Comprehensive Glycoproteomics and Glycomics Study of N-Linked Glycans and N-Glycoproteins

Xu Li

Follow this and additional works at: https://scholarworks.gsu.edu/chemistry_diss

Recommended Citation
doi: https://doi.org/10.57709/9002077

This Dissertation is brought to you for free and open access by the Department of Chemistry at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Chemistry Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.
COMPREHENSIVE GLYCOPROTEOMIC AND GLYCOMICS STUDY OF N-LINKED GLYCANS AND N-GLYCOPROTEINS

by

XU LI

Under the Direction of Peng George Wang, PhD

ABSTRACT

N-linked glycosylation is the most common post-translational modification (PTM) of proteins that exist in nature. N-glycosylation and change in cells serve as a criterion to monitor the activity of developmental stages and diseases severity. Currently, there is an increasing application of mass spectrometry on glycoprotein for malicious, chronic or acute diseases, such as cancers, rheumatoid arthritis (RA) or influenza.

In this dissertation, several mass spectrometric assays have been utilized to, quantitatively and qualitatively, characterize protein N-glycosylation at the glycan, glycopeptide and peptide levels. The goals are to identify serum-based RA biomarker (Chapter 2), or to determine possible glycan structures from monoclonal antibody (Chapter 3), or comprehensively to study one influenza glycoprotein, hemagglutinin (Chapter 4).
In Chapter 2, LC-MS/MS with CID as MS 2 is the primary technique that is applied to collect raw data for RA biomarker screening; western blot is the verification method for newfound biomarkers. This mass spectrometry based comparative analysis of N-glycoprotein in RA and healthy patients’ sera reveal 41 potential biomarkers for RA that can be applied in clinical research. Chapter 3 describes another LC-MS/MS based method developed for the structural analysis of N-glycan released from the monoclonal antibody, immunoglobulin G. Higher-energy collision dissociation (HCD) was the surpriour technique utilized to identify glycopeptide fragments. The results show that 19 and 23 N-glycan structures were determined from standard and modified mAb samples respectively by using SimGlycan software, while 38 and 35 glycan structures were recognized by manually mapping respectively. 13 N-glycoforms, out of 26 overlapped glycan structures, were identified with significant alterations by comparing standard sample (sample A) and modified mAb (sample B) utilizing our method. In Chapter 4, we comprehensively studied hemagglutinin by using LC-MS/MS and MALDI from both proteomic perspective and glycomics prospective. After confirmed and verified protein sequence and glycosylation sites, galactose-specific quantitation was performed with exoglycosidase digestion combined HPLC with fluorescence detection. The MALDI-MS/MS based method was utilized to confirm glycan structures.

The results in this dissertation provide insights into the significance of protein glycosylation alterations as RA biomarkers, and these quantitative methods can be reapplied to any other disease biomarkers screening for clinical researchers.

INDEX WORDS: LC-MS/MS, Glycosylation Occupancy, HILIC, Glycoform, Glycoproteomics N-Glycosylation Site, Rheumatoid Arthritis, Monoclonal Antibody, CID, HCD, MALDI-MS/MS
COMPREHENSIVE GLYCOPROTEOMICS AND GLYCOMICS STUDY OF N-LINKED GLYCANS AND N-GLYCOPROTEINS

by

XU LI

Committee Chair: Peng George Wang

Committee: Suazette Reid Mooring

Gangli Wang

Siming Wang

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
August 2016
DEDICATION

I dedicate this to my parents for loving me unconditionally and setting an example to aspire me to achieve. This dissertation is also dedicated to my boyfriend, Andrew, who has been supporting me through the challenges in graduate school and life.
ACKNOWLEDGEMENTS

I am very fortunate to study and perform my graduate work at Georgia State University. Thus, I would like to express my deepest gratitude to my advisor, Dr. Peng George Wang, for giving me a home in his lab and continuous supports these years. I’m very grateful for his insightful guidance and the opportunities he has offered me. His supervisions and fruitful suggestions were immensely helpful in moving my project forward. I could not have imagined a better advisor for my Ph.D. study.

I would also like sincerely to thank my dissertation committee members, Professor Suazette Reid Mooring, Professor Gangli Wang, and Professor Siming Wang, for their contributions to this work. Over the years, each has given me incredible scientific guidance, insightful comments, and encouragements. I am grateful to have such a strong group of intelligent scientists on my committee.

I would like to recognize the current and former colleagues in the Wang group who have all contributed to the progress I have made. Dr. Cheng Ma, Dr. Chun-kai Yang, Dr. Jingyao Qu, Dr. Jing Li and Dr. Wanyi Guan have offered their time and generous help in training and teaching me several techniques. I would like to acknowledge He Zhu, Kuan Jiang, Arya Aloor and Ebstesam Gashash for their help in my projects. Special thanks go to Angie Calderon and Kristina Garner for being great friends when something bothers me in my experiments or life. The Wang lab members are a wonderful group of people who I have enjoyed working with and learning from.

I also would like to thank Dr. Lisa P. Armistead and Dori Neptune from the office of the Associate Provost for Graduate Programs at Georgia State University. For the past one year, they have been offering me superb supports and clearing up any obstacles that stop me pursuing my
degree. I also would like to express my gratitude to the staffs and friends at Chemistry Department. The individuals I have met in graduate school that I consider friends are too many to name. Nikita Burrows and Dr. Jerry Smith have been great listeners when I have any stressful moments on my experiments or in life. Will Lovett, Kedayne King, and Graduate Director Dr. Donald Hamelberg have been there for me when the challenges of graduate school seem too burdensome to overcome. I will never forget the experiences we’ve shared and hope to stay in touch.

At last, I would like to give my deepest gratitude to my family. My parents unconditionally offer me the numerous love, supports, and encouragements in my dissertation work. I would like to thank my boyfriend, Andrew, for being continually supportive and considerate during my graduate education. I would love to thank my cat, Mia, for being my precious kitty.
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................ v

LIST OF TABLES .................................................................................................................. xiv

LIST OF FIGURES ................................................................................................................ xv

CHAPTER 1. GENERAL INTRODUCTION TO GLYCOMICS AND GLYCOPROTEOMICS ........................................................................................................................... 1

CHAPTER 2. COMPARATIVE PROTEOMIC STUDY OF N-GLYCOPROTEIN BIOMARKERS IN RHEUMATOID ARTHRITIS PATIENTS’ SERA ...................................... 3

2.1 Introduction ..................................................................................................................... 3

2.2 Materials and Experiments ............................................................................................ 6

2.2.1 Materials and Chemicals ........................................................................................... 6

2.2.2 Filter-Aided Sample Preparation (FASP) ................................................................. 7

2.2.3 N-Glycopeptides Enrichment and Simplification ..................................................... 8

2.2.4 LC-MS/MS Analysis ................................................................................................. 8

2.2.5 Data Searching and Data Annotation for N-Glycoprotein and Protein Analysis ............................................................................................................................... 9

2.2.6 Verification of Protein Lists with Significant Difference from N-Glycoprotein and Protein Data Analysis – Western Blot Assay (WBA) ................................................. 10

2.3 Results ............................................................................................................................ 11

2.3.1 Label-Free Quantification Strategy for N-Glycoprotein and Protein Sample ................................................................................................................................. 11
2.3.2 Database Searching and Data Comparison of Site-Specific Glycosylation Occupancy between N-Glycoproteins ................................................................. 11

2.3.3 Database Searching and Data Comparison for Proteins Identified with Alteration in Global Perspective ................................................................. 13

2.3.4 Protein Verification by Western Blot Assay (WBA) ............................. 14

2.4 Discussions ....................................................................................... 18

2.5 Concluding Remarks ........................................................................ 23

2.6 Complementary Materials ................................................................. 25

2.6.1 Clinical Characteristics of Patients with RA .................................... 25

2.6.2 N-glycosylation Variations Are Caused by Protein-Specific or Site-Specific .............................................................................................................. 26

2.6.3 Complete Lists of T-Test Results from N-Glycoprotein Comparisons Site-Specific and Globally................................................................. 28

CHAPTER 3. A STRUCTURAL ANALYSIS OF N-GLYCAN RELEASED FROM MONOCLONAL ANTIBODY USING HCD-MS/MS ............................................. 32

3.1 Introduction....................................................................................... 32

3.2 Materials and Experiments ............................................................... 35

3.2.1 Materials and Chemicals ............................................................... 35

3.2.2 Trypsin Digestion of Monoclonal Antibody ................................ 36

3.2.3 Enrichment and Simplification of N-Glycopeptides Using HILIC Tips .... 36

3.2.4 HCD – MS/MS Coupling with Nano-HPLC .................................. 37
3.2.5 Data Annotation of N-Glycan from N-Glycopeptides Using SimGlycan .. 38
3.2.6 Manual Sugar Structural Determination from Data of N-Glycopeptides . 38
3.2.7 Results Interpretation for Dominate Glycan Structure......................... 39
3.3 Results and Discussions ........................................................................ 40
3.3.1 N-Glycan Structural Profiling............................................................. 40
3.3.2 Results Analysis with Data Generated from SA and MI Approaches ...... 40
3.3.3 Initial Results Interpretation from SimGlycan................................... 41
3.3.4 A Comprehensive Explanation of MI Approach’s Procedures............ 43
3.3.5 Discussions and Suggestions in Choosing between MI and SA Methods. 45
3.3.6 The Influences of Different Glycoforms on Monoclonal Antibody During in vivo Distributions........................................................................................................... 47
3.4 Conclusions............................................................................................. 48
3.5 Complementary Materials ....................................................................... 49
3.5.1 The Glycan Structures Identified in Standard (A) and Modified Standard (B) ........................................................................................................................................................................... 49
3.5.2 Raw Data of the Glycan Structures with Significant Differences between A and B ........................................................................................................................................................................... 50

CHAPTER 4. A COMPREHENSIVE STUDY OF MODIFIED HEMAGGLUTININ FROM PROTEOMICS AND GLYCOPROTEOMICS ASPECTS . 52
4.1 Hemagglutinin Proteinaceous Sequence Mapping ................................. 52
4.1.1 Introduction............................................................................................. 52
4.1.2 Materials and Experiments

4.1.2.1 Materials and Chemicals

4.1.2.2 Desalting Procedures

4.1.2.3 Filter-Aided Sample Preparation with Trypsin (FASP-T)

4.1.2.4 In-Solution Digestion with Trypsin (ISD-T)

4.1.2.5 Enzymatic Deglycosylation in Peptide Preparation

4.1.2.6 Nano-LC-MS/MS Peptide-Mass Fingerprinting (PMF)

4.1.2.7 The Confirmation of the HA-1 and -2’s Coverages Using Designated Databases

4.1.3 Results and Discussions

4.1.3.1 The Coverage Determination of HA-1 and HA-2’s Sequences through PMF

4.1.3.2 FASP Technique Gives Better Coverage of Sequences than IS Digestion Method

4.1.3.3 Trypsin is More Capable for HA-1 PMF than HA-2 in Protein Digestion Level

4.2 Multi-Enzyme Digestion to Improve Proteinaceous Coverages for HA-1 and -2

4.2.1 Introduction

4.2.2 Materials and Experiments

4.2.2.1 Materials and Chemicals

4.2.2.2 Filter-Added Sample Preparation with Trypsin & Glu-C (FASP-T&GC)
4.2.2.3 In-Solution Digestion with Trypsin & Glu-C (IS-T&GC) .......... 63

4.2.3 Results and Discussion ................................................................. 64

4.2.3.1 The Comparison of Sequential Coverages between HA-1 and HA-2 under MED ................................................................. 64

4.2.3.2 MED Technique is Superior to Single-Enzyme Digestion Sample Treatment Based on Proteinaceous Coverage ................................ 65

4.3 The Verification of Theoretical Glycosylation Sites in HA and Determination of Glycosylation Occupancy for Each Specific Site .... 67

4.3.1 Introduction ....................................................................................... 67

4.3.2 Materials and Experiments ............................................................... 69

4.3.2.1 Materials and Chemicals ............................................................... 69

4.3.2.2 Enzymatic Deglycosylation in Isotope-Labeled Peptide Preparation 69

4.3.2.3 N-Glycopeptides Enrichment and Enzymatic Deglycosylation in Peptide Preparation ................................................................. 69

4.3.2.4 Data Processing and Determinations of Glycosylation Occupancy for Each Sites ................................................................. 70

4.3.3 Results and Discussion ....................................................................... 71

4.3.3.1 The Verification of Glycosylation Sites and Calculation for Each Glycan Occupancy ................................................................. 71

4.3.3.2 Comparisons between $^{18}$O and HILIC Approaches Relying on the Results from the Identifications of Glycosylation Sites ........ 72

4.3.3.3 The Determination of Occupancy Ratio for Glycosylation Sites in HA-1 and HA-2 ................................................................. 76
4.4 Quantitative Determination of Terminal \( \alpha \)-Galactose and Structural Analysis of Attached Sugar Moieties

4.4.1 Introduction

4.4.2 Materials and Experiments

4.4.2.1 Materials and Chemicals

4.4.2.2 Enzymatic Releasing of Glycans

4.4.2.3 Hydrolysis of Oligosaccharides into Monosaccharides

4.4.2.4 Monosaccharide Quantitation with HPAEC-PAD

4.4.2.5 Conjugation of Glycan to 2-Aminobenzamide (2-AB)

4.4.2.6 Exoglycosidase Digestion on 2-AB Derived Glycans

4.4.2.7 Chromatographic Separation and MALDI Analysis of Glycan Derivatives with 2-AB Labels

4.4.2.8 Enzymatic Activity Confirmation of \( \alpha \)-1-3,6 Galactosidase by Dot Blot

4.4.2.9 Removal of Terminal \( \alpha \)-Galactose with \( \alpha \)-1,3,6 Galactosidase

4.4.3 Results and Discussion

4.4.3.1 Experimental Design Scheme for Hydrolyzed Monosaccharides Quantitation and 2-AB Labeled Sugar Quantitation

4.4.3.2 \( \alpha \)-1,3,6 Galactosidase Activity Verification Using Dot Blot

4.4.3.3 Quantitation of Monosaccharides from Released Glycans with HPAEC-PAD

4.4.3.4 Exoglycosidase Treatment of 2-AB Derived Glycan

4.5 Conclusions
LIST OF TABLES

Table 2.1 The Expression Differences between N-Glycoproteins from Complement System

Table 2.2 Clinical Characteristics of Patients with RA

Table 2.3 Comparison Between N-Glycosylation Sites (NS) Analysis and Protein Expression (PE) Analysis

Table 2.4 Protein Comparison between N-Glycosylation Sites (NS) Analysis and Protein Expression (PE) Analysis

Table 2.5 Result List of T-Test from N-Glycosylation Sites Site-Specific Comparison

Table 2.6 Result List of T-Test from Protein Expression Comparison

Table 3.1 The Brief Analysis of Advantages and Disadvantages between These Two Techniques

Table 3.2 The List of Glycan Structures Identified from A and B

Table 3.3 Raw Data for These 15 Glycoforms that Shown Significant Differences Among Samples

Table 4.1 HA-1 and HA-2 Digestion Methodology Comprehensive Comparison

Table 4.2 Compare HILIC and \(^{18}\)O Isotope Labeled Approaches

Table 4.3 Average Occupancy Ratio Results for Five Potential Glycosylation Sites in HA-1 and HA-2

Table 4.4 Gradient Chart for Monosaccharide Separation on HPAEC-PAC

Table 4.5 Gradient Chart for Derived Sugars’ Separation on HPLC-FL

Table 4.6 The Chart of \(\alpha1-3,6\) Galactosidase Activity

Table 4.7 The List of Sugar Structures Identified from Derived Labeling Technique
LIST OF FIGURES

Figure 2.1 The Scheme to Identify the Difference in N-Glycoproteins and Serum Total Proteins .......................................................... 12

Figure 2.2 The Overlapped Lists between Rheumatoid Arthritis and Normal Volunteer in Both Assays .......................................................... 14

Figure 2.3 A List of Glycosylation Sites Identified with Expression-Alternations............ 16

Figure 2.4 A List of Proteins Identified with Hyper- /Hypo- Expressions ..................... 17

Figure 2.5 WBA Verification Results of CFB (Top) and SAA (Bottom) ....................... 17

Figure 2.6 Example Tandem Mass Spectra Showing Matches to One Peptide in Complement Factor B Protein .............................................. 20

Figure 2.7 Example Tandem Mass Spectra Showing Matches to One Peptide in Serum Amyloid A Protein ................................................................. 21

Figure 3.1 The Workflow Scheme of Profiling Immunoglobulin N-Glycan Glycoforms 39

Figure 3.2 The Mass Values in Daltons Used in Glycan Structure Interpretation .......... 40

Figure 3.3 The Comparison between Two Results from SA and MI .......................... 41

Figure 3.4 Glycan Structures with Significant Variations that Identified from MI....... 42

Figure 3.5 Carbohydrate Structural Comparison between Sample A and B ............... 43

Figure 3.6 An Example of MI Interpretation ...................................................... 45

Figure 4.1 The Workflow of Protein Sequence Coverage Determination .................. 58

Figure 4.2 A Fully Comparisons between FASP and IS Strategies.......................... 59

Figure 4.3 The Sequence Alignments between HA-1 and HA-2 .............................. 62

Figure 4.4 The Workflow Adopted for Multi-Enzyme Digestion .............................. 65

Figure 4.5 Improved Coverage Results for HA-1 and HA-2 After MED Technique...... 65
Figure 4.6 The Workflow of Enzymatic Digestion in Isotopic Labeled Water .................. 71
Figure 4.7 The Workflow of HILIC Enrichment Coupled Enzymatic Deglycosylation .. 72
Figure 4.8 Examples of Tandem Mass Spectra with Peptides included Glycosylation Site

N_32 for HA-1 (Top) and -2 (Bottom) .................................................................................. 74
Figure 4.9 A Scheme of Sugar Hydrolysis Coupled HAPEC-PAD Method .................. 84
Figure 4.10 A Diagram of 2-AB Derived Glycans Analysis with HPLC and MALDI-

MS/MS ........................................................................................................................................ 85
Figure 4.11 The Dot Plot Results of α1-3,6 Galactosidase Activity Determination ...... 86
Figure 4.12 Elution Times for 7-Sugar Standard and Monosaccharides ...................... 88
Figure 4.13 Standard Curve of Sialic Acid .............................................................................. 89
Figure 4.14 Calibration Curve for Neutral and Amino Monosaccharide ...................... 89
Figure 4.15 Spectra of HA-1 and HA-3 Obtained From HPAED-PAD ......................... 90
Figure 4.16 A Scheme that Shown Various Exoglycosidases’ Functions and Aim Sites 91
Figure 4.17 The Scheme of 2-AB Labelling Reaction ........................................................ 93
Figure 4.18 The Spectra from HPLC with a Fluorescent Detector ................................. 96
CHAPTER 1. GENERAL INTRODUCTION TO GLYCOMICS AND GLYCOPROTEOMICS

Glycosylation, as one of the most important post-translational modifications, determines the activities and functions of these carbohydrate-binding proteins, also known as glycoprotein [1]. There are two major types of glycosylation in animal glycosylation: N-Glycosylation, in which sugar is attached to an asparagine residue, and O-Glycosylation. N-glycosylation related study is more advance-developed because one enzyme, called PNGase F, was discovered to release N-glycan completely from their peptide backbone. Mass spectrometry is the dominate technology that is aggressively developing in order to comprehensively understand the structure-function relationship for glycoproteins [2].

Glycome is an entire collection of glycans synthesized by a cell, tissue, or organism under specific conditions of time, space and environment [2]. The glycomics is the comprehensive and systematic analysis of the glycome. To achieve sensitive and definitive glycan analysis, mass spectrometry, especially high-throughput mass spectrometry, is widely utilized to analyze glycan structures in small sample quantities [3]. A tandem mass spectrum offers a list of M/Zs (mass/charge) from the fragments of the carbohydrates. Meantime, the spectrum is capable for database searching by matching each fragment ion. However, the challenge in structural characterization of glycome is to identify the linkage between each monosaccharide. For this purpose, permethylated or isotope-labeled glycan generates a significant quality of fragments in the mass spectrometer as a perfect resource for determining glycan connectivity. Furthermore, an alternative method for requiring structural information is the stepwise removal of terminal monosaccharides through the use of exoglycosidases [4] (Chapter 4.3).
Glycoproteome is an entire collection of glycoproteins that exist in a cell, tissue, or organism under specific conditions of time, space and environment [2]. The field of glycoproteomics is designed to identify the peptides with sugar moiety attached. There is an increasing research interest in studying the changes in the expression of one glycosylation site or glycoprotein globally (Chapter 2). In general, a general design of glycoproteomic experiment involves enzyme digestion of either simple or complex sample, followed by enriching the glycopeptides. Sample are further treated with deglycosylation procedures and analyzed by a high-throughput mass spectrometer. Many proteomic techniques have been adopted in the glycoproteomic analysis, like SILAC (stable isotope labeling by amino acids in cell culture) [5] or $^{18}$O combine with enzymes.
CHAPTER 2. COMPARATIVE PROTEOMIC STUDY OF N-GLYCOPROTEIN BIOMARKERS IN RHEUMATOID ARTHRITIS PATIENTS’ SERA

2.1 Introduction

Rheumatoid arthritis (RA) is a chronic multisystem autoimmune disease, thus causing a systemic inflammation of many tissues and organs, but principally, flexible joints and surrounding tissues [6]. To date, no one has been able to determine the exact cause of this peculiar effect, but there is a theory that RA is most likely resulting from several factors combined, such as genetic factors or environmental triggers [7]. Up to now, the complicated problem is that it is challenging to find a way to stop the manifestation of this disease in the human body because various inducing factors are involved in the disease progression. So far, many classes of biomarkers have been discovered: they can come from clinical, imaging patterns, or biochemical parameters. However, none of them work properly as a monitoring factor for the disease process or as a most valuable biological signal for the response of treatments. According to the reason as mentioned above, the discovery of early stage diagnosis biomarkers is critical to relieving patients from suffering.

The valuableness of these biomarkers is noteworthy; however, the high specific RA biomarker was not included in the American College of Rheumatology (ACR) or the European League against Rheumatism (EULAR) diagnostic criteria until earlier 2000s. The reason is that several high expression proteins in arthritis have a single strong associations with one particular type of arthritis family disease, such as ACPA, related to patients with HLA-DRB1 genetic disorder [8] and C-reactive protein or cytokines (associated with inflammation) [9]. Without long-term experimental trial testing, those so-called biomarkers are not capable of fulfilling the specificity and sensitivity, the main criteria of biomarkers [9-11].
N-linked glycosylation is the most common post-translational modification (PTM) of proteins. In fact, the majority of serum proteins are believed to be glycosylated. N-glycosylation and its related changes serve as a criterion to monitor the activity of developmental stages and diseases severity [12-14]. Thus, an extensive correlation exists between many human diseases and the changes in the glycosylation of accompanying glycoproteins, such as cancers [15-18], cardiovascular diseases and immune deficiency diseases [19-22]. Hence, a thorough inspection of RA serum N-glycosylated proteins might endow the potentiality to reveal hyper- or hypo-expressed proteins associated with RA early diagnosis and pursue targets for selection or response monitoring of therapy [22, 23].

Label-free protein quantification is one of two major quantitative proteomics approaches beside stable isotope labeling techniques [24]. More recently, there is a booming interest in label-free protein quantification applications. Two major label-free quantification strategies, peak area or signal intensity measurement and spectral counting procedure, are widely applied into bottom-up proteomics. Even though label-free approaches offer the lowest accuracy among other quantification techniques, they superiorly have improved quantitative proteome coverage, simplified sample preparation procedures, and increased capability to quantitate multiple proteins in one experiment with relatively small sample quantity, regardless of the protein concentration [24, 25]. Thus, label-free quantitative techniques are more suitable for precious complex samples such as plasma or serum, body fluids, or tissue exactions. At the same time, label-free quantitative methods can offer a global view of understanding high expression proteins as a unique disease condition related change, instead of concentrating on one particular over-expressed protein [24]. Paralleled to label-free quantification, isotope labels are usually introduced into samples by using spiked synthetic peptides as external standards, or into systems metabolically, chemically or
enzymatically as an internal standard [26]. However, the cost of isotope labels for these techniques can be unreasonable, and the requirements, like specific software and training or expertise to process data can cost a fortune. Meanwhile, these stable isotope labeling techniques are not suitable for all sample types, and the capability to analyze multiple isotope-labeled samples in a single trial is constrained by those technologies. Theoretically, stable isotopic labels aim to ingrate with every reactive amino acid to generate a mass tag [27]; in contrast, the practical aspect of full-labeling is hardly reached [28].

Our study was structured to report a method to identify proteins and glycoproteins exhibiting significant alterations of their expression between RA patients and healthy human sera via label-free quantification technique. Two groups of samples were treated with filter added sample preparation (FASP) and trypsin digestion respectively, followed by zwitterionic hydrophilic interaction chromatography (ZIC-HILIC) enrichment and collision induced dissociation (CID) fragmentation by using Orbitrap Elite MS/MS. The generated raw files were mapped to peptide sequence and N-glycosylation sites through pFind software. The results showed that 92 proteins, including 39 glycoproteins, were identified with significant variations in RA patients’ serum samples. Meanwhile, 41 distinctly expressed (over 2-fold range) proteins, including 28 N-linked glycoproteins, were discovered by means of our method (complement factor B is overlapped). Further, this approach is an affordable, feasible and capable technique for pathological specimens’ studies. It offers a simple and fast manner for early stage diagnosing and disease progression monitoring for particular conditions that are associated with over-expressed proteins.
2.2 Materials and Experiments

2.2.1 Materials and Chemicals

Materials: Peripheral blood samples were received from 20 patients with RA at Peking University People’s Hospital in Beijing, China, and all were collected at 6 am and processed within 18 h of collection. The sera were split into aliquots and stored at −80°C until use. All patients met the ACR 1987 revised criteria for RA classification and had active arthritis at the time of sampling [29]. Detailed demographic and clinical characteristics were listed in supplementary material.

Chemicals and Enzymes: Peptide-N-glycosidase F (PNGase F) was purchased from New England Biolabs (Ipswich, MA), formic acid (FA) and urea and ammonia bicarbonate solids were obtained from Sigma-Aldrich Co. (St. Louis, MO). ZIC-HILIC media was bought from Merck (Merck, Germany). Sequencing grade porcine trypsin was purchased from Promega (Madison, WI); deionized water was produced by a Milli-Q A10 system from Millipore (Bedford, MA). Optimal LC-MS grade acetonitrile (ACN) and same grade water were purchased from Thermo Fisher Scientific, Inc. (Waltham, MA). Dithiothreitol (DTT) was obtained from Promega, and 98% Iodoacetamide (IAA) was acquired from Alfa Aesar (Ward Hill, MA). 3M Empore C8 disk was bought from 3M Bioanalytical Technologies (St. Paul, MN). The YM-30 (30 kD) and YM-3 (3 kD) filtration devices were purchased from Millipore. Western blot was performed by the Original iBlot® Gel Transfer Device from life technology, Thermo Fisher Scientific, Inc. The following commercial antibodies were purchased for Western blot analysis of result verification: rabbit anti-human complement factor B antibody, rabbit anti-human Serum Amyloid A (SAA) antibody and antibody horseradish peroxidase (HRP)-linked antirabbit IgG were purchased from Abcam plc. (Cambridge, MA). ECLplus reagent chemiluminescence detection system was obtained from GE Healthcare (Waukesha, WI).
2.2.2 Filter-Aided Sample Preparation (FASP)

Same portions of ten RA patients’ human sera were pooled to reduce the individual
difference. A same number of Healthy human sera were treated by following the same procedures.
Approximately 100 µg of serum (10 µL) was included in each sample according to Nanodrop 2000
Spectrophotometer manufactured by Thermo Fisher Scientific. Following the FASP protocol
previously reported by Wisnewski et al. [30]. Specimens were mixed thoroughly with 20 µL of
lysis buffer (20 mM Tris-HCl, 4% (v/v) SDS, 100 mM DTT, pH 7.6) followed by a 5-minute
incubation at 95 °C. An amount of 270 µL of Urea solution (8 M urea in 0.1 M Tris/HCl, pH 8.5)
was combined with each protein solution before transferring the entire volume of RA and healthy
human sample mixtures into two separate 30kDa Microcon filtration devices. After an hour of
14000g centrifugation, 100 µL of Urea solution was deposited into each filter unit, and then the
centrifugation was then resumed for 10 minutes twice. Followed by another centrifugation step,
the concentrates were combined with 100 µL of 50mM IAA in Urea solution and the units were
incubated in darkness at room temperature for 30 minutes. After twice of Urea solution
addition/centrifugation process, the systems were equilibrated twice with 100 µL of 50 mM
ammonia bicarbonate (Elution Buffer). Additionally, centrifuging the units to dead volume was
preferred, theoretically less than 20 µL remains inside the filter. 100 µL elution buffer was added
to each filter unit, followed by adding 2 µg trypsin. The digestion process was accomplished at 37
°C for 12 hours. To fully elute desired trypsinized peptides, the filter units were rinsed with 50 µL
elution buffer under centrifugation for five times. Millipore ZipTip pipette tips with C18 resin
were applied for desalting the samples before LC-MS/MS analysis. Desalted peptides were dried
in a vacufuge and stored in the -80°C refrigerator for future MS detection.
2.2.3 N-Glycopeptides Enrichment and Simplification

Approximately 50 µg of pre-dried digested peptides were added into 10 mg ZIC-HILIC media and enriched by following the previously published procedures [31]. The entire enrichment processes were in one self-made tip-size column. The detailed strategy is as follows: C8 extraction disk was loaded into a 200-µL pipet tip; afterward, ZIC-HILIC media was suspended in 100 µL acetonitrile and injected into tips. Digested peptides were re-dissolved in binding buffer (80% ACN, 0.5% FA) and loaded onto the ZIC-HILIC 200 µL-prepacked tip with six times equilibrating procedure by using100 µL binding buffer. Bounded peptides were eluted twice with 80 µL elution buffer (99% H₂O, 1% FA), then speed-vacuumed until dry. The enriched N-linked glycopeptides were resuspended in 50 µL G7 buffer (50 mM Sodium Phosphate, pH 7.5) and incubated with peptide-N-glycosidase F (N-glycanase F, PNGase F) for 12 hours under 37 °C for deeper enzymatic reaction. The well-prepared samples were dried by a vacufuge vacuum concentrator and stored in a refrigerator at -80 °C for further LC-MS/MS analysis.

2.2.4 LC-MS/MS Analysis

Reverse-phase nano-HPLC-MS/MS data were obtained on an LTQ-Orbitrap Elite mass spectrometer (Thermo Fisher) equipped with EASY-spray source and nano-LC UltiMate 3000 high-performance liquid chromatography system (Thermo Fisher). EASY-Spray PepMap C18 Column was a separation unit with a length of 50 cm capillary filled with a C18 reversed phase particles (EASY-Spray Column PepMap C18; particle size, 2 µm; pore size, 100 Å; ThermoFisher). Separation was achieved by a mixture of mobile phase A (99.9% H₂O, 0.1% FA) and phase B (79.95% ACN, 19.95% H₂O, 0.1% FA). The linear gradient was from 5 to 50% buffer B for 300 min at a flow rate of 300 nL/min. LTQ-Orbitrap Elite mass spectrometer was operated in the data-dependent mode. A full-scan survey MS experiment (m/z range from 375 to 1600;
automatic gain control target, 1e6 ions; resolution at 375 m/z, 60,000; maximum ion accumulation
time, 10 ms) was acquired by the Orbitrap mass spectrometer, and ten most intense ions were
fragmented by collision induced dissociation (CID) in the octopole collision cell. The CID
fragment ion spectra were acquired in the ion trap analyzer with a resolution of 60,000 at m/z 500
(automatic gain control target, 1e4 ions; maximum ion accumulation time, 10 ms). The MS² scan
model was set as the centroid. The other conditions used were: normalized collision energy is 35,
default charge state is 3, and activation time is 5 ms.

2.2.5 Data Searching and Data Annotation for N-Glycoprotein and Protein Analysis

In this work, the MS spectra were searched under the human database downloaded from
Uniprot_swissprot plus Uniprot_TrEMBL (Released on 2012-04, human, 65493 entries), linked
with reversed version of all sequences. According to Ma et al. (2013), all of the asparagine (N),
potential glycosylation sites, in sequence N-!P-[S|T|C] (where !P is not proline) were replaced with
J for database searching and easier data analysis purposes [31]. The definition of J was the
modification of N with 0.9850 mass tags, which is another symbol of aspartic acid (D). The raw
data was compared and matched to UniProt human database by using pFind 2.1 search engine [32]
with static modification of Carbamidomethyl (Cys), dynamic modification of J_0.9840 (N
position), Deamination (N), Oxidation (Met) and Acetylation (Lys). Trypsin was picked for
digestion with an allowance of two missed cleavages. The mass tolerance of the precursor ion was
adjusted to 20 ppm while 0.8 Da was applied for the fragment ion. A false discovery rate (FDR)
of 1% was estimated and implemented to all data sets at the total peptide level. Furthermore, the
pBuild software was assigned to remove redundant protein entries and to group related proteins
into a single group entry. All the data was assessed for the student test, and then a list of proteins
with significant variation based on p-value was generated. The DAVID functional analysis tool
(http://david.abcc.ncifcrf.gov/home.jsp) was applied to derive pathways and protein distribution to track identified serum N-glycosylated proteins [33]. The published work of Ma et al. (2013) was adopted in this study for searching and annotating proteins with some minor modification to reach higher proteome coverage. The modification-settings were optimized as following. The static modification is Carbamidomethyl (Cys), and the dynamic ones were Deamination (Asn), Oxidation (Met) and Gln to pyro-Glu (Any N-terminal started with Q).

2.2.6 Verification of Protein Lists with Significant Difference from N-Glycoprotein and Protein Data Analysis – Western Blot Assay (WBA)

The entire verification experiment was following the standard procedure of Western-blot assay [34]. The protein contents of sera were adjusted to 1.5 mg/ml with deionized water under Nanodrop 2000 monitoring. The units were boiled in 4X NuPAGE LDS protein sample buffer (Thermo Fisher) for 10 minutes and centrifuged under 12,000xg under 4°C. An equal volume of each diluted sample was loaded on 4-12% NuPAGE, Bis-Tris, 0.75 mm thick gels with 10 wells (Invitrogen). The electrophoresis was run in running buffer at 180 V for 1 h and transferred to Nitrocellulose Membrane by using iBlot® Transfer Stack (Thermo Fisher). The membrane was blotted in 5% nonfat dry milk in 0.1% Tween-20, 150mM NaCl and 50mM Tris, pH7.6(TBST) for 2 hours and incubated Over-Night with anti-complement factor b antibody, diluted 1:10000, or anti-serum amyloid A antibody with 1: 1000 dilution ratio, followed by three times 5-minute washing steps in TBST. Afterward, the entire membrane was incubated for one hour with antibody horseradish peroxidase (HRP)-linked antirabbit IgG under 1:5000 dilutions in blocking buffer. With carefully three-time rinsing, the membrane was detected under ImageQuant Las 4000 (GE Healthcare life sciences) after incubating with ECLplus reagent chemiluminescence detection
system (GE Healthcare) for 3 minutes. ImageQuant Las 4000 was applied in capturing and modifying the images.

2.3 Results

2.3.1 Label-Free Quantification Strategy for N-Glycoprotein and Protein Sample

In order to obtain the essential data, two sets of experiments were triplicated following the label-free quantitative technique. The entire experimental workflow used in this study is shown in Figure 2.1. Equal aliquots (200 µg) of 10 serum samples extracted from different healthy and RA-diseased individuals were separately mixed into one pooled sample. For N-glycoprotein analysis (in blue), the strategy of detection is to identify the glycopeptides with glycosylation sites. In the course of the procedure, proteins were digested with trypsin using FASP protocol followed by ZIC-HILIC enrichment to bind selectively digested glycosylated peptides [2]. Subsequently, glycopeptides were digested by PNGase F to remove N-linked glycans; it changed asparagine (Asn) to aspartic acid (Asp) and left a mass tag with 0.9840 Dalton difference. Simplified glycopeptides were analyzed with a high-precision Orbitrap Elite. Similar workflow (in green) was applied again on the total proteins without investigating glycosylation sites. The merged samples were studied under the collision-induced dissociation (CID) technique. All the spectra were matched with human protein database of using pFind studio 2.0. The obtained protein list was generated after statistic calculation and comparison; Western-blot assay was the primary technique to verify the selected abundant proteins among the list in different groups of samples.

2.3.2 Database Searching and Data Comparison of Site-Specific Glycosylation

Occupancy between N-Glycoproteins

On average, a total of 855 matched peptides of 653 proteins were identified in NHS. Of
those, 297 N-glycosylated sites in 134 glycoproteins with the sequence N-!P-[S|T] (where !P is not proline) were detected in NHS. Additionally, 540 proteins with 577 peptide fragments were recognized in RAHS from three trials for site-specific N-glycosylation expression analysis. Furthermore, a sum of 294 N-glycosylated sites in 132 glycoproteins was found in RAHS. In total, an intersection of 254 N-glycosylation sites from 119 N-linked glycoproteins was identified between RAHS and NHS categories (Figure 2.2). The enriched N-glycopeptides were evaluated through pFind software to confirm and score peptide sequences. From run to run, #2 and #3 sample runs were more alike according to spectra similarity between rheumatoid arthritis and healthy volunteer specimens (Negative Control).

Figure 2.1 The Scheme to Identify the Difference in N-Glycoproteins and Serum Total Proteins
The identifications workflow of N-glycoprotein and serum total protein with significant expressive variances between rheumatoid arthritis and normal human sera is shown above.

Based on the statistic comparisons, Student’s t-test, between RAHS data and NHS data, 53 N-glycosylation sites from 39 glycoproteins had significant differences between RA serum samples and normal serum samples (p-value< 0.05). Of those, three proteins were very highly
significant \((p < 0.001)\), while 12 proteins were identified with highly significant \((0.001 < p < 0.01)\). Furthermore, a total of 35 N-glycopeptide fragments within 28 glycoproteins listed in supplementary information had abundance variations significantly (over 2-fold difference) between RAHS and NHS. Of the information above, 27 N-linked glycopeptide fragments from 22 glycoproteins were hyper-glycosylated, or the expressions of these 22 proteins in RAHS are more than two times higher than the one in NHS. Meanwhile, 8 N-linked glycopeptides from 7 glycoproteins were hypo-glycosylated or down-expressed on the basis of the statistic result. The microarray result obtained by MeV software and the two-color T-Test image is attached below in Figure 2.3. During N-glycoprotein comparison study, two N-glycosylation sites were distinguished from four proteins respectively, complement C4-A or B, ceruloplasmin, apolipoprotein B-100 and \(\alpha\)-1-antitrypsin; while, four unique high expressed N-glycosylated peptides were identified from \(\alpha\)-1-antichymotryps in total. The entire list of 35 N-linked glycopeptides is listed in the supplementary material in the order of their alteration ratios.

### 2.3.3 Database Searching and Data Comparison for Proteins Identified with Alteration in Global Perspective

Without using the ZIC-HILIC enrichment technique, analysis of trypsin-digested proteins showed a broader and more accurate image of proteins with expression variation that are disease-related. In healthy volunteer human serum, an average of 786 proteins was identified. While, an average of 795 proteins was revealed in this step from rheumatoid arthritis patients’ sera with a total of 247 proteins in common between sample groups after removing unreview proteins. Based on the statistic comparisons between two sets of data, RAHS and NHS, 53 proteins were reliable \((p < 0.05)\); of the 53 proteins, three proteins were very highly significant \((p < 0.001)\), and 19 proteins were identified with high significance \((0.001 < p < 0.01)\). Besides, 14 proteins had major changes in
relative protein global expression ratios (Figure 2.4). The RA triplicates were more similar between sample-run #1 and #3; additionally, two triplicates in the negative control (NHS), run #2 and #3, were significantly alike. The variations between these 14 selected proteins were attached to the supplementary information.

![Venn Diagram](image)

**Figure 2.2 The Overlapped Lists between Rheumatoid Arthritis and Normal Volunteer in Both Assays**

Two lists of components shared that generated from the results from normal human (NHS) and rheumatoid arthritis (RAHS) is shown above. The one on the left is the pool where the candidates of N-Glycosylation sites chosen from; the right figure is the list where protein with global expression difference were selected.

2.3.4 **Protein Verification by Western Blot Assay (WBA)**

We selected complement factor B and serum amyloid A protein respectively as an N-linked glycoproteins group and a protein expression group to validate our assumptions, according to antibody commercially availability. The affinity of complement component C4 is inefficient despite the fact that one site is higher glycosylated. To pursue more reliable protein-concentration differences, both antibodies were selectively bound to protein, neglecting glycoforms. Since glycan observed from certain glycoproteins alter during diseases [35], therefore, antibodies targeting a specific glycoform are not reliable to verify the amount of protein when glycan
modifications take places simultaneously. During the procedure, different sets with the same protein amount (200 µg) of the ten RA patients’ sera were loaded and verified single with the same quantity of normal human sera by following Mahmood’s and Yang’s instructions [34]; protein concentrations were measured with Nanodrop after the samples thawed thoroughly. Since Western blot is a technique that is designed to qualify for semi-quantitative analysis, it offers a relative comparison of protein expression level on the same membrane [36]. While NHS and RAHS both contain complement factor B and serum amyloid A protein, it is helpful to compare them simultaneously for both interest proteins.

Since the antibody we chose, anti-complement factor B antibody, was specifically targeting the protein, marked in Figure 2.5 as complement factor B. Typically, glycoprotein forms several bands because of glycosylation occupancy variety, which was confirmed by Figure 2.5 underneath [2]. The differences in intensities between NHS and RAHS are not significant; based on our analysis, the protein expression of complement factor B in 10 pooled RAHS were 2.2 times more distinct than NHS. This observation might be because the antibody is aiming to complement factor B instead of glycan parts of targets, and these individual differences were balanced by merging sample preparation.

Serum amyloid A protein (SAA) is an inflammation associated protein that belongs to the acute phase protein category. During acute inflammation, the serum SAA level can dramatically increase over one 1000-fold in mice as an immune defense [37-39]. In our study, serum SAA level in 10-pooled RA sera was over 10-fold higher than its level in serum of volunteers. However, there are barely enough records about our donors concerning the disease statuses, which varies based on each health conditions, lifestyle, age, or gender individually. This information can explain that why some RA patients’ sera almost have the same band darkness as the healthy ones in Western
blot assay, which is shown in Figure 2.5 as well.

To confirm the accuracy and sensitivity of our techniques, two random spectra from complement factor B and serum amyloid A protein were chosen to conduct tandem mass spectra-peptide matching. The peptide-spectra mapping results are as following in Figure 2.6 and Figure 2.7.

![Figure 2.6 and Figure 2.7]

**Figure 2.3 A List of Glycosylation Sites Identified with Expression-Alternations**

35 N-glycosylation sites were identified with significant expression difference. 27 peptides were overly expressed, and 8 glycopeptides were down-regulated. (p<0.05) Complement factor B was selected for verification process accordingly. Multiple glycoproteins here were identified from complement system.
Figure 2.4 A List of Proteins Identified with Hyper-/Hypo- Expressions

14 serum protein expression ratios have a quite broad range from 14-fold over-expression to 4-fold down-expression; among those, 10 were overly expressed, and 4 proteins were down-regulated (p<0.05). Serum amyloid A protein was selective for validation procedure.

Figure 2.5 WBA Verification Results of CFB (Top) and SAA (Bottom)

Complement factor B (Top) was selected to verify the significance and our assumption of the complement system. Serum amyloid A protein (Bottom)’s result confirmed our theory. 10 samples were loaded independently; four normal sera were run on the left, and the rest four on the right were RA patients’ sera with two protein ladders. (One is on the left, and one is in the middle.) The apparent molecular masses of prominent bands in kilodaltons (kD) are marked on the left.
2.4 Discussions

Based on our initial statistical results on the occupancy of glycosylation sites in proteins, several sites were identified with hyper-/hypo-glycosylated between NHS and RAHS. However, judging singly by the difference of occupancy of N-glycosylation sites was hardly robust enough to support the conclusion that rheumatoid arthritis changes the distribution of glycan moieties in candidate proteins; this was the reason we designed the second part of the experiment. We hypothetically made two assumptions: first, the changes (increase/decrease) in glycan occupancy discrepancy might be related to rheumatoid arthritis. In more detail, the elevation or demotion of glycosylation among these sites is not consistent with the adjustments of identical protein expression. Thus, these glycosylation sites among glycoproteins are specific targets for clinical rheumatoid arthritis diagnostics. Second, RA prompted immune system, or any other body self-defense system results in globally increasing/decreasing expressions of proteins from a particular catalog, for example, cytokines and acute phase proteins [40, 41], in order to control disease progression. If so, the adjustments of proteins simultaneously alter the ratios of N-linked glycosylation sites. Some examples are given in supplement materials that explain how the results match our theory.

One protein, α-1-antitrypsin, had an elevated expression on one glycosylation site (N271) which is over 3.3 times. Simultaneously another peptide (N70) was decreased over 5-fold. In the meantime, the expression of α-1-antitrypsin globally elevated 1.39 times which is not consistent with either of the modifications stated above. Based on the above result, RA caused the occupancy of glycosylation site of α-1-antitrypsin to differentiate from NHS. In conclusion, α-1-antitrypsin was mainly modified in the glycosylation site occupancy that was unrelated with α-1-antitrypsin protein elevation, which was our first assumption. Two N-linked glycosylation sites identified with
noteworthy differences matched the result. Asn70 and Asn271 were known as the glycosylation sites for α-1-antitrypsin [42, 43], but the confirmation of RA-associated alteration of glycosylation occupancy has not been reported yet. Works of literature have qualified α-1-antitrypsin as a biomarker for multiple diseases, which are usually accompanied by inflammation symptoms, such as lung cancer, breast cancer, or arthritis-related disorders [44-48]. Focusing on glycosylation sites occupancy alternation for arthritis-related diseases is more accurate and observable than an emphasis on the protein backbone expression.

Alpha-1-antichymotrypsin is an acute phase protein generated in the liver during inflammation, and it is from the same protein subgroup as α-1-antitrypsin [49]. Based on our mass spectrometry results, there were three asparagine residues—Asn106, Asn127, and Asn271—out of six N-linked glycosylation sites with no confirmation of the other three positions —Asn33, Asn93, Asn186— in our analysis [50, 51]. Two of three glycosylation sites, Asn271 and Asn127 with elevated 4.70-fold and 4.15-fold respectively, had significantly higher glycan occupancies compared to protein backbone expressional elevation (1.88 times). Meanwhile, Asn106 was slightly elevated to 2.73 times. The latter is practically reliable with the elevated expression of α-1-antichymotrypsin.

Similar statistical treatment of the results was applied to protein global expression study. By combining the results from N-glycoprotein and total protein analysis together, a total of 7 proteins from the complement system were identified with significant alternations, which exceeded 19% of protein identifications. 6 glycoproteins from the complement system had glycosylation occupancies changes after patients were diagnosed with RA; at the same time, three proteins among these 7 were greatly expressed in RA patients’ sera (Table.1.2).
Figure 2.6 Example Tandem Mass Spectra Showing Matches to One Peptide in Complement Factor B Protein
Figure 2.7 Example Tandem Mass Spectra Showing Matches to One Peptide in Serum Amyloid A Protein
The complement system is comprised of plenty of unique plasma proteins that play a substantial role in immune system. A huge portion of these complement proteins is enzymes that are activated by proteolytic cleavage, and consequently, the activity triggers a chain acute inflammatory event [52]. A list of literature has approved that the levels of complement C1q-C4 proteins from classical pathway were significantly elevated in patients with RA compared to those patients with RA in clinical remission by 2-fold difference [53]. The result discovered here gives undeniable evidence that the complement component C4, a glycoprotein, has over 14-times higher glycosylation occupancy variance compared to a 1.5-fold increase in protein expression globally, which matches the number motioned before. Complement factor B is another example of the intermediate from the alternative pathway. During the alteration, complement factor B breaks into complement factor Ba and Bb, regulated by complement factor H; complement factor H is crucial to the homeostasis of complement system for protecting tissues from damage by complement activities. Complement factor B is highly expressed and -glycosylated while RA develops. Unlike complement factor B, factor H maintains its normal protein level but hypo-glycosylated, according to our results. Thus, we have assumed that up-glycosylated complement factor B decreases the associate activity to associated with factor C3b, which hinder the further activity [54]. Hypothetically, glycosylated factor B incompletes the pathway to form Ba or Bb, further impeding factor H to precede the protective function. Research has indicated that down-glycosylated factor H may lose regulation control of complement activation [55-57]. In the meantime, our study proves the previously mentioned theory; even though the level complement factor H maintained the same, the down-glycosylation of factor H might associate with the severity of RA. Complement component C7, C8, and C9 are the terminal components of the complement system, while C7 serves as an anchor for C8 and C9. Hence, the expression of these three proteins—C7, C8, and
C9—remain at the same level as literature has approved [58]. The variations of complement system proteins indicate that the progression of RA associated with complement system activation; besides, complement component C4 or complement factor B combined with factor H could be exploited as potential biomarkers for RA diagnosis. The changes of each protein from complement system are concluded here in Table 2.1.

**Table 2.1 The Expression Differences between N-Glycoproteins from Complement System**

<table>
<thead>
<tr>
<th>Protein name</th>
<th>Change (▲/▼)</th>
<th>Glycan Occupancy</th>
<th>Protein Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complement C4-A</td>
<td>▲</td>
<td>14.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Complement factor B</td>
<td>▲</td>
<td>9.8</td>
<td>2.2</td>
</tr>
<tr>
<td>Complement component C7</td>
<td>▲</td>
<td>2.2</td>
<td>1.4</td>
</tr>
<tr>
<td>Complement component C9</td>
<td>▲</td>
<td>7.9</td>
<td>2.7</td>
</tr>
<tr>
<td>Complement component C8</td>
<td>▲</td>
<td>-</td>
<td>2.0</td>
</tr>
<tr>
<td>gamma chain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complement factor H-related protein 3</td>
<td>▲</td>
<td>2.9</td>
<td>1.9</td>
</tr>
<tr>
<td>Complement factor H</td>
<td>▼</td>
<td>0.18</td>
<td>1.3</td>
</tr>
</tbody>
</table>

The comparisons of proteins from complement system are showing a significant difference in their glycosylation occupancies. The variances from our results follow the complement system functional mechanism, which is robust support for our study.

### 2.5 Concluding Remarks

In this study, we established a novel method to discover 41 potential biomarkers, among which 28 are glycoproteins, of rheumatoid arthritis. Complement system proteins showed significant changes during RA progression in our study; furthermore, the variations of these target
proteins matched complement pathway mechanism; this robustly supports our discovery and the hypothesis of complement system functions in RA. As one example, the proteins complement component C4 and complement factor B with distinct changes can be used as both a diagnostic technique and a monitor of therapeutic drug efficiency. Consequently, further study of these proteins may reveal a new method to diagnose rheumatoid arthritis, which may potentially revert RA from incurable to treatable when the patients are confirmed in the early stage of the disease. In addition, our method may offer a different prospect to predict thoroughly and screen potential biomarkers in any other diseases from acute life-threatening to chronic pain-suffering.
2.6 Complementary Materials

2.6.1 Clinical Characteristics of Patients with RA

Peripheral Clinical human sera, including healthy human serum samples and rheumatoid arthritis patient sera, were supplied by Peking University People’s Hospital (Beijing, P.R. China). Previous institutional ethical approval was obtained and the healthy volunteers in the study gave their written informed consent. Ten RA patients’ human sera and ten normal human sera were equally merged into two mixed sample units respectively in this study. The ethics committee approved the research protocol.

Table 2.2 Clinical Characteristics of Patients with RA

<table>
<thead>
<tr>
<th>Age, median (range) years</th>
<th>54 (34-77)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female, no. (%)</td>
<td>16 (80)</td>
</tr>
<tr>
<td>Smoker, no. (%)</td>
<td>4 (20)</td>
</tr>
<tr>
<td>Disease activity</td>
<td></td>
</tr>
<tr>
<td>RF(^{+}), no. (%)</td>
<td>14 (70)</td>
</tr>
<tr>
<td>anti-CCP(^{+}), no. (%)</td>
<td>14 (87.5)</td>
</tr>
<tr>
<td>ESR, median (range) mm/h</td>
<td>52 (17-111)</td>
</tr>
<tr>
<td>CRP, median (range) mg/dl</td>
<td>24.7 (6.12-229)</td>
</tr>
<tr>
<td>Tender joint count (TJC), median (range)</td>
<td>2 (0-22)</td>
</tr>
<tr>
<td>Swollen joint count (SJC), median (range)</td>
<td>6 (1-34)</td>
</tr>
<tr>
<td>Therapy</td>
<td></td>
</tr>
<tr>
<td>Prednisone, no. (%)</td>
<td>11 (55)</td>
</tr>
<tr>
<td>NSAIDs, no. (%)</td>
<td>10 (50)</td>
</tr>
<tr>
<td>Protein ID</td>
<td>Protein</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>P00751</td>
<td>Complement factor B</td>
</tr>
<tr>
<td>P02748</td>
<td>Complement component C9</td>
</tr>
<tr>
<td>P80108</td>
<td>Phosphatidylinositol-glycan-specific phospholipase D</td>
</tr>
</tbody>
</table>

Table 1.3 includes several glycoproteins (partial list) from N-glycosylation sites’ alteration over two-fold variances during the statistical comparison between patients and normal healthy donor. (RAHS/NHS > 2 – fold or RAHS/NHS < 0.5 – fold), meanwhile, the list on the right is from protein expression comparison. By paralleling these two lists, several conclusions can be made.

In Table 2.3, several glycoproteins are partially selected from the result list of N-Glycosylation sites’ alteration over two-fold variances during the statistical comparison between cDMARDs, no. (%) 13 (65)
bDMARDs, no. (%) 2 (10)

bDMARD, biological DMARD; CCP, cyclic citrullinated peptide;
cDMARD, conventional DMARD; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; NSAID, non-steroidal anti-inflammatory drug; RA, rheumatoid arthritis; RF, rheumatoid factor.

2.6.2 *N-glycosylation Variations Are Caused by Protein-Specific or Site-Specific*

Table 2.3 Comparison Between N-Glycosylation Sites (NS) Analysis and Protein Expression (PE) Analysis
patients and normal healthy donor (RAHS/NHS > 2 – fold or RAHS/NHS < 0.5 – fold). On the basis of our first assumptions, the alteration of glycosylation sites occupancy is associated with either protein expression or single glycosylation sites occupancy changes. Therefore, all three proteins – complement factor B, complement component C9, and Phosphatidylinositol-glycan-specific phospholipase D – are relatively hyper-glycosylated in these three glycosylation sites with slight changes in their protein expressions.

Partial result list from N-Glycosylation sites and protein expressions comparisons with significant difference (p-value<0.05) were chosen to support our protein-level specific alteration assumption. N-glycosylation site occupancies elevate in the approximate rate as the protein expressive ratio, which is consistent with the increasing in protein expression (the difference is less than or equal to 0.2). Therefore, these proteins are associated with rheumatoid arthritis progression that triggers immune reactions, i.e. inflammatory or auto-immune diseases behaviors.

Table 2.4 Protein Comparison between N-Glycosylation Sites (NS) Analysis and Protein Expression (PE) Analysis

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Protein</th>
<th>N-linked glycopeptide glycosylation sites</th>
<th>RA/Normal Ratio(NS)</th>
<th>RA/Normal Ratio (PE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P00738</td>
<td>Haptoglobin</td>
<td>MVSHHJLT</td>
<td>TTGATLINEQWL LTTAK</td>
<td>2.0 ± 0.2</td>
</tr>
<tr>
<td>P00751</td>
<td>Complement factor B</td>
<td>SPYYJVSDEISFHCYDGYT LR</td>
<td>2.0 ± 0.2</td>
<td>2.2 ± 0.2</td>
</tr>
<tr>
<td>P01871</td>
<td>Ig mu Chain C region</td>
<td>GLTFQQJASSMCVDPDQT AIR</td>
<td>1.9 ± 0.5</td>
<td>1.7 ± 0.3</td>
</tr>
</tbody>
</table>

Table 2.4 includes several proteins (partial list) from both methods with statistical significance. By paralleling these two lists, the results support our second assumption.
2.6.3 **Complete Lists of T-Test Results from N-Glycoprotein Comparisons Site-Specific and Globally**

In Table 2.5, we listed the overlapped glycosylation sites with significant differences between rheumatoid arthritis patients’ sera and normal healthy sera. 53 N-glycosylation sites from 39 glycoproteins have significant differences between RA serum samples and normal serum samples (p-value < 0.05); of those, 3 proteins were very highly significant (p < 0.001), while 12 proteins were identified with high significance (0.001 < p < 0.01).

**Table 2.5 Result List of T-Test from N-Glycosylation Sites Site-Specific Comparison**

<table>
<thead>
<tr>
<th>Protein</th>
<th>N-linked Glycosylation Sites</th>
<th>T-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Alpha-1-antichymotrypsin</td>
<td>TLJQSSDELQLSMGNAMFVK</td>
<td>4.25E-05</td>
</tr>
<tr>
<td>2 Plasma protease C1 inhibitor</td>
<td>VLSJNSDANLELINTWVAK</td>
<td>9.99E-05</td>
</tr>
<tr>
<td>3 Alpha-1-acid glycoprotein 1</td>
<td>SVQEIQATFFYFTPJKTEDTIFLR</td>
<td>0.000179</td>
</tr>
<tr>
<td>4 Carboxypeptidase B2</td>
<td>QVHFFVJASDVDNVK</td>
<td>0.003947</td>
</tr>
<tr>
<td>5 Attractin</td>
<td>DLMFIJASK</td>
<td>0.004057</td>
</tr>
<tr>
<td>6 Alpha-1-acid glycoprotein 1</td>
<td>SVQEIQATFFYFTPJK</td>
<td>0.004117</td>
</tr>
<tr>
<td>7 Corticosteroid-binding globulin</td>
<td>AQLLQGLGFJLTER</td>
<td>0.004803</td>
</tr>
<tr>
<td>8 Ig alpha-1 chain C region</td>
<td>LAGKPTHVJVSVMMAEDGVTCY</td>
<td>0.005974</td>
</tr>
<tr>
<td>9 Protein Z-dependent protease inhibitor</td>
<td>LPYQGJATMLVVLMEK</td>
<td>0.007457</td>
</tr>
<tr>
<td>10 Adipocyte plasma membrane-associated protein</td>
<td>AGPJWTLYLFWADAYK</td>
<td>0.007958</td>
</tr>
<tr>
<td>11 Metalloproteinase inhibitor 1</td>
<td>FVGTEVJQTTLYQR</td>
<td>0.007958</td>
</tr>
<tr>
<td>12 Kininogen-1</td>
<td>ITYSIVQTJCSKENFLFTPDCK</td>
<td>0.008506</td>
</tr>
<tr>
<td>13 Serotransferrin</td>
<td>CGLVPLAENYJK</td>
<td>0.009305</td>
</tr>
<tr>
<td>14 Alpha-1-antichymotrypsin</td>
<td>FJLTETSEAEIHQSFOHLLR</td>
<td>0.009363</td>
</tr>
<tr>
<td>15 Alpha-1-antitrypsin</td>
<td>QLAHSSTNFFPSVSIATAFAMLSLGTK</td>
<td>0.009572</td>
</tr>
<tr>
<td>16 Kallistatin</td>
<td>FLJDTHAVYEAK</td>
<td>0.01074</td>
</tr>
<tr>
<td>17 Apolipoprotein(a)</td>
<td>WEYCLTR</td>
<td>0.010884</td>
</tr>
<tr>
<td>18 Complement C4-A</td>
<td>JTTCDQLQIEVTVK</td>
<td>0.011478</td>
</tr>
<tr>
<td>19 Serum albumin</td>
<td>JECFLQHKDNPNLPR</td>
<td>0.012709</td>
</tr>
<tr>
<td>20 Ceruloplasmin</td>
<td>AGLQAFFVQECJK</td>
<td>0.013654</td>
</tr>
<tr>
<td>21 Complement factor B</td>
<td>SPYYVJSEISFHCYDGYTLR</td>
<td>0.014113</td>
</tr>
<tr>
<td>22 Alpha-1B-glycoprotein</td>
<td>EGDHEFLEVPEAQEDVEATFVHPGJYSCS YR</td>
<td>0.01421</td>
</tr>
<tr>
<td>23 Hemopexin</td>
<td>SWPAVGJCSSALR</td>
<td>0.0169</td>
</tr>
<tr>
<td>24 Kininogen-1</td>
<td>HGIQYFNJNTQHSSLFMLNEVK</td>
<td>0.019126</td>
</tr>
<tr>
<td>25 Haptoglobin</td>
<td>MVSHHJLTGATLINEQWLLTAK</td>
<td>0.019152</td>
</tr>
<tr>
<td>No.</td>
<td>Protein Name</td>
<td>Protein Description</td>
</tr>
<tr>
<td>-----</td>
<td>------------------------------------</td>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td>26</td>
<td>Alpha-2-HS-glycoprotein</td>
<td>KVCQDCPLLAPLJLDTR</td>
</tr>
<tr>
<td>27</td>
<td>Clusterin</td>
<td>QLEELFIQSSPFYFWMNGDRIDSLLENDR</td>
</tr>
<tr>
<td>28</td>
<td>Inter-alpha-trypsin inhibitor heavy chain H4</td>
<td>LPTQJITFQTESSVAEQAEEFQSPK</td>
</tr>
<tr>
<td>29</td>
<td>Prostaglandin-H2 D-isomerase</td>
<td>SVVAPATDGGGLJLTSTFLR</td>
</tr>
<tr>
<td>30</td>
<td>Alpha-1-acid glycoprotein 1</td>
<td>NEEYJKSVQEIQATFFYFTPJKTEDTIFLR</td>
</tr>
<tr>
<td>31</td>
<td>Alpha-1-antichymotrypsin</td>
<td>YTGJASALFILPDQDKMEEVEAMLLPETLK</td>
</tr>
<tr>
<td>32</td>
<td>Zinc-alpha-2-glycoprotein</td>
<td>EDIFMETLKDIVEYYJDSGHSVQLQR</td>
</tr>
<tr>
<td>33</td>
<td>Complement C4-A</td>
<td>FSDGLESJSSTQFKEV</td>
</tr>
<tr>
<td>34</td>
<td>C4b-binding protein alpha chain</td>
<td>LSVKDKQYVEPEIJTIQCDSGYGVGVGQCTYSGJR</td>
</tr>
<tr>
<td>35</td>
<td>Complement factor H</td>
<td>SPYEMFGDEEVMCLNGJWTEPPQCK</td>
</tr>
<tr>
<td>36</td>
<td>Phospholipid transfer protein</td>
<td>VSJVS Lebensweise</td>
</tr>
<tr>
<td>37</td>
<td>Ig mu chain C region</td>
<td>GLTFQQJASSMCVPDQDTAIR</td>
</tr>
<tr>
<td>38</td>
<td>Ig gamma-4 chain C region</td>
<td>TKPREEFQJSTYR</td>
</tr>
<tr>
<td>39</td>
<td>Ceruloplasmin</td>
<td>ELHLQEQJIVSNAPFLDK</td>
</tr>
<tr>
<td>40</td>
<td>Complement component C9</td>
<td>JETYQLFSLYSSK</td>
</tr>
<tr>
<td>41</td>
<td>Complement factor B</td>
<td>ALQAVYSSMSWPDDVPEGWJR</td>
</tr>
<tr>
<td>42</td>
<td>Alpha-1-antichymotrypsin</td>
<td>YTGJASALFILPDQDK</td>
</tr>
<tr>
<td>43</td>
<td>Haptoglobin</td>
<td>VVLHPJYSQVDGLK</td>
</tr>
<tr>
<td>44</td>
<td>Alpha-1-antitrypsin</td>
<td>YLGJATAIFFLPDEGK</td>
</tr>
<tr>
<td>45</td>
<td>Fibronectin</td>
<td>DQCVDDTNYVJDTFHK</td>
</tr>
<tr>
<td>46</td>
<td>Complement component C7</td>
<td>INNDFNYEFYJSTWSYVK</td>
</tr>
<tr>
<td>47</td>
<td>Leucine-rich alpha-2-glycoprotein</td>
<td>LPPGLLAJFTLLR</td>
</tr>
<tr>
<td>48</td>
<td>Clusterin</td>
<td>QLEELFIQSSPFYFWMNGDR</td>
</tr>
<tr>
<td>49</td>
<td>Apolipoprotein B-100</td>
<td>SSVITLNTAELFJQSDDAVHLSSSSSVIALQYK</td>
</tr>
<tr>
<td>50</td>
<td>Complement factor H-related protein 3</td>
<td>LGYNAJTSILSFQAVCR</td>
</tr>
<tr>
<td>51</td>
<td>Complement factor H</td>
<td>WDPEVJCSMAIQLCPHPPPQIPNSHJMFTTLNYR</td>
</tr>
<tr>
<td>52</td>
<td>Alpha-1-acid glycoprotein 1</td>
<td>QDQCIYJITLYNVR</td>
</tr>
<tr>
<td>53</td>
<td>Apolipoprotein B-100</td>
<td>NYLQVPLLGVLDLSTNVYNSLYJSASYSVGJSTDHFSLR</td>
</tr>
</tbody>
</table>

Based on our statistical study, 53 proteins were reliable (p<0.05); among aforementioned information, 3 proteins were very highly significant (p< 0.001), and 19 proteins were identified with high significance (0.001<p< 0.01). Besides, 14 proteins had significant alternation on relative protein expression ratio.
<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Protein Name</th>
<th>T-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P01011   Alpha-1-antichymotrypsin</td>
<td>0.000136</td>
</tr>
<tr>
<td>2</td>
<td>P02748   Complement component C9</td>
<td>0.000595</td>
</tr>
<tr>
<td>3</td>
<td>P02790   Hemopexin</td>
<td>0.000756</td>
</tr>
<tr>
<td>4</td>
<td>P00751   Complement factor B</td>
<td>0.000991</td>
</tr>
<tr>
<td>5</td>
<td>P01594   Ig kappa chain V-I region AU</td>
<td>0.001166</td>
</tr>
<tr>
<td>6</td>
<td>P02042   Hemoglobin subunit delta</td>
<td>0.002181</td>
</tr>
<tr>
<td>7</td>
<td>P01609   Ig kappa chain V-I region Scw</td>
<td>0.00239</td>
</tr>
<tr>
<td>8</td>
<td>Q02985   Complement factor H-related protein 3</td>
<td>0.002495</td>
</tr>
<tr>
<td>9</td>
<td>P0C0L5   Complement C4-B</td>
<td>0.002892</td>
</tr>
<tr>
<td>10</td>
<td>P07360   Complement component C8 gamma chain</td>
<td>0.002995</td>
</tr>
<tr>
<td>11</td>
<td>P0C0L4   Complement C4-A</td>
<td>0.003178</td>
</tr>
<tr>
<td>12</td>
<td>P01607   Ig kappa chain V-I region Rei</td>
<td>0.003228</td>
</tr>
<tr>
<td>13</td>
<td>P02751   Fibronectin</td>
<td>0.005603</td>
</tr>
<tr>
<td>14</td>
<td>P01600   Ig kappa chain V-I region Hau</td>
<td>0.006406</td>
</tr>
<tr>
<td>15</td>
<td>P80362   Ig kappa chain V-I region WAT</td>
<td>0.006406</td>
</tr>
<tr>
<td>16</td>
<td>P01621   Ig kappa chain V-III region NG9</td>
<td>0.006585</td>
</tr>
<tr>
<td>17</td>
<td>P02100   Hemoglobin subunit epsilon</td>
<td>0.007624</td>
</tr>
<tr>
<td>18</td>
<td>P69891   Hemoglobin subunit gamma-1</td>
<td>0.007624</td>
</tr>
<tr>
<td>19</td>
<td>P69892   Hemoglobin subunit gamma-2</td>
<td>0.007624</td>
</tr>
<tr>
<td>20</td>
<td>Q04756   Hepatocyte growth factor activator</td>
<td>0.007624</td>
</tr>
<tr>
<td>21</td>
<td>P01608   Ig kappa chain V-I region Roy</td>
<td>0.009575</td>
</tr>
<tr>
<td>22</td>
<td>P08603   Complement factor H</td>
<td>0.010102</td>
</tr>
<tr>
<td>23</td>
<td>P01610   Ig kappa chain V-I region WEA</td>
<td>0.012075</td>
</tr>
<tr>
<td>24</td>
<td>P02735   Serum amyloid A-1 protein</td>
<td>0.013095</td>
</tr>
<tr>
<td>25</td>
<td>P01834   Ig kappa chain C region</td>
<td>0.013605</td>
</tr>
<tr>
<td>26</td>
<td>P00450   Ceruloplasmin</td>
<td>0.014054</td>
</tr>
<tr>
<td>27</td>
<td>P04275   von Willebrand factor</td>
<td>0.014247</td>
</tr>
<tr>
<td>28</td>
<td>P01871   Ig mu chain C region</td>
<td>0.015397</td>
</tr>
<tr>
<td>29</td>
<td>P01599   Ig kappa chain V-I region Gal</td>
<td>0.01923</td>
</tr>
<tr>
<td>30</td>
<td>P02766   Transthyretin</td>
<td>0.019293</td>
</tr>
<tr>
<td>31</td>
<td>P01772   Ig heavy chain V-III region KOL</td>
<td>0.020374</td>
</tr>
<tr>
<td>32</td>
<td>P80748   Ig lambda chain V-III region LOI</td>
<td>0.021085</td>
</tr>
<tr>
<td>33</td>
<td>P04220   Ig mu heavy chain disease protein</td>
<td>0.021401</td>
</tr>
<tr>
<td>34</td>
<td>P01023   Alpha-2-macroglobulin</td>
<td>0.02217</td>
</tr>
<tr>
<td>35</td>
<td>P01764   Ig heavy chain V-III region 23</td>
<td>0.025756</td>
</tr>
<tr>
<td>36</td>
<td>P01593   Ig kappa chain V-I region AG</td>
<td>0.028236</td>
</tr>
<tr>
<td>37</td>
<td>P69905   Hemoglobin subunit alpha</td>
<td>0.028624</td>
</tr>
<tr>
<td>38</td>
<td>P22891   Vitamin K-dependent protein Z</td>
<td>0.030239</td>
</tr>
</tbody>
</table>

The table shows the result list of T-Test from Protein Expression Comparison.
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>Q6EMK4</td>
<td>Vasorin</td>
<td>0.030239</td>
</tr>
<tr>
<td>40</td>
<td>P80108</td>
<td>Phosphatidylinositol-glycan-specific phospholipase D</td>
<td>0.031395</td>
</tr>
<tr>
<td>41</td>
<td>P01009</td>
<td>Alpha-1-antitrypsin</td>
<td>0.035172</td>
</tr>
<tr>
<td>42</td>
<td>P04003</td>
<td>C4b-binding protein alpha chain</td>
<td>0.036065</td>
</tr>
<tr>
<td>43</td>
<td>P02654</td>
<td>Apolipoprotein C-I</td>
<td>0.036150</td>
</tr>
<tr>
<td>44</td>
<td>P00738</td>
<td>Haptoglobin</td>
<td>0.036588</td>
</tr>
<tr>
<td>45</td>
<td>P00739</td>
<td>Haptoglobin-related protein</td>
<td>0.036999</td>
</tr>
<tr>
<td>46</td>
<td>P35908</td>
<td>Keratin, type II cytoskeletal 2 epidermal</td>
<td>0.037195</td>
</tr>
<tr>
<td>47</td>
<td>P02538</td>
<td>Keratin, type II cytoskeletal 6A</td>
<td>0.037399</td>
</tr>
<tr>
<td>48</td>
<td>P07357</td>
<td>Complement component C8 alpha chain</td>
<td>0.040112</td>
</tr>
<tr>
<td>49</td>
<td>P05155</td>
<td>Plasma protease C1 inhibitor</td>
<td>0.040814</td>
</tr>
<tr>
<td>50</td>
<td>P02652</td>
<td>Apolipoprotein A-II</td>
<td>0.044276</td>
</tr>
<tr>
<td>51</td>
<td>P01031</td>
<td>Complement C5</td>
<td>0.044879</td>
</tr>
<tr>
<td>52</td>
<td>P02760</td>
<td>Protein AMBP</td>
<td>0.046114</td>
</tr>
<tr>
<td>53</td>
<td>P02787</td>
<td>Serotransferrin</td>
<td>0.047350</td>
</tr>
</tbody>
</table>
CHAPTER 3. A STRUCTURAL ANALYSIS OF N-GLYCAN RELEASED FROM MONOCLONAL ANTIBODY USING HCD-MS/MS

3.1 Introduction

Immunoglobulins, also known as antibodies, is a group of Y-shaped glycoprotein serving as parts of the immune system to control infections caused by pathogens, like virus and bacterium, or allergens [59]. Several classes of immunoglobulin exist in human bodies, such as immunoglobulin A (IgA), immunoglobulin D (IgD), immunoglobulin E (IgE), immunoglobulin G (IgG), and immunoglobulin M (IgM). The genetic-caused or pathological immunoglobulin deficiencies develop or promote long-term (chronic) infections, for example, autoimmune diseases [60], allergies [61], or inflammatory diseases [62]. Therefore, IgG or other antibodies are successfully applied in disease diagnoses for the conditions mentioned earlier. Additionally, there are wide-spread medical applications, including disease therapy or prenatal therapy, with antibodies delivered via intravenous, subcutaneous, or intramuscular routes of administration; like insulin, proteins cannot be given through ingestion. Commercial mass production of antibodies, also called monoclonal antibodies (mAb) because of its generating method, is well-developed in methodology and technology. Several mAb products have been approved by FDA for cancers or autoimmune disorders treatments, such as adalimumab (brand name: Humira) from Abbott Laboratories.

Glycosylation is one of the most common post-translational modifications (PTMs) of proteins and plays a significant role in the intercellular and intracellular communication processes [13]. According to research, the majority of serum proteins are believed to be glycosylated [12]. N-linked glycosylation not only increases the structural and functional diversity of the
glycoproteins but also serves as a criterion to monitor the activity of developmental stage and disease severity of a cell [14]. There is an association between many human diseases and the changes in the glycosylation of glycoproteins, such as cancer, cardiovascular disease, and immune deficiency diseases. Furthermore, some pathological influence on N-glycans of proteins, which contributes to the immune system, were proven to be related to the occurrence and development of many diseases, such as cancers [12, 17, 18], autoimmune disease [63, 64], and lymphocytic leukemia [65] or other diseases [19-21]. Additionally, glycans are highly regulated in the immune system, since glycans can perform as markers when any foreign infectious agents exist. Thus, the alternation of glycan structures is a pathogenic indicator for autoimmune diseases, such as rheumatoid and IgG immune complex to a diagnosis of rheumatoid arthritis [22]. Also, IgG is the smallest but most abundant immunoglobulin that offers the majority of antibody-based immunity in all humans [66]. It is engaged in several immune pathways, antibody dependent cell-mediated cytotoxicity (ADCC) and complement-dependent cytotoxicity (CDC), to fight against diseases like cancers [67, 68]. One conserved N-linked glycosylation site, Asn-297, is in the CH$_2$ domain of the heavy chain (Fc fragments) with relatively high glycosylation occupancy. It is crucial to have the FC region glycan to support and control the immune reactions mentioned above. However, IgG glycosylation alters during disease or during different stages of life, which means that it is influenced by changes in age or pregnancy. Hence, the alteration of IgG glycosylation affects the binding of IgG to Fc receptors and C1q and further regulatory functions of IgG [23].

There are two ways or levels to perform sugar structural analysis: glycan level or glycopeptides level. No matter which path is selected, isolation of glycan or glycopeptides can influence the coverage and discover the rate of possible sugar forms. Current progress in oligosaccharides and glycopeptides enrichments has shown that HILIC (hydrophilic interaction
chromatography) and affinity binding chromatography are the dominant methods utilized by research groups globally in order to increase glycopeptide signals in mass spectrometer [69, 70]. These two classes have several advantages over others: higher specificity and accuracy, high reproducibility and low affinity to non-target fragments. HILIC is the better among these two because of reasonable prices and simple operation [69]. Mapping the oligosaccharides at one or several specific glycosylation sites on a protein is a time-consuming procedure. However, analyzing glycopeptides in a single protein can obtain details for both heterogeneous glycan structures in targeting glycosylation sites and peptide containing glycosylation sites for sequence matching in one run. While there are several glycomics applications commercially available with reasonable accuracy; unfortunately, none of the software available includes functions to apply glycopeptide sequences into spectra mapping [71]. On the other hand, the software to analyze of intact glycopeptides with trustable accuracy has rarely been developed [72]. Additionally, HCD (higher-energy collision dissociation) is the main dissociation technique we applied in the analysis in glycopeptide level. Several advantages of HCD are qualified over CID. HCD offers more flexible collision energy that is proper for glycan analysis, since the bound of core-fucose is extremely sensitive to collision energy [73]. Further, CID has 1/3 cut-off effect during data-acquiring process, while, HCD obtains a more comprehensive picture of all the fragments generated from sugar moiety, even when peptide bone is attached [74].

In our study, we gave a detailed description of glycan structural analysis on the level of glycopeptides with a comparison between SimGlyan software and manual deciphering. Tryptic glycopeptides from two assigned mAb were enriched with self-packed zwitterionic hydrophilic interaction chromatography (ZIC-HILIC) tip column and prepared samples ran HCD applied in nano-LC-MS-MS method. The raw files generated were mapped onto the peptide sequence,
allowing us to identify N-linked glycan through two approaches: manual mapping and SimGlycan application with peptide sequence as a modification. The results show that 19 and 23 N-glycan structures were determined from standard and modified mAb samples respectively by using SimGlycan software, while 38 and 35 glycan structures were recognized by manually mapping respectively. A total of 6 and 26 overlapped respectively from these two methods. 13 N-glycoforms were identified with significant alterations by comparing standard sample (sample A) and modified mAb (sample B) by means of our method. Further, this method is a more affordable, feasible, and accurate technique for sugar structural studies from specific glycosylation sites with peptide sequence confirmation in one data collection cycle. It offers an approach for glycan structure study and also provides information to improve the currently available software.

3.2 Materials and Experiments

3.2.1 Materials and Chemicals

Materials: monoclonal antibodies (mAb), including one standard (sample A) and one modified (sample B) samples, were supplied by National Institute of Standard and Technology (Gaithersburg, MD) with the amount of 400 µg respectively.

Chemicals: Formic acid (FA), urea and ammonia bicarbonate solids were purchased from Sigma-Aldrich Co. (St. Louis, MO). ZIC-HILIC media was bought from Merck (Merck, Germany). Sequencing grade porcine trypsin and dithiothreitol (DTT) were obtained from Promega (Madison, WI); deionized water was produced by a Milli-Q A10 system from Millipore (Bedford, MA), YM-3(3kD) filtration devices were from the same company. Optimal LC-MS grade acetonitrile (ACN) and same grade water were acquired from Thermo Fisher Scientific, Inc. (Waltham, MA). 98% Iodoacetamide (IAA) was got from Alfa Aesar (Ward Hill, MA). 3M Empore C8 disk was bought from 3M Bioanalytical Technologies (St. Paul, MN).
3.2.2 Trypsin Digestion of Monoclonal Antibody

Approximately 200 µg of each sample (2 µL) were included in following procedures according to the measurements on Nanodrop 2000 Spectrophotometer manufactured by Thermo Fisher Scientific, Inc. Incubate Buffer (8.0 M urea and 10mM DTT in 50mM ABC, pH 8.0) were added respectively to obtain the final concentration 4 µg/ µL. The units were incubated at 37 °C for three hours. Additional alkylating reagent (500 mM IAA in 50 mM ABC) was mixed finely with the samples with 50mM IAA as the final concentration, which was followed by a 30-minute alkylation under darkness at room temperature. Then, each system was combined with 1.2 µl reducing reagent (1M DTT in DI water) to react with any leftover alkylation reagent. 450 µl of 50 mM ABC solution that contained 4 µg trypsin inside was added to each sample so as to dilute high concentration of urea. After being gently mixed on a vortex mixer, the sample systems were placed into 37 °C incubator for Over-Night (O/N) digestion. Treated peptides were dried under an Eppendorf Vacufuge Concentrator and stored in the -80°C refrigerator for future MS detection.

3.2.3 Enrichment and Simplification of N-Glycopeptides Using HILIC Tips

Roughly 100 µg of digested peptides were added into 25 mg ZIC-HILIC media and enriched by following the previously published procedures by Ma et al. [31]. The entire enrichment procedures were performed using a self-packed tip-size column. The detailed strategy is as follows: C8 extraction disk was tightly stuffed into a 200 µl pipet tip. Afterward, ZIC-HILIC media was suspended in 100 µl binding buffer (80% ACN, 5% FA) and injected into tips, and the systems were equilibrated with the same buffer twice. Digested peptides gently thawed were re-dissolved in the binding buffer and loaded into 200 µl prepacked tip with six times desalting and washing procedure with 100 µl binding buffer per round. Bounded glycopeptides were eluted twice with
100 µl elution buffer (99% H₂O, 1% FA). The well-prepared samples were dried by a Vacufuge vacuum concentrator and stored in a refrigerator at -80 °C for further LC-MS/MS analysis.

### 3.2.4 HCD – MS/MS Coupling with Nano-HPLC

Reverse-phase nano-HPLC-MS/MS data were required from an LTQ-Orbitrap Elite mass spectrometer (Thermo Fisher) equipped with EASY-spray source and nano-LC UltiMate 3000 high-performance liquid chromatography system (Thermo Fisher). EASY-Spray PepMap C18 Columns were a separation unit with a length of 15 cm capillary filled with a C18 reversed phase resin (EASY-Spray Column PepMap C18; particle size, 2 µm; pore size, 100 Å; ThermoFisher). Separation was achieved by a mobile phase from 99.9% H₂O, 0.1% FA (phase A) and 79.95% ACN, 19.95% H₂O, 0.1% FA (phase B), and the linear gradient was from 5 to 40% buffer B for 40 min at a flow rate of 300 nL/min. LTQ-Orbitrap Elite mass spectrometer was operated in the data-dependent mode. A full-scan survey MS experiment (m/z range from 400 to 1600; automatic gain control target, 1e6 ions; resolution at 400 m/z, 60,000; maximum ion accumulation time, 10 ms) was acquired by the Orbitrap mass spectrometer, and ten most intense ions were fragmented by High-Energy Collision Dissociation (HCD) in the octopole collision cell. The HCD fragment ion spectra were acquired in the ion trap analyzer with the resolution of 15,000 at m/z 500 (automatic gain control target, 1e4 ions; Default Charge State is 3; activation time, 0.5 ms). The MS2 scan model was set as the centroid. The other conditions used were: normalized collision energy is 27, charge state is ones larger than 1, and dynamic exclusion enabled in this method with 1 repeat count, 18 repeat duration, exclusion duration is 30ms. Each sample was triplicated to analyze the stability and repeatability of the methods.
3.2.5 Data Annotation of N-Glycan from N-Glycopeptides Using SimGlycan

The acquired MS/MS spectra from the instruments were submitted to SimGlycan 5.42 (Premier Biosoft, Palo Alto, CA, USA) for database searching using an in-house, licensed SimGlycan server. The enzyme specificity was set to trypsin with two missed cleavages allowed. The peptide mass tolerance was set to 1.0 Da and MS/MS mass tolerance was set to 0.5 Da for the data analysis. Only scores for the peptides defined by the software greater than "identity" were considered significant for peptide identification and modification site determinations. The percentage of each glycoform for each specific core peptide was defined as the peak area for each specific glycoform/sum of the peak areas for all the glycoforms associated with the same core peptide. The peaks that related to peptides were pre-selected and used to provide predictions of potential compositions for the masses determined from MS/MS analysis. The precursor ion mass tolerance was set to 0.8 Da and tandem mass(MS/MS) tolerance was to 20 ppm initially and increased to 0.8 Da to increase the selectivity. Ion mode was positive, and the adduct was sedated carbohydrates. The resulting assignments were also validated manually if the assigned structure was not considered to be a possible component of the human glycoproteome (Xylose-containing glycans, inappropriately degraded structures, biosynthetic impossibilities).

3.2.6 Manual Sugar Structural Determination from Data of N-Glycopeptides

In this method, each raw data was manually selected for further glycan structural analysis, because 204 (HexNAc), 366 (HexHexNAc) or 290 (NeuGc-H₂O) are the characteristic glycopeptide-marker ions in high-energy collision dissociation (HCD). Furthermore, ions at m/z 186 and m/z 168 were generated from losing of one or two water molecules from the HexNAc oxonium ions [75, 76] can indicate for N-glycans existence [77]. The spectra of these individual ions were selectively chosen and summarized into another file, for example, EXCEL file, for
further analysis as well as the corresponding peak list. The peak lists were matched with the database at Consortium for Functional Glycomics (CFG) to narrow down the candidate pool. Furthermore, the potential structures were additionally selected and verified with GlycoWorkBench [78]. Another Database, glycosciences, was chosen as a resource. No derivatization or labeling agents was added. Since HCD was utilized as the primary technique, only B/Y ions were generated through tandem MS. Accuracy was set to 2.0 Dalton. The same procedures were re-performed for other 5 raw data files with the manual interpretation of results.

3.2.7 Results Interpretation for Dominate Glycan Structure

Theoretically, the candidate list required were accurate and reliable, peak areas for specific parent ion were obtained based on m/z and retention time. Following by merged six files together respectively into sample A and sample B subgroups. Abundance and standard deviation of each structure were calculated using the equation: specific peak area was divided by a sum of peak areas in one file. The calculation was performed in EXCEL utilizing built-in functions.

![Figure 3.1 The Workflow Scheme of Profiling Immunoglobulin N-Glycan Glycoforms](image)

1. Trypsin Digestion
2. HiLIC Enrichment
3. LC-MS
4. Data Analysis
5. List Generation
Spectra from both samples were recorded for three times under identical mass spectrometry condition. Further results were manually interpreted to generate a list of candidates.

3.3 Results and Discussions

3.3.1 N-Glycan Structural Profiling

A schematic diagram of the N-linked glycan structural profiling is shown in Figure 3.1. The mass values in Dalton, in Figure 3.2, are utilized for glycan structure determination. These building blocks have identical molecular weights as the neutral lost fragments that collide off the non-reducing end of glycan in the mass spectrometer.

<table>
<thead>
<tr>
<th>Name</th>
<th>M/Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fucose</td>
<td>146.0579</td>
</tr>
<tr>
<td>Xylose</td>
<td>132.0423</td>
</tr>
<tr>
<td>Mannose</td>
<td>162.0528</td>
</tr>
<tr>
<td>Glucose</td>
<td>162.0528</td>
</tr>
<tr>
<td>Galactose</td>
<td>162.0528</td>
</tr>
<tr>
<td>N-Acetylglucosamine</td>
<td>203.0794</td>
</tr>
<tr>
<td>N-Acetylgalactosamine</td>
<td>203.0794</td>
</tr>
<tr>
<td>N-Glycolyl sialic acid</td>
<td>307.0903</td>
</tr>
<tr>
<td>Sialic acid</td>
<td>291.0954</td>
</tr>
</tbody>
</table>

Figure 3.2 The Mass Values in Daltons Used in Glycan Structure Interpretation

3.3.2 Results Analysis with Data Generated from SA and MI Approaches

In our study, a total of 19 glycan sugar structures was identified from standard sample A and 23 glycoforms were discovered in sample B, which is designated as unknown, from software
searching including 6 identical structures. On the other manual interpretation strategy, a total of 38 and 35 structures were identified from both sample respectively with 26 overlapping, meanwhile confirming with other published researches [79]; 6 structures are novel structures that haven’t been reported [79]. The structures determined by SimGlycan were contained in the list generated through manual interpretation strategy. The complete list of identified structures from standard (sample mAb-A) was listed in the supplement materials. Among these, two results from these samples share 15 glycan structures including 13 with significant differences in Figure 3.4. In the meanwhile, standard deviations were considered and labeled in Figure 3.5 to show the potential differences. For sample B (modified mAb), Table 3.2 lists the identified structures. Since NIST didn’t offer any biological or structural information regarding these samples, we only can predict the possible kinetics and dynamics on the basis of literature studies on sugar structures.

Figure 3.3 The Comparison between Two Results from SA and MI
By comparing two methods together, manual interpretation (MI) approach has approximate 1.5-fold higher identification number compared to SimGlycan software analysis (SA) method. The overlapping discovery comparison ratio between two techniques – SA to MI – is over 4 times.

3.3.3 Initial Results Interpretation from SimGlycan
Initially, a total of 25 and 44 glycan structures were identified from sample A and B respectively without sharing glycoforms, which caused our assumptions. Under further analysis,
some structures identified were irrational, for example, xylose broadly exists in bacteria or fungi in nature, and none has been determined in mammals [2]. On a basis of literature suggestions, the candidates – includes xylose, etc. were eliminated from our list. Additionally, filtered sequences were optimized and verified with published known structures resulting fewer reasonable identifications – 21 from A and 27 from B. The full set of results was matched with manual interpretation results with full inclusion. Based on this reason, only the first set of data was run by SimGlycan on a basis of low identification accuracy.

Figure 3.4 Glycan Structures with Significant Variations that Identified from MI
A total of 26 glycan structures overlapped between sample A and sample B by using manual interpretation strategy. Among those, 13 out of 26 glycans showed a significant difference, the first 6 structures were hyper-expressed in sample A, while, the other 7 showed a decrease between these two.
Figure 3.5 Carbohydrate Structural Comparison between Sample A and B
Carbohydrate forms with error bars were displayed in this figure. G0 structure (#1 in the list) is the most dominate glycoform. Several structures are with mild alternation between these two samples.

3.3.4 A Comprehensive Explanation of MI Approach’s Procedures

Without the invention of accurate and precise searching engineers, the current optimal strategy is to interpret manually spectra that are potentially from target glycopeptides. Regarding the preciousness of two samples, there was a set of experimental studies conducted with identical specimens, which is donated from Dr. Liu’ lab at Augusta State University. During the experimental practices, 100 ug of mAb was utilized to imitate the entire process; meantime, a glycopeptide list was generated accordingly to include all the miscleavage conditions. The detectable glycopeptides sequences – contains Asn 297– are EEQYNSTYR (MW: 1189.512 Daltons) and TKPREEQYNSTYR (MW:1671.8085 Daltons) with double charges. The formation of glycopeptides is accomplished by transferring of precursor oligosaccharides to polypeptides in the lumen of the endoplasmic reticulum (ER) membrane through an enzymatic reaction.
Oligosaccharyltransferase is the enzyme responsible for the transfer of the precursor glycan to an asparagine residue in polysaccharide acceptor and generation of 1 water molecule. There are several criteria for this specific N-glycosylation site, asparagine, to satisfy before the glycan transfer stated above:

1) Asn must follow the consensus sequence in Asn-X-Ser/Thr (X is not Pro).
2) Asn should locate on the surface of the fully folded protein acceptor.
3) Asn must be acceptable on the luminal side of the ER membrane for precursor oligosaccharides transfer [80].

Based on those principles, the entire MS2 spectra interpretation procedures were described in the following diagram (Figure 3.6). In the MS/MS spectrum, the target peptide was TKPREEQYNSTYR with one miscleavage at second amino acid, lysine (K). The peak (M/Z=937.95, Z=2) was the above peptide with one HexNAc; theoretically, N-glycopeptides is more likely to lose the entire sugar chain except the first HexNAc [81]. The differences between each peak were interpreted as double charged monosaccharides – monosaccharide molecular weight is shown in Figure 3.6 on the right. Moreover, the oligosaccharide chain is deduced through the molecular weight difference as N-glycan sugar core with two HexNAcs and one Hex. Additionally, this MS2 spectrum was fragmented from a precursor ion, M/Z 1093.80, and the charge state is 3 in this spectrum. Therefore, the glycopeptides molecular weight is the mass of both precursor oligosaccharides and polypeptides acceptor with losing one water molecule, then the molecular weight of this glycopeptide was calculated by hand, which is 3299.4 Daltons. The glycan sequence mass was drawn on the basis of these calculations, 1627.5915 with a hydrogen ion. The sugar information was analyzed and confirmed with glycan database Consortium for Functional Glycomics (CFG). Lastly, the sugar chain structure was proved as shown in Figure 3.6.
This strategy is capable of deducing the glycan structure with the knowledge of protein sequence. Without information about protein backbone, it would be extremely tough to decipher N-linked glycan structures. N-glycosylation sites are easier to verify and speculate by following the criteria described above. Preferably proteins including one N-glycosylation site, the mAbs in this study, is more practical compared to proteins with multiple N-glycosylation sites.

Figure 3.6 An Example of MI Interpretation
One spectrum was selected to explain the interpretation of oligosaccharide structure (on the left), each peak was marked with its theoretical glycopeptides structure. Meanwhile, the glycopeptide structure was determined as the diagram in the corner. The list on the right is a molecular mass of monosaccharide component in N-glycan structures.

3.3.5 Discussions and Suggestions in Choosing between MI and SA Methods

Simglycan, developed and designed by a tech company called Premier Biosoft, is one of high throughput glycan and glycopeptide identification software which is widely used around the world. On the basis of the MS/MS and multi-stage mass spectrometry (MS^n) data, SimGlycan predicts structures of glycans, however, the software is not easy to operate and requiring high expertise in glycomics. However, technology service offered from the company can offset these drawbacks. On the other aspect, SimGlycan functions under a strong dependence with the currently published database. Therefore, SimGlycan is not optimal for novel glycan structure determination.
Even though the updated database is continually offered and uploaded to their website, it requires a valid annual membership to access these information packages. For the institutions or industries have no intention to pursue in-depth understanding in glycomics, they are reluctant to invest their profits into software that they only need for once.

### Table 3.1 The Brief Analysis of Advantages and Disadvantages between These Two Techniques.

<table>
<thead>
<tr>
<th></th>
<th>SimGlycan</th>
<th>Manual Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Price</strong></td>
<td>$ 10,000 per year</td>
<td>$&lt; 100</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>Depends on database</td>
<td>Depends on researcher expertise</td>
</tr>
<tr>
<td><strong>Number of Identification (total)</strong></td>
<td>25 for A, 44 for B</td>
<td>38 for A, 35 for B</td>
</tr>
<tr>
<td><strong>Number of Identification (verified)</strong></td>
<td>19 for A, 23 for B</td>
<td>38 for A, 35 for B</td>
</tr>
<tr>
<td><strong>Time consumption (same spectra)</strong></td>
<td>Hours (up to 24 hrs)</td>
<td>Days (at least 1-2 days)</td>
</tr>
<tr>
<td><strong>Easy to operate</strong></td>
<td>Hard</td>
<td>Easy</td>
</tr>
<tr>
<td><strong>Database Dependence</strong></td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Consistency of Result Output</strong></td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

Before deciding the strategy to persist on, researchers are suggested to consider these factors so as to make a deliberate choice.

By comparing our MI results with the ones generated by SimGlycan, approximate 30 percent of their identifications are false or misleading because of xylose’s existence; however, the results generated from manual interpretation technique were verified with several database and literature that the false identification rate (FDR) is limited less than five percent. Secondly, the
identification consistency in SimGlycan is not acceptable with a standard deviation higher than 10.

Despite these disadvantages of SimGlycan, software involved data interpretation process is time-saving and efficient when a large number of data or complicated glycoproteins (more than 5 N-linked glycosylation sites) are involved in the researches. Researchers need to consider those facts before determining the techniques they will persist on.

3.3.6 The Influences of Different Glycoforms on Monoclonal Antibody During in vivo Distributions

Currently, there are approximately 40 therapeutic monoclonal antibodies approved by the Food and Drug Administration (FDA), and the number is rapidly increasing annually. Most of these approved mAbs are the treatments for cancer, autoimmune disorders or infectious diseases despite the reacting mechanisms. In the meantime, pharmaceutical companies are thriving in engineering novel monoclonal antibodies with traditional cancer treatments attaching, such as radioummunotherapy (mAbs carrying radioactive materials), chemoimmunotherapy (mAbs bind with cancer- targeting chemical) or immunoliposome therapy (mAbs attaching with liposome contains drugs or so) [82]. Further explorations are carrying out and gaining satisfactory achievements; bispecific antibodies, a new class of therapeutic antibodies structurally and functionally merging two monoclonal antibodies, is one of the novel engineered products and under clinical trials currently [83]. N-glycans on position Asn 297 in monoclonal antibodies serve not only crucial structural functions but also affects mAbs behaviors in a human body. The terminal sugars influence the half times of mAb confirmed in mice model and healthy volunteers. Sialic acid has a potential to increase the retention time in circulation; the removal of terminal galactose serves a significant function to prolong half-life as well [84, 85]. Additionally, galactose is the
most variable residue of the Fc sugar core by nature; it alternates relying on age. IgG is less galactosylated when people are in their childhood or elder [86]. However, monoclonal antibody with G0 structure (sugar core with two GlcNAc), which derived from the marine source, induces a potent antibody-dependent cell cytotoxicity even though G0 naturally exist in the human body [86]. Furthermore, modified mAb with high-mannose glycan indicates faster serum clearance in Goetze et al.’s research [87].

In this study, sample A and B were determined significant differences in 13 glycan structures; however, any information for these two samples are confidential that NIST was not willing to offer. Like this, we made our assumptions on a basis of limited sample information. In Figure 3.5, G0 structure was the most dominant structure in both samples, notwithstanding, G0 in sample A is significantly higher than in sample B. On the basis of the sample resource, mammal, we can make following assumptions: 1. The samples can naturally generate by a human, that these donors should be either young or senior. A lot of evidence has been showing that aging can trigger the elimination of terminal galactose in immunoglobulin G that generated in a human [64, 88, 89]. 2. Both mAbs doesn’t have a long circulation and can trigger ADCC in treatments because of the lacking of terminal sialic acid [90]. 3. Modified mAbs (B) may treat with some enzyme or stored in a different environment, compared with standard mAb (A). The unsuitable environment may cause the degradation of sugars and proteins. However, we have to confirm with NIST after the entire study finished.

3.4 Conclusions

In this study, we measured and compared two glycomics strategies, SimGlyan software, and manual deciphering, and gave a detailed description of glycan structural analysis on the level of glycopeptides. The results were showing that glycoform manual interpretation offers over 50 %
more sugar structure in total identification numbers and determines more reliable results. A total of 13 N-glycoforms were identified with significant alterations by comparing standard sample (sample A) and modified mAb (sample B) by means of MI technique. Further, this method is a more affordable, feasible, and accurate technique for sugar structural studies from specific glycosylation sites with peptide sequence confirmation in one data collection cycle. However, MI strategy is time- and man power-consuming, additionally it required adequate knowledge in glycoscience to decipher the structures. In the meantime, current commercially available software is continuously improving, and our method offers an approach for glycan structure study and also provides information to improve the current software.

### 3.5 Complementary Materials

#### 3.5.1 The Glycan Structures Identified in Standard (A) and Modified Standard (B)

**Table 3.2 The List of Glycan Structures Identified from A and B**

<table>
<thead>
<tr>
<th>Number</th>
<th>Structure</th>
<th>A</th>
<th>B</th>
<th>Number</th>
<th>Structure</th>
<th>A</th>
<th>B</th>
<th>Number</th>
<th>Structure</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1" alt="Structure 1" /></td>
<td>✗</td>
<td>✗</td>
<td>16</td>
<td><img src="image16" alt="Structure 16" /></td>
<td>✗</td>
<td>✗</td>
<td>31</td>
<td><img src="image31" alt="Structure 31" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>2</td>
<td><img src="image2" alt="Structure 2" /></td>
<td>✗</td>
<td>✗</td>
<td>17</td>
<td><img src="image17" alt="Structure 17" /></td>
<td>✗</td>
<td>✗</td>
<td>32</td>
<td><img src="image32" alt="Structure 32" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>3</td>
<td><img src="image3" alt="Structure 3" /></td>
<td>✗</td>
<td>✗</td>
<td>18</td>
<td><img src="image18" alt="Structure 18" /></td>
<td>✗</td>
<td>✗</td>
<td>33</td>
<td><img src="image33" alt="Structure 33" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>4</td>
<td><img src="image4" alt="Structure 4" /></td>
<td>✗</td>
<td>✗</td>
<td>19</td>
<td><img src="image19" alt="Structure 19" /></td>
<td>✗</td>
<td>✗</td>
<td>34</td>
<td><img src="image34" alt="Structure 34" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>5</td>
<td><img src="image5" alt="Structure 5" /></td>
<td>✗</td>
<td>✗</td>
<td>20</td>
<td><img src="image20" alt="Structure 20" /></td>
<td>✗</td>
<td>✗</td>
<td>35</td>
<td><img src="image35" alt="Structure 35" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>6</td>
<td><img src="image6" alt="Structure 6" /></td>
<td>✗</td>
<td>✗</td>
<td>21</td>
<td><img src="image21" alt="Structure 21" /></td>
<td>✗</td>
<td>✗</td>
<td>36</td>
<td><img src="image36" alt="Structure 36" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>7</td>
<td><img src="image7" alt="Structure 7" /></td>
<td>✗</td>
<td>✗</td>
<td>22</td>
<td><img src="image22" alt="Structure 22" /></td>
<td>✗</td>
<td>✗</td>
<td>37</td>
<td><img src="image37" alt="Structure 37" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>8</td>
<td><img src="image8" alt="Structure 8" /></td>
<td>✗</td>
<td>✗</td>
<td>23</td>
<td><img src="image23" alt="Structure 23" /></td>
<td>✗</td>
<td>✗</td>
<td>38</td>
<td><img src="image38" alt="Structure 38" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>9</td>
<td><img src="image9" alt="Structure 9" /></td>
<td>✗</td>
<td>✗</td>
<td>24</td>
<td><img src="image24" alt="Structure 24" /></td>
<td>✗</td>
<td>✗</td>
<td>39</td>
<td><img src="image39" alt="Structure 39" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>10</td>
<td><img src="image10" alt="Structure 10" /></td>
<td>✗</td>
<td>✗</td>
<td>25</td>
<td><img src="image25" alt="Structure 25" /></td>
<td>✗</td>
<td>✗</td>
<td>40</td>
<td><img src="image40" alt="Structure 40" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>11</td>
<td><img src="image11" alt="Structure 11" /></td>
<td>✗</td>
<td>✗</td>
<td>26</td>
<td><img src="image26" alt="Structure 26" /></td>
<td>✗</td>
<td>✗</td>
<td>41</td>
<td><img src="image41" alt="Structure 41" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>12</td>
<td><img src="image12" alt="Structure 12" /></td>
<td>✗</td>
<td>✗</td>
<td>27</td>
<td><img src="image27" alt="Structure 27" /></td>
<td>✗</td>
<td>✗</td>
<td>42</td>
<td><img src="image42" alt="Structure 42" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>13</td>
<td><img src="image13" alt="Structure 13" /></td>
<td>✗</td>
<td>✗</td>
<td>28</td>
<td><img src="image28" alt="Structure 28" /></td>
<td>✗</td>
<td>✗</td>
<td>43</td>
<td><img src="image43" alt="Structure 43" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>14</td>
<td><img src="image14" alt="Structure 14" /></td>
<td>✗</td>
<td>✗</td>
<td>29</td>
<td><img src="image29" alt="Structure 29" /></td>
<td>✗</td>
<td>✗</td>
<td>44</td>
<td><img src="image44" alt="Structure 44" /></td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>15</td>
<td><img src="image15" alt="Structure 15" /></td>
<td>✗</td>
<td>✗</td>
<td>30</td>
<td><img src="image30" alt="Structure 30" /></td>
<td>✗</td>
<td>✗</td>
<td>45</td>
<td><img src="image45" alt="Structure 45" /></td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
The “x” symbol marks the existences of the structures identified from sample A and B. 45 structures were identified in total, and a total of 26 structures were overlapped between these two samples. A sum of 15 structures was shown the significant differences (p-value< 0.05).

3.5.2 Raw Data of the Glycan Structures with Significant Differences between A and B

Table 3.3 Raw Data for These 15 Glycoforms that Shown Significant Differences Among Samples

A sum of 13, out of 15 structures, indicated more significant alterations as displayed in Figure 3.4. The MeV figure revealed the variations between these samples.

RAW DATA: GLYCAN QUANTIFICATION

<table>
<thead>
<tr>
<th>CFG Structure</th>
<th>Replicates of Sample A, %</th>
<th>Replicates of Sample B, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Replicate 1</td>
<td>Replicate 2</td>
</tr>
<tr>
<td></td>
<td>Abundance</td>
<td>Abundance</td>
</tr>
<tr>
<td></td>
<td>81.945326</td>
<td>82.937990</td>
</tr>
<tr>
<td></td>
<td>2.8594026</td>
<td>2.6074288</td>
</tr>
<tr>
<td></td>
<td>0.0476739</td>
<td>0.0506948</td>
</tr>
<tr>
<td></td>
<td>0.0476322</td>
<td>0.0498587</td>
</tr>
<tr>
<td></td>
<td>0.0307472</td>
<td>0.0339171</td>
</tr>
<tr>
<td></td>
<td>0.0601871</td>
<td>0.0538477</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td>0.2782378</td>
<td>0.2082737</td>
</tr>
<tr>
<td></td>
<td>0.3380569</td>
<td>0.2737635</td>
</tr>
<tr>
<td></td>
<td>0.2376010</td>
<td>0.2499111</td>
</tr>
<tr>
<td></td>
<td>0.0668847</td>
<td>0.0669907</td>
</tr>
<tr>
<td></td>
<td>1.5837620</td>
<td>1.0481779</td>
</tr>
<tr>
<td></td>
<td>2.3408574</td>
<td>1.0233581</td>
</tr>
<tr>
<td></td>
<td>0.0764650</td>
<td>0.0393770</td>
</tr>
<tr>
<td></td>
<td>0.0711352</td>
<td>0.0703319</td>
</tr>
</tbody>
</table>
CHAPTER 4. A COMPREHENSIVE STUDY OF MODIFIED HEMAGGLUTININ FROM PROTEOMICS AND GLYCOPROTEOMICS ASPECTS

4.1 Hemagglutinin Proteinaceous Sequence Mapping

4.1.1 Introduction

Influenza virus, broadly known as “Flu virus”, can cause a contagious disease, influenza, among human, avian, or porcine [91]. Usually, the virus is spread through coughing or sneezing into the air. The virus can circulate worldwide and infect several individuals in any age group seasonally [92]. The influenza virus has a higher potential and frequency to mutate, which is an extraordinary threat to the human population; for example, the flu outbreak in China in 2009 and the pandemics spreading in the United States (U.S.) in 1957. The influenza viruses, includes influenza virus A, influenza virus B and influenza virus C, were a type of RNA viruses that were included in Orthomyxoviridae family according to virus classification [93]. The type A influenza virus is the most common and widespread among these three subtypes. However, influenza A viruses are divided into several subtypes based on two glycoproteins, Hemagglutinin (HA) and neuraminidase (NA), located on the surface of the virus. HA is vital for influenza virus to attack and blend in with host cells. In more detail, HA recognizes target host cells through sialic acid ligand on host cells’ surfaces, usually in the upper respiratory tract or erythrocytes, and forms an endosome with enveloped virus. When pH drops, HA rearranges the structure and releases endosome, which contains viral internal RNA genome, into the cytoplasm of the host.

HA and NA are the only two glycoproteins, proteins with N-linked glycosylation sites, which exists on the influenza virus. Due to the evolution of the immune system, the capabilities of these glycoproteins alternate to react against existing vaccines when infections continuously appear through a pandemic era [91]. Meanwhile, the classification of influenza type A utilizes
these antigenic variations that are divided into 16 subtypes of HA and a total of 9 NAs. The nomenclature of influenza type A is based on these HA and NA subtypes. During the antigenic alternations, amino acid sequence modifies near receptor-binding sites in HA and consequently influence the specificity and affinity of viruses and host cells. Theoretically, N-linked glycosylation sites generally follow the [N-X-S/T] sequence motif (sequon) in which X denotes any amino acid except proline [94]. Besides, there is no evidence showing any O-linked glycosylation observed among these protein subtypes [95]. Hence, glycosylation of HA and NA can affect the host specificity, virulence and infectivity of an influenza strain either directly, by changing the biological properties of HA and NA; or indirectly, by attenuating receptor binding, masking antigenic regions of the protein, regulating catalytic activity or preventing proteolytic cleavage of the stalk of NA. Paulson’s group demonstrates that certain subtypes of HA in influenza has higher capability to bind sialic acids selectively in α2-6 linkage, which primarily exists in the human respiratory track [96].

Currently, vaccination is the most efficient method to prevent seasonal influenza infections other than avoiding any contact with patients [97]. However, the rapid mutation and drug resistance of these viruses urge scientists to predict potential flu strains and to engineer the suitable vaccines. The number and distribution of the N-glycosylation sites over the viral proteome can be computationally studied and theoretically predicted by scanning the sequences for these sequons. Therefore, the combination of computational prediction and virological engineering is timely essential for forecasting and preventing the outbreak of a new influenza virus mutant before a pandemic outbreak.

In this chapter, the major focus is on protein sequence verification of hemagglutinin influenza and N-glycosylation site identifications in these hemagglutinin isomers. The coverage
of each protein under trypsin digestion treatment is over 90% by utilizing our techniques. Furthermore, 5 out of 5 glycosylation sites in both hemagglutinins were identified and verified through our two-enzyme methods. There is an additional analysis of glycosylation occupancy for each glycosylation sites, which varies from site to site. The technique described in this chapter established a method that can be widely applied to any protein or glycoprotein analysis.

4.1.2 Materials and Experiments

4.1.2.1 Materials and Chemicals

Sequencing grade porcine trypsin and sequencing grade Endoproteinase GluC were purchased from Promega (Madison, WI); formic acid (FA) liquid, heavy-label H$_2^{18}$O water, urea, and ammonia bicarbonate solids were obtained from Sigma-Aldrich Co (St. Louis, MO). Peptide-N-glycosidase F (N-glycanase F, PNGase F) was purchased from New England BioLabs (Ipswich, MA). The YM-30 (30 kD) and YM-3(3kD) filtration devices were purchased from Millipore (Bedford, MA), deionized water was produced by a Milli-Q A10 system from the same company. Optimal LC-MS grade quality acetonitrile (ACN) and same grade quality water were purchased from Thermo Fisher Scientific, Inc. (Waltham, MA). Dithiothreitol (DTT) was obtained from Promega, and 98% 2-Iodoacetamide (IAA) was acquired from Alfa Aesar (Ward Hill, MA). HA1 and HA2 purified protein stored in phosphate buffered saline (PBS) were generously donated by our collaborator.

4.1.2.2 Desalting Procedures

Approximately 200 µg of each sample (sample volume around 100 µL) were included in following procedures according to the measurements on Nanodrop 2000 Spectrophotometer manufactured by Thermo Fisher Scientific, Inc. The samples were loaded into 3KD Microcon
filtration devices for buffer changing. The units were centrifuged to the dead volume under 14000g centrifugation at 4°C. 100 µL of 50 mM ammonium bicarbonate were gently mixed with protein pellet and centrifuged for 15 minutes under 14000g centrifugation at 4°C. In order to reach the optimal buffer condition, this buffer exchange step was repeated for 6 times.

4.1.2.3 Filter-Aided Sample Preparation with Trypsin (FASP-T)

The system was combined with 80 µl Incubate Buffer (8.0 M urea and 10mM DTT in 50mM ABC, pH 8.0), and incubated at 37 °C for three hours. 200 µl of Urea solution (8 M urea in 0.1 M Tris/HCl, pH 8.5) was combined with protein solution before 1 hour of 14000g centrifugation at room temperature. 100 µl of Urea solution was deposited into each filter unit and resumed centrifuged for 10 minutes continuously two times. Followed by another centrifugation step, the concentrates were combined with 100 µl of 50 mM IAA in Urea solution, and the units were incubated in darkness at room temperature for 30 minutes. After twice of Urea solution addition/centrifugation process, the systems were equilibrated twice with 100 µl of 50 mM ammonia bicarbonate (Elution Buffer) and centrifuged until the dead volume inside the filter was less than 20 µl. 100 µL Elution Buffer was added to each filter unit, followed by adding 2 µg trypsin. Digestion conditions were under process at 37 °C O/N. In order to elute desired tryptic peptides entirely, the filter units (applicable to both methods) were rinsed with 100 µl elution buffer under centrifugation for five times. Collected flow-through peptides were dried under Eppendorf Vacufuge Concentrator and stored in the -80°C refrigerator for future sugar release procedure.

4.1.2.4 In-Solution Digestion with Trypsin (ISD-T)

The unit from desalting procedure was removed by 50% ACN in 50mM ABC with the filter upside down. The protein mixture proceeded to further preparation methods as following. Incubate Buffer (8.0 M urea and 10inmM DTT in 50mM ABC, pH 8.0) were added respectively
to occur 50µL as the final volume L. The units were incubated under 37 °C for three hours. Additional alkylation reagent (500 mM IAA in 50 mM ABC) was mixed finely with the samples with 50mM IAA as final concentration, which followed by 30-minute alkylation under darkness in room temperature. Then, each system was combined with 1.2 µl reducing reagent (1M DTT in DI water) to react with any leftover alkylation reagent. 450 µl of 50 mM ABC solution that contained 4 µg trypsin inside was added to each sample so as to dilute high concentration of urea. After gently mixed on a vortex mixer, the sample systems were placed into 37 °C incubator for O/N digestion. Further, the units were incubated for 5 minutes at 95°C to denature trypsin. Treated peptides were dried under Eppendorf Vacufuge Concentrator and stored in the -80°C refrigerator for further preparations.

4.1.2.5 Enzymatic Deglycosylation in Peptide Preparation

The pellets were reconstituted in 50 µL G7 buffer (50 mM Sodium Phosphate, pH 7.5) and incubated with peptide-N-glycosidase F (N-glycanase F, PNGase F) for 12 hours under 37 °C for deeper digestion following the published procedure previously [98]. The samples were dried in a vacuum concentrator and stored in a refrigerator at -80 °C for LC-MS/MS analysis.

4.1.2.6 Nano-LC-MS/MS Peptide-Mass Fingerprinting (PMF)

Reverse-phase nano-HPLC-MS/MS data were required from an LTQ-Orbitrap Elite mass spectrometer (Thermo Fisher) equipped with EASY-spray source and nano-LC UltiMate 3000 high-performance liquid chromatography system (Thermo Fisher). EASY-Spray PepMap C18 Columns were a separation unit with a length of 15 cm capillary filled with a C18 reversed phase resin (EASY-Spray Column PepMap C18; particle size, 2 µm; pore size, 100 Å; ThermoFisher). Separation was achieved by a mobile phase from 99.9% H2O, 0.1% FA (phase A) and 79.95% ACN, 19.95% H2O, 0.1% FA (phase B), and the linear gradient was from 5 to 40% buffer B for
40 min at a flow rate of 300 nL/min. LTQ-Orbitrap Elite mass spectrometer was operated in the data-dependent mode. A full-scan survey MS experiment (m/z range from 400 to 1600; automatic gain control target, 1e6 ions; resolution at 400 m/z, 60,000; maximum ion accumulation time, 10 ms) was acquired by the Orbitrap mass spectrometer, and ten most intense ions were fragmented by High-Energy Collision Dissociation (HCD) in the octopole collision cell. The HCD fragment ion spectra were acquired in the ion trap analyzer with the resolution of 15,000 at m/z 500 (automatic gain control target, 1e4 ions; Default Charge State is 3; activation time, 0.5 ms). The MS2 scan model was set as the centroid. The other conditions used were: normalized collision energy is 27, charge state is ones except 1, and dynamic exclusion enabled in this method with 1 repeat count, 18 repeat duration, exclusion duration is 30ms. The full process description is attached in the supplement. Each sample was triplicated to analyze the stability and repeatability of the methods.

4.1.2.7 The Confirmation of the HA-1 and -2’s Coverages Using Designated Databases

This procedure has followed a publication from Dr. Ma [31] with some minor modification to reach the high results. The raw data were searched though pFind 2.1 in HA-1 and HA-2 protein database and its reverse sequences. The modifications were optimized: the static modification is Carbamidomethyl (Cys), and the dynamic modifications were generic: Deamination (Asn), Oxidation (Met) and Gln to pyro-Glu (Any N- terminal started with Q). The most significant modification is a 0.9840 Da m/z difference when Asn altered into Asp. Trypsin was selected as the enzyme, and two missed cleavages were allowed. The mass tolerance of the precursor ion was set to 20 ppm and the fragment ions mass tolerance was 0.8 Da. A false discovery rate (FDR) of 1% was estimated and applied to all data sets at the total peptide level. Further, the pBuild was
used to remove redundant protein entries and to group related proteins into a single group entry with the coverage information as well.

4.1.3 Results and Discussions

4.1.3.1 The Coverage Determination of HA-1 and HA-2’s Sequences through PMF

A schematic diagram of the protein coverage determination for both proteins are shown in Figure 4.1.

**Figure 4.1 The Workflow of Protein Sequence Coverage Determination**

Two techniques were applied into protein digestion, in-solution digestion and FASP, to obtain better coverages. The same techniques were used to both proteins.

Two samples were respectively performed under the workflow shown Figure 4.1 in order to obtain the essential and reliable data. Two purified HAs when received were performed using in-solution or filter-aided sample preparation (FASP) digestion with trypsin. Tryptic peptides, after digestion, were further analyzed with a high-precision Orbitrap Elite. All the spectra were confirmed with fasta files, attached to supplement material, through pFind software. Obtained peptide mass fingerprinting mapping information was generated with peptides molecular weight and coverages. The coverage results are attached in Figure 4.2.
4.1.3.2 FASP Technique Gives Better Coverage of Sequences than IS Digestion Method

Based on our result in Figure 4.2, FASP method has more comprehensive peptide sequence coverage, 90.94% for HA-1 and 69.18% for HA-2, over in-solution (IS) technique, 73.40% and 40.38% respectively for HA-1 and HA-2. There are several reasons to explain the superiority of FASP strategy. FASP method involves several spin wash processes after adding each reacting reagent, which can effectively remove detergents or salts, such as SDS or phosphate salt, to prevent contaminations in the mass spectrometer. However, in solution digestion is conducted under small volume enzyme-preferred buffer, which frequently contains Tris or ammonium bicarbonate. Since protein preserves in specific buffer contain SDS or glycerol, these compounds can increase noise signal and analysis difficulty in mass spectrometer detection and suppress proteolytic digestion by deactivating enzymes.

<table>
<thead>
<tr>
<th>FASP Method</th>
<th>IS Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA-1 Coverage: 90.94%</td>
<td>HA-1 Coverage: 73.40%</td>
</tr>
<tr>
<td>HA-2 Coverage: 69.18%</td>
<td>HA-2 Coverage: 40.38%</td>
</tr>
</tbody>
</table>

Figure 4.2 A Fully Comparisons between FASP and IS Strategies.
The results of HA-1 and -2 from FASP method are on the left and results from IS technique are on the right side. According to our results, FASP method has better peptide mapping coverage and excellent confirmation ratio of n-linked glycosylation sites. However, in solution digestion
method was still adopted in further analysis to confirm our assumption and increase the reliability of our initial results.

4.1.3.3 Trypsin is More Capable for HA-1 PMF than HA-2 in Protein Digestion Level

HA-1 and HA-2 are two analogs of the glycoprotein, hemagglutinin, with artificial modifications. HA-1’s molecular weight is 58895.26 Daltons, while HA-2 molecular weight is 63867.97 Daltons. The sequences of both proteins were identical for the first 530 amino acids besides 41 additional peptides, containing one polyhistidine tag (6 histidines) at C-terminal, in HA-2, alignment of both proteins was included in Figure 4.3. Two proteins were performed simultaneously under the same procedures avoiding man-made errors. The protein coverage for both proteins was surprisingly different: The coverage of HA-1 after trypsin digestion is 89.43%. However, HA-2 has a lower coverage of 69.18% as described (Figure 4.2).

A purification technique that involved polyhistidine affinity tags (usually contains 6-9 histidines) to interact with metal ions is widely utilized in target protein purifications. Generally, the affinity tags retain on targeted protein after purification process without being removed, because the placement of the tag is optimized specifically for the target protein [99]. However, immobilized metal-affinity chromatography (IMAC) is selectively efficient to enrich tagged protein expressed at relatively higher expression rate – higher than the naturally expressed protein with the his-tag. Thus, failure of placing his-tag into the optimal position may lower the enrichment of target protein. One of these assumptions is that the last 41 amino acids may form a shield to protect tryptic cleaving positions, lysine (K) and arginine (R); additionally, the cleavages of HA-1 and HA-2, especially the last 41 amino acids, the interval between each K or R is too compact to support the ability of trypsin. The spatial structure of glycan attached is generally quite large and could potentially block trypsin to access these two residues as well. Furthermore, IMAC technique is not compatible with chemical protease inhibitors, like EDTA or DTT, since these
compounds may weaken the affinity from nickel ions. Therefore, protease inhibitor cocktails are intended to add into cell lysate to sustain degradation and deactivation of aimed proteins [100]. Thus, this is one of these assumptions that HA-2 sample contains or failed to remove these protease inhibitors through desalting procedures (dialysis) and further hinders trypsin cleaving activity. The addition of protease inhibitors influenced HA-2 coverage.

4.2 Multi-Enzyme Digestion to Improve Proteinaceous Coverages for HA-1 and -2

4.2.1 Introduction

Trypsin is one of the most widely applicable proteases in the sample preparation for proteomics study which cleaves specifically at arginine (Arg/R) and lysine (Lys/K) [101]. However, it is not the most suitable enzyme in post-translational modifications since the attached carbohydrates sterically hinder the cleavage sites away from trypsin. Thus, miscleavages have frequently occurred in glycoprotein studies [50]. Further, trypsin is not absolutely optimal to some glycoproteins because of missing cleaving sites, Arg and/or Lys. For overcoming the limitations, proteases, such as pronase [102] or proteinase K [103] have been accepted for glycoprotein metabolism in nonspecific manners. Some other enzymes are applied into glycoprotein digestion to overcome the shortage of Arg and/or Lys in protein sequences, such as endoproteinase Glu-C [104]. The smaller size of digested peptides increases not only detection signals and resolution in mass spectrometer data collecting process, but also improves sequence coverage by analyzing peptides generated in multiple enzymes treatments [50].

In this part of the dissertation, more comprehensive sequence coverage of HA-1 and -2 are identified in over 85% respectively, compared to single-enzyme digestion, trypsin. Further, we confirmed the importance of changing protein preserved buffer, phosphate buffered saline (PBS), into ammonia bicarbonate (ABC) buffer to retain enzyme activities.
### Figure 4.3 The Sequence Alignments between HA-1 and HA-2

The alignment results from T-Coffee are shown that HA-1 and HA-2 are 100% matched besides the last 41 amino acids from HA-2.

#### 4.2.2 Materials and Experiments

##### 4.2.2.1 Materials and Chemicals

All the materials and chemicals are identical with section 4.1.2.1.
4.2.2.2 Filter-Added Sample Preparation with Trypsin & Glu-C (FASP-T&GC)

The system was combined with 80 µl Incubate Buffer (8.0 M urea and 10mM DTT in 50mM ABC, pH 8.0), and incubated at 37 °C for three hours. 200 µl of Urea solution (8 M urea in 0.1 M Tris/HCl, pH 8.5) was combined with protein solution before 1 hour of 14000g centrifugation at room temperature. 100 µl of Urea solution was deposited into each filter unit and resumed centrifuged for 10-minute continuously two times. Followed by another centrifugation step, the concentrates were combined with 100 µl of 50 mM IAA in Urea solution, and the units were incubated in darkness at room temperature for 30 minutes. After twice of Urea solution addition/centrifugation process, the systems were equilibrated twice with 100 µl of 50 mM ammonia bicarbonate (Elution Buffer) and centrifuged until the dead volume inside the filter was less than 20 µl. 100 µL of 50mM ammonium bicarbonate with Glu-C enzyme (enzyme: protein ratio is 1:100) added to the unit at 37 °C incubation overnight. Trypsin stock solution (enzyme: protein ratio is 1:25) was mixed with the solution in the unit; the system was incubated under 37 °C for the same period. Digestion conditions were under process at 37 °C overnight. In order to elute desired tryptic peptides entirely, the filter units (applicable to both methods) were rinsed with 100 µl elution buffer under centrifugation for five times. Collected flow-through peptides were dried under Eppendorf Vacufuge Concentrator and stored in the -80°C refrigerator for future sugar release procedure.

4.2.2.3 In-Solution Digestion with Trypsin &Glu-C (IS-T&GC)

The unit from the desalting process was removed by 50% ACN in 50mM ABC with the filter upside down. The protein mixture proceeded to further preparation procedures as following. Incubate Buffer (8.0 M urea and 10inmM DTT in 50mM ABC, pH 8.0) were added respectively to occur 50µL as the final volume L. The units were incubated at 37 °C for three hours. Additional
alkylating reagent (500 mM IAA in 50 mM ABC) was mixed finely with the samples with 50mM IAA as final concentration, which followed by 30-minute alkylation under darkness in room temperature. Then, each system was combined with 1.2 µl reducing reagent (1M DTT in DI water) to react with any leftover alkylating reagent. 450 µl of 50 mM ABC solution that contained Glu-C under 1:100 enzyme-protein ratio inside was added to each sample so as to dilute high concentration of urea. After gently mixed on a vortex mixer, the sample systems were placed into 37 °C incubator for Over-Night (O/N) digestion. Trypsin (under 1:25 ratio) was genterally vortexed in the mixture following by the same time-length of incubator under identical condition. Further, the units were incubated for 5 minutes at 95°C to deactivate trypsin. Treated peptides were dried under Eppendorf Vacufuge Concentrator and stored in the -80°C refrigerator for further preparations.

4.2.3 Results and Discussion

4.2.3.1 The Comparison of Sequential Coverages between HA-1 and HA-2 under MED

A schematic diagram of protein coverage determination for both proteins under multi-enzyme digestion (MED) is shown in Figure 4.4.

Two samples were performed under the workflow in Figure 4.4 in order to obtain the essential and reliable data respectively. Two purified HAs upon received were performed in-solution or filter-aided sample preparation (FASP) digestion with certain concentration Glu-C and trypsin. Proteolytic peptides were further analyzed with a high-precision Orbitrap Elite under the same method in section 4.1. All the sequences were identified and mapped through pFind software. Obtained peptide mass fingerprinting mapping (PMFP) information was generated with peptides’ molecular weight and coverages. The coverage results under FASP technique are attached in Figure 4.5.
Figure 4.4 The Workflow Adopted for Multi-Enzyme Digestion
The flowchart of multi-enzyme digestion is adopted in this study concerning improving protein sequence coverage to an acceptable range.

Figure 4.5 Improved Coverage Results for HA-1 and HA-2 After MED Technique
HA-1 and HA-2 were both treated with both MED and FASP techniques and achieved over 88% coverage respectively. Meanwhile, the results identified 5 out of 5 n-glycosylation sites which are discussed in chapter 4.3.

4.2.3.2 MED Technique is Superior to Single-Enzyme Digestion Sample Treatment Based on Proteinaceous Coverage
An available and comprehensive strategy, multi-enzyme digestion (MED), was designed in this section of the study. We did a further in-depth analysis with different factors that contain two digestion methods, single-enzyme digestion (SED) and MED, and two sample preparation techniques, FASP and IS. On the basis of our results, FASP is superior to IS in sample preparation technique level; MED gives better sequence coverage in the single tandem mass spectrometry.
experiment. By adding one more enzyme, proteolytic peptide sizes dropped from 44 amino acids to 29 amino acids, which were included in the detection range of electron spray ionization (ESI) mass spectrometer, especially peptide fragment are more likely to be double charged by losing two electrons. Furthermore, FASP technique offers 50% more coverage than IS sample treatments. However, MED in FASP and IS methods didn’t show a recognizable significance within 20% peptide coverage of 100 amino acids, and the results from MED in each sample preparation procedures has distinct peptides identified from HA-1 and HA-2 respectively. Therefore, we performed both sample preparation processes in section 4.3 to further confirm our assumption of FASP’s advantage.

There are several studies published to perform multi-step enzyme digestion (same enzyme) or multi-enzyme digestion (different enzymes) to increase protein sequence coverage and overcome sample complexity. In our study, we confirmed protein sequence coverage over 88% (Table 4.1) by applying MED and FASP techniques. Achieving consistent over 88% coverage offers a reliable and comprehensive identification for peptidemass fingerprinting technology. Furthermore, MED technique is usually flexible with enzyme selections as long as deactivation is included between each enzyme digestion.

<table>
<thead>
<tr>
<th>HA-1</th>
<th>SED</th>
<th>MED</th>
<th>HA-2</th>
<th>SED</th>
<th>MED</th>
</tr>
</thead>
<tbody>
<tr>
<td>FASP</td>
<td>90.94%</td>
<td>88.87%</td>
<td>FASP</td>
<td>69.18%</td>
<td>88.44%</td>
</tr>
<tr>
<td>IS</td>
<td>73.40%</td>
<td>73.21%</td>
<td>IS</td>
<td>40.38%</td>
<td>70.93%</td>
</tr>
</tbody>
</table>

Table 4.1 HA-1 and HA-2 Digestion Methodology Comprehensive Comparison
4.3 The Verification of Theoretical Glycosylation Sites in HA and Determination of Glycosylation Occupancy for Each Specific Site

4.3.1 Introduction

For the past several decades, proteomics by mass spectrometry, especially quantitative proteomics has been an important technique for unrevealing and understanding the biological interaction in all organisms. Quantitative proteomics, to be more accurate, includes two major approaches – label-free quantification method and stable isotope labeling quantification technique [24]. Label-free quantification method has been mentioned in Chapter 3, whereas isotope labeling quantification will be well-explained in this chapter.

Isotope-code labeling introduces mass shift to be compared to target proteins in biological samples, such as body fluids[105, 106], cancer tissues [107], and cells [108]. Up until now, an increasing number of approaches have been developed based on the selected pathway of isotopic label introductions. The ICAT (isotope-coded affinity tag) [109] is the first and well-developed chemical-isotopic labeling technique, frequently using deuterium or $^{13}$C. Besides many methods designed on the basis of ICAT principle, for example, ICPL (isotope-code protein labeling) [110] and SILAC (stable isotope labeling with amino acids in cell culture) [111], offering $^{15}$N-labeled amino acids like lysine and arginine through metabolic incorporation, are more frequently applied to cell cultures for a better functional understanding of target protein under in vivo condition. The technique we utilized in this chapter is enzymatic isotopic labeling strategy to provide a mass difference through the hydrolysis of peptide bonds in $^{18}$O-coded water [112]. Currently, glycoproteomics techniques are frequently the reapplications of proteomics methods previously mentioned, including ICAT, SILAC, etc.
In our study, we modified enzymatic isotope-code labeling pathway to introduce mass difference to asparagine in N-glycosylated peptides. In theory, PNGase F hydrolyses the covalent bond formed between glycan and asparagine – glycosylation site – in $^{18}$O water and introduces a molecular weight shift of 2.988. The isotopic difference on glycosylated asparagine is utilized as a variable modification to mark any glycosylation site. Furthermore, the routine enzymatic treatment offers a mass difference, 0.984, by changing asparagine to aspartic acid during hydrolysis. Meanwhile, the deamidation in protein occurs unavoidable and brings an identical mass change. Among these asparagines with mass shift, only the asparagines that follow a consensus pattern [N-X (X can’t be proline)-S/T] are generally considered as N-linked glycosylation sites. Thus, the advantage of our method is to prevent false positive identifications from occurring by introducing $^{18}$O.

Hydrophilic interaction liquid chromatography (HILIC) is one of various glycopeptide/glycan enrichment techniques besides lectin affinity chromatography [113] or immobilized boronic acid application [114]. Commonly, silicon-based HILIC media is utilized as a hydrophilic stationary phase to interact selectively with a hydroxyl group in sugar residues under the existence of relatively hydrophobic mobile phase. According to the mechanisms of functional groups, HILIC is classified into following categories based on their functions: cationic exchange, anionic exchange, zwitterionic (ZIC) interaction, and sepharose. HILIC is capable of separating glycosylated peptides from digested unglycosylated fragments through interaction with hydrophilic sugar moiety.

In this part, we combined multi-enzyme digestion with $^{18}$O isotope-code labeling and HILIC enrichment procedure consecutively for better identifications of N-glycosylation sites. The result shows that $^{18}$O isotopic labeling method offers more comprehensive coverages of all five
glycosylation sites in HA-1 and -2 and provides glycosylation occupancy information of each position within one run.

4.3.2 Materials and Experiments

4.3.2.1 Materials and Chemicals

18O labeled water was purchased from Sigma, and ZIC-HILIC resin was obtained from Merck. The entire procedures are identical with two-enzyme digestion procedures in section 4.2.

4.3.2.2 Enzymatic Deglycosylation in Isotope-Labeled Peptide Preparation

The pallets after two enzyme digestion were resuspended in 50 µL G7 buffer in H218O (50 mM Sodium Phosphate, pH 7.5) and incubated with peptide-N-glycosidase F (N-glycanase F, PNGase F) for 12 hours under 37 °C for deeper digestion following the published procedure previously [98]. The samples were dried in a vacuum concentrator and stored in a refrigerator at -80 °C for LC-MS/MS analysis.

4.3.2.3 N-Glycopeptides Enrichment and Enzymatic Deglycosylation in Peptide Preparation

Approximately 100 µg of digested peptides were added into 25 mg ZIC-HILIC media and enriched by following the previously published procedures by Ma et.al. [31]. The entire enrichment procedures were performed using a self-packed tip-size column. The detailed strategy is as follows: C8 extraction disk was tightly stuffed into a 200 µl pipet tip; afterward, ZIC-HILIC media was suspended in 100 µl binding buffer (80% ACN, 5% FA) and injected into tips, the systems were equilibrated with the same buffer twice. Digested peptides gently thawed were re-dissolved in binding buffer and loaded into 200 µl prepacked tip with six times desalting and washing procedure with 100 µl binding buffer per round. Bounded glycopeptides were eluted twice with 100 µl elution buffer (99% H2O, 1% FA). A vacufuge vacuum concentrator dried the well-
prepared samples. Furthermore, enriched glycopeptides have followed the procedure described in section 4.1.2.5 to remove the sugars and introduce a 0.9840 Dalton mass difference. Additionally, the reaction solution can be stored in a refrigerator at -80 °C for further LC-MS/MS analysis.

**4.3.2.4 Data Processing and Determinations of Glycosylation Occupancy for Each Sites**

The sample acquired from 3.3.2.1 is following the procedures as below:

This process has followed a publication from Dr. Ma [31] with some minor modification to reach the high results. The raw data were searched though the pFind 2.1 in HA-1 and HA-2 protein database and its reverse sequences. The modifications were optimized: the static modification is Carbamidomethyl (Cys), and the dynamic modifications were generic: Deamination (Asn), Oxidation (Met) and Gln to pyro-Glu (Any N-terminal started with Q). The most significant modification is a 2.9883 Da mass difference, which was introduced when sugar cleaving process occurred in H218O. Trypsin and Glu-C were selected as the enzyme, and two missed cleavages were allowed. The mass tolerance of the precursor ion and the fragment ions mass tolerance were identical with the other procedures under the same FDR (false discovery rate).

The peptides pallets from 4.3.2.2 were performed the same procedure as described in section 4.1.2.7. The database searching result was obtained from pBuild application with peptides mapping information. On basis of statistical accuracy, the peptides with score lower 1E-7 were chosen for the data pool. Spectrum number counting was selective for glycosylation occupancy quantitation. A list of potential peptide sequences covered glycosylation sites were generated with spectra number for isotope label and native Asp respectively. Further calculations were utilizing Microsoft Excel software.
4.3.3 Results and Discussion

4.3.3.1 The Verification of Glycosylation Sites and Calculation for Each Glycan Occupancy

Proteolytic peptides, including glycopeptides, were equally treated under two parallel strategies to ensure a full coverage of all five glycosylation sites. One approach is based on the idea of isotope-code labeling quantification by incubation within the $^{18}$O-labeled water. Meanwhile, glycopeptides were enriched through hydrophilic interaction with ZIC-HILIC media and further hydrolyzed with PNGase F. Sample slurries were ionized, detected, and then analyzed by reverse phase LC-ESI-MS/MS as shown in Figure 4.6. Raw data was analyzed with pFind to confirm both various and fixed modifications in target proteins, and initial PMFP result was collected with pBuild, including protein coverage information. Finalized results were screened in Excel sheet with selected peptides for further PMF analysis (Figure 4.8).

**Figure 4.6 The Workflow of Enzymatic Digestion in Isotopic Labeled Water**
The flowchart of PNGase F digestion in isotope label water combined with multi-enzyme digestion is adopted in this study. 5 out 5 glycosylation sites were mapped in these methods; that declared a higher sensitivity.
HA-1 and -2 both have five identical glycosylation sites in following positions: Asn32, Asn48, Asn251, Asn423, and Asn495. The glycopeptides with Asn495 were more challenging to identify during our study, which was caused by the undetectable size of proteomic fragments and/or low intensity for these glycopeptides. Nevertheless, five potential glycosylation sites were confirmed via our protocol and detailed peptide mass fingerprinting data was attached in Figure 4.7.

**Figure 4.7 The Workflow of HILIC Enrichment Coupled Enzymatic Deglycosylation**

The flowchart of HILIC enrichment combined with PNGase F digestion is adopted in this experiment. 4 out 5 glycosylation sites were identified in this method.

4.3.3.2 Comparisons between $^{18}$O and HILIC Approaches Relying on the Results from the Identifications of Glycosylation Sites

According to previous studies, HILIC method is widely applied into polar or hydrophilic analyte purifications in biopharmaceutical applications, for example, saccharides [115], glycopeptides/glycoproteins [31], amino acids [116], polar basic compounds [117], metabolites
[118] and related polar molecules. Additionally, HILIC is more suitable for complex samples with various proteins, like body fluids. Our HILIC self-pack tip is easy to operate and with higher flexibility according to sample complexity. Enzymatic isotopic labeling, in comparison, is a conventional technology applicable for purified homogeneous sample system.

In our study, two experiments were conducted with the two approaches under same conditions, such as sample amount, temperature, enzyme quantity, and incubation time. HA-1 and -2 were simultaneously treated with both approaches. Meanwhile, HA-1 and -2 showed similar results between two approaches: isotope-code labeling method is more appropriate and providing superior coverage of all five glycosylation sites than HILIC method. From Table 4.2, isotopically labeled quantitation has shown superior advantages in studying homogeneous protein system: Isotopic labeled quantitation required shorter sample preparation time relatively in comparison because HILIC demands one extra purified steps. The introduction of $^{18}$O improves the detection signal intensity of glycopeptides and offers better spectra to interpret each glycosylation sites. Detailed glycopeptides tandem mass matching graphs for N32 is attached in Figure 4.8; the rest of the graphs are enclosed in Appendix C.

The calculation of glycosylation occupancy is based on the spectrum number counting. An alternative technique, peak area, was conducted as well. However, the efficiency of separation in the column is not optimal, the results from peak area calculation is not consistent with our initial result from spectra number counting. However, the peak area spectra from HA-1 is attached in Appendix D as a reference. Enlarged and unique mass shift cannot only distinguish these glycosylated asparagines from the ones under some unavoidable modifications – deamination, but also trace the ratios alternation of glycosylation occupancy on each glycosylation sites. The occupancy of each glycosylation site will be discussed in next section. Furthermore, isotopic
labeling approaches offer more detailed sample information in single mass spec data collection to verify protein sequence.

**HA-1 N32GT Site Glycosylated Ratio: 22.2%**

![Tandem Mass Spectrum](image)

**HA-2 N32GT Site Glycosylated Ratio: 20.0%**

![Tandem Mass Spectrum](image)

**Figure 4.8 Examples of Tandem Mass Spectra with Peptides included Glycosylation Site N32 for HA-1 (Top) and -2 (Bottom)**

In contrast, HILIC method surpasses in contamination tolerance because of the designated wash process in the protocol. Considering the rigid limitation of salt existence in ESI mass spectrometer, isotopic labeling quantitation approach desires higher purity of samples in low-salt buffers. Thus, proper sample desalting and repurifying procedures may be considered before
PNGase F digestion. There is more to say that no water contamination is permissive during deglycosylation treatment to ensure properly labeling of $^{18}$O.

Table 4.2 Compare HILIC and $^{18}$O Isotope Labeled Approaches

<table>
<thead>
<tr>
<th></th>
<th>HILIC</th>
<th>$^{18}$O</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time Required for Sample Preparation</strong></td>
<td>3-4 Days</td>
<td>2-3 Days</td>
</tr>
<tr>
<td><strong>Hardness of Sample Preparation</strong></td>
<td>Medium</td>
<td>Easy</td>
</tr>
<tr>
<td><strong>Information Acquired from Single MS Run</strong></td>
<td>Limited</td>
<td>Triple</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td>&lt; $100 per sample</td>
<td>&lt;$100 per sample</td>
</tr>
<tr>
<td><strong>Contamination Tolerance</strong></td>
<td>High</td>
<td>Low (especially for water)</td>
</tr>
<tr>
<td><strong>Benefits to Data According and Processing</strong></td>
<td>Increase target peptide intensity</td>
<td>Larger and unique mass shift, easy to detect modification + HILIC advantages</td>
</tr>
<tr>
<td><strong>Application</strong></td>
<td>Complicated sample type, such as body fluid</td>
<td>Purified single protein (better) and complicated sample</td>
</tr>
</tbody>
</table>

Table 4.3 Average Occupancy Ratio Results for Five Potential Glycosylation Sites in HA-1 and HA-2

<table>
<thead>
<tr>
<th>Glycosylation Site Occupancy Ratio</th>
<th>HA-1</th>
<th>HA-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{32}GT$</td>
<td>22.2%</td>
<td>20.0%</td>
</tr>
<tr>
<td>$N_{48}AT$</td>
<td>63.6%</td>
<td>64.7%</td>
</tr>
<tr>
<td>$N_{251}DT$</td>
<td>25.0%</td>
<td>14.3%</td>
</tr>
</tbody>
</table>
### 4.3.3.3 The Determination of Occupancy Ratio for Glycosylation Sites in HA-1 and HA-2

The entire design of this section was due to the initial unsuccessful α-gal study in chapter 4.4. We assumed that the incapability to quantitate α-gal is the reason of incomplete interaction between oligosaccharides and peptides. In consequence, $^{18}$O-isotope code labeling approach offers the information of peptide fragments with/without glycan attached. The raw data files obtained from experiments in section 4.3.2 were reevaluated for glycosylation occupancies. The ratio of the formerly glycosylated asparagine over the non-glycosylated asparagine (and thus the site occupancy) is calculated by comparing the spectra number of the deglycosylated peptide against the non-glycosylated peptide. The glycosylation site occupancy result was included in Table 4.3. HA-1 and -2 have enormous significant differences of glycosylation site occupancy in N$_{251}$ and N$_{459}$ positions. In HA-1, none of the sites has occupancy higher than 70% that may cause the unsuccessful detection of terminal glycans. As mentioned in Chapter 1, disease-related aberrant glycosylation may be associated with protein expression or glycosylation changes. Therefore, quantitative evaluation of glycosylation change in site-specific level can provide better understanding abnormal disease-associate changes in protein level or in complicated specimens [119]. However, we didn’t receive enough background for us to make any hypothesis of HA-1 and HA-2. Furthermore, this method to determine glycosylation occupancy is appropriate to other glycoproteins.

<table>
<thead>
<tr>
<th></th>
<th>N$_{422}$WT</th>
<th>67.6%</th>
<th>63.0%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N$_{459}$NT</td>
<td>50.0%</td>
<td>87.5%</td>
</tr>
</tbody>
</table>
4.4 Quantitative Determination of Terminal \( \alpha \)-Galactose and Structural Analysis of Attached Sugar Moieties

4.4.1 Introduction

Alpha-1,3- galactose, typically called alpha-gal, is a terminal glycan structure in a definite consensus pattern of galactose-alpha-1,3- galactose-R [120]. Alpha-Gal is a basic sugar moiety that expressed in mammalian cells excluding human and higher evolved primates [121]. As a result of the non-existence in human, hyperacute allergy reaction, such as anaphylaxis [122], can be activated rapidly and lethal. On basis of the pieces of literature reported [122-124], tick bites are one of the causes of generating IgE antibody against alpha-Gal around the world. In the meantime, introducing alpha-gal into a human through xenotransplantation of organs, xenogenetic heart valves for example, from pigs can trigger similar severer reactions and consequently a rejection of transplantation, and a low percentage of IgG and IgM are commonly responsible for the complement cascade [125]. Alpha-Galactose allergy, or red meat allergy, is manageable by avoiding consuming any red meat. However, there is no absolute cure for alpha galactose allergy like any other autoimmune deficiencies.

In order to understand the functional connection between glycan moieties and glycoproteins, increasing attentions have been drawn into the field of glycomics and glycoproteomics, especially under the rapid development of mass spectrometer associated technologies. The configuration and structures of glycan moiety actively regulates the function and pharmacokinetics of glycoproteins. The terminal sugars on immunoglobulins are various in their natural forms or human modified monoclonal antibodies. When terminal sugar is GlcNAc, IgG has a short half-life in serum [126]; Sialic acid is a type of common terminal sugars that exists
in N-glycans from various serum glycoproteins, the appearance of terminal sialylation extends the half-life of these glycoproteins in serum [127].

Several glycan sequencing approaches were developed to sequence properly released glycans from glycoproteins. For example, permethylated oligosaccharides analyzed by mass spectrometer [1, 128], fluorescent tags labeled glycans combined with exoglycosidase treatment [129], or tandem mass spectrometry using LC-ESI-MS/MS or MALDI (matrix-assisted laser desorption ionization) [1]. In spite of any techniques, the beginning of sequencing carbohydrates starts from releasing the carbohydrate moieties from peptide backbone using appropriate endoglycosidases, especially N-glycans can be efficiently released by PNGase F. Among these glycan sequencing strategies, reductive amination is the most common reaction applied in the field. Most often, released oligosaccharides can be labeled at their reducing end using the labels with either fluorescent or UV-absorbance. 2-aminobenamide (2-AB) is the most widely employed labels in chromatography [130]: the strategy of 2-AB- labeled glycans in HILIC with fluorescence detection has been standardized for glycan structural study, extensive detailed knowledge – LC gradient, oven temperature, or elution time for each sugar – are accepted in these protocols [131]. This derivatization approach has a superior capability to tag and truly quantify each glycan on the basis of the fluorescence intensity.

High-performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) has been developed as a technique offering a sensitive detection and efficient separation for carbohydrate analysis for the past 30 years [132]. Because sugar molecules show the property as weakly acidic anions in basic solution [133], the combination of pulsed amperometric detection with a gold working electrode serves as an outstanding technology with highly sensitivity and selectivity for the determination of monosaccharides, oligosaccharides, and
glycolipids or so. Meanwhile, the excellent reliability of analysis from HAPEC-PAD makes it a usual tool for many fields, such as food, agriculture, medicine, and biotechnology. In comparison with HPLC, HPAEC-PAD provides remarkable separation of underived sugars in lower concentrations with significant shorter columns. Additionally, monosaccharides elute immediately without binding with a column in HPLC. However, HAPEC-PAD is not capable of being on-line with ESI-MS because of the high salt concentration in fractions. It is more common to analyze fractions in MALDI mass spectrometer after a brief desalting procedure. Meantime, HPAEC-PAD lacks commercial standards for oligosaccharides [134]. Furthermore, HPLC overcomes the low sensitivity and selectivity of oligosaccharides by deriving carbohydrates with labels that have fluorescent absorbance.

In this study, we examined multiple approaches to quantify the percentage of alpha-galactose in HA-1. The combination of 2-AB labeling and exoglycosidase hydrolysis was the final approach applied to our study. We experimentally developed a combined application of glycan acid hydrolysis and HPAEC-PAD. However, the uncomprehensive glycoconjugate in HA cannot offer sufficient amount of oligosaccharide to exceed the instrumental limit of detection (LOD). Therefore, our novel developed methods mentioned here are more applicable to the understanding of high-glycosylated proteins. The second we developed here can successfully identify terminal α-galactose and quantitate the percentage of galactose utilizing HPLC with fluorescence detector to determine by the fluorescent absorbance of 2-AB labeled sugars. After MALDI-MS/MS analysis, the results were shown that only 0.02% of target galactose among total oligosaccharides released from the HA-1 backbone. This technique is appropriate to quantitate and determine proteinaceous glycan structures quantitatively and qualitatively.
4.4.2 Materials and Experiments

4.4.2.1 Materials and Chemicals

The following chemicals were obtained the highest grade available from Sigma-Aldrich Co (St. Louis, MO, USA): 2-amino benzamide (2-AB), dimethyl sulfoxide (DMSO), acetic acid, trifluoroacetic acid (TFA) and sodium cyanoborohydride. HPLC-grade acetonitrile, methanol, and water were from Fisher Scientific (Fairlawn, NJ, USA). Nonporous graphitized carbon (Carbograph) solid-phase extraction (SPE) columns (150 mg/4mL) were obtained from Alltech Associates (Deerfield, IL, USA). PNGase F, trypsin, and Glu-C were purchased from New England Biolabs (Ipswich, MA, USA).

4.4.2.2 Enzymatic Releasing of Glycans

N-glycans were released from 600 µg of HA-1 with Peptide-N-glycosidase F (PNGase F) (200 units) in 50 µL of 50mM ammonia bicarbonate; the unit was incubated for 12 hours under 37 °C water bath. The mixture was treated with ice-cold methanol to deactivate and precipitate protein, and the supernatant was aliquoted into two portions and dried in a speed-vac concentrator afterward.

4.4.2.3 Hydrolysis of Oligosaccharides into Monosaccharides

Two parts of N-glycans (equivalent to the amount of sugar from 1mg HA-1 and HA-3) were respectively treated with neutral and amino sugars or sialic acid. Two approaches are using similar procedures: 2M TFA was added into neutral and amino sugars samples and incubated under 100°C for 4 hours; similar to this, one sialic acid solution was combined with 2M acetic acid and incubate under 80°C for 3 hours. To offer an accurate and consistent temperature, dry metal block bath is more preferred than a water bath. N₂ dried samples were reconstituted with 200 µL water for HPAEC-PAD analysis. 7 sugars standards (fucose, glucose, galactose, n-acetylgalactosamine,
n-acetylglucosamine, mannose, sialic acid) divided by their prosperities into two parts: neutral and amino sugars (fucose, glucose, galactose, n-acetylglucosamine, mannose), and sialic acid. The SA is under different concentrations (0.02 µM, 0.04 µM, 0.06 µM, 0.08 µM) for a calibration curve. The concentrations of neutral and amino sugars were (5 nM, 10 nM, 15 nM, 60nM, 20 nM) of glucose, internal standard, was spiked into HA-1 and -3 samples. Sugar standard calibrators were hydrolyzed under the same procedures as described above. 10 µg of each monosaccharide and 7 sugars mixture detected without hydrolysis.

4.4.2.4 Monosaccharide Quantitation with HPAEC-PAD

Monosaccharides were separated at 0.4 mL/min flow rate on a Dinox CarboPac PA-100 column (4 x 250mm) with compatible guard column under following multi-step gradients (Table 4.4). The mobile phase is 200mM NaOH, mobile phase B is 500mM NaAc, and C is HPLC-grade water. Each mobile phase is degassed and filtered before using. The temperature for the column oven is set to 30 °C. Back pressure of the column is usually around 900-1500 psi.

Table 4.4 Gradient Chart for Monosaccharide Separation on HPAEC-PAC

<table>
<thead>
<tr>
<th>Minutes</th>
<th>A%</th>
<th>B%</th>
<th>C%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>2.5</td>
<td>0</td>
<td>97.5</td>
</tr>
<tr>
<td>20.0</td>
<td>2.5</td>
<td>0</td>
<td>97.5</td>
</tr>
<tr>
<td>45.0</td>
<td>50</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>45.1</td>
<td>2.5</td>
<td>0</td>
<td>97.5</td>
</tr>
<tr>
<td>60.0</td>
<td>2.5</td>
<td>0</td>
<td>97.5</td>
</tr>
<tr>
<td>60.0</td>
<td></td>
<td></td>
<td>Stop Run</td>
</tr>
</tbody>
</table>
4.4.2.5 Conjugation of Glycan to 2-Aminobenzamide (2-AB)

One portion of dehydrated glycans was dissolved in 50 µL labeling agent (50 mg/mL 2-AB and 1 M NaBH₃CN in DMSO/acetic acid (7:3, v/v)) before incubation at 65 °C up to 2 hours. The labeled glycan mixture was loaded onto an SPE column that prewashed with 5 mL of ACN and 5 mL of water. Theoretically, the salts and other impurities were removed by 15 mL of water, and the target derivatives were eluted with 4 mL of 25% acetonitrile containing 0.1% TFA. The elutes dried for the following enzymatic reactions.

4.4.2.6 Exoglycosidase Digestion on 2-AB Derived Glycans

The entire procedure was carried out in a 37°C water bath. The dried glycans were resuspended in 30 µL volume of 1X glyco-buffer 1(G6 reaction buffer: 50mM CaCl₂ and 50mM Sodium acetate, pH5.5 @25°C). Each enzyme was thoroughly added and mixed after 10-minute boiling denature process and was incubated for 24 hours at 37°C by following the order: α2-3,6,8 neurominidase, β1-4 Galactosidase, α1-2,4,6 fucosidase. The mixture after digestion without purification was injected into HPLC.

4.4.2.7 Chromatographic Separation and MALDI Analysis of Glycan Derivatives with 2-AB Labels

Around 10 µL oligosaccharides were injected into Shimazu Nexera x2 ultra High Performance Liquid chromatography system with a fluorescence detector. The XBridge amide HILIC column with following configurations was applied: 3.5 um, 4.6 x 150 mm. Oven temperature was 40 °C. Fluorescence detector were set as λ excitation=320 nm, λ emission =420 nm. Phase A is 50mM ammonium formate solution with pH 4.5; phase B is HPLC-grade acetonitrile based on Tayi’s protocol [135]. Glycan fractions were separated as following gradients (Table 4.5) and collected for further analysis.
The sample fractions were dried in a speedvac vacuum concentrator and stored in -80°C refrigerator for further preparations. This procedure was performed on ultrafleXtreme MALDI-TOF-TOF mass spectrometer with flex method by using 2.4-dihydroxybenzoid acid (DHB) as the matrix. All 23 fractions were spotted on an MTP 384 target plate (ground steel) The tandem mass results acquired were analyzed using GlycoWorkBench software, glycan spectrum-structure mapping results were identified by matching MS2 ion peaks against CFG glycan databases. Further mapping information is discussed in result section 4.4.3.

**4.4.2.8 Enzymatic Activity Confirmation of Alpha 1-3,6 Galactosidase by Dot Blot**

Deactivated HA-1 was treated with different ratios of α1-3,6 galactosidase for 24 hours. The enzyme to protein ratios were 100: 32, 100: 16, 100: 8, 100:2, positive and negative control. Proper nitrocellulose membrane was marked with pencil and spotted 2 µl of samples in each spot. Dried membrane was blocked following the same procedures as Western Blot. The primary antibody (1:1000 for HA-1) and the secondary antibody conjugated with HRP (1:1500) were particularly applied to HA-1. The concentration was confirmed between samples.

**Table 4.5 Gradient Chart for Derived Sugars’ Separation on HPLC-FL**

<table>
<thead>
<tr>
<th>Minutes</th>
<th>%A</th>
<th>%B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>60</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>70</td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>71</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>80</td>
<td>Stop Run</td>
<td></td>
</tr>
</tbody>
</table>

**4.4.2.9 Removal of Terminal α–Galactose with α1-3,6 Galactosidase**

All the conditions and procedures are following section 4.4.2.6 with α1-3,6 galactosidase.
4.4.3 Results and Discussion

4.4.3.1 Experimental Design Scheme for Hydrolyzed Monosaccharides Quantitation and 2-AB Labeled Sugar Quantitation

Sequencing of complex glycan mixtures is usually handled from either monosaccharide perspective or sugar moieties level. Oligosaccharide derived from HA-1 were digested with alpha-1-3,6-galactosidase and checking reaction progression simultaneously by Dot-blot. Further, samples are hydrolyzed under the acid condition to monosaccharides. Monosaccharides were quantified with HPAEC-PAD (Figure 4.9). 2-AB labeled glycans were assigned as the second approach: released glycans were tagged with reductive labels initially, the slurry was incubated with a series of continuous exoglycosidases to remove terminal sugars. These fragments were separated utilizing HPLC with fraction collector in HILIC column. Dried fractions were detected with MALDI (Figure 4.10).

Figure 4.9 A Scheme of Sugar Hydrolysis Coupled HAPEC-PAD Method
Different units of α1-3,6 Galactosidase were incubated with 100 µg of HA-1 for 12 hours. The average concentration of 2U and the positive control was taken to as the number of positive control.

![Diagram of 2-AB Derived Glycans Analysis with HPLC and MALDI-MS/MS](image)

**Figure 4.10 A Diagram of 2-AB Derived Glycans Analysis with HPLC and MALDI-MS/MS**

Derived carbohydrates with 2-AB labels were hydrolyzed with exoglycosidases and separated by an HILIC column in a gradient method of HPLC with a fluorescence detector. The collected fractions were dried and reconstitute with DHB in 50% ACN. 23 fractions were spotted on a 384 polished MALDI plate. Results were interoperated manually and confirmed with CFG database.

### 4.4.3.2 α1-3,6 Galactosidase Activity Verification Using Dot Blot

Various units of α1-3,6-galactosidase were incubated with 100 µg of HA-1 as shown in Figure 4.11. The positive control was HA-1 with no enzyme, while negative control, HA-3, was provided by the collaborator. It was an analog of HA-1 without terminal galactose attached in alpha 1-3 configuration. From analyzing our result, the perfect reacting ratio is around 100:4 (enzyme to protein) in 12-hour incubation. Extended incubation time can proceed to completion with fewer units of the enzyme [136, 137]; thus we extended our incubation time to 24-hour in 37 °C water bath, which offers more steady temperature and humidity than incubation chamber. However, the reaction unit with the highest concentration of galactosidase was showing false positive results. Many reasons can trigger false positive results: poor blocking effect, strong binding between
galactosidase and substrate, or competitive inhibition between existing antigens in samples [138].

With comprehensive blocking procedure, the applicable possibility is that denatured galactosidase competitively binds with the primary antibody. Galactosidase removes the terminal galactose in α1-3 configuration, and may keep that monosaccharide in the active site and prevent completing the reaction. When primary antibody added in, it targets those alpha galactoses on both the subtract, HA-1, and inside catalytic pockets. The level of reactions was determined based on the darkness of spots (Table 4.6).

![Dot Plot Results of α1-3,6 Galactosidase Activity Determination](image)

**Figure 4.11 The Dot Plot Results of α1-3,6 Galactosidase Activity Determination**
The same quantity of HA-1 protein was combined with various concentrations of α1-3,6 galactosidase. The optimal ratio between protein and enzyme (100µg to multiple units) is from 100:4 - 100:8 with prolonged incubation time.

### 4.4.3.3 Quantitation of Monosaccharides from Released Glycans with HPACE-PAD

The intention of this experiment is to quantitate α-Gal in either percentage level or weight level. We artificially removed terminal galactose at α-position in HA-1 and confirmed reaction completion in Dot-blot. Then, derived oligosaccharides from intact HA-1(with α-gal) and treated HA-1 (without α-gal) were hydrolyzed into monosaccharides simultaneously as long as monosaccharides standards. The quantity of α-gal, hypothetically, is capable of being calculated by subtracting galactose concentration in treated HA-1 from the concentration from intact HA-1.
However, our data has not shown any traces of monosaccharides above the limit of detection (S/N=3) and/or limit of quantitation (S/N=5).

Table 4.6 The Chart of α1-3,6 Galactosidase Activity

<table>
<thead>
<tr>
<th>Q_Engyse</th>
<th>32U</th>
<th>16U</th>
<th>8U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial #</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Darkness</td>
<td>7.7</td>
<td>6.6</td>
<td>7.9</td>
</tr>
<tr>
<td>Mean</td>
<td>7.4</td>
<td>2.0</td>
<td>3</td>
</tr>
<tr>
<td>%</td>
<td>80</td>
<td>100</td>
<td>95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q_Engyse</th>
<th>2U</th>
<th>Positive Control</th>
<th>Negative Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial #</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Darkness</td>
<td>39.7</td>
<td>37.1</td>
<td>36.1</td>
</tr>
<tr>
<td>Mean</td>
<td>37.6</td>
<td>18.7</td>
<td>1.9</td>
</tr>
<tr>
<td>%</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
</tbody>
</table>

Released carbohydrates from intact and treated HA-1 protein were hydrolyzed into monosaccharides for quantification in HPAEC-PAD. Calibrators were treated under the same protocol before HPAEC-PAD quantitation.

As following, elution times for monosaccharide and the 7-sugar standard are shown in Figure 4.12. After hydrolysis treatment, the retention time for each sugar is shortened as shown in Figure 4.15. Galactose has shown a lower affinity and sensitivity compared to other monosaccharides in the picture. The linearity of each standard curve, neutral and amino sugars and SA that determined with their peak areas, is acceptable as larger than 0.9850 (Figure 4.13 and 4.14). A table is attached in Appendix B including the peak area and retention time for monosaccharides.

In each run of HA-1 hydrolysate, no monosaccharide was identified in the limit of detection besides the internal standard, glucose. In Figure 3.15, Peak 2 and Peak 3 respectively are from internal standard, glucose, in HA-1 and HA-3 samples. According to our collaborator’s description (Figure 4.16), detection of glucose from HA-1 and HA-3 is impossible because of the glucose.
abundance in mentioned carbohydrates. 25.7 nM and 18.1 nM of Glu were quantitated from HA-1 and -3 respectively. Fucose was determined from both samples with 16.9 nM and 10.9 nM in concentration. GlcNAc has detected from both proteins also relaying on the retention time, 21.9 nM and 16.3 nM. After triplet injections of each sample, we concluded that HA-1 is more core-fucosylated than HA-3; meantime, 4.78 nM of glycan, the most dominate biantennary structure in Figure 3.16, were quantitated from 15.0 nM of the intact protein, which means only 30% of HA-1, in average, were glycosylated; among these, 0.067% is galactose with α1-4 configuration regarding the results from 2-AB labeled quantitation method.

Figure 4.12 Elution Times for 7-Sugar Standard and Monosaccharides
Based on the bottom figure, the retention time of each monosaccharide can be confirmed and identified from the 7-sugar standard graph. In PA-100 column, the quality of buffer affects the retention time. However, the elution order of monosaccharides is consistent.
Figure 4.13 Standard Curve of Sialic Acid

However, the determination of galactose is unapproachable with this technology. In order to obtain acceptable peak separation and resolution from HPAEC-PAD, the desirable concertation
of a monosaccharide, especially galactose, should exceed 10 nM, which equals to 600 µg of HA-1 if we consider HA-1 as 100% fully glycosylated with 11 sugars in each glycosylation site. As mentioned in the previous chapter, we confirmed the glycosylation occupancies for HA-1 and HA-2 respectively. For example, the average glycosylation site occupancy for HA-1 is lower than 50%, and each is lower than 70% respectively. Therefore, using HPAEC-PAD as a detection technique requires a numerous amount of target protein, over 1 mg. However, it is no doubt that HPAEC-PAD offers a high sensitivity and selectivity with a proper column. Meanwhile, it serves as a sufficient tool to monitor glycosyltransferase reaction with an injection of a small quantity of supernatant.

![Figure 4.15 Spectra of HA-1 and HA-3 Obtained From HPAED-PAD](image)

No galactose was detected in the samples; glucose was spiked into the sample as internal standards.
4.4.3.4 Exoglycosidase Treatment of 2-AB Derived Glycan

Exoglycosidase digestion combined with MALDI-TOF mass spectrometry has been an effective and efficient way to sequence the N-linked glycans from a glycoprotein [129]. The technique that couples exoglycosidase treatment with HPLC is more sensitive and reliable with derived glycans: HPLC offers excellent separation and collects the fraction for further enzyme digestion and structure confirmation [78]; the use of exoglycosidase enzyme is essential to acquire the information of linkage and sequence of glycans simultaneously. Because of the inefficiency of β1-2,4 GlcNAcse, only enzyme 1, 2, and 4 were tested in our experiments (Figure 3.16). Following incubation, one-step enzyme deactivation in between adding each enzyme, these exoglycosidases were separately mixed into the reaction for 24-hour. The list of exoglycosidase is attached in Figure 4.16.

![Diagram of exoglycosidases](image)

**Figure 4.16 A Scheme that Shown Various Exoglycosidases’ Functions and Aim Sites**

5 major glycosidases were applied in this experiment, the two complex structure of glycans are the possible glycoforms provided by our collaborator.
In this study, our purpose is to confirm the quantity of terminal galactose, which is in alpha 1-4 configuration with galactose. The glycans were derived with the 2-AB label (Figure 4.10) after released from peptide backbone with following retention times. The molecular weight of the sugar increased 120.0687 Da following the chemical reaction mentioned in Figure 4.17 underneath. On the basis of the description from our collaborator, the potential structures of carbohydrates are drawn in Figure 4.16 with/out core fucose – fucose alpha1-6 linked with the reducing GlcNAc. Therefore, one-pot enzyme reaction is more applicable for this experiment instead of monitoring the extent of each reactions using HPLC. However, we prolonged the incubation time as long as the existence of excessive enzyme in order to ensure the comprehensive activity of each glycosidase [84, 139]. However, we discovered 13 types of glycans that include only a small portion of our target from 24 fractions we collected by HPLC (Figure 4.18). Theses carbohydrates are identified with 4 man, 5 man, hybrid and a majority of complex structures. 5 out of 18 glycoforms were not declared in the possibility list. Among these structures, the percentage of alpha-gal is 0.067% based on peak area obtained in a fluorescent detector (Table 4.7). In other words, only 0.067% is our target among released oligosaccharides from HA-1. This result is much lower than the assumption and expectation of our collaborator. The list of the glycan structure we identified was attached in Table 4.7, and MALDI spectra and related interpretation diagrams are attached in Appendix A.
Figure 4.17 The Scheme of 2-AB Labelling Reaction

In theory, the activity of each enzyme is not certain 100% even though we added excessive enzyme and prolonged the incubation time, especially β1-4 galactosidase, more efficient with simple sugar, like lactose. Hydrolysis is more likely to happen when a large excess of water exists in the environment. Several enzymes were showing low activity of removing the target glycans; the reasons were stated as following: 1. The entire enzymatic hydrolytic process was operated in the same buffer, glyco-buffer 1. Between steps, the denature procedure – hot water bath – would affect pH, salt concentration and the reaction volume of the reaction system. When enzyme precipitate, the subtract in the active sites may not be released, which in consequences lower subtract quantity; meanwhile, the protein can envelop salt or water in surrounding to influence the reaction environment. Since the hydrolytic approach is an equilibrium-controlled reaction, which means milder conditions [140], the minor alteration in the condition substantially differentiates the activity of each enzyme. α1-2,4,6 Fucosidase was the last exo-acting enzyme into the system, and the reason mentioned above may explain the lower clearance ratio of fucose on these glycosidic moieties. Furthermore, most of the exohydralases are double enzymes that control hydrolysis and reverse hydrolysis. The increasing amount of monosaccharides in the reacting solution could inhibit the activity of these enzymes and trigger the early coming of undesirable equilibrium point in hydrolytic reactions.
Table 4.7 The List of Sugar Structures Identified from Derived Labeling Technique

<table>
<thead>
<tr>
<th>Peak #</th>
<th>Glycan Structure</th>
<th>m/z in Dalton</th>
<th>Peak Area</th>
<th>Peak Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td><img src="image1" alt="Image" /></td>
<td>1255.8755</td>
<td>2795331</td>
<td>4.287</td>
</tr>
<tr>
<td>6</td>
<td><img src="image2" alt="Image" /></td>
<td>1459.1709</td>
<td>418648</td>
<td>0.642</td>
</tr>
<tr>
<td>7</td>
<td><img src="image3" alt="Image" /></td>
<td>1605.0888</td>
<td>3625416</td>
<td>5.56</td>
</tr>
<tr>
<td>8-1</td>
<td><img src="image4" alt="Image" /></td>
<td>1377.2113</td>
<td>16155770</td>
<td>24.777</td>
</tr>
<tr>
<td>8-2</td>
<td><img src="image5" alt="Image" /></td>
<td>1605.1021</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td><img src="image6" alt="Image" /></td>
<td>1662.1943</td>
<td>1092057</td>
<td>1.675</td>
</tr>
<tr>
<td>10-1</td>
<td><img src="image7" alt="Image" /></td>
<td>1362.2843</td>
<td>10540431</td>
<td>16.165</td>
</tr>
<tr>
<td>10-2</td>
<td><img src="image8" alt="Image" /></td>
<td>1808.3703</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10-3</td>
<td><img src="image9" alt="Image" /></td>
<td>1824.3480</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td><img src="image10" alt="Image" /></td>
<td>1808.2474</td>
<td>1016339</td>
<td>1.559</td>
</tr>
<tr>
<td>12</td>
<td><img src="image11" alt="Image" /></td>
<td>1565.2543</td>
<td>5970020</td>
<td>9.156</td>
</tr>
<tr>
<td>14</td>
<td><img src="image12" alt="Image" /></td>
<td>1768.5025</td>
<td>5057791</td>
<td>7.757</td>
</tr>
<tr>
<td>No.</td>
<td>Structure</td>
<td>Identification</td>
<td>Mass (Da)</td>
<td>Charge (m/z)</td>
</tr>
<tr>
<td>-----</td>
<td>-----------</td>
<td>----------------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>15-1</td>
<td><img src="image1" alt="Structure" /></td>
<td>1622.4119</td>
<td>56600</td>
<td>0.087</td>
</tr>
<tr>
<td>15-2</td>
<td><img src="image2" alt="Structure" /></td>
<td>1768.4876</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td><img src="image3" alt="Structure" /></td>
<td>1727.6112</td>
<td>30108</td>
<td>0.046</td>
</tr>
<tr>
<td>19</td>
<td><img src="image4" alt="Structure" /></td>
<td>1784.4987</td>
<td>13846</td>
<td>0.021</td>
</tr>
<tr>
<td>23</td>
<td><img src="image5" alt="Structure" /></td>
<td>2092.7482</td>
<td>437724</td>
<td>0.671</td>
</tr>
</tbody>
</table>

Two out of 18 structures were discovered with terminal α-galactose (Labeled in red) that comprises 0.067% of the total identifications based on peak area. The MALDI results and Mass-Structure Mapping with CFG database are contained in the appendix.

ESI offers better sensitivity and resolution in analyzing small molecules with clearer spectrum and lower noise; however, MALDI-MS/MS is the optimal choice for our study. Since 5 identical injections were introduced to HPLC, each fraction is over 1mL. This large volume of the sample exceeds the carrying capacity of a C18 column on one hand; on the other hand, ammonia formate in the elution buffer increases noise signal. Meanwhile, it may contaminate the ion source and block the ion source needle. In contrast, a small quantity of salt facilitates the ionization efficiency and spikes the sample signals. The only drawback of MALDI-MS/MS is the off-line connection from HPLC because the removal of solvent in fractions is essential for MALDI analysis. Meantime, this off-line design may increase the possibility to introduce random error into the system.
Figure 4.18 The Spectra from HPLC with a Fluorescent Detector
24 peaks were detected and collected through HPLC collector. A total number of 5 injections were made to assure reliable separation ability and determinate retention time in the HILIC column, each injection contains an identical amount of sample.

4.5 Conclusions

In this study, we initially mapped HA-1 and -2 proteinaceous sequences using PMF and achieved a nearly 90% coverage of peptide fragments. Further, a total of 5 potential glycosylation sites were completely confirmed with acceptable peptide-spectra mapping results. Additionally, glycosylation occupancy for each site was various that influenced the quantitation of alpha-galactose experiments. Applying the technique that exoglycosidases coupling with HPLC, 0.067%
of terminal galactose in α1-4 position was quantitated from 100 µg HA-1. A total of 17 carbohydrates were identified from the released sugar moieties. Among those, novel n-glycan structures, high mannose (two structures) and hybrid (three possible forms), were discovered in HA-1 besides the known complex structures in our assumption. The technologies developed in this chapter can offer a novel approach to understanding comprehensively and study the structural and functional properties of any glycoproteins.
CHAPTER 5. Conclusions

5.1 Project Conclusions

The overall aim of this Ph.D. dissertation is to study thoroughly the structures of N-glycans, and the potential outcomes of N-glycosylation to proteins or molecules in biological specimens or biopharmaceuticals (monoclonal antibody and HA-1), and further to comprehend the connection between structure and function. Modifications in N-glycosylation sites are triggered by environmental changes in cells when diseases cause physical and pathological differences. Frequently, these alterations have robust impacts on the structure, function, and pharmakinetics of therapeutic proteins or these markers. Currently, many glycoproteins have been recognized as biomarkers for many cancers and autoimmune deficiencies. However, because of the non-pattern nature of protein N-glycosylation, it has been considered as one of the most complicated post-translational modifications. Meanwhile, glycoprotein frequently has multiple glycosylation sites (HA-1), and various glycoforms in one particular site (mAb). Due to the correlation between N-glycosylation and diseases, there have been a rousing number of studies published on developing novel methods and technologies of mass spectrometer to improve identification and quantification of N-glycoprotein. Comprehensive quantitative analysis of protein N-glycosylation in this dissertation includes the quantitation of glycosylation alteration site-specifically and global or sugar-specific, determination of glycosylation occupancy at each site and identification of glycoforms in therapeutic proteins.

To increase the sensitivity and identified information related to N-glycosylation, analytical techniques with high resolving power are demanded in quantification. Furthermore, an improved sample preparation approach had to be developed to enable analysis of clinical specimens. In Chapter 2, trypsin digestion coupled with HILIC tip-enrichment was applied to rheumatoid
arthritis biomarkers’ screening. N-glycosylation alteration is studied globally and site-specifically, which means the variation of the glycosylation site may associate with either protein expression or glycosylation sites specific changes. Thus, verification of global expression and site-specific alteration is conducted in Western-blot. The complement system plays an influential role in the human immune system against RA. Over 17% of the glycosylation sites identified are from or associated with the complement system. The change of glycosylation occupancy indicates a possible means for distinguishing RA from other immune system deficiencies, and many other applications are available with this method.

The monoclonal antibody has one highly-identified glycosylation site, N274, with multiple glycoforms attached. In Chapter 3, we designed a fast and efficient method combined HILIC enrichment technique higher-energy dissociation- MS², without 1/3 cutoff influence of CID. In other words, HCD receives more essential information on the fragments generated from sugar moieties; this information is very crucial for identifying core-fucose structure. Furthermore, based on our study, manual interpretation is shown to offer more accuracy and various structures in glycan structural analysis than some commercially available software. A total of 13, from 26 overlapped glycoforms, were identified with significant alteration between standard and modified monoclonal antibody samples. Manual interpretation of glycan structure provides a reliable approach to research carbohydrate moiety fully and to establish a structure-function correlation between carbohydrate branches and glycoprotein. This technique can be further applied for site-specific glycan structure analysis. However, MI strategy is time- and man power-consuming, additionally it required adequate knowledge in glycoscience to decipher the structures. For glycoprotein with multiple glycosylation sites, the workload is dramatically increased since the miscleavages from enzymatic digestion exponentially rise the mass set of potential peptide
backbones. Our desire is to optimize the fragmentation database for N-glycan so that software companies are driven to improve accuracy and affordability of their products.

A comprehensive study of therapeutic protein offers better understanding and predicting the efficiency and safety of the product. In Chapter 4, we designed several approaches to verify the proteinaceous sequence of HA-1 and -2, to determine glycosylation sites and glycosylation occupancy for each site, and to quantitate the percentage of one sugar block, galactose. The entire procedure is described as Figure 3.6 in detail. Multi-enzyme digestion was the most comprehensive and highly reproducible technique to offer sufficient coverage, over 80%, in proteomic perspective. The nano-LC-MS/MS mass spectrometric method in combination with stable heavy isotopic labeling was advantageous compared to nano-LC-MS/MS combined with HILIC enrichment. In order to achieve α-galactose quantitation, 2-AB derived glycans associated with exoglycosidase digestion was optimized by sample preparation, HPLC separation, and MALDI detection. Furthermore, the method was qualified with respect to linearity, reproducibility, and robustness. 0.067% of target sugar building block were detected from HA-1.

5.2 Suggested Future Applications

Overall, this dissertation focuses on the development and application of mass spectrometric assays for characterization and quantitation protein N-glycosylation – globally and site-specifically – and a specific monosaccharide, and the techniques of glycan structural determination. These results are envisioned as a potentially valuable approach for uncovering more perspectives of the structure-function relationship between glycan moiety and protein backbone. Many future works are demanded to improve or expand the application of these assays. Currently, there is a high demand for establishing databases of signature signals from specific N-glycans or distinctive therapeutic proteins, especially commercially available software is continually
improving and desperately waiting for that information. The other techniques are potentially omnipotent to quantitative and characterize N-glycosylation and the related area in glycomics and glycoproteomics.
REFERENCES


79. Kawaguchi-Sakita, N., et al., *Serum immunoglobulin G Fc region N-glycosylation profiling by matrix-assisted laser desorption/ionization mass spectrometry can


134. Corradini, C., A. Cavazza, and C. Bignardi, *High-Performance Anion-Exchange Chromatography Coupled with Pulsed Electrochemical Detection as a Powerful Tool to


APPENDICES

Appendix A

24 fractions were collected and analyzed by MALDI-MS/MS, 19 sugar structures were detected from these. All the spectra and results from manual interpretation were attached underneath.

Peak 2

Peak 2-1255.8755
Peak 8

Peak 8-1377.2113
Peak 8-1605.1021
Peak 12

Peak 12-1565.2543
Peak 17

Peak 17-1727.6112
Peak 23

Peak 23-2092.7482
Appendix B

Table 1 Retention Time and Peak Area of Hydrolyzed Monosaccharide Using HPAEC-PAD

<table>
<thead>
<tr>
<th></th>
<th>Fucose</th>
<th>Galactose</th>
<th>Glucose</th>
<th>Mannose</th>
<th>GlcNAc</th>
<th>SA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Con. (nM)</strong></td>
<td><strong>RT</strong></td>
<td><strong>Area</strong></td>
<td><strong>RT</strong></td>
<td><strong>Area</strong></td>
<td><strong>RT</strong></td>
<td><strong>Area</strong></td>
</tr>
<tr>
<td>5</td>
<td>3.46</td>
<td>0.38</td>
<td>6.91</td>
<td>1.06</td>
<td>8.70</td>
<td>3.07</td>
</tr>
<tr>
<td>10</td>
<td>3.45</td>
<td>0.80</td>
<td>6.88</td>
<td>1.894</td>
<td>8.68</td>
<td>5.378</td>
</tr>
<tr>
<td>15</td>
<td>3.43</td>
<td>1.112</td>
<td>6.84</td>
<td>2.702</td>
<td>8.63</td>
<td>7.479</td>
</tr>
<tr>
<td>60</td>
<td>3.40</td>
<td>7.393</td>
<td>6.78</td>
<td>12.118</td>
<td>8.55</td>
<td>34.707</td>
</tr>
</tbody>
</table>

Appendix C

Figure 1 Tandem Mass Spectra Taken from pLabel Showing Matches to Peptides Included Glycosylation Site N₄₈ for HA-1 and -2
The glycosylation occupancies for N48 site is very similar between HA-1 (Top) and HA-2 (Bottom). However, the enzymatic-generate fragment in -1 is bigger than -2.

Figure 2 Tandem Mass Spectra Taken from pLabel Showing Matches to Peptides Included Glycosylation Site N_{251} for HA-1 and -2
Figure 3: Examples of Tandem Mass Spectra Taken from pLabel Showing Matches to Peptides Included Glycosylation Site N\textsubscript{423} for HA-1 and -2.
Appendix D

There are two ways to calculate glycosylation occupancy: 1. Spectra Number. 2. Peak Area. In Appendix D, I attached the graph obtained with peak area marked on top. These numbers are sufficient to calculate the glycosylation occupancy of each site in HA-1. However, the results from peak area are not consistent with the ones from spectra number counting method. Since our column is losing the ability to efficiently separate sample slurry, we believe the spectra number counting method is more accurate.
HA-1 N\textsubscript{32}GT Site Glycosylated Ratio: 16.8%

![Graph showing the calculation of N\textsubscript{32}GT glycosylation occupancy in HA-1 based on peak area.](image)

**Figure 1** The Calculation of N\textsubscript{32}GT Glycosylation Occupancy in HA-1 Based On Peak Area.

HA-1 N\textsubscript{48}AT Site Glycosylated Ratio: 86.2%

![Graph showing the calculation of N\textsubscript{48}AT glycosylation occupancy in HA-1 based on peak area.](image)

**Figure 2** The Calculation of N\textsubscript{48}AT Glycosylation Occupancy in HA-1 Based On Peak Area.
Figure 3 The Calculation of N$_{251}$DT Glycosylation Occupancy in HA-1 Based On Peak Area.

Figure 4 The Calculation of N$_{423}$WT Glycosylation Occupancy in HA-1 Based On Peak Area.
Figure 5 The Calculation of $N_{459}NT$ Glycosylation Occupancy in HA-1 Based On Peak Area.