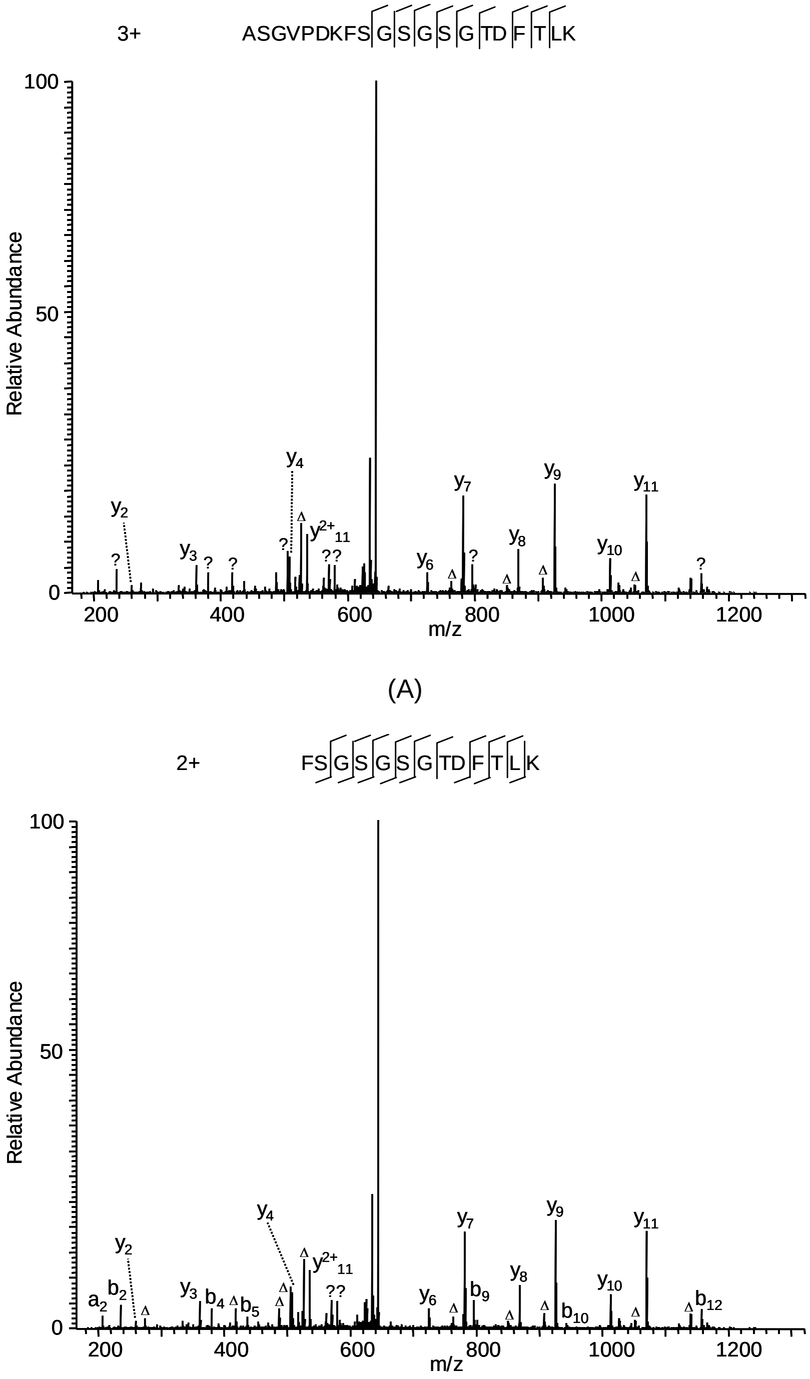


Figure 14 False peptide identification caused by enzymatic digestion at an abnormal site. (A) Assignment of an MS/MS spectrum with a triply charged peptide, VL^{Ox}MLPTLQNDPPSLETGVQDK, which was identified by Mascot with Mascot score of 36. (B) Assignment of the same MS/MS spectrum with a doubly charged peptide, QNDPPSLETGVQDK, identified by manual inspection.

The second category - assignment of incorrect charge states for peptide ions. The second common type of false positive in peptide identification is assigning the wrong charge state to peptide ions. Low-resolution mass spectrometers generate MS and MS/MS spectra with low mass accuracy, which sometimes prevents identification of the proper charge state of peptide ions, possibly leading to incorrect peptide identification. Protein sequence alignment of the MS/MS spectrum in Figure 15A led to the identification of a triply charged peptide, ASGVPDKFSGSGSGTDFTLK, with a Mascot score of 39. Careful inspection of the MS/MS spectrum suggested two problems with the peptide identification. First, no b ions are present, and second, several significant peaks in the high mass range (between m/z 560 and 1160) were not assigned. Repetition of the sequence alignment with Mascot after adjustment of the charge state from +3 to +2 led to identification of a doubly charged peptide, FSGSGSGTDFTLK, which can explain all major peaks in the MS/MS spectrum (Figure 15B), and the identification was confirmed by the fragmentation of the synthetic peptide (data not shown). Likewise, a Mascot search assigned a triply charged peptide, ^{P}Y^{P}TLVLTDPDAPSR, to the MS/MS spectrum in Figure 16A. Adjustment of the charge state from +3 to +2 led to identification of the doubly charged peptide VLTDPDAPSR (Figure 16B), which can also be confirmed by synthetic peptide (data not shown).



(A)

(B)

Figure 15 False peptide identification due to misinterpretation of charge states. (A) Assignment of MS/MS spectrum with a triply charged peptide, ASGVPDKFSGSGSGTDFTLK, identified by the Mascot algorithm with a Mascot score of 39. (B) Assignment of the same MS/MS spectrum with a doubly charged peptide, FSGSGSGTDFTLK by the Mascot algorithm with Mascot score of 71.

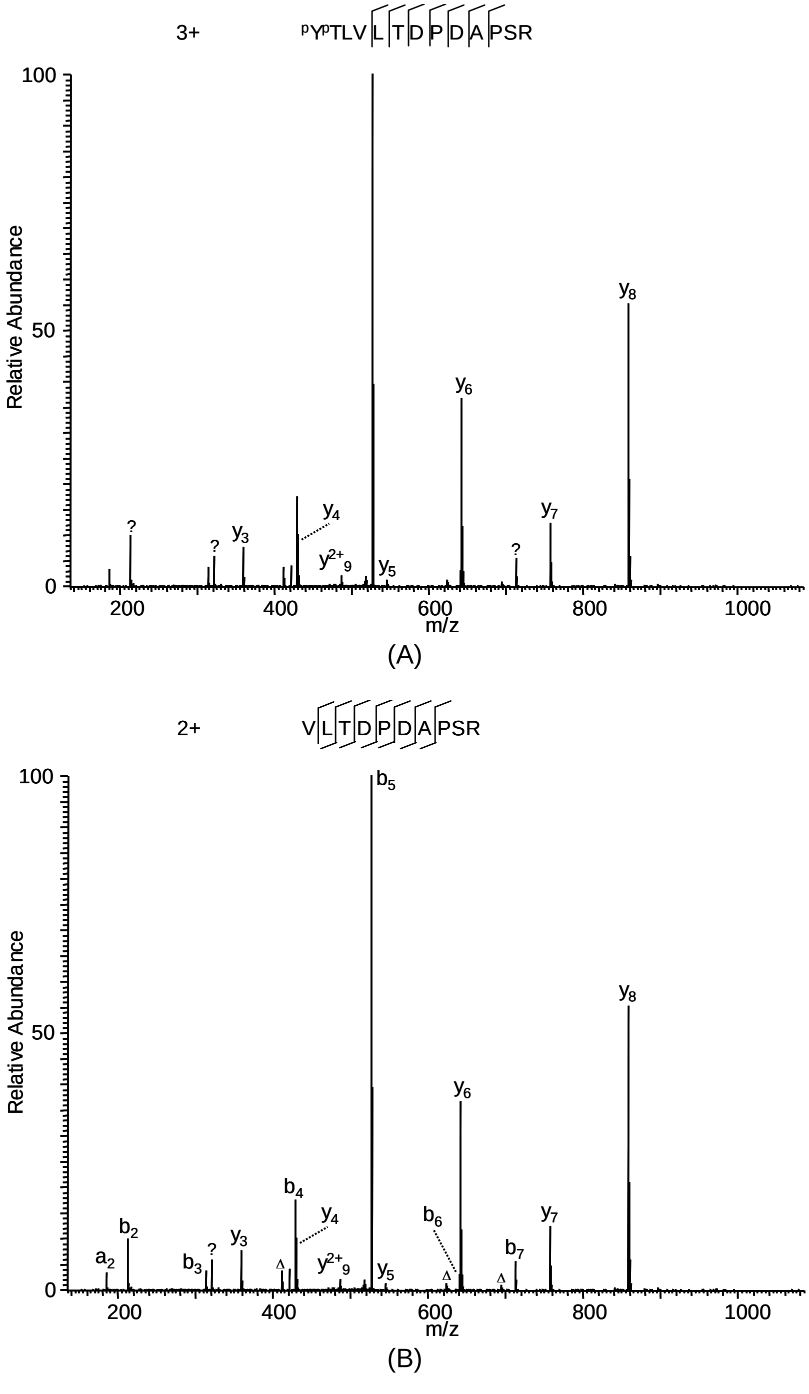
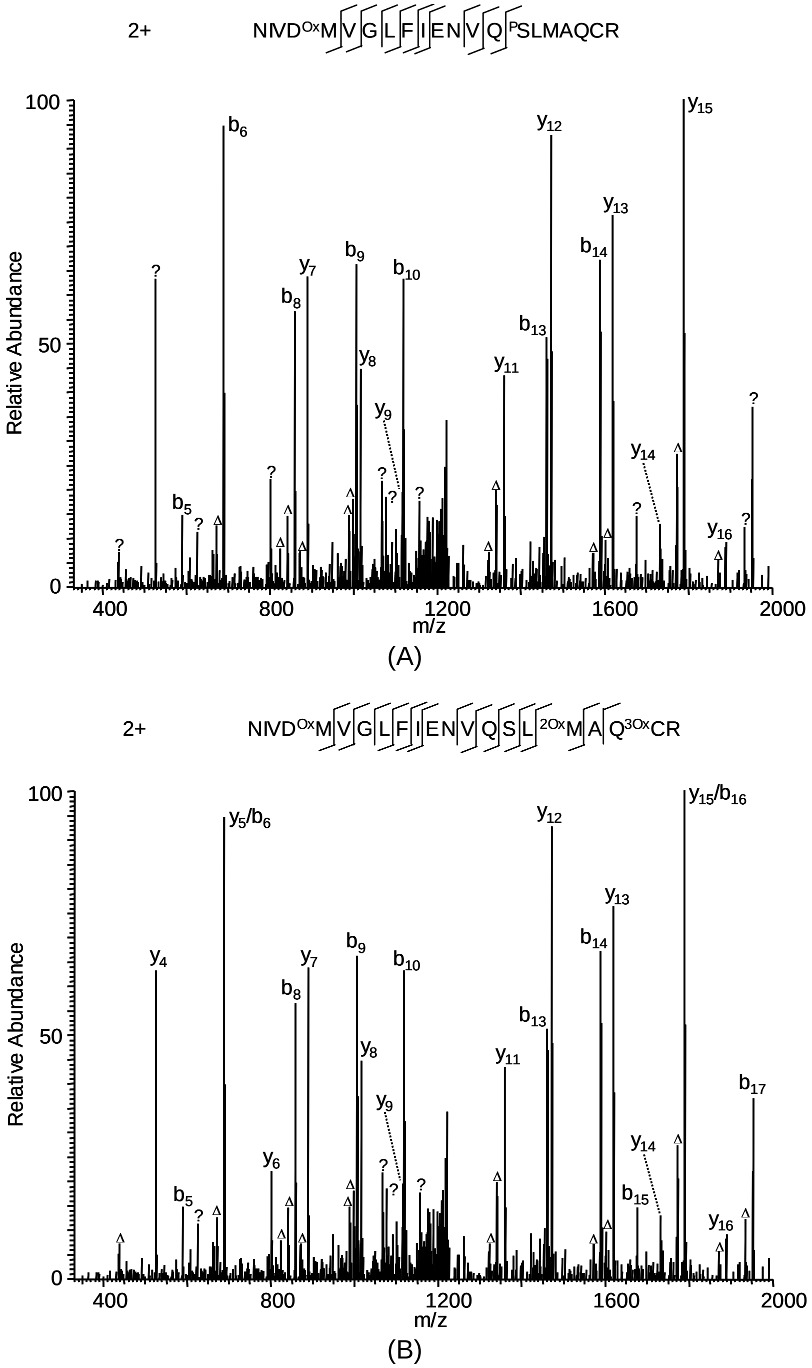


Figure 16 False peptide identification caused by misinterpretation of charge state and modifications. (A) Assignment of an MS/MS spectrum with a triply charged peptide, $^{P}Y^{P}$TLVLTDPDAPSR, by the Mascot algorithm with a Mascot score of 32. (B) Assignment of the same MS/MS spectrum with a doubly charged unmodified VLTDPDAPSR with a Mascot score of 51.

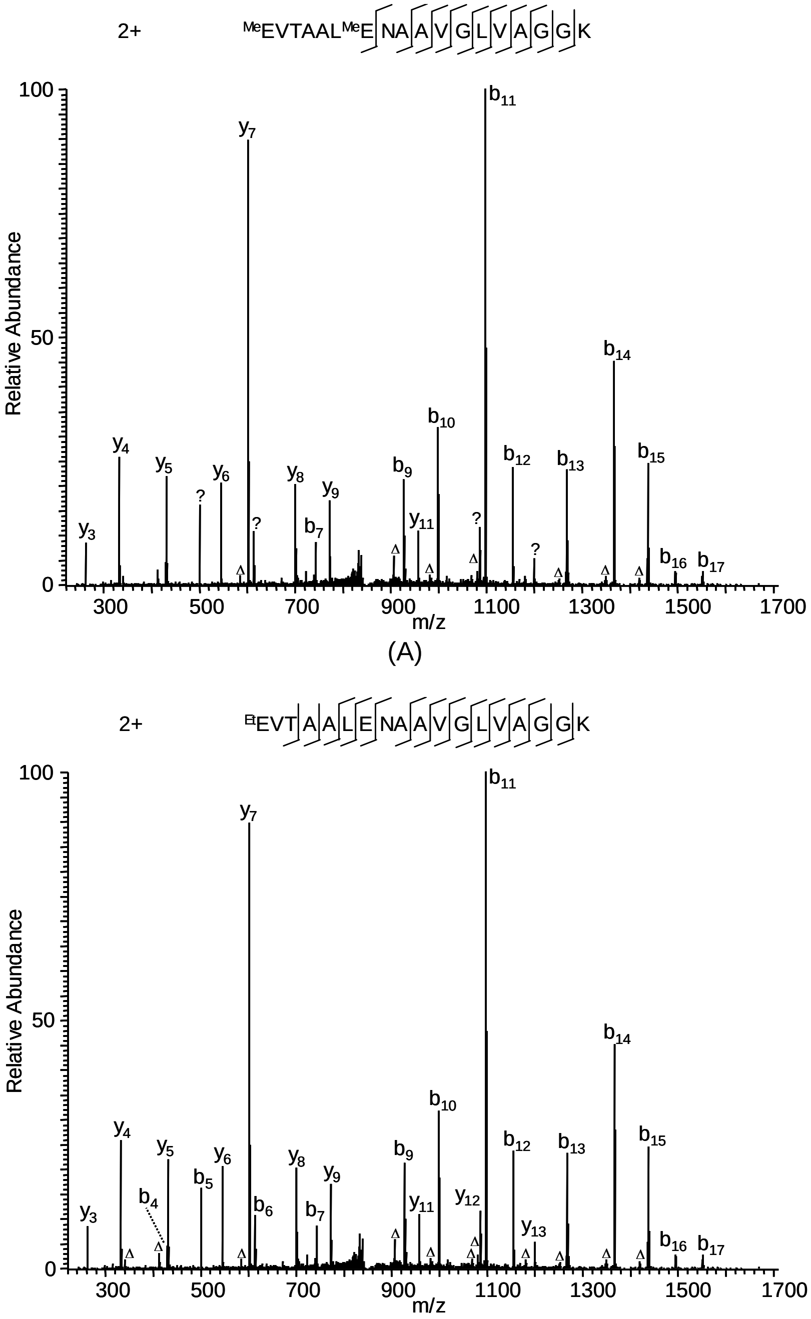
The third category - assignment of false protein modifications. A protein can potentially be modified by more than 300 different types of post-translational modifications, some of which have similar mass shifts [65]. In addition, the mass shift caused by a single protein modification can be similar to the sum of the shifts caused by two or more smaller modifications. As an example, a Mascot search of an MS/MS spectrum identified a doubly charged tryptic peptide, NIVD^{Ox}MVGLFIENVQ^{p}SLMAQCR (Figure 17A), with a Mascot score of 35. Nevertheless, the peptide sequence cannot explain two major peaks (m/z 525 and 1952.8) and several minor ones in the MS/MS spectrum. Careful manual inspection and some calculations led to identification of a doubly charged peptide, NIVD^{Ox}MVGLFIENVQSL^{2Ox}MAQ^{3Ox}CR (Figure 17B). The false alignment was caused by the two unexpected modifications of double oxidation at methionine and sulfation at cysteine. The two oxygen atoms added to Met-17 and the three oxygen atoms added to Cys-20 add a total of 80 units to the peptide's mass, the same value as the mass shift of a phosphate group. This example demonstrates that a large number of matched daughter ions (28 ions in Figure 17A) does not necessarily indicate a true peptide identification if unmatched peaks with high intensities exist in the spectrum. In another example, a Mascot search using the MS/MS spectrum in Figure 18A identified a doubly charged peptide, ^{Me}EVTAAL^{Me}ENAAVGLVAGGK, when D/E protein methylation was specified. Though most of the daughter ions in the spectrum could be explained by the peptide sequence, a series of minor peaks remained unassigned. In addition, no fragment ions (either b or y ions) related to the sequence were found between the modified residues (Glu-1 and Glu-7). Careful inspection of the mass spectrum suggested an unexpected modification, ethylation at the side chain of the first Glu residue. The new peptide sequence, ^{Et}EVTAALENAAVGLVAGGK, explained almost all the peaks in the MS/MS spectrum (Figure 18B). In addition, a series of b ions (b_3 to b_6) emerged in the N-terminal region of the peptide, and two more y ions (y_{12}, y_{13}) were assigned to the first six amino acid residues. Therefore, the correct peptide identification is ^{Et}EVTAALENAAVGLVAGGK. Ethylation of the Glu side chain likely occurred during gel staining, which involved incubation in the presence of ethanol.



(A)

(B)

Figure 17 False peptide identification caused by false protein modification assignment. (A) Assignment of an MS/MS spectrum with a doubly charged peptide, NIVD^{Ox}MVGLFIENVQ^{P}SLMAQCR, identified by Mascot with a score of 35. (B) Assignment of the same MS/MS spectrum with a doubly charged peptide, NIVD^{Ox}MVGLFIENVQSL^{2Ox}MAQ^{3Ox}CR, identified by manual inspection.



(A)

(B)

Figure 18 False peptide identification caused by false modification assignment and unexpected modification. (A) Assignment of an MS/MS spectrum with a doubly charged peptide, ^{Me}EVTAAL^{Me}ENAAVGLVAGGK, identified by Mascot with a score of 57. (B) Assignment of the same MS/MS spectrum with a doubly charged peptide, ^{Et}EVTAALENAAVGLVAGGK, identified by manual inspection.

The fourth category - assignment of ambiguous protein modification sites. Precise mapping of the sites of modification within a modified peptide can be challenging, because peptides that differ only by the modification site give highly similar theoretical fragmentation patterns and lead to similar statistical scores. Moreover, some types of protein modification can occur on different amino acid side chains. For example, protein methylation can be present at eight of the twenty ribosomally encoded amino acid residues (K, R, D, E, H, D, N, C); together these residues account for almost 50% of the residues in a typical peptide. In one analysis, Mascot identified an MS/MS spectrum as an E-methylated peptide "YPI^{Me}EHGIVTNWDDMEK" from human actin (gi|14250401) (Figure 19A) when D, E methylation were specified as variable modifications. Almost all the major peaks (~90%) can be assigned by the software with the exception of only three peaks. Such high-quality sequence alignment lead to very confident identification with Mascot score of 54. However, after careful examination of the peptide sequence and the MS/MS spectrum, we realized that the MS/MS spectrum cannot exclusively localize the +14 Da mass shift on the E-4 residue raising the possibility that the PTM assignment was false positive due to misassignment of PTM site. Indeed, manual verification suggested that the MS/MS spectrum comes from the peptide isoform "YPIE^{Me}HGIVTNWDDMEK" with methylation on H-5 instead of E-4, which can fully explain all the three unassigned peaks with significant intensity (Figure 19B).

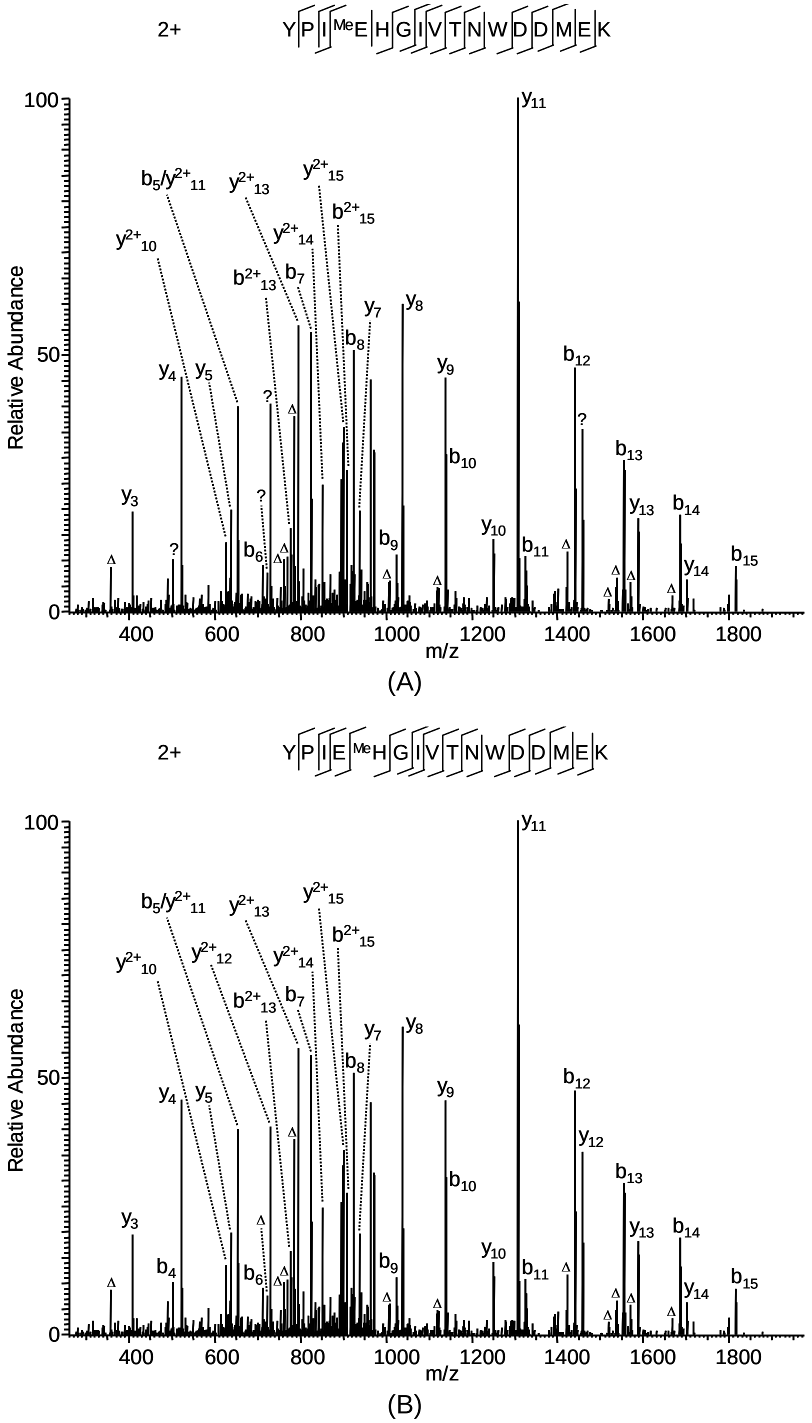
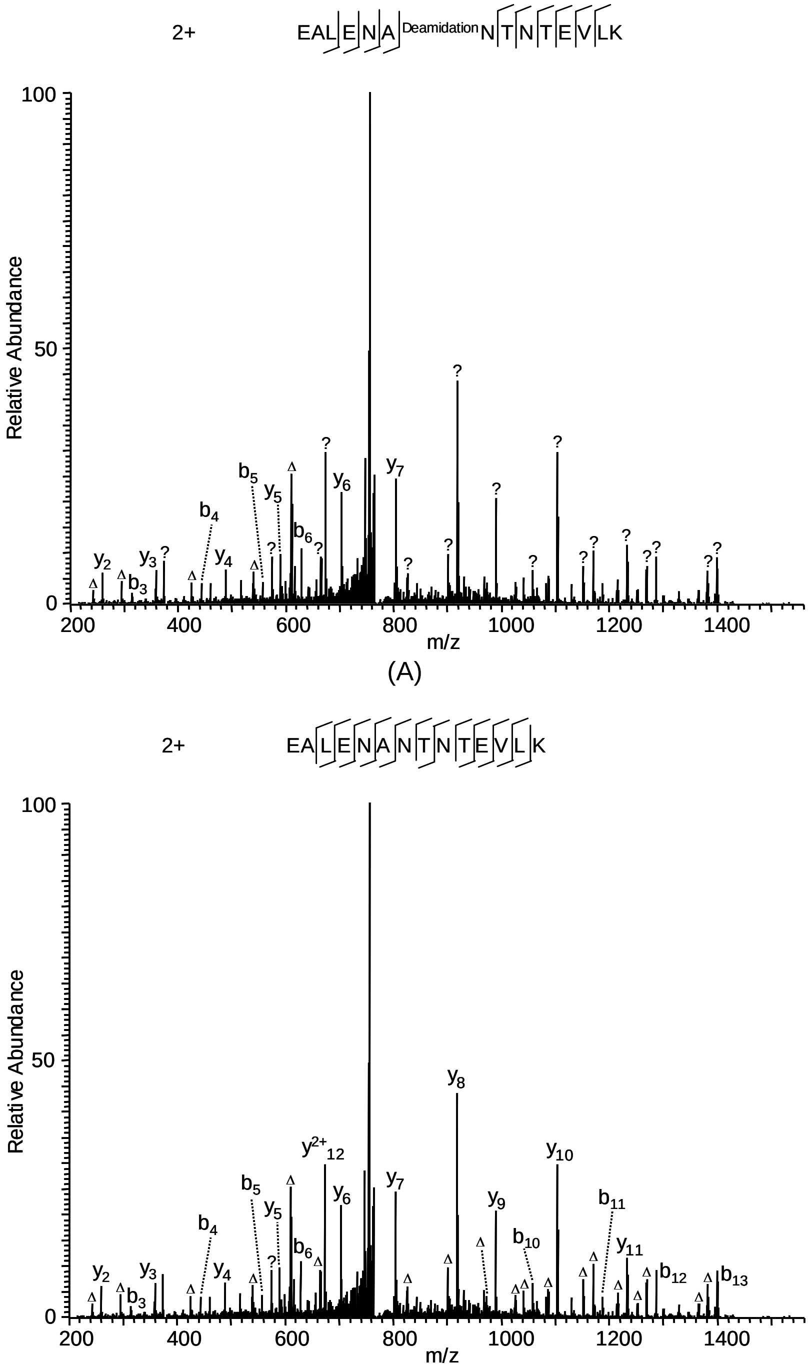


Figure 19 False peptide identification caused by misassignment of modification site. (A) Assignment of an MS/MS spectrum with a doubly charged peptide, YPI^{Me}EHGIVTNWDDMEK, identified by Mascot with a score of 54. (B) Assignment of the same MS/MS spectrum with a doubly charged peptide, YPIE^{Me}HGIVTNWDDMEK, identified by manual inspection.

The fifth category - incorrect use of isotopic peaks. For peptide identification, all protein sequence database search algorithms use monoisotopic peaks, which are one or two Da different from other peaks in the associated isotopic distribution. Unfortunately, some protein modifications only result in a mass shift of one or two Da, which cannot be distinguished from isotopic peaks in low-resolution mass spectrometers. For example, deamidation of asparagine and glutamine are common protein modifications, which result in a one-Da increase in the mass of the residues. In addition, some amino acid pairs differ in mass by only one or two Da. Accordingly, mistaking a peak within the isotopic distribution for the monoisotopic peak can lead to incorrect identification of protein modifications or peptide sequences. As an example, a Mascot sequence alignment using the MS/MS spectrum and allowing deamidation of N and Q residues identified a deaminated peptide, EALENA$^{\mathrm{Deamidation}}$NTNTEVLK (Figure 20A) with a Mascot score of 70. However, manual verification found that the peptide sequence could explain almost none of the peaks in the high mass region, unless the isotopic peaks were used (Figure 20A). This observation suggests that the algorithm incorrectly used higher isotopic peaks instead of monoisotopic peaks during the peptide identification. All the peaks in the MS/MS spectrum can be explained by the unmodified peptide, EALENANTNTEVLK (Figure 20B), and the identification was confirmed by the synthetic peptide (data not shown).



(A)

(B)

Figure 20 False peptide identification caused by use of peaks from the isotopic distribution. (A) Assignment of an MS/MS spectrum with a doubly charged peptide, EALENA$^{\text{Deamidation}}$NTNTEVLK, by Mascot with a score of 70. (B) Assignment of the same MS/MS spectrum with a doubly charged unmodified peptide, EALENANTNTEVLK, identified by manual inspection.

2.4 Discussion

The MASCOT score (or p value) has been used as a key criterion to distinguish between positive and false peptide identification [66-70]. Our studies described in this chapter suggest that there is no clear cutoff Mascot score (or p value) that can be used to distinguish correct peptide identifications from incorrect ones. Therefore, manual analysis is required to evaluate peptide identification. Our case studies suggest that these false positive can have a high statistical score, but cannot be completely eliminated by the popular statistics-based protein sequence alignment algorithms, such as Mascot. In all the cases described in this chapter, the incorrectly identified peptide sequences share significant sequence similarity to the correct peptide sequences. Accordingly, these misidentifications are not random events and their incidences cannot be estimated by the methods commonly used to evaluate false positive rates, such as reversing or scrambling protein sequence databases.

A feature common to all the incorrect peptide assignments is the existence of unmatched peaks with significant intensities. In each example presented here (except the last isotopic case), a significant proportion of fragment ions could be assigned by the false positive peptide (Table 2). Such high numbers of assigned daughter ions usually lead to high scores in the statistics-based protein identification algorithms. Nevertheless, a correctly identified peptide should be able to explain almost all the peaks in the MS/MS spectrum, except in special instances when irregular fragmentations occur. Therefore, we argue that it is more logical to use unmatched peaks rather than matched peaks as an objective matrix to remove false positives.

Table 2 Frequencies of matched and unmatched ions for the peptides identified by Mascot. All the fragment ions with relative intensity of more than 5% are counted.

Peptides Identified in Figures	Number of Matched Ions	Number of Unmatched Ions
Figure 1A	17	9
Figure 2A	15	8
Figure 3A	7	3
Figure 4A	28	10
Figure 5A	25	4
Figure 6A	36	4
Figure 7A	15	17

Reversed or scrambled protein sequence databases have been used to determine the false positive rate of peptide identification [71]. While useful, these methods are unlikely to reflect the true positive rates. Among all the five types of false positives described here, the peptide sequences of the false positives are over 50% identical to the sequences of the corresponding true hits. These false positives would not be included in the false positive rate calculated by searching a reversed or scrambled protein sequence database. Therefore, false positive rates determined by searching such control databases should be much lower than the actual false positive rate. We believe that this gap will be more significant in those protein database searches where protein modifications are included.

Protein sequence alignment in which the mass of possible protein modifications is unrestricted has been used to comprehensively map sites of protein modifications [55,72,73]. Aligning sequences in this way can easily increase the size of the protein sequence database 1,000- to 10,000-fold, which will in turn lead to exponentially increased false positive rates. An objective and reliable method for verifying peptide identifications will be critical to high quality of proteomics data. This is especially important when mapping multiple protein modifications. The manual verification methods developed here represent an important step toward this goal.

2.5 Concluding Remarks

In this chapter, I presented an integrated approach for evaluation of peptide identification. The utility of the approach was demonstrated by (1) exhaustive detection of incorrect peptide identifications when searching against protein databases of an inappropriate species or that had reversed protein sequences; and (2) verification of correct peptide identifications given low Mascot scores (<25). With the application of the manual verification rules, I identified five common types of false-positives with high scores that are generated from statistic-based database searching. Our case studies highlight a few future directions for improving algorithms for protein sequence database searching. First, unmatched peaks should be emphasized when evaluating the accuracy of peptide identification. Second, care should be taken to select monoisotopic peaks, and not peaks higher in the isotopic distribution, for protein sequence database searching. Third, for those MS/MS spectra with ambiguous charge states, care should be taken to select the charge state giving a fragmentation pattern that best matches the base peaks with significant intensity. Fourth, the identification of post-translational modifications should require the modification site to be completely mapped in a restricted or unrestricted database search. When the fragmentation pattern is not sufficient to accurately localize the site of modification, the sequence alignment score should be reduced accordingly. Incorporation of these features into search algorithms will improve the accuracy of peptide identification and mapping modification sites.

2.6 Material And Methods

2.6.1 Preparation of Cell lysates from HeLa Cells and E. coli

One dish (10 cm) of HeLa cells was grown in DMEM supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin in a humidified CO_2 atmosphere at 37℃. When the cells reached 80-90% confluence, they were washed with cold Dulbecco's phosphate buffered saline twice. To the resulting cell pellet was added 200 µL of cell lysis buffer (6 M urea, 2 M thiourea, 4% CHAPS, 50 mM Tris-HCl, pH 8.0) to lyse the cells. The cell lysate was harvested and sonicated three times for 5 seconds each with 20-second intervals between sonications using a 550 Sonic Dismembrator (Fisher Scientific Corp, CA). The lysate was centrifuged at 4℃ for 1 h at 21,000 x g. The debri s was discarded while the supernatant fluid was divided into aliquots and stored at -80℃ until use. *E. coli* DH5 was grown aerobically in LB medium at 37℃. The cultured cells were harvested a t log phase by centrifugation at 4,500 x g for 10 min and washed twice by resuspension of the pellet in ice-cold PBS buffer (0.1 M Na_2HPO_4, 0.15 M NaCl, pH 7.2). The cells were resuspended in chilled lysis buffer (50 mM Tris-HCl, pH 7.5, 100 mM NaCl, 5 mM DTT) and then sonicated with 12 short bursts of 10 sec followed by intervals of 30 sec for cooling. Unbroken cells and debris were removed by centrifugation at 4^{o}C for 30 min at 21,000 x g. The supernatant was divided into aliquots and stored at -80℃ until use.

2.6.2 Protein digestion

HeLa cell lysate solution was diluted with four volumes of 50 mM ammonium bicarbonate buffer (pH 8.0) to reduce the urea concentration. Trypsin in 50 mM ammonium bicarbonate buffer was added to the HeLa cell lysate at an enzyme-to-substrate ratio of 1:50. After overnight incubation at 37℃, peptide solutio ns were dried in a SpeedVac (ThermoSavant Corp, Holbrook, NY) and reconstituted in 0.1% (v/v) trifluoric acetic acid (TFA) solution. *E. coli* cell lysate was digested in a similar fashion. µC18 ZipTips were used to wash the tryptic peptides according to the manufacturer's directions before nano-HPLC/mass spectrometry.

2.6.3 Nano-HPLC Mass Spectrometry Analysis

HPLC/MS/MS analysis was performed in an LCQ DECA XP ion-trap mass spectrometer (ThermoFinnigan, San Jose, CA) equipped with a nano-electrospray ionization source. The source was coupled online to an Agilent 1100 series nano flow LC system (Agilent, Palo Alto, CA). Two µL of the peptide solution in buffer A (2% acetonitrile/97.9% water/0.1% acetic acid, v/v/v) was manually injected and separated in a capillary HPLC column (50 mm length X 75 µm ID, 5 µm particle size, 300 Å pore diameter) packed in-house with Luna C18 resin. Peptides were eluted from the column with a 60-min gradient of 5% to 80% buffer B (90% acetonitrile/9.9% water/0.1% acetic acid, v/v/v) in buffer A. The eluted peptides were electrosprayed directly into the LCQ DECA XP ion-trap mass spectrometer. Normalized energy for collision-induced dissociation is 35%. Each MS/MS spectrum obtained by averaging three micro-scans with maximum injection time of 110 ms for each micro-scan. The MS/MS spectra were acquired in a data-dependent mode, such that the masses and fragmentation patterns of the two strongest ions in each MS scan were determined. All spectra were acquired in centroid mode as we usually performed.

2.6.4 Protein Sequence Database Search

Tandem mass spectra were used to search the NCBI-nr database with the Mascot search engine (Matrix Science, London, UK). Trypsin was specified as the proteolytic enzyme. Oxidization of methionine residues (+16 Da) and 1 missed cleavage site per peptide were taken into account. The maximum allowable mass error was set to ±4 Da for parent ion masses and ±0.6 Da for fragment ion masses. Charge states of +1, +2 or +3 were considered for parent ions. For the HPLC-MS case analysis presented as examples in the five types of commonly-seen false-positive peptide identifications, oxidation of methionine and one or more of the following modifications were set as variable modifications: acetylation, propionylation and butyrylation of lysine; phosphorylation of serine, threonine and tyrosine; methylation of aspartic acid and glutamic acid; and deamidation of asparagine and glutamine. If more than one spectrum was assigned to a peptide, each was given a Mascot score and only the spectrum with the highest score was used for manual analysis. Peptides identified with a Mascot score higher than 20 were considered to be potential positive identifications and each was manually verified.

CHAPTER 3

IDENTIFICATION AND CHARACTERIZATION OF TWO NOVEL PROTEIN MODIFICATIONS, LYSINE PROPIONYLATION AND BUTYRYLATION

Positive PTM identification by tandem mass spectrometry requires correctly interpreting the tandem mass spectra of the modified peptides. The traditional sequence alignment approach with statistical analysis causes false positive identifications with high statistical scores. Our manual verification strategy improves the accuracy and sensitivity of PTM identification by eliminating false results with poor alignment quality or wrong modification sites, and retaining true positives with high sequence alignment quality. In this chapter, I will present our discovery of two novel protein modifications, lysine propionylation and lysine butyrylation. Careful and thorough manual analysis, based on the principle that a correct sequence should adequately explain the major mass spectrometric peaks in the MS/MS spectrum, identified co-eluting molecular species that carry similar PTMs on different residues of the same peptide, correctly interpreted the mass spectrometry data and allowed the identification of novel lysine propionylation and butyrylation sites in histone H4, p300, CBP and p53, *in vitro* and *in vivo*. The work was a collaborative effort with Dr. Wei Gu’s lab at Columbia University.

3.1 Summary

The positively charged lysine residue plays an important role in protein folding and function. Neutralization of the charge often has a profound impact on the substrate proteins. Accordingly, all the known post-translational modifications at lysine have pivotal roles in cell physiology and pathology. In this chapter, I report the discovery of two novel, *in vivo* lysine modifications in histones, lysine propionylation and butyrylation. We confirmed, by *in vitro* labeling and peptide mapping by mass spectrometry, that two previously known acetyltransferases, p300 and CBP, could catalyze lysine propionylation and lysine butyrylation in histones and p53 *in vitro*. Finally, p300 and CBP could carry out autopropionylation and autobutyrylation *in vitro*. Taken together, our results conclusively establish that lysine propionylation and lysine butyrylation are novel post-translational modifications. Given the unique roles of propionyl-CoA and butyryl-CoA in energy metabolism and the significant structural changes induced by the modifications, the two modifications are likely to have important but distinct functions in the regulation of biological processes.

To further confirm *in vivo* lysine propionylation and butyrylation, we used mass spectrometry to map the *in vivo* modification sites of the three proteins, p53, p300 and CBP. We identified the first two *in vivo* lysine propionyltransferases, p300 and CBP, and the first depropionylase, Sirt1, in eukaryotic cells. We further demonstrated that p300 can carry out auto-lysine propionylation in HeLa cells. Our results suggest that lysine propionylation, like lysine acetylation, is a dynamic and regulatory post-translational modification. We further showed that some lysine propionylation regulatory enzymes are shared with those for lysine acetylation. Our results therefore identified the first several important players in the lysine propionylation pathway.

3.2 Introduction

Molecular anatomy of post-translational modifications that regulate cellular processes and disease progression stands as one of the major goals of post-genomic biological research. To date, more than 200 post-translational modifications have been described, providing an efficient way to diversify a protein's primary structure and possibly its functions [8,10,15]. The remarkable complexity of these molecular networks is exemplified by modifications at the side chain of lysine, one of the fifteen ribosomally-coded amino acid residues known to be modified [8]. The electron-rich and nucleophilic nature of the lysine side chain makes it suitable for undergoing covalent post-translational modification reactions with diverse substrates. The residue can be potentially modulated by several post-translational modifications including methylation, acetylation, biotinylation, ubiquitination, and sumoylation, which have pivotal roles in cell physiology and pathology.

Lysine acetylation is an abundant, reversible, and highly regulated post-translational modification. While initially discovered in histones [74], the modification was later identified in non-histone proteins, such as p53 [6]. A recent proteomics screening showed that acetyllysine is abundant and present in substrates that are affiliated with multiple organelles and have diverse functions [14]. Interestingly, the modification is enriched in mitochondrial proteins and metabolic enzymes, implying its roles in fine-tuning the organelle's functions and energy metabolism [14]. The modification plays important roles in diverse cellular processes, such as apoptosis, metabolism, transcription, and the stress response [75-78]. In addition to their roles in fundamental biology, lysine acetylation and its regulatory enzymes (acetyltransferases and deacetylases) are intimately linked to aging [79] and several major diseases such as cancer, neurodegenerative disorders, and cardiovascular diseases [80-82].

Acetyl-CoA, a member of the high-energy CoA compounds, is the substrate used by acetyltransferases to catalyze the lysine-acetylation reaction. It remains unknown, however, if cells can use other short-chain CoAs, such as propionyl- and butyryl-CoA (which are structurally close to acetyl-CoA), to carry out similar post-translational modifications at lysine. Nevertheless, several lines of evidence suggest such a possibility. First, like acetyl-CoA, propionyl-CoA and butyryl-CoA are high-energy molecules, making it thermodynamically feasible to carry out a reaction with a lysine side chain. Second, propionyl-CoA and butyryl-CoA are structurally similar to acetyl-CoA, with a difference of only one or two CH_2. Third, propionyl-CoA and butyryl-CoA are present at high concentrations in cells. In the case of starved mouse liver, the two CoA's concentrations are only 1-3 times less than acetyl-CoA [83]. Finally, it appears, from structural studies on some HATs (such as Hat1), that the enzyme has ample space within the cofactor binding pocket to accept propionyl-CoA without steric interference [84]. Despite such evidence, the short-chain CoAs with the exception of acetyl-CoA have not been described as a substrate for protein modification.

In this chapter, I will report the identification and validation of two novel post-translational protein modifications, propionylation and butyrylation at lysine residues, by a proteomics study. The unbiased global screening involved exhaustive peptide identification by nano-HPLC/MS/MS analysis, protein sequence database search, and manual verification. The resulting propionylated and butyrylated peptides were verified by MS/MS of their corresponding synthetic peptides. Using *in vitro* labeling with isotopic propionyl-CoA and butyryl-CoA as well as mass spectrometry, we identified two acetyltransferases, p300 and CBP, could perform robust lysine modifications at histones *in vitro*. Furthermore, we demonstrated that p300 and CBP could carry out autopropionylation and autobutyrylation at lysine residues in a similar fashion as autoacetylation. Taken together, these results reveal that lysine propionylation and butyrylation are novel lysine modifications that can be catalyzed by acetyltransferases.

To examine if lysine propionylation and butyrylation also occurs on non-histone proteins, we evaluated the lysine propionylation and butyrylation on p53 by acetyltransferases p300 and CBP. p53 is the first non-histone substrate protein found to be lysine acetylated. Lysine acetylation regulates the protein’s stability (by competing the potential sites for ubiquitination), protein-protein interactions (*e.g.*, with Mdm2 and Mdmx), and DNA-binding of p53. Lysine acetylation status modulates p53-mediated effects in both cell cycle arrest and apoptosis. In addition to lysine acetylation, we recently demonstrated that p53 can be lysine propionylated and butyrylated *in vitro*, catalyzed by p300 and CBP. This observation suggests the possible existence of the two PTMs in p53 *in vivo*. In this chapter I will present evidence suggesting that p53 can be lysine propionylated and lysine butyrylated by HATs *in vivo* and we also mapped the *in vivo* modification sites of p53 by co-transfecting p53 with p300 or CBP vector.

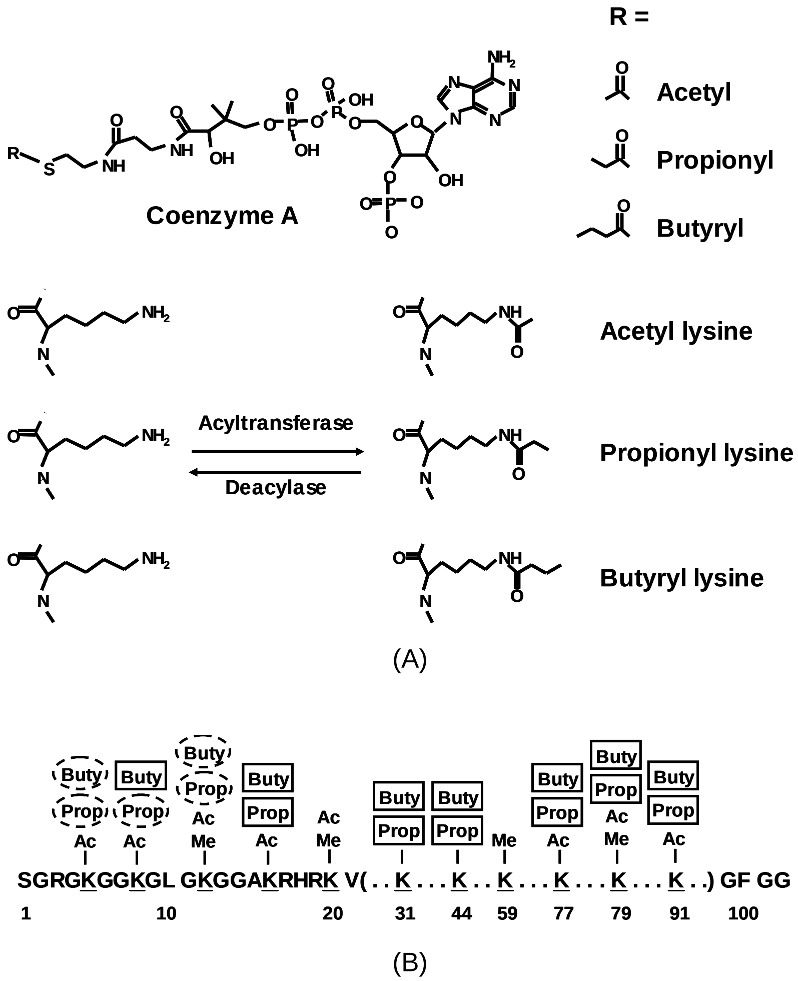


Figure 21 (A) Structures of three short-chain CoAs: the acetyl-CoA, propionyl-CoA, and butyryl-CoA, as well as the three modified lysines: acetyllysine, propionyllysine, and butyryllysine. (B) An illustration of novel lysine propionylation and butyrylation sites in histone H4. (unmarked labels: lysine-acetylation and -methylation sites identified previously; circled labels: novel, in vivo lysine-modification sites identified in this study; boxed labels: additional novel, in vitro lysine-modification sites identified in this study). The known sites of lysine acetylation and methylation were obtained from a histone website (www.histone.com).

Given the unique roles of propionyl-CoA and butyryl-CoA in energy metabolism [85], their distinct structure, and significant structural changes induced by the modifications, it is anticipated that lysine propionylation and butyrylation will have important, but likely distinct functions in the regulation of biological processes. The identification of lysine-propionylated and lysine-butyrylated substrates described here provides an entry point for future functional studies of the two modifications in cellular physiology and pathology.

3.3 Results

3.3.1 Initial identification of lysine propionylated and butyrylated peptides in histone H4 protein

We hypothesized that propionyl-CoA could be used by acetyltransferases for lysine propionylation. In addition, we further assumed that some propionylated, tryptic peptides could be affinity-purified by anti-acetyllysine antibody due to the close structural similarity between the acetyllysine residue and the propionyllysine residue. To identify lysine-propionylated peptides, we searched the MS/MS datasets of affinity-enriched acetyllysine-containing tryptic peptides acquired in nano-HPLC/LTQ mass spectrometry (the study on lysine-acetylation proteomics was published previously [14]). During the protein sequence database search, the lysine was considered as unmodified, acetylated, or propionylated. The database search and manual verification of peptide hits led to the identification of eleven lysine-propionylated histone H4 peptides (Table 3). The identification of lysine-propionylated peptides motivated us to hunt for the lysine-butyrylated peptides. The same datasets were searched again, with the lysine considered as unmodified, acetylated, or butyrylated. The analysis identified two additional histone H4 peptides with lysine butyrylation sites (Table 4).

Two examples of MS/MS spectra were presented to show the identification of lysine-propionylated and lysine-butyrylated peptides (Figure 22). In the initial peptide identification by protein sequence database search, the spectrum in Figure 22A led to the identification of a lysine-propionylated Peptide 1 (Table 3), while that in Figure 23A lysine propionylated Peptide 2 (Table 3). A careful inspection of the spectra revealed that there were a number of fragment ions in both spectra, with strong intensities, that could not be assigned to the fragment ions derived from the Peptide 1 and Peptide 2, respectively, suggesting the presence of additional peptide(s). Three lines of evidence suggest that the spectra were from peptide isomers: (i) multiple peak-pairs with mass difference of 14 Da were observed in both spectra; (ii) the peptides have the same molecular weights; and (iii) the peptides co-eluted. Indeed, the remaining peaks in spectrum Figure 22A could be accounted for by two additional lysine-butyrylated peptides: Peptide 12 and Peptide 13 (Table 4). Likewise, an additional lysine-propionylated peptide isomer, Peptide 3 (Table 3), could also be identified from the spectrum in Figure 23A.

Table 3 A list of lysine-propionylated peptides identified *in vivo*. Protein name, gene index number, modified peptide sequence and the number of modification sites are specified in two lists. Modified lysine residues are marked as: * - acetylated lysine, ^ - propionylated lysine, " - butyrylated lysine.

No.	Protein Name	gi#	Sequence	No. of Propionyl-Lys Site
1	Histone 4, H4	28173560	GK^GGK*GLGK^GGAK*R	2
2	Histone 4, H4	28173560	GGK^GLGK*GGAK*R	1
3	Histone 4, H4	28173560	GGK*GLGK^GGAK*R	1
4	Histone 4, H4	28173560	GK^GGK*GLGK*GGAK*R	1
5	Histone 4, H4	28173560	GK*GGK^GLGK*GGAK*R	1
6	Histone 4, H4	28173560	GK*GGK*GLGK^GGAK*R	1
7	Histone 4, H4	28173560	GK^GGK*GLGK*GGAK	1
8	Histone 4, H4	28173560	GK*GGK^GLGK*GGAK	1
9	Histone 4, H4	28173560	GK*GGK*GLGK^GGAK	1
10	Histone 4, H4	28173560	GGK*GLGK^GGAK	1
11	Histone 4, H4	28173560	GGK^GLGK*GGAK	1

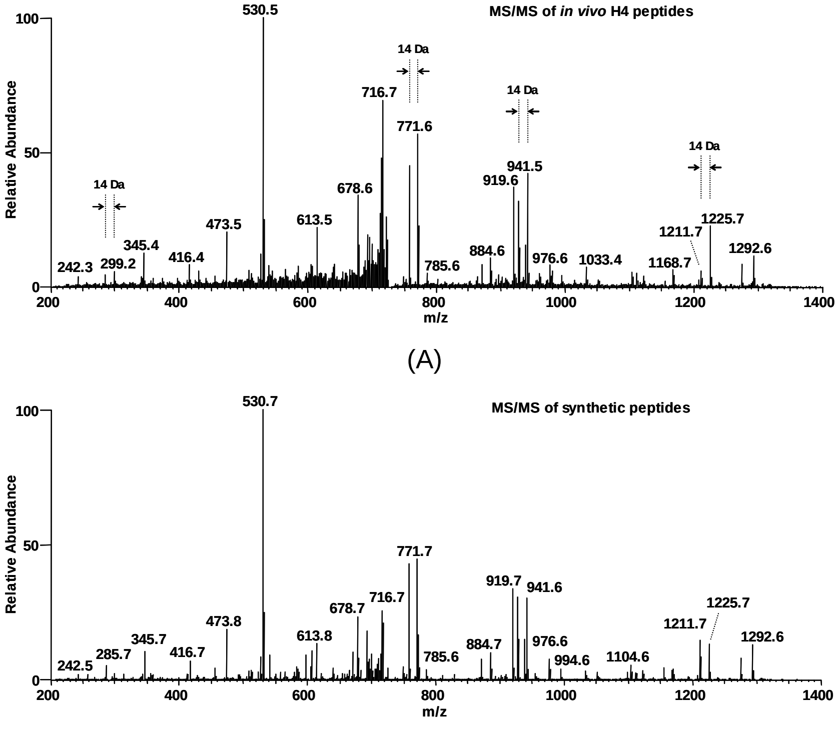
Table 4 A list of lysine butyrylated peptides identified *in vivo*. Protein name, gene index number, modified peptide sequence and the number of modification sites are specified in two lists. Modified lysine residues are marked as: * - acetylated lysine, ^ - propionylated lysine, " - butyrylated lysine.

No.	Protein Name	gi#	Sequence	No. of Butyryl-Lys Site
12	Histone 4, H4	28173560	GK"GGK*GLGK*GGAK*R	1
13	Histone 4, H4	28173560	GK*GGK*GLGK"GGAK*R	1

Table 5 The total number of lysine-propionylation and butyrylation sites identified in p53, histone H4, p300 and CBP *in vitro*.

In vitro analysis	p53	p300	Histone H4	CBP
Propionyl-Lys	11	21	9	12
Butyryl-Lys	9	11	9	7

The chemical nature of an identified peptide can be confirmed by MS/MS of their corresponding synthetic peptides, a gold standard for verification of peptide identification and chemical identity. To validate the identities of the propionylated and butyrylated peptides in Figure 22A, MS/MS of 3 synthetic peptides with the sequences and modification patterns corresponding to Peptide 1, 12 and 13 (Table 3 and Table 4) were analyzed. In order to validate the *in vivo* fragmentation spectrum by the peptide mixture observed in Figure 22A, three synthetic peptides were mixed at a molar ratio of 4:2:1 (Peptide 1: Peptide 12: Peptide 13). MS/MS analysis of the peptide mixture resulted in a fragmentation pattern (Figure 22B) that matched perfectly with the *in vivo* spectrum in Figure 22A, verifying the identification of one peptide with lysine propionylation and two peptides with lysine butyrylation. Likewise, in order to validate Peptides 2 and 3 (Table 3) identified from the *in vivo* spectrum in Figure 23A, two corresponding synthetic peptides were analyzed. The fragmentation pattern of a peptide mixture at a 2:1 molar ratio (Peptide 2:Peptide 3, Figure 23B) matched perfectly with the *in vivo* spectrum in Figure 23A, confirming the identification of two lysine-propionylated peptides.



(A)

(B)

Figure 22 Identification and verification of lysine-propionylated and lysine-butyrylated histone H4 peptides. (A) The tandem mass spectrum (MS/MS) of a tryptic peptide ion from a peptide mixture that was affinity-enriched with an anti-acetyllysine antibody from tryptic peptides of HeLa nuclear extracts. which identified one lysine-propionylated and two lysine-butyrylated peptides: Peptides 1, 12 and 13 (Table 3 and Table 4) from histone H4. (B) Tandem mass spectrum of a peptide mixture from the three synthetic peptides corresponding to the sequences identified in Panel A, showing similar ion intensity.

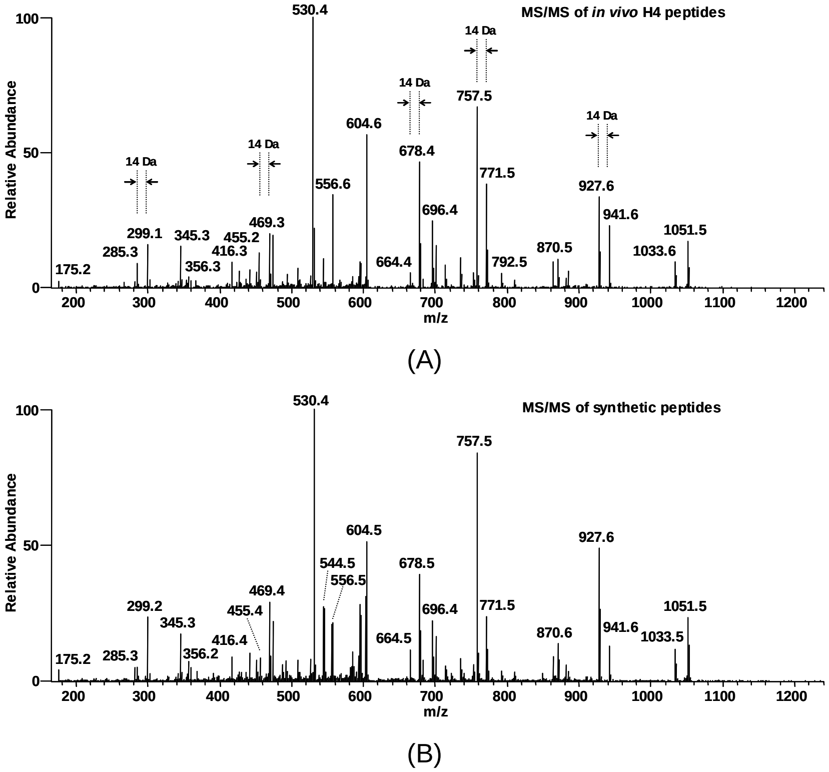


Figure 23 Identification and verification of lysine-propionylated and lysine-butyrylated histone H4 peptides. (A) MS/MS of a tryptic peptide ion led to the identification of two modified peptides, Peptides 2 and 3 (Table 3) from histone H4. (B) Tandem mass spectrum of a peptide mixture from the two synthetic peptides corresponding to the sequences identified in Panel A, showing similar ion intensity.

Together, we identified *in vivo* lysine propionylation at K5, K8, and K12, as well as *in vivo* lysine butyrylation at K5 and K12 of histone H4 (Table 3 and Table 4). The K5, K8, and K12 of histone H4 are known to be acetylated, while K12 is the subject of lysine methylation (www.histone.com). Lysine acetylation at the four H4 lysine residues is associated with transcriptional activation, transcriptional silencing, chromatin high-order structure, and DNA repair [86,87]. Some of the acetyllysine residues (*e.g.*, K8 of histone H4) provide a docking site to recruit a bromodomain-containing chromatin remodeling enzyme SWI/SNF [88]. While biological functions of lysine propionylation and butyrylation in histones remain unknown, it is intriguing to speculate that propionyllysine or butyryllysine would be involved in the interaction/recruiting of a distinct set of proteins/enzymes to control chromatin's structure and transcriptional activities.

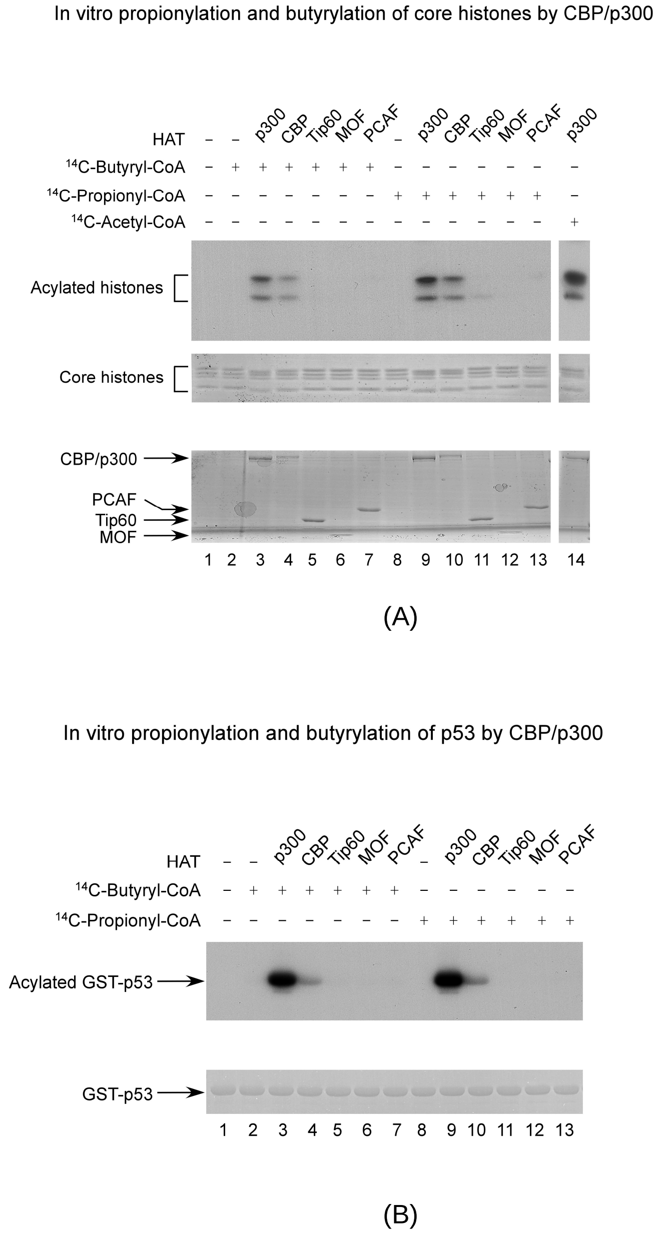
3.3.2 Propionylation and butyrylation of core histones can be catalyzed by p300/CBP

Since histone H4 can be propionylated and butyrylated *in vivo*, we next tested if core histones can be propionylated and butyrylated *in vitro* by acetyltransferases, using either ^{14}C-propionyl CoA or ^{14}C-butyryl CoA. Five acetyltransferases were tested, CBP, p300, Tip60, MOF and PCAF. CBP and p300 are known acetyltransferases for K5, K8, K12, and K16 of histone H4. The core histones were incubated with one of the CoAs and the acetyltransferase of interest. The protein mixture was then resolved in SDS-PAGE and visualized by autoradiography. CBP and p300 showed significant activities in catalyzing modifications on histone H4 (Figure 24A). On the other hand, no significant propionylation or butyrylation products were detected for the other three acetyltransferases, Tip60, MOF, and PCAF.

To corroborate evidence of an *in vitro* modification reaction at lysine residues, we used nano-HPLC/mass spectrometric analysis to map the CBP-catalyzed, lysine-modified residues in histone H4. K5, K8, K12, K16, K31, K44, K77, K79 and K91 were found to be both propionylated and butyrylated by CBP. Together, these data establish that histone H4 can be lysine-propionylated and -butyrylated directly by CBP and p300 *in vitro*.

3.3.3 p53 can be propionylated and butyrylated by p300/CBP in vitro

To examine if acetyltransferases can catalyze lysine propionylation and butyrylation reactions in non-histone proteins, we evaluated *in vitro* propionylation and butyrylation reactions in p53. CBP/p300, co-activators of p53, augment its transcriptional activity and modulate its biological functions [6,89]. Multiple lysine residues in p53 can be acetylated, some of which are known to be modified by CBP/p300 [6,90,91]. Given the fact that CBP/p300 are the acetyltransferases for p53 and that they have enzymatic activities for lysine propionylation and butyrylation in histones, we tested if the HATs could catalyze similar reactions in p53. Toward this aim, we repeated the *in vitro* enzymatic reactions for p53 using a procedure described above. Again, only two of the five acetyltransferases, CBP and p300, could carry out propionylation and butyrylation reactions at p53 at a significant reaction rate under our experimental conditions (Figure 24B). Interestingly, p300 shows higher catalytic activity than CBP for p53. In contrast, the two enzymes have comparable activities toward histones.


In vitro propionylation and butyrylation of core histones by CBP/p300

(A)

In vitro propionylation and butyrylation of p53 by CBP/p300

(B)

Figure 24 *In vitro* propionylation and butyrylation of core histones and p53 by acetyltransferases. (A) *In vitro* propionylation and butyrylation of core histones. (B) *In vitro* propionylation and butyrylation of p53.

CBP and p300 are acetyltransferases that can catalyze autoacetylation reactions. To test if the proteins can carry out autopropionylation and autobutyrylation reactions, we mapped the modification sites at p300 and CBP by mass spectrometry. Twenty-one lysine-propionylation sites and eleven lysine-butyrylation sites were localized in p300, while twelve lysine-propionylation sites and seven lysine-butyrylation sites were mapped in CBP (Table 5). Identification of propionylated and butyrylated peptides in non-histone proteins, p53, CBP, and p300, suggests the possibility that the two modifications are not restricted to histones.

3.3.4 p300 and CBP are propionyltransferases for p53 in vivo

Given the fact that the two enzymes can catalyze both *in vitro* acetylation and propionylation reaction and that they can carry out *in vivo* p53 acetylation, we rationalize that the enzymes are likely to be able to perform *in vivo* propionylation in p53. To test this, we used co-transfection experiments and Western blotting analysis by a pan-propionyllysine antibody. As expected, p53 lysine propionylation was dramatically increased when co-transfected with either p300 or CBP (Figure 25A), but not their HAT-dead mutants (Figure 25B). Co-transfection experiment using other four HATs, MOF, Tip60, PCAF, and HBO1, suggest that these enzymes have little enzymatic activities toward p53 lysine propionylation (Figure 25B).

To identify *in vivo* lysine propionylation sites in p53, we isolated p53 proteins using immunoprecipitation, when p53 was co-transfected with CBP (Figure 26A). The isolated proteins were subjected to in-gel digestion and HPLC/mass spectrometric analysis for mapping lysine acetylation and lysine propionylation sites. As expected, our analysis identified 15 lysine acetylation sites in p53. Interestingly, we also identified one lysine propionylation site (Figure 26B, C). The propionllysine sites were identified in CBP co-transfection cells, but not in cells without CBP co-transfection (data not shown), further suggesting that CBP is an *in vivo* propionyltransferase for p53.

3.3.5 p300 and CBP are lysine propionylated in vivo

We previously found that p300/CBP can catalyze auto-lysine propionylation reaction *in vitro* [42]. To examine if they are propionylated *in vivo*, we carried out transfection and Western blotting experiments. To increase lysine propionylation levels, we treated cells with HDAC inhibitors for 6 hours before they were harvested. Lysine propionylation in p300 was detected in cells that was further enhanced by HDAC inhibitors (Figure 27A), suggesting that depropionylation reaction is likely catalyzed by deacetylation enzymes. On the other hand, the modification in CBP was not detected under our experimental conditions (Figure 27B). To test if Sirt 1, a class III HDACs, can carry out depropionylation reaction in p300, we detected p300 lysine propionylation when co-transfected with wt Sirt1 and its HDAC-dead mutant. Lysine propionylation level was significantly decreased when co-transfected with wt Sirt1, but not its mutant (Figure 27C), implying that Sirt1 can catalyze *in vivo* lysine depropionylation reaction.

Next, we carried out mass spectrometric analysis of p300 and CBP immunoprecipitated directly from HeLa cells. To possibly boost lysine propionylation levels, the cells were treated with HDAC inhibitors for 6 hours before harvesting cells. HPLC/mass spectrometric analysis in combination with protein sequence database searching identified 27 lysine acetylation sites, 4 propionylation sites and 4 lysine butyrylation sites in p300 (Figure 28A), Likewise, we also identified 31 acetylating sites and 2 lysine butyrylation sites in CBP (Figure 28B). We were not able to identify propionylation sites in CBP with mass spectrometry possibly due to low stoichiometry of the modification. Nevertheless, together with Western blotting analysis, these results demonstrate that p300 is an *in vivo* lysine propionylation substrate while p300 and CBP are *in vivo* lysine butyrylation substrates, and SirT1 can catalyze *in vivo* lysine depropionylation reaction.

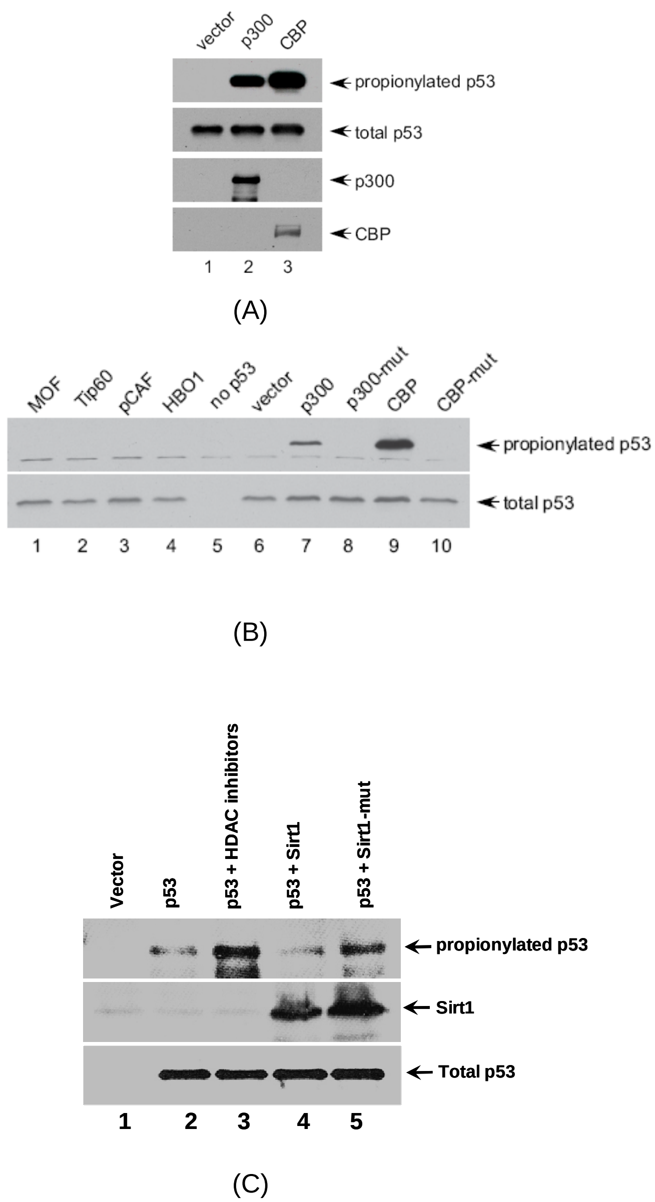
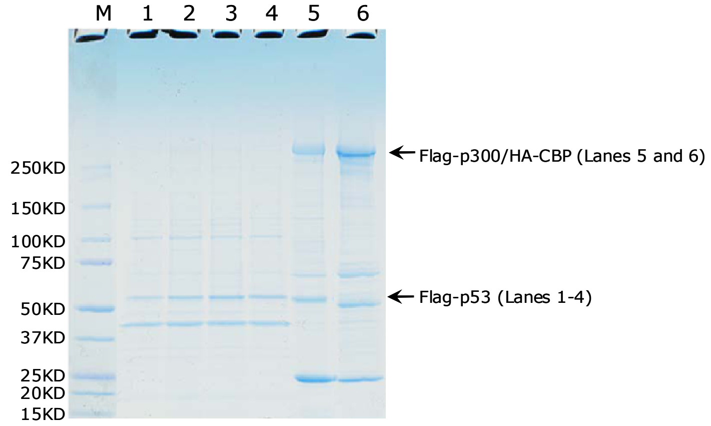


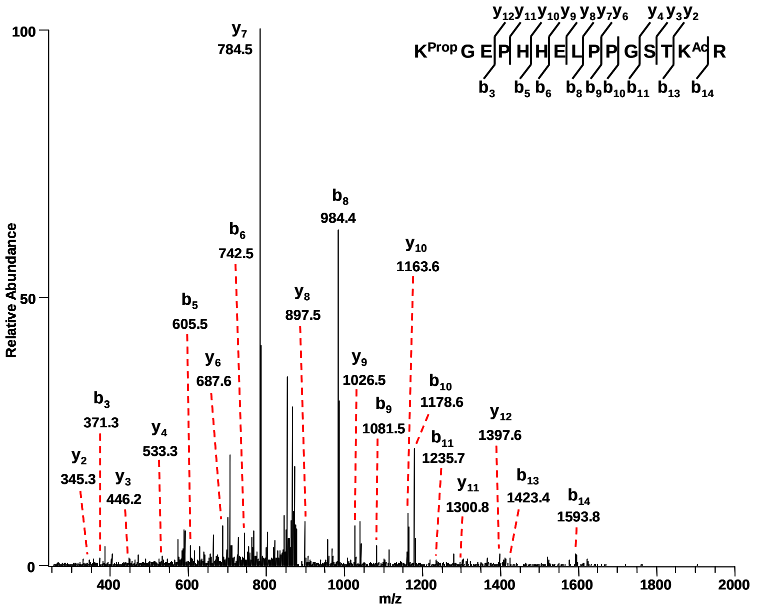
Figure 25 *In vivo* propionylation of p53 determined by p300/CBP and Sirt1. (A) p300/CBP catalyzes p53 propionylation *in vivo*. (B) p300/CBP are specific propionyl transferases in p53 propionylation. (C) Sirt1 de-propionylates p53 *in vivo*.



(A)

Protein	K^{Prop}/K^{Buty} Sites	Peptide Sequence	Co-expression with			
			control	p300	p300 + SIRT1	CBP
p53	K292	$\underline{\mathbf{K}}^{Prop}$GEPHHELPPGST$\underline{\mathbf{K}}^{Ac}$R	-	+	-	+
	K372	S$\underline{\mathbf{K}}^{Buty}\underline{\mathbf{K}}^{Ac}$GQSTSR	-	-	-	+
	K373	S$\underline{\mathbf{K}}^{Ac}\underline{\mathbf{K}}^{Buty}$GQSTSR	-	-	-	+
	K382	H$\underline{\mathbf{K}}^{Ac}\underline{\mathbf{K}}^{Buty}$LMFK	-	+	+	+

(B)



(C)

Figure 26 Mapping lysine propionylation sites in p53 by mass spectrometry. (A) Purification of p53, p300 and CBP *in vivo*. (B) *In vivo* lysine propionylated and butyrylated peptides and sites identified in p53 when it was co-transfected with (+) or without (-) an enzyme; (C) MS/MS spectrum of “$\underline{\text{K}}^{Prop}$GEPHHELPPGST$\underline{\text{K}}^{Ac}$R” that identified K292 as *in vivo* lysine propionylation site in p53.

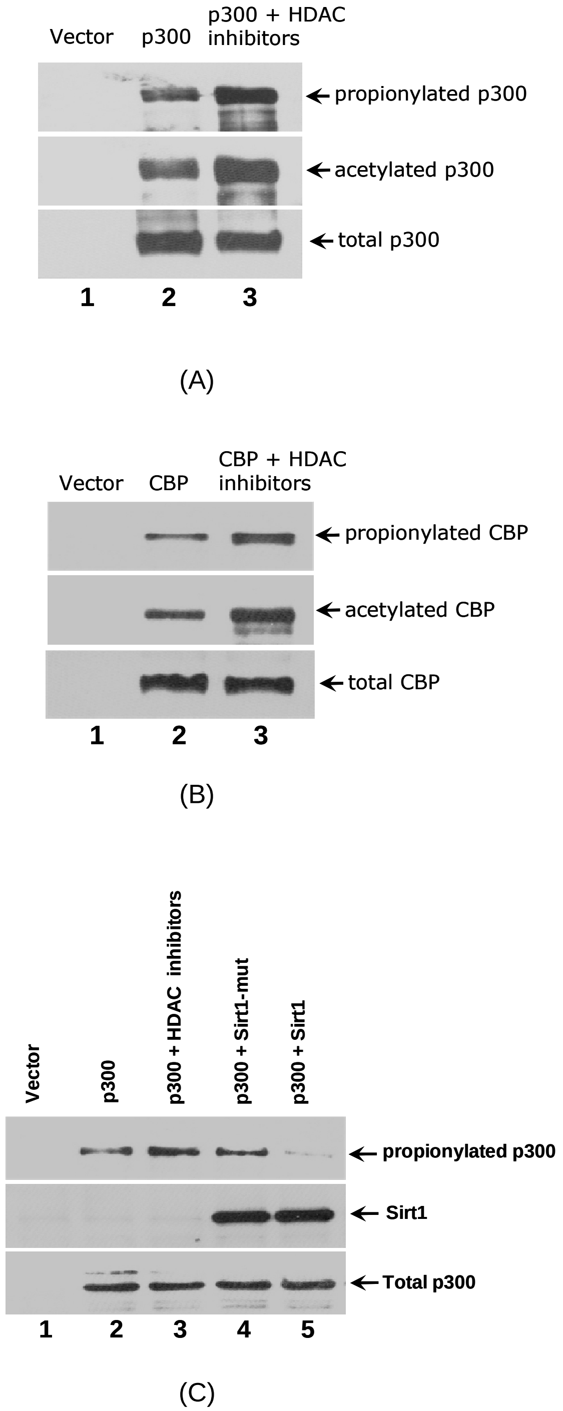


Figure 27 p300 and CBP are lysine propionylated *in vivo.* (A)The p300 propionylation was induced *in vivo* by HDAC inhibitor cocktail. (B). The CBP propionylation levels were enhanced *in vivo* under HDAC inhibitor cocktail treatment. (C) Sirt1 de-propionylates p300 *in vivo*. 293T cells were transfected with plasmid DNA expressing Flag-p300 with or without Sirt1-V5-His.

3.4 Discussion

Lysine is one of the two major ribosomally coded amino acid residues with a positively charged side chain at physiological pH, playing an important role in protein folding and functions. We do not know yet what will be the functional consequences of the two modifications identified in this study. Nevertheless, given the regulatory functions of every known lysine modification and significant structural changes induced by lysine propionylation and butyrylation, it is anticipated that the two modifications are likely to have biological functions in the same fashion as other lysine modifications.

Histones are known to be modified by an array of post-translational modifications, including methylation, acetylation, ubiquitination, small ubiquitin-like modification, and ribosylation [92]. A combinatorial array of post-translational modifications in histones, termed the "histone code", dictates the proteins' functions in gene expression and chromatin dynamics [45]. Post-translational modifications of histones have been elegantly studied by both biochemistry [45] and mass spectrometry [93-95]. Nevertheless, lysine propionylation and butyrylation at histones have not been reported before. Our results suggest that the complexity of histone modifications remains to be explored.

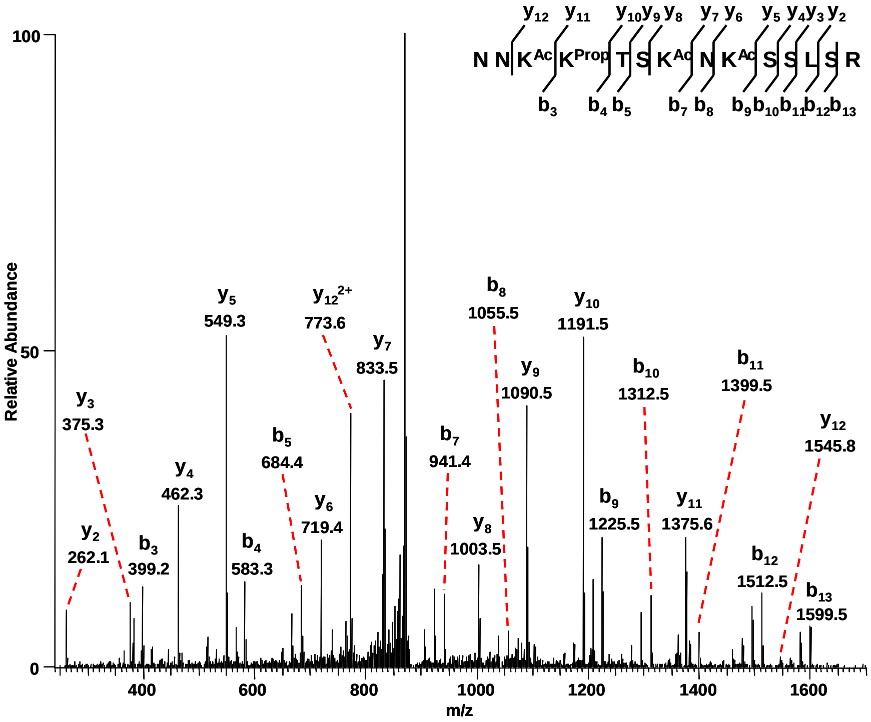
Acetyl-CoA can arise during the catabolism of sugars, fatty acids and amino acids. Propionyl-CoA derives only from odd-chain fatty acid and amino acid catabolism, while butyryl-CoA is a metabolic intermediate formed during the β-oxidation of fatty acids as well as a substrate for fatty acid elongation. The concentration of the short-chain CoAs fluctuates depending on diet and cellular physiological conditions [83]. If the rate of the modifications depends on the concentration of the short-chain CoAs, directly or indirectly, it would be appealing to speculate that lysine propionylation and butyrylation may regulate cellular metabolic pathways in response to cellular physiological conditions. Such a scenario then opens up the potential for the biochemical intermediates thus produced to lead to tissue-specific and environmentally-responsive regulatory programs.

Protein	K^{Prop} / K^{Buty} Site	Peptide Sequence
p300	K1554	NNK^{Prop}K^{Ac}TSK^{Ac}NK^{Ac}SSLSR
	K1555	NNK^{Ac}K^{Prop}TSK^{Ac}NK^{Ac}SSLSR
	K1558	NNK^{Ac}K^{Ac}TSK^{Prop}NK^{Ac}SSLSR
	K1560	NNK^{Ac}K^{Ac}TSK^{Ac}NK^{Prop}SSLSR
	K1554	NNK^{Buty}K^{Ac}TSK^{Ac}NK^{Ac}SSLSR
	K1555	NNK^{Ac}K^{Buty}TSK^{Ac}NK^{Ac}SSLSR
	K1558	NNK^{Ac}K^{Ac}TSK^{Buty}NK^{Ac}SSLSR
	K1560	NNK^{Ac}K^{Ac}TSK^{Ac}NK^{Buty}SSLSR

(A)

Protein	Modification Sites	Peptide Sequence
CBP	K1595	NNKKTNK^{Buty}NKSSISR
	K1597	NNKKTNKNK^{Buty}SSISR

(B)



(C)

Figure 28 Mapping lysine propionylation sites in p300 by mass spectrometry. *In vivo* lysine propionylated and butyrylated peptides and sites identified in p300 (A) and in CBP (B). (C) MS/MS spectrum of “NNK^{Ac}K^{Prop}TSK^{Ac}NK^{Ac}SSLSR” that identified K1555 as *in vivo* lysine propionylation site in p300.

The activities and specificities of regulatory enzymes responsible for propionylation and butyrylation might be influenced by other factors. For some acyltransferases (*e.g.*, acetyltransferases), the dimensions of the CoA-binding pocket might determine the binding affinity to the short-chain CoA. Alternatively, it is also possible that the binding affinity might be modulated by other regulatory molecules, such as association of co-factors or interaction proteins. Finally, identification of lysine propionylation and lysine butyrylation further suggests the possible existence of novel enzymes specific for the modifications (other than the known acetyltransferases).

Discovery of lysine propionylation and butyrylation raises many interesting questions. What are the enzymes responsible for controlling the status of lysine-propionylation and -butyrylation? Given the close structural similarity, some of the HATs and HDACs might be able to catalyze the reactions. What determines the specificities of a transferase among three types of lysine modifications: acetylation, propionylation, and butyrylation? What are the biological pathways regulated by lysine propionylation and butyrylation? The study described here is really the beginning of future research into addressing the fundamental questions of the novel lysine modifications.

3.5 Concluding Remarks

In this chapter, I report the identification of *in vivo* lysine propionylation and lysine butyrylaton in histones. The modifications were validated by synthetic peptides and *in vitro* enzymatic reactions. After testing five acetyltransferases, we showed that two previously known acetyltransferases, CBP and p300, could catalyze lysine modification reactions using either propionyl-CoA or butyryl-CoA, in histones and p53, at a significant reaction rate *in vitro*. With *in vivo* identification of lysine propionylation and butyrylation, our results further demonstrated that (i) lysine propionylation is present in three non-histone proteins, p53, p300 and CBP. (ii) p300 and CBP can catalyze lysine propionylation reaction in p53 *in vivo*. (iii) p300 and CBP have auto-lysine propionylation enzymatic activities. And (iv) Sirt1 is a depropionylase that removes lysine propionylation from p53 and p300. These results indicate that many substrates and regulatory enzymes are likely to be shared between lysine acetylation and lysine propionylation pathways. While three regulatory enzymes for lysine acetylation (Sirt1, p300 and CBP) are shared with lysine propionylation pathway, Tip60 seems to have much higher selectivity and is likely to restrict its enzymatic activity to lysine acetylation, at least in the case of p53 protein.

3.6 Material and Methods

3.6.1 In-gel Digestion and HPLC/MS/MS Analysis

Protein in-gel digestion, peptide extraction, and peptide cleaning using a µ-C18 Ziptip were carried out as previously reported [96]. HPLC/MS/MS analysis for mapping propionylation and butyrylation sites in histone H4, p53, p300, and CBP were carried out by nano-HPLC/LTQ mass spectrometry as previously described [14]. Briefly, each tryptic digest was dissolved in 10 µL HPLC buffer A (0.1% formic acid in water (v/v)) and 2 µL were injected into an Agilent HPLC system (Agilent, Palo Alto, CA). Peptides were separated on a home-made capillary HPLC column (50 mm length X 75 µm ID, 4 µm particle size, 90 Å pore diameter) with Jupiter C12 resin (Phenomenex, St. Torrance, CA) and directly electrosprayed into the mass spectrometer using a nano-spray source. The LTQ mass spectrometer was operated in the data-dependent mode acquiring fragmentation spectra of the ten strongest ions.

3.6.2 Protein sequence database search and manual verification

All MS/MS spectra were searched against NCBI-nr protein sequence database specifying lysine modifications using the MASCOT database search engine (version 2.1). A low cutoff peptide score of 20.0 was selected to maximize the identification of lysine-modified peptides. For each Mascot search, the peptide mass error was set to +/- 4 Da, fragment ion mass error was set to +/- 0.6 Da and six missing cleavages were allowed. All lysine-propionylated or -butyrylated peptides identified with MASCOT score > 20.0 were manually examined with the rules previously described [41] and all lysine propionylation or butyrylation sites had to be identified by consecutive b- or y- ions so that the possibility that propionylation (+56 Da) or butyrylation (+70 Da) occurred on adjacent residues was eliminated.

3.6.3 Synthesis of lysine-propionylated and lysine-butyrylated peptides

The peptides were synthesized on a Protein Technologies Symphony peptide synthesizer using Fmoc chemistry. All amino acids were purchased from Novabiochem (San Diego, CA) and the solvents were obtained from Thermo Fisher Scientific (Fair Lawn, NJ). Fmoc-Lys(Ac) was used for lysine residues with acetylated side-chains. For lysine residues requiring modification with either butyl or propionyl moieties, an orthogonally-protected Fmoc-Lys(Mtt) reagent was used. At the end of the synthesis, prior to removal of the N-terminal Fmoc protecting group, the methyltrityl (Mtt) side-chain protection was removed with 1% trifluoroacetic acid in dichloromethane. The resin was washed 10x in the acidic solution until the yellow color disappeared. The resin was then treated with 5% diisopropylethylamine to neutralize the TFA salt and the free amino group was reacted with either propionic acid or butyric acid, which had been preactivated with 2-(1H-Benzotriazole-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate (HBTU / HOBt). The coupling efficiency was monitored using a quantitative ninhydrin test. After derivatization the resin was treated with 20% piperidine in N-methyl-2-pyrrolidone (NMP) to remove the Fmoc group and cleaved with 95% TFA, containing thiol scavengers for 90 mins. The crude peptides were precipitated in diethyl ether and desalted on C-18 RP Sep-Pak columns before lyophilization to a dry powder.

3.6.4 In vitro propionylation and butyrylation assay

In vitro propionylation and butyrylation assays were carried out essentially as previously described [6] with some modifications. Tagged human FLAG-p300, HA-CBP, FLAG-MOF, and FLAG-PCAF proteins from transfected 293 cells and tagged human GST-Tip60 and GST-p53 expressed in bacteria were purified to homogeneity under stringent conditions (500 mM NaCl + 1% Triton X-100). Ten-microliter reactions contained 50 mM Tris pH 7.9, 10% glycerol, 1 mM DTT, 10 mM sodium butyrate, 1 μl of [^{14}C]-acyl-CoA (55 mCi/mmol; acetyl-CoA from Amersham and propionyl-CoA and butyryl-CoA from ARC, Inc.). Two and a half µg of substrates (human core histones and recombinant human Histone H4 [Upstate, Lake Placid, NY] or GST-p53) and 20 – 100 ng of the enzyme protein, as indicated, were incubated at 30^{o}C for 1 hour. The reaction mixture was then subjected to electrophoresis on SDS-PAGE gels, followed by either autoradiography or Coomassie Blue staining.

3.6.5 Cell culture, cell transfection and treatment for propionylation and butyrylation of p53, p300, and CBP in cells

293T cells were maintained in Dulbecco's Modification of Eagle's Medium (DMEM) in the presence of 10% fetal bovine serum at 37 ^{o}C in a humidified atmosphere of 5% CO_2. Twenty-µg plasmid DNA of interest, either Flag-p300 or HA-CBP, were transfected into 3 x 107 293T cells with Lipofectamine 2000 as described in manual instructions (Invitrogen). For co-transfection experiment, 20 µg of the HAT plasmid DNA was co-transfected with 10-ug His-Sirt1 DNA into 3 x 107 293T cells. Thirty-six hours after transfection, cells were treated with 2 µM TSA, 30 mM nicotinamide and 50 mM sodium butyrate for six hours to inhibit endogenous HDACs enzymatic activities before harvesting the cells.

24 hours after transfection with various plasmids, as indicated, H1299 cells were treated with 5 mM sodium butyrate for 6 hours, harvested, and lysed in Flag-lysis buffer (50 mM Tris-HCl pH 7.9, 137 mM NaCl, 10 mM NaF, 1 mM EDTA, 1% Triton X-100, 0.2% Sarkosyl, 10% glycerol, and fresh proteinase inhibitor cocktail (Sigma)) plus 10 mM Sodium butyrate. Propionylation of p53 was detected by western blot analysis using anti-propionyl-lysine antibody either in the total cell extracts or in the immunoprecipitated material by M2 agarose beads. Propionylation of p300 or CBP was detected by western blot analysis using anti-propionyl-lysine antibody in immunoprecipitated materials by M2 or HA agarose beads.

3.6.6 Purifications of p53, p300 and CBP proteins

293T cells were lysed by NETN buffer (20 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5% NP-40; containing 30 µM TSA, 50 mM sodium butryrate, 30 mM nicotinamide and cocktail protease inhibitors). Cell lysates were clarified by centrifugation at 100,000 x g for 20 minutes at 4 ^{o}C and incubated with either Flag-M2 beads or HA beads (Sigma-Aldrich, St. Louis, MO) at 4 ^{o}C for 1 hour. Antibody beads were washed five times with NETN buffer. The bound proteins were eluted with SDS sample buffer and resolved in SDS-PAGE for colloidal coomassie staining.

CHAPTER 4

DEVELOPMENT OF PTMAP, A NOVEL ALGORITHM FOR UNRESTRICTIVE SEQUENCE ALIGNMENT WITH HIGH ACCURACY

4.1 Summary

Molecular characterization of novel PTMs by mass spectrometry involves three essential aspects: confident identification of the peptide sequences bearing PTMs, precise localization of PTM sites, and accurate determination of PTM mass shifts. Based on our experience in the identification of novel lysine propionylation and butyrylation as well as manual analysis of tandem mass spectra, we found that statistics-based software for restrictive and unrestrictive PTM analysis such as Mascot, Sequest (with unified scores), X!Tandem, MS-Alignment and P-Mod lack emphasis on the spectral matching quality and rely heavily on the significance test of matched ions from statistical simulations. Such an approach introduces bias against peptide identifications with high-quality sequence alignment but low statistical significance and causes false positive PTM identifications that match poorly but have high statistical significance. These false positives could be identified with the wrong modification site, the wrong mass shift or even the wrong peptide sequence. Due to their high statistical significance, these results cannot be removed using existing methods for evaluating statistics-based false positive rates, including reverse/scrambled protein sequences or an empirical Bayesian model. Effective removal of these false positives often requires extensive and time-consuming manual verification. To address this issue, we developed PTMap software for accurate, unrestrictive identification of PTMs and polymorphisms. In this chapter, I will introduce the development and application of this new software to discover novel PTMs.

PTMap incorporates several novel features to improve searching speed and accuracy, including peak selection, adjustment of inaccurate mass shifts, and precise localization of PTM

sites. PTMap also automates rules, based mainly on unmatched peaks, for manual verification of identified peptides. To evaluate the quality of sequence alignment, we developed a scoring system that takes into account both matched and unmatched peaks in the mass spectrum. Incorporation of these features dramatically increased both accuracy and sensitivity of the peptide and PTM identifications. To our knowledge, PTMap is the first algorithm that emphasizes unmatched peaks to eliminate false positives. The superior performance and reliability of PTMap were demonstrated by confident identification of PTMs on 156 peptides from four proteins and by MS/MS verification with the synthetic peptides of the low-scoring peptides that were generally considered as false identifications by Mascot search engines. Our results demonstrate that PTMap is a powerful algorithm capable of identification of all possible protein PTMs with high confidence.

4.2 Introduction

Mass spectrometry is the method of choice for mapping sites of protein post-translational modifications (PTMs) that are known to have more than 300 types [65,97]. Efficient sequence alignment algorithms are essential for mapping PTM sites and peptide identification from mass spectrometric data [98]. Widely used programs such as Sequest [32,99], Mascot [33], and X!Tandem [35] identify PTM sites based on a restricted database search, in which tandem mass spectra are aligned with protein sequences bearing one or several specified PTMs at specified amino acid residues. This restricted database search strategy developed in early days has been very useful for identifying peptides bearing a limited number of specified PTMs, but lacks the flexibility to identify unexpected PTMs.

Recently, several algorithms have been developed to extend the capability of shotgun proteomics to allow identification of all possible PTMs and sequence polymorphisms [72,73,100-104]. These algorithms can carry out unrestricted database searches, identifying any PTM, whether previously known or unknown. To improve the accuracy of peptide identification, different statistical strategies have been proposed in attempts to reduce the number of false positives [60,72,73,104-106]. These strategies typically involve applying a statistical significance test to score the confidence level of each identification. While useful, the reliability of these strategies has not been critically evaluated, for example, by testing them with manual verification with high stringency, or with MS/MS of synthetic peptides, the gold standard for confirming peptide identification.

Identification of false positives using statistical methods is daunting in unrestricted sequence alignments for several reasons. First, the size of the protein sequence database is exponentially increased. For example, consider a peptide of 12 amino acid residues. If it is assumed that this peptide contains a single PTM at an unknown position, and that this PTM causes a mass shift that could range over the integers between -100 and +300, a list of 4800 possible modified peptides is generated from the single peptide sequence. The exponentially increased size of the peptide pool and the high similarity among peptide sequences derived from the full spectrum of possible PTMs make it difficult to remove most of the false positives, let alone all of them. Second, a PTM could happen at several residue (*e.g.*, protein methylation [107]), modified peptides with adjacent PTM sites are likely to have similar theoretical fragmentation patterns, and hence similar statistical scores. Third, in silico-generated peptide sequence pools used for sequence alignment are unlikely to include all of the peptide sequences generated by proteolytic digestion. Take, for example, trypsin, a protease often used for shotgun proteomics, generates not only tryptic peptides, but also semi-tryptic peptides, in which one terminus is generated by chymotryptic activity within the enzyme preparation [46].

The false positives were caused during protein database search, because of misinterpretation of charge states, abnormal enzymatic digestion sites, misinterpretation of protein modifications, wrong assignment of modification sites and modification types, and incorrect use of isotopic peaks [46]. Unfortunately, the decoy protein sequence database cannot accurately predicate the false discovery rates as many false positives are sequence-linked to the true hits, such as those cases caused by abnormal enzymatic digestion and wrong assignment of modification sites and types [46].

We have observed that the MS/MS spectra of false positive identifications commonly contain unmatched peaks with significant intensities. Accordingly, we argue that, while identification of candidate peptides should focus on matched peaks, to identify false positives comprehensively, emphasis should be placed on unmatched peaks. Our argument is based on the rationale that a true peptide identification should explain all major peaks in the spectrum, specially for those MS/MS data that are generated in ion trap types of mass spectrometers, which have relatively simple fragmentation patterns [41]. We further argue that the modification site should be unambiguously located to confidently identify peptides containing PTMs, and to eliminate false positive PTM assignments.

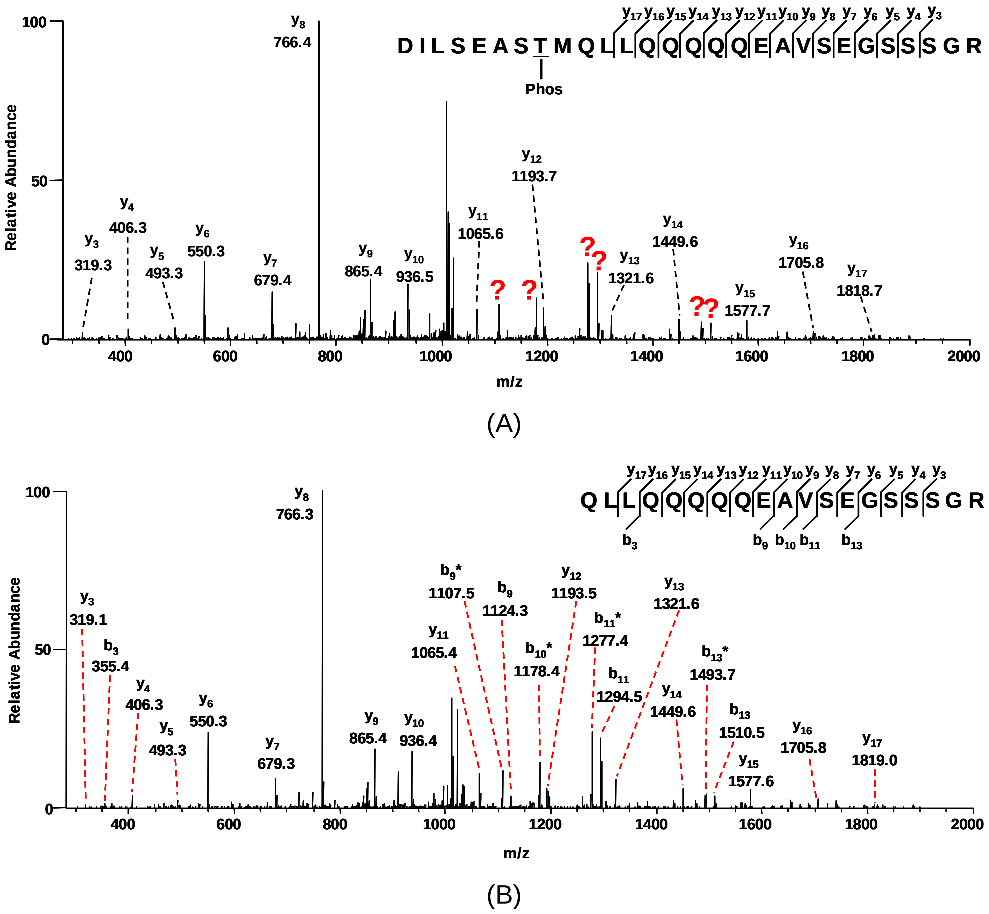
In this chapter, I will describe PTMap, a new sequence alignment algorithm for reliable, full-spectrum identification of mass shifts associated with unspecified PTMs or polymorphisms. To reduce the risk of obtaining false positives when identifying peptides that bear an unknown mass shift at an unknown amino acid residue, the algorithm adopts several novel strategies: selection of sequence-rich MS peaks, automation of mass-shift adjustment, precise localization of PTM sites, implementation of rules for manual verification, and development of a scoring system based on spectrum quality to remove false positive sequence alignments. PTMap also reduces searching time by aligning the sequences of only those proteins that are of interest, and by removing isotopic peaks and noise peaks. A unique feature of PTMap is identification of false positives by emphasizing unmatched peaks instead of matched peaks with statistically significant scores. The software is able to use MS/MS data from low-resolution tandem mass spectrometers by implementing automatic mass-shift adjustment. To demonstrate its usefulness, we used PTMap to identify PTM sites in human histone H4, HMG2, mouse SGK1, and bovine serum albumin. Accuracy of peptide identification by PTMap was confirmed by MS/MS of synthetic peptides for the corresponding peptide identifications with Mascot scores between 15-27 that were usually considered as false positives. Our data demonstrate that PTMap can remove almost all the false positives while maintaining high sensitivity for identification of peptides and PTM sites.

4.3 Results

4.3.1 The challenge for statistics-based sequence alignment algorithms: identifying false positives

Statistics-based algorithms have been used to analyze MS/MS data, resulting in putative peptide sequence matches that are statistically significant relative to the peptide pool of interest. Among these statistically significant matches will be both true and false positive hits. Such algorithms have been shown to be powerful for peptide identification and mapping PTM sites when a limited number of PTMs are involved. We argue that such statistical methods cannot efficiently eliminate false positives in an unrestrictive sequence alignment, because of the exponentially increased size of the protein sequence database, high sequence similarities among modified peptides that have the same sequence and PTM type but different PTM sites, and the existence of atypical proteolytic peptides (*e.g.*, semi-tryptic peptides). While statistical methods can efficiently calculate the statistical significance of randomly matched peptides, these methods are less capable of identifying false positives that are derived from non-random peptides, such as those arising from proteolytic digestion at peptide bonds other than those corresponding to an enzyme's specificity or peptides from incorrect assignment of a PTM type or PTM sites.

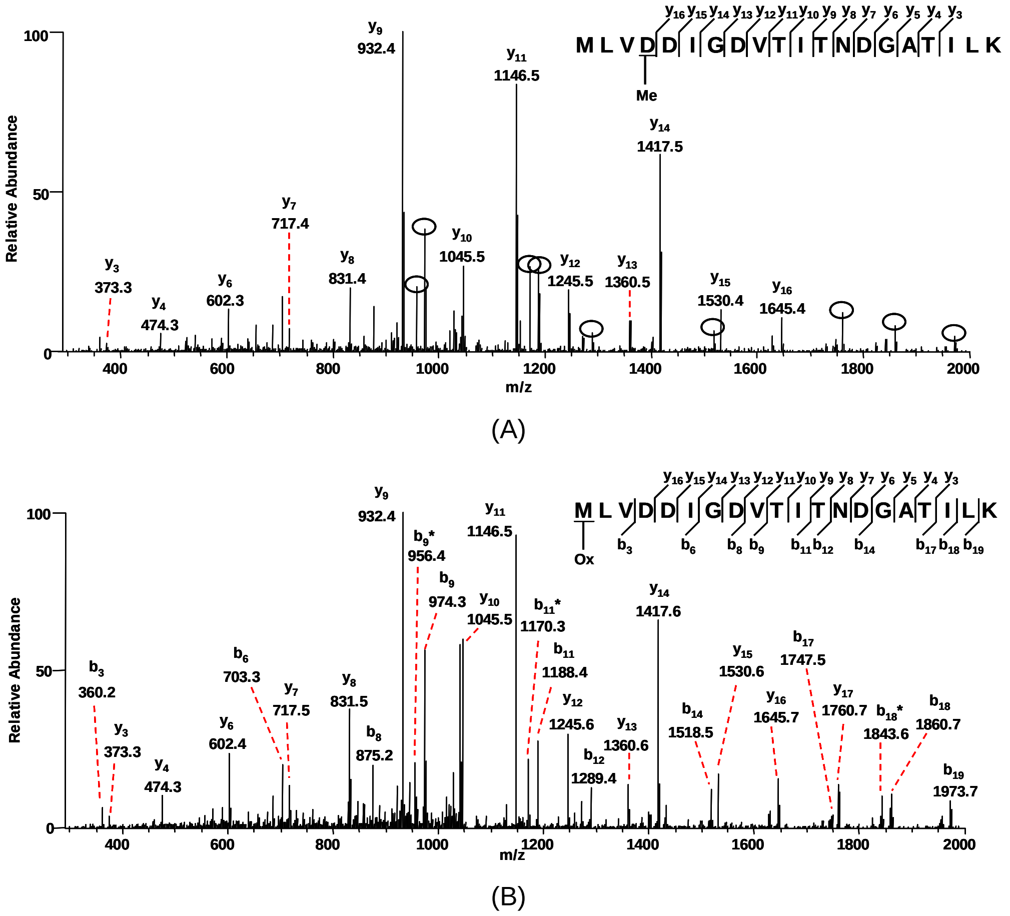
In one example from our lab, Mascot used an MS/MS spectrum to identify a peptide from human androgen receptor as the phosphopeptide "DILSEAS^{p}TMQLLQQQQQEAVSEGSSSGR" with a Mascot score of 66 (Figure 29A). Manual analysis showed that the spectrum was actually from the peptide "QLLQQQQQEAVSEGSSSGR" (Figure 29B), which resulted from non-specific enzymatic activity of trypsin. In a second example, Mascot identified a peptide from human t-complex polypeptide 1 as the D-methylated peptide "MLV^{Me}DDIGDVTITNDGATILK" with a Mascot score of 71 (Figure 30A) when Met oxidation was not specified as a variable modification, but in fact the spectrum is from the peptide “^{Ox}MLVDDIGDVTITNDGATILK” (Figure 30B), which bears an oxidation at M1 that was not specified in the Mascot search. In both cases, the false positive sequence alignments were statistically significant because of well-matched y ion series. Since the peptide sequences of the false positives shared with the true hits, these false identifications cannot be estimated by those strategy based on decoy database.



(A)

(B)

Figure 29 One example of annotated MS/MS spectra of true and false positive identifications of peptides bearing PTMs. (A) The MS/MS spectrum (precursor m/z =1030.87) that Mascot incorrectly matched (score = 66) with the peptide “DILSEAS^{p}TMQLLQQQQQEAVSEGSSSGR” (+3 charge) with a phosphorylation site at T8 in androgen receptor (gi|113830); (B) The MS/MS spectrum (precursor m/z =1030.87) of synthetic unmodified peptide “QLLQQQQQEAVSEGSSSGR” (+2 charge) showing the same fragmentation pattern as in (A).



(A)

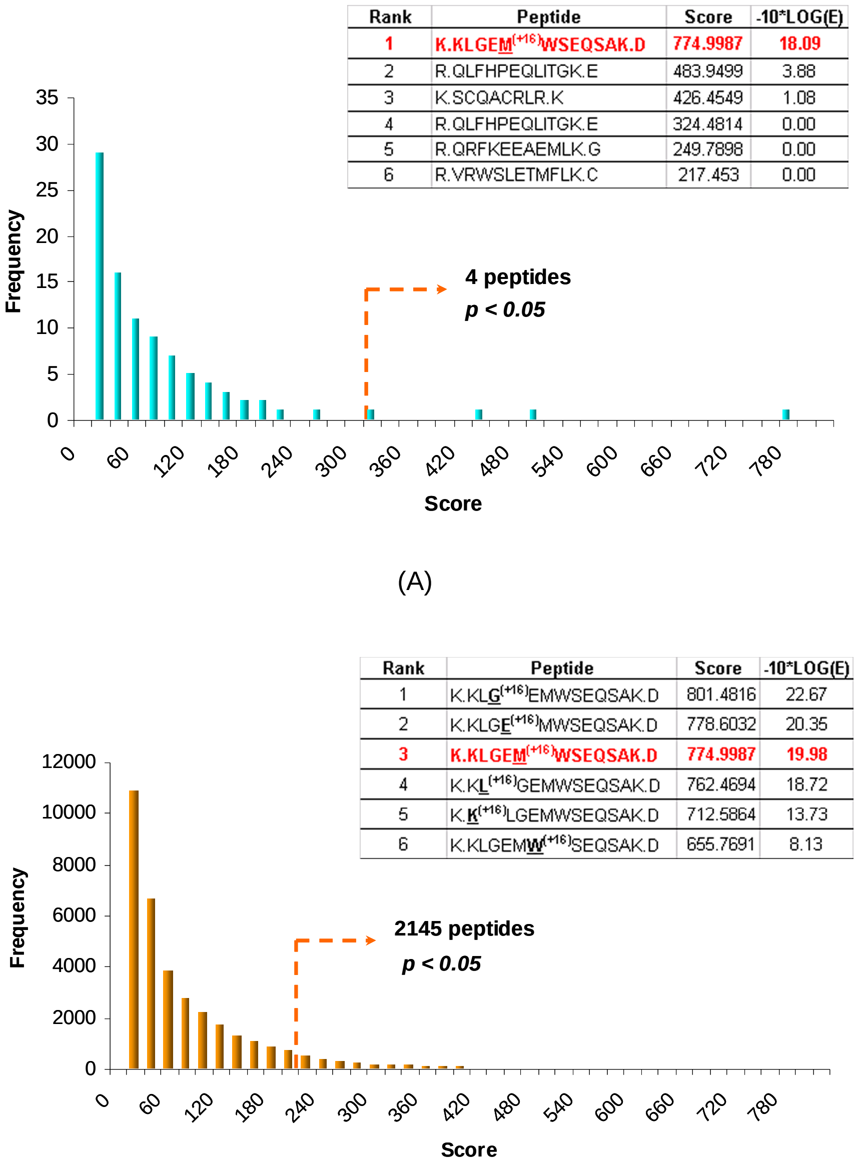
(B)

Figure 30 Another example of annotated MS/MS spectra of true and false positive identifications of peptides bearing PTMs. (A) The MS/MS spectrum (precursor m/z =1061.05) that Mascot incorrectly matched (score = 71) with the peptide "MLV^{Me}DDIGDVTITNDGATILK" (+2 charge) with a methylation site at D4 in t-complex polypeptide 1 (gi|36796); (B) The MS/MS spectrum (precursor m/z =1061.05) of the synthetic peptide "^{Ox}MLVDDIGDVTITNDGATILK" with M1 oxidation (+2 charge) showing the same fragmentation pattern as that in (A).

This example highlights a major problem in sequence alignment. Even with high statistical score, a peptide identification could be wrong. In unrestrictive sequence alignment, such false positive PTM identification could become much more prevalent. To illustrate this problem, we performed a simulation with an MS/MS spectrum of a synthetic peptide bearing a methionine oxidation. We searched the spectrum against a small protein database in a restrictive approach with the specification of only methionine oxidation. The quality of the sequence alignment is evaluated by a score calculated by the total number of matched ions multiplying and the total intensity of matched ions divided by the peptide length. All the search scores were plotted and statistically analyzed. Only four peptide identifications have raw scores above the threshold, where p value is less than 0.05 (Figure 31A). The correct peptide was identified as the top hit with expectation value score significantly higher than the second best peptide. When we performed an unrestrictive database search with the same spectrum and the protein database allowing any modification with mass shift ranging from -100 to +200 Da, we found three interesting results (Figure 31B). First, the total number of potential candidate peptide sequences with p value less than 0.05 is exponentially increased; Second, peptide isoforms with different oxidation site in the same peptide share very similar scores with high statistical significance. Third, the top hit is no longer the correct peptide. This result suggests that false positives can be exponentially increased in unrestrictive analysis.

4.3.2 Our strategy for removing false positives

The exponential increase in the number of peptide sequences that results from not restricting potential PTMs, and the high degree of similarity among the sequences, will undoubtedly lead to large numbers of false positives that cannot be efficiently eliminated by focusing on matched peaks alone. We reason that a correctly identified peptide should explain all major peaks in an MS/MS spectrum. Accordingly, we developed a new concept that identification of false positives should focus on unmatched peaks, while matched peaks can provide a list of candidate identifications.



Rank	Peptide	Score	-10*LOG(E)
1	**K.KLGEM$^{(+16)}$WSEQSAK.D**	**774.9987**	**18.09**
2	R.QLFHPEQLITGK.E	483.9499	3.88
3	K.SCQACRLR.K	426.4549	1.08
4	R.QLFHPEQLITGK.E	324.4814	0.00
5	R.QRFKEEAEMLK.G	249.7898	0.00
6	R.VRWSLETMFLK.C	217.453	0.00

(A)

Rank	Peptide	Score	-10*LOG(E)
1	K.KL**G**$^{(+16)}$EMWSEQSAK.D	801.4816	22.67
2	K.KLG**E**$^{(+16)}$MWSEQSAK.D	778.6032	20.35
3	**K.KLGEM$^{(+16)}$WSEQSAK.D**	**774.9987**	**19.98**
4	K.K**L**$^{(+16)}$GEMWSEQSAK.D	762.4694	18.72
5	K.**K**$^{(+16)}$LGEMWSEQSAK.D	712.5864	13.73
6	K.KLGEM**W**$^{(+16)}$SEQSAK.D	655.7691	8.13

(B)

Figure 31 Simulation of statistic analysis of sequence alignment in restrictive and unrestrictive PTM identifications. (A) Statistic distribution for Restrictive PTM analysis listed with the top 6 peptide identifications with the best statistic scores and the true peptide is ranked the 1st; (B) Statistic distribution for Unrestrictive PTM analysis listed with the top 6 peptide identifications with the best statistic scores and the true peptide is ranked the 3rd.

Toward this goal, we developed the PTMap sequence alignment algorithm for mapping sites of unspecified PTMs and identifying protein polymorphisms (Figure 32). This algorithm incorporates several approaches to improve search speed and reduce false positives, while identifying a large percentage of true positives. First, PTMap searches only protein sequences of interest that have already been identified by general software such as Sequest or Mascot. This targeted sequence analysis reduces the incidence of false matching between MS/MS data and irrelevant protein sequences. Second, PTMap identifies peptide candidates based on matched peaks and then removes false positives from the candidate list based on unmatched peaks. This filtering process is superior to statistical methods, which will typically sacrifice sensitivity of peptide identification for identification accuracy. Third, identification is only considered true if the PTM site is unambiguously determined. We reason that the precise location of a PTM site is critical in unrestricted sequence alignment because of the large number of possible modified peptides with the same sequence.

To further improve the performance of PTMap, we incorporated three additional strategies (Figure 32): (i) PTMap selects only those peaks with signal intensities that are significant relative to the local noise; (ii) PTMap uses only monoisotopic peaks for sequence alignment; and (iii) PTMap automates the adjustment of mass shifts. The last feature allows PTMap to analyze data from low-resolution mass spectrometers in addition to the data from high-resolution mass spectrometer.

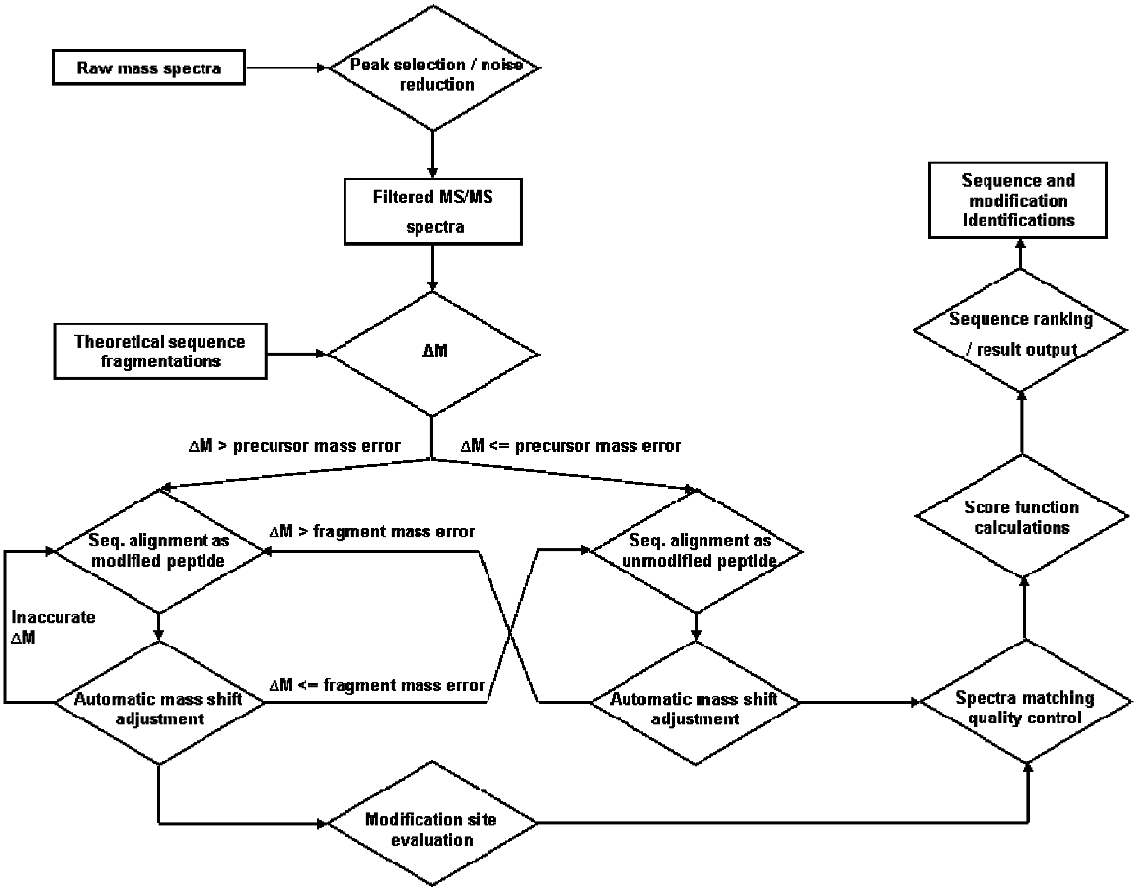

Raw mass spectra
Peak selection / noise reduction
Filtered MS/MS spectra
Theoretical sequence fragmentations
ΔM
ΔM > precursor mass error
ΔM <= precursor mass error
Seq. alignment as modified peptide
ΔM > fragment mass error
Seq. alignment as unmodified peptide
Inaccurate ΔM
Automatic mass shift adjustment
ΔM <= fragment mass error
Automatic mass shift adjustment
Spectra matching quality control
Modification site evaluation
Score function calculations
Sequence ranking / result output
Sequence and modification Identifications

Figure 32 Flow chart of the PTMap algorithm.

4.3.3 Selectivity and false discovery rate of PTMap analysis

A high $S_{Unmatched}$ score suggests a large number of unmatched peaks with significant intensities. For a singly charged precursor ion, only one $S_{Unmatched}$ score is used. For a precursor ion with two or more charges, two $S_{Unmatched}$ scores are used for each spectrum: one for the high mass range (higher than the precursor ion m/z) and one for the low mass range (lower than the precursor ion m/z). Both scores must be satisfactory for a positive identification. We routinely use $S_{Unmatched}$ scores of 4.0 and 10.0, for the high and low mass ranges, respectively (notated as 4.0:10.0), to filter out sequence-spectrum alignments of low quality. The $S_{Unmatched}$ score for the low mass range is not as stringent because a greater number of intense noise peaks are usually generated in the low mass range than in the high mass range. To check if the $S_{Unmatched}$ threshold scores will cause the loss of sensitivity of the analysis, we manually analyzed all the Mascot peptide identifications and plotted the $S_{Unmatched}$ score distributions for correct and incorrect IDs. The results demonstrated that the $S_{Unmatched}$ scores (4.0:10.0) were sufficient for the identification of all the correct unmodified peptides (Figure 33A, B).

A difficulty arises from spectra having few fragment ions, because these spectra result in low numbers of both matched and unmatched peaks. Such false positives cannot be identified by the $S_{Unmatched}$ score alone. This issue is addressed by the PTMap score.

To evaluate the usefulness of the PTMap score as a second parameter to remove false positives, we searched each MS/MS dataset against the true or randomly scrambled sequence of the corresponding protein. We used 4.0:10.0 as threshold $S_{Unmatched}$ scores, and then generated PTMap scores. PTMap scores for identifications of unmodified peptides from both normal and scrambled protein sequences were calculated (Figure 34A). When a PTMap score cutoff of 0.50 was used, the false discovery rate was about 1.6% for unmodified peptides (Figure 34A).

Next, we tested whether we could use this cutoff score to evaluate false positives of all peptide identifications, including those containing PTMs. We generated PTMap scores for all peptide identifications (Figure 34B). The number of peptides with scores below the 0.50 cutoff was much higher for modified peptides than for unmodified peptides (Figure 34A, B), suggesting a large number of low quality sequence-spectrum pairs derived from unspecified PTMs. Analysis of search results with scrambled protein sequences (Figure 34B) shows the false discovery rate for searches including modified peptides was 20.9% using a cutoff score of 0.50. To improve the accuracy of peptide identification, we increased the cutoff score for all peptide identifications to 1.00 for unrestrictive PTM analysis, which caused a loss of sensitivity of about 18% and an improvement in the false discovery rate to 3.6%.

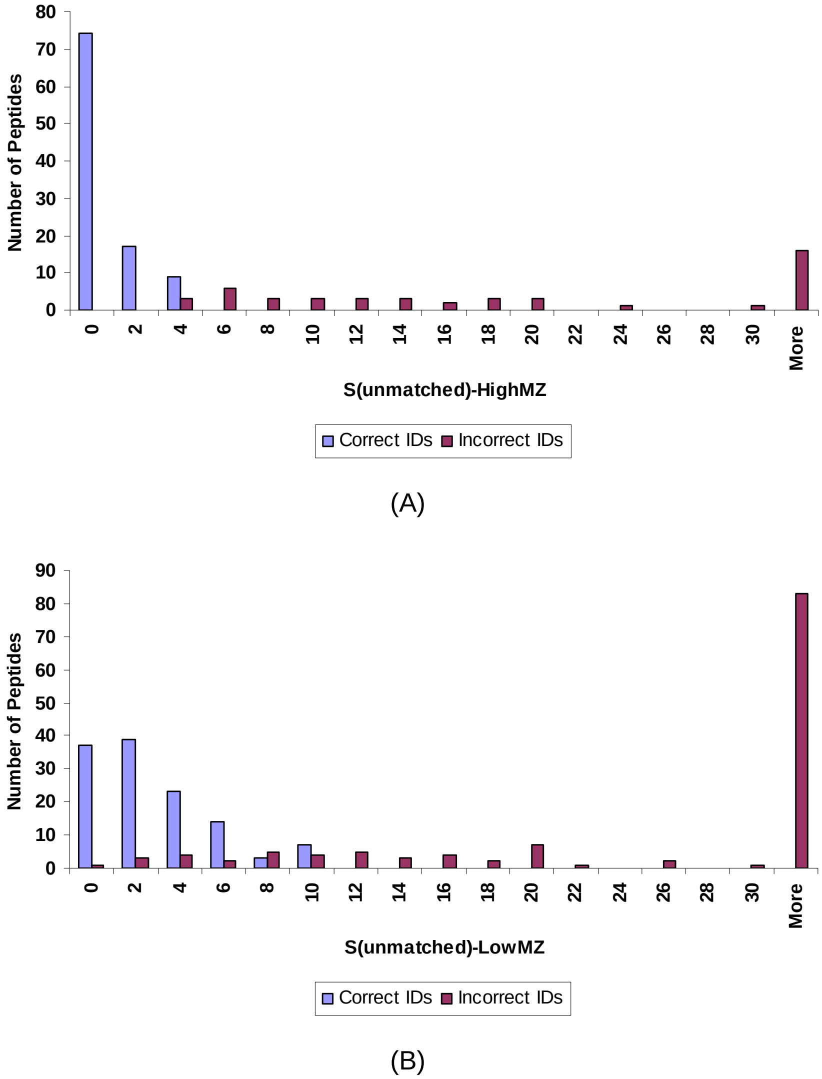
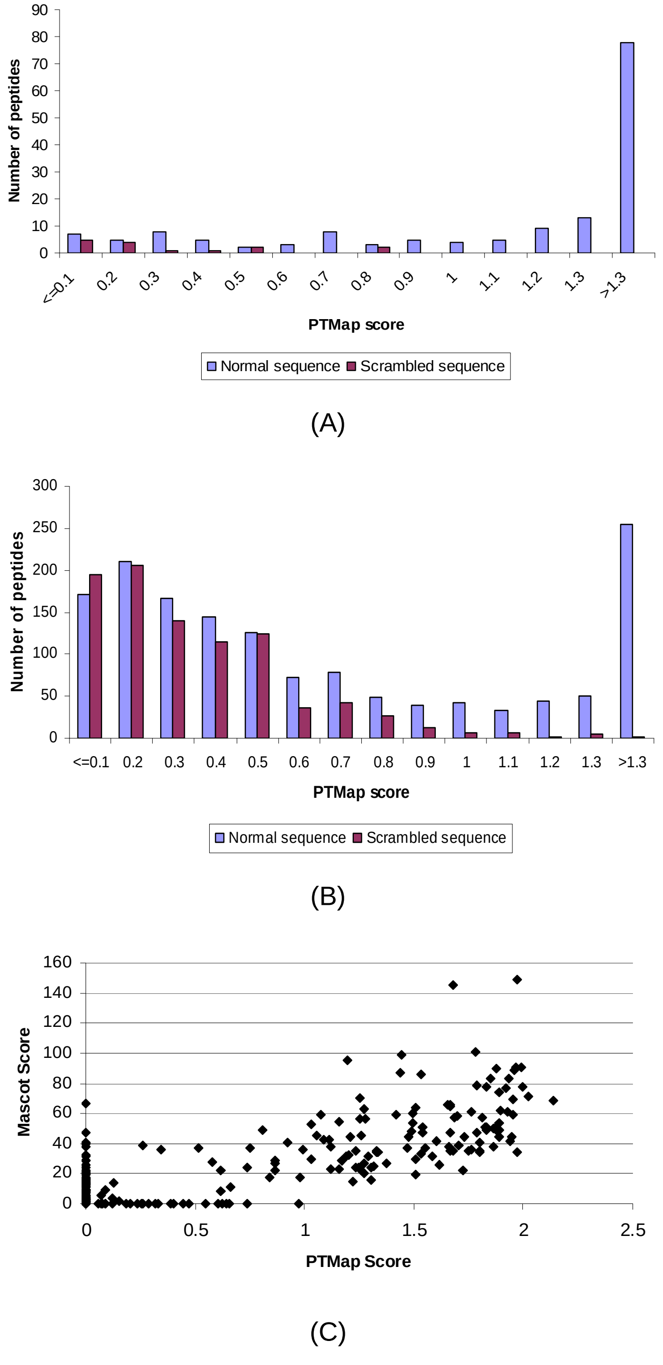


Figure 33 $S_{Unmatched}$ score distribution analysis for the Mascot peptide identifications. The bar diagrams of correct and incorrect $S_{Unmatched}$ distributions in high mass range (above precursor m/z) (A) and low mass range (below precursor m/z) (B) are shown.



(A)

(B)

(C)

Figure 34 (A) PTMap score distributions ($S_{Unmatched}$ = 4.0:10.0 in high and low mass ranges respectively) of unmodified peptides identified from normal and scrambled protein sequences of the four selected proteins – histone H4, SGK1, HMG2 and BSA using PTMap; (B) PTMap score distributions ($S_{Unmatched}$ = 4.0:10.0) of both unmodified and modified peptides identified from normal and scrambled of the same four protein sequences using PTMap; (C) Correlation of Mascot scores with PTMap scores ($S_{Unmatched}$ = 4.0:10.0) for identification of unmodified peptides.

4.3.4 Sensitivity of PTMap analysis

To evaluate the sensitivity of the PTMap score for detecting true peptide identifications, we compared the identification results of Mascot and PTMap for unmodified peptides in restrictive analysis (Figure 34C). Because each peptide was usually identified by multiple spectra, only the identifications with the best scores were plotted. When using Mascot, a score of 40 is typically used as a cutoff score for peptide identification. Mascot analysis identified 71 peptides with scores above the threshold score of 40 after manual verification. Over 90% of these peptides had high PTMap scores above 1.0 and all have PTMap score above 0.5 threshold for restrictive analysis. PTMap was able to identify an additional 40 peptides whose PTMap scores were above 1.0 and Mascot scores were below 40, suggesting that the PTMap algorithm was able to boost the sensitivity of the identification by 57%. To evaluate the accuracy with which peptides were identified, we performed MS/MS analysis of corresponding synthetic peptides for 8 of the 40 peptides with Mascot scores from 15 to 27. The results showed that the fragmentation patterns of synthetic peptides completely matched the experimental MS/MS spectra and confirmed the identification of these peptides by PTMap.

4.3.5 Strategies used by PTMap to increase accuracy of peptide identification

PTMap implements three main strategies to improve searching speed and accuracy of peptide identification – peak selection, automatic mass shift adjustment and precise localization of PTM sites. We systematically evaluated the effectiveness of each strategy. The results demonstrate that the application of these strategies significantly reduces the false discovery rate and boosts the sensitivity of peptide identification.

Peak selection PTMap selects only those monoisotopic peaks that are significant when compared with local noise levels. Because noise levels are usually not homogeneous across the whole mass range, the local noise level is used when selecting meaningful peaks. In addition, isotopic peaks do not contain extra sequence information and are therefore removed. To evaluate the effectiveness of peak selection for peptide identification, we compared the distribution of PTMap scores of the identified unmodified peptides with or without peak selection.

Because noise peaks are different with or without application of peak selection, we used PTMap score instead of $S_{Unmatched}$ to evaluate the utility of peak selection. The $S_{Unmatched}$ threshold was therefore relaxed to 400 in this analysis. PTMap analysis of four protein LC/MS/MS data sets without the application of peak selection identified 147 non-redundant peptides, compared to 126 peptides with the application of peak selection (Figure 35; PTMap cutoff score=0.50, for restrictive analysis at 1.6% false discovery rate). The 21 additional peptides identified in the absence of peak selection were manually inspected using the stringent rules described previously [41] and were found to be false positives.

Manual examination of the false positive spectra found that the false positive results were mainly caused by matching to (i) peaks in a mass range of 0 to -50 Da relative to the parent ion; (ii) noise signals with low intensities; and (iii) isotope peaks. Applying peak selection strategies allows PTMap to efficiently eliminate these types of false positives, thereby increasing the accuracy of identification. Additionally, by considerably reducing the total number of peaks, peak selection improves searching speed.

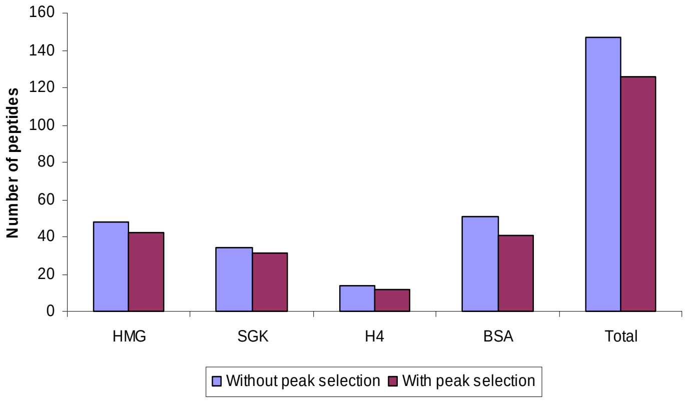
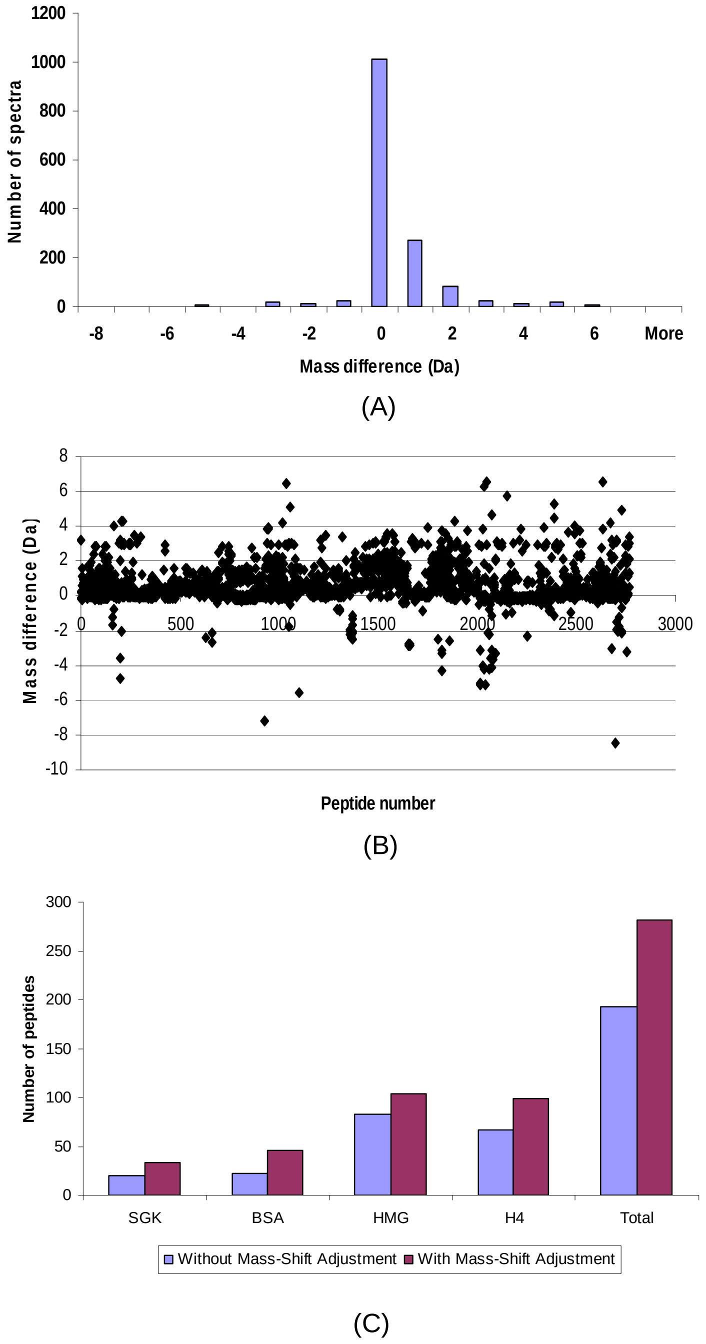


Figure 35 The number of unmodified peptides (PTMap score cutoff=0.5) identified with or without incorporation of the peak-selection function in PTMap



(A)

(B)

(C)

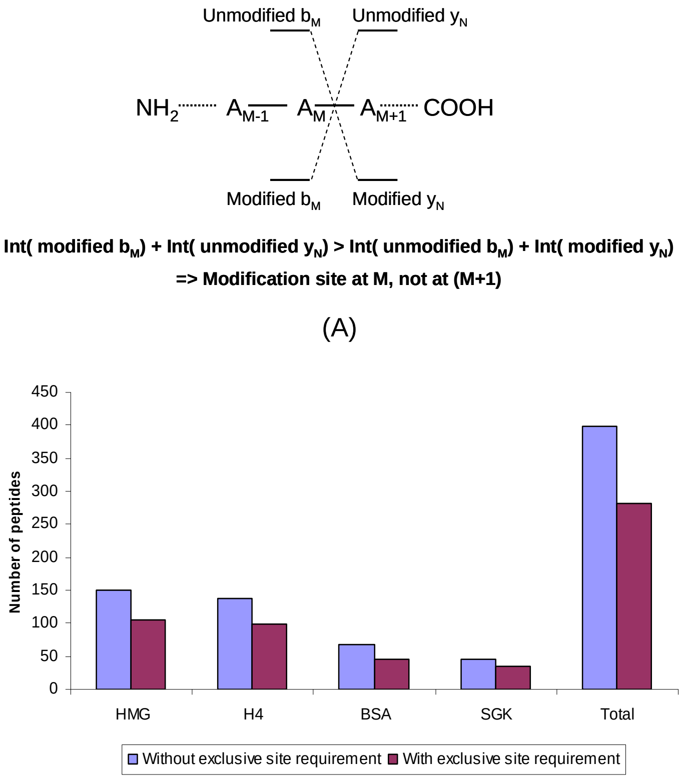
Figure 36 (A) The distribution of the mass changes (ΔM_{after} - ΔM_{before}) made by PTMap after automatic mass-shift adjustment for all identified spectra bearing PTMs (PTMap score cutoff=1.0); (B) The number of modified peptides identified with or without automatic mass-shift adjustment strategy in PTMap (PTMap score cutoff=1.0); (C) scatter plot showing the distribution of the mass errors of the spectra that identified unmodified peptides in the four proteins (PTMap score cutoff=0.5).

Automatic mass-shift adjustment To search for undefined PTMs, PTMap scans all mass shifts from -100 to +200 Da in 1-Da increments. In low-resolution mass spectrometers, such as those with the popular ion trap mass analyzer, mass errors are much lower in MS/MS spectra (usually <0.6 Da) than in the corresponding MS spectrum (usually <4.0 Da). High mass errors for precursor ions arise because of low-resolution instruments, space charge in ion trap mass spectrometers, use of isotopic peaks as precursor ions, and identification of parent ions from co-incident ions. A large mass error for a precursor ion, if not corrected, could result in identification of a PTM with an incorrect mass shift, or improper identification of the nature of the PTM. To address this concern, PTMap includes a function determining the consensus mass shift based on the MS/MS spectrum. Then the spectrum is re-aligned, using this consensus mass shift, before spectral alignment quality is evaluated. Automatic mass-shift adjustment allows identification of unmodified peptides with large precursor ion mass errors. We believe that this feature is also useful for analysis of MS/MS spectra from high-resolution instruments because of the possibility that precursor ions may be assigned to isotopic peaks.

To test the effectiveness of this strategy for identification of PTMs, we examined the MS/MS spectra of the identified peptides bearing PTMs (PTMap cutoff score=1.0, for unrestrictive analysis at 3.6% false discovery rate). The mass differences between the calculated mass shift and the adjusted mass shift were calculated and the distribution of all the mass differences was plotted (Figure 36A). Among the total number of 1476 identified spectra, 462 spectra or 31.3% were identified with the application of automatic mass shift adjustment. The total number of peptide identifications with PTMs increases by 45.6% after the application of this strategy (Figure 36B). We also examined the mass errors of precursor ions of unmodified peptides (Figure 36C). The results show that PTMap was able to identify unmodified peptides with large precursor ion mass errors.

Exclusive site localization Localization of the site of modification within a modified peptide can be difficult, because peptides that differ only by the site of modification will give highly similar theoretical fragmentation patterns. Incorrect localization of PTM sites will dramatically increase the number of modified peptides identified, a problem that has not been previously addressed. PTMap incorporates a two-step procedure enabling PTM sites to be exclusively located. First, it is assumed that the candidate PTM may reside on any residue of the peptide; theoretical fragmentation patterns of these peptides are aligned with the MS/MS spectrum iteratively. The resulting sequence alignments are compared with each other, and PTMap identifies the peptide isoform with the best PTMap score. If peptides with adjacent modification sites have identical PTMap scores, a strategy illustrated in Figure 37A is used to define the PTM site. For the modification site to be localized to position M and not M+1 (Figure 37), PTMap requires that the total intensity of the two PTM-relevant fragment ions, modified bM and unmodified yN, be higher than that of unmodified fragment bM and modified fragment yN. Second, PTMap requires the PTM site to be identified by consecutive ions in the b or y ion series, or by the simultaneous appearance of modified b and y ions in which the modified residue is the end residue of each fragment. The site of modification is considered identified if both conditions are met, while hits failing at least one of the filters are considered ambiguous and are removed.

To illustrate the effectiveness of this approach, we compared the peptide modification analysis with or without these criteria (Figure 37B, PTMap cutoff score=1.0, for unrestrictive analysis at 3.6% false discovery rate). Without exclusive site localization, PTMap identified a total of 399 modified peptides, in large contrast to only 282 peptides when the exclusive site localization requirements were applied. This represents a 29.3% reduction of false positives for modified peptides when ambiguous modification site assignments were removed.



(A)

(B)

Figure 37 (A) A strategy for precise mapping of a PTM site that will distinguish two peptide isoforms that are modified on adjacent sites; (B) The number of modified peptides identified by PTMap before and after the implementation of exclusive site localization.

4.3.5 Identification of modified peptides in HMG2, human histone H4, mouse SGK1, and bovine serum albumin

PTMap identified a total of 282 non-redundant modified peptides with a PTMap score higher than 1.0 from the MS/MS datasets of human histone H4, HMG2, mouse SGK1, and BSA. We removed peptides with PTM sites on the N- or C-terminal amino acid for two reasons. First, LTQ mass spectrometer used in our study has low mass cutoff that eliminates modified b1 or y1 ions whose m/z values are below the cutoff. Second, terminal modifications cannot be mapped through consecutive b or y ion series, raising the possibility that the m/z of identified b1 or y1 ions coincide with the nominal m/z of a combination of terminal amino acids (*e.g.* m/z of one Asn is equal to the total m/z of two Gly) and causing false-positive PTM identification. From this dataset, we further removed redundant peptide sequences that identified the same PTM sites.

The final results give 156 modified peptides, among which 56 modified peptides were identified by at least 3 spectra. Seventy-seven of these PTMs (~50%) can be annotated using the Unimod database (http://www.unimod.org). From the data, we found that the most commonly modified residues were methionine (M), lysine (K), cysteine (C), histidine (H) and glutamic acid (E).

For 18 out of the 156 peptides (~11%), the unmodified peptide counterparts were not observed, which can be explained by three possibilities. First, the modification may have been caused by common chemical reactions (*e.g.*, Cys alkylation with acrylamide or iodoacetamide) with high reaction yields. Second, the modification might have completely prevented cleavage by the protease (*e.g.*, lysine acetylation). Third, some identified mass shifts were the result of protein point mutations.

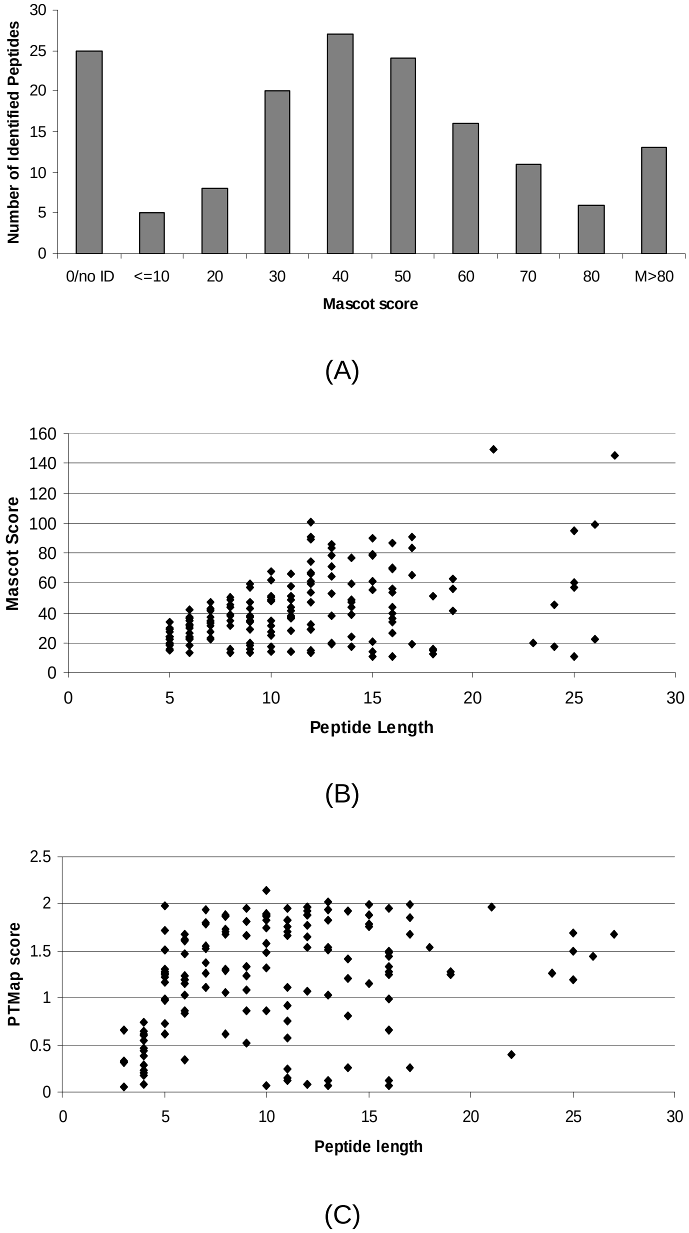
To distinguish among the three possibilities, we synthesized 17 peptides containing a mutation, based on the mass shifts, and then performed MS/MS analysis of the peptides. The MS/MS spectra of 6 protein-derived peptides and their synthetic counterparts were almost same, implying that these protein polymorphisms identified by PTMap are true. The remaining 11 MS/MS spectra did not match those of the synthetic peptides, suggesting that the identified mass shifts on these peptides come from unknown PTMs.

Post-translational modifications on histone proteins play critical roles in chromatin structure and gene regulation [92]. In this study, we identified a total of 110 peptides from histone H4, among which 99 peptides were identified as modified peptides through unrestrictive analysis with PTMap cutoff score of 1.00. After eliminating the PTMs on the N- or C- terminals of the peptides, whose PTM sites cannot be precisely located, we conclusively identified a total of 64 non-redundant PTMs on histone H4 with high confidence.

4.3.6 PTMap score reflects sequence alignment more accurately than Mascot score

To compare PTMap scores and Mascot scores as indicators of peptide identification, the four LC/MS/MS datasets were input into the Mascot algorithm to search a very small FASTA database with only about 200 protein sequences. The unmodified peptides identified with Mascot scores of over 40 were manually evaluated. All peptides confidently identified by Mascot were also identified by PTMap (Figure 34C). Mascot scores for unmodified peptides that were identified by PTMap showed a range from 0 (or unidentified) to over 100, with a sizable proportion of peptides falling between 20 and 60 (Figure 38A). The Mascot score distribution pattern is noticeably different from the PTMap score distribution pattern for the same peptide set (Figure 34A). PTMap scores for correct peptide identifications tend to aggregate in the high score range.

To gain insight into the differences between the two score distributions, we studied the correlation between peptide length and Mascot identification score. According to our experience with the Mascot algorithm, longer peptides usually give more fragmentation peaks, leading to a higher Mascot score and a greater degree of statistical significance. In agreement with our impressions, peptide length seems to positively correlate with Mascot score (Figure 38B). As a result, a statistical algorithm will lead to inaccurate evaluation of the strength of a match for both long and short peptides. Accordingly, true positives were identified for peptides given low or high Mascot scores, suggesting the need for manual verification [41]. In contrast, PTMap scores, which are based on both matched and unmatched peaks, are independent of sequence length except for very short peptides with fewer than 5 amino acids (Figure 38C). Accordingly, PTMap score provides a more accurate and objective parameter for evaluation of the accuracy of peptide identification.



(A)

(B)

(C)

Figure 38 (A) Mascot score distribution for unmodified peptides identified by PTMap; (B) and (C), correlation between peptide length and Mascot score or PTMap score, respectively. Only those peptide identifications with a minimum Mascot score of 10 (B) or minimum PTMap score of 0.01 (C) were included.

4.3.7 Comparison between PTMap with other unrestrictive PTM analysis softwares, MS-Alignment and P-Mod

P-Mod software To compare the P-Mod's performance with PTMap, we downloaded the P-Mod software (version II) from Vanderbilt software resource website and analyzed the LC-MS/MS data of tryptic digest of HMG. Three missing cleavages were allowed for tryptic peptides, the same specification for PTMap. The total time used for PTMap analysis is less than 5 minutes while P-Mod used over 3 hours for the same dataset on the same computer. P-Mod software identified a total of 296 non-redundant modified peptides with FDR <3.6%, while PTMap identified a total of 107 modified non-redundant peptides with the same FDR. We manually analyzed the top 50 best-scored peptides by the two softwares using a procedure we previously described [41]. Our results show that while top 50 hits by PTMap can be manually verified, only 16 modified peptides of the top 50 hits by P-Mod can pass the stringent manual verification procedure.

MS-Alignment software We evaluated MSAlignment (version 2008.0404) also using the LC-MS/MS data from tryptic digest of HMG. We carried out the sequence alignment, using the procedure described by Pevzner's group [108], for installing the software and performing data analysis. The number of MS/MS spectra in the dataset is large enough (>22,000 spectra) to allow the software to have a good statistic analysis. The total time used for PTMap analysis is less than 5 minutes while it took over 25 minutes for the same dataset on the same computer. The sequence alignment by MSAlignment identified a total of 315 modified non-redundant peptides with FDR <3.6%, while PTMap identified a total of 107 modified peptides with the same FDR. We chose to manually analyze the top 50 best hits bearing protein modifications using the same procedure as Part I. Our results showed that all top 50 hits identified by PTMap can be manually verified, while only 18 peptides among the top 50 peptides by MSAlignment passed stringent manual verification. Thus, over 60% of the top 50 identifications from MSAlignment are false positives.

Taken together, our evaluation of PTMap with P-Mod and MSAlignment suggests that PTMap have superior performance over two other known softwares for unrestrictive PTM analysis, MS-Alignment and P-Mod, in terms of identification accuracy and speed.

4.4 Discussion

Minimization of false positives is always accompanied by a sacrifice of sensitivity of peptide identification in sequence alignments based on statistical methods. In contrast, our results demonstrate that PTMap can identify 57% more peptides than statistical methods, while achieving higher accuracy as indicated by the identification of peptides with low Mascot scores. Therefore, PTMap addresses a major problem with statistics-based algorithms, such as Mascot.

Some PTMs lead to unique fragmentation patterns. For example, phosphopeptides tend to generate daughter ions with a neutral loss. PTMap can take such unique fragmentation patterns into consideration to boost the efficiency of identifying peptides with such properties.

While PTMap was used in the analysis of a single, pure protein in our case studies, the algorithm can be easily expanded to the analysis of an MS/MS dataset from complex protein mixtures (*e.g.*, >100 proteins) by initial protein identification and subsequent mapping of PTMs. While more than 200 types of PTM have been identified, the abundance and scope of these PTMs, and interrelationships between them, remain largely unknown. Recent identification of two novel PTMs, lysine propionylation and lysine butyrylation [42], suggest that important undescribed PTMs remain to be discovered. PTMap will provide a powerful technology platform for chemical dissection of cellular PTM networks.

4.5 Concluding Remarks

In this chapter, I described the development of PTMap software for accurate identification of full-spectrum PTMs and polymorphisms. PTMap incorporates three unique features to improve its function: MS peak selection, automatic mass-shift adjustment, and exclusive site localization. Two logical score systems, $S_{Unmatched}$ and PTMap score, were developed and were demonstrated to be accurate for evaluating peptide identifications. To remove false positives, the algorithm stresses unmatched peaks and incorporates stringent manual-verification rules. To our knowledge, this strategy has not been previously described for removing false positives. Accordingly, PTMap will provide a powerful tool for systematic analysis of protein modifications.

4.6 Material and Methods

4.6.1 Preparation of human histone H4, HMG2, mouse SGK1 and BSA

Human histone H4 was purified from HeLa cells as previously described [109], except that HDAC inhibitors (2 μM TSA, 50 mM sodium butyrate, and 30 mM nicotinamide) were added during preparation of the core histones to prevent histone deacetylation, depropionylation and debutyrylation. The mouse SGK1 protein was purified as previously described [110]. Human HMG2 and bovine serum albumin (BSA) were purchased from ProteinOne Inc. (Bethesda, MD) and Sigma Inc. (St. Louis, MO), respectively. Synthetic peptides were synthesized by GL Biochem (Shanghai, China) and Genemed Synthesis (San Antonio, TX).

4.6.2 In-gel and in-solution digestion with trypsin

Proteins of interest were resolved on SDS-PAGE and visualized by colloidal Coomassie blue staining. The protein bands were excised, in-gel digested, and extracted as previously described [96]. The extracted peptides were pooled, dried in a SpeedVac, and desalted using a μ-C18 Ziptip (Millipore, Billerica, MA) before HPLC/MS/MS analysis.

To generate tryptic peptides of BSA, about 2.5 μg BSA was solubilized in 50 μl of 50 mM ammonium bicarbonate solution (pH 8.0) and digested in solution with 100 ng of modified porcine trypsin for five hours at 37 ℃. The reacti on was stopped by adding 10 μl of 10% trifluoroacetic acid. The tryptic peptides were dried in a Speed-Vac and desalted with μ-C18 Ziptip before HPLC/MS/MS analysis.

4.6.3 HPLC/MS/MS analysis and protein sequence database searching

Peptide samples were dissolved in 3 μl HPLC buffer A (acetonitrile:water:acetic acid = 2:97.9:0.1, v/v/v) and injected into an Agilent 1100 nanoflow HPLC system (Agilent, Palo Alto, CA). Peptides were eluted with a gradient of 10% to 90% HPLC buffer B (acetonitrile:water:acetic acid = 90:9.9:0.1, v/v/v) on a home-made capillary column (10 cm length X 75 µm ID) packed with Jupiter C12 resin (4 µm particle size, 90 Å pore diameter, Phenomenex, St. Torrance, CA). Peptides were directly electrosprayed into a LTQ mass spectrometer using a nano-spray source; MS/MS spectra were acquired in centroid mode by a data-dependent instrument method with 4 m/z isolation window and 32% normalized collision energy. Tandem mass spectra were analyzed using PTMap or Mascot software with Met oxidation specified as a variable modification, trypsin as the proteolytic enzyme, 3 missing cleavages allowed, 4-Da precursor mass error and 0.6-Da fragment mass error.

4.6.4 Algorithm description

PTMap aligns tandem mass spectra with each peptide sequence from a single protein, each residue of which is theoretically modified by a mass shift ranging from -100 to +200 Da in 1-Da increments. The mass shift range can be adjusted as desired. Because neither the modified residue nor its modification (as defined by a mass shift) is specified, the algorithm can carry out unrestricted identification of PTMs. The processes can be reiterated so that the sequence alignment can be performed for multiple proteins of interest against the same MS/MS dataset.

We minimized false positive rates and improved the search speed of PTMap using four steps: MS/MS peak selection, automatic mass-shift adjustment, automatic evaluation of peptide identification results, and data output and visualization. Each step is described in detail below.

Part I. Peak selection

Because the masses of mass spectrometric peaks are the basis for peptide identification, peak quality is critical for accurate identification. In addition to monoisotopic peaks that are directly relevant to the peptide sequence of interest, an MS/MS spectrum also has other isotopic peaks and noise peaks that are derived from electronics and unrelated chemicals and do not contain additional sequence information. These isotopic and noise peaks should be removed to reduce false positives and searching time. In addition, peaks that are peptide-derived and hence significant should be distinguished from noise peaks.

Toward this goal, PTMap processes MS/MS spectra in three steps. First, an MS/MS spectrum is divided into small sections, each of which contains 60 peaks. Each section is typically less than 30 Da wide. A local noise level is determined by averaging the peak intensities of the 40 ions with the lowest intensities in the section. A user-defined signal-to-noise ratio is used along with the local section noise calculated above to calculate a minimum intensity required to select MS peaks that will be used for sequence alignment. Our simulation experiments suggest that a four-fold signal-to-local-noise ratio is sufficient to remove most noise peaks. Those ions that have low intensities are considered noise peaks and are therefore removed. Second, the software removes all MS peaks within a user-defined mass window surrounding the m/z of the precursor ion to eliminate the interference from unpredictable peptide neutral-loss ions and high noise background in this region. We recognize that the noise peak patterns in the m/z region close to the parent ions are different among singly, doubly and triply charged ions. The typical mass window within which sequence-independent peaks are usually abundant is 100 Da for singly charged precursor ions and 50 Da for doubly or triply charged precursor ions. Third, PTMap removes from the spectrum all isotopic peaks, and all peaks, usually caused by electronic sparks, that lack an adequate isotope pattern.

The average peak intensities are then calculated for those peaks in the high mass range (m/z above precursor m/z) or the low mass range (m/z below precursor m/z). All peaks with intensity greater than half the average peak intensity level are considered to be significant peaks and will be used to calculate the unmatched peak score ($S_{Unmatched}$) and spectrum-matching quality score (S_{match}) as shown in Formulas (1) and (4). These two scores are used to evaluate the quality of spectrum alignment in a process described in Part II below.

Part II. Spectral alignment

PTMap generates theoretical proteolytic peptides *in silico* based on enzyme specificity, the user-defined number of missed cleavages, and specified variable modifications. Incorporation of the specified variable PTMs enables the algorithm to identify more than one type of PTM in a peptide. Those theoretical peptides within the mass range of 350 to 6000 Da are used for spectral alignment. For each theoretical peptide, PTMap generates an m/z table of a, b and y fragment ions with +1, +2, or +3 charges.

Three parameters are used to align each spectrum with theoretical peptides: Δm, the mass difference between the peptide and the precursor ion, which represents a potential PTM-caused mass shift; E_p mass errors of precursor ions; and E_f, mass errors of fragment ions. Both E_p and E_f are dependent on the mass spectrometer and are specified by the user. In our study, we used a Δm range from -100 Da to +200 Da. The sequence-spectrum alignment involves two interconnected steps:

(i) If the calculated Δm is lower than E_p, the software aligns the spectrum with the theoretical peptide fragmentation assuming there is no PTM. After sequence alignment, PTMap evaluates the results and calculates an average mass error based on fragmentation ions. If the averaged mass error of the fragment ions is larger than E_f, it is rounded to the nearest integer and is considered to be a potential peptide modification. Then PTMap will re-align the spectrum to the modified peptide as described in part (ii).

(ii) If Δm is higher than E_p, the mass shift is rounded to the nearest integer and treated as a potential modification. To locate the modification site, theoretical fragmentation patterns are calculated based on the assumption that the mass shift can possibly occur at any position in the peptide. Then the software performs alignments with each fragmentation pattern iteratively. After sequence alignment, PTMap evaluates the results as described in part (i). If the averaged mass error of the fragment ions is larger than E_f, the mass shift is adjusted accordingly and the spectrum is re-aligned. However, if the averaged mass shift of the fragment ions after adjustment falls within the E_f window of the unmodified peptide mass, PTMap will re-align the spectrum to the unmodified peptide as described in part (i).

We call this iterative process based on Δm “automatic mass shift adjustment”. This process enables PTMap to obtain PTM masses based on the MS/MS spectrum, which has much higher mass accuracy than the precursor ion MS spectrum. This strategy allows efficient identification of modifications with small apparent mass shifts (*e.g.*, deamidation of N or Q) despite potentially large errors in the precursor ion spectrum. The described strategy allows identification of PTMs using MS/MS data generated in low-resolution mass spectrometers.

Part III. Sequence verification

PTMap uses two independent scoring systems to evaluate peptide identification: unmatched peak score ($S_{Unmatched}$) and PTMap score. $S_{Unmatched}$ is determined by the number and intensities of unmatched peaks. PTMap score emphasizes mutual explanation of both MS/MS spectrum and its candidate peptide sequence, providing an additional criterion for evaluating the accuracy of peptide identification.

The concept of using the factor $S_{Unmatched}$ to evaluate reliability of peptide identification stemmed from our earlier study on manual evaluation of identified peptides [41], which showed that both the number and intensities of unmatched peaks are the critical parameters for false peptide identification. The factor $S_{Unmatched}$ is related to both the number of unmatched peaks and the averaged peak intensities of all peaks in the spectrum (equation (1)). The latter parameter is intended to normalize variation among MS/MS peak intensities, which are influenced by peptide sequence, the instrument, and the instrument’s operating conditions.

$$S_{Unmatched} = \Sigma\ (\mathrm{Int}_{Unmatched} *2 / \mathrm{Int}_{Average}) \qquad (1)$$

$\mathrm{Int}_{Unmatched}$ represents the intensity of an unmatched peak, while $\mathrm{Int}_{Average}$ represents the average intensity of all selected peaks in the mass range either higher or lower than the precursor m/z. Two $S_{\text{Unmatched}}$ scores are calculated, one for the range of masses higher than the precursor m/z, and the other for the range of masses lower than the precursor m/z. The software requires that $S_{\text{Unmatched}}$ for neither the high mass range nor the low mass range exceed user-defined thresholds

To systematically evaluate PTMap performance, we developed the PTMap score, which is a function of sequence coverage score (S_{seq}) and spectrum score (S_{spec}) (see Formula (2)). S_{seq} represents the percentage of the peptide sequence that can be explained by spectrum fragment ions. S_{spec} (see Formula (3.1 and 3.2)) is a function of the spectrum matching quality scores S_{match} (see Formula (4)), discounted by the noise level factor P_{Noise} . S_{spec} is high when the relative intensity of the matched peaks is high and the spectrum noise level is low. For multiply-charged peptide identification, S_{match} in both high and low mass ranges (above or below precursor m/z ranges) are considered (see Formula (3.1), C1 and C2 are optimized constants), while for the singly-charged peptide identification, only S_{match} in low mass range is considered (see Formula (3.2)). P_{Noise} is determined by the averaged signal-to-noise ratio of the matched peaks.

$$\text{PTMap score} = S_{\text{seq}} \times S_{\text{spec}} \tag{2}$$

$$S_{\text{spec}} = (S_{\text{match(HighMZ)}} \times \text{C1} + S_{\text{match(LowMZ)}} \times \text{C2}) \times P_{\text{Noise}} \tag{3.1}$$

$$S_{\text{spec}} = S_{\text{match(LowMZ)}} \times P_{\text{Noise}} \tag{3.2}$$

$$S_{\text{match}} = \text{T}_{\text{Matched}} * \Sigma\ \text{Int}_{\text{Matched}} / (\text{T}_{\text{Matched}} * \Sigma\ \text{Int}_{\text{Matched}} + \text{T}_{\text{Unmatched}} * \Sigma\text{Int}_{\text{Unmatched}}) \tag{4}$$

In Formula (4), $\text{T}_{\text{Matched}}$ and $\Sigma\text{Int}_{\text{Matched}}$ represent the total number of matched peaks and the sum of their intensities, respectively, while $\text{T}_{\text{Unmatched}}$ and $\Sigma\text{Int}_{\text{Unmatched}}$ represent the total number of unmatched peaks and the sum of their intensities, respectively. When the matching peptide length is less than 5, PTMap applies additional discounts on sequence coverage score S_{seq} because of the increased chance for random peak matching for very short peptide sequences.

In addition to $S_{Unmatched}$ and PTMap score, PTMap requires several other criteria to be satisfied for any peptide identification. For example, before an unrestricted PTM is considered identified, PTMap requires the modification site to be exclusively localized.

Part IV. Result output

PTMap generates two search results: simple data output and complete data output. The former supplies only the spectra that most closely match each peptide sequence and the latter supplies a list of all matching spectra. The software output is written as a tab-delimited text file. When a result file is opened in the software, the user is able to display each peptide's fragmentation table and annotated tandem mass spectrum. The MS/MS spectrum of the unmodified counterpart can also be displayed, if available, to compare their fragmentation patterns and HPLC elution times. All results can be edited and deleted within the software and saved into a new file.

CHAPTER 5

IDENTIFICATION OF NON-BIOLOGICAL PROTEIN MODIFICATIONS

5.1 Summary

In this chapter, I will describe the application of PTMap software to unrestrictive PTM identification in order to discover unexpected non-biological modifications. Chemical modifications in proteins, introduced during sample preparation, can complicate mass spectra and could potentially lead to false-positive identifications. While several such artificial protein modifications were described in the past, it remains unknown whether additional ones exist. Here we report discovery of four types of non-biological protein modifications identified by HPLC/MS/MS analysis and non-restrictive protein sequence alignment by PTMap, an algorithm recently developed by our laboratory. These modifications include ethylation of aspartate and glutamate (+28 Da), esterification of aspartate and glutamate by glycerol (+74 Da), an unexpected lysine modification with a mass shift of -19 Da, and an unexpected cysteine modification with a mass shift of +108 Da. We confirmed the non-biological modifications using control experiments such as protein in-solution digestion and using an experimental procedure devoid of a solvent of interest. We proposed a possible molecular mechanism for lysine -19 Da modification. Our study therefore conclusively identifies several novel *in vitro* protein modifications, suggests solutions to avoid such modifications, and highlights precautions for assignment of protein modifications.

5.2 Introduction

The mass spectrometry technologies have been so advanced in the past two decades that low stoichometric protein modifications, produced *in vitro* or *in vivo*, can be detected. Mass analysis of proteolytic peptides by those mass spectrometers with high resolution and high mass accuracy makes it possible to correlate the detected peptide masses to the known protein modifications [47,65,98]. In addition, the recent parallel development of protein sequence alignment algorithms for MS/MS data enables to identify all the possible protein modifications, known or undescribed ones [44,55,56]. Upto today, more than three hundred types of protein modifications were described [65], some of which were generated *in vitro*.

Several types of protein modifications are known to be induced *in vitro* under some specific protein environments [111]: (1) Acrylamide adduct. SDS-PAGE is a popular method for resolving proteins. The protein separation procedure induces formation of covalent protein adducts between the protein of interest and chemicals in the gel. The side chain of a cysteine residue can react with the unpolymerized free acrylamide in the matrix to form cysteinyl-S-β-propionamide, with a mass increase of 71 Da [112-114]. In addition to cysteine residues, it was reported that acrylamide adduction in peptides could happen in N-terminal amine [114]. (2) β-mercaptoethanol adduct. The cysteine residue was also reported to be *in vitro* modified by β-mercaptoethanol, leading to a mass increase of 76 Da [115]. (3) Oxidation. Some amino acid residues are easily oxidized *in vitro*, such as methionine, cysteine and tryptophan. (4) Methylation. When methanol is used in staining and destaining buffers following SDS-PAGE, *in vitro* methylation could be induced at the side chains of aspartate and glutamate [107]. (5) Deamidation. Deamidation could happen at the side chains of asparagine and glutamine, leading to the formation of aspartate and glutamate, respectively, and the reaction can be dramatically induced at high pH (pH>10) [116].

In this chapter, I will report systematic analysis of protein modifications in BSA and histone H4 by HPLC/MS/MS analysis in combination of protein sequence alignment by PTMap, a recently developed program for identifying all the possible protein modifications with high accuracy. Our analysis discovered four types of *in vitro* protein modifications, ethylation of aspartate and glutamate, esterification of aspartate and glutamate by glycerol, an unexpected lysine modification with a mass shift of -19 Da, and an unexpected cysteine modification with a mass shift of +108 Da. We confirmed the four modifications as the *in vitro* ones by various control experiments. Our study therefore enrich our understanding of protein *in vitro* modifications, provide a procedure to avoid such *in vitro* modifications, and suggest the cautions to define *in vivo* protein modifications whose mass shifts are overlapped with the ones described here.

5.3 Results

I will report four types of *in vitro* protein modifications that occur during staining and destaining gels, or in-gel digestion, or protein storage. These modifications were identified by HPLC/MS/MS analysis of tryptic peptides using PTMap algorithm, a recently developed software that enables to carry out non-restrictive sequence alignment for identifying all the possible PTMs [44]. Thus, the algorithm allows us to identify undescribed protein modifications in addition to the known ones. A comparison of the protein modifications under different digestion methods (in-gel or in-solution digestion) or using different sample solutions (*e.g.*, various buffers for staining and destaining gel, and protein storage solution) allow to distinguish *in vitro* protein modifications from the *in vivo* ones.

To study possible *in vitro* protein modifications, we resolved BSA and histone H4 in SDS-PAGE. The gel was stained and destained with buffers containing ethanol. The resulting tryptic peptides were analyzed by nano-HPLC/MS/MS followed by non-restrictive sequence alignment by PTMap.

Analysis of BSA tryptic peptides identified 104 peptides with coverage of 77.8% of the protein sequence. Interestingly we identified 19 modified residues with diverse mass shifts (Table 6). Likewise, we identified 217 peptides with a sequence coverage of 78.6% in histone H4. The analysis identified 17 modified residues with variant mass shifts (Table 7). Some of the modifications were known to occur *in vitro*. Here we highlighted four types of novel *in vitro* modification identified in BSA and histone H4 (Table 8 and Table 9).

Table 6 List of identified mass shifts (Δ Da) and corresponding modified amino acids in BSA after SDS-PAGE separation and in-gel digestion.

Modified residue	Mass shift (Δ Da)
D	1, 65, 76, 128,
A	15, 30, 57, 102, 56
E	-71, -58, -18, -2, 16, 28, 48, 64, 78, 80, 94, 101, 110, 174
F	-49, -48, -18, 7, 26, 31, 44, 48, 80
G	40, 43, 58, 71, 174
H	-23, 12, 28, 64, 71, 156
K	-32, -17, -13, 28, 42, 43, 57, 71, 76, 174
L	-14, -2, 2, 108, 162, 195
M	2, 16, 110
N	-17, 1
P	1, 156
Q	-2, 1, 58, 128, 156, 171
R	18, 68
S	-18, 72
T	-37, -18, 128,
V	-37, -33, -18, 2, 12, 128,
W	6, 16, 32,
Y	-18, 2, 16, 71, 76,
C	-34, -32, -4, -2, 11, 16, 25, 30, 32, 37, 42, 44, 48, 54, 59, 64, 71, 72, 76, 77, 78, 79, 91, 96, 100, 103, 106, 108, 124, 130, 136, 147, 152, 153, 154, 184

Table 7 List of identified mass shifts (Δ Da) and corresponding modified amino acids in calf thymus Histone H4 after SDS-PAGE separation and in-gel digestion.

Modified residue	Mass shift (Δ Da)
A	-14, 14, 42, 46,
D	22, 26, 28, 42, 96, 139,
E	-52, -18, -1, 22, 28, 42, 96, 156
F	12, 16, 26, 136
G	28, 72, 170, 185,
H	14, 16, 26, 32, 54, 92, 182
I	-28, -18, -17, -1, 26, 28, 42, 136, 140, 154, 156, 195, 198,
K	-58, -49, -34, -33, -19, -18, -15, -1, 14, 20, 28, 42, 56, 64, 71, 81, 82, 85, 96, 98, 128, 149, 156, 169, 184
L	-44, -18, 1, 2, 12, 14, 22
M	-48, -29, -14, 2, 16, 32, 34, 50, 62, 98, 101, 149
N	1
Q	-17, 1, 46, 56, 128
R	-82, -60, -59, -55, -43, -28, -26, -17, -1, 2, 12, 14, 16, 28, 29, 39, 53, 74, 82, 113, 127, 115, 185
S	-54, 12, 14, 28, 96, 139, 141, 142, 156
T	-18, 12, 20, 28, 42, 55, 65, 109, 128, 130
V	-20, -18, 11, 14, 26, 28, 42, 83, 124, 138, 139, 154, 156, 168, 182, 186
Y	-58, -42, 1, 2, 12, 16, 26, 28, 34, 42, 96

Table 8 The numbers of modified peptides identified with either E-methylation or E-ethylation induced by various destaining/washing solutions during in-gel digestion.

Destaining solution	**Frequency of E-methylation**	**Frequency of E-ethylation**
EtOH/H_2O (50%/50%)	0	1
HOAc/H_2O/EtOH (10%/40%/50%)	0	2
MeOH/H_2O (50%/50%)	0	0
HOAc/H_2O/MeOH (10%/40%/50%)	4[17]	0

Table 9 The comparison of the peptide numbers identified with the modifications identified with any one of the four mass shifts — Lys -19 Da, Cys +108 Da, Cys +76 Da, Cys + 71 Da following in-gel or in-solution tryptic digestion.

Type of modifications		**Frequency of modifications in In-gel digestion**	**Frequency of modifications in In-solution digestion**
Modified amino acid	Mass shift (Δ Da)	8	0
Lys	- 19	6	0
Cys	+ 108	15	3
Cys	+ 76 (mercaptoethanol)	18	0
Cys	+ 71 (acrylamide)	8	0

5.3.1 Ethylation of aspartic acid and glutamic acid residues

Methanol is usually included in staining buffer for visualization of proteins in SDS-PAGE and in destaining buffer prior to in-gel digestion. Methanol could lead to methylation of aspartic acid and glutamic acid residues [107]. To distinguish between *in vitro* methylation and *in vivo* methylation of the acidic residues, ethanol was used to replace the methanol in the staining and destaining buffers [107,117]. However, such a change induces *in vitro* ethylation.

As an example, Figure 39 shows an MS/MS spectrum of an *in vitro* ethylated BSA tryptic peptides. The ethylated peptide was identified in the sample that was destained with buffers containing ethanol, but not in the buffers without ethanol (Table 8), suggesting that ethylation happens *in vitro* instead of *in vivo*. Since acetic acid can catalyze esterification reaction between alcohol and carboxylic acid side chains of acidic residues, as expected, acetic acid increases the number of BSA tryptic peptides that are ethylated (Table 8). We argue that the detected ethylation in BSA tryptic peptides were produced *in vitro* because all the ethylated BSA peptides are not present in the BSA tryptic peptide samples that were destained with buffers devoid of ethanol (Table 8).

Likewise, acetic acid in a destaining buffer containing methanol increases the number of BSA tryptic peptides that were methylated (Table 8).

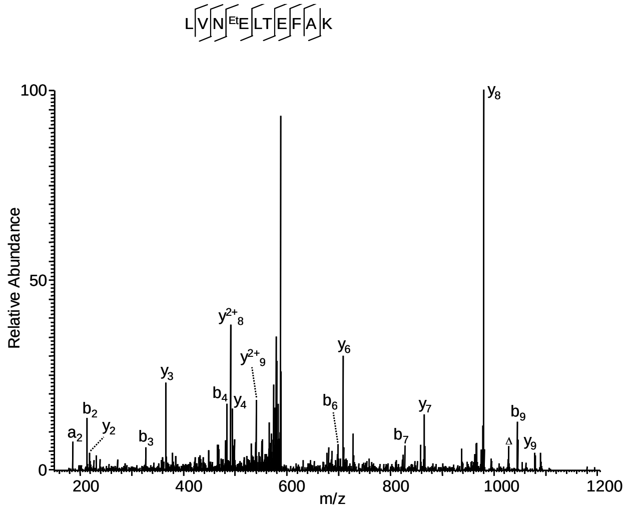
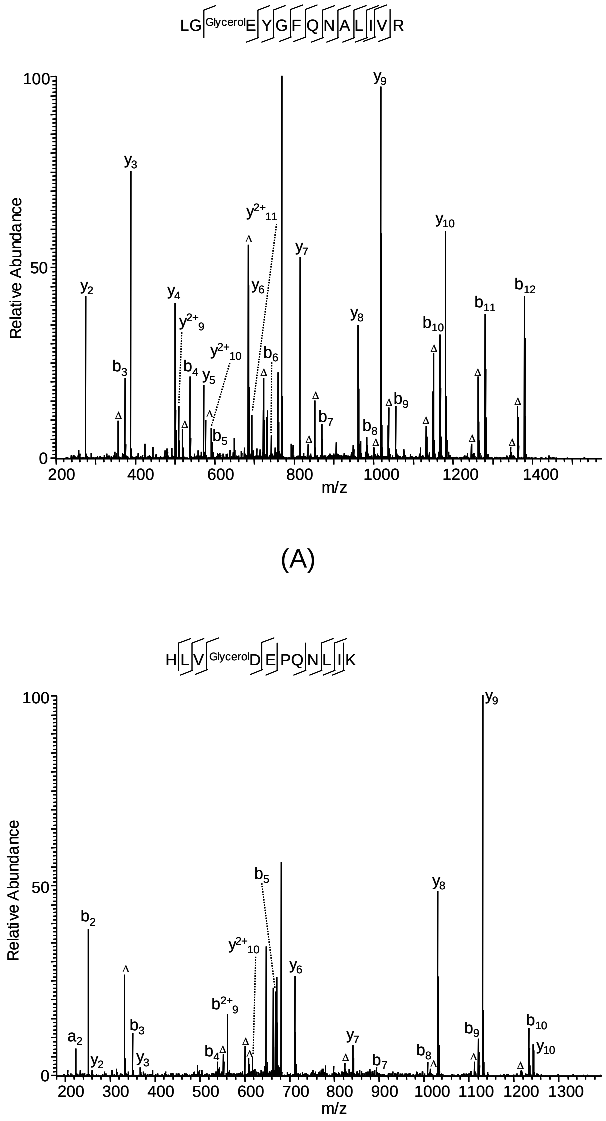


Figure 39 Identification of ethylated glutamic acid in BSA.

5.3.2 Esterification of aspartic acid and glutamic acid residues by glycerol

Glycerol is a common solvent used in buffers for purification and storage of proteins. Like methanol and ethanol, glycerol contains hydroxyl groups that can possibly participate in esterification reactions with acidic side chains of aspartic acid and glutamic acid residues. As expected, glycerol-modified BSA tryptic peptides were identified when the protein was incubated with a buffer containing 20% glycerol, which lead to a mass shift of 74 Da in aspartate or glutamate residues. As examples, glycerol-modified glutamate- and aspartate-containing peptides were shown (Figure 40). The detected glycerol-modification reaction occurs *in vitro* instead of *in vivo* because glycerol-modified BSA peptides were not detected with the BSA samples that have not treated with a buffer containing glycerol (Figure 41).



(A)

(B)

Figure 40 Modification of acidic residues by glycerol. BSA was incubated with 20% glycerol at 4^{o}C for 5 days before in-solution digested by trypsin. The tryptic peptides were analyzed by HPLC/MS/MS in an LTQ mass spectrometer. (A) An example of BSA peptides modified by glycerol at a glutamic acid residue. (B) An example of BSA peptides modified by glycerol at an aspartic acid residue.

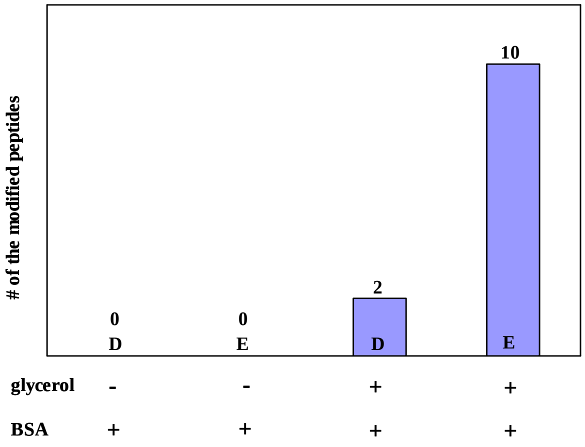


Figure 41 The number of glycerol modified BSA peptides identified in our analysis under the conditions with or without glycerol incubation. BSA was incubated with 20% glycerol at 4 ^{o}C for 5 days before in-solution digested by trypsin. The tryptic peptides were analyzed by HPLC/MS/MS in an LTQ mass spectrometer.

5.3.3 Modification of cysteine residue by 108 Da

Cysteine residues could be potentially subjected to diverse *in vitro* modifications, such as oxidation, addition reaction with iodoacetamide, acrylamide, DTT, and beta-mecaptoethanol [118,119]. Availability of PTMap algorithm at our laboratory enables us to identify all the possible modifications at cysteine residues produced *in vivo* or *in vitro*. Nevertheless, it remains unknown if additional undescribed modifications are present at cysteine residues.

Our mass spectrometric analysis and subsequent nonrestrictive sequence alignment identified 6 cysteine residues of H4 peptides that were modified by a chemical that leads to a mass shift of 108 Da (Table 9, Figure 42). To our knowledge, such modification has not been described at cysteine residue before. Since the modification was not detected in the H4 tryptic peptides that were produced by in-solution digestion, the modification was considered as an *in vitro* modification that happens during SDS-PAGE and gel-staining/destaining steps. The determined mass shift has low mass accuracy that could not lead to an elemental composition. The structure of the modification moiety remains to be determined.

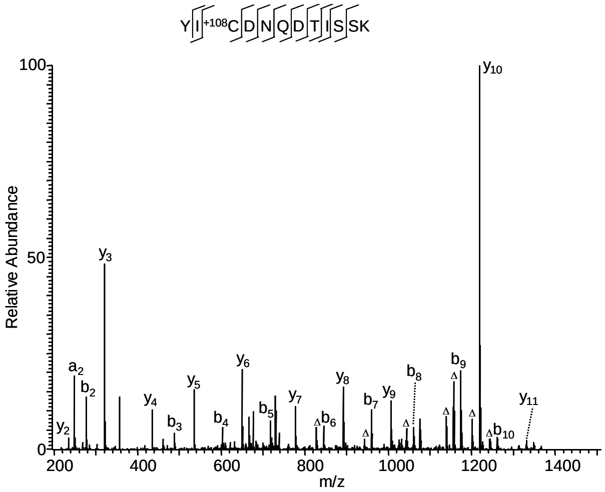


Figure 42 An example of a novel modified peptide with a mass shift of +108 Da at Cys residue.

5.3.4 Loss 19 Da in lysine residues

The side chain of lysine residue can be potentially modified *in vivo* by several post-translational modifications including methylation, acetylation, biotinylation, ubiquitination, and sumoylation, which have pivotal roles in cell physiology and pathology. Our study on the tryptic peptides from histone H4 identified 8 lysine residues that were modified by -19 Da at lysine residue (Figure 43). Since the modification was not identified in the H4 peptides that were produced by in-solution digestion, the modification should happen non-biologically. To the best of our knowledge, such *in vitro* modification has not reported before.

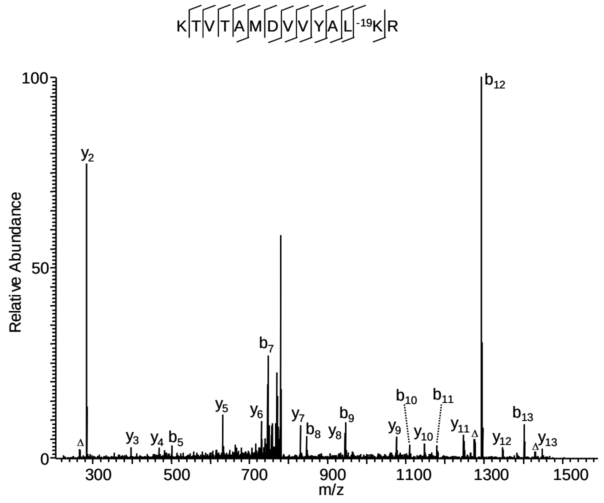


Figure 43 An example of a novel modified peptide with a mass shift of -19 Da at Lys residue.

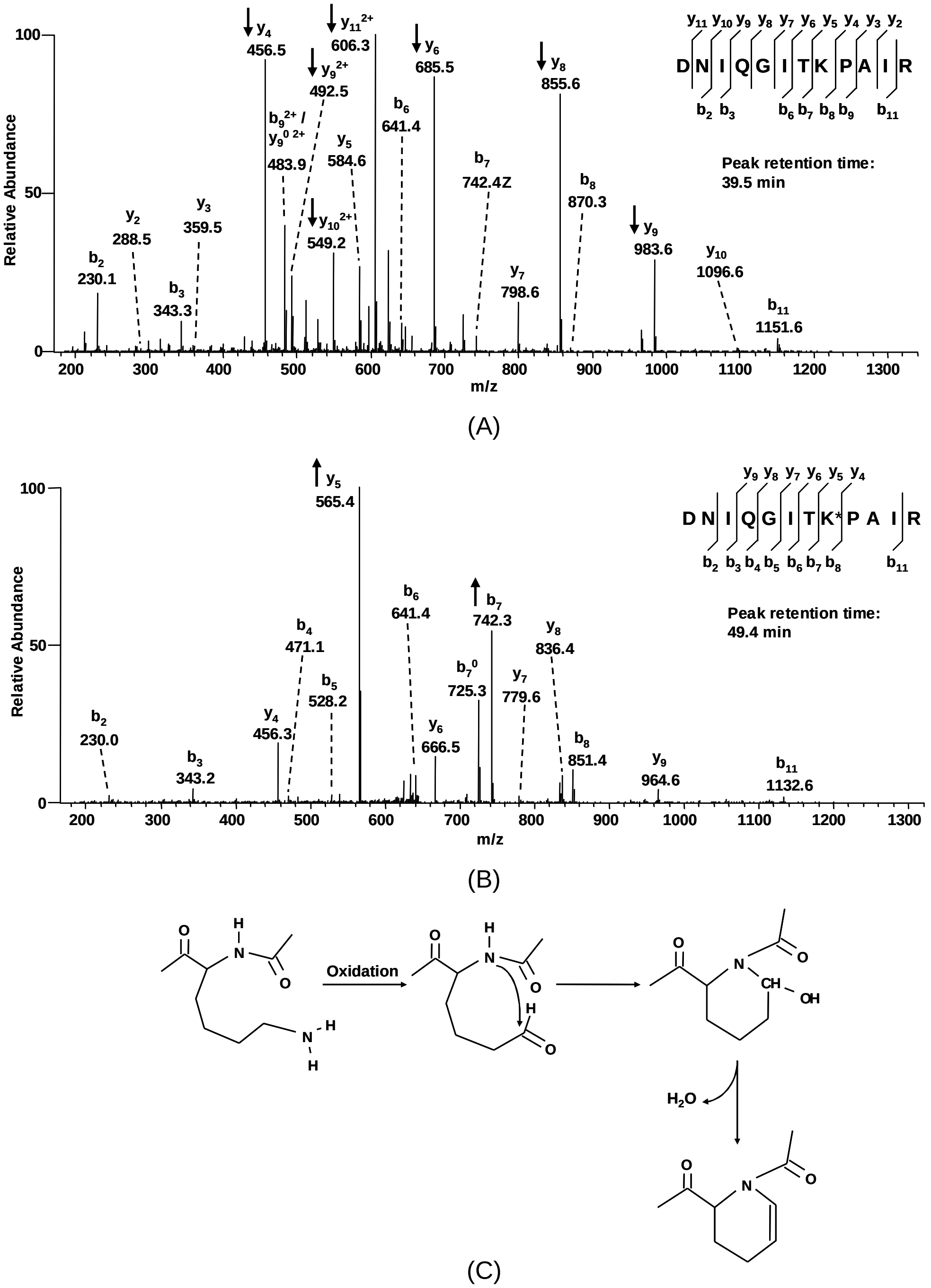


Figure 44 Mechanistic study of the artificial modification with a mass shift of -19 at a lysine residue. (A) MS/MS of the unmodified H4 peptide, DNIQGITKPAIR. (B) MS/MS spectrum of the modified H4 peptides with -19 Da loss at lysine residue, DNIQGITK*PAIR. And (C) a possible molecular mechanism of an *in vitro* protein modification with a mass shift of -19 at the lysine residue.

To explore the possible mechanism of this modification, we compared the peptide fragmentation patterns and retention times for the peptide of interest with and without this modification. The MS/MS spectra of the H4 peptide "DNIQITK*PAIR" with or without K7 modification (Figure 44A, B) showed an apparent change of daughter ions' intensity profiles among a series of y ions containing the modified lysine residue. After K-19 Da modifications, the ion intensities of y_6, y_8, y_9^{2+}, y_{10}^{2+}, and y_{11}^{2+} from the unmodified H4 peptide (Figure 44A) are dramatically decreased comparing to those from the K-19 Da modified peptide (Figure 44B). In addition, the relative intensity of y_5 and b_7 ions are the strongest in the spectrum of the modified peptide. Next, we noticed that the strongest ion, y_4 in the unmodified peptide's spectrum, which was produced by breakage at K-P peptide bond, became much weaker in modified peptide's spectrum, suggesting some significant structural change of K-P residues. On the other hand, y_5 became the strongest the ion in the modified peptide (Figure 44B). This change suggests that the most fragile peptide bond in the sequence has been possibly shifted from K-P bond to T-K bond, implying an increase of the steric hindrance in T-P bond. The HPLC retention time of the modified peptides was around 49.4 min that is significant delayed than that of the unmodified peptides (~39.5 min), suggesting an increase of hydrophobicity. These fragmentation pattern change and retention time delay were also observed for other histone H4 peptides with the same modification.

Taken together, our results suggest several possible peptide property changes upon the modification: first, the unknown modification on the side chain of Lys causes the loss of basicity on the ε-amino group. Second, the shift in most fragile peptide bond indicates that this modification probably introduced strong stereo hindrance on the T-K peptide backbone, which may lead to the same effect as proline. Third, the apparent increase in retention time shows that the modified Lys side chain is more hydrophobic. Based on these observations, the known mass shift (-19 Da), and the cellular Lys metabolism pathway, we proposed a possible *in vitro* reaction mechanism for this new modification (Figure 44C). The ε-amino group of lysine can be first oxidized by free-radical reaction [120] which may be initiated by the radicals in the SDS-PAGE gel. The reactive ε-aldehyde group was nucleophilic attacked by the α-amino group on the peptide backbone in a mechanism similar to glutamine / asparagine deamidation through succinimide intermediate [121]. The resulting 2-hydroxyl piperidine homolog loses one molecular of water to form a more stable structure with a double bond. The -19 Da mass shift is the net effect of the loss of one nitrogen and five hydrogen atoms. While this model remains to be further confirmed, our model can explain four experimental observations after the modification: (1) 19 Da is decreased. (2) The peptide becomes more hydrophobic. (3) The modified lysine residue introduces an additional steric constraint in a similar fashion as proline, therefore facilitating fragmentation. And (4) due to facile fragmentation at the N-terminal of the modified lysine, the fragmentation channel at the N-terminal of proline is significant compromised.

5.4 Discussion

The identified i*n vitro* modifications could be confused with other known protein modifications, and therefore could potentially lead to misassignment during protein identification and mapping modification sites by automatic sequence alignment. For example, ethylation of aspartate and glutamate leads to a mass shift of +28 Da, which is the same as the sum of two methylation sites or formylation. Therefore, when methylation is included as a variable modification, ethylation of aspartate and glutamate can be potentially misassigned as two methylation sites by protein sequence database algorithms. Such misassignment could happen more easily when the candidate methylation residues are adjacent to the ethylated aspartate and glutamate residues. Likewise, the modification with Lys -19 Da mass shift has similar mass shift as those with protein modifications associated with water loss ion, and a change from serine to dehydralanine, from cysteine to formylglycine, glutamic acid from pyrogluamic acid formed, and succinimide from aspartic acid, which has a mass shift of -18 Da. Accordingly, the six types of protein identifications could be mistakenly cross-assigned due to their similar mass shifts. Two peptide sequences that are modified by two types of protein modifications with same mass shifts will have similar fragmentation patterns and could probably lead to similar statistical scores, when identified by statistics-based protein alignment algorithms. Therefore, exclusive mapping of the protein modification site, complete assignment of all the major daughter ions in the MS/MS spectrum, and exact matching of mass shifts are critical to ensure the accuracy of peptide identification and mapping modification sites. This argument also suggest the importance to emphasize the unmatched to remove the false peptide identifications as we described previously [41].

5.5 Concluding Remarks

In this chapter, I reported the identification of four types of novel *in vitro* protein modifications with the application of PTMap software. Protein modifications with mass shifts of cysteine +108 Da and lysine -19 Da were not reported before and we proposed a possible reaction mechanism for lysine -19 modification based on mass spectrometry and chromatography analysis. Protein modification by glycerol and ethylation increases the masses of the substrate residues for 74 Da and 28 Da, respectively, which was reported previously in glutamate [122], but not at aspartate. Together, our results provide further insights into chemical-induced artificial protein modifications and highlight the importance of careful sample preparation and storage for biological analysis.

5.6 Material and Methods

5.6.1 Proteins

BSA was purchased from Sigma (St. Louis, MO) and calf thymus core histones were purified according to a procedure described previously[109]. Briefly, cold, fat free calf thymus (200 g) was sliced into 1 – 2 cm^3 small cubes, soaked in 160 mL of 0.5 M sucrose solution for 3 min, and then mixed with 1.44 L homogenizing buffer (0.25 M sucrose and 3.3 mM $CaCl_2$ solution). Every 250 mL of above solution was homogenized for 30 second twice in an Oster 12-speed blender at the speed of “easy clean”. The homogenate was filtered through two layers of cheesecloth. The filtrate was centrifuged at 1,000 x g for 10 min to obtain wet cell pellet. The cell pellet was re-suspended in 4 volumes of hypotonic buffer (50 mM Tris-Cl (pH7.9), 2.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, and 0.5 mM PMSF) with slow stirring for 30 min. The suspension was centrifuged at 1,600 x g for 10 min to collect nuclei pellet. The core histones were extracted twice from the nuclei pellet using ~ 4 volume of 0.4 N H_2SO_4 solution overnight. The extract was dialysed sequentially against H_2O and 50 mM Tris buffer (pH 7.3) for eight hours, respectively. The core histones were resolved in 15% SDS-PAGE and visualized by colloidal Commasie blue staining method. To isolate H4 protein, the core histone preparation was subjected to HPLC separation using C4 column. The H4 peak was collected, dried in SpeedVac, and resolubalized in water.

5.6.2 Protein tryptic digestion

Three sets of experiments were designed to compare the effects of various reagents used in protein storage and in-gel digestion solutions on *in vitro* protein modifications: 1) BSA was incubated in a buffer (50 mM ammonium bicarbonate, pH 8) with or without 20% glycerol at 4 ^{o}C for 5 days, followed by in-solution digestion using porcine trypsin (Promega, Madison, WI) at 1:50 enzyme-to-substrate ratio. 2) BSA or calf thymus core histone H4 was digested either in solution or in gel. 3) BSA was separated by SDS-PAGE and in-gel digested following a standard protocol with varied washing solutions. SDS-PAGE gels were stained by colloidal Coomassie staining solution composed of 9 volumes of g-250 stain solution (protoBlue, national diagnostics, Atlanta, Georgia) and 1 volume of ethanol overnight. Prior to in-gel digestion, the gels were subject to destaining with water.

5.6.3 Protein digestion

For protein in-gel digestion, the protein bands of interest were subject to 8 hour wash, with buffer exchange for 3 times, with a destaining solution in 1.5 ml microcentrifuge tubes on a Tomy MT-360 microtube mixer (Tomy Digital Biology, Japan) at medium speed. Four different types of destaining buffers were used depending on the experiment desgin: Destaining buffer I: ethanol/water (50%:50%); Destaining buffer II: acetic acid/ethanol/water (10%:50%:40%);

Destaining buffer III: methanol/water (50%:50%); and Destaining buffer IV: acetic acid/methanol/water (10%:50%:40%) [96]. After destaining, the protein bands were first rehydrated by a twenty-minute wash with 1 mL water, then cut into 1 mm^3 cubes, subsequently dehydrated by acetonitrile and dried in SpeedVac for 20 minutes. The dried gel pieces were swelled and covered by 10 ng/µL trypsin in 50 mM ammonium bicarbonate solution and subject to overnight digestion at 37 ^{o}C. The resulting tryptic peptides were extracted, dried, and cleaned in C18 Zip-tip (Millipore, Bedford, MA) as previously described [96]. Protein in-solution digestion was carried out by adding trypsin stock solution into a protein solution (in 50 mM ammonium bicarbonate (pH 8.0)) at a 1:50 enzyme-to-substrate ratio and allow the overnight digestion at 37 ^{o}C.

5.6.4 Nano-HPLC/MS/MS analysis

Tryptic peptides cleaned by C18 Zip-tip (Millipore, Bedford, MA) were reconstituted in buffer A solution (0.1% acetic acid/2% acetonitrile/97.9% H_2O (v/v/v)). Mass analysis were performed on a LTQ – 2D ion trap spectrometer (Thermo Scientific, San Jose, CA) equipped with a nano-electrospray ionization source, which is coupled with a Agilent 1100 nano flow HPLC system. Two micro liters of a peptide sample in buffer A was manually loaded onto a capillary column (10cm length x 75 μm ID) home packed with Jupiter C12 resin (4 μm particle size, 90 Å pore diameter) (Phenomenex, Torrance, CA). Peptides were eluted from the column using a gradient from 8% to 90% buffer B (0.1% acetic acid/90% acetonitrile/9.9% H_2O (v/v/v)) in a 100 min cycle. The eluted peptides were directly electro-sprayed into LTQ spectrometer with MS/MS spectra acquired in a data dependent mode that cycled between MS and MS/MS of the 10 strongest parent ions.

5.6.5 Protein sequence alignment and manual validation of peptide hits

Each LC/MS dataset was searched against the corresponding protein sequence with an in-house developed software, PTMap, for identifying all possible protein modifications, including previously undescribed PTMs. PTMap is able to confidently identify protein modifications with mass shifts ranging from -100 Da to +200 Da with 1 Da increment. In the search parameters, trypsin was specified as the proteolytic enzyme with 3 allowed missing cleavages. Parent ion mass error and fragment ion error mass were set as ±4 Da and ±0.6 Da, respectively. All peptide identifications were manually validated with high stringency according to the previously published criteria and each modification site was able to be exclusively localized in the peptide sequence by PTMap [41].

APPENDIX A

PRIOR PUBLICATIONS

Geiman, T. M., Sankpal, U. T., Robertson, A. K., **Chen, Y.**, Mazumdar, M., Heale, J. T., Schmiesing, J. A., Kim, W., Yokomori, K., Zhao, Y., and Robertson, K. D. Isolation and characterization of a novel DNA methyltransferase complex linking DNMT3B with components of the mitotic chromosome condensation machinery. Nucleic Acids Res, *32:* 2716-2729, 2004.

Chen, Y., Kim, S. C., and Zhao, Y. High-throughput identification of in-gel digested proteins by rapid, isocratic HPLC/MS/MS. Anal Chem, *77:* 8179-8184, 2005.

Chen, Y., Kwon, S. W., Kim, S. C., and Zhao, Y. An integrated approach for manual verification of peptides identified by searching protein sequence databases with tandem mass spectra. J. Proteome Research, *4:* 998-1005, 2005.

Sprung, R., Nandi, A., **Chen, Y.**, Kim, S. C., Barma, D., Falck, J. R., and Zhao, Y. Tagging-via-substrate strategy for probing O-GlcNAc modified proteins. J Proteome Res, *4:* 950-957, 2005.

Zhi, G., Ryder, J. W., Huang, J., Ding, P., **Chen, Y.**, Zhao, Y., Kamm, K. E., and Stull, J. T. Myosin light chain kinase and myosin phosphorylation effect frequency-dependent potentiation of skeletal muscle contraction. Proc Natl Acad Sci U S A, *102:* 17519-17524, 2005.

Binns, D., Januszewski, T., **Chen, Y.**, Hill, J., Markin, V. S., Zhao, Y., Gilpin, C., Chapman, K. D., Anderson, R. G., and Goodman, J. M. An intimate collaboration between peroxisomes and lipid bodies. J Cell Biol, *173:* 719-731, 2006.

Kim, S. C., **Chen, Y.**, Mirza, S., Xu, Y., Lee, J., Liu, P., and Zhao, Y. A clean, more efficient method for in-solution digestion of protein mixtures without detergent or urea. J Proteome Res, *5:* 3446-3452, 2006.

Kim, S. C., Sprung, R., **Chen, Y.**, Xu, Y., Ball, H., Pei, J., Cheng, T., Kho, Y., Xiao, H., Xiao, L., Grishin, N. V., White, M., Yang, X. J., and Zhao, Y. Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. Mol Cell, *23:* 607-618, 2006.

Qiu, Y., Zhao, Y., Becker, M., John, S., Parekh, B. S., Huang, S., Martinez, E. D., **Chen, Y.**, Lu, H., Adkins, N. L., Georgel, P. T., Schiltz, P. L., and Hager, G. L. HDAC1 Acetylation is linked to progressive modulation of steroid receptor induced gene transcription. Mol Cell, *22:* 669-679, 2006.

Lee, J., Xu, Y., **Chen, Y.**, Sprung, R., Kim, S. C., Xie, S., and Zhao, Y. Mitochondrial phosphoproteome revealed by an improved IMAC method and MS/MS/MS. Mol Cell Proteomics, *6:* 669-676, 2007.

Chen, Y., Sprung, R., Tang, Y., Ball, H., Sangras, B., Kim, S.C., Falck, J.R., Peng, J., Gu, W. and Zhao, Y. Lysine propionylation and butyrylation are novel post-translational modifications in histones. Mol Cell Proteomics, *6:* 812-819, 2007.

Bartz, R., Zehmer, J.K., Zhu, M., **Chen, Y.**, Serrero, G., Zhao, Y. and Liu, P. Dynamic activity of lipid droplets: protein phosphorylation and GTP-mediated protein translocation. J Proteome Res, *6:* 3256-3265, 2007.

Ganesan, A.K., Kho, Y., Kim, S.C., **Chen, Y.**, Zhao, Y. and White, M.A. Broad spectrum identification of SUMO substrates in melanoma cells. Proteomics, *7:* 2216-2221, 2007.

Chang, B., **Chen, Y.**, Zhao, Y. and Bruick, R.K. JMJD6 is a histone arginine demethylase. Science, *318:* 444-447, 2007.

Kang, J., **Chen, Y.**, Zhao, Y. and Yu H. Autophosphorylation-dependent activation of human Mps1 is required for the spindle checkpoint. Proc Natl Acad Sci U S A, *104:* 20232-20237, 2007.

Sprung, R., **Chen, Y.**, Zhang, K., Cheng, D., Zhang, T., Peng, J. and Zhao, Y. Identification and validation of eukaryotic aspartate and glutamate methylation in proteins. J Proteome Res, *7*, 1001-1006, 2008.

Jung, S.Y., Li, Y., Wang, Y., **Chen, Y.**, Zhao, Y. and Qin, J. Complications in the assignment of 14 and 28 Da mass shift detected by mass spectrometry as *in vivo* methylation from endogenous proteins. Anal Chem, *80*, 1721-1729, 2008.

Tang, Y., Zhao, W., **Chen, Y.**, Zhao, Y. and Gu, W. Acetylation is indispensable for p53 activation. Cell, *133*, 612-626, 2008.

Chen, Y., Chen, W., Cobb, M. and Zhao, Y. PTMap—a novel sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-translational modification sites. submitted.

Xing, G., Zhang, J., **Chen, Y.**, and Zhao, Y. Identification of four types of novel *in vitro* protein modifications in proteins. J Proteome Res. in press.

Chen, Y., Zhang, J., Xing, G. and Zhao, Y. Common types of false positives identified in shotgun proteomics. submitted.

Cheng, Z., Tang, Y., **Chen, Y.**, Kim, S.C., Liu, H., Li, S.S.C., Gu, W. and Zhao, Y. Molecular characterization of propionyllysines in non-histone proteins. Mol Cell Proteomics, in press.

REFERENCES

1. A. Bachmair, D. Finley, and A. Varshavsky, *Science* **234** (4773), 179 (1986).
2. B. Ackermann and J. Steinmeyer, *Osteoarthritis Cartilage* **13** (10), 906 (2005).
3. J. S. Kavanaugh, J. A. Weydert, P. H. Rogers et al., *Protein Sci* **10** (9), 1847 (2001).
4. M. H. Kubbutat, S. N. Jones, and K. H. Vousden, *Nature* **387** (6630), 299 (1997).
5. S. Y. Fuchs, V. Adler, T. Buschmann et al., *Genes Dev* **12** (17), 2658 (1998).
6. W. Gu, X. L. Shi, and R. G. Roeder, *Nature* **387** (6635), 819 (1997).
7. G. He, Z. H. Siddik, Z. Huang et al., *Oncogene* **24** (18), 2929 (2005).
8. C. T. Walsh, S. Garneau-Tsodikova, and G. J. Gatto, Jr., *Angew Chem Int Ed Engl* **44** (45), 7342 (2005).
9. J.A. Hoch and T.J. Silhavy, *ASM, Washington* (1995).
10. Christopher T. Walsh, *Posttranslational Modifications of Proteins: Expanding Nature's Inventory,*. (Roberts and Company Publishers, Greenwood Village, Co, 2006).
11. G. L. Johnson and R. Lapadat, *Science* **298** (5600), 1911 (2002).
12. R. L. Thurmond, S. A. Wadsworth, P. H. Schafer et al., *Eur J Biochem* **268** (22), 5747 (2001).
13. D. M. Ouwens, N. D. de Ruiter, G. C. van der Zon et al., *EMBO J* **21** (14), 3782 (2002).
14. S. C. Kim, R. Sprung, Y. Chen et al., *Mol Cell* **23** (4), 607 (2006).
15. R. E. Schweppe, C. E. Haydon, T. S. Lewis et al., *Acc Chem Res* **36** (6), 453 (2003).
16. F.A. Lipmann and P.A. Levene, *Journal of Biological Chemistry* **98** (1), 6 (1932).
17. R. Schoenheimer and D. Rittenberg, *Science* **82** (2120), 156 (1935).
18. G. Vidali, E. L. Gershey, and V. G. Allfrey, *J Biol Chem* **243** (24), 6361 (1968).
19. J. B. Fenn, M. Mann, C. K. Meng et al., *Science* **246** (4926), 64 (1989).
20. M. Karas, D. Bachmann, and F. Hillenkamp, *Analytical Chemistry* **57** (14), 2935 (1985).
21. K. Tanaka, H. Waki, Y. Ido et al., *Rapid Commun Mass Spectrom* **2** (20), 3 (1988).
22. M. Mann and M. Wilm, *Trends Biochem Sci* **20** (6), 219 (1995).
23. W. Stephens, *Bull. Am. Phys. Soc.* **21** (2), 1 (1946).
24. W. Paul and H. Steinwedel, *Z. Naturforschg.* **8a**, 3 (1953).
25. M. B. Comisarow and A. G. Marshall, *Chem. Phys. Lett.* **25** (2), 2 (1974).
26. A. Makarov, *Anal Chem* **72** (6), 1156 (2000).
27. A. Gruhler, J. V. Olsen, S. Mohammed et al., *Mol Cell Proteomics* **4** (3), 310 (2005).
28. S. A. Beausoleil, M. Jedrychowski, D. Schwartz et al., *Proc Natl Acad Sci U S A* **101** (33), 12130 (2004).
29. K. Biemann, *Annu Rev Biochem* **61**, 977 (1992).
30. V. H. Wysocki, G. Tsaprailis, L. L. Smith et al., *J Mass Spectrom* **35** (12), 1399 (2000).
31. R. S. Johnson and K. Biemann, *Biomed Environ Mass Spectrom* **18** (11), 945 (1989).
32. J. K. Eng, A. L. McCormack, and J. R. Yates, 3rd, *Journal of American Society for Mass Spectrometry* **5** (11), 14 (1994).
33. D. N. Perkins, D. J. Pappin, D. M. Creasy et al., *Electrophoresis* **20** (18), 3551 (1999).
34. J. R. Yates, 3rd, J. K. Eng, A. L. McCormack et al., *Anal Chem* **67** (8), 1426 (1995).
35. R. Craig and R. C. Beavis, *Bioinformatics* **20** (9), 1466 (2004).
36. L. Y. Geer, S. P. Markey, J. A. Kowalak et al., *J Proteome Res* **3** (5), 958 (2004).
37. J. Colinge, A. Masselot, M. Giron et al., *Proteomics* **3** (8), 1454 (2003).
38. K. R. Clauser, P. Baker, and A. L. Burlingame, *Anal Chem* **71** (14), 2871 (1999).
39. A. I. Nesvizhskii, O. Vitek, and R. Aebersold, *Nat Methods* **4** (10), 787 (2007).
40. O. N. Jensen, *Nat Rev Mol Cell Biol* **7** (6), 391 (2006).

41. Y. Chen, S. W. Kwon, S. C. Kim et al., *J Proteome Res* **4** (3), 998 (2005).
42. Y. Chen, R. Sprung, Y. Tang et al., *Mol Cell Proteomics* **6** (5), 812 (2007).
43. Z. Cheng, Y. Tang, Y. Chen et al., *Mol Cell Proteomics* **in press** (2008).
44. Y. Chen, W. Chen, M. Cobb et al., *Submitted* (2008).
45. T. Jenuwein and C. D. Allis, *Science* **293** (5532), 1074 (2001).
46. Y. Chen, J. Zhang, G. Xing et al., **Submitted.** (2008).
47. R. Aebersold and M. Mann, *Nature* **422** (6928), 198 (2003).
48. M. P. Washburn, D. Wolters, and J. R. Yates, 3rd, *Nat Biotechnol* **19** (3), 242 (2001).
49. H. I. Field, D. Fenyo, and R. C. Beavis, *Proteomics* **2** (1), 36 (2002).
50. N. Zhang, R. Aebersold, and B. Schwikowski, *Proteomics* **2** (10), 1406 (2002).
51. P. Hernandez, R. Gras, J. Frey et al., *Proteomics* **3** (6), 870 (2003).
52. R. Craig and R. C. Beavis, *Rapid Commun Mass Spectrom* **17** (20), 2310 (2003).
53. W. Bae and X. Chen, *Mol Cell Proteomics* **3** (6), 596 (2004).
54. A. J. Tackett, J. A. DeGrasse, M. D. Sekedat et al., *J Proteome Res* **4** (5), 1752 (2005).
55. P. A. Pevzner, Z. Mulyukov, V. Dancik et al., *Genome Res* **11** (2), 290 (2001).
56. D. C. Liebler, B. T. Hansen, S. W. Davey et al., *Anal Chem* **74** (1), 203 (2002).
57. S. P. Gygi, B. Rist, S. A. Gerber et al., *Nat Biotechnol* **17** (10), 994 (1999).
58. S. E. Ong, B. Blagoev, I. Kratchmarova et al., *Mol Cell Proteomics* **1** (5), 376 (2002).
59. M. J. MacCoss, C. C. Wu, and J. R. Yates, 3rd, *Anal Chem* **74** (21), 5593 (2002).
60. J. Eriksson, B. T. Chait, and D. Fenyo, *Anal Chem* **72** (5), 999 (2000).
61. J. Eriksson and D. Fenyo, *Proteomics* **2** (3), 262 (2002).
62. J. Eriksson and D. Fenyo, *J Proteome Res* **3** (1), 32 (2004).
63. K. A. Resing, K. Meyer-Arendt, A. M. Mendoza et al., *Anal Chem* **76** (13), 3556 (2004).
64. R. E. Moore, M. K. Young, and T. D. Lee, *J Am Soc Mass Spectrom* **13** (4), 378 (2002).
65. E. S. Witze, W. M. Old, K. A. Resing et al., *Nat Methods* **4** (10), 798 (2007).
66. Michael C. Giddings, Atul A. Shah, Ray Gesteland et al., *Proc. Natl. Acad. Sci.* **100** (1), 6 (2003).
67. A. Castegna, V. Thongboonkerd, J. B. Klein et al., *J Neurochem* **85** (6), 1394 (2003).
68. K. Laukens, P. Deckers, E. Esmans et al., *Proteomics* **4** (3), 720 (2004).
69. T. T. Calikowski, T. Meulia, and I. Meier, *J Cell Biochem* **90** (2), 361 (2003).
70. S. Kubis, R. Patela, J. Combea et al., *Plant Cell* **16**, 19 (2004).
71. J.E. Elias and S.P. Gygi, *Nature Methods* **4** (3), 207 (2007).
72. B. T. Hansen, S. W. Davey, A. J. Ham et al., *J Proteome Res* **4** (2), 358 (2005).
73. M. M. Savitski, M. L. Nielsen, and R. A. Zubarev, *Mol Cell Proteomics* **5** (5), 935 (2006).
74. V. G. Allfrey, R. Faulkner, and A. E. Mirsky, *Proc Natl Acad Sci U S A* **51**, 786 (1964).
75. G. Blander and L. Guarente, *Annu Rev Biochem* **73**, 417 (2004).
76. A. A. Sauve, C. Wolberger, V. L. Schramm et al., *Annu Rev Biochem* (2006).
77. S. Y. Roth, J. M. Denu, and C. D. Allis, *Annu Rev Biochem* **70**, 81 (2001).
78. T. Kouzarides, *Embo J* **19** (6), 1176 (2000).
79. M. C. Haigis and L. P. Guarente, *Genes Dev* **20** (21), 2913 (2006).
80. T. A. McKinsey and E. N. Olson, *Trends Genet* **20** (4), 206 (2004).
81. X. J. Yang, *Nucleic Acids Res* **32** (3), 959 (2004).
82. S. B. Hake, A. Xiao, and C. D. Allis, *Br J Cancer* **90** (4), 761 (2004).
83. M. T. King and P. D. Reiss, *Anal Biochem* **146** (1), 173 (1985).
84. R. N. Dutnall, S. T. Tafrov, R. Sternglanz et al., *Cell* **94** (4), 427 (1998).
85. David L. Nelson and Michael M. Cox, *Chapter 17 in Lehninger Principles of Biochemistry.* (W.H. Freemand and Company, New York, NY, 2005).
86. C. L. Peterson and M. A. Laniel, *Curr Biol* **14** (14), R546 (2004).
87. W. J. Shia, S. G. Pattenden, and J. L. Workman, *Genome Biol* **7** (5), 217 (2006).
88. T. Agalioti, G. Chen, and D. Thanos, *Cell* **111** (3), 381 (2002).
89. M. L. Avantaggiati, V. Ogryzko, K. Gardner et al., *Cell* **89** (7), 1175 (1997).
90. Y. Tang, J. Luo, W. Zhang et al., *Mol Cell* **24** (6), 827 (2006).

91. K. Sakaguchi, J. E. Herrera, S. Saito et al., *Genes Dev* **12** (18), 2831 (1998).
92. W. Fischle, Y. Wang, and C. D. Allis, *Curr Opin Cell Biol* **15** (2), 172 (2003).
93. B. A. Garcia, S. B. Hake, R. L. Diaz et al., *J Biol Chem* (2006).
94. M. T. Boyne, 2nd, J. J. Pesavento, C. A. Mizzen et al., *J Proteome Res* **5** (2), 248 (2006).
95. K. F. Medzihradszky, X. Zhang, R. J. Chalkley et al., *Mol Cell Proteomics* **3** (9), 872 (2004).
96. Y. Zhao, W. Zhang, Y. Kho et al., *Anal Chem* **76** (7), 1817 (2004).
97. O. N. Jensen, *Curr Opin Chem Biol* **8** (1), 33 (2004).
98. R. G. Sadygov, D. Cociorva, and J. R. Yates, 3rd, *Nat Methods* **1** (3), 195 (2004).
99. J. R. Yates, 3rd, J. K. Eng, and A. L. McCormack, *Anal Chem* **67** (18), 3202 (1995).
100. P. A. Pevzner, V. Dancik, and C. L. Tang, *J Comput Biol* **7** (6), 777 (2000).
101. B. T. Hansen, J. A. Jones, D. E. Mason et al., *Anal Chem* **73** (8), 1676 (2001).
102. D. Tsur, S. Tanner, E. Zandi et al., *Nat Biotechnol* **23** (12), 1562 (2005).
103. N. Bandeira, D. Tsur, A. Frank et al., *Proc Natl Acad Sci U S A* **104** (15), 6140 (2007).
104. M. Havilio and A. Wool, *Anal Chem* **79** (4), 1362 (2007).
105. S. A. Beausoleil, J. Villen, S. A. Gerber et al., *Nat Biotechnol* **24** (10), 1285 (2006).
106. S. Tanner, S. H. Payne, S. Dasari et al., *J Proteome Res* (2007).
107. R. Sprung, Y. Chen, K. Zhang et al., *J Proteome Res* **7** (3), 1001 (2008).
108. S. Tanner, P. A. Pevzner, and V. Bafna, *Nat Protoc* **1** (1), 67 (2006).
109. D. Shechter, H. L. Dormann, C. D. Allis et al., *Nat Protoc* **2** (6), 1445 (2007).
110. B. E. Xu, S. Stippec, P. Y. Chu et al., *Proc Natl Acad Sci U S A* **102** (29), 10315 (2005).
111. S. D. Patterson and R. Aebersold, *Electrophoresis* **16** (10), 1791 (1995).
112. K. R. Clauser, S. C. Hall, D. M. Smith et al., *Proc Natl Acad Sci U S A* **92** (11), 5072 (1995).
113. M. le Maire, S. Deschamps, J. V. Moller et al., *Anal Biochem* **214** (1), 50 (1993).
114. S. Haebel, C. Jensen, S. O. Andersen et al., *Protein Sci* **4** (3), 394 (1995).
115. K. Klarskov, D. Roecklin, B. Bouchon et al., *Anal Biochem* **216** (1), 127 (1994).
116. H. Sarioglu, F. Lottspeich, T. Walk et al., *Electrophoresis* **21** (11), 2209 (2000).
117. S. Y. Jung, Y. Li, Y. Wang et al., *Anal Chem* **80** (5), 1721 (2008).
118. D. M. Creasy and J. S. Cottrell, *Proteomics* **2** (10), 1426 (2002).
119. H. Steen and M. Mann, *J Am Soc Mass Spectrom* **12** (2), 228 (2001).
120. E. R. Stadtman and R. L. Levine, *Amino Acids* **25** (3-4), 207 (2003).
121. R. C. Stephenson and S. Clarke, *J Biol Chem* **264** (11), 6164 (1989).
122. S. S. Harwig, N. P. Chen, A. S. Park et al., *Anal Biochem* **208** (2), 382 (1993).

BIOGRAPHICAL INFORMATION

Yue Chen was born to Dehua Chen and Hui Huang on August 11, 1977 in Beijing, China. After graduation from the Shengli Oil Field No.2 High School in the city of Dongying in Shan Dong province, he began studying at Peking University in Beijing in 1995. He received a Bachelor of Science degree with a major in Chemistry in July, 1999. Undergraduate research in the laboratory of Prof. Youchang Xie and Biying Zhao inspired Yue to pursue an advanced degree in research in the United States. Between the fall of 1999 and the summer of 2001, he received his graduate training in Texas A&M University in College Station, TX, where he learned the fundamental knowledge of biological mass spectrometry under the guidance of Dr. David Russell. In the summer of 2001, he transferred to the University of Texas at Dallas to pursue a graduate degree in computer science, where he was trained extensively in computer programming and algorithm design and received the Master of Science degree in the fall of 2002. He joined the Protein Chemistry Technology Center (PCTC) at the University of Texas Southwestern Medical Center as a staff scientist and manager of protein ID facility lab which was surpervised by Dr. Yingming Zhao in the January of 2002. His persisting interest in biomedical research led him to continue graduate study at the University of Texas at Arlington, where he began his study in mass spectrometry and the application of proteomic analysis techniques under the guidance of Dr. Edward Bellion and Dr. Yingming Zhao in the Department of Chemistry and Biochemistry in the fall of 2005.