Mass Spectrometric Analysis of Biopharmaceutical and Follow-on Protein Drugs

A dissertation presented

by

Haitao Jiang

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 2010
Mass Spectrometric Analysis of Biopharmaceutical and Follow-on Protein Drugs

by

Haitao Jiang

ABSTRACT OF DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemistry in the Graduate School of Arts and Sciences of Northeastern University, June 2010
ABSTRACT

Protein drugs in biotechnology manufacturing are often complex because of the complexity of the manufacturing procedure and the chemical complexity of protein. During the past several years, there has been increasing interest in the development of follow-on protein drugs in light of advances in manufacturing technology, process control, and protein characterization. Biopharmaceutical and follow-on protein analysis is playing a critical rule in aiding the regulation of generic protein drugs.

This dissertation focuses on the characterization and comparison of different recombinant therapeutic proteins by the application of liquid chromatography coupled online with tandem mass spectrometry (LC-MS) technology. Method development for the characterization of disulfide-linked peptides in therapeutic proteins is also presented.

Chapter 1 reviews the development of biopharmaceutical and follow-on protein drugs and the technologies being used in the characterization of recombinant therapeutic proteins which include the LC-MS analysis and electron transfer dissociation (ETD)/collision induced dissociation (CID) methodology.

Chapter 2 described a detailed characterization of recombinant human growth hormone that included the identification of the entire sequence with disulfide linkages as well as subtle modifications by a sensitive LC-MS approach using the accurate peptide mass (Fourier Transform Ion Cyclotron Resonance(FTICR) MS) and sequence assignment (MS/MS measurement). The extent of oxidation, deamidation, and chain
cleavages were measured by the ratio of peak areas of the nonmodified peptide vs. the sum of peak area of the nonmodified and modified peptides in the same LC-MS analysis. The subtle but distinct differences were found in the recombinant human growth hormone from the three manufacturers (the follow-on, counterfeit, and the original innovator products).

In chapter 3, TNK-Tissue Plasminogen Activator (TNK-tPA) samples from the innovator and follow-on manufacturers were characterized and compared. All tryptic peptides including N-terminal, C-terminal and mutated peptides as well as the disulfide-linked peptides were identified, with the demonstration of the same primary sequence and disulfide linkages between the innovator and follow-on products. The three N-linked and one O-linked fucose glycosylation sites were identified. The two N-linked (N103 and N448) and one O-linked fucose (T61) sites were fully glycosylated in both innovator and follow-on products. The other N-linked site (N184) was partially glycosylated and was shown to have a ~2.5x difference between the innovator (60% occupancy) and follow-on (25% occupancy) products. The cleavage site for the conversion of the zymogen form to an active enzyme was identified as being between R275 and I276, with a cleavage of 40% for the innovator and 10% for the follow-on products.

Chapter 4 developed an online LC-MS strategy combining collision-induced dissociation (CID-MS2), electron-transfer dissociation (ETD-MS2), and CID of an isolated product ion derived from ETD (MS3) that has been used to characterize disulfide-linked peptides. Disulfide-linked peptide ions were identified by CID and ETD
fragmentation, and the disulfide-dissociated (or partially dissociated) peptide ions were characterized in the subsequent MS3 step. The online LC-MS approach is successfully demonstrated in the characterization of disulfide linkages of recombinant human growth hormone (Nutropin), a therapeutic monoclonal antibody, and tissue plasminogen activator (Activase). The characterization of disulfide-dissociated or partially dissociated peptide ions in the MS3 step is important to assign the disulfide linkages, particularly, for intertwined disulfide bridges and the unexpected disulfide scrambling of tissue plasminogen activator. The disulfide-dissociated peptide ions are shown to be obtained either directly from the ETD fragmentation of the precursors (disulfide-linked peptide ions) or indirectly from the charge-reduced species in the ETD fragmentation of the precursors. The simultaneous observation of disulfide-linked and disulfide-dissociated peptide ions with high abundance not only provided facile interpretation with high confidence but also simplified the conventional approach for determination of disulfide linkages, which often requires two separate experiments (with and without chemical reduction). The online LC-MS with ETD methodology represents a powerful approach to aid in the characterization of the correct folding of therapeutic proteins.

Chapter 5 described the identification of the unpaired cysteine status and mapping of the 17 disulfides of recombinant tissue plasminogen activator (rt-PA) using LC-MS with ETD/CID. The analysis was conducted using a multifragmentation approach consisting of ETD and CID, in combination with a multienzyme digestion strategy (Lys-C, trypsin, and Glu-C). The disulfide linked peptides, even those containing N- or O-linked
glycosylation, could be assigned since the disulfide bonds were still preferably cleaved over the glycosidic cleavages under ETD fragmentation. The use of a multiple and sequential enzymatic digestion strategy was important in producing fragment sizes suitable for analysis. For the analysis of complex intertwined disulfides, the use of CID-MS3 to target partially disulfide-dissociated peptides from the ETD fragmentation was necessary for linkage assignment. The ability to identify the exact location and status of the unpaired cysteine (free or blocked with a glutathione or cysteine) could shed light on the activation of rt-PA, upon stimulation by either oxidative or ischemic stress.
ACKNOWLEDGEMENTS

I am happy and honored to have finally completed this dissertation. To all the people who gave me collective help, support and encouragement, I want to express my sincere gratitude.

My first and foremost acknowledgment goes to my advisor, Professor Bill Hancock for accepting me as his student and fostering my strong interest in protein drug analysis. All the projects I have finished with him are critical for understanding of the development of protein therapeutics and they are extremely useful for my future professional career. Professor Hancock has given me invaluable scientific insights and inspirations. Without his sound advice and guidance, I would not have come this far. I will be forever grateful to him for his mentoring throughout my doctoral studies.

I sincerely thank my supervisor, Dr. Shiaw-lin Wu, for his great help and thoughtful guidance. Not only supporting and mentoring me for the research, he also shared his life experiences to encourage me in pursuing a promising career.

To my dissertation committee members: Professor Paul Vouros, Professor Zhaohui Zhou and Professor David Forsyth, I deeply thank them for all the work they did for me to make this defense successful.

I would like to thank Professor Barry Karger for proving an excellent research environment at the Barnett Institute. I would like to acknowledge all the members of the
the Barnett Institute with whom I had opportunities to work with. In Hancock’s lab, I appreciate the help and support of Dr. Marina Hincapie, Dr. Xiaoyang Zheng, Qiaozhen Lu, Dr. Yonghui Wang, Majlinda Kullolli, Zhi Zeng, Xiaomei He, Yi Wang, Suli Liu, Yue Zhang and Fangfei Yan. In Karger’s lab, Dr. Tomas Rejtar, Dr. Shujia Dai, Dr. Quanzhou Luo, Dr. Guihua Yue, Dr. Jianmin Ren, Dr. Ye Gu, Dongdong Wang and Zhenke Liu have been very helpful. Also, I thank Dr. Jim Glick from Vouros’s lab, Tianzhu Zang from Zhou’s lab, Dr. Edward Hua from Hanson’s lab.

I would like to express my gratitude to Dr. Graham Jones, Dr. Robert Hanson, Jean Harris and Nancy Weston from Department of Chemistry and Chemical Biology, Jeffrey Kasilman, Jana Volf and Felicia Hopkins from Barnett Institute for their kind help.

I really thank my precious friends Gao Chen, Yongzhong Wang, Shuhui Ma, Li Zhou, Guangquan Gu, Qinyan Liu, Xinwei Lin, Jinsheng Jiang, Jingxing Zhang, Naibing Du, Ming Yuan and Xiaoyu Guo for their numerous support and care.

At last, special thanks to my family with their unconditional support and love. My parents, my grandparents and my sisters have unwavering faith in me and were always there when I needed them most. This dissertation is dedicated to my family!
# TABLE OF CONTENTS

Abstract .......................................................................................................................... 3

Acknowledgements ....................................................................................................... 7

Table of Contents ......................................................................................................... 9

List of Figures ................................................................................................................ 17

List of Tables ................................................................................................................ 22

Abbreviation and Convention .................................................................................... 24

Chapter 1 Overview of Protein Drugs and Proteomics

1.1 Introduction .............................................................................................................. 29

1.2 Connection between Proteomics and Biotech ....................................................... 30

1.2.1 Overview of Protein Drugs .................................................................................. 30

1.2.2 Follow-on Protein Drugs ..................................................................................... 34

1.2.3 Proteomics as an Advanced Analytical Tool for Protein Drug Analysis ......... 34

1.3 Proteomics ............................................................................................................... 36
1.3.1 Overview of Proteomics ................................................................. 36

1.3.2 Current Methodologies of Proteomic Analysis ............................... 37

1.3.2.1 Mass Spectrometers ................................................................. 37

1.3.2.1.1 Ionization Methods ............................................................ 37

1.3.2.1.2 Mass Analyzers ................................................................. 39

1.3.2.2 Separation Technologies ........................................................... 47

1.3.2.2.1 Gel Based Separation ......................................................... 47

1.3.2.2.2 Gel Free Based Separation ................................................... 48

1.3.2.3 Proteomic Approaches .............................................................. 49

1.3.2.3.1 Bottom-up Approach-Shotgun Sequencing ............................. 50

1.3.2.3.2 Top-down Approach-Protein Fragmentation in the Gas Phase .......... 51

1.3.2.4 Tandem MS (Fragmentation Method) ......................................... 55

1.3.2.4.1 Collision Induced Dissociation (CID) ..................................... 55

1.3.2.4.2 Electron Transfer Dissociation (ETD)/Electron Capture Dissociation (ECD) ................................................................. 56
1.3.2.5 Bioinformatics……...60

1.3.2.6 Quantitation……..63

1.3.2.6.1 Targeted Absolute Quantitation……..64

1.3.2.6.2 Relative Quantitation……..67

1.3.2.6.3 Label-Free Quantitation……..72

1.4 The Application of Proteomic Analysis to the Recombinant Protein Drugs……..74

1.4.1 Overview of Post Translational Modifications (PTM)……..74

1.4.2 Glycosylation……..75

1.4.3 Stability of Protein Drugs……..79

1.4.3.1 Deamidation……..79

1.4.3.2 Oxidation……..83

1.4.4 Disulfide-linkage Determination……..84

1.5 Conclusions and Perspectives……..86

1.6 References……..87
Chapter 2  Mass Spectrometric Analysis of Innovator, Counterfeit, and Follow-On Recombinant Human Growth Hormone

Abstract .........................................................................................................................103

2.1 Introduction ..............................................................................................................104

2.2 Experimental ..........................................................................................................106

2.2.1 Materials .............................................................................................................106

2.2.2 Enzymatic Digestion ..........................................................................................107

2.2.3 LC-MS ..................................................................................................................107

2.2.4 MALDI-TOF/TOF .............................................................................................109

2.2.5 Peptide Assignment ...........................................................................................109

2.3 Result and Discussion ............................................................................................111

2.3.1 Primary structure identification ..........................................................................111

2.3.2 Oxidation ............................................................................................................117

2.3.3 Deamidation .......................................................................................................124

2.3.4 Chain Cleavage ..................................................................................................127

2.4 Conclusions ............................................................................................................131
Chapter 3 The Characterization of the Glycosylation Occupancy and the Active Site in the Follow-on Protein Therapeutic: TNK-Tissue Plasminogen Activator

Abstract

3.1 Introduction

3.2 Experimental
  3.2.1 Materials
  3.2.2 Enzymatic Digestion
  3.2.3 SDS-PAGE and In-gel Digestion
  3.2.4 LC-MS
  3.2.5 MALDI-TOF/TOF
  3.2.6 Peptide Assignment

3.3 Results and Discussion
  3.3.1 Primary Structure identification
  3.3.2 Glycosylation occupancy

References
Chapter 4 Mass Spectrometric Determination of Disulfide Linkages in Recombinant Therapeutic Proteins using On-line LC-MS with Electron Transfer Dissociation

Abstracts.................................................................167

4.1 Introduction..........................................................168

4.2 Experimental Procedures...........................................170

4.2.1 Reagents............................................................170

4.2.2 Enzymatic Digestion..............................................171

4.2.3 LC-MS.............................................................171
Chapter 5 Identification of the Unpaired Cysteine Status and Complete Mapping of the 17 Disulfides of Recombinant Tissue Plasminogen Activator Using LC-MS with Electron Transfer Dissociation/Collision Induced Dissociation

Abstracts

5.1 Introduction
5.2 Materials and Methods..........................................................................................204

5.2.1 Materials...........................................................................................................204

5.2.2 Enzymatic Digestion (without Reduction and Alkylation)..............................205

5.2.3 LC-MS.............................................................................................................205

5.3 Results and Discussion........................................................................................207

5.3.1 Identification of Peptide #2 with O-Linked Fucose and Unpaired Cysteine....208

5.3.2 Identification of Disulfide Linkages of Peptides #4 and #7 with N-Linked

Glycosylation.............................................................................................................213

5.3.3 Identification of Disulfide Linkages of Peptide #3: A Large Glycopeptide.....215

5.3.4 Analysis of TNK-tPA.......................................................................................219

5.4 Conclusions.........................................................................................................220

5.5 References..........................................................................................................221
List of Figures

Chapter 1

Figure 1. Illustration of the complexity of a protein product and processes involved…32

Figure 2. Mass spectrometers used in proteome research…………………………43

Figure 3. Scheme of the newly developed LTQ-Orbitrap XL with ETD……………45

Figure 4. Classical versus electron-based methods for fragmentation of protein and
peptide ions in tandem mass spectrometry……………………………………..54

Figure 5. Nomenclature for peptide fragment……………………………………..55

Figure 6. CID fragmentation pathway………………………………………………57

Figure 7. ETD fragmentation pathway……………………………………………..59

Figure 8. Schematic of ion selection and fragmentation in a triple-quadrupole mass
spectrometer, and the principle behind MRM analysis…………………………66

Figure 9. Mechanism of asparagine deamidation via formation of a cyclic imide
Intermediate……………………………………………………………………81

Figure 10. Methionine oxidation…………………………………………………….84

Chapter 2

Figure 1. Structural diagram of human growth hormone with disulfide bonds………112
Figure 2. LC-MS analysis of the N-terminal (T1) peptide from the tryptic digest of rhGH (Nutropin AQ)……………………………………………………………………113

Figure 3. LC-MS analysis of the disulfide-linked peptide (Cys53–Cys165, T6–T16) from the tryptic digest of rhGH (Nutropin AQ)………………………………………116

Figure 4. Comparison of LC-MS analysis of the tryptic map of rhGH from three different manufacturers…………………………………………………………118

Figure 5. LC-MS analysis of the T2 and the oxidized T2 peptides from the tryptic digest of rhGH (Nutropin AQ)…………………………………………………………120

Figure 6. Extract ion chromatography (XIC) of the T2 and the oxidized T2 peptides from the tryptic digest of rhGH (Nutropin AQ)……………………………………121

Figure 7. LC-MS analysis of the T15 and the deamidated T15 peptides from the tryptic digest of rhGH (Nutropin AQ)……………………………………………………125

Figure 8. LC-MS analysis of the T10 and the cleaved T10 peptides from the tryptic digest of rhGH (Nutropin AQ)…………………………………………………………129

Chapter 3

Figure 1. Primary structure of TNK-tPA…………………………………………………………145
Figure 2. Determination of the mutation of KHRR to AAAA (296-299) at T27 peptide from the trypsin digest of TNK-tPA by the LC-MS analysis………………..146

Figure 3. Determination of the sequence of the tryptic peptide containing the mutation of T103N and N184Q at plus PNGase F digest of TNK-tPA by the LC-MS analysis……………………………………………………………………….149

Figure 4. Analysis of the glycosylation occupancy at N184 of T17 peptide from the trypsin plus PNGase F digest of TNK-tPA………………………………….150

Figure 5A. Comparison of the reduced and alkylated TNK-tPA from the innovator and follow-on manufacturers by SDS-PAGE separation…………………………..152

Figure 5B. Comparison of the reduced and alkylated TNK-tPA (with and without PNGase F treatment) from the innovator and follow-on manufacturers by SDS-PAGE separation…………………………………………………………153

Figure 6. Determination of the extent of 2-chain cleavage from the Lys-C digest of TNK-tPA………………………………………………………………………...156

Chapter 4

Figure 1. Data acquisition scheme with CID-MS2, ETD-MS2, and MS3 steps……….173
Figure 2. Diagram illustrating the cleavage sites of a disulfide-linked peptide by CID and ETD………………………………………………………………………………176

Figure 3. CID, ETD, and MS3 analysis of the disulfide-linked peptide (Cys182-Cys189, T20-T21) from the tryptic digest of human growth hormone (hGH)…………178

Figure 4. CID, ETD, and MS3 analysis of the disulfide-linked peptide (Cys53-Cys165, T6-T16) from the tryptic digest of hGH………………………………………………………………………181

Figure 5. Structural diagram of a monoclonal antibody with disulfide bonds………………184

Figure 6. CID, ETD, MS3, and MS4 analysis of the disulfide-linked peptide (Cys229-Cys229 and Cys232-Cys232 in the hinge region of the Fc) from the Lys-C digest of a therapeutic monoclonal antibody………………………………………185

Figure 7. CID, ETD, and MS3 analysis of the disulfide-linked peptide (Cys264-Cys324 in the first loop of the Fc) from the Lys-C digest of a therapeutic monoclonal antibody……………………………………………………………………187

Figure 8. CID, ETD, and MS3 analysis of the disulfide-linked peptide (Cys6-Cys36 and Cys34-Cys43) from the Lys-C plus trypsin digest of Activase…………………190

Figure 9. CID, ETD, and MS3 analysis of the disulfide-linked peptide (Cys457-Cys36 and Cys34-Cys43) from the Lys-C plus trypsin digest of Activase…………………193

Chapter 5

Figure 1. Theoretical nonreduced tryptic peptides of rt-PA, including linked disulfides
(lines), cysteine position (parentheses), the peptides with glycosylation (cycles), and glycosylation sites (underlines)………………………………………………208

Figure 2. ETD-MS2 spectrum of the disulfide-linked tryptic peptide with an O-linked fucose and unassigned cysteine paired with either a glutathione (A) or cysteine (B)………………………………………………….……… ..……212

Figure 3. Analysis of the disulfide-linked tryptic peptide with an N-linked glycan by CID-MS2 (A), ETD-MS2 (B), and CID-MS3 (C)……………………………………215

Figure 4. Structure of the disulfide-linked peptides from a multidigestion of rt-PA (Lys-C plus trypsin plus Glu-C digestion)………………………………………………..216

Figure 5. Analysis of the disulfide-linked peptide A (two disulfides) from Figure 4 by CID-MS2 (A) and ETD-MS2 (B)………………………………………………..217

Figure 6. ETD-MS2 spectrum of the disulfide-linked peptide B (five disulfides) from Figure 4……………………………………………………………………………….219
List of Tables

Chapter 1

Table 1. Product quality attributes for a biological product…………………………33

Table 2. Performance comparisons of the mass spectrometry instruments……………46

Table 3. Approaches in mass spectrometry proteomics……………………………..53

Chapter 2

Table 1. Summary of the identification of rhGH peptide sequences…………………..115

Table 2. The comparison of the extent of rhGH oxidation for different

manufacturers……………………………………………………………………123

Table 3. The comparison of the extent of rhGH deamidation for different

manufacturers……………………………………………………………………128

Table 4. The comparison of the chain cleavages of rhGH for different

manufacturers……………………………………………………………………130
Chapter 3

Table 1. Determination of glycosylation occupancy of TNK-tPA at N184 site using either PNGase F treatment for tryptic glycopeptides or in-gel digestion of the resolved glycosylated and non-glycosylated intact proteins………………………… …154

Table 2. Determination of the extent of 2-chain cleavage (between R275 and I276) of TNK-tPA using either Lys-C digested peptide fragments or in-gel digestion of the resolved single and 2-chain forms……………………………………157

Table 3. Determination of the extent of methionine oxidation of TNK-tPA products…159

Table 4. Determination of the extent of deamidation of TNK-tPA products………………160
Abbreviation and Convention

aa: amino acid;
aEST: expressed sequence tag;
ADCC: antibody-dependent cell-mediated cytotoxicity;
AQUA: absolute quantitation;
AUC: analytical ultracentrifugation;
CDC: complement-dependent cytotoxicity;
CE: capillary electrophoresis;
CEX: cation exchange;
CV: coefficients of variation;
ECD: electron capture dissociation;
EPO: erythropoietin;
ETD: electron transfer dissociation;
EST: expressed sequence tag;
FFT: fast fourier transform;
Gal: galactose;

GalNac: N-acetylagalactosamine;

GlaNAc: N-acetylglucosamine;

Glc: glucose;

GPI: glycosylphatidylinositol;

HILIC: hydrophilic interaction chromatography;

HPLC: high-pressure liquid chromatography;

HUPO: human proteome organization;

ICR: ion cyclotron resonance;

ICAT: isotope-coded affinity tags;

IE: ion exchange;

IEF: isoelectric focusing;

IFN: interferon;

iMALDI: immuno-MALDI;

iTRAQ: isobaric tags for relative and absolute quantification;

LIT: linear ion trap;
MRM: multiple reaction monitoring;

MS: mass spectrometry;

MS/MS: tandem mass spectrometry;

MudPIT: multidimensional protein identification technology;

NeuAc: sialic acid;

PAGE: polyacrylamide gel electrophoresis;

PMF: peptide mass fingerprinting;

PTM: posttranslational modifications;

QIT: quadrupole ion trap;

Q-TOF: quadrupole time-of-flight;

RF: radio frequency;

RP: reverse phase resins;

rhGH: recombinant human growth hormone;

rt-PA: recombinant tissue plasminogen activator;

SE: size exclusion;

SILAC: stable isotope labeling by amino acids in cell culture;

SISCAPA: stable isotope capture by anti-peptide antibodies;
SIM: selected ion monitoring;

SRM: selected reaction monitoring;

TMT: tandem mass tag;

XCorr: cross-correlation;

Xyl: xylose.
Chapter 1

Overview of Protein Drugs and Proteomics
1.1 Introduction

Traditional drugs in most cases are small molecules that are chemically synthesized, highly purified and well characterized. Biotechnology-derived pharmaceutical products are typically macromolecules derived from living sources that can result in complex heterogeneous mixtures that are difficult to characterize and widely diverse in their form and function.¹ Recombinant protein production is the expression of proteins that have been produced by recombinant DNA techniques.

“Follow-on” protein products refer to proteins and peptides that are intended to be sufficiently similar to an approved product to permit the applicant to rely on certain existing scientific knowledge about the safety and effectiveness of the approved protein product.²

Product comparability is an important issue which one needs to be concerned about. Complexity of protein products with respect to the numerous quality attributes and complexity of biotechnology processes add to the challenges. The ability of non-clinical tools for predicting clinical safety and efficacy is required.³

The research presented in this dissertation focused on the characterization and comparison of different recombinant therapeutic proteins by the application of LC-MS based proteomics technology. Chapter 1 gives an overview of pharmaceutical protein and follow-on drugs, proteomics and related methodologies for pharmaceutical protein analysis.
1.2 Connection between Biotech and Proteomics

In this part, we discuss the connection between biotech/pharmaceutical drugs and proteomic research. The challenges from follow-on protein drugs are also being reviewed.

1.2.1 Overview of Protein Products

Conventional pharmaceutical agents are small-molecule chemicals with a defined molecular weight, typically in the range of 50–1000 Daltons (Da). In contrast, biopharmaceuticals are large, complex and heterogeneous proteins with more variable molecular weights between 5 and 200 kDa. For example, growth hormone is around 22 kDa, and tissue plasmanogin activator (tPA) is around 70 kDa. Compared to the manufacture of small-molecule entities, the manufacture of biopharmaceuticals requires a greater number of batch records (>250 vs. <10), a large number of product quality tests (>2000 vs. <100), increased amount of critical process steps (>5000 vs. <100) and significantly larger process data entries (>60 000 vs. <4000). Additionally, protein products often consist of a heterogeneous mixture of biomolecules that can vary slightly in their molecular weight. The molecular weight of a traditional small-molecule drug can be measured accurately within 1/100 Da. In contrast, the molecular weight of biopharmaceuticals can vary by as much as 1 kDa because of the heterogeneity of the production processes and of products originating from the same process.

Protein products also exhibit great sensitivity towards the particular manufacturing process used for their production. Lot to lot variability in product quality is observed
commonly, even when manufacturing has been performed using the exact same process, and variability of source material has also been known to have an impact on product quality. Thus, the product is affected both by the host cell and the processing steps that follow. In addition, protein molecules can be degraded during processing steps that are performed after their purification, including freezing, thawing, formulation, sterile filtration, filling and freeze drying. Any impurities created in these steps can contribute to decreased potency and/or increased immunogenicity of the protein product. Figure 1 illustrates a typical platform process utilized for manufacturing of monoclonal antibody products.

Product handling can also have a major influence on potential immunogenicity, as it might affect relevant factors, such as purification, exposure to light, cold-chain distribution, excipients, as well as route and frequency of administration.

Unlike traditional small-molecule drugs, post-translational modifications (PTM) of protein products such as, glycosylation, carboxylation, hydroxylation, sulfation and amidation, have been proposed to have a significant effect on protein properties that are relevant to their therapeutic application. The stability of protein products are affected by deamidation and oxidation also (see chapter 2 and chapter 3). Table 1 lists some of the commonly observed modifications to protein products.
Figure 1. Illustration of the complexity of a protein product and processes involved.

(a) Structure of a typical monoclonal antibody along with the various sites of heterogeneity that are commonly observed. (b) Typical platform process utilized for production of a monoclonal antibody. Each of the steps shown here is associated with potential variability, which contributes to the overall variability in the product quality.³
<table>
<thead>
<tr>
<th>Class</th>
<th>Sample quality attributes (commonly used tests)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity</td>
<td>Primary sequence (peptide map and amino acid sequence analysis), immunogenicity (immunoassay), other identity indicators (IE HPLC, gel electrophoresis)</td>
</tr>
<tr>
<td>Potency</td>
<td>Potency (cell-based bioassay, gene expression bioassay, ADCC, CDC)</td>
</tr>
<tr>
<td>General characteristics</td>
<td>Appearance/color (visual), clarity (turbimetric), sub-visible particles, pH (potentiometric), osmolality (freezing-point depression), conformation (near/far UV circular dichroism spectroscopy, Fourier transform infrared spectroscopy, X ray crystallography and differential scanning calorimetry), effector function (receptor binding), concentration (UV absorbance)</td>
</tr>
<tr>
<td>Desired host cell modifications</td>
<td>Glycosylation (monosaccharide composition analysis, oligosaccharide profile, CE, LC-MS, MS/MS, ESI, MALDI-TOF), phosphorylation (peptide mapping with MS)</td>
</tr>
<tr>
<td>Undesirable host cell modifications</td>
<td>Truncation (SE HPLC, gel electrophoresis, analytical ultracentrifugation (AUC), peptide mapping with MS, RP HPLC), glycation (peptide mapping with MS, HPLC), methylation, isomerization (RP HPLC)</td>
</tr>
<tr>
<td>Undesirable process modifications (product variants and impurities)</td>
<td>Aggregation (SE-HPLC, gel electrophoresis, light scattering and AUC), oxidation (peptide mapping with MS), deamidation [capillary isoelectric focusing (cIEF), peptide mapping with MS, and cation exchange (CEX)-HPLC], C-terminal lysine (cIEF, peptide mapping with MS, and CEX-HPLC), misfolds (RP-HPLC)</td>
</tr>
<tr>
<td>Host cell related impurities</td>
<td>Host cell proteins (ELISA), DNA, endotoxin (Limulus amebocyte lysate assay)</td>
</tr>
<tr>
<td>Process related impurities</td>
<td>Antifoam (compendial), extractables (miscellaneous), leachables (miscellaneous), bioburden (membrane filtration test), pluronic (compendial), protein A (protein A ELISA)</td>
</tr>
</tbody>
</table>

Table 1. Product quality attributes for a biological product.³
1.2.2 Follow-on Protein Drugs

The term follow-on protein drugs refer to “proteins and peptides that are intended to be sufficiently similar to a product already approved, or licensed, to permit the applicant to rely for approval on certain existing scientific knowledge about the safety and effectiveness of the approved protein product. Follow-on protein products may be produced through biotechnology or derived from natural sources.” (FDA 5/30/06 response the 5/13/04 Citizen’s Petition)

In addition to variations that are inherent to a specific manufacturing process, the manufacturer of a follow-on protein drug is also likely to use different DNA vectors, different cell culture and purification processes, and perhaps even different formulations compared to the innovator protein. Although the manufacturing process itself might be robust towards small variations in operating conditions of a single process step, the cumulative effect of these minor changes can have an adverse effect on the product quality attributes.

For a follow-on protein drug, similarity to the innovator drug needs to be established with respect to quality, safety and efficacy, with the objective that the follow-on drug shows similar efficacy and comparable or improved immunogenicity compared with the innovator drug.

1.2.3 Proteomics as an Advanced Analytical Tool for Protein Drug Analysis
Market needs lower cost drugs like follow-on drugs but no big risk. During the past two decades, the biotechnology industry has made great advances in improving our understanding of manufacturing processes and their starting materials, as well as of the impact they have on product quality.

The recent advances in attaining a detailed understanding of biotechnology processes have been accompanied by the development of sophisticated analytical tools that are capable of a thorough biochemical characterization of the protein product. As shown in Table 1, it is a common practice to characterize the product using a variety of methods for different attributes, or alternatively, with orthogonal methods for the characterization of a given attribute, thus developing a comprehensive fingerprinting of a protein product.\textsuperscript{9,10} For example, glycosylation, which is frequently considered a critical quality attribute, can be characterized using a number of orthogonal tools, such as monosaccharide composition analysis, oligosaccharide profile, capillary electrophoresis (CE), liquid chromatography-mass spectroscopy (LC-MS), tandem mass spectrometry (MS/MS), electrospray ionization (ESI) MS, and matrix-assisted laser desorption ionization time of flight (MALDI TOF) MS.\textsuperscript{11,12} This enables a thorough comparison of the biochemical characteristics of the follow-on protein drug with that of the innovator drug. Other recent improvements in analytical technology include the use of reverse phase (RP)-HPLC/MS followed by in-line electrospray ionization time-of-flight (ESI-TOF) MS for quantifying structural heterogeneity of monoclonal antibodies IgG1 and IgG2\textsuperscript{5} or the use of LC-MS after trypsin digestion in combination with either ESI-TOF mass spectroscopy or ion chromatography for identifying different forms of a C-terminal
lysine variant\textsuperscript{13} and chip-based gel electrophoresis for separating the non-covalently linked heavy chains of an IgG4 monoclonal antibody from the native form.\textsuperscript{14} Moreover, advances in proteomics techniques have led to the development of biomarkers that are able to improve diagnosis, or to identify disease subsets that might differ in drug response by defining variability in the drug target, thus providing a prediction of response to therapy.\textsuperscript{15,16} This is crucial for establishing comparability of efficacy of the follow-on drug with respect to the innovator drug. In the next part of this chapter, we will discuss the methodologies of the proteomics.

### 1.3 Proteomics

#### 1.3.1 Overview of proteomics

Proteomics is a term in the study of genetics which refers to all the proteins expressed by a genome. The term "proteomics" was first coined in 1997 to make an analogy with genomics, the study of the genes.\textsuperscript{17}

Proteomics is the systematic analysis of protein expression patterns, including identifying protein modifications, quantitating protein expression levels, determining the structure of specific proteins, and elucidating their biological functions.\textsuperscript{18} Research in proteomics can be divided into three areas: expression proteomics, structural proteomics and functional proteomics. In biological research, proteomics has been used to identify
biological relevant proteins like biomarkers. In medicine, it has been used in the
manufacture of biopharmaceutical and evaluating the effective use of the protein drugs.

1.3.2 Current Methodologies of Proteomic Analysis

Proteomics studies the protein complement of cells, including identification,
modification, quantification and localization. Mass spectrometry uses mass analysis for
protein characterization, and it is the most comprehensive and versatile tool in
proteomics.  

1.3.2.1 Mass Spectrometer

Mass spectrometers usually consist three parts: the ion source and optics, the mass
analyzer and the data processing electronics.

1.3.2.1.1 Ionization Methods

Protein MS has enjoyed rapid growth in the past two decades owing to important
developments in experimental methods, instrumentation, and data analysis approaches.
One of the most important developments in instrumentation is the introduction of soft
ionization methods that allow for proteins and peptides to be analyzed by MS. Proteins
and peptides are polar, nonvolatile, and thermally unstable species that require an
ionization technique that transfers an analyte into the gas phase without extensive
degradation. Two such techniques paved the way for the modern bench-top MS
proteomics, matrix-assisted laser desorption ionization (MALDI)\textsuperscript{20-22} and electrospray ionization (ESI).\textsuperscript{23}

MALDI

MALDI is an improvement of the laser desorption ionization (LDI) technique.\textsuperscript{24} The MALDI matrix absorbs laser energy and transfers it to the acidified analyte, whereas the rapid laser heating causes desorption of matrix and [M+H]\textsuperscript{+} ions of analyte into the gas phase. MALDI-generated ions are predominantly singly charged. This makes MALDI applicable to top-down analysis of high-molecular-weight proteins with pulsed analysis instruments. The drawbacks are low shot-to-shot reproducibility and strong dependence on sample preparation methods.\textsuperscript{25, 26} An important development in MALDI ionization is atmospheric pressure MALDI (AP-MALDI).\textsuperscript{27} This interface allows easy interchange between MALDI and ESI sources. The concept of MALDI has led to techniques such as surface-enhanced laser desorption ionization (SELDI)\textsuperscript{28} that introduce surface affinity toward various protein and peptide molecules.

ESI

Unlike MALDI, the ESI source produces ions from solution. Electrospray ionization is driven by high voltage (2–6 kV) applied between the emitter at the end of the separation pipeline and the inlet of the mass spectrometer. Physicochemical processes of ESI involve creation of electrically charged spray, Taylor cone,\textsuperscript{29} followed by formation and desolvation of analyte-solvent droplets. Formation and desolvation of the droplets is
aided by a heated capillary, and in some cases, by sheath gas flow at the mass spectrometer inlet. What distinguished ESI from other ionization method is its ability to produce multiple charged ions from large biological molecules like proteins, which makes it possible to analyze large biological molecules with instrument having a small mass range. An important development in ESI technique includes nano-ESI\textsuperscript{30} in which the flow rates are lowered to a nanoliter-per-minute regime to improve the method’s sensitivity. Nano-ESI is the primary technique we used in this dissertation. An ESI source is usually coupled to the continuous analysis instruments and it is readily interfaced online with liquid separation techniques such as reverse phased HPLC.

1.3.2.1.2 Mass Analyzers

Mass analyzers are an integral part of each instrument because they can store ions and separate them based on the mass-to-charge (m/z) ratios. Ion trap (IT), Orbitrap, and ion cyclotron resonance (ICR) mass analyzers separate ions based on their m/z resonance frequency, quadrupoles (Q) use m/z stability, and time-of-flight (TOF) analyzers use flight time. Each mass analyzer has unique properties, such as mass range, analysis speed, resolution, sensitivity, ion transmission, and dynamic range. Hybrid mass spectrometers have been built that combined more than one mass analyzer to answer specific needs during analysis.\textsuperscript{31-35}

Time of Flight (TOF)
TOF analyzer separates ions based on the time it takes for the ions generated in the source to fly the length of the analyzer and strike the detector. The speed, and therefore the time, at which the ions fly down the analyzer tube is proportional to their m/z value. Larger ions have a slower speed compared to smaller ions and therefore take a longer time to reach the detector.

TOF analyzer has been used primarily to generate peptide fingerprints for identifying individual proteins. As a pulsed –source mass analyzer, it is ideal to couple with MALDI technique. MALDI-TOF instruments are high throughput and offer high sensitivity. Another advantage of MALDI-TOF is that there is virtually no upper mass limitation.

Quadrupole/Triple Quadrupole

The quadrupole analyzer has been the most commonly used mass analyzer with ESI. A quadrupole consists of four metal rods arranged in parallel. Direct current and radiofrequency (rf) voltages are applied to these rods to guide and manipulate ions through the mass analyzer. Altering the voltage allows a specific m/z range of ions to pass through the quadrupole region of the analyzer and onto the detector. The two most common types of quadrupole instrument mass spectrometers are single –stage and triple quadrupoles. Single quadrupole analyzer has limited use since it lacks true tandem MS capabilities. Triple quadrupole analyzer can switch between two different scan modes and has the ability to produce an ion, precursor ion and neutral loss scanning. For example, it
can be used to identify and quantify proteins extracted from 2D-Polyacrylamide Gel Electrophoresis (PAGE) gels\textsuperscript{38} in phosphopeptide characterization.\textsuperscript{39}

**Ion Trap**

An ion-trap mass spectrometer functions just as its name implies: it traps ions. The popularity of this analyzer lies in the discovery and development of ways to manipulate the ions after they are trapped.\textsuperscript{40} The ion trap conducts repeated iterations of collecting, storing and ejecting ions out of the trap. The true power of the ion-trap analyzer is its ability to isolate and fragment peptide ions (i.e., conduct tandem MS) from complex mixtures. To perform tandem MS analysis, specific ions are selected and the trapping voltages are adjusted to eject all other ions from the trap. The applied voltages are then increased to cause an increase in the energy of the remaining ions. These high-energy ions undergo collisions with He\textsubscript{2}, causing them to fragment. These fragments are then trapped and scanned out according to their m/z value. Daughter ions resulting from the fragmentation of large ions can also be retained within the trap and subjected to further rounds of MS/MS. Ion trap instruments\textsuperscript{41,42} are the high-throughput workhorses in proteomics. These versatile instruments feature fast scan rates, MS\textsubscript{n} scans, high-duty cycle, high sensitivity, and reasonable resolution and mass accuracy. The LTQ ion trap from Thermo Scientific combines a tenfold-higher ion storage capacity than 3D traps and high resolution at a fast scanning rate. In addition, the LTQ radial ion ejection offers higher sensitivity than other two-dimensional (2D) ion-trap instruments.\textsuperscript{43} Stand-alone ion trap instruments are best suited for the bottom-up LC-MS protein identification
studies from complex samples and whole cell lysates for which the fast scanning rates and high sensitivity of Linear Ion Trap (LIT)s offer high proteome coverage. LIT is used as the front end of hybrid instruments, such as LTQ-Orbitrap and LTQ-FTICR, where it is used for trapping, ion selection, and ion reactions.

**Fourier Transform Ion Cyclotron Resonance (FTICR)**

FTICR mass analyzer is based on a Penning trap, a device that can store charged particles using a combination of electric and magnetic fields. In an FTICR mass analyzer, ions are captured in a region of a vacuum system using a constant magnetic field and a constant electric field. Ions within the trap resonate at their cyclotron frequency due to the presence of the magnetic field. A uniform electric field that oscillates at or near the cyclotron frequency of the trapped ions is applied to excite the ions into a larger orbit that can be measured as they pass by detector plates on opposite sides of the trap. The energy applied can also be adjusted to dissociate the ions or eject them from the trap. The detector measures the cyclotron frequencies of all the ions in the trap and a Fourier transform is used to convert these frequencies into m/z values. Working at higher magnetic fields benefits many of the parameters related to FTICR performance. The two most critical parameters that are improved through the use of FTICR are resolution and mass accuracy. These analyzers have been proved experimentally to provide the highest resolution and mass accuracy. It is widely used in the bottom-up and top-down
proteomic approaches. Also it can work with ECD technology. We use FTICR in the research for the precursor measurement.

Fig 2 shows the ion source and mass analyzer we covered so far.

**Figure 2. Mass spectrometers used in proteome research.**

Orbitrap
As a new type of mass spectrometer, Orbitrap was invented by Makarov.\textsuperscript{48, 49} Orbitrap uses orbital trapping of ions in its static electrostatic fields in which the ions orbit around a central electrode and oscillate in axial direction. Both Orbitrap and ICR instruments use a fast Fourier transform (FFT) algorithm\textsuperscript{50} to convert time-domain signal into m/z spectrum.

The Orbitrap mass analyzer features high resolution, high mass accuracy a m/z range of 6000, and a dynamic range greater than $10^3$.\textsuperscript{51, 52} When coupled to an LTQ ion trap, the hybrid instrument has the advantages of both high resolution and mass accuracy of the Orbitrap and the speed and the sensitivity of the LTQ. Furthermore, one can operate LTQ-Orbitrap in a parallel fashion: the Orbitrap acquires MS full scans while the LTQ carries out fragmentation reactions. Also Orbitrap can work with ETD technology. Several papers reviewed the performance of the Orbitrap for bottom-up\textsuperscript{53, 54} and top-down\textsuperscript{55, 56} proteomic applications. A next generation LTQ-Orbitrap instrument is termed the LTQ Orbitrap Velos. It has significantly increased sensitivity and scan speed.\textsuperscript{57} The LTQ-Orbitrap offers mass accuracy comparable to the LTQ-FTICR at a lower price tag and a lower maintenance cost for many proteomic applications.

Although the LTQ-Orbitrap is used in top-down experiments, one of the FTICR benefits is broader m/z range, which is best suited for top-down protein analysis and has the ability to carry out gas-phase reactions in the ICR cell.\textsuperscript{58}

Figure 3 shows a scheme the newly developed LTQ-Orbitrap XL with ETD.
Figure 3. Scheme of the newly developed LTQ-Orbitrap XL with ETD.\textsuperscript{59}

Overall, The following instrument configurations are the most widely used solutions in the field of proteomics: ion traps (QIT, LIT), triple quadrupoles (TQ), LTQ-Orbitrap\textsuperscript{49} hybrid instrument (Thermo Scientific), LTQFTICR\textsuperscript{60,61} (Thermo Scientific), and the TQ-FTICR hybrid instruments Q-TOF\textsuperscript{62} and IT-TOF (Shimadzu).\textsuperscript{63-66} Table 2 highlights comparative features and applications of the instruments most commonly used in proteomics.
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Applications</th>
<th>Resolution</th>
<th>Mass accuracy</th>
<th>Sensitivity</th>
<th>Dynamic range</th>
<th>Scan rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIT (LTQ)</td>
<td>Bottom-up protein identification in high-complexity, high-throughput analysis, LC MSn capabilities</td>
<td>2000</td>
<td>100 ppm</td>
<td>Femtomole</td>
<td>$10^4$</td>
<td>Fast</td>
</tr>
<tr>
<td>TQ (TSQ)</td>
<td>Bottom-up peptide and protein quantification; medium complexity samples, peptide and protein quantification (SRM, MRM, precursor, product, neutral fragment monitoring)</td>
<td>2000</td>
<td>100 ppm</td>
<td>Attomole</td>
<td>$10^6$</td>
<td>Moderate</td>
</tr>
<tr>
<td>LTQ-Orbitrap</td>
<td>Protein identification, quantification, PTM identification</td>
<td>100,000</td>
<td>2 ppm</td>
<td>Femtomole</td>
<td>$10^4$</td>
<td>Moderate</td>
</tr>
<tr>
<td>LTQ-FTICR, Q-FTICR</td>
<td>Protein identification, quantification, PTM identification, top-down protein identification</td>
<td>500,000</td>
<td>&lt;2 ppm</td>
<td>Femtomole</td>
<td>$10^4$</td>
<td>Slow, slow</td>
</tr>
<tr>
<td>Q-TOF, IT-TOF</td>
<td>Bottom-up, top-down protein identification, PTM identification</td>
<td>10,000</td>
<td>2–5 ppm</td>
<td>Attomole</td>
<td>$10^6$</td>
<td>Moderate, fast</td>
</tr>
<tr>
<td>Q-LIT</td>
<td>Bottom-up peptide and protein quantification; medium complexity samples, peptide and protein quantification (SRM, MRM, precursor, product, neutral fragment monitoring)</td>
<td>2,000</td>
<td>100 ppm</td>
<td>Attomole</td>
<td>$10^6$</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Performance comparisons of the mass spectrometry instruments.\textsuperscript{19}
1.3.2.2 Separation Technologies

Protein MS is tightly linked and highly dependent on separation technologies that simplify incredibly complex biological samples prior to mass analysis. Because proteins are identified by the m/z of their peptides and fragments, sufficient separation is required for unambiguous identifications. Front-end separation is also required to detect low-abundance species that would otherwise be overshadowed by a higher abundance signal. Therefore, both accuracy and sensitivity of a mass spectrometric experiment rely on efficient separation. Selection of appropriate separation methods is often the first step in designing the proteomic application.

Two major approaches to separation widely used in proteomics are gel based and gel free.

1.3.2.2.1 Gel Based Separation

Two-dimensional Polyacrylamide Gel Electrophoresis (2D PAGE) is the historic centerpiece of the gel-based separation methods.\(^{67}\) 2D-PAGE of proteins is essentially a stepwise combination of two electrophoretic techniques: isoelectric focusing (IEF) and SDS-PAGE. First, proteins are fractionated in a first-dimension pH gradient according to their electric charges. Being amphoteric, protein molecules migrate in an electric field along a continuous pH gradient until they reach their pI, that is, when their net charge is zero. Next, gel tubes or strips are then applied to the top of slab SDS-PAGE gels and the proteins are resolved on the basis of their molecular weight, yielding to a two-
dimensional protein separation. There are many excellent reviews that cover 2D PAGE and gel-based approaches to proteomics. Gel-based methods have been traditionally used with pulsed ionization MALDI instruments in which the protein band can be excised, digested, and off-line sampled with MALDI source.

1.3.2.2.2 Gel Free Based Separation

**High-pressure liquid chromatography (HPLC)**

HPLC is usually directly coupled to instruments with an ESI source. Continuous separation via HPLC is conceptually and technologically compatible with a continuous ionization source such as ESI, and both are usually interfaced with scanning or trapping mass analyzers (LTQ, QqLIT, QqTOF, LTQ-Orbitrap and LTQ-FTICR).

HPLC has become a standard front end for many biological applications and gave rise to several LC/MS setups. The following types of HPLC chromatographic materials are most commonly used in MS-based proteomics: ion exchange (IEX), reverse phase (rp), hydrophilic interaction chromatography (HILIC), affinity, and hybrid materials. However, the high-pressure reverse phase chromatography is as essential to LC/MS as 2D PAGE is to gel-based proteomics. Reverse phase resins (RPLC or RP) separate compounds based on their hydrophobicity, and a significant advantage of RPLC is that the buffers used are compatible with ESI. Given high resolution, efficiency, reproducibility, and mobile phase compatibility with ESI, the analytical RPLC is used as
the single phase and as the last dimension of multidimensional separation\textsuperscript{74-76} before mass analysis. Significant effort goes into increasing peak capacity,\textsuperscript{77} sensitivity,\textsuperscript{78,79} reproducibility, and analysis speed of reverse phase chromatography.\textsuperscript{80,81}

**Multidimensional Separation**

Another common way to address limited peak capacity is to integrate RPLC as part of a multidimensional separation approach. High-complexity large-scale proteomic samples contain thousands of proteins that can range upward of five orders of magnitude in their abundance.\textsuperscript{82} Multidimensional separation is used to address this high sample complexity. By definition, the multidimensional separation approach combines several separation techniques coupled to improve the resolving power. An important consideration for multidimensional separation is the orthogonality of the individual separation methods in which each dimension uses different (orthogonal) molecular properties of molecules as a basis for separation. The 2D SCX chromatography followed by RPLC has become a popular method in shotgun proteomics known as multidimensional protein identification technology (MudPIT).\textsuperscript{83,84}

**1.3.2.3 Proteomic Approaches**

Several general strategies of protein identification have emerged since there are many technical options available for proteome analysis. Gel-based or chromatographic
separation is used to reduce sample complexity prior to mass analysis. Mass spectrometric data acquisition is usually implemented in a data-dependent manner in which information from a current mass spectrometric scan determines the parameters of subsequent scans. Another feature of proteomic analysis is tandem MS, whereby mass analysis is carried out on intact molecular ions (full-scan MS) or on fragmented precursor ions (MSn scans). In most cases, full scans produce masses of the proteins or peptides, and fragmentation scans yield the primary sequence information.

A proteomic analysis begins with the sample preparation in which proteins are either enzymatically digested into peptides (bottom-up analysis)\textsuperscript{85,86} or analyzed intact (top-down analysis).\textsuperscript{87-89}

1.3.2.3.1 Bottom-up Approach-Shotgun Sequencing

The bottom-up approach is the most popular method when tackling high-complexity samples for large-scale analyses. The term shotgun proteomics\textsuperscript{83} is the protein equivalent to shotgun genomic sequencing. Shotgun sequencing and bottom-up proteomics are often used interchangeably. Bottom-up proteomics is an approach in which proteins are proteolytically digested into peptides prior to mass analysis, and the ensuing peptide masses and sequences are used to identify corresponding proteins. Most bottom-up applications require tandem data acquisition in which peptides are subjected to CID.
The most widely used method for bottom-up tandem MS data identification is the database search\textsuperscript{90,91} in which experimental MS\textsuperscript{n} data are compared with the predicted, in silico generated fragmentation patterns of the peptides under investigation.

The bottom-up approach is also well suited for chemical modification of peptides, with the aim of peptide and protein quantification. Techniques such as isotope-coded affinity tags (ICAT),\textsuperscript{92} O\textsuperscript{18} labeling\textsuperscript{93} work best with flexible and accessible peptides. Some of the advantages of the bottom-up approach include better front-end separation of peptides compared with proteins and higher sensitivity than the top-down method. Disadvantages of the bottom-up approach include limited protein sequence coverage by identified peptides, loss of labile PTMs, and ambiguity of the origin for redundant peptide sequences.

1.3.2.3.2 Top-down Approach-Protein Fragmentation in the Gas Phase

Top-down methods use masses of intact proteins and their fragments for successful identifications. In the top-down approach, the protein sample is not subjected to enzymatic digestion, but instead transferred into the gas phase intact. Subsequent measurement of the protein molecular weight and fragmentation of the intact protein using various techniques, combined with database searching, lead to identification of the source protein.\textsuperscript{94} Alternative fragmentation reactions, such as electron capture dissociation (ECD)\textsuperscript{95} and electron transfer dissociation (ETD)\textsuperscript{96} that yield a more
complete backbone sequencing and retain labile Post Translational Modification (PTM)s\textsuperscript{97} are the preferred fragmentation methods of the top-down approach.

Top-down data are usually analyzed using the expressed sequence tag (EST) method\textsuperscript{98} or the de novo method.\textsuperscript{99} Some of the advantages of the top-down approach include higher sequence coverage of target proteins\textsuperscript{100} and better characterization of the posttranslational modifications.\textsuperscript{101,102} Compared with bottom-up approaches, the higher sequence coverage of top-down experiments reduces the ambiguities of the peptide-to-protein mapping, which allows for identification of the specific protein isoforms.\textsuperscript{103} Another reported advantage of the top-down approach is improved reliability of protein quantification\textsuperscript{104-106} when protein abundances are measured directly instead of using abundances of peptides.

However, there are several technological limitations to the top-down method, which keeps it from widespread use. Front-end separation of intact proteins is more challenging than the separation of peptide mixtures. This means that larger quantities of protein and higher mass accuracy instruments such as FTMS and LTQ-Orbitrap are required to resolve isotopic envelopes of coeluting proteins. Furthermore, generic and efficient methods to fragment large proteins are not available yet. Because of these technical limitations, the top-down approach has been limited to the analysis of single proteins and simple protein mixtures so far.

Table 3 summarizes the top-down and bottom-up methods.
<table>
<thead>
<tr>
<th>Prerequisite</th>
<th>Advantages</th>
<th>Drawbacks</th>
<th>Applications</th>
<th>Front end</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top down</strong></td>
<td>High mass accuracy instruments</td>
<td>High sequence coverage</td>
<td>Precursor ion charge state limitations</td>
<td>Ion-exchange chromatography</td>
<td>EST</td>
</tr>
<tr>
<td>LTQ-Orbitrap</td>
<td>PTM and Protein-protein complexes information</td>
<td>Separation limitations</td>
<td>Proteome analysis</td>
<td>RP chromatography</td>
<td>De novo</td>
</tr>
<tr>
<td>LIT-ICR</td>
<td>Multiple PTM identification</td>
<td>Low sensitivity</td>
<td>Alternative splicing</td>
<td>Two-dimensional separation</td>
<td></td>
</tr>
<tr>
<td>Large sample amount</td>
<td>Soft fragmentation with ECD, ETD</td>
<td>Protein identification issues related to the charge-state ambiguity</td>
<td>Multiple PTM analysis</td>
<td>ESI ionization</td>
<td></td>
</tr>
<tr>
<td>Better Quantification compared with bottom up</td>
<td>Better Quantification compared with bottom up</td>
<td>Better Quantification compared with bottom up</td>
<td>Better Quantification compared with bottom up</td>
<td>Better Quantification compared with bottom up</td>
<td>Better Quantification compared with bottom up</td>
</tr>
<tr>
<td><strong>Bottom up</strong></td>
<td>Wide variety of instruments</td>
<td>Large-scale data acquisition</td>
<td>Narrow mass range</td>
<td>Gel based</td>
<td>PMF</td>
</tr>
<tr>
<td>Q-TOF, LIT, LTQ-Orbitrap, etc.</td>
<td>High-complexity samples</td>
<td>Front-end separation required</td>
<td>Protein quantification, PTM analysis</td>
<td>Gel free</td>
<td>Database search</td>
</tr>
<tr>
<td>Sample digest prior to analysis</td>
<td>Sample digest prior to analysis</td>
<td>Sample digest prior to analysis</td>
<td>Sample digest prior to analysis</td>
<td>Sample digest prior to analysis</td>
<td>Sample digest prior to analysis</td>
</tr>
<tr>
<td>Good front-end separation</td>
<td>High sensitivity</td>
<td>Oversampling of high-abundance peptides</td>
<td>RP chromatography</td>
<td>De novo</td>
<td></td>
</tr>
<tr>
<td>Chemical derivatization</td>
<td>Loss of labile PTMs</td>
<td>Affinity chromatography</td>
<td>Library search</td>
<td>Library search</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 Approaches in mass spectrometry proteomics.19
Fig 4 showed the Fragmentation methods in top-down and bottom-up approaches which will be covered in the next part.

Figure 4. Classical versus electron-based methods for fragmentation of protein and peptide ions in tandem mass spectrometry.\textsuperscript{101}
1.3.2.4 Tandem MS (Fragmentation Method)

CID and ETD/ECD are the most popular fragmentation methods in the tandem MS application. CID generates y- and b-ions whereas ETD/ECD provides c- and z-ions. As CID spectra often show high signal intensity of the ion corresponding to the amide bond adjacent to the proline ring, however, upon ETD/ECD this ring remains intact due to secondary bonds that have to be cleaved and results in respective gaps within the spectra. Fig 5 shows the nomenclature for CID, ECD and ETD.

![Figure 5: Nomenclature for peptide fragment](image)

1.3.2.4.1 CID
A commonly used fragmentation technique in MS-based peptide sequencing is CID. A protease typically used for the digestion of protein samples is trypsin, which generally cleaves C-terminal of the basic amino acid residues arginine and lysine if not followed by proline. As a result, peptides that contain at least two charged groups per sequence are generated (including the basic aa side chain as well as the amino group at the N-terminus of the peptide). Upon collision with a neutral gas, those low charged peptides are fragmented within the mass analyzer to produce homolog series of singly and to a minor degree doubly charged b- and y-type ions, which correspond to the cleavage of the backbone amide bonds. Figure 6 shows the fragmentation pathway.

Although CID has been widely adopted for peptide characterization, it has several shortcomings. These include the following: (a) CID often promotes the loss of labile PTMs (e.g., phosphorylation), making PTM site-mapping difficult; (b) CID fails to generate random cleavage along the backbone of peptides that contain multiple basic residues and (c) CID provides limited sequence information for large, highly charged peptides and intact proteins.

1.3.2.4.2 ETD/ECD

In 1998 Zubarev et al.\textsuperscript{95} introduced a novel method for peptide/protein fragmentation: ECD. In this approach, multiply protonated peptides (or proteins) are captured within a penning trap of an FT-ICR MS and exposed to electrons with near-thermal energy. This reaction is an exothermic (6 eV) and nonergodic (i.e., no transfer of
intramolecular oscillation energy) process and induces peptide backbone cleavages to produce homolog series of c- and z-ions.

Figure 6: CID fragmentation pathway.\textsuperscript{107}

One major advantage of ECD is to promote a lower energy pathway than CID. Although fragmentation events occur along the peptide backbone, labile PTMs are preserved.\textsuperscript{97} However, ECD is not widely available due to the high-cost mass
spectrometers needed for the reaction of precursor sample ions with dense populations of near-thermal electrons. To transfer these conditions to a common mass analyzer that uses RF for ion capture, it has to be considered that thermal electrons become kinetically excited due to the influence of the electrodynamic field; thus, these electrons are no longer available for ECD. Furthermore, high-quality ECD spectra often require averaging of data resulting from large numbers of scans acquired over a period of time. This precludes a widespread usage of ECD for the analysis of peptides and proteins in complex mixtures when working with MS interfaced to HPLC.

To avoid the above named limitations Hunt and coworkers introduced a new technique, termed ETD in 2004. It is for usage apart from FT-ICR MS in radio frequency (RF) quadrupole ion trapping mass spectrometers, which are comparatively inexpensive, widely accessible and of lower maintenance. During ETD, protonated peptides are fragmented via electron transfer (ET) from a radical anion.

ETD results when multiply charged peptide cations in the gas phase are allowed to react with radical anions of a polyaromatic hydrocarbon such as fluoranthene (Fig. 7). During the reaction, the radical anion transfers an electron to the multiply charged peptide cation.

\[
[M+3H]^+ + C_{16}H_{10}^- \rightarrow [M+3H]^+ + C_{16}H_{10}
\]

This process is exothermic and triggers highly selective fragmentation of N-Cα bonds along the peptide/protein amide backbone, and fragment ions of type c’ and z’ result. ETD is likely a nonergodic process, not a threshold energy dissociation process such as CID.
ETD can be performed with femtomole quantities of sample and on a timescale that
is compatible to both separations by liquid chromatography and analysis using MS. This method is adaptable to widely spread mass spectrometers that contain the architecture of a LIT or Orbitrap mass analyzers. In comparison to CID, ETD preserves PTMs (which are often labile upon CID), hence conserving all relevant sequence information for peptide identification.\textsuperscript{108,109} Also, ETD is the method of choice for the analysis of intact proteins and large peptides obtained from other proteases than trypsin or resulting from missed cleavage sites.\textsuperscript{107}

1.3.2.5 Bioinformatics

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. At the beginning of the "genomic revolution", a bioinformatics concern was the creation and maintenance of a database to store biological information, such as nucleotide and amino acid sequences. (http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html)

Identification algorithms can be roughly divided into two distinct categories: database search algorithms and de novo sequencing algorithms.

Protein identification by peptide mass fingerprinting (PMF)
Database search algorithms are most commonly used and they rely on the information in protein sequence databases to assign a sequence to a spectrum. To this end, the sequences in the database are digested in silico according to the known cleavage pattern for the protease that was used for the in vitro digest. This yields a collection of proteolytic peptides per database entry, which can directly be used to create a theoretical, ideal PMF spectrum. An experimental PMF can subsequently be compared with this theoretical spectrum. Two of most popular PMF algorithms are Mascot and ProFound. Algorithms that identify peptide fragmentation spectra perform an additional step on the in silico generated peptides before constructing theoretical spectra. Based on the peptide sequence, a list of all the potential fragment ions is calculated which is then used to create a theoretical fragmentation spectrum. This theoretical spectrum is then matched with an experimental fragmentation spectrum in the next step. Popular fragmentation spectrum search engines include Mascot, SEQUEST and X!Tandem.

Protein identification by analysis of uninterpreted tandem MS data. De novo

De novo algorithms extract sequence information directly from the mass differences between the fragment peaks in fragmentation spectra. This approach is particularly useful in cases where the sequence of the peptide precursor that led to the fragmentation spectrum is either not present in any sequence database or when that sequence carries unexpected modifications that alter its mass signature.

The databases are being used in majority of proteomics experiments are: UniProt knowledgebase, (UniProtKB, http://www. Uniprot.org), the NCBI nonredundant (NCBI
Ebi.ac.uk/IPI).

**SEQUEST algorithm**

SEQUEST algorithm was developed by Yates in 1994 and licensed to Thermo
Corporation. It uses a cross-correlation (XCorr) scoring routine to match experimental
tandem spectra of peptides to theoretical mass spectra. In order to submit the
experimental spectra for SEQUEST matching, certain parameters including the mass
range of the peptides, the number of peaks in the spectra, tolerance for mass variation, the
enzyme used in the digestion process, the database to be searched against and any
modifications in the amino acids have to be specified.

SEQUEST utilizes multiple parameters such as XCorr, ΔCn, Sp and Ions. The XCorr
value is dependent on the number of charges present on the peptide ions, the
physiochemical properties of the peptides, database size, searching parameters and
sequence homologies. The XCorr values are used to generate the final ranking of the
candidate peptides in the search. XCorr values are usually higher for well-matched, large
peptides, and lower for small peptides. The human Proteome Organization (HUPO) has
published XCorr criteria (single charge XCorr >1.9, double charge XCorr >2.2, triple
charge XCorr >3.75) for identifications performed with an LTQ instrument. ΔCn is the
delta correlation value between the first and second-ranked hits in the search results. A
general rule is that the best hit should have a ΔCn is at least 0.1 greater between the best
and the second best hits. ΔCn is generally smaller when the data is searched against a
bigger database because the possibility of random sequence similarity is higher in such case. Sp is a score based on the number of theoretical ions derived from the peptides which match with the experimental ions from the tandem MS spectrum within a specified mass tolerance. “Ions” represents the number of ions matched between the experimental ions of a specific tandem MS spectrum and the theoretical ions listed for the peptide. The selection of suitable filters is very important for the quality of searching results. Probability models are a valuable supplement to XCorr-based data filtering.

1.3.2.6 Quantitation

The ability to accurately quantitate changes in protein expression in response to a variety of changes is one of the most important goals of proteomics.\textsuperscript{114} The development of methods for accurate protein quantitation is currently one of the most challenging and rapidly changing areas of proteomics.\textsuperscript{115}

Quantitative proteomics can be categorized into two types: absolute and relative. Absolute quantitation determines changes in protein expression in terms of an exact amount or concentration (e.g. ng or nmoles per gram of tissue, or ng or nmoles/ml of plasma) of each protein present, whereas relative quantitation determines the up- or down-regulation of a protein relative to the control sample, and the results are expressed as “fold” increases or decreases. Relative quantitative proteomics can compare two or more samples using either stable isotope–labeling methods or label-free methods. Isotope labels can be introduced (a) metabolically, (b) chemically, or (c) enzymatically.\textsuperscript{114}
Metabolic labeling represents the earliest point of marking proteins with the stable isotopes of elements ($^{15}$N) or stable isotopes of amino acids (heavy Arg, Lys, Leu, and Ile). For the stable isotope labeling by amino acids in cell culture (SILAC) approach, cell media contain $^{13}$C6-Lys and $^{13}$C6, and $^{15}$N4-Arg for comprehensive labeling of tryptic cleavage products. Examples of chemical derivatization techniques for quantitative proteomics include isotope-coded affinity tags (ICAT), used for the labeling of free cysteine, and isobaric tags for relative and absolute quantification (iTRAQ), used for the labeling of free amines. Chemical derivatization procedures can be applied to any sample at either the protein or the peptide level. Enzymatic labeling usually incorporates $^{18}$O either during or after digestion. Absolute measurements of protein concentrations can be achieved with spiked synthetic peptides.

1.3.2.6.1 Targeted Absolute Quantitation

**Multiple Reactions Monitoring (MRM):**

In its simplest form, quantitation can be done by comparing key characteristic ions from the labeled and unlabeled standard with respect to the peak heights or peak areas of labeled versus unlabeled analytes. Online separation can be added to increase the specificity of the assay. Targeted quantitation of analytes based on isotopically labeled internal standards has been used for LC-MS since 1987. To improve the sensitivity of the assay, instead of scanning all of the possible m/z values; only selected ions can be measured. This selected ion monitoring (SIM) technique can improve the limits of
detection for an analyte by several orders of magnitude. The development of tandem mass spectrometry greatly enhanced the accuracy and specificity of analyte quantitation through the use of selected reaction monitoring (SRM). In this form of quantitation, precursor/fragment ion pairs are monitored. In a triple-quadrupole mass spectrometer (Fig. 8), high sensitivity and specificity are achieved by only allowing a selected peptide to pass through the first quadrupole and enter the collision cell (Q2). Inside the collision cell, the peptide dissociates into fragments specific to the amino acid sequence of the precursor peptide. A second stage of specificity is added in the second MS (Q3), and only a specific fragment is allowed to pass through and strike the detector. By repeatedly cycling through a list of SRM ion pairs associated with a set of specific retention times, multiple peptides can be targeted in a single MRM experiment.

The application of the MRM technique for the absolute quantitation (AQUA) of proteins was first introduced by Steven Gygi in 2003. Gygi et al. proposed that protein concentrations could be determined by quantitating their component tryptic peptides. This means that instead of protein quantitation, one is actually performing peptide quantitation. There are three steps to designing a MRM assay for proteomics biomarker studies: (a) the target proteins have to be selected, e.g. from the set of differentially expressed proteins found in the biomarker “discovery” experiments. (b) peptides have to be selected which have good MS responses and uniquely identify the target protein. (c) the fragment ions for each peptide that provide the best signal intensity and discriminate the targeted peptide from other species present in the sample have to be selected.
MRM analysis with labeled internal standards is capable of sensitive (attomole) and absolute determination of peptide concentrations across a wide dynamic range of $10^3$–$10^4$.\textsuperscript{38, 120-122} One of the strengths of this quantitation method is the low coefficients of variation (CVs) which can be obtained.

Sample complexity can limit the sensitivity of the MRM assay for the detection of low-abundance proteins in the presence of high-abundance proteins. An alternative to immunodepletion is to use antibodies to enrich the sample in target proteins or peptides.
Two methodologies that have shown promise in improving the sensitivity of the MRM assay by enriching the target from a complex background are Stable Isotope Capture by Anti-Peptide antibodies (SISCAPA)\textsuperscript{123} and immuno-MALDI (iMALDI).\textsuperscript{124}

1.3.2.6.2 Relative Quantitation

Relative quantitation is based on the introduction of a chemically equivalent differential mass tag that allows the comparative quantitation of proteins in one sample to another. The labels change the mass of a protein or peptide without affecting the analytical or biochemical properties.\textsuperscript{125} Differential isotopic labels can be introduced metabolically, enzymatically or chemically and depending on the method used at either the peptide or the protein level.

**Metabolic labeling**

Metabolic or in vivo labeling involves the incorporation of stable isotopes during protein biosynthesis. During cell culture, cells are provided with media containing isotopically labeled amino acids which will be incorporated into the proteome during normal cell growth and division. The benefit is that samples grown with different labeled amino acids can be pooled prior to the sample preparation and analysis steps. This technique, however, is not applicable to all sample types but is limited to situations where
the cells are metabolically active. As a result, metabolic labeling cannot be used on tissue samples or biofluids.

\(^{15}\text{N}\)

The first type of metabolic labeling applied to MS-based proteomic analysis was \(^{15}\text{N}\) labeling. Metabolic labeling of yeast using \(^{15}\text{N}\)-labeled growth medium, followed by in-gel digestion and LC-ESI-MS/MS identification of the proteins and determination of the ratios was performed by the Chait group in 1999.\(^{126}\) \(^{15}\text{N}\) labeling can be a good choice for autotrophic organisms such as plants and bacteria. One drawback to this approach is that only two samples can be compared within a single experiment.

\textbf{SILAC}

This method was popularized by Mann and co-workers in 2002.\(^{127}\) The SILAC technique relies on the incorporation of isotopically labeled amino acids into proteins formed by the growing organism. Isotopically labeled amino acids are usually added to the growth medium,\(^{128}\) or the labeled amino acids can be generated by the organism through the addition of isotopically labeled salts to the growth medium. Lysine and arginine are the two most commonly used labeled amino acids which are added to the medium, because each tryptic peptide from proteins synthesized from these amino acids will contain an isotopically labeled lysine or arginine, which increases the quantitative coverage of the experiment.
Unlike $^{15}\text{N}$, SILAC allows for more comparisons within a single experiment due to the availability of several labels. In addition, SILAC has the advantage of a predictable mass shift. However, complete incorporation of isotopic amino acids is not the same for all cell lines.\textsuperscript{129} SILAC cannot be used on cell types that are unable to incorporate certain amino acids.

**Enzymatic labeling**

**Proteolytic labeling ($^{18}\text{O}$)**

$^{18}\text{O}$ is a technique where proteolytic labeling and stable isotope incorporation occurs simultaneously during digestion.\textsuperscript{130} Samples are digested, usually with trypsin, in the presence of $\text{H}_2^{18}\text{O}$ resulting in a 2–4 Da mass shift from the incorporation of one or two $^{18}\text{O}$ atoms on the carboxyl terminus of each peptide. The presence of the label on the carboxyl termini of peptides is advantageous because it facilitates the assignment of ‘y’ ions in the spectra.\textsuperscript{131} In order to prevent contamination with $^{16}\text{O}$ water, samples must be completely dry prior to labeling.

**Chemical Labeling**
ICAT

ICAT introduced by Gygi and Aebersold in 1999, is one of the earliest chemical reagents introduced for quantitative proteomics. The original ICAT label included a thiol-specific reactive group (iodoacetamide), biotin, and either 0 or 8 deuterium atoms, leading to a difference in mw of 8 Da between the two different forms of the tag. A new cleavable version of this reagent was developed in 2003. This cICAT reagent contained nine $^{13}$C instead of eight deuteriums, and an acid-cleavable biotin moiety. One advantage of ICAT is the reduction in sample complexity because the label specifically targets cysteines, a relatively rare amino acid making up only 1.42% of all amino acids. However, this simultaneously reduces the reliability of the quantification as the experiment is based on a limited number of peptides per protein. It also makes it impossible to detect changes in the 10–13% of proteins that do not contain cysteine residues.

iTRAQ

One limitation of ICAT is that there are only two labels available. This could result in multiple experiments if more than two treatments need to be compared, and would increase the cost accordingly. The need for comparisons of larger numbers of treatments led to the development of the 2-or-4-plex ICPL, the 4-or-8-plex iTRAQ, and the 2-or-6-plex Tandem Mass Tag (TMT) labeling techniques, which can compare up to four, eight, or
six samples in a single analysis, respectively. The iTRAQ technique was first described by Ross et al. in 2004 and was subsequently commercialized by Applied Biosystems.

The iTRAQ label is an isobaric tagging compound consisting of a reporter group (variable mass of 114–117 Da or 113–121 Da), a balance group and an amino-reactive group that introduces a highly basic group at lysine side chains and at peptide N-termini. During the initial MS scan, labeled peptides appear as a single peak due to the isobaric masses. The isobaric nature of iTRAQ-labeled peptides allows the signal from all peptides to be summed in both MS and MS/MS modes thus enhancing the sensitivity of detection. During MS/MS, the label releases the reporter group as a singly charged ion of masses at m/z 114–117 (4-plex) or m/z 113, 114, 115, 116, 117, 118, 119 and 121(8-plex). iTRAQ can be used to analyze up to four different samples using the 4-plex kit, or up to eight samples using the new 8-plex kit. These isobaric tagging techniques can also be manipulated to perform absolute quantitation by adding SIS peptides. This then involves comparing peptides from a target protein to a known amount of labeled standard peptide spiked into the sample.

In iTRAQ and similar labels, the labeling is done at the peptide level, and, because every tryptic peptide should be labeled, multiple peptides can be detected for the same protein, thus giving multiple quantitation measurements per protein and increasing the confidence of protein identification. This potential benefit – to identify and quantify low-abundance proteins in complex samples – coupled with the ability to multiplex up to eight samples in parallel suggests that iTRAQ and similar mass balanced labels holds the most promise for quantitative biomarker discovery. Disadvantages of the technique
are the variability in labeling efficiencies, the variability in the initial protein digestion and the expense of the reagents.

1.3.2.6.3 Label-Free Quantitation

Because of the expense and effort required for use of some of the labeling techniques, there now seems to a trend away from isotopic labeling, toward methods of label-free quantitation. Several types of label-free quantitation methods are currently in use. Label-free quantitation is exciting because it holds the promise of shotgun quantitation. Label-free quantitation seems at first to be easy, requiring no labeling steps, only “standard” LC-MS or LC-MS/MS as is typically done for the identification of mixtures of proteins. It has the advantage of eliminating the need for costly labeling reagents, and does not require the multi-step labeling protocols which can lead to irreproducibility and loss of target peptides. It also eliminates the need for the synthesis of expensive isotopically labeled reference peptides. However, on closer examination, label-free quantitation is not so easy. There is a greater need for reducing potential interferences that could lead to suppression effects that could have been compensated for by using isotopically labeled peptide standards. This leads to an increased need for high-resolution or multidimensional chromatography to reduce suppression effects and to allow the detection of low-abundance proteins.

LC/MS based peak area measurement method
A LC/MS based method without isotope labeling has been developed by several groups for global proteomics quantitation.\(^{138,139}\) This technique quantitated a large number of proteins in complex matrixes based on the reproducibility of sample preparation and the linearity of signal versus corresponding peptide concentration.\(^{140,141}\) The relative amount of proteins of interest is calculated by comparing the peak areas (integrated ion counts given mass over the peptide elution time) of peptides selected from the interested protein to a spiked standard protein or house-keeping proteins. This method is simple, can be fully automated, and has the ability to quantitate a large number of proteins in a large set of complex samples.

Q-MEND, a refinement of LC/MS based peak area measurement has been developed for label-free quantitation of relative protein abundance across several complex samples.\(^{142}\) This method requires the use of high-resolution and high-mass-accuracy mass spectrometer such as hybrid LTQ-FT MS or LTQ-Orbitrap MS to minimize false quantitation, In high resolution and mass accuracy data, precursor ions can be assigned to corresponding peptide sequences using MS/MS data recorded form other runs, even if those precursor ions are no selected for data-dependent MS/MS in that run. Thus, the number of peptides selected for quantitation in all runs is nearly doubled by adding those peptides identified by data-dependent MS/MS analysis in at least one of the runs. It has been claimed that Q-MEND can perform relative quantitation of a set of LC-MS runs with a mean quantitation accuracy of 7% and mean precision of 15%, and without manual intervention.
An inherent drawback of LC/MS based peak area measurement technology is the lack of sensitivity to detect low abundance proteins in biological sample of high complexity due to high background interferences.

1.4 The Application of Proteomic Analysis to the Recombinant Protein Drugs

The field of proteomics is important to gain better insights into protein function and structure. One major proteomic approach involves the MS-based analysis of proteins and peptides in order to solve amino acid (aa) sequence as well as site and identity of Post Translational Modifications (PTMs).\textsuperscript{107}

1.4.1 Overview of PTMs

Proteins are converted to their mature form through a complicated sequence of post-translational protein processing and “decoration” events. As many as 300 post-translational modifications of proteins are known to occur physiologically. MS is a central technology in the protein chemist’s toolkit, enabling site mapping and quantification of chemical modifications on proteins, as well as detection of new types of structures. Key to analyzing PTMs by MS is an understanding of their solution and gas-phase reactivities, given that the range in chemical behavior of amino acids and functional groups causes significant differences among peptides with variable composition.
The advent of follow-on therapeutic proteins further highlights the significance of PTMs, because such modifications can influence product equivalence and immunogenicity, issues central to the debate of whether biopharmaceuticals can ever be classified as generic products. The potential effect of upstream and downstream processing upon PTM profiles, may influence the properties of protein drug candidates.\(^{143}\)

1.4.2 Glycosylation

Overview

Glycosylation represents the most widespread PTM found in natural and biopharmaceutical proteins; ~50% of human proteins are glycosylated.\(^{144}\) The presence and nature of the oligosaccharide (glycoform) may impact glycoprotein folding, stability, trafficking and immunogenicity as well as its primary functional activity.\(^{145}\) Glycosylation is characterized by its heterogeneity, in that several glycoforms are usually generated.\(^{146}\) Changes in the glycosylation profiles of specific proteins have been recognized as disease markers. More than one-third of approved biopharmaceuticals are glycoproteins, and an absolute requirement for glycosylation is the main reason that most glycoprotein therapeutics are produced in eukaryotic systems (for example, CHO) rather than prokaryotes. The glycoform profile and functional activity of a glycoprotein may differ depending on the tissue in which it is expressed in the same organism. A focus of interest is the glycosylation of recombinant glycoprotein therapeutics that are produced in
various cell lines and/or species under varying production conditions. Biopharmaceuticals with functions potentially dependent on particular glycoforms include erythropoietin (EPO), antibodies, blood factors, some interferons (IFNs) and some hormones (for example, the gonadotrophins).143

**Various Glycosylation structures**

Glycoproteins consist of proteins covalently linked to carbohydrates. The predominant sugars found in glycoproteins are glucose (Glc), Galactose (Gal), mannose (Man), fucose (Fuc), N-acetylagalactosamine (GalNac), N-acetylglucosamine (GlcNAc), sialic acid (NeuAc) and xylose (Xyl).

There are three major glycan structures attached to proteins: N-linked, O-linked and glycosylphatidylinositol (GPI) anchored structures. N-linked glycans are attached to the side chain of an amino acid group of asparagines, through GlcNAc, or some GalNAc. There is a consensus sequence for N-linked glycosylation, -Asn-X-Ser/Thr-, where X can be any amino acid except Pro. N-linked glycans have a common pentasaccharide structure, Man3GlcNAc2. A variety of other sugars can be attached to this core and comprise three major N-linked families: high mannose type, hybrid type and complex type. O-linked glycans are attached to the hydroxyl group of a serine or threonine residue, through an O-glycosidic bond, O-linked glycosylation is very complex to predict and does not have a consensus sequence. There is no common core structure for O-linked glycans. GPI anchored
proteins are membrane bond proteins. With their carboxyl-terminus linked to a trimannosyl-non-acetylated glucosamine (Man3-GlcN) core.

**Analysis of Glycoproteins**

Analyses of glycoproteins and glycopeptides are analytically challenging, in spite of the emergence of new analytical techniques and approaches over the past few years. For example, glycopeptides are known to suffer from substantial ion suppression when analyzed in complex mixtures. To reduce the adverse effects of competitive ionization, a common strategy is to enrich glycopeptides prior to MS analysis. One such approach utilizes lectins, which are proteins that are selective toward carbohydrates. Glycoproteins can be enriched by lectins immobilized on a solid-phase support. The trapped glycoproteins are then released with the proper eluent, enzymatically digested, and subjected to MS analysis. This method can be very effective for determining glycoproteins present in a complex sample, resulting in multiple peptide identifications for a single protein, thus increasing the confidence of a correct identification. However, due to the competitive ionization, some glycopeptides may still not be identified. To circumvent this problem, glycoproteins can first be digested, while the resulting peptide pool can then be subjected to lectin enrichment. Alternatively, glycopeptides can be chemically immobilized on specialty resins after first oxidizing the glycan. To release the peptide, a subsequent enzymatic cleavage is accomplished through a treatment with PNGase F. While both techniques can be very effective in enriching glycopeptides from complex mixtures, the information pertaining to the glycosylation sites is generally lost.
and alternative techniques must be designed to fully characterize the glycopeptide structures.

Although MS is widely considered to be the ultimate technique in proteomic identification, characterization of glycopeptides remains a challenge because the fragmentation patterns produced through CID tandem mass spectrometry (MS/MS) are often not easily interpretable. During the CID process, glycopeptides undergo preferential glycan fragmentation, as it represents a lower energy fragmentation pathway. Only minimal fragmentation of the peptide backbone is observed. While CID is effective in the characterization of the glycan moieties, it does not provide diagnostic information about the peptide backbone.

The minimal fragmentation of the peptide backbone has previously provided the justification for the removal of the glycan from the peptide in some experiments. Recently, electron-based methods such as ECD and, more recently, ETD, where a radical anion transfers an electron to a peptide cation, have been introduced as alternative methods of peptide fragmentation. Both ECD and ETD induce extensive fragmentation of the peptide backbone with minimal fragmentation of PTMs. This is due to the radical anion induced fragmentation and the absence of vibrational excitation. Moreover, this type of fragmentation results in retaining the PTM on the peptide backbone, thus easily allowing for the site of modification to be identified. When CID and ETD were used in tandem, both the glycan structure and the amino acid sequence of the glycopeptide under investigation could be easily deduced.
In general, the fragmentation characteristics of ETD are similar to ECD, with the intact glycan moiety usually being retained and a radical driven c- and z-ion fragmentation of the peptide backbone observed. These findings provide information about peptide sequence, glycosylation site and glycan mass and are complementary to the glycan composition and branching information obtained by CID. Thus, by alternating CID and ETD a rather comprehensive characterization of glycopeptides might be achieved.\textsuperscript{163,164}

1.4.3 Stability of protein therapeutics

A primary challenge in the commercialization of therapeutic proteins is to maintain its safety and efficacy throughout its manufacture processing, storage, and administration. As with any drug, the stability and purity of protein pharmaceuticals must be monitored carefully. Because proteins present a wide variety of functional groups, they are susceptible to a diverse number of degradation processes\textsuperscript{165}, such as deamidation, oxidation. An understanding of these representative protein degradation reactions and the analytical techniques to characterize them is a prerequisite to any discussion of biopharmaceutical development.

1.4.3.1 Deamidation

Overview
Deamidation and subsequent isoaspartate formation can occur over time in purified proteins, potentially affecting their purity and stability. Deamidation is the loss of an amine group from the side chain of asparagines (Asn) and glutamine (Gln). It is a spontaneous process, one of the most widely occurring covalent modifications in protein and peptides. The deamidation of Asn will result in the formation of protein variants bearing aspartate (Asp), isoaspartate (IsoAsp), or the intra-molecular succinimide intermediate (Asu).

Deamidation of asparagine residues is a major source of amino acid sequence heterogeneity and spontaneous degradation, and may well be the single greatest source of protein damage under conditions of neutral pH. The liability of asparagine residues is highly sequence- and conformation-dependent, arising as the result of the intramolecular reaction shown in Figure 9(IonSource.com). Deamidation results in the conversion of asparagine to a mixture of isoaspartate and aspartate, usually in the ratio of 3:1 to 4:1.

Isoaspartate can arise not only from deamidation of asparagine (Asn), but also from direct isomerization of aspartate (Asp). Whereas the deamidation reaction is greatly accelerated by alkaline pH, the direct isomerization of aspartate is optimal at pH 5. While glutamine residues can also undergo deamidation, they are much less labile than asparagine residues because they are unable to form the same 5-membered cyclic imide intermediate.
Figure 9. Mechanism of asparagine deamidation via formation of a cyclic imide intermediate.

All of the confirmed major degradation sites in proteins have been found to share two common and important features: (a) they occur at Asn-Gly, Asn-Ser, or Asp-Gly sequences and (b) they occur in regions of the polypeptide that are known or predicted to be highly flexible. Deamidation is a leading pathway for spontaneous degradation of purified proteins under mild conditions. The major product of deamidation, isoaspartate, can be easily detected by PIMT-catalyzed radiolabeling of the isoaspartyl sites.
Analysis of Deamidation

Mass spectrometric identification of deamidated peptides is relatively straightforward, as deamidation adds to the mass of intact molecule +0.984 Da (the mass difference between –OH and –NH₂ groups). Since deamidation is a modification stable in the gas phase, CID MS/MS spectra can reveal the position of deamidation even in the presence of several potential deamidation sites.¹⁶⁸

Unlike Asn deamidation, Asp isomerization analysis presents a significant challenge for mass spectrometry, as isomerization is the sublest posttranslational modification of amino acids in polypeptides.¹⁶⁹ Identification of the iso-Asp sites by mass spectrometry requires MS/MS method specific to iso-Asp. As a more reliable criterion, Cournoyer et al. have found iso-Asp-specific fragments in ECD¹⁹⁵ of synthetic peptides.¹⁷⁰ Besides the conventional in ECD N–Cα bond cleavage leading to complementary c and z fragments, these authors detected a diagnostic cleavage giving cₙ•+58.0054 (C₂H₂O₂) and zᵢ₋ₙ−56.9976 (C₂HO₂), where n is the position of the isoaspartyl residue and l is the peptide length. These diagnostic fragments are usually less abundant than the adjacent conventional cₙ and zᵢ₋ₙ species. The same fragments are observed with ETD⁹⁶ the technique that is similar to ECD but that employs radical anions instead of free electrons. Andreazza et al. have recently reported that collisional dissociation of negative peptide ions does not provide Asp/iso-Asp differentiation.¹⁷¹ This means that ECD/ETD are the only MS/MS technologies applied to electrospray-produced ions that provide reliable diagnostic ions for iso-Asp detection.
1.4.3.2 Oxidation

Overview

Proteins are well known to be sensitive to oxidative damage, often with important biological effects. Protein oxidation has been suggested as a causative or contributory factor in many diseases.\textsuperscript{172}

Oxidation is another major chemical degradation pathway of protein and peptides, which can occur during all processing steps in biotechnology manufacturing, from protein isolation to purification and storage\textsuperscript{173}. Amino acid residues that may undergo oxidation include methionine, cysteine, histidine, tryptophan and tyrosine. Oxidation will change the biological activity of proteins such as enzymatic activity, receptor binding ability, antigenicity and sensitivity to in vivo proteases.\textsuperscript{174-175} The reactants and catalysts of the oxidation reaction are numerous and complex. For example, methionine is readily oxidized, by atmospheric oxygen to methionine sulfoxide, as shown in Fig10. The oxidation reaction can be catalyzed by trace amount of metal ions or by hydrogen peroxide, which may be present in organic solvents or pharmaceutical excipients.\textsuperscript{176}

Analysis of Oxidation

Traditionally oxidation has been monitored by reversed-phase HPLC (peptide map), hydrophobic interaction chromatography and amino acid analysis. It has been
shown that mass spectrometry has the ability to measure oxidation by the change in molecular weight on oxidation.

\[
\begin{align*}
\text{CH}_3 & \quad \text{CH}_2 & \quad \text{CH}_3 \\
\text{S} & \quad \text{S} & \quad \text{O} \\
\text{CH}_2 & \quad \text{CH}_2 & \quad \text{CH}_2 \\
\text{CH}_2 & \quad \text{CH}_2 & \quad \text{CH}_2 \\
\text{-NH-CH-CO-} & \quad \text{-NH-CH-CO-} & \quad \text{-NH-CH-CO-} \\
\text{Methionine} & \quad \text{Methionine Sulfoxide} & \quad \text{Methionine Sulfone}
\end{align*}
\]

**Figure 10. Methionine oxidation (IonSource.com)**

1.4.4 Disulfide-linkage determination

**Overview**

Most of the current therapeutic proteins contain disulfide linkages. Disulfide linkages are involved in protein structure formation, protein folding and function.\(^{177}\) This PTM organizes structures of higher order and therefore directs biological function. Disulfide linkages and the location and status of unpaired cysteines (free or blocked) are critical structural features that need to be determined. The free cysteines can be reactive
(most often in redox reactions), leading to covalent association, enzyme catalysis, or, alternatively, disulfide scrambling. Thus, it is clear that disulfide linkage and unpaired cysteine location and status are important structural features that need to be determined as part of the comprehensive characterization of proteins for structure and function correlation.

Analysis of Disulfide Linkage

Significant efforts continue to be devoted to map disulfide bonds to ensure drug quality. Mapping methods generally involve the use of Edman sequencing or mass spectrometry to obtain disulfide-linked peptide information in the first step, followed by determination of disulfide-dissociated peptide sequences after chemical reduction. However, the confidence of assignment can often be limited, particularly when multiple disulfide bonds exist in a protein.

Tools for disulfide bond analysis have improved dramatically in the past two decades, especially in terms of speed and sensitivity. This improvement is largely due to the development of MALDI and ESI, and complementary analyzers with high resolution and accuracy. In general, the characterization of the primary aa sequence demands in solution chemical reduction of the cysteine bridges followed by alkylation of the thiol group. Since these additional steps prior MS analysis are a perquisite, it is of high interest to develop techniques able to cleave disulfide linkages directly in the gas phase during the mass spectrometric analysis. Disulfide bond cleavage has been shown to be favored over peptide backbone breakage for ECD and more recently for ETD. This preferred
cleavage is likely due to the fact that free electrons can be more easily captured by sulfur (disulfides) than by a backbone (amides) during the electron capture or transfer process. In chapter 4 and 5, we have successfully mapped disulfide linkages using online LC-MS with ETD based on the favorable breakage of the disulfide bond. The cleaved or partially cleaved disulfides were further fragmented by CID-MS3 to obtain specific linkage locations.

1.5 Conclusions and Perspectives

This chapter reviews the definition of protein drugs and follow-on drugs. The current technologies in the proteomics research were summarized and the application of these technologies in the analysis of protein and follow-on drugs was reviewed.

In Chapter 2 and Chapter 3, we used LC-MS platform to extensively characterize and compared two different therapeutic protein drugs, recombinant human growth hormone (rhGH) and recombinant tissue plasminogen activator (rt-PA).

Disulfide linkage determination is the topic for the Chapter 4 and Chapter 5. The online LC-MS approach with ETD technology successfully demonstrated in the characterization of disulfide linkages of recombinant human growth hormone (Nutropin), a therapeutic monoclonal antibody, and tissue plasminogen activator (Activase). Also,
the exact location and status of the unpaired cysteine (free or blocked with a glutathione or cysteine) in rt-PA has been identified.

1.6 References:


Yan, J. X.; Gooley, A. A.; Hughes, G.; Smith, I. H.; Williams, K. L.;


49–79.


(21) Hillenkamp, F.; Karas, M.; Beavis, R.C.; Chait, B.T. Anal. Chem 1991, 63,
1193–1203.


(23) Fenn, J.B.; Mann, M.; Meng, C.K.; Wong, S.F.; Whitehouse, C.M. Science

2006, 3-18.

(25) Silvertand, L.H.; Torano, J.S.; de Jong, G.J.; van Bennekom, W.P.


(26) Zheng, J.; Li, N.; Ridyard, M.; Dai, H.; Robbins, S.M.; Li, L. J. Proteome

(27) Laiko, V.V.; Baldwin, M.A.; Burlingame, A.L. Anal. Chem 2000, 72, 652–
657.


(70) Eckerskorn, C.; Strupat, K.; Karas, M.; Hillenkamp, F.; Lottspeich, F.


Chapter 2

Mass Spectrometric Analysis of Innovator, Counterfeit, and Follow-On Recombinant Human Growth Hormone

Publication:

Abstract

We have performed a detailed characterization of recombinant human growth hormone that included the identification of the entire sequence with disulfide linkages as well as subtle modifications by a sensitive liquid chromatography coupled online with tandem mass spectrometry (LC-MS) approach using the accurate peptide mass (FTICR MS) and sequence assignment (MS/MS measurement). The extent of oxidation, deamidation, and chain cleavages were measured by the ratio of peak areas of the nonmodified peptide vs. the sum of peak area of the nonmodified and modified peptides in the same LC-MS analysis. The subtle but distinct differences were found in the recombinant human growth from the three manufacturers (the follow-on, counterfeit, and the original innovator products). In relative comparison, the follow-on product had the highest degree of oxidation at methionine residues, followed by the counterfeit product, and the original innovator product had the least amount of oxidation at all three sites with the similar oxidation order. In cases, the oxidation order was Met14 > Met125 > Met170. In contrast, the follow-on had the least amount of deamidation at asparagine (Asn149), and the counterfeit had the highest degree of deamidation at this site. For the chain cleavage, the follow-on product had the highest cleavage occurring at T 10 peptide (between Asn99 and Ser100), the counterfeit had the highest cleavage on T4 peptide, (between Glu30 and Phe31), and the original innovator product with the least amount of cleavages on both sites. These subtle but distinct differences are likely because of nonidentical manufacturing, formulation procedures, and storage conditions.
2.1 Introduction

Protein drugs in biotechnology manufacturing are often complex in nature due to the complexity of the manufacture procedures and the chemical complexity of proteins. The advent of generic biologicals present even a higher challenge for the analysis of these copy drugs to assure the same quality, efficacy, and safety as compared to the original innovator drugs.

The generic version of recombinant human growth hormone (e.g. Omnitrope) was approved recently in Europe and US through extensive pre-clinical and clinical studies. However, even as the agencies approved the copy drugs as effective and safe as the original innovator drugs, they recognized that exact copies are impossible, and the US named such products as “follow-on” biologics and the Europe as “biosimilar”, recognizing that these are similar but not identical drugs.

Lately, with the increasing globalization of the biotech industry and its high marketing potential, the counterfeit recombinant human growth hormone preparation has become the target in recent fraud scandals and criminal investigations among athletes and action actors. It should be noted these counterfeit drugs, not like the follow-on or biosimilar copies, are often not approved through legal procedures by US or European regulatory authorities and also with limited or no clinical / pre-clinical studies. It is clear in this context that a powerful analytical technique, which is capable of characterizing and comparing the identity of such protein drugs at the detailed molecular level, either similar or not similar, is needed to assure the same drug quality. While we acknowledge
that multiple analytical and biological assays are often needed to assess drug quality.\textsuperscript{9} Nevertheless, a single powerful analytical tool, which can obtain the extensive information in time with high degree of confidence and sensitivity to identify the subtle difference, is our goal in this study as a valuable initial screen.

Currently, liquid chromatography coupled on-line with tandem mass spectrometry (LC-MS) has become the major analytical tool to identify and characterize proteins.\textsuperscript{3,10-15} Lately we have used this LC-MS technology with new generation of mass spectrometers of high detection sensitivity to study a biosimilar version of beta interferon. After extensive characterization, we discovered a low level, but potentially harmful modification, due to glycation at lysine amino acid residues on beta interferon, as well as on the carrier protein, human serum albumin. We also concluded that the glycation modification was due to the use of a slightly different chemical component of excipient in the formulation of the biosimilar version.\textsuperscript{3}

In this report, we used similar LC-MS approach with an additional fragmentation capability in the mass spectrometer, electron transfer dissociation,\textsuperscript{16} to further enhance the detailed characterization of recombinant human growth hormone from three different manufacturers: (1) the original biopharmaceutical, which is in liquid formulation, (2) the biosimilar or follow-on, which is a lyophilized product, and (3) the counterfeit, which is in liquid formulation. As shown in the following, a different and greater extent of oxidation, deamidation, aspartyl isomerization, and chain cleavages were observed in the
counterfeit and follow-on products as compared to the original biopharmaceutical version of recombinant human growth hormone.

2.2 Experimental

2.2.1 Materials

Nutropin AQ (Genentech, So. San Francisco, CA), a liquid formulation product, which consists of 10 mg of recombinant human growth hormone (somatropin, the active drug ingredient), 17.4 mg sodium chloride, 5 mg phenol, 4 mg polysorbate 20, and 10 mM sodium citrate in 2 mL sterile liquid. Omintrope (Sandoz, Holzkirchen, Germany), a lyophilized product, which consists of 5.8 mg of somatropin (~17.4 IU), 27.6 mg of glycine, 2.09 mg of disodium hydrogen phosphate heptahydrate, 0.56 mg of sodium dihydrogen phosphate dehydrate, and a vial with 1.14 mL of bacteriostatic water (1.5% benzoic alcohol and preservatives) for dilution. The counterfeit (unknown vendor), a liquid formulation product in a vial with 2 mL liquid (suspicious to imitate Nutropin AQ), was obtained from Genentech security. A total of 5 vials for each drug product were used for the following LC-MS analysis. Achromobacter protease I (Lys-C) was obtained from Wako Co. (Richmond, VA), and trypsin (sequencing grade) was purchased from Promega (Madison, WI). Guanidine hydrochloride, α-cyano-4-hydroxycinnamic acid (CHCA), ammonium bicarbonate, trifluoracetic acid (TFA) and formic acid (FA) were from Sigma-Aldrich (St. Louis, MO). The CHCA matrix was recrystallized before
use. LC-MS grade water was purchased from JT baker (Phillipsburg, NJ), HPLC grade acetonitrile was purchased from ThermoFisher Scientific (Fairlawn, NJ). Microcon YM-3 Centrifugal Filter Unit was obtained from Millipore (Bedford, MA).

2.2.2 Enzymatic digestion

Protein solution (5 mg/mL) was buffer exchanged with 0.1 M ammonium bicarbonate over an YM-3 Centrifugal Filter Unit, and then endoproteinase Lys-C (1:100 w/w) was added to digest the protein for 4 hr at 37 °C. For tryptic digestion, trypsin (1:100 w/w) was added at room temperature for 8 hr and then added a second time (1:100 w/w) for 12 hr again at room temperature. Digestion was stopped by addition of 1% formic acid.

2.2.3 LC-MS

LC-MS experiments were performed on an LTQ-FT MS instrument (ThermoFisher Scientific, San Jose, CA). An Ultimate 3000 nanoLC pump (Dionex, Mountain View, CA) and a self-packed C18 column (Magic C8, 200Å pore and 5 µm particle size, 75 µm i.d. x 10 cm) (Michrom Bioresources, Auburn, CA) was coupled on-line to the mass spectrometer through a nanospray ion source (New Objective, Woburn, MA). Mobile phase A was 0.1% formic acid in water, and mobile phase B was 0.1% formic acid in acetonitrile. The gradient consisted of: (i) 20 minutes at 2% B for sample loading; (ii) linear from 2 to 60% B over 60 min; (iii) linear from 60 to 80% B over 10 min; and finally (iv) isocratic at 80% B for 10 min. The flow rate of the column was maintained at
200 nL/min. The LTQ-FT MS was operated as in the following conditions: survey full-scan MS spectra with 2 microscans (m/z 400 – 2000) were acquired in the FTICR cell with mass resolution of 100,000 at m/z 400 (after accumulation to a target value of 2x10^6 ions in the linear ion trap), followed by 8 sequential MS2 scans. Dynamic exclusion was utilized with an exclusion duration of 30 sec and no repeat counts. The total cycle time (1 FTICR survey scan with 2 µscans plus 8 sequential linear ion trap MS2 scans) was ~ 2.7 sec.

For the experiments using electron transfer dissociation in the mass spectrometry: an LTQXL with ETD mass spectrometer (Thermo Fisher Scientific, San Jose, CA), consisting of a newly developed linear ion trap with an additional chemical ionization source to generate fluoranthene anions. The mass spectrometer was operated in the data-dependent mode to switch automatically between MS (scan 1), CID-MS2 (scan 2), ETD-MS2 (scan 3), and MS3 (scan 4). Briefly, after a survey full-scan MS spectrum from m/z 400 to 2000 in the linear ion trap (at a target value of 30,000 ions), subsequent CID-MS2 (target value of 30,000 ions and 35% normalized collision energy) and ETD-MS2 (target value of 30,000 ions) activation scan steps were performed on the same precursor ion. Each precursor ion for each of these scans was isolated using the data-dependent acquisition mode with a ± 2.5 m/z isolation width to select automatically and sequentially a specific ion (starting with the most intense ion) from the survey scan. Finally, an additional MS3 step, which isolated the highest intensity ion from the prior ETD spectrum for further CID fragmentation (± 5 m/z isolation width) was performed. This MS3 step used 15 % normalized collision energy relative to CID-MS2, with the decrease of the activation Q value from 0.25 to 0.15, and with the ion target value at
10,000 ions (2 microscans). Scans 2, 3, and 4 were repeated in sequence for an additional 2 times to select for fragmentation of the second and third highest intensity precursor ions from the first survey scan. The total cycle (10 scans), approximately 3 seconds, was continuously repeated for the entire LC-MS run under data-dependent conditions with dynamic exclusion.

2.2.4 MALDI-TOF/TOF

MALDI experiment was performed on a 4700 Proteomics Analyzer (Applied Biosystems Framingham, MA). An aliquot of ~1 µg of rhGH was mixed with matrix solution (7 mg/mL CHCA, 0.1% (v/v) TFA in 50% (v/v) ACN/water) at a ratio of 1:1, and the mixture was then deposited onto a stainless steel MALDI plate for MS and MS/MS measurement. Each spectrum was obtained from a total of 1000 laser shots across the entire spot (50 laser shots at each of 20 random positions within the spot). A maximum of 15 MS/MS spectra were acquired from each spot. The spectra were calibrated using a single internal standard with the 4700 Explorer software (Applied Biosystems), resulting in roughly 30 ppm mass accuracy across the entire MALDI plate.

2.2.5 Peptide Assignment

Spectra generated on LTQ-FT MS were first filtered using BioWorks software (3.3.1, Thermo Fisher Scientific) that had the Sequest algorithm incorporated to assign fragmentation spectra to the most probable peptide sequence. Briefly, the spectra generated in the CID-MS2 step were searched against spectra of theoretical
fragmentations (b and y ions) of recombinant human growth hormone with a mass tolerance ± 1.4 Da (for both precursor and fragment ion) and with either trypsin or Lys-C specificity (2 missed cleavages). The resultant spectra were then filtered using Xcorr (1+ precursor ion ≥1.0, 2+ ≥ 2.0, and 3+ and above ≥ 2.5). The spectra generated in the ETD-MS2 were searched against spectra of theoretical fragmentations (c and z ions) of recombinant human growth hormone but filtered using Xcorr (≥1). Final confirmation of the most probable peptide assignment was obtained by manual inspection of (i) the mass accuracy (< 5 ppm), and (ii) preferred fragmentation patterns in the observed CID-MS2, ETD-MS2 and MS3 spectra. A disulfide-linked peptide was assigned by assuming that the cysteine was modified with a polypeptide chain. For example, if two polypeptides, labeled as P1 and P2, were linked by a disulfide bond from the two cysteine residues, one cysteine would initially be assumed to be modified with the molecular weight of the P1 peptide to search against spectra of theoretical fragmentations. The search was then repeated using the other cysteine assumed to be modified with the P2 peptide. Both searches were then combined to assign the cleavages for the disulfide-linked peptide. It should be noted that if a disulfide-dissociated peptide ion was isolated from ETD for fragmentation, the spectra generated in this MS3 step were searched against spectra of theoretical fragmentations (b and y ions), similar to the CID-MS2 step but with no modification on cysteine residues. Spectra generated on MALDI-TOF/TOF were assigned manually.
2.3 Results and Discussion

Human growth hormone has been used to treat children with hypopituitarism or growth hormone deficiency. Native human growth hormone derived from pituitary gland is a single polypeptide (monomer) with 191 amino acids and two intra-disulfide linkages. Recombinant human growth hormone (rhGH) was expressed in an E.coli cell line with the identical sequence gene and recovered from down-stream purification processes. The amino acid sequence and disulfide linkages for either tryptic or Lys-C fragments of rhGH are shown in Figure 1A and 1B. Since the correct amino acid sequence and disulfide linkages are critical to assess the recombinant DNA process, the primary structure of rhGH was extensively characterized by our LC-MS analysis, from the products of Lys-C or tryptic digestion without reduction, in the following.

2.3.1 Primary Structure Identification

As shown in Figure 2, the N-terminal peptide (T1) was identified at 31.18 min (Figure 2A), with the accurate mass assignment for the peptide, m/z 465.7723 with 2+ charge (Figure 2B). This precursor ion (m/z 465.7723) was isolated using the data-dependent acquisition mode and subjected to CID-MS2 fragmentation in the linear ion trap (Figure 2C). The characteristic fragmentation pattern of CID was observed with the high abundant product ions derived from cleavages at the N-terminal part of proline residues, i.e. y7, y4, b2, and b4 ions, as indicated in the figure. It should be noted that this characteristic fragmentation pattern was very similar to our previous identification of the T1 peptide of rhGH in human plasma.
Native Human Growth Hormone

A single polypeptide with 191 amino acids and two intra-disulfide linkages at Cys53 - Cys165 and Cys182 – Cys189

Fig 1A

Tryptic and Lys-C peptides of rhGH (no reduction)

Tryptic

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1(1-8):</td>
<td>FPTIPLSR</td>
</tr>
<tr>
<td>T2(9-16):</td>
<td>LFDNAML</td>
</tr>
<tr>
<td>T3(17-19):</td>
<td>AHR</td>
</tr>
<tr>
<td>T4(20-38):</td>
<td>LHQLAFDTYOEFEAYIPK</td>
</tr>
<tr>
<td>T5(39-41):</td>
<td>EQK</td>
</tr>
<tr>
<td>T6(42-64, 159-167):</td>
<td>YSFLONPQTSLCFSESIPTPSR (T6) NYGLLYCFRK (T16)</td>
</tr>
<tr>
<td>T7(65-70):</td>
<td>EETQQK</td>
</tr>
<tr>
<td>T8(71-77):</td>
<td>SNLELLR</td>
</tr>
<tr>
<td>T9(78-94):</td>
<td>ISLLIQSWLEPVQFLR</td>
</tr>
<tr>
<td>T10(95-115):</td>
<td>SVFANSLVYGASDNSVYDNLK</td>
</tr>
<tr>
<td>T11(116-127):</td>
<td>DLEEGIQTLMGR</td>
</tr>
<tr>
<td>T12(128-134):</td>
<td>LEDGSFR</td>
</tr>
<tr>
<td>T13(135-140):</td>
<td>TGGIFK</td>
</tr>
<tr>
<td>T14(141-145):</td>
<td>QTYSK</td>
</tr>
<tr>
<td>T15(146-158):</td>
<td>FDTNHNDALLK</td>
</tr>
<tr>
<td>T17(168):</td>
<td>K</td>
</tr>
<tr>
<td>T18(169-172):</td>
<td>DMDK</td>
</tr>
<tr>
<td>T19(173-178):</td>
<td>VETFRL</td>
</tr>
<tr>
<td>T20-T21: (183-191)</td>
<td>IVOCR (T20) SVEGS CGF (T21)</td>
</tr>
</tbody>
</table>

Lys-C

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1(1-38):</td>
<td>FPTIPLSRFDNAMLRAHRHLQAFDTYQEFEEAYPK</td>
</tr>
<tr>
<td>L2(39-41):</td>
<td>EQK</td>
</tr>
<tr>
<td>L3(42-70)</td>
<td>YSFLONPQTSLCFSESIPTPSNREETQQK (L3) NYGLLYCFRK (L8)</td>
</tr>
<tr>
<td>L4(71-115):</td>
<td>SNLELLRISLLIQSWLEPVQFLRSVFANSLVYGASDNSVYDNLK</td>
</tr>
<tr>
<td>L5(116-140):</td>
<td>DLEEGIQTLMGRLEDGSPRTGQIFK</td>
</tr>
<tr>
<td>L6(141-145):</td>
<td>QTYSK</td>
</tr>
<tr>
<td>L7(146-158):</td>
<td>FDTNHNDALLK</td>
</tr>
<tr>
<td>L9(169-172):</td>
<td>DMDK</td>
</tr>
<tr>
<td>L10(173-191):</td>
<td>VETFRLIVOCR SVEGS CGF</td>
</tr>
</tbody>
</table>

Fig 1B

Figure 1. Structural diagram of human growth hormone with disulfide bonds (Figure 1A), tryptic and Lys-C peptide sequences (Figure 1B).
Figure 2. LC-MS analysis of the N-terminal (T1) peptide from the tryptic digest of rhGH. A: Base peak ion chromatogram. B: Precursor mass scan at 31.18 min using FTICR. For illustration purpose, only m/z 464 – 468 region is shown. C: MS2 scan of the m/z 465.7723 (2+) ion. The peptide sequences with the observed fragment ions are shown in the insert.
All other peptides of rhGH were identified in a similar way as shown in Figure 2. The identification of tryptic and Lys-C peptides of rhGH were summarized in Table 1A and 1B, respectively. As shown in Table 1A, all tryptic peptides were identified except 4 small tryptic peptides (underlined), while all Lys-C peptides were identified except 3 small Lys-C peptides (Table 1B) (underlined). The combination of both tryptic and Lys-C peptide identification, only 2 very small peptides (EQK and QTYSK) were not identified. These two very small peptides, probably not retained in the LC-MS chromatogram, were identified by MALDI-TOF analysis. Thus, a total of 100% sequence coverage of rhGH was achieved which allowed us to monitor all possible structural change.

For the examination of the rhGH disulfide linkages, as shown in Figure 3, the disulfide-linked peptide (T6-T16) was identified at 38.57 min (Figure 3A), with the accurate mass assignment for the peptide, m/z 941.7143 with a 4+ charge (Figure 3B), and with assignment of the peptide sequence using cleavages of the peptide such as y6 and b17 (Figure 3C). The other disulfide-linked peptide (T20-T21) was identified in a similar way. It should be noted that the disulfide linkages (the disulfide bond between two cysteines) was not typically fragmented by collision-induced dissociation in LC-MS.21
1A Trypsin digestion

Note 1: trypic and Lys-C peptides identified (amino acids in bold) and not identified (amino acids with underline) by LC-MS analysis were indicated in 1A and 1B, respectively.

Note 2: Combined the results of LC-MS analysis of tryptic and Lys-C peptides (1A and 1B), only EQK and QTYSK peptides were not identified. These two peptides (un-retained) were identified by MALDI-TOF analysis (see Supplementary Material). A total of 100% sequence coverage of rhGH was achieved.

Table 1. Summary of the identification of rhGH peptide sequences.

An alternative means of fragmentation relative to CID is electron transfer dissociation (ETD), which has been shown to cleave polypeptide ions preferentially at disulfide bonds. Thus these two disulfide-linked peptides of rhGH were further characterized by on-line LC-MS with ETD and confirmed that these two disulfide linkages of rhGH were indeed correct.
Figure 3. LC-MS analysis of the disulfide-linked peptide (Cys53 – Cys165, T6-T16) from the tryptic digest of rhGH. A: Base peak ion chromatogram. B: Precursor mass scan at 38.57 min using FTICR. For illustration purpose, only m/z 940 - 943 region is shown. C: MS² scan of the m/z 941.7143 (4+) ion.

After we achieved the identification of 100% primary sequence with two disulfide linkages between the expected cysteine residues, we then compared the rhGH samples
from different manufacturers. From the analysis of Lys-C and tryptic peptide fragments, we confirmed that the rhGH preparations from these three manufacturers indeed have the same primary sequence and disulfide linkages.

However, subtle differences of modifications were observed between the manufacturers, as illustrated for the following tryptic maps. As shown in Figure 4, the three tryptic maps were very similar except for the regions indicated by the circles. Therefore, we focused on the identification of these minor differences as described in the following text.

2.3.2 Oxidation

In the past, one of the key concerns with the manufacture of therapeutic protein drugs was the drug stability during the manufacturing and formulation process, which could lead to the degradation or modification of drugs to cause safety and efficacy concerns.\textsuperscript{24, 25} To assess the stability of protein drugs, the degradation products, such as oxidation, deamidation, and chain cleavages, must be examined.\textsuperscript{26-28} In the first part of our analysis we examined the oxidation of rhGH.
Figure 4. Comparison of LC-MS analysis of the tryptic map of rhGH from three different manufacturers. A: Base peak ion chromatogram of the tryptic map of rhGH from Nutropin AQ. B: Base peak ion chromatogram of the tryptic map of rhGH from the counterfeit. C: Base peak ion chromatogram of the tryptic map of rhGH from the follow-on product. The regions with the differences are indicated by the circles.
As shown in Figure 5, the oxidized and non-oxidized peptides (T12) were identified by accurate mass assignment of the peptides shown in the insert of Figure 5A and 5B. The oxidized peptide should have an extra oxygen atom, and in theory it should have the addition of 15.9944 Da for 1+ charge ion, or 7.9972 Da for a 2+ ion (as in this case) than the non-oxidized counterpart. As shown in the insert of Figure 5A and 5B, the difference of the monoisotopic ion (2+ charge) between the oxidized and non-oxidized peptide was 7.9973 (= 489.2516 – 490.2543), which matched the mass difference of oxygen atom for a 2+ charge ion (within 2 ppm mass accuracy). The position of oxidation was identified by fragmentation (CID-MS2) of the oxidized peptide (Figure 5B), which was localized at the methionine (M) residue, as indicated in the insert. After the identification of the T2 oxidized peptide, we also identified the oxidized T11 and T18-T19 peptides of rhGH using a similar approach (data not shown).

Since both the oxidized and non-oxidized peptides were identified in the same LC-MS analysis, the relative quantitation for the extent of oxidation can be easily performed, similar to our previous studies for glycation and phosphorylation modifications. Figure 6 presents the extracted ion chromatogram (XIC) of the non-oxidized T2 (6A) and the oxidized T2 (6B). The percent of oxidation on this T2 peptide can be estimated by the peak area of the oxidized peptide divided by the sum of peak area of the oxidized and non-oxidized peptides as in the following:
Figure 5. LC-MS analysis of the T2 and the oxidized T2 peptides from the tryptic digest of rhGH. A: MS² scan of the T2 peptide, m/z 490.2543 (2+) ion, with the FTICR MS measurement of the precursor ion in the insert. B: MS² scan of the oxidized T2 peptide, m/z 498.2516 (2+) ion, with the FTICR MS measurement of the precursor ion in the insert.
Figure 6. Extract ion chromatography (XIC) of the T2 and the oxidized T2 peptides from the tryptic digest of rhGH. A: Base peak XIC of the T2 peptide, m/z 490.2543 (2+) ion, with the FTICR MS measurement of the precursor ion in the insert. B: Base peak XIC of the oxidized T2 peptide, m/z 498.2516 (2+) ion, with the FTICR MS measurement of the precursor ion in the insert.
% of T2 Oxidation = Peak area of \( \frac{(\text{Oxidized T2})}{(\text{Oxidized T2} + \text{Non-oxidized T2})} \) x 100 %.

It should be noted that the percent oxidation was calculated based on the assumption that the response factor in mass spectrometry for the oxidized and non-oxidized counter parts were similar. In addition, some extent of oxidation may be artificially occurred in the LC-MS analysis when the electrospray ion source is operated under atmospheric pressure. Nevertheless, the assumption of equal response factors and the amount of artificial oxidation can be precisely corrected with appropriate synthetic standards (e.g. chemically synthesized T2 and oxidized T2 peptides for calibration). In addition, the relative comparison (not the absolute values) between the rhGH preparations from different manufacturers is still valid without using calibration with synthetic standards. Using the equation described above, we then compared the relative percent oxidation of rhGH from the three manufacturers. Table 2 presents the % oxidation of T2, T11, and T18-T19 peptides of rhGH from the three manufacturers. The reported percentages were the average of 5 measurements from 5 different vials with %CV less than 10% in every measurement. As shown, the follow-on product had the highest amount of oxidation at all three sites, followed by the counterfeit product, and the original innovator product with the least amount of oxidation.

It has been suggested that proteins in a solid state (e.g. lyophilized products) can exhibit higher oxidation than proteins in liquid solution (e.g. the liquid formulation)\(^{30-34}\) due to
Oxidized peptides and sites:

T2: LFDNAMLR (Met14)
T11: DLEEGIQTLMGR (Met125)
T18-T19: DMDKVETFLR (Met170)

<table>
<thead>
<tr>
<th>Oxidation (%)a</th>
<th>T2</th>
<th>T11</th>
<th>T18-T19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nutropin AQ (liquid formulation)</td>
<td>3.3 (± 0.2) %b</td>
<td>1.8 (± 0.1) %</td>
<td>1.1 (± 0.1) %</td>
</tr>
<tr>
<td>Counterfeit (liquid formulation)</td>
<td>7.7 (± 0.6) %</td>
<td>4.4 (± 0.4) %</td>
<td>1.4 (± 0.1) %</td>
</tr>
<tr>
<td>Follow-on (lyophilized product)</td>
<td>11.7 (± 1.1) %</td>
<td>7.8 (± 0.6) %</td>
<td>5.7 (± 0.5) %</td>
</tr>
</tbody>
</table>

a: the % of oxidation was calculated used the equation as described in the text.
b: the reported values were the average of 5 measurements from 5 different vials with the CV in the bracket.

Table 2. The comparison of the extent of rhGH oxidation for different manufacturers.

the ability of trapping and stabilizing a trace amount of peroxide species for protein and excipient in the solid state relative to the liquid formulation. Thus, the oxidation could
happen due to the close contact to the drug, either in the lyophilization process (dehydration) or during the storage after lyophilization.\textsuperscript{30, 32, 33} The higher amount of oxidation observed in the follow-on product as compared to the other two products could be explained by the difference between a lyophilized and liquid formulation.

2.3.3 Deamidation

Using the similar approach for identification of oxidation, the deamidation of rhGH was examined. As shown in Figure 7, deamidated and non-deamidated peptides (T15) were first identified through the accurate mass assignment of the peptides, as shown in the insert of Figure 7 A and 7B. The deamidated peptide should have the side chain of asparagine, NH\textsubscript{2}, be substituted by OH,\textsuperscript{26, 35} as the result of asparagine to aspartic acid. Thus, in theory, the deamidation should result in the addition of 0.9840 Da for a 1+ charge ion or 0.4920 Da for a 2+ ion (as in this case). As shown in the insert of Figure 7A and 7B, the observed difference for the monoisotopic ion (2+ charge) was 0.4921 (= 745.8419 − 745.3498) between the deamidated and non-deamidated peptides, which matched the deamidation mass difference for a 2+ charge ion (within 2 ppm). The position of deamidation was further identified by the fragmentation (CID-MS2) observed for the deamidated peptide, asparagine (N) at position 149, as indicated in the insert of Figure 7B. It should be noted that the deamidation, similar to the dehydration, goes
Figure 7. LC-MS analysis of the T15 and the deamidated T15 peptides from the tryptic digest of rhGH. A: MS² scan of the T15 peptide, m/z 745.3498 (2+) ion, with the FTICR MS measurement of the precursor ion in the insert. B: MS² scan of the deamidated T15 peptide, m/z 745.8457 (2+) ion, with the FTICR MS measurement of the precursor ion in the insert.
through the loss of amine in the side chain of aspargine residue to form a 5-membered ring (cyclic imide) with the next amino acid residue, which is usually a Gly or Ser residue in the sequence. This cyclic imide form of aspargine residue usually rehydrates quickly to the aspartic acid residue (i.e. either as α or β aspartic acid or both) in a liquid solution with alkali or acidic condition.\textsuperscript{26,35} We observed very little the cyclic imide form but mostly the aspartic acid forms (both α and β) for the aspargine (N149). This could be explained by the trypic digestion condition (at pH8 for 20 hr as described in the Experimental). In addition, we did not observe any deamidation of the aspargine at 152 position in the same T15 peptide, and that could be explained by the next amino acid residue to N152 (not Gly or Ser).

After the characterization of the deamidation reaction, we used a similar approach as for monitoring oxidation to relatively quantitate the % deamidation of rhGH from the three manufacturers, i.e. the peak areas of the deamidated peptides (both forms; α and β aspartic acids) divided by the sum of peak areas of the deamidated and non-deamidated peptides. Table 3 presents the % deamidation of T15 peptides of rhGH from these three manufacturers. As shown in the table, the follow-on product had the least amount of deamidation, and the counterfeit has the highest level of deamidation. Opposite to the observed extent of oxidation, proteins in a liquid formulation can exhibit higher amount of deamidation than proteins in a solid state.\textsuperscript{36} as observed by the higher amount of deamidation on the two products with liquid formulation.
Deamidated peptide and site:

T15: FDT\textbf{NS}HNDDALLK (N149)

<table>
<thead>
<tr>
<th>Deamidation (%)\textsuperscript{a}</th>
<th>T15</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nutropin AQ</strong></td>
<td>5.7 (± 0.5) %\textsuperscript{b}</td>
</tr>
<tr>
<td>(liquid formulation)</td>
<td></td>
</tr>
<tr>
<td><strong>Counterfeit</strong></td>
<td>10.6 (± 1.1) %</td>
</tr>
<tr>
<td>(liquid formulation)</td>
<td></td>
</tr>
<tr>
<td><strong>Follow-on</strong></td>
<td>2.1 (± 0.2) %</td>
</tr>
<tr>
<td>(lyophilized product)</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{a} the % of deamidation was calculated using the equation as described in the text.

\textsuperscript{b} the reported values were the average of 5 measurements from 5 different vials with the CV in the bracket.

Table 3. The comparison of the extent of rhGH deamidation for different manufacturers.

2.3.4 Chain cleavage
Using the similar approach as above, we next characterized the chain cleavages in the rhGH preparations. As shown in Figure 8, the T10 and the cleaved T10 peptides were first identified through the accurate mass assignment of the peptides, as shown in the insert of Figures 8A and 8B. A similar identification process was used as for Figures 6 and 7, the position of the cleavage was identified after the asparagine residue (between N99 and S100 of T10, see Figure 8B). Similarly, another cleavage site of T4 tryptic peptide was also identified between E30 and F31 of T4 (data not shown).

After the characterization of chain cleavages, we quantitated the % chain cleavage of rhGH from the three manufacturers (the similar approach as above, the peak area of the cleaved peptide divided by the sum of peak areas of the cleaved and non-cleaved peptides). It should be noted that the response factor in mass spectrometry for peptides with different length (e.g. the intact and cleaved peptides) are far greater than the peptides with similar length (e.g. the deamidated and non-deamidated counterpart). Nevertheless, the relative quantitation comparison (not the absolute values) between the manufacturer samples should still be valid without using synthetic standards for calibration. The relative comparison of the % cleavage at T4 and T10 peptides of rhGH from the three manufacturers is shown in Table 4. As shown, the follow-on had the highest level of cleavage for T10 peptide, the counterfeit had the highest cleavage for T4 peptide, and the original innovator product had the least amount of cleavage for both T10 and T4 peptides. We do not know the reason for the high cleavages of the follow-on and
Figure 8. LC-MS analysis of the T10 and the cleaved T10 peptides from the tryptic digest of rhGH. A: MS² scan of the T10 peptide, m/z 1132.0865 (2+) ion, with the FTICR MS measurement of the precursor ion in the insert. B: MS² scan of the cleaved T10 peptide, m/z 872.4508 (2+) ion, with the FTICR MS measurement of the precursor ion in the insert.
Peptides with chain cleavages:

$\downarrow$

T10: **SVFAN**<sub>SLVYGASDNSVYDLLK</sub> (N99 and S100)

$\downarrow$

T4: **LHQLAFDTYQE**<sub>FEEAYIPK</sub> (E30 and F31)

<table>
<thead>
<tr>
<th>Chain Cleavages (%)&lt;sup&gt;a&lt;/sup&gt;</th>
<th>T10</th>
<th>T4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nutropin AQ</strong>&lt;br&gt;(liquid formulation)</td>
<td>11.6 (± 1.1) %&lt;sup&gt;b&lt;/sup&gt;</td>
<td>6.8 (± 0.4) %</td>
</tr>
<tr>
<td><strong>Counterfeit</strong>&lt;br&gt;(liquid formulation)</td>
<td>14.7 (± 1.2) %</td>
<td>11.8 (± 1.1) %</td>
</tr>
<tr>
<td><strong>Follow-on</strong>&lt;br&gt;(lyophilized product)</td>
<td>47.6 (± 4.5) %</td>
<td>7.7 (± 0.7) %</td>
</tr>
</tbody>
</table>

<sup>a</sup>: the % of chain cleavage was calculated used the equation as described in the text.

<sup>b</sup>: the reported values were the average of 5 measurements from 5 different vials with the CV in the bracket.

Table 4. The comparison of the chain cleavages of rhGH for different manufacturers.

counterfeit products but suspect that the observed differences could be due to the different manufacturing and formulation process.\textsuperscript{32, 36}
2.4 Conclusions

Our novel LC-MS approaches with new mass spectrometry instrumentation has successfully identified the entire recombinant human growth hormone sequence with disulfide linkages and subtle modifications. The modification sites for the corresponding peptides of oxidation, deamidation, and chain cleavages were precisely determined by accurate mass measurement (FTICR MS) and peptide sequence assignment (MS/MS measurement). The extent of oxidation, deamidation, and chain cleavages were measured by the ratio of peak areas of the modified and non-modified peptides.

In relative comparison, the follow-on product had the highest amount of oxidation on all three sites (Met14, Mat125, and Met170), with the order of Met14 > Met125 > Met170, followed by the counterfeit product, and the original innovator product with the least amount of oxidation. In all products, the relative oxidation at the three sites occurred with similar order. In contrast, the follow-on product had the least amount of deamidation, and the counterfeit had the highest level of deamidation. For the chain cleavage, the follow-on product had the highest level of cleavage at the site in the T10 peptide, while the counterfeit had the highest level of cleavage on T4 peptide, and the original innovator product with the least amount of cleavages on both peptides.

These modifications in the rhGH may or may not cause any safety and efficacy concerns. The extensive animal and human clinical studies, such as for the innovator and follow-on products, certainly alleviated these concerns. In conclusion, we have demonstrated that a powerful analytical tool as described here can allow one to
extensively characterize and quantitate different versions of human growth hormone and
with high degree of confidence and sensitivity to pinpoint the subtle and distinct
differences due to non identical manufacturing and formulation procedures. We believe
that the regulation of generic biopharmaceuticals can be benefit from the detailed analysis
of their products with such a powerful analytical technology as a valuable initial screen.

2.5 References

(4) The US Food and Drug Administration (FDA) on May 31 approved Omnitrope, a
recombinant human growth hormone and the country's first so-called follow-on
(7) Sylvester Stallone charged with illegally importing 48 vials of human growth
(8) Marion Jones gets 6-month prison term for fraud scan of EPO, human growth


Chapter 3

The Characterization of the Glycosylation Occupancy and the Active Site in the Follow-on Protein Therapeutic: TNK-Tissue Plasminogen Activator

Publication:

Abstract:

TNK-tPA products from the innovator and follow-on manufacturers were characterized and compared. All tryptic peptides including N-terminal, C-terminal and mutated peptides as well as the disulfide linked peptides were identified, with the demonstration of the same primary sequence and disulfide linkages between the innovator and follow-on products. The three N-linked and one O-linked fucose glycosylation sites were identified. The two N-linked (N103 and N448) and one O-linked fucose (T61) sites were fully glycosylated in both innovator and follow-on products. The other N-linked site (N184) was partially glycosylated and exhibited a ~2.5 fold difference between the innovator (60% occupancy) and follow-on (25% occupancy) products. Since the glycosylation occupancy at this site is known to affect biological activity in the clot lysis assay, this observed difference could cause a concern as to their bioequivalence. The cleavage site for the conversion of the zymogen form to active enzyme was also identified between R275 and I276, with a cleavage of 40% for the innovator and 10% for the follow-on products. Both the % glycosylation occupancy and the chain cleavage were determined by two independent approaches, starting from either the peptide or intact protein separation, with consistent results by both methods. Subtle differences of modifications such as deamidation and oxidation between innovator and biosimilar were shown at M207, M445, M490 and N58, N184. The observation of different extent of oxidation at M207 and deamidation at N184, which could influence the clot lysis activity,
were also of potential concern in drug efficacy between the follow-on and innovator products.

3.1 Introduction

Protein therapeutics are often complex in nature, particularly for glycosylated proteins. The production of generic or follow-on biologicals with glycosylation thus presents a high degree of challenge as to reproduce a similar if not identical structure to their original innovator drugs. Therefore, to assure a similar or even same structure, a powerful analytical technique is needed to comprehensively characterize the structure of such protein drugs to the detailed molecular level. The recent advancement of proteomic technology, such as liquid chromatography coupled online with tandem mass spectrometry (LC-MS), has been shown to be able to comprehensively characterize targeted proteins. For example, our laboratory has extensively studied a biosimilar version of beta interferon using LC-MS. After comprehensive characterization, we discovered a low level, but potentially harmful modification, due to glycation at lysine amino acid residues on beta interferon as well as the carrier protein, human serum albumin. In addition, we further applied this approach to study recombinant human growth hormone obtained from the follow-on, counterfeit, and innovator manufacturers. This characterization included the identification of disulfide linkages using LC-MS with electron transfer dissociation. We were able to distinguish subtle but distinct differences in oxidation, deamidation and chain cleavages from the three manufacturers (the original innovator, biosimilar and counterfeit products). These subtle but distinct
differences were likely due to a combination of nonidentical manufacturing, formulation procedures, and storage conditions.

In this work, we used a similar LC-MS approach combined with a multi-separation and multienzymatic digestion strategy for a more complex glycoprotein, TNK-tPA, a generic variant of tissue plasminogen activator (t-PA), which was approved for treatments of acute myocardial infarction and ischemic stroke.\textsuperscript{7-9} TNK-tPA has the same amino acid sequence as natural human t-PA except with the three substitutions, at T103 to N, at N117 to Q, and at KHRR (296-299) to AAAA. These substitutions lead to a longer half life and higher fibrin specificity than t-PA.\textsuperscript{10-1} In this study, differences in glycosylation occupancy and chain cleavage at the activation site of the enzyme have been observed between the innovator and follow-on products.

3.2 Experimental

3.2.1 Materials

TNK-tPA (Tenecteplase) (Genentech, So. San Francisco, CA), a lyophilized product, consisted of 50 mg of recombinant tissue plasminogen activator. The follow-on product, Elaxim (Gennova Biopharmaceuticals Ltd, Hinjwadi, Pune, India), a lyophilized product, consisted of 52.5mg of recombinant tissue plasminogen activator and similar excipients as for Genentech. Achromobacter protease I (Lys-C) was obtained from Wako Co. (Richmond, VA), endoproteinase Glu-C from Roche (Indianapolis, IN), and trypsin (sequencing grade) from Promega (Madison, WI). PNGase F, guanidine hydrochloride,
α-cyano-4-hydroxycinnamic acid (CHCA), ammonium bicarbonate, trifluoroacetic acid (TFA) and formic acid (FA) were from Sigma-Aldrich (St. Louis, MO). The CHCA matrix was recrystallized before use. LC-MS grade water was purchased from JT Baker (Phillipsburg, NJ), and HPLC grade acetonitrile was from Thermo Fisher Scientific (Fairlawn, NJ). Microcon YM-10 Centrifugal Filter Unit was obtained from Millipore (Bedford, MA).

3.2.2 Enzymatic Digestion

Protein solution (2.5 mg/mL) was denatured with 6M guanidine hydrochloride containing 100mM ammonium bicarbonate, reduced with 5mM dithiothreitol (DTT) for 30 min at 37 °C and alkylated with 20mM of iodoacetamide (IAA) in the dark for 30min at room temperature. The reduced and alkylated protein was buffer exchanged with trypsin digestion buffer (100mM ammonium bicarbonate, pH 8) using a 10kDa molecular weight cut-off (MWCO) filter. The endoproteinase Lys-C (1:100 w/w) was added to the protein solution for 4 hr at 37 °C. For trypsin digestion, trypsin (1:100 w/w) was added to the protein solution at room temperature for 8 hr and then added a second time (1:100 w/w) for 12 hr at room temperature. Digestion was stopped by addition of 1% formic acid. For Glu-C digestion, Glu-C (1:50 w/w) was added to the protein solution at 37 °C for 8 hrs. For PNGase F digestion (after trypsin digestion), PNGase F (10units/mg) was added to the trypsin digest solution for an additional 4 hr at 37 °C. In all cases, digestion was terminated by addition of 1% formic acid. For digestion without disulfide reduction,
the same digestion protocol as above was applied except skipping the reduction and alkylation steps.

3.2.3 SDS-PAGE and In-gel Digestion

An aliquot of protein solution (20 or 40 µg) diluted with the running buffer of a SDS-PAGE gel (10% Tris-HEPES-SDS gel) for separation. The gel bands of interest were cut for digestion. Briefly, the gel slices, after removed Coomassie stain, were reduced with dithiothreitol (DTT) by the addition of 250 µL of 10 mM DTT in 0.1 M ammonium bicarbonate and incubated for 30 min at 56 °C, and alkylated with 250 µL of 55 mM iodoacetamide (IAA) in 0.1 M ammonium bicarbonate at room temperature for 1 hr in the dark. Then, a trypsin digestion reagent (12.5 ng/µL trypsin in 50 mM ammonium bicarbonate, pH 8.0) was added for 30-35 min at 4 °C, followed by further incubation overnight at 37 °C. The digested peptides were extracted from the gel with 25 mM ammonium bicarbonate, then acetonitrile (37 °C for 15 min), and further extracted with 5% formic acid at 37 °C for 5 min. All supernatants were collected and concentrated for the subsequent LC-MS analysis.

3.2.4 LC-MS

An Ultimate 3000 nanoLC 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. x 15 cm) (Michrom
Bioresources, Auburn, CA) was coupled on-line to an LTQ-FT mass spectrometer (Thermo Fisher Scientific, San Jose, CA) through a nanospray ion source (New Objective, Woburn, MA).

Mobile phase A was 0.1% formic acid in water, and mobile phase B was 0.1% formic acid in acetonitrile. The gradient consisted of: (i) 20 minutes at 2% B for sample loading; (ii) linear from 2 to 60% B over 60 min; (iii) linear from 60 to 80% B over 10 min; and finally (iv) isocratic at 80% B for 10 min. The flow rate of the column was maintained at 200 nL/min. The LTQ-FT MS was operated as follows: survey full-scan MS spectra (m/z 400 – 2000) were acquired in the FTICR cell with mass resolution of 100,000 at m/z 400 (with the ion target value of 2x106 ions), followed by 8 sequential MS2 scans using LTQ portion. For disulfide linkage determination, an LTQXL with ETD mass spectrometer (Thermo Fisher Scientific) was operated in the data-dependent mode to switch automatically between MS (scan 1), CID-MS2 (scan 2), ETD-MS2 (scan 3), and MS3 (scan 4). Briefly, after a survey MS spectrum from m/z 400 to 2000, subsequent CID-MS2 and ETD-MS2 steps were performed on the same precursor ion with a ± 2.5 m/z isolation width. The CID-MS3 step was performed on the highest intensity ion from the ETD-MS2 spectrum (± 5 m/z isolation width).

3.2.5 MALDI-TOF/TOF

An aliquot of ~1 µg of TNK-tPA was mixed with matrix solution (7 mg/mL CHCA, 0.1% (v/v) TFA in 50% (v/v) ACN/water) at a ratio of 1:1, and the mixture was then
deposited onto a stainless steel MALDI plate for MS and MS/MS measurement using a 4700 Proteomics Analyzer (Applied Biosystems Framingham, MA). Each spectrum was obtained from a total of 1000 laser shots across the entire spot (50 laser shots at each of 20 random positions within the spot). A maximum of 15 MS/MS spectra were acquired from each spot. The spectra were calibrated using a single internal standard with the 4700 Explorer software (Applied Biosystems), resulting in roughly 30 ppm mass accuracy across the entire MALDI plate.

3.2.6 Peptide Assignment

The spectra generated in the CID-MS2 step were searched against spectra of theoretical fragmentations (b and y ions) of TNK-tPA with a mass tolerance ± 1.4 Da (for both precursor and fragment ions) and with either trypsin or Lys-C specificity, using Xcorr (1+ precursor ion ≥1.9, 2+ ≥ 2.2, and 3+ and above ≥ 3.8 as the initial filter. The spectra generated in the ETD-MS2 were searched against spectra of theoretical fragmentations (c and z ions) of TNK-tPA but filtered using Xcorr (≥1) initially. Final confirmation of the assignment was obtained by manual inspection to match all high abundant product ions with the precursor ion mass accuracy (< 5 ppm for LTQ FT MS).
3.3 Results and Discussion

Recombinant tissue plasminogen activator (rt-PA) is a glycoprotein composed of 527 amino acids, which are homologous with five different protein families, namely finger, growth factor, two kringle regions, and a serine protease (see Figure 1). TNK-tPA has the same number of amino acids as rt-PA except for mutation at the three locations as described in the introduction and Figure 1 (amino acids with circles). Since the correct amino acid sequence is critical to assess the success of a recombinant DNA process, the primary structure of TNK-tPA was first extensively characterized by LC-MS analysis from the products of Lys-C, trypsin and Glu-C digestion.

3.3.1 Primary Structure identification

The entire amino acid sequence was identified by LC-MS tryptic mapping except for the 6 small peptides (RPDAIR, NPDR, HNYCR, NRR, TYR, and SDSSR), which were not retained in the LC separation but were further identified with MALDI-TOF analysis. In addition, we used Lys-C or Glu-C digestion and the corresponding peptide fragments with overlapped sequences at these six regions were also identified.
Figure 1: Primary structure of TNK-tPA. The positions of amino acid mutation from rt-PA and the chain cleavage for activation are indicated in the figure.

The peptide (T27) which contains the mutation sites of AAAA, as shown in Figure 2, was identified at 41.90 min (Figure 2A), with the accurate precursor mass measurement (m/z 895.1342, 2+ charge) (Figure 2B) and CID-MS2 of the precursor ion (Figure 2C). The high abundance product ions, as the results of characteristic fragmentation by CID, were indicated in the figure. In addition, compared with original rt-PA digestion, there was no un-mutated peptide in the TNK-tPA digest, which further confirmed the mutation
Figure 2: Determination of the mutation of KHRR to AAAA (296-299) at T27 peptide from the trypsin digest of TNK-tPA by the LC-MS analysis. The LC-MS elution profile (A), the precursor mass of T27 by FTICR (B), and the CID-MS2 of the precursor (C), with the annotation of the identified peptide sequence indicated in the inserts.

is complete. Both the follow-on and the innovator drugs have the same primary and mutated sequence at the same sites.
3.3.2 Glycosylation occupancy

The other two mutation sites occurred in the same tryptic peptide (T11), GN(103)WSTAESGAECTNWQ(117)SSALAQKPYSGR, in which T103 was replaced by N and N117 by Q. As seen, these two mutations moved the glycosylation site from N117 to N103, which resulted in the elimination of high-mannose structure at N117 and with the gain of a complex-type glycan at N103. This change was believed to have a longer half life and higher fibrin-binding specificity than rt-PA. The full glycan structures of TNK-tPA will be extensively characterized and compared in a separate report. To confirm the gene mutations at these two sites, the protein was deglycosylated by PNGase F, and the remaining peptide backbone (T11) was analyzed. It should be noted that the asparagine (N103) was converted to aspartic acid (D103) after the removal of the glycans by PNGase F. Thus, as shown in Figure 3, the peptide backbone of T11 was identified using the same approach for the identification of T27. If this site (N103) is partially glycosylated, we should observe the non-glycosylated counterpart of T11 as N103, not D103 (which derived from the deglycosylation step). The mass of ~1 Da difference (similar to deamidation) should be readily differentiated by FTMS measurement. We did not observe any non-glycosylated counterpart of T11, which suggests that this site is indeed fully glycosylated. The N-linked site at N448 found as fully glycosylated, was also determined by this approach. The O-linked fucose was identified at T61, based on the addition of fucose to the peptide mass as well as the
characteristic neutral loss of fucose by CID-MS2. The site of fucose attachment was identified by ETD as described in our previous report. Both the follow-on and the innovator drugs have the same mutated sequence on the T11 peptide, with the glycosylation occupancy (full) on both N103 and N448 sites as well as at T61 (fucose) on both drugs.

Similarly, the remaining N-linked site (N184) was found partially glycosylated since both N184 and D184 of the corresponding peptides (T17) were found after PNGase F treatment (see Figure 4). The % of glycosylation occupancy was determined by the ratio of the D184 peptide divided by the sum of N184 and D184 peptide intensities, using a similar method as described in our previous paper. In this calculation, the observed intensity of D184 was corrected from the D184 intensity derived from the sample without PNGase F treatment (caused by the deamidation of N184, see the next deamidation section). As a result, approximately 60% glycosylation occupancy for innovator and 25% for the follow-on drug were found (~2.5x difference). The glycosylation occupancy at this site has been reported to exhibit differences in biological activity, as the type II t-PA (non-glycosylated at this site) has higher clot lysis activity than the type I t-PA (fully glycosylated at this site). Thus, the major difference in this glycosylation
Figure 3: Determination of the mutation of T103N and N184Q at T11 peptide from the trypsin plus PNGase F digest of TNK-tPA by the LC-MS analysis. The LC-MS elution profile (A), the precursor mass of the deglycosylated T11 by FTICR (B), and the CID-MS2 of the precursor (C), with the annotation of the identified peptide sequence indicated in the inserts.
Figure 4: Determination of the glycosylation occupancy at N184 of T17 peptide from the trypsin plus PNGase F digest of TNK-tPA. The extracted ion chromatography of D (aspartic acid) form representing the glycosylation occupancy and N (asparagine) form representing the nonglycosylation occupancy of T17 are shown as Figures 4A and 4B, respectively, with the accurate precursor masses representing for D and N forms indicated in the inserts. The % of glycosylation occupancy was estimated by the ratio of D divided by the sum of D and N peptide intensities, as indicated in the bottom of the figure.
occupancy, which is related to biological activity, should again raise a concern as to if these products are biosimilar.

To further confirm the difference detected by the PNGase F treatment of tryptic glycopeptides, we used the SDS-PAGE to separate the glycosylated and non-glycosylated forms of the intact protein for analysis. As shown in Figure 5, both the follow-on and innovator drugs were first reduced and alkylated and then separated by SDS-PAGE side by side as indicated in the figure. It should be noted that alkylation of the reduced form of intact TNK-tPA is necessary prior to running the SDS-PAGE. As shown in the right (enlarged) panel of Figure 5A, the type I t-PA (full glycosylated at N184) eluted with a higher molecular weight than the type II t-PA (nonglycosylated at N184), along with the associated cleaved (2-chain) and intact (single-chain) forms of TNK-tPA. We will discuss analysis of the cleavage of TNK-tPA in the next paragraph. These type I or type II forms, as indicated in the gel of Figure 5A, were further confirmed by the PNGase F treatment. After deglycosylation, the gel bands were simplified to one band (single chain) or two bands (2 chains), as shown in Figure 5B. We also further confirmed their identities by analyzing these gel bands with the observation of the corresponding tryptic peptides with either N184 or D184 (with or without PNGase F treatment). In addition, we compared the ratio of type I and II from these in-gel peptide intensities isolated from the corresponding gel bands. Again, we observed the similar trend, with the follow-on product at ~20%
Figure 5A: Comparison of the reduced and alkylated TNK-tPA from the innovator and follow-on manufacturers by SDS-PAGE separation. The resolved single and 2-chain forms with their associated glycosylation such as type I (3 glycosylation sites) and type II (2 glycosylation sites) are indicated with squares and arrows in the figure.

occupancy and the innovator drug with ~50% occupancy (~2.5x difference). The results, as shown in Table 1, are consistent by these two independent approaches, starting from either the peptide or intact protein separation.
Figure 5B: Comparison of the reduced and alkylated TNK-tPA (with and without PNGase F treatment) from the innovator and follow-on manufacturers by SDS-PAGE separation. The resolved single and 2-chain forms with their associated glycosylation such as type I (3 glycosylation sites) and type II (2 glycosylation sites) or deglycosylated forms (with PNase F treatment) are indicated with squares and arrows in the figure.
### Table 1 Determination of glycosylation occupancy of TNK-tPA at N184 site using either PNGase F treatment for tryptic glycopeptides or in-gel digestion of the resolved glycosylated and nonglycosylated intact proteins.

<table>
<thead>
<tr>
<th>Product / Method</th>
<th>Innovator</th>
<th>Follow-on</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNGase F treatment</td>
<td>60 (±3) %</td>
<td>35 (±2) %</td>
</tr>
<tr>
<td>(% occupancy at N184)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-Gel Analysis</td>
<td>50 (±5) %</td>
<td>20 (±3) %</td>
</tr>
<tr>
<td>(% occupancy at N184)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Non-glycosylated and deglycosylated tryptic peptides and sites:

T17 (163-189): YSSEFCSTPACSEGNSDCYFGDGSA YR (deglycosylated)

T17 (163-189): YSSEFCSTPACSEGNSDCYFGNGRSA YR (non-glycosylated)

Average of 3 measurements

3.3.3 2-chain cleavage

In addition to the variation in glycosylation occupancy, the major difference in the single-chain (intact) and 2-chain (cleaved) forms of the two products was readily seen in the gel, as the follow-on product with the higher amount of single-chain and lower 2-chain forms than the innovator drug. The intact (zymogen) form of t-PA (1-527 chain length) needs to be cleaved between R275 and I276 positions to become an active enzyme, which resulted in two separated chains as the 1-275 and 276-527 forms after the linked disulfides were reduced (see Figure 1). The peptide molecular weight of 276-527
is smaller than the 1-275 form. In addition, the 276-527 form has only one glycosylation site (N448) while the 1-275 has 2 glycosylation sites with fully glycosylated at N103 and partially glycosylated at N184. The glycosylated and nonglycosylated forms at N184 show a difference in molecular weight, with the glycosylated (type I) at a higher molecular than the non-glycosylated form (type II) as indicated at the bottom panel of the gel. Similarly, the single-chain form (1-527) with the extra glycosylation (type I) was observed with a higher molecular than the non-glycosylated form (type II) as indicated at the top panel of the gel. These bands which represent different 2-chain forms (i.e. 1-275 and 276-527) were analyzed by LC-MS to confirm their identities. The representative peptides from the 2-chain and single-chain forms were compared to estimate the relative % of the 2-chain cleavage, measured as 40% for the innovator and 10% for the follow-on product.

In addition, rather than separating the two intact forms by SDS-PAGE, the mixtures of the intact and 2-chain forms could be digested directly by Lys-C (after reduction and alkylation) since the 2-chain form produced the peptide fragment up to R (cleaved between R275 and I276) and the intact form generated the corresponding peptide length extended to IK (Lys-C could not cleave at a R residue position), as shown in Figures 6A and 6B. Thus, the % of 2-chain cleavage could be determined by the ratio of 2-chain peptide divided by the sum of 2-chain and single-chain peptide intensities. Again, the results, as shown in Table 2, were consistent with the results of SDS-PAGE in which the intact protein forms were separated.
Figure 6: Determination of the extent of 2-chain cleavage from the Lys-C digest of TNK-tPA. The extracted ion chromatography of the cleaved (2-chain) and un-cleaved (single-chain) peptides are shown as Figures 6A and 6B, respectively. The % of 2-chain was estimated by the ratio of 2-chain divided by the sum of 2-chain and single-chain peptide intensities, as indicated in the bottom of the figure.
Lys-C digested peptide with and without 2-chain cleavage:

NRRLTWYEYCDVPSCSTCGLRQYSQPQFR (248-275)

NRRLTWYEYCDVPSCSTCGLRQYSQPQFRIK (248-277)

Average of 3 measurements

Table 2 Determination of the extent of 2-chain cleavage (between R275 and I276) of TNK-tPA using either Lys-C digested peptide fragments or in-gel digestion of the resolved single and 2-chain forms.

<table>
<thead>
<tr>
<th>Product / Method</th>
<th>Innovator</th>
<th>Follow-on</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lys-C digestion (% 2-chain form)</td>
<td>40 (±2) %</td>
<td>8 (±1) %</td>
</tr>
<tr>
<td>In-gel Analysis (% 2-chain form)</td>
<td>40 (±5) %</td>
<td>10 (±3) %</td>
</tr>
</tbody>
</table>

Although a significant difference in the amount of 2-chain forms was observed between the innovator and follow-on products, any single-chain should convert into the 2-chain forms on contact with plasmin either with \textit{in vitro} or \textit{in vivo} experiments, and thus, the different levels should cause no variation in the biological activity.\textsuperscript{17} Instead, the observed difference could be used to represent the different manufacturer processes, and
the consistency of the % 2-chain production should be a gauge of the manufacturer’s reproducibility.

3.3.4 Oxidation of methionine residues

Oxidation of methionine residues was determined by the observation of +16 Da in the precursor ion and the MS/MS of the precursor ion was used for site assignment, similar to our previous analysis of recombinant human growth hormone.\textsuperscript{5} Three oxidation sites at Met 207 (T18), Met 445 (T43), and Met 490 (T45) were identified, and the relative % of oxidation (the ratio of nonoxidized and oxidized counterpart) were estimated as shown in Table 3. The relative % of oxidation at these three sites exhibited higher amount in the follow-on as compared to the innovator products.

Oxidation of t-PA with chloramine-T has been shown to abolish ~40% of fibrin binding and clot lysis activity of t-PA.\textsuperscript{18} The specific methionine oxidation sites at the Finger (Met 13) and Kringle II regions (Met 207) of t-PA were shown to weaken the interaction between fibrin and t-PA and thus reduce the activation for the subsequent thrombolytic activity. We did not observe Met 13 but did observe Met 207 oxidation, with a higher amount of Met 207 in the follow-on product which should raise the concern of lower efficacy relative to the innovator product.
Oxidation / Products | M 207 (T18) | M 445 (T46) | M 490 (T48) |
--- | --- | --- | --- |
Innovator | 11 (±1) % | 3 (±0.8) % | 6 (±1) % |
Follow-on | 21 (±2) % | 6 (±1) % | 7 (±1) % |

Oxidized tryptic peptides and sites:

T18 (190-212): GTHSLTESGASCLPWNSMILIGK (Met 207)
T46 (450-462): TVTDNMLCAGDTR (Met 455)
T48 (490-505): MTLVGIISWGLGCGQK (Met 490)

Average of 5 measurements

**Table 3 Determination of the extent of methionine oxidation of TNK-tPA products.**

3.3.5 Deamidation

Similarly, the deamidation of asparagines residues were determined by the observation of +1 Da in the precursor ion and the MS/MS of the precursor ion for site assignment. Two asparagines at N58 (T8) and N 184 (T17) were identified with deamidation. The site at N184 is also a glycosylation site, which should not be deamidated once it is glycosylated. However, this site was partially glycosylated and the non-glycosylated counterpart at this site is prone for deamidation (next to glycine
As described in the section of glycosylation occupancy, the follow-on TNK-tPA, with more of the non-glycosylated counterpart at this site, resulted in more deamidation than the innovator product (see the results in Table 4). Similar to oxidation, the relative % of deamidation was calculated from the deamidated species divided by the sum of the non-deamidated and deamidated species. Although the non-glycosylated form at this site (type II) has higher clot lysis activity than the glycosylated form (type I), the presence of a higher level of deamidation in this region could change the activity. Thus, not only the glycosylation occupancy but also deamidation at this site should be consistent between different manufacturing lots or different manufacturers.

<table>
<thead>
<tr>
<th>Deamidation / Products</th>
<th>N 58 (T8)</th>
<th>N 184 (T17)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Innovator</td>
<td>13 (±2) %</td>
<td>29 (±4) %</td>
</tr>
<tr>
<td>Follow-on</td>
<td>28 (±3) %</td>
<td>48 (±5) %</td>
</tr>
</tbody>
</table>

Deamidated tryptic peptides and sites:

T8 (56-82): CFNGGT(Fu)CQQALYFSDFVCQCPEGFAGK (N58)

T17 (163-189): YSSEFCSTPACSEGNSDCYFGN(non-glycol)GSA(YR) (N184)

Average of 5 measurements

Table 4 Determination of the extent of deamidation of TNK-tPA products.
3.3.6 Disulfide linkages

In our previous report we successfully achieved mapping the connectivity of disulfide linkages using mass spectrometry with electron transfer and collision induced dissociation, in which several disulfide linkages of t-PA were assigned. In this study, the disulfide-linked peptides have the identical sequences in TNK-tPA as in t-PA. Thus, these assignments were straightforward for TNK-tPA on both innovator and follow-on products. The complete mapping of all 17 disulfide linkages has just been reported in a separate paper but based on our identifications so far, both the innovator and follow-on drugs have the same disulfide-linkages.

3.4 Conclusions

The comprehensive characterization and comparison of complex glycoprotein products from innovator and follow-on manufacturers were achieved using LC-MS analysis of the separated intact proteins as well as their enzymatic peptide mixtures. For such a complex molecule, to evaluate the comparability between the innovator and follow-on products, the analysis strategy was focused on regions that could impact the biologic activity. A main concern was the observation of different glycosylation occupancy at N184, which could influence the clot lysis activity.

Although the other modifications such as the deamidation at N184 and oxidation at M207 could also impact the activity, the capability of downstream purification and the formulation process to minimize the deamidation and oxidation should lessen the concern. However, the amount of glycosylation occupancy is not easy to be changed by
the downstream processes or even at the beginning step of cell culture condition. Thus, this occupancy difference raises the question of whether these two products are indeed similar or not. The % chain cleavage at the activation site (R275 and I276) also greatly varied, but the nature of this protein is to convert all remaining single chain to 2-chain rapidly once it is administrated to a patient, and thus the observed difference in this case should be less of a concern.

The quantitative comparison of the % glycosylation occupancy and 2-chain cleavage was performed by two independent approaches, starting from either the peptide or intact protein separation. The quantitation by the peptide approach, which directly digested the mixture of various intact protein forms (type I, II, single, and 2-chain forms), was simple but required the identification and quantitation of the characteristic peptides representing the various forms, e.g., the peptides containing the glycosylation or cleavage sites. On the other hand, the direct protein analysis required the separation of various intact forms, a more complicated procedure than peptide separation. Nevertheless, once protein separation was achieved (using SDS-PAGE in this study), quantitation could be simplified since any peptide fragments from the corresponding gel bands could be used for relative comparison (no need to use the characteristic peptides). Thus, one of these representing peptides (highly reproducible from in-gel and LC-MS analysis) from each of the corresponding gel bands was used for relative comparison. In addition to the complementary (orthogonal) comparison, the intact protein separation (gel image) readily revealed the major differences between the innovator and follow-on products, and that directed our focus for this analysis. Other chromatographic approaches, such as
hydrophobic interaction chromatography, have been shown to effectively separate the various intact forms (type I, II, single, and 2-chain) of rt-PA, and these approaches should be able to be adopted for the separation of TNK-tPA as well.\textsuperscript{20} We used SDS-PAGE initially, since the separation conditions should be more general for many glycoproteins in comparison to chromatographic approaches (rather specific). The detailed characterization and comparison of all glycan structures at each site may not be necessary, since several other major differences have already been defined in this work. Nevertheless, the full characterization and quantitative comparison of the glycan structures will be presented shortly in a separate report, for the purpose of evaluating our capability to differentiate structures using state of the art proteomic techniques. The application of the recent advanced proteomic techniques is proven again to be valuable for the comparison of follow-on or biosimilar products, even for complicated glycoproteins.

### 3.5 References


(13) Wu, S.L.; Hühmer, A.F.; Hao, Z.; Karger, B.L. J Proteome Res. 2007, 6, 4230-44.


Chapter 4

Mass Spectrometric Determination of Disulfide Linkages in Recombinant Therapeutic Proteins using On-line LC-MS with Electron Transfer Dissociation (ETD)

Publication:
Abstract

In the biotechnology industry, the generation of incorrectly folded recombinant proteins, either from an E.coli expression system or from an over expressed CHO cell line (disulfide scrambling), is often a great concern as such incorrectly folded forms may not be completely removed in the final product. Thus, significant efforts have been devoted to map disulfide bonds to ensure drug quality. Similar to ECD, disulfide bond cleavages are preferred over peptide backbone fragmentation in ETD. Thus, an online LC-MS strategy combining collision-induced dissociation (CID-MS2), electron-transfer dissociation (ETDMS2), and CID of an isolated product ion derived from ETD (MS3) has been used to characterize disulfide-linked peptides. Disulfide-linked peptide ions were identified by CID and ETD fragmentation, and the disulfide-dissociated (or partially dissociated) peptide ions were characterized in the subsequent MS3 step. The online LC-MS approach is successfully demonstrated in the characterization of disulfide linkages of recombinant human growth hormone (Nutropin), a therapeutic monoclonal antibody, and tissue plasminogen activator (Activase). The characterization of disulfide-dissociated or partially dissociated peptide ions in the MS3 step is important to assign the disulfide linkages, particularly, for intertwined disulfide bridges and the unexpected disulfide scrambling of tissue plasminogen activator. The disulfide-dissociated peptide ions are shown to be obtained either directly from the ETD fragmentation of the precursors (disulfide-linked peptide ions) or indirectly from the charge-reduced species in the ETD fragmentation of the precursors. The simultaneous observation of disulfide-linked and
disulfide-dissociated peptide ions with high abundance not only provided facile interpretation with high confidence but also simplified the conventional approach for determination of disulfide linkages, which often requires two separate experiments (with and without chemical reduction). The online LC-MS with ETD methodology represents a powerful approach to aid in the characterization of the correct folding of therapeutic proteins.

4.1 Introduction

Recombinant proteins, when expressed in an *Escherichia coli* cell line, initially generate unfolded forms. These unfolded proteins with free cysteines form disulfide bonds during the refolding process in a cell culture medium prior to downstream purification.\(^1,2\) In the early days of biotechnology, the generation of incorrectly folded recombinant proteins from an *E. coli* expression system was of great concern as such forms may not be completely removed in the final product.\(^3-5\) Lately, the expression of secreted proteins (folded proteins with disulfide linkages), such as from Chinese hamster ovary (CHO) cell lines, to produce monoclonal antibodies has lessened the concern of incorrectly folded forms. However, when such proteins are over expressed in the CHO cell line to improve the protein yield, disulfide scrambling is still possible.\(^6\) Thus, significant efforts continue to be devoted to map disulfide bonds to ensure drug quality.\(^7\) Mapping methods generally involve the use of Edman sequencing or mass spectrometry to obtain disulfide-linked peptide information in the first step, followed by determination of disulfide-dissociated peptide sequences after chemical reduction.\(^8,9\) However, the
confidence of assignment can often be limited, particularly when multiple disulfide bonds exist in a protein.

Currently, liquid chromatography coupled online with tandem mass spectrometry (LC-MS) has become the major analytical tool to identify and characterize proteins. 10-14 Collision-induced dissociation (CID) in LC-MS is the most common means of fragmentation to derive polypeptide structure information. 15-18 However, modifications such as disulfide bonds are not typically fragmented by CID. 19 In some cases, the fragmentation by CID in the negative ion mode can lead to the molecular weight of the disulfide-dissociated peptides, however, with limited or no peptide backbone sequence information, in addition to the low ionization efficiency in the negative ion mode. 20-23

An alternative means of fragmentation to CID is electron capture dissociation (ECD), which has been shown to cleave polypeptide ions preferentially at disulfide bonds. 24 Analogous to ECD, a recent paper demonstrated that disulfide bonds in peptides can be broken by electron-transfer dissociation (ETD) using a three-dimensional quadrupole ion trap mass spectrometer with SO2– as the reagent anion, with the disulfide-dissociated peptides being further characterized in a subsequent MS3 step. 25 Currently, ETD in a two-dimensional linear ion trap mass spectrometer using fluoranthene as the reagent anion has been introduced. 26 ETD fragmentation is now commercially available in both twodimensional 27 and three-dimensional ion traps. 28

Recently, we employed a linear ion trap ETD system with CID, ETD, and CID of an isolated charge-reduced species (MS3 or CRCID) for the characterization of proteolytically digested proteins with glycosylation and phosphorylation modifications. 29
The charge-reduced species is likely mainly an ETD fragmented peptide held together by intramolecular noncovalent forces that can be broken apart into an ETD fragmentation pattern (i.e., c and z ions) by addition of kinetic energy. The exact phosphorylation and N-linked glycosylation sites of the epidermal growth factor receptor, and the O-linked glycosylation site of recombinant tissue plasminogen activator, were identified using online LC-MS.29

The purpose of this paper is to apply the above online LC-MS approach with a linear ion trap ETD instrument to determine the disulfide linkages for therapeutic proteins derived from recombinant DNA technology and also to simplify the conventional procedure, which often requires two separate experiments (i.e., with and without chemical reduction). In particular, this method has the potential to determine complicated intertwined disulfide bridges and to identify disulfide scrambling; both are difficult to characterize by conventional methods. In the following, the successful characterization of disulfide linkages of three important biotechnology products, recombinant human growth hormone (Nutropin), a therapeutic monoclonal antibody, and tissue plasminogen activator (Activase), are demonstrated.

4.2 Experimental Procedures

4.2.1 Reagents

Achromobacter protease I (Lys-C) was obtained from Wako Co. (Richmond, VA), and trypsin (sequencing grade) was purchased from Promega (Madison, WI).
Fluoranthene, guanidine hydrochloride, and ammonium bicarbonate were from Sigma-Aldrich (St. Louis, MO). Recombinant human growth hormone (Nutropin), a therapeutic monoclonal antibody, and tissue plasminogen activator (Activase) were obtained through Mass General Hospital (Boston, MA). Formic acid, acetone, and acetonitrile were purchased from Fisher Scientific (Fair Lawn, NJ), and HPLC-grade water, used in all experiments, was from J.T. Baker (Bedford, MA).

4.2.2 Enzymatic Digestion

Protein solution (1 mg/mL) was buffer exchanged with 0.1 M ammonium bicarbonate (pH 8) over a Microcon spin column (10 kDa MWCO; Millipore, Bedford MA). For Lys-C digestion, the protein solution (after buffer exchanged) was added with endoproteinase Lys-C (1:50 w/w) for 4 h at 37°C. For Lys-C plus tryptic digestion, trypsin (1:50 w/w) was added to an aliquot from the Lys-C digestion to digest the protein further for 12 h at room temperature. For tryptic digestion, trypsin (1:50 w/w) was added to the protein solution (after buffer exchanged) at room temperature for 8 h followed by a second addition of trypsin (1:50 w/w) for an additional 12 h at room temperature. In all cases, digestion was stopped by addition of 1% formic acid.

4.2.3 LC-MS

LC-MS experiments were performed on an LTQXL with ETD mass spectrometer (Thermo Fisher Scientific, San Jose, CA), consisting of a linear ion trap with an additional chemical ionization source to generate fluoranthene anions. An Ultimate 3000 nanoLC pump (Dionex, Mountain View, CA) and a self-packed C8 column (Vydac C8,
300-Å pore and 5-µm particle size, 75 µm i.d. × 10 cm) were coupled online to the mass spectrometer through a nanospray ion source (New Objective, Woburn, MA). Mobile phase A was 0.1% formic acid in water, and mobile phase B was 0.1% formic acid in acetonitrile. The gradient consisted of the following: (i) 20 min at 0% B for sample loading; (ii) linear from 0 to 40% B over 40 min; (iii) linear from 40 to 80% B over 10 min; and finally (iv) isocratic at 80% B for 10 min. The flow rate of the column was maintained at 200 nL/min.

Figure 1 shows the general survey scheme of the mass spectrometer, which was operated in the data-dependent mode to switch automatically between MS (scan 1), CID-MS2 (scan 2), ETD-MS2 (scan 3), and CID of isolated species in the MS3 steps (scan 4). Briefly, after a full-scan MS spectrum from m/z 400 to 2000 in the linear ion trap (at a target value of 30 000 ions or a maximum of 100 ms), CID-MS2 (target value of 30 000 ions or a maximum of 200 ms, with 28% normalized collision energy and activation Q at 0.25), and ETD-MS2 (target value of 30000 ions or a maximum of 200 ms), activation scan steps were performed on the same precursor ion. Each precursor ion for the CID and ETD scans was isolated using the data-dependent acquisition mode with a ±2.5 m/z isolation width to select automatically and sequentially a specific ion (starting with the most intense ion) from the first MS scan. Finally, an additional MS3 step, which isolated the highest intensity ion from the prior ETD spectrum for further CID fragmentation (±5 m/z isolation width, target value of 30 000, or a maximum of 200 ms, with 28% normalized collision energy and activation Q at 0.25), was performed.
Figure 1. **Data acquisition scheme with CID-MS\(^2\), ETD-MS\(^2\), and MS\(^3\) steps.** Using an LTQXL MS with ETD, the first survey MS (scan 1) is followed by three consecutive ion activation steps: CID-MS\(^2\) (scan 2), ETD-MS\(^2\) (scan 3), and MS\(^3\) (CRCID or CID-MS\(^3\)) (scan 4). Scans 2, 3, and 4 are repeated two additional times to fragment the second and third highest precursor ion generated from the first MS scan. The total cycle (10 scans) takes ~3 s and is continuously repeated for the entire LC-MS run under data-dependent conditions with dynamic exclusion.

In contrast to our previous paper,\(^{29}\) the normalized collision energy in this work was identical to that of CID-MS\(^2\). The major fragment ions were found to be similar in the MS\(^3\) step using normalized collision energies between 15 and 35%. In our previous paper, we used 10% normalized energy with a decrease of the activation Q value from 0.25 to 0.15 in order to minimize glycan fragmentation as much as possible in the MS\(^3\) step.

Scans 2, 3, and 4 in Figure 1 were repeated in sequence for two additional times for fragmentation of the second and third highest intensity precursor ions from the first scan. The total cycle (10 scans), lasting ~3 s, was continuously repeated for the entire LC-MS run under data-dependent conditions with dynamic exclusion after three repeats of the same precursor ion within 30 s. Moreover, the acquisition scheme of Figure 1 can be programmed to select desired ions either at MS3 or MS4 after the ETD-MS\(^2\) step, if
needed. Frequently, the CID-MS2 step can be eliminated in such follow-up (targeted) runs. The chemical ionization (CI) source parameters for fluoranthene, such as ion optics, filament emission current, anion injection time (anion target value set at $3 \times 10^5$ ions), fluoranthene gas flow, and CI gas flow, were optimized automatically after the procedure for tuning the instrument. The duration time of the ion/ion reaction was maintained constant throughout the experiment at 100 ms. In most cases, the generation of several charge-reduced species with high intensity in the ETD spectrum allowed the determination of the charge state of the precursor ion. The intensity of the charge reduced species could be further enhanced, if necessary (e.g., decreased ion/ion reaction time to 30 ms).

4.2.4 Disulfide Peptide Assignment

A disulfide-linked peptide was assigned by assuming that the cysteine residue was modified with a polypeptide chain. For example, as shown in Figure 2, if two polypeptides, labeled as P1 and P2, were linked by a disulfide bond from the two cysteine residues, one cysteine would initially be assumed to be modified with the molecular weight of the P1 peptide to search against spectra of theoretical fragmentations of the given protein (b and y ions for CID-type fragmentation and c and z ions for ETD-type fragmentation) using Xcorr (g1). The search was then repeated using the other cysteine assumed to be modified with the P2 peptide. Both searches were then combined to assign the cleavages for the disulfide-linked peptide. Final confirmation of the most probable peptide assignment was made by manual inspection of individual spectra with the
preferred fragmentation patterns in the CID-MS2, ETD-MS2, and MS3 spectra, as
detailed in the Results and Discussion section. Any internal cleavages, i.e., simultaneous
cleavages at both the P1 and P2 polypeptides, were assigned manually. If a disulfide-
dissociated peptide ion was isolated from the ETD spectrum for fragmentation, the
spectra generated in this MS3 step were searched against spectra of theoretical
fragmentations of the protein, similar to the CID-MS2 step but with no modification on
cysteine residues. Any unmatched fragment ions in a spectrum, especially relatively high
intensity ions, were manually assigned, taking into account side chain cleavages or
multiple types of fragmentation (e.g., not just b and y but also c and z ions due to
potentially mixed populations). The cleavage sites for peptides with two or more
disulfides (≥4 cysteines with three or more linked peptides) were also assigned manually.
Figure 2. Diagram illustrating the cleavage sites of a disulfide-linked peptide by CID and ETD. The two polypeptides, labeled as P1 and P2, are linked by a disulfide bond from their cysteine residues. The cleavages by CID-MS² or CID-MS³ to produce b and y ions, and by ETD or CID of charge-reduced species (CRCID) to produce c and z ions, as well as disulfide bond cleavages, i.e. -SH and -S*, are indicated in the figure.

4.3. Results and Discussion

In the following, we examine three recombinant proteins, Nutropin, Herceptin, and Activase, by online LC-MS using the strategy shown in Figure 1. Simple (Nutropin), moderate (a monoclonal antibody), and complicated (Activase) disulfide linkages are used to illustrate the effectiveness of the approach.

4.3.1. Data Acquisition Strategies Using the LTQXL with ETD
The various types of fragmentation of a disulfide-linked polypeptide by CID-MS\(^2\), ETD-MS2, and CID in the MS\(^3\) steps are illustrated in Figure 2. As shown, CID mainly cleaves peptide amide bonds (NH-C=O) to produce b and y ions, and ETD fragments NH-C\(\alpha\) bond to produce c and z ions. In addition, ETD can break the disulfide bond to produce two polypeptides, labeled as a free cysteine-containing peptide for one polypeptide (e.g., P1-SH) and a free cysteine containing peptide with an odd electron replacing its proton as the other polypeptide (e.g., P2-S'). As described for the mechanism in ECD and ETD, “H” serves as a donor to the disulfide bond (Cys-S-S-Cys), breaking the bond into a protonated (Cys-SH) and an odd electron (Cys-S') species.\(^{24,25}\) Each disulfide-dissociated polypeptide thus can consist of two populations, either as the Cys-SH (proton transfer) or Cys-S' (electron transfer) forms. Depending on the polypeptide sequence and charge state, one form can be more dominant (stable). In the MS\(^3\) step, CID of the charge-reduced species (i.e., electron-transfer form) will generate an ETD cleavage pattern (c and z ions), and CID of the noncharge-reduced species (i.e., proton-transfer form) will produce typical b and y ions.

4.3.2 Identification of Disulfide-Linked Peptides from Recombinant Human Growth Hormone (Nutropin)

Human growth hormone has been used to treat children with hypopituitarism or growth hormone deficiency.\(^{32}\) Native human growth hormone derived from the pituitary gland consists a single polypeptide (monomer) with two intradisulfide linkages.\(^{33}\)
Figure 3. CID, ETD, and MS$^3$ analysis of the disulfide-linked peptide (Cys182-Cys189, T20-T21) from the tryptic digest of human growth hormone (hGH). (A) CID-MS$^2$ spectrum of the m/z 468.0 (3+) ion. (B) ETD-MS$^2$ spectrum of the m/z 468.0 (3+) ion. (C) MS$^3$ (CID-MS3) scan of the m/z 785.0 ion (from the ETD-MS$^2$ spectrum as indicated by the circle). The peptide sequences with the observed fragment ions are shown in the inset. If the cleavage results from the P2 peptide, a specific P2 label is used in the figure, such as y$^7$(P2). Otherwise, for simplicity, no specific P label (e.g., y7) means cleavage at the P1 peptide. If the product ions contain a disulfide bond or disulfide-dissociated polypeptides (P1 or P2), red labels are used; otherwise, black labels are selected for the product ions with expected cleavages. The product ions with unexpected cleavages (e.g., y ions occurring with typical c and z cleavages for ETD or CRCID fragmentation) are labeled in blue.
Recombinant human growth hormone (Nutropin) was expressed in an E. coli cell line with the identical gene and recovered from the downstream purification process. Since the correct disulfide linkages are critical to assess the recombinant DNA process, the four cysteines linked together as two disulfide bonds of Nutropin are the focus of this work.

Nutropin was digested with trypsin without reduction and then analyzed, as described in Figure 1. As shown in Figure 3, a disulfide-linked peptide ion (m/z 468.0, 3+), was selected from the MS scan for CID-MS$^2$ (Figure 3A), and ETD-MS$^2$ (Figure 3B), and one of the highest intensity ions (m/z 785.0) from the ETDMS$^2$ spectrum was automatically selected for CID-MS$^3$ (Figure 3C).

Using CID (Figure 3A), the disulfide bond was not broken, and only a few characteristic b and y fragmentation ions were observed. On the other hand, with ETD (Figure 3B), the disulfide bond was found to dissociate into two separate peptide ions (P1 and P2), along with a typical ETD fragmentation pattern of the backbone cleavages (c and z ions) with several high-intensity ions consisting of charge-reduced species of the precursor ion ([M +3H]$^{2+}$, [M- NH$_3$ + 3H]$^{2+}$, [M+ 3H]$^{2+}$, [M- NH$_3$ + 3H]$^{2+}$, and [M- 2NH$_3$ + 3H]$^{2+}$). The loss of NH$_3$ (17 Da) from the N-terminus, common in ETD fragmentation, is due to the NH-CR bond at the N-terminus (see Figure 2). Thus, the loss of two NH$_3$ (34 Da) could result from the disulfide-linked precursor ion, which contained both peptides (P1 and P2). The P1 and P2 ions were observed as the highest intensity ions in the ETD spectrum, indicative of preferred cleavage. One of these peptide ions, P2 (circled in Figure 3B), was automatically isolated for further fragmentation in
the MS\(^3\) step (Figure 3C). This peptide ion was found to be fragmented into b and y ions, along with characteristic side-chain losses of amino acid residues, such as loss of 18 (water), 34 (SH\(_2\)), or 46 Da (SCH\(_2\)) from the cysteine residue. These losses could be explained by this peptide containing a mixed population, as P2-SH and P2-S\(^*\), with the protonated form generating b and y ions and the odd-electron (electron transferred) form generating characteristic side-chain losses of SH\(_2\) and SCH\(_2\), along with c and z ions. Similar observations of the side chain losses have been described by others.\(^{25}\) Cysteine-containing peptides can often undergo these types of side-chain losses when they are ionized in the gas phase over a longer period of time, as evident by the observation of increased side-chain losses in the cysteine-containing product ions (e.g., [b6 - SCH\(_2\)] and [b7 - SCH\(_2\)]) in the MS\(^3\) spectrum of Figure 3.

The sequence information of the P2 peptide generated in Figure 3C provided the identification of this peptide without the assumption of a molecular weight modification on the cysteine residue. In a similar manner, the P1 peptide, which was another high-abundant ion in Figure 3B, was next selected for MS\(^3\) and backbone cleavages with a similar fragmentation pattern was generated. In summary, from Figures 3, both the disulfide-linked and disulfide-dissociated peptides were obtained and simultaneously characterized, in contrast to the widely used two-step protocol to obtain the same information (i.e., with and without chemical reduction).
Figure 4. CID, ETD, and MS$^3$ analysis of the disulfide-linked peptide (Cys53-Cys165, T6-T16) from the tryptic digest of hGH. (A) CID-MS$^2$ spectrum of the m/z 941.8 (4+) ion. (B) ETD-MS$^2$ spectrum of the m/z 941.8 (4+) ion. (C) MS3 (CRCID) scan of the m/z 1883.0 ion (from the ETD-MS$^2$ spectrum as indicated by the circle). The peptide sequences with the observed fragment ions are shown in the inset. See Figure 3 for ion label details.

The second disulfide-linked peptide in growth hormone was next examined, as shown in Figure 4. The CID-MS$^2$ spectrum (Figure 4A) of the precursor ion, m/z 941.8 (4+), was observed with a few characteristic fragmentation ions (e.g., y6 and b17 ions at proline residues). The ETD-MS$^2$ spectrum (Figure 4B) of the same precursor ion showed
a typical ETD fragmentation (c and z ions), along with several high intensity ions, including charge-reduced species (labeled as [M + 4H]^{3+} and [M + 4H]^{2+2•}), and side-chain loss ions ([M + 4H - H_2O]^{2+2•}). For clarity, several characteristic side-chain losses are not labeled in Figure 4B. One of the high-intensity charge-reduced species, [M + 4H]^{2+2•}, m/z 1881.8, was automatically isolated for further fragmentation in the MS3 step (Figure 4C). The disulfide dissociated P2 peptide ion along with the backbone cleavage ions (c and z ions) were observed. The P2 ion was further fragmented (MS^4) in an additional LC-MS run to obtain the backbone sequence information (data not shown).

As recently described,^{29-31} charge-reduced species become dominant product ions in the ETD spectrum for precursor ions with m/z >900, as evident for the two disulfide-linked peptides (compare Figure 4B to Figure 3B). The generation of several charge-reduced species with high intensities in the ETD spectrum allowed the determination of the charge state of the precursor ion (4+). As noted, the disulfide-dissociated peptides became the major ions in the MS^3 step. It has been suggested that the two disulfide-linked peptides, even with the disulfide bond dissociated, are still held together by noncovalent forces in the charge-reduced species.^{24,25} As a consequence, the charge-reduced species could be a mixture of two populations, one peptide species that is backbone-dissociated (with the electron already transferred but the disulfide is still intact) and the other a disulfide-dissociated peptide species (either P1-SH or P1-S•, or both). With additional kinetic energy in the MS^3 step, the peptide backbone-dissociated species (with the disulfide still linked) would yield c and z ions, and the disulfide-dissociated peptide (held together by noncovalent forces) would result in two separated polypeptides (i.e., P1 and
P2). The P2 ion is seen in the MS3 spectrum (Figure 4C) while the other peptide ion P1 (m/z 2617 with 1+ charge), not seen in Figure 4C, could appear beyond the mass detection window of this experiment.

The charge-reduced species can also be dissociated using supplemental activation without isolation.\textsuperscript{35-37} In this case, we would anticipate that the two fragmentation steps, ETD-MS\textsuperscript{2} and MS\textsuperscript{3} (CRCID), merge to a single step with product ions from both ETD-MS\textsuperscript{2} and CRCID and with minimal charge-reduced species. While supplemental activation may reduce the instrument cycle time, the observation of product ions in two separate spectra can be useful for simpler interpretation of complicated product ions. Moreover, the MS\textsuperscript{3} step is still required for the fragmentation of P1 or P2 peptides in the analysis of disulfides.

Since the generally larger Lys-C fragments, relative to tryptic fragments, provide more choices of multiple higher charge states for improved ETD/CRCID fragmentation,\textsuperscript{29} growth hormone was next digested with Lys-C to examine the influence of the size and charge states of the disulfide-linked peptides on ETD fragmentation.

There are only two disulfide bonds in rhGH; CID fragmentation alone could be sufficient to identify the disulfide linkages even though an incomplete ion series is produced.\textsuperscript{38} Nevertheless, rhGH is a good model to illustrate the determination of the linkages by the fragmentation strategy of Figure 1. With this determination strategy in mind, a more complicated example, a recombinant monoclonal antibody with multiple disulfide bonds was next selected for analysis.
4.3.3. Identification of Disulfide-Linked Peptides from a Monoclonal Antibody

Figure 5. Structural diagram of a monoclonal antibody with disulfide bonds. The Fab and Fc domains, the light (L) and heavy (H) chains, and the labels of disulfide bonds (\(-S-S-\)) with the location of cysteines in Fc, which consists of two intra (Cys264 with Cys324 and Cys370 with Cys428) and one inter (Cys229 with Cys229 plus Cys232 with Cys232) disulfide linkages, are indicated.

A therapeutic monoclonal antibody, similar to a typical antibody, has constant (Fc) and variable (Fab) domains with inter- and intra-disulfide bonds between the heavy and light chains, as illustrated in Figure 5. The identification of the peptide sequences with disulfide bonds in the Fc region is the focus in this paper.
Figure 6. CID, ETD, MS$^3$, and MS$^4$ analysis of the disulfide-linked peptide (Cys229-Cys229 and Cys232-Cys232 in the hinge region of the Fc) from the Lys-C digest of a therapeutic monoclonal antibody. (A) CID-MS$^2$ spectrum of the m/z 907.8 (6+) ion. (B) ETD-MS$^2$ spectrum of the m/z 907.8 (6+) ion. (C) MS$^3$ (CRCID) scan of the m/z 1814.9 ion (from the ETD-MS$^2$ spectrum as indicated by the circle). (D) MS$^4$ (CID-MS4) scan of the m/z 1361.1 ion (from the MS$^3$ spectrum as indicated by the circle). See Figure 3 for ion label details.

The monoclonal antibody was first digested with Lys-C without reduction. The precursor ion of the disulfide-linked peptide (m/z 907.8, 6+), which was associated through two inter-disulfide bonds between two heavy chains in the hinge region of the Fc domain (Cys229 with Cys229 and Cys232 with Cys232, see Figure 5), was selected for
analysis, and the LC-MS results are shown in Figure 6. As expected, the disulfide bonds were not dissociated by the CID-MS$^2$ fragmentation (Figure 6A), and the characteristic CID cleavages at proline residues were observed with high abundance (i.e., y25, y19, and y5). It should be noted that the two disulfides are very close to each other, and no cleavages were observed for the peptide residues inside these two disulfides even with the presence of proline residues. Similar observations were reported by others as well.$^{39,40}$

The same precursor ion produced a typical ETD spectrum (c and z ions), along with several high-intensity ions, consisting mainly of charge-reduced species of the precursor ion; see Figure 6B. The highest intensity ion ([M + 6H]$^{3+3}$, m/z 1814.5) was automatically fragmented in the MS$^3$ step (Figure 6C). Significantly, while the disulfide-dissociated peptide ions were not detected in the ETD spectrum (Figure 6B), they were the dominant ions in the MS$^3$ spectrum (Figure 6C). The lack of disulfide-dissociated peptide ions (P1 or P2) in the ETD spectrum could be due to the structure of this peptide (two disulfide bonds) with a high m/z (>900). However, fragmentation of the charge-reduced species (CRCID) produced the P1 or P2 peptide ion. To obtain the sequence information, the P1/P2 peptide ion was further fragmented in a CID-MS$^4$ step (an additional LC-MS run) to confirm the correct assignment (Figure 6D). The multi-fragmentation strategy provided good complementary information for this disulfide-linked peptide.
Figure 7. CID, ETD, and MS\textsuperscript{3} analysis of the disulfide-linked peptide (Cys264-Cys324 in the first loop of the Fc) from the Lys-C digest of a therapeutic monoclonal antibody. (A) CID-MS\textsuperscript{2} spectrum of the m/z 960.7 (5+) ion. (B) ETD-MS\textsuperscript{2} spectrum of the m/z 960.7 (5+) ion. (C) MS\textsuperscript{3} (CID-MS\textsuperscript{3}) scan of the m/z 1599.4 ion (from the ETD-MS\textsuperscript{2} spectrum as indicated by the circle). See Figure 3 for ion label details.

Turning to a second disulfide-linked peptide, connected through an intra-disulfide bond (Cys264 and Cys324) in the first loop of the Fc domain (see Figure 5), a precursor ion of (m/z 960.7, 5+) was detected in the same LC-MS run as in Figure 6. Again, the disulfide bond was not dissociated by CID-MS\textsuperscript{2} fragmentation, and only a few high-abundant ions
with characteristic CID cleavages were observed (Figure 7A). For ETD-MS\(^2\), the same precursor ion yielded a typical ETD fragmentation pattern (c and z ions), along with several high-intensity ions, of the charge-reduced species. One of the charge-reduced species ([M + 5H]\(^{3+2}\), m/z 1599.4) was automatically fragmented in the MS\(^3\) step. The disulfide-dissociated peptide ions (P1) were not observed in the ETD spectrum (Figure 7B), but they became the high-abundant ions in the MS3 fragmentation (Figure 7C), similar to that shown in Figure 6C. The lack of any observable P1 or P2 peptide ions in the ETD spectrum could again be due to the sequence structure of this peptide ion with a high m/z (>900). The very small P2 peptide (less than 250 Da) was not in the mass detection window, and the large P1 peptide ion (4547 Da) may have existed as 2+ or 1+ charge state (m/z >2000), which was beyond the mass detection window.

For the remaining disulfide-linked peptide (labeled as Cys370 and Cys428 in the second loop of the Fc domain), NQVSLTCLVK (P1) attached to SRWQEGNVFSCSVMSHELHNHYTQK (P2), a precursor ion with m/z 682.8 (6+), was detected and analyzed as above. Given the high charge and low m/z, the disulfide-dissociated peptides were first identified with high abundance by ETD-MS\(^2\) (P1 and P2), and the sequence information of the dissociated P2 peptide was obtained in the subsequent MS\(^3\) step. Finally, the Fab domain of the antibody (after Lys-C digestion) produced a similar number of the disulfide-linked peptides (2 intra- and 1 inter-disulfide for each light and heavy chain), as for the Fc domain (see Figure 5). Only one disulfide bond was linked to the two peptides, and thus, the assignment was straightforward (data not shown). In summary, the results for the monoclonal antibody further demonstrate that
multiple disulfide-linked peptides can be readily characterized by the multifragmentation steps shown in Figure 1.

4.3.4. Identification of Disulfide-Linked Peptides from Recombinant Tissue Plasminogen Activator (Activase)

We next turn to examine another therapeutic protein (for acute ischemic stroke) with the complicated disulfides, recombinant tissue plasminogen activator (rt-PA), with 17 potential intertwined disulfide bonds that are quite difficult to assign.\textsuperscript{7-9} These disulfide bonds are distributed in 7 tryptic peptides, with 4 of the 7 peptides glycosylated (with N-linked and O-linked glycosylation).\textsuperscript{41,42} Before proceeding, it should be noted that we used Lys-C plus trypsin digestion for this analysis since the Lys-C digestion alone would produce only 2 peptide fragments, one with 2 disulfide bonds and the other with 15 disulfide bonds (>50 kDa), which would be too large to analyze. The use of trypsin alone could not efficiently digest rt-PA since a domain of the protein is known to be resistant to tryptic digestion.\textsuperscript{43}
We first examine peptide A in which there are two disulfide linkages intertwined between three peptide backbones. This peptide with the precursor ion of 610.9 (5+) was fragmented by CID-MS\(^2\) (Figure 8A) and ETD-MS\(^2\) (Figure 8B), and one of the high-abundant product ions from ETD-MS\(^2\) was isolated and further fragmented by CID-MS\(^3\) (Figure 8C). Since there are three peptides linked by disulfide bonds, the assignments are
difficult to determine from the CID-MS$^2$ spectrum. On the other hand, the high-abundant product ions can be assigned in the ETD-MS$^2$ spectrum since the disulfide bond cleavages are the dominant ions (i.e., P1, P2, P3, P1-P2, and P2-P3 ions in Figure 8B). From these partially disulfide-dissociated peptide ions, one can readily assign the linked peptides as P1 with P2, and P2 with P3, but not P1 with P3, see Figure 8B. The exact linkages (i.e., Cys6 with Cys36, and Cys34 with Cys43) required MS$^3$ determination on the partially disulfide-dissociated peptide. In this case, the disulfide bond linked to P3 was broken, but the bond between P1 and P2 remained intact (circled as P1-P2$^{2+}$ in Figure 8B). The isolation of the P1-P2$^{2+}$ ion (m/z 1041.6 in Figure 8B) for further CID fragmentation generated the information that Cys6 was indeed linked to Cys36, and Cys34 was in the free form (see Figure 8C). This information was critical since the ETD-MS$^2$ spectrum did not allow determination of the specific linkages. Similarly, the isolation of the other partially disulfide-dissociated peptide (labeled as P2-P3$^{2+}$ in Figure 8B) generated additional evidence that Cys34 was linked to Cys43, and Cys36 was in the free form.

To our knowledge, this is the first direct evidence to assign these linkages of tissue plasminogen activator by LC-MS, and the result is consistent with the prediction from homology.$^{43,44}$

As can be seen from the above, the isolation of the partially dissociated peptide ions, such as the P1-P2 or P2-P3 peptide ions, followed by CID fragmentation (MS$^3$) was essential to determine the linkage sites. The backbone sequence information generated separately from the P1, P2, or P3 ions (or from the charge-reduced species) was not
sufficient to determine the intertwined disulfide linkages. Importantly, the partially dissociated peptides would be difficult to obtain by a chemical approach (e.g., partial reduction by DTT) but are readily generated with high abundance in the ETD spectrum.

Interestingly, an extra shoulder with a different precursor mass was found on the main chromatographic peak of peptide A. However, the ETD spectrum of this precursor was similar to that of the main peak shown in Figure 8B. For this ETD spectrum, as seen in Figure 9B, the high-abundant ions could be partially assigned to P2 and P3. Based on the theoretical molecular weight of the tryptic peptide fragments of Activase, the linkage with a Cys457-containing tryptic peptide (T43 peptide) was assigned as P1, replacing the T1 peptide containing Cys6. The rest of the fragment ions were consistent with this assignment (Cys457 linked to Cys36). In this example, the initial CID-MS\(^2\) spectrum could not be analyzed; however, based on the results derived from the ETD-MS\(^2\) spectrum, the CID-MS\(^2\) as well as the MS\(^3\) spectra could then be interpreted, reinforcing the assignment made from the ETD spectrum.
Figure 9. CID, ETD, and MS$^3$ analysis of the disulfide-linked peptide (Cys457-Cys36 and Cys34-Cys43) from the Lys-C plus trypsin digest of Activase. (A) CID-MS$^2$ spectrum of the m/z 716.5 (5+) ion. (B) ETD-MS$^2$ spectrum of the m/z 716.5 (5+) ion. (C) MS$^3$ (CID-MS$^3$) scan of the m/z 968.5 ion (from the ETD-MS$^2$ spectrum as indicated by the circle). See Figure 3 for ion label details.

The above disulfide scrambling (i.e., Cys457 linked to Cys36 instead of Cys6 to Cys36) is most likely a consequence of the digestion conditions (see Experimental Section), since higher amounts of the scrambled peptide were found with longer digestion times. In rt-PA, there are 35 cysteines, an odd cysteine without a pairing disulfide (likely
at the Cys83 position based on homology), which could facilitate the disulfide scrambling under the digestion conditions.\textsuperscript{45,46} The influence of digestion conditions on disulfide scrambling is under further study. Nevertheless, the multifragmentation strategy proved successful in revealing the scrambling and identifying the disulfide linkages in this example.

The assignments of the other two nonglycosylated disulfide-linked peptides were achieved using the same strategy as for Figure 8. The multifragmentation results for Peptide C, containing 4 cysteines, Cys307 linked to Cys323 and Cys315 connected to Cys384 are shown in Figure S6 (SI). As seen in the figure, the lack of significant fragmentation between C307 and C325 (inside the circle) indirectly established the disulfide linkages between these two residues. The other potential linkage (i.e., C307 to C315 and C325 to C384) would produce the fragmentation between C315 and C325 in the CID or ETD (or MS\textsuperscript{3}) spectrum. The other disulfide-linked peptides with multiple glycosylation forms are currently being investigated and will be reported later. The assignments of fragment ions with multiple glycosylation and disulfide linkage forms are time-consuming and difficult to ascertain using the LTQ with ETD system. We anticipate that the newly developed ETD with Orbitrap mass spectrometer may facilitate the analysis.\textsuperscript{36}
4.4 Conclusions

To produce therapeutic recombinant proteins in the biotech industry, the disulfide linkages are usually well established through the study of their crystal structures, if possible, in the initial development stage or through the expression of smaller domains of the protein to establish the overall assignment. In rare cases, such as for tissue plasminogen activator, the disulfide linkages are not directly assigned but are based on sequence homology to other protein families. Nevertheless, even if the disulfide linkages are assigned in the initial stage, in later development stages such as for large-scale production or when the cell culturing conditions have been changed, the disulfide linkages often need to be reassigned, particularly when the biological activity has changed. Thus, the confirmation of the disulfide linkages is required, and the method in this paper should provide valid and convenient evidence of confirmation. We can anticipate that this approach can have a major application in the biotech industry, assuring that recombinant protein therapeutics are folded correctly, either with different expression systems or with the same expression system but in multiple production lots.

In this study, the disulfide-linked peptides were found to produce several different types of product ions by the multifragmentation steps (i.e., disulfide bond cleaved peptide, c and z, b and y, and side-chain loss ions), derived from mixed populations of electron-transfer and proton-transfer products. Although some spectra required manual interpretation, our strategy simplified the assignments by first focusing on the highly abundant product ions of disulfide-dissociated or partially dissociated peptide
ions. This initial assignment could identify directly the linkage sites for a peptide with only two cysteines (or one disulfide bond) since there is only one possibility for the linkage. For a peptide with multiple disulfide bonds, the assignments of backbone cleavages are necessary to determine the exact disulfide linkages. The disulfide-dissociated (or partially dissociated) peptides could be obtained directly from the ETD fragmentation of the disulfide-linked peptide precursors. However, when the m/z value of the precursor ion is above roughly 900, the charge-reduced species, rather than the disulfide-dissociated peptides, will likely be the dominant form(s) in the ETD fragmentation. An additional targeted LC-MS run (in the context of data-dependent analyses) will often be needed for further structure elucidation.

For even more complex linkages of disulfides than studied in this paper, ETD with higher resolution mass spectrometry will likely be required (e.g., ETD with Orbitrap or Q-TOF). However, the multifragmentation approach in combination with a high-resolution mass spectrometer may not exclude all potential linkages, particularly when two disulfides are very close to each other. For such cases, the enzymatic digestion efficiency without reduction of disulfides may be low. Excluding these possibilities, the approach in this paper should be able to identify the disulfide linkages in recombinant proteins as long as there is sufficient digested material (e.g., > fmol). The results of the LC-MS approach can be complemented using non mass spectrometric methods, such as point mutation with recombinant DNA technology.

4.5 References


(27) Syka, J. E.; Coon, J. J; Schwartz, J; Starford, G; Hunt, D. F. Proceedings of
The 54th ASMS Conference, Seattle, WA, June 2006.


(39) Loo, J. A.; Edmonds, C. G.; Udseth, H. R.; Smith, R. D. Anal. Chem. 1990,
62, 693–698.


Chapter 5

Identification of the Unpaired Cysteine Status and Complete Mapping of the 17 Disulfides of Recombinant Tissue Plasminogen Activator Using LC-MS with Electron Transfer Dissociation/Collision Induced Dissociation

Publication:

Abstracts

Recombinant tissue plasminogen (rt-PA) with 35 cysteine residues has been completely assigned by mapping the 17 disulfide linkages and the unpaired cysteine. The result is consistent with the prediction from homology except for the unassigned cysteine, which was identified at Cys83. This cysteine was found to be blocked and paired with either a glutathione or cysteine residue in a 60:40 ratio, respectively. The analysis was conducted using a multifragmentation approach consisting of electron transfer dissociation (ETD) and collision induced dissociation (CID), in combination with a multienzyme digestion strategy (Lys-C, trypsin, and Glu-C). The disulfide-linked peptides, even those containing N- or O-linked glycosylation, could be assigned since the disulfide bonds were still preferably cleaved over the glycosidic cleavages under ETD fragmentation. The use of a multiple and sequential enzymatic digestion strategy was important in producing fragment sizes suitable for analysis. For the analysis of complex intertwined disulfides, the use of CID-MS3 to target partially disulfide-dissociated peptides from the ETD fragmentation was necessary for linkage assignment. The ability to identify the exact location and status of the unpaired cysteine (free or blocked with a glutathione or cysteine) could shed light on the activation of rt-PA, upon stimulation by either oxidative or ischemic stress.

5.1. Introduction
The comprehensive characterization of protein therapeutics or protein targets is a continual challenge, given the complexity of these biopolymers.\textsuperscript{1-5} In addition to post-translational modifications, disulfide linkages and the location and status of unpaired cysteines (free or blocked) are critical structural features that need to be determined. Disulfide linkages are a controlling factor in the three-dimensional structure of proteins and are, thus, intimately involved in protein folding.\textsuperscript{6} The free cysteines can be reactive (most often in redox reactions), leading to covalent association,\textsuperscript{7} enzyme catalysis,\textsuperscript{8} or, alternatively, disulfide scrambling.\textsuperscript{9} It is also well-known that unpaired cysteines on a protein may be blocked by the redox “buffer” glutathione, prior to secretion.\textsuperscript{10} In some cases, the reversible reaction of glutathione with cysteine or other oxidizing species can be a mechanism for pathway signaling.\textsuperscript{11} Thus, it is clear that disulfide linkage and unpaired cysteine location and status are important structural features that need to be determined as part of the comprehensive characterization of proteins for structure and function correlation.

The need to analyze disulfide linkages and the status of unpaired cysteines has led to the introduction of a variety of methods, such as Edman degradation\textsuperscript{12,13} and mass spectrometry using collision induced dissociation (CID) in either the positive or negative ion modes.\textsuperscript{14-16} With respect to the analysis of unpaired cysteines, the use of the Ellman reagent to quantitate the free cysteines with tagged fluorophores is often employed.\textsuperscript{17} However, the determination of intertwined disulfide linkages and the location of cysteines, either free or possibly bonded with other molecules, are still difficult to accomplish.
Disulfide bond cleavage has been shown to be favored over peptide backbone breakage for electron capture dissociation (ECD) and more recently for electron transfer dissociation (ETD). This preferred cleavage is likely due to the fact that free electrons can be more easily captured by sulfur (disulfides) than by a backbone (amides) during the electron capture or transfer process. Recently, we have successfully mapped disulfide linkages using online LC-MS with ETD based on the favorable breakage of the disulfide bond. The cleaved or partially cleaved disulfides were further fragmented by CID-MS3 to obtain specific linkage locations.

In this work, we use multiple and sequential enzymatic digestion in conjunction with ETD to study the complex glycoprotein, recombinant tissue plasminogen activator (rt-PA) which has 35 cysteine residues (17 disulfides plus one potential free cysteine). Up to now, the disulfide assignments have been based on homology, and no direct evidence has been presented for the complete structure. Using this LC-MS with ETD approach, not only are all intertwined disulfide linkages determined but also the location of the unpaired cysteine is identified and found to be blocked by either glutathione or an additional cysteine.

5.2. Materials and Methods

5.2.1. Materials

Achromobacter protease I (Lys-C) was obtained from Wako Co. (Richmond, VA), Glu-C from Roche (Indianapolis, IN), and mass spectrometric grade trypsin from Promega (Madison, WI). Fluoranthene, guanidine hydrochloride, and ammonium
bicarbonate were purchased from Sigma-Aldrich (St. Louis, MO). Recombinant tissue plasminogen activator (Activase), consisting of 50 mg of the active protein, was the gift of Genentech Inc. (South San Francisco, CA). Additionally, a genetically engineered t-PA mutant (TNK), which has the same amino acid sequence as the wild type except with substitutions at three locations, T103N, N117Q, and KHRR (296-299) to AAAA,\textsuperscript{22} was also kindly provided by Genentech. Formic acid and acetonitrile were purchased from Thermo Fisher Scientific (Fair Lawn, NJ), and HPLC-grade water, used in all experiments, was from J.T. Baker (Bedford, MA). A Microcon YM-10 centrifugal filter unit was obtained from Millipore (Bedford, MA).

5.2.2. Enzymatic Digestion (without Reduction and Alkylation)

Protein solution (1 \(\mu\)g/\(\mu\)L) was buffer exchanged with 0.1 M ammonium bicarbonate (pH 8) over a Microcon spin column (10 kDa MWCO, Millipore). For Lys-C plus trypsin digestion, the protein solution (after buffer exchange) was added with endoproteinase Lys-C (1:50 w/w) for 8 h at room temperature. Then, trypsin (1:50 w/w) was added to an aliquot from the Lys-C digestion for further reaction for 12 h at room temperature, with the pH adjusted to 6.8. In order to minimize potential disulfide scrambling, a slightly less than alkaline pH was used in the enzymatic digestion protocol. For Lys-C plus trypsin, followed by Glu-C digestion, Glu-C (1:50 w/w) was added to an aliquot from the Lys-C plus trypsin digestion for an additional 8 h at room temperature, with the pH adjusted to 5.8. In all cases, digestion was terminated by addition of 1% formic acid.

5.2.3. LC-MS
LC-MS experiments were performed on an LTQXL with ETD mass spectrometer (Thermo Fisher Scientific, San Jose, CA), consisting of a linear ion trap with an additional chemical ionization source to generate fluoranthene anions. An Ultimate 3000 nanoLC pump (Dionex, Mountain View, CA) and a self packed C8 column (Vydac C8, 300 Å pore and 5 µm particle size, 75 µm i.d. × 10 cm) were coupled online to the mass spectrometer through a nanospray ion source (New Objective, Woburn, MA). Mobile phase A was 0.1% formic acid in water, and mobile phase B was 0.1% formic acid in acetonitrile. The gradient consisted of the following: (i) 20 min at 0% B for sample loading; (ii) linear from 0 to 40% B over 40 min; (iii) linear from 40 to 80% B over 10 min; and finally (iv) isocratic at 80% B for 10 min. The flow rate of the column was maintained at 200 nL/min. In an analogous manner to our recent publication, the mass spectrometer was operated in the data-dependent mode to switch automatically between MS (scan 1), CID-MS$^2$ (scan 2), ETD-MS$^2$ (scan 3), and CID of isolated species in the MS$^3$ steps (scan 4). The normalized collision energy used in CID (scans 2 and 4) was set at 35%. Briefly, after a precursor ion scan, the CID-MS$^2$ and ETD-MS$^2$ activation steps were performed on the same precursor ion. Each precursor ion was isolated using the data dependent acquisition mode with a ±2.5 m/z isolation width to select automatically and sequentially a specific ion (starting with the most intense ion) from the first MS scan. Finally, an additional MS$^3$ step, which isolated the highest intensity ion (or targeted ion) from the prior ETD spectrum for further CID fragmentation at 35% normalized collision energy (±5 m/z isolation width), was implemented. To further confirm the peptide masses and charge states, an LTQ-FT MS (Thermo Fisher Scientific) was employed, as
necessary, to acquire full mass spectra in the FTICR (400-2000 m/z) at 100,000 mass resolution to determine the accurate mass and charge states of the precursor ions generated under similar conditions to those in the LTQXL-ETD MS instrument. If two or more precursor ions with close m/z values appeared at overlapped retention times, the CID-MS\(^2\) spectrum pattern was then added to track the correct m/z precursor ion between the LTQETD and LTQ-FT MS runs.

### 5.3. Results and Discussion

The recombinant tissue plasminogen activator (rt-PA) has 17 disulfide bonds and one unpaired cysteine, with linkages assigned based on sequence homology with other related protein families (e.g., epidermal growth factor and serine protease).\(^{22,23}\) When rt-PA is digested by trypsin (without disulfide reduction), the disulfide bonds are distributed in seven tryptic peptides, with four out of the seven peptides being glycosylated (three N-linked and one O-linked), as shown in Figure 1. The three nonglycosylated peptides have been successfully analyzed in our previous paper using the ETD/CID strategy.\(^{21}\) The remaining four peptides with glycosylation are examined in this study. Initially, Lys-C plus trypsin digestion was employed to improve the digest of rt-PA.\(^{21}\) Additionally, Lys-C plus trypsin, followed by Glu-C, was used to further trim down the size of the peptides with multiple disulfides (e.g., peptides #2 and #3 in Figure 1) to produce sufficient fragmentation cleavages for assignment, as discussed in the following.
<table>
<thead>
<tr>
<th>Peptide</th>
<th>Sequence</th>
<th>Cysteine Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 (1-7)</td>
<td>SYQVIR (Cys6)</td>
<td></td>
</tr>
<tr>
<td>T5 (31-40)</td>
<td>VEGCWNSGR (Cys34, Cys36)</td>
<td></td>
</tr>
<tr>
<td>T6 (41-49)</td>
<td>AQCHSPVPK (Cys43)</td>
<td></td>
</tr>
<tr>
<td>T7 (50-55)</td>
<td>SCSEPR (Cys51)</td>
<td></td>
</tr>
<tr>
<td>T8 (56-82)</td>
<td>CNGGCQQALYSDFVKQCPEGFAGK (Cys56, Cys62, Cys73, Cys75)</td>
<td></td>
</tr>
<tr>
<td>T9 (83-89)</td>
<td>CCEIDTR (Cys83, Cys84)</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1.** Theoretical nonreduced tryptic peptides of rt-PA, including linked disulfides (lines), cysteine position (parentheses), the peptides with glycosylation (cycles), and glycosylation sites (underlines).

### 5.3.1. Identification of Peptide #2 with O-Linked Fucose and Unpaired Cysteine

As indicated in Figure 1, peptide #2 contains the unpaired cysteine along with an O-linked fucose. There has been a significant effort to pinpoint the exact position of the unpaired cysteine which is adjacent to a second cysteine, as well as to determine the status of this cysteine (i.e., whether the thiol is free or blocked). To date, no direct evidence has been presented on the location and state of the unpaired cysteine. This is
likely due in part to the attached O-linked fucose, which results in mainly neutral loss by CID-MS2. In previous work, we found that the fucose was still attached to the peptide backbone after ETD fragmentation, and this result allowed us to identify the fucose linkage position on the peptide (after the protein was reduced).

In the present study with the disulfide intact, we further analyzed the peptide which contained the unpaired cysteine.

Interestingly, we could not find the expected mass of the peptide, assuming the unpaired cysteine was free. Instead, two other masses were found, and both were closely related to each other, based on their fragmentation patterns and difference in precursor masses. One precursor could be matched to the peptide paired with a glutathione and the other with a cysteine residue, as shown in panels A and B of Figure 2. It should be noted the free cysteine, once forming a disulfide (i.e., paired with a glutathione or cysteine), prevents the enzyme (either Lys-C or trypsin) from cleaving the amino acid “K” next to the disulfide. (In Figure 1, the assumption was made that the unpaired cysteine was free, and thus, the lysine was cleaved.) Without digestion of this K, 18 Da (loss of water) must be subtracted from the resulting peptide mass to account for the miscleavage, as shown in panels A (for glutathione blockage) and B (for cysteine blockage) of Figure 2.

The precursor mass of 811.9 (6+) (for the peptide paired with a glutathione residue and with one miscleavage) was analyzed by ETD-MS^2 (panel A of Figure 2). The glutathione residue (P3 ion) was readily observed, along with ions for P1, P2, P1-P2, and P2-P3. The partially disulfide-dissociated peptides (P1-P2 and P2-P3) allowed direct assignment of the linked peptides as P1 associated with P2 and P2 associated with P3 but
not P1 associated with P3. The exact linkages within P2 (i.e., Cys56 with Cys73 and Cys75 with Cys84) typically require CID-MS³ to further target the peptide ion for analysis. However, since the fucose was also located on P2, the additional CID-MS³ step produced, as expected, mainly the neutral loss of the fucose from the P2 ion. Since the P2 peptide ion, after the MS³ step, was still large (due to the miscleavage), the use of CID-MS⁴ to fragment the P2 ion further generated limited cleavages that were insufficient for structure assignment. As a result, we employed an additional Glu-C digestion step to cut the disulfide-linked peptide into smaller fragments. With these smaller fragments, even taking into account the neutral loss of fucose, the limited backbone cleavage for CID-MS² provided sufficient information for structure assignment. From the fragmentation observed, we were able to locate the unpaired cysteine to Cys83 and to determine that it was paired with a glutathione residue.

The other precursor mass of 781.0 (6+), for the peptide paired with a cysteine residue, was assigned in a similar manner. In this case, the cysteine residue (P3 ion) was too small to be observed in the ETD-MS² spectrum (panel B of Figure 2). Nevertheless, the cysteine disulfide (paired with another cysteine containing peptide, P2) was clearly found (P2-P3). In addition, as with the glutathione assignment, the limited backbone cleavage in CID-MS² was also used for linkage determination. Furthermore, identical fragment ions as in panel A of Figure 2 (e.g., P1, P2, and P1-P2), which were independent of the glutathione or cysteine linkage, were observed in the spectrum. In summary, the complementary spectra of these two peptide molecules allowed the assignment of the two
adjacent cysteines in the peptide backbone and differentiated the unpaired cysteine associated with either glutathione or cysteine. After characterization, the same m/z values of the two species (precursors), measured by FTICR MS, were extracted for comparison. The ratio of glutathione to cysteine on the unpaired cysteine was estimated to be 60/40, assuming that both species had similar response factors, given the identical long peptide backbone.

It should be noted that the unpaired cysteine with a free sulfhydryl group is reactive, and the pairing with glutathione or cysteine could be protective. Under oxidative or ischemic (hypoxia) stress, the redox of glutathione or cysteine could possibly trigger the pairing molecule to become free (sulfhydryl group) for protein activation, as found for other proteins in biological systems.25
Figure 2. ETD-MS2 spectrum of the disulfide-linked tryptic peptide with an O-linked fucose and unassigned cysteine paired with either a glutathione (A) or cysteine (B), with the peptide structure and annotation of the fragment ions indicated in the inset.
5.3.2. Identification of Disulfide Linkages of Peptides #4 and #7 with N-Linked Glycosylation

There are three disulfide linked peptides which each contain one N-linked glycosylation, peptides #3, 4, and 7; see Figure 1. In principle, PNGase F can be used to remove the N-linked glycosylation or the nonglycosylated counterpart (if partial glycosylation exists) can be directly analyzed to determine the disulfide linkages. However, without reduction, the deglycosylation step by PNGase F can often be inefficient, and two of the three sites in rt-PA are known to be fully glycosylated. Nevertheless, assuming the mass (including glycosylation) of the peptide is not too large (e.g., <10 kDa), our initial strategy was to try to analyze the disulfide-linked peptides with the glycosylation intact.

Since ETD fragmentation does not normally produce a glycan fragmentation pattern, the CID-MS² fragmentation for glycan structure determination for the glycopeptide can be included in the analysis, as shown in Figure 3 for peptide #7. The precursor ion at 906.4 (5+) was fragmented by CID-MS² (top panel), ETD-MS² (middle panel), and CID-MS³ (bottom panel). From the glycosidic bond cleavages in the CID-MS² along with the precursor mass, a biantennary glycan with a terminal sialic acid was assigned, in agreement with our previous results.²⁴ In the ETD-MS² spectrum, the favorable disulfide bond cleavages were readily seen (i.e., P1 and P2 ions), along with charge-reduced species and peptide backbone fragmented ions (c and z). One of the charge-reduced
species ([M + 3H]^{3+}, m/z 1509.6) was isolated and automatically fragmented in the MS$^3$ step, which further confirmed the disulfide linkage assignments from the observation of P1 and P2 ions (bottom panel). Similarly, for the other N-linked glycopeptide which also contained only one disulfide bond linking the two peptides (peptide # 4 in Figure 1), the disulfide assignment was straightforward. In both cases, the disulfide bonds were preferably cleaved by ETD even though the peptides contained different types of glycosylation, i.e., the complex type (Figure 3) or the high-mannose type.

In the above analysis, the glycan structure determinations took advantage of the fact that the glycans on rt-PA have been previously studied.$^{26}$ In the case of unknown glycan structures, the use of PNGase F could simplify the determination of the disulfide assignment, even if the deglycosylation step was not efficient. It is interesting to note that we recently showed that glycan removal by PNGase F can be facilitated using pressure cycling technology,$^{27}$ and this approach could potentially aid deglycosylation to simplify disulfide linkage assignments of glycopeptides.
5.3.3. Identification of Disulfide Linkages of Peptide #3: A Large Glycopeptide

We next examined the remaining disulfide peptide, a glycopeptide with seven disulfides (peptide #3 in Figure 1). This peptide was difficult to analyze because of the large mass, even ignoring glycosylation (~14 kDa). The inefficient fragmentation and wide distribution of charge states for the fragment ions created complexity in the
analysis. As a result, we again employed an additional Glu-C digestion step to cut the peptide into smaller pieces. However, as was the case for trypsin digestion, Glu-C could not digest the glutamic acid (E) adjacent to the disulfide (Cys 409 with Cys 484 in Figure 1). As a result, only two peptide fragments were generated, as shown in Figure 4, one with two disulfides (~2.7 kDa) and the other with five disulfides (~9.5 kDa).

**Figure 4. Structure of the disulfide-linked peptides from a multidigestion of rt-PA (Lys-C plus trypsin plus Glu-C digestion).** The breakage of the peptide with seven disulfides to two fragments is shown: one with two disulfides labeled as peptide A, and the other with five disulfides labeled as peptide B with the glycosylation site (underlined).

The peptide containing the two disulfides (peptide A), m/z of 669.4 (4+), was fragmented by CID-MS² (top panel) and ETD-MS² (bottom panel), as shown in Figure 5. As previously shown, the partially disulfide-dissociated peptides, P1-P3 and P2-P3 in

<table>
<thead>
<tr>
<th>T10(90-94): ATCYE</th>
<th>Peptide A</th>
</tr>
</thead>
<tbody>
<tr>
<td>T13(136-145): LGLGNHNYCR</td>
<td></td>
</tr>
<tr>
<td>T17(167-175): FCSTPACSE</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T17(176-189): GNSDCYFGNGSAYR</th>
<th>Peptide B</th>
</tr>
</thead>
<tbody>
<tr>
<td>T20(229-233): HNYCR</td>
<td></td>
</tr>
<tr>
<td>T24(255-267): YCDVPCSTCGLR</td>
<td></td>
</tr>
<tr>
<td>T37(393-410): TVCPLPADLQLPDWTECE</td>
<td></td>
</tr>
<tr>
<td>T44(463-489): SGGPQANLHDACQGDGGGPLVCNDGR</td>
<td></td>
</tr>
<tr>
<td>T45(490-505): MTLVGIIWGLGCQK</td>
<td></td>
</tr>
</tbody>
</table>

As previously shown, the partially disulfide-dissociated peptides, P1-P3 and P2-P3 in
ETD-MS², could readily be assigned to the linked peptides as P1 with P3 and P2 with P3. The exact linkages (Cys 92 with Cys 173 and Cys 144 with Cys 168) were confirmed by specific fragment ions found in CID-MS² (i.e., y5 and b4). In this case, the additional confirmation by CID-MS³ was, thus, not required.

Figure 5. Analysis of the disulfide-linked peptide A (two disulfides) from Figure 4 by CID-MS2 (A) and ETD-MS2 (B), with the peptide structure and annotation of the fragment ions indicated in the inset.
The other peptide with five disulfides and an N-glycosylation site (peptide B) was more complex. Since it is known that the glycosylation occupancy at this site is only 50%,\textsuperscript{26} the nonglycosylated peptide was chosen for analysis. The peptide precursor m/z, 904.5 (11+), was fragmented by ETD-MS\textsuperscript{2}, as shown in Figure 6. The partially disulfide-dissociated peptides (P1-P2-P3, P2-P3-P4, P3-P4-P5, P4-P5-P6, P3-P4-P5-P6, and P2-P3-P4-P5-P6) could be assigned to the linked peptides, as indicated in the figure. Since many partially linked disulfides were found, the possibilities of other linkages were, thus, eliminated. It should be noted that we also observed several additional cleavages for this peptide (e.g., further cleaved at D after Glu-C digestion), which could result in the same mass with the addition of water (if the cleaved peptides could still be linked by the disulfides) or additional molecular weight loss (if the cleaved peptides could not be held together by the disulfides). In this case, the assignment of the combination of several miscleavages was difficult, initially by matching the precursor masses. The CID fragmentation of these precursor ions was not informative because many fragment ions contained intertwined disulfides still linked together. Nevertheless, the ETD fragmentation provided the clue for assignment since the fragment ions for the nonmiscleaved portion were identical, and also, these partially cleaved disulfides from the ETD spectrum could be fragmented by CID-MS\textsuperscript{3} for further confirmation. Thus, as demonstrated here, the analysis of multiple intertwined disulfides in a peptide requires the combination of multiple enzymes with sequential digestion, coupled with ETD-MS\textsuperscript{2} and targeted CID-MS\textsuperscript{3}. 
Figure 6. ETD-MS² spectrum of the disulfide-linked peptide B (five disulfides) from Figure 4, with the peptide structure and annotation of the fragment ions indicated in the inset.

5.3.4. Analysis of TNK-tPA

After the complete characterization of disulfides in rt-PA, we also examined the disulfide linkages and the status of the free cysteine in the rt-PA mutant (TNK-tPA or Tenecteplase), which is also approved for the same indications as rt-PA (myocardial infarction and acute ischemic stroke). TNK-tPA has been shown to have a longer half-life
and higher fibrin binding specificity than rt-PA because of the difference in glycosylation patterns for the two therapeutics.\textsuperscript{22,23} Although TNK-tPA has amino acid substitutions at three sites (see Materials and Methods), the amino acid sequences remain identical for all disulfide-linked peptides for both forms of t-PA. Because the masses of the disulfide-linked peptides are identical, the determination of TNK-tPA disulfides was straightforward by comparing the masses to those found for rt-PA. Importantly, we again found that the free sulfhydryl group for TNK-tPA was fully blocked by glutathione and cysteine at the similar 60/40 ratio.

It is important to note that rt-PA and TNK-tPA were obtained from different clones and, thus, from different cell cultures and downstream manufacturing conditions. However, neither the glycosylation (caused by the point mutation) nor the manufacturing conditions appear to influence the disulfide formation; rather, the amino acid sequences in close proximity to the cysteine residues appear to dictate disulfide formation during protein folding. Thus, TNK-tPA is a well engineered mutant, where it is only mutated sufficiently to change the outer surface glycans while not influencing the folding disulfides, even at the location of the unpaired cysteine (which may influence activation upon oxidative or ischemic stress).

5.4. Conclusions

The 17 disulfides of rt-PA were completely assigned, for the first time, by the combination of multiple and sequential enzymatic digestion along with the CID/ETD and CID-MS\textsuperscript{3}. The result is consistent with homology predictions. Additionally, the
unassigned (unpaired) cysteine, identified at position 83, was found to be paired with a glutathione or cysteine molecule. As we have noted, the identification of the location and status of the unpaired cysteine should enable the study of the activation or signaling of rt-PA upon stimulation by oxidative or hypoxia stress, such as for patients under ischemic stroke (hypoxia) or myocardio-infarction bleeding (oxygen exposure).

In the production of therapeutic recombinant proteins, the analysis demonstrated in this paper should provide important confirmatory evidence that the protein is folded properly. In continuation of this work, we are at present developing with colleagues a software tool to simplify disulfide assignments and also to guide users for predicting critical cleavages for confirmation.

5.5. References

211–235.


