The application of LC/MS to exploration of potential biomarkers for disease and protein therapeutics

by Yanjun Liu

B.Sc. in Applied Pharmacy, Peking University
M.Sc. in Chemical Biology, Peking University

A dissertation submitted to

The Faculty of
the College of Science of
Northeastern University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

July 17th, 2017

Dissertation directed by

William S. Hancock
Professor of Chemistry and Chemical Biology
Acknowledgements

To all the people who helped, supported and encouraged me during my PhD life, I would like to express my sincere appreciation.

My deepest gratitude goes first and foremost to my advisor, Professor William Hancock for accepting me as his student and providing me several exciting projects which stimulated my strong interest in protein drug analysis. Professor Hancock has provided me guidance and priceless scientific insights which is extremely useful for my future career. I will be forever grateful to him for his mentoring during my doctoral studies.

I would like to express my heartfelt gratitude to my supervisor Dr. Shiaw-lin Wu, for his excellent instruction for my research. Without his valuable suggestion and thoughtful guidance, I may not have come this far.

I sincerely thank my dissertation committee members, Professor Alexandros Makriyannis, Professor George O’Doherty, Professor Roman Manetsch and Dr. Jared Auclair, for all the work they did for me to make the defense successful.

I would like to thank Dr. Barry Karger for proving an excellent research environment at the Barnett Institute. I would like to acknowledge all the members of the Barnett Institute with whom I had opportunities to work with. In Hancock’s lab, I appreciate the help and support of Dr. Fan Zhang, Dr. KyOnese Taylor, Yu Wang, and Di Wu. In Karger’s lab, I appreciate the help of Dr. Jason Wang, Dr. Siyuan Liu, Dr. Yuanwei Gao and Dr. Simion Kreimer. I also appreciate the help of Zhidan Chen from Vouros’s lab.

I deeply appreciate my collaborators: Professor Jordan Kreidberg and Dr. Yunjoon Jung from Boston Children’s Hospital; Professor J. Christopher Love and his labs from Massachusetts
Institute of Technology. They not only provide us samples generously but give us useful discussion as well.

I would like to express my gratitude to Professor Jeffrey Agar from Barnett Institute and want to acknowledge the staff in the department of chemistry and chemical biology and Barnett institute: Nancy Carbone, Andrew Bean, Cara Shockley, Emanuelle Hestermann and Richard Pumphrey for their kind help.

Finally, special thanks to my family especially for my mother and father. I thank them from the bottom of my heart for all the love and support they give to me. This dissertation is dedicated to my family!
Abstract

Due to the improvement of instrumentation, computer technology and associated software, mass spectrometry has become a powerful tool in proteomics research and protein analysis. It is a method of choice for the identification of proteins in large-scale proteomic research because of the high sensitivity of the mass spectrometric measurement. Furthermore, it is widely used in quality control of recombinant proteins and characterization of post translational modifications.

This thesis focuses on introducing the application of liquid chromatography coupled online with tandem mass spectrometry (LC-MS) technology to discover potential biomarkers for disease and analyze the posttranslational modifications of protein therapeutics.

In the Chapter 1, we introduce the commonly used mass spectrometers and ionization technologies for the analysis of large biomolecules. We also reviewed the current methodologies utilized in proteomics research and the preferred techniques used in protein therapeutic characterization.

Chapter 2 describes a study that used 1D electrophoresis followed by ultra-performance liquid chromatography coupled with tandem mass spectrometry (UPLC-MS/MS) to recognize polycystic kidney disease related changes in glycoproteins. Lectin Sambucus nigra bark (SNA) was used to enrich glycoproteins which have changes in sialic acid content in a mouse disease model (wild type (WT) and disease (pkd1/-) cell lines. Subsequently, 1D SDS PAGE was performed followed by in-gel digestion and UPLC-MS/MS analysis. As an example of our results, clinically interesting glycoproteins, such as Thrombospondin-1 (TSP-1) were detected only in the disease sample.

Chapter 3 represents the complete analysis of a complex glycoprotein. We used glycoproteomic analysis based on the combination of selected protease enzymes and nanoLC
tandem MS to characterize the glycosylation of Influenza hemagglutinin (H1/A/California/04/2009). PNGaseF treatment combined with trypsin or pepsin digestion were used to determine the glycosites and glycan occupancy. Three enzymes trypsin, AspN and pepsin were used separately to generate suitable glycopeptides for on-line LC tandem MS analysis. The glycan structure of a given glycopeptide was determined by collision-induced dissociation MS/MS fragmentation and the peptide backbone information was provided by CID-MS3 fragmentation. With this approach 100% sequence coverage of the hemagglutinin sample was obtained. Six glycosylation sites fitting the sequon N-X-S/T were successfully confirmed and the glycan heterogeneity as well as the ratios of glycoforms were determined at each site.

Chapter 4 developed a state-of-the-art analytical procedure for analyzing a polyethylene glycol(PEG)ylated peptide that is 8 residues long and conjugated with 20 kDa linear PEG. Multiple enzyme digestion, followed by HPLC-tandem mass spectrometry was applied to analyze the PEGylation. After comparing the result between the un-PEGylated peptide and PEGylated counterpart, the exact location of PEGylation sites at the peptide was confirmed as well as the stoichiometry of PEGylation.
# Table of Contents

Acknowledgements ............................................................................................................. ii

Abstract ................................................................................................................................ iv

Table of Contents .................................................................................................................. vi

List of Figures ......................................................................................................................... xii

List of Tables .......................................................................................................................... xv

List of Abbreviations ............................................................................................................ xvii

Chapter 1: Overview of LC/MS applications in Proteomics and protein therapeutics ........ 1

1.1 Introduction of Mass Spectrometry .................................................................................. 2

1.1.1 Introduction .................................................................................................................. 2

1.1.2 Types of Mass Analyzer ............................................................................................... 4

1.1.2.1 Time of Flight ........................................................................................................ 4

1.1.2.2 Transmission Quadrupole ..................................................................................... 5

1.1.2.3 Quadrupole Ion Traps ........................................................................................... 6

1.1.2.4 FTICR .................................................................................................................... 7

1.1.2.5 The Orbitrap ......................................................................................................... 8

1.1.2.6 Hybrid mass analyzer ........................................................................................... 11

1.1.3 Soft Ionization Techniques .......................................................................................... 12

1.1.3.1 Matrix-assisted laser desorption/ionization ......................................................... 12
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1.3.2</td>
<td>Electrospray</td>
<td>13</td>
</tr>
<tr>
<td>1.2</td>
<td>Mass spectrometry-based Proteomics</td>
<td>15</td>
</tr>
<tr>
<td>1.2.1</td>
<td>Proteomics</td>
<td>15</td>
</tr>
<tr>
<td>1.2.2</td>
<td>Separation technologies</td>
<td>15</td>
</tr>
<tr>
<td>1.2.2.1</td>
<td>Gel based separation: Two-dimensional gel electrophoresis</td>
<td>15</td>
</tr>
<tr>
<td>1.2.2.2</td>
<td>Unidimensional liquid-based separation techniques</td>
<td>16</td>
</tr>
<tr>
<td>1.2.2.3</td>
<td>Multidimensional liquid-based separation techniques</td>
<td>17</td>
</tr>
<tr>
<td>1.2.3</td>
<td>Tandem mass spectrometry (fragment method)</td>
<td>18</td>
</tr>
<tr>
<td>1.2.3.1</td>
<td>CID</td>
<td>18</td>
</tr>
<tr>
<td>1.2.3.2</td>
<td>ETD</td>
<td>19</td>
</tr>
<tr>
<td>1.2.4</td>
<td>Bioinformatics in Proteomic Analysis—Data Interpretation</td>
<td>19</td>
</tr>
<tr>
<td>1.2.4.1</td>
<td>Peptide Mass Fingerprinting</td>
<td>19</td>
</tr>
<tr>
<td>1.2.4.2</td>
<td>Peptide Fragmentation Fingerprinting (sequence Tagging)</td>
<td>20</td>
</tr>
<tr>
<td>1.2.4.3</td>
<td>Search Algorithms</td>
<td>20</td>
</tr>
<tr>
<td>1.2.4.4</td>
<td>De Novo Sequencing</td>
<td>21</td>
</tr>
<tr>
<td>1.2.5</td>
<td>Proteomics Strategy</td>
<td>21</td>
</tr>
<tr>
<td>1.2.5.1</td>
<td>Bottom Up Strategies</td>
<td>22</td>
</tr>
<tr>
<td>1.2.5.2</td>
<td>Top-Down Proteomics</td>
<td>22</td>
</tr>
<tr>
<td>1.2.6</td>
<td>Quantitation</td>
<td>24</td>
</tr>
<tr>
<td>1.2.6.1</td>
<td>Metabolic labeling</td>
<td>24</td>
</tr>
</tbody>
</table>
1.2.6.2 Post-biosynthetic labeling ..............................................................25
1.2.6.3 Absolute quantification using internal standards.................................26
1.2.6.4 Label-free quantification ..................................................................26
1.2.7 Application: Proteomics as a Tool for Biomarker Discovery .................26

1.3 Application of Proteomic methodologies to the characterization of protein therapeutics

1.3.1 Overview of biologics and biosimilars .................................................27
1.3.2 Issues or concerns of biologics and biosimilar ....................................28
1.3.3 Analysis of Molecular weight and amino acid sequence .........................29
1.3.4 Analysis of Protein Posttranslational Modifications ............................30
  1.3.4.1 Overview of Protein Post Translational Modifications (PTM) ............30
  1.3.4.2 PEGylation (Chemical modification) ................................................30
  1.3.4.3 Glycosylation ...............................................................................34
  1.3.4.4 Disulfide bond linkage analysis .....................................................35
  1.3.4.5 Stability related modification .........................................................36
1.4 Conclusion and perspective ......................................................................38
1.5 References ..............................................................................................38

Chapter 2: UPLC-MS/MS analysis of glycoprotein changes in Mouse Kidney cells with the phenotype of autosomal dominant polycystic kidney disease ..................................................55

2.1 Abstract .................................................................................................56
3.3.1 Samples and Reagents ................................................................. 81
3.3.2 Gel electrophoresis (SDS-PAGE) analysis ........................................... 82
3.3.3 Enzyme digestion ............................................................................... 82
3.3.4 NanoLC tandem analysis ................................................................. 83
3.3.5 Data analysis ...................................................................................... 84
3.4 Result and Discussion ........................................................................... 84
   3.4.1 Six out of eight possible glycosylation sites were glycosylated .............. 86
   3.4.2 Tandem MS/MS data used to confirm that N28 is the major glycosylated site in peptide 22 to 34 ................................................................. 87
   3.4.3 Glycan occupancy analysis of HA antigen ........................................... 90
   3.4.4 Oligosaccharide Heterogeneity characterization of site N28 as an example ...... 90
   3.4.1 The glycan ratios at all six sites of HA antigen and major glycan types ........ 91
3.5 Conclusions ........................................................................................... 96
3.6 Reference ............................................................................................... 96
3.7 Supplementary Data .............................................................................. 100

Chapter 4: An Approach to Site-Specific Characterization and Quantitation of Polyethylene glycol (PEG) Derivatives of Polypeptides ........................................ 110

   4.1 Abstract .............................................................................................. 111
   4.2 Introduction .......................................................................................... 111
   4.3 Material and methods .......................................................................... 113
4.3.1 Reagents and materials.................................................................113
4.3.2 Direct-infusion and ESI-LTQ MS analysis ........................................114
4.3.3 MALDI-TOF analysis ....................................................................114
4.3.4 Enzyme digestion ........................................................................115
4.3.5 LC-MS analysis ...........................................................................115
4.3.6 LC-UV sample collection ...............................................................116
4.4 Results and Discussion .....................................................................116
4.4.1 Intact analysis of the PEGylated KXLXQAQX. .................................116
4.4.2 LC-UV-MALDI-TOF off line analysis of the PEGylated KTXSQAQS ..........120
4.4.4 Quantitation analysis of the PEGylated KXLXQAQX.........................127
4.5 Conclusion .......................................................................................130
4.6 Supplementary Data ........................................................................130
4.7 Reference .......................................................................................135
Conclusion ............................................................................................139
List of Figures

Chapter 1: Overview of LC/MS applications in Proteomics and protein therapeutics

Figure 1.1. Soft ionization technology and mass analyzers used in proteome research........8
Figure 1.2. Cross-section of the C-trap and Orbitrap analyzer ........................................9
Figure 1.3. Schematic representation of a hybrid ion trap-Orbitrap mass spectrometer........12
Figure 1.4. Peptide backbone fragment notation .............................................................18
Figure 1.5. Strategies for MS-based protein identification and characterization.............23
Figure 1.6. Metabolic stable isotope coding of proteomes .............................................24
Figure 1.7. Enzyme and Chemical stable isotope-coding of proteomes ......................25
Figure 1.8. Sources of variation between manufacturing steps in a recombinant protein
production ....................................................................................................................28
Figure 1.9. Deamidation of asparaginyl residues and isomerization of aspartyl residues via a
succinimide intermediate .........................................................................................37

Chapter 2: UPLC-MS/MS analysis of glycoprotein changes in Mouse Kidney cells with
the phenotype of autosomal dominant polycystic kidney disease

Figure 2.1. The number of proteins and glycoproteins identified in wide type or mutant cell
(KO) lysates ..................................................................................................................62
Figure 2.2. gel image for proteins by SNA pull down .....................................................64
Figure 2.3. Representative LC/MS spectra of an unique peptide from glutamyl aminopeptidase
........................................................................................................................................67
Figure 2.4. Representative tandem MS spectra of unique peptides from the other two potential
candidates ....................................................................................................................68
Chapter 3: Site-Specific glycosylation characterization of Influenza A Virus

Hemagglutinin produced by Spodoptera frugiperda insect cell line using multienzyme
digestion with nanoLC mass spectrometry

Figure 3.1. HA antigen (H1/A/California/04/2009) is a glycoprotein with 8 possible
glycosylation sites

Figure 3.2. MS and tandem MS analysis of the peptide IGYHANDSTDTVD (residues 22-34)
containing the site N28 prepared by pepsin digestion

Figure 3.3. Representative glycopeptide analysis of HA antigen digests

Figure 3.4. 100% sequence coverage of HA antigen

Figure 3.5. MS and tandem MS analysis of peptide WMCSNGSLQC, 553-562

Figure 3.6. MS and tandem MS analysis of GluC digestion peptide containing the glycosylation
site, residue 28

Figure 3.7. MS and tandem MS analysis of Tryptic peptide containing the residue 28

Figure 3.8. LC/MS analysis of glycopeptides containing N40, N293 and N304

Figure 3.9. LC/MS analysis of glycopeptides containing N104 and N498

Chapter 4: An Approach to Site-Specific Characterization and Quantitation of
Polyethylene glycol (PEG) Derivatives of Polypeptides

Figure 4.1. ESI-LTQ MS spectrum of PEGylated KXLXQAQX (~20 kDa PEG)

Figure 4.2. MALDI-TOF intact analysis of the PEGylated peptide using different matrices.

Figure 4.3. MALDI TOF intact mass analysis of the PEGylated peptide using the reflectron
mode

Figure 4.4. HPLC-UV analysis of PEGylated KXLXQAQX on a Vydac C4 column
Figure 4.5. MALDI-TOF analysis of fraction 3 (PEGylated KXLXQAQX) in the LC-UV separation.................................................................122

Figure 4.6. LC tandem MS analysis of trypsin digests of PEGylated KXLXQAQX and its unPEGylated counterpart .................................................................124

Figure 4.7. Mass spectra of pepsin digests of PEGylated KXLXQAQX and its unPEGylated counterpart .................................................................126

Figure 4.8. LC-MS analysis of trypsin digests of PEGylated MKXLXQAQXK and its unPEGylated counterpart .................................................................127

Figure 4.9. Quantitative analysis of PEGylated KXLXQAQX by standard addition ..........128

Figure 4.10. Quantitative analysis of PEGylated KXLXQAQX ........................................129

Figure 4.11. MALDI TOF intact analysis of the PEGylated MKXLXQAQXK. ...............132

Figure 4.12. LC-MS analysis of pepsin digests of PEGylated MKXLXQAQXK and its unPEGylated counterpart .................................................................133

Figure 4.13. XIC of peptide KXLXQAQX m/z 431.737, (charge 1+) ..........................133

Figure 4.14. Quantitative analysis of PGEylated KXLXQAQX by the standard addition approach (duplicate) .................................................................134

Figure 4.15. Linear curve generated by the digests (XLXQAQX) of standard peptides with PEGylated peptide mixtures (duplicate) ........................................135
List of Tables

Chapter 1: Overview of LC/MS applications in Proteomics and protein therapeutics

Table 1.1. Performance comparisons of the mass spectrometry instruments ..................11
Table 1.2. Structures and properties of commonly used MALDI Matrices .......................13
Table 1.3. PEGylated biopharmaceuticals on the market .............................................33

Chapter 2: UPLC-MS/MS analysis of glycoprotein changes in Mouse Kidney cells with the phenotype of autosomal dominant polycystic kidney disease

Table 2.1. Upregulated glycoproteins in Pkd1-/- cells .............................................63
Table 2.2. Glycoproteins identified by LC/MS analysis of the SNA pull-down ...............65
Table 2.3. KEGG pathway analysis of differently expressed proteins ...........................72
Table 2.4. Potential glycoprotein candidates for future study .....................................73

Chapter 3: Site-Specific glycosylation characterization of Influenza A Virus

Hemagglutinin produced by Spodoptera frugiperda insect cell line using multienzyme digestion with nanoLC mass spectrometry

Table 3.1. Glycopeptides and Non-glycosylated peptides identified by trypsin or pepsin after PNGaseF treatment .................................................................86
Table 3.2. Quantitation of glycan occupancy at each glycosylation sites .....................89
Table 3.3. Relative quantitation of different glycoforms at each glycosylation sites ........95
Table 3.4. Glycan names and their structures mentioned in the study .......................101
Chapter 4: An Approach to Site-Specific Characterization and Quantitation of Polyethylene glycol (PEG) Derivatives of Polypeptides

Table 4.1. The mass and the intensity of PEGylated KXLXQAQX used for calculation of Mn, Mw and PD. .......................................................... 130

Table 4.2. UnPEGylated peptide percentage and the LC/MS peak area used to generate the standard curve by the standard addition approach. ............................................................... 134

Table 4.3. UnPEGylated peptide percentage and the LC/MS peak area used to generate the standard curve by the digests of standard peptides with PEGylated peptide mixtures. .......... 135
List of Abbreviations

MS: Mass spectrometry

GC: Gas chromatography

EI: Electron ionization

CI: Chemical ionization

FAB: Fast atom bombardment

MALDI: Matrix-assisted laser desorption/ionization

ESI: Electrospray ionization

TOF: Time-of-flight

FT-ICR: Fourier transform ion cyclotron resonance

FWHM: full width at half maximum

QMF: Quadrupole mass filter

DC: Direct current

RF: Radio frequency

3D QIT: The three-dimensional quadrupole ion trap

AGC: Automatic gain control

LIT: Linear ion trap

FTMS: Fourier transform-based mass spectrometers
HPLC: High performance liquid chromatography

API: Atmospheric pressure ionization

APCI: Atmospheric Pressure Chemical Ionization

APPI: Atmospheric Pressure Photoionization

PTMs: post-translational modifications

2D-PAGE: Two-dimensional polyacrylamide gel-electrophoresis

MW: molecular weight

IEC: Ion exchange chromatography

SEC: Size exclusion chromatography

HILIC: Hydrophilic interaction chromatography

CE: capillary electrophoresis

Pc: Peak capacity

UPLC: Ultra-high-pressure liquid chromatography

CID: Collision-induced dissociation

ECD/ETD: Electron capture and transfer dissociation

HCD: Higher energy collisional dissociation

PFF: Peptide fragmentation fingerprinting

PMF: Peptide mass fingerprinting
ppm: parts-per-million, $10^{-6}$

Xcorr: cross-correlation

AP: Affinity purification

SILAC: Stable Isotope Labeling by/with Amino acids in Cell culture

ICAT: Isotope-coded affinity tag

iTRAQ: isotope tags for relative and absolute quantification

FDA: US Food and Drug Administration

AA: amino acid

PEG: Polyethylene glycol

NMR: Nuclear magnetic resonance spectroscopy
Chapter 1: Overview of LC/MS applications in Proteomics and protein therapeutics
1.1 Introduction of Mass Spectrometry

1.1.1 Introduction

In 1886, the German physicist Eugen Goldstein discovered Kanalstrahlen or canal rays in a gas discharge tube [1]. Wilhelm Wien later showed that these positive ions of the canal rays can be deflected in magnetic fields [2, 3]. These discoveries became the fundamental principles for mass spectrometry, leading J.J. Thompson to build the first mass spectrometer in 1912. At that time, the major use of early mass spectrometers was to measure the atomic weight of the elements and the relative abundance of their isotopes [4].

By the 1940s, the application of mass spectrometry (MS) was extended to organic chemistry, where it was used to monitor the concentration of small hydrocarbons for the petroleum industry [5]. During the 1950s, the mass spectrometer was used as the detector of gas chromatography (GC), creating the GC/MS tool [6]. The emergence of GC/MS expanded the utilization of mass spectrometry in organic chemistry. Organic compound mixtures are separated to individual constituents by GC and delivered to mass spectrometer, where a beam of high-energy electrons converts the analytes to molecular ions (M+). This ionization method is called Electron ionization (EI), which was first introduced by Arthur J. Dempster in his report "A new method of positive ray analysis" [7]. The GC-EI-MS system gave chemists structural information for unknown compounds by providing fragmentation patterns in the mass spectrum. The complimentary ionization method for EI is called Chemical Ionization (CI), which was invented in the 1960s by Burnaby Munson and Frank Field. CI is the first soft ionization technique where molecular ions are more likely to remain intact [8]. The EI or CI GC/MS helped chemists to better understand the structure of organic compounds, making MS a powerful tool in organic chemistry study.
The nature of MS requires that the analytes are charged and gaseous, however, biomolecules such as peptides and proteins are large and polar and thus the ionization process and transfer to the gas phase is not as easy as for the small molecules. It was not until the introduction of the Fast atom bombardment (FAB) method in 1981 that a large biomolecule, such as a protein, could be successfully ionized [9, 10]. The advent of Electrospray ionization (ESI) and Matrix-assisted laser desorption/ionization (MALDI) further improved the ionization of these biomolecules. ESI technique was firstly introduced by Masamichi Yamashita and John Fenn in 1984 [11]. Then, in 1985 MALDI was reported by Franz Hillenkamp, Michael Karas and their colleagues [12]. A breakthrough happened in 1987, Koichi Tanaka and his co-workers, using the combination of a 337-nm laser with 30 nm cobalt particles in glycerol, successfully ionized the large protein, carboxypeptidase-A (~34 kDa) [13].

The MS instruments which employed magnetic fields or a combination of magnetic and electric fields (e.g. double focusing Mass spectrometer) dominated advancements for many years. Time-of-flight (TOF) mass analyzer proposed by Stephens in 1946, was quickly replaced by a quadrupole mass analyzer due to its slow data recording system and limited resolution [5]. Later, the emergence of the soft ionization technology largely stimulated the application of MS into biomolecule studies and brought TOF MS back into use. The development of ionization technology stimulated the improvement of the mass analyzer. Different types of mass analyzers were developed, commercialized and substituted rapidly to meet the sensitivity requirement for proteomic studies. Today, MS has become a powerful tool widely used in biology, chemistry, pharmacology, medical science, and environmental studies. [14].
1.1.2 Types of Mass Analyzer

No area of MS updates more rapidly than the innovations in the mass analyzer [15]. Different mass analyzers take advantage of different ion behaviors in electric or/and magnetic fields, to provide the identity and quantity of the individual ion group. There are several types of mass analyzers, such as magnetic-sector instruments (single-focusing and double-focusing), TOF m/z analyzer, transmission quadrupole, quadrupole ion trap, Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR MS) and the latest development of trapping devices, Orbitrap [16]. This section will introduce five commonly used mass analyzers in the proteomic research.

1.1.2.1 Time of Flight

The time of flight (TOF) mass spectrometer measures the time that an ion travels from ion source to a detector after acceleration [17]. The ions dissociate from the material and go into the electric acceleration field. Ideally, they share the same kinetic energy and the field free region allows the ion drift from ion source to the detector without any interference. Ions with a different m/z obtain a corresponding velocity and hit the detector sequentially. Because the drift time is mass dependent, the time spectrum is directly converted to a mass spectrum.

The benefit of the TOF mass analyzer is that it does not have an upper m/z limitation. However, the disadvantages of the early TOF instrument are clear. First, in the actual situation, it is difficult to ensure all ions start at the same position and obtain identical kinetic energy. A high distribution of the ions initial kinetic energy results in broad spectral peaks and low resolution. Second, the ions will travel approximately 1 to 2 meters from ion source to detector, increasing the possibility that they interact with any neutral molecules. This means that TOF requires a higher vacuum than other mass analyzers with shorter travel distances.
**Linear**

The low resolving power of the early linear TOF, caused by high energetic variation of the initial ion packet, is a problem. The technique commonly used for improving the resolution of linear TOF MS is known as delayed extraction (also called time-lag focusing or pulsed ion extraction) which is invented by Wiley & McLaren in 1955 [18]. Delayed extraction means that there is a time difference between the end of the ionization pulse and start of the extraction pulse. This enables slower ions to catch up to the faster ions during the delay time by receiving enough potential energy after travelling sufficient distance from the pulsed acceleration system. Although simple linear TOF spectrometers are still considered to be low resolution mass analyzers, they are still used coupled with MALDI in biological laboratories, being used to solve structural problems with biological molecules [19].

**Relectron**

An electrostatic mirror consumes the kinetic energy differential where the opposite electric field causes the ion velocity to decrease to zero. Ions with higher velocity penetrate deeper into the mirror before they reflect to the detector. This ensures that the same mass ions hit the detector together [20]. The most conspicuous advantage of the re TOF is the higher resolving power [commercial instruments for MALDI can reach or over 20,000 (full width at half maximum, FWHM) and for GC/MS, LC/MS typically 7000-15000 (FWHM)].

**1.1.2.2 Transmission Quadrupole**

Transmission Quadrupole is also known as the quadrupole mass filter (QMF) which employs both direct current (DC) and radio frequency (RF) electric fields. The QMF consists of four parallel electrical rods. A RF voltage and a DC offset voltage are applied to an opposing rod pair. The RF
(V) and DC (U) voltages are scanned at a fixed ratio and the RF frequency is constant. By a given set of DC and RF potentials, the nearly hyperbolic electric fields allow only certain m/z value ions to pass the QMF and hit the detector [21].

The merits of quadrupole analyzers include low cost, robustness, easy maintenance, small size and compatibility with the high vacuum, which allows it to connect with the GC/MS [22]. However, its limited mass range (<4000Da) and relatively low resolving power are two drawbacks in its usage. The design of triple quadrupole (three quadrupoles linked together—the first and the third serve as mass analyzers while the second as the collision cell) compensates for the limited capability of a quadrupole to perform MS/MS analysis. The RF only quadrupole can be used as an ion focusing device and/or an ion filter in a hybrid mass spectrometer [23]. The use of hexapoles or octapoles significantly improve the ion focusing function but with the loss of performance to be an ion filter.

### 1.1.2.3 Quadrupole Ion Traps

The quadrupole ion trap serves not only as an ion storage device restricting gaseous ions for a period but also as a mass analyzer differentiating charged ions based on their m/z. It is a versatile technique with high sensitivity and specificity.

### 3D Quadrupole Ion Trap m/z Analyzers

The 3D quadrupole ion trap (3D QIT) uses the three-dimensional quadrupole electric field to trap charged particles. It consists of two end-cap electrodes and one ring electrode. These three electrodes form an ideal quadrupole electric field by the hyperboloidal geometric design of all the electrodes, which will produce a parabolic potential well to confine the ions. In the device, two end-cap electrodes are normally held at ground and a unipolar RF potential plus any DC voltage
are applied to the ring electrode. After ramping the amplitude of the RF, the ions with a higher m/z become unstable and travel through the exit end cap electrode and then reach the detector [24].

The high scan rate of QIT (5K-10K m/z units per second) makes this mass analyzer cover a large range of m/z values in a short period. A minimum number of ions must be stored in the trap analyzer for yielding good quality spectra, but too many ions will cause the space charge effect that results in anomalies in mass spectra. Automatic gain control (AGC) is an operational parameter to alleviate the results of space charge effect. It was developed at Finnigan Corp. to give a limitation of the maximum ions in the trap. AGC describes the process that ion flux is determined during a pre-ionization run (microscan) and then the ionization period is adjusted to generate the prescribed number of ions from the scanned ion flux [25].

**Linear Quadrupole Ion Trap (LIT) m/z Analyzer**

Unlike 3D QIT, linear ion trap (LIT) is two-dimensional quadrupole ion trap (2D QIT). It adopts a set of quadrupole rods, the basic structure of QMF which applies a static electrical potential at the end of electrodes to trap ions radially. It can serve as a mass filter or an ion trap. LIT has superior ion storage capability compared to 3D QIT because it can accommodate a larger number of ions [25].

**1.1.2.4 FTICR**

FTICR (Fourier transform ion cyclotron resonance) mass spectrometer is derived from a technique called ion cyclotron resonance (ICR). In the mass spectrometer, the ion enters the ICR cell and is confined by the combination of a magnetic field and electric field in a Penning trap. At first, ions have very little thermal motion and then a RF potential is applied to a pair of opposite plates in the ICR cell, exciting ions with different m/z values [26].
Figure 1.1. Soft ionization technology and mass analyzers used in proteome research [27].

Note: Reproduced from “Mass spectrometry-based proteomics” by permission of Nature Publishing Group.

After removing the oscillating electric field, the ions rotate at their cyclotron frequency, inducing an oscillating charge (image current). The frequency of the image current is detected and amplified by the receiver plate and the detected signal is “translated” to a mass spectrum by performing a Fourier transform. Frequency is the physical parameter of the ion that can be measured very accurately. This allows FTMS to outperform other mass spectrometers in terms of the maximum mass resolution and mass accuracy [28]. Figure 1.1 displays the schematics of two ionization methods and different mass analyzers that are used in proteomic research.

1.1.2.5 The Orbitrap

Orbitrap is a new type of mass analyzer which includes two electrodes; an outer barrel-shaped surface and an inner spindle shaped electrode form coaxial axisymmetric electrodes [29, 30]. A
constant electric potential is applied between the two coaxial axisymmetric electrodes. Stable ion routes in the Orbitrap mass analyzer involve a combination of orbiting motion around the central electrode and simultaneous oscillations along the z-axis.

**Figure 1.2. Cross-section of the C-trap and Orbitrap analyzer [29].** Note: Reproduced from “Orbitrap Mass Spectrometry” by permission of American Chemical Society.

Injected ions are constrained between the field of the two electrodes and follow a circular orbit around the inner electrode (Figure 1.2). The radius of the circular movement is determined by balancing an electrodynamic centripetal force and the centrifugal forces associated with the ion initial tangential velocity. Due to mass independence of the radius of ion orbital migration, Orbitrap stores ions of all masses radially.
The electric field between the coaxial electrodes is inhomogeneous in a symmetric manner. It accelerates the ions from the point of ion entry to the center of the Orbitrap, but decelerates them when they continue to the opposite end of the Orbitrap. After consuming the kinetic energy gained before, the ions are accelerated back to the center of the trap. The frequency of the axial oscillation (z-direction) is mass dependent, allowing the Orbitrap to serve as a mass spectrometer. The image current detection senses these oscillations as a time domain signal. The frequencies of oscillating image current are converted to mass spectra via a Fourier transform, which is similar to the manner used in FTICRMS [31].

FTICR and Orbitrap analyzers are commonly categorized to Fourier transform-based mass spectrometers (FTMS). Both exhibit outstanding performance in mass resolution and accuracy even for small numbers of ions. Identification with adequately high mass accuracy is useful to determine composition and thus identity of a compound [31]. FTMS instruments have begun to dominate the area of proteomics research because of their performance in resolving power, mass accuracy, and dynamic range. Compared with FTICR, the Orbitrap mass analyzer is a small and simple device with a larger tapping volume. However, it is also the highest vacuum requirement of a mass analyzer [25]. Table 1.1 summarizes the performance of different MS instruments.
Table 1.1. Performance comparisons of the mass spectrometry instruments [32]. Note:

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Applications</th>
<th>Resolution</th>
<th>Mass accuracy</th>
<th>Sensitivity</th>
<th>Dynamic range</th>
<th>Scan rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIT (LTQ)</td>
<td>Bottom-up protein identification in high-complexity, high-throughput analysis, LC-MS² capabilities</td>
<td>2000</td>
<td>100 ppm</td>
<td>Femtomole</td>
<td>1e4</td>
<td>Fast</td>
</tr>
<tr>
<td>TQ (TSQ)</td>
<td>Bottom-up peptide and protein quantification; medium complexity samples, peptide and protein quantification (SRM, MRM, precursor, product, neutral fragment monitoring)</td>
<td>2000</td>
<td>100 ppm</td>
<td>Attomole</td>
<td>1e6</td>
<td>Moderate</td>
</tr>
<tr>
<td>LTQ-Orbitrap</td>
<td>Protein identification, quantification, PTM identification</td>
<td>100,000</td>
<td>2 ppm</td>
<td>Femtomole</td>
<td>1e4</td>
<td>Moderate</td>
</tr>
<tr>
<td>LTQFTICR, Q-FTICR</td>
<td>Protein identification, quantification, PTM identification, top-down protein identification</td>
<td>500,000</td>
<td>&lt;2 ppm</td>
<td>Femtomole</td>
<td>1e4</td>
<td>Slow; slow</td>
</tr>
<tr>
<td>Q-TOF, IT-TOF</td>
<td>Bottom-up, top-down protein identification, PTM identification</td>
<td>10,000</td>
<td>2-5 ppm</td>
<td>Attomole</td>
<td>1e6</td>
<td>Moderate, fast</td>
</tr>
<tr>
<td>Q-LIT</td>
<td>Bottom-up peptide and protein quantification; medium complexity samples, peptide and protein quantification (SRM, MRM, precursor, product, neutral fragment monitoring)</td>
<td>2000</td>
<td>100 ppm</td>
<td>Attomole</td>
<td>1e6</td>
<td>Moderate, fast</td>
</tr>
</tbody>
</table>

1.1.2.6 Hybrid mass analyzer

To fulfill different separation, resolution and detection demands, hybrid mass spectrometers employ the best features of each mass analyzer to maximize precision, resolution and robustness. Commonly combined mass analyzers include Q-TOF (quadrupole plus time of flight), Q-Trap (quadrupole plus ion trap), LTQ-Orbitrap (linear ion trap plus Orbitrap) and Q Exactive (quadrupole plus Orbitrap).

Thermo Scientific™ LTQ Orbitrap™, includes a hyperbolic QMF (better selects precursor ions) and an Orbitrap mass analyzer (high resolution). This hybrid FTMS instrument contains an HCD collision cell to fragment larger ions effectively, improving detection and confirmation of molecular identities [31]. A schematic representation of a hybrid ion trap Orbitrap mass spectrometer is displayed in Figure 1.3.
Figure 1.3. Schematic representation of a hybrid ion trap-Orbitrap mass spectrometer [31].
Note: Reproduced from “Fourier Transform Mass Spectrometry” by permission of American Society for Biochemistry and Molecular Biology.

1.1.3 Soft Ionization Techniques

1.1.3.1 Matrix-assisted laser desorption/ionization

Matrix-assisted laser desorption/ionization (MALDI) is a soft ionization technique especially for large and/or labile molecules. In this ionization technique, analytes co-deposit with a large excess of matrix (typically weak organic acid) on a plate (molar ratio of analyte to matrix is approximately 1:5000) and generate gas phase, protonated or deprotonated molecules by laser irradiation. The co-crystallization of the analytes and the matrix is the key element for a successful experiment. The matrix is responsible for absorbing radiation and inducing the nondestructive vaporization of the analyte indirectly [33, 34]. The matrix also serves as a proton donor and receptor during ionization. Typically used matrices in protein or peptide analyses are displayed in Table 1.2.
Table 1.2. Structures and properties of commonly used MALDI Matrices

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Structure</th>
<th>Sample Type</th>
<th>Solvent</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,5-dihydroxy benzoic acid (DHB)</td>
<td><img src="image" alt="Structure" /></td>
<td>Peptides, proteins, polymers and carbohydrates</td>
<td>10 mg/ml, 30:70 [v/v] acetonitrile: water with 0.1% TFA</td>
</tr>
<tr>
<td>3,5-dimethoxy-4-hydroxycinnamic acid (sinapic acid, SA)</td>
<td><img src="image" alt="Structure" /></td>
<td>Good choice for larger proteins (10–150 kDa) and some polar polymers</td>
<td>10 mg/ml, 30:70 [v/v] acetonitrile: water with 0.1% TFA</td>
</tr>
<tr>
<td>α-Cyano-4-hydroxycinnamic acid (CHCA)</td>
<td><img src="image" alt="Structure" /></td>
<td>First choice for peptides and small proteins (&lt; 10 kDa)</td>
<td>10 mg/ml, 50:50 [v/v] acetonitrile: water with 0.3% TFA</td>
</tr>
</tbody>
</table>

1.1.3.2 Electrospray

Electrospray (ESI) is another widely used ionization method to study the weakly bound, noncovalent complexes such as carbohydrates, peptides and proteins[35]. In the ionization process, a high-voltage is applied to the liquid containing the analytes and the liquid is dispersed to a cone-shaped aerosol. For the high flow ESI, an inert gas such as nitrogen or carbon dioxide is utilized for additional nebulization.

The aerosol flows into a heated capillary carrying a potential difference of 3000V for further solvent evaporation to form charged droplets. A charged droplet becomes unstable when it reaches
its Rayleigh limit and deforms due to the electrostatic repulsion of charges becomes more powerful than the surface tension of the droplet. ESI uses extensive solvent evaporation and typically the composition of the solvents includes water and volatile organic compounds. These compounds increase the conductivity of the solution and facilitate the ionization process. ESI generates multiple charged ions, which effectively extend the mass range of the analytes that can be analyzed. ESI-MS has become a crucial technique in various laboratories because it provides sensitive, robust and reliable analysis even in microliter sample volumes. By coupling with various separation technologies such as high-performance-liquid-chromatography (HPLC), ESI-MS has emerged as a powerful tool for complex biological sample analysis.

**Atmospheric Pressure Ionization (API)**

The versatile detection power gives MS the possibility to serve as a universal detector for HPLC. However, large amounts of liquid solvent and nonvolatile and/or polar compounds eluted from the HPLC separation are two major problems limiting the application of LC/MS technique. The availability of atmospheric pressure ionization (API) sources improves the ability to couple HPLC with MS. The API interface includes Atmospheric Pressure Chemical Ionization (APCI), ESI and Atmospheric Pressure Photoionization (APPI) [9]. The APCI source commonly consists of a nebulizer probe (350-500 °C), a corona discharge needle and an ion-transfer region. APCI is a complimentary ionization technology to ESI, usually generates singly-charged ions and is typically used for less polar, thermally stable molecules whose molecular weight is under 1500 Da [36, 37]. APPI is another ionization technique which is used for compounds that possess ionization energies (IEs) below 10 eV, are ionized by 10 eV protons which are emitted by a krypton discharge lamp [38]. It extends the LC-MS combinations to molecules that are poorly amenable to APCI or ESI [39].
1.2 Mass spectrometry-based Proteomics

1.2.1 Proteomics

Proteomics is defined as the large-scale study of the entire set of proteins in a cell line, tissue or organism. The major goal of proteomics is to generate an integrated view of biological systems by analyzing all the proteins in the system rather than each one individually [40, 41]. The proteomics methodology not only involves the separation, identification, and quantification of proteins but includes the characterization and analysis of protein post-translational modifications (PTMs) and sequence variants (mutants and spliced isoforms) as well. Today, mass spectrometry has become the essential tool in the proteomics field based on its high resolution, high mass accuracy and sufficient sensitivity for many biological problems.

1.2.2 Separation technologies

1.2.2.1 Gel based separation: Two-dimensional gel electrophoresis

Two-dimensional polyacrylamide gel-electrophoresis (2D-PAGE) is a classical technique developed for protein separation. It played dominant role for many years in the proteomics studies when it was introduced in the mid-70’s. Its high efficiency of protein separation comes from the combination of independent parameter in each dimension. In the first dimension, proteins are separated based on their different isoelectric point (pI) by isoelectric focusing and in the second dimension, a discontinuous SDS-PAGE system trap the proteins according to their size [42-44]. Traditionally, two-dimensional gel electrophoresis (2-DE) is the standard method for comparing protein expression between normal and disease-perturbed state. Nevertheless, the low resolution and sensitivity are the major issue of this technique as it can only be used to resolve and study the abundant proteins. Furthermore, the reproducibility of the protein spots observed in the gel is not
high and certain type of proteins, such as hydrophobic membrane protein, very acidic or basic proteins and high MW proteins, are difficult to analyze using gel based technique.

1.2.2.2 Unidimensional liquid-based separation techniques

The mass spectrometry can only tolerate certain complex samples because it is sensitive to ion suppression[45, 46] and it has limited dynamic range [47]. Reducing the component overlap plays an important role in effective MS analysis and the subsequent protein identification and quantitation. Substantial effort has been conducted for improving the resolution of such unidimensional separations [48, 49]. The physicochemical diversity of peptides, including charge, isoelectric point, hydrophobicity and size can be utilized in the separation technology. Various types of liquid based separation such as reversed-phase LC, ion exchange chromatography (IEC) [50], size exclusion chromatography (SEC) [51], hydrophilic interaction chromatography (HILIC) [52], affinity chromatography [53], and capillary electrophoresis (CE) [54] were used in both in quality control of biotherapeutic manufacture and proteomic research to resolve peptide mixtures.

Reversed phase HPLC is the most widely used technique for peptide mapping since it was introduced at the mid-1980s [55, 56]. Peak capacity (Pc) is the best assessment of the resolution of separation under gradient conditions [57] and it can be defined as the number of peaks that can be separated in a defined timeframe at unit resolution and calculated using the total timeframe that peptides elute divided by the average peak width (4σ). Conventional HPLC is operated at pressure < 600 bar and the particle size of the column is 2.5–5 microns. Ultra-high-pressure liquid chromatography (UPLC) is designed to operate at pressures over 600 bar. The particle size of column for UPLC system can decrease to less than 2.5μm (sub 2-micron). Small particle size generates short diffusion path between the stationary phase and analytes providing a high efficiency separation that doesn’t diminish at increased flow rates, according to the Van Deemter
equation [58-60]. The application of UPLC system can extend the speed and peak capacity of a separation to new limits to a high sensitivity analysis. Tolley et al. demonstrated that better signal-to-noise ratio and higher peak capacity were obtained by performing a separation at pressures ranging from 790 to 930 bar in 22-cm-long RP columns packed with 1.5-µm nonporous silica particles [61]. An effort has also been made to maximize the peak capacity by using conventional HPLC instruments (≤ 400 bar) with packed columns in one-dimensional separations of complex samples. It has been reported that comparable peak capacities to those obtained in UPLC for complex peptides mixture were obtained by employing a 60 cm column set packed with 5 µm pellicular (superficially porous) particles [62].

1.2.2.3 Multidimensional liquid-based separation techniques

It is generally recognized that only one chromatographic or electrophoretic separation is not capable of resolving a large number of peptides resulting from a proteolytic digest of a complex proteome. Thus, combination of two or more orthogonal separation techniques dramatically increases the overall resolution and peptide identification from digests of complex proteomes or high MW biopharmaceuticals. There are several good reviews that introduce the multidimensional separation system and as well as applications [63-66]. Unlike the parent proteins, peptides are more soluble and easier to separate. Hence, in the chromatographic and capillary electrophoresis multidimensional separation approaches proteins are commonly digested first before the separation step [67]. However, multidimensional techniques also successfully applied to separate protein mixtures [68, 69].
1.2.3  **Tandem mass spectrometry (fragment method)**

Soft ionization techniques primarily yield little or no fragmentation of protonated or deprotonated molecular ions. Therefore, single-stage mass spectrometers usually cannot provide structural information of analytes. Tandem mass spectrometry has become a vital technique for the structural analysis of polar, thermally labile molecules, such as pharmaceutical peptide or protein drugs and nucleic acids. Collision-induced dissociation (CID) and electron capture and transfer dissociation (ECD/ETD) are the most popular ion dissociation approaches applied in tandem mass spectrometry.

![Peptide backbone fragment notation](image-url)

**Figure 1.4. Peptide backbone fragment notation.**

1.2.3.1 **CID**

The selected ion (precursor ion) is directed into a collision cell and after interactions with the atoms of the neutral collision gas (like nitrogen helium or argon), the precursor ion is converted to the product ion. In the process, part of the kinetic energy is converted to internal energy and the vibrational energy will redistribute over the molecular ion that can result in amide bond breakage along the peptide.  Collision Induced dissociation usually generate b- and y-type product ions or
loss of small neutral molecules, such as water and/or ammonia from side chains. Ions obtaining energy from an electric field is correlated to the charges they carry, so ESI-CID MS/MS often efficiently decomposes a precursor ion because of the multiply protonated molecules ESI produces [70].

**HCD**

The higher energy collisional dissociation (HCD) is a CID technique available for the LTQ Orbitrap. Ions are fragmented in a collision cell instead of an ion trap and sent back to the Orbitrap through the C-trap for high resolution analysis. Compared with the traditional CID fragmentation pattern, HCD is featured with no low-mass cutoff, high resolution ion detection, and higher quality MS/MS spectra resulting from the increased ion fragmentation [71]. HCD also provides b- and y ions, while the higher energy induces a predominance of y-ions and b-ions can undergo further fragmented forming a-ions or smaller species [72].

1.2.3.2 ETD

Electron transfer dissociation (ETD) is a new fragmentation technique by which electron was transferred to a multiply protonated peptide/protein. This technology results in the cleavage of the N-Cα backbone bonds and generates c- and z-type fragment ions. The nomenclature for CID and ETD is listed in Figure 1.4. ETD fragmentation can provide information about CID labile posttranslational modifications (PTMs) such as phosphate groups. Theoretically, this fragmentation method can yield both the sequence information and the sites of modification.

1.2.4 Bioinformatics in Proteomic Analysis—Data Interpretation

1.2.4.1 Peptide Mass Fingerprinting
One strategy for proteins identification involves 1D or 2D SDS-PAGE separations coupled with in-gel digestion and MALDI-TOF MS. The identification of a protein is achieved by a set of experimentally determined peptide masses, which is called a peptide mass fingerprint. The measured masses profile is searched against the theoretical masses generated from an in-silico digestion by applying the same enzyme specificity to all protein sequences selected in the target database. The identified protein sequences are ranked by the number of matched peptide masses with a given mass error tolerance. The mass accuracy of a peak identification (ppm), the relation between assigned and unassigned peaks and the size of the database are used as criteria for evaluation of a successful protein identification.

1.2.4.2 Peptide Fragmentation Fingerprinting (sequence Tagging)

The strategy of peptide fragmentation fingerprinting (PFF) is similar to peptide mass fingerprinting (PMF). With in silico digestion of proteins listed in the database, PFF software creates theoretical fragmentation spectra of all peptides. Instead of presenting the overview of a whole protein, the strategy of peptide fragmentation is to provide primary sequence information of each peptide. A length of eight amino acid is necessary in a sequence tag to avoid ambiguous protein identification. PFF and PMF are complementary and can be used in combination to increase the protein identification accuracy. PFF is preferable in the identification of the constituents of protein mixtures and analysis of post translational modifications (PTMs).

1.2.4.3 Search Algorithms

SEQUEST [73] and MASCOT [74] are two commonly used search algorithms for peptide identification. The peptide amino acid sequence is determined by matching the experimental peptide tandem MS spectra to the in silico created database of peptide spectra. SEQUEST is a type
of heuristic algorithm in which a set of candidate peptides are chosen based on their measured m/z that match closely to theoretical predictions. The best candidates were selected to maximizes the score function [termed cross-correlation (Xcorr)] [75]. MASCOT belongs to the type of probabilistic algorithms and gives each peptide a probability “Ions Score” calculated by $-10\log_{10}(p)$ and $p$ is defined as the probability that “the observed match between experimental data and the database sequence is a random event” [76].

1.2.4.4 De Novo Sequencing

A protein identification using strategies of PMF and PFF naturally requires the sequence of the protein of interest to be in the database. Nevertheless, if the genome of the protein is not sequenced or alternate RNA splicing was occurred, de novo sequencing from the peptide fragmentation spectra is necessary. Automated de novo sequencing can be performed by two approaches. The global one encompasses uses a procedure that generates all theoretically predicted amino acid sequences and the corresponding spectra. The spectrum best matched to the measured data will provide the most probable peptide sequence. The local approach is a graph theory approach that is applied in newer de novo sequencing algorithms. The spectrum graph is generated from measured spectrum and the peaks serve as nodes and two nodes will be connected by an edge if the mass difference between the two nodes corresponds to the mass of an amino acid.

1.2.5 Proteomics Strategy

Protein identification via mass spectrometry is usually carried out in the form of intact-protein analysis (“top-down” proteomics) or the analysis of enzymatically or chemically produced peptides (“bottom-up” proteomics) [77]. In this section, these two common proteomics strategies will be introduced.
1.2.5.1 Bottom Up Strategies

In the bottom up strategy, the detection of peptides is used to provide evidence for existence of a given protein existence, which is a high throughput strategy for highly complex samples. There are two workflows. One is called “sort then break” which means off-line protein fractionation and separation prior to protein digestion are followed by direct peptide determination by PMF or by further LC separation of the digests combined with tandem mass spectrometry [78, 79]. The other is termed “break-then-sort” approaches (commonly referred to as “shotgun proteomics”), in which protein digestion is accomplished without any pre-fractionation/separation. The enzymatic digests are separated by multi-dimensional chromatography followed by tandem mass spectrometry. Ion trap (IT) mass spectrometers are typically used in this approach for its rapid scanning capability and this approach has gained popularity in the proteomics field[80].

1.2.5.2 Top-Down Proteomics

The top-down approach starts with the gas-phase ionization of intact proteins without any prior digestion and followed by high-resolution mass measurement of the intact protein ions and direct ion fragmentation in the mass spectrometer (mainly FTMS) [81]. Sometimes in the bottom up strategy, the identified digests may not be specific to an individual protein and sequence variation may occur resulting the loss of the relationship of one peptide to one another. Theoretically, top-down approach allows for full characterization of proteins including 100% sequence coverage and specific molecular forms (causing from genetic variation, alternative splicing, and post-translational modifications) [82].

However, the physiochemical variety of intact proteins is greater than that of peptides, which cause large-scale separation of intact proteins is complicated and challenging. Furthermore, the
capability to fragment the intact protein is the key element for the MS based top-down approach, which requires for better mass spectrometry instrumentation and fragmentation strategy. The technical difficulty of top-down proteomics makes it lag behind bottom up strategy in the aspects like proteome coverage, sensitivity, and throughput, but top-down approach is beneficial for investigation of single protein or simple mixtures of important biological sample typically gained via affinity purification (AP) [83, 84]. Figure 1.5 summaries the strategies for MS-based protein determination.

Figure 1.5. Strategies for MS-based protein identification and characterization[77]. Note:
Reproduced from “Mass spectrometry for proteomics” by permission of Elsevier.
1.2.6 Quantitation

1.2.6.1 Metabolic labeling

Metabolic labeling introduces stable isotopes (e.g. 13C, 15N, or 18O) during cell growth and division and this technology has been developed for accurate protein quantitation [85]. Metabolic incorporation of heavy isotopes by using 15N-enriched (>96%) cell growth media was described for investigation of protein differences in wild-type versus mutant cells [86]. Stable Isotope Labeling by/with Amino acids in Cell culture (SILAC) is an alternative metabolic labeling technique which has achieved wide popularity after the report by Mann and co-workers [87]. Selected amino acids, usually $^{13}$C$_6$-arginine and $^{13}$C$_6$-lysine were added into the medium to ensure that all tryptic cleavage peptides include at least one labeled amino acid. The workflow is presented in Figure 1.6.

![Figure 1.6. Metabolic stable isotope coding of proteomes][88]. Note: Reproduced from “Stable isotope-coded proteomic mass spectrometry” by permission of Elsevier.
1.2.6.2 Post-biosynthetic labeling

In contrast to metabolic labeling, post-biosynthetic labeling strategy is performed post lysis in vitro by chemical or enzymatic derivatization (Figure 1.7). Enzymatic derivatization, incorporating $^{18}\text{O}$ by using $\text{H}_2^{18}\text{O}$ during the protein digestion, is a specific way to introduce an isotope signature [89]. Chemical derivatization primarily uses the side chain of lysine and cysteine to introduce an

---

**Figure 1.7. Enzyme and Chemical stable isotope-coding of proteomes [88].** Note: Reproduced from “Stable isotope-coded proteomic mass spectrometry” by permission of Elsevier.
isotope-coded mass tag such as in isotope-coded affinity tag (ICAT) which uses chemical labeling reagents to tag the protein [90]. Labeling reagents commonly consist of a reactive group to react with the amino acid side chain, a linker with an isotopic code and a tag (such as biotin) for an affinity isolation step and the initial pioneering work specifically derivatized cysteine residues. Another type of labeling reagent targets the protein N-terminus and the side chain amines of lysine residues with isotope tags for relative and absolute quantification (iTRAQ) [91].

1.2.6.3 Absolute quantification using internal standards

Absolute quantification can be achieved by the addition of a known quantity of a stable isotope-labeled standard peptide to a protein digest and subsequent comparison of the mass spectrometric signal of the standard to the endogenous peptide in the sample.

1.2.6.4 Label-free quantification

Samples for label-free quantification are prepared separately and undergo individual LC-MS/MS or LC/LC-MS/MS analysis. Relative protein quantitation is typically generated from the measurements of peptide signal intensity (peak areas, peak heights) in chromatography or spectral counting of peptide fragment spectra in MS/MS analysis. Early studies showed that peak areas of identified peptides in myoglobin digests correlated linearly to protein concentration. The correlation remained after the myoglobin standard was spiked into complex mixture (such as human serum) and normalization of the calculated peak areas can further improve the results of quantitative profiling.

1.2.7 Application: Proteomics as a Tool for Biomarker Discovery

Mass spectrometry based proteomics plays a crucial role in the discovery of thousands of potential biomarkers for cancer and other diseases. Novel and powerful mass spectrometry related
technologies have been developed to study intricate proteomes. A variety of isotopic labeling techniques such as isobaric tagging e.g., iTRAQ (Ross et al., 2004) and label-free quantitation techniques including spectral counting can be utilized for relative quantitation in these studies. Nassar et al. reported a feasible approach to quantify proteins in complex samples without isotope labeling. Nano-ultra-performance liquid chromatography coupled to a hybrid Q-TOF ion-mobility mass spectrometry (MS) and followed by label-free quantitation is used in a comparative study between cancer and non-cancer samples.

1.3 Application of Proteomic methodologies to the characterization of protein therapeutics

1.3.1 Overview of biologics and biosimilars

The term biologics has almost become the synonym of protein-based therapeutics produced by living organisms. In contrast to the traditional small molecule drug, whose molecular weight typically lower than 1,000 Da, protein therapeutics are large and complex molecules whose molecular weight range is generally from 5 to 200 kDa. The first therapeutic protein created by recombinant DNA technology, was approved by US Food and Drug Administration (FDA) in 1982, is human insulin (Humulin®, created at Genentech). After that, more and more recombinant human proteins appeared as therapeutic agents and used to complement the deficiency of natural proteins (e.g., growth hormone) or augment existing pathways (e.g., interferon-α) [95, 96]. At present, there are around 200 biopharmaceuticals marketed in the United States and/or the European Union [97]. However, the high cost and risk of developing new biologics is a major issue in the control of the healthcare costs. The expiration of numerous patents for blockbuster biologics between 2012 and 2019 gives biosimilars (i.e., follow-on biologics) an improved chance to impact the current biopharmaceuticals market [98, 99]. Biosimilars describes the development of replicas of innovator biologics. Unlike the generic versions of small molecule drugs, biosimilars are
“highly similar” to a branded drug instead of the identical copy of the reference products [100, 101].

1.3.2 Issues or concerns of biologics and biosimilar

The manufacturing process of biologics is complex with production in specially engineered living cells, and the requirement of several stages of cell culture and purification (Figure 1.8). These complicated procedures used during the production and the sensitivity of the biosynthetic process to the fermentation conditions will introduce variability into the final product with the production of a multitude of product variants. The heterogeneity of the products can affect the biological efficacy, clearance, safety and immunogenicity of the final drugs.

Figure 1.8. Sources of variation between manufacturing steps in a recombinant protein production[102]. Note: Reproduced from “The challenge of biosimilars” by permission of Oxford University Press.
To illustrate the complications in quality control of biotechnology products, we will discuss the factors that can result in immunogenicity of the final product in patients [103, 104]. For example, sequence variation of protein drugs due to clonal variation, different host cells such as yeast and insect cell lines can induce the glycosylation of biopharmaceuticals that is divergent from human glycans, host cell contaminants and aggregated species can be associated with immunogenicity and severe clinical consequences [105]. Process-related and product-related product variants that can alter the protein three-dimensional structure are also related to immunogenicity of biopharmaceuticals such as insulin and growth hormone [106].

Understanding of how formulation and process factors affect product quality and activity is important for the quality control of the biologics and biosimilars. Different from a large-scale proteomics study, the biologics characterization focuses on detailed information of a protein including its amino acid (AA) sequence, molecular weight (MW), post-translational modifications (PTMs), and aggregation or degradation products. Mass spectrometry is a key technique for biologics characterization due to its sensitivity, resolution, selectivity and specificity [107].

1.3.3 Analysis of Molecular weight and amino acid sequence

The initial analysis of a therapeutic protein usually includes molecular weight determination and amino acid sequence confirmation [107]. Sequence variation between the innovator and a biosimilar therapeutic can be determined initially by MW measurement. Furthermore, the mutation(s) can be situated by amino acid sequencing [108]. SDS-PAGE analysis can provide approximate molecular weight information. Mass spectrometry such as MALDI TOF and ESI FTMS gives accurate mass information of protein drugs due to their relatively high resolution and mass accuracy. Bottom up and top down methods are two approaches utilized to perform amino acid sequencing of the target protein. In the bottom up strategy the protein sample undergoes
denaturation, reduction, alkylation, enzyme digestion and then the digests are analyzed by a LC-tandem MS system. This approach is also referred to as peptide mapping. Conversely the top down approach comprises much less sample handling and the protein is sequenced directly by tandem MS or multiple-stage MS. However, the top down approach has difficulty in sequencing a large protein and it has been reported that complete sequence coverage can be obtained for proteins less than 5 kDa by using CID and the range of protein size can be increase to 20 kDa by utilizing ETD fragmentation.

1.3.4 Analysis of Protein Posttranslational Modifications

1.3.4.1 Overview of Protein Post Translational Modifications (PTM)

Post translational modifications (PTMs) are chemical alterations that occur on a protein during the biosynthetic process. There are more than 300 known PTMs that happen physiologically and new types of PTMs are discovered regularly. Mass spectrometry (MS) plays a pivotal role in the protein PTM investigation studies, such as site characterization and quantification of chemical modifications and detection of new types of structures. Understanding of the solution and gas-phase reactivity of PTMs is important to the successful analysis of the protein PTMs by MS [109].

This discussion section is mainly restricted to the application of mass spectrometry to characterize protein PEGylation (non-enzymatic, designer PTM), glycosylation (enzymatic), disulfide bond linkage (enzymatic) or two stability related PTMs, namely deamidation and oxidation (non-enzymatic).

1.3.4.2 PEGylation (Chemical modification)

*Overview*
PEGylation describes the process of attaching polyethylene glycol (PEG) polymer chains to molecules such as therapeutic protein or peptides. This modification will alter the physiochemical properties of the therapeutics. The linkage of PEG to a therapeutic protein will add to a hydrophobic moiety a hydrophilic property and the attachment of PEG will also increase hydrodynamic size, which prolongs the drug circulating half-life in the body. Compared with its unPEGylated counterpart, PEGylated therapeutics exhibit decrease immunogenicity and the large PEG polymer chain can mask an antigenic region from the immune system of the host [110, 111].

The early PEGylated therapeutic proteins were synthesized via a non-specific reaction involving the side-chain of lysines and/or the protein N-terminus. However, this method leads to a heterogeneous mixture of PEGylated products and each conjugate has individual activity and stability properties. Site-Specific PEGylation creates a homogeneous product and decreases the influence of PEG on biological activity of the protein [112]. The typical reactive amino acids of a protein for its PEG derivative include lysine, cysteine, histidine, arginine, aspartic acid, glutamic acid, serine, threonine, tyrosine, N-terminal amino group and C-terminal carboxylic acid [112, 113] and the functional groups involved in the reaction including thiol, alpha amino group, epsilon amino group and carboxylate hydroxylate[114]. However, a thiol group rarely exists in proteins, furthermore it is often related to active sites and also the carboxylic group is not easily activated. Thus, amine-terminus or epsilon-amino groups of lysines are the typical site for PEG linkage due to their highly reactive activity and the mild reaction conditions [111, 115]. The reactive functional group of the synthesized PEG such as aldehyde is largely selective for the α-amine of protein N-terminus because of the relatively low pKa compared with the α-amine of lysine side chain [116-119]. Site-specific conjugates like N-terminally PEGylated human recombinant granulocyte
colony-stimulating factor are commercially available [112]. The PEGylated biopharmaceuticals on the market are listed in the Table 1.3.

**Analysis of PEGylated therapeutics**

Development of state of art analytical methods to characterize the PEG-protein conjugates contributes to the selection of an efficient PEGylation process to produce a safe PEGylated therapeutic[113]. However, characterization of proteins and peptides with polyethylene glycol (PEG) modification is a challenge, due to the large mass of the PEG molecule (such as the 20 kDa PEG polymer) and the heterogeneity of the PEG (polydispersity). Colorimetric titration of the non-conjugated amino groups of the proteins or peptides provides the degree of the modification but the method cannot precisely evaluate the number of polymer chains attached to a peptides or protein. The colorimetric reaction of PEG with iodine gives direct evaluation of the PEGylated product, however, it has relatively low-sensitivity with high blank values. MALDI-TOF MS offers a determination of the molecular weight which provides a picture of the mixtures present in the conjugation reaction, since it typically generates a single charge molecular ion and possesses a high mass range, but it is not quantitative because different weight compounds show different extraction yield [120-122]. Conventional ESI-MS analysis of PEGylated products leads to an overlap of differently charged ions in the mass spectra because PEG is normally heavily charged and its charge distribution is always complicated by the polydispersity of the species [123, 124]. For a high mass resolution spectrometer like an Orbitrap, the PEG heterogeneity makes the charge state pattern of the analyzed product outside the m/z range for the mass analyzer. It has also been reported that the large size PEG chain may prevent the PEGylated peptides from traversing the C-trap in an Orbitrap mass analyzer [125]. The addition of amines (such as triethylamine) is beneficial for reducing the charge of PEGylated proteins and make the charge state pattern of the
product shift into the m/z range of an Orbitrap mass analyzer, to enable intact mass analysis of a PEGylated protein [126, 127].

Table 1.3. PEGylated biopharmaceuticals on the market

<table>
<thead>
<tr>
<th>Brand</th>
<th>PEG Conjugate</th>
<th>PEG Chain</th>
<th>Company</th>
<th>Year</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adagen</td>
<td>PEG-adenosine deaminase</td>
<td>Linear, 1*5 KDa</td>
<td>Enzon</td>
<td>1990</td>
<td>severe combined immunodeficiency disease (SCID)</td>
</tr>
<tr>
<td>Oncaspar</td>
<td>PEGylated L-asparaginase</td>
<td>Linear, 1*5 KDa</td>
<td>Enzon</td>
<td>1994</td>
<td>acute lymphoblastic leukemia</td>
</tr>
<tr>
<td>PegIntron</td>
<td>PEGylated interferon alpha 2b</td>
<td>Linear, 1*12 KDa</td>
<td>Schering-Plough</td>
<td>2001</td>
<td>chronic hepatitis C and hepatitis B</td>
</tr>
<tr>
<td>Pegasys</td>
<td>PEGylated interferon alpha 2a</td>
<td>Branched, 1*40 KDa</td>
<td>Hoffmann-La Roche</td>
<td>2002</td>
<td>chronic hepatitis C and hepatitis B</td>
</tr>
<tr>
<td>Somavert</td>
<td>Pegvisomant*</td>
<td>Linear, 4-6*5 KDa</td>
<td>Pfizer</td>
<td>2003</td>
<td>acromegaly</td>
</tr>
<tr>
<td>Neulasta</td>
<td>PEGylated GCSF*</td>
<td>Linear, 1*20 KDa</td>
<td>Amgen</td>
<td>2002</td>
<td>severe cancer chemotherapy-induced neutropenia</td>
</tr>
<tr>
<td>Macugen</td>
<td>PEG aptanib sodium</td>
<td>Branched, 1*40 KDa</td>
<td>Pfizer</td>
<td>2004</td>
<td>neovascular age-related macular degeneration</td>
</tr>
<tr>
<td>Mircera</td>
<td>PEG epoetin beta</td>
<td>Linear, 1*30 KDa</td>
<td>Roche</td>
<td>2007</td>
<td>anemia associated with chronic kidney disease</td>
</tr>
<tr>
<td>Cinzio</td>
<td>certolizumab pegol*</td>
<td>Branched, 1*40 KDa</td>
<td>UCB Pharma</td>
<td>2008</td>
<td>severe rheumatoid arthritis and Crohn's disease</td>
</tr>
<tr>
<td>Krystexxa</td>
<td>PEGylated uricase</td>
<td>Linear, ~9*10 KDa (each monomer)</td>
<td>Savient Pharmaceuticals</td>
<td>2010</td>
<td>chronic gout in adult patients refractory to conventional therapy</td>
</tr>
<tr>
<td>Omontys</td>
<td>PEGylated synthetic peptide (Peginesatide)</td>
<td>Branched, 1*40 KDa</td>
<td>Affymax</td>
<td>2012</td>
<td>anemia due to chronic kidney disease (CKD) in adult patients on dialysis</td>
</tr>
<tr>
<td>Pleggridy</td>
<td>PEGylated Interferon Beta 1a</td>
<td>Linear, 1*20 KDa</td>
<td>BioGen</td>
<td>2014</td>
<td>patients with relapsing forms of multiple sclerosis</td>
</tr>
<tr>
<td>Adynovate</td>
<td>PEGylated Antithrombotic Factor VIII</td>
<td>1 or more 20 KDa</td>
<td>Baxalta US</td>
<td>2015</td>
<td>hemophilia A (congenital Factor VIII deficiency)</td>
</tr>
</tbody>
</table>

Note: GCSF* is short for recombinant methionyl human granulocyte colony-stimulating factor. Pegvisomant* is PEGylted human growth hormone receptor antagonist. certolizumab pegol* is PEGylation of the fragment of a monoclonal antibody specific to tumor necrosis factor alpha (TNF-α)
1.3.4.3 Glycosylation

Glycosylation describes the process that a carbohydrate is covalently attached to a macromolecule such as proteins. This modification plays important role in protein properties relevant to their therapeutic application and mediates correct protein folding [128, 129]. N-linked and O-linked glycans are two common types of glycosylation. N-linked glycosylation which describes when the glycan is attached to an Asn residue must fit into the sequon Asn-X-Ser/Thr (X cannot be Pro). O-linked glycosylation means the glycan is attached to a Ser or Thr residue.

There is a wide range of branching N-linked glycans in human cells. They all share a trimannosyl core (Man3GlcNAc2) and are categorized as high mannose, complex and hybrid types [130]. High mannose contains five to nine mannose residues linked to the trimannosyl core while the complex type has branching antennae, which often consists of polylactosamine chains and this type can contain a range of the other types of saccharides. The hybrid type includes antennae of both high mannose and complex types.

In contrast to the amino acid sequence of protein therapeutics, a glycosylation pattern is determined by the host cell line instead of the inserted gene [131]. Consequently, the glycosylation variation from host to host or even from clone to clone because of different fermentation procedure, specific productivity and the physiological stage of the host cell is the major issue during the biologics manufacture process. The glycans of recombinant glycoproteins produced by various type of host cells such as bacteria, yeasts, plants, and insects are different compared to that produced by mammalian cells. The glycan moieties chiefly discovered in plants, insects and parasites, which are made up of xylose and core-3-linked fucose are absent in mammals and can cause human immunogenic response.
Oligosaccharide profiling and glycopeptide analysis are two common approaches in the analytical studies of carbohydrates. [132]. In oligosaccharide profiling, the glycans are released by chemical or enzymatic techniques and analyzed by mass spectrometry. [133, 134]. This procedure provides glycan structural information and simplifies the mass spectra by removing the protein moiety from the glycan. In contrast, glycopeptide analysis gives both the glycan structure and site attachment information. In the process, the glycoprotein is enzymatically digested and the resulting glycopeptides are then analyzed by LC-tandem MS analysis.

1.3.4.4 Disulfide bond linkage analysis

Disulfide bond linkages play a pivotal role in stabilization of the three-dimensional structure of proteins. Mislinkages of cysteines can prevent a protein refolding to its native structure. Traditional method such as NMR or Edman degradation are widely used to characterize the disulfide connections, however, these methods require high concentration of a homogeneous protein or well purified digested peptides [135]. MS is a powerful tool in the disulfide bond linkage analysis. Localization of disulfide linkages in a protein involves alkylation of free sulfhydryls, enzyme digestion of the protein and LC tandem MS analysis of the digests [136]. However, sometimes, the digestion conditions (such as pH 8.0) will trigger disulfide bond rearrangement by a thiol–disulfide exchange reaction. Pepsin is a favorable enzyme to generate disulfide linked peptides due to its ideal pH optimum of an acidic pH to minimize the possibility of disulfide bond scrambling [137]. The linkage site can be located by the CID spectra and the sequence of peptides containing the disulfide bond can be provided by the ETD spectra because this fragment method can break the disulfide bond.
1.3.4.5 Stability related modification

1.3.4.5.1 Deamidation

Deamidation is a common degradation production product which can occur at an asparagine (Asn, N) or a glutamine (Gln, Q) residue. Based on the deamidation mechanisms, asparagine is converted to an isoaspartyl (isoAsp) or n-aspartyl (Asp) residue via a succinimide intermediate (Figure 1.9) and glutamine deamidates to γ-glutamyl (γ-Glu) and α-glutamyl (α-Glu) residue via a glutaramide intermediate or deamidates by directly via hydrolysis to α-Glu [138, 139]. Gln undergoes deamidation at a much slower rate due to an extra methylene group in the side chain which causes the formation of a much less favorable intermediate of a 6-membered glutarimide ring instead of the 5-membered succinimide ring [140].

Deamidation can happen spontaneously not only in vivo affecting protein structure and function which is related to human degenerative diseases and natural aging, but also occurring in therapeutic proteins biosynthesis and/or the storage process which is associated with altering the products purity, stability, bioactivity, and antigenicity [141, 142]. Mass spectrometry has already become a powerful tool for monitoring the deamidation products. High-resolution mass spectrometry, which can detect a mass increase of 0.984, is used for the top-down approach for identification of the deamidation reaction. Collision-induced dissociation (CID), which is used to fragment the peptide, is the method of choice to characterize deamidation sites by generating specific fragments. Electron transfer dissociation (ETD) can be used to identify Asp and isoAsp structures by generating diagnostic c and z ions (c+57 and z-57).
Figure 1.9. Deamidation of asparaginyl residues and isomerization of aspartyl residues via a succinimide intermediate [143]. Note: Reproduced from “Recent advances in mass spectrometric analysis of protein deamidation” by permission of John Wiley and Sons.

It had been reported that five deamidation sites on ribonuclease A (13.7 kDa) were successful identified by Zabrouskov and co-workers. Furthermore, the isoforms were differentiated by employment of isoelectric focusing and top-down mass spectrometry approach on a 6T FTMS [144, 145]. Li et al. reported that using top-down MS successfully differentiate Asp to isoAsp using top-down MS [139]. Yu et al. introduced an integrated online–offline top-down ECD mass spectrometry platform which is able to detect low levels of isoAsp degradation products in therapeutic polypeptides and proteins [146].

1.3.4.5.2 Oxidation

Protein oxidation is a reactive oxygen induced covalent modification which can occur on several amino acids (such as methionine (Met), cysteine (Cys), tryptophan (Trp), lysine (Lys)) by
different mechanisms [147]. Methionine is the residue that is most sensitive to oxidation and is a primary concern in the analysis of oxidation reactions. The oxidation of Met has been discovered in a wide variety of proteins and can lead to the result that biological activity decreases and with an increase in protein aggregation or immunogenicity. There are many assays used to detect protein oxidation, such as amino acid analysis, capillary electrophoresis, HPLC/UV or MS analysis and fluorescence assays [148]. Some of these methods are routinely used for structure characterization and others for quality control.

1.4 Conclusion and perspective

This chapter introduces the commonly used mass spectrometer and ionization technologies used in the analysis of large biomolecules, the current methodologies used in proteomics research and the techniques used in protein therapeutic characterization. Chapter 2 used 1D electrophoresis combined with mass spectroscopy to determine the glycan structures of glycoproteins of a model system for the study of polycystic kidney disease (Pkd1+/+ and Pkd1-/- cells). Chapter 3 describes the development of a method for the characterization a novel viral antigen which is a complex glycoprotein. In Chapter 4, with the goal of the location of the site of modification of a PEGylated peptide, a multiple enzyme digestion strategy, followed by UPLC-tandem mass spectrometry was applied to the analysis of model PEGylated peptides.

1.5 References


70. Loo, J., et al., Tandem mass spectrometry and MS/MS/MS of multiply charged ions from large biomolecules. 1991, Pacific Northwest Lab., Richland, WA (United States).


Chapter 2: UPLC-MS/MS analysis of glycoprotein changes in Mouse Kidney cells with the phenotype of autosomal dominant polycystic kidney disease

*Some of the raw data used in this study were created by Fan (Anna) Zhang and analyzed by Yanjun Liu

Manuscript in preparation for submission to Journal of Proteome Research

Title: UPLC-MS/MS analysis of glycoprotein changes in Mouse Kidney cells with the phenotype of autosomal dominant polycystic kidney disease

Author: Yan-Jun Liu, Shiaw-Lin Wu, Fan (Anna) Zhang, Jordan A. Kreidberg and William S. Hancock*
2.1 Abstract

Polycystic Kidney Disease is a genetic disease that can occur in predominantly in adults (autosomal dominant, ADPKD) and in children. In a variant of the disease known as autosomal recessive the cysts grow in both the kidney and liver. ADPKD is one of the most common causes of kidney failure and at present has no known treatment. In this disease, numerous cysts filled with fluid grow on the kidney and approximately 50% of the cases lead to kidney failure. Our research collaboration with Boston Children’s hospital focused on exploring the molecular basis for ADPKD and using mass spectroscopy to analyze differences in protein glycosylation in a mouse model which focuses on polycystins which have been related to cystogenesis (wild type (WT) and pkd1-/- cells). In our previous study, some biomolecular changes have been observed in this disease, such as abnormal glycosylation of integrin α3β1 which appears to be a causal factor in PKD. Glycoproteins participate in many key biological processes such as cell adhesion, receptor activation, and signal transduction. Determination of ADPKD related glycoprotein differences can reveal significant information for the study of disease mechanisms, which is essential to improve diagnosis and treatment strategies. In this study, ultra-performance liquid chromatography coupled with tandem mass spectrometry (UPLC-MS/MS) was used to recognize ADPKD related changes in glycoproteins. Lectin Sambucus nigra bark (SNA) was used to enrich glycoproteins which have the changes in sialic acid content in WT and pkd1-/- cells. Subsequently, 1D SDS PAGE was performed followed by in-gel digestion and UPLC-MS/MS analysis. As an example of our results, clinically interesting glycoproteins, such as Thrombospondin-1 (TSP-1) were only detected in the disease sample and with changes in glycan composition.
2.2 Introduction

Autosomal Dominant Polycystic Kidney Disease (ADPKD) is the most frequent inherited kidney disorder and cause of End Stage Renal Disease (ESRD)[1] and requires dialysis treatment or renal transplant [2-4]. The primary characteristics of ADPKD is cyst formation and enlargement in the kidney and other organs, such as liver, pancreas and spleen) [5]. ADPKD has two predisposing loci, PKD1 and PKD2 which encode polycystin-1 (PC1) and polycystin-2 (PC2) respectively. Polycystins and associated proteins are related to cystogenesis[6, 7]. Mutation occurs in either or both principal genes, PKD1 and PKD2, can cause ADPKD and, in clinically identified populations, PKD1 mutations were responsible for around 85% of the cases [8-10]. It has been reported that cysts start to develop at the embryonic stage in the Pkd1 homozygous knockout mice. However, the results in Pkd1 heterozygous cells shows that a germ line mutation is insufficient to cause the cystic phenotype in the heterozygous mouse [11].

Although the precise molecular mechanisms of cyst formation are still unclear, a lot of publications reported that integrins, heterodimeric transmembrane receptors, are involved in kidney development and disease [12]. Our previous study characterized abnormal glycosylation of integrin α3 in Pkd1 homozygous knockout mice cells [13].

Glycosylation is a diverse and highly regulated process in which cellular glycans are attached to proteins and lipids primarily in the Golgi body. Glycoproteins participate in many key biological processes such as cell adhesion, receptor activation, and signal transduction. Glycosylation not only governs physiology but also contributes to disease [14, 15]. It has been reported that integrin α3β1 is retained in the Golgi apparatus in the Pkd1-/- model, and this aberration is thought to be related to abnormal glycosylation of the integrin α3 subunit. An E3 ubiquitin ligase, casitas B-lineage lymphoma (c-Cbl), whose function is to ubiquitinate tyrosine-protein kinase Met (c-Met)
was also found to be sequestered in the Golgi apparatus with integrin α3β1, with the result that c-Met cannot be ubiquitinated. Since c-Met is an upstream regulator of PI3K/AKT/mTOR signal pathway this lack of regulation of c-Met results in an increased activation of the c-Met dependent PI3K/AKT/mTOR pathway. Increase of the activity of mTOR signaling pathway has been suggested to contribute to cyst formation in mice and in humans. Inhibition of this signal pathway reverses renal cystogenesis in PKD [16-19].

Further work was done to study the cause of the abnormal glycosylation of integrin α3 subunit in the Pkd1-/- cells. The abnormal integrin α3 was immunoprecipitated from the Pkd1-/- cells and exhibited anomalous mobility in SDS-PAGE electrophoresis compared with that from the Pkd1+/+ cells. A western blot showed that the integrin α3 protein band from the mutant cells (Pkd1-/-) had a faster migration compared with the band from the wild type cells (Pkd1+/+). After treating the samples with PNGase F, the molecular weight difference was removed and indicated that N-linked glycosylation was abnormal in the Pkd1-/- model. Mass spectroscopy was utilized to analyze differences in integrin α3 glycosylation between Pkd1+/+ and Pkd1-/- cells. In the mutant Pkd1-/- cells, it was found that lower molecular weight glycans with different monosaccharide compositions were present at site Asn-925 and Asn-928 in the integrin α3 subunit, as well as an unusual disialic acid glycan which were observed solely in Pkd1-/- cells [13]. An increase in sialic acid density in glycoproteins is associated with cell surface changes and altered growth patterns that is quite possibly related to induction of malignant cells. Turner and co-workers reported that increase of sialic acid in serum of patient with cancer is correlated with an acute phase response [20, 21].

In order to determine that whether there are specific and global differences in protein glycosylation in PKD and provide potential glycoprotein targets for further glycan structure
analysis, our study used 1D electrophoresis combined with mass spectroscopy to determine the glycoproteins present in the Pkd1+/+ and Pkd1-/- cells with sialic acid differences. Lectin Sambucus Nigra Lectin (SNA) was chosen to pull down the glycoprotein with α2–6 linked sialic acid. Then SDS PAGE was used as the first dimension of separation, followed by in-gel digestion and LC tandem MS to identify the targeted glycoproteins. After bioinformatic analysis three glycoproteins were selected as candidates for future study.

2.3 Material and methods

2.3.1 Cell lines

Pkd1+/+ and Pkd1-/- were described previously [22]. Mouse kidney epithelial cells (Pkd1 -/-) were isolated from embryonic kidneys, day 15.5 from a cross of Pkd1 -/- mice that also carry a temperature-sensitive simian virus 40 (SV40) large T-antigen transgene. Thus, the resulting cell lines were conditionally immortalized, and the expression of the SV40 large T-antigen was regulated by temperature and IFN-γ [11].

2.3.2 Reagents

Trypsin (Sequencing grade) was purchased from Promega (Madison, WI). Ammonium bicarbonate, formic acid, dithiothreitol (DTT) and iodoacetamide (IAA) were purchased from Sigma-Aldrich (St. Louis, MO). LC-MS grade water and acetonitrile were purchased from Fisher Scientific (Fairlawn, NJ). NuPAGE® 4-12% Bis-Tris gels, Novex® sharp pertained protein standard, SimplyBlueTM safe stain were purchased from InvitrogenTM (Carlsbad, CA).

2.3.3 Sambucus Nigra Lectin (SNA) pull down
Immunoprecipitation were performed as previous description [13]. Confluent cells were washed with PBS, lysed with lysis buffer (20 mM Tris/HCl, 1 mM EDTA, 150 mM NaCl, 1% Triton X-100), which contains proteinase inhibitor cocktail tablet (Roche 1697498), at 4 °C for 30 minutes. After centrifugation at 15 700g for 15 min, supernatants were incubated with lectin SNA at 4 °C for 4 h.

2.3.4 Gel electrophoresis (SDS-PAGE) analysis

1 µL DTT and 4 µL loading buffer (4X) were added to 10 µL protein solution. The mixture was heated at 70 °C for 10 min before gel electrophoresis. The samples were loaded on a SDS-PAGE gels (4-12% Bis-Tris gel) for separation by molecular weight. The constant voltage of 160 V is applied for 60 min to run the gel. After the gel electrophoresis, remove the gel from the cassette and wash the gel with the Deionized water. Then, stain the gel for 1 hour at room temperature and de-stain the gel overnight using Deionized water at the same condition.

2.3.5 In gel digestion

Each gel lane was cut into five pieces and then chopped into small 1mm3 cubes. Briefly, the small gel pieces were destained by adding 200 µL 50% acetonitrile (ACN) in water. After removing the Coomassie blue stain, the gel pieces were reduced by the addition of 10 mM DTT dissolved in 0.1 M ammonium bicarbonate and incubated at 56°C for 30 minutes. Then, the gel pieces were alkylated with a fresh 10 mg/ml iodoacetamide (IAA) in 0.1 M ammonium bicarbonate in dark for one hour to avoid exposure to light. Trypsin digestion buffer (50 mM NH₄HCO₃, 5 mM CaCl₂, and 12.5 ng/µL sequencing grade Trypsin) were added into each sample to cover the small chopped gel pieces at 4 for 30-35 min and followed by further incubation overnight at 37 °C.
2.3.6 NanoLC tandem analysis

All the digested samples were loaded on a self-packed 75 μm i.d. x 15 cm C18 column (Magic C18 beads, 200Å pore and 5 μm particle size, Michrom BioResources, Auburn, CA). The peptides or glycopeptides analysis is performed on an Ultimate 3000 nano LC pump (Dionex, Mountain View, CA) on-line coupled to a LTQ-Orbitrap-XL mass spectrometer (Thermo Fisher Scientific, San Jose, CA) interfaced with a nano spray ion source (New Objective, Woburn, MA). Mobile phase A (0.1% formic acid in water) and mobile phase B (0.1% formic acid in acetonitrile) were used for the gradient consisting of linear from 2% to 40% B in 60 min, then linear from 40% to 80% B for 10 min and finally isocratic at 80% B for 5 min. The mass spectrometry was operated in the data-dependent mode by alternatively switching between MS survey with the scan mass range 300-2000 and CID tandem MS fragmentation on the ten most high intensity precursor ions. The parameters used to operate the mass spectrometers are: mass resolution is 30,000 for 400 m/z, precursor ion isolation width is ± 1 m/z, 35% normalized collision energy, 2.3 kV spray voltage, capillary temperature of 270 ℃ and dynamic exclusion was 30 s for 2 repeat counts.

2.3.7 Data analysis

The spectra generated in the CID-MS2 step were searched against SEQUEST algorithm integrated in the software Proteomic Discovers (version 1.4 Thermo Fisher Scientific), the Xcorr score was set as +1 > 1.5; +2 > 2.0; +3> 2.5 for the initial search. In addition, manual inspections were performed to exclude the false positive or false negative resulted induced by the software.
2.4 Results

2.4.1 A comparison between the proteomes of wild type and mutant cell lines.

1D-gel electrophoresis was employed as the first dimensional separation and using the image of the gel stain we minimized interference of the high abundance proteins in the subsequent LC tandem MS analysis of the tryptic digest.

![Venn diagram](image)

Figure 2.1. The number of proteins and glycoproteins identified in wide type or mutant cell (KO) lysates. Part A. Number of proteins identified in WT or KO lysate, B. Number of glycoproteins identified in wide type or mutant cell lysate. WT indicates wild type cells and KO describes the mutant cell model (Pkd1 gene knock out).

The experimental spectra were searched against SEQUEST (V1.4) integrated with mouse database (Mus musculus). In the wild type mouse kidney epithelial cells (Pkd1+/+), 1221 proteins were identified, while in the mutant cell line (Pkd1-/-), 1269 proteins were identified and 797 common proteins were shared by both cell lines. After removal of proteins which are unannotated in UniProt, 650 common proteins were shared by both cell lines and 321 and 318 unique proteins identified in the mutant and wild type cell lines respectively. We further explored predicted glycoprotein differences between cell lysates prepared from the two cell lines with the following result; 74 glycoproteins were common; 55 and 40 unique glycoproteins were identified in the
mutant and wild type cell lines respectively (results summarized in the Figure 2.1). In addition, 416 proteins were upregulated and 278 proteins are downregulated in Pkd1-/- cells compared with the proteins in Pkd1+/- cells. The upregulated glycoproteins in Pkd1-/- cells were listed in the Table 2.1

### Table 2.1. Upregulated glycoproteins in Pkd1-/- cells

<table>
<thead>
<tr>
<th>Uniprot Accession</th>
<th>Ratio</th>
<th>Protein names</th>
</tr>
</thead>
<tbody>
<tr>
<td>P20152</td>
<td>0/38</td>
<td>Vimentin</td>
</tr>
<tr>
<td>A2ARV4</td>
<td>0/23</td>
<td>Low-density lipoprotein receptor-related protein 2</td>
</tr>
<tr>
<td>Q91ZB0</td>
<td>0/19</td>
<td>Cadherin EGF LAG seven-pass G-type receptor 3</td>
</tr>
<tr>
<td>P70206</td>
<td>0/17</td>
<td>Plexin-A1</td>
</tr>
<tr>
<td>O88799</td>
<td>0/16</td>
<td>Zonadhesin</td>
</tr>
<tr>
<td>Q9WU62</td>
<td>0/16</td>
<td>Inner centromere protein</td>
</tr>
<tr>
<td>Q9JVF9</td>
<td>0/15</td>
<td>Sodium channel protein type 5 subunit alpha</td>
</tr>
<tr>
<td>P02468</td>
<td>0/15</td>
<td>Laminin subunit gamma-1</td>
</tr>
<tr>
<td>Q66K08</td>
<td>0/15</td>
<td>Cartilage intermediate layer protein 1</td>
</tr>
<tr>
<td>Q3TRM4</td>
<td>0/15</td>
<td>Neuropathy target esterase</td>
</tr>
<tr>
<td>G3X9E5</td>
<td>0/14</td>
<td>Atrial natriuretic peptide receptor 3</td>
</tr>
<tr>
<td>F8VQN5</td>
<td>0/14</td>
<td>Integrin alpha-1</td>
</tr>
<tr>
<td>P11276</td>
<td>0/13</td>
<td>Fibronectin</td>
</tr>
<tr>
<td>P84309</td>
<td>0/13</td>
<td>Adenylate cyclase type 5</td>
</tr>
<tr>
<td>O35797</td>
<td>0/13</td>
<td>Multidrug resistance-associated protein 1</td>
</tr>
<tr>
<td>Q8B8S4</td>
<td>0/12</td>
<td>Melanoma inhibitory activity protein 3</td>
</tr>
<tr>
<td>P08122</td>
<td>0/12</td>
<td>Collagen alpha-2(IV) chain [Cleaved into: Canstatin]</td>
</tr>
<tr>
<td>Q80WJ6</td>
<td>0/12</td>
<td>Multidrug resistance-associated protein 9</td>
</tr>
<tr>
<td>O89001</td>
<td>0/11</td>
<td>Carboxypeptidase D</td>
</tr>
<tr>
<td>Q9WTL4</td>
<td>0/11</td>
<td>Insulin receptor-related protein</td>
</tr>
<tr>
<td>Q8R2Y2</td>
<td>0/10</td>
<td>Cell surface glycoprotein MUC18</td>
</tr>
<tr>
<td>Q8BKG3</td>
<td>0/10</td>
<td>Inactive tyrosine-protein kinase 7</td>
</tr>
<tr>
<td>Q91ZK7</td>
<td>0/10</td>
<td>Prolow-density lipoprotein receptor-related protein 1</td>
</tr>
<tr>
<td>O88572</td>
<td>0/10</td>
<td>Low-density lipoprotein receptor-related protein 6</td>
</tr>
<tr>
<td>P55906</td>
<td>0/9</td>
<td>ATP-binding cassette sub-family D member 3</td>
</tr>
<tr>
<td>Q9CQF9</td>
<td>0/9</td>
<td>Prenylcysteine oxidase</td>
</tr>
<tr>
<td>Q9EPR5</td>
<td>0/9</td>
<td>VPS10 domain-containing receptor SorCS2</td>
</tr>
<tr>
<td>Q8CFX1</td>
<td>0/9</td>
<td>GDH/6PGL endoplasmic bifunctional protein</td>
</tr>
<tr>
<td>P55302</td>
<td>0/9</td>
<td>Alpha-2-macroglobulin receptor-associated protein</td>
</tr>
<tr>
<td>Q0VF58</td>
<td>0/9</td>
<td>Collagen alpha-1(XIX) chain</td>
</tr>
<tr>
<td>Q9R069</td>
<td>0/8</td>
<td>Basal cell adhesion molecule</td>
</tr>
<tr>
<td>Q9WVL3</td>
<td>0/8</td>
<td>Solute carrier family 12 member 7</td>
</tr>
<tr>
<td>P46978</td>
<td>0/7</td>
<td>Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit STT3A</td>
</tr>
<tr>
<td>Q9HJS8</td>
<td>0/7</td>
<td>Solute carrier family 12 member 4</td>
</tr>
<tr>
<td>Q9Z226</td>
<td>0/7</td>
<td>Contactin-4</td>
</tr>
<tr>
<td>P35441</td>
<td>0/7</td>
<td>Thrombospondin-1</td>
</tr>
<tr>
<td>O88704</td>
<td>0/7</td>
<td>Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 1</td>
</tr>
<tr>
<td>Q8BM96</td>
<td>0/6</td>
<td>Adhesion G-protein coupled receptor G7</td>
</tr>
<tr>
<td>P19070</td>
<td>0/6</td>
<td>Complement receptor type 2</td>
</tr>
<tr>
<td>Q3TDQ1</td>
<td>0/6</td>
<td>Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit STT3B</td>
</tr>
<tr>
<td>P14483</td>
<td>0/6</td>
<td>H-2 class II histocompatibility antigen, A beta chain</td>
</tr>
</tbody>
</table>
2.4.2 SDS-PAGE analysis of lectin affinity pull-down samples and LC/MS analysis

![Figure 2.2. gel image for proteins by SNA pull down](image)

The carbohydrate binding features of plant lectins were commonly used to detect, isolate and characterize glycoconjugates. Sambucus Nigra Lectin (SNA) binds specifically to terminal
Neu5Ac (α2–6) Gal/GalNAc of glycan conjugates [23]. In our study, the SNA purified cell lysates of wide type and mutant cell lines were obtained using previously published procedures [13]. SDS-PAGE analysis was utilized as the first-dimension separation technology to monitor and separate proteins in the sample. The gel image in Figure 2.2 shows the fractions purified from the two cell lines with the SNA lectin. The gel was cut into five pieces based on the band intensity, followed by in-gel trypsin digestion and LC tandem MS analysis.

Table 2.2. Glycoproteins identified by LC/MS analysis of the SNA pull-down

<table>
<thead>
<tr>
<th>Uniprot Accession</th>
<th>Protein Name</th>
<th># Unique Peptides&lt;sup&gt;a&lt;/sup&gt;</th>
<th>WT</th>
<th>KO</th>
</tr>
</thead>
<tbody>
<tr>
<td>P11438</td>
<td>Lysosome-associated membrane glycoprotein 1</td>
<td>2</td>
<td>ND</td>
<td></td>
</tr>
<tr>
<td>P29533</td>
<td>Vascular cell adhesion protein 1</td>
<td>6</td>
<td>ND</td>
<td></td>
</tr>
<tr>
<td>Q91ZX7</td>
<td>Prolow-density lipoprotein receptor-related protein 1</td>
<td>2</td>
<td>ND</td>
<td></td>
</tr>
<tr>
<td>Q07797</td>
<td>Galectin-3-binding protein</td>
<td>10</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P17047</td>
<td>Lysosome-associated membrane glycoprotein 2</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P17439</td>
<td>Glucosylceramidase</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P11679</td>
<td>Keratin, type II cytoskeletal 8</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>P09055</td>
<td>Integrin beta-1</td>
<td>4</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Q62470</td>
<td>Integrin alpha-3</td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Q02248</td>
<td>Catenin beta-1</td>
<td>ND</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>P16406</td>
<td>Glutamyl aminopeptidase</td>
<td>ND</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>P35441</td>
<td>Thrombospondin-1</td>
<td>ND</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Q8R2Y2</td>
<td>Cell surface glycoprotein MUC18</td>
<td>ND</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P01897</td>
<td>H-2 class I histocompatibility antigen, L-D alpha chain</td>
<td>ND</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Glycoproteins were identified using the criteria described in Data Analysis (see Material and Methods)

<sup>b</sup> In the analysis, proteins that were not glycosylated were also identified and the presence of these proteins may be caused by nonspecific binding or interaction with the captured glycoproteins.
The glycoproteins identified after the SNA lectin binding are listed in the Table 2.2 and six glycoproteins displayed significant abundance increases in the Pkd1-/- cells. Integrin alpha-3 and catenin beta-1 were identified with 6 unique peptides while glutamyl aminopeptidase and thrombospondin-1 were identified with 5 unique peptides. Cell surface glycoprotein MUC18 and H-2 class I histocompatibility antigen, L-D alpha chain were identified with 1 and 3 unique peptides. In the proteomics study of the cell lysate, we did not find a significant difference in amount of integrin alpha-3 and catenin beta-1 between Pkd1-/- cells and Pkd1+/+ cells. However, after the SNA binding purification, we only detected catenin beta-1 in the Pkd1-/- cells and the number of unique peptides detected for integrin alpha-3 increased to six in Pkd1-/- cells. These results were consistent with our previous study which had characterized a novel disialic acid motif present in integrin alpha-3 isolated from the Pkd1-/- cells and also illustrates the effectiveness of the SNA affinity step to concentrate sialic acid containing glycoproteins. As examples of proteins of interest that will included in the Discussion section, a representative extracted ion chromatogram and mass spectrometry spectra of an unique peptide for catenin beta-1, glutamyl aminopeptidase and thrombospondin-1 are presented in the Figure 2.3 and Figure 2.4 respectively.
Figure 2.3. Representative LC/MS spectra of a unique peptide from glutamyl aminopeptidase A. Extracted ion chromatogram of peptide ETNLLYDPLLSASSNQQR, (m/z 1025.525, charge 2+). B. Precursor ion mass spectrum of peptide ETNLLYDPLLSASSNQQR. C. CID-MS2 of peptide ETNLLYDPLLSASSNQQR.
Figure 2.4. Representative tandem MS spectra of unique peptides from the other two potential candidates. (A) CID-MS2 of peptide HQEAEMAQNAVR (B) CID-MS2 of peptide GDVNDNFQGVLQNVR
2.5 Discussion

A hallmark of polycystic kidney disease is the displacement of normal renal tubules with fluid-filled cysts and results in a massive enlargement of the kidney. Autosomal-dominant polycystic kidney disease (ADPKD) is a disease of defective tissue homeostasis resulting in active remodeling of nephrons and bile ducts to form fluid-filled sacs called cysts. Numerous studies have identified the proteins polycystin-1 (PC-1), polycystin-2 (PC-2) which are responsible for most cases of the disease. However the mechanism by which mutations that occur in the genes result in the formation of cysts is poorly understood [24]. However, cyst formation and cystic change is also observed in metastatic lymph nodes such as those observed in head and neck cancer with squamous cell carcinoma, as well as fibrocystic breast disease [25, 26].

In a previous study of the plasma/serum sample from breast cancer patients using our multiple lectin affinity chromatography approach we observed changes both in abundance and glycosylation of proteins including thrombospondin -1 and -5 [27]. It was therefore of interest to explore both the proteome and glycoproteins observed in the model system of Pkd1+/+ (WT) and Pkd1-/- cell lines (KO) to determine if there more global changes that in glycosylation. In this study, the initial separation was achieved with 1D electrophoresis combined with mass spectroscopy to determine the glycoproteins present in the Pkd1+/+ and Pkd1-/- cells.

In our previous work we identified the unusual glycan motif of terminal diastialic acid on the N925 and N928 glycosites of the alpha-3 subunit of integrin in the Pkd1-/- cell model (KO). In addition, the KO model cell line was associated with the upregulation of glycosyltransferase alpha-2,8-sialyltransferase 8B (ST8SIA2, O35696) in the Pkd1-/- cells. ST8SIA2 is involved in the protein glycosylation and can transfer sialic acid via alpha-2,8-linkages to alpha-2,3-linked and alpha-2,6-linked sialic acid of N-linked oligosaccharides of glycoproteins and is related to
polysialic acid expression. Thus, the existence of extra sialic acid in glycoproteins of the transformed and malignant cells further suggests that abnormal protein glycosylation may play a role on the pathogenesis of cyst formation in ADPKD.

To further explore the glycoproteins with unusual sialic acid motifs in this model system for polycystic disease we selected an affinity chromatography purification step which utilized Sambucus Nigra Lectin (SNA) which is known to bind specifically to terminal Neu5Ac (α2–6) Gal/GalNAc of glycan conjugates.

Unveiling disease related signal pathways based on differences in protein levels can help elucidate disease mechanisms which can be helpful for biomarker and drug target discovery. In the proteomic analysis, the identified unique peptides were used to determine the identity of proteins in the cell lines and the total peptide count was used for label free quantitation. The threshold for protein identifications required a minimum of two peptide observations. The average number of identified peptides for each protein was calculated based on the average of three independent runs. We selected proteins that were three times upregulated or downregulated in Pkd1-/- cells to do the following pathway analysis by uploading the protein UniProt accession number to the DAVID bioinformatics Resources (https://david.ncifcrf.gov/home.jsp) and pathways were selected by gene count >= 5 and p value <= 0.05.

In our analysis, nine proteins are downregulated while twelve proteins are upregulated in Pkd1-/- cells related to tight junction pathway. It has been reported that tight junction dysfunction is related to kidney disease [28-30]. We observed Ras-related protein R-Ras upregulated in the mutant cell model. Also, five proteins related to N-Glycan biosynthesis are upregulated in Pkd1-/- cells which supports the conclusion that glycosylation is abnormal in the Pkd1-/- cells. We listed the upregulated glycoproteins found in the mutant cells in the Table 2.1 and after the KEGG
pathway analysis (see Table 2.3) we found that five glycoproteins, laminin subunit gamma-1 (P02468), thrombospondin-1 (P35441), collagen alpha-2(IV) chain (P08122), fibronectin (P11276) and integrin alpha-5 (P11688) were related to extracellular matrix (ECM)-receptor interaction and focal adhesion. In addition, KEGG analysis identified that the PI3K-Akt/ mTOR signaling pathway was upregulated in the KO cells which was consistent with the association of c-MET/ sequestration in the Golgi and activation of this signaling pathway which contributes to cyst formation.

Three possible candidates identified from the SNA affinity purification and bioinformatics analysis were selected for future glycosylation analysis based on their protein function (see Table 2.4). They are catenin beta-1, thrombospondin-1 (TSP-1) and glutamyl aminopeptidase. Studies have reported dysregulation of the Wnt/β- catenin signaling in autosomal dominant polycystic kidney disease (ADPKD) [31] and we observed an upregulation of catenin beta-1 in Pkd1-/- cells. TSP-1 is an adhesive glycoprotein that mediates cell-to-cell and cell-to-matrix interactions. It has been reported that TSP-1 is related to cell-cell adhesion and its overexpression will cause increased tumor progression and tumor cell invasion [32, 33] and was observed in our lectin studies of the breast cancer serum glycoproteome (see earlier discussion). Glutamyl-aminopeptidase (GAP) is involved in many biological processes and is related to the development of kidney structures [34]. It has been reported that altered GAP expression and activity were observed in renal cell carcinoma cell lines and in primary renal cancers [35, 36]. Thrombospondin-1 (TSP-1) and Glutamyl aminopeptidase are also upregulated in the cell lysate of Pkd1-/- cells without the lectin binding step.
### Table 2.3. KEGG pathway analysis of differently expressed proteins

#### Downregulated proteins related KEGG pathway in KO cells

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>P Value</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tight junction</td>
<td>9</td>
<td>0.00</td>
<td>Tjp2, Gna13, Ppp2cb, Yes1, Actn1, Ptkckd, Rab3b, Ybox3, Shroom2</td>
</tr>
<tr>
<td>Carbon metabolism</td>
<td>8</td>
<td>0.01</td>
<td>Tald1, Pgd, Mcj, Pdhb, Pstl1, Got1, Esd, Pcx</td>
</tr>
<tr>
<td>Metabolic pathways</td>
<td>34</td>
<td>0.03</td>
<td>Pgd, Cpoa, Me3, Pcyt1a, Dde, Pdhh, Gart, P4ha2*, Aldh1a1, Aptr, Cyes, Imdh2, Hal, Napt, Ugt1a7c*, Pstl1, Ugdh, Cdol, Alad, Tald1, Bdh1*, Nif5c1a, Pd1, P4cb4, Ida1, Got1, Alg2, Atp5c1, Pnp, Pcx, Galnt14, Rpn2, Ugt1a5, Atp5a1</td>
</tr>
<tr>
<td>Cell adhesion molecules (CAMas)</td>
<td>8</td>
<td>0.03</td>
<td>Vcam1*, H2-K1*, H2-D1*, H2-L*, Itgai, Cdh2*, Alocam, H2-Ab1*</td>
</tr>
<tr>
<td>Endocytosis</td>
<td>11</td>
<td>0.03</td>
<td>H2-K1*, Dmn2, H2-D1*, H2-L*, Rbsn, Rab8a, Clic, Pdk1, Arfgef2, Hspa5, Igvec1</td>
</tr>
<tr>
<td>eGMP-PKG signaling pathway</td>
<td>8</td>
<td>0.04</td>
<td>Adey6, Vdac2, Adey7*, Gna13, Pdtb4, Rock2, Pde2a, Ins4</td>
</tr>
</tbody>
</table>

#### Upregulated proteins related KEGG pathway in KO cells

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>P Value</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focal adhesion</td>
<td>16</td>
<td>0.00</td>
<td>Myrl2a, Itgai1*, Isgf7r, Col4a2*, Pik3r1, Fni1*, Igb5, Igav*, Lancl1*, Actn1, Thbs1*, Flna, Actn4, Mapk1, Igs5*, Ppp1r2a</td>
</tr>
<tr>
<td>Regulation of actin cytoskeleton</td>
<td>16</td>
<td>0.00</td>
<td>Myrl2a, Rms, Itgai1*, Insr*, Gsn, Shi1, Pik3r1, Fni1*, Igb5, Igav*, Nckap1, Actn1, Actn4, Mapk1, Igs5*, Ppp1r2a</td>
</tr>
<tr>
<td>Tight junction</td>
<td>12</td>
<td>0.00</td>
<td>Myrl2a, Rms, Myb7b, Lgl2, Ppp2rlb, Actn1, Actn4, Ctn, Eplb4, Igy5, Ctma2, Ppp2r2b</td>
</tr>
<tr>
<td>ECM-receptor interaction</td>
<td>9</td>
<td>0.00</td>
<td>Igb5, Igav*, Lancl1*, Itgai1*, Thbs1*, Cd44, Col4a2*, Fni1*, Igs5*</td>
</tr>
<tr>
<td>Adherens junction</td>
<td>7</td>
<td>0.01</td>
<td>Farp2, Actn1, Actn4, Isgf7r, Ctma2, Mapk1, Ctncl1</td>
</tr>
<tr>
<td>Endocytosis</td>
<td>15</td>
<td>0.02</td>
<td>Rab8a, Nedd4, Clic, Cap2b, Ap2a1, Arf5, Ldh*, Cap41, H2-Q7*, Ap2m1, Isgf7r, Igvec2, Asap2, Rab1lfyp5, Dab2</td>
</tr>
<tr>
<td>N-Glycan biosynthesis</td>
<td>5</td>
<td>0.04</td>
<td>Stt3a*, Rpn2, Stt3b*, Dpm1, Alg2</td>
</tr>
<tr>
<td>Phagosome</td>
<td>10</td>
<td>0.04</td>
<td>H2-Q7, M6pr*, Igb5*, Igav*, Thbs1*, Sec61a1, H2-E2b, H2-Ab1*, Igs5*, H2-Aa*</td>
</tr>
<tr>
<td>PI3K-Akt signaling pathway</td>
<td>16</td>
<td>0.04</td>
<td>Igai1*, Col4a2*, Pik3r1, Fni1*, Ppp2rb2, Cdc37, Igav*, Lancl1*, Thbs1*, Mapk1, Igs5*, Wwphq, Isgf7r, Igb5, Ppp2rb1, Jak2</td>
</tr>
</tbody>
</table>

Note: the italic gene is unreviewed in Uniprot (computer-annotated TrEMBL section) * means the gene codes for a protein identified as being glycosylated in Uniprot. The glycoproteins identified in SNA pull down analysis is labeled with underline. KO represents of Pkd1−/− cells and the protein level (downregulated or upregulated) is compared with Pkd1−/− cells (WT). Count refers to the number of genes involved in the term. P Value refers to modified fisher exact P-Value, EASE Score. The smaller, the more enriched. Usually P-Value is equal or smaller than 0.05 to be considered strongly enriched in the annotation categories.
Table 2.4. Potential glycoprotein candidates for future study

<table>
<thead>
<tr>
<th>Uniprot Accession</th>
<th>Protein Name</th>
<th>Theoretical Glycosylation sites</th>
<th>Protein Score</th>
<th>Coverage</th>
<th>M.W.</th>
<th>Calc pI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q02248</td>
<td>Catenin beta-1</td>
<td>1, O-linked</td>
<td>42.77</td>
<td>12.80%</td>
<td>85.4</td>
<td>5.86</td>
</tr>
<tr>
<td>P16406</td>
<td>Glutamyl aminopeptidase</td>
<td>9, N-linked</td>
<td>34.53</td>
<td>9.21%</td>
<td>107.9</td>
<td>5.44</td>
</tr>
<tr>
<td>P35441</td>
<td>Thrombospondin-1</td>
<td>4, N-linked</td>
<td>34.79</td>
<td>7.61%</td>
<td>129.6</td>
<td>4.96</td>
</tr>
</tbody>
</table>

2.6 Conclusion

In our study, 1D electrophoresis combined with mass spectroscopy were used to determine the protein and glycoprotein difference in the Pkd1+/+ and Pkd1-/- cells. Lectin Sambucus Nigra Lectin (SNA) was chose to pull down the glycoprotein with α 2–6 linked sialic acid. Three glycoproteins, Catenin beta-1, Thrombospondin-1 (TSP-1) and Glutamyl aminopeptidase will be targeted for future study of glycan motifs associated with ADPKD.

2.7 References


Chapter 3: Site-Specific glycosylation characterization of Influenza a Virus Hemagglutinin produced by Spodoptera frugiperda insect cell line using multienzyme digestion with nanoLC mass spectrometry

Manuscript is submitted to Analytical Chemistry

Title: Site-Specific glycosylation characterization of Influenza A Virus Hemagglutinin produced by Spodoptera frugiperda insect cell line using multienzyme digestion with nanoLC mass spectrometry

Author: Yan-Jun Liu, Shiaw-Lin Wu, Kerry R. Love and William S. Hancock*
3.1 Abstract

Influenza hemagglutinin is a surface glycoprotein which is related to virus invasion and host immune system response. Site specific glycosylation investigation of hemagglutinin will increase our knowledge about virus evolution and can improve the design and quality of vaccines. In our study, we used glycoproteomic analysis based on multienzyme digestion followed by LC tandem MS analysis to determine the glycosylation of Influenza hemagglutinin (H1/A/California/04/2009) using the following steps. PNGaseF treatment combined with trypsin or pepsin digestion were used to determine the glycosites and glycan occupancy. Three enzymes trypsin, AspN and pepsin were used separately to generate suitable glycopeptides for on-line LC tandem MS analysis. The glycan structure of a given glycopeptide was determined by collision-induced dissociation MS/MS fragmentation and the peptide backbone information was provided by CID-MS3 fragmentation. With this approach 100% sequence coverage of the hemagglutinin sample was obtained. Six glycosylation sites fitting the sequon N-X-S/T were successfully confirmed and the glycan heterogeneity as well as the ratios of glycoforms were determined at each site.

3.2 Introduction

Influenza hemagglutinin (HA) is an integral membrane glycoprotein, which is located on the surface of the viral particle [1]. The glycoprotein plays crucial part in the process by which viruses attach and penetrate host cells and it is also the main antigenic protein that simulates the host immune system responses [2-4]. Cell-based production of influenza vaccines containing recombinant HA proteins are more efficient than the traditional embryonated hen egg based production systems due to the absence of potential contaminants such as pathogenic viruses and egg derived proteins. A variety of cell based systems are used to produce vaccines, such as mammalian, insect and bacterial cells [5]. However, a production system that uses insect cells and
the baculovirus expression vector (BEVS) can provide a quick access to biologically active proteins, in weeks instead of months or even years. FluBlok is the first recombinant HA influenza vaccine which is produced by an insect system [6].

In vaccine development, glycosylation is one of the most important posttranscriptional modifications. Previous studies have produced evidence that glycan moieties on the HA antigen can shield or modify its antigenic sites [7-9]. It has also been reported that glycosylation in specific sites of HA can participate in the formation of the active trimeric structure [10-12]. Thus, development of an analytical method to precisely monitor site-specific glycosylation will facilitate the characterization and production of a more consistent, higher quality HA antigen preparation. In addition, the methods developed in this study for the characterization of site specific glycans can be used to observe glycosylation changes in the HA antigen in nature, which may help the understanding of the process of virus evolution and the design of better vaccines for the prevention of Influenza virus pandemics.

A comprehensive analysis of protein glycosylation includes determination of glycosylation sites, quantitation of glycosite occupancy, determination and relative quantitation of major glycan structures and thus assessment of glycan heterogeneity. However, the macroheterogeneity (glycans attached to multiple glycosylation sites) and the microheterogeneity (a single glycosylation site embraces different type of glycans) make the study of glycosylation profiles a major challenge. In addition, due to relatively low ionization efficiency of the glycopeptides, the MS signal of glycopeptides may be suppressed by coeluting peptides [13].

Oligosaccharide profiling and glycopeptide analysis are two common approaches in analytical studies of glycoproteins [14]. In the oligosaccharide profiling, the glycans are released by chemical or enzymatic techniques and analyzed by mass spectrometry [15, 16]. This procedure provides
glycan structural information and simplifies the mass spectra by removing the protein moiety from the glycan; however, in this process information about the glycan site attachment site is lost.

In contrast, glycopeptide analysis gives both the glycan structure and site attachment information. The glycoprotein is enzymatically digested and the resulting glycopeptides are then analyzed by LC-tandem MS analysis [17-19]. Sometimes, a single enzyme, such as trypsin, is sufficient to completely digest a given protein [20, 21]; however, for a complex glycoprotein such as HA antigen, other enzymes with broader specificity are required for a better sequence coverage and to allow isolation of glycopeptides suitable for MS analysis [22, 23].

In our study, two enzymes with high specificity, trypsin and AspN, as well as pepsin an enzyme with broader specificity, were used to achieve 100% sequence coverage and profile glycosylation of the HA antigen. The digests were analyzed by reversed phase liquid chromatography coupled with tandem mass spectrometry. The glycan structure attached to a given glycopeptide was determined by collision-induced dissociation (CID) MS/MS (MS2) fragmentation and the peptide backbone information was provided by CID-MS3 fragmentation. In this manner, we confirmed six glycosylation sites, determined the major glycan structures at each site as well as the ratios.

3.3 Material and Method

3.3.1 Samples and Reagents.

HA antigen (H1/A/California/04/2009) was bought from Protein Sciences Corporation (Meriden, CT). Trypsin (Sequencing grade) was purchased from Promega (Madison, WI). Asp-N Endoproteinase (Pierce, MS Grade), Tris (2-carboxyethyl) phosphine hydrochloride (TCEP-HCl) was from Thermo Scientific (Waltham, MA). Pepsin (Porcine stomach mucosa) was from MP Biomedicals (Solon, OH). PNGaseF, Ammonium bicarbonate, formic acid, dithiothreitol (DTT)
and iodoacetamide (IAA) were purchased from Sigma-Aldrich (St. Louis, MO). 1.0 N hydrochloric acid (HCl) solution, LC-MS grade water and acetonitrile were purchased from Fisher Scientific (Fairlawn, NJ). NuPAGE® 4-12% Bis-Tris gels, Novex® sharp pertained protein standard, SimplyBlueTM safe stain were purchased from InvitrogenTM (Carlsbad, CA).

3.3.2 Gel electrophoresis (SDS-PAGE) analysis.

1 μL DTT and 4 μL loading buffer (4X) were added to 10 μL protein solution. The mixture was heated at 70 °C for 10 min before gel electrophoresis. The samples were loaded on a SDS-PAGE gel (4-12% Bis-Tris gel) for separation by molecular weight. The constant voltage of 160 V is applied for 60 min to run the gel. After the gel electrophoresis, remove the gel from the cassette and wash the gel with the de-ionized water. Then, stain the gel for 1 hour at room temperature and de-stain overnight using deionized water under the same conditions.

3.3.3 Enzyme digestion.

Samples containing 20 μg proteins were digested by three enzymes separately with or without PNGaseF using different protocols. Trypsin digestion. HA is buffer exchanged with 6 M guanidine hydrochloride in 100 mM ammonium bicarbonate by ultracentrifugation (11000 g, 5 min per cycle, 3 cycle). The denatured protein solution is reduced by DTT at 37 °C for 30 min (DTT final concentration is 2 mM) and alkylated by IAA in dark at room temperature for 1 hour (IAA final concentration is 10 mM). The reduction and alkylation sample was transferred to a 10 kDa MWCO filter and buffer exchanged with the digestion buffer (100 mM ammonium bicarbonate, pH 7.8) to remove the guanidine hydrochloride and excess DTT and IAA. Add trypsin (enzyme weight to substrate weight = 1:50) to the protein solution and incubate at 37 °C for overnight. 5% formic acid was used to stop the reaction. Trypsin plus PNGaseF digestion. Using
exactly the same conditions as the trypsin digestion we added 5 units PNGaseF (500 units/ml) with trypsin to the protein solution. **AspN digestion.** The procedures for the reduction, alkylation and buffer exchange of HA is the same as for trypsin digestion. The weight ratio of AspN to substrate is 1:100. **Pepsin digestion.** HA is buffer exchanged with 50 mM TCEP in 10 mM HCl by 10 kDa MWCO filter. The buffer exchange solution is stored at room temperature for 1 hour for reduction. Add pepsin (enzyme weight to substrate weight = 1:10) to the protein solution and incubate at 37 °C for 30 min. **Pepsin plus PNGaseF digestion.** 20 μg HA solution (pH 7.0) was incubated with 5 units of PNGaseF for overnight at 37 °C. The deglycosylated sample was reduced and digested using the exactly same conditions as employed in the pepsin digestion.

### 3.3.4 NanoLC tandem analysis.

All the digested samples were loaded on a self-packed 75 μm i.d. x 15 cm C18 column (Magic C18 beads, 200Å pore and 5 μm particle size) (Michrom BioResources, Auburn, CA). The peptides or glycopeptides analysis is performed on an Ultimate 3000 nano LC pump (Dionex, Mountain View, CA) coupled on-line to a LTQ-Orbitrap-XL mass spectrometer (Thermo Fisher Scientific, San Jose, CA) interfaced with a nano spray ion source (New Objective, Woburn, MA). Mobile phase A (0.1% formic acid in water) and mobile phase B (0.1% formic acid in acetonitrile) were used for the gradient consisting of linear from 2% to 40% B in 60 min, then linear from 40% to 80% B for 10 min and finally isocratic at 80% B for 5 min. The mass spectrometer was operated in the data-dependent mode by alternatively switching between MS survey with the scan mass range 300-2000 and CID tandem MS fragmentation for the ten most high intensity precursor ions. The parameters used to operate the mass spectrometer were: mass resolution of 30,000 for 400 m/z, precursor ion isolation width is ± 1 m/z, 35% normalized collision energy, 2.3 kV spray voltage, capillary temperature of 270 °C and dynamic exclusion was 30 s for 2 repeat counts.
3.3.5 Data analysis.

The spectra generated in the CID-MS2 analysis were searched against spectra of theoretical fragmentations (b and y ions) of a digest of the HA sequence (Uniprot ID: C3W5S1) by Biopharm Finder with a mass tolerance ≤5 ppm of precursor ions and the peptide confidence > 99% as the filter.

3.4 Result and Discussion

In our study, the HA antigen (H1/A/California/04/2009), produced by the Spodoptera frugiperda insect cell line, was used as a model system to develop an analytical protocol for the analysis of the influenza hemagglutinin (H0A). Firstly, we confirmed the reported sequence [24-26], which contains 566 amino acids by achieving 100% sequence coverage in our analysis (see Figure 3.4). In addition, we confirmed the expected cleavage of the signal peptide (17 amino acids) from the precursor HA0 when N-linked glycosylation occurs during the co-translational process [27]. In our study, the expected peptide was observed with cleavage between residues 17 and 18 in the GluC (Figure 3.6) or trypsin (Figure 3.7) digests.

We then characterized the glycosylation site, glycosite occupancy as well as glycan types and ratios of different forms. In order to determine the glycosylation status of the HA antigen, PNGaseF treated and untreated samples were analyzed by SDS-PAGE. The gel image (Figure 3.1A) displays an altered migration to a slightly lower molecular weight in the PNGaseF treated
Figure 3.1. **HA antigen (H1/A/California/04/2009) is a glycoprotein with 8 possible glycosylation sites.** (A) SDS-PAGE analysis of HA antigen with or without PNGaseF treatment. (B) The predicted 3D structure of HA antigen was performed in the website (https://swissmodel.expasy.org/) based on the sequence downloaded from Uniprot (Uniprot ID: C3W5S1). The theoretical glycosylation sites are labeled with green squares.

The evidence confirms that the HA antigen contains N-linked glycosylation. N-linked glycosylation is restricted to the sequon N-X-S/T, in which X cannot be a proline and based on this feature, the theoretical glycosylation sites of HA are N27, N28, N40, N104, N293, N304, N498 and N557. Figure 3.1B exhibits the predicted 3D structure of the HA antigen based on the protein sequence from Uniprot and the theoretical glycosylation sites are labeled with green squares.
Table 3.1. Glycopeptides and Non-glycosylated peptides identified by trypsin or pepsin after PNGaseF treatment

<table>
<thead>
<tr>
<th>Theoretical Glycosylation site</th>
<th>Peptide identified by Trypsin map</th>
<th>Peptide identified by Pepsin map</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Identification</td>
<td>m/z (Da)</td>
</tr>
<tr>
<td>N27/N28</td>
<td>DTLC*IGYHAN⁄N⁄DST DVDTVLEK, 18-39</td>
<td>1234.573</td>
</tr>
<tr>
<td>N-40</td>
<td>N/DVTVTHSVNLLEDK, 40-53</td>
<td>785.414</td>
</tr>
<tr>
<td>N-104</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>N-293</td>
<td>NAGSGIIISDTPVHDC<em>N/DTTC</em>QTPK, 277-300</td>
<td>863.401</td>
</tr>
<tr>
<td>N-304</td>
<td>GAIN/DTSLPFQNIHPITIGK, 301-319</td>
<td>675.042</td>
</tr>
<tr>
<td>N-498</td>
<td>N/DGTYDYPK, 498-505</td>
<td>479.714</td>
</tr>
<tr>
<td>N-557</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Note: N/D represents the conversion of Asn to Asp with PNGaseF deglycosylation. N^ means the glycosylation sites is this asparagine or the adjacent one. NA means the glycopeptides cannot be detected. *Carbamidomethylated cysteine.

3.4.1 Six out of eight possible glycosylation sites were glycosylated.

The software Biopharma Finder was used to process the spectrometric data. For data processing of the PNGaseF treatment of pepsin digests the precursor mass tolerance is set at 20 ppm with a signal to noise threshold is 1.00+E3. The mass increase of 0.984 Da, which is caused by the removal of oligosaccharides with PNGaseF treatment, was used to profile the glycosylation sites. In this manner, 7 out of 8 possible glycosylation sites which contain the N-X-T/S sequon were
present in the pepsin digest and 5 of them (N27/N28, N40, N104, N293, N498) were occupied by glycans.

No deamidated form was observed for the peptide WMCSNGSLQC, residues 553-562 after PNGaseF treatment and its precursor mass 1128.431, (charge 1+) was identified in both the PNGaseF treated and untreated samples with the exactly same MS2 fragmentation pattern (Figure 3.5). Thus, there is no evidence from our results to support the conclusion that N557 is glycosylated which is consistent with the location of this region of the protein in the intracellular region of HA [28]. In a similar manner, we used mass spectrometric analysis of trypsin or trypsin combined with PNGaseF digests to confirm that the N304 site is glycosylated. Only one deamidation and thus glycosylation site was found in the peptide IGYHANNSTDTVD (see below for details of the analysis). In Table 3.1 we summarize these results by showing the peptides containing the N-X-T/S sequon and the glycosylation status as identified by the trypsin or pepsin map.

3.4.2 Tandem MS/MS data used to confirm that N28 is the major glycosylated site in peptide 22 to 34.

In this peptide, there are two theoretical glycosylation sites, N27 and N28 which are adjacent to each other. With our digestion protocols, it was not possible to isolate these two residues into different pep-tides by enzymatic digestion. In order to determine the exact glycosylation site, we pretreated the HA antigen with PNGaseF at pH 7.0 and then digested the sample with pepsin at pH 2.0 to avoid possible artificial induced deamidation by enzymes such as trypsin which require higher pH values. The theoretical monoisotopic mass to charge ratio predicted for the doubly deamidated IGYHANNSTDTVD pep-tide i.e. (m/z) 1408.586 (charge 1+), 704.797 (charge 2+), 470.200 (charge3+) were not observed in the corresponding extracted ion chromatograms.
Figure 3.2. MS and tandem MS analysis of the peptide IGYHANDSTDTVD (residues 22-34) containing the site N28 prepared by pepsin digestion. (A) Precursor ion mass spectrum of peptide 22-34 (m/z 704.309, charge 2+). (B) CID-MS2 of peptide 22-34. The diagnostic b and y ions for determination of the deamidation site is zoomed in for a clearer view.

Furthermore, the observed precursor mass (m/z value 704.305 (charge 2+)), indicates that there is only one deamidation site in this peptide as measured by PNGaseF treatment. Then tandem MS/MS data gave the location of the glycosylation site in this peptide. In the MS/MS analysis of the corresponding peptic peptide a 0.985 mass defect was observed in the y7 fragment while the mass of b6 corresponded to the undeamidated form of the peptide (Figure 3.2). Further confirmation that the N28 site is indeed the major glycosylation site was provided by the analysis
### Table 3.2. Quantitation of glycan occupancy at each glycosylation sites

<table>
<thead>
<tr>
<th>Glycosylation site</th>
<th>Occupancy (Avg. n=3)</th>
<th>Std. dev. (n=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-28</td>
<td>Fully glycosylated</td>
<td>N/A</td>
</tr>
<tr>
<td>N-40</td>
<td>Fully glycosylated</td>
<td>N/A</td>
</tr>
<tr>
<td>N-104</td>
<td>88.7%</td>
<td>0.6%</td>
</tr>
<tr>
<td>N-293</td>
<td>87.2%</td>
<td>0.5%</td>
</tr>
<tr>
<td>N-304</td>
<td>98.7%</td>
<td>0.3%</td>
</tr>
<tr>
<td>N-498</td>
<td>95.6%</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

Note: In the quantitation analysis of partially glycosylated sites, because of the significant difference of the mass spectrometry response between the glycopeptide and its non-glycosylation form, PNGaseF was utilized to totally remove the glycan and convert the Asn to Asp. What is more, the peptides contain multiple enzyme cleavage sites or include multiple deamidation sites were excluded from the analysis.

- **a** Average glycan occupancy from three individual runs. The calculation is based on formula that the glycan occupancy is equal to the amount of deglycosylated peptides divided by the sum of the value of deglycosylated peptides and their corresponding nonglycosylated counterparts. We used the peak area of the extracted chromatogram of the most abundant charge state of the peptides to do the quantitation and the peak area is integrated over the time scale.

- **b** Standard deviation. This analysis was performed in three replicates and the standard deviation is below 1%, showing the good reproducibility of the analysis.

- **c** Calculation based on pepsin digests. The sequence containing N104 is not covered by trypsin digestion of the molecule and the peptide NAGSGIISDTPVHDC*NTTC*QTPK, 277-300 includes an easily deamidated site N277, which will affect the exact quantitation of the glycosylation occupancy in the site N293, so we used pepsin digests to do the glycan occupancy of these two sites. The occupancy of the site N104 and N293 are 88.7% and 87.2% separately.

- **d** Calculation based on trypsin digests. The occupancy of the site N304 and N498 are 98.7% and 95.6% separately, which is calculated from the trypsin digests.
of the corresponding digest results. The mass defect observed in y11 from a GluC digest (Figure 3.6) and in y12 from trypsin digestion (Figure 3.7) with no mass change in b10 confirmed that deamidation occurred in N28 after PNGaseF treatment.

3.4.3 Glycan occupancy analysis of HA antigen

Determination of the glycan occupancy in each specific site is necessary for the full characterization of a glycoprotein. We firstly identified any non-glycosylated peptides which contained the glycosylation motifs via database searching of the tryptic or peptic digests. The non-glycosylated form was identified in the peptides containing the N104, 293, 304 and 498, but not found in the peptides containing N28 and 40. We also manually checked these peptides and the extracted ion chromatograms match to confirm the result. Based on these data, we draw the conclusion that sites 28 and 40 were fully glycosylated while the sites 104, 293, 304 and 498 were partially glycosylated. The percentage of glycan occupancy at each glycosylation site is displayed in Table 3.2.

3.4.4 Oligosaccharide Heterogeneity characterization of site N28 as an example

Oligosaccharide Heterogeneity characterization of site N28 as an example. LC tandem MS is a valuable tool for characterization of glycan structures and ratios at a specific site. Firstly, CID fragmentation of glycopeptides will generate low molecular weight diagnostic oxonium ions (m/z 366, 528 and 657) [29, 30] and a series b ions (peptide moiety linked with a glycan fragment), which allow us to deduce a glycan structure. However, CID-MS2 fragmentation gives little or no information on the peptide backbone structure which was then confirmed by CID-MS3 fragmentation. In this report, the Asn 28 is used as the representative site for analysis of the glycan
heterogeneity. The software Biopharma Finder was used to perform the database search for the sample without PNGaseF treatment for initial identification of glycopeptides. The precursor mass tolerance is set as 5 ppm for an accurate mass identification, then the identified structure was confirmed by manual annotation of the experimental glycopeptide MS2 and MS3 spectra. The identified glycans at residue 28 were high mannose and hybrid types which is presented in the Figure 3.3A and thus show the glycan heterogeneity at this site. As expected the retention time of these glycopeptides is similar (from 17.5 to 18.2 min, data not shown) because the major factor affecting reversed phase HPLC retention is the hydrophobicity of the peptide backbone rather than the nature of the carbohydrate moiety. Figure 3.3B displays the tandem MS spectrum of the ion 1312.525, charge 2+ (N28 with 5 high mannose residues). We observed the diagnostic oxonium ions m/z 366 (Hex-HexNAc+) and 528 (Hex-Hex-HexNAc+) in the spectrum. The precursor ion m/z 1609.61(charge 1+) which carries GlcNAc plus the intact peptide from the CID MS2 was selected to perform CID MS3 fragmentation. The spectrum in Figure 3.3C shows the peptide backbone information.

3.4.1 The glycan ratios at all six sites of HA antigen and major glycan types

The glycan ratios at all six sites of HA antigen and major glycan types. The baculovirus-insect cell system is widely utilized to produce recombinant proteins. Unlike its vertebrate counterparts, glycoproteins produced by the insect cells contains simpler, shorter N-glycans with little sialylation. The oligosaccharide side chains carry high-mannose-type and truncated trimannosyl N-glycans, frequently with a fucose residue attached to the Asn-linked GlcNAc [31]. The development of analytical methods to monitor the oligosaccharide present in proteins produced by recombinant DNA technology in this cell line is crucial due to the potential immunogenicity of such glycans which are not present in human glycoproteins.
Figure 3.3. Representative glycopeptide analysis of HA antigen digests. (A) Average mass spectrum of glycopeptide IGYHANN*STDTVD including all the detectable glycans. (B) CID-MS/MS (MS2) spectrum of the ion 1312.525 (charge 2+). (C) CID-MS/MS/MS (MS3) spectrum of fragment ion 1609.582 (charge 1+). Green round circles, mannose; blue squares N-acetylglucosamine (GlcNAc); red triangle, fucose.
Glycopeptides, especially those containing a large glycan (such as mannose 8), generally do not ionize well with electrospray. In addition, the heterogeneity due to the different glycan species further decreases the signal intensity observed in MS analysis. Sometimes, a single specific enzyme digestion (usually trypsin) is sufficient for glycoprotein characterization, however, in many cases, using a single enzyme will result in incomplete protein sequence information due to a partial digestion or the generation of peptides not suitable for MS analysis. A solution to this challenge is the use of suitable combinations of enzymes but such combinations may generate too many small peptides (3-5 amino acids) and result in ion suppression of the glycopeptides [32-34]. In this study, we used separate digestions with trypsin, AspN and pepsin to generate a full set of glycopeptides for MS analysis.

Trypsin digestion generates suitable glycopeptides for glycan structure analysis in the sites N40, N293 and N304. The major glycan identified in site N40 is of the high mannose type and we observed five glycoforms, ranging from Man 3 to Man 8 but missing the Man 4 species. Four glycoforms, Man 3, 5, 6, 7 were identified in the pepsin digest and 2 glycoforms Man 3,5 were identified by AspN digestion. The reason for the incomplete identification by pepsin or AspN in this site is unclear, but this may be related to recovery issues or ionization efficiency of the different glycopeptides. The glycoforms observed for site N293 for trypsin digestion shows the same pattern as the site N40. In the site N304, with trypsin digestion we detected three hybrid oligosaccharide structure (A1G0, A1G0M4 and A1G0M5) besides the high mannose type. No glycopeptide information was provided by either the AspN or pepsin digestion for sites N293 and N304. Figure 3.8 displays the elution time window and all the glycoforms identified in these three sites.

The site N104 is not covered by peptides produced by the trypsin digestion and we did not detect any glycopeptides containing this site in the pepsin digests. Another enzyme, AspN was chosen to
solve the problem based on the protein sequence. AspN generates glycopeptide DN(104)GTCYPG with three high mannose glycoforms (Man 3, Man5 and Man 6) (Figure 3.9B). The elution time window for these glycopeptides is from 17.5 to 19.5 min which is labeled by red squares in the Figure 3.9A panel. For the site N498, seven glycoforms were determined in AspN digests (Figure 3.9C). Besides the common high mannose glycans type, two hybrid types (A1G0, A1G0F) and one complex type (A2G0F) were identified at this site and both fucosylated N-glycans were observed to contain core fucosylation based on the CID MS2 analysis (data not shown). The interesting discovery is mannose-4 glycan was not observed for any of the glycosylation sites. We also targeted the theoretical mass for any of the glycopeptides that contained the mannose-4 glycan but again no MS2 data can be observed for a glycopeptide containing this glycan.

The glycoform ratios were quantitated by peak areas measured for each glycoform at a specific site divided by the sum of peak areas of all the glycoforms identified at that site (Table 3.3). The major glycoform observed at site N304 is mannose-8 and we detected two hybrid glycan A1G0M4 and A1G0M5 with an abundance of only 0.93%, 0.69% respectively. The key intermediate in the protein N-glycosylation pathway is Man5GlcNAc2 [31] and correlates with the most abundant glycan structures in glycopeptides containing Asn 28, Asn 40 and Asn 498. The major insect processed N-glycan products Man3GlcNAcGlcNAc [31] were the most abundant glycoform observed for glycopeptides containing Asn104 and Asn 293.
## Table 3.3. Relative quantitation of different glycoforms at each glycosylation sites

<table>
<thead>
<tr>
<th>Site</th>
<th>Enzyme (optimized) and peptide sequence</th>
<th>Glycan</th>
<th>Observed m/z</th>
<th>charge</th>
<th>Ratio (Avg, n=3)</th>
<th>Std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>N28</td>
<td>pepsin IGYHAN VSSTDTVD</td>
<td>A1G0F</td>
<td>1325.039</td>
<td>2</td>
<td>20.3%</td>
<td>0.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3</td>
<td>1150.470</td>
<td>2</td>
<td>6.8%</td>
<td>0.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M5</td>
<td>1312.526</td>
<td>2</td>
<td>38.0%</td>
<td>2.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6</td>
<td>1393.551</td>
<td>2</td>
<td>15.5%</td>
<td>0.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M7</td>
<td>1474.578</td>
<td>2</td>
<td>12.5%</td>
<td>1.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M8</td>
<td>1555.604</td>
<td>2</td>
<td>6.9%</td>
<td>0.5%</td>
</tr>
<tr>
<td>N40</td>
<td>TrypsinNVTVTHS VNLLEDK</td>
<td>M3</td>
<td>821.396</td>
<td>3</td>
<td>29.1%</td>
<td>0.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M5</td>
<td>929.434</td>
<td>3</td>
<td>31.7%</td>
<td>0.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6</td>
<td>983.452</td>
<td>3</td>
<td>24.9%</td>
<td>0.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M7</td>
<td>1037.470</td>
<td>3</td>
<td>12.2%</td>
<td>0.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M8</td>
<td>1091.487</td>
<td>3</td>
<td>2.1%</td>
<td>0.1%</td>
</tr>
<tr>
<td>N104</td>
<td>AspN DNGTCYPG</td>
<td>M3</td>
<td>888.328</td>
<td>2</td>
<td>58.3%</td>
<td>2.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M5</td>
<td>1050.880</td>
<td>2</td>
<td>22.0%</td>
<td>2.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6</td>
<td>1131.407</td>
<td>2</td>
<td>19.6%</td>
<td>0.5%</td>
</tr>
<tr>
<td>N293</td>
<td>Trypsin NAGSHIII DTPVH DCNTT CQTPK</td>
<td>M3</td>
<td>1160.520</td>
<td>3</td>
<td>36.7%</td>
<td>2.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M5</td>
<td>1268.890</td>
<td>3</td>
<td>20.4%</td>
<td>1.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6</td>
<td>1322.910</td>
<td>3</td>
<td>10.8%</td>
<td>1.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M7</td>
<td>1376.930</td>
<td>3</td>
<td>13.7%</td>
<td>1.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M8</td>
<td>1430.950</td>
<td>3</td>
<td>18.4%</td>
<td>1.7%</td>
</tr>
<tr>
<td>N304</td>
<td>Trypsin GAINTSLPFQNIHPITIGK</td>
<td>A1G0</td>
<td>1039.840</td>
<td>3</td>
<td>4.4%</td>
<td>0.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1G0M4</td>
<td>1093.850</td>
<td>3</td>
<td>0.9%</td>
<td>0.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1G0M5</td>
<td>1148.200</td>
<td>3</td>
<td>0.7%</td>
<td>0.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3</td>
<td>972.145</td>
<td>3</td>
<td>11.3%</td>
<td>2.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M5</td>
<td>1080.180</td>
<td>3</td>
<td>15.2%</td>
<td>2.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6</td>
<td>1134.200</td>
<td>3</td>
<td>9.5%</td>
<td>2.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M7</td>
<td>1188.220</td>
<td>3</td>
<td>19.9%</td>
<td>5.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M8</td>
<td>1242.230</td>
<td>3</td>
<td>35.5%</td>
<td>5.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M9</td>
<td>1296.580</td>
<td>3</td>
<td>2.6%</td>
<td>0.7%</td>
</tr>
<tr>
<td>N498</td>
<td>AspN ESVK NGTY</td>
<td>A1G0</td>
<td>996.919</td>
<td>2</td>
<td>2.8%</td>
<td>0.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1G0F</td>
<td>1070.448</td>
<td>2</td>
<td>29.9%</td>
<td>2.7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2G0F</td>
<td>1171.989</td>
<td>2</td>
<td>12.5%</td>
<td>1.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3</td>
<td>895.379</td>
<td>2</td>
<td>14.2%</td>
<td>1.4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M5</td>
<td>1057.433</td>
<td>2</td>
<td>35.4%</td>
<td>3.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M6</td>
<td>1138.959</td>
<td>2</td>
<td>2.2%</td>
<td>0.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M7</td>
<td>1219.986</td>
<td>2</td>
<td>3.2%</td>
<td>0.7%</td>
</tr>
</tbody>
</table>

Note: a The glycosylation site in the peptide is labeled in red. b The corresponding glycan structure for all the glycan abbreviations are listed in the table 3.4. c Average glycoform ratio is calculated based on three separate runs. d Standard deviation.
3.5 Conclusions

A multienzyme digestion strategy coupled with nano-UPLC-tandem MS has been successfully applied for the sensitive site-specific glycosylation analysis of hemagglutinin A. 100% sequence coverage, six N-linked glycosites and glycan occupancy on each site are determined by PNGaseF combined with trypsin or pepsin digestion. Three enzymes trypsin, AspN and pepsin were used separately to generate glycopeptides suitable for direct LC tandem MS analysis. A total of thirty-five glycoforms were determined and the glycan ratio in each site was quantitated by three separate data-dependent LC-MS runs. This study provides the methodology for sensitivity, reliable, robust and rapid glycan profiling of hemagglutinin. In the future, this state-of-art analytical method can also be applied to the analysis of other complex glycoproteins.

3.6 Reference


### 3.7 Supplementary Data

![Signal peptide](image)

**Figure 3.4. 100% sequence coverage of HA antigen.** HA antigen molecule was pretreated with PNGaseF to convert the glycosites Asn to Asp, which is labeled in red (the reported sequence is asparagine linked with glycan, after treated with PNGaseF the glycan will be removed and the asparagine will change to aspartic acid, N—>D). The sequence labeled with green color is observed by trypsin digestion while the sequence labeled with blue only can be detected by pepsin digestion only. The sequence covered overlap by two digestions is represented by trypsin digests. The sequence (M1 to A17) with black color is missing by both digestion and this peptide is the signal peptide that has been cleaved during the co-translational process.
### Table 3.4. Glycan names and their structures mentioned in the study

<table>
<thead>
<tr>
<th>Glycan name</th>
<th>Structure</th>
<th>Glycan name</th>
<th>Structure</th>
<th>Glycan name</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-Mannose</td>
<td>Hybrid</td>
<td>Complex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td><img src="image" alt="M3 Structure" /></td>
<td>A1G0</td>
<td><img src="image" alt="A1G0 Structure" /></td>
<td>A2G0F</td>
<td><img src="image" alt="A2G0F Structure" /></td>
</tr>
<tr>
<td>M5</td>
<td><img src="image" alt="M5 Structure" /></td>
<td>A1G0M4</td>
<td><img src="image" alt="A1G0M4 Structure" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M6</td>
<td><img src="image" alt="M6 Structure" /></td>
<td>A1G0M5</td>
<td><img src="image" alt="A1G0M5 Structure" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M7</td>
<td><img src="image" alt="M7 Structure" /></td>
<td>A1G0F</td>
<td><img src="image" alt="A1G0F Structure" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M8</td>
<td><img src="image" alt="M8 Structure" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M9</td>
<td><img src="image" alt="M9 Structure" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.5. MS and tandem MS analysis of peptide WMCSNGSLQC, 553-562. (A) Precursor ion mass spectrum of peptide WMCSNGSLQC (m/z 1128.431, charge 1+) in PNGaseF treated and untreated samples. (B) CID-MS2 of peptide WMCSNGSLQC in PNGaseF treated and untreated samples. The identical precursor ion and tandem MS spectra between PNGaseF treated and untreated samples shows the identity of the peptide WMCSNGSLQC without any glycan modification.
Figure 3.6. MS and tandem MS analysis of GluC digestion peptide containing the glycosylation site, residue 28. (A) precursor ion mass spectrum of peptide DTLCIGYHANDSTDVTVDVLE (m/z 1170.519, charge 2+). (B) CID-MS2 of peptide DTLCIGYHANDSTDVTVDVLE
Figure 3.7. MS and tandem MS analysis of Tryptic peptide containing the residue 28. (A) precursor ion mass spectrum of peptide DTLCIGYHANDSTDTVDTVLEK (m/z 1234.572, charge 2+). (B) CID-MS2 of peptide DTLCIGYHANDSTDTVDTVLEK
Figure 3.8. LC/MS analysis of glycopeptides containing N40, N293 and N304. The elution time window for glycopeptides containing the sites N40 N293 and N304 is labeled with blue, red and green squares separately.
Figure 3.9. LC/MS analysis of glycopeptides containing N104 and N498. The elution time window for glycopeptides containing the sites N104 and N498 is labeled with red and blue squares separately.
## Trypsin Peptide Mapping Summary

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Sequence</th>
<th>Time</th>
<th>Observed m/z</th>
<th>Charge</th>
<th>Theoretical Mono Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>N40</td>
<td>K53</td>
<td>NVTVTHSVNLLEDK (N40+M3)</td>
<td>28.6</td>
<td>821.396</td>
<td>3</td>
<td>2460.138</td>
</tr>
<tr>
<td>N40</td>
<td>K53</td>
<td>NVTVTHSVNLLEDK (N40+M5)</td>
<td>28.0</td>
<td>929.434</td>
<td>3</td>
<td>2784.244</td>
</tr>
<tr>
<td>N40</td>
<td>K53</td>
<td>NVTVTHSVNLLEDK (N40+M6)</td>
<td>27.9</td>
<td>983.452</td>
<td>3</td>
<td>2946.296</td>
</tr>
<tr>
<td>N40</td>
<td>K53</td>
<td>NVTVTHSVNLLEDK (N40+M7)</td>
<td>27.7</td>
<td>1037.470</td>
<td>3</td>
<td>3108.244</td>
</tr>
<tr>
<td>N40</td>
<td>K53</td>
<td>NVTVTHSVNLLEDK (N40+M8)</td>
<td>27.7</td>
<td>1091.487</td>
<td>3</td>
<td>3270.402</td>
</tr>
<tr>
<td>G63</td>
<td>K71</td>
<td>GVAPLHLGK</td>
<td>22.6</td>
<td>446.280</td>
<td>2</td>
<td>890.534</td>
</tr>
<tr>
<td>E120</td>
<td>R130</td>
<td>EQLSSVSSFER</td>
<td>28.9</td>
<td>634.817</td>
<td>2</td>
<td>1267.605</td>
</tr>
<tr>
<td>F131</td>
<td>K136</td>
<td>FEIFPK</td>
<td>34.4</td>
<td>780.437</td>
<td>1</td>
<td>779.422</td>
</tr>
<tr>
<td>T137</td>
<td>K147</td>
<td>TSSWPNHDSNK</td>
<td>14.5</td>
<td>636.792</td>
<td>2</td>
<td>1271.553</td>
</tr>
<tr>
<td>G148</td>
<td>K159</td>
<td>GVTAACPHAGAK</td>
<td>13.7</td>
<td>570.292</td>
<td>2</td>
<td>1140.584</td>
</tr>
<tr>
<td>S160</td>
<td>K163</td>
<td>SFYK</td>
<td>15.0</td>
<td>543.278</td>
<td>1</td>
<td>543.269</td>
</tr>
<tr>
<td>N164</td>
<td>K170</td>
<td>NIWLVK</td>
<td>41.5</td>
<td>443.283</td>
<td>1</td>
<td>443.269</td>
</tr>
<tr>
<td>S181</td>
<td>K186</td>
<td>SYINDK</td>
<td>15.5</td>
<td>739.370</td>
<td>1</td>
<td>738.355</td>
</tr>
<tr>
<td>E189</td>
<td>R222</td>
<td>EVLVLWGIHPSTSAQQSLQYQNDYRVFVGSSR</td>
<td>51.9</td>
<td>952.734</td>
<td>4</td>
<td>3803.854</td>
</tr>
<tr>
<td>F227</td>
<td>K236</td>
<td>FKPEIAIRPK</td>
<td>27.5</td>
<td>400.253</td>
<td>3</td>
<td>1197.723</td>
</tr>
<tr>
<td>M244</td>
<td>K256</td>
<td>MNYYWTLVEPGDK</td>
<td>43.0</td>
<td>808.388</td>
<td>2</td>
<td>1614.739</td>
</tr>
<tr>
<td>I257</td>
<td>R269</td>
<td>ITFEATGNLVPR</td>
<td>38.4</td>
<td>709.403</td>
<td>2</td>
<td>1415.777</td>
</tr>
<tr>
<td>Y270</td>
<td>R276</td>
<td>YAFAMER</td>
<td>28.5</td>
<td>444.213</td>
<td>2</td>
<td>886.401</td>
</tr>
<tr>
<td>N277</td>
<td>K300</td>
<td>NAGSGIIIDTPVHDCTNCTCQTPK (N293+M3)</td>
<td>27.1</td>
<td>1160.525</td>
<td>3</td>
<td>3477.502</td>
</tr>
<tr>
<td>N277</td>
<td>K300</td>
<td>NAGSGIIIDTPVHDCTNCTCQTPK (N293+M5)</td>
<td>26.7</td>
<td>1268.893</td>
<td>3</td>
<td>3801.608</td>
</tr>
<tr>
<td>N277</td>
<td>K300</td>
<td>NAGSGIIIDTPVHDCTNCTCQTPK (N293+M6)</td>
<td>26.7</td>
<td>1322.912</td>
<td>3</td>
<td>3963.661</td>
</tr>
<tr>
<td>N277</td>
<td>K300</td>
<td>NAGSGIIIDTPVHDCTNCTCQTPK (N293+M7)</td>
<td>26.5</td>
<td>1376.932</td>
<td>3</td>
<td>4125.714</td>
</tr>
<tr>
<td>N277</td>
<td>K300</td>
<td>NAGSGIIIDTPVHDCTNCTCQTPK (N293+M8)</td>
<td>26.5</td>
<td>1430.950</td>
<td>3</td>
<td>4287.766</td>
</tr>
<tr>
<td>G301</td>
<td>K319</td>
<td>GAIYSLPFQNIHPITIGK (N304+M3)</td>
<td>40.3</td>
<td>972.164</td>
<td>3</td>
<td>2912.414</td>
</tr>
<tr>
<td>G301</td>
<td>K319</td>
<td>GAIYSLPFQNIHPITIGK (N304+M5)</td>
<td>39.6</td>
<td>1080.201</td>
<td>3</td>
<td>3236.517</td>
</tr>
<tr>
<td>G301</td>
<td>K319</td>
<td>GAIYSLPFQNIHPITIGK (N304+M6)</td>
<td>39.4</td>
<td>1134.219</td>
<td>3</td>
<td>3398.566</td>
</tr>
<tr>
<td>G301</td>
<td>K319</td>
<td>GAIYSLPFQNIHPITIGK (N304+M7)</td>
<td>39.3</td>
<td>1188.237</td>
<td>3</td>
<td>3560.623</td>
</tr>
<tr>
<td>G301</td>
<td>K319</td>
<td>GAIYSLPFQNIHPITIGK (N304+M8)</td>
<td>39.1</td>
<td>1242.588</td>
<td>3</td>
<td>3722.679</td>
</tr>
<tr>
<td>G301</td>
<td>K319</td>
<td>GAIYSLPFQNIHPITIGK (N304+A1G0)</td>
<td>39.1</td>
<td>1296.604</td>
<td>3</td>
<td>3884.719</td>
</tr>
<tr>
<td>G301</td>
<td>K319</td>
<td>GAIYSLPFQNIHPITIGK (N304+A1G0M4)</td>
<td>40.5</td>
<td>1093.874</td>
<td>3</td>
<td>3277.534</td>
</tr>
<tr>
<td>G301</td>
<td>K319</td>
<td>GAIYSLPFQNIHPITIGK (N304+A1G0M5)</td>
<td>39.3</td>
<td>1148.228</td>
<td>3</td>
<td>3439.592</td>
</tr>
<tr>
<td>L331</td>
<td>R336</td>
<td>LATGLR</td>
<td>17.6</td>
<td>630.400</td>
<td>1</td>
<td>629.386</td>
</tr>
<tr>
<td>N337</td>
<td>R344</td>
<td>NIPSQSR</td>
<td>22.3</td>
<td>457.762</td>
<td>2</td>
<td>913.498</td>
</tr>
<tr>
<td>S384</td>
<td>K395</td>
<td>STQNAIDEITNK</td>
<td>28.4</td>
<td>667.341</td>
<td>2</td>
<td>1332.652</td>
</tr>
<tr>
<td>Y396</td>
<td>K402</td>
<td>VNSVIEK</td>
<td>15.7</td>
<td>394.734</td>
<td>2</td>
<td>787.444</td>
</tr>
<tr>
<td>M403</td>
<td>K412</td>
<td>MNTQFTAVGK</td>
<td>24.7</td>
<td>548.757</td>
<td>2</td>
<td>1095.499</td>
</tr>
<tr>
<td>Start</td>
<td>End</td>
<td>Sequence</td>
<td>Time</td>
<td>Observed m/z</td>
<td>Charge</td>
<td>Theoretical Mono Mass</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-----------------------------------</td>
<td>------</td>
<td>-----------------</td>
<td>--------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>D18</td>
<td>T30</td>
<td>DTLCIGYHANNST (N28+M5)</td>
<td>26.1</td>
<td>1342.037</td>
<td>2</td>
<td>2681.053</td>
</tr>
<tr>
<td>D18</td>
<td>T30</td>
<td>DTLCIGYHANNST (N28+M7)</td>
<td>25.9</td>
<td>1504.088</td>
<td>2</td>
<td>3008.176</td>
</tr>
<tr>
<td>D18</td>
<td>T30</td>
<td>DTLCIGYHANNST (N28+M15)</td>
<td>26.3</td>
<td>1354.553</td>
<td>2</td>
<td>3029.106</td>
</tr>
<tr>
<td>D103</td>
<td>G110</td>
<td>DNGTCYPG (N104+M6)</td>
<td>18.4</td>
<td>1131.407</td>
<td>2</td>
<td>2262.814</td>
</tr>
<tr>
<td>D103</td>
<td>G110</td>
<td>DNGTCYPG (N104+M5)</td>
<td>18.7</td>
<td>1050.880</td>
<td>2</td>
<td>2098.761</td>
</tr>
<tr>
<td>D103</td>
<td>G110</td>
<td>DNGTCYPG (N104+M3)</td>
<td>19.2</td>
<td>888.328</td>
<td>2</td>
<td>1776.655</td>
</tr>
<tr>
<td>D111</td>
<td>R119</td>
<td>DFIDYEELR</td>
<td>37.3</td>
<td>600.284</td>
<td>2</td>
<td>1198.567</td>
</tr>
<tr>
<td>E120</td>
<td>F131</td>
<td>EQLSSVSFERF</td>
<td>36.3</td>
<td>708.346</td>
<td>2</td>
<td>1416.693</td>
</tr>
<tr>
<td>E132</td>
<td>H143</td>
<td>EIFPKTSSWPNH</td>
<td>29.5</td>
<td>481.575</td>
<td>3</td>
<td>1441.699</td>
</tr>
<tr>
<td>D185</td>
<td>A203</td>
<td>DKGKEVLVLWGIHHPSTSA</td>
<td>32.3</td>
<td>519.533</td>
<td>4</td>
<td>2073.101</td>
</tr>
<tr>
<td>D204</td>
<td>A212</td>
<td>DQQRSLYQNA</td>
<td>25.6</td>
<td>533.745</td>
<td>2</td>
<td>1067.490</td>
</tr>
<tr>
<td>D239</td>
<td>G254</td>
<td>DQEGRMNYYWTLVEPG</td>
<td>43.6</td>
<td>979.943</td>
<td>2</td>
<td>1959.886</td>
</tr>
<tr>
<td>D255</td>
<td>F259</td>
<td>DKITF</td>
<td>27.1</td>
<td>312.174</td>
<td>2</td>
<td>624.348</td>
</tr>
<tr>
<td>E260</td>
<td>M274</td>
<td>EATGNLVVPYAFAM</td>
<td>41.1</td>
<td>819.921</td>
<td>2</td>
<td>1637.842</td>
</tr>
<tr>
<td>E275</td>
<td>S285</td>
<td>ERNAGSGIIS</td>
<td>25.7</td>
<td>558.805</td>
<td>2</td>
<td>1115.610</td>
</tr>
<tr>
<td>D363</td>
<td>A380</td>
<td>DGWGYGHQNEQGSGYAA</td>
<td>27.1</td>
<td>680.948</td>
<td>3</td>
<td>2042.844</td>
</tr>
<tr>
<td>D381</td>
<td>I389</td>
<td>DLKSTQNAI</td>
<td>23.2</td>
<td>495.268</td>
<td>2</td>
<td>990.536</td>
</tr>
<tr>
<td>D390</td>
<td>I400</td>
<td>DEITNKNSVI</td>
<td>34.0</td>
<td>616.331</td>
<td>2</td>
<td>1232.662</td>
</tr>
<tr>
<td>E401</td>
<td>L417</td>
<td>EKMNQFTAVGKEFNLH</td>
<td>33.3</td>
<td>665.666</td>
<td>3</td>
<td>1992.998</td>
</tr>
<tr>
<td>E418</td>
<td>D429</td>
<td>EKRIENLNKVD</td>
<td>15.8</td>
<td>495.951</td>
<td>3</td>
<td>1484.831</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>------------</td>
<td>----</td>
<td>----</td>
<td>------------</td>
<td>----</td>
</tr>
<tr>
<td>D429</td>
<td>L433</td>
<td>DDGFL</td>
<td>35.7</td>
<td>566.246</td>
<td>1</td>
<td>565.238</td>
</tr>
<tr>
<td>D434</td>
<td>A440</td>
<td>DIWTYNA</td>
<td>38.1</td>
<td>882.400</td>
<td>1</td>
<td>881.392</td>
</tr>
<tr>
<td>E441</td>
<td>L452</td>
<td>ELLVLENERTL</td>
<td>43.6</td>
<td>721.416</td>
<td>2</td>
<td>1440.819</td>
</tr>
<tr>
<td>D456</td>
<td>Y463</td>
<td>DSNVKNL</td>
<td>24.6</td>
<td>476.741</td>
<td>2</td>
<td>951.466</td>
</tr>
<tr>
<td>E464</td>
<td>K475</td>
<td>EKVRSQLKNNAK</td>
<td>12.5</td>
<td>472.276</td>
<td>3</td>
<td>1413.805</td>
</tr>
<tr>
<td>E476</td>
<td>F482</td>
<td>EIGNGCF</td>
<td>29.8</td>
<td>796.329</td>
<td>1</td>
<td>795.322</td>
</tr>
<tr>
<td>E483</td>
<td>C488</td>
<td>EFYHKC</td>
<td>15.3</td>
<td>442.193</td>
<td>2</td>
<td>882.369</td>
</tr>
<tr>
<td>E494</td>
<td>Y501</td>
<td>ESVKNGTY (N498+M7)</td>
<td>14.9</td>
<td>1219.986</td>
<td>2</td>
<td>2436.953</td>
</tr>
<tr>
<td>E494</td>
<td>Y501</td>
<td>ESVKNGTY (N498+M6)</td>
<td>15.1</td>
<td>1138.959</td>
<td>2</td>
<td>2274.900</td>
</tr>
<tr>
<td>E494</td>
<td>Y501</td>
<td>ESVKNGTY (N498+M5)</td>
<td>15.4</td>
<td>1057.433</td>
<td>2</td>
<td>2112.847</td>
</tr>
<tr>
<td>E494</td>
<td>Y501</td>
<td>ESVKNGTY (N498+M3)</td>
<td>15.8</td>
<td>895.379</td>
<td>2</td>
<td>1788.741</td>
</tr>
<tr>
<td>E494</td>
<td>Y501</td>
<td>ESVKNGTY (N498+A1G0)</td>
<td>15.7</td>
<td>996.919</td>
<td>2</td>
<td>1991.821</td>
</tr>
<tr>
<td>E494</td>
<td>Y501</td>
<td>ESVKNGTY (N498+A1G0F)</td>
<td>15.9</td>
<td>1070.448</td>
<td>2</td>
<td>2137.879</td>
</tr>
<tr>
<td>E494</td>
<td>Y501</td>
<td>ESVKNGTY (N498+A2G0F)</td>
<td>15.9</td>
<td>1171.989</td>
<td>2</td>
<td>2340.958</td>
</tr>
<tr>
<td>D502</td>
<td>I517</td>
<td>DYPKYSEEAKLNREEI</td>
<td>27.8</td>
<td>662.328</td>
<td>3</td>
<td>1982.959</td>
</tr>
</tbody>
</table>

109
Chapter 4: An Approach to Site-Specific Characterization and Quantitation of Polyethylene glycol (PEG) Derivatives of Polypeptides

Manuscript in preparation for submission to *molecular pharmaceutics*

Title: An Approach to Site-Specific Characterization and Quantitation of Polyethylene glycol (PEG) Derivatives of Polypeptides

Author: Yan-Jun Liu, Shiaw-Lin Wu, William S. Hancock *
4.1 Abstract

The characterization of proteins and peptides with a polyethylene glycol (PEG) modification is a challenge due to the large mass and the heterogeneity of the PEG molecule as well as the potential multiple attachment sites to the polypeptide sequence. Typical electrospray mass spectrometry (ESI-MS) spectra of PEGylated proteins or peptide therapeutics are extremely complex and hard to deconvolute. In our study, a state-of-the-art analytical procedure was developed for analyzing a PEGylated peptide that is 8 residues long and conjugated with 20 kDa linear PEG. Multiple enzyme digestion, followed by HPLC-tandem mass spectrometry was applied to identify the PEGylation. After comparing the result between the un-PEGylated peptide and PEGylated counterpart, the exact location of PEGylation sites at the peptide was confirmed as well as the stoichiometry of PEGylation. Using this approach, we expect that the locations of PEGylation sites in recombinant proteins could be pinpointed as well.

4.2 Introduction

Since human insulin (Humulin®) approved by FDA in 1982, more and more proteins and peptides have marketed in the USA and/or European Union as efficient drugs to treat various pathophysiological states [1]. However, some biopharmaceuticals such as small proteins or polypeptide therapeutics suffer from problems including low solubility, short circulating half-life, physicochemical instability and immunogenicity [2]. Several strategies, including peptide and protein PEGylation, arose to overcome these problems, improving the pharmacokinetic and pharmacodynamic properties of biopharmaceuticals [3, 4]. PEGylation is a term used for the procedure of attaching polyethylene glycol (PEG) polymer chains to molecules. Properly linking the PEG with protein or peptide therapeutics can increase the solubility of the drug and decrease the renal clearance rate or the immunogenicity [5, 6].
The early PEGylated therapeutic proteins were prepared by non-specific PEGylation via reactions involving the side-chain of lysines residues and the protein N-terminus. However, this method leads to formation of a heterogeneous mixture of PEGylated products and each conjugate may have a unique activity and stability properties. Site-Specific PEGylation creates a more homogeneous product and decreases the influence of the PEG moiety on the biological activity of the protein [7]. The common reactive groups used for PEGylation is the thiol, alpha amino group, epsilon amino group, carboxylate and hydroxylate[8]. However, a thiol group rarely exists in proteins as a free species and the carboxylic acid group is not easily activated. Thus, amino-terminus or epsilon-amino groups of lysine are the typical site for PEG linkage due to their reactivity and the mild reaction conditions [9, 10]. The reactive functional group of the PEG reagent such as an aldehyde moiety is largely selective for the α-amine of the protein N-terminus because of the relatively low pKa compared with the lysine side chain [11-14]. Site-specific conjugations such as a N-terminally PEGylated human recombinant granulocyte colony-stimulating factor are commercially available [7].

Developing state of the art analytical methods to monitor the PEGylation product is necessary to improve the product quality and decease any therapeutic side effects. However, characterization of proteins and peptides with polyethylene glycol (PEG) modifications is a challenge, due to the large mass of the PEG molecule (such as the 20 kDa PEG polymer) and the heterogeneity of the PEG (polydispersity). Conventional ESI-MS analysis of a PEGylated product is complicated by a complex overlapping set of different charge states of the peptide/protein combined with the polydispersity of the PEG derivative [15, 16]. For a high mass resolution spectrometer like an Orbitrap, the PEG heterogeneity makes the charge state pattern of the analyzed product outside the m/z range for the mass analyzer. It has also been reported that the large size PEG chain may
prevent the PEGylated peptides from traversing the C-trap in an Orbitrap mass analyzer [17]. The addition of amines (such as triethylamine) is beneficial for reducing the charge of PEGylated proteins and make the charge state pattern of the product shift into the m/z range of an Orbitrap mass analyzer, to enable intact mass analysis of a PEGylated protein [18, 19]. Matrix-assisted laser desorption/ionization coupled with time of flight mass analyzer (MALDI-TOF) is an alternative easy and simple analysis technique since it typically generates single charged molecule and uses the high mass range of the TOF mass spectrometer [20-22]. However, MALDI-TOF analysis is not quantitative and intact mass analysis cannot provide the PEGylation site information of the product.

In our study, MALDI-TOF MS was utilized for intact molecular weight determination. Since PEGylated peptides were poorly detected in the LC-MS chromatogram we used an offline LC-UV method combined with MALDI-TOF MS to locate the PEGylated peptide. To determine the PEGylation site, a multiple enzyme approach was used to release the un-PEGylated peptide sequence for determination of the PEGylation site and the enzymatic digests were further analyzed by LC-MS/MS. After determination of the PEGylation site, the stoichiometry of PEGylation was further determined by the analysis of mixtures of PEGylated peptide and non-PEGylated peptide standards.

4.3 Material and methods

4.3.1 Reagents and materials

The standards used in analysis are PEGylated KXLXQAQX, PEGylated MKXLXQAQXK, un-PEGylated KXLXQAQX and un-PEGylated MKXLXQAQXK (X represents the amino acid and the amino acid is not K or R). The two PEGylated products are dissolved in buffer containing 100
mM potassium phosphate, 0.005% (v/v) Polysorbate 80, 4 mM trans-cinnamic acid and 2% (v/v) glycerol. Trypsin (Sequencing grade) was bought from Promega (Madison, WI). Pepsin (Sequence grade) α-Cyano-4-hydroxycinnamic acid (CHCA), sinapinic acid (SA) and formic acid were purchased from Sigma-Aldrich (St. Louis, MO). Water, LC-MS grade was purchased from J.T. Baker (Philipsburg, NJ). HPLC grade acetonitrile was purchased from Thermo Fisher Scientific (Fairlawn, NJ).

4.3.2 Direct-infusion and ESI-LTQ MS analysis

The PEGylated KXLXQAQX was analyzed with direct injection into the Linear Trap Quadrupole mass spectrometer interfaced with an ESI source. A flow of 5 μl/min was provided by a 250 μL syringe pump integrated with the MS (Hamilton, Reno, NV, USA). The following parameters were applied: sheath gas flow 10 arbitrary units, auxiliary gas flow 8 arbitrary units, a spray voltage of 3.5 kV, capillary temperature of 275 °C, tube lens 80 V, and mass range of m/z 400 to 4000.

4.3.3 MALDI-TOF analysis

10 mg/ml CHCA solution is prepared in 50% acetonitrile (ACN) in water and a saturated solution of SA is prepared in 30% ACN in water. The PEGylated KXLXQAQX (1 mg/ml) was diluted by HPLC water to form a 10 pmol/μl sample solution. The sample solution and the Matrix solution (CHCA, SA) were mixed at 1:1 ratio (v:v). 1 μL of the mixture was loaded onto the target plate to form a thin-layer after the solvent evaporation. A Bruker Microflex MALDI-TOF mass spectrometer equipped with a 20 Hz standard nitrogen laser with 337 nm wavelength was used for the MALDI spectra acquisition. It was operated in a linear or reflectron mode using the standard instrument parameters. The instrument was calibrated with a protein calibration standard I and II
(Bruker Daltonics, Germany) for m/z of 8000-45000 and 1000 laser shots were acquired for each spectrum. The MALDI spectra were displayed using the software FlexAnalysis 3.3.

4.3.4 Enzyme digestion

Samples containing 100 ug PEGylated peptide or unmodified peptide were separately digested by trypsin or pepsin using different protocols. **Trypsin digestion.** The PEGylated peptide was transferred to a 3 kDa MWCO filter and buffer exchanged with the digestion buffer (100 mM ammonium bicarbonate, PH 7.8). Add trypsin (enzyme weight to substrate weight = 1:50) to the solution and incubate at 37 ℃ for overnight. 5% formic acid was used to stop the reaction. **Pepsin digestion.** PEGylated peptide is buffer exchanged with 10 mM HCl by a 3 kDa MWCO filter. Add pepsin (enzyme weight to substrate weight = 1:10) to the solution and incubate at 37 ℃ for 30 min. The peptide standard is used with the same digestion protocol but without the buffer exchange process.

4.3.5 LC-MS analysis

Peptides were analyzed by an Agilent 1200 HPLC pump and a Phenomenex Jupiter® Proteo HPLC column (4.0 μm particle size, 2.1mm i.d. × 250 mm) coupled online with a 6520-quadrupole time-of-flight (QTOF) mass spectrometer (Agilent Technologies). 20 μg of the enzyme-digested samples was used for each LC-MS analysis. Mobile phase A was 0.1% formic acid in water, and mobile phase B 0.1% formic acid in acetonitrile. The separation gradient started from a linear increase from 2% B to 5% in 2 min; increased to 80% B in 37 min and then isocratic at 85% B for 5 min. The flow rate was maintained at 200 μL/min. The Q-TOF mass spectrometer was operated in the positive-ion mode with a capillary voltage of 4 KV, gas temperature of 325 ℃, drying gas
at 11 L/min, nebulizer at 45 psi, and m/z scan range of 350-1800 with an acquisition rate of 7 spectra/sec.

4.3.6 **LC-UV sample collection**

The PEGylated KXLXQAQX were chromatographed with an Agilent 1200 HPLC pump and a Vydac C4 HPLC column (5 μm particle size, 4.6 mm i.d. × 250 mm) coupled online with a UV detector (G1315D photodiode array) and 80 μg of the enzyme-digested samples was used per each LC-UV analysis. Mobile phase A was 0.1% formic acid in water, and mobile phase B was 0.1% formic acid in acetonitrile. The separation gradient started from a linear increase from 2% B to 5% in 2 min; increased to 85% B in 29 min and isocratic at 85% B for 5 min. The flow rate was maintained at 500 μL/min, all the peaks were collect and used for MALDI-TOF analysis. The instrument parameter is the same as above.

4.4 **Results and Discussion**

4.4.1 **Intact analysis of the PEGylated KXLXQAQX.**

Polyethylene glycol (PEG) polymer consists of repeat units (monomers) of ethylene oxide. The synthetic oligomer is polydisperse with a distribution of chain lengths and thus a polydisperse molecular mass. ESI-MS typically generates multiple charged ions and the heterogeneity of PEG combined with distribution of different charge states increase complexity of the spectra and difficulty of deconvolution of the signal. In our study, we used a synthetic PEGylated peptide, KXLXQAQX linked with an approximate 20 kDa PEG to develop an analytical method to characterized the PEGylation site and amount of modification. Figure 4.1 shows the mass spectra obtained by direct infusion of PEGylated KXLXQAQX (20 kDa, molecular weight, Mₐ 16,000-24,000) into a linear ion trap with electrospray ionization.
In contrast to ESI, MALDI ionization typically generates a singly charged ion. Since the test sample we used is a small peptide attached with a large PEG chain, we evaluated the TOF signal for two different matrices: α-Cyano-4-hydroxycinnamic acid (CHCA), commonly used for the analysis of the peptides or small proteins (< 10 kDa) and 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid, SA) which is used for analysis of larger proteins (10–150 kDa)). The test PEG-peptide (0.1 mg/ml) was mixed with the two different matrixes and the result of the MALDI-TOF analysis is displayed in Figure 4.2. By using the matrix CHCA, we observed a peak centered around 21 kDa (mass range between 19 kDa to 23 kDa), however, we did not obtain any corresponding PEGylated peptide signal by using the matrix SA. Based on this result, CHCA was selected for intact mass analysis of the PEGylated peptide.
Figure 4.2. MALDI-TOF intact analysis of the PEGylated peptide using different matrices. The red line is the MALDI-TOF signal obtained by using matrix sinapinic acid, SA. The blue line is the signal obtained by using matrix α-cyano-4-hydroxycinnamic acid (CHCA). The TOF is used in the linear mode of operation.

Unlike its unPEGylated counterpart, the PEGylated product always displays a polydispersity with mass differences of 44 Da (monomer unit -O−CH₂−CH₂−) between adjacent molecular ions. Figure 4.3 displays the spectrum of PEGylated KXLXQAQX operated by MALDI-TOF in reflectron mode. The observed mass of 21607.3 Da corresponds to the mass of peptide KXLXQAQX with 471 repeat oxyethylene units plus a sodium ion adduct. The MNa⁺ form is commonly observed in the MALDI analysis of polymers such as polyethylene glycols and polymethylmethacrylate. The mass difference between 21563.4 Da and 21607.3 Da corresponds to the PEG monomer. Number average weight (M_N), weight-average molecular weight (M_W) and
polydispersity (PD) are parameters typically used to characterize the polymer distribution [23, 24].

The formula is listed below:

\[
M_N = \sum \frac{M_i N_i}{N_i} \quad (1)
\]

\[
M_W = \sum \frac{(M_i)^2 N_i}{M_i N_i} \quad (2)
\]

\[
PD = \frac{M_W}{M_N} \quad (3)
\]

In these equations, \(M_i\) and \(N_i\) refer to the mass and the intensity observed with PEG oligomer respectively and MALDI mass spectroscopy is the best strategy to evaluate the PD, \(M_N\) and \(M_W\) values. Based on the spectrum shown in Figure 4.3, formulas 1-3 and the following parameters: \(M_N = 21489.3\), \(M_W = 21504.1\) and the polydispersity (PD) of 20 kDa PEG is 1.001 (data for calculation are listed in Table 4.1) and thus we determined that only one PEG unit (not multiple PEGs) was attached to the peptide. To further investigate the actual site of PEGylation, we then used the multiple enzyme approach to release the un-PEGylated peptide sequence for determination.
Figure 4.3. MALDI TOF intact mass analysis of the PEGylated peptide using the reflectron mode. The mass range from 21450 Da to 21800 Da is zoomed in for a clearer view. The mass difference is labeled in red.

4.4.2 LC-UV-MALDI-TOF off line analysis of the PEGylated KTXSQAQS.

A typical LC-MS analysis of a PEGylated peptide is very insensitive and hard to deconvolute. Nevertheless, the absorbance of the peptide moiety could be readily observed in the corresponding LC-UV map. Thus, a subtractive offline HPLC-UV analysis coupled with a MALDI TOF approach was developed to characterize PEGylated KXLXQAQX. The sample is dissolved in the buffer containing trans-cinnamic acid (t-CA) (see Methods section) and separated by HPLC. Analysis of the resulting chromatograms at 214 and 280 nm followed by peak collection and MALDI TOF analysis demonstrated that the peaks with retention times of 4.3, 19.1 and 21.5 min corresponds to the buffer, cinnamic acid and PEG-peptide respectively (see Figure 4.4). Each fraction was
collected based on the UV spectrum and further analyzed by MALDI-TOF MS. We saw the PEGylated KXLXQAQX signal in the fraction 3 (Figure 4.5). Thus LC-UV offline analysis combined with MALDI-TOF MS successfully located the PEGylated peptide and this strategy gives us direct information about the PEGylated peptide.

Figure 4.4. HPLC-UV analysis of PEGylated KXLXQAQX on a Vydac C4 column. (A) LC-UV spectra of the sample buffer at 214 and 280 nm. The buffer contains: 100 mM potassium phosphate, 0.005% (v/v) Polysorbate 80, 4 mM trans-cinnamic acid and 2% (v/v) glycerol. (B) LC-UV spectra of PEG-KXLXQAQX at 214 and 280 nm. The peak labeled 1 (4.3 min) corresponds to the buffer gradient (mobile phase A was 0.1% formic acid in water, and mobile phase B 0.1% formic acid in acetonitrile) peaks; 2 (19.1 min) is trans-cinnamic acid peaks and 3 is PEGylated-KXLXQAQX peak.
Figure 4.5. MALDI-TOF analysis of fraction 3 (PEGylated KXLXQAQX) in the LC-UV separation. Peaks in the HPLC-UV analysis were collected for MALDI-TOF detection. Only the fraction 3 (21.5 min) gives the signal of PEGylated KXLXQAQX.

4.4.3 PEGylation site determination for PEG-KXLXQAQX.

The determination uses the specificity of trypsin for C-terminal cleavage at lysine (K) and arginine (R) residues. Based on the PEGylated peptide sequence (PEG-KXLXQAQX), if the PEG group is attached to the N-terminus, MS analysis should show the mass of the ion XIXQAQX generated by the trypsin digestion. However, if the PEG is linked to the side chain of lysine, the trypsin digestion will be blocked by the bulky 20 kDa PEG. The theoretical monoisotopic mass of XIXQAQX, charge 1+ should be 734.367 and in fact the peptide XIXQAQX was identified in
both the PEGylated peptide and its unPEGylated counterpart after the trypsin digestion. While we observed a minor shift in the retention times observed for peptide XLXQAQX in the two tryptic digests, the peptide was identified unambiguously by the monoisotopic mass and CID tandem mass spectra. The extracted ion chromatograms, precursor ion mass spectra and CID-MS-MS spectra of the ion XLXQAQX are listed in the Figure 4.6. Therefore, the results of the trypsin digest analysis indicate that the PEG chain was attached at the N-terminus of the peptide.
Figure 4.6. LC tandem MS analysis of trypsin digests of PEGylated KXLXQAQX and its unPEGylated counterpart. In this figure, the first panel is the data of PEGylated KXLXQAQX digests while the second panel is the data of KXLXQAQX digests. (A) Extracted ion chromatogram of peptide XLXQAQX, m/z 734.363, charge 1+. (Theoretical peptide monoisotopic mass was calculated by MS-product http://prospector.ucsf.edu/prospector/cgibin/msform.cgi?form=msproduct). We extracted this mass in the experimental data (set the mass accuracy to 10 ppm) and compared with the experimental monoisotopic mass of XLXQAQX, mass accuracy is approximately 5 ppm.) (B) Isotopic mass distribution of the ion 734.363. (C) CID-MS/MS (MS2) spectrum of ion 734.363 (charge 1+).

In order to confirm the PEGylation site, another enzyme with broad specificity, pepsin, was utilized to digest the PEGylated peptide. With pepsin digestion of KXLXQAQX we observed
peptide sequence XQAQX (m/z 520.232, charge 1+) in the digest. The same monoisotopic mass was observed with a pepsin digest of the PEGylated peptide (Figure 4.7A). The identical MS2 data further supported the identity of XQAQX in both PEGylated and unPEGylated peptide digest samples (Figure 4.7B). The pepsin digestion result demonstrated that the PEG group is linked in the KTL sequence. Therefore, based on the peptide sequences released by trypsin and pepsin digestions, we draw the conclusion that the PEGylation site is located at the N-terminus.
Figure 4.7. Mass spectra of pepsin digests of PEGylated KXLXQAQX and its unPEGylated counterpart. In this figure, the first panel is the data of PEGylated KXLXQAQX digests while the second panel is the data of KXLXQAQX digests. (A) Isotopic mass distribution of the precursor ion 502.234, charge 1+ (peptide sequence XQAQX). (B) CID-MS/MS (MS2) spectrum of ion 502.234 (charge 1+).

To test the generality of our method, we applied the multienzyme strategy to another synthetic PEGylated peptide, PEG-MKXLXQAQXK. The initial analysis of this PEGylated standard by MALDI-TOF in the reflectron mode showed that there was only one PEG molecule attached on the peptide (Figure 4.11). Based on the peptide sequence, we chose trypsin and pepsin to do the digestion and generate fragments suitable for analysis. Trypsin digestion of the peptide MKXLXQAQXK generated the fragment ion XLXQAQXK, 431.737, charge 2+. However, this peptide was not found in the trypsin digest of PEGylated MKXLXQAQXK (Figure 4.8). The same pattern was observed on pepsin digestion of the PEGylated MKXLXQAQXK and its unPEGylated counterpart (Figure 4.12). By comparing the results between the PEGylated KXLXQAQX and
PEGylated MKXLXQAQXK, we deduced the PEG chain is attached on the lysine side chain of the MKXLXQAQXK.

![Graph showing LC-MS analysis of trypsin digests of PEGylated MKXLXQAQXK and its unPEGylated counterpart.](image)

**Figure 4.8.** LC-MS analysis of trypsin digests of PEGylated MKXLXQAQXK and its unPEGylated counterpart. The first panel is the Extracted Ion Chromatogram (XIC) of peptide XLXQAQXK (m/z 431.373, charge 2+) from the peptide MKXLXQAQXK digests and the second panel is the XIC of XLXQAQXK from the PEGylated MKXLXQAQXK digests. No peptide signal was found in the second panel.

### 4.4.4 Quantitation analysis of the PEGylated KXLXQAQX.

In our preliminary study, we injected PEGylated KXLXQAQX and KXLXQAQX into the LC/MS system and did not found a peptide mass in the PEGylated KXLXQAQX sample (Figure 4.13). Based on our earlier characterization study, the PEGylated peptide corresponds to completely modified N-terminal PEGylated product. In order to develop an analytical method for quantitation of future PEGylated products, we used the fully N-terminal PEGylated product and its unPEGylated counterpart to generate standard curves to test the possibility by using mass
spectrometry to do the quantitation. We spiked different amount of standard peptides into the 100% PEGylated product to form 0%, 10%, 15%, 30%, 50% or 75% (w:w) peptide in the resulting PEGylated product mixtures and the samples were analyzed by LC/MS. The peak area and the unPEGylated peptide percentage were used to generate the plot shown in Figure 4.9 and the experiment was run in duplicate (Figure 4.14). Both data used for plotting were listed in Table 4.2. The synthetic peptide standard was mixed with the PEGylated counterpart to form a 30%, 50%, 75% 100% PEGylated mixture, digested by trypsin and investigated by LC-MS analysis. The extracted ion chromatograms of the tryptic digest XLXQAQX were listed in the Figure 4.10A. Peak area and the unPEGylated peptide percentage are used to generate the standard curve (Figure 4.10B). The results of the replicate experiment are shown in Figure 4.15 (data were listed in Table 4.3) and both studies show good linearity for the measurement (for the standard addition approach and for trypsin digestion), the $R^2$ value is 0.98, 0.95 (replicates) and 0.98, 0.99 respectively. Based on these results we propose that this approach can be used for the quantitation of PEGylation of polypeptides in the future.

![Image of Figure 4.9](image.png)

**Figure 4.9.** Quantitative analysis of PEGylated KXLXQAQX by standard addition.
Figure 4.10. Quantitative analysis of PEGylated KXLXQAQX. (A) Extracted ion chromatogram of peptide XLXQAQX with different percentage of PEGylation. (B) Linear curve generated by the digests of standard peptides with PEGylated peptide mixtures.
4.5 Conclusion

In this study, a multiple enzyme digestion strategy followed by UPLC-tandem mass spectrometry was applied to locate the site of modification of a PEGylated peptide. In the initial intact molecular weight determination, only one PEG chain was found in the model PEGylated peptide KXLXQAQX. Then, trypsin and pepsin were used to digest the un-PEGylated moiety of the peptide, and further LC MS/MS analysis located the PEGylation site at the N-terminus of the modified peptide. Another PEGylated product, PEGylated MKXLXQAQXK, was used to evaluate the generality of the method. After comparison of the analytical results for PEGylated KXLXQAQX and PEGylated MKXLXQAQXK, the PEGylation site of the MKXLXQAQXK was deduced as the side chain of lysine. The stoichiometry of PEGylation at the N-terminus was then determined by the use of mixtures of the PEGylated peptide and non-PEGylated peptide standards. The method developed in this study demonstrated that one could determine both the site of modification as well as the stoichiometry and will have applicability to the analysis of larger PEGylated polypeptides or proteins in future studies.

4.6 Supplementary Data

Table 4.1. The mass and the intensity of PEGylated KXLXQAQX used for calculation of \( \text{Mn, Mw and PD} \).

<table>
<thead>
<tr>
<th>( i )</th>
<th>Average Mass (Na+)</th>
<th>Intensity</th>
<th>( \text{M}i^*\text{N}i )</th>
<th>( (\text{M})_i^*\text{N}i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>444</td>
<td>20411.1</td>
<td>605</td>
<td>1.23E+07</td>
<td>2.52E+11</td>
</tr>
<tr>
<td>445</td>
<td>20452.5</td>
<td>655</td>
<td>1.34E+07</td>
<td>2.74E+11</td>
</tr>
<tr>
<td>446</td>
<td>20496.3</td>
<td>779</td>
<td>1.60E+07</td>
<td>3.27E+11</td>
</tr>
<tr>
<td>447</td>
<td>20548.1</td>
<td>827</td>
<td>1.70E+07</td>
<td>3.49E+11</td>
</tr>
<tr>
<td>448</td>
<td>20580.0</td>
<td>624</td>
<td>1.28E+07</td>
<td>2.64E+11</td>
</tr>
<tr>
<td>449</td>
<td>20633.1</td>
<td>655</td>
<td>1.35E+07</td>
<td>2.79E+11</td>
</tr>
<tr>
<td>450</td>
<td>20674.2</td>
<td>694</td>
<td>1.43E+07</td>
<td>2.97E+11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>---------</td>
<td>-----</td>
<td>-----</td>
<td>-----------</td>
</tr>
<tr>
<td>451</td>
<td>20710.7</td>
<td>701</td>
<td>1.45E+07</td>
<td>3.01E+11</td>
</tr>
<tr>
<td>452</td>
<td>20764.6</td>
<td>827</td>
<td>1.72E+07</td>
<td>3.57E+11</td>
</tr>
<tr>
<td>453</td>
<td>20808.1</td>
<td>824</td>
<td>1.71E+07</td>
<td>3.57E+11</td>
</tr>
<tr>
<td>454</td>
<td>20854.1</td>
<td>703</td>
<td>1.47E+07</td>
<td>3.07E+11</td>
</tr>
<tr>
<td>455</td>
<td>20890.7</td>
<td>792</td>
<td>1.65E+07</td>
<td>3.46E+11</td>
</tr>
<tr>
<td>456</td>
<td>20938.8</td>
<td>825</td>
<td>1.73E+07</td>
<td>3.62E+11</td>
</tr>
<tr>
<td>457</td>
<td>20986.0</td>
<td>1010</td>
<td>2.12E+07</td>
<td>4.45E+11</td>
</tr>
<tr>
<td>458</td>
<td>21026.9</td>
<td>995</td>
<td>2.09E+07</td>
<td>4.40E+11</td>
</tr>
<tr>
<td>459</td>
<td>21071.3</td>
<td>1036</td>
<td>2.18E+07</td>
<td>4.60E+11</td>
</tr>
<tr>
<td>460</td>
<td>21114.3</td>
<td>958</td>
<td>2.02E+07</td>
<td>4.27E+11</td>
</tr>
<tr>
<td>461</td>
<td>21157.7</td>
<td>1136</td>
<td>2.40E+07</td>
<td>5.08E+11</td>
</tr>
<tr>
<td>462</td>
<td>21201.2</td>
<td>1099</td>
<td>2.33E+07</td>
<td>4.94E+11</td>
</tr>
<tr>
<td>463</td>
<td>21246.0</td>
<td>1221</td>
<td>2.59E+07</td>
<td>5.51E+11</td>
</tr>
<tr>
<td>464</td>
<td>21291.0</td>
<td>1389</td>
<td>2.96E+07</td>
<td>6.30E+11</td>
</tr>
<tr>
<td>465</td>
<td>21335.2</td>
<td>1260</td>
<td>2.69E+07</td>
<td>5.73E+11</td>
</tr>
<tr>
<td>466</td>
<td>21378.0</td>
<td>1058</td>
<td>2.26E+07</td>
<td>4.84E+11</td>
</tr>
<tr>
<td>467</td>
<td>21426.7</td>
<td>1051</td>
<td>2.25E+07</td>
<td>4.82E+11</td>
</tr>
<tr>
<td>468</td>
<td>21473.5</td>
<td>1021</td>
<td>2.19E+07</td>
<td>4.71E+11</td>
</tr>
<tr>
<td>469</td>
<td>21519.0</td>
<td>1168</td>
<td>2.51E+07</td>
<td>5.41E+11</td>
</tr>
<tr>
<td>470</td>
<td>21563.4</td>
<td>1154</td>
<td>2.49E+07</td>
<td>5.36E+11</td>
</tr>
<tr>
<td>471</td>
<td>21607.3</td>
<td>1339</td>
<td>2.89E+07</td>
<td>6.25E+11</td>
</tr>
<tr>
<td>472</td>
<td>21651.2</td>
<td>1209</td>
<td>2.62E+07</td>
<td>5.67E+11</td>
</tr>
<tr>
<td>473</td>
<td>21695.2</td>
<td>1139</td>
<td>2.47E+07</td>
<td>5.36E+11</td>
</tr>
<tr>
<td>474</td>
<td>21740.4</td>
<td>1199</td>
<td>2.61E+07</td>
<td>5.67E+11</td>
</tr>
<tr>
<td>475</td>
<td>21784.4</td>
<td>1265</td>
<td>2.76E+07</td>
<td>6.01E+11</td>
</tr>
<tr>
<td>476</td>
<td>21827.6</td>
<td>1478</td>
<td>3.22E+07</td>
<td>7.04E+11</td>
</tr>
<tr>
<td>477</td>
<td>21869.9</td>
<td>1114</td>
<td>2.43E+07</td>
<td>5.32E+11</td>
</tr>
<tr>
<td>478</td>
<td>21911.3</td>
<td>1058</td>
<td>2.32E+07</td>
<td>5.08E+11</td>
</tr>
<tr>
<td>479</td>
<td>21951.1</td>
<td>1088</td>
<td>2.39E+07</td>
<td>5.24E+11</td>
</tr>
<tr>
<td>480</td>
<td>21997.4</td>
<td>1025</td>
<td>2.25E+07</td>
<td>4.96E+11</td>
</tr>
<tr>
<td>481</td>
<td>22041.7</td>
<td>997</td>
<td>2.20E+07</td>
<td>4.84E+11</td>
</tr>
<tr>
<td>482</td>
<td>22080.7</td>
<td>723</td>
<td>1.60E+07</td>
<td>3.53E+11</td>
</tr>
<tr>
<td>483</td>
<td>22128.5</td>
<td>1105</td>
<td>2.44E+07</td>
<td>5.41E+11</td>
</tr>
<tr>
<td>484</td>
<td>22169.3</td>
<td>895</td>
<td>1.99E+07</td>
<td>4.40E+11</td>
</tr>
<tr>
<td>485</td>
<td>22217.7</td>
<td>844</td>
<td>1.87E+07</td>
<td>4.16E+11</td>
</tr>
<tr>
<td>486</td>
<td>22258.3</td>
<td>736</td>
<td>1.64E+07</td>
<td>3.65E+11</td>
</tr>
<tr>
<td>487</td>
<td>22309.7</td>
<td>770</td>
<td>1.72E+07</td>
<td>3.83E+11</td>
</tr>
<tr>
<td>488</td>
<td>22356.8</td>
<td>677</td>
<td>1.51E+07</td>
<td>3.38E+11</td>
</tr>
<tr>
<td>489</td>
<td>22394.4</td>
<td>759</td>
<td>1.70E+07</td>
<td>3.80E+11</td>
</tr>
<tr>
<td>490</td>
<td>22441.6</td>
<td>522</td>
<td>1.17E+07</td>
<td>2.63E+11</td>
</tr>
<tr>
<td>491</td>
<td>22486.9</td>
<td>747</td>
<td>1.68E+07</td>
<td>3.78E+11</td>
</tr>
<tr>
<td>492</td>
<td>22525.4</td>
<td>657</td>
<td>1.48E+07</td>
<td>3.35E+11</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td>46413</td>
<td>9.97E+08</td>
<td>2.14E+13</td>
</tr>
</tbody>
</table>

Note: i represents the number of repeat oxyethylene units. Ni is the value of peak intensity. Number average weight (MN), weight-average molecular weight (MW) and polydispersity (PD) is calculated based on the equation 1-3 which is listed in the paragraph. Mn =21489.3, Mw =21504.1 and the polydispersity (PD) is 1.001

Figure 4.11. MALDI TOF intact analysis of the PEGylated MKXLXQAQXK. The signal range shows that there is only one PEG chain attached on the peptide.
Figure 4.12. LC-MS analysis of pepsin digests of PEGylated MKXLXQAQXK and its unPEGylated counterpart. The first panel is the Extracted Ion Chromatogram (XIC) of peptide MKTLSQ (m/z 354.191.373, charge 2+) from the peptide MKXLXQAQXK digests and the second panel is the XIC of MKTLSQ from the PEGylated MKXLXQAQXK digests. No peptide signal was found in the second panel.

Figure 4.13. XIC of peptide KXLXQAQX m/z 431.737, (charge 2+). A. XIC of mass 431.737 in the synthesis peptide sample. B. XIC of mass 431.737 from the PEGylated peptide product data.
Table 4.2. UnPEGylated peptide percentage and the LC/MS peak area used to generate the standard curve by the standard addition approach

<table>
<thead>
<tr>
<th>UnPEGylated Peptide Percentage</th>
<th>Peak Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>0%</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>10%</td>
<td>7.77E+07</td>
</tr>
<tr>
<td>15%</td>
<td>8.48E+07</td>
</tr>
<tr>
<td>30%</td>
<td>1.64E+08</td>
</tr>
<tr>
<td>50%</td>
<td>2.12E+08</td>
</tr>
<tr>
<td>75%</td>
<td>3.06E+08</td>
</tr>
</tbody>
</table>

Figure 4.14. Quantitative analysis of PEGylated KXLXQAQX by the standard addition approach (duplicate).
Table 4.3. UnPEGylated peptide percentage and the LC/MS peak area used to generate the standard curve by the digests of standard peptides with PEGylated peptide mixtures.

<table>
<thead>
<tr>
<th>UnPEGylated Peptide Percentage</th>
<th>Peak Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>70%</td>
<td>1.41E+07</td>
</tr>
<tr>
<td>50%</td>
<td>1.13E+07</td>
</tr>
<tr>
<td>25%</td>
<td>6.11E+06</td>
</tr>
<tr>
<td>0%</td>
<td>2.12E+06</td>
</tr>
</tbody>
</table>

Figure 4.15. Linear curve generated by the digests (XLXQAQX) of standard peptides with PEGylated peptide mixtures (duplicate).

4.7 Reference


Conclusion

Proteomics is the large-scale identification of an entire set of proteins in a cell line, tissue or organism. The most challenging proteomics study is comparison of two or more physiological states of a biological system. However, these kinds of studies can reveal significant information about cellular and molecular mechanisms of diseases and play a pivotal role in discovery of biomarkers which are essential to disease detection and evaluation of therapeutic strategies.

After the approval of human insulin by FDA, more and more recombinant human proteins have appeared as therapeutic agents. In contrast to small molecule drugs, protein therapeutics are produced by living organisms which can generate complex heterogeneous mixtures causing difficulty in quality control process. For that reason, proteomics strategy extended its application to the characterization of recombinant proteins. Different from large-scale proteomics studies, the biologics characterization of biosimilars focuses on detailed information of a protein including its amino acid (AA) sequence, molecular weight (MW), post-translational modifications (PTMs), and aggregation or degradation products.

LC-MS technology has become a powerful tool in both large-scale protein identification and protein characterization due to its sensitivity, resolution, selectivity and specificity. Significant improvements have been made in separation technology, ionization techniques, instrumentation, and associated software to fulfill the needs of different separation, resolution and detection demands. In this thesis, 1D electrophoresis followed by ultra-performance liquid chromatography coupled with tandem mass spectrometry (UPLC-MS/MS) was utilized to recognize polycystic kidney disease related changes in the glycan composition of glycoproteins. These techniques utilized a multienzyme digestion strategy and when combined with a LC tandem MS procedure
were used to determine the modifications present in a broad range of analytes ranging from a complex glycoprotein to a small peptide modified with the polydisperse PEG moiety.

State-of-the-art analytical methods not only benefit the disease-related study but also contribute to providing high quality biologics and biosimilars. LC-MS technology has become a vital analytical tool in biology, pharmacology and medical science studies.