

Automatic annotation of N-glycans in MALDI-TOF
spectra for rapid glycan profiling and comparison

Chuan-Yih, Yu

May 2010

Abstract

Being one of the most common protein post-translational modifications occurring in humans, glycosylation plays a crucial role in the onset of various diseases such as cancer. Mass spectrometry is often used to acquire glycan profile data in order to provide a quantitative assessment of variations in glycan abundance between cancer and healthy patients, with the aim of identifying biomarkers of the disease. In this project, we develop a fully automatic glycan annotation and comparison software, which allows users to identify possible glycan biomarkers accurately. Initially, we identify potential glycans from the profile data using spectral searching based on a dedicated list of glycan compositions. The method then identifies subsets of glycans that are significant in the fact that they show similar patterns within a group (either disease or health) but exhibit large variance across groups. We illustrate the efficacy of the method using previously studied data for discovering biomarkers in hepatocellular carcinoma using N-glycan serum markers. We were able to identify the major biomarkers and several new glycans that could be potential glycan biomarkers for Hepatocellular carcinoma.

Contents

1	Introduction	2
1.1	Post-translational modification(PTM)	2
1.2	Glycosylation	2
1.3	Glycans analyzing strategies	5
1.3.1	Microarray	5
1.3.2	Mass Spectrometry	5
2	Methodology	8
2.1	Annotation of N-glycan	8
2.2	Glycan profile comparison	10
3	Implementation	12
4	Result	14
4.1	Annotation of N-glycan	14
4.2	Glycan profile comparison	14
5	Conclusion & Future Direction	17
6	Acknowledge	18

Chapter 1

Introduction

First, I will briefly introduce some background knowledge, which is related to my project.

1.1 Post-translational modification(PTM)

Post-translational modification is an enzyme-catalyzed protein modification after protein being synthesized. The figure 1.1 shows the process of the protein synthesis and post-translational modifications. affect. It can either enable or disable the biological function of modified proteins. There are different kinds of PTMs such as Acetylation, Glycosylation, Methylation, Phosphorylation, Prenylation and, etc. The previous study [2] reports that there are more than 50% of all eukaryotic proteins are glycosylated. Therefore, we can say that the glycosylation is one of the common post-translational modifications within humans. We cannot fully understand the biological system without studying the PTM.

1.2 Glycosylation

The glycosylation is an attachment of a glycan to the peptide chain. There are two types of glycosylations, O-linked glycosylation and N-linked glycosylation. Glycosylation also has been verified to relate to some diseases [3][4][5]. Thus, the glycan profiling and glycoprotein profiling could be a way to discover biomarkers [6][7][8].

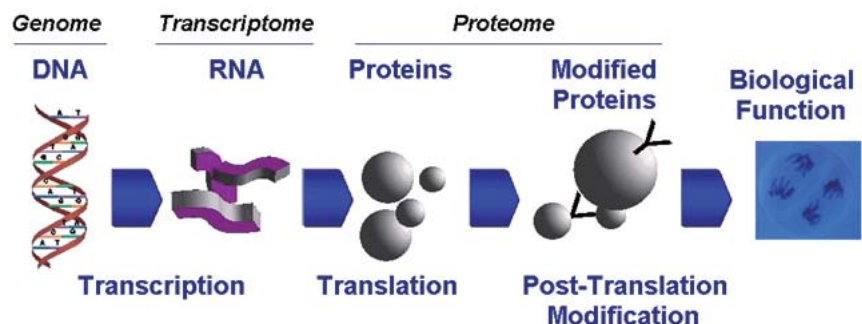


Figure 1.1: Overview of protein synthesis and post-translational modifications

[1]

The O-linked glycans are covalently linked to the oxygen atom on the Thr(T) or Ser(S) within a peptide chain. The O-linked glycosylation has various structures, which are different from N-linked glycosylation. We only illustrate some of them in figure 1.2. The analysis complexity of O-linked glycosylation is higher than N-linked glycosylation due to their diverse core structures. The O-linked glycosylation happens after protein folding.

The N-linked glycans are covalently linked to the nitrogen atom on the Asn (N) on the peptide. It recognizes certain peptide patterns, which are Asn-X-Ser(NXS) and Asn-X-Thr(NXT), X can be any amino acid but Pro. These peptide sequences are called glycosylation sequon. The N-linked glycosylation has very conserved core structure, which contains two GlcNac and three Mannose. The figure 1.2 shows the N-linked glycosylation core structure through common nomenclature [9]. This type of glycosylation happens while protein folding. It means without correct modification the protein may not have corrected folding, which can affect the protein function directly.

The N-linked glycan can be viewed as the tree structure. We can regard the monosaccharides, which is the building block of the glycan as nodes, and the linkage which has up to four out degree link as branches. The figure 1.3 shows three distinct N-linked glycan structures, Oligomannose, Complex and Hybrid.

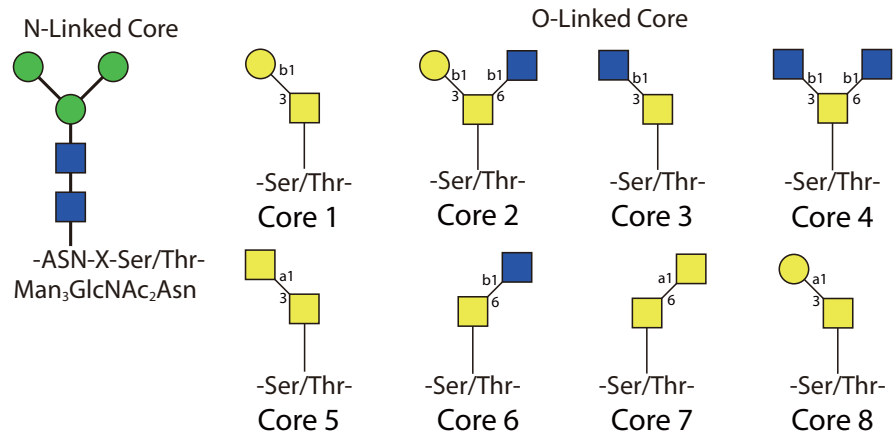


Figure 1.2: Core structure of glycosylation

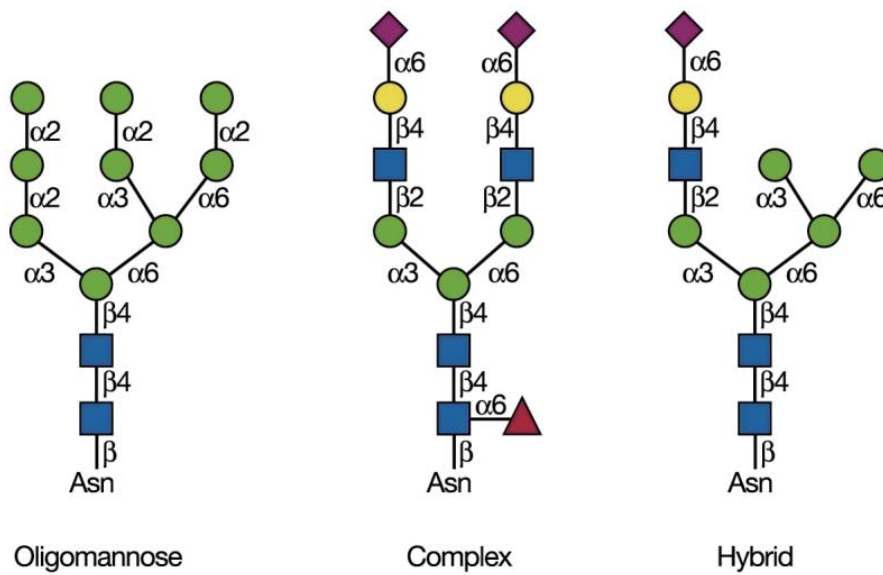


Figure 1.3: Three types of N-linked glycans

1.3 Glycans analyzing strategies

Both instrument and experimental technique have had major improvement for past decades. This helps scientists gather better data with various perspectives by different techniques. We can understand more in the whole biological system by integrating complementary datasets [10][11]. The figure 1.4 shows the common technique used in glycomics.

1.3.1 Microarray

Because the microarray is not the main focus in my study, I will only briefly describe it. The microarray is one of the commonly used tools to analyze the protein interaction. We also can use microarray to study the glycoproteomics and glycomics. There are three different types of microarraies, antibody microarray, glycan microarray and lectin microarray. First, the antibody microarray uses antibodies as probes to identify which proteins are glycosylated by certain glycan. Second, the glycan microarray contains distinct glycans as probes to interact with glycan binding protein (GBP). Finally, the lectin microarray prints various lectin on the microarray and uses lectin to bind with different glycan.

1.3.2 Mass Spectrometry

Mass spectrometry has high throughput and high sensitivity characteristics. It is widely used not only in proteomics but also glycoproteomics and glycomics. We can use mass spectrometry to analyze glycopeptides and glycans.

MALDI-TOF and MALDI-FTICR are two different mass spectrometry, which allow us to gather information from the sample for glycan profiling. The glycan profiling is to know what kind of glycan (annotation) and how abundance (quantification) it is within the sample (figure 1.5).

The isotope phenomenon happens in the nature due to some chemical elements have different number of neutron, and we will always observe series of peaks, which represent the same compound in mass spectra. So, if the mass of these two compounds is closed,

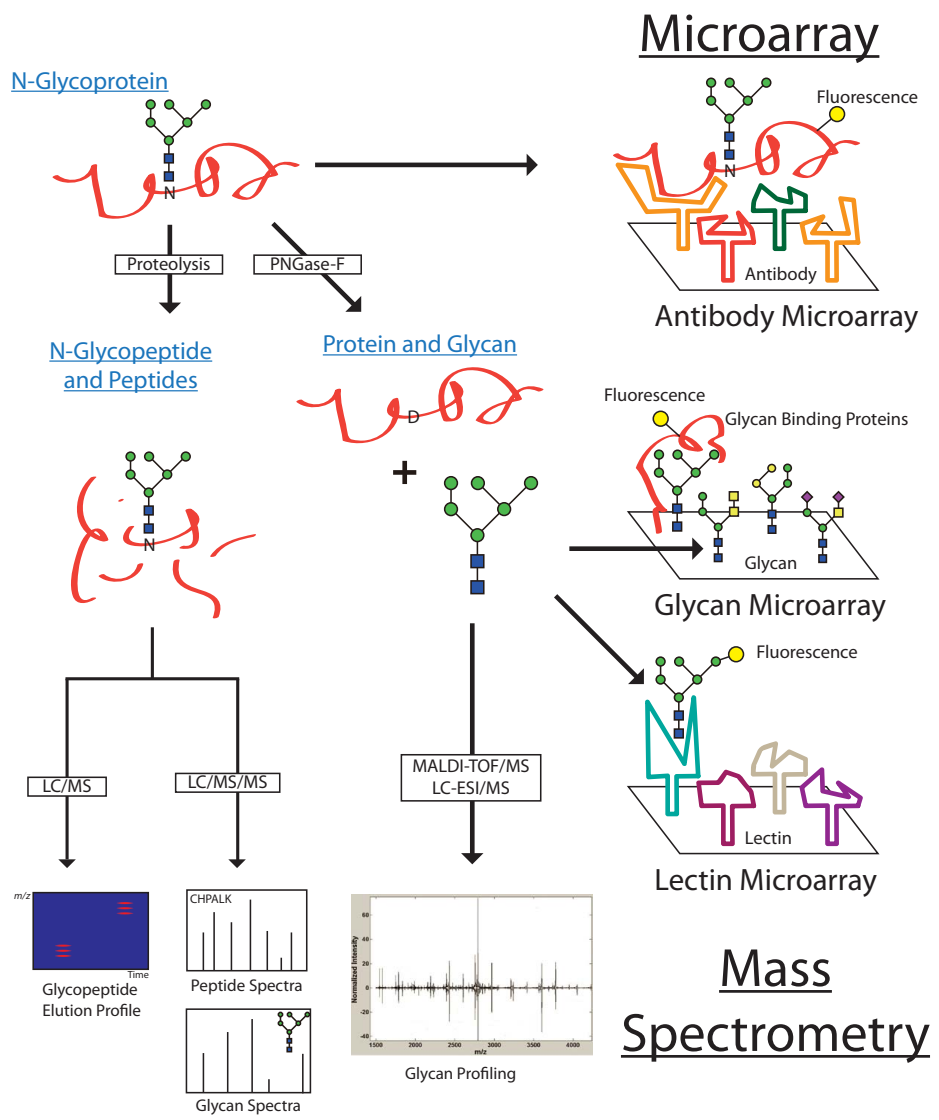


Figure 1.4: Overview of analytical strategies for analyzing glycans and glycopeptides

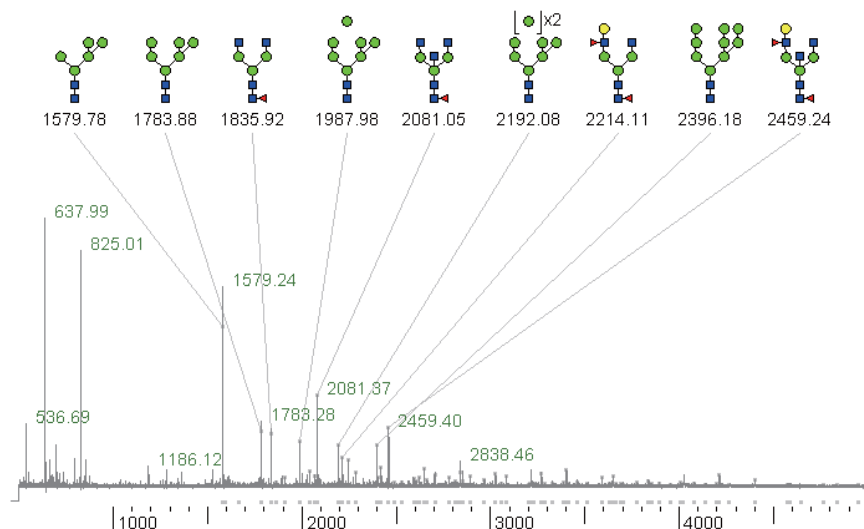


Figure 1.5: Example of N-glycan profiling

two isotope envelopes will overlap with each other. As we know the glycans can have various combinations, and sometime two glycans will have very close mass. Therefore, we will observe an overlapping isotope envelope. These two reasons could affect the accuracy of glycan profiling. Table 1.1 demonstrates a sample of this problem.

Table 1.1: Example of overlapping glycans

2 GlcNac + 9 Man = 2374.5960		7 GlcNac + 3 Man = 2375.63	
Mass	%	Mass	%
2371	0	-	-
2372	84.3	2372	0.0
2373	100.0	2373	82.4
2374	68.5	2374	100.0
2375	34.3	2375	68.8
2376	13.9	2376	34.4

Chapter 2

Methodology

2.1 Annotation of N-glycan

Couple years ago, Krambeck and colleagues derived common monosaccharide sequences of N-glycans based on the N-glycan synthetic pathways in human cells [12]. We use this model to generate a glycan list containing 412 different combinations. This list is used as default glycan combination list in our program. We test input spectra against this glycan list. In previous section, we talked about the overlapping isotope envelope problem that will happen when we want to annotate the spectra which has some compounds mixed up together. Therefore, we design three scenarios to deal with this kind of problem (figure 2.1). The first one contains only glycan; the second one contains glycan mixed with another glycan, which has molecular weight close to each other; the third one contains glycan mixed with unknown compound, which might be contamination or other molecular. We generate the non-overlapping and overlapping theoretical isotope envelope and apply linear fitting according to these scenarios. For unknown compound, we use Mercury algorithm to generate the unknown isotope envelopes [13]. Each potential isotope envelope will assign the correlation scores with three theoretical isotope envelopes (equation 2.1) and decompose the relative abundance.

$$Score = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (2.1)$$

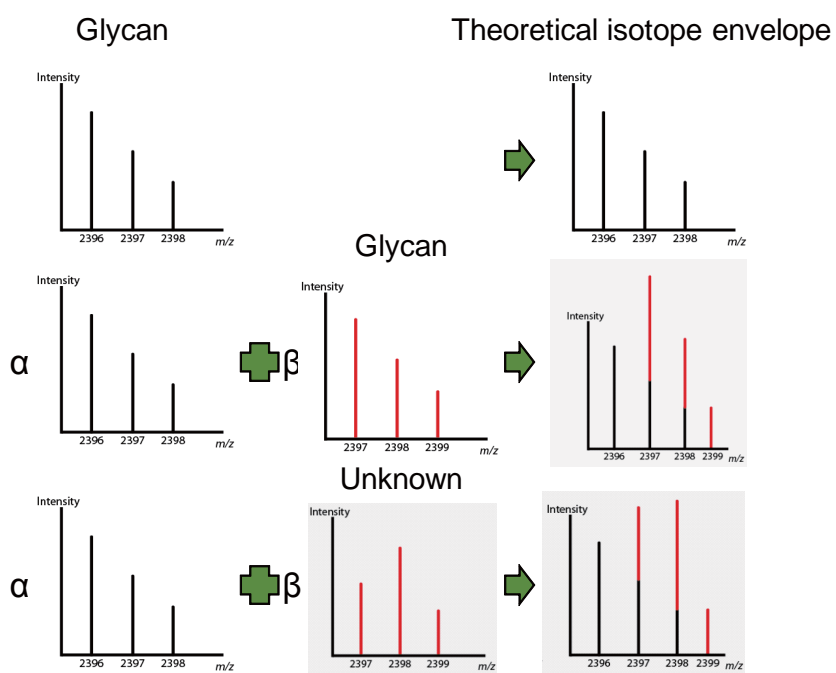


Figure 2.1: Three scenarios

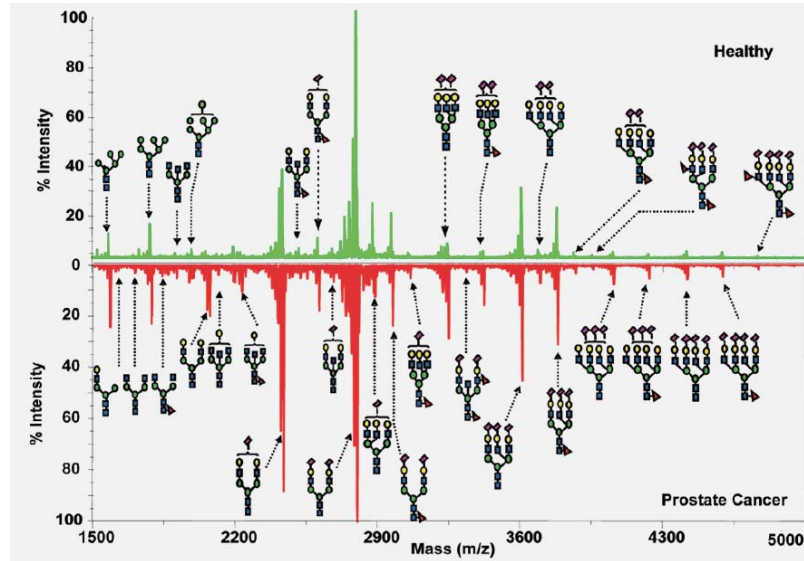


Figure 2.2: Mass spectra for cancer and health data with annotated glycans. [14]

n = Number of peaks in the series

2.2 Glycan profile comparison

In previous research [14], the glycomic profile is used to discover the potential biomarker by using principal component analysis(PCA). The figure 2.2 shows an example of mass spectra for cancer and healthy data with annotated glycans. A computational method is developed for identifying potential biomarkers for hepatocellular carcinoma(HCC) and chronic liver disease(CLD)[15][16]. They build Support Vector Models(SVMs) to isolate importance spectra and to identify glycans that show considerable change among HCC, CLD and health. Because SVMs approach becomes tedious with complicated glycan peak-picking algorithm, we use another approach with less sophisticated.

This approach works as following 2.3. We will annotate all the spectrum within two groups, then filter out glycans have low cor-

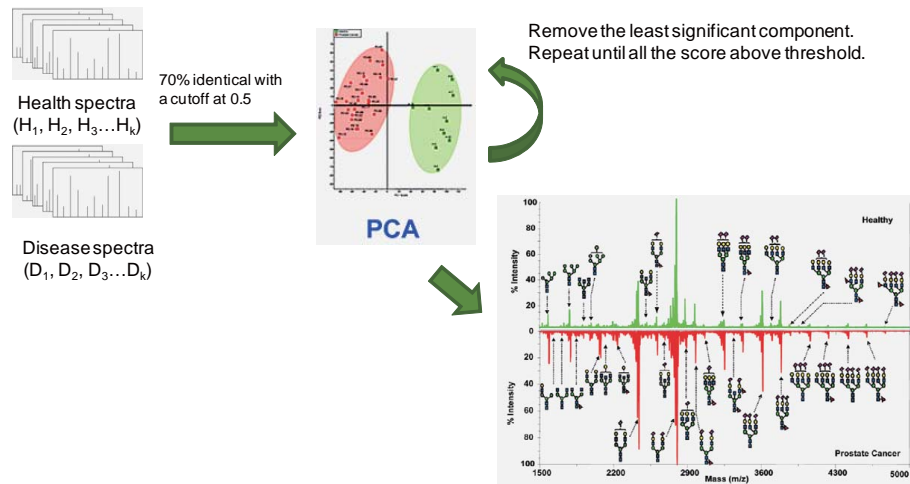


Figure 2.3: The workflow of glycan profile comparison

relation score and less than 30% of the spectrum within the same group. Afterward, PCA is performed and discarded the least significant component; repeating this step until all the score above the threshold.

Chapter 3

Implementation

To achieve the goals of glycan annotation and profile comparison, we present Multi N-Glycan. A software can annotate different N-glycans by examining the mass spectra automatically. This program is written by C# and developed under Microsoft .NET framework 2.0 environment. It supports various spectrum formats such as plain text, mzXML [17] and RAW file. Users can provide their own glycan list in CSV format. The output can be either in html or CSV format. The html output contains not only the glycan score and abundance but also the graphic isotope envelope, but CSV output only export the score and abundance. The Multi N-Glycan has flexibility to incorporate with the different experimental procedures, and also provides a user friendly interface.

Software Link:

<http://mendel.informatics.indiana.edu/~chuyu/MultiNGlycan/>

- Software Requirements
 - .NET framework 2.0
 - C++ runtime (Mercury algorithm)
 - [R] for PCA analysis
 - Thermo Scientific Xcalibur
- Input
 - Spectrum: Plain text (Peak list), mzXML [17], or RAW file (Thermo Scientific raw file)

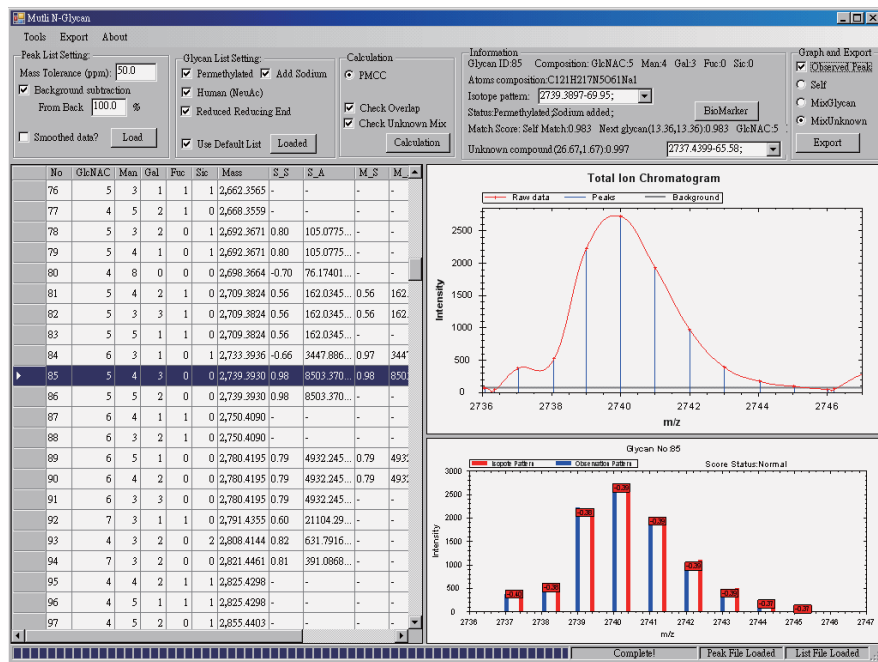


Figure 3.1: Software interface

- N-Glycans list: CSV file (User-defined); default list contains 412 common glycans [12]
- Output: List of glycans with scores and graphs.

Chapter 4

Result

4.1 Annotation of N-glycan

In a MALDI-TOF experiment, first we only can identify 86 N-glycans in manual inspection, but Multi N-Glycan can identify an extra 34 N-glycans. Among these 120 N-glycans, 40 N-glycans received scores above 0.7; 16 N-glycans were detected to have the isotopic envelope overlapping with unknown molecule; and two N-glycans were detected to have overlapping isotopic envelopes with other glycans 4.1.

4.2 Glycan profile comparison

We take two groups, health and hepatocellular carcinoma(HCC), of dataset from a published paper [16]. Health group has 78 individuals and HCC group has 73 individuals respectively. The filter criteria are set as following: glycan correlation score < 0.5 and glycan not present at least 30% spectrum in the same group will be discarded. In the figure 4.2, the left side is the top 10 distinct glycan across the health and the HCC group result generating by Multi N-Glycan; the right side is the partial result from the original paper [16] using SVMs method. In the figure, we can find out that three out of ten N-glycans have the similar trend (m/z 1580, 2192 and 2850). There is a very interesting case, we should take a close look. In the SVMs method, they identify 2187 instead of 2192, but in Multi N-Glycan reports that there is an overlapping isotope envelope in 2187 and 2192. This case represents a good

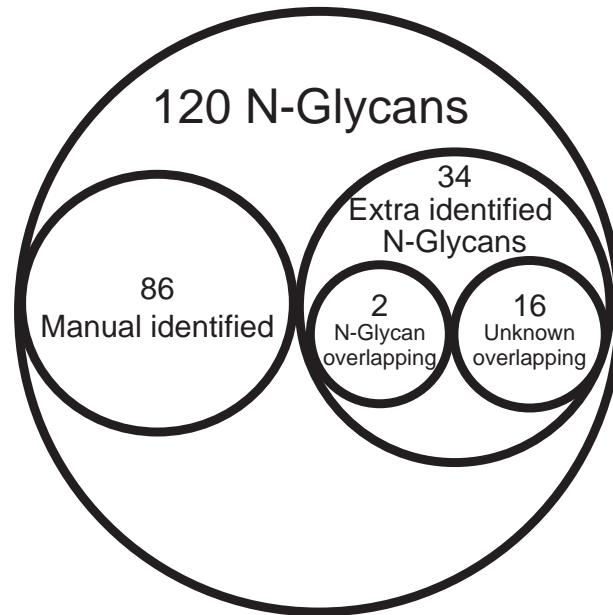


Figure 4.1: The result of Glycan annotation

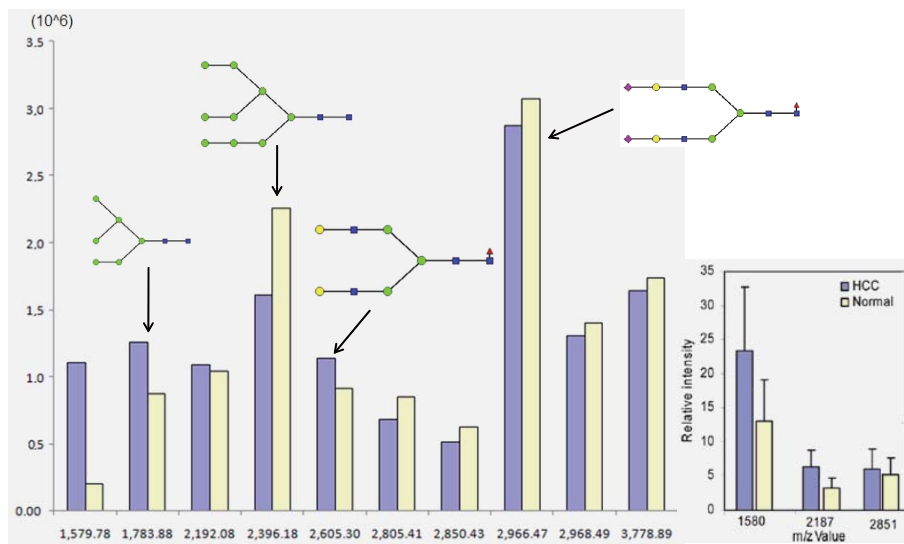


Figure 4.2: The result of Glycan profile comparison

example of two glycan overlapping with each other, which cannot be identified by other tools but can be found in my program. We take our result glycans to search against CFG database. Four out of seven in the total 10 N-glycans are found in the database (m/z 1784, 2396, 2605 and 2966). These glycans m/z 1784, 2396, 2605 and 2966) might be considered to be potential biomarkers for further investigation.

Chapter 5

Conclusion & Future Direction

We only use a simple computational method to annotate and identify potential glycan biomarkers using mass spectra acquired from a MALDI-TOF platform. The method was applied to data from previous research and validated glycan biomarkers for hepatocellular carcinoma. Three glycans are confirmed both in the previous research and Multi N-glycan; there are some new glycans only identified by our program.

There are other important glycan characteristics, the linkage between monosaccharides and glycan structure, have not been considered in this program yet. This kind of informations could increase accuracy of glycan profiling. Furthermore, extending this program to deal with O-linked glycosylation will be another future work.

Chapter 6

Acknowledge

I would like to thank my primary advisor Prof. Haixu Tang. He provides many valuable directions and recommendations. My co-worker, Anoop, he gives lots of help in this project, as will as the collaborator, Prof. Yehia Mechref, Department of Chemistry, who provides plenty suggestion and data sets and all Computational Omics Lab members also provide many helps. The administrators, faculties and staffs in school of informatics and computing give me a good opportunity to achieve this goal during the past two years, especially Linda and Rachel help me a lot. Finally, I want to thank all of my friends and family. I cannot finish this project without their support.

Bibliography

- [1] Csar M. Garca. 2007 investor's conference five-year long range plan, 2007.
- [2] R. Apweiler, H. Hermjakob, and N. Sharon. On the frequency of protein glycosylation, as deduced from analysis of the swiss-prot database. *Biochim Biophys Acta*, 1473(1):4–8, 1999. Apweiler, R Hermjakob, H Sharon, N Comparative Study Review Netherlands Biochimica et biophysica acta Biochim Biophys Acta. 1999 Dec 6;1473(1):4-8.
- [3] Enca Martin-Rendon and Derek J. Blake. Protein glycosylation in disease: new insights into the congenital muscular dystrophies. *Trends in Pharmacological Sciences*, 24(4):178 – 183, 2003.
- [4] T. F. Scanlin and M. C. Glick. Terminal glycosylation and disease: influence on cancer and cystic fibrosis. *Glycoconj J*, 17(7-9):617–26, 2000. Scanlin, T F Glick, M C Historical Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. Review United States Glycoconjugate journal Glycoconj J. 2000 Jul-Sep;17(7-9):617-26.
- [5] G.A. Turner. N-glycosylation of serum proteins in disease and its investigation using lectins. *Clinica Chimica Acta*, 208(3):149 – 171, 1992.
- [6] T. M. Block, M. A. Comunale, M. Lowman, L. F. Steel, P. R. Romano, C. Fimmel, B. C. Tennant, W. T. London, A. A. Evans, B. S. Blumberg, R. A. Dwek, T. S. Mattu, and A. S. Mehta. Use of targeted glycoproteomics to identify serum glycoproteins that correlate with liver cancer in woodchucks and humans. *Proc Natl Acad Sci U S A*, 102(3):779–84, 2005. Block, Timothy M Comunale, Mary Ann Lowman, Melissa Steel, Laura F Romano, Patrick R Fimmel, Claus Tennant, Bud C London, W Thomas Evans, Alison A Blumberg, Baruch S Dwek, Raymond A Mattu, Tajinder S Mehta, Anand S Comparative Study Research

Support, Non-U.S. Gov't United States Proceedings of the National Academy of Sciences of the United States of America Proc Natl Acad Sci U S A. 2005 Jan 18;102(3):779-84. Epub 2005 Jan 10.

- [7] G. Durand and N. Seta. Protein glycosylation and diseases: blood and urinary oligosaccharides as markers for diagnosis and therapeutic monitoring. *Clin Chem*, 46(6 Pt 1):795–805, 2000. Durand, G Seta, N Review United states Clinical chemistry Clin Chem. 2000 Jun;46(6 Pt 1):795-805.
- [8] P. Hongsachart, R. Huang-Liu, S. Sinchaikul, F. M. Pan, S. Phutrakul, Y. M. Chuang, C. J. Yu, and S. T. Chen. Glycoproteomic analysis of wga-bound glycoprotein biomarkers in sera from patients with lung adenocarcinoma. *Electrophoresis*, 30(7):1206–20, 2009. Hongsachart, Piyorot Huang-Liu, Rosa Sinchaikul, Supachok Pan, Fu-Ming Phutrakul, Suree Chuang, Yu-Min Yu, Chong-Jen Chen, Shui-Tein Research Support, Non-U.S. Gov't Validation Studies Germany Electrophoresis Electrophoresis. 2009 Apr;30(7):1206-20.
- [9] A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, J. D. Marth, C. R. Bertozzi, G. W. Hart, and M. E. Etzler. Symbol nomenclature for glycan representation. *Proteomics*, 9(24):5398–9, 2009. Varki, Ajit Cummings, Richard D Esko, Jeffrey D Freeze, Hudson H Stanley, Pamela Marth, Jamey D Bertozzi, Carolyn R Hart, Gerald W Etzler, Marilynn E Comment Germany Proteomics Proteomics. 2009 Dec;9(24):5398-9.
- [10] S. Fukui, T. Feizi, C. Galustian, A. M. Lawson, and W. Chai. Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions. *Nat Biotechnol*, 20(10):1011–7, 2002. Fukui, Shigeyuki Feizi, Ten Galustian, Christine Lawson, Alexander M Chai, Wengang Research Support, Non-U.S. Gov't United States Nature biotechnology Nat Biotechnol. 2002 Oct;20(10):1011-7. Epub 2002 Sep 3.
- [11] T. Patwa, C. Li, D. M. Simeone, and D. M. Lubman. Glycoprotein analysis using protein microarrays and mass spectrometry. *Mass Spectrom Rev*, 2010. Journal article Mass spectrometry reviews Mass Spectrom Rev. 2010 Jan 13.
- [12] F. J. Krambeck and M. J. Betenbaugh. A mathematical model of n-linked glycosylation. *Biotechnol Bioeng*, 92(6):711–28, 2005. Krambeck,

Frederick J Betenbaugh, Michael J United States Biotechnology and bioengineering Biotechnol Bioeng. 2005 Dec 20;92(6):711-28.

- [13] Alan L. Rockwood, Steven L. Van Orden, and Richard D. Smith. Rapid calculation of isotope distributions. *Analytical Chemistry*, 67(15):2699–2704, 1995. doi: 10.1021/ac00111a031.
- [14] Z. Kyselova, Y. Mechref, M. M. Al Bataineh, L. E. Dobrolecki, R. J. Hickey, J. Vinson, C. J. Sweeney, and M. V. Novotny. Alterations in the serum glycome due to metastatic prostate cancer. *J Proteome Res*, 6(5):1822–32, 2007. Kyselova, Zuzana Mechref, Yehia Al Bataineh, Mohammad M Dobrolecki, Lacey E Hickey, Robert J Vinson, Jake Sweeney, Christopher J Novotny, Milos V GM24349/GM/NIGMS NIH HHS/United States RR018942/RR/NCRR NIH HHS/United States Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't United States Journal of proteome research *J Proteome Res*. 2007 May;6(5):1822-32. Epub 2007 Apr 14.
- [15] H. W. Resson, R. S. Varghese, L. Goldman, Y. An, C. A. Loffredo, M. Abdel-Hamid, Z. Kyselova, Y. Mechref, M. Novotny, S. K. Drake, and R. Goldman. Analysis of maldi-tof mass spectrometry data for discovery of peptide and glycan biomarkers of hepatocellular carcinoma. *J Proteome Res*, 7(2):603–10, 2008. Resson, Habtom W Varghese, Rency S Goldman, Lenka An, Yanming Loffredo, Christopher A Abdel-Hamid, Mohamed Kyselova, Zuzana Mechref, Yehia Novotny, Milos Drake, Steven K Goldman, Radoslav R01 CA115625/CA/NCI NIH HHS/United States R03 CA119288/CA/NCI NIH HHS/United States R03 CA119313/CA/NCI NIH HHS/United States R03 CA119313-01A2/CA/NCI NIH HHS/United States Research Support, N.I.H., Extramural United States Journal of proteome research *J Proteome Res*. 2008 Feb;7(2):603-10. Epub 2008 Jan 12.
- [16] Z. Tang, R. S. Varghese, S. Bekesova, C. A. Loffredo, M. A. Hamid, Z. Kyselova, Y. Mechref, M. V. Novotny, R. Goldman, and H. W. Resson. Identification of n-glycan serum markers associated with hepatocellular carcinoma from mass spectrometry data. *J Proteome Res*, 9(1):104–12, 2010. Tang, Zhiqun Varghese, Rency S Bekesova, Slavka Loffredo, Christopher A Hamid, Mohamed Abdul Kyselova, Zuzana Mechref, Yehia Novotny, Milos V Goldman, Radoslav Resson, Habtom W R03CA119313/CA/NCI NIH HHS/United States R21CA130837/CA/NCI NIH HHS/United States Research Support,

N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. United States Journal of proteome research J Proteome Res. 2010 Jan;9(1):104-12.

- [17] P. G. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, and R. Aebersold. A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol*, 22(11):1459–66, 2004. Pedrioli, Patrick G A Eng, Jimmy K Hubley, Robert Vogelzang, Mathijs Deutsch, Eric W Raught, Brian Pratt, Brian Nilsson, Erik Angeletti, Ruth H Apweiler, Rolf Cheung, Kei Costello, Catherine E Hermjakob, Henning Huang, Sequin Julian, Randall K Kapp, Eugene McComb, Mark E Oliver, Stephen G Omenn, Gilbert Paton, Norman W Simpson, Richard Smith, Richard Taylor, Chris F Zhu, Weimin Aebersold, Ruedi 1R33CA93302/CA/NCI NIH HHS/United States N01-HV-28179/HV/NHLBI NIH HHS/United States Evaluation Studies Research Support, U.S. Gov't, P.H.S. United States Nature biotechnology Nat Biotechnol. 2004 Nov;22(11):1459-66.