Mass spectrometric methods for analysis of therapeutic products and advanced proteomic analysis

by Di Wu

B.Sc. in Applied Chemistry, Beijing technology and Business University
M.Sc. in Chemistry, University of Massachusetts Dartmouth

A dissertation submitted to

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

March 29th 2018

Dissertation directed by

Dr. William Hancock
Professor of Department of Chemistry and Chemical Biology

Acknowledgments

The journey to complete my Ph.D. study is a hard but an enjoyable time due to many people who always support me during this five years. It's impossible for me to achieve the milestones without their encouragement. I would like to thank everyone who helped with my Ph.D. studies.

Foremost, I would like to express my sincere gratitude to my advisor Prof. Hancock for the continuous support of my Ph.D. study and research from the first day I joined the group, for his patience, motivation, enthusiasm, and immense knowledge. His profession guidance and life philosophy helped me in all the time of research and my life. I could not have imagined having a better advisor and mentor for my Ph.D. study. His words will always inspire me in my future career.

I would like to thank Dr. Shiaw-lin (Billy) Wu for his mentoring and instructing with my research details. His academic and industrial experience and advice made me learn a lot. I would like to thank Dr. Jeff Agar for his kindness help and training at the beginning of my Ph.D. study. His passion and attitude to research encouraged me all the time. I would like to thank Dr. Jared Auclair and Dr. Alexander Ivanov for their help with my projects and lab management. I would like to thank all my committee members Dr. George O'Doherty, Dr. Jared Auclair, Dr. Jeff Agar for their help and suggestions about my projects and thesis work.

I would like to thank my collaborators, the Love Lab from MIT, Koch Institute of Integrative Cancer Research for this great opportunity on the Biomedicine on Demand project. I would like to thank the funding support from Defense Advanced Research Projects Agency (DARPA) and Bill and Melinda Gates Foundation.

I would like to thank the staff in the Department of Chemistry and Chemical Biology, and the Barnett Institute, for their help during my study in Northeastern University. I would like to thank my colleagues in

Hancock Lab, and colleagues in Agar lab, for their help during my study. I would like to thank all my dear friends in Northeastern University, Yu Wang, Yanjun Liu, Catie Rawlins, Zhidan Chen, Daniel Donnelly, Jeniffer Quijada, Yuanwei Gao, Shanshan Liu, Wanlu Qu and Xiaofan Liu, etc. for their support and help in my study and life. I'm so grateful for time spent with my friends during these years.

Nobody has been more important to me in the pursuit of this degree than the members of my family. I would like to thank my parents for their unconditional love and support on my life and study. And most of all for my loving, encouraging, supportive, and patient husband Boyan Lin whose faithful support during the final stages of this Ph.D. is so appreciated. Thank you.

Abstract of Dissertation

The biopharmaceutical industry has played an important role in human lives in both academia and industry. Biopharmaceutical drugs have much higher development costs and complexity but with studies of basic science and biological, genomic research, there is continual progress in these important fields which have great potential to create commercial value in this market place. On the other hand, the basic sciences of manufacturing and analytics also are attracting many research efforts. Other key issues of concern include the manufacture of vaccines, and safe transportation of the drug product under refrigeration. It has become a global challenge to efficiently produce high quality drugs in remote areas or under emergency situations. In this dissertation, the team headed up by MIT, designed a novel manufacturing platform and have applied novel yeast (*Pichia*) genomics to rapidly produce high quality biopharmaceutics which are more readily purified due to less host cell contamination. The InSCyT (Integrated and Scalable Cyto Technologies) for Flexible microbial manufacturing platform is both portable and convenient to produce biologics in remote locations.

In Chapter 2, we described the characterization of purified recombinant drugs from this novel system with full sequence coverage, ranging from the primary structure identification to possible post translational modifications by sensitive liquid chromatography coupled with tandem mass spectrometry (LC-MS). With the analysis of any product degradations that can occur during the fermentation or purification steps or during product storage, we observed that the manufactured biopharmaceuticals are at a high-quality level and comparable with the innovator drugs. The quality analysis was used to guide the upstream team to optimize their manufacturing protocol. We have built the resulting methodology for product identification into a robust approach for quality control and demonstrated that this final manufacturing platform is efficient, convenient and consistent in terms of product manufacture.

In Chapter 3, we provided the application of the InSCyT platform on to the production of the important biopharmaceutical, namely glargine an insulin analog. We also describe the characterization of the product with both bottom-up and top-down methods and the results from these various methods are comparable and consistent. The results of the analysis demonstrated that the products were prepared in a good quality in the initial upstream product. The characterization of glargine required a variety of techniques, including MALDI-TOF, Q-TOF, and LC-MS and we found that the products were secreted with full sequence coverage and with few modifications. In the future these analytical technologies will assist fermentation development and improve the subsequent production by a more robust system.

In Chapter 4, we demonstrated several analytical methods and techniques to characterize the protein therapeutics after incubation under stressed conditions. The resulting degradations from the accelerated stability study were monitored with LC-MS-MS. In addition, real time Raman spectroscopy techniques and electrokinetic measurements were utilized in this study and demonstrated a powerful ability to monitor degradation reactions, as well as determine the identity of a given biosimilar therapeutic product and any product variants. The study indicates these techniques are sensitive and robust and will provide an alternative method to conventional analytical methods for the monitoring of the manufacture of protein therapeutics. Such methods also have the potential to provide quality assessment of protein therapeutics manufactured in remote locations.

In Chapter 5, we described a proteomics study of breast cancer biomarkers with 1D gel and enzyme digestion coupled with LC-MS. The early detection of disease is always an important research area for the diagnosis and effective treatment of patients. In this study, we have characterized four cell lines including the secretome of cancer non-aggressive and aggressive cell lines using appropriate conditional medium to obtain information on potential biomarkers. We have applied HPLC tandem MS techniques and a proteomic strategy to characterize the cell lines. With a serial data analysis using information listed in data bases of

GeneCards VarElect, we have successfully curated a proteomic protein list. The evaluation of biomarker candidates derived from the searching of proteomic databases was then processed for bioactivity measurement, and as a promising lead we have found significant difference in expression of fibronectin in the aggressive cell lines. Future studies using RNA-sequence analysis, signal pathway analysis and bioactivity will be explored to confirm the disease associations of this biomarker.

In Chapter 6, we summarized and concluded our findings and provided a future view of the global health biopharmaceutical manufacturing and production studies.

Table of Contents

List of Figures

List of Tables

List of Acronyms

aa amino acid

ABC amnion bicarbonate

ACN acetonitrile

Asn asparagine

Asp aspartic acid

BRCA2 breast cancer type 2 susceptibility protein

CE capillary electrophoresis

CHO Chinese Hamster Ovary

C-HPP chromosome-centric human proteome project

CID collision induced dissociation

Cys cysteine

Da Dalton

DARPA Defense Advanced Research Projects Agency

DTT Dithiothreitol

EGFR epidermal growth factor receptor

ESI electrospray ionization

ETD electron transfer dissociation

FDA Food and Drug Administration

FT-ICR Fourier-transform ion cyclotron resonance

GCSF Granulocyte-colony stimulating factor

HEK human embryo kidney

hGH Human growth hormone

HPLC High Performance Liquid Chromatography

IAA Iodoacetic acid

IE ion exchange

IEF isoelectric focusing

IgG immunoglobulin G

IsoAsp isoapartic acid

kDa kilodalton

LC liquid chromatography

LIT or LTQ linear ion trap

mAb monoclonal antibody

MALDI matrix-assisted laser desorption/ionization

Met methionine

mg milligram

MS mass spectrometry

MS/MS tandem mass spectrometry

m/z mass to charge ratio

PD pharmacodynamics

pI isoeletric point

PK pharmacokinetics

ppm part-per million

Pro proline

PTM post-translational modification

Q quadrupole

QE Quadrupole Exactive

Q-TOF quadrupole time-of flight

RP reverse phase

SDS-PAGE sodium dodecyl sulfate-polyacrylamide gel electrophoresis

SE size exclusion

SRM selected reaction monitoring

Tiam1 T-lymphoma invasion and metastasis-inducing protein 1

TP53 DNA damage response/usually called tumor suppressor

Tris 2-amino-2hydroxymethyl-propan-1,3-diol

XCorr cross-correlation

Chapter 1

Overview of the development and the analytical qualification of biopharmaceutical products

1.1 Abstract

The biopharmaceutical industry has played an important role in human lives in both academia and industry. Biopharmaceutical drugs have much higher development costs and complexity but with studies of basic science and biological, genomic research, there is continual progress in these important fields with potential to create commercial value in this market place. On the other hand, the basic sciences of manufacturing and analytics also are attracting many research efforts. Other key issues of concern include the manufacture of the vaccines, and safe transportation of the drug product under refrigeration.

This chapter describs the development of protein therapeutics from the cultivation and the quality control aspects and an overview of the background. Other important topics include the importance of global health and proteomics to characterize disease biomarkers. Also introduced is the significance of analytical developments in the biopharmaceutical arena and the overview introduces the nature of protein modifications, which are crucial for drug structure and functions. In the last part, this chapter briefly describes the techniques and applications of liquid chromatography coupled with mass spectrometry.

1.2 Global health and biosimilars

Global health is always a topic for scientists and socialists to explore due to the rapid development of medical and chemical technologies.[1] Decades ago, scientists discussed the arrival of a new generation of health treatments especially from the 1990s.[2, 3] Lancet has recently reported on health issues and opportunities in 1993 to 2035 from national and international viewpoints and from the financial viewpoint.[4] Global burden of diseases was estimated. The mortality and burden of disease by region and cause were serious issues of WHO's concern. The deaths causes from tuberculosis, malaria, and AIDS; that is the communicable

diseases, then maternal, perinatal and nutritional health; the deaths from non-communicable disease; and the deaths caused by accident injury are all taken into consideration. For the baseline scenario, the trends in total deaths from 2002 to 2030 are summarized in Fig.1.1 by WHO scientists.[5] Global deaths are increasing in the coming years, not only for the aging population reason, but the increase in aging related mortalities. The biopharmaceutical area is now prevailing the whole world. China and India are remarkable mew emerging markets.[6, 7] The development of the biopharmaceutical area is regardless of the economic development of a country, but has become a worldwide system for solving the health problem effectively.[8] As far as we've seen, with the dramatic gains in global health in these years and with an optimistic trend, scientific advances will continue into the future if accompanied with a matched growth in technology and industry policies.

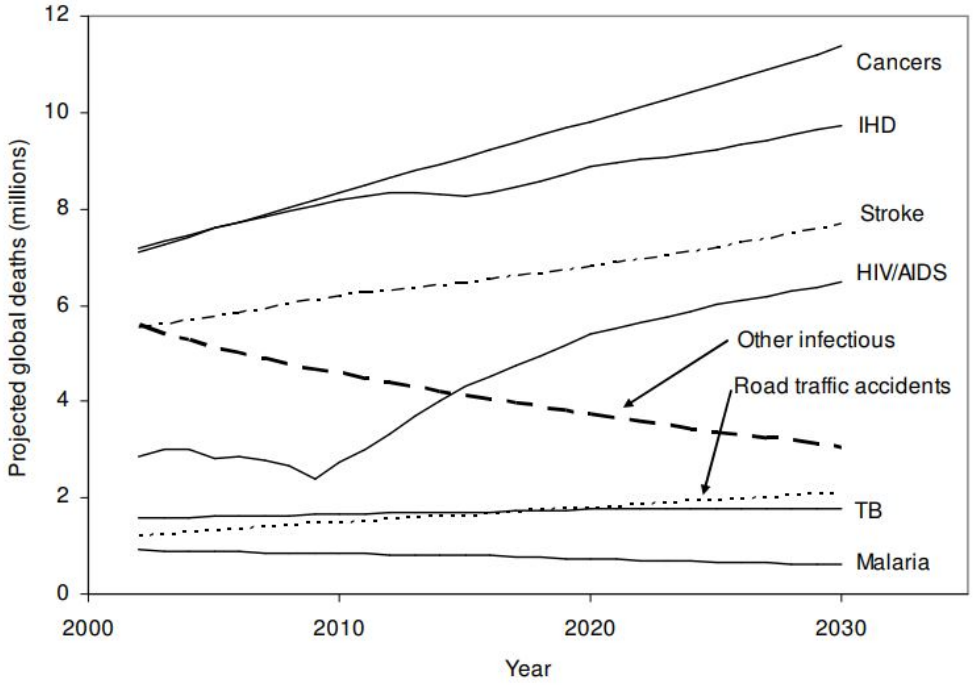

Fig 1.1: Projected trends in total deaths for selected causes, baseline scenario, world, 2002-2030

In 2015, the government has addressed the topic of Precision Medicine[9] and encouraged research on pro-

teomics, metabolomics, and genomics to guide clinical practice. Oncology prevention, diagnosis and other effective treatments by precision medicine are expected to provide better understanding of the disease process. Therefore, many novel therapies will be expected to obtain benefits from this program. The biopharmaceutical industry is a growth market, with both high financial impact and with reduction of health issues.

From 1996, Bill and Melinda Gates Foundation has contributed millions of dollars to investigate the supply of medicines and health education worldwide which has resulted in the production of new vaccines, including monoclonal antibodies. From the year 1982, with the approval of the first biopharmaceutical, insulin, research including biology, chemistry, analytical sciences that are related to biosimilars has accumulated. The reduced development expense of biosimilars from both time and cost has appealed to the industry in recent decades.[10, 11] The biosimilar drug definition quoted from the FDA 351(k) is "A biosimilar product is a biological product that is approved based on showing that it is highly similar to an FDA-approved biological product, known as a reference product, and has no clinically meaningful differences in terms of safety and effectiveness from the reference product. Only minor differences in clinically inactive components are allowable in biosimilar products."[12, 13]

The definition of biosimilars has varied in different countries and regions,[14, 15] as indicated in Table.1.1. The drug analogs are referred to in different terms, which are described as similar biotherapeutic product (SBP) by WHO, subsequent-entry biologics by Canada, follow-on biologics (FOB) by US FDA, and only called biosimilars in Korea. Even the definition are not interpreted in a uniform manner, however, the important common disciplines are centered around the demonstration of safety and efficiency.

Tab 1.1: The definition of biosimilar products in the world

Term	By	Definition
SBP	WHO	A biotherapeutic product similar to an already licensed reference biotherapeutic product in terms of quality, safety and efficacy
SEB	Canada	A biologic drug that enters the market subsequent to a version previously authorized in Canada with demonstrated similarity to a reference biologic drug
FOB	US FDA	A product highly similar to the reference product without clinically meaningful differences in safety, purity and potency
Biosimilar	Korea	Biological products which demonstrated its equivalence to an already approved reference product with regard to quality, safety, and efficacy

The identification and quality control of the biotherapeutics requires a long and complex procedure. Generally speaking, as long as the amino acid sequence and 3D structure of active region of the biosimilar drug is same with the original patented medicine, the biosimilar would not require extensive clinical trials for different indications.[16, 17, 18] However, the investigation process for a new drug is still a long and expensive way to achieve.[19] For example, protein activity is often controlled by amino acid modifications which may effect structural stability and thus activity of the drug. Millions of dollars have been invested into the biopharmaceutical medical industry, but according to the safety, stability, toxicity, and other concerns, the approvals of innovative biopharmaceuticals are not very frequent. The FDAs Center for Drug Evaluation and Research (CDER), has approved a number of cancer related therapeutics in 2015 and recent years, as shown in Fig1.2 and in fact oncology drugs have taken one third of the total approvals.

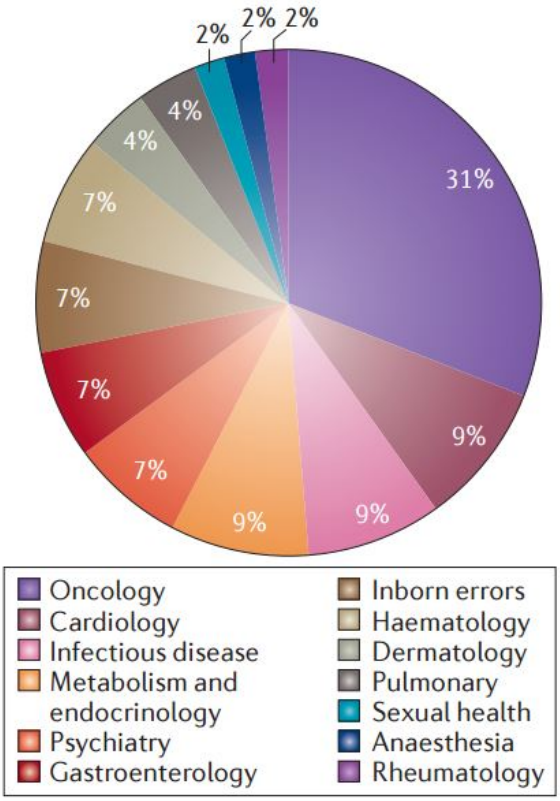

Fig 1.2: Drug approvals by therapeutics

Among the past 15 years,[20] FDA[21] and EMA[22] have granted 76 unique pharmaceuticals approval without the randomized control studies. The EMA has approved the first antibody in 2013,[23] which has become a milestone in this area. To approve biopharmaceuticals, there are many controversial opinions around, especially focusing on clinical treatment studies. Randomized controlled trails are sometimes unnecessary for biosimilar pharmaceuticals due to a high degree of structural confidence in these drugs, however, the quality of the manufacturing process and the resulting of biopharmaceuticals still remain important issues. Hence, improvements in manufacturing and quality control testing are crucial to the field. The product safety is always the most important factor to consider when producing biosimilars.[24, 25] As an example, the efficacy

and safety of many biosimilar and clinical trails have been studied,[26] such as infliximab biosimilar.[27, 28] While the industry and market for biosimilars is a worldwide field the use of definitions and guidelines have differences in different regions. Wang[14] has reviewed the regulatory approval procedures of biosimilar products by regions. The main guidelines are based on EMA and WHO, while at the same time scientists in Japan have reviewed their industry view of biosimilar development,[29] and the regulation process in America is always important.[30]

1.3 Biosimilar manufacture and delivery

The recombinant DNA technique that is used to produce a biosimilar was created in 1982,[31, 32] and during the past decades, the technologies have developed and advanced significantly. Based on special medical needs in emergency situations, DARPA has launched a program for medical treatments in remote areas especially focusing on the battlefield, which aims to provide a spontaneous cure based on the patients need. Although there are ways to produce therapeutics on small scale in the laboratories, the needs of a larger scale in short time with high quality are not easy as we can imagine.[10, 16] How the drug could be produced within one week or even overnight became a serious issue to save lives in the battlefield. There are many issues to manufacture vaccines and drugs in remote areas based on the traditional highly resourced pharmaceutical manufacturing process. Under stressed manufacturing conditions, how can one develop a process to efficiently produce a biosimilar by a suitable cell line and with high purity and safety and with suitable storage conditions that protect the drug? Clearly, there remain a lot of questions.

Besides the quality issues, the scale of the fermentation reactors is an important issue and portable small scale reactors are trending the research studies. For example, some institutes are making efforts on the development of portable devices to produce and evaluate biopharmaceuticals to more efficiently research

new medical treatments. Rao's group[33, 34] has worked on innovating a portable fluorescence sensing bioreactor for many years with documentation in patents and publications,which focused on the measurement of amount and identification of biopharmaceuticals in relation to their bio-process techniques. However, while fluorescence is an effective technique it is not discriminating with large molecules such as proteins or antibodies. A number of novel techniques were applied in a related research program funded by DARPA, where the Love labs at MIT assembled a team to investigate a platform named InSCyT that focused on on-demand drug production in an efficient time period. The research aim of the team was to create a whole integrated platform to produce, purify, evaluate and release the biosimilar product for use at remote location. The DARPA mission is extremely challenging with the complexity of the biosimilar drug, the need for thorough characterization and the avoidance of safety issues.[35] This challenging goal of assuring quality and safety of such new drugs, has been mitigated by the contributions of scientists for over a century, with novel technologies, clinical trial statistics, regulations, and other aspects. In our modern academic atmosphere and society, scientists already have contributed a lot.

The traditional manufacturing fermentation process for a recombinant DNA technique requires a long time to produce the final products where, the expression system mostly used mammalian (such as CHO) and *E.coli* cell lines. These methods have been well studied and there are numerous researches on novel expression technologies traditional cell cultivation approaches. However, there are drawbacks with these techniques, for example with *E.coli* production, in spite of an economical and timesaving fermentation process, the cellular secretory process may not be fully understand which can result in inefficient generation of a large protein.[36] On the other hand, the most crucial aspect is a possibility of serious contamination with adventitious agents (endotoxins) and other host cell related impurities. With these considerations, the MIT team decided not explore more opportunities with the commonly used cell lines for innovations and development of the familiar upstream and downstream processes.

In recent decades, increasing studies have focused on exploring alternative expression systems, and the FDA has approved some novel products using alternative cultivation systems.[37, 38, 39] The *P. pastoris* produced Kalbitor has been approved by FDA because of the high tolerance in a condensed fermentation environment compared with the similar and traditional *S. cerevisiae*. The yeast expression system has advantages over *E.Coli* on purification and can result in isolation procedures, and the sufficient production in large proteins as well. With the InSCyT platform, research focused on using yeast production, *Pichia*, which is easier to control in a fermenter, gave good yields of secreted bioactive product with minimal contamination of host cell protein. The Love lab produced different strains of *Pichia* including *K. pastoris*, *K. phaffii* wildtype, and *K. phaffii* GS115 and compared the proteins produced in the three strains.[40, 41] During the manufacturing of protein drugs, the engineered modifications of the process such as changes in the amino acid sequence and undesired processing modifications such as chemical or physical degradations are all monitored in our quality analysis.

1.4 InSCyT platform

The Love lab[42] at MIT has developed an effective manufacturing platform named Integrated and Scalable Cyto-Technologies for flexible microbial manufacturing (InSCyT) for the Biologically-derived Medicines On Demand (BioMOD) program, aiming to produce high quality biopharmaceutics on demand in a short time period. The overview of the manufacturing process is shown in Fig1.3, containing upstream production and fermentation process, and downstream purification, polishing and ultrafiltration process. This platform also applied both on-line and off-line analytics to assure the quality. The cultivation fermentation only consumes less than 24 hours for one dose of therapeutic protein in 1.5L in volume. With further innovation, the team can now produce the therapeutic target in a fermentation system that contains three tanks to produce

within one day triplicates doses for emergency situations.

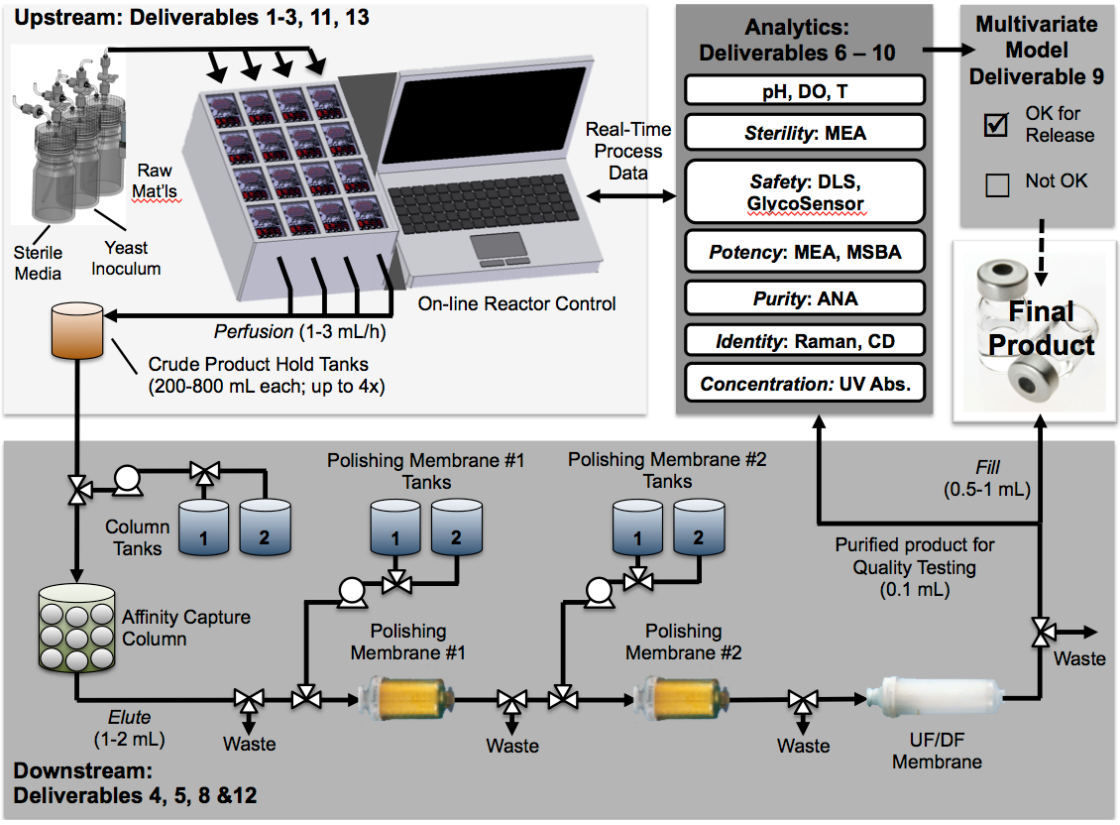

Fig 1.3: Overview of InSCyT biomanufacturing platform

The InSCyT platform has innovated the manufacturing bioprocess with the use of yeast rather than mammalian CHO cell or plant cells productions. In this way, multiple drugs will be produced in the same time period by an easily programmed switch with chemical induction. For example, Tim Lu's group[43] has already developed and demonstratedthe exposure of the production cells to methanol as an inducer to produce hGH or interferon.

This platform will generate selected pharmaceutical products with the use of controlled transcription factors. Another advantage of this system is the ability to use lyophilized yeast strain samples to minimize the cold

chain burden which is particularly advantageous in remote locations. The MIT team also innovated and developed a closed-loop process control with on-line techniques in order to monitor the fermentation process and if necessary adjust the production cycle. The platform utilized of novel affinity reagents, produced with a fusion tag system to efficiently capture and purify the products. Subsequently, product profiling utilized off-line mass spectrometry to discover product variants produced by fermentation issues in the development phase and then provide the evidence of a high level purity and low level degradations for products produced by the final platform. To monitor the quality of the products, several of analytical technologies and strategies have been utilized, including real-time monitoring Raman spectroscopy, UV detection, NMR, spectrofluorimetry,[42, 44] in addition to including off-line liquid chromatography with mass spectrometry. A fundamental need of the product testing stage is to determine the primary and secondary structure of the protein products. Then for measurement of the required biological properties, further testing with binding and bioassays were performed by the MIT team or contract testing laboratories but will not be further discussed in this report.

As discussed earlier the expiration of original pharmaceutical patents has allowed the development of biosimilar products and these were the focus of the MIT team. In this dissertation, we will demonstrate the characterization of the recombinant biopharmaceuticals produced from the InSCyT system, including human growth hormone, GCSF, glargine, interferon and antibodies. Aspects of the recombinant proteins such as the primary structures, stability, degradation variants and post-translational modifications will be discussed.

1.5 Proteomic studies

Proteomics is a challenging area of global study due to the complexity of the human set of proteins and much research must be performed before a targeted method for a disease specific biomarker can be developed.

High sensitivity, specificity and resolution are needed in the LC/MS analysis to well characterize the proteins or peptides in a complex samples such as human serum or plasma. To goal of such research is to identify biomarkers that can detect the early stage of a disease or to guide treatment of the patients. However, these aims will encounter the obstacles of protein post-translational modifications, degradations, and the presence of mutant variants, which can misdirect the analysis and the discovery of a biomarker suitable for the evaluation of the illness. If one looks at the history of medicine its important even urgent to investigate the advanced analytical methods and then differentiate the human proteins and their modifications precisely. The study of cancer biomarker proteins began in 1847 with discoveries by Henry Bence-Jones.[45] The nowadays proteomic researchers have a huge advantage of the sequencing of the human genome and their studies rely on the derived amino acid sequence databases and predicted mass spectrometry data sets,[46] however, not all potential sequences are expressed and determined. The identification of specific biomarkers from thousands of secreted proteins is still a challenge and termed the secretome.[47]

Although predicted mass spectrometric datasets have been applied to many advanced proteomic studies with promising data results, the use of this tool has remained restricted by the complexity of clinical samples. The field continues to advance with innovative MS technologies which are well developed and updated in a fast pace. The high sensitivity, resolution and multi fragmentation functions of the latest mass spectrometers are designed and marketed by companies such as Thermo, Waters and Agilent. Mass spectrometry techniques, nowadays are more and more powerful than previous decades.[48, 49] Consequently, these novel instruments have been used in the research studies of biomarker discoveries and human diseases understandings. For these studies, the accuracy and tolerance of measurement of the protein ions are crucial to the identifications. From the other side, the identification of random mutations or post translational modifications must rely on sophisticated instrumentation, both in the liquid chromatography separation and an on-line mass analyzer. However, low abundance data sets are difficult to process and thus accurately diagnose the con-

stituent proteins or genes. MALDI-TOF[50] has been applied in many research studies and aided the detection and evaluation of cancer biomarkers. While many advanced and sophisticated techniques have been applied in this field, however, to distinctly predict the biomarkers then directly help the patients in clinical trials there are still the need for continued progress in the analytical techniques. Research studies focusing on the molecular level of cancer cells and upregulated disease mechanisms are essential for scientists to comprehensively understand cancers. However, the barriers to identify the true biological indicators still exist and the field requires additional high resolution studies of cancer cells with different phenotypes.

1.6 Protein degradation and PTMs

the amino acid sequence of a protein is determined by the corresponding gene sequence, however, variants of expression may occur as a result of alternative splicing mechanisms. Then the biosynthetic process introduces post-translational modifications (PTMs) of the structure which may effect function by the inducing conformational changes in the 3D structure, altered cellular locations and modified protein-protein interactions. The diversity of PTMs is shown in Fig.1.4[51] The PTMs described in this study include the most common ones, asparagine (Asn) deamidation, methionine (Met) oxidation, cysteine (Cys) disulfide bond linkages and glycosylation. The efficacy of a protein therapeutic may be influenced by dynamic changes of PTMs, which result in functional changes.[52] Previous reports have monitored the existence of degradation reactions in samples corresponding to an innovator drug substance, biosimilar and counterfeit forms. In the reported study the approved biosimilar version contained similar level of the variants to the innovator product, however, a significantly higher level of variants existed in the unauthorized medication.[53] The consequences of the variants on the drugs still remain to be studied in biological researches, but the quality control to accurately identify and characterize them are very necessary and can be accurately performed.[54, 55, 56]

Aiming to detect the subtle differences of the biosimilar with the original product, a rigorous stability study is very necessary. The evaluation of protein stability will illustrate any changes in structure on storage that can effect protein dynamics and kinetics. These studies will also guide the development of appropriate formulations and storage conditions. Thus in order to well understand the function and efficacy of the biosimilar pharmaceuticals, stability test must be carefully designed, not only from a research view, but also in product development which is very important in both industry and clinical studies.

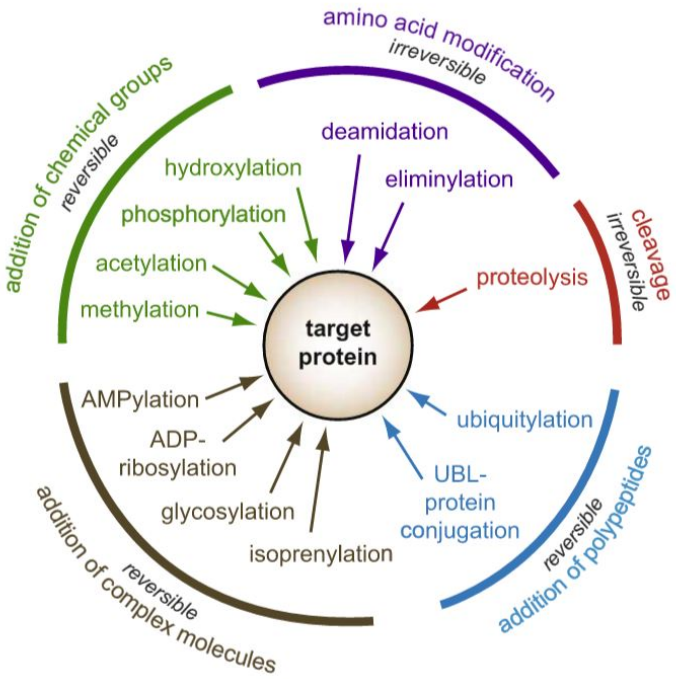

Fig 1.4: Diversity of post-translational modifications

There are many aspects of product quality to concern the development scientist. Beside the results of clinical trails, only from the analytical aspects, one determines the product degradations over time such as aggregations, oxidations under regular storage, which results in changes in the structures and functions of the protein, which could be crucial for predicting safety concerns for the patients. As for the complexity

of the protein drugs, since the biosimilar is not completely identical with the innovator product, a small differences could impact hugely in the stability of the final products. In some researches,[57] scientists did stability examinations for a series of temperatures for 4 °C , 25°C , 37°C . In other studies, the shelf storage conditions were explored, and under freeze- thaw cycles. Moreover, stability evaluations have been taken under additional stressed conditions such as high oxygen, or extreme pH levels. These tests are compared with the innovator products and then the performance of biosimilar can be indicated. The reason for product instability can originate from many manufacture aspects, such as the upstream process with cell culturing, the cultivation environments, the level of gene product expression, also the downstream process including issues with purification or the polishing system that adds the final formulation buffer. Based on the theoretical three-dimensional conformation of a given proteins, the design of the stability study will include various conditions. For specific amino acid locations, the outside residues can be modified more easily than the buried core. In this study the stability detections will be focused on oxidation, deamidation, aggregation, as well as other modifications. As a consequence, the approval of a final product is not simply based on the presence of a correct sequence but will be evaluated from many other aspects to ensure the appropriate level of safety in the clinic.

1.6.1 Proteolysis

Proteolysis is a process used to enzymatically digest proteins into small peptides with proteases in order to effectively separated all protein fragments by HPLC and successfully analyze them by MS.[58] Each enzyme has its best conditions for specificity and efficiency such as temperature, pH, duration of reaction, ratio of enzyme to substrate and other aspects. Peptide mapping also has limitations by the efficiency of proteolysis step.[59] Enzyme selection is one of key parts of the experimental optimization process which is based on

both the structure of the protein of interest and the goal of the research. The enzymes can include specific and non-specific proteases, both of which have their unique limitations in the sequence mapping and their use will be applied based on known digestion specificity.[59] Trypsin is the most commonly used specific protease that precisely cleaves at the carboxyl side of the amino acids arginine and lysine, with predictable peptides. Other well used enzymes in this study include LysC to cleave at the carboxyl side of lysine, GluC to cleave at the carboxyl side of glutamate or aspartate acids. On the other hand, non-specific enzymes such as pepsin aims to cleave at the carboxyl side of hydrophilic amino acids, and chymotrypsin to cleave at the aromatic amino acids. Additionally, multiple enzymes can be applied in one experiment to generate more cleavage sites to obtain peptides more appropriate for identification. For example, in a glycosylation study, PNGaseF is a specific enzyme that releases N-linked oligosaccharides and is added after a trypsin digestion to identify the corresponding peptide backbones of glycopeptides.

1.6.2 Deamidation

Deamidation is one of the most common degradation reactions of proteins that occurs under high pH or temperature conditions in which the amide group is converted to a carboxylate group in the asparagine orrarely a glutamine residue. The mechanism[60] is shown in Fig1.5. The symmetric succinimide intermediate results in two products from its hydrolysis, either forming an aspartate, or isoaspartate residue. Deamidation occurs readily if the Asn or Gln residue is followed by a glycine residue that provides minimal steric hindrance to the degradative reaction.[?]

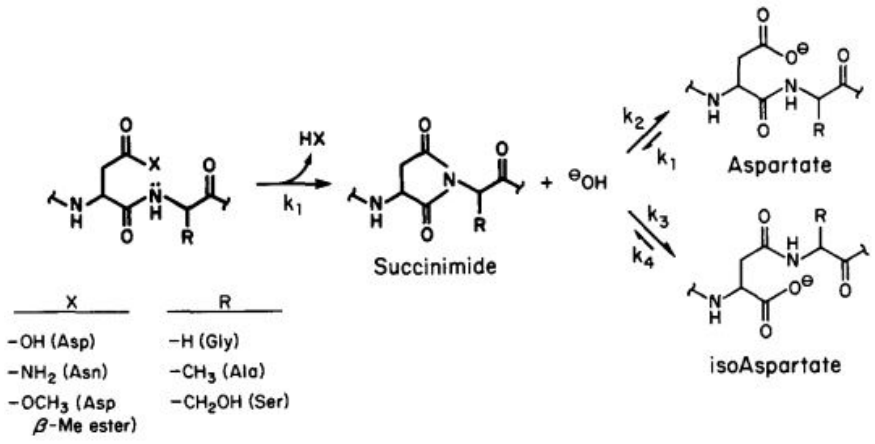

Fig 1.5: Mechanism of deamidation

The strategy to characterize and quantitate the deamidation reaction has been developed in several ways even if there is only a 1 Da mass change, however, this small mass shift between asparagine and aspartate or isoaspartate can be detected and differentiated by a high resolution mass spectrometer. At the same time, the elution order of Asn, Asp and isoAsp are different according to their isoelectric point and to monitor the deamidation reaction, SDS-PAGE with IEF is commonly applied.

The MS-based method with either ECD or ETD fragmentation method have been studied and applied to differentiate between Asp and isoAsp residues which as isomers present unusual detection challenges. In this approach, the isomeric residues containing the respective side chain are cleaved with different fragmentation patterns and then generate c and z ions of different mass. After the fragmentation, isoAsp will generate c+57 and z-57 fragments which are unique from the Asp cleavage reactions.[60, 61] Isotope labeling of oxygen is an alternative method that has been used to monitor and quantitate deamidation.[62]

1.6.3 Oxidation

Oxidation may occur on the residues of methionine, tyrosine, tryptophan, and histidine, and the most common site is methionine. This residue in the protein sequence will be oxidized to methionine sulfoxide or even methionine sulfone and the chemical mechanism was shown in Fig1.6. The sulfone is formed when a harsh oxidized condition occurs. The 3-D structure is the major determinant for the site of the degradations as steric accessibility often limits such reactions.[63] With our previous experience, we expect that the amino acids that located on the outer region of the protein are easily exposed and degraded.

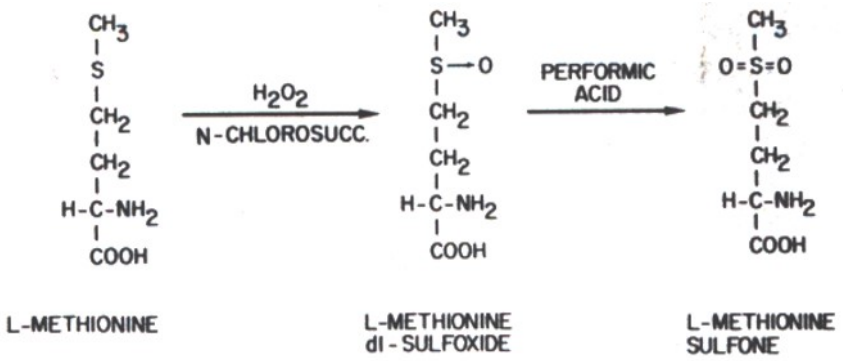

Fig 1.6: Mechanism of methionine oxidation

Since oxygen is everywhere, the oxidation reaction may occur in any stage of the procedures of protein therapeutic productions, such as fermentation, purification, formulation or transportation. Oxidation of residues of a protein can then promote aggregation processes, deactivation of function, and dynamic alteration of structure. The oxidation differences in original drug, biosimilars and counterfeit have been analyzed by Wu,[53] which indicated the higher oxidation level in the counterfeit version. To monitor and evaluate the level of oxidation is essential to assure the quality and stability of a biopharmaceutical. In some cases, the

stability of the proteins under stressed condition will be investigated as well to further evaluate the oxidation level under extreme conditions and the drug capacity to resist degradation.

The oxidized peptide containing methionine and the normal peptide are easily differentiated by the large mass shift due to an additional oxygen which can be detected on a medium resolution mass spectrometer.as well as elution order changes in the reversed phase LC separation as a result of the decrease in hydrophobicity of the peptide that contains the oxidized methionine, Under severe oxidized conditions, such as a hydrogen peroxide incubation, the sulfone moiety will formed as well.[64] Of course, the structure and sequence of the protein decide the possibility of conformation changes. Based on previous studies,[65] a methionine residue located on outer loop is more readily oxidized than a residue located in the inner loop of the protein structure.

1.6.4 Disulfide bond linkages

The covalent linkage between cysteine residues plays an important role in stabilizing protein structure. The formation of the disulfide linkage has been described in Fig1.7, which is critical to maintain the structural composition of a protein. To characterize and identify a disulfide bond, mass spectrometry has become an essential technique, which determines not only the mass difference due to the linkage, but also the residue site and quantitation. For a MS analysis, ETD has the power to cleave the disulfide bond while CID cannot. For a small protein and simple structure, CID can be applied to the characterization of the peptide with cysteine residue based on the predicted sequence. ETD enables the breakage of the cysteine linkage and generates disulfide-dissociated species as well as charge-reduced species which can be determined with a high resolution and accuracy mass spectrometer.

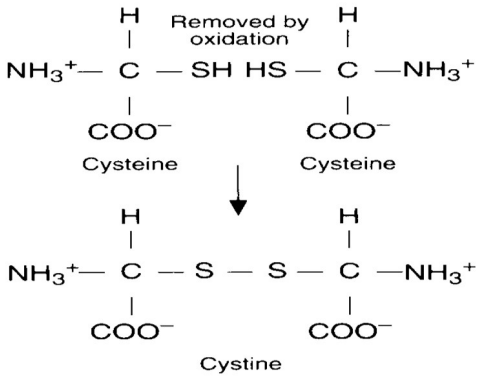

Fig 1.7: Mechanism of disulfide bond formation

1.6.5 Glycosylation

Glycosylation is an important PTM that occurs in the Golgi body and endoplasmic reticulum of a eukaryotic cell. As well studied, glycosylation has been separated into two major types, N- and O-glycosylation. While the motif that determines O-glycosylation if less predictable than for N-glycosylation the glycans are always attached to the residues of serine or threonine. For the N-glycosylation, there are many studies that have resulted in the determination of the NXS/T (the X is not proline) motif that allows the prediction of these structures. The research studies focusing on the functions of glycans and the safety concerns due to glycosylation changes, is summarized in the following review.[66]

In addition, compositional and linkage analysis are necessary to well characterize glycans. The pipeline of glycan qualification and evaluation always includes releasing glycans from the backbones and identification of glycopeptides with enzymatic methodologies. With the drawbacks of limited information of site-specificity of O-linked glycosylation, its complicated to identify the glycan structures with a straightforward method and thus multiple steps sample preparation are needed. The most widely used enzyme to

release N-linked glycans is PNGaseF.

Since the glycosylation structures are complex, the recognition and identification of the site location of individual structures is always determined by MS analysis of enzymatically generated glycopeptides (bottom-up techniques). The characterization of glycopeptides are basically estimated from the parent mass and abundance differences of the backbone peptides before and after the glycan release. Several ionization techniques could be applied to identify the structures, such as collision induced dissociation (CID),[67] electron transfer dissociation (ETD) and high-energy collision dissociation (HCD).[68, 69, 70]

In recent years, the bioinformatics technologies related to glycan analysis have undergone rapid exploration and put in to service with huge databases. The data generated will assist the discovery of biomarkers and the structure predictions. The evaluation and monitoring of glycans are significant in clinical researches, in which the past has been an ignored field. Although the characterization of glycans is challenging, the utilization of online bioinformatic tools together with MS software are readily applied and effective.

There are several types of glycosylation, including N-linked glycosylation,[71] O-linked glycosylation[72] C-glycosylation and S-linked glycosylation (only found in bacteria). Aberrant patterns of N- and O-linked oligosaccharide modification in proteins are found in most of human cancers. This project will focus on N-linked glycosylation, and in order to fully understand the characterizations of glycans and the mutations in associated diseases, more research on this area needs to be achieved.

The source of the biological material used for analysis and structural detection provides different difficulties which can compromise results of the analysis. Blood or serum of the patients is the most commonly used, while tumor tissues, pleural and saliva samples also are used for diagnosis and validations. Glycosylation proteomics[73] has become a new trend for cancer monitoring, since accumulating studies have indicated can-

cer cell shows a significantly different level in amount of glycans or mutated structures. Since the prediction of early stage cancers are not positive enough from characterization of these biomarkers,[74] we must conclude that the detection of the disease stages is a challenge that will require better analysis tools. In recent years, although increasing attention has been focused on glycan biomarkers and glycomics, limitations stillexist in most aspects, because of the complex nature of glycosylation modification and abnormal patterns in disease.

1.7 Analytical aspects

In todays research, many analytical techniques are used to identify chemical structures, such as fluorescence, UV-Vis, and Raman spectroscopy, etc. These techniques are effective in specific situations, however, they are not capable to determine the complex protein structural changes or quantitation of variants. Mass spectrometry is considered as one of the most powerful analytical techniques to identify and profile protein biomarkers.[75, 76] With the progress of methodology and techniques, the matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) methods have been commonly used many research studies in the past decades.[77, 78] The sensitivity, efficiency and resolution of these two ionization techniques have had extensive use by scientists in this field.

MALDI was first mentioned/invented by Tanka, Hillenkamp and Karas in late 1980s[79] and is a commonly applied ionization technique to analyze proteins and other biomolecules. MALDI ionization is usually coupled with time-of-flight (TOF) mass spectrometer and has the advantages of tolerances to moderate amounts of salts in millimolar concentrations, unlike ESI as well as sample preparation of complex mixtures with minimal steps. ESI offers a more gentle ionization procedure than MALDI, and the development of an online separation with LC is convenient. Desalting the sample is a crucial step needed to be considered before the MS analysis. A scheme of the mechanism[80] of MALDI and ESI is shown in the Fig.1.8. Considering

the convenience and effectiveness of this ionization method for both protein therapeutic characterization and clinical proteomics measurements, ESI has been selected in this project.

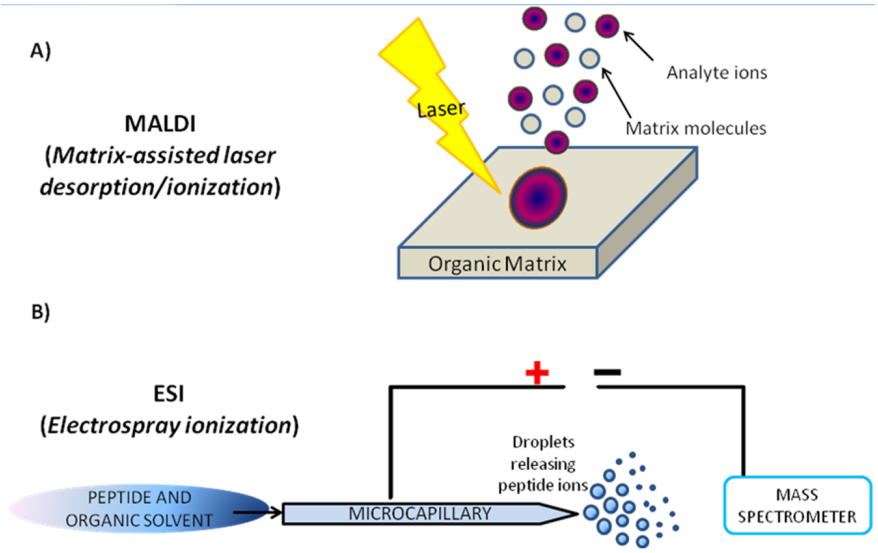

Fig 1.8: Mechanism scheme of MALDI and ESI

In most cases, MALDI will be applied with a time of flight mass spectrometers, named MALDI-TOF which used a laser pulse to desorb the samples from the matrix and then for ionization. The analyte was co-crystallized with an excess of the matrix such as α-Cyano-4-hydroxycinnamic acid (CHCA), and sinapinic acid (SA), etc. The matrix plays an essential role to vaporize the analytes by absorbing the laser energy. The laser pulse was applied to desorb and ionize the samples then the sample ions was accelerated by the high voltage in the vacuum system. The analyte ions pass through the evacuated tube arriving to the reflectron then to be analyzed.

The characterization of protein is a complex process and LC-MS is the most powerful way to accomplish that goal. This technique will lead the protein characterization into a well understood way.[81] The proteins sequence, modifications, degradation variants are required to be determined, for the reason of the purity,

quality, efficacy, safety or toxicity of the biopharmaceutical. In the biological market and pharmaceutical industry, the analysis of bioproducts with LC-MS is very critical to success of the industry. In the recent decades, the analysis has majorly divided into top-down and bottom-up analysis, in other words, which the protocols are separated to intact protein and peptide fingerprint analysis, respectively. The two mass analysis tools have their own advantages and shortcomings to monitor the protein quality. In this section, the advanced LC-MS techniques for both top-down and bottom-up method will be described.

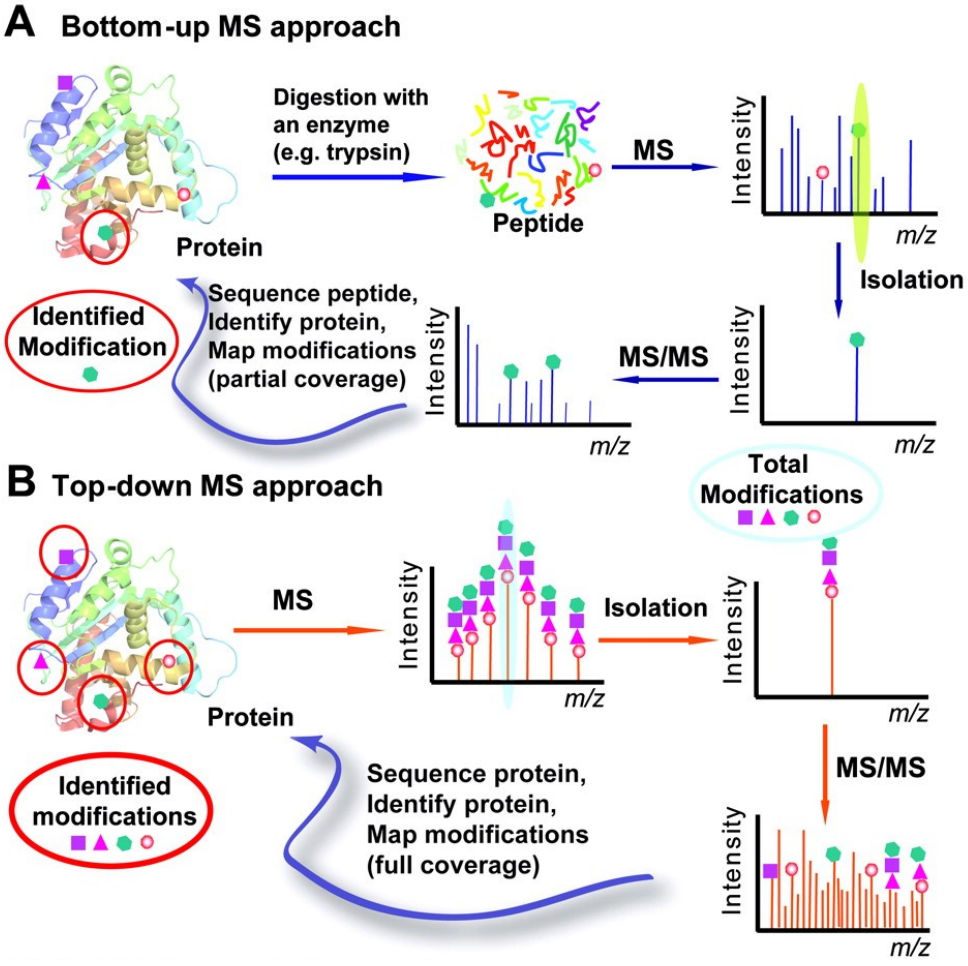

Fig 1.9: Characterization comparison of bottom-up and top-down methods; A, Bottom-up MS approach; B, Top-down MS approach

The bottom-up method is more commonly used to characterize the protein structure with digested peptide information derived from mass spectrometry analysis. With more detailed information from additional enzymatic digestions and repeated LC/MS analysis, this analysis can give a complete picture of protein structures but the preparation and post-analysis consumes significantly amounts of time and the digestion procedures can generate artifacts. Fig.1.9A indicates the process of the bottom-up approach.[82] The proteins, or bioproducts will be digested into peptide fragments by a selected protease and then for analysis by a LC separation and MS analysis. The peptides will then be identified by the mass detector with advanced MS/MS determinations. The amino acids of the protein sequence will be specifically identified in this process. Although there are several limitations such as enzyme efficacy or resolution of instrument, peptides could still be precisely predicted and monitored because the analysis could be adjusted for a selected multi-enzymes digestion and development of an optimized LC-MS method for each digest. For a biotherapeutic characterization a 100% sequence coverage is required to ensure that there are no mutations in the sequence or PTMs. Proteins are much more complex to be measured than small molecules or short peptides due to the large structures and lower solubility, moreover, the modifications are also challenging to be easily detected by top-down techniques.

As shown in Fig.1.10, the peptide backbone will be cleaved in the indicated pattern.[83] By mass spectrometry, in general, this will determine the amino-acid sequence according to the mass differences between the mass peaks. Peptide sequencing is a sensitive and robust way to identify the proteins. The peptide fragmentations are induced by the collisions and the amide bonds are the main breakage site due to the preferred lowest energy pathway.[84] This pathway will provide b-ions by the amino-terminal fragment or y-ions by the carboxy terminal fragment.

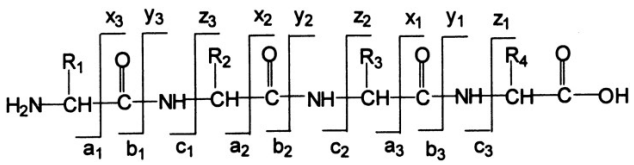

Fig 1.10: MSMS peptide fragmentation

Top-down and bottom-up methodologies are both applied in modern proteomics researches. The top-down techniques are widely used[85, 86, 87] for intact protein determination with high mass accuracy measurement. The proteins or bioproducts will be analyzed directly as the intact molecule without digestion which saves preparation time as well as artifacts generated by the digestion process. In our research, the bottom-up method is mainly applied to evaluate the quality of the proteins, and top-down instruments are used as well to accompany the digested analysis for small proteins drugs. At the level of a small protein level, intact protein analysis can reach full sequence coverage and successful determination usually can be achieved with a molecular weight less than 70kDa.[88] However, the peptide fragmentation of the intact protein in the mass spectrometer at a large molecule level is less predictable than for analysis of an enzyme digested protein. In summary top-down applications require a short time period from sample preparation to final analysis, however, the detailed information of post translational modifications may not be specifically obtained. For traditional methods, the FT-MS is more applicable for intact protein analysis based on its resolving power while so far, Orbitrap, TOF and more sophisticated instruments and software are under explored.[89]

In this dissertation, the technology used is proposed to be optional to characterize the biological products in both crude and purified stages. The peptide mapping digests and the natural post-translational modifications in the proteins are required to be evaluated. In this analytical perspective, the intact protein analysis still remain challenging such as for variants identification, amino acid substitution, and quantitation, etc. The

intact analysis methodology may not provide an indicative strategy to the manufacturing process as the crude biologics may contain a high portion of buffer, which will give a large background signal that interferes with identification. Additionally, the remained host cell protein determination is another concern from the manufacturing process, which is challenge to a top-down analysis.[90] Although intact protein analysis is a meaningful tool to rapidly evaluate the quality of a complex protein, our team decided to majorly apply the bottom-up analytical strategies in this dissertation.

References

[1] Jeffrey P Koplan, T Christopher Bond, Michael H Merson, K Srinath Reddy, Mario Henry Rodriguez, Nelson K Sewankambo, and Judith N Wasserheit. Towards a common definition of global health. *The Lancet*, 373(9679):1993–1995, 2009.

[2] World Health Organization. *Global health risks: mortality and burden of disease attributable to selected major risks*. World Health Organization, 2009.

[3] Pamela Y Collins, Vikram Patel, Sarah S Joestl, Dana March, Thomas R Insel, Abdallah S Daar, Isabel A Bordin, E Jane Costello, Maureen Durkin, Christopher Fairburn, et al. Grand challenges in global mental health. *Nature*, 475(7354):27–30, 2011.

[4] Dean T Jamison, Lawrence H Summers, George Alleyne, Kenneth J Arrow, Seth Berkley, Agnes Binagwaho, Flavia Bustreo, David Evans, Richard GA Feachem, Julio Frenk, et al. Global health 2035: a world converging within a generation. *The Lancet*, 382(9908):1898–1955, 2013.

[5] Colin D Mathers and Dejan Loncar. Updated projections of global mortality and burden of disease, 2002-2030: data sources, methods and results. *Geneva: World Health Organization*, 2005.

[6] Gonghuan Yang, Yu Wang, Yixin Zeng, George F Gao, Xiaofeng Liang, Maigeng Zhou, Xia Wan, Shicheng Yu, Yuhong Jiang, Mohsen Naghavi, et al. Rapid health transition in china, 1990–2010: findings from the global burden of disease study 2010. *The lancet*, 381(9882):1987–2015, 2013.

[7] BWKP Stewart, Christopher P Wild, et al. World cancer report 2014. *Health*, 2017.

[8] Pavel A Nalimov, Dmitry Y Rudenko, and Djamilia F Skripnuk. Big pharma: Trick or treat for global health. *Mediterranean Journal of Social Sciences*, 6(1):216, 2015.

[9] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795, 2015.

[10] Steven Simoens. Biosimilar medicines and cost-effectiveness. *Clinicoecon Outcomes Res*, 3:29–36, 2011.

[11] Henry G Grabowski, Rahul Guha, and Maria Salgado. Regulatory and cost barriers are likely to limit biosimilar development and expected savings in the near future. *Health Affairs*, 33(6):1048–1057, 2014.

[12] Food, Drug Administration, et al. Biologics price competition and innovation act, 2014.

[13] Shein-Chung Chow, Laszlo Endrenyi, and Peter A Lachenbruch. Comments on the fda draft guidance on biosimilar products. *Statistics in medicine*, 32(3):364–369, 2013.

[14] Jun Wang and Shein-Chung Chow. On the regulatory approval pathway of biosimilar products. *Pharmaceuticals*, 5(4):353–368, 2012.

[15] Teruyo Arato. Recent regulations of biosimilars in japan. In *2011 The Drug Information Association Conference. Chicago, Illinois, USA. http://www. pmda. go. jp/regulatory/file/english_presentation/biologics/B-E1arato. pdf (accessed October, 2014)*, 2014.

[16] Huub Schellekens. Biosimilar therapeuticswhat do we need to consider? *NDT plus*, 2(suppl 1):i27–i36, 2009.

[17] Huub Schellekens and Ellen Moors. Clinical comparability and european biosimilar regulations. *Nature biotechnology*, 28(1):28, 2010.

[18] Hakan Mellstedt, D Niederwieser, and H Ludwig. The challenge of biosimilars. *Annals of oncology*, 19(3):411–419, 2008.

[19] Thomas Dörner, Vibeke Strand, Gilberto Castañeda-Hernández, Gianfranco Ferraccioli, John D Isaacs, Tore K Kvien, Emilio Martin-Mola, Thomas Mittendorf, Josef S Smolen, and Gerd R Burmester. The role of biosimilars in the treatment of rheumatic diseases. *Annals of the rheumatic diseases*, pages annrheumdis–2012, 2012.

[20] Anthony J Hatswell, Gianluca Baio, Jesse A Berlin, Alar Irs, and Nick Freemantle. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999-2014. *BMJ open*, 6(6):e011666, 6 2016.

[21] Asher Mullard. 2015 FDA drug approvals. *Nature Reviews Drug Discovery*, 15(2):73–76, 2 2016.

[22] Asher Mullard. EMA recommended 39 new drug approvals last year. *Nature Reviews Drug Discovery*, 15(2):77–77, 2 2016.

[23] Alain Beck and Janice M Reichert. Approval of the first biosimilar antibodies in europe: a major landmark for the biopharmaceutical industry. In *MAbs*, volume 5, pages 621–623. Taylor & Francis, 2013.

[24] Simon D Roger. Biosimilars: How similar or dissimilar are they?(review article). *Nephrology*, 11(4):341–346, 2006.

[25] Food, Drug Administration, et al. Quality considerations in demonstrating biosimilarity of a therapeutic protein product to a reference product. guidance for industry, 2015.

[26] Ivo Abraham and Karen MacDonald. Clinical safety of biosimilar recombinant human erythropoietins. *Expert opinion on drug safety*, 11(5):819–840, 2012.

[27] Dae Hyun Yoo, Nenad Prodanovic, Janusz Jaworski, Pedro Miranda, Edgar Ramiterre, Allan Lanzon,

Asta Baranauskaite, Piotr Wiland, Carlos Abud-Mendoza, Boycho Oparanov, et al. Efficacy and safety of ct-p13 (biosimilar infliximab) in patients with rheumatoid arthritis: comparison between switching from reference infliximab to ct-p13 and continuing ct-p13 in the planetra extension study. *Annals of the rheumatic diseases*, 76(2):355–363, 2017.

[28] Sang Hyoung Park, Young-Ho Kim, Ji Hyun Lee, Hyeok Jin Kwon, Suck-Ho Lee, Dong Il Park, Hyung Kil Kim, Jae Hee Cheon, Jong Pil Im, You Sun Kim, et al. Post-marketing study of biosimilar infliximab (ct-p13) to evaluate its safety and efficacy in korea. *Expert review of gastroenterology & hepatology*, 9(sup1):35–44, 2015.

[29] Hiroshi Horikawa, Mina Tsubouchi, and Koji Kawakami. Industry views of biosimilar development in japan. *Health policy*, 91(2):189–194, 2009.

[30] Jonathan Kay. Biosimilars: a regulatory perspective from america. *Arthritis research & therapy*, 13(3):112, 2011.

[31] Irving S Johnson. Human insulin from recombinant dna technology. *Science*, 219(4585):632–637, 1983.

[32] Harry Keen, JC Pickup, RW Bilous, A Glynne, GC Viberti, RJ Jarrett, and R Marsden. Human insulin produced by recombinant dna technology: safety and hypoglycaemic potency in healthy men. *The Lancet*, 316(8191):398–401, 1980.

[33] R. S. Sai Murali, R. Basavaraju, and G. Nageswara Rao. Liquid Chromatography Coupled to Mass Spectrometry Based Identification of Elite Chemotypes of <I>Adhatoda vasica</I> Nees for Profitable Agronomy A Farmer Centric Approach. *Indian Journal of Science and Technology*, 9(26), 7 2016.

[34] Nacole D Lee, Bhargavi Kondragunta, Shaunak Uplekar, Jose Vallejos, Antonio Moreira, and Govind Rao. Studies of protein oxidation as a product quality attribute on a scale-down model for cell culture process development. *PDA Journal of Pharmaceutical Science and Technology*, 69(2):236–247, 2015.

[35] Robert L Garnick. Safety aspects in the quality control of recombinant products from mammalian cell culture. *Journal of pharmaceutical and biomedical analysis*, 7(2):255–266, 1989.

[36] Klaus Graumann and Andreas Premstaller. Manufacturing of recombinant therapeutic proteins in microbial systems. *Biotechnology journal*, 1(2):164–186, 2006.

[37] Gary Walsh. Biopharmaceutical benchmarks–2003. *Nature biotechnology*, 21(8):865, 2003.

[38] Gary Walsh. Biopharmaceutical benchmarks 2006. *Nature biotechnology*, 24(7):769, 2006.

[39] Gary Walsh. Biopharmaceutical benchmarks 2014. *Nature biotechnology*, 32(10):992–1000, 2014.

[40] Kerry R. Love, Kartik A. Shah, Charles A. Whittaker, Jie Wu, M. Catherine Bartlett, Duanduan Ma, Rachel L. Leeson, Margaret Priest, Jonathan Borowsky, Sarah K. Young, and J. Christopher Love. Comparative genomics and transcriptomics of Pichia pastoris. *BMC Genomics*, 17:550, 2016.

[41] Tillman U Gerngross. Advances in the production of human therapeutic proteins in yeasts and filamentous fungi. *Nature biotechnology*, 22(11):1409, 2004.

[42] J. Christopher Love, Kerry Routenberg Love, and Paul W. Barone. Enabling global access to high-quality biopharmaceuticals. *Current Opinion in Chemical Engineering*, 2(4):383–390, 2013.

[43] Amos E. Lu, Joel A. Paulson, and Richard D. Braatz. pH and conductivity control in an integrated biomanufacturing plant. In *2016 American Control Conference (ACC)*, pages 1741–1746. IEEE, 7 2016.

[44] Wei Ouyang, Sung Hee Ko, Di Wu, Annie Yu Wang, Paul W Barone, William S Hancock, and Jongyoon Han. Microfluidic platform for assessment of therapeutic proteins using molecular charge modulation enhanced electrokinetic concentration assays. *Analytical Chemistry*, 88(19):9669–9677, 2016.

[45] H Bence Jones. Papers on chemical pathology;: Prefaced by the gulstonian lectures, read at the royal college of physicians, 1846. *The Lancet*, 50(1245):32–35, 1847.

[46] Richard M. Neve, Koei Chin, Jane Fridlyand, Jennifer Yeh, Frederick L. Baehner, Tea Fevr, Laura Clark, Nora Bayani, Jean-Philippe Coppe, Frances Tong, Terry Speed, Paul T. Spellman, Sandy DeVries, Anna Lapuk, Nick J. Wang, Wen-Lin Kuo, Jackie L. Stilwell, Daniel Pinkel, Donna G. Albertson, Frederic M. Waldman, Frank McCormick, Robert B. Dickson, Michael D. Johnson, Marc Lippman, Stephen Ethier, Adi Gazdar, and Joe W. Gray. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, 10(6):515–527, 2006.

[47] Martin Sjöström, Reto Ossola, Thomas Breslin, Oliver Rinner, Lars Malmström, Alexander Schmidt, Ruedi Aebersold, Johan Malmström, and Emma Niméus. A Combined Shotgun and Targeted Mass Spectrometry Strategy for Breast Cancer Biomarker Discovery. *Journal of Proteome Research*, 14(7):2807–2818, 7 2015.

[48] Shao-En Ong and Matthias Mann. Mass spectrometry–based proteomics turns quantitative. *Nature chemical biology*, 1(5):252–262, 2005.

[49] Emanuel F Petricoin, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, et al. Use of proteomic patterns in serum to identify ovarian cancer. *The lancet*, 359(9306):572–577, 2002.

[50] Haleem J Issaq, Timothy D Veenstra, Thomas P Conrads, and Donna Felschow. The seldi-tof ms

approach to proteomics: protein profiling and biomarker identification. *Biochemical and biophysical research communications*, 292(3):587–592, 2002.

[51] David Ribet and Pascale Cossart. Post-translational modifications in host cells during bacterial infection. *FEBS letters*, 584(13):2748–2758, 2010.

[52] Sven Frokjaer and Daniel E Otzen. Protein drug stability: a formulation challenge. *Nature reviews drug discovery*, 4(4):298, 2005.

[53] Shiaw Lin Wu, Haitao Jiang, William S. Hancock, and Barry L. Karger. 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. *Analytical Chemistry*, 82(12):5296–5303, 2010.

[54] RL Garnick, NJ Solli, and PA Papa. The role of quality control in biotechnology: An analytical perspective. *Analytical chemistry*, 60(23):2546–2557, 1988.

[55] Roberto Sitia and Ineke Braakman. Quality control in the endoplasmic reticulum protein factory. *Nature*, 426(6968):891–894, 2003.

[56] M Marcia Federici. The quality control of biotechnology products. *Biologicals*, 22(2):151–159, 1994.

[57] V Vieillard, A Astier, C Sauzay, and M Paul. One-month stability study of a biosimilar of infliximab (remsima®) after dilution and storage at 4 c and 25 c. In *Annales Pharmaceutiques Françaises*, volume 75, pages 17–29. Elsevier, 2017.

[58] Randall W King, Raymond J Deshaies, Jan-Michael Peters, and Marc W Kirschner. How proteolysis drives the cell cycle. *Science*, 274(5293):1652, 1996.

[59] Don W Cleveland, STUART G Fischer, MARC W Kirschner, and ULRICH K Laemmli. Peptide mapping by limited proteolysis in sodium dodecyl sulfate and analysis by gel electrophoresis. *Journal of Biological Chemistry*, 252(3):1102–1106, 1977.

[60] Terrence Geiger and S Clarke. Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. succinimide-linked reactions that contribute to protein degradation. *Journal of Biological Chemistry*, 262(2):785–794, 1987.

[61] Rainer Bischoff and Hanno VJ Kolbe. Deamidation of asparagine and glutamine residues in proteins and peptides: structural determinants and analytical methodology. *Journal of Chromatography B: Biomedical Sciences and Applications*, 662(2):261–278, 1994.

[62] Xiaojuan Li, Jason J. Cournoyer, Cheng Lin, and Peter B. O'Connor. Use of 18O Labels to Monitor Deamidation during Protein and Peptide Sample Processing. *Journal of the American Society for Mass*

Spectrometry, 19(6):855–864, 6 2008.

[63] ML Houchin and EM Topp. Chemical degradation of peptides and proteins in plga: a review of reactions and mechanisms. *Journal of pharmaceutical sciences*, 97(7):2395–2404, 2008.

[64] Gerrit Toennies and Thomas P Callan. Methionine studies iii. a comparison of oxidative reactions of methionine, cysteine, and cystine. determination of methionine by hydrogen peroxide oxidation. *Journal of Biological Chemistry*, 129(2):481–490, 1939.

[65] Walther Vogt. Oxidation of methionyl residues in proteins: tools, targets, and reversal. *Free Radical Biology and Medicine*, 18(1):93–105, 1995.

[66] Alain Beck, Elsa Wagner-Rousset, Marie-Claire Bussat, Maryline Lokteff, Christine Klinguer-Hamour, Jean-Franois Haeuw, Liliane Goetsch, Thierry Wurch, Alain V. Dorsselaer, and Nathalie Corvaia. Trends in Glycosylation, Glycoanalysis and Glycoengineering of Therapeutic Antibodies and Fc-Fusion Proteins.

[67] Johannes Stadlmann, Martin Pabst, Daniel Kolarich, Renate Kunert, and Friedrich Altmann. Analysis of immunoglobulin glycosylation by LC-ESI-MS of glycopeptides and oligosaccharides. *PROTEOMICS*, 8(14):2858–2871, 7 2008.

[68] Samnang Tep, Marina Hincapie, and William S. Hancock. A general approach for the purification and quantitative glycomic analysis of human plasma. *Analytical and Bioanalytical Chemistry*, 402(9):2687–2700, 3 2012.

[69] Samnang Tep, Marina Hincapie, and William S. Hancock. A MALDI-TOF MS method for the simultaneous and quantitative analysis of neutral and sialylated glycans of CHO-expressed glycoproteins. *Carbohydrate Research*, 347(1):121–129, 1 2012.

[70] Neeley Remmers, Judy M Anderson, Erin M Linde, Dominick J DiMaio, Audrey J Lazenby, Hans H Wandall, Ulla Mandel, Henrik Clausen, Fang Yu, and Michael a Hollingsworth. Aberrant expression of mucin core proteins and o-linked glycans associated with progression of pancreatic cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 19(8):1981–93, 4 2013.

[71] Barbara Imperiali and Sarah E OConnor. Effect of n-linked glycosylation on glycopeptide and glycoprotein structure. *Current opinion in chemical biology*, 3(6):643–649, 1999.

[72] Akira Seko, Takashi Ohkura, Hiroko Ideo, and Katsuko Yamashita. Novel O-linked glycans containing 6'-sulfo-Gal/GalNAc of MUC1 secreted from human breast cancer YMB-S cells: possible carbohydrate epitopes of KL-6(MUC1) monoclonal antibody. *Glycobiology*, 22(2):181–95, 2 2012.

[73] Sergei a. Svarovsky and Lokesh Joshi. Cancer glycan biomarkers and their detection past, present and future. *Analytical Methods*, 2014.

[74] Markus Aebi. N-linked protein glycosylation in the ER. *Biochimica et biophysica acta*, 1833(11):2430–7, 11 2013.

[75] Francisca Owusu Gbormittah, Brian B. Haab, Katie Partyka, Carolina Garcia-Ott, Marina Hancapie, and William S. Hancock. Characterization of glycoproteins in pancreatic cyst fluid using a high-performance multiple lectin affinity chromatography platform. *Journal of Proteome Research*, 13(1):289–299, 2014.

[76] K Biemann. Mass spectrometry of peptides and proteins. *Annual review of biochemistry*, 61(1):977–1010, 1992.

[77] Neil L Kelleher, Hong Y Lin, Gary A Valaskovic, David J Aaserud, Einar K Fridriksson, and Fred W McLafferty. Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. *Journal of the American Chemical Society*, 121(4):806–812, 1999.

[78] Matthias Mann and Matthias Wilm. Electrospray mass spectrometry for protein characterization. *Trends in biochemical sciences*, 20(6):219–224, 1995.

[79] Michael Karas and Franz Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10000 daltons. *Analytical chemistry*, 60(20):2299–2301, 1988.

[80] Victor Manuel Bautista-de Lucio, Mariana Ortiz-Casas, Luis Antonio Bautista-Hernández, Nadia Luz López-Espinosa, Carolina Gaona-Juárez, Ángel Gustavo Salas-Lais, Dulce Aurora Frausto-del Río, and Herlinda Mejía-López. Diagnostics methods in ocular infections- from microorganism culture to molecular methods. In *Common Eye Infections*. InTech, 2013.

[81] Ursula Gassenschmidt, Klaus D Jany, Tauscher Bernhard, and Heinz Niebergall. Isolation and characterization of a flocculating protein from moringa oleifera lam. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1243(3):477–481, 1995.

[82] Han Zhang and Ying Ge. Comprehensive analysis of protein modifications by top-down mass spectrometry. *Circulation: Cardiovascular Genetics*, 4(6):711–711, 2011.

[83] Hanno Steen and Matthias Mann. The abc's (and xyz's) of peptide sequencing. *Nature reviews Molecular cell biology*, 5(9):699–711, 2004.

[84] P. Roepstorff and J. Fohlman. Letter to the editors. *Biological Mass Spectrometry*, 11(11):601–601, 1984.

[85] Zachery R Gregorich and Ying Ge. Top-down proteomics in health and disease: Challenges and

opportunities. *Proteomics*, 14(10):1195–1210, 2014.

[86] Caroline J DeHart, Ryan T Fellers, Luca Fornelli, Neil L Kelleher, and Paul M Thomas. Bioinformatics analysis of top-down mass spectrometry data. 2016.

[87] Bifan Chen, Ying Peng, Santosh G Valeja, Lichen Xiu, Andrew J Alpert, and Ying Ge. Online hydrophobic interaction chromatography–mass spectrometry for top-down proteomics. *Analytical chemistry*, 88(3):1885–1891, 2016.

[88] John C Tran, Leonid Zamdborg, Dorothy R Ahlf, Ji Eun Lee, Adam D Catherman, Kenneth R Durbin, Jeremiah D Tipton, Adaikkalam Vellaichamy, John F Kellie, Mingxi Li, et al. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature*, 480(7376):254–258, 2011.

[89] Nertila Siuti and Neil L Kelleher. Decoding protein modifications using top-down mass spectrometry. *Nature methods*, 4(10):817–821, 2007.

[90] Daniel G Bracewell, Richard Francis, and C Mark Smales. The future of host cell protein (hcp) identification during process development and manufacturing linked to a risk-based management for their control. *Biotechnology and bioengineering*, 112(9):1727–1737, 2015.

Chapter 2

Advanced analysis of a series of recombinant therapeutics from the InSCyT platform by LC tandem mass spectrometry

Contributions:
Di Wu: concept contribution, experimental design, experiment procedures on recombinant hGH, GCSF, interferon, data analysis, manuscript writing and revision; Annie Y. Wang: Co-worker, experiment procedures on different recombinant proteins; MIT team: supply of samples from InSCyT platform; William Hancock: goal of the study, concept contribution, manuscript revision

Publications:
Wu.D, Crowell,L.E., Lu,A.E., Love,K.R., Hancock,W.S., Demonstration of Performance and reproducibility of the InSCyT platform manufacture by characterization of multiple protein therapeutic products., *Biotechnology Progress*, (Manuscript in preparation)

Crowell,L.E., Lu,A.E., Love,K.R., Stockdale, A., Timmick, S.M., Wu,D., Wang, Y.A., Hancock, W.S., Braatz, R.D., Cramer, S.M., Love, J.C., etc. (These authors contributed equally to this work)
Pharmacy-scale, on-demand manufacturing of high-quality biologic drugs, *Nature Biotechnology*, (Under Review)

2.1 Abstract

Biosimilar therapeutics are marketed worldwide nowadays and stability assessment is necessary to validate the long term safety, purity and potency of the products. In addition, the manufacture of biopharmaceuticals are not easily controlled from either the upstream or downstream processes. In this study, we have applied yeast (*Pichia*) to produce biopharmaceutics instead of the traditional method using *E.coli* manufacture, to improve product stability with less host cell contamination. The InSCyT (Integrated and Scalable Cyto-Technologies) for Flexible microbial manufacturing platform is both portable and convenient to produce biologics. This generation platform is designed on a demand basis, so the quality obtained from a short time production is important. Our study has described the characterization of purified recombinant drugs from this novel system with coverage from the primary structure identification to possible post translational modifications by sensitive liquid chromatography coupled with tandem mass spectrometry (LC-MS). With the analysis of any product degradations, we observed that the manufactured biopharmaceuticals are at a high-quality level and comparable with the innovator drugs. The quality analysis was used to guide the upstream team to optimize their manufacturing protocol. We have built the resulting methodology for product identification and demonstrated that this final manufacture platform is efficient, convenient and consistent in terms of product manufacture.

2.2 Introduction

Biological therapeutics are a large market for both the academic arena and the clinical and financial field in modern society. However, the global health market for biopharmaceuticals is facing serious issues including drug innovations, methodology developments, quality control and transportation and storage conditions. To efficiently produce new clinical products to meet specific medical needs, however, traditional drug development requires long innovation times and huge financial support.[1] In nowadays, biosimilar drug development has taken an important role in the market with short approval times with FDA review, minimizing cost as long as a highly similar structure and function is demonstrated with regard to the innovator therapeutics.[2, 3] To develop a stable formulation platform is important to produce high quality biopharmaceuticals and minimizing modifications.[4] Recognizing the limitation of traditional manufacture, our goal aims to develop a small-scale platform to produce biological drugs efficiently and convenient for transportation to remote areas. The recombinant proteins are usually produced by *E.coli*, however, the DARPA team has focusing on minimizing host cell contamination, increasing cell density and expression accuracy, and thus have applied the production process to the yeast *Pichia*.

Our collaborators at the MIT lab[1] have developed an effective manufacturing method named Integrated and Scalable Cyto Technologies for Flexible microbial manufacturing (InSCyT) for the Biologically-derived Medicines On Demand (BioMOD) program, aiming to produce high quality biopharmaceutics on demand in a short time. The recombinant DNA manufacturing process contains the upstream production and fermentation process, downstream purification, final polishing and an ultrafiltration process. This platform also applied both on-line and off-line analytics to assure the quality, however, traditional manufacturing of therapeutics require long cycle time of several days. The cultivation fermentation used by our team only consumed less than 24 hours for one dose. In fermentation development, the team evaluated different *Pichia*

strains to produce in high yield therapeutic proteins including, *K. pastoris*, *K. phaffii* wildtype, and *K. phaffii* GS115.[5, 6]

Each step utilized during the production of recombinant proteins will possibly induce modifications and any PTMs and degradations reactions occurring during the cultivation are of concern and must be monitored. The safety, toxicity of resulting product analogs are required to be evaluated based on the structural characterization.[7] As an example, the characterization of recombinant human growth hormone has been traced back to 1970s.[8] The prediction of any post-translational modification and degradation reactions of the proteins target will depend both on the protein structure and manufacturing conditions.[9] For example, in the sequence of human growth hormone, deamidation and oxidation are the major concern for the clinical efficacy and the safety of the therapeutic product.

In nowadays, the high resolution of modern analytical techniques as well as significant market opportunities are attracting more research studies. In this project, we have used liquid chromatography coupled with tandem mass spectrometry to evaluate the quality of yeast based recombinant protein therapeutics. From primary structure to modifications analysis to residual host cell proteins, the follow on products were compared with innovator marketed product samples. We observed that the new DARPA platform has a high level of reproducibility and stability to produce high quality biopharmaceuticals and that the platform could generate multiple protein products efficiently and with high purity. In this chapter, we will demonstrate result from the analysis of three purified protein drugs produced from InSCyT platform to show success of the system.

2.3 Experimental

2.3.1 Chemicals and materials

The chemicals including acetonitrile, hydrogen peroxide, ammonium carbonate sodium phosphate, dithiothreitol, and enzymes (trypsin, GluC and LysC) were analytical grade, obtained from Thermo (Thermo Fisher, USA). The reference standard was provided by commercial companies. The strain recombinant human growth hormone(hGH), GCSF, interferon α2b (IFN) samples were produced by InCSyT platform by our MIT collaborators. Stock samples were stored in -80°C.

2.3.2 SDS-PAGE

The strain hGH samples were standardized to 15 μg per loading based on the concentration provided by the collaborators in order to equalize the loading amount for comparison of proteins samples. The samples were diluted with deionized water and adding 5 μL sample buffer for non-reduced conditions. At the same time, the other batch of samples were prepared with additional 2 μL 1M DTT solution incubated at 90°C for 30 min for reduced condition. The protein mixtures were separated on SDS-PAGE at 160V for 45 min and stained with Coomassie blue. After the destaining process, the bands were monitored by BioRad Image system for further evaluation of the protein quality. In this experiment, an in-gel digestion was not performed.

2.3.3 In-solution digestion

Proteins were dialyzed via 10kD membrane Amicon centrifugal filter at 13,000 rpm for 15 minute and three times in ammonium carbonate sodium phosphate buffer (pH=7).The following in-solution digestion process with Trypsin for hGH and IFN samples, GluC and LysC for GCSF was kept overnight at 37°C. The digestion was terminated by addition of 20μL 5% formic acid. Proteins were then ready for LC-MS analysis, the remained materials were aliquot to 20 μL and stored in -80°C for further analysis.

2.3.4 Peptide assignment by LC-MS

LC-MS analysis used an Ultimate 3000 nano LC pump (Dionex, Mountain View, CA) and self-packed C18 column (Magic C18, 200Å pore and 5 μm particle size, 75μm internal diameter by 100 mm) connected to a coated 10μm internal diameter emitter (New Objective, Woburn, MA). LTQ-Orbitrap XL mass spectrometer was connected (Thermo Fisher Scientific, San Jose, CA) through a nanospray ion source (New Objective, Woburn, MA). Mobile phase A was using 0.1% formic acid in HPLC grade water and mobile phase B was using 0.1% formic acid in acetonitrile. During sample injection, the flow rate was set 250 nL/min with 2% B for 25 min. The flow rate of the gradient was set at 200 nL/min, with mobile phase B, 0-60 min 2-40%, 60-70 min to 90% , 70-75 min 90% and 75-78 min 2%. The mass spectrometer was operated in a data dependent mode to switch between MS and CID-MS^2. Briefly, after a full-scan MS spectrum from m/z 400-2000 in the ion-trap, 8 CID-MS^2 activation steps were performed on the 8 most intense precursor ions from the full scan.

For peptide identification, raw data were searched against human Growth Hormone, GCSF, interferon α 2b sequence using in BioPharma Finder 2.0 software (Thermo Fisher Scientific). For peptide mapping,

searches were performed using a single-entry protein FASTA database with oxidation and deamidation set as variable modifications, 20 ppm mass accuracy, and a confidence level of 0.8 for MS/MS spectra. Final confirmation of the peptide identification was determined by manual inspection, extracting the base peak from the chromatogram and matching the MS^2 fragmentation data with theoretical prediction. The modification percentage was calculated by peptide peak area. Both non-degraded and degraded peptides were evaluated in the same LC-MS analysis with characteristic m/z differences and elution time shifts. The quantitation of the degradation could be calculated as the following: ratio of Degradation (such as oxidation, deamidation, etc) = peak area of [(degraded peptide)/(degraded peptide) + (Non-degraded peptide)] $\times$ 100%.

In this study, the biosimilar products from the advanced InSCyT platform were analyzed for not only the primary structures, as well as comparison of degradation products vs the standard to understand the purity profile to reach the FDA approval. For further stability studies, we developed an integrated stability method development and analytics which will be described in later Chapters. As the post-translational modifications are related with the efficacy,[10, 11] stability or safety of the therapeutic drugs, the evaluation of the oxidation, deamidation and other variants are commonly measured.[12] As concerned, the N-terminal truncation of GCSF and IFN and the two-chain variants and oxidations of hGH are on our analysis priority.

2.4 Results for GCSF products

The first drug produced from InSCyT platform is the recombinant GCSF, which is a protein with 175 amino acids. The GCSF products were digested with GluC combined with LysC to improve the digestion efficiency of GluC.

2.4.1 Primary structure identification

The N-terminus of GCSF product has two significant modifications, the truncation resulting from gene expression in the host organism and oxidation of the first residue methionine. The HPLC comparison of each products and reference material are shown as Fig.2.10. There protein samples are observed with full sequence coverage and comparable with the control. There are a total 12 cleaved peptides that were detected. The N-terminal peptide MTPLGPASSLPQSFLLKCLE with a m/z 1067.07($^{2+}$) eluted at 29.64 min and was detected with the highest abundance in the reference sample, however, the G1 peptide (first peptide from GluC digestion) in GCSF products was not observed with the highest signal due to modifications occurring in the biosynthetic process. As an example of a less successful fermentation the product GCSF#2 was produced with a low concentration and its MS signal abundance was lower and with incomplete sequence coverage. The separated peptides were eluted based on their hydrophobicity, and with a consistent retention time in the four samples, and the results of peptide mapping has been summarized in the supplementary data Table.2.1. The products were produced at different times but the reproducibility of the manufacture platform was proven to be robust.

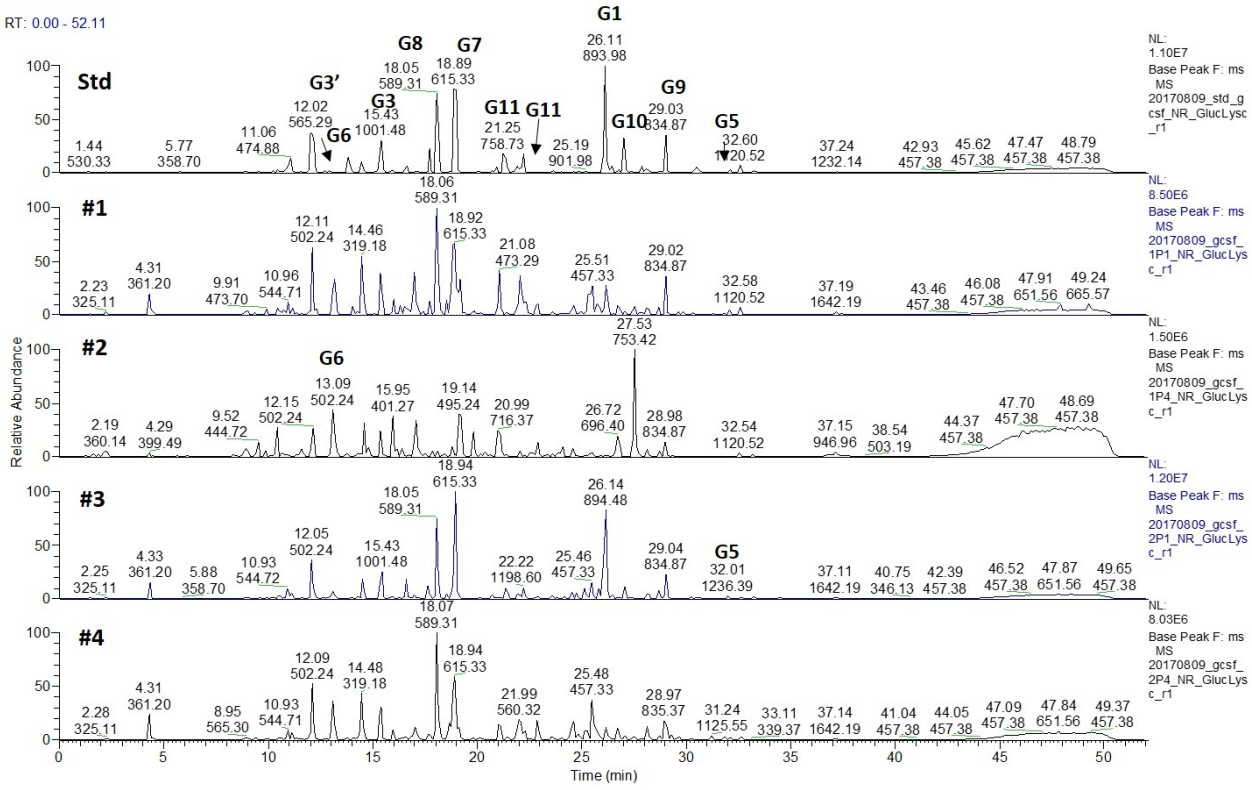

Fig 2.1: HPLC base peak of strain GCSF samples

2.4.2 Variants of GCSF products

2.4.2.1 Determination of oxidation

The sequence of GCSF starts from a N-terminal methionine residue that is located on the outer space of the structure. This N-terminal methionine is the easiest residue to be oxidized. There are three other methionine residues, Met122, Met127 and Met138 located at the inner side of the structure. The tendency of oxidation of these three methionines are various based on different buffer formulations and storage conditions and Met127 and Met138 may switch in the tendency for higher oxidation levels depending on the conditions. However, Met1 is the most likely oxidized residue,[13] with the crucial first residue position, and thus the N-terminal peptide measurement was more complex than for other modified peptides.

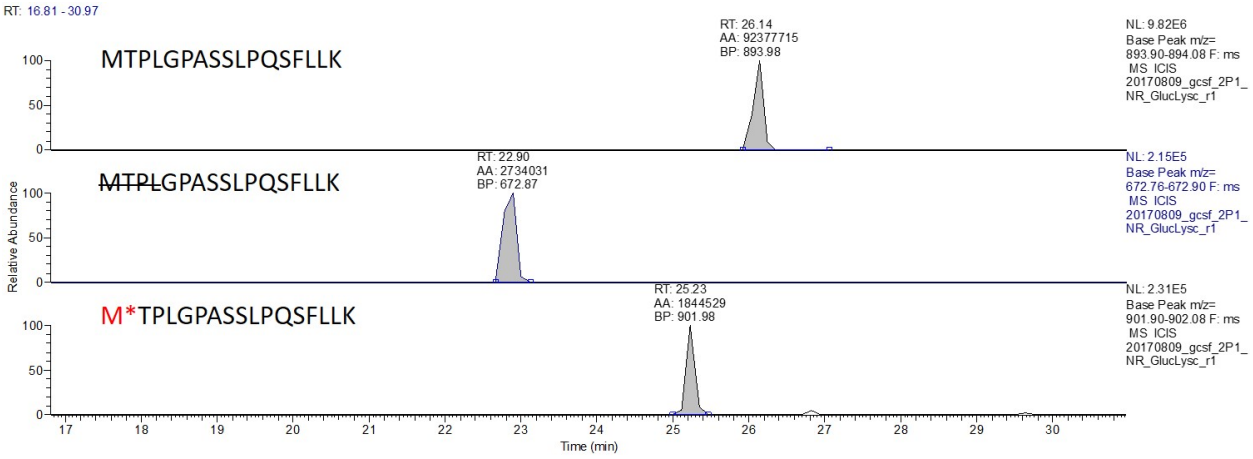

Fig 2.2: HPLC base peak of strain GCSF

As shown in Fig.2.2, the HPLC base peaks were shown for the N-terminal peptide. Not only the modification was demonstrated, but also the truncation of N-terminal residue was determined. The expected N-terminal

peptide with the mass of 893.98(2+) was eluted at 26.14 min, and the oxidized variant was eluted at 25.23 min, while the truncated N-terminal peptide with the m/z of 672.87(2+) was eluted at 22.90 min. The integrated peak area were calculated to give a 2.9% variant composition and used to guide the upstream team to optimize the generation process. The extracted fragment ion peaks were illustrated in Fig.2.3 and show peaks with a high confidence of identification.

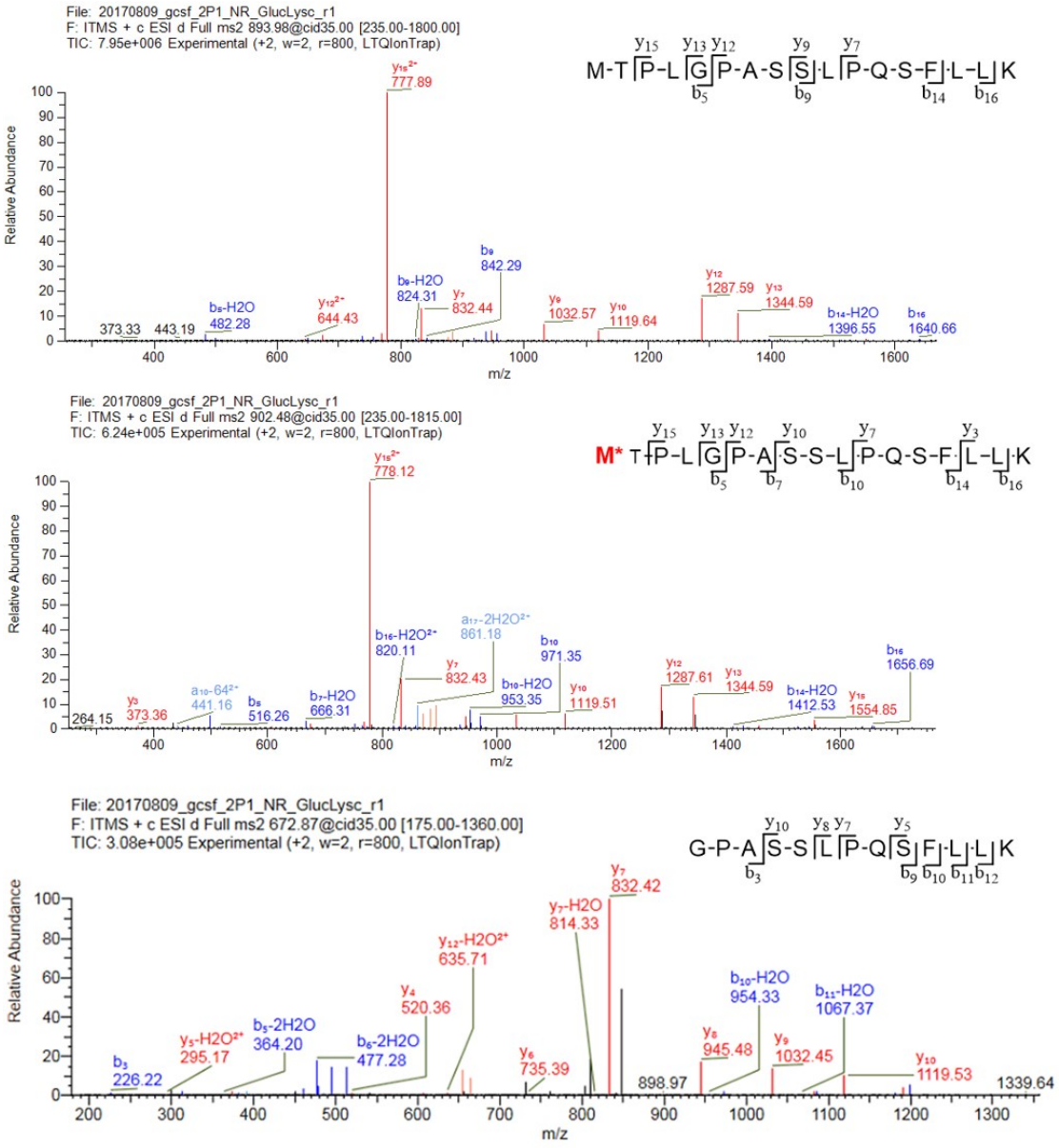

Fig 2.3: N-terminal peptide characterization, MS2 identification of peptide MTPLGPASSLPQSFLLK

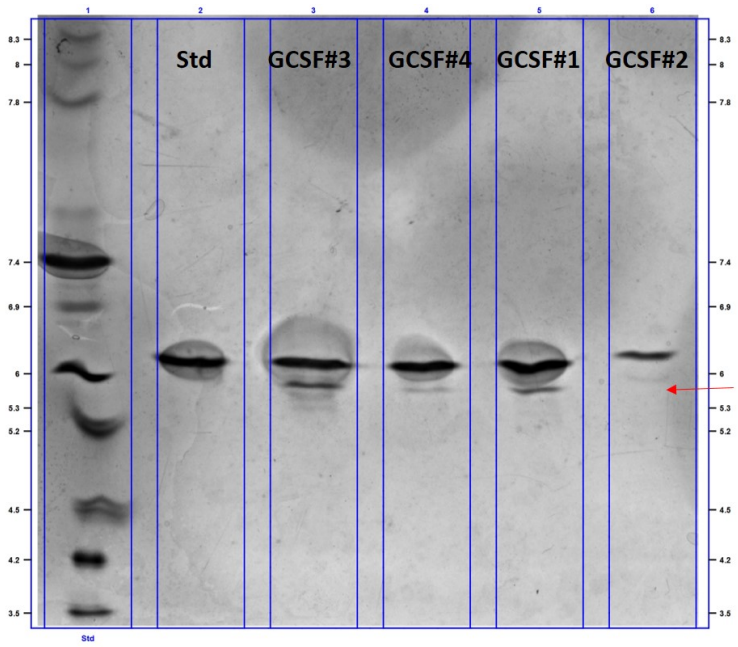

Fig 2.4: IEF gel image of strain GCSF samples; Lane1: marker; Lane2: standard GCSF; Lane3: GCSF#3; Lane4: GCSF#4; Lane5: product GCSF#1; Lane6: GCSF#2

The missing residue which resulted in a protein variant with a lower pI is shown in Fig.2.4 with the detection of significant additional bands. This variant band indicated lower pI of the product, illustrating the N-terminal truncation may induce the acidic variants or there are other impurities in this protein. From this point of view, the manufacturing team should be aware of the contamination in the fermentation process and to minimize the impurity level.

2.4.2.2 Determination of O-glycan variants

During the fermentation production, the reference GCSF sample corresponding to Lenograstim was derived from Chinese hamster ovary (CHO) cells and glycosylated and the sugar chain has been first studied by Ohdea.[14] Two structures were obtained and analyzed by NMR. The glycosylation has been proved to be of limited clinical relevance.[15] In our products from InSCyT platform, the existence of possible O-mannose

have been studied. The experimental data was shown in Fig.2.5, which illustrated the residue T134 on peptide LGMAPALQPTQGAMPAFA which has potential glycosylation. The extracted ion peak for the non glycosylated peptide with the m/z of 866.44(4+) and the detected glycosylated peptide with m/z 1048.49(4+) was eluted at 25.81 min, earlier than the non-glycosylated peptide eluting at 27.06 min, according to the relative polarity.

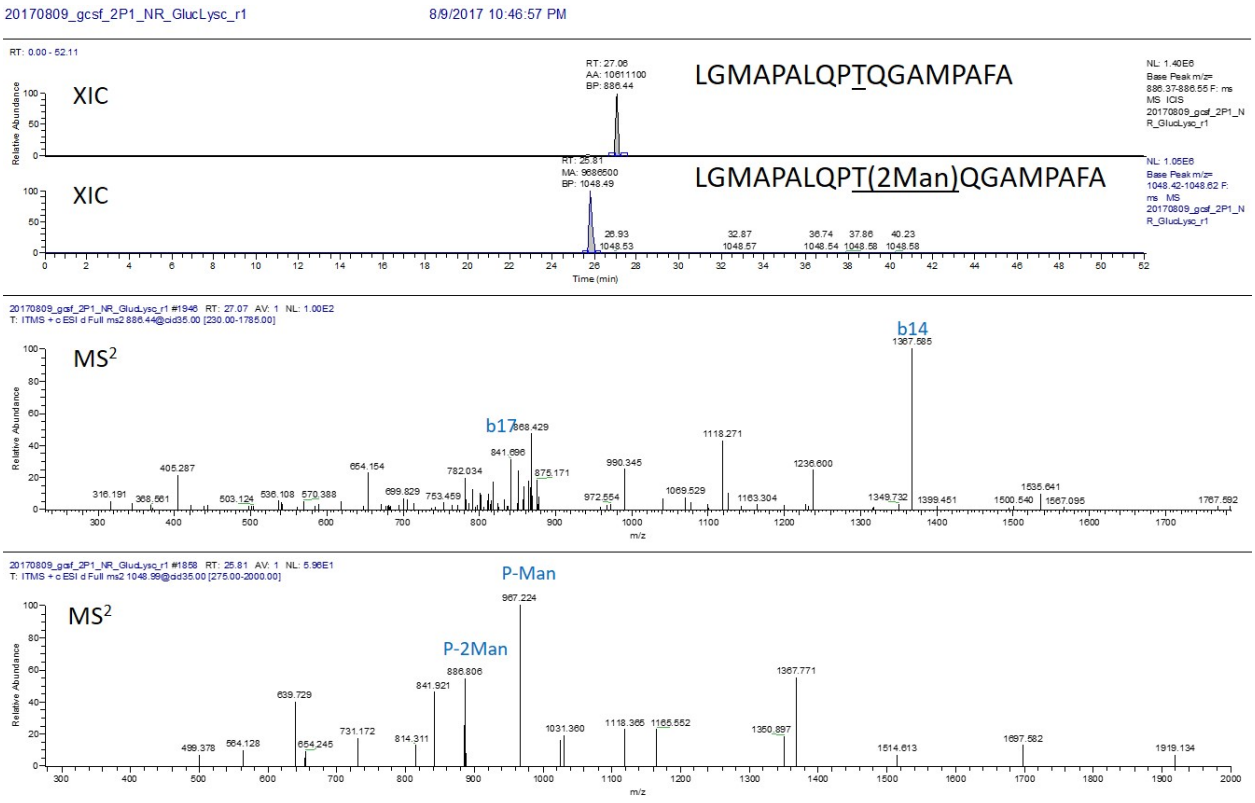

Fig 2.5: Analysis on O-glycosylation of strain GCSF products; XIC base peak and CID-MS2 of normal form and glycosylated form of peptide "LGMAPALQPTQGAMPAFA"

The figure shows a labeled fragmentation pattern for the fragment ions of the non-glycosylated peptide and the glycosylated peptide with two mannose residues. The structural heterogeneity of the peptide leads to a relatively low signal of the target peptide, however, we have detected this modification consistently in all

strain products except the #4 sample. Therefore, we could conclude that the potential glycosylation structures produced by the MIT platform contained two mannose residues in the GCSF products and for final product characterization further analysis are needed for a precise structural confirmation and ratio of the glycans. The existing glycan ratios in each product were varied, which will provide guidance to the manufacturing team to optimize the control conditions. All the modifications are summarized in supplementary data Table.2.2. The T134 mannosylation was at a high level, present in above 50% in the strain samples.

2.5 Results for Interferon α2b products

The second product from the final test of the InSCyT platform is Interferon α2b (IFN). The purified samples were concentrated and collected by centrifugal filters to obtain a suitable amount for HPLC analysis.

2.5.1 Primary structure identification

Interferon α2b is a protein with 165 amino acids, including two pairs of disulfide bond, the Cys1-Cys98 and Cys29-Cys138. The samples and the standard were treated under identical conditions as same with trypsin in-solution digestion. The N-terminal sequence often contains the signal peptide of a protein and also know as the leader peptide. In the biosynthetic process this short sequence will promote the cell to translocated the newly synthesized protein to the cellular membrane and help secretion. In the strain interferon products, residual leader peptide EEGVSLEK with a mass of 445.72 (2+) was identified. This leader peptide was eluted at 12.72 min in the strain products, and was not detected in the reference material. The HPLC separation chromatogram is shown in Fig.2.6, and the peptide mapping result is shown in supplementary data Table.2.4.

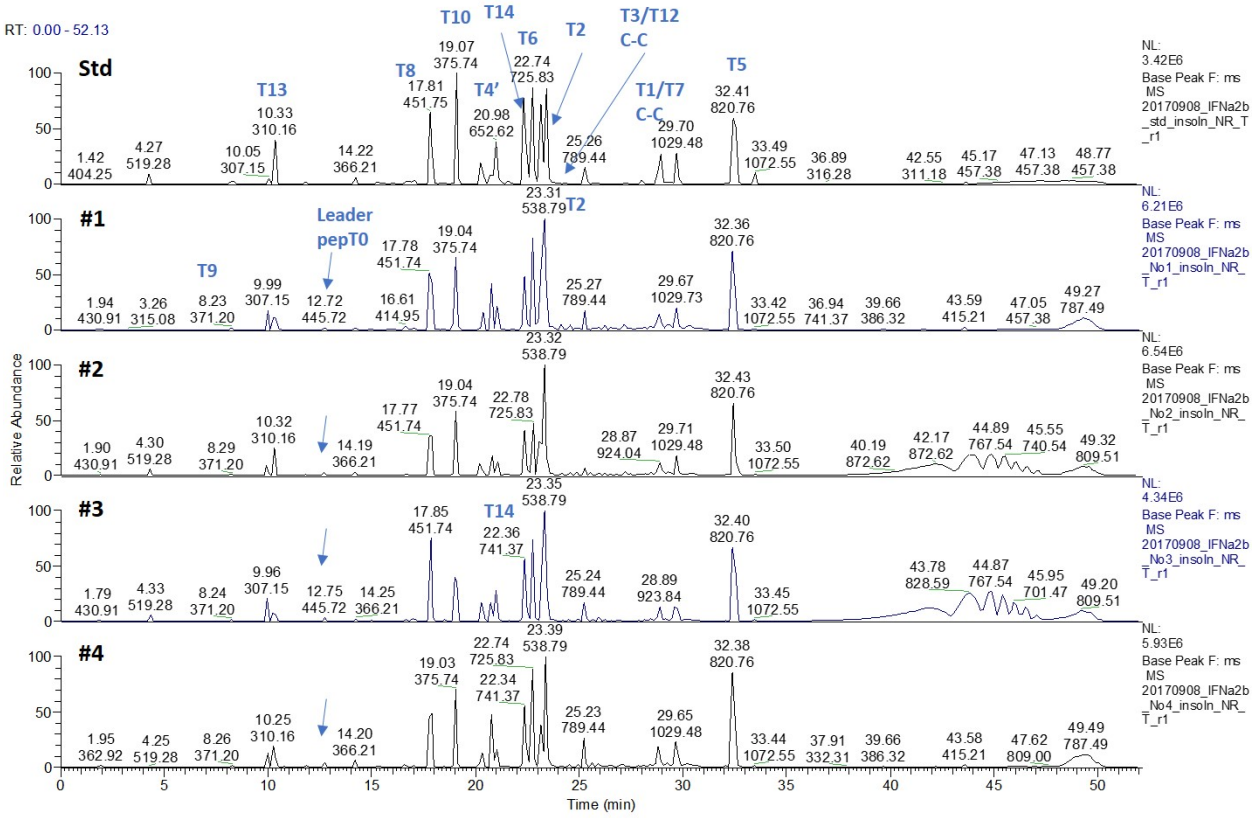

Fig 2.6: HPLC base peak of strain Interferon samples

2.5.2 Variants of IFN products

2.5.2.1 Determination of leader peptide sequence

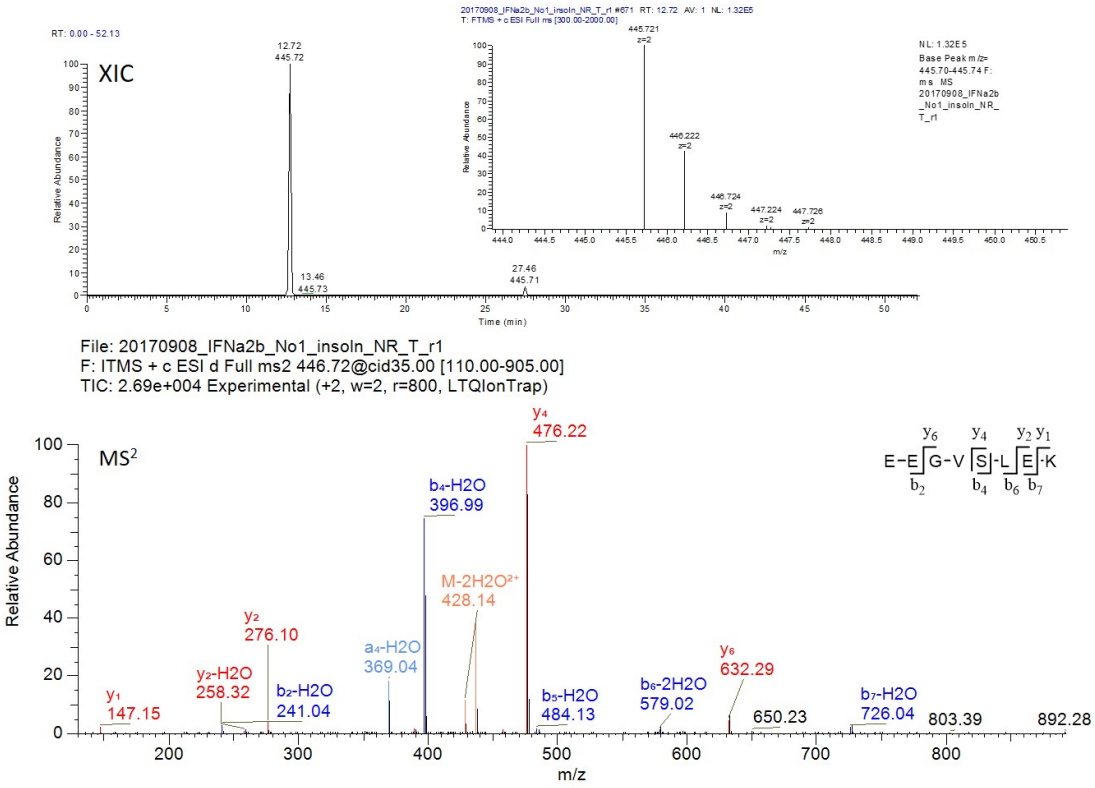

Fig 2.7: LC-MS analysis of leader peptide EEGVSLEK for the strain Interferon sample; Top figure: XIC of leader peptide of 446.72^{2+}; Bottom figure: CID-MS2 assignment of leader peptide

In our previous study, the leader peptide was identified as EEGVSLEKR, however, in this study trypsin was used to cleave the protein to prepare a digest for LCMS analysis and the terminal amino acid R was not detected. Therefore for a successful analysis we needed to select another enzyme to apply to a further study to specifically monitor the existence of an additional residue in the signal peptide. Even if the length of leader peptide has no effect of the function of a protein,[16] the upstream manufacturing team for effective fermentation development needs to monitor any changes in N-terminal sequence. In Fig.2.7, the extracted

ion chromatogram of the m/z 445.72($^2+$) ion is demonstrated, and the fragment ion peaks are shown. The y_1, y_2, y_4 and y_6 were detected, and the b_2, b_4, b_6 and b_7 were monitored as well. The leader peptide was measured with a level of 2.1% in strain interferon IFN#1, and raised up to 4.6% in IFN#4. This accumulation could be caused by an increase in the fermentation time to achieve an increased product yield.

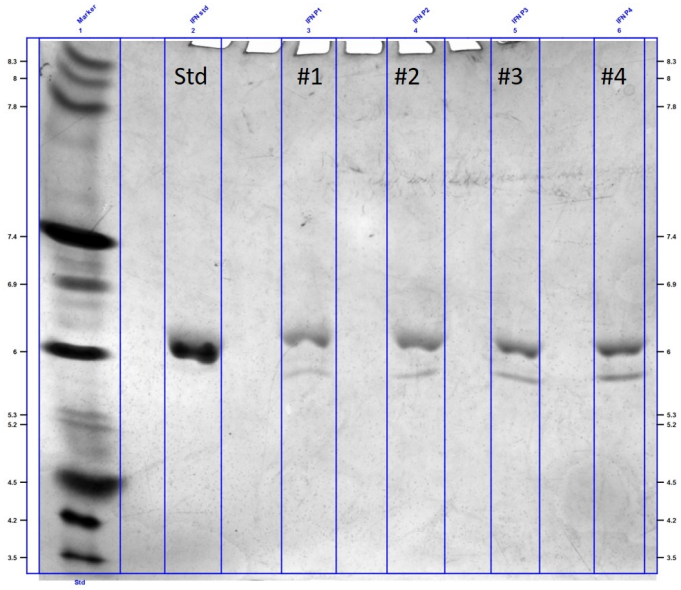

Fig 2.8: IEF gel image of strain Interferon samples

The variant of the leader peptide sequence resulted in a change of the pI of the protein and variant bands were detected in a gel analysis, shown in Fig.2.8. The additional N-terminal residue Glu (E) with an additional acidic carboxyl group lead to lower observed pI. The theoretical pI reported on Drugbank of interferon α 2b is 5.99. The major band of the biologic interferon products was observed with a pI at 6.0 and was comparable with the drug standard although migration values varied slightly for each lane. The amount of acidic bands observed from #1 to #4 was based on the darkness of the bands, and the observed result could be related to the date of each production batch and longer fermentation times could result in accumulated variants and thus be an issue in the future cultivation process. Besides, due to the lack of a potential deamidation residue

in interferon sequence in the appearance of an additional acidic bands must be due to another mechanism such as an altered amino acid sequence.

2.5.2.2 Characterization of Oxidation

There are total of five methionine residues in the interferon sequence, with locating ranging from the surface to the hydrophobic core, and the entire sequence contains residue Met16, Met21, Met59, Met111 and Met148. The peptide T2 of interferon TLMLLAQMR contains two potential oxidized methionine residues, Met16 and Met21. The mass spectrometric data was used to differentiate the site of a specific oxidized residue and as shown in Fig.2.9, a peptide with the m/z 538.79(2+) was eluted at 23.31 min and two oxidized peptides with the same m/z 546.79(2+) were eluted at 20.50 min and 21.47 min. The assigned MS2 ion fragments confirm the presence of an oxidized residue. The Met16 oxidized peptide was not detected as the 2+ ion, however, the 1+ fragment was monitored. The detected b and y ion illustrated the modification site which allowed determination of the oxidized site and the amount.

The other potential oxidized residues have been characterized with the same data analysis method and the data are summarized in supplementary data Table.2.4. The residue Met111 is the least likely candidates to be modified.[17] As a result of our analyses we have concluded that the purified products from InSCyT platform have comparable quality to the standard reference material.

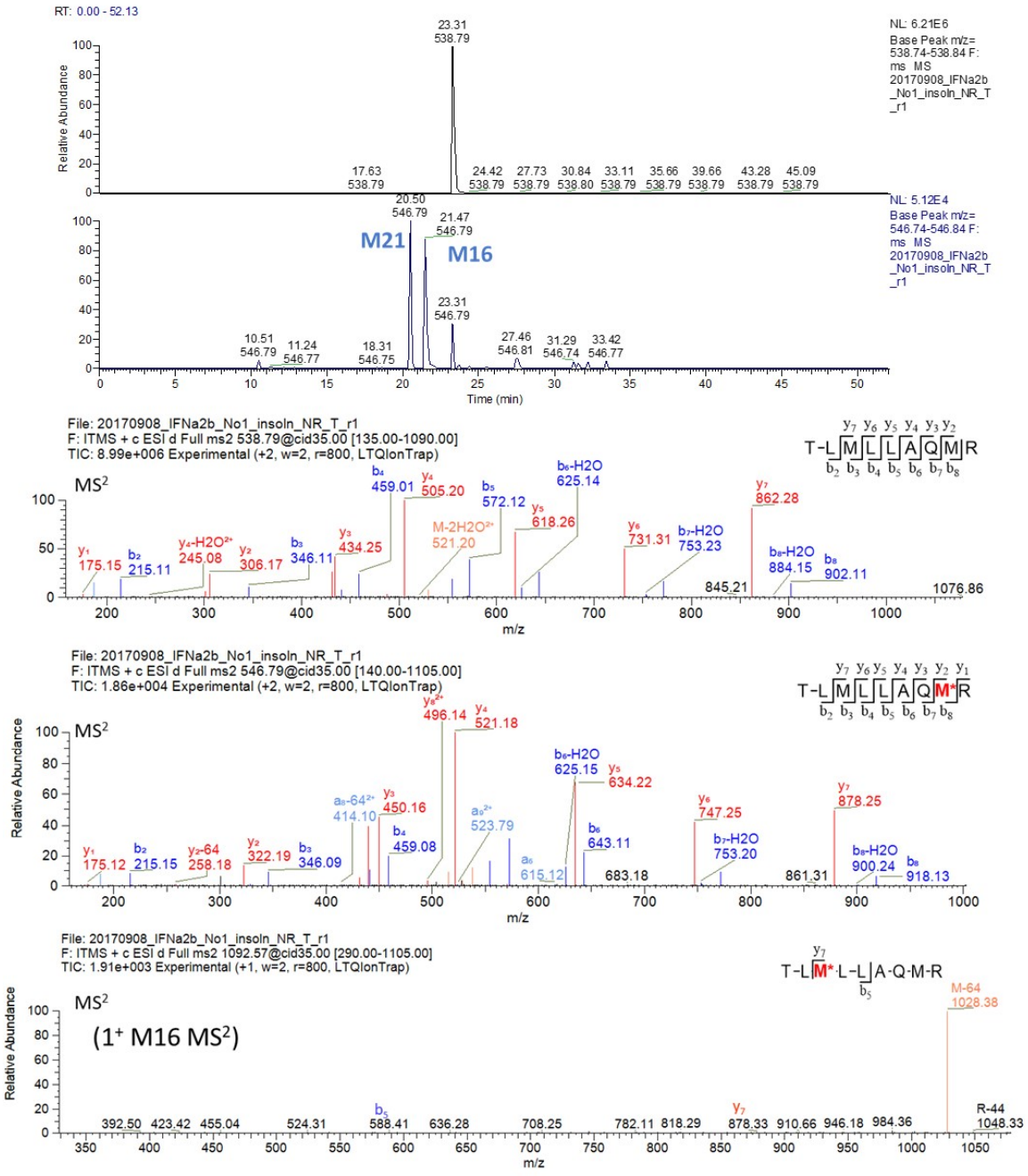

Fig 2.9: LC-MS analysis of strain Interferon oxidation at Met16 and Met21; Base peak of non-oxidized peptide TLMLLAQMR and oxidized peptide; CID-MS2 peak assignments of non-oxidized and oxidized peptide

2.6 Results for strain hGH products

The last protein product for evaluation of the platform is human growth hormone (hGH) which is essential for the treatment to children with growth hormone deficiency or hypopituitarism. hGH has been studied for a very long time for this manufacturing and production system. For the final product characterization, the MIT platform was able to manufacture hGH to achieve a product level of 5mg/mL that was the same unit level as with the commercial drugs. The products have been treated with tryptic digestion and without reduction of the disulfide linkages.

2.6.1 Primary structure identification

Recombinant human growth hormone (rhGH) was produced by yeast expression (*Pichia*) with an identical sequence for human growth hormone with 191 amino acids. The intra-disulfide linkages are Cys53-Cys156 and Cys182-Cys189[18] as the precise sequence and disulfide linkage are critical to the function and toxicity of biopharmaceuticals.[19] In the analysis, the 191 amino acids sequence would be cleaved to 21 peptides by trypsin digestion.

We summarized the peptide mapping in Table.2.5 (supplementary data) for the primary structure identification. In the analysis all the expected peptide sequences may not be cleaved theoretically, and thus the missing residues are observed adjuncted with adjacent peptides, such as with peptides T3-T5, T17-T19 and need to be identified manually. In this section the hGH amino acids sequence is provided to show the connection of the intra disulfide bonds and potential modification sites.

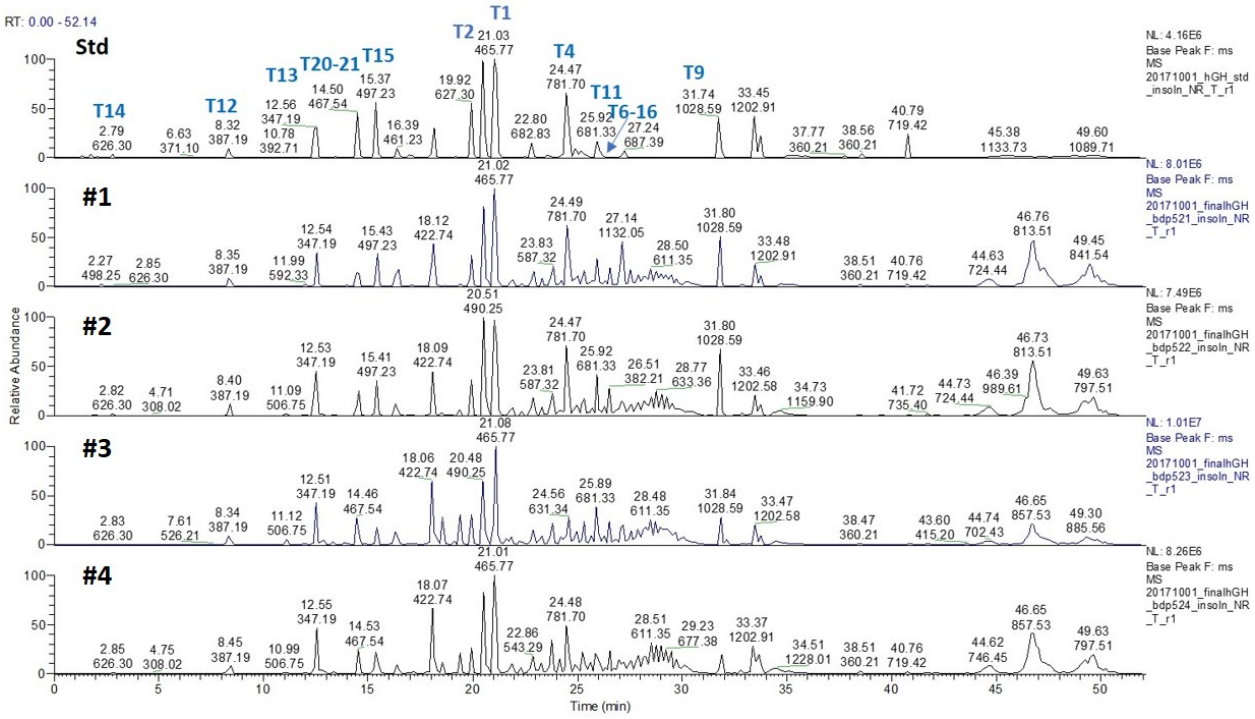

Fig 2.10: HPLC base peak for the rhGH samples

The HPLC comparison of each products and reference material are shown as Fig.2.10. The peptides were eluted based on their size and polarity, the least RPLC retained peptides are T3 and T5, which only contain three amino acids and would not retained on the reversed phase column. The peptides were eluted with a consistent retention time in the four samples, and the peptide mapping results have been summarized in Table.2.5. The N-terminal peptide FPTIPLSR which was observed with a m/z 465.77($^2+$) and eluted at 30.19 min and was detected with the highest abundance. The manufacture platform was proven to be robust as the growth hormone products were produced with a high degree of reproducibility at different times of fermentation.

2.6.2 Variants of hGH products

For the concerns about the effects of degradation reactions on toxicity and efficacy for patient therapy it necessary to not only confirm primary structure but also measure degradative reactions such as oxidation.

2.6.2.1 Characterization of Oxidation

There are three methionine residues in the hGH sequence, which are the Met14, Met125 and Met170. Located at the surface area, the residue Met14 was the mostly easily oxidized methionine and the Met170 is located in the core area and resistant to oxidation and thus a crucial indicator of the protein structure.

The oxidized peptides will have additional mass of 15.994 Da for a 1+ oxygen ion or 7.997 Da for 2+ oxygen ion than the corresponding non-oxidized peptides. For the characterization method, the base peak from HPLC separation as well the extracted ion peaks for the oxidized species were analyzed. As shown in Fig.2.11, with the different polarity, the oxidized peptides have eluted earlier than non-oxidized peptides on reverse phase liquid chromatography, which could be differentiated by the base peak as well.

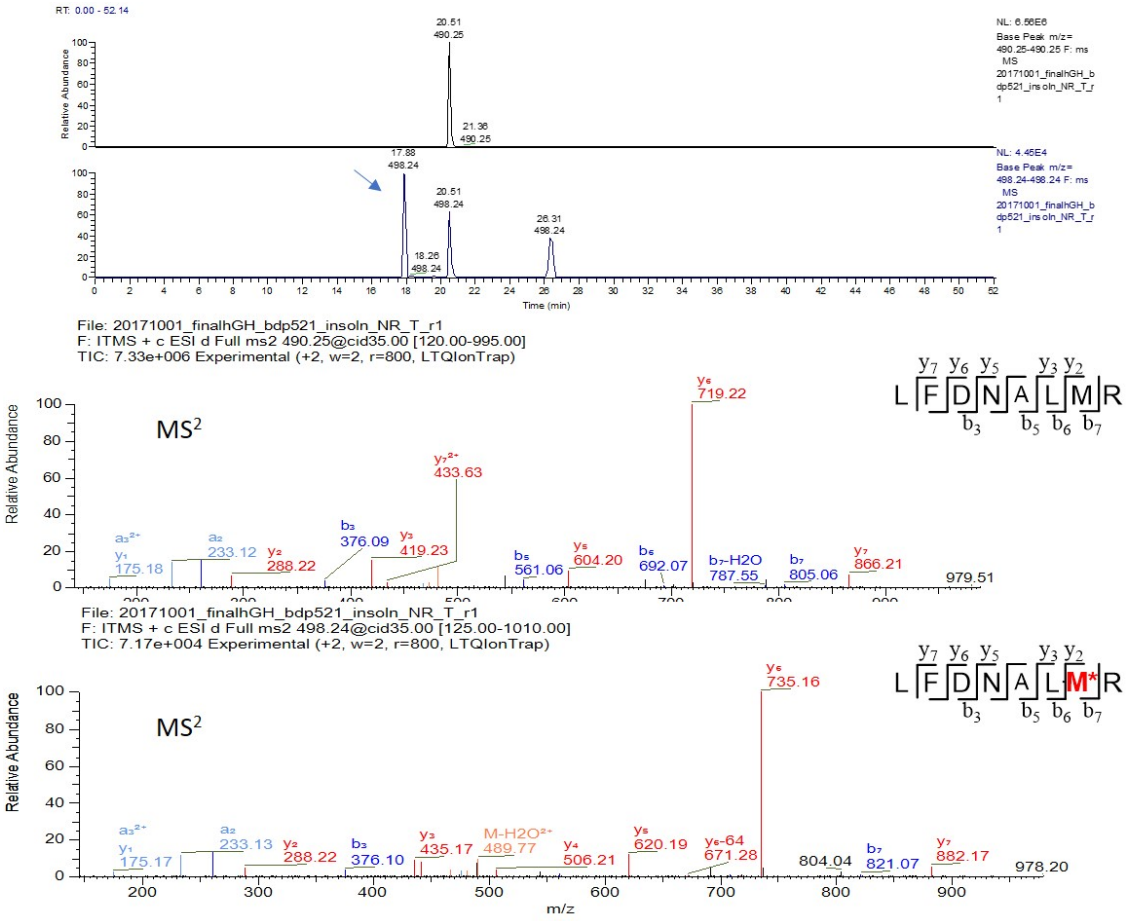

Fig 2.11: LC-MS analysis of the oxidized peptide (T2) from the tryptic digest rhGH

2.6.2.2 Characterization of Deamidation

Deamidation of the residue Asn149 was well characterized in our previous study[20] and we have developed an optimized analytical method to determine the nature of the degradation reaction. The amount of deamidation monitored in this study based on the peak areas was low at 0.2% in the drug standard and lower than 1% in the products and corresponded to the retention time and mass observed in the previous studies (Table.2.6). However, the low levels of this modification resulted in the inability to observe the corresponding MS/MS fragmentation pattern. In addition, the corresponding IEF gel image (Fig.2.13) supported this conclusion with the observation of an additional minor band at a more acidic pI of 5.2.

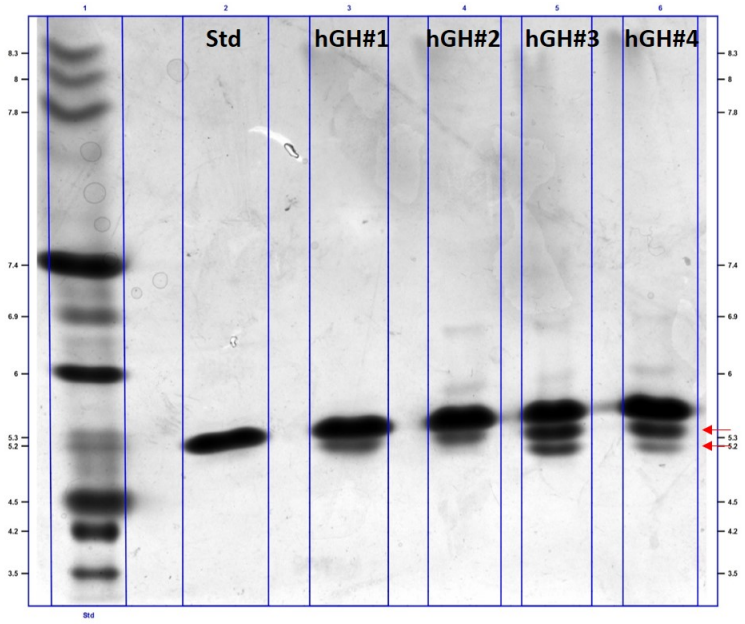

Fig 2.12: IEF gel image of strain hGH products

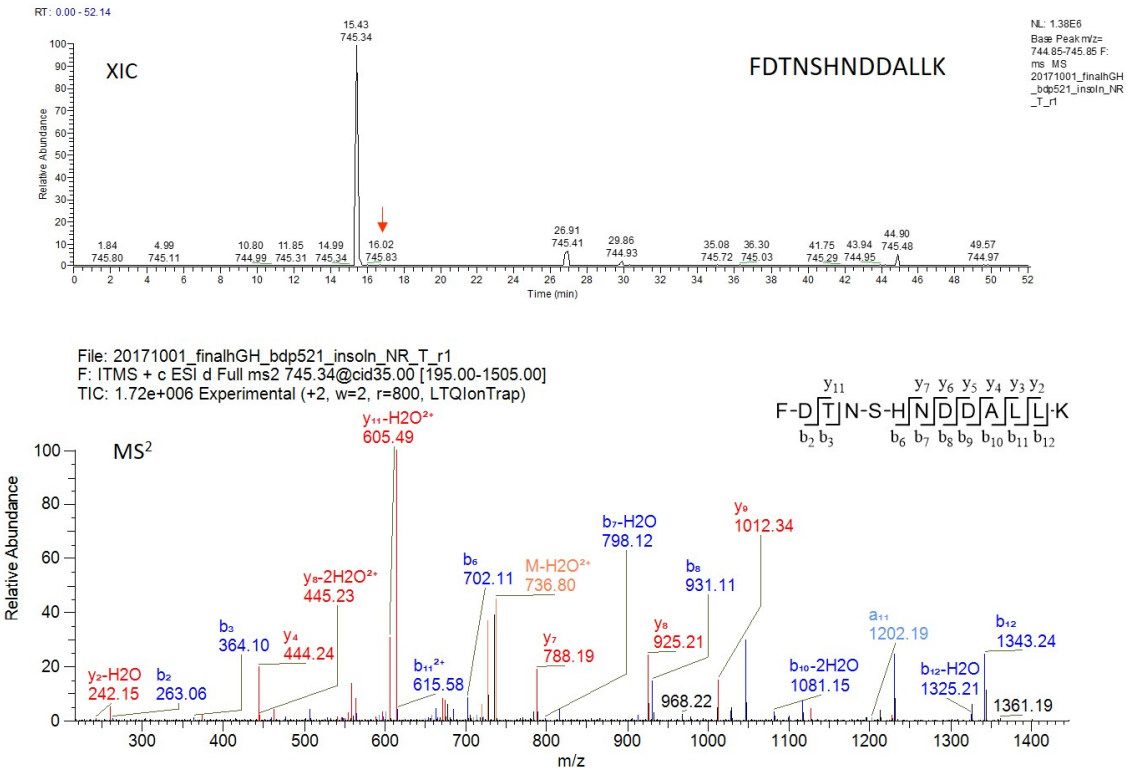

Fig 2.13: LC-NS analysis of the peptide T15 (FDTNSHNDDALLK); Top: XIC of peptide with m/z 745.34($^{2+}$); Bottom: CID-MS2 peaks assignments

In this particular biological product sample set, we didn't detect fragment ion peaks but the variant could be present at a trace amount and a lower level than the mass spectrometer detection setting range. Because the extracted base peak of the target peptide was detected based on previous experience, and the detection of CID MS2 assignment from previous study is shown in Fig.2.18(supplementary data). The variant data was summarized in Table.2.6 (supplementary data). The amount of deamidation monitored in this study based on the peak area measurement was 0.2% in the drug standard and lower than 1% in the products. The IEF method was applied as well and the gel image was shown as Fig.2.13, the deamidation of asparagine will decrease the iso electric point of the protein. The additional faint band of variants at pI 5.2 were detected and reported by order of batch time.

In our previous study, the nondeamidated and deamidated peptides (T15) were identified through the accurate mass assignment by LC-MS. The deamidation mass shifts are the result of an increase of 0.9840 Da for 1+ ion and 0.4920 Da for 2+ ion as a result of the side chain changing of the deamidated peptide from asparagine to aspartic acid, NH_2 substituted by OH. The peptide LFDNAMLR was detected with the most abundance 2+ with 745.35, and the second high abundance 3+ with 490.25 and the deamidated peak was eluted later than the nondeamidated peptide. For specific hGH deamidation analysis, there are still ETD fragmentation techniques that could be applied in the future study.

2.6.2.3 Characterization of Disulfide linkage and two-chain cleavage

There are two disulfide linkages that exist in the hGH structure, that is the linkage of Cys53-Cys165 and Cys182-Cys189 . To precisely characterize the connection, non-reduced conditions for enzyme digestion should be applied to assure the presence of the intact structure. In this research, CID fragmentation was used instead of ETD. The mass assignment of the strain samples was compared with the commercial standard.

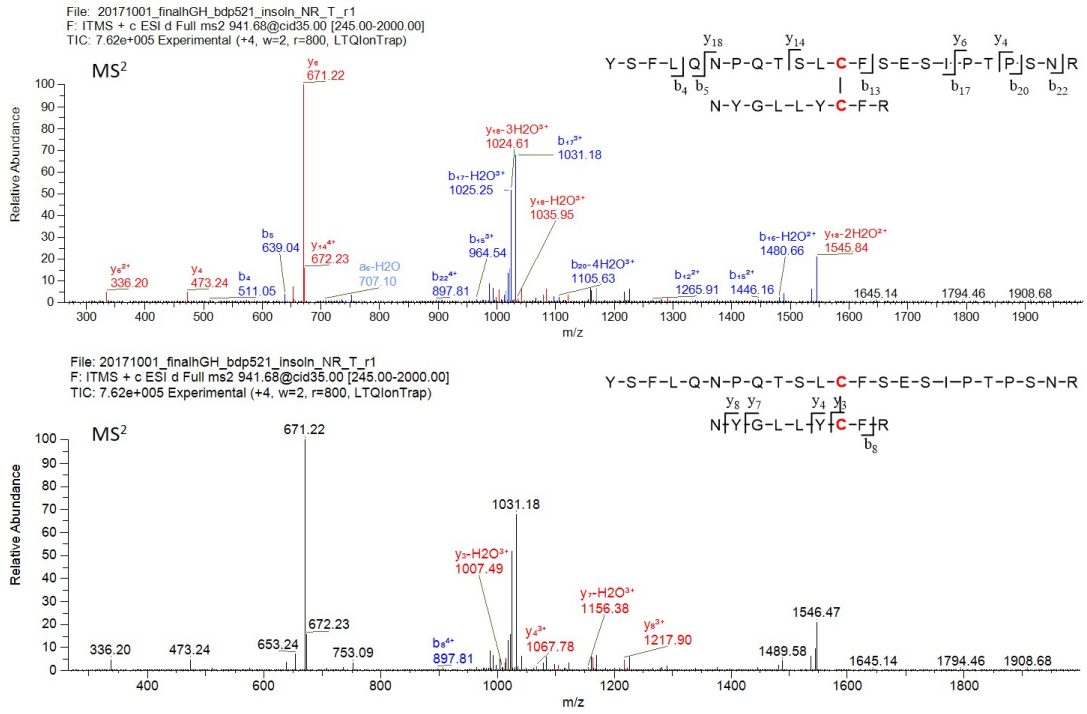

Fig 2.14: LC-MS analysis of the disulfide linked peptide (Cys53-Cys165, T6-T16) from the tryptic digest rhGH

The connecting peptide with the m/z 941.71 for the 4+ ion, was eluted at 26.05 min. With the intra disulfide bond, the mass will be expected for the theoretical total mass minus 2 hydrogen atoms. The ion fragments were assigned to each backbone of the disulfide linkage in the CID fragmentation mode. In Fig.2.14, the major peaks of mass 671.22, 1031.18, 1545.84 are y6 and b17(3+) from T6 as the backbone. Fragments with the mass 897.81, 1067.78, and 1217.90 were detected as b8(4+), y4(2+) and y8(3+) from T16 as the backbone. Therefore, the observed fragment ions from the expected connecting peptides confirms the correct structure for the protein.

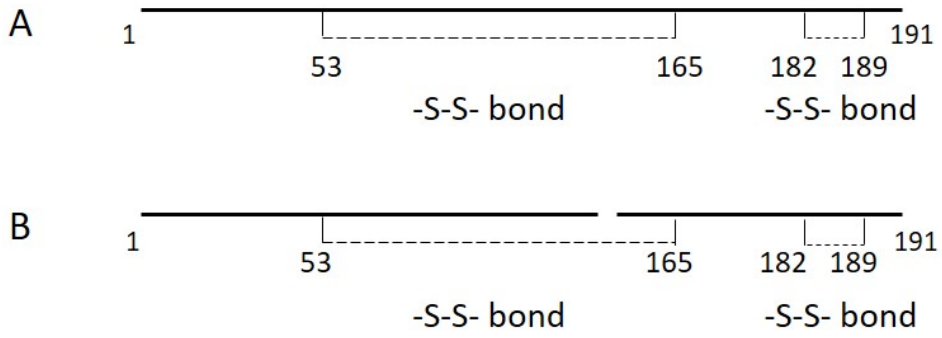

Fig 2.15: Scheme of hGH sequence structure; A: Normal hGH structure; B: Two-chain hGH structure

There are reported studies on the observation of two-chain variants of growth hormone and the effects this modification on protein function, safety or clinical efficacy. A scheme of hGH sequence and two-chain structure is compared shown in Fig.2.15. The two-chain cleavage sites may not be reproducible by different cultivation methods due the secretion of different proteases. In previous studies,[21] the cleavage site was determined as residue T142, however, in this experiment with *Pichia* as the host organism, we detected the cleavage site is at residue Q141, which is reproducible in different fermentation samples, as shown in Fig.2.17.

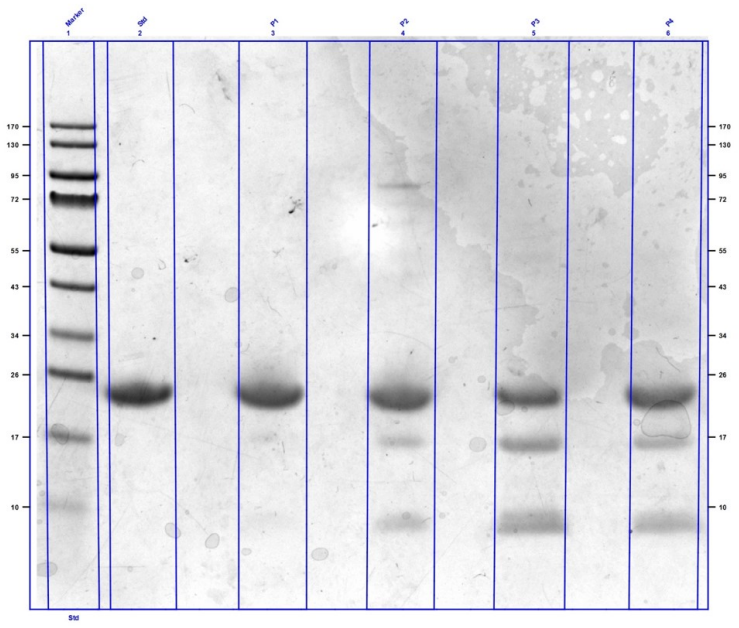

Fig 2.16: SDS-PAGE gel image of strain hGH products

The SDS-PAGE gel images of the strain hGH samples with reduced condition are shown in Fig.2.16. As the gel image indicated, the protein samples did not contain a significant level of covalent aggregates, but as the bands showed in the figures, there are fragments or two-chain variants observed. The identification of the variants and cleavages were described in detail in the following discussion. The gel images showed the major band with a molecular weight of approximately 22kDa with both standard and strain products. With the disulfide bonds cleaved by the reducing reagent, the major bands representing the intact protein sequence exhibit a slight increase in apparent MW due to an increase in hydrodynamic volume while the fragment bands at around 15kDa and 7kDa are observed as faint bands in the nonreduced gel, and get darker under the reduced conditions.

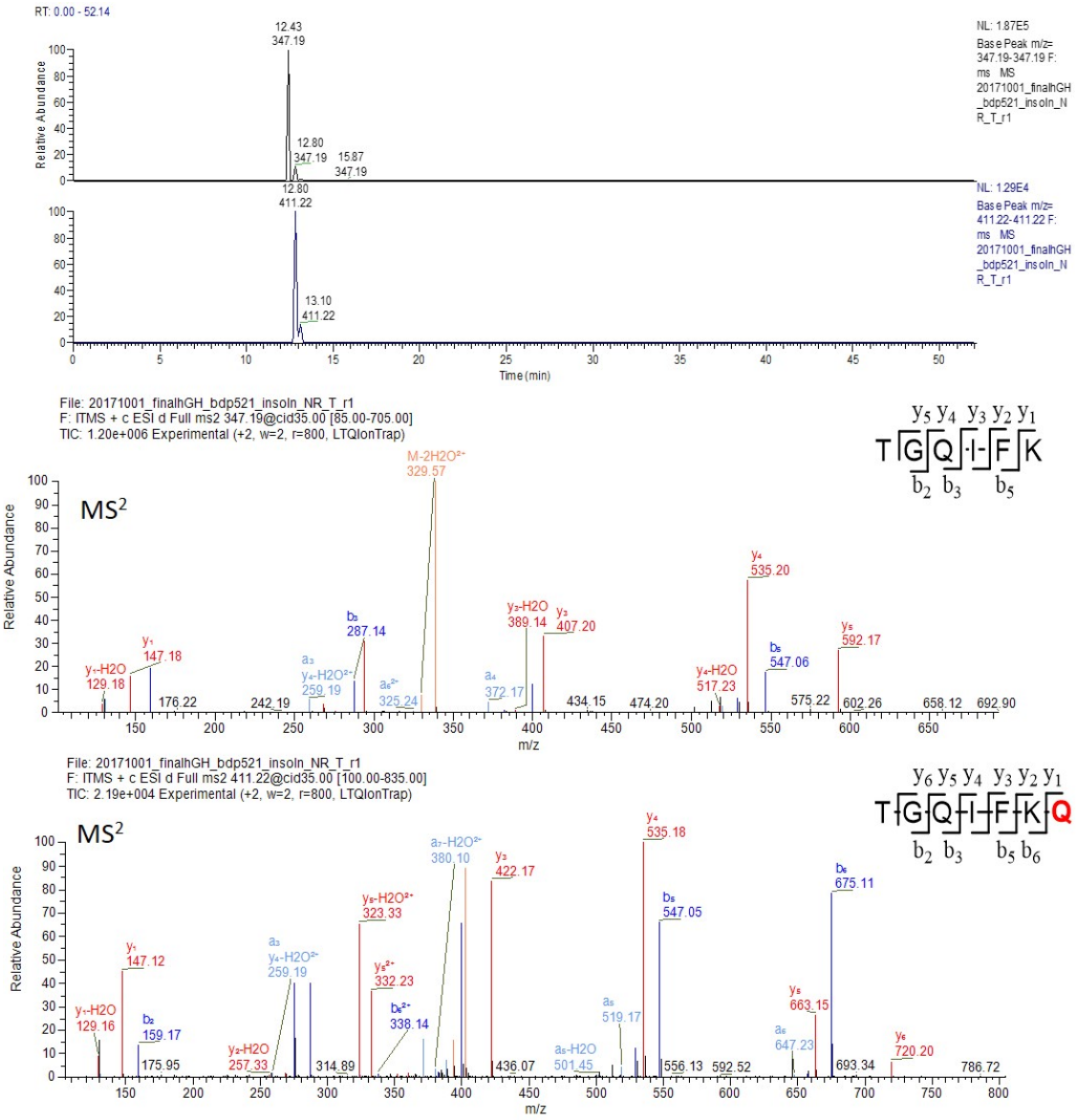

Fig 2.17: LC-MS analysis of the two-chain cleaved peptides from the tryptic digest strain hGH; XIC base peak of T13 with m/z 347.19(2+); T13+Q 411.22(2+); and the corresponding CID MS2 assignments

The precursor ion of T13 is 347.19(2+), eluted at 12.43 min; . The variant cleaved at Q134 peptide has the precursor ion with 411.22(2+) and elutes at 12.80 min. The fragmentation pattern of CID with b ion and y ion were labeled and indicated in the figure.

To characterize these variants, we needed to collect the peptide data of the predicted peptide with trypsin,

that is peptides T13 and T14, and the non-specific cleaved peptide. The specific protease site is the lysine (K133), on the peptide TGQIFK, however, the peptide of TGQIFKQ, which cleaved at residue Q134 has also been detected. From the CID-MS2 spectroscopy, the b ions and y ions of the peptides have been identified with high confidence level. The peak area of the variant and the amount of the variant are calculated with the assumption of a similar MS sensitivity.

2.7 Conclusions

Recombinant protein production and manufacture is a popular topic with many research studies. The manufacturing stage is a huge issue that needs the development of efficient techniques with high quality. In remote areas or at emergency situations, to requirement support the patient in their urgent need and to avoid fatal issues are very important. We have demonstrated the capability of the *Pichia* based production on the of InSCyT platform to operate with consistency and reproducibility. This bench size system is capable of generating multiple biopharmaceutics in a continuous production mode. Comparing with the drug quality level required for FDA approved, our products have been produced in a sufficient amount, full sequence coverage, and low level of host cell proteins. Although modifications such as oxidation were generated in the fermentation process, the quantity was within safety range and not clinically relevant. From the view point of extended manufacturing, the amount of protein modifications observed in each sample lot was observed to accumulate from beginning to the end, which will suggest the need for continuous analytical monitoring of the manufacturing process to monitor the freshness of the system. For other properties of the drug such as toxicity, biofunction and pharmacokinetics, further studies will need to be performed to ensure comparability with the innovator product.

References

[1] J. Christopher Love, Kerry Routenberg Love, and Paul W. Barone. Enabling global access to high-quality biopharmaceuticals. *Current Opinion in Chemical Engineering*, 2(4):383–390, 2013.

[2] Ronald A Rader. FDA Biopharmaceutical Product Approvals and Trends in 2012 Up from 2011, but Innovation and Impact Are Limited. *18 BioProcess International*, 11(3), 2013.

[3] Anthony J Hatswell, Gianluca Baio, Jesse A Berlin, Alar Irs, and Nick Freemantle. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999-2014. *BMJ open*, 6(6):e011666, 6 2016.

[4] Anurag S Rathore and Helen Winkle. Quality by design for biopharmaceuticals. *Nature biotechnology*, 27(1):26, 2009.

[5] Kerry R. Love, Kartik A. Shah, Charles A. Whittaker, Jie Wu, M. Catherine Bartlett, Duanduan Ma, Rachel L. Leeson, Margaret Priest, Jonathan Borowsky, Sarah K. Young, and J. Christopher Love. Comparative genomics and transcriptomics of Pichia pastoris. *BMC Genomics*, 17:550, 2016.

[6] Geoff P Lin Cereghino, Joan Lin Cereghino, Christine Ilgen, and James M Cregg. Production of recombinant proteins in fermenter cultures of the yeast pichia pastoris. *Current opinion in biotechnology*, 13(4):329–332, 2002.

[7] David Gervais. Protein deamidation in biopharmaceutical manufacture: understanding, control and impact. *Journal of Chemical Technology and Biotechnology*, 91(3):569–575, 2016.

[8] WILLIAM S Hancock, ELEANOR Canova-Davis, ROSANNE C Chloupek, SL Wu, IP Baldonado, JOHN E Battersby, MICHAEL W Spellman, LOUISETTE J Basa, and JOHN A Chakel. *Characterization of Degradation Products of Recombinant Human Growth Hormone*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1988.

[9] Khorshed SM Alam, Takahiko Fujikawa, Hideo Yoshizato, Minoru Tanaka, and Kunio Nakashima. Synthesis and purification of a deleted human growth hormone, hghδ135–146: sensitivity to plasmin cleavage and in vitro and in vivo bioactivities. *Journal of biotechnology*, 78(1):49–59, 2000.

[10] JB Calixto. Efficacy, safety, quality control, marketing and regulatory guidelines for herbal medicines (phytotherapeutic agents). *Brazilian Journal of Medical and Biological Research*, 33(2):179–189, 2000.

[11] Robert L Garnick. Safety aspects in the quality control of recombinant products from mammalian cell culture. *Journal of pharmaceutical and biomedical analysis*, 7(2):255–266, 1989.

[12] Glen Teshima, JT Stults, Victor Ling, and Eleanor Canova-Davis. Isolation and characterization of a succinimide variant of methionyl human growth hormone. *Journal of Biological Chemistry*, 266(21):13544–13547, 1991.

[13] H S Lu, P R Fausset, L O Narhi, T Horan, K Shinagawa, G Shimamoto, and T C Boone. Chemical modification and site-directed mutagenesis of methionine residues in recombinant human granulocyte colony-stimulating factor: effect on stability and biological activity. *Archives of biochemistry and biophysics*, 362(1):1–11, 2 1999.

[14] Masayoshi Oheda, Sumihiro Hase, Masayoshi Ono, and Tokuji Ikenaka. Structures of the sugar chains of recombinant human granulocyte-colony-stimulating factor produced by chinese hamster ovary cells. *The Journal of Biochemistry*, 103(3):544–546, 1988.

[15] Martin Höglund. Glycosylated and non-glycosylated recombinant human granulocyte colony-stimulating factor (rhg-csf)what is the difference? *Medical oncology*, 15(4):229–233, 1998.

[16] James R. Roesser and Charles Yanofsky. The effects of leader peptide sequence and length on attenuation control of the trp operon of e.coli. *Nucleic Acids Research*, 19(4):795–800, 1991.

[17] Rodney Pearlman and Y John Wang. *Formulation, characterization, and stability of protein drugs*, volume 9. Springer Science & Business Media, 1996.

[18] Anne Munk JESPERSEN, Thorkild CHRISTENSEN, Niels Kristian KLAUSEN, Per Franklin NIELSEN, and Hans Holmegaard SØRENSEN. Characterisation of a trisulphide derivative of biosynthetic human growth hormone produced in escherichia coli. *The FEBS Journal*, 219(1-2):365–373, 1994.

[19] Rajender K Chawla, John S Parks, and Daniel Rudman. Structural variants of human growth hormone: biochemical, genetic, and clinical aspects. *Annual review of medicine*, 34(1):519–545, 1983.

[20] Shiaw Lin Wu, Haitao Jiang, William S. Hancock, and Barry L. Karger. 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. *Analytical Chemistry*, 82(12):5296–5303, 2010.

[21] ELEANOR CANOVA-DAVIS, IDA P BALDONADO, JEROME A MOORE, CHRISTOPHER G RUDMAN, WILLIAM F BENNETT, and WILLIAM S HANCOCK. Properties of a cleaved two-chain form of recombinant human growth hormone. *Chemical Biology & Drug Design*, 35(1):17–24, 1990.

2.8 Supplementary Data

Tab 2.1: Peptide mapping of GCSF products

	Sequence	m/z	Std	#1	#2	#3	#4
G1 (1-17)	MTPLGPASSLPQSFLLK	893.99^{2+}	26.11	26.18	26.14	26.14	26.15
G2 (18-24)	CLEQVRK	875.44^{+}	18.57	18.55	-	18.55	18.57
G3 (25-34)	IQGDGAALQE	1002.48^{+}	15.43	15.38	15.38	15.43	15.42
G3 (25-35)	IQGDGAALQEK	565.28^{2+}	12.02	12.11	12.15	12.05	12.09
G4 (36-47)	LCATYKLCHPEE	469.54^{3+}	14.55	14.54	14.59	14.58	14.57
G5 (48-94)	LVLLGHSLGIPWAPLSS-CPSQALQLAGCLSQLHS-GLFLYQGLLQALE	1236.4^{4+}	32.1	32.08	32.03	32.01	32.11
G6 (95-99)	GISPE	502.255^{+}	12.1	12.11	12.15	12.05	12.09
G7 (100-105)	LGPTLD	615.34^{+}	18.89	18.92	18.91	18.94	18.94
G8 (106-110)	TLQLD	589.31^{+}	18.05	18.06	18.1	18.05	18.07
G9 (111-124)	VADFATTIWQQMEE	834.89^{2+}	29.03	29.02	28.98	29.04	28.97
G10 (125-142)	LGMAPALQPTQGAMPAFA	886.46^{2+}	27.03	27.07	-	27.06	27.07
G11 (143-163)	SAFQRRAGGVLVASHLQSFLE	758.75^{3+}	21.25	21.5	21.53	21.37	21.44
G12 (164-175)	VSYRVLRHLAQP	360.465^{4+}	15.32	15.57	15.6	15.43	15.42

Tab 2.2: Modification summary of strain GCSF products

	Std	#1	#2	#3	#4
Sequence coverage	100%	100%	86.3%	100%	100%
N-terminal truncation	N/D	35.9%	N/D	2.9%	63.8%
Met 1 oxidation	2.5%	2.6%	N/D	2.0%	2.0%
Met 122 oxidation	$\sim$ 0.5%	N/D	N/D	$\sim$ 0.5%	N/D
Met 127 oxidation	N/D	N/D	N/D	N/D	N/D
Met 138 oxidation	$\sim$ 0.5%	N/D	N/D	$\sim$ 0.5%	N/D
Glu12 deamidation	N/D	N/D	N/D	N/D	N/D
T134 Mannosylation	N/D	67.3%	N/D	47.7%	62.9%

Tab 2.3: peptide mapping of strain Interferon products

peptide	Sequence	m/z	std	#1	#2	#3	#4
T0 (Leader)	EEGVSLEK (R)	445.72 (2+)	n/d	12.72	12.71	12.75	12.72
T1-T7 (1-12/84-112)	CDLPQTHSLGSR/ FYTELYQQLNDLEAC- VIQGVGVTETPLMK	923.84 (5+)	28.94	28.84	28.87	28.89	28.82
1-13/84-112	CDLPQTHSLGSRR/ FYTELYQQLNDLEAC- VIQGVGVTETPLMK	796.05 (6+)	28.65	28.57	28.59	28.6	28.54
T2(14-22)	TLMLLAQMR	538.79^{2+}	23.42	23.31	23.32	23.35	23.39
T3-T12 (24-31/135-144)	ISLFSCLK/YSPCAWEVVR	530.26^{4+}	23.15	23.17	23.05	23.21	23.14
T4 (32-49)	DRHDFGFPQEEFGNQFQK	557.50^{4+}	20.26	20.36	20.21	20.3	20.34
34-49	HDFGFPQEEFGNQFQK	652.62^{3+}	20.98	21.05	21.06	21.01	21.03
T5 (50-70)	AETIPVLHEMIQQIFNLFSTK	820.76^{3+}	32.41	32.36	32.43	32.4	32.38
T6 (71-83)	DSSAAWDETLLDK	725.83^{2+}	22.74	22.78	22.78	22.78	22.74
T8 (113-120)	EDSILAVR	451.75^{2+}	17.81	17.78	17.77	17.85	17.89
T9 (121-125)	KYFQR	371.20^{2+}	8.33	8.23	8.29	8.24	8.26
122-125	YFQR	307.15^{2+}	10.05	9.99	9.92	9.96	9.98
T10 (126-131)	ITLYLK	375.74^{2+}	19.07	19.04	19.04	19.00	19.03
T11 (132-134)	EKK	N/D	n/d	n/d	n/d	n/d	n/d
T13 (145-149)	AEIMR	310.16^{2+}	10.33	10.37	10.32	10.24	10.25
T14 (150-162)	SFSLSTNLQESLR	741.37^{2+}	22.33	22.36	22.37	22.36	22.34
T15 (163-165)	SKE	n/d	n/d	n/d	n/d	n/d	n/d

Tab 2.4: Summary of modifications observed in Interferon products

	Sequence	Std	#1	#2	#3	#4
Leader peptide (T0)	EEGVSLEK	n/d	2.1%	2.6%	3.7%	4.6%
Met16	TLMLLAQMR	0.3%	0.7%	1.5%	1.0%	1.0%
Met21	TLMLLAQMR	1.1%	0.8%	0.5%	0.2%	0.9%
Met59	AETIPVLHEMIQQIFNLFSTK	1.9%	1.1%	0.7%	0.9%	2.6%
Met111	FYTELYQQLNDLEACVIQGVGVTETPLMK	N/D	N/D	N/D	N/D	N/D
Met148	AEIMR	N/D	N/D	N/D	N/D	N/D

Tab 2.5: Peptide mapping of strain hGH products

peptide	Sequence	m/z	Std	#1	#2	#3	#4
T1(1-8)	FPTIPLSR	465.77^{2+}	21.03	21.02	21.02	21.08	21.01
T2(9-16)	LFDNAMLR	490.25^{2+}	20.49	20.51	20.51	20.48	20.5
T2-T3	LFDNAMLRAHR	448.20^{3+}	12.83	12.95	12.93	12.89	12.93
T4(20-38)	LHQLAFDTYQEFE-EAYIPK	781.38^{3+}	24.47	24.49	24.47	24.56	24.89
T4-T5	LHQLAFDTYQEFEE AYIPKEQK	682.83^{4+}	22.8	22.79	22.89	22.89	22.86
T6(42-64) T16(159-167)	YSFLQNPQTSLCFSE-SIPTPSNR NYGLLYCFR	941.703^{4+}	26.05	26.05	26.02	26.01	26.07
T7(65-70)	EETQQK	762.36^{+}	1.77	1.98	2.09	1.95	n/d
T8(71-77)	SNLELLR	422.74^{2+}	18.15	18.12	18.09	18.06	18.07
T7-T8	EETQQKSNLELLR	529.94^{3+}	15.65	15.72	15.7	15.7	15.8
T9(78-94)	ISLLLIQSWLEPVQFLR	1028.10^{2+}	31.74	31.8	31.8	31.84	31.88
T10(95-115)	SVFANSLVYGASDSN-VYDLLK	754.71^{3+}	27.24	27.14	27.21	27.18	27.1
T11(116-127)	DLEEGIQTLMGR	681.34^{2+}	25.92	25.94	25.92	25.89	25.85
T12(128-134)	LEDGSPR	387.19^{2+}	8.32	8.35	8.4	8.34	8.45
T13(135-140)	TGQIFK	347.20^{2+}	12.56	12.54	12.53	12.51	12.55
T14(141-145)	QTYSK	626.31^{+}	2.79	2.85	2.82	2.83	2.85
T15(146-158)	FDTNSHNDDALLK	497.23^{3+}	15.37	15.43	15.41	15.43	15.39
T17-19(168-178)	KDMDKVETFLR	691.36^{2+}	16.39	16.31	16.29	16.29	16.38
T19(173-178)	VETFLR	382.71^{2+}	16.52	16.45	16.54	16.53	16.51
T20(179-183) T21(184-191)	IVQCR SVEGSCGF	701.82^{2+}	14.5	14.56	14.55	14.46	14.53

Tab 2.6: Summary of observed modifications of strain hGH products

	Sequence	STD	#1	#2	#3	#4
Sequence Coverage	1-191	100%	100%	100%	100%	100%
M14 (oxidation)	LFDNAMLR	0.77%	0.67%	0.77%	1.55%	1.41%
M125 (oxidation)	DLEEGIQTLMGR	1.89%	0.72%	0.48%	0.90%	1.67%
M170 (oxidation)	KDMDKVETFLR	n/d	n/d	n/d	n/d	n/d
N149 (deamidation)	FDTNSHNDDALLK	0.2%	0.6%	0.2%	0.9%	1.3%
Q141 (Two-chain)		n/d	0.13%	1.29%	3.90%	3.30%

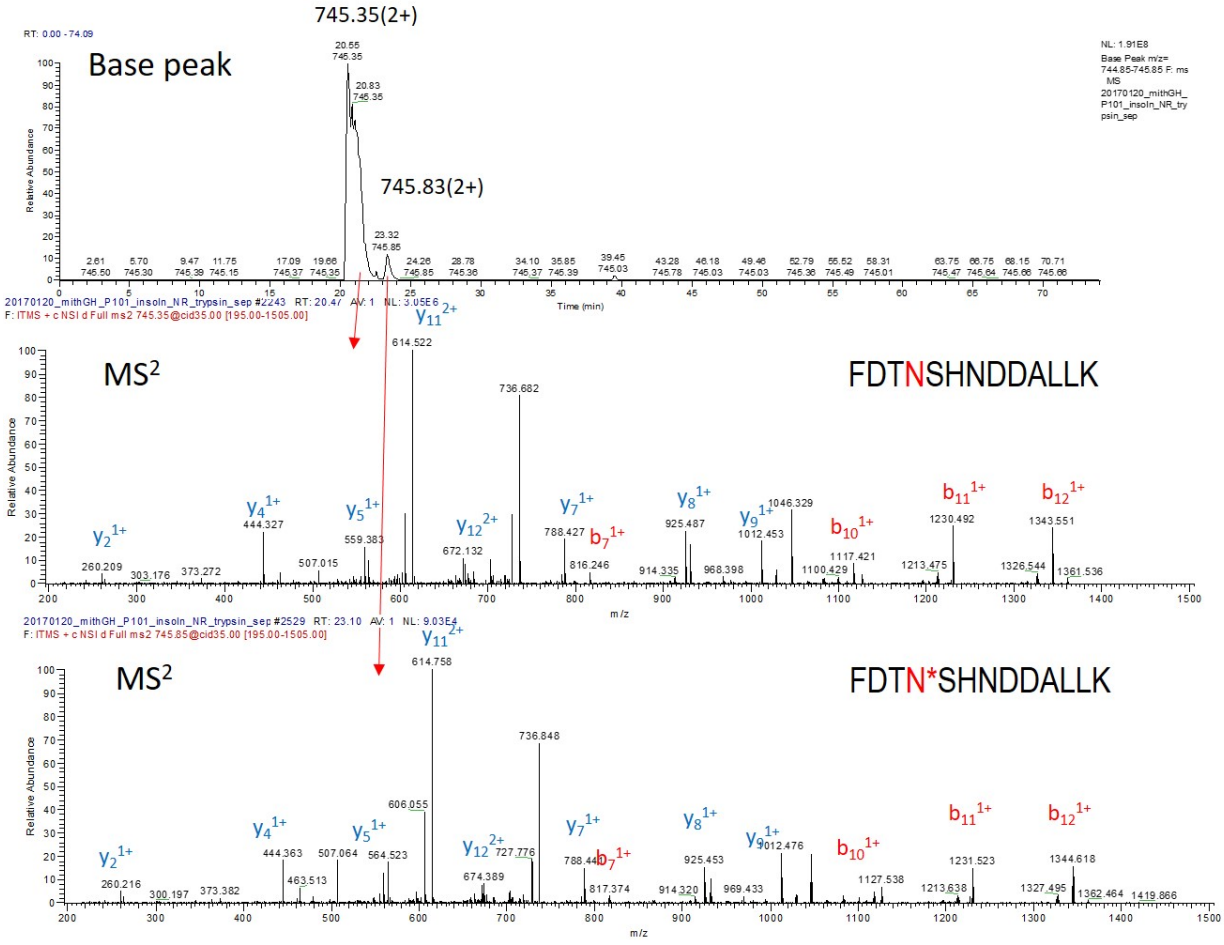

Fig 2.18: LC-MS analysis of the deamidated peptide (T15) from the tryptic digest of rhGH

Chapter 3

Combined top-down and bottom-up mass spectrometric approach to recombinant glargine product characterization approaches

Contribution:

Di Wu: concept contribution, experiment procedures on recombinant glargine, data analysis and interpretation, manuscript writing and revision;

Di Liu: MIT team collaborator, strain sample preparation

William Hancock: goal of the study, manuscript revision

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and SPAWAR Systems Center Pacific (SSC Pacific) under Contract No. N66001-13-C-4025.

3.1 Abstract

The industry of biopharmaceuticals has been studied in great detail in humanity in recent decades. To efficiently produce the high quality drugs in remote areas or under emergency situations has become a global challenge. Insulin is very commonly required and when available used for the treatment of diabetes. To produce high quality biosimilar drug with a improved treatment profile and lower cost, the development of glargine would be significant for both the medical and biological field. The research funded by DARPA and developed by MIT has developed a small scale production technique, namely the InSCyT platform. We describe the characterization of the product with both bottom-up and top-down methods and the results from these various methods are comparable and consistent. The results of the analysis demonstrated that the products were prepared in a good quality even in the initial upstream product. The characterization of insulin glargine includes applying a variety of techniques, including MALDI-TOF, Q-TOF, and LCMS and we found that the products were secreted with full sequence coverage and with few modifications. These analytical technologies will assist fermentation development and improve the subsequent production by a more robust system in the future.

3.2 Introduction

Diabetes is one of the worst diseases globally, and due to co-morbidities patients with diabetes may induce cancer as well.[1, 2] There are many studies focusing on the chemical, biological or clinical views of diabetes and insulin[3] therapy. With strong clinical evidence to indicate that insulin will significantly decrease the glucose level of patients with both type I and II diabetes.[4, 5] In the insulin pharmaceutical market, the original insulin patents will expire in the coming future with the promise of biosimilar versions. This is a potentially welcome development as the global market has increased rapidly and reached a high cost level of 50 millions.[3]

As well known, insulin is essential for the treatment of diabetes type I and II. As the first time to produce biosimilar of insulin (Humulin) was in the 1980s,[6, 7] this industry has appealed accumulated number of studies.[8] The therapeutic from the successful insulin production, the market is continuously expanding from the past decades.[9] Both industrial companies or academic institutes are focusing on more opportunities to introduce and explore the development of high quality biosimilar insulins to the market. However, the approval of analogs are not easy in any countries, such as in the U.S and EU, the application and approval process are complicated and the failed follow-on applications will be detailed. The safety, quality, stability, manufacturing efficiency, efficacy and many other aspects must be evaluated to obtain an approval. Sequence modification between of the biosimilar (Lantus glargine, Novolin, Levemir, and etc.) with the innovator insulin may affect the pharmacokinetics and pharmacodynamics. To date, the modifications are designed to improve the functions for patients need to change the glucose level, such as rapid acting insulin analogues (Novolog, etc) and long duration drug (Lantus glargine, etc).[10]

The traditional cultivation of recombinant proteins with *E.coli* expression induce contamination to the man-

ufacturing process, other bioactivities, or even inaccuracy on expression.[11] To minimize these issues, our team applied yeast *Pichia* cultivation to enhance to gene expression and improve product quality.[12, 13] This platform called Integrated and Scalable Cyto-Technologies for Flexible microbial manufacturing (InSCyT) was developed by our team for the Biologically-derived Medicines On Demand (BioMOD) program,[14] aiming to efficiently produce high quality biopharmaceutics in remote locations. The platform and products have been introduced in Chapter II. The quality of the final purified biopharmaceuticals such as recombinant human growth hormone and GCSF were performed at a high analytical level. In fact, during the development of the upstream and downstream processes, there is long development cycle to achieve the required level of purified product.[15] The fermentation, manufacture and progress of each production steps are important and in this context the crude protein drug is the beginning of the final success. Its hard to control the manufacture quality and reproducibility and difficult to determine the quality of the resulting crude proteins. In this chapter, we have not only introduced both top-down and bottom-up technologies on to characterize the reference materials we will also report on the analytical data of the crude protein product and illustrate the guidance and suggestions we gave to the manufacturing teams. In the future, the advanced purified drugs from the platform will be further studies by our techniques to provide a detailed analysis relative to the corresponding reference drug.

The bottom-up method is more commonly used to characterize the protein structure with digested peptide information from mass spectrometry analysis. With more detailed information from additional enzymatic digestions, the analysis would be more accurate but the preparation and post-analysis consumes significantly more time. . The proteins, or bioproducts will be digested into small pieces by a selected protease and then for analysis by a LC separation and MS analysis.[16] The peptides will be identified by the mass detector with advanced MS/MS determination. The amino acids of the protein sequence will be specifically identified in this process. Proteins are much more complex than small molecule or short peptides to be measured due

to the large structures and lower solubility, moreover, the modifications are also challenging to be easily detected by top-down techniques. Top-down and bottom-up methodologies are both applied in analytical researches. The top-down techniques are widely used[17, 18, 19] for intact protein determination with high mass accuracy measurement. The proteins or bioproducts will be analyzed directly as the intact molecule without digestion which saves preparation time as well as artifacts generated by the digestion process. At the level of a small protein level, intact protein analysis can reach full sequence coverage and successful determination usually can be achieved with a molecular weight less than 70kDa.[20] However, the peptide fragmentation of the intact protein in the mass spectrometer at a large molecule level is less predictable for analysis of an enzyme digested protein. In summary top-down applications require a short time period from sample preparation to final analysis, however, the detailed information of post translational modifications may not be specifically obtained. For traditional methods, the FT-MS is more applicable for intact protein analysis based on its resolving power while so far, Orbitrap, TOF and more sophisticated instruments and software are under explored.[21]

In this chapter, we will introduce the characterization of recombinant glargine with both top-down and bottom-up approaches. We also applied the method to crude strain glargine protein samples.

3.3 Experimental

3.3.1 Chemicals and materials

The glargine standard was obtained from Lantus (refer as glargine), the compared insulin standard was obtained from Humulin (refer as insulin), and the strain glargine samples were produced by MIT laboratory. The chemicals including acetonitrile, Tris-HCl sodium phosphate, guanidinum chloride, dithiothreitol,

iodoacetamide, and trypsin and GluC were analytical grade, obtained from Thermo (Thermo Fisher, USA).

3.3.2 In-solution digestion

50μg glargine sample were dissolved in 6M guanidinium chloride, reduced by 10 mM dithiothreitol under 70°C for 30 minute and followed by alkylation with 55mM iodoacetamide under room temperature in dark condition. Proteins were dialyzed via 10kD membrane Amicon centrifugal filter at 13,000 rpm for 15 minute and three times in Tris-HCl buffer (pH=4.0).The following in-solution digestion process with Trypsin and GluC was kept overnight at room temperature to avoid artificial oxidation. Proteins were then ready for LC-MS analysis, the remained materials were aliquot to 20 μL and stored in -80°C for further analysis.

3.3.3 In-gel digestion

The strain samples were prepared to 15 μg as the concentration provided by the collaborators in order to equalize the loading amount of proteins for comparison. At the same time, the other batch of samples were prepared with additional 2 μL 1M DTT solution incubated at 90°C for 30 min for reduced conditions. The protein mixtures were separated on SDS-PAGE at 160V for 45 min and stained with Coomassie blue. After a destaining process, the bands were monitored for further evaluation of the protein quality.

The gel slices were washed with 500 μL ACN and 0.1 M NH_4CO_3 for 45 min shaking and then centrifuged, the supernatant were removed. Proteins were followed by a reducing process by adding 500 μL of 10 mM dithiothreitol in 0.1 M NH_4HCO_3 and incubated for 30 min at 56°Cwith centrifuging and removing supernatant before the alkylation process. Proteins were alkylated with 500 μL of 55 mM iodoacetamide in 0.1 M NH_4HCO_3 under room temperature and kept in dark for 60 min. All supernatant were removed after

spinning the gel pieces down and the following digestion process. Trypsin was added to protein solution and the concentration was based enzyme : protein ratio of 1:100 to 1:20 (w/w). The samples were covered by trypsin and GluC buffer and incubated overnight at 37°C to yield peptides. The digestion process was stopped with 50 μL 5% formic acid and extracted with acetonitrile. All supernatants were combined and stored for the following LC-MS analysis.

3.3.4 Isoelectric Focusing measurement

Insulin Humulin and Insulin glargine standard samples were prepared under stressed condition at pH 9, 37°C incubation for 3 days.This stability tests the variation of deamidation level with the protein sequence difference as the variants will be observed with different isoelectric focusing points relative to the test samples. After the stressed incubation, we prepared the humulin and glargine samples to 1 mg/ml total protein concentration and mixed 10 μL of the samples 1:1 with Novex IEF sample buffer pH 3-10 (2x (Thermo Fisher, USA)). After loading the humulin and glargine standard and the samples which has been with stressed under alkaline conditions, we set the program in the power supply at constant voltage of 100V for 1hr, 200V for 1h and 500V for 30min. When the program is complete, the gel was fixed with 12% trichloroacetic acid. The staining and destaining will terminate the whole process to evaluate the bands.

3.3.5 Measurement of HPLC mass spectrometry

LC-MS analysis was used an Ultimate 3000 nano LC pump (Dionex, Mountain View, CA) and a self-packed C18 column (Magic C18, 200Å pore and 5 μm particle size, 75μm internal diameter by 100 mm) connected to a coated 10μm internal diameter emitter (New Objective, Woburn, MA). A LTQ-Orbitrap XL mass spectrometer was connected (Thermo Fisher Scientific, San Jose, CA) through a nanospray ion source

(New Objective, Woburn, MA). Mobile phase A was using 0.1% formic acid in HPLC grade water and mobile phase B was using 0.1% formic acid in acetonitrile. During sample injection, the flow rate was set 250 nL/min with 2% B for 25 min. The flow rate of the gradient was set at 200 nL/min, with mobile phase B, 0-60 min 2-40%, 60-70 min to 90% , 70-75 min 90% and 75-78 min 2%. The mass spectrometer was operated in a data dependent mode to switch between MS and CID-MS^2. Briefly, after a full-scan MS spectrum from m/z 400-2000 in the ion-trap, 8 CID-MS^2 activation steps were performed on the 8 most intense precursor ions from the full scan. All control and variants samples were run in triplicate.

BioPharma Finder 2.0 software (Thermo Fisher Scientific) was used for analysis of all data acquired on the peptide and protein level. For peptide mapping, searches were performed using a single-entry protein FASTA database with oxidation and deamidation set as variable modifications, 20 ppm mass accuracy, and a confidence level of 0.8 for MS/MS spectra. Final confirmation of the peptide identification was determined by manual inspection, extracting the base peak from the chromatogram and matching the MS^2 fragmentation data with theoretical prediction. The modification percentage was calculated by peptide peak area.

3.3.6 Q-TOF measurement of intact protein analysis

Protein profiles were analyzed using reverse-phase liquid chromatography (RPLC) and a Xevo G2-S Q-ToF (Waters Corp, Milford, MA). Liquid chromatography was performed at 0.2 mL/min using an H-Class Acquity ultra-high pressure liquid chromatography system (UPLC) (Waters Corp, Milford, MA) on a BEH300-C4, 2.1 mm x 100 mm column, with pore size of 1.7μm (Waters Corp, Milford, MA). Buffer A consisted of 0.1% formic acid (v/v) in HPLC grade water and buffer B consisted of 0.1% formic acid (v/v) in 100% HPLC grade acetonitrile (v/v). A 26 minute gradient was used: 0-5 min 5% B, 6-15 min 5-90% B, 16-20 min isocratic at 90% B and reducing to 5% B from 21 min to the end, samples were introduced via an

electrospray ion source in-line with the Xevo G2-S Q-ToF. External calibration of m/z scale was performed using sodium cesium iodide. The Q-ToF parameters were run in resolution mode, scanning m/z 400-4000, 3.00kV capillary voltage, 40V cone voltage, 120°C source temperature, 350°C de-solvation temperature, and 800L/h de-solvation gas. MS/MS data were collected at a scan speed of 0.1s. Data were manually interpreted using the UNIFI software package (Waters Corp, version 1.7.1).

3.3.7 MALDI measurement of intact protein analysis

Sinapic acid (Sigma Aldrich, 85429) at 10 mg/mL in ACN:wate (1:1) was mixed with sample solution in 40:1 ratio, and 0.5 mL of the mixture was deposited on MALDI plate and air-dried. The MALDI-TOF/TOF-MS was a model 5800 from AB-SCIEX (Framingham, MA) with a frequency-tripled Nd:YAG laser at 349 nm, and operated in the positive ion mode. The delay time for reflectron MS mode was 750 ns. TOF/TOF was in a positive 1 kV mode with air (1×10^{-6} Torr) as the collision gas, and a mass window of $\pm$5 Da.

3.4 Results and discussion

Insulin, as we know, as a treatment for diabetes, is very important for adults to lower levels of glucose in the blood. The sequence of insulin and glargine has differences from the native human insulin by 3 amino acid residues, which lead changes in the stability of the protein and degradations pathways. In glargine vs. insulin the amino acids glycine in substituted in the place of asparagine at chain A site21, and the two arginines attached to the chain B at sites 31 and 32 as the C- terminal difference, see Fig.3.1, for the sequence of glargine and amino acid substitutions. The identification and characterization include intact protein identification with bottom-up identification with HPLC-MS/MS and the top-town strategy such as

HPLC-QTOF, and MALDI-TOFTOF.

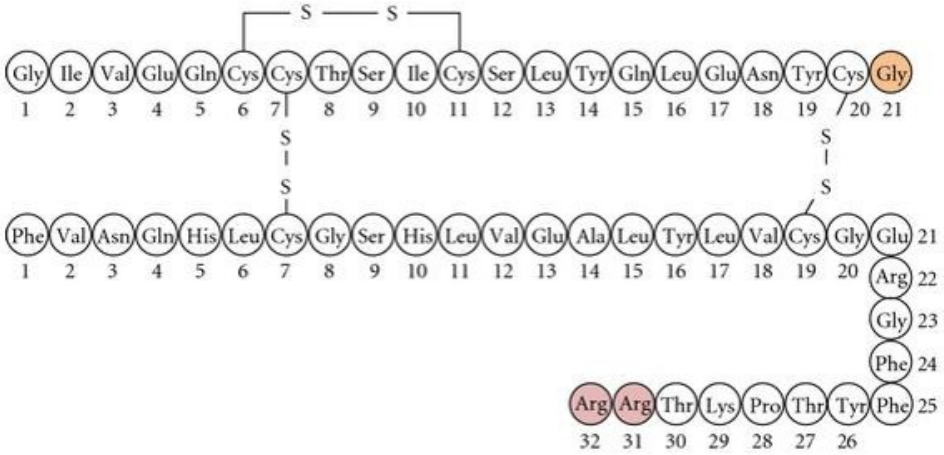

Fig 3.1: Glargine sequence (The highlighted amino acids Gly21 in Chain A and C-terminal Arg31 and Arg32 are the different residues vs. Insulin)

As shown in Fig.3.1, the sequence of glargine and differences of the amino acids. The sequence of insulin and glargine has differences with the native human insulin by 3 amino acids, which lead to the stability and degradations reactions changing.

3.4.1 Bottom-up analysis

3.4.1.1 Primary structure identification

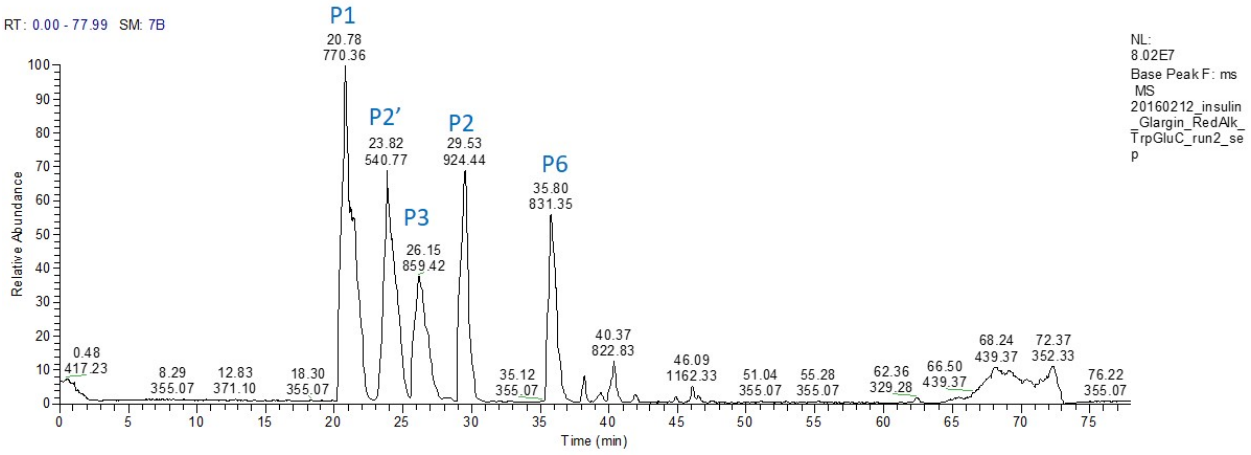

Fig 3.2: Base peak ion chromatogram of the digested glargine

Tab 3.1: Peptide mapping of glargine standard

Peptide	amino acid	sequence	m/z	t_R (min)
P1	1-13	(-)FVNQHLCGSHLVE(A)	770.37 (2^+)	22.09
P2	14-21	(E)ALYLVCGE(R)	924.45 (1^+)	29.75
P2'	14-22	(E)ALYLVCGER(G)	540.78 (2^+)	25.14
P3	23-29	(R)GFFYTPK(T)	859.42 (2^+)	26.15
P4	30-32	(K)TRR(G)		N/A
P5	33-49	(R)GIVEQCCTSICSLYQLE(Q)	1030.46 (2^+)	39.11
P6	37-49	(E)QCCTSICSLYQLE(N)	831.35 (2^+)	24.81
P7	50-53	(E)NYCG(-)		N/A

With 53 amino acids, glargine was cleaved by trypsin combined GluC, the multi-enzyme strategy and the theoretical digested peptide mapping data are shown as below in Table.3.1 and the base peaks were shown in Fig.3.2. The multi enzyme strategy worked well in this experiment and the peptides were observed in high abundance. The protein was processed with reduction and alkylation so that the disulfide bonds were cleaved. The N-terminal peptide FVNQHLCGSHLVE was shown with the highest signal with m/z 770.37 (2^+) was eluted at 22.09 min. With the connecting and cleaved residue K, R and E in the sequence, several peptides contained miscleaved residues such as P2'. The peptide P5 with a high m/z was not shown in a significantly high signal, but the extracted m/z has been identified. The residue difference of glargine and insulin in peptide P4 "TRR" and P7"NYCG" were not determined in bottom-up experiment due to generation of small digested fragments. The intact mass difference was identified by top-down mass spectrometry and will be demonstrated in following sections.

3.4.2 Disulfide bond identification

For the linkage identification, with the non-reduced conditions, the protein was digested and remained as the intact disulfide bonds, so the linked peptides would be detected. The characterized fragmentation mass assignments were shown in Fig.3.3 and Fig. 3.4. In this project, CID was applied to identify the linkage with the whole connected peptide information. Since the chain A and chain B of glargine is connected the disulfide bonds,, the proteolysis may not generate unique form peptides with two enzymes. The fragmentation has enabled the determination of the intra and inter disulfide linkage with the predicted structures and mass with six forms of linked peptides identified in the control samples. The demonstration of forms 2 and 3 are shown in detail with extracted ion peaks in Fig.3.5 and Fig.3.6.

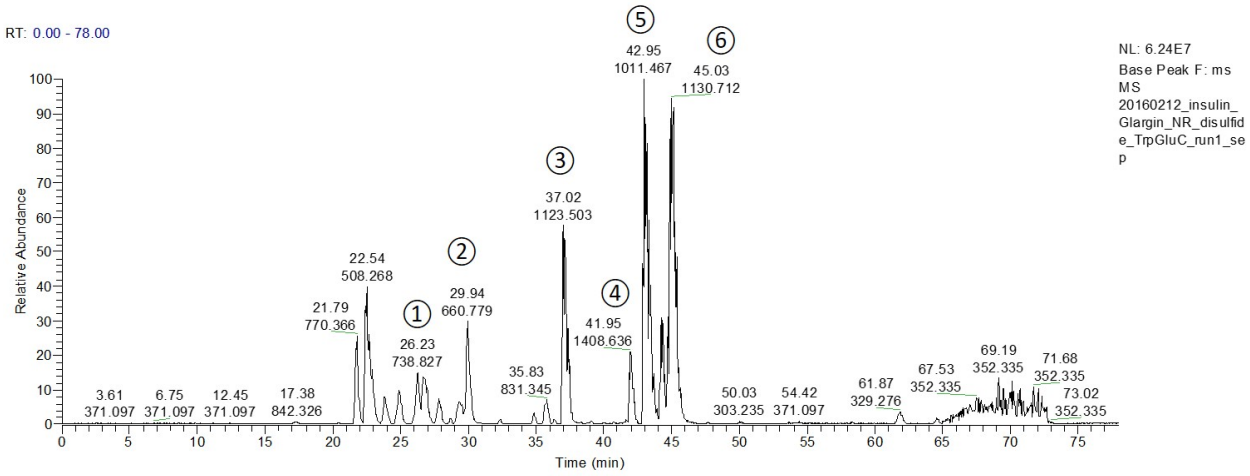

Fig 3.3: Base peak of glargine standard disulfide bond assignment

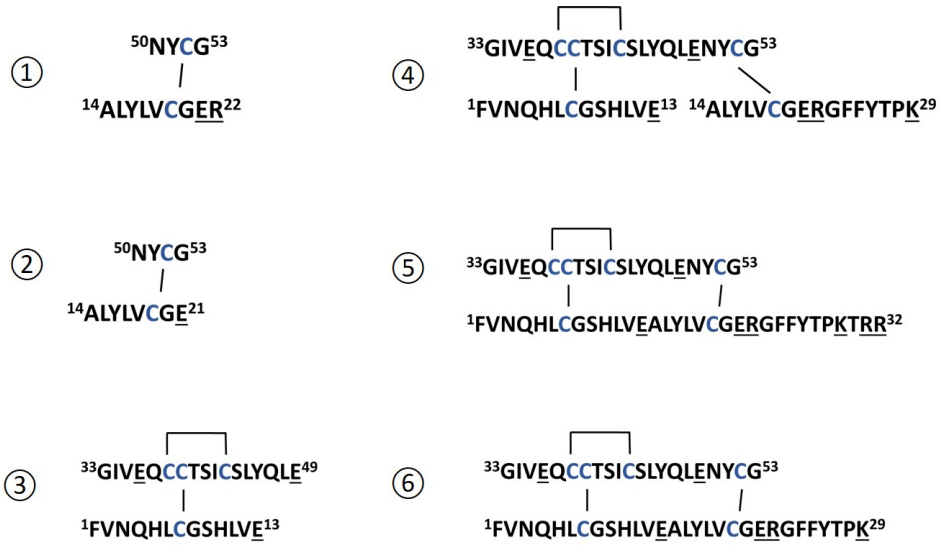

Fig 3.4: Structures of glargine disulfide bond assignment

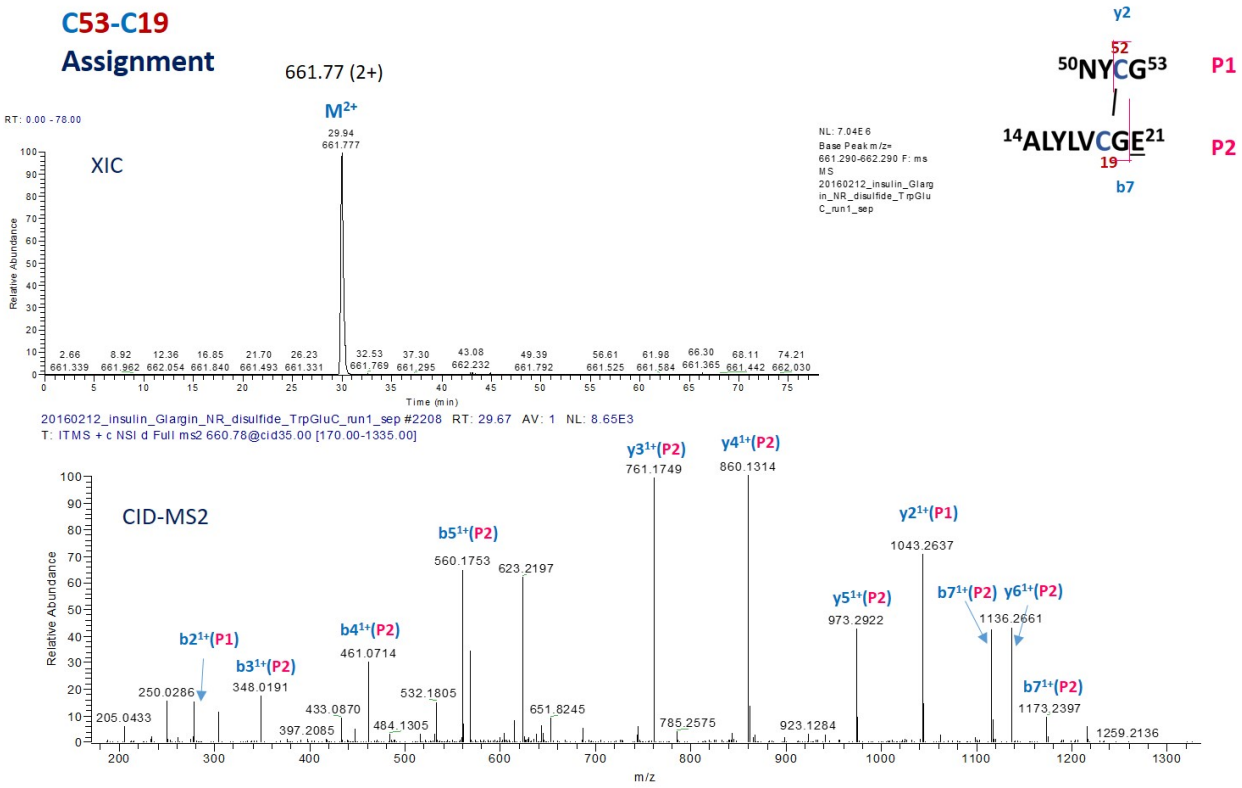

Fig 3.5: Disulfide bond assignment of glargine standard Cys53-Cys19

Fig.3.5 is an example of disulfide bond connected peptides, the linkage is between Cys53 and Cys19, the m/z is 661.77(2^+), which was form 2 in Fig.3.4. The b ion and y ion from each peptide (P1 and P2) indicated in the figure were identified separately. Fig.3.6 illustrated another inter and intra disulfide bonds of Cys38-Cys43 and Cys7, the m/z is 1123.50(3^+), indicated form 3. In Fig.3.5, the accurate mass matched the mass of the two enzymatically generated peptides linked together with 1 disulfide bond, and the corresponding MS2 spectra matched the fragment ions of the expected peptide sequence.

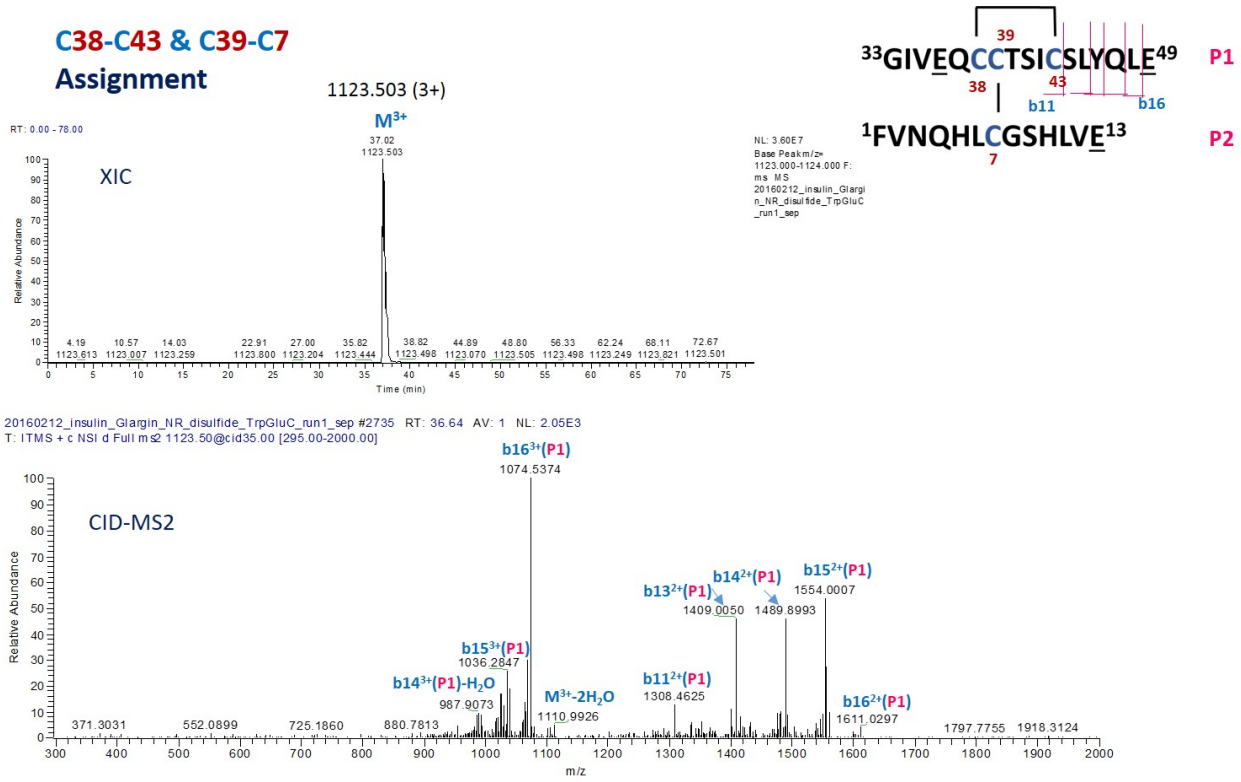

Fig 3.6: Di-sulfide bond assignment of glargine standard Cys38-Cys43, C39-C7

In Fig.3.6, the accurate mass matched the expected mass of the two enzymatic peptides linked together with 2 disulfide bonds were the peptide mass with the loss of 4H, and the corresponding MS2 spectra matched the fragment ions of the expected peptide sequence. The fragment ion patterns indirectly indicated the C38-C43 linkage even though no fragment ions was observed inside the ring as expected. Nevertheless, the exact Cys linkages can be confirmed by further ETD identification in future analyses.

3.4.2.1 Degradation evaluation

The changing of amino acids on human insulin to glargine can make the protein more soluble in an acidic environment but insoluble in neutral conditions. The carboxyl terminal of chain B changes leads to isoelectric point close to neutral and more inactive to degradation in a high pH environment. The glycine substitution of asparagine on chain A prevents deamidation of the acid-sensitive asparagine at acidic pH.[22] As shown in Fig. 3.7, the deamidation of glargine has been studied under stressed conditions where the glutamine (Q37) in glargine is a potential deamidation site. While the Q deamidation is not common, we would like to observe if indeed any deamidation occurred to generate a glutamate residue.

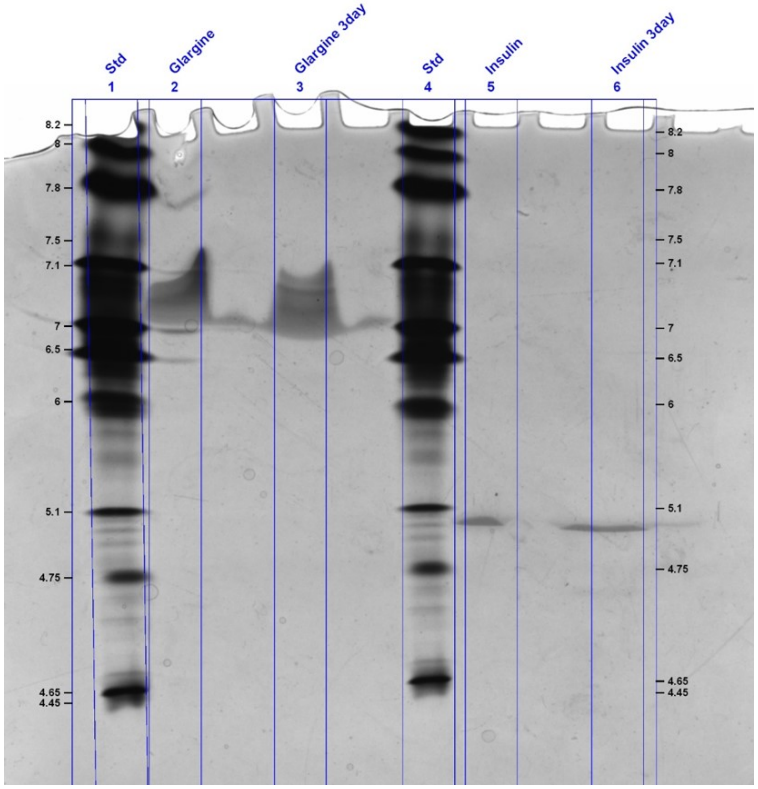

Fig 3.7: IEF gel image of glargine samples

The glargine standard has been dialyzed into sodium borate at pH 9 condition, with the 37°C incubation for

three days. From the gel image, this gel image shows electrophoresis migration and broad bands, however, the isoelectric point of glargine standard and the stressed sample were obviously indicate around 7.0. As lane 5 and lane 6 shows, the stressed insulin had variants bands besides the major bands indicating pH around 5.1. Thus we concluded that glargine has more structural stability than insulin, due to the absence of asparagine residues and glutamine deamidation.

Besides IEF gel observation, a digested procedure has been performed by LC/MS and the resulting extracted ion peak shown in Fig.3.8. The potential glutamate residue is located in P1 FVNQHLCGSHLVE, while the +1 mass shift was not found. The detected m/z was 770.36(2^+), and the +0.5 amu difference at charge 2 was not detected. This peptide under stressed condition has been found to be the same as with the normal conditions. Also the isotopic peak of this peptide has no signal shift at the second abundance, 770.86(2^+) peak. From these different analyses we cannot conclude there is any deamidation that occurred in the glargine sample with our stressed conditions. Thus we conclude that the stability of glargine is better than insulin.

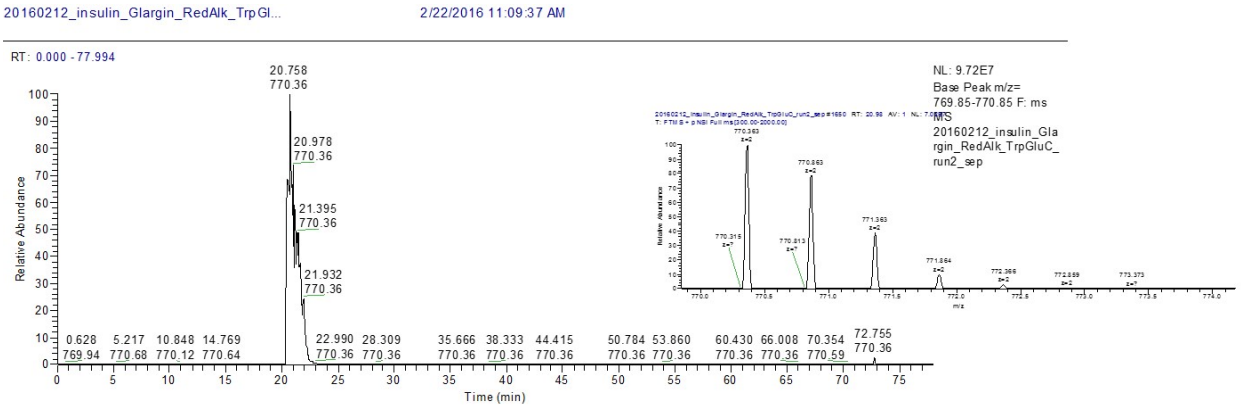

Fig 3.8: XIC peak of Q4 deamidation evaluation of glargine

3.4.3 Top-down analysis

Top-down analysis is a very efficient and convenient way to measure and evaluate the product at the first stage. This may guide the cultivation process in a correct direction although for the strain samples, further purification and polishing procedures are still necessary to achieve the analytic quality required for a therapeutic product. In our research, we have performed the intact protein analysis by several instruments including Thermo Orbitrap, Waters Q-TOF and Sciex MALDI.

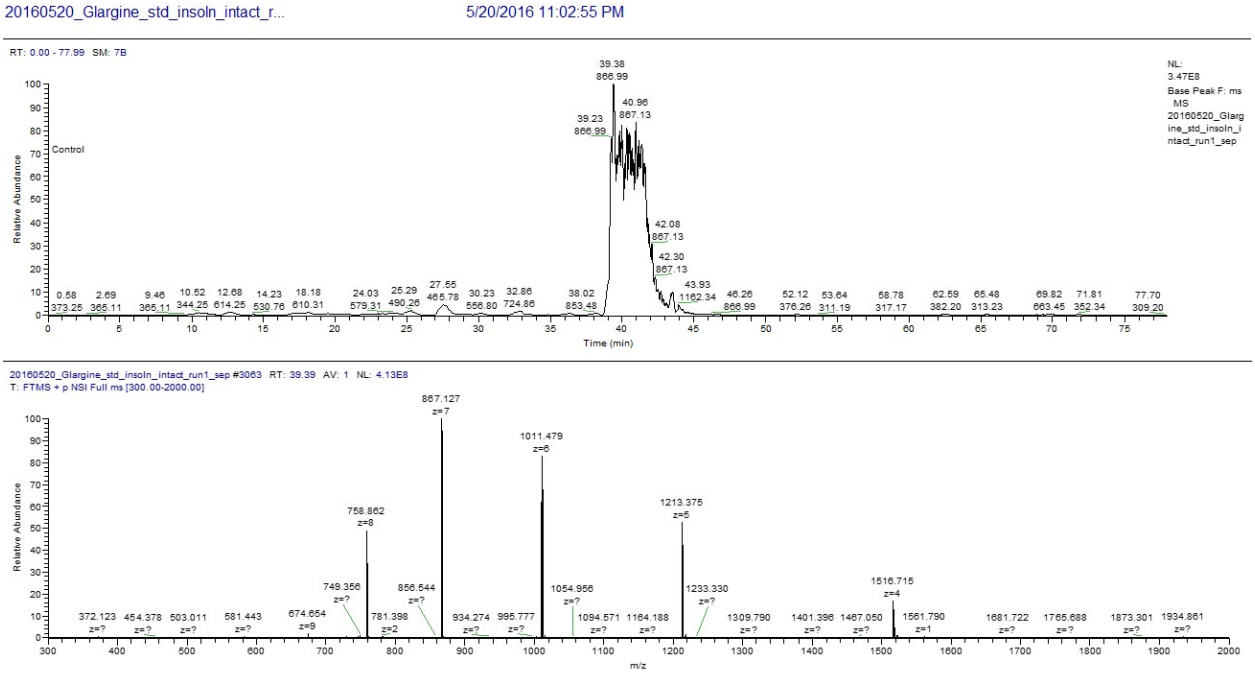

Fig 3.9: Base peak ion chromatogram of the intact glargine analysis

As shown in Fig.3.9, the protein was analyzed by Orbitrap detector, eluted at 39.38min. The peak shape was not smoothly shown due to the amount of protein and the column capacity. At this elution time, the observed m/z fragments included 758.862(8^+), 867.127(7^+)), 1011.479(6^+), and 1516.715(4^+). This instrument has

been set the limitation mass range of m/z 2000, then the ion of 3^+, 2^+ or 1^+ were not detected. The deconvoluted mass is 6062.889Da, which matched with the theoretical molecular average mass for Lantus Glargine 6063Da.Thus we concluded that the Orbitrap is a powerful mass detector to measure small protein.

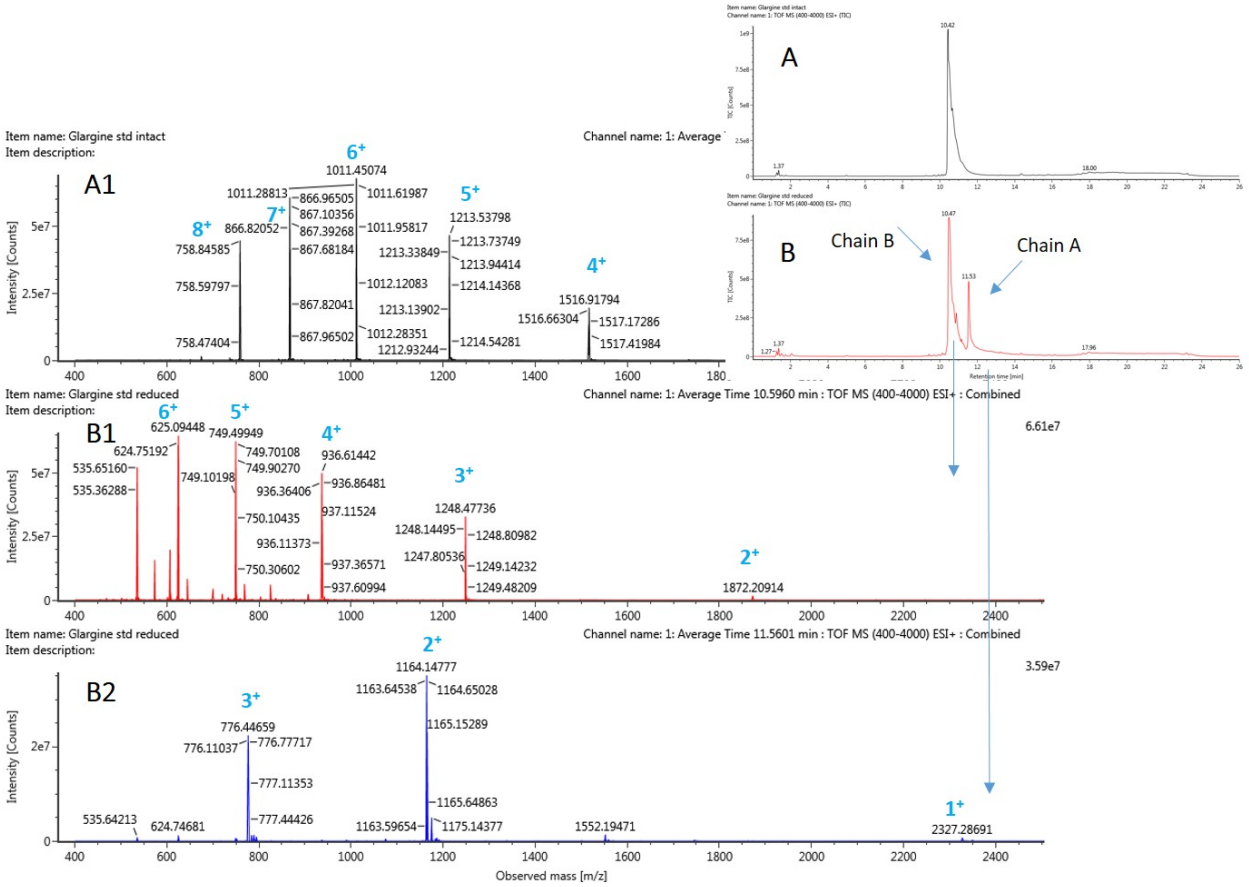

Fig 3.10: QTOF result of glargine with non-reduced and reduced conditions

Using the QTOF technology to identify the intact protein in non-reduced and reduced condition. The drug samples were reduced with dithiothreitol and not alkylated prior to analysis. Thus the addition of hydrogen mass will be observed under reduction conditions. In the Fig.3.10, a has indicated the intact glargine elution on the LC column, as well as the reduced A chain and B chain have been separated and eluted

by the order shown as Fig.3.10B. The intact protein peak was eluted at 10.42 min, and for the reduced sample, chain A was eluted at 11.56 min while chain B was eluted earlier at 10.59 min. As the figure indicated, the m/z of intact glargine shown in Fig.3.10A1 from charge 4^+ ion of mass 1516.9179, to charge 8^+ ion with mass 758.845, and the accurate mass has been deconvoluted at 6062.7Da. The fragments ion of chain A were shown in Fig.3.10B2 with 2327.2869(1^+), 1164.1478(2^+) and 776.4466(3^+). As the result of chain B, shown in Fig.3.10B1, the ions were observed as 1872.2091(2^+), 1248.4774(3^+), 936.6144(4^+), 749.4995(5^+), 625.0945(6^+) and 535.6510(7^+). After deconvolution, the mass has been calculated to 2326.2869Da for chain A and 3742.2091Da for Chain B, which could be identified as the reduced products, comparable with theoretical molecular weight of 2326Da for chain A and 3742Da for chain B.

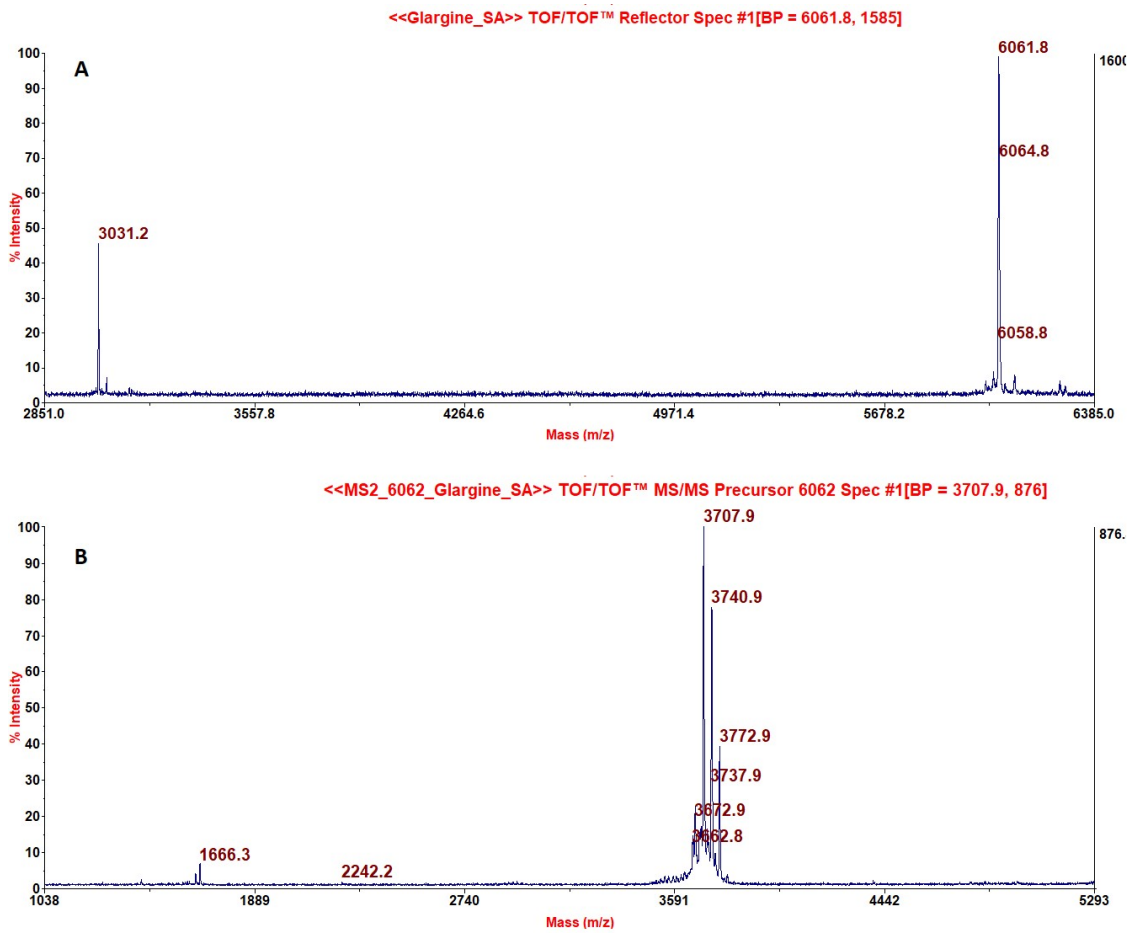

Fig 3.11: MALDI TOFTOF data of glargine with non-reduced and reduced conditions; A: MS of Intact glargine; B: Reduced glargine MS/MS with zoomed in Chain B, from the asymmetric and symmetric cleavage of disulfide bonds

In this study, MALDI-TOF[23] has been applied to measure the intact protein molecular mass as well, and our data was demonstrated. Fig.3.11A, illustrated the 6061.8(1+) peak and the corresponding 3031.2 (2+) peak of the protein. As the Fig.3.11B indicating, the chain B with the quintet peak centring with 3740.9 Da with the difference of 32 amu, as a result of the asymmetric and symmetric cleavage of disulfide bonds and chain A and chain B connection.[24] However, the strain glargine sample didn't provide distinctive peaks for structure determination by MALDI. Although MALDI is capable with relative high salt tolerance, the desalting and purification of crude protein samples are necessary prior to initial analysis.

3.4.4 Evaluation of strain proteins

Because the cultivation of glargine does not result in a strain product that is not mature enough for detailed analysis at the analytical level, we performed a bottom-up in-gel analysis on the crude recombinant protein sample for sequence and structure determination and top-down analysis on a QTOF for intact protein mass. The strain glargine samples were tested and compared with standard insulin and glargine by SDS-PAGE, and the gel image was shown in Fig. 3.12. Lane 2 and 3 were Lantus glargine standard with 10 ug and 2 μg loading. The lane 4 to 6 were the triplicates of strain glargine sample and. The strain glargine with a relative concentration around 0.05mg/mL, and the loading volume were adjusted to 25 μL. Due to the buffer components in the strain sample, the glargine bands exhibit a change in migration under high voltage, which is highlighted by a red box. The bands under the box were cut and combined into one vial and processed with an in-gel digestion with trypsin and GluC for primary structure identification.

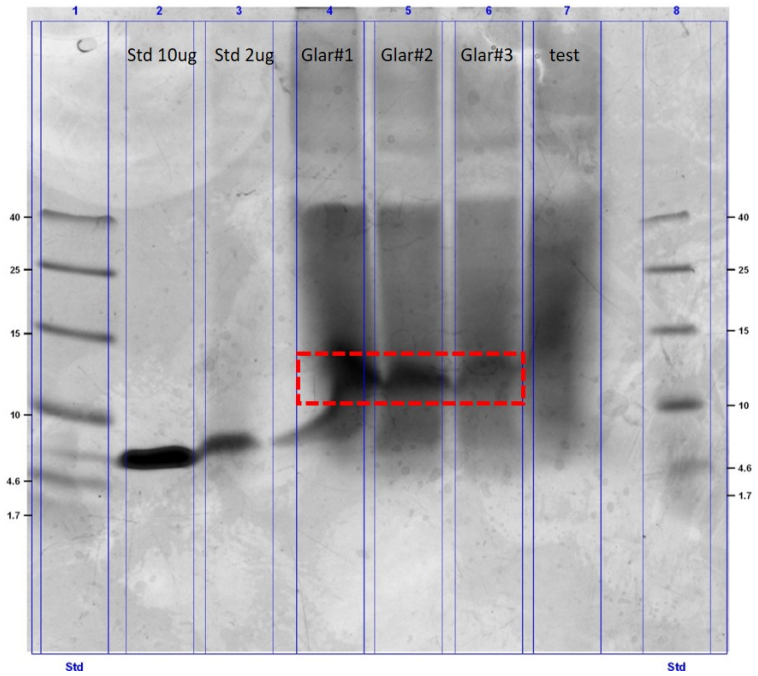

Fig 3.12: SDS-PAGE gel image of strain glargine samples; Lane 1 and 8: marker; Lane 2, 10μg control; Lane3, 2μg control; Lane 4-6, strain glargine with adjusted volume of 25 μL; Lane 7, another strain test sample.

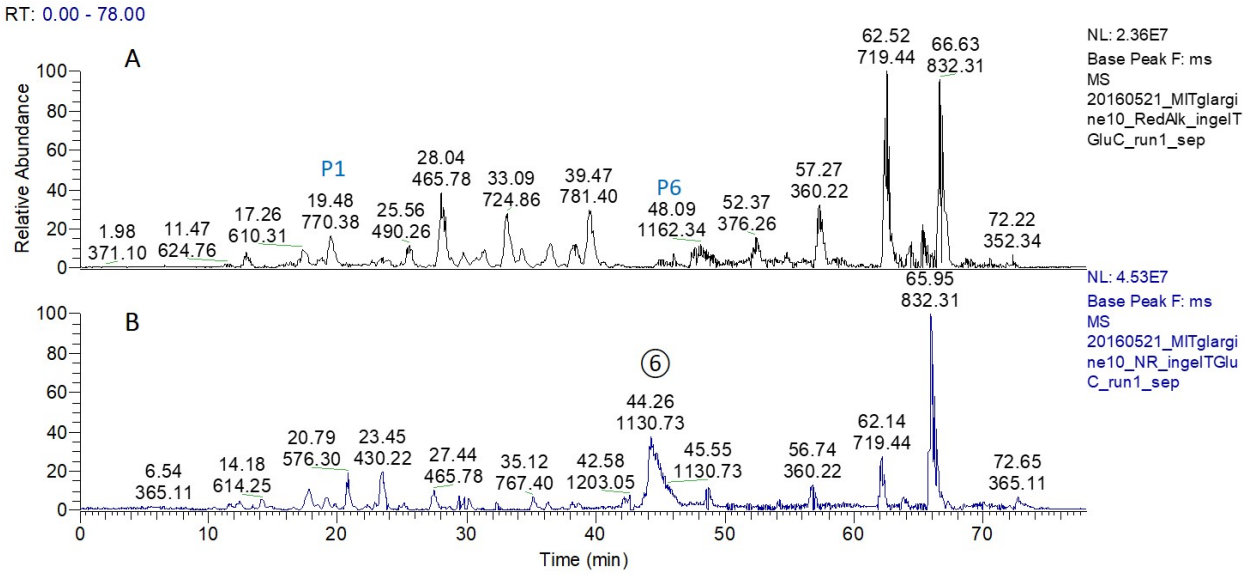

Fig 3.13: LCMS base peak of crude glargine samples; A: Reduced and alkylated condition; B: Non-reduced condition

At the pre-purified stage, the strain protein samples were examined with full sequence coverage with medium confidence. As shown in Fig.3.13, the digested protein with reduction and without digestion result were shown as A and B. The N-terminal peptide of glargine FVNQHLCGSHLVE was detected in a relative high confidence and marked P1. The 1^+ P6 QCCTSICSLYQLE was observed. These peptides had an elution time shifting due to sample impurities such as salt. With non-reduced conditions, form 6 was detected in a high abundance, shown in Fig.3.13B. With extraction of each expected peptide m/z, only form 4 and 6 were identified. Prior to the purification and polishing stage, the crude sample from the fermentation supernatant was in a good condition with sufficient concentration to allow us to determine accurate gene expression. However, the crude samples were not present at a purity level to enable the evaluation of in fermentation sample degradation or stability evaluation.

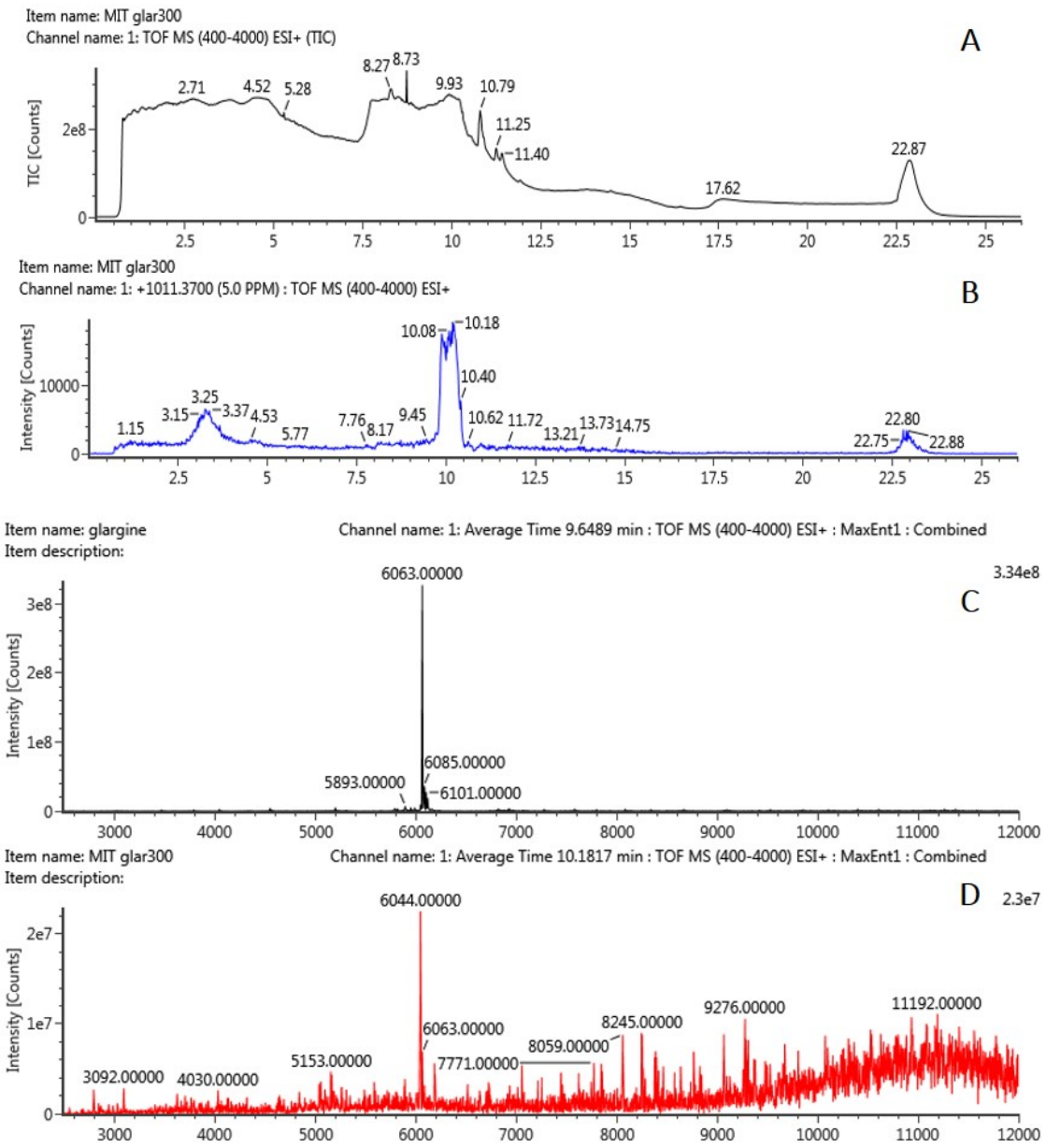

Fig 3.14: QTOF result of crude glargine samples

The crude glargine samples were analyzed on QTOF detector as well and the result compared with the standard and shown in Fig.3.14. The buffer gave a high background signal shown as Fig.3.14A, and the expected elution time 10.08min with mass of 1011.376^{+} was extracted and the chromatogram was shown in

Fig.3.14B compared with reference material elution time 9.67min. The deconvoluted mass peak of 6063Da of standard was shown in Fig.3.14C, and the crude sample was observed with 6044Da shown in Fig.3.14D. The mass shift could be as a result of dehydration (loss of H_2O), however, the purfied samples may need further polishing procedures to obtain mone conclusive analytical information.

3.5 Conclusion

As diabetes is one of the most serious disease in the world, scientists from clinical, pharmaceutical or biological areas are focusing on methods to improve the treatment of the rapidly increasing diabetic population. In our research, a series of efficient methods have been established to characterize the biopharmaceutical products from pre and post purified stage. The methodology is effective to evaluate the sequence structures and monitor the degradation. Different mass spectrometers were applied and compared, which will guide further analysis based on protein properties to select the best instrument. The crude protein drug produced from InSCyT system has been analyzed in this research and the data was demonstrated. Even the purification still need polishing by downstream teams, our data could illustrate the high capability for the production team and provide guidance to manufacture optimization.

References

[1] Roberto Zoncu, David M Sabatini, and Alejo Efeyan. mtor: from growth signal integration to cancer, diabetes and ageing. *Nature reviews. Molecular cell biology*, 12(1):21, 2011.

[2] Yuankai Shi and Frank B Hu. The global implications of diabetes and cancer. *The lancet*, 383(9933):1947, 2014.

[3] UK Prospective Diabetes Study (UKPDS) Group et al. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (ukpds 33). *The lancet*, 352(9131):837–853, 1998.

[4] Kenneth Hershon, Tom Blevins, David Donley, and Christy Littlejohn. Lower fasting blood glucose (fbg) and less symptomatic hypoglycemia with qd insulin glargine (lantus®) compared to bid nph in subjects with type 1 diabetes. *Diabetes*, 50:A116–A117, 2001.

[5] E Schober, E Schoenle, Jacobus Van Dyk, K Wernicke-Panten, Pediatric Study Group of Insulin Glargine, et al. Comparative trial between insulin glargine and nph insulin in children and adolescents with type i diabetes mellitus. *Journal of Pediatric Endocrinology and Metabolism*, 15(4):369–376, 2002.

[6] Irving S Johnson. Human insulin from recombinant dna technology. *Science*, 219(4585):632–637, 1983.

[7] Harry Keen, JC Pickup, RW Bilous, A Glynne, GC Viberti, RJ Jarrett, and R Marsden. Human insulin produced by recombinant dna technology: safety and hypoglycaemic potency in healthy men. *The Lancet*, 316(8191):398–401, 1980.

[8] Lutz Heinemann and Marcus Hompesch. Biosimilar insulins: how similar is similar? *Journal of diabetes science and technology*, 5(3):741–754, 2011.

[9] Gary Walsh. Therapeutic insulins and their large-scale manufacture. *Applied microbiology and biotechnology*, 67(2):151–159, 2005.

[10] Stanislav Goykhman, Andjela Drincic, Jean Claude Desmangles, and Marc Rendell. Insulin glargine: a review 8 years after its introduction. *Expert opinion on pharmacotherapy*, 10(4):705–718, 2009.

[11] Thomas Kjeldsen. Yeast secretory expression of insulin precursors. *Applied microbiology and biotechnology*, 54(3):277–286, 2000.

[12] Suma Sreenivas, Sateesh M Krishnaiah, Nagaraja Govindappa, Yogesh Basavaraju, Komal Kanojia, Niveditha Mallikarjun, Jayaprakash Natarajan, Amarnath Chatterjee, and Kedarnath N Sastry. En-

hancement in production of recombinant two-chain insulin glargine by over-expression of kex2 protease in pichia pastoris. *Applied microbiology and biotechnology*, 99(1):327–336, 2015.

[13] DF Steiner, S-Y Park, J Støy, LH Philipson, and GI Bell. A brief perspective on insulin production. *Diabetes, Obesity and Metabolism*, 11(s4):189–196, 2009.

[14] J. Christopher Love, Kerry Routenberg Love, and Paul W. Barone. Enabling global access to high-quality biopharmaceuticals. *Current Opinion in Chemical Engineering*, 2(4):383–390, 2013.

[15] Nabih A Baeshen, Mohammed N Baeshen, Abdullah Sheikh, Roop S Bora, Mohamed Morsi M Ahmed, Hassan AI Ramadan, Kulvinder Singh Saini, and Elrashdy M Redwan. Cell factories for insulin production. *Microbial cell factories*, 13(1):141, 2014.

[16] Han Zhang and Ying Ge. Comprehensive analysis of protein modifications by top-down mass spectrometry. *Circulation: Cardiovascular Genetics*, 4(6):711–711, 2011.

[17] Zachery R Gregorich and Ying Ge. Top-down proteomics in health and disease: Challenges and opportunities. *Proteomics*, 14(10):1195–1210, 2014.

[18] Caroline J DeHart, Ryan T Fellers, Luca Fornelli, Neil L Kelleher, and Paul M Thomas. Bioinformatics analysis of top-down mass spectrometry data. 2016.

[19] Bifan Chen, Ying Peng, Santosh G Valeja, Lichen Xiu, Andrew J Alpert, and Ying Ge. Online hydrophobic interaction chromatography–mass spectrometry for top-down proteomics. *Analytical chemistry*, 88(3):1885–1891, 2016.

[20] John C Tran, Leonid Zamdborg, Dorothy R Ahlf, Ji Eun Lee, Adam D Catherman, Kenneth R Durbin, Jeremiah D Tipton, Adaikkalam Vellaichamy, John F Kellie, Mingxi Li, et al. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature*, 480(7376):254–258, 2011.

[21] Nertila Siuti and Neil L Kelleher. Decoding protein modifications using top-down mass spectrometry. *Nature methods*, 4(10):817–821, 2007.

[22] GB Bolli, RD Di Marchi, GD Park, S Pramming, and Veikko A Koivisto. Insulin analogues and their potential in the management of diabetes mellitus. *Diabetologia*, 42(10):1151–1167, 1999.

[23] Zhaoyang Liu and Kevin L Schey. Optimization of a maldi tof-tof mass spectrometer for intact protein analysis. *Journal of the American Society for Mass Spectrometry*, 16(4):482–490, 2005.

[24] Michael D Jones, Scott D Patterson, and Hsieng S Lu. Determination of disulfide bonds in highly bridged disulfide-linked peptides by matrix-assisted laser desorption/ionization mass spectrometry with postsource decay. *Analytical chemistry*, 70(1):136–143, 1998.

Chapter 4

Stability analysis of biopharmaceutical products by LC tandem mass spectrometry

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and SPAWAR Systems Center Pacific (SSC Pacific) under Contract No. N66001-13-C-4025.

Contributions:Northeastern University, Di Wu:LC-MS experiment procedures on protein drugs, data analysis and interpretation, manuscript writing and revision; William Hancock: goal of the study, concept contribution;MIT, William Herrington; Raman Spectroscopy experimental procedures and manuscript writing; MIT Wei Ouyang: Electrokinetic concentration binding assay experiment operating and manuscript writing

Wei Ouyang, Sung Hee Ko, Di Wu, Annie Yu Wang, Paul W. Barone, William S. Hancock, and Jongyoon Han Rapid Activity Assessment for Biologics using Electrokinetic Concentration Binding Assays with Charge-Modulated Receptors, *Analytical Chemistry*, 2016, 88 (19)

William F. Herrington, Gajendra Pratap Singh, Di Wu, Paul Barone, , William S. Hancock, Ram Rajeev Optical Detection of Degraded Therapeutic Proteins, *Scientific Reports* (Accepted)

4.1 Abstract

Biosimilar therapeutics are marketed worldwide nowadays and stability assessment is necessary to validate the long term safety, purity and potency of the products. In this study, a range of controlled stressed conditions, including temperature, humidity and pH, have been applied to a set of therapeutic proteins. The resulting degradations from the accelerated stability study are monitored with LC-MS-MS. With the high efficiency, accuracy, and sensitivity of mass spectrometry, deamidation and oxidation variants were detected. The differences in structural integrity and conformational stability profiles were monitored and potential degradations were evaluated by the MS analysis. In addition, real time Raman spectroscopy techniques were utilized in this study and demonstrated a powerful ability to monitor degradation reactions as well as determine the identity of a given biosimilar therapeutic products as well as to detect product variants. Data obtained from Raman spectrometry and electrokinetic measurement are comparable with mass spectrometry, which can differentiate modified protein from the normal form. The study indicates these techniques are sensitive and robust and will provide an alternative method to conventional analytical methods for the monitoring of the manufacture of protein therapeutics. Such methods have the potential to provide quality assessment of protein therapeutics manufactured in remote locations.

4.2 Introduction

Biopharmaceuticals have captured global attention and increasing market share in recent decades especially from 1980s when the first biosynthetic insulin has been produced and approved.[1] The high sensitivity and specificity of protein therapeutics are important for safety and efficacy of these drugs in patients. After accurate diagnosis and effective treatment programs biologic drugs can be set up to treat difficult diseases. Although an increasing panel of such drugs have been discovered and studied, there are still many challenges in the biopharmaceutical market, such as the cost, time of treatment and targeting specificity and to assure the efficacy of biosimilars solutions are crucial.[2, 3] The safety and quality of a given biopharmaceutical may change significantly with the generation of subtle differences in the molecular species of the drug during the steps of production, transportation and storage conditions. In previous studies, we characterized the highly complex protein structures produced by the novel production system, InCSyT (Integrated and Scalable Cyto-Technologies), which is designed to produce high quality biologic medicine in the field. We have demonstrated that the biologics from this platform are with high quality produced under optimized conditions.[4] However, the post-production quality will have additional quality concerns which are based on formulation and stability issues and thus we need to detect and evaluate further variations in the products. Stability is always a key feature of biological medicines. Furthermore the knowledge of potential toxicity caused by subtle changes in the therapeutic can be used to improve the production system. In this study, we have designed a set of experiments forcing the generation of stability degradations and then measured the resulting degradation products by different techniques.

Mass spectrometry is a powerful technique to characterize and quantitate biological products. Bottom-up analysis is commonly applied in many studies to identify and evaluate the structure, sequence or any mutations in the sequence of the protein product. To perform a bottom-up analysis we use enzymes with

high specificity and efficiency to digest the large protein molecule into small peptides. However, while a completed analysis can be achieved by mass spectrometry, the challenges of on-line samples preparation mean that timing of the sampling will be a challenge for real time monitoring of the fermentation. Therefore, for the InCyST program, other measurement techniques were considered for on-line measurement of the intact protein during the fermentation.

As we know, the drugs bioactivity was determined by the binding of the drugs to their physiological targets or receptors.[5] Biologics are efficacious only when they are bound to their physiological targets (receptors). Both the binding equilibrium and binding kinetics, which correspond to the in vivo efficacy and duration of efficacy, are deemed as key metrics in drug discovery and pharmacokinetics.[6] In this project, one of our collaborator teams has applied the molecular charge modulation (MCM) and electrokinetic concentration (EC) to assess the protein activity.[7] They have obtained distinctive results of the correlation of degradation and bindings.

Another collaborator team working with Raman spectroscopy has also provided effective experimental data. Raman spectroscopy is a technology to verify the primary and secondary structure of proteins since 1950s.[8] Previous work on proteins characterization by Raman spectroscopy has focused on high concentration in solution or in solid powder form. Our technique was with sensitive detection and to powerfully monitor the single molecule level biologics by Surface Enhanced Raman Scattering technique (SERS).[9] If the biologics was not in control of its orientation, the absorbance will be differed by the sensitive SERS technology.[10] Our Raman collaborator team has reported on the accurate identification and quantification of various therapeutic proteins by monitoring the SERS signal shifts according to the protein backbone or amino acids changes.

4.3 Experimental

4.3.1 Chemicals and materials

The recombinant human growth hormone was purchased as Humatrope (Lilly, USA). The recombinant GCSF was purchased from Neupogen (Amgen, USA). The chemicals including acetonitrile, hydrogen peroxide, Tris-HCl sodium phosphate, guanidinum chloride, dithiothreitol, iodoacetamide, and trypsin were analytical grade, obtained from Thermo (Thermo Fisher, USA).

4.3.2 Stressed conditions for oxidation

hGH was initially reconstituted in sample buffer (10mM sodium phosphate, pH=7), for a concentration of 2 mg/mL. The oxidation variants were prepared by addition of hydrogen peroxide to a final concentration of 0.5% (v/v), then incubated under 37°C overnight. After the incubation, the oxidized hGH were dialyzed back into sodium phosphate sample buffer, to avoid further oxidation.

G-CSF, initial concentration at 0.6 mg/mL was reconstituted into sample buffer (20mM glutamic acid with 5% sorbitol, w/v, pH=4.4). The oxidation variants were prepared by addition of hydrogen peroxide to a final concentration of 0.5% (v/v), then incubated at 37°C for 2 hour. The oxidized G-CSF was then dialyzed back into glutamic acid sample buffer and was executed by 200μL buffer at 13,000 rpm for 20 minutes, repeated three times in a 500μL 10kDa Amicon (EMD Millipore Corporation, Merck, Germany) centrifugal filter.

The test samples were produced by proportionally mixing of control and oxidized variants, with oxidation percentage at 10%, 40%, 70% and 100% (v/v). All test samples and control were lyophilized overnight and reconstituted again to give a final concentration 2 mg/mL of hGH, 0.6mg/mL of G-CSF, and stored at -80°C

for further analysis.

4.3.3 Stressed conditions of deamidation

100μL hGH was reconstituted into 0.1M sodium borate buffer with pH 8 and incubated under 37 degree for total four weeks. 20μL sample was collected at each time point for control, 1 week, 2 week, three week and four week. The samples were stored at -80°Cafter collection and all samples were performed in-solution digestion and analysis at same time.

4.3.4 In-solution digestion

hGH sample were dissolved in 6M guanidinium chloride, reduced with 10 mM dithiothreitol at 70°C for 30 minutes and followed by alkylation with 55mM iodoacetamide under room temperature in dark conditions. Proteins were dialyzed via a 10kD membrane Amicon centrifugal filter at 13,000 rpm for 15 minute and three times in Tris-HCl buffer (pH=6.8). The following in-solution digestion process was used trypsin and the solution was kept overnight at room temperature to avoid artificial oxidation. The digestion was terminated by addition of 20μL 5% formic acid.

GCSF samples were adjusted pH to 3 by HCl, and followed by digestion with pepsin at 37°C for 30 minutes. The digestion process was terminated by adjusting pH to 8 by 0.1M NH_4HCO_3. Protein samples were then ready for LC-MS analysis and the remained materials were aliquotted to 20 μL sample and stored at -80 °C for further analysis.

4.3.5 LC-MS measurement

LC-MS analysis used an Ultimate 3000 nano LC pump (Dionex, Mountain View, CA) and a self-packed C18 column (Magic C18, 200Å pore and 5 μm particle size, 75μm internal diameter by 100 mm) connected to a coated 10μm internal diameter emitter (New Objective, Woburn, MA). LTQ-Orbitrap XL mass spectrometer was connected (Thermo Fisher Scientific, San Jose, CA) through a nanospray ion source (New Objective, Woburn, MA). Mobile phase A used 0.1% formic acid in HPLC grade water and mobile phase B was using 0.1% formic acid in acetonitrile. During sample injection, the flow rate was set at 250 nL/min with 2% B for 25 min. The flow rate of the gradient was set at 200 nL/min, with mobile phase B, 0-60 min 2-40%, 60-70 min up to 90% , 70-75 min kept for 90% and 75-78 min 2%B. The mass spectrometer was operated in a data dependent mode to switch between MS and CID-MS^2. Briefly, after a full-scan MS spectrum from m/z 400-2000 in the ion-trap, 8 CID-MS^2 activation steps were performed on the 8 most intense precursor ions from the full scan. All control and variants samples were run in triplicate.

For peptide identifications, raw data were searched against human growth hormone and Granulocyte-colony stimulating factor sequence using BioPharma Finder 2.0 software (Thermo Fisher Scientific). For peptide mapping, searches were performed using a single-entry protein FASTA database with oxidation and deamidation set as variable modifications, 10 ppm mass accuracy, and a confidence level of 0.8 for MS/MS spectra. Final confirmation of the peptide identification was determined by manual inspection, extracting the base peak from the chromatogram and matching the MS^2 fragmentation data with theoretical prediction. The modification percentage was calculated by peptide peak area. Both non-degraded and degraded peptides were evaluated in the same LC-MS analysis with characteristic m/z differences and elution time shifts. The quantitation of the degradation could be calculated as the following: ratio of Degradation (such as oxidation, deamidation, etc) = peak area of [(degraded peptide)/(degraded peptide) + (Non-degraded peptide)] $\times$

100%.

4.3.6 Raman techniques

Raman spectra were collected using a purpose built system designed around the low volume sample holder illustrated in the Figure. The sample holder featured a 100m thick fused silica sampling window used for both excitation and collection of the Raman signal. A microscope objective was used to focus the excitation light and collect the Raman scattered light. A gold mirror at the back plane of the sample holder, positioned near the focal point of the microscope objective (NA=0.75, 40x magnification), increased the excitation energy by redirecting the excitation light back through the sample. Additionally, the mirror increased the systems Raman signal collection by redirecting forward scattered Raman light back towards the microscope objective for collection.

Raman signal levels increase with excitation intensity, so most of the Raman signal in the system was generated close to the focal point of the microscope objective. The liquid containing well of the sample holder was 4mm wide, 6mm long, and 0.8mm deep along the optical path for a total volume of approximately 19L. To use the holder a 20-25L drop of material was added to the well, over filling it slightly, and then the well was capped with the mirror. This prevented the formation of an air bubble between the sample and mirror. The sample holder was designed so that the focal point could be placed near the mirror for increased collection as described above, and the depth of the sample holder was chosen so that the excitation spot was far from the fused silica window to minimize the background signal from the window. The walls and mirror in the sample holder were metallic, so they added no background Raman signal. The inverted configuration ensured that neither the sample holder nor the microscope objective needed to move during sample loading or cleaning operations. This allowed the system to be used without re-aligning the optics each time a new

sample was presented to the system.

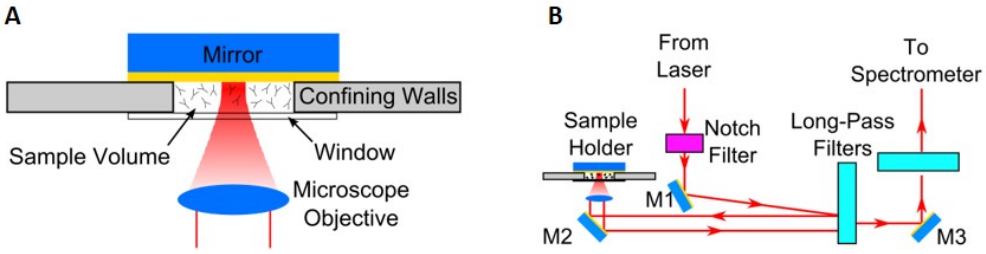

Fig 4.1: Scheme of Raman spectroscopy mechanism; A: Raman spectroscopy sample holder; B: Raman spectroscopy mechanism

A schematic diagram of our sample holder is shown in Fig.4.1A and the Raman system shown in Fig.4.1B. The total volume is kept low by recognizing that the volume of sample directly illuminated by the excitation light will produce the strongest Raman signal. The microscope objective used in this system is a Nikon Plan Fluorite, 40x 0.75NA objective.

4.3.7 Electrokinetic Concentration (EC) Binding Assays

Our partner group[7] has previously reported on EC devices for simultaneous concentration and separation of charged species based on their mobility. Under the voltage configuration shown in Fig.4.2(a), a spatially decaying electric field is created in the vicinity of the cation-selective Nafion membrane due to the ion concentration polarization phenomenon, which imposes an electrophoretic force on charged species.

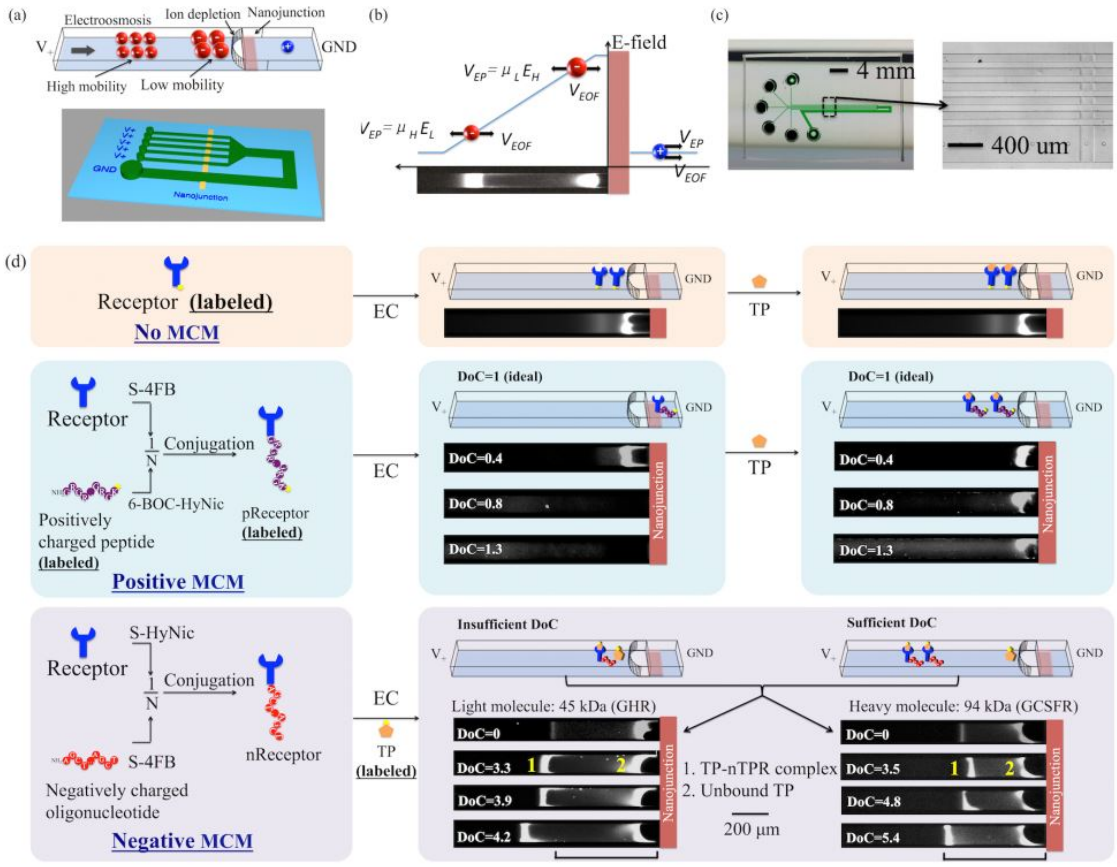

Fig 4.2: Principle of MCM-EC

Meanwhile, all species are brought downstream by the electroosmotic flow. Therefore, negatively charged species are concentrated where electrophoresis and electroosmosis balance, with high mobility species being farther and low mobility species being closer to the nanojunction. It is also to be noted that positively charged species go downstream without being concentrated due to the cation-selectivity of the nanojunction. Fig.4.2(b-c) shows the design of the device, with five channels in parallel for testing five samples at the same time. EC can easily separate negative species with high mobility difference, such as aptamers and aptamer-IgE complex, at voltages $<$ 100 V.37 However, as with other mobility-based separation technologies, it

is incapable of separating species with minimal mobility difference, including many proteins and protein-protein complexes.Fig.4.2 shows the positive and negative MCM for biologic-receptor binding assays.

4.4 Results and discussion

4.4.1 LC-MS results of hGH

4.4.1.1 Primary structure identification

Human growth hormone is essential for the treatment of children with growth hormone deficiency or hypopituitarism.[11, 12] The recombinant human growth hormone (rhGH) produced by the expression in the Pichia yeast strain had the identical sequence to natural growth hormone with 191 amino acids. The 191 amino acids sequence is cleaved to 21 peptides by proteolysis with the enzyme trypsin. Under non-reducing conditions, the intra disulfide linkages at Cys53-Cys156 and Cys182-Cys189 are maintained. The precise sequence and disulfide linkage are critical to the function and toxicity of biopharmaceuticals.[13] Prior to any degradation analysis, peptide mapping is the fundamental procedure to assure a completed protein structure was detected. The tryptic peptide mapping was shown in Table4.2

The samples measured are all obtained by purchase of the corresponding FDA approved drugs with known primary sequences and primary structures were confirmed. The project aims were to design and control a reproducible set of stressed conditions and to subsequently evaluate the stability. At the same time, this experiment aims to generate specific degradation reactions and not denature the protein structure. The three dimensional structure of human growth hormone and the sequence information are shown in Fig. 4.3. The major degradation pathways of growth hormone were observed to be oxidation of methionine and deamidation of asparagine residues respectively. In this study, these variants were primarily evaluated by MS and correlated with Raman measurements.

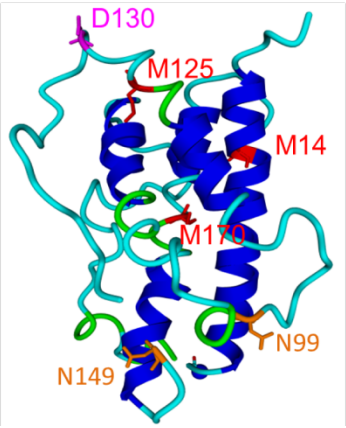

Fig 4.3: 3-D structure of hGH

The amino acid residues located in the flexible outer loop of hGH, the outer loop of growth hormone (residue M14) will be more easily oxidized, especially under stressed conditions. However, the methionine residue located in the hydrophobic inner core will not. A structural change induced by unfavourable manufacturing or storage conditions can therefore have a significant effect on the quality of a protein drug. It is important, therefore, to keep a consistent temperature, moisture, oxygen proportions or other conditions at remote supply areas or during the transportation of the product. The stability of the biopharmaceuticals will be thus crucial for the drug quality, then for the patient's safety. Our work is to effectively evaluate the drug stability and measure any degradation reactions with our optimized methods.

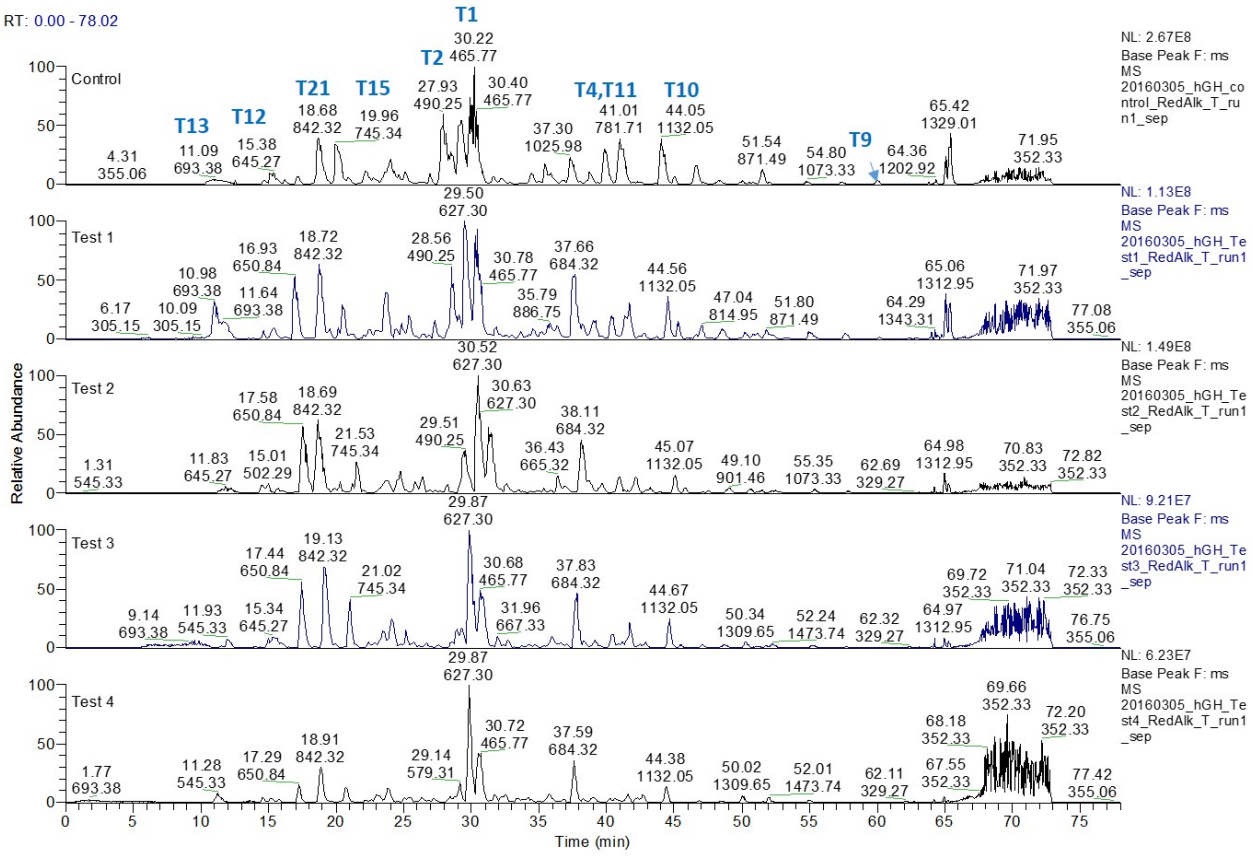

Fig 4.4: Base peak ion chromatogram of the tryptic map of hGH test samples

The stressed oxidized condition was primarily targeted to the surface methionine residues instead of denaturing the protein and exposing the buried methionine residue at position 170. To ensure that each sample as shown in Table.4.2, contained the correct sequence of amino acids we used trypsin digestion for identification, although in another study a combination of trypsin and LysC were used and gave similar results.[14] As Tab. 4.2 indicated, the T3,T5, T17 and T18 are small peptides and difficult to detect but were observed as the result of incomplete trypsin digestion, such as in the peptide T3-T5, T17-T19.

4.4.1.2 Oxidation determination of hGH

We observed that artificial oxidation of methionine residues occurred when the electrospray ion source is operated under high voltage. While a peptide with oxidized methionine would be expected to elute earlier in RPLC due to the greater polarity of the sulfoxide moiety, the artifactual oxidation that occurs after separation, would result in the oxidized peptide eluting at the same time as the non-oxidized peptides but with an increased m/z value. The artificial oxidation may not occur all the times, however, this calculation will be used to obtain an accurate oxidation ratio. In this study, we applied the coated tip for ESI to minimize any artifactual oxidation in the analysis so that the two types of oxidation variants will be differentiated by elution order differences, which shown in Fig.4.5. The voltage will approach the mobile phase solvent via the emitter that was equipped with conductive coating. The artifact oxidation induced by the high voltage will be differentiated from the real oxidation according to the elution time.

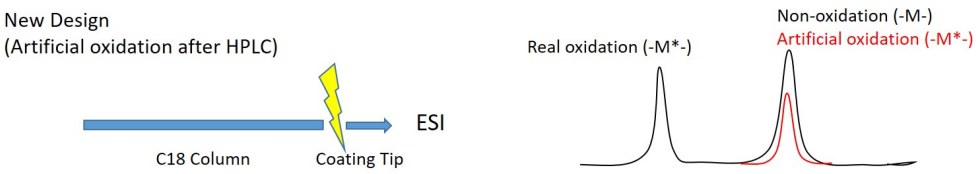

Fig 4.5: Scheme of coated tips

Considering the product stability and the type of formulation, previous researchers have indicated that higher oxidation levels may occur in lyophilized proteins than with the liquid products which usually contained a trace amount.[15, 16, 17, 18] The oxidation reaction could occur readily with oxygen existing in the product sample and is promoted by the presence of a less stable denatured structures either generated during the lyophilization process or any storage conditions. In this project, the samples were lyophilized to a stable form and then diluted to the required concentration for evaluation. All samples were prepared using the

same conditions, so the levels of observed oxidation in the reference standard could used as a control.

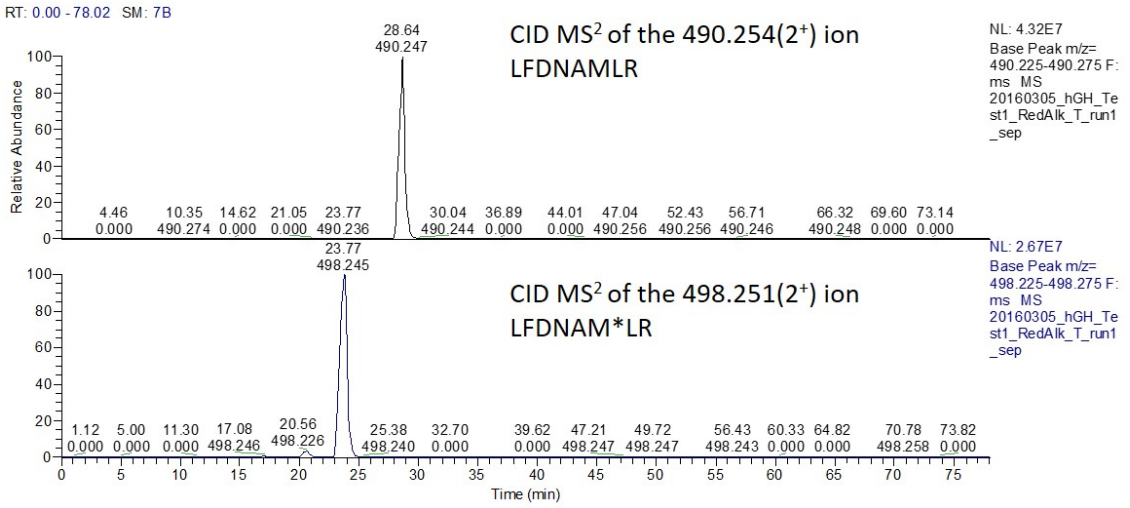

Fig 4.6: Extracted ion peak of T2 and the oxidized T2 peptides from the tryptic digested rhGH test1 sample

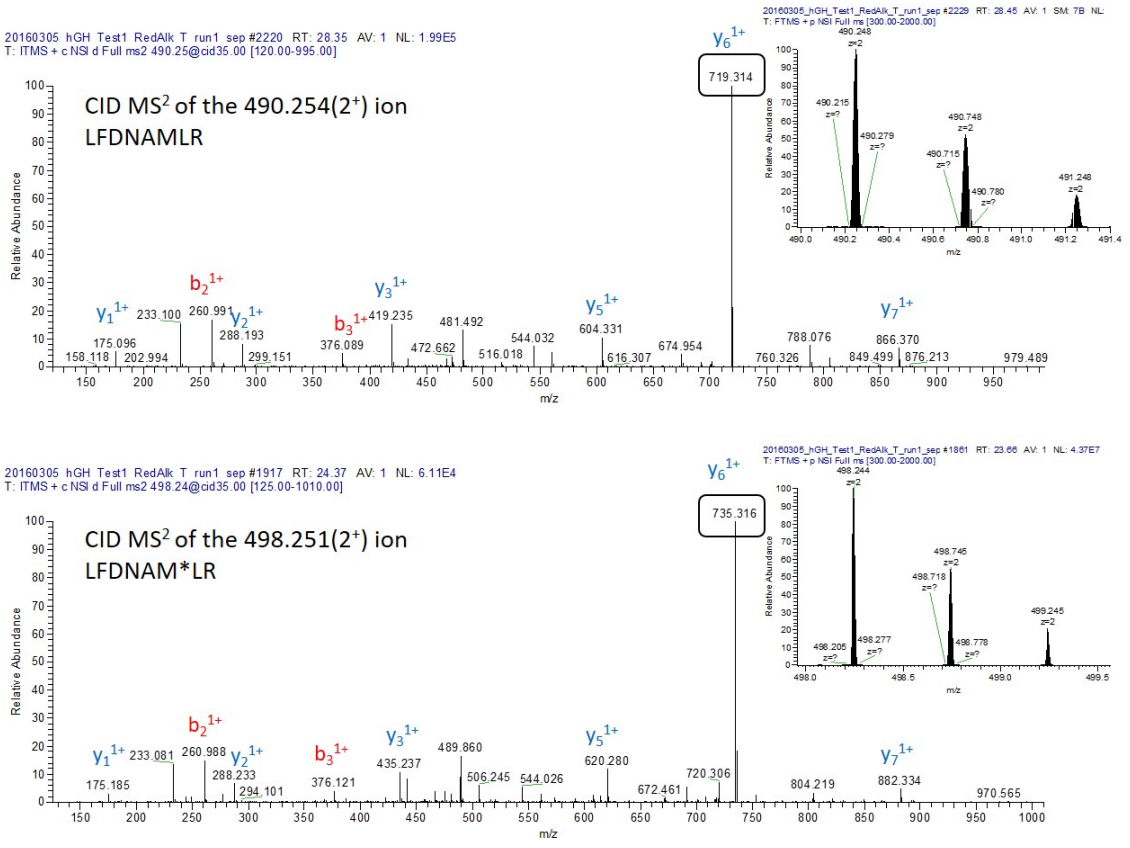

Fig 4.7: LC-MS analysis of the T2 and the oxidized T2 peptides from the tryptic digested rhGH test1 sample

As shown in Fig.4.6, the elution order of oxidized peptide and non-oxidized peptide could be differentiated by the retention times due to differences in hydrophobicity. With oxidation, the peptide T2 LFDNAMLR, the oxidized methionine will bring a more hydrophilic property to this peptide, which will elute earlier than the normal peptide on the reversed phase separation.

Both non-oxidized and oxidized peptides were evaluated in the same LC-MS analysis with m/z difference and elution time shifts. Fig.4.7 presents one example of the extracted ion chromatogram (XIC) of non-oxidized T2 and oxidized T2. The methionine residue locates at position 6 in this peptide T2. By CID fragmentation, the b6, b7ions and y3 to y7 ions will have a mass change due to the methionine oxidation. As shown in Fig.4.7, it is indicated here that y6 ion 719.31^{+} has been oxidized to 735.31^{+}, with the change of amu as did the y7 ion from 866.37^{+} to 882.33^{+}.

This base peak chromatogram was made up of the extracted ion peak from sample Test1, with a 20% mixing of control and harshly stressed oxidized sample Test4. The abundance shown in Fig4.8 illustrates the oxidation amount, which indicates the trend of oxidation accumulation with an increased mixing ratio of oxidizing agent to sample. However, the residue Met14 in Test1 had a 36.3% oxidation, while in Test2 a value of 30.4%. Such minor variations are more likely caused by differences in sample handling. The data was summarized in Supplementary information.

The residue Met125 is not as easily oxidized as Met14 due to the conformation location difference. In this study, with overnight incubation, the residue Met125 was detected 96.20% oxidation in Test4, with a higher ratio than Met14. However, the oxidation on Met14 accelerated faster than on Met125 during the overall stressed conditions study. As the residue Met170 is located at the inner core of the protein it was resistant to oxidation, the data illustrated for the protein used in this study indicates a robust structure under stressed conditions, and that the developed method can effectively induce specific degradations of the biological

drug samples. Under very harsh oxidation conditions the reaction of residue Met170 indicated loss of native protein structure and thus the stability of this residue allows an analyst to monitor the structural stability of recombinant human growth hormone drugs. We conclude that the stressed oxidation design used in this study is effective for evaluating product stability. To evaluate other storage stability potentials, we would adjust temperature, humidity or incubation time in the future.

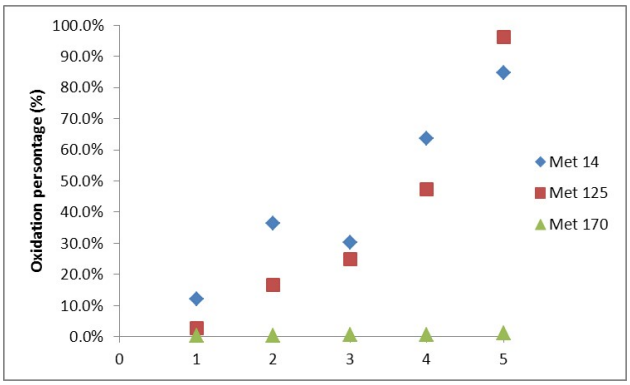

Fig 4.8: hGH oxidation result

4.4.1.3 Deamidation result of hGH

The generation of deamidation used incubation at high pH levels and higher temperature than normal drug storage to forced more degradation reactions.[19] In this study, we incubated the protein for a total of one month and analyzed each week time point to monitor the tendency of deamidation in hGH.

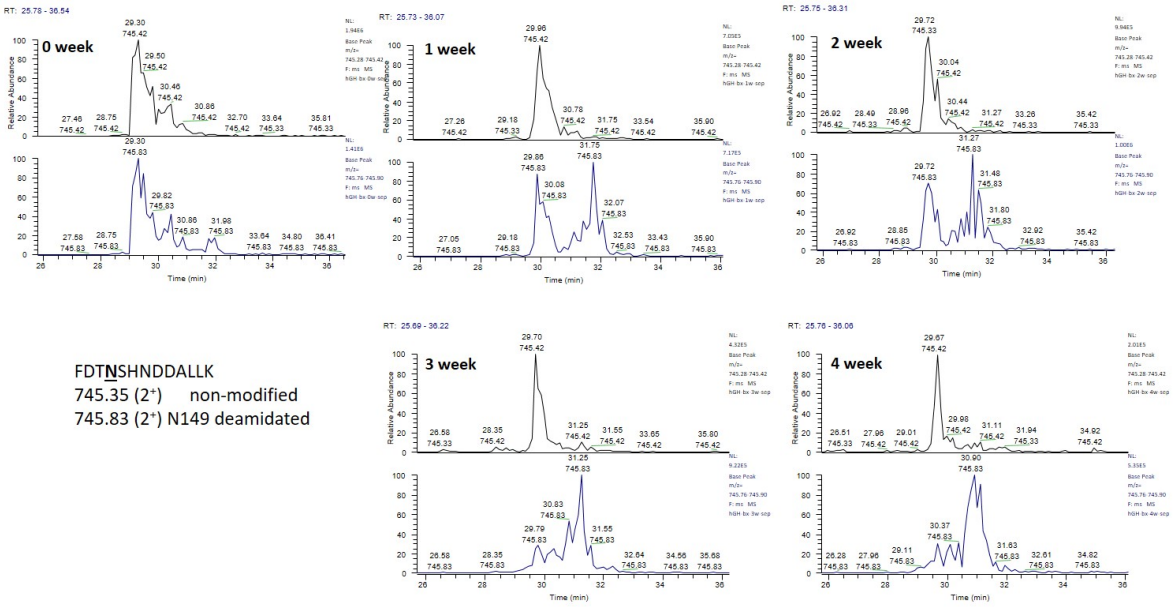

Fig 4.9: Deamidation characterization; XIC of peptide FDTNSHNDDALLK of each test sample (745.35^{2+} and 745.83^{2+});

The peptide FDTNSHNDDALLK with the most labile residue Asn149 is the target peptide site.[20] The extracted ion peak with m/z 745.33^{2+} eluted at 29.30 min was shown as Fig.4.9, while the peak with m/z 745.83^{2+} eluted 31.75 min contained the expected deamidation site. The changes of the structure resulted in a 1 Da mass change and a polarity difference which led the two minutes elution delay in the RPLC separation. The results showed the increasing level of deamidation of peptide T15 from the control(0 week) with 5.16% variant and accumulated to each week point 40.34%(1 week), 43.53% (2 weeks), 74.13% (3 weeks) and 82.00% (4 weeks). As mentioned before, this harsh condition may induce a large amount deamida-

tion, indicating other potential residues such as Asn99 and Asn152 deamidation or Asp130 isomerization to isoAsp. however, in this study we didn't characterize these additional sites of deamidation or differentiate between Asp and isoAsp which were not the focus of this study.

4.4.2 Results for GCSF

4.4.2.1 Primary structure identification

The test samples to be used to evaluate of the Raman spectroscopy were produced by proportionally mixing of the control and oxidized variants, with a resulting oxidation percentage at 10%, 40%, 70% and 100% (v/v). With experimental condition limitations, the preparation of a GCSF oxidation standard focused on a harsh condition with the expectation of a fully oxidized sample for the subsequent preparation of a well controlled oxidation sample set.

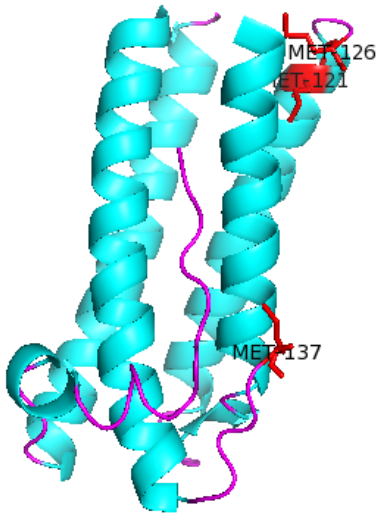

Fig 4.10: 3-D structure of GCSF

We have illustrated the three-dimensional structure of GCSF in Fig.4.10. The potential modification site amino acids have been highlighted. As similar with human growth hormone, the methionine residues are reported as being oxidized, but the tendency of reaction at each methionine positions are different. The sequence of GCSF starts from a N-terminal methionine residue that is located on the outer space of the structure. This N-terminal methionine is the easiest one to be oxidized. There are three other methionines, Met122, Met127 and Met138 located at the inner side of its structure. The tendency of oxidation of these three methionine residues are variable based on different formulation buffers and storage conditions and Met127 and Met138 may switch with the tendency for oxidation levels depending on the conditions.

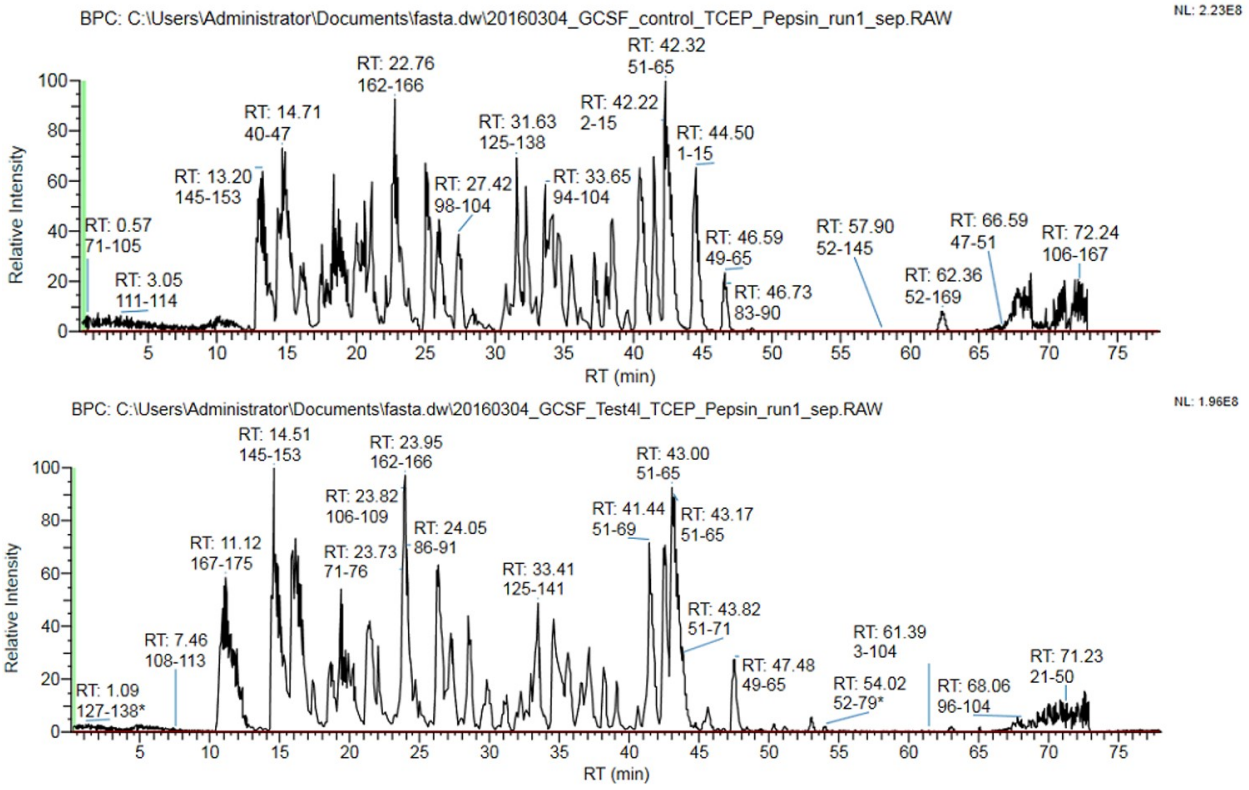

Fig 4.11: Base peak ion chromatogram of the pepsin map of GCSF test samples; Top: Control standard; Bottom: Test sample

As indicated, the elution and base peaks of GCSF test samples have been shown in Fig. 4.11. The upper one

is the elution base peaks of the reference material, and the lower one is the test sample. The peptides were labeled by the residue amino acids position numbers.

4.4.2.2 Oxidation of GCSF

The GCSF analyte was digested with pepsin under acidic conditions according to the optimal pH for this enzyme activity. The N-terminal peptide MTPLGPASSLPQSFL is the primary methionine oxidation peptide. Both non-oxidized and oxidized peptides were evaluated in the same LC-MS analysis with m/z difference and elution time shifts. Fig.4.12 presents one example of the extracted ion chromatogram (XIC) from the control sample of non-oxidized P1 and oxidized P1.

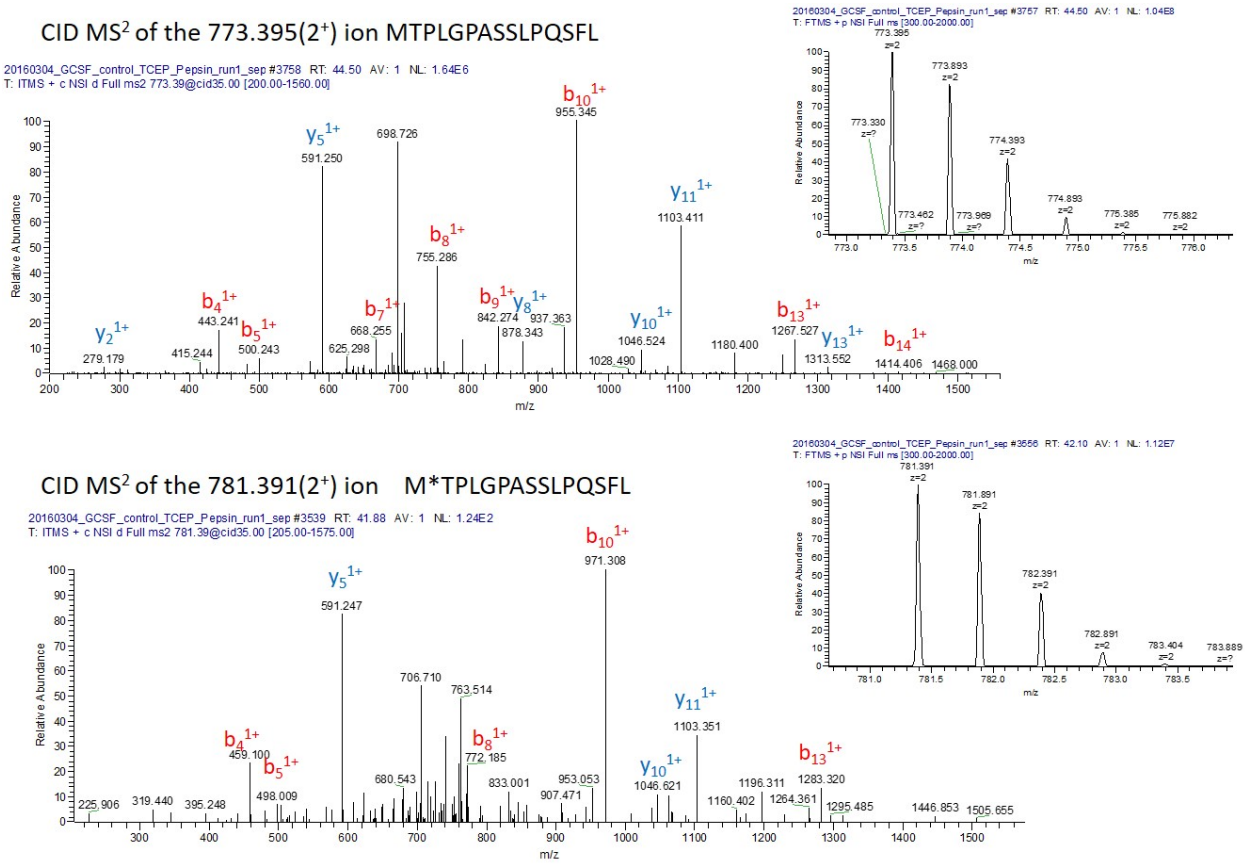

Fig 4.12: LC-MS analysis of the N-terminal and the oxidized N-terminal peptides from the pepsin digested GCSF sample

As shown in the figure, the diagnostic peptide was detected with m/z 773.39^{2+}, and the oxidized form 781.39^{2+} indicating that the N-terminal residue methionine was the most readily oxidized site. The CID fragmentation MS2 data was shown as Fig.4.12. All the b ions will add an oxygen mass in the oxidized status, while the y ions will keep the same mass since the residue methionine located the first position. The oxidation data of all four methionine residues in the control and test samples were shown in Fig.4.13, which illustrated the percentage of the oxidation level. The test sample has indicated the significant increasing of the oxidation of each methionine residue. In this case, our designed stressed method is effective. This

strategy is also unique to target only methionine oxidation instead of inducing other modifications, such as there were no increasing deamidation was detected in this test sample. The levels of the oxidized base peak area increased significantly over the harsh conditions. The data here indicated that the stressed sample Test4 has 92.5% oxidation and the control has only 19.30% on Met1. As previous studies described, the oxidation would occur M1> M138> M127> M122,[21] and our data shows comparable results with that reported study. The data was provided in the supplementary information Table4.5.

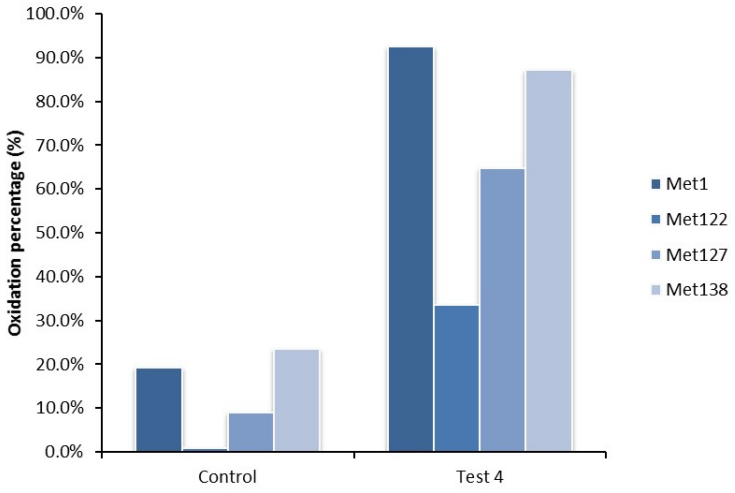

Fig 4.13: Stability measurement of GCSF oxidation result

4.4.3 Raman measurement

The Raman spectra of a protein will be composed of peaks due to the content of amino acids, the secondary structure of the protein, and bonds such as the disulfide bond.[22] The individual amino acids which make up the proteins have their own possible vibrational modes and contribution to the proteins Raman spectra, but in a protein the strongest amino acid Raman signals are seen in the amino acids which have an aromatic ring structure in their side-chain including Tyrosine, Phenylalanine, Tryptophan, and Histidine as well as sulfur containing side-chain.[23]

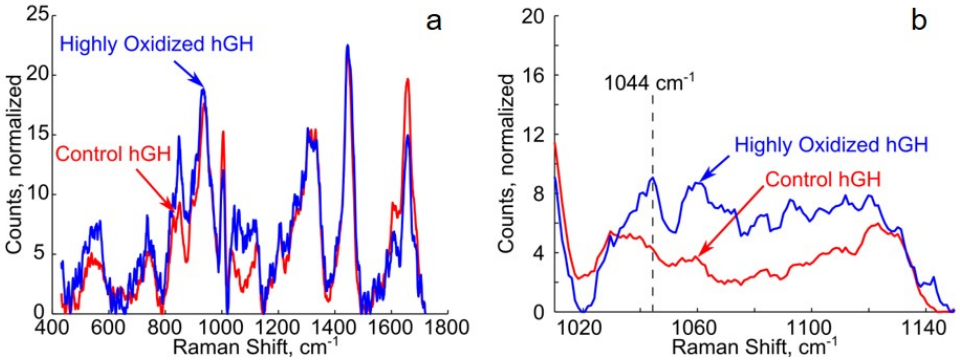

Fig 4.14: Raman spectra of control and oxidized hGH; a, Raman spectra shift from 400 to 1800 cm^{-1}; b, zoomed in spectra Raman shift from 1020 to 1140 cm^{-1}; control in red, test hGH in blue

The Raman peaks due to Tyrosine, Phenylalanine, and Tryptophan occur at a small enough Raman shift to be visible in the system which captures signal from 450 cm^{-1} to 1700 cm^{-1}. The peaks associated with the aromatic side-chain in Histidine (around 3100cm-1) and the sulfhydryl bond in cysteine (2500-2600cm^{-1}) occur beyond the range of the system. The Raman signal from disulfide bridges appears at a shift around 500cm^{-1}, which is near the lower limit of the systems range in a region impacted by the signal from the fused silica of the sampling window. The Amide I band due to the protein backbone produces a visible and relatively strong signal for all four proteins of interest in the region 1500-1700cm^{-1}.

Raman spectra for reference and highly oxidized hGH Data is shown as normalized counts per second after background subtraction. Changes are seen in the 847 cm^{-1} Tyrosine band, between 1100 and 1150 associated with the oxidized methionine residue, and in the Amide I band. The spectra shown in Fig.4.14b has indicated significant shift differences between the test hGH and the control. Our collaborator then applied Circular Dichroism Spectrometer (Aviv Model 202) to measure the same sample. They didn't detect major changes in the circular dichroism spectrum in both control and stressed sample, which indicated the alpha-helix segment in hGH was still folded. As reported, α-helix is about 45% of the whole structure of hGH.[24] Thus, the result from the Raman study indicated the shift changes may be induced by structural changes in the non-helix regions in addition to the oxidation of methionine residues during the stressed condition process.

Although Raman spectroscopy is not a technique for quantitation in our project, it can differentiate the oxidized sample from the standard material, which is powerful to be applied in the on-line QC measurements for InSCyT platform in the future. This will allow the scientists to monitor the product quality efficiently at the early stage after manufacture in a remote area with medical need.

4.4.4 Electrokinetic Concentration Binding Assay

The platform designed by our collaborator for the electrokinetic concentration binding assay enabled the application to separate the bound and unbound species with a mobility difference, so that to differentiate the inactive degradated biologics from the active control sample. We have demonstrated comparable results for our mass spectrometry methods with the distinctive measurement of the equilibrium and kinetic binding behavior of both hGH and GCSF degradation samples and the correlation between different levels of biologics analysis technologies has been shown in Fig.4.15.

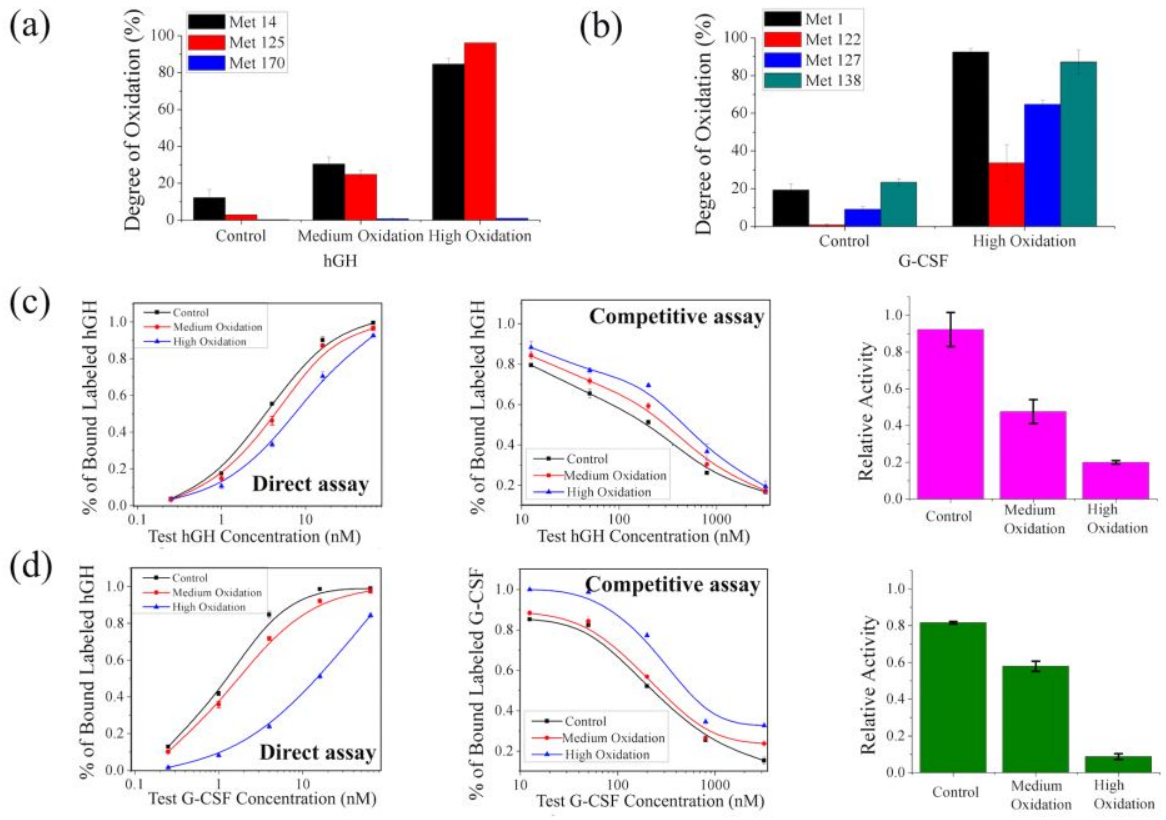

Fig 4.15: Correlation between different levels of biologics analysis technologies[7](a) Degrees of oxidation in hGH samples; (b) Degrees of oxidation in GCSF samples; (c) Results of testing the oxidized hGH samples by MCM-EC platform; (d) Results of testing the oxidized GCSF samples by the MCM-EC platform; (e) Bioassay results of oxidized hGH; (f) Bioassay results of oxidized GCSF.

Our collaborator has designed and applied three different assays to evaluate a biologic drug by this MCM-EC platform,[7] which included direct assay, competitive assay and degradation determination assay. The degraded drug samples were assessed by the MCM-EC platform, which indicated the decreased product activity when oxidation increased which demonstrated the capability of this system. The MCM-EC platform applied in this project is efficient, economical and with easy-control.

4.5 Conclusions

In this chapter, we have introduced a series of characterizations analytics tools that can be used to monitor the stability of protein drugs. The products under stressed condition illustrate the stability of a given bio-therapeutic and the susceptibility of different residues to degradative reaction and provide an example of the evaluation of pre-commercial products. The results shown here indicated that the refinements made by our MIT collaborators have produced a sophisticated Raman technology which has potential to detect the modifications in protein therapeutics both in the manufacturing process and on storage.

Bottom-up LCMS techniques may require more offline times for sample preparation and analysis, however, by this way, there will be more detailed information on the exact structural changes in the biologic product, such as the level of change in specific amino acids. On the other hand, other techniques used in this study such as Raman spectrometry and electrokinetic concentration binding assays will differentiate the modifications directly on intact proteins in a short assay time period. These technologies are very efficient for a quick product evaluation or diagnosis of product quality.

We have introduced the design of stressed modifications to generate degradation products and the required characterization methodology. Our methods are effective in analyzing changes at specific amino acid residues, and the oxidation and deamidation harsh conditions did not interfere with the detection of the different set of variants. In addition, we detect that the observed modifications were specific to a set of forced degradation conditions. We are therefore confident that our designed methods could be applied in newly produced biosimilars and to monitor specific stability risks.

References

[1] Harry Keen, JC Pickup, RW Bilous, A Glynne, GC Viberti, RJ Jarrett, and R Marsden. Human insulin produced by recombinant dna technology: safety and hypoglycaemic potency in healthy men. *The Lancet*, 316(8191):398–401, 1980.

[2] Jeffrey Benjamin, Robert J Lilieholm, and David Damery. Challenges and opportunities for the north-eastern forest bioindustry. *Journal of Forestry*, 107(3):125–131, 2009.

[3] Tony Grift, Qin Zhang, Naoshi Kondo, and KC Ting. A review of automation and robotics for the bioindustry. *Journal of Biomechatronics Engineering*, 1(1):37–54, 2008.

[4] Kerry R. Love, Kartik A. Shah, Charles A. Whittaker, Jie Wu, M. Catherine Bartlett, Duanduan Ma, Rachel L. Leeson, Margaret Priest, Jonathan Borowsky, Sarah K. Young, and J. Christopher Love. Comparative genomics and transcriptomics of Pichia pastoris. *BMC Genomics*, 17:550, 2016.

[5] Robert A Copeland, David L Pompliano, and Thomas D Meek. Drug–target residence time and its implications for lead optimization. *Nature reviews Drug discovery*, 5(9):730, 2006.

[6] Peter J Tummino and Robert A Copeland. Residence time of receptor- ligand complexes and its effect on biological function. *Biochemistry*, 47(20):5481–5492, 2008.

[7] Wei Ouyang, Sung Hee Ko, Di Wu, Annie Yu Wang, Paul W Barone, William S Hancock, and Jongyoon Han. Microfluidic platform for assessment of therapeutic proteins using molecular charge modulation enhanced electrokinetic concentration assays. *Analytical Chemistry*, 88(19):9669–9677, 2016.

[8] David Garfinkel and John T Edsall. Raman spectra of amino acids and related compounds. x. the raman spectra of certain peptides and of lysozyme1-3. *Journal of the American Chemical Society*, 80(15):3818–3823, 1958.

[9] Vladimir P Drachev, Mark D Thoreson, Eldar N Khaliullin, V Jo Davisson, and Vladimir M Shalaev. Surface-enhanced raman difference between human insulin and insulin lispro detected with adaptive nanostructures. *The Journal of Physical Chemistry B*, 108(46):18046–18052, 2004.

[10] AG Brolo, Z Jiang, and DE Irish. The orientation of 2, 2-bipyridine adsorbed at a sers-active au (1 1 1) electrode surface. *Journal of Electroanalytical Chemistry*, 547(2):163–172, 2003.

[11] Franco Salomon, Ross C Cuneo, Richard Hesp, and Peter H Sönksen. The effects of treatment with recombinant human growth hormone on body composition and metabolism in adults with growth hormone deficiency. *New England Journal of Medicine*, 321(26):1797–1803, 1989.

[12] RC Cuneo, F Salomon, GA McGauley, and PH Sönksen. The growth hormone deficiency syndrome

in adults. *Clinical endocrinology*, 37(5):387–397, 1992.

[13] Judith C Juskevich, C Greg Guyer, et al. Bovine growth hormone: human food safety evaluation. *Science*, 249(4971):875–884, 1990.

[14] Shiaw Lin Wu, Haitao Jiang, William S. Hancock, and Barry L. Karger. 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. *Analytical Chemistry*, 82(12):5296–5303, 2010.

[15] MJ Pikal, Karen Dellerman, and ML Roy. Formulation and stability of freeze-dried proteins: effects of moisture and oxygen on the stability of freeze-dried formulations of human growth hormone. *Developments in biological standardization*, 74:21–37, 1991.

[16] Yuh-Fun Maa, Phuong-Anh T Nguyen, and Selena W Hsu. Spray-drying of air–liquid interface sensitive recombinant human growth hormone. *Journal of pharmaceutical sciences*, 87(2):152–159, 1998.

[17] F Franks. Freeze-drying: from empiricism to predictability. the significance of glass transitions. *Developments in biological standardization*, 74:9–18, 1991.

[18] James D Andya, Yuh-Fun Maa, Henry R Costantino, Phuong-Anh Nguyen, Nancy Dasovich, Theresa D Sweeney, Chung C Hsu, and Steven J Shire. The effect of formulation excipients on protein stability and aerosol performance of spray-dried powders of a recombinant humanized anti-ige monoclonal antibody1. *Pharmaceutical research*, 16(3):350–358, 1999.

[19] May Y Kwong and Reed J Harris. Identification of succinimide sites in proteins by n-terminal sequence analysis after alkaline hydroxylamine cleavage. *Protein Science*, 3(1):147–149, 1994.

[20] Rodney Pearlman and Y John Wang. *Formulation, characterization, and stability of protein drugs*, volume 9. Springer Science & Business Media, 1996.

[21] Jin Yin, Jhih-Wei Chu, Margaret Speed Ricci, David N Brems, Daniel IC Wang, and Bernhardt L Trout. Effects of excipients on the hydrogen peroxide-induced oxidation of methionine residues in granulocyte colony-stimulating factor. *Pharmaceutical research*, 22(1):141–147, 2005.

[22] Pablo Perez-Pinera, Ningren Han, Sara Cleto, Jicong Cao, Oliver Purcell, Kartik A Shah, Kevin Lee, Rajeev Ram, and Timothy K Lu. Synthetic biology and microbioreactor platforms for programmable production of biologics at the point-of-care. *Nature Communications*, 7, 2016.

[23] Gajendra P Singh, Shireen Goh, Michelangelo Canzoneri, and Rajeev J Ram. Raman spectroscopy of complex defined media: biopharmaceutical applications. *Journal of Raman Spectroscopy*, 46(6):545–550, 2015.

[24] L Chantalat, ND Jones, F Korber, J Navaza, and AG Pavlovsky. The crystalstructure of wild-type growth hormone at 2.5 aresolution. *Protein Pept Lett*, 2:333–340, 1995.

4.6 Supplementary Data

Tab 4.1: Raman spectroscopy peak assignments of hGH samples

Pixel Number	Raman Shift	Delta
70 to 120	515.8 to 576.2	2.3364
252	735.7	3.1951
345	848.0	5.8277
417	935.0	1.6181
476	1005.9	-3.2405
500-580	1032.9 to 1124.4	3.3718
653	1204.5	0.4824
745 to 780	1302.9 to 1340.3	-0.2884
878	1445.1	0.3125
1025-1040	1602.3 to 1618.3	-2.6589
1077	1657.9	-4.7180

Tab 4.2: Tryptic peptide mapping of hGH control and test samples

Peptide	Sequence	m/z	t_R (min)				
			Std	Test1	Test2	Test3	Test4
T1 (1-8)	FPTIPLSR	465.77^{2+}	30.12	30.41	31.37	30.82	30.72
T2(9-16)	LFDNAMLR	490.25^{2+}	27.93	28.64	29.51	28.95	28.72
T2-T3(9-19)	LFDNAMLR AHR	671.80^{2+}	20.26	20.62	21.62	21.11	20.96
T4(20-38)	LHQLAFDTYQEFEEAYIPK	781.38^{3+}	41.03	41.73	42.28	41.87	41.68
T4-T5(20-41)	LHQLAFDTYQEFEEAYIPK EQK	682.58^{4+}	38.91	39.18	39.72	39.19	39.13
T6-T7(42-70)	YSFLQNPQTSLCFSESIPT PSNREETQQK	854.91^{4+}	37.56	37.82	37.05	37.83	36.65
T8(71-77)	SNLELLR	422.74^{2+}	24.11	24.61	25.32	24.91	24.64
T9(78-94)	ISLLLIQSWLEPVQFLR	1028.10^{2+}	60.02	60.19	60.40	60.19	59.99
T10(95-115)	SVFANSLVYGASDSNVYD LLK	1132.05^{2+}	44.07	44.53	44.97	44.69	44.43
T11(116-127)	DLEEGIQTLMGR	681.34^{2+}	41.01	41.56	42.09	41.53	41.33
T12(128-134)	LEDGSPR	387.19^{2+}	10.90	11.16	11.42	9.95	10.79
T13-T14(135-145)	TGQIFKQTYSK	650.84^{2+}	17.25	17.08	17.61	17.59	17.41
T15(146-158)	FDTNSHNDDALLK	497.23^{3+}	20.26	20.62	21.62	21.11	20.96
T16-T17(159-168)	NYGLLYCFRK	667.33^{2+}	31.81	31.99	32.73	32.10	31.84
T17-19(168-178)	KDMDKVETFLR	691.36^{2+}	24.81	25.06	26.09	25.44	25.33
T20-T21(179-191)	IVQCRSVEGSCGF	749.84^{2+}	21.01	21.18	22.16	21.71	21.36
T21(184-191)	SVEGSCGF	842.33^{+}	18.77	18.91	18.75	19.19	19.01

Tab 4.3: GCSF peptide mapping

peptide	sequence	m/z	t_R (min) control	t_R (min) Test4
1-15	MTPLGPASSLPQSFL	773.39^{2+}	44.51	45.72
15-21	LLKCLEQ	423.74^{2+}	22.18	23.49
22-32	VRKIQGDGAAL	564.32^{2+}	13.23	14.81
17-42	KCLEQVRKIQGDGAALQEKLCATYKL	727.865^{4+}	44.73	44.74
39-47	TYKLCHPEE	560.26^{2+}	16.1	16.09
40-48	YKLCHPEEL	566.27^{2+}	23.2	24.35
49-69	VLLGHSLGIPWAPLSSCPSQA	1067.55^{2+}	44.64	45.59
70-76	LQLAGCL	717.39^{+}	34.66	35.56
77-84	SQLHSGLF	444.73^{2+}	25.4	25.56
80-84	HSGLF	560.27^{+}	14.87	15.64
84-90	FLYQGLL	853.47^{+}	41.51	42.54
91-104	QALEGISPELGPTL	1424.74^{+}	38.11	39.08
105-109	DTLQL	589.31^{+}	25.14	26.30
110-114	DVADF	566.24^{+}	20.06	21.40
115-122	ATTIWQQM	978.46^{+}	34.11	35.17
125-138	LGMAPALQPTQGAM	1385.67^{+}	31.58	32.66
139-144	PAFASA	563.27^{+}	10.87	12.98
145-153	FQRRAGGVL	502.29^{2+}	13.12	14.51
154-161	VASHLQSF	888.44^{+}	19.35	20.52
162-166	LEVSY	305.65^{2+}	22.81	23.95
167-175	RVLRHLAQP	545.33^{2+}	10.1	11.12

Tab 4.4: hGH oxidation result

	Met 14	std.v	Met 125	std .v	Met 170	std.v
Control	12.10%	0.045	2.90%	0.001	0.30%	0.0002
Test 1	36.30%	0.037	16.80%	0.026	0.40%	0.0022
Test 2	30.40%	0.039	24.80%	0.022	0.60%	0.0056
Test 3	63.70%	0.04	47.40%	0.054	0.60%	0.0018
Test 4	84.70%	0.031	96.20%	0.01	1.00%	0.0007

Tab 4.5: GCSF oxidation result

	Met1	std.v	Met122	std.v	Met127	std.v	Met138	std.v
Control	19.30%	0.033	0.80%	0.006	9.00%	0.014	23.40%	1.80%
Test 4	92.50%	0.02	33.60%	0.097	64.80%	0.022	87.20%	6.60%

Chapter 5

Proteomics analysis of the secretome of a cancer cell line (Tiam-1) regulated cell medium by HPLC-MS

Contribution: Northeastern University, Di Wu: LC-MS experiment procedures on cell medium, data analysis, William Hancock: goal of the study, concept contribution; Tufts Medical Center, Kun Xu: cell medium preparation, bioassay; Rachel Buchsbaum: concept contribution

5.1 Abstract

Breast cancer is a serious disease among the constellation of global health issues and the early detection is always an important research area for the correct disease diagnosis and effective treatment of these patients. In this study, we have characterized the secretome of four cell lines including non-aggressive and aggressive cancer cell lines using appropriate conditioned media to obtain information on potential biomarkers. We have applied HPLC tandem MS techniques and a proteomic strategy to study these cell lines. With a serial data analysis protocol using information listed in data bases of GeneCards VarElect, we have successfully curated a protein list derived from the proteomic study. The proteomic database was then evaluated with relevant bioactivity measurements, and as a promising lead we have found a significant difference in the expression of fibronectin relative to other gene products in the protein list. Future studies using RNA-sequence analysis, signal pathway analysis and bioactivity will be explored to confirm the disease associations of this biomarker.

5.2 Introduction

Cancer has become one of the most serious threats amongst all the diseases.[1, 2] In the past decades, knowledge, research, and health information focusing on cancer has accumulated tremendously, however, clinical detection of a patient status still has unknown issues.[3] There are various topics on cancer biomarkers, as Sawyers[4] has pointed out three types of biomarkers in his review which are prognostic, predictive and pharmacodynamic biomarkers. It's not easy to choose the correct associated cell line since the fist breast cancer cell line was established, BT-20 in 1958.[5] A commonly used breast cancer cell lines, the SUM series, are derived at the University of Michigan and licensed and distributed on to the market.[6] The SUM cell lines are known as aggressive cancer cell lines and have allowed scientists to learn many new aspects of breast cancer biology including early stage detection of the disease, biomarker identification, and drug discovery for treatment.[7]

Cancer spreading from the primary site to other parts of the body is termed metastasis and is often fatal.[8] The cancer cell can be spread to the whole body through the lymphatic system or the bloodstream when tumor cells break away from the primary tumor. There is urgent patient need for the discovery of cancer biomarkers for early detection of cancer and prognostic markers of those who may have risk of developing tumor metastases.[9] The metastasis associated cells will have unregulated activities which been partially characterized cancer detection studies.[10]

Mass spectrometry technology is important to proteomic research and a popular approach is to use a gel based digestion method followed by LC/MS analysis and protein database searching. The accuracy of measurement depends on the control of experimental conditions, which can be challenging, including resolution and sensitivity of the mass spectrometer, concentration of secretome sample from cell lines, contamination

of the preparation with components from the fermentation media, and so on.[11] The bottom-up methodology with enzyme digestion and a MS-based strategy is commonly used nowadays in biomarker discovery. In this way, controlled identification and quantitation are very important to the validation process. The protein variants that exist naturally in the cell is also a challenge for identification and quantitation studies.[12] To identify well-curated focused set of potential target biomarkers from a long proteomic list is challenging.[13] Data dependent acquisition (DDA), data independent acquisition (DIA) and selected reaction monitoring (SRM) are three main acquisition modes applied in the biomarker proteomic studies. We used the data dependent mode in our experiment and determined the peptides and proteins with confident MSMS measurements.

In this study we focused on the discovery of predictive biomarkers, whereas, in our previous studies, we have characterized several cell lines by the proteomic analysis to evaluate potential disease biomarker candidates and identified, differential gene mutations with resultant amino acid substitutions or different post-translational modifications, such as glycosylation than observed in the normal cell.[14, 15] In this study, we are comparing cancer aggressive cell lines to non-aggressive cell lines and will describe the differences between the cell lines and findings with disease associations. We assumed that the gene activities are different from the normal human cell line HMEC to pleural effusions MCF7 to the aggressive SUM cell line, to the primary breast cancer cell line SUM159 and to the mouse xenograft of the metastatic nodules cell line SUM1315.[16, 17] The ideal expectation is to discover the genes that have an unique existence in the cancer aggressive secretome as well as differences in gene product abundance in the aggressive cell lines. In this project, we will introduce the experimental design and characterization of the cell line by our mass spectrometry proteomic strategy and the comparative results will guide the resulting search of promising cancer biomarkers for the early detection of this disease.

5.3 Experimental section and methods

5.3.1 Cell medium and cell lysates

Cancer aggressive cell line secretome samples derived from the SUM159 and SUM1315 cell lines as well as the cancer non-aggressive cell lines HMEC (Human mammary epithelial cell) and MCF7 (Michigan Cancer Foundation-7) were prepared by our collaborators at the Tufts Medical Center (Boston, MA).

5.3.2 In-gel digestion

All four cell lines have been aliquotted to fifty microliters and kept in -80°C for further analysis. Twenty microliters of cells secretome sample were loaded onto a gel (SDS-PAGE, 4-12% gradient) to separate proteins by molecular weight. The gel was followed by the Coomassie blue staining process, and each gel lane was cut into five individual slices as shown in supporting information. The gel procedure has been duplicated, and the slices of same cell line with same position have been combined in order to get adequate concentration for a LC-MS validation.

Each slice was then minced into tiny pieces (approximately 1 mm^2), and transferred into a 1.5 mL micro-centrifuge tube. The gel slices were washed with 500 μL ACN and 0.1 M NH_4CO_3 for 45 min shaking and then centrifuged, the supernatant were removed. Proteins were analyzed under a reduced process by adding 500 μL of 10 mM dithiothreitol in 0.1 M NH_4HCO_3 and incubated for 30 min at 56°Cfollowed by centrifugation and removing supernatant before the alkylation process. Proteins were alkylated with 500 μL of 55 mM iodoacetamide in 0.1 M NH_4HCO_3 under room temperature and kept in dark for 60 min. All supernatant was removed after spinning gel pieces down and the following digestion process. Trypsin was

added in a protein solution and the concentration was based on an enzyme : protein ratio of 1:100 to 1:20 (w/w). The samples were covered by a trypsin buffer (200 μL 10 ng/mL of trypsin in 50 mM NH_4HCO_3, pH=8) and incubated 30 min at 4°Cto saturate gel pieces. Then the gel pieces were covered with the solution of 50 mM NH_4HCO_3 and incubated overnight at 37°C to yield peptides. Digestion process was stopped with 50 μL 5% formic acid and extracted with ACN. All supernatant were combined and stored for the following LC-MS analysis.

5.3.3 LC-MS/MS methods

The in-gel digested peptides were analyzed by LC coupled to a Quadrupole Exactive (QE) plus mass spectrometer (Q Exactive Plus, Thermo Electron, San Jose, CA) with a Dionex nano-LC instrument (Utimate 3000, Sunnyvale, CA) and a 4.6 mm $\times$ 150 mm packed with stable bond C-18 column (Zorbax 300 Å pore size, SB C18) (Agilent Technologiees, Santa Clara, CA). The analytical LC separation was carried out using a three step linear gradient, starting from 0.2% B to 40% B in 50 minute (A: water with 0.1% formic acid; B: ACN with 0.1% formic acid), increased to 80% in 5 minute, and kept for 5 minute. The flow rate was maintained at 200 nL/min. The following QE plus mass analysis was operated in the data-dependent mode to switch automatically between MS and MS/MS acquisition. The mass analyzer provide full-scan MS spectra with two microscans (m/z 300-2000) with a resolution of 70 000.

5.3.4 Protein identification

Thermo Proteome Discover 1.4 was using to identify peptide sequences, with database *Homo sapiens*, full trypsin specificity and up to three internal missed cleavages. The static modification was carbamidomethylation for cysteine, and dynamic modifications were for deamidation of asparagine and oxidation of methio-

nine residues. The tolerance has been set at 20 ppm for precursor ions and 1.0 Da for product ions. Peptides were identified with high confidence level and Xcorr scores above the following thresholds: $\geq$ 3.8 for 3^+ and for higher charge state ions, $\geq$ 2.2 for 2^+ ions, and $\geq$ 1.9 for 1^+ ions. Several housekeeping proteins, such as glyceraldehyder-3-phosphate dehydrogenase (GAPDH), and b-actin (ACTB) have been selected as internal standards for relative quantification to minimize variations in the amount of samples loaded on the gel, which provide consistent ratios among all four cell medium samples with the required criteria of high abundance.

5.4 Results and Discussion

5.4.1 Characterization of the secretome (cell line medium)

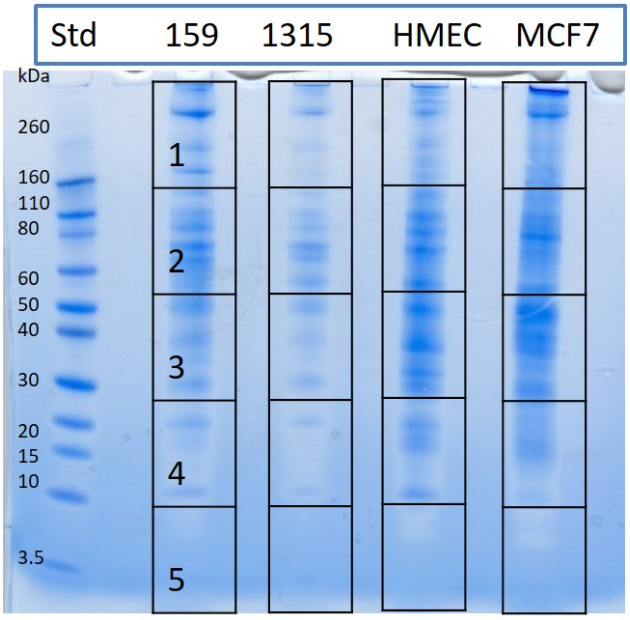

Fig 5.1: SDS-PAGE image of four secretome samples

The cell line secretome samples were analyzed by SDS-PAGE and followed by a gel based digestion. The gel image was shown as Fig.5.1, which indicated the concentration and protein expression of the cell culture samples. The 1D gel method is effective to separate the proteins secreted from the cell lines and help to remove the contaminations from the media components used to enable cell growth.

As shown, the cell line SUM1315 was observed with a relative low protein concentration. As marked in the gel image, the gel segments for each cell line were separated into five segments for enzyme based digestion and LC/MS analysis and the proteomic/genomic data from each segment were then combined a proteome of a whole cell line. The replicate observation of high abundance proteins was eliminated from the whole protein list and the final proteogenomic lists were then analyzed by bioinformatic tools. The cell medium gel based analysis was repeated and consistent observations from both analysis were marked with high confidence for further analysis.

The preliminary study was performed on a linear ion-trap mass spectrometer and then the optimized study was performed in duplicate with the SDS-PAGE procedure on a more powerful instrument the Quadrupole Exactive (QE) plus ion trap mass spectrometer. The optimized study gave a better result with higher confidence and larger scale identifications due to the improved sensitivity and resolution of the instrument. The observed gene products that were observed consistently with cancer associations are shown in Table5.1. For purposes of comparison the top 25 gene products from each cell line is shown in the supplementary data and it can be seen that some cancer related proteins such as fibronectin and thrombospondin-1 are also present at high levels.

Tab 5.1: Results of cancer-related proteins

Gene Name	Description	Novoseek Hit	SUM159	SUM1315	MCF7	HMEC
AHSG	Alpha-2-HS-glycoprotein	3^a	++	−+	−+	−+
C3	Complement C3	4^a	++	−+	−−	−+
CLEC3B	Tetranectin	17^a, 4^b	++	−−	−−	−−
FN1	Fibronectin	56^a, 6^c, 43^m	++	++	−+	−+
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	39^a	++	−+	−+	−+
HSPG2	Basement membrane-specific heparan sulfate proteoglycan core protein	14^a, 1^c	−+	++	−	−+
IGFBP4	Insulin-like growth factor-binding protein 4	11^a, 22^b, 15^c	−+	++	−+	−+
LDHA	L-lactate dehydrogenase A chain	15^a, 8^b	++	−+	−+	−+
LDHB	L-lactate dehydrogenase B chain	5^a, 4^b	++	−+	−+	−+
LTF	Lactotransferrin (Fragment)	3^c	++	−	−−	−+
NQO1	NAD(P)H dehydrogenase [quinone] 1	92^a, 4^m	+−	−−	−+	−−
THBS1	Thrombospondin-1	143^a,114^m	++	−+	−+	++
TIMP1	Metalloproteinase inhibitor 1	153^a,115^b, 28^c, 241^m	++	−+	−+	−+
TIMP2	Metalloproteinase inhibitor 2	81^a, 37^b,160^m	++	++	−+	−+
TPI1	Triosephosphate isomerase	3^a, 1^b,1^m	++	−+	−+	−+
VIM	Vimentin	87^m	++	−+	−+	−+

[a]Cancer

[b]Breast cancer

[c]Colon cancer

[d]Matastasis

The total amount of protein candidate numbers observed in both preliminary study and the optimized study were compared, the gene observed was shown "+" and the nonexistence was shown as "−". The number of total gene products with high confidence for each cell line in both studies are, in SUM159 are 101 and 167; in SUM1315 are 36 and 190; in HMEC are 155 and 268; in MCF7 are 207 and 471. Table5.1 provides the comparable cancer-related protein results from the four cell secretome samples derived from two batches. The protein names are the symbols used in UniProt and GeneCards, which are listed alphabetical order.

Biomarker proteomics does not allow a straightforward analysis for biomarker discovery as numerous proteins will be searched in a large data set, however, only a small set of proteins will be targeted for subsequent clinical studies. Novoseek Hits from GeneCards indicated biomedical literature numbers in which both the gene symbol and the protein product and disease associations appear. We have shown the relevance of the observed proteins with general associations to breast, cancer, colon cancer and metastatic disease in this table.

Tab 5.2: Gene data from search of VarElect GeneCards

Symbol	Description	GIFTs	Global Rank (of 10931)	Score
TIMP1	TIMP Metallopeptidase Inhibitor 1	55	210	44.97
TIMP2	TIMP Metallopeptidase Inhibitor 2	53	249	42.08
VIM	Vimentin	61	347	34.72
FN1	Fibronectin 1	61	428	30.69
THBS1	Thrombospondin 1	55	524	26.25
LTF	Lactotransferrin	53	671	21.27
LDHA	Lactate Dehydrogenase A	62	778	19.14
IGFBP4	Insulin Like Growth Factor Binding Protein 4	52	1122	14.56
GAPDH	Glyceraldehyde-3-Phosphate Dehydrogenase	58	1172	14.05
HSPG2	Heparan Sulfate Proteoglycan 2	57	1297	13.01
CLEC3B	C-Type Lectin Domain Family 3 Member B	49	1326	12.85
C3	Complement C3	59	1452	12.01
LDHB	Lactate Dehydrogenase B	57	2653	7.85
YWHAZ	14-3-3 protein Zeta	57	3052	7.04
YWHAB	14-3-3 protein Beta	58	4189	5.37
RPLP0	Ribosomal Protein Lateral Stalk Subunit P0	50	4693	4.79

The VarElect searching engine has functions to classify the identified gene products based on the gene data and the queried phenotypes. The scores are usually in a range of 1-200 to illustrate the strength of the

connection. Table.5.2 shows the input of our gene list together with the additional inquiry of breast cancer, and the ranking of the strength score. The GIFT number, with the range 1 to 100, indicated the GeneCards Inferred Functionality Score uses the wealth of gene annotations within GeneCards to produce the degree of knowledge about the functionality of more than 169k human genes entries.[18] The Global rank number has indicated the rank of a giving gene product out of the total 10931 genes related with breast cancer. In our result, the top five genes from the list are TIMP1 (TIMP Metallopeptidase Inhibitor 1), TIMP2 (TIMP Metallopeptidase Inhibitor 2), VIM (Vimentin), FN1 (Fibronectin 1) and THBS1 (Thrombospondin 1).

5.4.2 Pathway analysis of targeted proteins

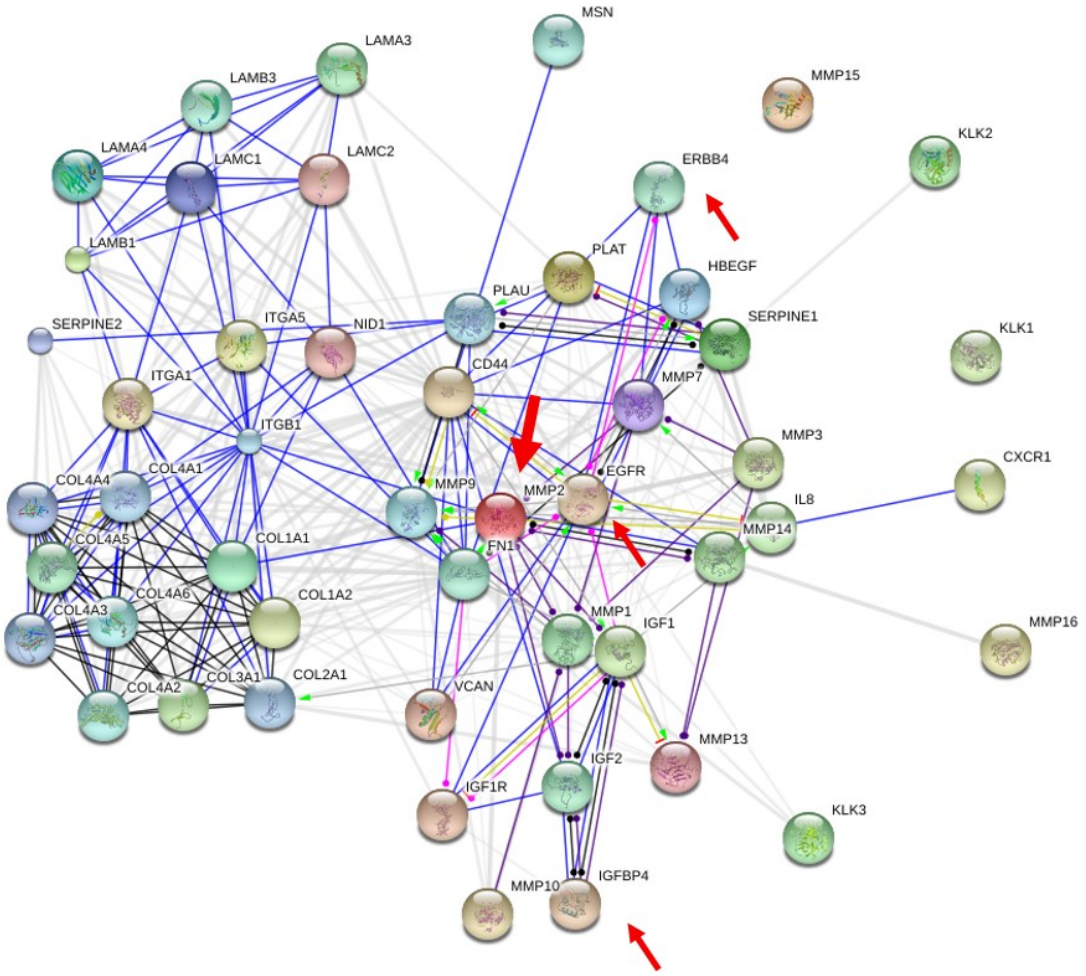

Fig 5.2: Pathway Cell adhesion ECM remodeling

The candidate protein list with 16 genes was analyzed by Gene A La Cart tool on the GeneCards analysis suite. The relative pathways were calculated and the best matched one was the cell adhesion ECM remodeling signal pathway. The FN1 gene was shown in red and in the center of the pathway picture. EGFR was shown in this pathway as well. As we know, EGFR is an important gene of breast cancer signaling as well as a growth factor receptor. From the data collected so far, fibronectin has attracted our interest as a biomarker

as it matched well our discovery criteria and will be the subject of a clinical study by our collaborator at Tufts Medical center.

5.5 Discussion

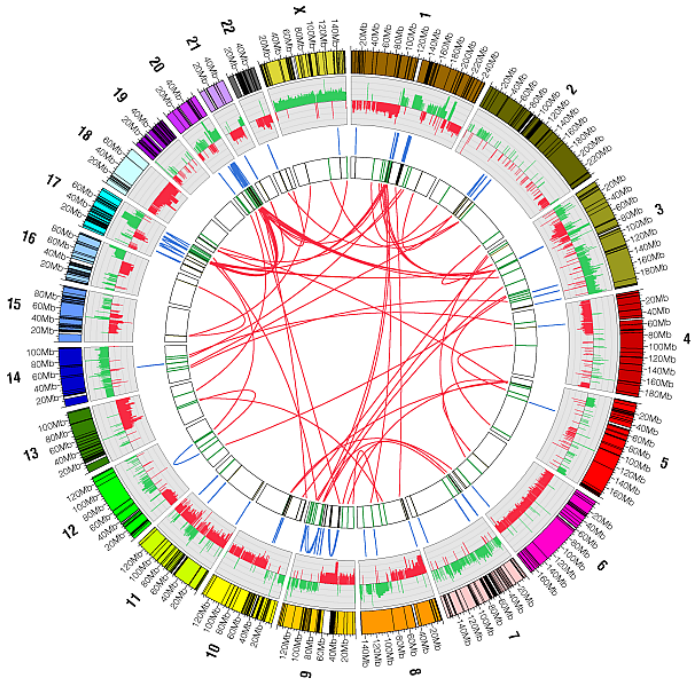

Fig 5.3: The circular visualization of MCF7 genome

Its a complicated process to discover and confirm a gene as a cancer biomarker, and then even more difficult to find a treatment for the resulting disease. The MCF7 cell line with estrogen, progesterone and glucocorticoid receptors is a human breast cancer cell line,[19, 20] which was derived from the pleural effusion in 1970s.[21] MCF7 was recognized as a fundamental cell line serving as reference in many breast cancer and genomic research studies, according to its ability to generate RNA/DNA leads to support downstream

validation studies. For example a unique property of this cell line is associated with the estrogen receptor in the cell cytoplasm which made MCF7 a useful model for hormone responsive breast cancer research.[22] As we show in Fig.5.3, the large set of data from these studies can be visualized by the software package Circos where the visual circle can be used to track genomic rearrangements and gene duplications.[23, 24] In this manner the relationship of cancer tumor cell lines with the cell line MCF7 can be significantly validated by a comparison of the Circos plots of the cell lines.

From our gene list, our collaborator and us found several interesting genes that are reported association with breast cancer, such as LDHA (L-lactate dehydrogenase A chain), LDHB (L-lactate dehydrogenase B chain),[25] CLEC3B (Tetranectin), IGFBP4 (Insulin-like growth factor binding protein 4),[26] TIMP1 (Metalloproteinase inhibitor 1), TIMP2 (Metalloproteinase inhibitor 2),[27, 28] THBS1 (Thrombospondin 1)[29] and FN1 (Fibronectin).[30] Based on our analysis and the curation by our collaborator of the potential biomarker list they have performed bioassays to validate promising gene candidates.

Our collaborator has selected FN1 and IGFBP4 from the list and monitored for activity associated with Tiam1 (T-Cell Lymphoma Invasion And Metastasis 1) regulation. The two genes were connected in the ECM (Extracellular matrix) pathway. The IGFBP4 had no effect on the fibroblast Tiam- osteopontin (OPN) pathway, however, with the increase of fibronectin secretion in aggressive breast cancer cell lines, our collaborators observed that exogenous fibronectin at physiological concentrations induces a decrease in Tiam1 expression and increased osteopontin levels in treated fibroblasts. As far back to 1980's, there are findings on fibronectin with breast cancer linkages,[31] where the scientists discovered the loss of fibronectin in cancer cells on comparing with normal cells and has cancer connections.[32] Fibronectin is found in the extracellular matrix of all cells as linear and branched networks that surround and connect neighbouring cells.[33] Fibronectin peptides can also mediate HMEC (Human Mammary Epithelial Cells) adhesion a to

porcine-derived extracellular matrix.[34]

Another promising lead is that tumor lactate levels are associated with increasing metastasis and tumor recurrence. Thus the LDH proteins serve as indicators of lactate metabolism genes and were therefore targeted for cancer research. The inhibitors of LDHA have been shown to suppress tumor progression in prostate cancer.[35] Although lactate was considered a byproduct from glycolysis, it has emerged as a critical regulator of cancer development and metastasis.

5.6 Conclusions

In this study, we have demonstrated that an initial proteomic analysis of the secretome (cell medium) characterized proteins of biological importance which can be correlated with breast cancer etiology. This study was based on advanced mass spectrometry technology, which gave a detailed view of the proteins secreted by the cell lines. The study of proteins that are related to metastasis correlated genes will yield insights into the biology of aggressive cancer cell lines. Subsequent studies can include the associated pathways, the metabolism of the gene products, RNA-sequence measurement to explore alternative splicing mechanisms, as well as other bioinformatic tools such as gene set enrichment and pathway analysis as well as targeted animal model studies using new genomic tools such as CRISPER. However, proteomic technology gives a unique view on the actual gene expression product with the additional complexity of post translational modifications that is produced and secreted by cancer cell and has a higher probability of generating a signal with a indicative prognosis of metastatic potential in an individual patient. With the development of the bioinformatics and biostatistics, the enriched genomic/proteomics data sets will be a promising approach to discover the cancer biomarkers for diagnosis and treatment in the future.

References

[1] Keith W Singletary and Susan M Gapstur. Alcohol and breast cancer: review of epidemiologic and experimental evidence and potential mechanisms. *Jama*, 286(17):2143–2151, 2001.

[2] Independent UK Panel on Breast Cancer Screening et al. The benefits and harms of breast cancer screening: an independent review. *The Lancet*, 380(9855):1778–1786, 2012.

[3] JENNIFER L Kelsey. A review of the epidemiology of human breast cancer. *Epidemiologic reviews*, 1:74–109, 1979.

[4] Charles L Sawyers. The cancer biomarker problem. *Nature*, 452(7187):548–552, 2008.

[5] Etienne Y Lasfargues and Luciano Ozzello. Cultivation of human breast carcinomas. *Journal of the National Cancer Institute*, 21(6):1131–1147, 1958.

[6] U-M Comprehensive Cancer Center. *SUM cell line*, 2006 (accessed February 3, 2018).

[7] Christine M Fillmore and Charlotte Kuperwasser. Human breast cancer cell lines contain stem-like cells that self-renew, give rise to phenotypically diverse progeny and survive chemotherapy. *Breast cancer research*, 10(2):R25, 2008.

[8] Gaorav P. Gupta and Joan Massagué. Cancer Metastasis: Building a Framework. *Cell*, 127(4):679–695, 2006.

[9] Britta Weigelt, Johannes L Peterse, and Laura J Van't Veer. Breast cancer metastasis: markers and models. *Nature reviews cancer*, 5(8):591, 2005.

[10] Saravana M Dhanasekaran, Terrence R Barrette, Debashis Ghosh, Rajal Shah, Sooryanarayana Varambally, Kotoku Kurachi, Kenneth J Pienta, Mark A Rubin, and Arul M Chinnaiyan. Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412(6849):822, 2001.

[11] Joshua E Elias, Wilhelm Haas, Brendan K Faherty, and Steven P Gygi. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature methods*, 2(9):667, 2005.

[12] Ludovic C Gillet, Alexander Leitner, and Ruedi Aebersold. Mass spectrometry applied to bottom-up proteomics: entering the high-throughput era for hypothesis testing. *Annual review of analytical chemistry*, 9:449–472, 2016.

[13] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, 4(3):207, 2007.

[14] Emma Yue Zhang, Massimo Cristofanilli, Fredika Robertson, James M Reuben, Zhaomei Mu,

Ronald C Beavis, Hogune Im, Michael Snyder, Matan Hofree, Trey Ideker, Gilbert S Omenn, Susan Fanayan, Seul-ki Jeong, Young-ki Paik, Anna Fan Zhang, Shiaw-lin Wu, and William S Hancock. Genome Wide Proteomics of ERBB2 and EGFR and Other Oncogenic Pathways in In fl ammatory Breast Cancer. 1, 2013.

[15] Shiaw-Lin Wu, Allen D Taylor, Qiaozhen Lu, Samir M Hanash, Hogune Im, Michael Snyder, and William S Hancock. Identification of potential glycan cancer markers with sialic acid attached to sialic acid and up-regulated fucosylated galactose structures in epidermal growth factor receptor secreted from A431 cell line. *Molecular & cellular proteomics : MCP*, 12(5):1239–49, 5 2013.

[16] Richard M Neve, Koei Chin, Jane Fridlyand, Jennifer Yeh, Frederick L Baehner, Tea Fevr, Laura Clark, Nora Bayani, Jean-Philippe Coppe, Frances Tong, et al. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer cell*, 10(6):515–527, 2006.

[17] Marc Lacroix and Guy Leclercq. Relevance of breast cancer cell lines as models for breast tumours: an update. *Breast cancer research and treatment*, 83(3):249–289, 2004.

[18] Arye Harel, Aron Inger, Gil Stelzer, Liora Strichman-Almashanu, Irina Dalah, Marilyn Safran, and Doron Lancet. Gifts: annotation landscape analysis with genecards. *BMC bioinformatics*, 10(1):348, 2009.

[19] Adrian V Lee, Steffi Oesterreich, and Nancy E Davidson. Mcf-7 cellschanging the course of breast cancer research and care for 45 years. *JNCI: Journal of the National Cancer Institute*, 107(7), 2015.

[20] Anait S Levenson and V Craig Jordan. Mcf-7: the first hormone-responsive breast cancer cell line. *Cancer research*, 57(15):3071–3078, 1997.

[21] KB Horwitz, ME Costlow, and WL McGuire. Mcf-7: a human breast cancer cell line with estrogen, androgen, progesterone, and glucocorticoid receptors. *Steroids*, 26(6):785–795, 1975.

[22] Deborah L Holliday and Valerie Speirs. Choosing the right cell line for breast cancer research. *Breast cancer research*, 13(4):215, 2011.

[23] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.

[24] Oliver A Hampton, Petra Den Hollander, Christopher A Miller, David A Delgado, Jian Li, Cristian Coarfa, Ronald A Harris, Stephen Richards, Steven E Scherer, Donna M Muzny, et al. A sequence-level map of chromosomal breakpoints in the mcf-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome research*, 19(2):167–177, 2009.

[25] Joanne R Doherty and John L Cleveland. Targeting lactate metabolism for cancer therapeutics. *The Journal of clinical investigation*, 123(9):3685–3692, 2013.

[26] Gabriela Dontu, Muhammad Al-Hajj, Wissam M Abdallah, Michael F Clarke, and Max S Wicha. Stem cells in normal breast development and breast cancer. *Cell proliferation*, 36(s1):59–72, 2003.

[27] Zheng-Sheng Wu, Qiang Wu, Jiu-Hua Yang, Hong-Qun Wang, Xiang-Dong Ding, Feng Yang, and Xiao-Chun Xu. Prognostic significance of mmp-9 and timp-1 serum and tissue expression in breast cancer. *International journal of cancer*, 122(9):2050–2056, 2008.

[28] DC Jinga, A Blidaru, Ileana Condrea, Carmen Ardeleanu, Cristina Dragomir, Geza Szegli, Maria Stefanescu, and Cristiana Matache. Mmp-9 and mmp-2 gelatinases and timp-1 and timp-2 inhibitors in breast cancer: correlations with prognostic factors. *Journal of cellular and molecular medicine*, 10(2):499–510, 2006.

[29] Cecilia Williams, K Edvardsson, SA Lewandowski, A Ström, and J-Å Gustafsson. A genome-wide study of the repressive effects of estrogen receptor beta on estrogen receptor alpha signaling in breast cancer cells. *Oncogene*, 27(7):1019, 2008.

[30] Jennifer J Choate and Deane F Mosher. Fibronectin concentration in plasma of patients with breast cancer, colon cancer, and acute leukemia. *Cancer*, 51(6):1142–1147, 1983.

[31] J Labat-Robert, P Birembaut, JJ Adnet, F Mercantini, and L Robert. Loss of fibronectin in human breast cancer. *Cell biology international reports*, 4(6):609–616, 1980.

[32] Heena Kumra and Dieter P Reinhardt. Fibronectin-targeted drug delivery in cancer. *Advanced drug delivery reviews*, 97:101–110, 2016.

[33] Purva Singh, Cara Carraher, and Jean E Schwarzbauer. Assembly of fibronectin extracellular matrix. *Annual review of cell and developmental biology*, 26:397–419, 2010.

[34] Jason Hodde, Rae Record, Robert Tullius, and Stephen Badylak. Fibronectin peptides mediate hmec adhesion to porcine-derived extracellular matrix. *Biomaterials*, 23(8):1841–1848, 2002.

[35] Zhi-Yong Xian, Jiu-Min Liu, Qing-Ke Chen, Han-Zhong Chen, Chu-Jin Ye, Jian Xue, Huan-Qing Yang, Jing-Lei Li, Xue-Feng Liu, and Su-Juan Kuang. Inhibition of ldha suppresses tumor progression in prostate cancer. *Tumor Biology*, 36(10):8093–8100, 2015.

5.7 Supplementary Data

Tab 5.3: Top 25 genes list from MCF7 cell line

UniProt	Gene	Description	Xcorr Score	Cov %	Uniq. Pep.	PSMs #	AAs #	MW (kDa)]
P02751	FN1	Fibronectin	1848.08	40.70	54	237	2386	262.5
P60709	ACTB	Actin, cytoplasmic 1	1597.51	56.53	8	181	375	41.7
P49327	FASN	Fatty acid synthase	1494.88	35.72	55	194	2511	273.3
P06733	ENO1	Alpha-enolase	1371.23	73.50	23	168	434	47.1
P08123	COL1A2	Collagen alpha-2(I) chain	1169.09	47.73	37	152	1366	129.2
P14618	PKM	Pyruvate kinase PKM	1109.73	49.53	18	127	531	57.9
P07900	HSP90AA1	Heat shock protein HSP 90-alpha	1068.54	38.93	12	122	732	84.6
P08238	HSP90AB1	Heat shock protein HSP 90-beta	980.94	41.57	14	103	724	83.2
P07996	THBS1	Thrombospondin-1	904.00	29.23	28	120	1170	129.3
P04792	HSPB1	Heat shock protein beta-1	823.12	76.10	10	94	205	22.8
P07437	TUBB	Tubulin beta chain	788.03	56.76	4	81	444	49.6
P04406	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	780.26	62.09	14	109	335	36.0
P68371	TUBB4B	Tubulin beta-4B chain	755.40	53.26	3	74	445	49.8
P02452	COL1A1	Collagen alpha-1(I) chain	727.51	31.83	27	85	1464	138.9
P21333	GLNA	Filamin-A	698.88	18.62	29	64	2647	280.6
P68363	TUBA1B	Tubulin alpha-1B chain	683.34	48.56	2	91	451	50.1
Q05639	EEF1A2	Elongation factor 1-alpha 2	643.67	30.02	3	65	463	50.4
P04075	ALDOA	Fructose-bisphosphate aldolase A	643.67	58.24	12	75	364	39.4
P68032	ACTC1	Actin, alpha cardiac muscle 1	633.81	24.93	2	51	377	42.0
Q71U36	TUBA1A	Tubulin alpha-1A chain	619.88	45.45	1	82	451	50.1
P68104	EEF1A1	Elongation factor 1-alpha 1	606.21	24.89	2	63	462	50.1
O43707	ACTN4	Alpha-actinin-4	562.44	36.33	14	68	911	104.8
P11142	HSPA8	Heat shock cognate 71kDa	524.62	26.63	12	52	646	70.9
P08727	KRT19	Keratin, type I cytoskeletal 19	493.97	54.50	17	65	400	44.1
P63104	YWHAZ	14-3-3 protein zeta/delta	483.23	62.04	12	67	245	27.7

Tab 5.4: Top 25 genes list from HMEC cell line

UniProt	Gene	Description	Xcorr. Score	Cov. (%)	Uniq. Pep.	PSMs #	AA #	MW (kDa)
P06733	ENO1	Alpha-enolase	1264.28	60.83	19	162	434	47.1
P04406	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	963.67	72.24	17	113	335	36.0
P14618	PKM	Pyruvate kinase PKM	646.79	49.15	17	74	531	57.9
B0YJC4	VIM	Vimentin	584.64	38.52	12	54	431	49.6
P07900	HS90AA1	Heat shock protein HSP 90-alpha	530.47	22.13	7	59	732	84.6
P08238	HSP90AB1	Heat shock protein HSP 90-beta	497.01	22.24	6	54	724	83.2
P11021	HSPA5	78 kDa glucose-regulated protein	431.55	33.79	16	60	654	72.3
P11142	HSP7C	Heat shock cognate 71kDa protein	404.81	28.33	12	43	646	70.9
P07996	THBS1	Thrombospondin-1	394.79	16.92	15	40	1170	129.3
P60174	TP11	Triosephosphate isomerase	390.09	56.64	10	56	286	30.8
P60709	ACTB	Actin, cytoplasmic 1	359.78	22.67	5	22	375	41.7
P04075	ALDOA	Fructose-bisphosphate aldolase A	356.23	41.76	7	37	364	39.4
P00558	PGK1	Phosphoglycerate kinase 1	348.64	33.33	8	25	417	44.6
P00338	LDHA	L-lactate dehydrogenase A chain	317.29	44.58	8	42	332	36.7
P09486	SPARC	SPARC	309.40	23.10	5	19	303	34.6
O00391	QSOX1	Sulfhydryl oxidase 1	270.10	19.01	10	34	747	82.5
Q08380	LGALS3BP	Galectin-3-binding protein	258.42	27.52	11	36	585	65.3
P04083	ANXA1	Annexin A1	250.66	39.88	10	37	346	38.7
Q5H9A7	TIMP1	Metalloproteinase inhibitor 1	246.98	72.03	6	37	143	16.0
O14773	TPP1	Tripeptidyl-peptidase 1	236.76	3.73	1	29	563	61.2
P62258	YWHAE	14-3-3 protein epsilon	227.81	36.47	6	34	255	29.2
Q15582	TGFBI	Transforming growth factor-beta-induced protein ig-h3	218.58	13.32	5	11	683	74.6
P02751	FN1	Fibronectin	214.57	6.71	9	17	2386	262.5
O43707	ACTN4	Alpha-actinin-4	212.07	11.31	8	23	911	104.8
P12109	COL6A1	Collagen alpha-1(VI) chain	210.69	10.21	7	17	1028	108.5

Tab 5.5: Top 25 genes list from SUM159 cell line

UniProt	Gene	Description	Xcorr Score	Cov. %	Uniq Pep #	PSMs #	AAs #	MW (kDa)
P02751	FN1	Fibronectin	2625.83	45.31	62	309	2386	262.5
Q15582	TGDBI	Transforming growth factor-beta-induced protein ig-h3	2018.70	65.74	29	256	683	74.6
P07996	THBS1	Thrombospondin-1	1031.96	32.82	28	119	1170	129.3
P08253	MMP2	72 kDa type IV collagenase	1007.69	43.79	20	109	660	73.8
P01024	C3	Complement C3	797.74	24.11	31	94	1663	187.0
P23142	FBLN1	Fibulin-1	727.83	34.71	5	85	703	77.2
P08238	HSP90AB1	Heat shock protein HSP 90-beta	504.28	26.80	7	54	724	83.2
Q08380	LGALS3BP	Galectin-3-binding protein	493.91	35.73	13	71	585	65.3
Q12805	EFEMP	EGF-containing fibulin-like extracellular matrix protein 1	488.38	39.76	12	70	493	54.6
P06733	ENO1	Alpha-enolase	408.06	47.24	12	45	434	47.1
O14773	TPP1	Tripeptidyl-peptidase 1	376.96	3.73	1	58	563	61.2
P07900	HSP90AA1	Heat shock protein HSP 90-alpha	369.52	16.67	4	41	732	84.6
O00468	AGRN	Agrin	340.03	8.71	11	29	2067	217.1
P29401	TKT	Transketolase	308.91	20.55	7	30	623	67.8
O94985	CLSTN1	Calsyntenin-1	268.43	15.60	7	25	981	109.7
Q5H9A7	TIMP1	Metalloproteinase inhibitor 1	257.87	72.03	6	30	143	16.0
P04264	KRT1	Keratin, type II cytoskeletal 1	241.13	20.65	9	42	644	66.0
B0YJC4	VIM	Vimetin	239.47	14.85	6	13	431	49.6
P05121	SERPINE1	Plasminogen activator inhibitor 1	234.65	20.90	5	22	402	45.0
P09486	SPARC	SPARC	222.45	13.53	2	6	303	34.6
O15230	LAMA5	Laminin subunit alpha-5	211.81	3.82	9	13	3695	399.5
P11142	HSPA8 PE	Heat shock cognate 71 kDa protein	190.72	19.97	10	17	646	70.9
Q15113	PCOLCE	Procollagen C-endopeptidase enhancer 1	189.48	17.15	5	18	449	47.9
O43707	ACTN4	Alpha-actinin-4	185.20	7.79	6	28	911	104.8
P60709	ACTB	Actin, cytoplasmic 1	155.54	15.20	3	9	375	41.7

Tab 5.6: Top 25 genes list from SUM1315 cell line

UniProt	Gene	Description	Xcorr. Score	Cov %	Uniq. Pep.	PSMs #	AAs #	MW (kDa)
P07996	THBS1	Thrombospondin-1	385.57	12.99	12	45	1170	129.3
O14773	TPP1	Tripeptidyl-peptidase 1	315.32	3.73	1	44	563	61.2
P08253	MMP2	72 kDa type IV collagenase	272.62	19.70	8	20	660	73.8
P23142	FBLN1	Fibulin-1	209.65	7.68	1	5	703	77.2
P05121	SERPINE1	Plasminogen activator inhibitor 1	113.22	8.21	2	10	402	45.0
Q15582	TGFBI	Transforming growth factor-beta-induced protein ig-h3	90.33	5.71	3	7	683	74.6
P02751	FN1	Fibronectin	90.24	1.26	2	7	2386	262.5
Q5H9A7	TIMP1	Metalloproteinase inhibitor 1	87.51	29.37	3	9	143	16.0
P06733	ENO1	Alpha-enolase	71.59	17.28	4	7	434	47.1
Q08380	LGALS3BP	Galectin-3 binding protein	70.69	2.22	1	2	585	65.3
P63104	YWHAZ	14-3-3 protein zeta/delta	64.19	11.84	1	11	245	27.7
P10909	CLU	Clusterin	56.70	12.47	4	6	449	52.5
P07602	PSAP	Prosaposin	54.05	8.40	2	7	524	58.1
P08238	HSP90AB1	Heat shock protein HSP 90-beta	52.12	5.11	3	8	724	83.2
C9JKR2	ALB	Albumin, isoform CRA_k	51.26	3.60	1	3	417	47.3
P61769	B2M	Beta-2-microglobulin	50.32	18.49	1	4	119	13.7
Q12805	EFEMP1	EGF-containing fibulin-like extracellular matrix protein 1	49.36	5.88	1	7	493	54.6
O00468	AGRN	Agrin	46.66	1.69	3	6	2067	217.1
P04406	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	34.81	8.66	2	4	335	36.0
P60709	ACTB	Actin, cytoplasmic 1	33.98	4.27	1	2	375	41.7
P12109	COL6A1	Collagen alpha-1(VI) chain	33.70	0.88	1	2	1028	108.5
Q9UBP4	DKK3	Dickkopf-related protein 3	29.53	4.86	1	2	350	38.4
O00391	QSOX1	Sulfhydryl oxidase 1	28.31	3.08	1	2	747	82.5
Q5SR54	CLSTN1	Calsyntenin-1	25.32	1.66	1	2	782	88.0
P16035	TIMP2	Metalloproteinase inhibitor 2	24.50	6.82	1	3	220	24.4

Conclusions

Global health is always a topic for natural, health and social scientists to explore due to the rapid development of medical and chemical technologies. Recombinant protein production and manufacture is a popular topic with many researches. The requirement of low cost manufacturing is a huge issue that needs efficient techniques with high quality. In remote areas or emergency situations, to support the patient in their need and avoid fatal issues is very important. We have demonstrated the capability and consistency of *Pichia* based production and the reproducibility of InSCyT platform. This bench top system is capable to generate multiple biopharmaceutics with continuous running. Comparing with the quality level of FDA approved biotherapeutics, the InSCyT products have been produced in sufficient amount, full sequence coverage, and low levels of host cell proteins. Although modifications such as oxidation were generated in the fermentation process, the quantity of degradation products and process impurities was within the required safety range and not clinically relevant. We have developed protocols for the stability studies that generate degradation products as well as the required characterization methodology. Our methods are effective in analyzing changes at specific amino acid residues, and the harsh conditions developed to study oxidation and deamidation did not interfere with the detection of different sets of variants. In addition, we detected that the modifications were specific to a set of forced degradation conditions. We are therefore confident that our designed methods could be applied in newly produced biosimilars and to monitor specific stability

risks. During a fermentation run, the observed modification for each sample will indeed be continuously accumulated from beginning to the end, which suggests that the manufacturing process should be monitored continuously for product impurities. For other properties of the drug such as toxicity, biological function or in vivo stability, further studies will need to be performed.

In summary, the methodology we developed is effective to evaluate the sequence structures and monitor any degradation reactions. Different mass spectrometers were applied and compared, which will guide further analysis to select the best instrument based on the properties of the protein product . The crude protein drug produced from InSCyT system has been analyzed in this research and consistent data was demonstrated. Bottom-up LC-MS techniques may require more offline time for sample preparation and analysis, however, by this way, there will be more detailed information on the exact structural changes in the biologic product, down to the level of amino acids. On the other hand, the other techniques used in this study such as Raman spectrometry and electrokinetic concentration binding assays will differentiate the modifications directly on intact proteins in a short time period. These technologies are very efficient for a quick product evaluation or diagnosis of product quality.

For the cancer proteomic study, we performed a deep study on the genes and the corresponding proteins associated with metastasis and well as a comparison of aggressive vs non aggressive cancer cell lines. These results will help direct subsequent studies that can include associated pathways, the metabolism of the genes, RNA-sequence measurements, as well as other bioinformatic tools such as in depth gene pathway analysis and targeted animal model studies using new genomic tools such as CRISPER. However, proteomic technology gives a unique view on the actual expression product that is produced and secreted by cancer cell and has a higher probability of prognostic signals of metastatic potential in an individual patient. With the further development of the bioinformatics and biostatistics tools, the genomic data will be further explored

to discover informative cancer biomarkers for the diagnosis as well as treatment in the future.