Comprehensive Characterization of Protein Structures by Liquid Chromatography Coupled with Mass Spectrometry (LC-MS)

A dissertation presented by

Chen Li

to

The Department of Chemistry and Chemical Biology

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the field of

Chemistry

Northeastern University

Boston, Massachusetts

June 3, 2013
Comprehensive Characterization of Protein Structures by Liquid Chromatography Coupled with Mass Spectrometry (LC-MS)

By

Chen Li

ABSTRACT OF DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemistry in the College of Science of Northeastern University

June 3, 2013
Abstract

The characterization of protein structure plays an important role in understanding the biological functions of proteins. Especially in the biotechnological industry, the determination of the protein structure expressed in cell lines can ensure the quality of the product. This dissertation is focused on the application of LC-MS to characterize protein structures.

In Chapter 1, methods that determine the protein structure are introduced. LC-MS, the major tool used in the thesis, is presented with a focus on the key steps involved in the method, including sample preparation, LC separation, and MS detection. Finally, major applications are described.

In Chapter 2, an anti-CD20 monoclonal antibody developed using RNA interference to decrease core fucosylation is comprehensively characterized by LC-MS and compared to commercially-available anti-CD20 Rituximab (MabThera®). As anticipated, < 30% core fucose was found within the RNAi-produced molecule (compared to > 90% with Rituximab). Two mutations, S258Y (fully mutated) and F174I/L (partially mutated), however, were detected in the product of the RNAi-mediated molecule. An alternative LC-MS approach using dimethyl labeling (i.e., 2CH₂ for Rituximab and 2CD₂ for the RNAi-mediated molecule) was developed to further compare the two mAbs and confirm the full sequence with the two mutation sites. The approaches described provide a strategy for comparing biosimilars to innovator products.

In Chapter 3, glycan structures of recombinant glucocerebrosidase (Cerezyme®) are comprehensively characterized and compared to another recombinant glucocerebrosidase
molecule produced by using small interfering RNA (siRNA) to silence the *MGAT1* gene, which expresses N-acetylglucosaminytransferase I (GlcNAc-TI) which catalyzes the formation of complex and hybrid-type glycans. The analysis was achieved by determination of the glycans attached to peptides of the four Asn sites by studying the four glycopeptides with the LC-MS approach. The glycosylation site, occupancy, and glycan distribution were determined by this approach and used for comparison between the two products.

In Chapter 4, the cysteines in the CxxC motifs of CD3 ε, δ, and γ are characterized by LC-MS. CD3ε, δ, and γ contain the common RxCxxCxE motif in the membrane proximal segment, in which the cysteine status is expected to change with the biological function. The CD3s were purified from human T-cells by our collaborators and separated by SDS-PAGE before in-gel digestion. Various enzymatic strategies were tested, and a four-enzyme protocol, including trypsin, LysC, GluC, and PNGase F, was optimized to detect the three CxxC motifs in one run by LC-MS. The disulfide-linked form (oxidized) was detected to be the major species in the three CD3s. Additionally, alkylated forms were identified from reducing and alkylating the sample.

Finally, in Chapter 5, the dissertation is summarized and future works in each project are described.
Acknowledgements

I would like to express the deepest appreciation to all those who provided me help to complete this dissertation.

First of all, I am very grateful to my advisor, Prof. Barry L. Karger, for his advice, guidance, and support in every stage of this research and so to this dissertation. He guides me to be “progressive, professional, (nicely) aggressive, and don’t be afraid”. These words greatly inspired me in my Ph.D career and will continue to encourage me in my future career life.

I am highly indebted to Dr. Shiaw-lin (Billy) Wu, for his supervision, assistance, and support in my dissertation projects. His expertise and knowledge greatly accelerate my growth in the Ph.D career.

My special thanks to the committee members: Dr. William S. Hancock, Dr. Roger W. Giese, and Dr. Michael P. Pollastri for reviewing my dissertation and the valuable suggestions.

I would like to thank the former and present people in our group: Dr. Sangwon Cha, Dr. Dongdong Wang, Dr. Dipak Thakur, Dr. Shujia (Daniel) Dai, Dr. Alexander Ivanov, Dr. Sandor Spisak, Dr. David Bush, Dr. Wenqin Ni, Zhenke (Jack) Liu, Siyang (Peter) Li, Victoria Berger, Somak Ray, Siyuan Liu, Yuanwei Gao, Xianzhe Wang for their help, scientific discussion, and friendship.
Thanks to people in Barnett Institute: Dr. Jim Glick Dr. Suli Liu, Yue Zhang, Dr. Yi Wang, Dr. Janet Lau, Fan Zhang, Fateme Tousi, Dr. Haitao Jiang, Dr. Qiaozhen (Cheryl), Dr. Fangfei Yan for their help.

I would also acknowledge the help from the staff in Department of Chemistry and Chemical Biology and Barnett Institute: Nancy Carbone, Jean Harris, Andrew Bean, Cara Shockley, and Richard Pumphrey.

Finally, I would like to thank my husband and my parents. I’m sincerely grateful for their support, encouragement, patience, and love.
# Table of Contents

ABSTRACT .............................................................................................................................................. 2

ACKNOWLEDGEMENTS ......................................................................................................................... 5

TABLE OF CONTENTS .......................................................................................................................... 7

LIST OF FIGURES .................................................................................................................................... 11

LIST OF TABLES ...................................................................................................................................... 16

CHAPTER 1 OVERVIEW: COMPREHENSIVE CHARACTERIZATION OF PROTEIN STRUCTURES BY LIQUID CHROMATOGRAPHY COUPLED WITH MASS SPECTROMETRY (LC-MS) .................................................................................................. 17

1.1 Abstract ............................................................................................................................................. 18

1.2. Introduction .................................................................................................................................... 18

1.3 Key steps in protein structure characterization by LC-MS ......................................................... 26

1.3.1 Sample preparation .................................................................................................................... 27

1.3.2 LC separation methods ............................................................................................................. 31

1.3.3 Mass spectrometer .................................................................................................................... 39

1.3.4 Data analysis ............................................................................................................................. 46

1.3.4.2 Software data processing .................................................................................................... 52
1.4 Applications of LC-MS in the protein characterization ........................................53

1.4.1 Top-down protein characterization ......................................................................53

1.4.2 Bottom-up method .................................................................................................54

1.4.3 Quantitation ............................................................................................................63

1.5 References ..................................................................................................................67

CHAPTER 2 COMPARABILITY ANALYSIS OF ANTI-CD20 COMMERCIAL
(RITUXIMAB) AND RNAI-MEDIATED FUCOSYLATED ANTIBODIES BY TWO
LC-MS APPROACHES .....................................................................................................83

2.1 Abstract .......................................................................................................................84

2.2 Introduction ..................................................................................................................85

2.3 Results ..........................................................................................................................86

2.4 Discussion ....................................................................................................................105

2.5 Acknowledgment .........................................................................................................107

2.6 Materials and Methods ..............................................................................................107

2.7 References ..................................................................................................................111
CHAPTER 3 CHARACTERIZATION OF SITE-SPECIFIC GLYCAN STRUCTURES IN GLUCOCEREBROSIDASE FROM COMMERCIAL (CEREZYME®) AND RNAI-MEDIATED MOLECULES BY LC-MS.................................................................116

3.1 Abstract .................................................................................................................................117

3.2 Introduction ............................................................................................................................117

3.3 Results and Discussion .........................................................................................................119

3.3.1 Glycosylation site analysis ...............................................................................................119

3.3.2 Glycan structure analysis .................................................................................................121

3.3.3 Characterization of Glycans in RNAi-Mediated Glucocerebrosidase .........................127

3.4 Conclusion ............................................................................................................................134

3.5 Acknowledgment ..................................................................................................................135

3.6 Experimental .......................................................................................................................135

3.7 References ...........................................................................................................................138

CHAPTER 4 ANALYSIS OF DISULFIDE STATUS IN THE TRANSMEMBRANE PROTEINS CD3 E, Δ, AND Γ OF HUMAN T-CELLS..........................................................141

4.1 Abstract .................................................................................................................................142

4.2 Introduction ...........................................................................................................................142
4.3 Results and Discussion ........................................................................................................ 143

4.3.1 CxxC motif in mouse recombinant CD3 delta-1 and delta fragments ........................ 143

4.3.2 CxxC motifs in human CD3 ε, γ, and δ ..................................................................... 146

4.4 Conclusion ...................................................................................................................... 152

4.5 Acknowledgements ........................................................................................................ 153

4.6 Materials and Methods ................................................................................................. 153

4.7 References ..................................................................................................................... 154

CHAPTER 5 SUMMARY AND FUTURE STUDIES ................................................................ 156

APPENDIX ............................................................................................................................ 159
List of Figures

Figure 1.1 Edman reaction. Step one: the reaction of the N-terminal amino acid residue with phenylisothiocyanate (PTC); step two: hydrolysis of the phenylthiocarbamoyl derivative of the peptide to yield a protein of N-1 residues and a free “labeled” amino acid residue. ................................................................. 23

Figure 1.2. Nomenclature for peptide fragmentation in mass spectrometry. ...................... 25

Figure 1.3. The mechanism of the asparagine residue deamidation in H$_2$O$_5$. .................... 29

Figure 1.4 Relationship between band broadening and flow rate in different particle sizes: 5, 3, and 1.7 mm $^6$3 ........................................................ .................................................. 34

Figure 1.5 An example of RP-HPLC gradient for proteomic analysis................................. 36

Figure 1.6 Components of a mass spectrometer. Note that the ion source does not have to be within the vacuum of the mass spectrometer................................................................. 39

Figure 1.7 Schematic of the Orbitrap XL MS$^9$. ................................................................. 45

Figure 1.8 Resolution: The hypothetical (M+2H)$_2$+ m/z clusters of peptide

GTVGTLSSDITR. A. R=100,000; B. R=10,000; C. R=1,000........................................... 49
Figure 1.9 Peptide identification by mass spectrometry: A peptide identification from mass; B theoretical m/z values of CID fragment ions; C peptide identification from CID-MS2.

Figure 2.1 Tryptic map of RNAi-mediated mAb. T1H stands for the first tryptic peptide from the N-terminal heavy chain, T1L for the first tryptic peptide from the N-terminal light chain, pyro-T1H for the pyro-Glu form of T1H, GoF for the glycopeptide with core fucose, T24H* for the deamidated T24H. Similar nomenclature is used for the other peptides.

Figure 2.2 A. Precursor mass of normal (left) and S258Y mutated tryptic peptide (right), theoretical and observed monoisotopic mass are indicated in the figure; B CID-MS2 of the precursor ion from A, normal (top) and S258Y mutated tryptic peptide (bottom).

Figure 2.3 A. Precursor mass of normal and F174L/I mutated tryptic peptide; B CID-MS2 of the precursor ion from A, normal (top) and F174L/I mutated tryptic peptide (bottom).

Figure 2.4 A. Precursor mass of the peptide with CH2 dimethyl labeling (left) and the precursor mass of the peptide with CD2 dimethyl labeling (right). The mass difference is not a constant 4 Da but 80 Da, accounting for the difference between S and Y (76 Da), as indicated in the figure; B CID-MS2 of the precursor ion from A, CH2 dimethyl labeling (top) and CD2 dimethyl labeling (bottom).
Figure 2.5 A precursor mass of the peptide with CH2 dimethyl labeling (left) and the precursor mass of the peptide with CD2 dimethyl labeling (right). The mass difference is not constant 4*3=12 Da but 22 Da, accounting for the difference between F and I or L (34 Da), as indicated in the figure; B CID-MS2 of the precursor ion from A, CH2 dimethyl labeling (top) and CD2 dimethyl labeling (bottom) .......95

Figure 2.6 A precursor mass of 1029.1143 (3+); B CID-MS2 of the precursor in A; C ETD-MS2 of the precursor for the disulfide-linked peptide (Cys22-Cys96). For ETD-MS2, a different charge of the precursor ion, 772.59 (4+), was used for fragmentation ..97

Figure 3.1 The primary structure of Cerezyme® with glycosylation sites in bold. .......119

Figure 3.2 Identification of deglycosylated peptide DLGPTLAD(270)STHHNVR. A. MS of precursor (2+); B. CID-MS2 of precursor 816.91, 2+ ........................................121

Figure 3.3 Identification of glycopeptide T22 (263-277) DLGPTLANSTHHNVR (Man3GlcNAc2Fuc) in Cerezyme. A. MS of glycopeptide (1335.6041, 2+); B. CID-MS2 of 1336.11, 2+ ........................................................................123

Figure 3.4 The glycopeptide assignment for Asn 270 DLGPTLANSTHHNVR in Cerezyme: sum of mass spectrum 24-26.8 Min .................................................................124

Figure 3.5 The glycopeptides assignment in T2 (8-39) SFGYSSVVCVCN(Asn 19)ATYCDSDPPTFPALGTFSR: A. MS of glycopeptide (Man3GlcNAc2) 1501.3118, 3+; B. CID-MS 2 of 1501.97, 3+; C. glycopeptides assignment in T2, sum of mass spectrum 57-58.5 Min .................................................................125
FIGURE 3.6 THE GLYCOPEPTIDES ASSIGNMENT IN T5 (49-74) MELSMGPIQAN(ASN

59)HTGTGLLLTLQPEQK: A. MS OF GLYCOPEPTIDE (MAN3GlcNAc2Fuc) 1282.6150, 3+; B. CID-MS 2 OF 1283.28, 3+; C. GLYCOPEPTIDES ASSIGNMENT IN T5, SUM OF MASS SPECTRUM 50-53 MIN. .................................................................

FIGURE 3.7 THE GLYCOPEPTIDES ASSIGNMENT T11 (132-155) TYTYADTPDFQLHN(ASN

146)FSLPEEDTK: A. MS OF GLYCOPEPTIDE (MAN3GlcNAc2Fuc) 1295.8896, 3+; B. CID-MS 2 OF 1296.55, 3+; C. GLYCOPEPTIDES ASSIGNMENT IN T11, SUM OF MASS SPECTRUM 45-49 MIN. .................................................................

FIGURE 3.8 THE GLYCOPEPTIDES ASSIGNMENT IN T22 (263-277) DLGPTLAN(ASN270)STHHNVR OF RNAI-GCase: A. MS OF GLYCOPEPTIDE (MAN5GlcNAc2) 1424.6275, 2+; B. CID-MS 2 OF 1425.13, 2+; C. GLYCOPEPTIDES ASSIGNMENT IN T22, SUM OF MASS SPECTRUM 24.5-26.5 MIN. .................................................................

FIGURE 3.9 THE GLYCOPEPTIDES ASSIGNMENT IN T2 (8-39) SFGYSSVVCVCN(ASN

19)ATYCDSDFDPPALGTFSR OF RNAI-GCase: A. MS OF GLYCOPEPTIDE (MAN3GlcNAc2) 1501.3075, 3+; B. CID-MS 2 OF 1501.97, 3+; C. GLYCOPEPTIDES ASSIGNMENT IN T2, SUM OF MASS SPECTRUM 57.5-58.5 MIN. .................................................................

FIGURE 3.10 THE GLYCOPEPTIDES ASSIGNMENT IN T5 (49-74) MELSMGPIQAN(ASN

59)HTGTGLLLTLQPEQK: A. MS OF GLYCOPEPTIDE (MAN5GlcNAc2) 1314.9626, 3+; B. CID-MS 2 OF 1314.63, 3+; C. GLYCOPEPTIDES ASSIGNMENT IN T5, SUM OF MASS SPECTRUM 50.5-53.1 MIN. .................................................................
FIGURE 3.11 THE GLYCOPEPTIDES ASSIGNMENT T11 (132-155) TYTYADTPDDFQLHN(ASN146)FSLPEEDTK of RNAi-GCase: A. MS of glycopeptide (Man5GlcNAc2) 1355.2382, 3+; B. CID-MS2 of 1355.90, 3+; C. GLYCOPEPTIDES ASSIGNMENT IN T11, SUM OF MASS SPECTRUM 46.2-48.6 MIN

133

FIGURE 4.1 AMINO ACID SEQUENCE: A. MOUSE RECOMBINANT CD3 DELTA-1; B MOUSE RECOMBINANT CD3 DELTA

144

FIGURE 4.2 CHARACTERIZATION OF PEPTIDE MCQNCVE IN MOUSE RECOMBINANT CD3 DELTA-1: A. PRECURSOR MASS OF MCQNCVE; B. CID-MS2 OF THE PRECURSOR ION FROM A

145

FIGURE 4.3 CHARACTERIZATION OF MMTS ALKYLATED MCQNCVE IN MOUSE RECOMBINANT CD3 DELTA-1: A. PRECURSOR MASS MMTS ALKYLATED MCQNCVE; B. CID-MS2 OF THE PRECURSOR ION FROM A

146

FIGURE 4.4 AMINO ACID SEQUENCES OF HUMAN CD3 E, Γ AND Δ

147

FIGURE 4.5 CHARACTERIZATION OF DISULFIDE-LINKED CXXC PEPTIDES IN HUMAN CD3S: A. MS AND CID-MS2 OF DISULFIDE-LINKED VCENCME IN CD3 E; B. MS AND CID-MS2 OF DISULFIDE-LINKED MCQNCIE; C. MS AND CID-MS2 OF DISULFIDE-LINKED MCQSCVE

149

FIGURE 4.6 CHARACTERIZATION OF IODOACETAMIDE (IAM) ALKYLATED CXXC PEPTIDES IN HUMAN CD3S: A. MS AND CID-MS2 OF ALKYLATED VCENCME IN CD3 E; B. MS AND CID-MS2 OF ALKYLATED MCQNCIE; C. MS AND CID-MS2 OF ALKYLATED MCQSCVE

151

15
List of Tables

TABLE 1.1 Functional groups found in RP-HPLC stationary phases for proteomic analysis. ................................................................. 32

TABLE 2.1 Summary of the identified peptides in the heavy chain and light chain of anti-CD20............................................................................................................. 87

TABLE 2.2 Summary of disulfide identification in rituximab and RNAi-mediated mAb. ... 98

TABLE 2.3 Free cysteine determined at different digestion pHs for rituximab and RNAi-mediated mAb. ................................................................................................. 99

TABLE 2.4 Percent scrambled disulfides at different digestion pHs for rituximab and RNAi-mediated mAb.................................................................................................................. 100

TABLE 2.5 Comparison of glycan distribution for rituximab and RNAi-mediated mAb. . 103

TABLE 2.6 Comparison of common modifications in two anti-CD20 mAb ......................... 105

TABLE 3.1 Glycan comparison between Cerezyme and RNAi-GCase ................................. 134

TABLE 4.1 Human CD3 ε γ δ peptide mapping using four-enzymatic digestion ................. 150
Chapter 1 Overview: Comprehensive Characterization of Protein Structures by Liquid Chromatography Coupled with Mass Spectrometry (LC-MS)
1.1 Abstract

This chapter provides an overview of methods involved in the characterization of protein structures by LC-MS. First, the significance of characterizing protein structures are introduced, as well as technologies commonly used to detect primary, secondary, tertiary, and quaternary structures. Second, key steps in the characterization of protein structures by LC-MS are reviewed, including the sample preparation, LC separation, MS detection, and data analysis. Finally, the application of LC-MS in the characterization of protein structures is presented.

1.2. Introduction

Proteins, large molecules consisting of amino acid residues, carbohydrates, phosphates, etc., perform multiple functions in living organisms. To understand the function of proteins, the characterization of their structures plays a critical role. In this section, the definition of protein structure is introduced and then analytical methodologies involved in the characterization of protein structure are described.

A protein has a primary, secondary, tertiary, and even quaternary structure. The primary structure is the amino acid sequence. Most proteins consist of 20 different L-amino acids, which are connected by peptide bonds with different combinations. The amino acid sequence of a protein is mainly determined by the gene which expresses that protein. For example, the human protein amino acid sequence database was converted from the human genome. The secondary structure consists of local three dimensional interactions between amino acid residues; there are three basic units of secondary structure: α helix, β strand, and turns, which
are formed based on the amino acid residues. For example, proline and glycine are common in turns but seldom in a helical conformation. Hence, the secondary structure can often be predicted from the primary structure. There is software available to predict secondary structures, such as DEFINE\(^1\), DSSP\(^2\), and STRIDE (protein)\(^3\). In the tertiary structure, the overall shape of a protein is formed by the further folding of the \(\alpha\) helix, \(\beta\) strand, and turns. The conformation of a protein fluctuates; it is stabilized by disulfides bridges, hydrophobic interactions, charge-charge interactions, van der Waals interactions, and hydrogen bonds. At the beginning of 2013, \(~82,000\) sets of native protein tertiary structures have been deposited into Protein Data Bank (PDB)\(^4\). In PDB files, the position of every atom in the protein is recorded by measuring the \(x\), \(y\), and \(z\) coordinates of that atom. The quaternary structure refers to the interactions between multiple polypeptide chains. Proteins that contain more than one chain possess quaternary structures.

The correct structure of a protein, including amino acid sequence, post translational modifications (PTMs), and three-dimensional structure, is essential to its function. If mutation occurs in a gene, the amino acid sequence of the protein expressed from the gene would be changed, and the protein may not fold into its native structure; hence, it can lose its biological function. For example, the mutations in collagen genes give rise to many defects in their proteins, resulting in severe diseases, such as Osteogenesis imperfecta\(^5\) and Ehlers-Danlos syndrome\(^6\). The gene determines the amino acid sequence of a protein; however, the translation product of the gene may not be the final form of the mature protein, since the polypeptide chain can experience additional reactions called PTMs. These PTMs include the removal of “pre” and “pro” sequences, the addition of new groups such as carbohydrates and phosphate, and the modification of existing amino acid residues such as the formation of
disulfide linkages. Almost all of these PTMs are critical to the structure and function of a protein. For example, if the disulfide linkage is not correct, the protein would not fold into its native structure; hence, it may lose its function. As an example, aspergillus niger phytase is deactivated if its five disulfides are reduced. Soybean Bowman-Birk inhibitor (BBI), in contrast to most proteins, has exposed hydrophobic regions and charged residues containing bound water molecules in its interior; however, it is stable against heat and chemical denaturants because of its seven disulfides.

In the biopharmaceutical industry, structure is essential to the safety and efficacy of a recombinant therapeutic protein. Unlike chemically synthesized products, recombinant proteins are derived from living sources which can result in heterogeneities of the products. Furthermore, approaches for metabolic engineering of host cells are applied to enhance the therapeutic usages of products or improve their manufacturing processes, such as adding synthetic small interfering siRNA (siRNA) directly to the manufacturing cell line in the bioreactor to initiate RNA interference (RNAi). Details of products expressed under RNAi are explained in chapter 2, 3, and appendix of this dissertation. All of these can bring difficulties to characterize structures of protein products. There are many concerns about the structure which might affect the safety and efficacy of a recombinant product, such as the variation of amino acid sequence, disulfide scrambling, undesired glycosylation, deamidation, and oxidation. All of these modified structures may inactivate the function of the product and result in immunogenic responses; hence, the product would not be approved by the U.S. Food and Drug Administration (FDA). On March 23, 2010, President Obama signed a law to “create an abbreviated licensure pathway for biological products that are demonstrated to be ‘biosimilar’ to or ‘ interchangeable’ with FDA-licensed biological product”. Since
recombinant therapeutic proteins are expressed in living organisms which can result in heterogeneities, it is hard to manufacture the exact same product in different batches. Additionally, with the limitation of current analytical technologies, it can be difficult to demonstrate that a protein is the same as any other one. Hence, the term "biosimilar" emerged, which means that the proposed product should be as similar as possible to the reference drug in structure, function, toxicity, human pharmacokinetics (PK), human pharmacodynamics (PD), efficacy, and safety\textsuperscript{13}. Among the parameters, structure characterization is the first and one of the most important steps to support the comparability of the reference drug to the biosimilar product. Well-characterized and comparable biosimilar candidates can potentially reduce the requirements of clinical data and approval costs.

Currently, there are many analytical tools to characterize protein structure. For the primary structure, DNA/RNA sequencing is one of the most effective methods to determine the sequence of a protein. By translating the order of codon sequences from the known DNA sequence of a gene, one can obtain the primary amino acid sequence of the protein. Many protein sequences in the UniProtKB/Swiss-Prot database have been generated in this way. Direct sequencing of a protein is also widely used. Edman sequencing and mass spectrometry are two major methods of the direct protein sequencing.

Traditionally, Edman sequencing was the method of choice\textsuperscript{14}. The Edman reaction consists of two steps, shown in Figure 1.1. First, the free amino group at the N-terminus of a polypeptide reacts with phenylisothiocyanate (PITC) to form a phenylthiocarbamyl (PTC) derivative under basic conditions (~pH 9.0), and then the phenylthiocarbamoyl derivative hydrolyses in aqueous acidic solution (trifluoroacetic acid) to yield a phenylthiohydantion (PTH)-N-
terminal amino acid derivative. The remaining polypeptide chain with the new N-terminal amino acid is available for the next cycle reaction with PITC. The PTH-amino acid derivative is subsequently separated using chromatography or electrophoresis, compared against a standard, and identified by sequencer software\(^\text{15}\). The Edman sequencer is able to work with 10-100 picomoles of peptides. The major drawback of this technology, and the reason it is used far less today, is that this level is much higher than what is achievable by mass spectrometry.
Figure 1.1 Edman reaction. Step one: the reaction of the N-terminal amino acid residue with phenylisothiocyanate (PTC); step two: hydrolysis of the phenylthiocarbamoyl derivative of the peptide to yield a protein of n-1 residues and a free “labeled” amino acid residue.

Mass spectrometry is the other major method to directly sequence a protein. With the development of new techniques, mass spectrometry can achieve increasingly accurate mass determination, sensitive MS/MS, and rapid speed. Though it has difficulties with isobaric
masses, such as leucine and isoleucine, it can analyze modified N-termini, which is not possible in Edman sequencing.

De novo sequencing by mass spectrometry can determine a peptide without prior knowledge of its sequence. This method requires preenzymatic-cutting the protein into smaller peptide pieces. Sequencing is mainly dependent on interpretation of fragment ions which are generated by low energy collision-induced dissociation (CID) or electron-transfer dissociation (ETD). In CID fragmentation, peptide bonds are broken into b and y ions: the positively charged ion extending from the N-terminus is the b ion, and that from C-terminus is the y ion, as shown in Figure 1.2. The nomenclature for fragment ions was first proposed by Roepstorff and Fohlman\textsuperscript{16}, and subsequently modified by Johnson et. al\textsuperscript{17}. Peptides do not typically fragment in CID sequentially, but randomly, and every peptide bond is possible to break; therefore, a CID-MS2 spectrum is the sum of all fragment possibilities. The sequence of a peptide can be predicted by differences of masses between the fragment ions. On the other hand, ETD transfers an electron from a radical anion to a protonated peptide, inducing fragmentation of that peptide. In ETD fragmentation, N-Cα in the peptide backbone is broken to produce c and z ions, as shown in Figure 1.2. Manual de novo sequencing requires expert analysis, and it is time consuming; some software has been developed, such as PEAKS\textsuperscript{18}, Sequit\textsuperscript{19} and PepNovo\textsuperscript{20}. The accuracy of the result is limited by the quality of the fragment ions, which is the major concern for de novo sequencing.
The most important application for mass spectrometry in the primary structure determination is to compare the results with known amino acid sequences. Basically, mass spectrometers determine the mass of a protein/peptide and MS/MS fragment ions, and then the experimental data are compared with the known primary structure of that protein, or the data is searched against a database containing that protein. When the mass of a protein/peptide matches a certain protein/peptide in the database, fragment ions in the MS/MS scan are compared with the theoretical fragmentation spectrum of that protein/peptide; software usually lists the protein/peptide that best fits the precursor mass and fragment ions. In a complex digest, the software finds the peptide first and then the protein that contains it. For a single protein, if the gene sequence is known, scientists can quickly examine if the amino acid sequence is correct by MS.

The three-dimensional structure of a protein can be determined at different levels by many physical methods, including X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, hydrogen/deuterium exchange mass spectrometry (H/DX-MS), circular dichroism (CD), UV-visible absorbance spectrometry, and fluorescence. X-ray crystallography and NMR spectroscopy determine the structure from an atomic level; these
methods can measure the position of atoms, such as carbon, nitrogen, and oxygen. By combining the positions of most atoms, the “picture” of a protein can be established. The other methods listed above provide mainly information on specific regions of a protein. CD reveals the distribution of different secondary structures. UV-visible absorbance helps to identify metal ions, aromatic groups or co-factors attached to proteins. Fluorescence is used to probe tryptophan or other attached fluorophores. These methods cannot yield the position of each atom in a protein; however, they are quick, easy to operate, and do not have extensive data processing. Thus, they are widely used.

In this section, methods involved in the protein structure characterization are summarized. In the following sections, the in-depth analysis of protein structure by LC-MS will be presented. In the title of this dissertation, the “comprehensive characterization” of protein structure is narrowed to apply LC-MS based methods; it should be pointed that these methods cannot cover every structural feature of a protein.

1.3 Key steps in protein structure characterization by LC-MS

LC-MS is a powerful technique which combines LC separation and MS detection. In current mass spectrometers, m/z values of analytes can be determined in a full MS scan, and their structures can be analyzed in a tandem MS scan. LC-MS plays an increasingly significant role in the analysis of complex samples. It was reported that 10,000 proteins were detected in a 110 nL sample of microdissected cells from human colonic adenoma by Matthias Mann’s group. The number of analytes detected in this experiment was even higher, since they digested proteins into peptides, which were then detected by LC-MS.
The experimental workflow of LC-MS is designed based on the goal of study. To detect analytes of interest, the key steps for characterization of protein structure by LC-MS are discussed next, including sample preparation, separation methods, mass spectrometers, and data processing.

1.3.1 Sample preparation

The purpose of sample preparation is to optimize a sample for LC-MS analysis, in order to obtain the information of interest. Therefore, the procedures of sample preparation are mainly dependent on the properties of the samples and the experimental goals. For example, glycans in a monoclonal antibody (mAb) can be determined by LC-MS after removing glycans either on-line or off-line by PNGase F\textsuperscript{32-34}. However, if analyzing glycans in a complex sample, such structures in low abundant proteins might not be determined; hence, additional sample preparations may be required. In this section, sample preparation is introduced with respect to enrichment or fractionation, derivatization, and proteolysis.

1.3.1.1 Enrichment and fractionation

The purpose of enrichment is usually to increase the relative concentration of the analytes of interest, remove or reduce compounds that might interfere with the analysis. Enrichment can be based on affinity interaction or other separation principles that are orthogonal to the LC method directly coupled with the mass spectrometer. Additionally, these methods can also be used as fractionation. The complex sample is separated, fractions are collected, and then each fraction is run by LC-MS, respectively. As mentioned above, the selection of enrichment or fractionation is dependent on the experimental goal.
Characterization of phosphoproteins can be an example to show the importance of enrichment. Since phosphorylation is a transient and reversible modification, the abundance of phosphoproteins is typically low. In order to study the phosphoproteome, most researchers apply phosphospecific enrichment in sample preparation. Thingholm et al. summarized the affinity enrichment methods used in phosphoproteomics, including immunoaffinity chromatography for phosphotyrosines, titanium dioxide (TiO2) immobilized chromatography, and immobilized metal affinity chromatography (IMAC). As another example, glycoproteins can be enriched from human serum using a multi-lectin affinity column, e.g. agarose-bound ConA, WGA, and JAC.

Other methods, including gel separation, reversed phase (RP), ion exchange (IEC), size exclusion chromatography (SEC), and their combinations, are also widely used in both enrichment and fractionation. As an example, gel separation is usually applied at the protein level, including 1D SDS-PAGE and 2D-PAGE, and then the protein bands are digested in the gel and detected by LC-MS. For fractionation, the entire gel is usually cut in rectangular sections, and then each section is digested and analyzed by LC-MS, separately.

1.3.1.2 Derivatization

Derivatization is a widely used technique in LC-MS analysis. The derivative has new properties which improve the LC-MS detection qualitatively or quantitatively.

Alkylation is one of the most popular derivatization methods in the preparation of protein samples. Protein alkylation irreversibly derivatizes free sulfhydryls using chemical reagents,
such as iodoacetamide\textsuperscript{50}, iodoacetic acid\textsuperscript{51}, and N-ethylmaleimide\textsuperscript{52}. In proteomics, alkylation is a routine step applied after reduction to prevent cysteines from reforming disulfides. Additionally, it is also used to reduce thiol-disulfide exchange reactions, and to locate free cysteines in a protein\textsuperscript{53}.

Isotopic labeling is a method used to track the passage of a species through a reaction, metabolic pathway, or location in a cell. An example of isotopic labeling is to monitor sample preparation induced deamidation using $^{18}$O water\textsuperscript{54}. Asparagine (Asn) residues may be converted to the mixture of aspartic acid and isoaspartic acid (isoAsp) during sample preparation\textsuperscript{55}. Since $^{18}$O makes the mass shift +3 Da from Asn to Asp or isoAsp, while the pre-existing deamidation in the protein increases the mass of 1 Da, the sample preparation induced Asp or isoAsp can be recognized and monitored, as shown in Figure 1.3\textsuperscript{56}.

**Figure 1.3.** the mechanism of the asparagine residue deamidation in H$_2^{18}$O\textsuperscript{55}.
Moreover, labeling of proteins and peptides with stable heavy isotopes, such as deuterium, $^{13}$C, $^{15}$N, and $^{18}$O, is widely used in quantitative proteomics. Such isotopes can be incorporated metabolically in cells and small organisms, or post-metabolically in proteins and peptides by chemical or enzymatic reactions$^{42}$. Discussion of isotopic labeling is described in a later section.

### 1.3.1.3 Proteolysis

To obtain detailed information of structure, proteins are enzymatically digested into smaller peptides using proteases, such as trypsin, GluC, pepsin, etc. Subsequently these peptides are separated and analyzed by LC-MS.

The selection of enzymes is determined by the experimental goal and the protein structure. The most commonly used enzyme is trypsin, which specifically cleaves at the carboxyl side of the amino acids lysine and arginine, generating peptides with appropriate size for LC-MS detection. It is often the first choice for characterization of a single protein or a complex protein mixture, since trypsin specifically cleaves at two charged amino acid residues (R and K), which generates peptides with appropriate sizes and charges. Other enzymes are also applied to produce desired peptides. For instance, to increase charges or sizes of peptides, LysC might be used. Our group digested proteins by LysC to detect PTMs using the ETD-MS$^2$ method, since ETD generates more fragment ions from a highly charged peptide$^{57}$. Additionally, multiple enzymes may be applied in one sample, such as PNGase F, which specifically releases N-linked oligosaccharides. This digestion procedure is used together with trypsin to identify peptide backbones of glycopeptides. Our group used multi-enzyme
digestion to map disulfides in a recombinant protein with a cystine knot\textsuperscript{58}, in order to generate appropriate disulfide-linked species for the MS analysis.

### 1.3.2 LC separation methods

Separating analytes before MS can provide time to analyze the structure of analytes, increase the dynamic range and sensitivity of detection, and generate more comprehensive data. In a complex sample, such as human plasma, the concentration range of proteins can be $\sim$10-12 orders of magnitude; there are 15,747 proteins reported in Plasma Proteome Database by March 25, 2013\textsuperscript{50}. LC separation, which elutes different analytes at different times, is critical to identify more proteins. In a single protein sample, LC separation is also necessary, since the tryptic digest usually contains multiple peptides, plus PTMs, the number of analytes is very large. For example, there are $\sim$60 peptides in the tryptic digest of a mAb, plus glycosylation, N-terminal pyroglutamation, C-terminal processing, oxidation, deamidation, etc\textsuperscript{59,60}, the number of peptides can be well over 100 in the sample.

In this section, LC separation methods for protein analysis are introduced from two aspects: 1. LC applications in directly coupling with MS; 2. applications as multi-dimensional separations before LC-MS.

#### 1.3.2.1 LC separation methods directly coupled with MS

In proteomic analysis, the most common mode of LC coupled with MS is reversed phase HPLC (RP-HPLC). RP-HPLC uses a hydrophobic stationary phase. Analytes interact with stationary phases are based on their hydrophobicities, and they are eluted by increasing
concentration of the organic solvent. The most widely used RP-HPLC columns contain packed stationary phase. The column packings are formed by chemically bonding functional groups to a silica surface; the commonly used functional groups in proteomic analysis are listed in Table 1.1.

**Table 1.1** Functional groups found in RP-HPLC stationary phases for proteomic analysis.

<table>
<thead>
<tr>
<th>Functional Group</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3, C4, C5</td>
<td>Used primarily for separations of proteins</td>
</tr>
<tr>
<td>C8</td>
<td>Used in both proteins and peptides; most commonly used columns</td>
</tr>
<tr>
<td>C18</td>
<td>Used primarily for separations of peptides; most commonly used columns</td>
</tr>
<tr>
<td>Phenyl</td>
<td>Used mainly for a change in selectivity</td>
</tr>
</tbody>
</table>

The sizes of packed beads in a RP-HPLC analytical column are usually 3 and 5 µm; sub 2 µm is widely used in ultra high performance liquid chromatography (UHPLC). The relationship between the resolving power of a RP-HPLC and various parameters that cause peak broadening is described by Van Deemter equation

\[ HETP = A + \frac{B}{\mu} + C\mu \]  

**Equation 1.1**

Where HETP, height equivalent to a theoretical plate, describes the resolving power, \( \mu \) is the linear velocity, A is eddy-diffusion parameter, which is related to channeling through a non-ideal packing. In a packed RP-HPLC column, a smaller diameter of the packing beads with
less dispersion of particle size can greatly reduce the contribution of eddy-diffusion to HETP. B is the mobile phase diffusion coefficient of the solute, which describes the longitudinal diffusion of analytes. The contribution of longitudinal diffusion can be minimized by increasing the linear velocity. C, consisting of mobile and stationary phase mass transfer coefficients, can be minimized by using low viscous mobile phase, a "thin" coating of the stationary phase on the solid support, and small packing beads. The van Deemter equation can be further expanded to

\[ HETP = 2\lambda d_p + \frac{2GD_m}{\mu} + \frac{\omega d_p^2\mu}{D_m} + \frac{Rd_f^2\mu}{D_s} \]

Equation 1.2

where \( \lambda \) is particle shape; \( d_p \) is particle diameter; G, \( \omega \), and R are constants; \( D_m \) is the diffusion coefficient of the mobile phase; \( d_f \) is the film thickness; \( D_s \) is the diffusion coefficient of the stationary phase. From Equation 1.2, the column with smaller packing beads (\( d_p \)) will have a smaller HETP, and separates better. However, the increment of back pressure is inversely proportional to square of the particle diameter,

\[ \Delta P \propto \frac{1}{d_p^2} \]

Equation 1.3

The smaller packing beads will lead to higher back pressures. Ultra-performance liquid chromatography (UPLC), which was introduced by Waters Corporation in 2004, is the first and most popular instrument to perform high pressure separation. In Waters’ ACQUITY UPLC system, columns are packed with small particles (1.7 – 1.8 \( \mu \)m), and the mobile phase can be precisely delivered at pressures up to 1030 bar (15, 000 psi). The theory behind
UHPLC is also the van Deemter equation, shown in Equations 1.1 and 1.2. The relationship between HETP and flow rate can also be described in Figure 1.4. The optimum flow rate is reached when the value of HETP is at the lowest point. By decreasing the particle size of packing beads, the resolving power of a column is increased; meanwhile, the optimum flow rate can be in a wider range, as shown in Figure 1.4 for the particle size of 1.7 μm. Hence, the resolving power of the column is not compromised when running in a higher flow rate, which is important to decrease separation time. The major concern about UPLC so far is the high cost for many users. The upgrading from HPLC to UHPLC will be a trend in the future.

![Figure 1.4](image.png)

**Figure 1.4** Relationship between band broadening and flow rate in different particle sizes: 5, 3, and 1.7 μm.63

In proteomic analysis, the mobile phase of RP-HPLC mainly consists of water and acetonitrile (ACN). Other components may also be added, such as formic acid (FA), phosphoric acid, acetic acid, trifluoroacetic acid (TFA), or salts, to assist the separation of samples. For LC-
MS, the most suitable additive is FA or acetic acid in the positive mode; salts such as ammonium formate or ammonium acetate may be applied, when the analytes are sensitive to acids, or the optimal separation does not occur at the low pH. Salts also permit both the positive and negative ion mode detection. Some additives are, however, not compatible with MS, such as phosphoric acid, sulfuric acid, TFA, and surfactants. The inorganic acids such as phosphoric acid and sulfuric acid may cause corrosion of metal parts in the ion source, while surfactants tend to suppress the ionization of analytes. Although, TFA is very suitable for peptide separation, it also shows suppression of signals in MS, because of its high conductivity and ion-pairing ability with basic analytes\textsuperscript{64,65}. Hence, researchers try to avoid using TFA. If TFA must be used, it is usually applied in a low percentage 0.05 or even 0.02 %.

In a typical RP-HPLC-MS proteomic analysis, mobile phase A consists of HPLC grade water and 0.1% formic acid, and mobile phase B contains HPLC grade ACN and 0.1% formic acid. The gradient is important because of the complexity of proteomic samples; it usually starts with 2-5% B, and then mobile phase B linearly increases to 60% in 60 or more minutes; analytes are eluted in the gradient, as shown in Figure 1.5. At the end of the gradient, mobile phase B drops to the starting point (2-5% B), and column equilibration begins. The equilibration usually takes 10 to 20 column volumes of the solvent.
Figure 1.5 An example of RP-HPLC gradient for proteomic analysis.

The time of the gradient is dependent on the complexity of the sample. For a complex sample, the longer gradient provides sufficient time for MS and tandem MS detection, which can increase the detection sensitivity and identify more analytes. Matthias Mann’s group used a 480 min gradient to identify proteins from yeast and HEK-293 cells.

In RP-HPLC-MS system, analytes are eluted according to the hydrophobicity; hydrophilic analytes have less retention, or may not be retained at all. Hence, RP-HPLC-MS has a limitation in separation and detection of hydrophilic analytes.

Porous graphitized carbon (PGC) chromatography has advantages for separation of hydrophilic analytes. PGC has unique properties as a stationary phase. Compared to conventional LC packings, such as bonded-silica beads (C18), PGC’s retention is very long for low polar analytes; however, PGC has good separation and selectivity towards polar and structurally related compounds. Glycans which are not well retained in the C18 column can be separated and detected by PGC-HPLC-MS. Additionally, researchers also use PGC columns
to separate short glycopeptides digested by pronase. This enzyme can digest the protein to nearly individual amino acids; hence, the properties of these glycopeptides are then mainly determined by attached glycans. Recently, PGC is also used on Agilent’s mAb-glyco-chip for trapping deglycosylated proteins and separating and analyzing by MS the glycans.

Hydrophilic interaction chromatography (HILIC) is also widely used to separate hydrophilic analytes. It combines a polar stationary phase with an aqueous / polar organic mobile phase, such as water / ACN, in which water is the stronger eluting solvent. HILIC is well compatible with ESI-MS. It obtains sensitive detection in ESI-MS, together with effective on-column retention for highly hydrophilic, ionic, and polar analytes. Moreover, it is possible for HILIC to have higher flow rates because of the lower back pressure using low aqueous / high organic solvent mobile phase. The application of HILIC in protein analysis is not as popular as RP-HPLC, since the protein separation is well established in the RP-HPLC system, and proteins or peptides are not readily dissolved in the mobile phase with predominantly the polar organic solvent. However, researchers usually use HILIC-ESI-MS to determine glycans or glycopeptides. As such, it can be an orthogonal method to PGC. In contrast to PGC chromatography, hydrophobic analytes are unretained or eluted earlier in HILIC. The retention in HILIC can be correlated to the monosaccharide compositions of the glycans by multiple linear regressions, whereas no adequate model is obtained for PGC chromatography, indicating the significance of the three-dimensional structure of the analytes for retention in the PGC column. Furthermore, HILIC based ultra-high performance liquid chromatography (UHPLC) is commercially available; hence, HILIC can be a choice for glycan analysis.
1.3.2.2 LC methods as the multiple dimensional separation for LC-MS

Multiple dimensional separations before LC-MS for peptide mixture are widely used. These methods usually have orthogonal separation principles in LC; hence, a higher level of separation and loading capacity can be achieved, which is helpful when analyzing complicated mixtures, such as the proteomics from a cell lysis. Besides the conventional off-line separation methods, Yates’ group developed multidimensional protein identification technology (MudPIT) for on-line analysis of proteins from a wide range of sources\textsuperscript{72-76}. This method has been used to identify post-translational modifications (PTMs)\textsuperscript{76}, as well as quantitatively compare of proteins expressed in cell cultures using stable isotope labeling by amino acids (SILAC)\textsuperscript{72,77}. In this technology, the first dimension is normally a 3-5 cm strong cation exchange (SCX) column, and the second dimension is RP-HPLC; the system is efficient at removing salts and is well compatible with MS. Since SCX has a better loading capacity than RP-HPLC, it behaves as not only a second separation, but also a peptide enrichment step. By running the salt gradient in SCX, peptides are loaded onto the RP-HPLC column and then further separated by an ACN gradient. A single MudPIT analysis is run between 6 and 24 hours\textsuperscript{78}.

Furthermore, the second dimension separation can be the same chromatography as the LC method directly coupled with MS, but run under different conditions. For example, researchers have used C18 RP-HPLC in a basic condition to fractionate peptides prior to C18 RP-HPLC-MS analysis, operated in an acidic condition\textsuperscript{40}. Peptides contain different charges under different pH conditions, resulting in different hydrophobicities; hence, the resolution of separation can be increased.
1.3.3 Mass spectrometer

The mass spectrometer measures the mass to charge ratio of ionic species. MS has shown a great number of applications for the analysis of proteins, peptides, drugs, metabolites, carbohydrates, DNA, and other relevant molecules. There are multiple mass spectrometer instruments available in the market; the selection of the instrument is based on the experimental goal. Many of the latest mass spectrometers are designed to be applied in multiple experimental goals; therefore, the selection of MS scan methods in one instrument is also significant.

In this section, the mass spectrometer is introduced in terms of its operation, which includes: 1. introducing samples; 2. vaporizing and ionizing molecules; 3. separating charged ions; 4. detecting ions; 5. recording signals as a mass spectrum, shown in Figure 1.6.

**Figure 1.6** Components of a mass spectrometer. Note that the ion source does not have to be within the vacuum of the mass spectrometer
1.3.3.1 Sample introduction and ionization

The most common methods of sample introduction include direct insertion of a probe or plate, used in matrix-assisted laser desorption/ionization (MALDI)-MS, and infusion or injection of samples into the ion source, used in ESI-MS.

Sample ionization is realized in the ion source. There are multiple ionization methods, such as electron ionization, atmospheric pressure chemical ionization, and atmospheric pressure photo ionization, all of which are comparatively "hard" and mainly used to ionize small molecules. Proteins, peptides, or other biomolecules (especially large molecules) are ionized by "soft" ion sources, such as MALDI and ESI.

MALDI-MS was first introduced in 1988 by Tanaka, Karas, and Hillenkamp\textsuperscript{79,80}. It is a widely used analytical technique for peptides, proteins, and other biomolecules. The exact mechanism of MALDI is still unknown. Sample ions produced from this ion source often contain one charge. MALDI can ionize molecules with the practical mass range up to 300 kDa without fragmentation. It is sensitive to low femtomoles, tolerant to salts in millimolar concentrations, and suitable for the analysis of complex mixtures\textsuperscript{52,81}. The most widely used mass spectrometer coupled with MALDI is time-of-flight mass spectrometer (TOF). The downside of MALDI is mainly the interference of matrix background when the molecular weight of analysis is less than 700 Da\textsuperscript{82}.

ESI is also a commonly used ion source in the analysis of proteins, peptides, and other biomolecules. Different from MALDI, it generates not only singly-charged ions but also
multiply-charged ions, and it is compatible with LC for online analysis, as well with capillary electrophoresis (CE) and gas chromatography (GC). ESI produces gaseous ionized molecules directly from a liquid solution\textsuperscript{83}. It is operated by creating a fine spray of highly charged droplets in the presence of an electric field. The sample solution eluting from LC is sprayed from a region of strong electric field at the tip of a nozzle maintained at a potential from 700 V to 5000 V, and then dispersed into charged droplets. Meanwhile, either dry gas, heat, or both are applied at the atmospheric pressure to help evaporate the droplets\textsuperscript{84, 85}. In a low flow rate LC, nanoelectrospray ionization (nanoESI) is applied, in which the spray needle is very small (around 10 µm orifice), and it is positioned close to the entrance of the mass analyzer. NanoESI allows for a higher sensitivity and higher efficiency of the ion transmission to a mass analyzer.

1.3.3.2 Separation of charged ions

The separation of charged ions is realized in a mass analyzer by the m/z ratio. For many years, researchers have been improving the performance of mass analyzers to obtain higher mass accuracy, higher resolution, wider mass range, faster scan speed, and improved capabilities of tandem mass analysis. Currently, there are basically two types of mass analyzers: one is operated to separate ions by space, such as quadrupole (Q) and time of flight (TOF), and the other is operated to separate ions by time, such as an ion trap, Fourier transform ion cyclotron resonance (FT-ICR), and Orbitrap. Additionally, to maximize the performance, more and more commercial mass spectrometers consist of multiple mass analyzers, such as triple quadrupole (QQQ), Q-TOF, linear ion trap (LTQ)-FT-ICR, and LTQ-Orbitrap XL.
**Quadrupole**

The quadrupole, one of the most common mass analyzers, consists of multiple electrodes. When direct current (DC) and radio frequency (RF) voltages are applied to the electrodes, an oscillating electric field is produced, and ions with selected m/z values will have the correct oscillatory path in the RF field to pass through the quadrupole and be either detected or move to a new mass analyzer region. Others ions will be rejected. Hence, by varying the DC/RF voltages, ions are scanned. The quadrupole mass spectrometer has good reproducibility, and it is a relatively low-cost instrument. However, it has a low resolution, and it is not compatible with pulsed ionization methods such as MALDI. Additionally, the single quadrupole by itself is not suitable for tandem mass analysis.

Some mass spectrometers are designed to combine a quadrupole with a TOF to establish a Q-TOF, or attach a quadrupole with two additional quadrupoles to form a triple quadrupole mass spectrometer. These combinations make mass analyzers complementary to each other. As an example, in the triple quadrupole mass spectrometer, each quadrupole has a separate function: the first quadrupole (Q1) is used to scan ions across a pre-set m/z range to select an ion of interest. The second quadrupole (Q2), also known as the collision cell, fragments ions. The third quadrupole (Q3) serves to analyze the fragment ions generated in the collision cell (Q2). In proteomic analysis, triple quadrupole mass spectrometer is powerful in quantitation under the scan method of select reaction monitoring (SRM), in which Q1 and Q3 is set to scan selected masses and Q2 acts as a collision cell; hence, only specific fragment ions from a certain precursor are detected in this mode. The Q1 and Q3 work as double filters to reduce
noises, which can improve the limit of detection and quantitation. The major drawback of triple quadrupole is that it cannot realize high resolution mass detection.

**TOF**

The TOF is a widely used mass analyzer. It measures the time that ions take to move from the ion source to the detector in the same amount of kinetic energy. See equation 1.4 below:

$$ t = L \sqrt{\frac{m}{z}} \frac{1}{2U} $$

Equation 1.4

In this equation, $t$ is the time that an ion takes to move from the ion source to the detector, $L$ is the distance over which the ion travels, $U$ is the electric potential difference of the time-of-flight tube, $m$ is the mass of the ion, and $z$ is the charge state of the ion. $L$ and $U$ are constants in an instrument; therefore, the time is only determined by the m/z of the ion. The TOF is one of the fastest MS analyzers, and it has the highest practical mass range. It is also the choice for the majority of MALDI-based mass spectrometer systems, since it is suitable for pulsed ionization methods. For applying the tandem MS analysis, an additional TOF is attached with the first TOF to establish a TOF/TOF, or the TOF is hybrid with a quadrupole to develop a Q-TOF. In the Q-TOF, the quadrupole acts the same as any simple quadrupole analyzer to scan ions across a specified m/z range, and it can also selectively isolate a precursor ion and direct it into the collision cell. Compared with the single TOF or quadrupole mass analyzer, Q-TOF gains more stable ion injection, and higher resolution and accuracy.
Ion trap

The ion trap mass analyzer consists of a ring electrode and two end caps. It was developed at the same time as the quadrupole mass analyzer by Wolfgang Paul. In an ion trap, ions are dynamically stored in the analyzer. With scanning of DC and RF, ions are resonantly excited and therefore ejected through small holes to a detector. Meanwhile, by ejecting all other ions except the desired ones, the ion trap can isolate ions and then fragment them. Hence, the major advantage of ion trap is that tandem MS can be performed quickly without having multiple analyzers. Additionally, the ion trap can be coupled with either a continuous injection, such as ESI or EI, or a pulsed injection, such as MALDI. In a linear ion trap (LTQ), which confines ions radially by a two-dimensional RF and axially by stopping potentials applied to end electrodes, both CID and ETD are compatible, such as LTQ XL ETD developed by Thermo Fisher Scientific. In my dissertation work, LTQ XL ETD was used to characterize disulfide linkages. In addition, the high resolution mass analyzers, such as Fourier transform ion cyclotron resonance (FT-ICR) and Orbitrap, were developed to couple with the linear ion trap. These instruments were employed to characterize protein structures in my dissertation work, such as peptide mapping and post translational modifications. Figure 1.7 shows the schematic of LTQ Orbitrap XL. The maximum resolution of this instrument can be 100K; CID and high-energy collision-induced dissociation (HCD) are integrated in one mass spectrometer instrument, as shown in Figure 1.7. The CID fragmentation takes places in LTQ; fragment ions are detected there. If high-resolution detection is desired, fragment ions can be sent to Orbitrap. HCD is a new fragmentation technique, which dissociates precursors in the HCD cell, and the fragment ions are detected in the Obitrap. Orbitrap XL is ETD.
upgradeable. For ETD fragmentation, the radical anions generated in reagent ion source are transferred to LTQ, where ETD occurs.

Figure 1.7 Schematic of the Orbitrap XL MS$^{93}$.

The limitations of ion trap mass spectrometers are obvious. The first is that fragment ions with m/z values < one third of the precursor's cannot be detected, since they are not stable in the ion trap$^{94}$. For example, a precursor of a peptide is m/z 1200, its fragment ions with m/z < 400 will not be detected, which presents a significant limitation for de novo sequencing$^{55}$. The second limitation is that it is difficult to perform high sensitivity triple-quadrupole type SRM.

1.3.3.3. Detector

Once the ions are separated by the mass analyzer, they travel to the ion detector, which generates a current signal from the incident ions. The most commonly used detectors include the electron multiplier, the Faraday cup, and microchannel plate detectors$^{95}$. The choice of detector depends on the type of mass spectrometer. Basically, the detection of ions involves multiplying signals and then recording them.
1.3.4 Data analysis

The computer records data generated from the mass spectrometer as the m/z and intensity. Then, analytes can be identified by the m/z of precursor and fragment ions, and quantitated through intensities or peak areas. In the data analysis, there are some important terms involved.

m/z

The calculation of a theoretical m/z is shown below:

\[
\frac{m}{z} = \frac{MW + z \cdot H}{z} \quad \text{Equation 1.5}
\]

This equation is used in the positive mode when adducts are not considered. In this formula, MW is the molecular weight of the analyte, H is the mass of hydrogen and z is the charge. For example, the theoretical MW of peptide GTVGTLSSDITR is 1231.68 Dalton, if its charge state is +1, the theoretical m/z is 1232.68; if its charge state is +2, then the theoretical m/z is 616.85. By searching the theoretical m/z of an analyte, a peak with m/z close to the theoretical value would be found. For a low resolution mass spectrometer, the experimental m/z can be ± 1 of the theoretical value; for a high resolution mass spectrometer, the experimental m/z is usually less than ± 0.01 of the theoretical one.
**Monoisotopic mass and average mass**

When calculating the MW, the monoisotopic mass and average mass are important. The monoisotopic mass of an element refers to the lightest stable isotope of that element. For example, the monoisotopic mass of carbon is 12.000000. There are two major isotopes of carbon in nature: $^{12}\text{C}$ and $^{13}\text{C}$ ($^{14}\text{C}$ is in the trace amount: 1 part per trillion in the atmosphere), with masses of 12.000000 and 13.003355, and natural abundances of 98.92% and 1.08% respectively. The average mass of an element is the abundance-weighted sum of each isotope. Hence the average mass of carbon is

$$12.000000 \times 98.92\% + 13.003355 \times 1.08\% = 12.0108$$

Similarly, the monoisotopic mass of nitrogen is 14.003074, which is the atom weight of $^{14}\text{N}$ (natural abundance 99.63%). Another isotope of nitrogen is $^{15}\text{N}$ (15.000109; natural abundance 0.37%); hence the average mass of nitrogen is 14.0078.

The monoisotopic mass of an analyte is determined by summing the monoisotopic mass of each element. The average mass of an analyte is the sum of the average mass of each element. In a high resolution mass spectrometer, the monoisotopic mass, as well as other heavier isotopes that contribute to the mass cluster would be obtained. In a low resolution mass spectrometer, the experimental m/z value is usually close to the average mass of the analyte.
Resolution

The resolution measures the ability to distinguish two peaks with similar, but not identical m/zs. The IUPAC definition of resolution in mass spectrometry is the ratio of m/Δm, in which m is the average m/z of the two peaks, and Δm is the m/z difference of them. A high resolution in mass spectrometry can greatly decrease the noise of detection and help the identification of analytes. In a high resolution mass spectrometer, such as an Orbitrap or FT-ICR, the charge states of analytes can be read directly from their isotopic peaks. As shown in Figures 1.8 A and B, the mass peaks of peptide GTVGTLSSDITR, the charge state of 616.85 can be obtained by 1 divided by the subtracting value of the adjacent m/zs in the isotopic mass cluster: \[ z = \frac{1}{617.35 - 616.85} = 2. \] However, in a low resolution instrument, such as LTQ or quadrupole, resolution (R) =1K, the monoisotopic m/z and its isotopic m/zs cannot be resolved, as shown in Figure 1.8C. Hence, the charge state of this ion cannot be read from the mass spectrum.
Figure 1.8 Resolution: The hypothetical (M+2H)2+ m/z clusters of peptide GTVGTLSSDITR. A. R=100,000; B. R=10,000; C. R=1,000.

Mass accuracy

The accuracy of a mass measurement describes the correctness of the experimental mass of an analyte. It is usually expressed as the difference between the theoretical mass and the experimental mass and can be calculated by \((\text{theoretical mass} - \text{experimental mass}) \times 10^6 / \text{theoretical mass}\) with the unit in ppm. In a high mass accuracy instrument, such as an Orbitrap, the mass accuracy can reach 2 ppm or less. For example, the experimental monoisotopic m/z of peptide GTVGTLSSDITR will be 616.8459 ± 0.0012. Hence, the confidence will be higher when identifying the peptide. Mass accuracy is related to resolution. In a low resolution instrument, the mass accuracy is usually calculated from the theoretical average mass and the peak top mass (617.20 in Figure 1.8 C). To maintain the optimal mass accuracy, mass spectrometers are required to be calibrated routinely by reference standards.
For the high accuracy instruments, such as Orbitrap and Q-TOF, besides for the routine calibration, internal calibration can make results more accurate.

1.3.4.1 Manual data analysis

Although there are many software programs available to process MS data, manual analysis is usually the most straightforward and reliable method. By calculating theoretical m/zs, analytes can be extracted from the raw MS data. For the data obtained from a high resolution instrument, monoisotopic m/z values can be used to calculate the difference between the experimental and the theoretical values. If the mass difference is within the expected mass accuracy of the instrument, the MS peak might be the analyte. As shown in Figure 1.9A, the MS spectrum is from Orbitrap XL, and the mass accuracy of precursor (950.0909, 3+) is less than 5 ppm, which is the expected value of this instrument; hence, the ion 950.0909, 3+ is probably the precursor for peptide TYTYADTPDQFQLDHDFLSPEDTK. Then, MS2 can be examined to further confirm the identification. In Figure 1.9B, the theoretical fragment ions in different charge states are calculated and listed. According to the theoretical fragment ions, the CID-MS2 of the precursor in Figure 1.9A is interpreted and shown in Figure 1.9C. Additionally, peptides determined by ETD or HCD can be identified by the same method. Peptides are broken in the Cα-N bond of the backbone in an ETD spectrum to generate c and z ions, shown in Figure 1.2, and HCD generates high resolution b and y fragment ions.
A. MS

Theoretical monoisotopic mass = 950.0872, 3+
Mass accuracy = 4 ppm

B. Theoretical m/z values of CID fragment ions

<table>
<thead>
<tr>
<th>Seq</th>
<th>B (1+)</th>
<th>Y (1+)</th>
<th>B (2+)</th>
<th>Y (2+)</th>
<th>B (3+)</th>
<th>Y (3+)</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>1</td>
<td>102.05</td>
<td>2848.25</td>
<td>51.53</td>
<td>1424.63</td>
<td>34.69</td>
<td>24</td>
</tr>
<tr>
<td>Y</td>
<td>2</td>
<td>265.12</td>
<td>2747.2</td>
<td>133.06</td>
<td>1374.1</td>
<td>89.04</td>
<td>23</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
<td>366.17</td>
<td>2584.14</td>
<td>183.59</td>
<td>1292.57</td>
<td>122.73</td>
<td>22</td>
</tr>
<tr>
<td>Y</td>
<td>4</td>
<td>529.23</td>
<td>2483.09</td>
<td>265.19</td>
<td>1242.05</td>
<td>177.08</td>
<td>21</td>
</tr>
<tr>
<td>A</td>
<td>5</td>
<td>600.27</td>
<td>2320.02</td>
<td>300.64</td>
<td>1160.52</td>
<td>200.76</td>
<td>20</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>715.29</td>
<td>2248.99</td>
<td>358.15</td>
<td>1125</td>
<td>239.1</td>
<td>19</td>
</tr>
<tr>
<td>T</td>
<td>7</td>
<td>816.34</td>
<td>2133.96</td>
<td>408.67</td>
<td>1067.48</td>
<td>272.78</td>
<td>18</td>
</tr>
<tr>
<td>P</td>
<td>8</td>
<td>913.39</td>
<td>2032.91</td>
<td>457.2</td>
<td>1016.96</td>
<td>305.14</td>
<td>17</td>
</tr>
<tr>
<td>D</td>
<td>9</td>
<td>1028.42</td>
<td>1935.86</td>
<td>514.71</td>
<td>968.43</td>
<td>343.48</td>
<td>16</td>
</tr>
<tr>
<td>T</td>
<td>10</td>
<td>1143.45</td>
<td>1820.83</td>
<td>572.23</td>
<td>910.92</td>
<td>381.82</td>
<td>15</td>
</tr>
<tr>
<td>P</td>
<td>11</td>
<td>1290.57</td>
<td>1705.81</td>
<td>645.76</td>
<td>853.41</td>
<td>430.84</td>
<td>14</td>
</tr>
<tr>
<td>Q</td>
<td>12</td>
<td>1418.57</td>
<td>1558.74</td>
<td>709.79</td>
<td>779.87</td>
<td>473.53</td>
<td>13</td>
</tr>
<tr>
<td>L</td>
<td>13</td>
<td>1531.66</td>
<td>1430.68</td>
<td>766.33</td>
<td>715.84</td>
<td>511.22</td>
<td>12</td>
</tr>
<tr>
<td>H</td>
<td>14</td>
<td>1668.72</td>
<td>1317.6</td>
<td>834.86</td>
<td>659.3</td>
<td>556.91</td>
<td>11</td>
</tr>
<tr>
<td>D</td>
<td>15</td>
<td>1783.74</td>
<td>1180.54</td>
<td>892.38</td>
<td>590.77</td>
<td>595.25</td>
<td>10</td>
</tr>
<tr>
<td>F</td>
<td>16</td>
<td>1930.81</td>
<td>1065.51</td>
<td>965.91</td>
<td>533.26</td>
<td>644.28</td>
<td>9</td>
</tr>
<tr>
<td>S</td>
<td>17</td>
<td>2017.84</td>
<td>918.44</td>
<td>1009.43</td>
<td>459.72</td>
<td>673.29</td>
<td>8</td>
</tr>
<tr>
<td>L</td>
<td>18</td>
<td>2130.93</td>
<td>831.41</td>
<td>1065.97</td>
<td>416.21</td>
<td>710.98</td>
<td>7</td>
</tr>
<tr>
<td>P</td>
<td>19</td>
<td>2272.98</td>
<td>718.32</td>
<td>1114.49</td>
<td>359.67</td>
<td>743.33</td>
<td>6</td>
</tr>
<tr>
<td>E</td>
<td>20</td>
<td>2357.02</td>
<td>621.27</td>
<td>1179.02</td>
<td>311.14</td>
<td>786.35</td>
<td>5</td>
</tr>
<tr>
<td>E</td>
<td>21</td>
<td>2486.07</td>
<td>492.23</td>
<td>1243.54</td>
<td>246.62</td>
<td>829.36</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>22</td>
<td>2601.09</td>
<td>363.19</td>
<td>1301.05</td>
<td>182.1</td>
<td>867.7</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>23</td>
<td>2702.14</td>
<td>248.16</td>
<td>1351.57</td>
<td>124.58</td>
<td>901.39</td>
<td>2</td>
</tr>
<tr>
<td>K</td>
<td>24</td>
<td>830.24</td>
<td>147.11</td>
<td>1415.62</td>
<td>74.06</td>
<td>944.08</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 1.9 Peptide identification by mass spectrometry: A peptide identification from mass; B theoretical m/z values of CID fragment ions; C peptide identification from CID-MS2

1.3.4.2 Software data processing

Manual analysis can obtain a large amount of useful information; however, considering the efficiency, software data processing is gaining more and more applications. There are multiple search engines available on the market, such as SEQUEST and Mascot. The basic point is that these software programs generate theoretical peptides according to the enzyme digest that the user sets. For each measured MS/MS spectrum, theoretical peptides which have masses matching the precursor mass of this spectrum within a given precursor mass tolerance are selected to generate theoretical fragment ions. The more the theoretical fragment masses match the measured ones, the better a peptide matches the spectrum. Depending on the search-engine-specific scoring model, one or more of the best matching peptides are assigned to the spectrum. Each assignment is given a score, reflecting how well the measurement matches the theoretical fragments; this score depends, for example, on the existence of experimental
fragment masses, their intensities, or their completeness. The major difference between different search engines is the scoring method, which has been systematically reviewed\textsuperscript{98}.

1.4 Applications of LC-MS in the protein characterization

LC-MS is a powerful technology in protein characterization. After the translation of the protein database from genes, protein characterization has become increasingly feasible by LC-MS. However, proteins are much more complicated than genes, since they usually experience various expression levels and PTMs in the cells. The characterization of proteins can lead to important understandings of biological activities or processes. Additionally, the analysis of recombinant proteins plays a critical role to ensure the quality, purity, efficacy, and safety of products in the biotechnology industry. In this section, the characterization of proteins is introduced from the top-down and bottom-up levels. In addition, PTM identification and MS quantitation are discussed.

1.4.1 Top-down protein characterization

Similar to the story of the blind men and the elephant, when a protein is digested into small pieces, the understanding of it will start from small peptides and site-specific modifications. Therefore, the knowledge of the whole protein will be incomplete. The top-down method determines a protein from the intact molecular level and reaches 100\% sequence coverage for proteins <70 kDa. Ideally, top-down proteomics can identify proteins and also their coding polymorphisms and PTMs; however, due to the limitation of current instruments and software, this method is still being developed\textsuperscript{99}.
Kelleher et al. identified 60 proteins in the yeast whole cell lysate with the MW ranging from 14 to 35 kDa\textsuperscript{100}. They applied ProSight, the first and important software program assisting top-down proteomic analysis, to identify proteins from both their masses and fragment ions. Furthermore, the characterization of PTMs was also reported by the top-down approach. Over 150 differentially modified forms of histone H3.2 were identified in the asynchronously grown and butyrate-treated HeLa cells by HILIC-Orbitrap MS-ETD-MS\textsuperscript{2}. The major PTMs identified in this protein were acetylation and methylation\textsuperscript{101}.

1.4.2 Bottom-up method

The bottom-up method refers to cleaving a protein into small pieces by enzymes or chemical reagents and then analyzing these pieces by LC-MS. The commonly used workflow is shown in Figure 1.10. This method provides useful information of the amino acid sequence, as well as glycosylation, disulfide linkages, and other modifications. It is also employed in the study of protein stabilities and the relative quantitation. In this section, the details of these applications will be presented.
1.4.2.1 Amino acid sequence

There are mainly two goals to determine the amino acid sequence of proteins: 1. identify the existence of a protein in a protein mixture; 2. verify if the protein is correctly expressed. The workflow of the sample preparation shown in Figure 1.10 can be employed in both cases.

For the first goal, researchers don't need to identify every amino acid residue in the protein; the observation of one or more unique peptides in a protein is thought to determine the presence of that protein. The data analysis usually starts from searching against a proteomic database by a software engine. Details of the software data processing were introduced in...
section 1.3.4.2. However, this peptide-centric proteomic approach has issues. There was a paper published on Natural Biotechnology in 2010, which deeply revealed the problems behind this methodology\textsuperscript{102}, including limitations of peptide-matching algorithms, incomplete databases, interference among proteins, etc.

For the second goal, every amino acid residue has to be identified. In biotechnology, gene sequences of recombinant proteins are inserted into host cells, and then they are expressed with host cell proteins. Finally, the desired proteins are extracted and purified. The amino acid sequences of these product proteins have to be confirmed, since mutations or unexpected modifications might occur in the host cells. To reach 100\% sequence coverage, multiple enzymes are usually used, since LC-MS can only detect peptides with appropriate sizes and hydrophobicities. Peptides, too small or hydrophilic, are not retained in the LC system; peptides, too big or hydrophobic, may not be eluted from the LC column. Peptides, inappropriate sizes under one protease, may be easy to detect when digested by another protease. The full sequence coverage can then be reached by combining the results from different enzymatic digests. In 1978, Hancock initially applied RP-HPLC-UV to efficiently separate and detect peptides, which were prepared by the tryptic digestion\textsuperscript{103}. This technique was then extensively used for the sequence determination of proteins. Coupled with the MS detection, this methodology became more powerful, and it is routinely applied in the biopharmaceutical industry nowadays\textsuperscript{96, 104, 105}.
1.4.2.2. Glycosylation

Protein glycosylation is a PTM that occurs in the endoplasmic reticulum (ER) and Golgi of a eukaryotic cell. In N-glycosylation, carbohydrates attach to the side chain of asparagine in a conserved sequence NXS/T. In O-glycosylation, carbohydrates usually attach to the side chain of serine or threonine. The biological roles of glycans can be divided into two aspects: 1. the structural and modulatory role; 2. the specific recognition of glycans by other molecules – most commonly, glycan-binding proteins (lectins). For example, the external location of glycans on most glycoproteins can provide a general shield, which protects the underlying polypeptide from recognition by proteases or antibodies. Researchers reported that glycans were different between diseases and normal cells\textsuperscript{106-108}. The FDA requires biopharmaceutical companies to provide the glycan profiles in recombinant therapeutic proteins, because glycans can have a profound influence on the safety and effectiveness of a biopharmaceutical product.

The characterization of glycans includes compositional and linkage analysis, since one structure might have multiple linkages of glycans. The application of MS for the glycosylation analysis has been reviewed\textsuperscript{109-114}. In this section, glycan analysis will be introduced from two aspects: glycopeptides and released glycans.

1.4.2.2.1. Glycopeptides

From glycopeptides, researchers can obtain glycan compositions, glycosites, and relative abundances of glycoforms. The workflow that reflects the use of MS methods to analyze glycopeptides was shown in Figure 1.11\textsuperscript{110}.
Glycans in a glycoprotein can be determined by a bottom-up approach, in which the glycan composition is estimated by the mass difference between the protein with and without glycans. Most frequently, scientists digest a glycoprotein into peptides, in which glycopeptides are contained, and then the sample is detected by LC-MS. The mass of a glycopeptide is calculated by adding the mass of the peptide backbone to the mass of the glycan. The composition of a glycan can be estimated from the mass and confirmed by CID-MS2, and the glycosites can be obtained from ETD-MS2. In a CID-MS2 spectrum, the major fragment ions are produced from cleavages of glycosidic bonds. In an ETD-MS2 spectrum, peptide backbones, instead of glycosidic bonds, are usually broken. HCD generates cleavages in both glycosidic bonds and peptide bonds; however, the major fragment ions are oxonium ions of glycans, which can only help the recognition of glycopeptides. In the

**Figure 1.11** Workflow for glycopeptide analysis.
glycomic analysis, CID, ETD, and HCD are usually applied in one run\textsuperscript{115-117}, which determines glycan compositions and glycosites from glycopeptide, as well as other peptides. Additionally, both N- and O-glycosylation can be studied by this glycopeptide-based method. However, the glycan linkages are difficult to study from glycopeptides.

1.4.2.2.2. Released glycans

On the other hand, researchers release glycans from proteins and detect them separately. The N-glycans can be removed by enzymes, e.g., N-glycosidase F and N-glycosidase A, which cleave between the innermost GlcNAc and asparagine residues. The removal of O-glycans relies mainly on the chemical release by either reductive or non-reductive β-elimination, since there is no enzyme that can universally remove O-glycans\textsuperscript{110}. The released glycans can be detected in native forms, aminated forms, or permethylated forms.

The common reductive amination groups include 2-aminopyridine (2AP), 2-aminobenzamide (2AB), 2-aminobenzoic acid (2AA), and 2-aminoacridone (AMAC), all of which are fluorophores. Aminated glycans can be determined by either fluorescence detector or MS. The former is suitable for quantification; researchers use capillary electrophoresis coupled with laser induced fluorescence (CE-LIF)\textsuperscript{118-121} or LC-LIF\textsuperscript{119, 121} to determine the structures of glycans. For the LC separation, HILIC and RP are commonly used. The glycan composition can be predicted by comparing the retention time of unknowns with standards, and glycan linkage can be estimated from gradual exoglycosidase digests. Besides LIF, MS is widely used to detect native glycans and aminated glycans. ESI-LC-MS is preferred to glycan analysis, since MALDI results in the dissociation of labile glycosidic bonds, which is a great
concern for acidic monosaccharides, such as sialic acids, sulfate, and phosphate\textsuperscript{122, 123}. The most popular nomenclature for fragment ions was developed in 1988 by Domon and Costello\textsuperscript{124}, as shown in Figure 1.12. C and y ions correspond to the mass of the intact carbohydrate; a and x ions are generated from the cross-ring fragmentation, which can help determine the linkages. In the positive mode, the major fragment ions are generated from the cleavage of glycosidic bonds, which do not provide enough information for the prediction of glycan linkages from native glycans and aminated glycans. It was reported that cross-ring fragmentation was preferred in the negative mode; hence, glycan linkages could be predicted\textsuperscript{109, 125-127}. Our group published on the characterization of N-glycans using fluoride-mediated negative ion microfluidic chip LC-MS. The negative charge results in the generation of c-type and a-type cross-ring fragment ions, which lead to in-depth oligosaccharide characterization of linkage and positional isomers\textsuperscript{128}. However, the sensitivity of neutral glycans is lower in the negative mode, relative to the positive mode\textsuperscript{127}.

![Diagram](image)

**Figure 1.12** Nomenclature for glycoconjugate mass spectrometry\textsuperscript{124}.
Glycan linkages can be obtained by MS after permethylation. Generally, the free hydroxyl groups in the glycan are first derivatized with one reagent. The product is hydrolyzed to release other hydroxyl groups which are then derivatized by a different reagent. The positions of hydroxyl groups derivatized in the later step, which are involved in glycan linkages, are subsequently located by GC-MS. In GC-MS, the derivatized monosaccharide can be identified via the retention and molecular weight. Glycan linkage analysis, which is one of the major applications of GC-MS, has been reported and reviewed in many articles.\textsuperscript{129-131}

1.4.2.3 Disulfides

The formation of a disulfide bond is a frequently encountered PTM of proteins, which is critical to establish and maintain the three-dimensional structure of a protein. Characterization of disulfide linkages can help understand the structure and function of a protein. In the Uniprot protein database, disulfide linkages are available for many proteins. The data might come from X-crystallography, NMR, or MS; it might also be predicted via homology. Currently, MS has become increasingly important in the disulfide characterization. MS determines the MW of disulfide-involved peptides. Then, researches can use CID to characterize disulfide bonds. CID doesn’t break disulfide bonds, one peptide can be thought as the modification of the other one on the cysteine residue. This method relies on the high quality fragmentation of CID. However, large or highly charged peptides usually cannot generate sufficient CID fragment ions. Hence, researchers also predict the disulfide linkages by comparing the result before and after reducing disulfides.\textsuperscript{132,133} The confidence of the disulfide assignment is reduced in a complicated linkage.
ETD is able to break disulfide bonds. Our group developed a CID and ETD combined method to characterize disulfides. Generally, the disulfide-involved peptides are fragmented by ETD to generate disulfide-dissociated species and charge-reduced species, which are then isolated for analysis by CID-MS3\textsuperscript{134-136}. Taking the advantage of high resolution and high accuracy mass spectrometers, the confidence of the disulfide assignment can be greatly increased, especially in a complicated linkage.

1.4.2.4 Chemical modifications

Other modifications, such as deamidation, isomerization, and oxidation, are commonly observed in proteins. They are not enzymatic reactions; they may increase as people age and induce human diseases. For example, the amount of isomerized or deamidated proteins is increased in the brains of patients with Alzheimer's diseases\textsuperscript{48}. In the biotechnological industry, these modifications influence drug stabilities, and may trigger potential immune responses\textsuperscript{137, 138}. Hence, the determination of these modifications plays a significant role in the protein structural characterization.

Deamidation, which proceeds through the formation of a succinimide intermediate followed by hydrolysis, results in the formation of isoaspartate acid (iso-Asp) and aspartate (Asp). The asparagine (Asn) followed by glycine is the most susceptible site for deamidation. Succinimide is stable only in the weak acidic condition, while isoaspartate acid and aspartate can stably exist in the protein under a wide pH range and induce the changing of a protein’s secondary and tertiary structures. The mass shifts -17 Da from Asn to succinimide, and +1 Da from Asn to iso-Asp and Asp. Deamidated species can be separated in RP-HPLC, and
detected by MS\textsuperscript{139}. The iso-Asp and Asp, which have the same mass, can be distinguished by ETD or ECD\textsuperscript{140,141}. Additionally, other sites like glutamine\textsuperscript{142} or asparagines, followed by other residues might also be deamidated; however, the reaction rate is much lower compared with asparagine which is followed by glycine. Deamidation is sensitive to buffer composition, ionic strength, and pH, artificial modifications generated in the sample preparation might be a concern, especially in quantitation.

Isomerization is prone to occur when Asp residue is followed by glycine. Asp isomerizes through a succinimide intermediate to form an iso-Asp. As was mentioned, Asp and iso-Asp can be differentiated by ETD and separated in RP-HPLC.

Methionine (Met) is one of the most susceptible residues to oxidation. The mass usually increases 16 Da in the oxidation, and the oxidized species is eluted earlier in RP-HPLC\textsuperscript{143}.

### 1.4.3 Quantitation

Protein quantitation is as significant as identification and is widely applied in clinical and pharmaceutical research. In MS-based quantitation, the amount of a protein or peptide is expressed by the peak area, the intensity, or spectral counting in a mass spectrum. There are two types of quantitation: relative and absolute. The relative quantitation is usually used to compare differences among samples, such as the changing of specific proteins or PTMs between samples from disease and control. The absolute quantitation determines the precise amounts of specific proteins, such as determining the concentration of a therapeutic protein in serum for the pharmacokinetic analysis, or drug target proteins for the pharmacodynamic
analysis. In this section, the relative and absolute methods are introduced for the MS-based quantitation.

1.4.3.1 Relative quantitation

Relative quantitation is usually realized from the label-free or isotopic labeling methods. Label-free quantitation is simple for the sample preparation, and it is widely used. It is applied to compare PTMs among samples, such as an innovator and biosimilar product, control and heat-stressed proteins, as well as complex protein mixtures. Liu et al.\textsuperscript{137} characterized the stability of a mAb after prolonged incubation at an elevated temperature by quantitatively comparing deamidation, fragmentation, and N-terminal glutamate cyclization. Furthermore, protein mixtures from the cell lysate or serum were also reported to be quantitated by the label-free method; the proteins were digested into peptides, which were then determined by LC-MS. Spectral counting, ion intensity, and integrated peak areas were used to relatively quantitate the protein amount. In this peptide-centric method, a protein amount was obtained from the spectral counting of peptides identified, or the top three highest abundant peptides in that protein if using ion intensity or integrated peak area\textsuperscript{144-146}. For comparison, internal standards with high reproducibility between runs were applied to normalize the concentration differences.

In order to minimize variations among LC-MS runs, isotopic-labeling protein quantitation was developed, which labels peptides or proteins from different samples into “light” and “heavy” forms. Multiple isotopic labeling methods were reported, such as stable isotope labeling by amino acids in cell culture (SILAC), isotope-coded affinity tags (ICATs), tandem mass
tags (TMTs), isobaric tags for relative and absolute quantitation (iTRAQ), and dimethyl labeling.

SILAC was originally developed by Mann\textsuperscript{147}. Generally, two populations of cells are cultivated in cell culture. One is fed with growth medium containing normal amino acids; the other containing amino acids with stable heavy isotopes, such as arginine labeled with six \textsuperscript{13}C which will be incorporated to all the proteins expressed in this cell culture. Then, the proteins in each culture can be combined and analyzed by LC-MS. They can be distinguished by 6 Da heavier in mass and relatively compared by the ratio of peak intensities. This method is applied to study key factors that affect proteins expressed in cell cultures, biological functions of specific proteins, and also PK and PD in biopharmaceutical areas. However, this method requires starting to prepare samples from the cell culture, and this process might not be available to some labs.

ICAT, TMT, iTRAQ, and dimethyl labeling quantitate proteins from labeling proteins or peptides. ICAT labels cysteines as heavy and light versions, and then they are combined, digested, and analyzed by LC-MS\textsuperscript{148}. However, it is not practical for proteins containing low or no cysteines. TMT, iTRAQ, and dimethylation label primary amine; they are usually applied after digestion. After TMT and iTRAQ labeling, peptides can be differentiated by the fragment ions, but not precursors; hence, the shift of retention time caused by the isotopic labeling can be eliminated\textsuperscript{149}. However, the report ions of the labeling in MS2 have masses < 150 Da, which might not be detected in the ion trap because of the 1/3 cut-off rule; hence, pulsed Q dissociation (PQD) or HCD are usually required in these methods. Furthermore, TMT can compare up to 6 groups of samples at once, and iTRAQ compares up to 4 groups\textsuperscript{150}.\textsuperscript{150}
both are widely used. Dimethylation also labels primary amines, which has a smaller tag than TMT and iTRAQ. Peptides from different samples are distinguished from precursor masses, and this method can compare 3 groups of samples at once\textsuperscript{151}. In my dissertation work, dimethylation was developed to compare the amino acid sequences between the innovator and biosimilar in one LC-MS run.

\textbf{1.4.3.2 Absolute quantitation}

For the absolute quantitation, standards are often required. The standards can be chemically synthesized analogs, or the same peptides with the isotopic labeling, and also SILAC is used in the absolute quantitation. The major roles of standards are to correct variations in experiments and establish standard curves. Basically, the standard curve is made by the concentration vs. MS response, and then the concentration of unknown can be obtained. The selected reaction monitoring (SRM) or multiple reactions monitoring (MRM) scanning mode is usually employed to quantitate in the mass spectrometer, such as triple quadrupole. The parent ions of interest are selected to fragment in CID, and then specific product ions are scanned and recorded. The peak areas of product ions are used to quantitate the amounts of the analytes. This method has advantages when analyzing complicated samples, such as proteins from the serum, since it can exclude interferes of co-eluted analytes.
1.5 References


57. Wu SL, Huhmer AF, Hao Z, Karger BL. On-line LC-MS approach combining collision-induced dissociation (CID), electron-transfer dissociation (ETD), and CID of an isolated charge-reduced species for the trace-level characterization of proteins with post-translational modifications. Journal of proteome research 2007; 6:4230-44.


62. van Deemter JZ, FJ; Klinkenberg, A. Longitudinal diffusion and resistance to mass transfer as causes of non ideality in chromatography. chemical engineering science 1956; 5:271-89.


79. Yoshida; KTHYISAYYT. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Communications in Mass Spectrometry 1988; 60:2299-301.


Chapter 2 Comparability analysis of anti-CD20 commercial (rituximab) and RNAi-mediated fucosylated antibodies by two LC-MS approaches

Chen Li, Anthony Rossomando, Shiaw-Lin Wu, and Barry L. Karger. Comparability analysis of anti-CD20 commercial (rituximab) and RNAi-mediated fucosylated antibodies by two LC-MS approaches. mAbs, volume 5, issue 4, 565-575, 2013

Co-authors’ work in this chapter: Chen Li: concept contribution, experimental design, experiment procedures, data analysis and interpretation, manuscript writing and revision; Anthony Rossomando: supply of samples; Shiaw-Lin Wu: concept contribution, experimental design, data interpretation, manuscript revision: Barry L. Karger: goal for the study, concept contribution, manuscript revision, and grant support.
2.1 Abstract
In developing biosimilar or biobetter products, comparability to the reference product is required to claim similar integrity or intended purpose. In this work, an anti-CD20 monoclonal antibody developed using RNA interference to decrease core fucosylation (RNAi-mediated) was comprehensively characterized by LC-MS and compared to the commercially-available anti-CD20 rituximab (MabThera®). As anticipated, < 30% core fucose was found within the RNAi-produced molecule (compared to > 90% in rituximab), and the reduction in fucose resulting in a significant improvement in FcγRIIIa binding and antibody-dependent cell-mediated cytotoxicity. Two mutations, S258Y (fully mutated) and F174I/L (partially mutated), however, were detected in the production of the RNAi-mediated molecule. An alternative LC-MS approach using dimethyl labeling (i.e., 2CH₂ for rituximab and 2CD₂ for the RNAi-mediated molecule) was developed to additionally compare the two mAbs and confirm the full sequence with the two mutation sites. Furthermore, disulfide linkages were found to be the same for the two antibodies, with a small portion of unpaired cysteines in both products. Disulfides were correctly linked if the samples were prepared at low pH (i.e., enzymatic digestion by pepsin at pH 2); however, trace amounts of scrambling were found by trypsin digestion at pH 6.8, and this scrambling increased significantly at pH 8. Typical modifications, such as pyro-Glu formation at the N-terminus, K clipping at the C-terminus, oxidation at Met, and deamidation at Asn, were also detected with no significant differences between the two products. Using the LC-MS approaches for the comparability study, product integrity with critical structure information was revealed for confirmation of intended purpose (core fucosylation), identification of critical parameters (e.g., sample pH), and correction as needed (amino acid mutation).
2.2 Introduction
Rituximab, a therapeutic monoclonal antibody (mAb) targeting CD20 in B cells, is used to treat B-cell non-Hodgkin lymphoma and rheumatoid arthritis.1-3 The product’s brand names are Rituxin® (in US) and MabThera® (in Europe). One important function for the antibody is to induce antibody-dependent cell-mediated cytotoxicity (ADCC), in which the Fc domain, including the glycans, binds specifically to Fc receptors in human effector cells, such as macrophages and natural killer cells, to induce ADCC.4-6 Because the glycan structure, particularly the core fucose, is important to mediate ADCC, the reduction of the core fucose (i.e., by RNAi) should enhance the effect.7-11 Thus, RNAi-mediated fucosyltransferase (FUT8) and GDP-man-4,6-dehydratase (GMDS) was used to produce anti-CD20 mAb for this purpose.12 Although the aim was for a bio-better product, the overall structure, except for the level of the core fucose, was intended to be as similar as possible to the reference product to maintain the drug integrity.

In this study, we first used state of the art mass spectrometric methods to characterize the structure of the newly developed RNAi-mediated anti-CD20 mAb and then compared it to the structure of the commercial rituximab molecule. As expected, reduction of core fucosylation was observed for the RNAi-mediated molecule. On the other hand, the primary structure, disulfide linkages, and common modifications such as pyro-Glu formation at the N-terminus, K clipping at the C-terminus, oxidation at Met, and deamidation at Asn were found to be similar between the two products. The liquid chromatography–mass spectrometry (LC-MS) used for full sequence analysis, however, identified two amino acid residues mutated on the RNAi-mediated molecule. An alternative LC-MS method, using dimethyl labeled with 2CH₂ for rituximab, and isotopically-dimethyl labeled with 2CD₂ for the RNAi-mediated molecule
confirmed the amino acid changes on the RNAi-mediated molecule. Moreover, both approaches were in agreement that one variant was fully mutated and the other partially mutated. Small amounts of free cysteines in both molecules were also observed. Disulfide scrambling, which could be caused by the free cysteines, was detected in both mAbs. At pH 6.8 or pH 8, which are typical enzymatic digestion conditions, a small amount of disulfide scrambling was observed (a trace amount at pH 6.8 and significantly more at pH 8), but no scrambling was seen at pH 2. The pH used for sample preparation is shown to be critical to measure correctly the free cysteines and disulfide linkages.

2.3 Results
To establish identity, the sequence of the newly developed RNAi-mediated molecule was compared to the amino acid sequence of rituximab found in US Patent 5736137. Additionally, disulfide linkages, glycosylation structure, and amino acid modifications in the two mAbs were characterized and compared as described in the following sections.

Peptide mapping

Enzymatic peptide mapping was used for the primary sequence identification. A typical trypsin peptide map of the RNAi-mediated mAb is illustrated in Figure 2.1, with the identifications of all peptides summarized in Tables 2.1A for the heavy chain and 2.1B for the light chain. As listed in the tables, several small peptides were identified through miscleavage or by digestion using different enzymes, such as Lys-C or pepsin. Importantly, many peptides with overlapping amino acids were repeatedly identified in the different enzymatic maps.
Thus, complete sequence coverage (100%) was successfully achieved by the combined analysis of these enzymatic peptide fragments.

**Figure 2.1** Tryptic map of RNAi-mediated mAb. T1H stands for the first tryptic peptide from the N-terminal heavy chain, T1L for the first tryptic peptide from the N-terminal light chain, pyro-T1H for the pyro-Glu form of T1H, GoF for the glycopeptide with core fucose, T24H* for the deamidated T24H. Similar nomenclature is used for the other peptides.

**Table 2.1** Summary of the identified peptides in the heavy chain and light chain of Anti-CD20

<table>
<thead>
<tr>
<th>#</th>
<th>Sequence position</th>
<th>Sequence</th>
<th>Enzyme</th>
<th>MabThera (observed)</th>
<th>Alnylam (observed)</th>
<th>[MH]+ (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1H</td>
<td>1-19</td>
<td>QVQLQQPGAELVKPGASVK</td>
<td>Trypsin</td>
<td>32.54</td>
<td>33.14</td>
<td>1977.1128</td>
</tr>
<tr>
<td>T2H</td>
<td>20-23</td>
<td>MSCK</td>
<td>Trypsin native, T2H-T9H</td>
<td>32.79</td>
<td>33.81</td>
<td>468.1945</td>
</tr>
<tr>
<td>T3H</td>
<td>24-38</td>
<td>ASGYTFTSYNMHWVK</td>
<td>trypsin</td>
<td>37.83</td>
<td>38.41</td>
<td>1791.816</td>
</tr>
<tr>
<td>T4H</td>
<td>39-43</td>
<td>QTPGR</td>
<td>trypsin mis cleavage, T3H-T4H</td>
<td>34.53</td>
<td>35.01</td>
<td>558.2994</td>
</tr>
<tr>
<td>T5H</td>
<td>44-63</td>
<td>GLEWIGAIYPGNGDTSYNQK</td>
<td>trypsin</td>
<td>46.35</td>
<td>46.65</td>
<td>2183.0404</td>
</tr>
<tr>
<td>T6H</td>
<td>64-65</td>
<td>FK</td>
<td>trypsin mis cleavage, room temperature , T5H-T6H</td>
<td>43.64</td>
<td>45.40</td>
<td>294.1812</td>
</tr>
<tr>
<td>T7H</td>
<td>66-67</td>
<td>GK</td>
<td>trypsin – mis cleavage, T7H-T8H</td>
<td>7.88</td>
<td>7.82</td>
<td>204.1343</td>
</tr>
<tr>
<td>T8H</td>
<td>68-74</td>
<td>ATLTADK</td>
<td>trypsin</td>
<td>9.12</td>
<td>9.45</td>
<td>719.3934</td>
</tr>
<tr>
<td>T9H</td>
<td>75-98</td>
<td>SSSTAYMQSLSSLTSESDAVYYCAR</td>
<td>trypsin</td>
<td>43.12</td>
<td>43.15</td>
<td>2620.154</td>
</tr>
<tr>
<td>T10H</td>
<td>99-125</td>
<td>STYYGGDWYFNVWGAGTTTVSAASTK</td>
<td>trypsin</td>
<td>53.70</td>
<td>54.05</td>
<td>2889.3366</td>
</tr>
<tr>
<td>T11H</td>
<td>126-137</td>
<td>GPSVFPLAPSSK</td>
<td>trypsin</td>
<td>33.27</td>
<td>34.22</td>
<td>1186.647</td>
</tr>
<tr>
<td>T12H</td>
<td>138-151</td>
<td>STSGGTAALGCLVK</td>
<td>trypsin</td>
<td>30.06</td>
<td>30.77</td>
<td>1264.657</td>
</tr>
<tr>
<td>T13H</td>
<td>152-214</td>
<td>DVFPEPVTVSWNSGALTSGVHTFP AVLQSSGLYSLSSVVTVPSSSLGTQ TYICNVNHKPSNTK</td>
<td>trypsin</td>
<td>56.12</td>
<td>56.86</td>
<td>6656.2930</td>
</tr>
<tr>
<td>T14H</td>
<td>215-217</td>
<td>VDK</td>
<td>Trypsin miscleavage, room temperature, T13H-T14H-T15H</td>
<td>54.79</td>
<td>55.63</td>
<td>361.2082</td>
</tr>
<tr>
<td>T15H</td>
<td>218</td>
<td>K</td>
<td>Trypsin miscleavage, room temperature, T13H-T14H-T15H</td>
<td>54.79</td>
<td>55.63</td>
<td>147.1128</td>
</tr>
<tr>
<td>T16H</td>
<td>21-222</td>
<td>AEPK</td>
<td>trypsin</td>
<td>10.24</td>
<td>10.72</td>
<td>444.2453</td>
</tr>
<tr>
<td>T17H</td>
<td>223-226</td>
<td>SCDK</td>
<td>lysC native, T17H-T17L-T18L</td>
<td>16.72</td>
<td>18.58</td>
<td>452.1810</td>
</tr>
<tr>
<td>T18H</td>
<td>227-252</td>
<td>THTCPPCPAPELGGPSVFLFPPKP</td>
<td>trypsin</td>
<td>45.50</td>
<td>46.48</td>
<td>2730.4146</td>
</tr>
<tr>
<td>T19H</td>
<td>253-259</td>
<td>DTLMISR</td>
<td>trypsin</td>
<td>24.14</td>
<td>31.21</td>
<td>835.4342</td>
</tr>
<tr>
<td>T19H'</td>
<td>253-259</td>
<td>DTLMIYR</td>
<td>trypsin</td>
<td>35.64</td>
<td>36.38</td>
<td>911.4655</td>
</tr>
<tr>
<td>T20H</td>
<td>260-278</td>
<td>TPEVTCVVVDVSHEDEPKV</td>
<td>trypsin</td>
<td>35.64</td>
<td>36.38</td>
<td>2082.0060</td>
</tr>
<tr>
<td>T21H</td>
<td>279-292</td>
<td>FNWYVDGVEVHNAK</td>
<td>trypsin</td>
<td>35.64</td>
<td>36.38</td>
<td>1677.8020</td>
</tr>
<tr>
<td>T23H</td>
<td>297-305</td>
<td>EEQYNSTYR</td>
<td>trypsin</td>
<td>16.02</td>
<td>15.74</td>
<td>1189.5144</td>
</tr>
<tr>
<td>T24H</td>
<td>306-321</td>
<td>VVSVLTVLHQDWLNGK</td>
<td>trypsin</td>
<td>50.41</td>
<td>51.24</td>
<td>1808.0065</td>
</tr>
<tr>
<td>T26H</td>
<td>325-326</td>
<td>CK</td>
<td>trypsin native, T20H-T26H</td>
<td>22.80</td>
<td>25.20</td>
<td>250.1220</td>
</tr>
<tr>
<td>T28H</td>
<td>331-338</td>
<td>ALPAPIEK</td>
<td>trypsin</td>
<td>21.73</td>
<td>22.46</td>
<td>838.5033</td>
</tr>
<tr>
<td>T29H</td>
<td>339-342</td>
<td>TISK</td>
<td>trypsin</td>
<td>18.05</td>
<td>18.25</td>
<td>448.2766</td>
</tr>
<tr>
<td>T30H</td>
<td>343-344</td>
<td>AK</td>
<td>Trypsin miscleavage, room temperature, T30H-T31H-T32H</td>
<td>23.66</td>
<td>24.64</td>
<td>218.1499</td>
</tr>
<tr>
<td>T31H</td>
<td>345-348</td>
<td>GQPR</td>
<td>Trypsin miscleavage, room temperature, T30H-T31H-T32H</td>
<td>23.66</td>
<td>24.64</td>
<td>457.2518</td>
</tr>
<tr>
<td>T32H</td>
<td>349-359</td>
<td>EPQVYTLPPSR</td>
<td>trypsin</td>
<td>26.39</td>
<td>26.83</td>
<td>1286.6739</td>
</tr>
<tr>
<td>T33H</td>
<td>360-364</td>
<td>DELTK</td>
<td>trypsin</td>
<td>6.50</td>
<td>6.45</td>
<td>605.3141</td>
</tr>
<tr>
<td>T34H</td>
<td>365-374</td>
<td>NQVSLTCLVK</td>
<td>trypsin</td>
<td>32.98</td>
<td>33.87</td>
<td>1104.6082</td>
</tr>
<tr>
<td>T35H</td>
<td>357-396</td>
<td>GFYPSDIAVEWESNGQPENNYK</td>
<td>trypsin</td>
<td>44.03</td>
<td>44.23</td>
<td>2544.1314</td>
</tr>
<tr>
<td>T36H</td>
<td>397-413</td>
<td>TTPVLDSDGFSFLYSK</td>
<td>trypsin</td>
<td>44.78</td>
<td>45.66</td>
<td>1873.9218</td>
</tr>
<tr>
<td>T37H</td>
<td>414-418</td>
<td>LTVDK</td>
<td>trypsin</td>
<td>10.29</td>
<td>10.72</td>
<td>575.3399</td>
</tr>
<tr>
<td>T38H</td>
<td>419-420</td>
<td>SR</td>
<td>Trypsin miscleavage, T37H-T38H</td>
<td>8.85</td>
<td>8.96</td>
<td>262.1510</td>
</tr>
<tr>
<td>T39H</td>
<td>421-443</td>
<td>WQQGNVFSCVMHEALHNYTQK</td>
<td>trypsin</td>
<td>35.09</td>
<td>35.77</td>
<td>2744.2456</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>------------------------</td>
<td>---------</td>
<td>-------</td>
<td>-------</td>
<td>------------</td>
</tr>
<tr>
<td>T40H</td>
<td>444-450</td>
<td>SLSLSPG</td>
<td>trypsin</td>
<td>26.21</td>
<td>26.64</td>
<td>660.3563</td>
</tr>
<tr>
<td>T1L</td>
<td>1-18</td>
<td>QIVLSQSPAILSAAPGEK</td>
<td>trypsin</td>
<td>35.56</td>
<td>36.34</td>
<td>1825.0066</td>
</tr>
<tr>
<td>T2L</td>
<td>19-24</td>
<td>VTMTCR</td>
<td>trypsin</td>
<td>11.55</td>
<td>11.86</td>
<td>710.3324</td>
</tr>
<tr>
<td>T3L</td>
<td>25-26</td>
<td>ASSSVSYIHFQKPGSPPKPWYATSNLASSGPVR</td>
<td>trypsin</td>
<td>44.11</td>
<td>45.19</td>
<td>3933.0082</td>
</tr>
<tr>
<td>T4L</td>
<td>61-76</td>
<td>FSGSGGTSYS2LTISR</td>
<td>trypsin</td>
<td>29.66</td>
<td>30.07</td>
<td>1606.7707</td>
</tr>
<tr>
<td>T5L</td>
<td>77-102</td>
<td>VEAEDAATYYCQWTSNPPTFGGGTK</td>
<td>trypsin</td>
<td>40.64</td>
<td>40.62</td>
<td>2821.241</td>
</tr>
<tr>
<td>T6L</td>
<td>103-106</td>
<td>LEIK</td>
<td>trypsin</td>
<td>14.02</td>
<td>14.89</td>
<td>502.3235</td>
</tr>
<tr>
<td>T7L</td>
<td>107</td>
<td>R</td>
<td>lysC, T7L-T8L</td>
<td>40.79</td>
<td>40.44</td>
<td>175.1190</td>
</tr>
<tr>
<td>T8L</td>
<td>108-125</td>
<td>TVAAPSVFIFPSDEQLK</td>
<td>trypsin</td>
<td>45.05</td>
<td>46.26</td>
<td>1946.027</td>
</tr>
<tr>
<td>T9L</td>
<td>126-141</td>
<td>SGTA5VVCCLNNFYPR</td>
<td>trypsin</td>
<td>52.60</td>
<td>53.19</td>
<td>1740.8738</td>
</tr>
<tr>
<td>T10L</td>
<td>142-144</td>
<td>EAK</td>
<td>lysC, T9L-T10L</td>
<td>48.21</td>
<td>47.69</td>
<td>347.1925</td>
</tr>
<tr>
<td>T11L</td>
<td>145-148</td>
<td>VQWK</td>
<td>trypsin</td>
<td>15.35</td>
<td>15.99</td>
<td>560.3191</td>
</tr>
<tr>
<td>T12L</td>
<td>149-168</td>
<td>VDNALQSGNSQESVTEQDSK</td>
<td>trypsin</td>
<td>19.46</td>
<td>19.41</td>
<td>2135.9687</td>
</tr>
<tr>
<td>T13L</td>
<td>169-182</td>
<td>DSTYSLSSTTLTSDK</td>
<td>trypsin</td>
<td>37.00</td>
<td>36.99</td>
<td>1502.7584</td>
</tr>
<tr>
<td>T14L</td>
<td>183-187</td>
<td>ADYEK</td>
<td>trypsin</td>
<td>4.15</td>
<td>3.44</td>
<td>625.2828</td>
</tr>
<tr>
<td>T15L</td>
<td>188-189</td>
<td>HK</td>
<td>Trypsin misclavage, T15L-T16L</td>
<td>20.60</td>
<td>20.77</td>
<td>284.1717</td>
</tr>
<tr>
<td>T16L</td>
<td>190-206</td>
<td>VYACEVTHQGLSSPVTK</td>
<td>trypsin</td>
<td>24.73</td>
<td>25.00</td>
<td>1818.9055</td>
</tr>
<tr>
<td>T17L</td>
<td>207-210</td>
<td>SFNR</td>
<td>trypsin</td>
<td>6.97</td>
<td>6.93</td>
<td>523.2623</td>
</tr>
<tr>
<td>T18L</td>
<td>211-213</td>
<td>GEC</td>
<td>lysC native, T17H-T17L-T18L</td>
<td>16.72</td>
<td>18.58</td>
<td>308.0911</td>
</tr>
</tbody>
</table>

In the comparison of the two molecules, an amino acid mutation of S258Y was found from the tryptic peptide T19H of the RNAi-mediated molecule (Figure 2.2). The reason for the mutation was likely due to an error in the PCR procedure (clonal error) as direct DNA
sequencing confirmed an X to Y nucleotide substitution at position XXX (data not shown), which changed the serine codon (XXX) to a tyrosine (YYY). Also in the RNAi-mediated molecule, a second mutation F174I or F174L, which is partially mutated (~40%), was found for the T13H tryptic peptide (Figure 2.3). This mutation could be attributable to extensive passaging (>21 passages) of the CHO cell line producing the RNAi-mediated molecule. To confirm the exact site of the mutation, a non-specific cleavage form of the T13H peptide, which has a shorter length, was chosen for collision induced dissociation-tandem mass spectrometry (CID-MS2) to obtain sufficient backbone cleavage for the site assignment (Figure 2.3B). The 100% sequence coverage is critical to confirm the sequence or identify a variation, as in this case. Once the position of variation is identified, the site can be pinpointed for correction.
Figure 2.2 A. Precursor mass of normal (Left) and S258Y mutated tryptic peptide (right), theoretical and observed monoisotopic mass are indicated in the figure; B  CID-MS2 of the precursor ion from A, normal (top) and S258Y mutated tryptic peptide (bottom).
Figure 2.3 A Precursor mass of normal and F174L/I mutated tryptic peptide; B CID-MS2 of the precursor ion from A, normal (top) and F174L/I mutated tryptic peptide (bottom).

Because two amino acid mutations were found, additional potential mutations were explored.

For this study, we used dimethyl labeling of the digest peptides, with dimethyl (2CH₂) for
rituximab, and isotopically-labeled dimethyl (2CD₂) for the RNAi-mediated molecule to examine more carefully the entire amino acid sequence. With this approach, the difference of the mass for each tryptic peptide between the two products should be constant (2CD₂ – 2CH₂ = 4 Da per primary amine) if there is no amino acid mutation. For the mutation of S258Y, the delta mass would be 80 Da (Figure 2.4A), and with the corresponding MS2 spectrum indicating the mutation site (Figure 2.4B). For the mutation of F174I or F174L, the delta mass would be -22 Da (Figure 2.5A), and the MS2 spectrum provides site confirmation (Figure 2.5B). Importantly, we did not observe any other mutation sites following the complete assignment of the peptides.
Figure 2.4 A Precursor mass of the peptide with CH2 dimethyl labeling (left) and the precursor mass of the peptide with CD2 dimethyl labeling (right). The mass difference is not a constant 4 Da but 80 Da, accounting for the difference between S and Y (76 Da), as indicated in the figure; B CID-MS2 of the precursor ion from A, CH2 dimethyl labeling (top) and CD2 dimethyl labeling (bottom).
Figure 2.5 A Precursor mass of the peptide with CH2 dimethyl labeling (left) and the precursor mass of the peptide with CD2 dimethyl labeling (right). The mass difference is not constant $4 \times 3 = 12$ Da but 22 Da, accounting for the difference between F and I or L (34 Da), as indicated in the figure; B CID-MS2 of the precursor ion from A, CH2 dimethyl labeling (top) and CD2 dimethyl labeling (bottom).
Stable isotope labeling by amino acids in cell culture (SILAC) to compare two different manufacturing processes for detecting sequence variation,\textsuperscript{14,15} labeling light SILAC in a reference material in comparison to another manufacturing lot with heavy SILAC labeling. For the sequence variation in biosimilar or biobetter products, however, it is not practical to use the metabolic approach (i.e., SILAC) to reproduce the cell culture conditions used to make the reference product because these conditions may be known by only the company that markets the product. Thus, we purchased the reference product (rituximab) and used the above external chemical approach (i.e., dimethyl labeling). It should be noted that dimethyl labeling is typically used for relative quantitation,\textsuperscript{16} but, in this case, we adopted the labeling method for mutation analysis. The labeling approach provides an additional measure of assurance to the results of peptide mapping sequence analysis. Moreover, the delta mass shift can provide a reasonable indication of the potential sequence variation, which can be subsequently determined by MS2 analysis.

**Disulfide linkage analysis**

Enzymatic digestion without reduction followed by LC-MS was used for disulfide analysis. All the expected disulfides were identified by accurate precursor mass measurement and CID or electron transfer dissociation (ETD) fragmentation of the precursor ion. The assignment approach is illustrated in Figure 2.6, with panel A the MS spectrum for the precursor ion, panel B for the CID-MS2 spectrum, and panel C for the ETD-MS2 spectrum. Using this approach, all the disulfide linkages were successfully identified. Both anti-CD20 mAbs were found to contain identical disulfide linkages, as shown in Table 2.2.
Figure 2.6 A precursor mass of 1029.1143 (3+); B CID-MS2 of the precursor in A; C ETD-MS2 of the precursor for the disulfide-linked peptide (Cys22-Cys96). For ETD-MS2, a different charge of the precursor ion, 772.59 (4+), was used for fragmentation.
Table 2.2 Summary of disulfide identification in rituximab and RNAi-mediated mAb.

<table>
<thead>
<tr>
<th>Site</th>
<th>Sequence</th>
<th>Disulfide linkages</th>
<th>Enzyme</th>
<th>Rituximab (RT/min)</th>
<th>RNAi-mediated (RT/min)</th>
<th>[MH]+ (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2H</td>
<td>20-23</td>
<td>MSCK (C22)</td>
<td>LysC plus trypsin</td>
<td>40.41</td>
<td>40.69</td>
<td>3085.3257</td>
</tr>
<tr>
<td>T9H</td>
<td>75-98</td>
<td>SSSTAYMQLSSLTEDSAVYYCAR (C96)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T13H</td>
<td>152-215</td>
<td>DYPFPEPVTSWNSGALTGTVHFTPAVLQSSGLYS LSSVTTPSSSLGTQITYKGVNHKPSNTK (C204)</td>
<td>LysC plus trypsin</td>
<td>57.40</td>
<td>57.69</td>
<td>7917.9266</td>
</tr>
<tr>
<td>T12H</td>
<td>138-151</td>
<td>STSGTGTAALGCVVK (C148)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T17H</td>
<td>223-226</td>
<td>SCDK (C224)</td>
<td>LysC</td>
<td>1.72</td>
<td>3.21</td>
<td>1261.4936</td>
</tr>
<tr>
<td>T1718L</td>
<td>207-213</td>
<td>SFNRC (C213)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T18H</td>
<td>227-250</td>
<td>THTCPPCPAPELGGPSVFLFPPK (C230C233)</td>
<td>LysC plus trypsin</td>
<td>56.83</td>
<td>57.12</td>
<td>5005.4952</td>
</tr>
<tr>
<td>T18H</td>
<td>227-250</td>
<td>THTCPPCPAPELGGPSVFLFPPK (C230C233)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T20H</td>
<td>260-267</td>
<td>TPEVTCVVVDVSHDEPEVK (C265)</td>
<td>LysC plus trypsin</td>
<td>32.29</td>
<td>33.44</td>
<td>2329.1050</td>
</tr>
<tr>
<td>T26H</td>
<td>325-326</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T34H</td>
<td>365-374</td>
<td>NQVSLTCLVK (C371)</td>
<td>LysC plus trypsin</td>
<td>39.70</td>
<td>40.00</td>
<td>3845.8309</td>
</tr>
<tr>
<td>T39H</td>
<td>421-443</td>
<td>WQQGNVFCSCVMHEALHNHYTQK (C429)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2L</td>
<td>19-23</td>
<td>VTMTCR (C23)</td>
<td>LysC plus trypsin</td>
<td>38.94</td>
<td>39.39</td>
<td>3528.5505</td>
</tr>
<tr>
<td>T5L</td>
<td>77-102</td>
<td>VEAEDATTPYCQWTSNPPFTGGGTK (C87)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T9L</td>
<td>126-141</td>
<td>SGTAQVCLLNNFYPYR (C133)</td>
<td>LysC plus trypsin</td>
<td>45.40</td>
<td>45.80</td>
<td>3556.7563</td>
</tr>
<tr>
<td>T16L</td>
<td>190-206</td>
<td>VYACEVTHQGLSSPVTK (C193)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the analysis, unpaired cysteines were also identified. These unpaired cysteines were found as free cysteines at pH 6.8 with trypsin digestion and at pH 2 with pepsin digestion (Table 2.3), but they were not observed at pH 8. It is likely that the free cysteines became linked disulfides (scrambled) in the alkaline pH (see Table 2.4). For accurate results, it is thus important to measure the free cysteine levels at low pH, and, especially, the digestion pH needs to be optimized to minimize disulfide scrambling. 17
Table 2.3 Free cysteine determined at different digestion pHs for rituximab and RNA-mediated mAb.

<table>
<thead>
<tr>
<th></th>
<th>pH 8.0</th>
<th></th>
<th>pH 6.8</th>
<th></th>
<th>pH 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rituximab (%)</td>
<td>RNAi-mediated (%)</td>
<td>Rituximab (%)</td>
<td>RNAi-mediated (%)</td>
<td>Rituximab (%)</td>
</tr>
<tr>
<td>T12H</td>
<td>N.D.</td>
<td>N.D.</td>
<td>5.4 ± 0.9</td>
<td>0.4 ± 0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>(C148)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T13H</td>
<td>N.D.</td>
<td>N.D.</td>
<td>1.3 ± 0.3</td>
<td>N.D.</td>
<td>0.2</td>
</tr>
<tr>
<td>(C204)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T20H</td>
<td>N.D.</td>
<td>N.D.</td>
<td>3.0 ± 0.3</td>
<td>0.2 ± 0.0</td>
<td>3.4</td>
</tr>
<tr>
<td>(C265)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T26H</td>
<td>--</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>3.1</td>
</tr>
<tr>
<td>(C325)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T34H</td>
<td>N.D.</td>
<td>N.D.</td>
<td>10.7 ± 1.8</td>
<td>0.2 ± 0.1</td>
<td>1.6</td>
</tr>
<tr>
<td>(C371)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T39H</td>
<td>N.D.</td>
<td>N.D.</td>
<td>4.7 ± 0.3</td>
<td>0.1 ± 0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>(C429)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2L</td>
<td>N.D.</td>
<td>N.D.</td>
<td>0.4 ± 0.2</td>
<td>0.0 ± 0.0</td>
<td>N.D.</td>
</tr>
<tr>
<td>(C23)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5L</td>
<td>N.D.</td>
<td>N.D.</td>
<td>0.2 ± 0.0</td>
<td>0.0 ± 0.0</td>
<td>N.D.</td>
</tr>
<tr>
<td>(C87)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T9L</td>
<td>N.D.</td>
<td>N.D.</td>
<td>1.7 ± 0.2</td>
<td>0.1 ± 0.0</td>
<td>N.D.</td>
</tr>
<tr>
<td>(C133)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T16L</td>
<td>N.D.</td>
<td>N.D.</td>
<td>1.1 ± 0.1</td>
<td>0.1 ± 0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>(C193)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T17L</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>0.2</td>
</tr>
<tr>
<td>(C213)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: the percent of free cysteine at pH 8.0 and 6.8 was obtained by the measured amount of the free cysteine form divided by the total cysteines on the specific tryptic peptide, assuming each cysteine is alkylated. The percent of free cysteine at pH 2 was obtained by the ratio of the observed free cysteines divided by the total cysteines after reduction with TCEP. “N.D.” represents not detectable (the values is too low to be observed), and “--” means not measurable due to the tryptic peptide length (too short to retain in the LC chromatogram).
Table 2.4 Percent scrambled disulfides at different digestion pHs for rituximab and RNAi-mediated mAb.

<table>
<thead>
<tr>
<th></th>
<th>pH 8.0</th>
<th>pH 6.8</th>
<th>pH 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rituximab (%)</td>
<td>RNAi-anti-CD20 (%)</td>
<td>Rituximab (%)</td>
</tr>
<tr>
<td>T2H(C22)-T20H(C265)</td>
<td>2.2</td>
<td>0.4</td>
<td>1.6</td>
</tr>
<tr>
<td>T2H(C22)-T34H(C371)</td>
<td>0.4</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>T9H(C96)-T26H(C325)</td>
<td>0.2</td>
<td>0.1</td>
<td>N.D.</td>
</tr>
<tr>
<td>T12H(C148)-H26H(C325)</td>
<td>0.3</td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td>T17H(C224)-T20H(C265)</td>
<td>0.3</td>
<td>0.6</td>
<td>N.D.</td>
</tr>
<tr>
<td>T18H(C230C233) – intra-linked</td>
<td>9.4</td>
<td>0.4</td>
<td>5.8</td>
</tr>
<tr>
<td>T20H(C265)-T20H(C265)</td>
<td>4.9</td>
<td>1.4</td>
<td>1.5</td>
</tr>
<tr>
<td>T20H(C265)-T39H(C429)</td>
<td>1.1</td>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td>T20H(C265)-T16L(C193)</td>
<td>1.0</td>
<td>1.7</td>
<td>N.D.</td>
</tr>
<tr>
<td>T26H(C325)-T34H(C371)</td>
<td>1.6</td>
<td>1.7</td>
<td>0.6</td>
</tr>
<tr>
<td>T26H(C325)-T39H(C429)</td>
<td>0.7</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>T26H(C325)-T16L(C193)</td>
<td>0.2</td>
<td>0.8</td>
<td>N.D.</td>
</tr>
</tbody>
</table>

Note: the percent scrambled disulfide was obtained by the measured amount of scrambled disulfide divided by the total disulfide-linked forms.

To prove that the scrambled disulfides were caused by the sample preparation and were not in the original samples, a study was conducted to examine the correlation of scrambling as a function of pH (pH 8, 6.8, and 2). For optimum digestion efficiency, trypsin was used at pH 8.
and pH 6.8, and pepsin at pH 2. Scrambled disulfides were found in trace amounts for digestion at pH 6.8, and significant amounts at pH 8, but no scrambling was observed at pH 2 (Table 2.4). Rituximab appears to have a higher amount of scrambling than the RNAi-mediated molecule due to a higher amount of free cysteines in rituximab (Table 2.3). Nevertheless, scrambling was not observed in either product at pH 2. It can be concluded that if scrambled disulfides are observed when examining a biopharmaceutical, a good test is to explore if scrambling decreases with digestion pH. Any “scrambling” observed at pH 2 likely represents the true structure of the molecule and not an artifact of sample preparation. In this case, a careful study should be conducted to obtain additional data points with different digestion pH to extrapolate the extent of scrambling at the intercept (i.e., extrapolate to pH 0) to reflect the true scrambling.

The Ellman reaction, coupling a chromophore (dithionitrobenzoic acid or DTNB) to a thiol group, has been used to detect free cysteine in a protein. This method, however, is insensitive and often requires hundreds of milligrams of mAb if the free cysteine is at low levels (i.e. ≤ 1%). Instead, we used the LC-MS approach for free cysteine analysis because it uses minute amounts of mAb (2 µg) and is quite sensitive for detecting free cysteine at low levels. Because the ionization efficiency in the LC-MS analysis could be different for free and disulfide linked cysteines, the relative proportion of the various cysteines would, however, only be an estimate unless synthetic standards are used for correction of potentially different responses. To minimize the difference in responses in the LC-MS experiment, we first converted the free cysteines at pH 8.0 and 6.8 to the alkylated forms, and then divided that value by the value derived from reduction and alkylation of all cysteines in a separate
experiment. Because the alkylation experiment is difficult to perform at pH 2, the percentage of free cysteine at this pH was obtained by the ratio of the observed free cysteine divided by the total free cysteines; the latter was derived from the reduction with TCEP at low pH in a separate experiment (see the section of enzymatic digestion in Materials and Methods). These results are presented in Table 2.3. It should be noted that disulfide scrambling caused by the free cysteine could be a concern for drug quality or safety. LC-MS has been known to be able to detect structural variants with a high degree of sensitivity and precision; however, proper sample handling and precise quantitation of low-level free cysteine residues have not been well documented. Thus, it is important to study the disulfide linkages carefully with a sensitive technique, as described here.

It should be noted that most therapeutic antibodies contain low levels of free cysteines.\textsuperscript{22-25} It has also been reported that human IgGs in serum possess low levels of free sulphhydryl as well (most likely associated with human IgG2, with some evidence in IgG1).\textsuperscript{26-28} Thus, a proper formulation buffer to control the pH is needed to avoid scrambling while still maintaining high solubility. The pH of the formulation buffer is between 5.5 and 6 for both rituximab and the RNAi-mediated molecules.

**Fc glycosylation**

The common glycoforms at the conserved Fc site (Asn300) were identified at the tryptic peptide T23H, e.g., glycan without galactose (G0F), with one galactose (G1F), with two galactoses (G2F), without galactose and without core fucose (G0), with one galactose and
without core fucose (G1), with two galactoses and without core fucose (G2), without
galactoses and without core fucose and without one N-acetyl glucosamine (G0-NGlc).

The reduction of core fucosylation, as expected, was found for the RNAi-mediated molecule.
The glycan distribution for the two mAbs are shown in Table 2.5, the a-fucosylated glycans
(without core fucose such as G0, G1, and G0-NGlc) were increased to > 70% for the RNAi-
mediated molecule, while being < 10% in rituximab. The percent distribution was obtained
through the amount of each glycopeptide, divided by the sum of all glycopeptides including
the non-glycosylated form. The reduction of the core fucose induced a 17-fold improvement
in FcγRIIIa binding, and an increase in specific cell lysis by up to 30%, as determined in an
ADCC assay.12

Table 2.5 Comparison of glycan distribution for rituximab and RNA-mediated mAb.

<table>
<thead>
<tr>
<th>Glycans</th>
<th>Rituximab ± SD (%)</th>
<th>RNAi-mediated ± SD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G0F</td>
<td>34.2 ± 1.3</td>
<td>14.6 ± 0.8</td>
</tr>
<tr>
<td>G0F-NGlc</td>
<td>7.6 ± 0.6</td>
<td>2.3 ± 0.6</td>
</tr>
<tr>
<td>G1F</td>
<td>35.6 ± 2.7</td>
<td>4.2 ± 0.6</td>
</tr>
<tr>
<td>G1F-NGlc</td>
<td>4.7 ± 0.6</td>
<td>0.4 ± 0.1</td>
</tr>
<tr>
<td>G2F</td>
<td>9.6 ± 0.3</td>
<td>0.7 ± 0.1</td>
</tr>
<tr>
<td>G0</td>
<td>2.8 ± 0.1</td>
<td>46.8 ± 1.6</td>
</tr>
<tr>
<td>G0-NGlc</td>
<td>1.2 ± 0.1</td>
<td>14.6 ± 2.1</td>
</tr>
<tr>
<td>G1</td>
<td>0.9 ± 0.1</td>
<td>9.1 ± 0.6</td>
</tr>
<tr>
<td>G1-NGlc</td>
<td>0.1 ± 0.0</td>
<td>1.1 ± 0.1</td>
</tr>
<tr>
<td>Man5</td>
<td>3.1 ± 0.1</td>
<td>5.0 ± 0.4</td>
</tr>
<tr>
<td>Man4</td>
<td>0.1 ± 0.0</td>
<td>0.2 ± 0.1</td>
</tr>
<tr>
<td>Man3</td>
<td>0.1 ± 0.0</td>
<td>0.9 ± 0.0</td>
</tr>
</tbody>
</table>

Note: the percent glycosylation was obtained by the measured amount of the specific
glycopeptide divided by the total of all forms of the glycopeptides including the non-
glycosylated form. The SD was determined from 6 measurements.
Other modifications

MAbs with “Q” or “E” N-terminal amino acids could easily convert to pyro-Glu. The N-terminus of the heavy and light chains of rituximab contains the Q amino acid residue. Thus, loss of NH3 (minus 17 Da) was examined in the N-terminal peptides, and both the non-modified and pyro-Glu forms were found. The C-terminus of the heavy chain could also be cleaved by carboxypeptidase present during mAb production in the CHO cell. Thus, the K clipping at the C-terminus of the heavy chain was examined. Both mAbs have the same pyro-Glu formation at the N-terminus, and near complete K clipping at the C-terminus of the heavy chain. Other typical modifications, including Asn deamidation and Met oxidation, were also found. No significant differences were observed for these modifications, except Met256 and pyro-Glu at the light chain (see t-test results in table 2.6). The minor difference at Met256 could be due to this amino acid residue being very close to the mutation site (S258Y), and the difference of pyro-Glu on the light chain could be due to the difference in storage time and conditions for the two products.
Table 2.6 Comparison of common modifications in two anti-CD20 mAb

<table>
<thead>
<tr>
<th></th>
<th>MabThera ± SD / %</th>
<th>RNAi-Fuc-antiCD20 ± SD / %</th>
<th>T-Test (P-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Oxidation (n=3)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met34</td>
<td>1.9 ± 0.2</td>
<td>2.5 ± 0.3</td>
<td>0.053</td>
</tr>
<tr>
<td>Met81</td>
<td>3.6 ± 0.2</td>
<td>3.7 ± 0.8</td>
<td>0.8866</td>
</tr>
<tr>
<td>Met256</td>
<td>1.3 ± 0.2</td>
<td>6.2 ± 2.1</td>
<td><strong>0.0052</strong></td>
</tr>
<tr>
<td><strong>Succinimide (n=6 for mabThera, n=9 for RNAi-Fuc-antiCD20)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N55</td>
<td>2.3 ± 0.4</td>
<td>2.7 ± 0.3</td>
<td>0.0667</td>
</tr>
<tr>
<td>N319</td>
<td>4.1 ± 2.0</td>
<td>5.1 ± 1.9</td>
<td>0.3493</td>
</tr>
<tr>
<td>N388</td>
<td>3.1 ± 0.3</td>
<td>2.7 ± 0.6</td>
<td>0.177</td>
</tr>
<tr>
<td>N394</td>
<td>0.4 ± 0.1</td>
<td>0.4 ± 0.0</td>
<td>0.588</td>
</tr>
<tr>
<td><strong>N-terminal and C-terminal modifications (n=3 for MabThera, n=6 RNAi-Fuc-antiCD20)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-N-term Pyro</td>
<td>100 ± 0.0</td>
<td>99.6 ± 0.6</td>
<td>0.251</td>
</tr>
<tr>
<td>L-N-term Pyro</td>
<td>97.6 ± 0.2</td>
<td>80.4 ± 2.1</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>C-term Processing</td>
<td>0.4 ± 0.1</td>
<td>0.6 ± 0.1</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Note: percent of modification (i.e. oxidation) at a particular site was obtained by the observed modification at that site divided by the total forms at that site (i.e. both oxidized and nonoxidized forms). Same type of measurement was used for other modifications. Deamidation was measured at low pH (2 and 6.8), and there is no aspartate forms (either alpha or beta) except succinimide form. The t-test value with more than 99% confidence (≤ 0.01) is shown in red. As shown, there is no statistic difference for the majority of modification sites, except Met256 and pyro-glu at the light chain. The minor difference at Met256 could be due to this Met is very close to the mutation site (S258Y), and the difference of pyro-glu at the light chain could be due to the difference in storage length and conditions for the two products.

2.4 Discussion
The full structure of the two anti-CD20 mAbs was compared using LC-MS analysis of the enzymatic peptide mixtures. As shown, the full amino acid sequence, including N-terminal,
C-terminal, disulfide linkages, and glycosylation site and structure, were successfully identified. The methodology can detect amino acid modifications, mutations, free cysteines, and scrambled disulfides. In such a comparative study, the identification of the entire amino acid sequence is critical. A multi-enzyme digestion strategy is necessary, along with complete LC-MS analysis of all the enzymatically-produced peptides. As shown in this work, dimethyl labeling of H and D isotopically labeled derivatizing agents adds a valuable dimension for verification.

We believe that sensitive LC-MS analysis should detect the structural variants as described. However, it is important not only to detect variants but also to have a strategy to rule out other possibilities in order to provide a high degree of assurance that the results are correct. The two orthogonal techniques used in this work not only independently confirm the findings but also can determine the degree of similarity of the two products. We believe that such an extensive comparability study by mass spectrometric techniques should be performed early in the R&D process in order to obtain an essentially identical sequence and the correct up-stream and down-stream process conditions to generate a highly similar antibody. As suggested in the FDA biosimilar draft guideline (February, 2012), a highly similar antibody could reduce the extent of clinical studies needed for approval and thus decrease the cost of biosimilar drugs which would have a significant impact on the cost of healthcare.

For a comparability analysis, it should be emphasized that even 80 to 90% sequence coverage would not be satisfactory for comparison of two antibodies (e.g., reference vs. biosimilar).
Although it may be time-consuming and labor-intensive, all peptides must be identified and, where appropriate, quantitated. In the future, top down mass spectrometry\textsuperscript{29-31} may be able to achieve 100\% sequence coverage on the intact protein, without enzymatic digestion. With improving sensitivity and enhanced resolution of mass spectrometric instrumentation, the top down approach could ultimately have a great impact in the characterization of biopharmaceuticals.\textsuperscript{32} However, detecting many modifications at once is presently too challenging, as is the identification of all the modifications with associated sites using current software. The combination of both peptide mapping and top down MS may alleviate constraints in both approaches. The middle down, also known as the extended range proteomic analysis approach,\textsuperscript{33-35} which takes advantage of large peptide fragments for high sequence coverage and with less sophisticated modifications than the entire protein, is at present a useful alternative.

2.5 Acknowledgment
This work was supported in part by NIH grant GM 15847. The authors thank Alnylam Pharmaceuticals for providing the RNAi-mediated anti-CD20 antibody and rituximab for this study.

2.6 Materials and Methods
**Samples:** RNAi-mediated mAb was manufactured at Alnylam Pharmaceuticals (Cambridge, MA, USA) and provided as 1 mg/mL x 0.8 mL as previously described.\textsuperscript{12} Rituximab was purchased by Alnylam from Imperial College, London and provided as 10 mg/mL x 0.5 mL. The samples were aliquoted as 10 µL per vial for rituximab (100 µg) and 100 µL per vial for RNAi-mediated mAb (100 µg) and stored at -80 °C before analysis.
**Reagents:** Trypsin was purchased from Promega, endoproteinase Lys-C from Wako Chemicals USA, pepsin from MP Biomedicals, and endoproteinase Glu-C from Roche Diagnostics. Formaldehyde (CH$_2$O), labeled formaldehyde (CD$_2$O), sodium cyanoborohydride, triethylammonium bicarbonate (TEAB), ammonium hydroxide solution, guanidine hydrochloride (Gn-HCl), dithiothreitol (DTT), iodoacetamide (IAM), ammonium bicarbonate and 1.0M hydrochloric acid (HCl) solution were obtained from Sigma-Aldrich. Tris buffer was from GE Healthcare, and tris(2-carboxyethyl)phosphine hydrochloride and formic acid from Fisher Scientific. LC-MS grade water and acetonitrile were purchased from VWR. Amicon centrifugal filters (10 kDa molecular weight cutoff) were obtained from EMD Millipore.

**Enzymatic Digestion:** An aliquot of 10 µL of rituximab and 100 µL of RNAi-mediated mAb solution (100 µg) was denatured with 6 M guanidine hydrochloride containing 100 mM ammonium bicarbonate (pH 8), reduced with 5 mM DTT for 30 min at 37 °C, and then alkylated with 20 mM IAM in the dark for 45 min at room temperature. The reduced and alkylated protein was buffer exchanged with 100 mM ammonium bicarbonate (pH 8) or 50 mM Tris (pH 6.8) using a 10 kDa molecular weight cutoff filter to a concentration of 1 mg/mL (100 µL). For tryptic digestion, trypsin (1:50, w/w) was added to the protein solution at room temperature. After 8 h, the enzyme was added a second time (1:50, w/w) and the digestion continued at room temperature for 12 h. For Lys-C digestion, the endoproteinase Lys-C (1:50, w/w) was added to the protein solution for 4 h at 37 °C. For Lys-C plus trypsin digestion, the protein solution was added with endoproteinase Lys-C (1:50 w/w) for 4 h at 37 °C, and then trypsin (1:50 w/w) for 20 h at room temperature. For pepsin digestion, the protein solution was dissolved in 10 mM HCl (pH 2). Pepsin (1:10, w/w) was added to the
protein solution and incubated at 37 °C for 30 min. The reaction was quenched by adjusting the pH to 5 with 100 mM ammonium bicarbonate. For digestion without reduction (for disulfide assignment), the same digestion protocol as above was applied but without the reduction and alkylation steps. For quantitation of free cysteines at pH 6.8, 100 µg sample was alkylated with 20 mM IAM (in 6M Gn-HCl) in pH 6.8 for 2 hours, and then half of the protein solution was digested by trypsin without reduction, and the other half reduced, alkylated and digested by trypsin. For quantitation of free cysteines at pH 2, the protein solution was digested in pepsin, and then half was reduced by 10 mM TCEP at pH 5.0 for an hour at room temperature. In all cases (except pepsin digestion), digestion was terminated by addition of 1% formic acid. An aliquot of 2 µg of the enzyme digest was analyzed per LC-MS run.

**Dimethyl labeling:** After digestion, the digests (20 µL, nearly completely dried) were reconstituted in 20 µL 0.1 mM TEAB, and 1 µL of 37% (v/v) CH₂O was added for the rituximab sample, while 1.9 µL of 20% (v/v) CD₂O was added for the RNAi-Fuc-anti-CD20 mAb sample. Each sample was then added 5.8 µL 1M NaBH₃CN, and the solution was incubated in a fume hood for 1 h at room temperature. Finally, the reaction was quenched by addition of 4 µL 10% ammonia solution, and then 8 µL of formic acid on an ice top to prevent frothing and/or heating of the sample¹⁶. The samples of rituximab and RNAi-Fuc-anti-CD20 mAb were mixed equally for subsequent LC-MS analysis.

**LC-MS:** An Ultimate 3000 nano-LC pump (Dionex, Mountain View, CA) and a self-packed C18 column (Magic C18, 200 Å pore and 5 µm particle size, 75 µm i.d. × 15 cm) (Michrom Bioresources, Auburn, CA) was coupled online to an LTQ-Orbitrap-ETD XL mass spectrometer (Thermo Fisher Scientific, San Jose, CA) through a nanospray ion source (New
Objective, Woburn, MA). Mobile phase A (0.1% formic acid in water) and mobile phase B (0.1% formic acid in acetonitrile) were used for the gradient consisting of (i) 20 min at 2% B for sample loading and 5 min for desalting at 0.3 µL/min (20 min desalting for pepsin or dimethylated digests); (ii) linear from 2 to 5% B for 2 min; (iii) linear from 5 to 35% B for 60 min; (iv) linear from 35 to 90% B for 3 min; and finally (v) isocratic at 90% B for 5 min. The flow rate of the column was maintained at 0.2 µL/min, and the mass spectrometer started to record data after 5 min of the gradient. The LTQ-Orbitrap-ETD XL mass spectrometer was operated initially in the data-dependent mode as follows: survey full-scan MS spectra (m/z 300-2000) were acquired in the Orbitrap with a mass resolution of 30,000 at m/z 400 (with an ion target value of 5 \times 10^5 ions), followed by nine sequential MS2 scans using the LTQ. For disulfide mapping, the MS was switched automatically between MS (scan 1 in the Orbitrap), CID-MS2 (scan 2 in the LTQ), and ETD-MS2 (scan 3 in the LTQ). Briefly, after a survey MS spectrum from m/z 300 to 2000, subsequent CID-MS2 and ETD-MS2 steps were performed on the same precursor ion with a ±2.5 m/z isolation width. Any incomplete assignment in the CID-MS2 and ETD-MS2 spectra was repeated by targeting the desired ions, e.g., the same precursor but with a different charge state, to gain additional linkage information. This targeted approach was repeated (e.g., targeting multiple charges of a precursor ion or the same disulfide-linked peptide but with different enzymatic cleavages or miscleavages) until the linkage information was complete. In addition, a targeted CID-MS3 after ETD for ions of interest was performed as necessary.

**Peptide Assignment:** The spectra generated in the CID-MS2 step were searched against spectra of theoretical fragmentations (b and y ions) of rituximab sequence with a mass tolerance of ≤5 ppm for the precursor ions and with enzyme specificity, using a Sequence
probability score (> 95% confidence) as the filter. For peptides with miscleavages or a mass tolerance >5 ppm (but less than 20 ppm) of the precursor ion, confirmation required manual inspection to match all highly abundant product ions. The sequence coverage generated by each enzymatic map was combined as a total coverage.

**Disulfide assignment:** The anticipated disulfide-linked tryptic or multi-enzyme digested peptide masses with different charges were first calculated and then matched to the observed masses in the LC-MS chromatogram. The matched masses (with < 5 ppm mass accuracy) were further confirmed by the corresponding CID-MS2 and ETD-MS2 fragmentation, as well as by CID-MS3 fragmentation, as needed.

**Glycan structure identification:** Theoretical masses of glycan structures such as G0F, G1F and G2F were added to the tryptic peptide backbone (EEQYNSTYR). The anticipated glycopeptide masses with different charges were thus obtained to match the observed masses in the LC-MS chromatogram. The matched masses (with ≤ 5 ppm mass accuracy) were further confirmed by the corresponding CID-MS2 fragmentation.

### 2.7 References


13. Anderson DR; Hanna, N; Leonard, GE; Newman, RA; Reff, ME; Rastetter, WH. Therapeutic application of chimeric and radiolabeled antibodies to human B lymphocyte restricted differentiation antigen for treatment of B cell lymphoma 1998. US patent 5736137


Chapter 3 Characterization of Site-Specific Glycan Structures in Glucocerebrosidase from Commercial (Cerezyme®) and RNAi-mediated Molecules by LC-MS
3.1 Abstract

Glycan structures of recombinant glucocerebrosidase (Cerezyme®), at Asn19, Asn59, Asn146, and Asn270 sites, are comprehensively characterized and compared to a recombinant glucocerebrosidase molecule produced by adding a small interfering RNA (siRNA) to silence $MGAT1$ gene, which expresses N-acetylglucosaminyltransferase I (GlcNAc-TI) that catalyzes the formation of complex and hybrid-type glycans. The analysis was achieved by determination of glycans attached to enzymatically digested peptides containing these four Asn sites by LC-MS. The glycosylation site, occupancy, and glycan distribution are determined by this approach and used as a basis of comparison between the two products. At the Asn19 site, Man3GlcNAc2 was found as the major glycan species for both products. The other three glycosylation sites (Asn 59, 146, and 270) contained similar mannose cores with additional fucose, Man3GlcNAc2Fuc, being identified as the major glycan for Cerezyme®. However, in the RNA interference (RNAi)-mediated glucocerebrosidase, the most abundant glycan was found to be a slightly larger mannose, Man5GlcNAc2 glycan. The slight difference could be due to the production of the glycans using different stages of the glycan biosynthesis pathway. As for Cerezyme®, the mannose core was obtained after the complex-type glycans have been first synthesized and then removed later by glycosidases. While for the RNAi-mediated molecule, the mannose core was obtained by silencing the gene to encode GlcNAc-TI before the formation of the complex-type glycans.

3.2 Introduction

Cerezyme®, a recombinant glucocerebrosidase expressed in Chinese hamster ovary (CHO) cells, is used to treat Gaucher’s disease $^{1,2}$, a disease caused by a hereditary deficiency of the
enzyme glucocerebrosidase³.⁴. When the enzyme is defective, glucosylceramide accumulates in macrophages to eventually cause an enlarged spleen and liver, liver malfunction, low blood platelets, and anemia⁵.⁶. However, to achieve the therapeutic effect with recombinant glucocerebrosidase (for effective delivery to the macrophages), the glycan structures of the recombinant glucocerebrosidase needs to be modified. Thus, Cerezyme® is produced by the remodeling of glycans, in which the complex glycans are sequentially digested to the mannose core by three glycosidases in order to enhance the delivery of the therapeutic via mannose receptors⁷.⁸. To achieve the same goal but with a different approach to modify the glycans, a recombinant glucocerebrosidase can be produced by exogenously adding small interfering RNA (siRNA) to obtain the similar mannose core; the siRNA specifically will silence the MGAT1 gene which encodes N-acetylglucosaminyltransferase I (GlcNAc-TI), a key enzyme to initiate complex glycan formation⁹.¹⁰.

As described in this chapter, glycans in both proteins, Cerezyme® and RNA interference (RNAi)-mediated glucocerebrosidase, were characterized and compared by analysis of the glycopeptides on both proteins using LC-MS approach. Results demonstrate that the major glycan in Cerezyme® is the mannose core (Man3GlcNAc2, with and without core fucose), while the most abundant glycans in the RNAi-mediated product is Man5GlcNAc2. Glucocerebrosidase with Man5GlcNAc2 was reported to have similar targeting behavior and uptake by macrophages as it with the mannose core¹¹.
3.3 Results and Discussion

Characterization and comparison of glycopeptides of Cerezyme® and RNAi-mediated glucocerebrosidase are described in the following sections.

3.3.1 Glycosylation site analysis

Cerezyme® contains 4 glycosylation sites, Asn19, Asn59, Asn146, and Asn270, as shown in the primary structure of Cerezyme® (see Figure 3.1). After digestion by trypsin, the four glycosylation sites are distributed into 4 separate tryptic peptides (peptide sequences, 8-39, 49-74, 132-155, and 263-277), which are underlined in Figure 3.1.

Figure 3.1 The primary structure of Cerezyme® with glycosylation sites in bold.
The characterization of glycans at each site was conducted in each corresponding tryptic peptide. The amino acid sequence for each glycopeptide was first determined from the deglycosylated counterparts, which were generated by removing glycans using PNGase F. As an example, the identification of deglycosylated tryptic peptide of DLGPTLAD(270)STHHNVR is shown in Figure 3.2. In the assignment, the glycan attached to the Asn residue was converted to Asp after release of the glycans by PNGaseF. The combination of analysis, with and without PNGase F treatment, confirmed the glycosylation status and site of attachment. In which, the peptide backbone could not be observed before PNFase F treatment but with the detection of the peptide backbone (with Asn converted to Asp) after PNFase F treatment. This observation indicated that the peptide is glycosylated. For determination of the glycosylation site, as shown in Figure 3.2A, the deglycosylated peptide (observed m/z 816.9092, 3+) accurately matched to the mass of the theoretical peptide, m/z 816.9081 (with Asn converted to Asp). The peptide fragmentation by CID-MS2 further confirmed the sequence of the peptide and the site of Asn to Asp at amino acid 270 location (Figure 3.2B).
3.3.2 Glycan structure analysis

Once the peptide backbone (deglycosylated peptide) was identified, the glycan attached to the peptide at this site was determined by LC-MS (without PNGase F treatment). In the examination, the masses of the predicted glycopeptides were calculated by adding the masses of peptide backbones to the predicted glycans. Once the masses of the predicted glycopeptides were matched, the CID-MS2 of these precursor ions were examined for confirmation. The identification of the glycan attached to the tryptic peptide containing Asn270 is illustrated in Figure 3.3. The predicted mass of the mannose core glycan, Man3GlcNAc2Fuc, attached to the backbone peptide matched the observed mass within 5
ppm mass accuracy (Figure 3.3A), and the corresponding CID-MS2 spectrum matched to the fragment ions of the predicted Man3GlcNAc2Fuc structure (Figure 3.3B). In assignment of fragment ions in the CID-MS2 spectrum, the mass differences of the oligosaccharides from the precursor should be considered as 146 Da for fucose (Fuc), 162 Da for mannose (Man) or galactose (Gal), 242 Da for mannose phosphate (Man-6-P), 203 Da for N-acetylgalcosamine(GlcNAc), and 292 Da for sialic acid (SA). As shown in Figure 3.3B, the difference in mass between the precursor and m/z 1262.69 indicated loss of a fucose (146 Da difference), loss of a mannose (162 Da difference between the precursor and m/z 1255.20), loss of one mannose plus one fucose (308 Da difference between the precursor and m/z 1181), loss of two mannoses plus one fucose (470 Da difference between the precursor and m/z 1100), and loss of three mannoses plus one fucose and one GlcNAc (835 Da difference between the precursor and m/z 918). Thus, the glycan structure of Man3GlcNAc2Fuc is deduced from the fragmentation pattern.
Figure 3.3 Identification of glycopeptide T22 (263-277) DLGPTLANSTHNNVR (Man3GlcNAc2Fuc) in Cerezyme. A. MS of glycopeptide (1335.6041, 2+); B. CID-MS2 of 1336.11, 2+.

Since glycan variants attached to the same peptide backbone are co-eluted, these glycans can be identified by matching the mass with the sugar difference at the similar retention time. For example, the masses in the retention time between 24.00 and 26.80 min can be extracted to match the glycan variants derived from the Man3GlcNAc2Fuc structure, such as plus additional mannose or minus the core fucose as shown in Figure 3.4. Again, these matched glycans can be further confirmed by their corresponding CID-MS2 spectra, if needed.
Figure 3.4 The glycopeptide assignment for Asn 270 DLGPTLANSTHHNVR in Cerezyme: sum of mass spectrum 24-26.8 min.

Using the same strategy as above, the assignments of glycan structures at different glycopeptides (Asn 19, Asn 59, and Asn 146) are shown in Figures 3.5, 3.6, 3.7, respectively.
Figure 3.5 The glycopeptides assignment in T2 (8-39) SFGYSSVVCVCN(Asn 19)ATYCDSDPPTFPALGTFSR: A. MS of glycopeptide (Man3GlcNAc2) 1501.3118, 3+; B. CID-MS 2 of 1501.97, 3+; C. glycopeptides assignment in T2, sum of mass spectrum 57-58.5 min.
Figure 3.6 The glycopeptides assignment in T5 (49-74) MELSMGPIQAN(Asn 59)HTGTGLLLLTLQEQQK: A. MS of glycopeptide (Man3GlcNAc2Fuc) 1282.6150, 3+; B. CID-MS 2 of 1283.28, 3+; C. glycopeptides assignment in T5, sum of mass spectrum 50-53 min.
3.3.3 Characterization of Glycans in RNAi-Mediated Glucocerebrosidase

As discussed in the introduction, a recombinant glucocerebrosidase was produced by using siRNA to suppress the *MGAT1* gene, which encodes N-acetylglucosaminytransferase I (GlcNAc-TI) – the key enzyme to biosynthesize complex glycans. In this approach, the protein (same sequence as Cerezyme®) was expressed in CHO cells with the addition of small interfering RNA MGAT1. The product was then purified (done by Alnylam). The desired protein along with other co-purified CHO proteins were further separated by SDS-PAGE.
The gel band of RNAi-mediated molecule was cut for in-gel tryptic digestion and LC-MS analysis.

In the LC-MS analysis, the glycan structure of RNAi-mediated glucocerebrosidase was analyzed as for Cerezyme®. As an example, identification of a glycan attached to tryptic peptide containing Asn270 is illustrated in Figure 3.8. The predicted mass of Man5GlcNAc2 attached to the backbone peptide matched to the observed mass within 5 ppm mass accuracy (Figure 3.8A), and the corresponding CID-MS2 spectrum matched to the fragment ions of the predicted Man5GlcNAc2 structure (Figure 3.8B). It should be noted that a slightly larger mannose (Man5) than Man3 was found as the major glycan for the RNAi-mediated molecule at Asn 270 site. Similarly other glycans attached to the same site (Asn 270) can be identified by matching the mass with the sugar difference at the similar retention time (Figure 3.8C).
The glycan structures at Asn 19, Asn 59, and Asn 146 sites are analyzed, shown in Figure 3.9, 3.10, and 3.11, and compared with Cerezyme®, see Table 3.1. In the comparison, similar glycans were found at Asn19. Both contain Man3GlcNAc2 as the major glycan species. The other three glycosylation sites (Asn 59, 146, and 270), a similar core with additional core fucose, Man3GlcNAc2Fuc, is found as the major glycan for Cerezyme®. However, in the RNAi-mediated molecule, the most abundant glycan is Man5GlcNAc2 glycan. The difference could be due to the fact that the remodeling of the glycans used different glycan
biosynthesis pathways. Here, glycan was initially formed as the Glc3Man9GlcNAc2 precursor in the endoplasmic reticulum, then trimmed down to Man5GlcNAc2 as second stage in Golgi\textsuperscript{12, 13}. In the late processing steps, N-acetylglucosamine could be added to Man5GlcNAc2 by the catalysis of GlcNAcT-I, which initiates the formation of hybrid and complex glycan structure. GlcNAcMan5GlcNAc2 is further trimmed down to GlcNAcMan3GlcNAc2 structure. Once GlcNAcMan3GlcNAc2 glycan forms, a second GlcNAc is added to produce the precursor of all biantennary, complex-type glycans\textsuperscript{9, 14}. The incomplete removal of mannose from GlcNAcMan5GlcNAc2 generates GlcNAcMan4GlcNAc2, which results in the hybrid glycans. The maturation of complex-type vertebrate glycans can be the addition of terminal sialic acid and/or fucose as last step to stop the glycan biosynthesis\textsuperscript{15}.

For Cerezyme\textsuperscript{®}, the complex-type glycans have been synthesized first, and then removed later (remove sialic acid, galactose, and N-acetylglucosamine in the complex-type glycans). Thus, Man3GlcNAc2Fuc is the major remaining glycan for Cerezyme\textsuperscript{®}. For the RNAi-mediated glucocerebrosidase, the mannose core was achieved by silencing \textit{MGAT1}, the gene to encode GlcNAcT-I, before the formation of the complex-type glycans. Once GlcNAcT-I is inactivated, the glycan biosynthesis could likely stop at the Man5GlcNAc2 stage before the formation of the mannose core for complex-type glycans\textsuperscript{16}. It should be noted that the biosynthesis pathway is a general guide for glycan structure formation. Some portion of the molecules could be advanced through at any stage of the biosynthesis process (depending on growth media or culture conditions), leading to a distribution of glycan variants. Nevertheless, the major glycans of the molecule should provide the clue as to how the glycan transferase or glycosidase has been utilized in its expression system.
Figure 3.9 The glycopeptides assignment in T2 (8-39) SFGYSSVVCVCN(Asn 19)ATYCDSDPPTFPALGTFSR of RNAi-GCase: A. MS of glycopeptide (Man3GlcNAc2) 1501.3075, 3+; B. CID-MS 2 of 1501.97, 3+; C. glycopeptides assignment in T2, sum of mass spectrum 57.5-58.5 min.
Figure 3.10 The glycopeptides assignment in T5 (49-74) MELSMGPIQAN(Asn 59)HTGTGLLLTLQPEQK: A. MS of glycopeptide (Man5GlcNAc2) 1314.9626, 3+; B. CID-MS 2 of 1314.63, 3+; C. glycopeptides assignment in T5, sum of mass spectrum 50.5-53.1 min.
Figure 3.11 The glycopeptides assignment T11 (132-155) TYTYADTPDDFQLHN(Asn 146)FSLPEEDTK of RNAi-GCase: A. MS of glycopeptide (Man5GlcNAc2) 1355.2382, 3+; B. CID-MS 2 of 1355.90, 3+; C. glycopeptides assignment in T11, sum of mass spectrum 46.2-48.6 min.
Table 3.1 Glycan comparison between Cerezyme and RNAi-GCase

<table>
<thead>
<tr>
<th>Glycosites</th>
<th>Cerezyme</th>
<th>RNAi-GCase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Major</td>
<td>others</td>
</tr>
<tr>
<td>Asn 19</td>
<td>Man3GlcNAc2</td>
<td>Man3GlcNAc3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Man3GlcNAc2Fuc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Man4GlcNAc2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Man5GlcNAc2</td>
</tr>
<tr>
<td>Asn 59</td>
<td>Man3GlcNAc2Fuc</td>
<td>Man3GlcNAc3Fuc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phos1Man5GlcNAc2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phos1Man6GlcNAc2</td>
</tr>
<tr>
<td>Asn 146</td>
<td>Man3GlcNAc2Fuc</td>
<td>Man3GlcNAc3Fuc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Man3GlcNAc4Fuc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phos1Man5GlcNAc2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phos1Man6GlcNAc2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phos1Man5GlcNAc3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phos1Man6GlcNAc3</td>
</tr>
<tr>
<td>Asn 270</td>
<td>Man3GlcNAc2Fuc</td>
<td>Man3GlcNAc3Fuc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Man3GlcNAc4Fuc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phos1Man5GlcNAc2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phos1Man6GlcNAc2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phos1Man5GlcNAc3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phos1Man6GlcNAc3</td>
</tr>
</tbody>
</table>

3.4 Conclusion

Glycans in Cerezyme® and RNAi-mediated molecule were site-specific characterized and compared by LC-MS analysis. The major glycan for Cerezyme® is found to be Man3GlcNAc2 with and without core fucose. For the RNAi-mediated molecule, the major
glycan is identified as Man5GlcNAc2 for Asn59, Asn146, and Asn270, and as Man3GlcNAc2 for Asn19 site.

The glycosylation site, occupancy, and glycan distribution are determined by the LC-MS approach and used for comparison between the two products. As a proof of concept, the production of mannose core can be achieved by the RNAi approach, although there are slightly different core mannoses produced when compared to the glycosidase sequentially digested glycans for Cerezyme®. From an engineering point of view, the RNAi approach seems to be advantageous for manufacturing as compared to using expensive glycosidases after the production of the molecule. Nevertheless, the therapeutic benefits or side-effects (i.e. immunogenicity) are difficult to assess without conducting human clinic studies.

3.5 Acknowledgment
I would like to thank Alnylam Pharmaceuticals for providing samples. This work was supported in part by NIH grant GM 15847.

3.6 Experimental
Samples. Cerezyme® was purchased by Alnylam (Cambridge, MA, USA) and provided as 0.933 mg / mL × 100 μL in each vial and 25 vials total. The vials were stored at -80 °C before analysis. RNAi-mediated glucocerebrosidase was expressed by Alnylam and provided as gel bands.
Materials. Trypsin (sequencing grade) was purchased from Promega (Madison, WI). N-Glycosidase F (PNGase F, from *elizabethkingia meningoseptica*), guanidine hydrochloride, ammonium bicarbonate, dithiothreitol (DTT), and iodoacetamide (IAA) were obtained from Sigma-Aldrich (St. Louis, MO). LC-MS grade water was from J.T. Baker (Phillipsburg, NJ), and HPLC grade acetonitrile (ACN) and formic acid (FA, optima LC/MS) from Thermo Fisher Scientific (Fairlawn, NJ). The 10K centrifugal filter units were ordered from Millipore (Bedford, MA). NuPAGE® Novex 4-12% Bis-Tris sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) gels, NuPAGE MES SDS running buffer (20 ×), NuPAGE LDS sample buffer (4×), NuPAGE sample reducing agent, and SimplyBlue SafeStain were all from Invitrogen (Carlsbad, CA).

Enzymatic Digestion. 100 µL Cerezyme® (0.933 mg/mL) was buffer exchanged to 0.1 M ammonium bicarbonate using a 10K molecular weight centrifugal filter and then denatured with 6 M guanidine hydrochloride. Sequentially, the protein solution was reduced with 4 mM dithiothreitol (DTT) for 30 min at 37 °C and alkylated with 10 mM iodoacetamide (IAA) in the dark for 45 min at room temperature. The excess IAA was removed by a 10K molecular weight centrifugal filter. Finally, 2 µg trypsin was added to the protein solution for 4h at 37ºC, and then additional 2 µg for overnight at 37 °C. For removing glycans, half of the digest was added PNGase F (1 unit) for 4 h at 37 ºC. The reactions were stopped by addition of 5% formic acid.

For RNAi-mediated glucocerebrosidase in gel bands, in-gel enzymatic digestion was applied. Meanwhile, Cerezyme® was run on a SDS-PAGE gel and digested side by side. According to the darkness of gel bands for RNAi-mediated glucocerebrosidase, 2.5 µg (2.7 µL) Cerezyme®
was mixed with 2.5 µL NuPAGE LDS sample buffer (4×), 1 µL NuPAGE sample reducing agent (10×) and 3.8 µL DI water to achieve the final volume of 10 µL. The mixture was incubated at 70 °C for 10 min, and then loaded onto SDS-PAGE gel at 200 V for 45 min and stained with Coomassie blue. The gel band of Cerezyme® was excised and destained with RNAi-GCase side by side with 50% ACN and 50% water to remove most of color, and washed then 0.1 M ammonium bicarbonate for 15 min and ACN 15 min. After drying in a SpeedVac (Labconco, Centrivap Cold Trap), trypsin (12.5 ng/µL) was added to cover the dried gel pieces at 4 °C for 30 min. The remaining supernatant was removed, and 25 mM ammonium bicarbonate was added to cover gel pieces for overnight at 37 °C. Finally, the digests were extracted by ACN and concentrated to ~5 µL (if < 5 µL, buffer A was added). All 5 µL were injected to run LC-MS/MS. Later, according to the data of the RNAi-GCase, the injection amount of Cerezyme® was reduced to 1 µg according to the amount of protein loaded on the SDS-PAGE gel.

**LC-MS.** An Ultimate 3000 nano-LC pump (Dionex, Mountain View, CA) and a self-packed C18 column (Magic C18, 200 Å pore and 5 µm particle size, 75 µm i.d. × 15 cm) (Michrom Bioresources, Auburn, CA) was coupled online to an LTQ-Orbitrap-ETD XL mass spectrometer (Thermo Fisher Scientific, San Jose, CA) through a nanospray ion source (New Objective, Woburn, MA). Mobile phase A (0.1% formic acid in water) and mobile phase B (0.1% formic acid in acetonitrile) were used for the gradient consisting of (i) 20 min at 2% B for sample loading and 5 min for desalting at 0.3 µL / min; (ii) linear from 2 to 5% B for 2 min; (iii) linear from 5 to 35% B for 60 min; (iv) linear from 35 to 90% B for 3 min; and finally (v) isocratic at 90% B for 5 min. The flow rate of the column was maintained at 0.2 µL/ min, and the mass spectrometer started to record data after 5 min of the gradient. The
LTQ-Orbitrap-ETD XL mass spectrometer was operated initially in the data-dependent mode as follows: survey full-scan MS spectra (m/z 300-2000) were acquired in the Orbitrap with a mass resolution of 30,000 at m/z 400 (with an ion target value of $5 \times 10^5$ ions), followed by nine sequential MS2 scans using the LTQ.

**Data analysis.** The spectra generated in the CID-MS2 step were searched against spectra of theoretical fragmentations (b and y ions) of Cerezyme® sequence with a mass tolerance of $\leq 5$ ppm for the precursor ions and with enzyme specificity, using a Sequence probability score (> 95% confidence) as the filter. For peptides with miscleavages or a mass tolerance $> 5$ ppm (but less than 20 ppm) of the precursor ion, confirmation required manual inspection to match all highly abundant product ions. The sequence coverage generated by each enzymatic map was combined as a total coverage.

### 3.7 References


Chapter 4  Analysis of Disulfide Status in the Transmembrane Proteins

CD3 ε, δ, and γ of Human T-Cells
4.1 Abstract

CD3 εδ and CD3 εγ are non-covalently linked heterodimers. Each consists of an Ig-like extracellular domain, an intracellular immunoreceptor tyrosine-based activation motif (ITAM), and a transmembrane domain. CD3ε, CD3δ, and CD3 γ contain the common RxCxxCxE motif in the membrane proximal segment, in which the cysteine status is expected to change with the biological function.

In this study, the cysteines in the RxCxxCxE motif were characterized by LC-MS. The CD3s were purified from human T-cells and separated by SDS-PAGE before in-gel digestion. Various enzymatic strategies were tested, and a four-enzyme protocol, including trypsin, LysC, GluC, and PNGase F, was optimized to be the final strategy to detect the three CxxC motifs in one run by LC-MS. The disulfide-linked form (oxidized) was detected to be the major species in the three CD3s. Additionally, alkylated forms were identified from reducing and alkylating the sample.

4.2 Introduction

CD3 εδ and CD3 εγ, non-covalently linked heterodimers in the αβ T cell receptor (TCR)\textsuperscript{1-3}, recognize antigens bound to the major histocompatibility complex (MHC). Antigen binding to the receptors triggers downstream signaling via the immunoreceptor tyrosine-based activation motif (ITAM) in the CD3 subunits\textsuperscript{4-6}. It has been speculated that the redox state of conserved CxxC motifs in the membrane proximal region of CD3ε, γ, and δ could be regulated during T cell development or activation\textsuperscript{7}. Hence, the determination of cysteine status in CxxC motifs was deemed to be important.
There are challenges to determine a specific post translational modification in or proximal to the membrane region of a protein by liquid chromatography coupled with mass spectrometry (LC-MS) method. First, the recovery of this region is usually low after enzymatic digestion. Second, CxxC motifs are adjacent to the transmembrane regions, which increases the difficulty for enzyme accessibility.

In this study, a four-enzyme digestion, followed by a liquid chromatography coupled to mass spectrometry (LC-MS) workflow was established to detect the redox status of CxxC motifs in CD3 ε, γ, and δ. The challenges encountered in the characterization of this site-specific post translational modification are discussed and overcome.

4.3 Results and Discussion

The detection of the redox status in CxxC motifs was designed to be two steps: 1. Detection of the oxidized and reduced CxxC in mouse recombinant CD3 fragments; 2. Determination of the cysteine status of CxxC in human CD3 δ, γ, and ε.

4.3.1 CxxC motif in mouse recombinant CD3 delta-1 and delta fragments

The recombinant CD3 delta-1 consisted of the CxxC motif, transmembrane region, and intracellular region, while CD3 delta contained an uncleaved histidine tag, the extracellular region, and the CxxC motif, as shown in Figure 4.1 with their amino acid sequences. The aim of this experiment is to demonstrate that CxxC motifs can be detected by LC-MS, and also to optimize the detection conditions.
Based on the amino acid sequences, as shown in Figure 4.1, trypsin plus Glu C was applied to cleave CD3 delta-1 into peptides of appropriate sizes, and trypsin was used for CD3 delta. Because trysin cleaves at the carboxylic side of arginine and lysine, and GluC cleaves at the carboxylic side of glutamate, after digestion, the CxxC was in peptide MCQNCVE for delta-1 and MCQNCVELDSGTMAGV for CD3 delta, both of which were identified in LC-MS/MS analysis based on their mass and CID-MS2. The experimental monoisotopic mass of MCQNCVE is 824.2755 with a charge state 1+, which is 2 Da less than the theoretical mass of the reduced form, as shown in Figure 4.2A. It indicates the formation of a disulfide between the two Cys in CxxC. Figure 4.2B shows CID-MS2 of this peptide: fragment ions between the two Cys are not detected, since CID usually cannot break peptide bonds within the disulfide ring. In addition, peptide MCQNCVELDSGTMAGV in CD3 delta was also detected with the mass 2 Da less than the theoretical mass of reduced form, and the CID-MS2 further demonstrated the formation of a disulfide bond (data was not shown).
Figure 4.2 Characterization of peptide MCQNCVE in mouse recombinant CD3 delta-1: A. precursor mass of MCQNCVE; B. CID-MS2 of the precursor ion from A.

On the other hand, CD3 delta-1 was reduced and then alkylated by methylmethane thiosulfonate (MMT). The sample was also digested by trypsin plus GluC. The MMT alkylated CxxC peptide was detected from the mass and CID-MS2. The fragment ions within CxxC region are observed, since the disulfide bond is broken, as shown in Figure 4.3.
4.3.2 CxxC motifs in human CD3 ε, γ, and δ

CD3εγ and CD3εδ are non-convalent heterodimers in the TCR; each contains a CxxC motif. The CxxC motifs are proximal to N-termini of transmembrane regions. The amino acid sequences of CD3 ε, γ, and δ are shown in Figure 4.4. There is another disulfide bond in the extracellular region of each CD3, and a free cysteine in the transmembrane region of CD3 ε and δ. CD3 γ and δ are glycoproteins, which contain glycosite N(70)KS and N(17)TS⁸, respectively. In order to clarify if cysteines in CxxC are free, disulfide-linked, or both existing in human T-cells, CD3 δ, γ and ε were expressed in Jurkat cells, and then they were alkylated by N-ethylmaleimide (NEM) and immunoprecipitated by anti-CD3 antibody. The purified human CD3s were provided as gel bands by our collaborator.
Figure 4.4 Amino acid sequences of human CD3 ε, γ and δ

To detect the CxxC motifs in CD3s, the selection of enzymes is critical, since peptides in all the three CD3s need to be cleaved into appropriate sizes, especially the CxxC motif. The peptides cannot be too small; otherwise, they will not be well retained in the LC system. On the other hand, the peptides cannot be too large, since large peptides often have recovery issues with the in-gel digestion as well as the C18 column separation. Four enzymes were applied simultaneously to digest CD3s, including LysC, trypsin, GluC, and PNGase F, in which trypsin and GluC were required to cleave CxxC motifs into desired sizes, LysC was to enhance the function of trypsin, and PNGase F released glycans which might decrease the efficiencies of enzymes. After digestion, disulfide-linked CxxC-peptides were detected in the three CD3s, as shown in Figure 4.5 for the characterization of CxxC peptides. Other peptides detected in CD3 are listed in Table 4.1 with their theoretical m/z values, intensities, and retentions. The NEM alkylated species were not detected.
The intensities of CxxC peptides are almost 2 orders of magnitude lower than the high intensity peptides detected in CD3s, as shown in Table 4.1 due to two possible reasons. First, oxidized CxxC peptides were observed in CD3s; however, these peptides were not well retained on the LC column. Hence, their amounts were difficult to estimate. Second, GluC didn’t digest well in the sites that were adjacent to the membrane regions, since the peptide adjacent to the CxxC peptide in each CD3 had a high intensity; however, the peptide after it, which is in the membrane region, was not detected in CD3ε or δ. It was detected to be a low intensity in CD3γ.
Figure 4.5 Characterization of disulfide-linked CxxC peptides in human CD3s: A. MS and CID-MS2 of disulfide-linked VCENCME in CD3 ε; B. MS and CID-MS2 of disulfide-linked MCQNCIE; C. MS and CID-MS2 of disulfide-linked MCQSCVE.
Table 4.1 Human CD3 ε γ δ peptide mapping using four-enzymatic digestion

<table>
<thead>
<tr>
<th>#</th>
<th>Sequence</th>
<th>Theoretical monoisotopic mass, charge</th>
<th>Intensity</th>
<th>Retention (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>CD3 ε</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-15</td>
<td>MGGITQTPYK</td>
<td>548.2788, 2+</td>
<td>2.67E+07</td>
<td>31.16</td>
</tr>
<tr>
<td>16-34</td>
<td>VSISSGVTLTCQYPGSEQSGYYVVCYPR</td>
<td>1062.1704, 3+</td>
<td>2.06E+06</td>
<td>50.77</td>
</tr>
<tr>
<td>35-42</td>
<td>ILWQHNDK</td>
<td>527.2774, 2+</td>
<td>1.37E+07</td>
<td>30.41</td>
</tr>
<tr>
<td>43-51</td>
<td>NIGGDEDK</td>
<td>481.7067, 2+</td>
<td>2.43E+06</td>
<td>17.44</td>
</tr>
<tr>
<td>52-63</td>
<td>NIGSGDEHLSLK</td>
<td>664.3281, 2+</td>
<td>2.97E+07</td>
<td>33.53</td>
</tr>
<tr>
<td>80-93</td>
<td>GSKPEDANFYLRL</td>
<td></td>
<td>7.08E+06</td>
<td>48.24</td>
</tr>
<tr>
<td>85-93</td>
<td>DANYFYLRL</td>
<td>587.7982, 2+</td>
<td>1.77E+06</td>
<td>52.83</td>
</tr>
<tr>
<td>96-102</td>
<td>VCENCME</td>
<td>825.2576, 1+</td>
<td>5.37E+05</td>
<td>24.63</td>
</tr>
<tr>
<td>157-172</td>
<td>ERPPPPVNPDYIEPIRK</td>
<td>635.3390, 3+</td>
<td>6.61E+06</td>
<td>33.78</td>
</tr>
<tr>
<td>176-184</td>
<td>DLYSGLNQQR</td>
<td>533.2698, 2+</td>
<td>2.39E+07</td>
<td>35.75</td>
</tr>
<tr>
<td></td>
<td><strong>CD3 γ</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-10</td>
<td>GNHLVK</td>
<td>667.3886, 1+</td>
<td>6.50E+04</td>
<td>14.08</td>
</tr>
<tr>
<td>11-16</td>
<td>VYDYQE</td>
<td>816.3410, 1+</td>
<td>8.03E+04</td>
<td>28.97</td>
</tr>
<tr>
<td>17-29</td>
<td>DGSVLLTCDAEAKGMYQCK</td>
<td>1024.4603, 2+</td>
<td>4.06E+05</td>
<td>46.85</td>
</tr>
<tr>
<td>30-35</td>
<td>NITWFK</td>
<td>808.4352, 2+</td>
<td>1.07E+07</td>
<td>47.64</td>
</tr>
<tr>
<td>39-47</td>
<td>MIGFLTEDK</td>
<td>527.2679, 2+</td>
<td>1.30E+07</td>
<td>45.30</td>
</tr>
<tr>
<td>50-57</td>
<td>WNLGNSNAK</td>
<td>445.2300, 2+</td>
<td>9.34E+06</td>
<td>32.48</td>
</tr>
<tr>
<td>72-80</td>
<td>SKPLQVYYR</td>
<td>577.3218, 2+</td>
<td>5.79E+06</td>
<td>31.76</td>
</tr>
<tr>
<td>81-87</td>
<td>MCQNCIE</td>
<td>838.2892, 1+</td>
<td>1.82E+05</td>
<td>27.20</td>
</tr>
<tr>
<td>88-100</td>
<td>LNAATISGFILFAE</td>
<td>677.3561, 2+</td>
<td>7.21E+05</td>
<td>65.46</td>
</tr>
<tr>
<td>129-142</td>
<td>QTLLPNDQLYQPLK</td>
<td>835.9592, 2+</td>
<td>8.01E+06</td>
<td>47.10</td>
</tr>
<tr>
<td>143-158</td>
<td>DREDDQYSHLQGNQLQR</td>
<td>658.6402, 3+</td>
<td>7.20E+06</td>
<td>30.75</td>
</tr>
<tr>
<td></td>
<td><strong>CD3 δ</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-7</td>
<td>FKIPHE</td>
<td>438.2473, 2+</td>
<td>6.90E+06</td>
<td>44.36</td>
</tr>
<tr>
<td>12-24</td>
<td>VFVNGDTSTWVE-CNGTDIYK</td>
<td>808.6920, 3+</td>
<td>3.80E+06</td>
<td>49.37</td>
</tr>
<tr>
<td>25-36</td>
<td>GTVGTLSSDITR</td>
<td>616.8459, 2+</td>
<td>1.56E+07</td>
<td>54.19</td>
</tr>
<tr>
<td>63-70</td>
<td>STVQVHYR</td>
<td>481.7325, 2+</td>
<td>6.41E+06</td>
<td>25.34</td>
</tr>
<tr>
<td>71-77</td>
<td>MCOSCE</td>
<td>797.2627, 1+</td>
<td>9.25E+04</td>
<td>23.55</td>
</tr>
<tr>
<td>112-123</td>
<td>LSGAADQTOALLR</td>
<td>608.3382, 2+</td>
<td>3.37E+07</td>
<td>37.78</td>
</tr>
<tr>
<td>124-132</td>
<td>NDQVYQPLR</td>
<td>566.7909, 2+</td>
<td>1.89E+07</td>
<td>32.06</td>
</tr>
<tr>
<td>133-148</td>
<td>DRDDAQYSHLQGNWAR</td>
<td>930.9222, 2+</td>
<td>3.55E+05</td>
<td>36.09</td>
</tr>
<tr>
<td>135-148</td>
<td>DDAQYSHLQGNWAR</td>
<td>795.3582, 2+</td>
<td>2.66E+06</td>
<td>37.84</td>
</tr>
</tbody>
</table>
The NEM alkylated CxxC peptides can be detected if reduction and alkylation are performed before the four-enzymatic digestion, shown in Figure 4.6.

**Figure 4.6** Characterization of iodoacetamide (IAM) alkylated CxxC peptides in human CD3s: A. MS and CID-MS2 of alkylated VCENCME in CD3 ε; B. MS and CID-MS2 of alkylated MCQNCE; C. MS and CID-MS2 of alkylated MCQSCVE.
Meanwhile, a three-enzymatic digestion (LysC + GluC + PNGase F) was also tested; CxxC peptides were detected in CD3γ and δ, as SKPLQVYYRM_CNIE in CD3 γ and ESTVQVHYRMQSCVE in CD3 δ. The retention of CxxC-containing peptides was significantly improved as well as the retention of their oxidized forms. According to their intensities, assuming equal response factors, the oxidized form was 31% in CD3 γ and 42% in CD3 δ, demonstrating that oxidation is a one factor that can decrease the intensities of CxxC peptides. The CxxC-containing peptide in CD3 ε was not observed in the three-enzymatic digestion, because of the low efficiency of GluC.

Other enzymatic strategies were also tried, including two-enzymatic digestions (LysC + GluC and trypsin + GluC) and pepsin digestion. By the two-enzymatic digestions, the detected peptides were mainly at the N-termini and C-termini of the CD3s. Pepsin, which usually cleaves peptide bonds between hydrophobic and preferably aromatic amino acid residues, did not work on CD3s; only one peptide at the N-termini was detected for each CD3.

4.4 Conclusion

The workflow of sample preparation and LC-MS detection was built up to analyze the redox status of CxxC peptides in the membrane proximal region of CD3s. By applying this method, CxxC-peptides were detected to form intra-chain disulfides in the three CD3s. Multiple-enzymatic digestion is critical to determine a site-specific post translational modification in transmembrane proteins. The efficiency of enzymes plays an important role in the analysis peptides in membrane proteins.
With the successful detection, we plan to stimulate the cells under environmental factors, such as hypoxia, metabolic stress, antigen activation, Interleukin-2, etc, and investigate if the redox status of CxxC motifs changes, and how it will change in CD3s.

4.5 Acknowledgements

I would like to thank Dr. Ellis Reinherz’s group of Dana Farber Cancer Center for providing sample and experimental discussion. This work was supported in part by NIH grant GM 15847.

4.6 Materials and Methods

**Samples.** The mouse recombinant and human CD3 samples were provided by Ellis Reinherz’s group in Dana Farber Cancer Institute, Boston, MA. The mouse recombinant CD3 delta fragment was 0.794 mg/mL, and the mouse recombinant CD3 delta-1 fragment was 0.742 mg/mL. Both were dissolved in 30 mM Tris + 100 mM NaCl buffer. The human CD3 epsilon, gamma and delta were expressed in T-REx™ Jurkat Cell Line, alkylated by NEM, enriched by anti-CD3 antibody column, separated by SDS-PAGE, and cut according to the molecular weight and the Western blotting result. The human CD3 samples were provided as gel bands.

**Reagents.** Trypsin (sequencing grade) was purchased from Promega (Madison, WI), endoproteinase Lys-C (MS grade) from Wako Chemicals USA (Richmond, VA), pepsin (from Porcine Stomach Mucosa) from MP Biomedicals (Solon, OH), and endoproteinase Glu-
C (sequencing grade from Staphylococcus aureus V8) from Roche applied science (Indianapolis, IN). PNGase F (proteomics grade), Guanidine hydrochloride (Gn-HCl), ammonium bicarbonate, L-dithiothreitol (DTT), iodoacetamide (IAM), N-Ethylmaleimide (NEM), formic acid (FA, optima LC/MS) and hydrochloric acid (HCl) solution 1.0 M were obtained from Sigma-Aldrich (St. Louis, MO). LC-MS grade acetonitrile (ACN) and water was purchased from J. T Baker (Phillipsburg, NJ, USA).

Native in-solution digestion. Trypsin was added into the mouse recombinant CD3 fragment sample as 1:50 by weight, then the solution was incubated at 37 °C for 4 hours, and finally the reaction was stopped by adding 0.5% formic acid.

Native in-gel digestion. The gel bands were cut into 1 mm × 1 mm pieces, destained by 50% ACN and 50% water, and dried in a Speedvac (Labconco, Centrivap Cold Trap). The gel pieces were then re-hydrated into the enzyme buffer which was 12.5 ng / µL desired enzyme in 25 mM ammonium bicarbonate (if multiple enzymes were applied, each was 12.5 ng/ µL) at 4 °C for 30 min. Subsequently, the enzyme buffer was replaced by 25 mM ammonium bicarbonate, and incubated at 37 °C overnight. Finally, the digest was extracted by ACN, and concentrated to be almost dry.

4.7 References


Chapter 5 Summary and Future Studies
In this dissertation, protein structures characterized by LC-MS based methods have been presented. The recombinant therapeutic protein, Rituximab, was comprehensively characterized by LC-MS, including amino acid sequence, carbohydrate, disulfide linkage, N-terminal pyroglutamation, C-terminal processing, deamidation, oxidation, et al. In addition, an siRNA-treated anti-CD20 mAb, designed as a biobetter of Rituximab, was compared with Rituximab from the structural features listed above. The afucocylated glycans, below 10% in Rituximab, were increased to 70% in the siRNA-treated anti-CD20 mAb. However, two mutation sites were detected: one was fully mutated at S256Y; the other was ~40% mutated at F174I/L.

The comprehensive characterization of a protein structure, as well as the comparison of one protein with the other, requires using multiple analytical methods. In the biosimilar / biobetter analysis, multiple methods should be applied, among which LC-MS plays an important role to determine how similar / better one protein is to the other. In my study, I used the bottom-up method, which studies a protein from the peptide level, leading to the structure of the protein by the addition of all of the peptides together; some minor structures, however, may be missing. In the future, characterization of proteins from the intact level should be applied (i.e., top down). The molecular weight of a protein can be determined by the mass spectrometer; hence, amino acid variation in the candidate biosimilar / biobetter will be obvious. Currently, the fragmentation of intact proteins has been improved. The characterization and comparison of anti-CD20 mAbs can be studied from the top-down level. By combining the results from the peptide and intact levels, the protein structure can be more complete.
The comparison of a recombinant therapeutic protein Cerezyme with an siRNA-treated glucocerebrosidase mainly focused on the glycan. The glycans of Cerezyme were remodeled to be the glycan core by glycosidases in order to improve their uptake by macrophages via mannose receptors. siRNA-treated glucocerebrosidase was expressed with the same amino acid sequence as Cerezyme, but siRNA was added to silence the gene to encode GlcNAc-T1, the key enzyme to generate complex glycans; hence, glycan structure was ended at Man5GlcNAc2. In our study, glycans were analyzed by glycopeptides using LC-MS. The major glycan in Cerezyme was the core structure (Man3GlcNAc with and without core fucose), while the major glycan in siRNA-treated glucocerebrosidase, as was designed, was Man5GlcNA2. Other glycans in these two proteins were detected and listed.

In the future, there are two areas for further study. First, the glycans can be determined by releasing them from the protein. The glycans identified in this way can assist the assignment of glycosylation at each site. Second, siRNA-treated glucocerebrosidase was designed to be a biobetter of Cerezyme. Other structures should be further characterized and compared.

The characterization of human CD3 ε, γ, and δ focused on the redox status of cysteines in the conserved CxxC motif, the status of which was predicted to change in the signaling process. In our study, the CxxC motifs in CD3s were detected to be oxidized (disulfide-linked) in the normal state. Additionally, a sample preparation and LC-MS platform to characterize site-specific modifications in transmembrane proteins was established.

In the future, stimulation of cells under environmental conditions, such as hypoxia, metabolic stress, antigen activation, interleukin-2, etc, can evaluate whether the redox status of CxxC motifs changes and how such changes will affect in CD3s. Additionally, our well-established platform can be applied to characterize other membrane proteins.
Appendix

Evaluation of Exogenous siRNA Addition as a Metabolic Engineering Tool for Modifying Biopharmaceuticals

**Chen Li’s work**

The aim of this project was to increase the afucosylated glycans in anti-CD20 monoclonal antibody (mAb) by silence genes involved in the expressing of enzymes that catalyze the adding of core fucose to glycans. My work in this project was to monitor the increment of afucosylated glycans in siRNA-treated anti-CD20 mAb and compare its other structures with untreated anti-CD20 by LC-MS, including amino acid sequence, N-terminal and C-terminal modifications, disulfides, oxidation, deamidation, et al. The analytical strategies were introduced in Chapter 2, and the data were shown in this paper.
1 Abstract

Traditional metabolic engineering approaches, including homologous recombination, zinc finger nucleases, and short hairpin RNA (shRNA), have previously been employed to generate biologics with specific characteristics that improve efficacy, potency, and safety. An alternative approach is to exogenously add soluble small interfering RNA (siRNA) duplexes, formulated with a cationic lipid, directly to cells grown in shake flasks or bioreactors. This approach has the following potential advantages: no cell line development required, ability to tailor mRNA silencing by adjusting siRNA concentration, simultaneous silencing of multiple target genes, and potential temporal control of down regulation of target gene expression. In this study, we demonstrate proof of concept of the siRNA feeding approach as a metabolic engineering tool in the context of increasing monoclonal antibody afucosylation. First, potent siRNA duplexes targeting fut8 and gmds were dosed into shake flasks with cells that express an anti-CD20 monoclonal antibody. Dose response studies demonstrated the ability to titrate the silencing effect. Furthermore, siRNA addition resulted in no deleterious effects on cell growth, final protein titer, or specific productivity. In bioreactors, antibodies produced by cells following siRNA treatment exhibited improved functional characteristics compared to antibodies from untreated cells, including increased levels of afucosylation (63%), a 17-fold improvement in FcγRIIIa binding, and an increase in specific cell lysis by up to 30%, as determined in an ADCC assay. In addition, standard purification procedures effectively cleared the exogenously added siRNA and transfection agent. Moreover, no differences were observed when other key product quality structural attributes were compared to untreated controls. These results establish that exogenous addition of siRNA represents a potentially novel metabolic engineering tool to improve biopharmaceutical function and quality that can complement existing metabolic engineering methods.
2 Introduction

With an ever increasing number of biologics in pharmaceutical company pipelines, researchers continue to explore novel technologies to modify host cell lines to improve productivity, safety, efficacy, and potency of biologics. An important area of study for host cell modification is gene inactivation\textsuperscript{1,2}. Currently, gene inactivation tools such as homologous recombination\textsuperscript{3-6}, zinc-finger nucleases \textsuperscript{7-9}, and short hairpin RNA (shRNA) \textsuperscript{10-12} are utilized to alter host cell gene expression. These gene inactivation strategies can be effective; however, they cannot tailor the degree of gene silencing which can be important \textsuperscript{13}. Moreover, these gene inactivation approaches can significantly increase the bioprocess development time, as cell line engineering requires significant time and resources. The length of development time is further increased if several targets are to be simultaneously inactivated. In addition, non-specific effects can occur due to the somewhat random nature of genetic insertion within the host cell chromosome \textsuperscript{14,15}.

An alternative approach for metabolic engineering of host cells is to add synthetic small interfering siRNA (siRNA) in a cationic lipid formulation directly to the manufacturing cell line in the bioreactor to initiate RNA interference (RNAi) \textsuperscript{16}. This strategy could, in principle, allow for rapid, transient, silencing of target genes, as no cell line engineering/selection is required. Moreover, by choosing the siRNA concentration, titration of the level of gene silencing could be possible, in contrast to gene knockout strategies. Furthermore, combining siRNA duplexes to target multiple genes in several cellular pathways could enable simultaneous modulation of key effect(s) critical to cell growth, protein production, and product quality. Also, by feeding at critical time points, the siRNA approach could provide temporal control of gene expression, which is currently not available with existing metabolic engineering strategies. Finally, using
genomic and transcriptomic data currently available, all expressed genes could, in principle, be targeted. Thus, exogenous siRNA addition directly to a bioprocess has the potential to accelerate biologics development and to generate products with very specific product profile(s) for enhanced biological activity, quality, and safety with improved productivity.

To demonstrate the potential of the exogenous siRNA addition approach, the fut8 and gmds genes, well known components of the de novo fucosylation pathway, were targeted for down regulation using exogenously added siRNA. Fucosyltransferase (FUT8) and GDP-man-4,6-dehydratase (GMDS) are important enzymes responsible for core fucose Fc carbohydrate on therapeutic monoclonal antibodies. Removal of the core fucose on glycosylation sites on monoclonal antibodies is known to enhance activity by improving FcγRIIIa binding, leading to increased antibody dependent cellular cytotoxicity. For this study, potent siRNA duplexes targeting fut8 and gmds were dosed into shake flasks with cells that express an anti-CD20 monoclonal antibody. Using optimal conditions determined from shake flask studies, exogenous siRNA addition was applied to bioreactors. Antibody generated from siRNA treatment was then compared to untreated controls to determine the extent of afucosylation and modification of biological activity, as well as whether any other product or process related modifications occurred, due to siRNA treatment. Our data demonstrate that RNAi-mediated metabolic engineering can be a powerful new tool in the control of a bioreactor process.

3 Materials and Methods
**Cell line development and anti-CD20 monoclonal antibody expression.** Using previously published sequences \(^{21}\), the heavy and light chains of an anti-CD20 monoclonal antibody (MAb) were separately cloned into a eukaryotic expression vector containing the cytomegalovirus (CMV) intermediate early promoter (pGV90). Cloning was followed by a double transfection of both the heavy and light chain plasmids into dihydrofolate reductase (DHFR) deficient DG44 Chinese hamster ovary (CHO) host cells (Invitrogen, Carlsbad CA) using the FuGene 6 transfection reagent (Roche Applied Sciences, Indianapolis IN). An anti-CD20 MAb producing clonal cell line was identified by iterative screening of transfected cells for high specific productivity by flow-assisted cell sorting (FACS) using an anti-human IgG1 monoclonal antibody, as previously reported \(^{22}\).

**Cell culture.** For dose response studies, cultures were seeded at \(1.0 \times 10^6\) cells/mL in 50 mL of CD DG44 cell culture media (Invitrogen, Carlsbad, CA) in 125 mL shake flasks, supplemented with Pluronic F68 (Invitrogen, Carlsbad, CA) and Glutamax (Invitrogen, Carlsbad, CA) to final concentrations of 0.18% and 8 mM, respectively. Cultures were grown in Multitron II incubators (Appropriate Technical Resources, Laurel, MD) at 37ºC, 5% CO\(_2\), 80% relative humidity, and 130 RPM (pitch = 25 mm) for 3 days. Cultures from seed flasks were then split back to \(0.5 \times 10^6\) cells/mL in fresh media to initiate production stage cultures. Seed cultures were transfected on day 0, while production stage cultures were transfected every 3 days starting on day 0. All transfections used a proprietary transfection reagent\(^{23}\) and a proprietary dosing strategy. For both seed and production stage cultures, the siRNA concentration for FUT8 and GMD varied for each culture from 0 - 50 picomoles/\(10^6\) cells, based on the experimental condition. Production stage cultures were also fed 5% of both Efficient Feeds A and B.
(Invitrogen, Carlsbad, CA) on days 3 and 6 and 0.5% of 50x RPMI amino acids (Sigma-Aldrich, St. Louis, MO) daily from day 7 to harvest. 1M glucose and Glutamax were also supplemented after day 7 to maintain culture concentrations between 2 - 6 g/L and > 0.5 mM, respectively. Cell culture supernatants were harvested on day 12 by centrifugation at 200 x g for 30 minutes. Measurements of cell density and viability were performed almost daily using a Vicell XR according to manufacturer protocols (Beckman Coulter, Brea, CA). Metabolite analysis was performed using a Bioprofile Flex, according to manufacturer recommendations (Nova Biomedical, Waltham, MA). Retain samples for \textit{fut8} and \textit{gmd} mRNA analyses were also taken daily. For these samples, 50 µL of each culture was added to 450 µL of 1X PBS (Invitrogen, Carlsbad, CA). The mixture was then centrifuged at 300 x g for 10 min at room temperature. Supernatants were discarded, and pellets stored at -80°C until mRNA analysis was initiated.

For bioreactor cultures, approximately 300 mL of culture were seeded in a 1L shake flask at an initial cell concentration of 1.0 x 10^6 cells/mL. The culture was grown for 3 days using the same cell line, media, and incubator conditions as in the dose response study. For the siRNA treated seed condition, cultures were transfected with a proprietary transfection reagent using a proprietary dosing strategy with a siRNA concentration ranging from 4-10 picomoles/10^6 cells for both \textit{fut8} and \textit{gmd} siRNA. Prior to day 3 of seed cultures, bioreactors (either 3L working volume in glass bioreactors (Chemglass, Vineland, NJ; CLS-1406-03) or 1.5L working volume in Millipore (Billerica, MA) disposable bioreactors (CR0003L200)) were aseptically prepared with a sterilized pH probe, dissolved oxygen probe, and aseptically charged with 2.7 and 1.2L of CD DG44 cell culture media, respectively. Bioreactors were inoculated with the appropriately treated seed culture to achieve an initial cell concentration of 0.6 x 10^6 cells/mL and controlled
as follows: temperature: 37°C, pH allowed to decrease from 7.4 to 6.8 with 1M sodium carbonate, and dissolved oxygen: 50% of air saturation using O₂ sparging. All siRNA treated bioreactors were transfected every three days starting on Day 0 utilizing the same strategy as for the respective seed culture. The feed strategy and sample analysis for all bioreactors were identical to those utilized for shake flasks in the dose response study. For harvests, between 0.2 - 1L of culture supernatant was collected on day 12 by filtering through a Millistak+ POD COHC clarifying filter (Millipore, Billerica, MA), followed by a SHC Opticap XL150 PES 0.2µm sterilizing filter (Millipore, Billerica, MA), using manufacturer recommendations. Supernatants were stored at --20°C until needed.

**Identification and synthesis of CHO cell-specific FUT8 and GMDS siRNA duplexes**

Based on previously published CHO cell-specific transcript sequences for *fut8* and *gmds* (Genbank, Accession# AF523564.1), 20 siRNA duplexes were designed targeting each gene using proprietary siRNA design algorithms. FUT8 and GMDS single-stranded RNAs were generated by small scale synthesis on a MerMade oligo synthesizer (BioAutomation, Plano TX), purified, and then annealed according to previously established protocols. All 20 candidate siRNA duplexes for each gene were screened for highest target mRNA knockdown by quantitative PCR (qPCR) in CHO-S cells (Invitrogen, Carlsbad CA) using a single siRNA concentration of 10 nM. The initial screen was followed by further testing of the top 5 identified candidates using a range of 10-fold serial dilutions of siRNA ranging from 0.01 to 100 nM. The top candidate for each gene was then chosen based on the lowest siRNA concentration required to generate 50% target mRNA silencing compared to the cells only controls. For dose response and bioreactor cell culture experiments, synthesis of the top FUT8 and GMDS siRNA duplexes
were scaled-up to produce approximately 200 mg of each duplex using previously described methods.

**mRNA quantitation by qPCR.** Total RNA isolation from mRNA retain cell pellets and subsequent cDNA synthesis were performed using commercial kits: MagMax Total RNA Isolation Kit and ABI High capacity cDNA reverse transcription kit (Life Technologies, Carlsbad, CA), respectively, according to manufacturer protocols. Probes and primers were designed using CLC Main Workbench 4 for both genes as follows:

<table>
<thead>
<tr>
<th>Target</th>
<th>Forward Primer</th>
<th>Reverse Primer</th>
<th>Probe</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUT8</td>
<td>TGGGAGACTGTGTGTTAGA</td>
<td>AAGGGTAAGTAAGGAGGA</td>
<td>CAGGTGAAGTGAAGGACAAAAA</td>
</tr>
<tr>
<td>GMDS</td>
<td>GAG ACC GGC AAA ATT CAT GT</td>
<td>TTC TCA TGA GCT CCA CAT CG</td>
<td>ACC GAC CAA CTGAAG TGG AC</td>
</tr>
</tbody>
</table>

Target probes were synthesized with a 5’FAM and 3’ Black Hole Quencher 1 (BHQ1).

For real time PCR, 2 µL of cDNA were added to a master mix containing 0.5 µL 18s TaqMan Probe (Life Technologies, Carlsbad, CA), 0.5 µL of gene specific probe, and 5 µL Lightcycler 480 probe master mix (Roche Applied Science, Indianapolis, IN) per well in a 384 well plates (Roche Applied Science, Indianapolis, IN) and performed in a LC 480 Real Time PCR machine (Roche Applied Science, Indianapolis, IN). To calculate relative fold change, real time data were analyzed by calculating differences in cycle time between assays performed with test samples and their respective cell only controls. All samples were normalized to either 18S RNA or GAPDH to account for differences in cell numbers.

**Anti-CD20 MAb Purification.** Filtered shake flask or bioreactor cell culture supernatants were thawed at 37°C and supplemented with 1M sodium phosphate pH 7.4 to a final concentration of
20 mM. Anti-CD20 MAb from siRNA-treated cell or control cultures were purified using a HiTray MabSelect SuRe column (GE Healthcare, Piscataway, NJ) followed by HiTray SP Sepharose chromatography (GE Healthcare, Piscataway, NJ) to remove remaining process-related contaminants. Following the elution from SP Sepharose, the proteins were concentrated and diafiltered using phosphate-buffered saline (PBS; Invitrogen, Carlsbad, CA) with a 9 kDa Orbital Biosciences (Topsfield, MA) centrifugal concentrator. Protein quantitation was performed utilizing a Bradford assay kit (Bio-Rad Laboratories, Hercules, CA) with a previously purified and quantitated (ultraviolet absorbance at 280 nm) anti-CD20 MAb as a standard.

**Bioanalytical characterization.** Anti-CD20 MAb glycan analysis was performed by capillary electrophoresis-laser induced fluorescence (CE-LIF) and LC/MS/MS as previously described. Protein samples were reduced and alkylated prior to tryptic digestion or Lys-C digestion, Achromobacter protease I (Roche, Nutley, NJ; 1:50, w/w) was added to the protein in place of trypsin for 4 hr at 37°C. LC-MS analysis was performed using an Ultimate 3000 nano-LC pump (Dionex, Mountain View, CA) and a self-packed C18 column (Magic C8, 200 Å pore and 5 μm particle size, 75 μm i.d. × 15 cm) (Michrom Bioresources, Auburn, CA) coupled online to an LTQ-Orbitrap-ETD XL mass spectrometer (Thermo Fisher Scientific, San Jose, CA) through a nanospray ion source (New Objective, Woburn, MA), as previously described. Full experimental details including analytical data will be published on anti-CD20 MAb characterization in a separate paper (manuscript in preparation).
**FcγRIIIa binding assay.** Anti-CD20 MAb binding to FcγRIIIa was monitored using a previously described capture-type ELISA. Briefly, NUNC 96-well flat-bottom Immuno MaxiSorp plates (Thermo Fisher Scientific, Fair Lawn, NJ) were coated with an anti-human IgG1 monoclonal antibody (10 µg/mL; R&D systems, Minneapolis, MN) and incubated at 37°C for 1 hour. Three consecutive washes were performed following each capture or blocking step with TBST (200 µL/well; 25mM Tris, 0.15 M NaCl, 0.05% Tween 20, pH 7.0). Plates were then blocked with 5% non-fat milk in TBST (200 µL/well), incubated for ½ hour at 37°C, and washed. Afterwards, a fixed amount (100 µL/well) of anti-CD20 monoclonal antibody (1 µg/mL) from each experimental condition was added to each well and incubated for 1 hour at 37°C. After washing the plate, human His-tagged FcγRIIIa (R&D Systems, Minneapolis, MN) was serially diluted to 5 µg/mL, 1.5 µg/mL, 0.5 µg/mL, 0.15 µg/mL, 0.05 µg/mL, 0.015 µg/mL, and 0.005 µg/mL from stock (500 µg/mL). 100 µL of each dilution of FcγRIIIa was added into each well containing the different monoclonal antibody preparations and incubated at room temperature for 1 hour. The plates were washed again with TBST. His-tagged FcγRIIIa was detected using an anti-His HRP-labeled antibody (R&D Systems, Minneapolis, MN) diluted 1/5000 (100 µL/well) and incubated at room temperature for 1 hour. The plates were washed a final three times. TMB substrate (Thermo Fisher Scientific, Fair Lawn, NJ) was added to each well (100 µL/well) and allowed to develop for less than 10 minutes at room temperature. The reaction was stopped by adding 100 µL/well stop buffer (1M H2SO4) and the optical density at 450 nm (OD450) measured using a Victor 4 (Perkin Elmer, Waltham, MA).

**ADCC assay.** The ADCC assay was contracted out to Eureka Therapeutics (Emeryville, CA). Briefly, Jeko-1 cells (ATCC, Cat# CRL-3006,) were used as the assay target cell and maintained
at 37°C and 5% CO₂ in F-12K media containing 10% fetal bovine serum. Fresh human normal PBMCs were purchased from AllCells (Emeryville, CA). Approximately 1 x 10⁵ Jeko-1 cells were pre-incubated with RNAi-treated or control antibodies at varying concentrations ranging from 0.01 to 10 µg/mL in 96-well flat bottom plates at 37°C with 5% CO₂ for 30 minutes. Following incubation, freshly isolated human PBMCs (2.5x10⁶ cells) were added (50 µL) into each well at an effector/target ratio of 25:1. The plates were incubated for 16 hr at 37°C with 5% CO₂. Next, the plates were centrifuged at approximately 300 x g for 15 minutes to pellet cells and debris. 50 µL of supernatant was transferred from each well to a 96-well round bottom assay plate. The amount of specific cell lysis was monitored by lactate dehydrogenase (LDH) release using a CytoTox 96® Non-Radio Cytotoxicity LDH Assay (Promega, Madison, WI) performed according to the manufacturer’s recommendation. Specific cell lysis was calculated as follows:

\[
\text{Sample Cell Lysis (\%) = \left( \frac{\text{Sample Reading} - \text{Spontaneous Release}}{\text{Maximal Release} - \text{Spontaneous Release}} \right) \times 100.}
\]
4 Results

No significant effects on growth/productivity due to siRNA addition. Based on a proprietary dosing strategy, CHO cells, expressing an anti-CD20 monoclonal antibody in shake flask cultures, were transfected directly with increasing concentrations of FUT8 and GMDS siRNA duplexes up to 50 picomoles/10^6 cells. Growth profiles, final titer, and final specific productivities (i.e., Final titer / Final IVCC) are summarized in Figure 1. After normalization of cell concentration to the initial culture volume in order to account for differences in feed and transfection volume (i.e., cell concentration x culture volume ÷ initial volume), growth profiles for all conditions were similar with maximum cell concentration occurring on day 9. Average maximum cell concentrations for all cultures were also consistently between 18 - 19 x 10^6 cells/mL. Thus, time course profiles of viable cell concentration for the range of siRNA concentrations tested indicated no deleterious effects on growth due to increasing siRNA concentration. Final titer and specific productivities varied from 84 - 100 µg/mL and 0.94 – 1.11 picograms/cell days, respectively. T-tests of final titer and specific productivity, comparing each dose group to control, were not significant (p > 0.05), which suggests no significant effect of siRNA concentration on antibody titer or specific productivity. These results demonstrate that siRNA transfection into cells grown in shake flasks does not affect cell growth characteristics or protein productivity.
Figure 1 siRNA treatment does not affect growth/productivity. A) Time course profiles of viable cell concentration for the range of siRNA concentrations. All cell concentrations were normalized based on initial working volume to adjust for differences in feed and transfection volumes between conditions. Error bars represent ± SEM (n=9 for control and 6 for other conditions). B) No significant effect of siRNA concentration on antibody production and specific productivity. Specific productivity was determined by dividing final titer by final integral viable cell concentration. To generate the final integral viable cell concentration, an integral analysis using a standard trapezoidal rule was performed on time course profiles of normalized cell concentrations. Error bars represent ± SEM (n=6 for control and 3 for other conditions).

mRNA silencing is dose dependent and titratable. The effect of FUT8 and GMDS siRNA on silencing of fut8 and gmds expression was evaluated using qPCR on daily samples from shake flask cultures with increasing concentrations of siRNA. Time course profiles of both fut8 and gmds expression are outlined in Figures 2A and 2B, respectively. For both genes, initial expression levels for siRNA treated cultures were lower than for the cells only control condition due to the transfection of their respective seed cultures. For all siRNA concentrations tested, fut8 expression, in general, decreased until day 2, increased on day 3, and then gradually decreased...
throughout the rest of the culture. Similarly, *gmds* expression decreased on day 1, increased until day 3 and then maintained this level until day 9. After day 9, *gmds* expression decreased and remained at low levels through day 12. These variations in expression profiles between *fut8* and *gmds* are most likely due to differences in target message levels and turnover rates.

Despite the variability in qPCR data, a decrease in gene expression with increasing siRNA concentration can be observed throughout the cultures for both *fut8* and *gmds* expression. To further elucidate the potential relationship between siRNA concentration and gene silencing, an integral analysis was performed for both *fut8* and *gmds* mRNA expression profiles to determine the cumulative mRNA remaining for the various siRNA concentrations (Figure 2C). The results indicate that silencing is dose dependent between 0 - 10 picomoles/10^6 cells for both siRNA with saturation (~16% and ~25% of cells only for *fut8* and *gmds* expression, respectively) occurring after siRNA concentrations greater than 10 picomoles/10^6 cells. This result suggests that the desired level of mRNA knockdown can be achieved by choosing siRNA concentration appropriately and that the optimal siRNA concentration for silencing both *fut8* and *gmds* expression is between 5-10 picomoles/10^6 cells.
Figure 2 Fut8 and gmds silencing is dose dependent. A) Fut8 mRNA time course profile shows a decrease in fut8 expression with increasing siRNA concentration. mRNA concentrations were determined via quantitative PCR and normalized based on respective daily cells only mRNA concentration average. Error bars represent ± SEM (n=9 for control and 6 for other conditions). B) Gmds mRNA time course profile shows a decrease in gmds expression with increasing siRNA concentration. mRNA concentrations were determined via quantitative PCR and normalized based on daily cells only mRNA concentration average. Error bars represent ± SEM (n=9 for control and 6 for other conditions). C) Cumulative mRNA remaining over the entire time course of the culture for both fut8 and gmds demonstrate desired level of mRNA knockdown can be achieved by choosing siRNA concentration appropriately. To calculate cumulative mRNA remaining, an integral analysis using a standard trapezoidal rule was performed on time course mRNA profiles that were normalized based on non-treated cells only daily mRNA concentration averages. Error bars represent ± SEM (n=9 for control and 6 for other conditions).
**fut8 and gmds mRNA expression is reduced in bioreactors.** To demonstrate the benefit of exogenous siRNA addition in bioreactors, CHO cells expressing anti-CD20 monoclonal antibody were grown in bench scale bioreactors and dosed based on conditions identified in shake flask cultures. Similar to shake flask cultures, siRNA addition and transfection do not affect growth, as shown by no significant difference in integral viable cell concentrations for both siRNA treated and control cultures throughout culture duration (Figure 3A).

Furthermore, titer is also not affected [siRNA treated (n=3) – 157 ± 4 mg/L and untreated controls (n=3) - 161 ± 1 mg/L]. Next, fut8 and gmds mRNA expression were determined for each day of culture via qPCR (Figure 3B). The results demonstrate significant silencing of both genes throughout the bioreactor runs, with fut8 and gmds expression below 20 and 30%, respectively, of untreated cells only control levels for most of culture duration. Interestingly, the expression patterns for both genes are very similar until day 6. The increase in expression on days 3 and 6 is most likely due to the inability of a transfection frequency of 3 days to maintain silencing with the substantial cell growth occurring at this stage. Thus, there may be an opportunity for further improvement of silencing by increasing transfection frequency during this culture stage. After day 6, fut8 expression continued to decrease with time to below 10% of control levels, while gmds expression remained relatively constant between 20-30% for the remainder of culture duration.

For comparison to expression patterns in shake flask cultures, cumulative mRNA knockdown was calculated for both genes in the bioreactor experiments (Figure 3C). Cumulative mRNA
remaining for *fut8* and *gmds* were 14 and 23%, respectively, of cell only controls which is very similar to averages for *fut8* and *gmds* cumulative mRNA remaining in shake flasks (16 and 25%, respectively, of cell only controls). This result demonstrates reproducibility in silencing between shake flask and bioreactor conditions.

Figure 3  siRNA dosed in a bioreactor results in drastic *fut8* and *gmds* mRNA knockdown with minimal effects on growth. A) Integral viable cell concentration profiles for control and transfected bioreactors show minimal effects of exogenously added siRNA on growth. Cell concentrations were normalized based on initial working volume to adjust for differences in feed and transfection volumes between conditions. To calculate integral viable cell concentration, an integral analysis using a standard trapezoidal rule was performed on time course cell concentration profiles. B) *Fut8* and *gmds* mRNA time course profile show significant *fut8* and *gmds* silencing in bioreactors. mRNA concentrations were determined via quantitative PCR and normalized based on daily cells only mRNA concentration average. Error bars represent ± SEM (n=3). C) Cumulative mRNA remaining over the entire time course of the culture demonstrates significant *fut8* and *gmds* silencing occurs in bioreactors. mRNA concentrations were determined via quantitative PCR and normalized based on daily cells only mRNA concentration average. To calculate cumulative mRNA knockdown, an integral analysis using a standard trapezoidal rule was performed on time course mRNA profiles. Error bars represent ± SEM (n=3).
FUT8 and GMDS siRNA specifically increase levels of core afucosylation without affecting other product quality attributes. Antibody was purified from harvests of siRNA treated and control bioreactor runs and examined for glycosylation profile as shown in Figure 4A. Addition of FUT8 and GMDS siRNA dramatically increased total levels of core afucosylation of antibody to 63% compared to 0% for purified material from control bioreactor harvests. For the agalactosylated form (G0), 59% of purified antibody from siRNA treated cultures was found without core fucose, compared to 88% with core fucose for control cultures. For total G1 forms, there was no afucosylated G1 form identified with control antibody, compared to 4% in siRNA treated material. For the completely galactosylated form (G2), roughly 50% of siRNA treated material was afucosylated. These data demonstrate that core afucosylation was dramatically enhanced with FUT8 and GMDS siRNA addition.
Figure 4 siRNA treatment only modifies afucosylation of monoclonal antibody and does not alter other product quality attributes. A) Summary of glycosylation profile for both siRNA treated and control bioreactors. Error bars represent ± SEM (n=3). B) Summary of N and C terminal modifications. Values represent average of six measurements from the same sample ± SEM. C) Antibody oxidation summary. Values represent average of six measurements from the same sample ± SEM.. D) Antibody deamidation summary. Values represent average of six measurements from the same sample ± SEM.

To determine if there were any other effects on product quality due to siRNA addition, product quality attributes including N-terminal and C-terminal modifications, oxidation, deamidation, and disulfide bond formation were examined. A summary of all the tested product quality attributes is shown in Figures 4B-D. Importantly, no significant differences were observed between purified material from siRNA treated and untreated control bioreactors for modifications to the N-terminal pyroglutamate, C-terminal lysine, and deamidation of asparagines. In addition, disulfide linkages were confirmed to occur at the appropriate sites and to be identical between conditions (data not shown). A slight increase in oxidation of methionine 34 and methionine 81 was observed from the antibody produced from siRNA treated cultures; however, this result is most likely due to sample handling as the buffer/conditions of these samples were not optimized for long term storage and manipulation. Furthermore, FUT8 and GMD siRNA and transfection lipid concentrations were determined for samples from various stages of the antibody purification procedure to demonstrate efficient removal of potential process-related contaminants. No significant levels above the background levels observed in control cultures were detected for any of these components after SP Sepharose purification (Figure 5). These data strongly suggest that siRNA addition should have no effect on a biologics manufacturing process, other than the intended target(s) silencing effect(s).
Figure 5  Lipid and siRNA added to the bioreactor are efficiently removed from the bioprocess using standard purification procedures.  A) siRNA concentrations at different stages of purification.  B) Lipid concentrations at different stages of purification (n=2); Error bars represent ± SEM.

siRNA treated bioreactors produce antibody with improved functional characteristics. A predominant effect of core afucosylation on an antibody is to improve binding with FcγRIIIa which then promotes increased ADCC. Thus, purified antibody from both siRNA treated and control bioreactor runs were further subjected to FcγRIIIa binding and ADCC assays to determine if there indeed was improved functionality with antibodies produced from siRNA treated bioreactors. As shown in Figure 6A, antibody purified from siRNA treated bioreactor runs exhibited significantly higher binding to FcγRIIIa compared to antibody produced from control bioreactors. Peak OD_{450} for antibody produced from siRNA treated runs were roughly 17-fold higher than material from control runs. Moreover, significantly less FcγRIIIa was
required (~500-fold) for material generated from siRNA treated cultures to produce the same peak binding activity generated from material derived from control cultures.

Figure 6 Purified antibody from siRNA-treated bioreactors have increased FcyRIIIa receptor binding and enhanced ADCC. A) FcyRIIIa receptor binding profiles exhibit improved binding characteristics with purified antibody from siRNA treated bioreactors than material from non-siRNA treated bioreactors. Average values of 3 independent assays for 3 bioreactor runs from each condition are shown. Error bars represent ± SEM (n=3). B) Improved ADCC kinetics observed with purified antibody generated from siRNA treated bioreactors than non-siRNA treated bioreactors. Error bars represent ± SEM (n=3).

The ADCC assay also demonstrated improved functionality with antibody generated from siRNA treated bioreactors (Figure 6B). Antibody from siRNA treated bioreactors exhibited roughly 10-30% higher specific cell lysis at any given antibody concentration. Additionally, peak specific cell lysis was also higher (66%) relative to antibody from control bioreactor runs (53%). Furthermore, to achieve a specific cell lysis of 50%, approximately 25-fold less antibody produced from siRNA treated cultures was required compared to control. In conjunction with FcγRIIIa binding assay data, these results demonstrate that addition of FUT8 and GMDS siRNA directly to bioreactor runs can significantly improve ADCC of a biologic
without requiring any cell line engineering and thus, provides proof of concept that exogenous siRNA addition can modify a biopharmaceutical for enhanced functionality.

5 Discussion
This study demonstrates that potent siRNA duplexes formulated with a lipid carrier and dosed into shake flasks or bioreactors with cells can produce a modified biopharmaceutical with improved biological function without the need for cell line engineering. Furthermore, exogenous siRNA addition resulted in no deleterious effects on cell growth, final protein titer, or specific productivity. In addition, dose response studies demonstrated the ability to titrate the silencing effect, which is unique to this metabolic engineering approach. Concerns regarding process clearance of the siRNA and transfection agent should be alleviated as standard purification procedures effectively cleared the exogenously added siRNA and transfection agent to levels below the lower limit of detection. Moreover, no differences were observed when other key product quality structural attributes were compared to untreated controls. These data clearly demonstrate that RNAi-mediated metabolic engineering via exogenous siRNA feeding is a feasible approach for modifying biopharmaceuticals.

The exogenous addition of FUT8 and GMDS siRNA duplexes directly to anti-CD20 MAb expressing CHO cells specifically and potently decreased fut8 and gmds mRNA to produce an anti-CD20 monoclonal antibody with decreased Fc carbohydrate core fucose, enhanced FcγRIIIa binding, and improved ADCC activity. These results compare well with initial studies expressing shRNA in CHO cells against fut8 which exhibited similar amounts of afucosylated
antibody (60%). However, more recent studies with shRNA \(^{17}\) and other metabolic strategies including homologous recombination \(^6\), zinc-finger nucleases \(^7-9\), and expression of fucose conversion enzymes \(^{32}\) yield complete antibody afucosylation. Nonetheless, complete afucosylation may not be required to maximize immune response, as suggested in a recent study \(^{33}\). These results demonstrate the potential of adding siRNA duplexes targeting fucose to existing biopharmaceutical manufacturing processes such as the cetuximab process \(^{34}\) to decrease Fc carbohydrate fucose and produce therapeutic antibodies with maximum ADCC activity in a relatively short time.

While RNAi-mediated metabolic engineering via exogenous addition of siRNA is novel, there are several areas in which the technology can be further improved. Since the largest scale used in this study was 3L (40 L bioreactor proof of concept target knockdown completed against lactate dehydrogenase; data not shown), a more thorough understanding of cost and scalability of transfection is important for implementation in large-scale manufacturing processes. Current cost estimates for an entire run suggest that the cost of materials is \(~\$5/L\) of initial working volume, which is on par with the cost of off the shelf media additions such as insulin that are readily available from commercial vendors. However, this cost needs to be further reduced to make the siRNA feeding approach commercially feasible. Scalability considerations for robust operation in commercial processes using this technology include: (1) siRNA and transfection reagent storage optimization to enhance reagent stability, (2) further understanding of compatibility and extractable/leachable profiles of siRNA and transfection reagent with process tubing/piping, filters, and containers (steel and disposable) typically used at commercial scale, and (3) mixing/shear effects on siRNA and transfection reagent individually, as well as on the...
siRNA/transfection reagent feed during bioreactor addition. Formulation optimization of the current transfection reagent and/or development of new transfection reagents, as well as chemical modifications to siRNA, may also lead to increased silencing stability, decreased dose levels, and improved silencing potency, which will further decrease costs and enhance process robustness. Furthermore, efforts for simplifying the entire transfection dose into a lyophilized form could also be useful and may provide a long-term storage alternative that leads to reproducible transfections over many runs. All of these modifications will better align this technology with large-scale manufacturing.

The ability to specifically target and down regulate multiple cellular pathways simultaneously by the addition of siRNA directly to cells growing in shake flasks or bioreactors has significant promise for the future of biologics development and manufacturing. This technology may enable pre-clinical efforts to rapidly identify cellular modifications that improve biologic structural characteristics that enhance biological activity without the time-consuming step of generating numerous cell lines. Titratable down-regulation of potential genetic bottlenecks (e.g. secretory pathway genes) may significantly increase expression levels for enzyme replacement therapies (e.g., α-anti trypsin) that are currently generated from alternative methods (e.g. purification from blood products) and lead to commercially viable bioprocesses. The use of siRNA duplexes to down-regulate genes responsible for immunogenic epitopes [e.g., Galα1-3Galβ1-4GlcNAc-R (α-Gal) linkage and N-glycolyneuraminic acid (Neu5Gc; NGNA) sialic acid] observed on marketed products such as Orencia and Erbitux could provide a means to improve biologic product quality without re-engineering of the manufacturing cell line. Moreover, undesired host cell proteins that reduce product purity or result in unnecessary product modifications (e.g.,
endogenous proteases) can be targeted to further improve product quality and process consistency. In summary, targeting critical cellular pathway genes with soluble siRNA duplexes added directly to bioreactors has the potential to give highly specific control over cell growth, productivity, and product quality that is currently unavailable with existing technologies.

6 Acknowledgements
The authors thank NIH for support in part of this study under NIH GM 15847 (BK). The authors also thank Nessan Bermingham and Dan Shufrin for their valuable discussions. Contribution number 1012 from the Barnett Institute.

7 References


This is a License Agreement between Chen Li ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number: 3095180014109
License date: Feb 24, 2013
Licensed content publisher: John Wiley and Sons
Licensed content publication: Biotechnology Progress
Licensed content title: Evaluation of exogenous siRNA addition as a metabolic engineering tool for modifying biopharmaceuticals
Licensed copyright line: Copyright © 2013 American Institute of Chemical Engineers (AIChE)
Licensed content author: Seshu Tummala, Michael Titus, Lee Wilson, Chunhua Wang, Carlo Ciatto, Greg Thill, Donald Foster, Chen Li, Zoltan Szabo, Andras Guttman, Brian Bettencourt, Muthuswamy Jayaraman, Jack Deroot, David Kocisko, Stuart Pollard, Klaus Charisse, Satya Kuchimanchi, Greg Hinkle, Stuart Milstein, Rachel Meyers, Shiaw-Lin Wu, Barry L. Karger, Anthony Rossomando
Licensed content date: Feb 13, 2013
Start page: n/a
End page: n/a
Type of use: Dissertation/Thesis
Requestor type: Author of this Wiley article
Format: Print and electronic
Portion: Full article