

**Imperial College
London**

Glycoproteomic Research using Mass Spectrometry

A thesis submitted for the Degree of Doctor of Philosophy of Imperial College
of Science, Technology and Medicine, London

Submitted by

Laura Bouché

Imperial College London
Department of Life Sciences
London SW72AZ
United Kingdom

April 2017

To my parents

*“Science, for me, gives a partial explanation for life.
In so far as it goes, it is based on fact, experience and experiment.”*

*Rosalind Franklin
(1920 – 1958)*

Abstract

The development of problem-specific mass spectrometric (MS) glycoproteomic strategies has allowed the discovery of previously unknown protein glycosylation in both eukaryotic and prokaryotic organisms. The research in this thesis focuses on the identification and structural characterisation of novel glycan structures in ADAMTS13 and *Clostridium difficile*.

ADAMTS13 is a large multi-domain protein which regulates thrombogenesis by cleavage of the adhesive blood glycoprotein, VWF, so generating smaller less thrombogenic fragments. Glycoproteomic strategies were employed to investigate a secretion-enhancing mutant by comparing the O-glycome of wild type (WT) with synonymous substitution, P118P, and a non-synonymous control, P118F. Identical post-translational modifications (PTMs) but several novel PTMs were discovered in ADAMTS13 including TSR1 O-glycosylation, C-mannosylation of W387 and DiSialyl Core-1 O-glycosylation of S1170.

Clostridium difficile is one of the main organisms responsible for morbidity in hospitalised patients, and is the etiological agent of antibiotic-associated diarrhoea and pseudomembranous colitis. The *C.difficile* cell wall is surrounded by an S-layer composed of two proteins, high molecular weight (HMW) and low molecular weight (LMW) SLPs. 12 slpA gene cassettes have been recently described and cassette-11 carries an insert containing 19 ORFs. Combining different biochemical and ES- and MALDI-MS approaches, including the ETD technique, with genetic experiments, it was demonstrated that LMW SLP in strain Ox247 is glycosylated with a surprisingly large linear pentose-branched oligosaccharide of more than forty sugar residues, and a collaborative NMR study suggests a Phospho- and Acetyl- substituted non-reducing terminal rhamnose.

Analysing different *C.difficile* hypervirulent strains, novel flagellar sulphonated peptidylamido-glycan structures not previously observed in sugar or amino acid chemistry were identified. High resolution mass measurement and negative-ion nanospray MS/MS of cone-voltage-induced fragment ions were crucial in allowing the discovery of a unique terminal Taurine (aminoethyl-sulphonic acid) peptidylamidoglycan unit which could provide a novel strategy to escape the immune system, by the *C.difficile* becoming more virulent.

List of publications

Bouché L., Panico M., Hitchen P., Binet D., Sastre F., et al (2016) The type B Flagellin of Hypervirulent *Clostridium difficile* is Modified with Novel Sulphonated Peptidylamido-Glycans. J Biol Chem 2016 Oct 7. pii: jbc.M116.749481.

Valiente E., **Bouché L.**, Hitchen P., Faulds-Pain A., Songane M., et al. (2016) Role of glycosyltransferases modifying type B flagellin of emerging hypervirulent *Clostridium difficile* lineages and their impact on motility and biofilm formation. J Biol Chem 2016 Oct 4. pii: jbc.M116.749523.

Panico M., **Bouché L.**, Binet D., O'Connor M.J., Rahman D., et al. (2016) Mapping the complete glycoproteome of virion-derived HIV-1 gp120 provides insights into broadly neutralizing antibody binding. Sci Rep. 2016 Sep 8;6:32956. doi: 10.1038/srep32956.

Stansell E., Panico M., Canis K., Pang P.C., **Bouché L.**, Binet D., et al. (2015) Gp120 on HIV-1 Virions Lacks O-Linked Carbohydrate. PLoS ONE 10 (4): e0124784. doi:10.1371/journal.pone.0124784.

Originality Declaration

I hereby declare that the work presented in this thesis has not been previously or concurrently submitted for any other degree, diploma or other qualification at any other university. It is the result of the author's own independent investigation unless otherwise stated.

Laura Bouché

April 2017

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial no Derivatives licence. Researchers are free to copy, distribute, or transmit the thesis on the conditions that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Acknowledgements

I could not walk through my PhD alone.

I would like first of all to express my deepest gratitude to Professor Howard Morris and Professor Anne Dell. You simply tirelessly guided me through the world of mass spectrometry and glycoproteomics. Your brilliant approach to leadership and scientific research have been enormously inspirational and I would like to thank you for reinforcing my confidence as a scientist.

I am beyond gratitude to Dr Maria Panico, for her constant support, her appreciation of my Neapolitan sense of humor and her invaluable advice in my personal and professional life. She listened to me for countless hours, helped me not only to carry my research on, but also to build a better myself during these years away from home. I could not have wished for a better role model.

It was a joy to work with the highly professional and dedicated biopolymer group at Imperial College London: Dr. Stuart Haslam, Dr. Valeria Ventura, Dr. Paola Grassi, Dr. Poh Choo Pang, Dr. Paul Hichen, Dr. Matthew Choo, Dr. Aristotelis Antonopolous, Dr. Grigorji Sutov, Dr. Gang Wu, Dr. Nan Jia, Dr. Qiushi Chen, Dr. Simon North, Dr. Francesco Piacente, Dr. Tiandi Yang, Dinah Rahman, Dongli Lu, Linda Ibeto, Matteo Gaglianone and Khadija El Jellas. Thank you all for the interesting discussions, friendliness and welcome assistance in more taxing times that have made this experience enormously enjoyable and memorable. I will never forget you!

To our collaborators and supervisors outside of the laboratory, I would like to thank Professor Neil Fairweather and his team based at Imperial College London, and in particular a big thank you to Emma Richards, not only a collaborator, but over these years she has become a friend to me; Professor Brendan Wren and his group at the London School of Hygiene and Tropical Medicine; Dr Susan Logan based at National Research Council, Ottawa, Canada and Dr. Chava Kimchi-Sarfaty and Dr. Ryan Hunt at the United States Food and Drug Administration (FDA), it has been such a pleasure to learn from you and explore these projects with you all. My sincere thanks go to all people working at BioPharmaSpec Ltd in Jersey, especially to Daniel Binet, Michael-John O'Connor and Dr. Christina Morris, thank you very much for all your patience and appreciated assistance.

An enormous thank you to all my Londoner friends, who have offered amazing breaks, memorable chats, uncountable G&Ts and kept me thoroughly entertained, throughout my PhD. I love you so much guys!! A big shout out goes to Giovanna, Gabriele, Emanuele,

Velia, Noemi, Carmela, Paola, Saria, Danilo, Umberto, Tiziana, Loredana, Alfonso, Virginia, Pietro, Maria, Pasquale e Luciano. Life would not be the same without you all!!!

And also a big thank you to all my friends spread around the world, but always next to me when I needed to! Thank you to Luigi and Nico, Laura, Genny and Luca, Xenia, Eleonora, Erika, Vera, Luigino, Mario, Maximilian, Annabel, Flavia and o'regista. I am so grateful for all the crazy, loud and happy moments over the past years and those to come.

Thank you to my amazing family: my hardworking and super special grandparents, I know you have been keeping an eye on me all the time; my super aunty Silvana and uncle Antonio and my incredible cousin Alessandra, my sister in law Marina and everyone else that makes up my wonderful family.

Finally and most importantly, the inner circle. I am unbelievable grateful to share my life's journey with you all. Each of you inspires me everyday. Mammina and Papi, thank you so much for all your support at every stepping stone of my life, your endless love and affection, thank you for your hugs and sweetness, and also for your patience and tolerance, thank you for always being there and giving me the strength to go forward on the days of weakness. Thank you to my big brother Stefano, thank you for all the laughter and fun, thank you for being with me, because if you weren't, I couldn't go any further. And simply for letting me become the aunty of our stunning Daria! I am very proud to be a Bouché!!!

Just a note: thesis soundtrack "Voodoo Child" by Jimi Hendrix!! Thanks Jimi!!

Table of Contents

Abstract.....	4
List of publications	5
Originality Declaration	5
Copyright Declaration.....	5
Acknowledgements.....	6
Table of Contents.....	8
List of Figures.....	12
Chapter 1:.....	30
Introduction.....	30
1. Introduction.....	31
1.1 Overview of Structural Glycobiology.....	31
1.2 Glycosylation in Eukaryotes	33
1.2.1 N-glycosylation in Eukaryotes.....	33
1.2.2 O-glycosylation in Eukaryotes.....	37
1.3 Glycosylation in Prokaryotes	39
1.3.1 N-glycosylation in Bacteria	40
1.3.2 O-glycosylation in Bacteria	43
1.3.3 N- and O-glycosylation in Archaea	45
1.4 Mass spectrometry.....	47
1.4.1 Historical background of mass spectrometry.....	48
1.4.2 Ionisation techniques	49
1.4.3 Mass Analysers	59
1.4.4 Tandem mass spectrometry.....	63
1.4.5 Fragmentation & Interpretation	69
1.5 Research Project Overview	77
1.5.1 Overview of ADAMTS13 Structural Biology.....	77

1.5.2	Overview of <i>Clostridium difficile</i>	80
1.5.2.1	<i>Clostridium difficile</i> genome	81
1.5.2.2	<i>Clostridium difficile</i> cell envelope and its components involved in host interactions	82
1.5.2.3	<i>Clostridium difficile</i> flagella	88
1.6	Project aims	91
Chapter 2:		92
Materials & Methods		92
2.	Materials & Methods	93
2.1.	Materials	93
2.2.	Equipment & Consumables	94
2.3.	Sample preparation	95
2.4.	Methods	95
2.4.1	Purification of <i>Clostridium difficile</i> S-layer proteins	95
2.4.2	SDS-PAGE electrophoresis of ADAMTS13 and <i>Clostridium difficile</i> S-layer proteins 96	
2.4.3	In-gel digestion of protein bands and elution of peptides	96
2.4.4	In-solution digestion	97
2.4.5	Electrotransfer: tank transfer	97
2.4.6	Release of O-glycans from <i>Clostridium difficile</i> S-layer glycopeptide	97
2.4.6.1	Reductive elimination and purification of O-glycans	98
2.4.6.2	Ion exchange clean-up	98
2.4.7	Hydrofluoric acid (HF) hydrolysis	98
2.4.8	Chemical derivatisation of carbohydrates for MS analysis	98
2.4.8.1	Permethylation	99
2.4.8.2	Deutero-permethylation	99
2.4.8.3	Sep-Pak [®] C ₁₈ purification of derivatised glycans	100
2.4.8.4	Alditol acetates	100

2.4.8.5	Linkage analysis	100
2.4.9	Mass spectrometric analysis	101
2.4.9.1	GC-MS	101
2.4.9.2	Nanospray MS and MS/MS with API Q-STAR ABSciex 5600 and Xevo G2 mass spectrometers	101
2.4.9.3	On-line LC-ES MS and MS/MS with an API Q-STAR ABSciex 5600 and Xevo G2 mass spectrometers	102
2.4.9.4	Off-line LC-MALDI-TOF and TOF/TOF MS/MS	102
2.4.9.5	MALDI-TOF MS and TOF/TOF MS/MS analysis of glycans	103
2.4.9.6	Waters Synapt G2-S	103
2.4.10	Data processing and interpretation	104
Chapter 3:	106
Characterisation of ADAMTS13 PTMs	106
3.	Characterisation of ADAMTS13 PTMs	107
3.1	Introduction	107
3.2	Experimental strategy	108
3.3	Results & Interpretation	109
3.3.1	TSRs	109
3.3.2	Other Discoveries	132
3.4	Discussion	135
Chapter 4:	138
Type B Flagellin of Hypervirulent <i>Clostridium difficile</i> Strains	138
4.	Type B Flagellin of Hypervirulent <i>Clostridium difficile</i> Strains	139
4.1	Historical Perspective on Type A and Type B flagellin	139
4.2	Experimental Strategy	140
4.3	Results	143
4.4	Conclusion	163
Chapter 5:	166

<i>Clostridium difficile</i> S-layer Research	166
5. <i>Clostridium difficile</i> S-layer Research	167
5.1 Introduction	167
5.2 Glycoproteomic Results	169
5.3 ETD fragmentation.....	181
5.4 β -elimination strategies	183
5.5 Analysis of the β -eliminated sample	185
5.6 Composition data:	191
5.7 NMR analysis.....	201
5.8 Overall Cross-correlation of the NMR and MS Data.....	202
Chapter 6:.....	216
Conclusion	216
6. Conclusion	217
Chapter 7:.....	219
References.....	219

List of Figures

Figure 1.1 Attachment of an N-linked glycan, specifically an N-acetylglucosamine, to an asparagine residue.....	33
Figure 1.2 The common core structure of N-linked glycans comprising three mannoses and two N-acetylglucosamines attached to an Asn residue within the consensus sequence Asn-X-Ser/Thr (Man α 1-6(Man α 1-3)Man β 1-4GlcNAc β 1-4GlcNAc β 1-Asn).....	34
Figure 1.3 O-linked glycosylation in Eukaryotes. The diagram representation shows the variety of protein-glycan linkages used by Eukaryotes in O-linked glycosylation.	37
Figure 1.4 The diagram shows an Electron Impact mechanism where the red circles are the gaseous volatile sample and the red circles with a plus are the highly energetic radical molecular cations produced following the bombardment by the electrons coming from a tungsten filament. The product ion is a radical cation (M ⁺).....	51
Figure 1.5 Example of linkage analysis. The chart displays the procedure of linkage analysis of a α 1,4-linked glucose branch. The terminal α 1,6 glucose, which is acetylated at C-1 and C-5, while the two α 1,4 glucose units on either side of the branching monosaccharide are acetylated at C-1, C-4 and C-5. The branching glucose is acetylated at C-1, C-4, C-5 and C-6. The schematic has been adapted from Varki and coworkers (Varki 2009).....	52
Figure 1.6 The diagram representation shows the Chemical Ionisation mechanism where the red circles are the gaseous volatile sample and the black circles are the reagent gas (methane). Following a collision between a sample molecule and a reagent ion, there is the formation of a quasi-molecular ion (M+H ⁺).....	54
Figure 1.7 The diagram representation shows the Fast Atom Bombardment mechanism in which a beam of high energy atoms/ions strikes a surface to create sample ions. Once generated, the gas-phase ions can reach the mass analyser, which can be either a magnetic sector or a triple quadrupole.	56
Figure 1.8 The diagram representation shows the Electrospray Ionisation mechanism. In positive ion mode, a positive potential is applied to the capillary tip with the following formation of positively charged droplets. Solvent evaporation of the charged droplets via a weak nitrogen countercurrent leads to gas-phase ions. The sample ions are driven by an electric potential and pressure difference and the potential difference over the cone orifice and downstream ion optical elements transports the sample ions into the high-vacuum region of the mass analyser chamber.	57

Figure 1.9 Typical ES-MS spectrum showing a distribution of signals carrying varying numbers of net positive charges. Specifically this MS spectrum shows two isoforms of haemoglobin, with the peak at m/z 616 being the heme group.58

Figure 1.10 The diagram representation displays the Matrix Assisted Laser Desorption Ionisation mechanism in positive ion mode. The sample is mixed with a crystalline matrix which presents an absorption maximum near the wavelength of the pulsed laser used to ionise the sample. Once the matrix absorbs the pulsed laser energy, flash evaporated and push the sample into the gas phase. Then the gas-phase ions can fly towards the mass analyser.59

Figure 1.11 The diagram representation displays the Magnetic Sector mechanism. In a magnetic sector analyser the ions pass through an electrostatic analyser (ESA) which produce a mono-energetic beam (orange line) and once reached a magnetic field, is deflected to a circular motion of a unique radius in a direction perpendicular to the applied magnetic field.60

Figure 1.12 The diagram representation displays the Ion Trap mechanism.....61

Figure 1.13 The diagram shows a quadrupole analyser where the red ion has the selected m/z value, therefore passing out of the quadrupole into the mass detector, while the path of the blue ion is intercepted by a quadrupole rod, which results in the ion being de-charged before reaching the mass detector and so not being detected.62

Figure 1.14 The above diagram depicts a Voyager-DE STR System with a Time-Of-Flight analyser. The TOF analyser separates the ions based on their differences in velocities therefore the ions reach the detector at different times. However, all ions have to enter the field-free flight tube at the same time. The ion mirror are necessary to improve the resolution.....63

Figure 1.15 The above diagram depicts a Triple Quadrupole mass spectrometer. The Quadrupole labelled Q_2 is just an ion guide around the gas cell.65

Figure 1.16 The diagram depicts a quadrupole analyser hexapole collision cell and orthogonal acceleration Time-Of-Flight (TOF). The hexapole collision cell makes the collision more efficient and the orthogonal geometry minimises the level of chemical noise coming from the collision cell.67

Figure 1.17 The MALDI TOF/TOF 4800 Analyser from Applied Biosystems is a double focusing instrument with linear and reflector flight path options, permitting the selection of individual precursor ions (TIS) for further fragmentation in the collision cell. Moreover fragmentation is facilitated by the presence of an inert gas, i.e. argon, in the collision cell (CID) and gas pressure is controlled to provide a satisfactory fragmentation.....69

Figure 1.18 The peptide fragmentation nomenclature is shown. Protein backbone cleavages lead to two families of potential fragments, called a, b and c ions (if the charge is retained by the N-terminal fragment) and x, y and z ions (if the charge is on the C-terminal fragment). As the CO-NH bonds result in the most susceptible cleavage site, the b and/or y ions are generally the most abundant fragments observed during MS/MS experiments on peptides. Considering the ETD mechanism, the cleavage of N-C α bonds occurs and produces c- and z-type fragments. Nevertheless, the mass difference between two adjacent a/b/c or x/y/z ions provides the mass to identify the amino acid residue and thus allows sequencing of the selected peptide (Roepstorff 1984). 71

Figure 1.19 Systematic nomenclature of carbohydrate fragmentation. Cleavage of the oligosaccharide takes place by breaking the glycosidic bonds (B, C, Y, Z) and across rings (A and X). Fragment ions containing the non-reducing terminal are A, B and C types, while the reducing end fragments are X, Y and Z types. The subscripts of A and X denote the bonds broken in order to form the respective fragments. 73

Figure 1.20 Fragmentation and cleavages of oligosaccharides (positive ion). A type cleavages form oxonium ions. Glycosidic cleavages with hydrogen transfer are known as β -cleavages, where the fragment ions can be either reducing (top panel) or non-reducing (lower panel). Examples of ring cleavages are shown in the bottom three panels. 74

Figure 1.21 Domain structure of ADAMTS13: The domain structure of ADAMTS13 includes of a metalloprotease domain (red box), a disintegrin domain (green box), TSR1 (light blue oval), a cysteine rich domain (yellow box), a spacer domain (white box), seven additional TSRs and two CUB domains (orange boxes). Potential consensus sequences for O-fucosylation within the TSRs are indicated with *. 78

Figure 1.22 Scheme of cell envelopes of Gram-positive bacteria. The S-layer is attached to the peptidoglycan layer and is anchored to the the peptidoglycan via lipoteichoic acids, since they have a lipid component that can assist in anchoring as the lipid component is embedded in the plasma membrane. 85

Figure 1.23 The cwp cluster of strains 630 and Ox247 are slightly different. The Ox247 strain lacks cwp2 and the cwp66 and cd2790 have been rearranged. This strain also includes a novel insertion of 23.8 kb containing 19 predicted open reading frames (ORFs) within one of the S-layer cassettes, SLC 11, between the cwp66 and the cd2790. 86

Figure 3.1 Separation of the ADAMTS13 glycoprotein samples by SDS-PAGE gel (Invitrogen NuPAGE 4-12% Bis Tris Gel) and visualised by Coomassie blue staining. Molecular size markers are indicated on the left. Three different kinds of samples were run,

wild type (WT), synonymous P118P mutation ADAMTS13 glycoprotein and non-synonymous P118F control. Lane 1: bench mark (5 µg); lanes 2-5-8: empty; lanes 3-4: WT; lanes 6-7: P118P; lanes 9-10: P118F. 108

Figure 3.2 Experimental strategy employed in ADAMTS13 PTMs characterisation. Following trypsin in gel digestion of the samples, mass mapping was carried out. The mass spectrometric analysis was done in positive ion mode and using on-line LC-ES-MS and MS/MS..... 109

Figure 3.3 TIC (Total Ion Current) for the 90 minute separation of the ADAMTS13 WT digest *online*-nano-LC-ES-MS/MS experiment. The TIC represents the summed intensity across the entire range of masses being detected at every point during the run. 113

Figure 3.4 MS LC profiles and MS and MS/MS spectra with the 447 signal highlighted in red. The top left hand panel corresponds to the TIC; the top right hand panel shows the MS spectrum; the bottom left hand panel is the XIC (Extracting Ion Current) and the bottom right hand panel consists in the MS/MS spectrum of the m/z 447..... 114

Figure 3.5 MS/MS spectrum of m/z 601²⁺ in WT ADAMTS13. Peptide fragmentation provides very strong evidence for the sequence SJGGGVVTR, all the y ions, expect for y₁, and some b ions have been found. 115

Figure 3.6 MS/MS spectrum of m/z 601²⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in **Figure 3.5**. 115

Figure 3.7 MS/MS spectrum of m/z 601²⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both **Figure 3.5** and **Figure 3.6**. 116

Figure 3.8 MS/MS spectrum of m/z 765.8²⁺ in WT ADAMTS13. Peptide fragmentation provides very strong evidence for the sequence GPJSVVSJGAGLR, the majority of y ions and some b ions have been found..... 117

Figure 3.9 MS/MS spectrum of m/z 765.8²⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in **Figure 3.8**. 118

Figure 3.10 MS/MS spectrum of m/z 765.8²⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both **Figure 3.8** and **Figure 3.9**..... 118

Figure 3.11 MS/MS spectrum of m/z 1381.2⁴⁺ in WT ADAMTS13. Peptide fragmentation provides evidence for the sequence ELVETVQJQGSQQPPAWPEAJVLEPJPYWAVGDFGPJSASJGGGLR..... 119

Figure 3.12 MS/MS spectrum of m/z 1380.5⁴⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in **Figure 3.11**..... 120

Figure 3.13 MS/MS spectrum of m/z 1380.65 ⁴⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both Figure 3.11 and Figure 3.12 .	120
Figure 3.14 MS/MS spectrum of m/z 857.2 ³⁺ in WT ADAMTS13. Peptide fragmentation provides evidence for the sequence TGAQA AHVWTPVAGSJSVSJGR.	121
Figure 3.15 MS/MS spectrum of m/z 857.2 ³⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in Figure 3.14 .	122
Figure 3.16 MS/MS spectrum of m/z 857.2 ³⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both Figure 3.14 and Figure 3.15 .	122
Figure 3.17 MS/MS spectrum of m/z 695.90 ²⁺ in WT ADAMTS13. Peptide fragmentation provides evidence for the sequence LAAJSVSJGR.	123
Figure 3.18 MS/MS spectrum of m/z 695.90 ²⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in Figure 3.17 .	124
Figure 3.19 MS/MS spectrum of m/z 695.90 ²⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both Figure 3.17 and Figure 3.18 .	124
Figure 3.20 LC-MS chromatogram of 936.4 ²⁺ (top panel) and 1017.4 ²⁺ (bottom panel) in WT ADAMTS13.	125
Figure 3.21 MS/MS spectrum of m/z 936.40 ²⁺ in WT ADAMTS13. Peptide fragmentation provides evidence for the sequence VM SLGPJSASJGLGTAR. The peak at m/z 1726 corresponds to a loss of 146 Da (fucose) from m/z 1871 [M+H] ⁺ .	126
Figure 3.22 MS/MS spectrum of m/z 1017.4 ²⁺ in WT ADAMTS13. Peptide fragmentation provides evidence for the sequence VM SLGPJSASJGLGTAR.	126
Figure 3.23 MS/MS spectrum of m/z 936.40 ²⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in Figure 3.21 .	127
Figure 3.24 MS/MS spectrum of m/z 1017.4 ²⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in Figure 3.22 .	128
Figure 3.25 MS/MS spectrum of m/z 1017.4 ²⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both Figure 3.22 and Figure 3.24 .	129
Figure 3.26 MS/MS spectrum of m/z 859.0 ³⁺ in WT ADAMTS13. Peptide fragmentation provides evidence for the sequence WHVGTWMEJSVSJGDGIQR.	130
Figure 3.27 MS/MS spectrum of m/z 859.0 ³⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in Figure 3.26 .	131
Figure 3.28 MS/MS spectrum of m/z 859.0 ³⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both Figure 3.26 and Figure 3.27 .	131

Figure 3.29 Glycosylation of W-387 in ADAMTS13. The quasi-molecular ion is 162 Da, a hexose unit, higher than the theoretical peptide mass, and the position of substitution of the hexose is shown to be on the first tryptophan, W-387, as proven by the interpretation of the fragment ions observed, and illustrated in schematic with suggested mechanisms. The principal fragment ions observed correspond to the novel formation of a 2-Ethynyl-Indole (Acetylenic substituent) on the side chain of W-387 resulting from loss of 138 Da caused by water loss and partial cleavage of the hexose ring, in preference to the normal β -elimination (162 Da) seen in O-linked glycosylation chemistry. 133

Figure 3.30 The partial MS spectrum (m/z 550-1200) at the corresponding LC-MS elution time where the MS/MS of a signal at 1065.5²⁺ gave b and y” signals attributable to the sequence GLLFSPAPQPR in WT ADAMTS13. The spectrum indicates an ion at 1065.5²⁺ and a corresponding quasi-molecular ion for the peptide component of 1182.6 (591.8²⁺) which calculates for a NeuAc₂HexHexNAc unit attached to Ser-1170 of the peptide sequence assigned. From an understanding of mammalian biosynthetic pathways, this likely corresponds to a DiSialyl Core-1 structure. 134

Figure 3.31 The partial MS spectrum (m/z 500-1200) where the MS/MS of a signal at 1065.5²⁺ gave b and y” signals attributable to the sequence GLLFSPAPQPR in P118P ADAMTS13. 134

Figure 3.32 The partial MS spectrum (m/z 500-1200) where the MS/MS of a signal at 1065.5²⁺ gave b and y” signals attributable to the sequence GLLFSPAPQPR in P118F ADAMTS13. 135

Figure 4.1 Type A and B flagellin in *C.difficile*. The *C.difficile* flagellin is formed of a single flagellin protein which is post-translationally modified by an O-linked sugar. Two types of flagellin PTM have been identified in *C.difficile*, named Type A and B. 139

Figure 4.2 Experimental strategy employed in *Clostridium difficile* Type B flagellin structural characterisation. The sample was digested with trypsin, then mass mapping was carried out. The mass spectrometric analysis was done in both positive and negative ion modes using different on-line LC-ES-MS and MS/MS instrumentations. 141






Figure 4.3 *Clostridium difficile* flagellin sequence 143

Figure 4.4 LC-MS/MS spectrum of m/z 991²⁺ 145

Figure 4.5 LC-MS/MS spectrum of m/z 998²⁺ 146

Figure 4.6 Fragment suggestion for the m/z 396 to give m/z 128 and m/z 268. 147

Figure 4.7 A-B: Nanospray of 8-hydroxyadenosine. The A spectrum is the MS spectrum of the commercial adenosine standard, whereas the B spectrum is the MS/MS of m/z 152. 148

Figure 4.8 MS/MS spectrum of 998 ²⁺ from m/z 60 to 450.....	149
Figure 4.9 Possible β-elimination mechanism (red arrows) with consequently loss of 128 from 396 where the red box at this stage remains unknown, and an alternative elimination mechanism (blue arrows) with formation of highly stable triply conjugated cyclic ion structure at m/z 111, respectively.	150
Figure 4.10 Possible fragmentation mechanism to lose NH ₃ and CO from m/z 268.	151
Figure 4.11 Positive ion online nano-LC MS/MS high resolution CID mass spectrum of m/z 998 ²⁺	151
Figure 4.12 Positive ion online nano-LC MS/MS high resolution CID mass spectrum of m/z 998 ²⁺ from 50-100 m/z.....	152
Figure 4.13 Positive ion online nano-LC MS/MS high resolution CID mass spectrum of m/z 998 ²⁺ from 100-200 m/z.....	153
Figure 4.14 Positive ion online nano-LC MS/MS high resolution CID mass spectrum of m/z 998 ²⁺ from 200-300 m/z.....	153
Figure 4.15 Positive ion online- nano-LC MS/MS high resolution CID mass spectrum of m/z 998 ²⁺  : peptide;  : HexNAc;  -Me :MethydeoxyHex;  : deoxyHex;  -NH ₂ : Amino-dideoxyHex.....	156
Figure 4.16 Possible fragmentation mechanism to form m/z 152.	157
Figure 4.17 Positive ion nanospray CID MS/MS spectrum of m/z 152 produced via cone-voltage induced in-source fragmentation of 998 ²⁺ . Signals at m/z 70 and 108 correspond to losses of 82 and 44 mass units respectively.....	159
Figure 4.18 Negative ion nanospray CID MS/MS spectrum of m/z 996 ²⁻ (m/z 100-1200). 160	
Figure 4.19 Suggested fragmentation mechanism to produce key signals in the positive and negative ion spectra of respectively the 998 ²⁺ and 996 ²⁻ glycopeptide quasi molecular ions.	161
Figure 4.20 Negative ion nanospray CID MS/MS spectrum of m/z 124 produced via cone voltage induced in-source fragmentation of m/z 996 ²⁻	162
Figure 4.21 Chemical structure of the novel Sulphonated Peptidylamido-Glycan found in hypervirulent strains of <i>Clostridium difficile</i> . The structure drawn produces the tryptic glycopeptide m/z 991 ²⁺ variant and the [CH ₃] above the glycine indicates the other variant observed at m/z 998 ²⁺ in the work described here.	163
Figure 5.1 Simplified representation of the experimental strategy employed in this section which were two-fold involving glycomics and glycoproteomics. S-layer glycoprotein of <i>Clostridium difficile</i> Ox247 and the S-layer proteins from several gene-deletions of the strain	

under study were purified by gel electrophoresis, enzymatically digested, separated by liquid chromatography and analysed using Q-TOF methodology. Fragmentation data were acquired using multiple different complementary techniques including Collision Induced Dissociation (CID) and Electron Transfer Dissociation (ETD). Moreover, a range of analysis techniques comprising GC-MS, MALDI-MS and MS/MS, ES-MS and MS/MS as well as chemical derivatisation and hydrolysis were employed to gain a further understanding of the overall structures discovered..... 168

Figure 5.2 The putative 23.8 kb locus inserted between CD2790 and *cwp66* in *Clostridium difficile* Ox247 has been analysed by BLAST and the putative functions are indicated by different colour. White (1, 5, 15): unknown; Blue (2, 4): initiating glycosyltransferases; Green (3, 6- 10): glycosyltransferases; Orange (11, 12): ABC transporters; Purple (13, 14, 16- 18): rhamnose biosynthetic genes; Red (19): O-specific ligase..... 169

Figure 5.3 Separation of the *C.difficile* proteins by SDS-PAGE gel (Invitrogen NuPAGE 4- 12% Bis Tris Gel) and visualized by Coomassie blue staining. Molecular size markers are indicated on the left. Six different kinds of samples were run, wild type (WT) and mutants of the SLC-11 strain Ox247 (*Orf2*, *Orf3*, *Orf4*, *Orf7* and *Orf16*). In the Ox247 strain HMW SLP and LMW SLP (defined in chapter 1) migrate between 50 and 40 kDa. Each mutant, instead, contains a band at ~20 kDa shown by mass spectrometry to be glycosylated LMW SLP. Lanes 1-4-7-9: bench mark (5µg); lane 2: wild type (WT); lane 3 $\Delta orf2$; lane 5 $\Delta orf3$; lane 6 $\Delta orf4$; lane 8 $\Delta orf7$ and lane 10 $\Delta orf16$; ** see text. 170

Figure 5.4 Amino acid sequence of *Clostridium difficile* Ox247 S-layer. Residues in red correspond to the signal peptide. Residues in black represent the LMW SLP “L” - 20 kDa and residues in blue indicates the HMW SLP “H” – 45 kDa. The tryptic glycopeptide DILAAQNLTGAVILNK has been shown in bold and it is part of the LMW SLP sequence. 171

Figure 5.5 TIC (Total Ion Current) for a 90 minute separation of the Ox247 48 kDa (marked in ** in **Figure 5.3**) digest *online-nano-LC-ES-MS/MS* experiment. The shaded box highlights a retention time at approximately 37 minutes. MS spectra of components at this time reveal a pattern of doubly charged ions centred around m/z 1178 with cone-voltage-induced fragment intervals corresponding to deoxyHex, Hex or Pentose..... 171

Figure 5.6 Pattern of doubly charged ions centred around m/z 1178 at intervals corresponding to deoxyhexose, hexose or pentose ($Hex_3deoxyHex_8Pent_1$). On the bottom right of the spectrum there is a magnification of the peak m/z 1771.71 showing that it is a real signal extending to $PentdHex_8Hex_3$. ○ : hexose; △ : deoxyhexose; ☆ : pentose. 172

Figure 5.7 MS/MS spectrum of m/z 1178 confirming the DILAAQNLTGAVILNK peptide via signals at m/z 229 (b₂), 342 (b₃), 413 (b₄), 484 (b₅), 612 (b₆), 726 (b₇), 839 (b₈), 261 (y₂), 342 (b₃), 374 (y₃), 413 (b₄), 487 (y₅), 586 (y₆)..... 173

Figure 5.8 The list of the ions identified and related composition. Each number corresponds to the peptide plus the sugar/s attached and possible interpretation of the final structure schematic of the oligosaccharide attached at the DILAAQNLTGAVILNK peptide..... 173

Figure 5.9 Summed MS data acquired at 39.5 min in the online nanoLC-ES-MS of a tryptic digest of *C.difficile Orf2::erm* mutant. The expanded middle mass region to show doubly-charged and singly-charged peaks that correspond to the DILAAQNLTGAVILNK peptide. 174

Figure 5.10 Summed MS data acquired at 44.0 min in the *online* nanoLC-ES-MS of a tryptic digest of *C.difficile Orf3::erm* mutant. Expanded middle mass region to show doubly-charged peaks that corresponds to glycans attached to the LMW SLP..... 175

Figure 5.11 Summed MS data acquired at 40.2 min in the *online* nanoLC-ES-MS of a tryptic digest of the band at 20 kDa of *C.difficile Orf4::erm* mutant. Expanded middle mass region to show doubly-charged peaks that corresponds to glycan compositions ranging from a single hexose up to deoxyHex₃Hex. m/z 897.95 is from elsewhere in the digest..... 175

Figure 5.12 Summed MS data acquired at 40.6 min in the *online* nanoLC-ES-MS of a tryptic digest of the band at 25 kDa of *C.difficile Orf7::erm* mutant. Expanded middle mass region to show doubly-charged peaks that corresponds to glycan compositions ranging from a single hexose up to PentdeoxyHex₅Hex..... 176

Figure 5.13 Summed MS data acquired at 42.9 min in the *online* nanoLC-ES-MS of a tryptic digest of the band at 20 kDa of *C.difficile Orf16::erm* mutant. Expanded middle mass region to show doubly-charged peaks that corresponds to glycans compositions ranging from a single hexose up to deoxyHex₄Hex. 177

Figure 5.14 MALDI-TOF mass spectrum of *Clostridium difficile Orf2::erm* mutant. The peak at 1754.94 m/z is the DILAAQNLTGAVILNK peptide in accordance with the data obtained by ES-MS (LC/MS) (**Figure 5.7**). 178

Figure 5.15 MALDI-TOF mass spectrum of *Clostridium difficile Orf3::erm* mutant. The peak at 2063.14 m/z is the DILAAQNLTGAVILNK peptide with an hexose and a deoxyhexose attached (1754 + 162 + 146 = 2063) in accordance with the data obtained by ES-MS (LC/MS) (**Figure 5.10**). 178

Figure 5.16 MALDI-TOF/TOF MS/MS spectra of m/z 2063.14 derived from the MS spectrum of <i>Clostridium difficile</i> Orf3:: <i>erm</i> mutant. Assignment of the fragment ions are shown.	179
Figure 5.17 Expanded MALDI-TOF/TOF MS/MS spectra of m/z 2063.14 derived from the MS spectrum of <i>Clostridium difficile</i> Orf3:: <i>erm</i> mutant. Peptide fragmentation provides very strong evidence for the sequence DILAAQNLTGAVILNK. b-ions are labelled in red and y-ions are labelled in blue. The peptide fragmentation is shown. Protein backbone cleavages involved lead to the formation of b ions (the charge is retained by the N-terminal fragment) and y ions (the charge is on the C-terminal fragment). The mass difference between two adjacent b or y ions provides the mass and identity of the amino acid residue and thus allows sequencing of the DILAAQNLTGAVILNK peptide. For site of attachment interpretation see text.	180
Figure 5.18 TIC for the 50 minute elution profile of Orf3:: <i>erm</i> 20 kDa Tryptic digest chromatogram for an ETD experiment. The DILAAQNLTGAVILNK peptide is eluted at approximately 26 minute, whereas the glycopeptide is found at approximately 25 minute. Both are seen in a quadruply charged state.	181
Figure 5.19 ETD spectra of m/z 439.52 ⁴⁺ and 516.55 ⁴⁺ derived from the <i>Clostridium difficile</i> Orf3:: <i>erm</i> mutant. Peptide fragmentation provides very strong evidence for the sequence DILAAQNLTGAVILNK. c-ions are labelled in orange and z-ions are labelled in purple. A: The peptide fragmentation is shown. Considering the ETD mechanism, the cleavage of N-C α bonds occurs and produces c- and z- type fragments. The mass difference between two adjacent c or z ions provides the mass and identify the amino acid residue and thus allows sequencing of the selected peptide. B: ETD spectrum of the DILAAQNLTGAVILNK peptide: most of the peaks have been assigned at the “c” or “z” series. C: ETD spectrum of the glycopeptide.	183
Figure 5.20 Separation of <i>C.difficile</i> Ox247 on an Immobilon PVDF transfer membrane and visualized after transfer by Ponceau-S red staining. Molecular size markers are indicated on the left.	184
Figure 5.21 MS spectra of WT SLP <i>C.difficile</i> Ox247 from m/z 800 to 2000 and from m/z 2000 to 4000. Clearly, these spectra show the presence of a series of peaks corresponding to the mass differences of monosaccharide residues, specifically hexoses, deoxyhexoses and pentoses, belonging to the SLP of <i>C.difficile</i> Ox247.	185
Figure 5.22 MS/MS spectra of m/z 2452 and 3512 belonging to WT SLP <i>C.difficile</i> Ox247.	186

Figure 5.23 MS/MS spectrum m/z 3339.....	186
Figure 5.24 MS spectrum from m/z 5000 to 10000 of the WT strain provides very strong evidence for the presence of a long glycan chain of almost fifty sugars belonging to the S-layer protein of <i>C.difficile</i> Ox247.....	187
Figure 5.25 MS/MS spectra of respectively m/z 6744 and 7632 coming from the MS showed in Figure 5.24	188
Figure 5.26 Cleavage between the Rha and the Glc residues within the glycan chain under study. This cleavage may be facilitated as a result of the α or β configuration of the anomeric carbon affecting the relative position in space of the oxygen and hydrogen atoms involved in the Hydrogen transfer.	189
Figure 5.27 MS/MS spectrum of m/z 6744 at the high end on the mass spectrum and showing the unusual fragmentation pattern.	190
Figure 5.28 MS/MS spectrum of m/z 7632 at the high end on the mass spectrum and showing the unusual fragmentation pattern.	190
Figure 5.29 MS spectrum from m/z 5000 to 10000 of <i>C.difficile</i> WT SLP after β -elimination and deuteromethylation procedures have been performed. This MS spectrum of the WT strain confirms the presence of a long and fragile glycan chain of up to more than forty sugars belonging to the S-layer protein of <i>C.difficile</i> Ox247.....	191
Figure 5.30 MS/MS spectra of m/z 6991 and 7746 of <i>C.difficile</i> WT SLP after β -elimination and deuteromethylation using CDI_3 procedures.	191
Figure 5.31 GC-MS chromatogram of alditol acetate derivatised glycans of <i>C.difficile</i> Ox247 WT SLP. The glycan composition of the <i>C.difficile</i> Ox247 S-layer (top panel) has been run alongside commercially obtained standards, starting from the top ribose, rhamnose, glucose and galactose respectively.....	192
Figure 5.32 Mass spectrometric fingerprint of alditol acetate rhamnose. The fingerprint is from the GC-MS chromatogram of the alditol acetate commercially obtained standard for rhamnose moiety (top panel) and that of the <i>C.difficile</i> Ox247 WT SLP (bottom panel).....	193
Figure 5.33 Mass spectrometric fingerprint of alditol acetate ribose. The fingerprint is from the GC-MS chromatogram of the alditol acetate commercially obtained standard for ribose moiety (top panel) and that of the <i>C.difficile</i> Ox247 WT SLP (bottom panel).	194
Figure 5.34 Mass spectrometric fingerprint of alditol acetate galactose. The fingerprint is from the GC-MS chromatogram of the alditol acetate commercially obtained standard for galactose moiety (top panel) and that of the <i>C.difficile</i> Ox247 WT SLP (bottom panel).....	194

Figure 5.35 Mass spectrometric fingerprint of alditol acetate glucose. The fingerprint is from the GC-MS chromatogram of the alditol acetate commercially obtained standard for glucose moiety (top panel) and that of the <i>C.difficile</i> Ox247 WT SLP (bottom panel).	195
Figure 5.36 GC-MS linkage analysis of <i>C.difficile</i> Ox247 WT SLP. The samples were run on a Bruker SCION SQ 456-GC fitted with a br-5ms column. The annotation has been completed using the mass spectrometric fingerprints and fragmentation pathways of the monosaccharide structures.	196
Figure 5.37 GC-MS chromatogram of linkage analysis of <i>C.difficile</i> Ox247 WT SLP (top panel), mass spectrometric fingerprint of a terminal ribose (middle panel) and mass spectrum of a terminal ribose available on the CCRC website (bottom panel).	196
Figure 5.38 GC-MS chromatogram of linkage analysis of <i>C.difficile</i> Ox247 WT SLP (top panel), mass spectrometric fingerprint of a 4-linked rhamnose (middle panel) and mass spectrum of a 4-linked rhamnose available on the CCRC website (bottom panel).	197
Figure 5.39 GC-MS chromatogram of linkage analysis of <i>C.difficile</i> Ox247 WT SLP (top panel), mass spectrometric fingerprint of a 3-linked rhamnose (middle panel) and mass spectrum of a 3-linked rhamnose available on the CCRC website (bottom panel).	198
Figure 5.40 GC-MS chromatogram of linkage analysis of <i>C.difficile</i> Ox247 WT SLP (top panel), mass spectrometric fingerprint of a 3,4-linked rhamnose (middle panel) and mass spectrum of a 3,4-linked rhamnose available on the CCRC website (bottom panel).	199
Figure 5.41 GC-MS chromatogram of linkage analysis of <i>C.difficile</i> Ox247 WT SLP (top panel), mass spectrometric fingerprint of a 4-linked glucose (middle panel) and mass spectrum of a 4-linked glucose available on the CCRC website (bottom panel).	200
Figure 5.42 Suggested NMR structure describing sugar and linkage compositions of the O-glycan chain attached to the LMW SLP of <i>C.difficile</i> Ox247 of interest (Logan and Vinogradov personal communication). * identity of peptide unknown.	201
Figure 5.43 A: Suggested glycan structure from Q-STAR data shown in Figure 5.6. B: Suggested NMR structure. ●:Galactose; ►:Rhamnose; ●: Glucose; ○: Hexose; ▽: deoxyHexose; ☆:Pentose.	202
Figure 5.44 MS spectrum of the S-layer of <i>C.difficile</i> Ox247 obtained using Xevo G2 Q-TOF. This data set shows a clear signal at m/z 1317 ²⁺ corresponding to ribosylation at Rha-4, and at m/z 1390 ²⁺ , corresponding to cleavage at the next rhamnose.	204
Figure 5.45 MS spectrum of the S-layer of <i>C.difficile</i> Ox247 sample after proteinase K digestion. This data set is consistent with the data shown in Figure 5.24 and a series of peaks	

corresponding to the mass differences of monosaccharide residues, specifically hexoses, deoxyhexoses and pentoses, are seen all along the spectrum.204

Figure 5.46 MS/MS spectrum of the m/z 845 peak seen in the MS spectrum shown in **Figure 5.44**. In this MS/MS, there is a clear loss of 126 from the molecular ion suggesting that the phosphate group has been dimethylated in the permethylation step. There are also further cross ring-type fragment ions in between the loss of 126 and the loss of 282, which corresponds to the loss of the complete capping residue.205

Figure 5.47 MS/MS spectrum of m/z 878, which can be correlated with the MS/MS at m/z 845 (**Figure 5.46**). There is a clear loss of 132 from the molecular ion (878-132=746) suggesting that the phosphate group is not naturally dimethylated or monomethylated, but the dimethyl groups have been added during the deuteropermethylation step.205

Figure 5.48 Various possible ring substitution positions for the dimethyl phosphate attached to the deoxyHex ring structure and the following loss of the dimethyl phosphate fragment (126) to produce m/z 719 and the subsequent ring cleavage to produce m/z 647. Considering the mechanism shown, only position 2 will allow the mechanistic formation of a m/z 647 ion.206

Figure 5.49 Considering a different mechanism, the dimethyl phosphate group can be attached to either carbon-2 or carbon-3 to allow the mechanistic formation of a m/z 647 ion.207

Figure 5.50 ES-MS spectrum in negative ion cone voltage fragmentation mode (spectrum from our Canadian collaborators).208

Figure 5.51 MALDI spectrum of HF digestion of the proteinase K digested S-layer glycopeptide followed by permethylation. The insert is a zooming-in of the main spectrum and specifically from m/z 5000 to 7500.210

Figure 5.52 MALDI MS spectrum in negative ion mode from m/z 5000 to 10000.210

Figure 5.53 The overall predicted structure of the O-link glycopeptide decorating the S-layer of *Clostridium difficile* Ox247.211

Figure 5.54 Q-STAR MS spectrum of the S-layer of *Clostridium difficile* Ox247 showing weak data around the 2000 position in the m/z scale.212

List of Tables

Table 1.1 Amino acid residues, compositions, structures and monoisotopic masses.	72
Table 1.2 Monosaccharide residues, compositions, structures and masses.	75
Table 2.1 Masses and sequences of the peptides contained in the calibration peptide mix used for MALDI-MS and MS/MS.	94
Table 3.1 <i>in-silico</i> Tryptic Map for the 1,427 residue ADAMTS13 protein (J = Cmc = carboxymethylcysteine).	112
Table 4.1 Strains of <i>C.difficile</i> studied in Professor Wren’s laboratories.	142
Table 4.2 <i>C.difficile</i> flagellin sequence <i>in-silico</i> tryptic digest.	144
Table 4.3 Possible atomic compositions for the accurate masses observed: 175.1192, 204.0873 and 288.2034. The compositions highlighted in yellow are the one being accepted.	155
Table 4.4 Atomic compositions assigned for key signals in the High Resolution MS/MS data. This table shows the deduced atomic compositions for the experimental and measured masses observed for key signals in the High Resolution MS/MS data (Figure 4.11) in association with the theoretical masses of those compositions. Crucial discoveries from these data comprised the finding of sulphur in the m/z 152 and higher mass fragments, thus confirming a novel structural entity and allowing the interpretation of a clear fragmentation pathway between the m/z 396 and 152 ions.	155

Abbreviations

Abs: Antibodies

ADAMTS13: A Disintegrin and Metalloproteinase with a Thrombospondin type 1 motif, member 13

API: Atmospheric Pressure Ionisation

Bac: Bacillosamine

CAD: Collisionally Activated Dissociation

CCRC: Complex Carbohydrate Research Centre

CDAD: *Clostridium difficile* Associated Disease

CDI: *Clostridium difficile* Infection

CPS: Capsular Polysaccharide

CI: Chemical Ionisation

CID: Collision-Induced Dissociation

CWP: Cell Wall Proteins

CZE: Capillary Zone Electrophoresis

Da: Daltons

DABP: Diaminobenzophenone

DATDH: 2,4-diacetoamido-2,4,6-trideoxyhexose

deoxyHex: deoxyhexose (dHex)

DC: Direct Current

DDA: Data Direct Analysis

DHB: 2,5-dihydroxy benzoic acid

DTT: Dithiothreitol

ECD: Electron Capture Dissociation

EI-MS: Electron Impact Mass Spectrometry

ER: Endoplasmic Reticulum

ESI-MS: Electrospray Ionisation

ETD: Electron Transfer Dissociation

FAB-MS: Fast Atom Bombardment Mass Spectrometry

FD: Field Desorption

FDA: Food and Drug Administration

GC-MS: Gas Chromatography Mass Spectrometry

GalNAC: N-acetylgalactosamine

GlcNAC: N-acetylglucosamine

HABA: 2-(4-hydroxy benezo) benzoic acid

HCCA: α -cyano-4hydroxy cinnamic acid

Hep: Heptose

Hex: Hexose

HMW: High Molecular Weight

HPLC: High Performance Liquid Chromatography

IAA: Iodoacetic acid

IDA: Automatic Information-dependent Acquisition

IS: Insertion Sequences

Kdo: 3-deoxy-D-manno-oct-2-ulosonic acid

LC: Liquid Chromatography

Leg: Legionaminic acid

LLO: Lipid-linked Oligosaccharide

LMW: Low Molecular Weight

LPS: Lipopolysaccharide

LTA: Lipoteichoic Acid

m/z: Mass to charge ratio

MALDI-MS: Matrix Assisted Laser Desorption Ionisation Mass Spectrometry

m-NBA: m-Nitrobenzyl Alcohol

MS: Mass Spectrometry

MS/MS: Tandem Mass Spectrometry

MurNAc: N-acetylmuramic acid

MWCO: Molecular Weight Cut Off

NMR: Nuclear Magnetic Resonance

N-OTase: N-Oligosaccharyltransferase

ORF: Opening Reading Frame

O-OTase: O-Oligosaccharyltransferase

PD-MS: Plasma Desorption Mass Spectrometry

PEG: Polyethylene Glycol

Pent: Pentose

PG: Peptidoglycan

PMAA: Partially Methylated Alditol Acetate

PMC: Pseudomembranous Colitis

ppm: parts per million

Pse: Pseudaminic acid

PTM: Post-Translational Modification

PVDF membrane: Polyvinylidene difluoride membrane

QIT: Quadrupole Ion Trap

Q-TOF: Quadrupole Orthogonal Acceleration Time of Flight Mass Spectrometer

RF: Radio Frequency

SA: Sinapinic Acid

SDS-PAGE: Sodium Dodecyl Sulphate PolyAcrylamide Gel

Ser: Serine

S-layer: Surface Layer

SLC: S-layer Cassette

SLPs: S-layer Proteins

TFA: Trifluoroacetic Acid

Thr: Threonine

TIC: Total Ion Current

TIS: Timed-Ion-Selector

TOF: Time-of-flight

TTP: Thrombotic Thrombocytopenic Purpura

UDP-GlcNAc: Uridine Diphosphate-activated N-acetylglucosamine

VWF: von Willebrand Factor

WGS: Whole-genome Sequencing

WT: wild type

XIC: Extracting Ion Current

Chapter 1:
Introduction

1. Introduction

1.1 Overview of Structural Glycobiology

At the beginning of the 20th century carbohydrates were mainly seen as a source of energy or as structural materials without any other biological activities. Since that time things have changed, and the field of glycobiology now encompasses a wide range of disciplines and occupies an important position within the scientific community, in research areas such as the enzymology of glycan formation and degradation, carbohydrate chemistry and biochemistry, the recognition of glycan epitopes by specific protein receptors, and in general making a significant impact to many areas of basic research in biotechnology and biomedicine.

Glycans attached to both proteins and lipids are involved in different ways in the cell life-cycle, having more than one main function. In fact they can be part of structural components, like cell walls or the extracellular matrix, or modify protein physico-chemical properties or be responsible for a wide range of events between cell-cell, cell-matrix and cell-molecule interactions, crucial to the growth and function of complex multicellular organisms (Varki 2009).

In contrast to the biosynthesis of proteins, glycan structures are the products of transcription and translation of genes that encode glycosyltransferases, which are the main machinery for the synthesis and assembly of glycan chains. The glycan repertoire of a cell comprises a subset of the entire glycome that the organism is capable of making, and in addition it depends not only on which genes have been transcribed and which transcripts have been translated, but also may change depending on the state of differentiation and the physiological environment. All of this helps to generate biological diversity and complexity of apparent benefit to the organism.

In nature, different types of post-translational modification exist, but protein glycosylation is surely the most abundant and the most diverse, with more than two-thirds of all eukaryotic proteins predicted to be glycosylated (Apweiler, Hermjakob et al. 1999). Glycosylation occurs in all domains of life and plays a fundamental role in prokaryotes as well as in eukaryotic organisms (Abu-Qarn, Eichler et al. 2008, Dell, Galadari et al. 2010). Two wide classes of protein glycosylation exist: (a) “N-linked”, that consist of a sugar chain covalently attached to an asparagine side-chain primary amide group in the protein and this residue has to be part of the consensus peptide sequence Asn-X-Ser/Thr (where X can be any amino acid except proline), and (b) “O-linked”, characterised by attachment to the side-chain hydroxyl of

serine or threonine residues in the protein sequence where these residues do not necessarily form part of a formal consensus sequence. Moreover, two different mechanisms exist by which the glycans are transferred to proteins. The first mechanism requires the transfer of carbohydrates directly from nucleotide-activated sugars to acceptor proteins. Examples are protein O-glycosylation in the Golgi apparatus in eukaryotic cells (see **section 1.2.2**) and flagellin O-glycosylation in many bacterial species (see **section 1.3.2**). The second mechanism, instead, is characterised by the assembly of an oligosaccharide on a lipid carrier before being transferred *en bloc* to protein acceptors by an oligosaccharyltransferase (OTase) and this mechanism is expressed in the N-glycosylation pathway of eukaryotes (see **section 1.2.1**) and in the N-glycosylation system of *Campylobacter jejuni* (Faridmoayer, Fentabil et al. 2007).

Glycosylation in prokaryotes shows a greater diversity of compositions and structures compared with that found in eukaryotic cells. In fact prokaryotic glycan chains present a major challenge for structural characterisation and are often composed of novel entities rarely seen elsewhere in nature.

Currently, particular attention has been paid to bacterial glycosylation, primarily because of the discovery of this type of post-translational modification in pathogenic species (Szymanski and Wren 2005). The precise role of glycosylation in bacteria and the reason why only selected proteins present glycosylation is still unknown due to the problem in obtaining purified glycoproteins and an incomplete knowledge of the biosynthetic pathways involved. Over the past twenty years, several research studies have served to better decipher the entire N-glycosylation pathway of the intestinal pathogen *Campylobacter jejuni* and to demonstrate the presence of O-glycosylation in many other bacterial pathogens such as *Helicobacter pylori* and *Neisseria gonorrhoeae* (Szymanski, Logan et al. 2003, Szymanski and Wren 2005, Weerapana and Imperiali 2006). Most of these pathogens are responsible for the biosynthesis of new and rare glycoprotein sugars not present in humans, making them possible novel therapeutic targets (Apweiler, Hermjakob et al. 1999).

In the next section, eukaryotic glycosylation and its peculiar characteristics are discussed first to better highlight similarities and differences compared with prokaryotic glycosylation.

1.2 Glycosylation in Eukaryotes

1.2.1 N-glycosylation in Eukaryotes

In eukaryotic glycoproteins N-linked glycans are covalently attached to the amide nitrogen of an asparagine (Asn) residue by an N-glycosidic bond, in which the most common N-glycan linkage is N-acetylglucosamine to asparagine (GlcNAc β 1-Asn) (**Figure 1.1**).

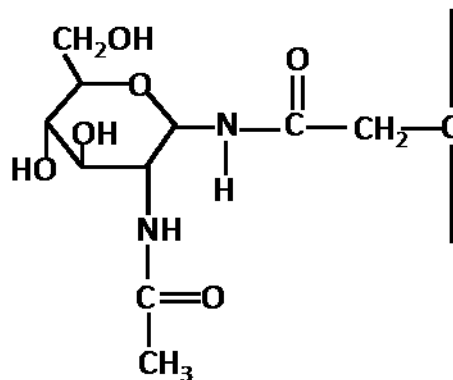


Figure 1.1 Attachment of an N-linked glycan, specifically an N-acetylglucosamine, to an asparagine residue.

Not all the asparagine residues in a peptide backbone can accept an N-glycan, but to do so they must fulfil three conditions. Firstly, the presence of a minimal amino acid sequence (“consensus sequence”) starting with an asparagine and then followed by any amino acid except proline and ending with serine or threonine (Asn-X-Ser/Thr) (Marshall 1974). Secondly, the acceptor sequence must be located in an accessible position in the eventual 3D structure of the protein, because the polar nature of the sugars would not be compatible with the hydrophobic interior of proteins, and finally, the acceptor protein must be found in the proper intracellular compartment (Taylor 2011).

All N-glycans comprise a common core (trimannosyl core) constituting three mannoses and two N-acetylglucosamines (Man α 1-6(Man α 1-3)Man β 1-4GlcNAc β 1-4GlcNAc β 1-Asn-X-Ser/Thr) (**Figure 1.2**) and the overall biosynthetic products can be classified in three main types:

1. High mannose, where only mannose residues are attached to the core sugar sequence;
2. Complex N-glycans, where antennae initiated by N-acetylglucosaminyltransferases (GlcNAcTs) are attached to the core;

- Hybrid N-glycans, in which only mannoses are linked to the Man α 1-6 arm of the core and one or two complex antennae are on the Man α 1-3 arm.

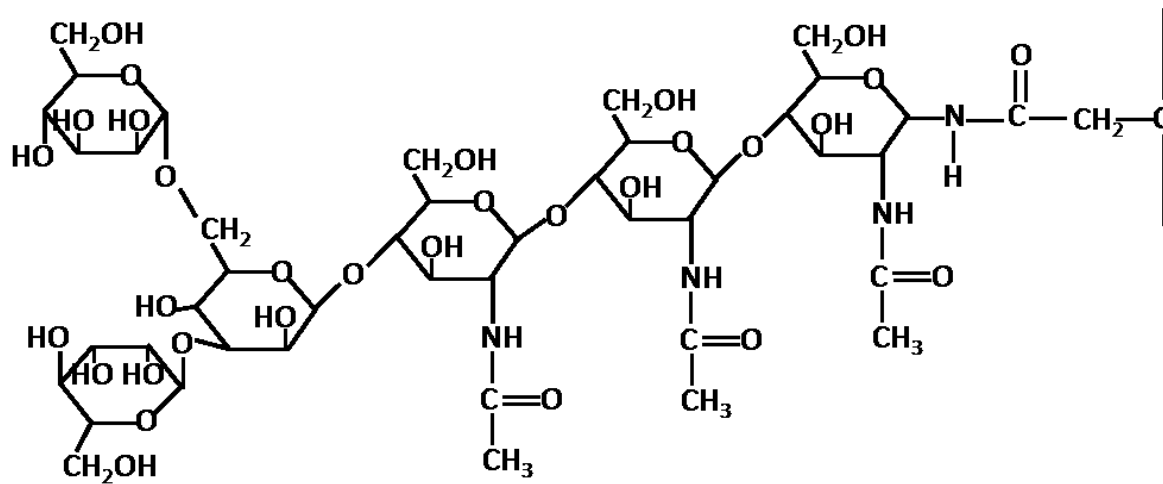


Figure 1.2 The common core structure of N-linked glycans comprising three mannoses and two N-acetylglucosamines attached to an Asn residue within the consensus sequence Asn-X-Ser/Thr (Man α 1-6(Man α 1-3)Man β 1-4GlcNAc β 1-4GlcNAc β 1-Asn).

Several enzymatic reactions occurring on both sides of the endoplasmic reticulum (ER) membrane allow the biosynthesis of a lipid-linked oligosaccharide and its transfer to selected asparagine residues of nascent polypeptide chains. The initial reactions take place on the cytoplasmic face of the ER membrane starting with the transfer of GlcNAc-P from UDP-GlcNAc to the lipid-like precursor dolichol-phosphate (Dol-P) forming N-Acetylglucosamine-pyrophospho-dolichol (GlcNAc-PP-Dol) via the enzyme GlcNAc-1-phosphotransferase (Helenius and Aebi 2002). The structure of dolichol is the result of its synthesis by condensation of five-carbon isoprene units assembled end-to-end and not cyclised. The number of these isoprene units in dolichol may vary within cells and between cell types and organisms and may have up to 19 isoprene units, meaning that the hydrophobic portion of this lipid is much longer than the fatty acid tails on membrane phospholipids even though it is part of the lipid bilayer in a helical or folded conformation (Varki 2009, Taylor 2011). Afterwards a second GlcNAc and five mannoses are transferred from UDP-GlcNAc and GDP-Man respectively to form the heptasaccharide intermediate, Man₅GlcNAc₂-PP-Dol, on the cytoplasmic side of the ER. Then the “flipping” of this intermediate across the ER membrane bilayer allows the exposure of the glycans to the lumen of the ER (Helenius and Aebi 2002). Once on the luminal face of the ER, the Man₅GlcNAc₂-PP-Dol is extended with another four mannoses transferred from Dol-P-Man and finally, the addition of further three

glucose residues donated from Dol-P-Glc complete the Dol-PP-glycan, with a final structure corresponding to $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2\text{-PP-Dol}$. The final step corresponds to the transfer of the pre-assembled tetradecasaccharide *en bloc* from the glycolipid to the consensus sequon Asn-X-Ser/Thr of a secretory protein that has been synthesised and translocated across the ER membrane. The responsible unit for catalysing this transfer is a multisubunit complex termed oligosaccharyltransferase (OTase) and is associated with components of the endoplasmic reticulum membrane. The OTase complex binds to the 14-sugar glycan and, cleaving the high-energy GlcNAc-P bond, transfers the glycan to the nascent protein with release of Dol-PP in the process (Silberstein and Gilmore 1996, Weerapana and Imperiali 2006). This enzyme complex has been largely investigated in yeast *Saccharomyces cerevisiae* and includes at least eight different membrane-bound protein subunits, divided in three sub-complexes. One of these, Stt3p, is a transmembrane protein, found in all eukaryotic organisms and involved in the catalytic process. Moreover all homologues of Stt3p in organisms that present an N-linked glycosylation system contain a highly conserved amino acid sequence, WWDYG. Mutations within this motif are implicated with the loss of glycosylation activity, demonstrating its essential role during catalysis (Weerapana and Imperiali 2006).

After the transfer to the polypeptide, a cascade of reactions trims the N-glycan chain in the ER. The first reactions consist of sequential removal of glucose residues by α -glucosidases I and II in the lumen of the ER, acting on the terminal α 1-2Glc and removing the two inner α 1-3Glc residues respectively. Subsequently an ER α -mannosidase I removes the terminal α 1-2Man from the central arm of $\text{Man}_9\text{GlcNAc}_2$ to give a $\text{Man}_8\text{GlcNAc}_2$ isomer. Then a second α -mannosidase I-like protein, called EDEM (ER Degradation-Enhancing α -mannosidase I-like protein) recognises possible misfolded glycoproteins and thus targets them for ER degradation. Further trimming in the *cis*-Golgi yields $\text{Man}_5\text{GlcNAc}_2$, which corresponds to a key intermediate for hybrid and complex N-glycans. However, not all N-glycans are fully trimmed, therefore resulting in between five and nine mannose residues which are called high mannose N-glycans (Varki 2009).

Alternatively, hybrid and complex N-glycans biosynthesis is initiated in the medial-Golgi thanks to the action of an N-acetylglucosaminyltransferase, GlcNAcT-1, responsible for add a GlcNAc to the carbon in position 2 of the mannose α 1-3 in the core $\text{Man}_5\text{GlcNAc}_2$. At this point the N-glycan chain can reach two possible destinies: either an α -mannosidase II can remove two mannose residues to form $\text{GlcNAcMan}_3\text{GlcNAc}_2$ and then a second GlcNAc is added to C-2 of the mannose α 1-6 in the core via the GlcNAcT-II to produce the precursor of

all biantennary complex N-glycans, or the GlcNAcMan₅GlcNAc₂ glycan is not acted on by α -mannosidase II and thus forms the hybrid N-glycans. Considering complex N-glycans, they can comprise more than two antennae and tri- and tetra-antennary glycans are the result of the catalytic activity of two N-acetylglucosaminyltransferases, GlcNAcT-IV and GlcNAcT-V, which add at C-4 of the α 1-3 core mannose and at C-6 of the α 1-6 mannose respectively. Moreover both complex and hybrid N-glycans may be characterised by the presence of a “bisecting” GlcNAc residue attached to the β -mannose of the core, a reaction catalysed by the GlcNAcT-III transferase. However the complexity of an extensive N-glycan array in terms of branch number, composition, length, capping arrangements and core modifications, is the result of supplementary sugar addition, mostly occurring in the trans-Golgi compartment, a process that can be divided into three steps: i) sugar addition to the core; ii) elongation of branching N-acetylglucosamine residues; iii) “capping” of elongated branches. These three components may present differences between vertebrates and invertebrates. In fact, regarding the core modification, vertebrates mainly modify the core with a fucose α 1-6 linked to the GlcNAc attached to the asparagine; instead in invertebrates the core may be modify with fucose, in both α 1-6 and α 1-3 linkages on the two N-acetylglucosamines, and also with a xylose in β 1-2 linkage to the β -mannose of the core (Schiller, Hykollari et al. 2012). Most of the complex and hybrid N-glycans present branches made by the addition of a β -linked galactose residue to the initiating N-acetylglucosamine producing in this way the building block Gal β 1-4GlcNAc, named “LacNAc”. Rarely, β -linked N-acetylglucosamine is extended with an N-acetylgalactosamine, forming in this case antennae with a GalNAc β 1-4GlcNAc or “LacdiNAc” extension. Finally, α -linked sialic acid, fucose, galactose, GalNAc and sulphate are example of the main “capping” or “decorating” reactions to extend the antennae (Varki 2009).

It is important to highlight that a single glycoprotein having more than one Asn-X-Ser/Thr N-glycosylation consensus sequence may be decorated with different glycans and this is referred to as microheterogeneity. Moreover, glycoproteins with a common peptide chain but carrying different glycans are named glycoforms. The significance of the heterogeneity may be varied and there are many ways to look at this biological event. It can reflect a neutral process of evolution or it can be a consequence of a positive selection affecting the properties of some proteins. In addition glycosylation of cell surfaces might offer a less uniform target for pathogens or in contrast glycans can be seen as a means of attacking cells (Taylor 2011). In conclusion, in all eukaryotes it is possible to find a set of different N-glycans, ranging from high mannose structures to complex ones. Furthermore eukaryotic glycoproteins show

heterogeneity in the degree of branching, core fucosylation and bisecting GlcNAc residues, the presence or absence of polylactosamine extensions, and a great diversity of terminal epitopes. All of this is at the base of generating complex glycans, involved in recognition events in multicellular organisms.

1.2.2 O-glycosylation in Eukaryotes

O-glycosylation is a stepwise process, starting with the attachment of a monosaccharide to the acceptor serine or threonine. Differently from N-linked glycoproteins, there is not a consensus sequence, even though an ideal glycosylation substrate consists of the reactive residue surrounded on both sides by Ser, Thr, Pro, Ala and Gly residues without any specific order (Elhammer, Poorman et al. 1993). Most eukaryotic O-glycans are mucin-type, which are covalently α -linked via an N-acetylgalactosamine (GalNAc) moiety to the hydroxyl group of serine or threonine by an O-glycosidic bond. Nevertheless other classes of non-mucin O-glycans exist, including α -linked O-fucose, β -linked O-xylose, α -linked O-mannose, β -linked O-GlcNAc, α - or β -linked O-galactose and, α - or β -linked O-glucose (Spiro 2002, Varki 2009, Dell, Galadari et al. 2010) (**Figure 1.3**).

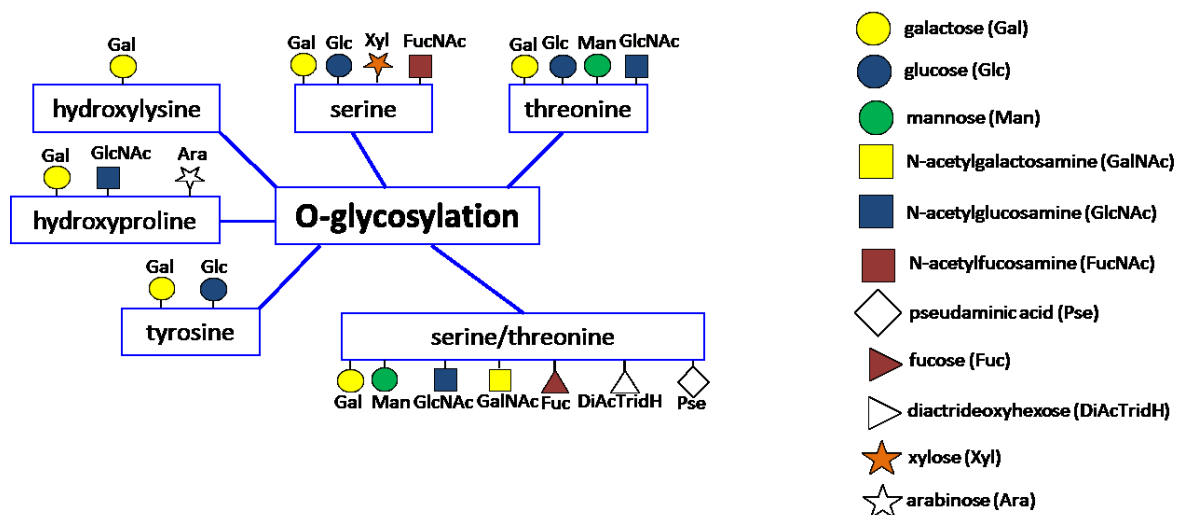


Figure 1.3 O-linked glycosylation in Eukaryotes. The diagram representation shows the variety of protein-glycan linkages used by Eukaryotes in O-linked glycosylation.

Mucins are high molecular weight ubiquitous glycoproteins in mucous secretions on cell surfaces, like surfaces of the gastrointestinal, genitourinary and respiratory tracts, and in body

fluids. They shelter the epithelial surfaces against physical and chemical damage and protect against infections by pathogens. Mucin glycosylation machinery uses glycosyltransferases analogous to those in the N-glycosylation pathway, albeit their organisation differentiates in two important ways. First of all, as mentioned above, one sugar at a time is added in a stepwise series of reactions and the glycan moiety starts with a GalNAc residue attached to a serine or threonine residue via an α -linkage: there is no *en bloc* transfer. Secondly, there are no target sequences for O-glycosylation analogous to the Asn-X-Ser/Thr sequences that define N-glycosylation sites. Compensating for the lack of a consensus glycosylation sequence is that there are numerous oligosaccharyltransferases that can attach GalNAc to the serine and threonine residues. The glycosyltransferases involved in the O-linked mechanism are distributed through the Golgi apparatus, like the enzymes that create the terminal elaborations of the N-linked structures. Therefore this distribution provides a parallel to the terminal modification of the N-glycans. This interesting parallelism suggests that O-glycans are in some way the terminal parts of N-linked sugars conjugated directly to proteins (Taylor 2011).

The GalNAc residue derives from UDP-GalNAc and this transfer is catalysed by a polypeptide-N-acetyl-galactosaminyltransferase (ppGalNAcT), of which there are at least 21 encoded by different genes. Once the first sugar has been attached, it creates the Tn antigen (GalNAc-Ser/Thr). The O-GalNAc glycan can be extended with different types of sugars including galactose, N-acetylglucosamine, fucose or sialic acid. There are eight common core structures, named core 1 through 8, most of which may be further substituted by other sugars. The most common core, found in many glycoproteins and mucins, is core 1 (Gal β 1-3GalNAc-), obtained by the transfer of a galactose by a core 1 β 1-3 galactosyltransferase. Core 1 is antigenic and it is called the T antigen. Both Tn and T antigens may be modified by sialic acid to form sialylated-Tn or -T antigens, respectively.

Core 2 is a common structure present in both glycoproteins and mucins of several cells and tissues, including the intestinal mucosa. It corresponds to O-glycan core 1 structure plus a branching N-acetylglucosamine. The enzyme responsible for core 2 synthesis is core 2 β 1-6 N-acetylglucosaminyltransferase.

Core 3 (GlcNAc β 1-3GalNAc) and its branched core 4 O-GalNAc glycans are typically present in secreted mucins from the gastrointestinal and respiratory tracts and salivary glands. All of the core structures may be modified and these modifications include O-acetylation of sialic acid and O-sulphation of galactose and N-acetylglucosamine. Therefore, mucin O-glycans are often very heterogenous, with hundreds of different chains (Varki 2009).

As mentioned earlier, there are other classes of O-glycans and one of these is the Fuc- α -Ser/Thr, primarily found in epidermal growth factor domains of multidomain proteins, i.e. coagulation and fibrinolytic factors (Harris 1993), but also in urokinase (Buko, Kentzer et al. 1991), human coagulation factors VII (Bjoern, Foster et al. 1991), IX (Nishimura, Takao et al. 1992) and XII (Harris, Ling et al. 1992). Fucose can be present in the mature glycoprotein in two ways, either alone or as the inner component of short oligosaccharide. Another example is GlcNAc- β -Ser/Thr which is widely dispersed among eukaryotes, from protozoa to higher mammals, and found in both nuclear and cytoskeletal proteins. Differently from other peptide-linked monosaccharides, the β -linked GlcNAc-Ser/Thr is not usually further substituted by other sugars, remaining a simple monosaccharide modification of the glycopeptide (Spiro 2002). Finally a Gal- α -Ser/Thr O-glycan was reported to be present in the cuticle collagens of the earthworm, *Lumbricus terrestris* (Muir and Lee 1970) and clamworm, *Nereis virens* (Spiro and Bhoyroo 1980), where it appears as di- and tri- α -linked galactose oligosaccharides. Single Gal residues α -linked to Ser are also reported to be present in the cell wall of *Phaseolus coccineus* (O'Neill and Selvendran 1980) and several higher plants (Lamport, Katona et al. 1973).

1.3 Glycosylation in Prokaryotes

Only after the work of Mescher and Strominger on surface layer (S-layer) glycoproteins of the gram-negative halophile *Halobacterium salinarium* (Mescher and Strominger 1976) and the work of Sleytr and Thorne which demonstrates the presence of glycoproteins in the cell walls of gram-positive bacteria (Sleytr and Thorne 1976), was the question of whether glycans are integral components of prokaryotic proteins convincingly approached (Messner 2004), with an ever-growing number of glycoproteins being identified (Hitchen and Dell 2006). In fact, prokaryotic glycosylation seems to lead to a greater diversity of glycan compositions and structures compared with that found in eukaryotic organisms (Abu-Qarn, Eichler et al. 2008). These glycans include several unusual sugars not found in vertebrates, like 3-deoxy-D-manno-oct-2-ulosonic acid (Kdo), heptose (Hep), the amino- and deoxy-monosaccharides, pseudoaminic acid, bacillosamine 2,4-diacetoamido-2,4,6-trideoxyhexose (DATDH) and modified hexoses, which are central in the biology and pathogenicity of bacterial cells, and may also in addition present other types of modifications, such as sulphation, acetylation or methylation.

Even though prokaryotic glycans are more complex than their eukaryotic counterparts and the topology of protein glycosylation in bacteria and archaea presents additional challenges, as there is an absence of the intracellular compartments fundamental to organise most protein glycosylation in eukaryotes, there are many similarities in the structures and in the biosynthetic pathways, in particular if referring to mucosal pathogens, indicating that these modifications might have analogous roles in multiple organisms (Szymanski and Wren 2005).

Over the last decade, the majority of reported prokaryotic glycosylation has been best demonstrated in S-layers, pilins, flagellins and a selection of cell surface and secreted proteins implicated at the functional level in adhesion and/or biofilm formation (Hitchen and Dell 2006, Dell, Galadari et al. 2010), protection against proteolytic cleavage, antigenic variation and protective immunity (Szymanski, Burr et al. 2002).

1.3.1 N-glycosylation in Bacteria

At the beginning of the 2000s, Szymanski and coworkers characterised the first N-linked glycosylation reported in bacteria and in particular they identified an N-glycan attached via the eukaryotic sequon Asn-X-Ser/Thr on multiple proteins in the Gram-negative bacterium *Campylobacter jejuni*, a human gut pathogen and one of the main causes of bacterial gastroenteritis worldwide (Szymanski, Yao et al. 1999, Young, Brisson et al. 2002, Szymanski and Wren 2005). As in the eukaryotic system, the N-linked protein glycosylation is characterised by a β -glycosylamide linkage to asparagine residue, but in this case, the glycan structure transferred is strikingly different between eukaryotic and bacterial systems (Burda and Aebi 1999, Wacker, Linton et al. 2002, Young, Brisson et al. 2002). The biosynthetic machinery follows a similar progression, whereby an oligosaccharide is assembled in a stepwise fashion on a polyisoprenyl-pyrophosphate carrier and then finally transferred to the protein (Weerapana and Imperiali 2006).

Several species of *Campylobacter* contain gene clusters encoding a glycosylation system able to modify various proteins, and the genes whose products modify a specific protein are located adjacent to the gene encoding the protein itself (Varki 2009). The gene locus involved in the biosynthesis of a number of highly immunogenic glycoproteins is termed the “*pgl* gene cluster” (Szymanski, Yao et al. 1999). It contains genes that present homologies to enzymes implicated in bacterial lipopolysaccharide (LPS) and capsular polysaccharide (CPS) biosynthesis. A computational analysis of the *pgl* cluster suggests that the locus encodes five

putative glycosyltransferases (PglA, PglC, PglH, PglI and PglJ) and three enzymes involved in sugar biosynthesis (PglD, PglE and PglF) (Linton, Dorrell et al. 2005, Weerapana and Imperiali 2006). Subsequent work by Linton and coworkers identified two highly immunoreactive glycoproteins in *C.jejuni*, PEB3 and CgpA, which are glycosylated by components of this pathway and these glycoproteins are able to bind the GalNAc-specific lectin, soybean agglutinin (SBA) (Linton, Allan et al. 2002). Using mass spectrometry and NMR spectroscopy, the N-linked glycan structure was elucidated and it corresponds to the heptasaccharide GlcGalNAc₅Bacβ1 N-Asn where Bac is bacillosamine, 2,4-diacetamido-2,4,6-trideoxyglucose (DATDH) (Wacker, Linton et al. 2002, Young, Brisson et al. 2002). Furthermore, it has been demonstrated that the structure is highly conserved throughout all *C.jejuni* and *C.coli* strains (Szymanski, Logan et al. 2003) and the *pgl* gene cluster can function in *E.coli* to glycosylate proteins, proving that the *pgl* cluster contains all of the genes necessary for the biosynthesis of the polyisoprenylpyrophosphate-linked heptasaccharide and its eventual transfer to protein (Wacker, Linton et al. 2002).

The key enzyme in the *pgl* gene cluster is PglB and it is the first example of a bacterial N-linked oligosaccharyltransferase. It demonstrates significant homology to the staurosporine- and temperature-sensitive yeast protein 3 (Stt3p) subunit of the N-linked oligosaccharyltransferase complex of *Saccharomyces cerevisiae* (Szymanski, Yao et al. 1999, Linton, Allan et al. 2002) (see **section 1.2.1**). In fact PglB contains a highly conserved carboxy-terminal catalytic motif (WWDYG) typical of all Stt3p orthologues. Moreover, this *pgl* gene cluster in *C.jejuni* also shows significant homology to a cluster found in the genome of *Neisseria meningitides*, known to be responsible for the O-linked glycosylation of pilin (Szymanski and Wren 2005).

Linton et al. carried out several mutational studies of the *pgl* gene cluster in *E.coli* to clarify the exact roles of various *pgl* genes with the help of structural analysis, including mass spectrometric studies in this laboratory, of the glycan transferred to the protein (Linton, Dorrell et al. 2005). They demonstrated that the *pglA*, *pglJ*, *pglH* and *pglI* genes encode specific glycosyltransferases responsible for sequential addition of monosaccharides to form the ultimate heptasaccharide donor in the cytosol. The N-glycan biosynthesis starts with uridine diphosphate (UDP)-activated N-acetylglucosamine (UDP-GlcNAc). The conversion of UDP-GlcNAc to bacillosamine occurs via sequential modifications by PglF (dehydratase), PglE (aminotransferase) and PglD (acetyltransferase) in the cytoplasm. Then PglC, the first glycosyltransferase, attaches the bacillosamine residue to a lipid carrier. Successively PglA adds the α-1,3-linked GalNAc moiety to bacillosamine and both PglH and PglJ are involved

in transferring the next four α -1,4-linked GalNAc moieties. Finally, PglI is the GTase responsible of the branching glucose moiety. Once the heptasaccharide has been assembled, the entire chain is flipped across the cytosolic side of the bacterial inner membrane toward the periplasm by an ATP-dependent flippase PglK. At this point PglB is the only protein necessary to transfer *en bloc* the glycan to the common N-linked sequon and there is no evidence for further processing of the heptasaccharide (Szymanski and Wren 2005). PglB also releases the heptasaccharide as free oligosaccharides into the periplasmic space and this hydrolase activity is influenced by the osmotic environment of the cell (Nothaft, Liu et al. 2009). In the *C.jejuni* the eukaryotic sequon is N-terminally extended to Asp/Glu-X₁-Asn-X₂-Ser/Thr, where X₁ and X₂ can be any amino acid except proline (Kowarik, Young et al. 2006) and even though D-Q-N-A-T is the optimal bacterial acceptor sequence, not all sequons are glycosylated (Chen, Glover et al. 2007). The N-glycosylation pathway in *C.jejuni* can modify several proteins and disruption of this pathway has pleiotropic effects for the bacterium, embracing reduced protein immunoreactivity with both human and animal sera (Szymanski, Yao et al. 1999, Szymanski and Wren 2005), a reduced ability to adhere and to invade human epithelial cells in vitro and a decrease in mouse and chicken colonisation in vivo (Szymanski, Burr et al. 2002, Hendrixson and DiRita 2004, Jones, Marston et al. 2004, Nothaft, Liu et al. 2009).

Even though the best studied N-glycosylation pathway remains the one in *Campylobacter jejuni*, another N-glycosylation system has been described in the Gram-negative gammaproteobacterium *Haemophilus influenza* (Gross, Grass et al. 2008). In 2003 St Geme and coworkers highlighted that one of the two *H.influenza* high-molecular-weight adhesins HMW1, which interacts with sialylated N-glycoproteins on the human epithelial cells (St Geme 1994), was modified with mono- or dihexoses on 31 asparagine residues and all but one of the asparagines is within the conventional sequon Asn-X-Ser/Thr (Grass, Buscher et al. 2003, Gross, Grass et al. 2008). The glycosylation is essential for HMW1 stability and translocation to the bacterial surface and it is carried out by a unique GTase, HMW1C. In a following study (Grass, Lichti et al. 2010), it has been demonstrated that HMW1C has substrate specificity for UDP- α -D-glucose and UDP- α -D-galactose, but the first hexose to be linked to asparagine containing dihexoses must be the glucose. As HMW1C orthologs can be found in other Gram-negative pathogens, like enterotoxigenic *E.coli*, *Yersinia pseudotuberculosis*, *Y.enterocolitica*, *Y. pestis*, *Burkholderia* spp and many others, these proteins may constitute a new family of bacterial enzymes capable of forming N-linked glycan-protein and glycan-glycan bonds in the cytoplasm without assembling

monosaccharide units on a lipid-linked intermediate (Gross, Grass et al. 2008). Moreover, because HMW1C does not share any sequence identity with PglB and because these two enzymes are a member of two structurally unrelated GTase families, these N-glycosylation systems may have evolved separately in two genera within two bacterial classes and in different compartments of the cell (Alemka, Nothaft et al. 2013).

1.3.2 O-glycosylation in Bacteria

As in the case of eukaryotic O-glycosylation (see **section 1.2.2**), bacterial O-glycosylation also does not seem to have a consensus sequence for the serine or threonine residues to which carbohydrate is attached. In bacteria, O-glycosylation is more widespread than N-glycosylation and a panoply of O-glycosylation mechanisms have been identified from the most diverse genera, including several important human pathogens. In particular, two classes of protein glycosylation systems are the most representative. One is organised based on the utilisation of an OTase, which catalyses the *en bloc* transfer of an oligosaccharide from a lipid donor to an acceptor molecule, usually a protein, and it is named OTase-dependent. The other one, instead, is OTase-independent, characterised by a sequential transfer of an individual monosaccharide to a protein acceptor (Iwashkiw, Vozza et al. 2013).

The degree of glycosylation is highly variable, as are the sugar compositions. The predominant O-glycans attached to the *Campylobacter* flagellum are derivatives of pseudaminic acid (Pse) or legionaminic acid (Leg) which are 9 carbon sugars related to sialic acid (Nothaft and Szymanski 2010), whereas *Helicobacter pylori* and *Pseudomonas aeruginosa* that O-glycosylate their flagellin with just a single pseudaminic acid (Castric, Cassels et al. 2001, Schirm, Soo et al. 2003). In contrast the pili of *Neisseria meningitides* and *Neisseria gonorrhoeae* contain serine O-glycans, the first of which was reported from this laboratory as Gal- β 1,3-Gal- α 1,3-2,4-diacetoamido-2,4,6-trideoxyhexose (DATDH) (Stimson, Virji et al. 1995), the same sugar necessary to attach the *C.jejuni* N-glycans to Asn (see **section 1.3.1**).

The first examples of OTase-dependent glycosylation are the O-glycosylation of the type IV major pili from *Pseudomonas aeruginosa* 1244 by the PilO system and the pilin protein in *Neisseria meningitides* by PglL more than 20 years ago (Castric 1995, Stimson, Virji et al. 1995). Research carried out in engineered *E.coli* has shown that the biosynthesis of the O-glycan presents similarities to its N-linked counterpart. The O-glycosylation pathways involve a lipid carrier (the so-called lipid-linked oligosaccharide or LLO) and the glycans are

transferred *en bloc* by the OTases from the LLOs carrier onto the protein (Faridmoayer, Fentabil et al. 2007). Furthermore, there is an evolutionary relationship between O-glycosylation in bacteria and the lipopolysaccharide (LPS) synthesis (Hug and Feldman 2011). For example, considering the pathogen *P.aeruginosa*, PilO catalyses the transfer of the glycan from its undecaprenol-pyrophosphate (Und-PP) carrier to the C-terminal serine 148 residue of *P.aeruginosa* 1244 pilin. The O-glycan [α -5N β OHC(4)7NFmPse-2,4 β -Xyl-1,3 β -FucNAc] is one of the products of the O antigen biosynthetic pathway and has the same structure as the O antigen of LPS (Castric, Cassels et al. 2001, DiGiandomenico, Matewish et al. 2002). After the lipid-linked glycan has been formed by subsequent addition of monosaccharides on the inner face of the plasma membrane, it flips to the periplasmic face by a flippase where it serves as a substrate for its respective enzyme. Both the OTase and the O-antigen ligase WaaL transfer a lipid-linked glycan from the periplasmic face of the inner membrane onto the hydroxyl groups on acceptor protein, in the case of the O-OTase, and to the lipid A-core in the case of WaaL (Hug and Feldman 2011). The glycans naturally transferred by PilO and PglL are short oligosaccharides (Castric 1995, Stimson, Virji et al. 1995, Castric, Cassels et al. 2001). DiGiandomenico has demonstrated that PilO presents relaxed glycan specificity, since different O antigens can be attached to pilin by heterologously expressing PilO in nonglycosylating *P.aeruginosa* strains (DiGiandomenico, Matewish et al. 2002), with a limit of maximum length of around 10 sugar residues (Faridmoayer, Fentabil et al. 2007).

Instead the OTase-independent system is primarily employed to decorate flagellar structural proteins and adhesins (Logan 2006, Dell, Galadari et al. 2010). In this system, “protein O-glycosyltransferases” (POGTases) mediate the linkage of individual monosaccharides to acceptor proteins at the cytoplasm-inner membrane interface and afterwards these are transported to the outer membrane or secreted by the flagellum. *Campylobacter* species are just one of the many bacteria using this system to modify their flagellar proteins with O-glycans (Nothaft and Szymanski 2010). These modifications are necessary for flagellum assembly and so affect secretion of virulence-modulating proteins, bacterial colonisation, auto agglutination and biofilm formation (Guerry 2007).

Another class of prokaryotic O-glycoproteins are the surface-layer (S-layer) proteins, identified over 40 years ago (Mescher and Strominger 1976, Sleytr and Thorne 1976), even though the “bacteria” studied then were later classified as archaea. It is now known that, for authentic bacteria, it is mainly Gram-positive organisms that have O-glycans attached on their S-layers, although a few examples of Gram-negative species have been recently

identified (Ristl, Steiner et al. 2011). The next section briefly reviews archaea glycosylation because research on these prokaryotes lays the foundations for studying S-layer glycosylation in bacteria.

1.3.3 N- and O-glycosylation in Archaea

Archaea is the third domain of life and they are single-cell bacteria-like organisms. They occupy “harsh” environments, typically characterised by high temperature or pressure (i.e. deep sea, thermal vents) or extremes in salinity, alkalinity or acidity. However, they also reside in “normal” biological niches, including seawater, soil and our intestinal flora (Varki 2009, Calo, Guan et al. 2011). The reason why they succeed in such habitats is the expression of proteins able to remain folded and functional in all those situations that would generally lead to protein denaturation, loss of solubility and aggregation (Eichler and Adams 2005). They also play fundamental roles in the ecosystem, for example in the Earth’s nitrogen cycle or in global warming (Francis, Beman et al. 2007, Galperin 2007), in technology, with methanogens used to provide biogas and as part of sewage treatment, and in biotechnology, in particular exploiting enzymes from extremophiles which are able to resist high temperature and organic solvents (Eichler and Adams 2005).

Archaea present features that distinguish them from both bacteria and eukarya. For example, their membranes contain unusual lipids comprising polyisoprenyl groups linked to glycerol and an S-layer of glycoproteins in a lattice-like arrangement directly attached to the membrane. Despite the fact that archaea and bacteria are size- and shape-similar, and their genome is organised as a single, circular chromosome (Koonin and Wolf 2008), archaea possess genes and several metabolic pathways that are more close to those in eukarya, like the enzymes involved in transcription and translation (Sandman 2000, Bell and Jackson 2001).

A detailed review of glycosylation in archaea is outside the scope of this thesis. Some common themes are exemplified by the examples given below.

In contrast to bacterial N-glycosylation, still believed to be a relatively rare event, N-glycosylation of archaeal proteins is more widespread and in the last decade impressive advances has been made in our understanding of several archaeal organisms, like halophiles, methanogens and thermoacidophiles, combining glycan structure information obtained by mass spectrometry with bioinformatic, genetic, biochemical and enzymatic data. Distinctive

hallmarks of archaeal N-glycosylation embrace the presence of unusual dolichol lipid carriers, a range of linking sugars as glycan constituents, two different N-linked glycans attached to the same protein and the ability to vary the N-glycan composition under various growth conditions (Jarrell, Ding et al. 2014).

Considering all the archaeal N-linked glycan structures determined, a wide diversity has been seen in terms of size, degree of branching, identity of the linking sugar, modification of sugar components by amino acid, sulphate and methyl groups and the presence of unique sugars. Several researchers have discovered the existence of glycoproteins in a large variety of archaea thanks to many different analytical procedures, like glycoprotein-specific staining techniques, lectin binding, deglycosylation experiments and study with glycosylation inhibitors (Jarrell, Ding et al. 2014).

As in eukaryotes, archaeal N-glycosylation occurs on the Asn-X-Ser/Thr consensus sequence and it has been suggested that the process responsible for N-glycosylation in eukaryotes derived from a simpler archaeal system, although it is also possible to find important differences between those two N-glycosylation pathways. For example, the archaeal N-linked oligosaccharide is assembled on a dolichol carrier, such as the eukaryotic one, but archaea include both dolichol phosphate (Dol-P) and dolichol pyrophosphate (Dol-PP) bearing either mono- or polysaccharides (Calo, Kaminski et al. 2010). Moreover, the linking sugar necessary to attach an oligosaccharide to the Asn residue also differs in glycoproteins across the two domains of life. In fact, as described in **section 1.2.1**, in the vast majority of eukaryotic N-glycoproteins, GlcNAc serves as the linking sugar (Spiro 1973). Instead, in archaea a variety of sugars have been shown to serve this role, such as glucose, GlcNAc and GalNAc and in addition, two diverse linking sugars can be involved within the same archaeal glycoprotein, as in the case of *Halobacterium salinarum* S-layer glycoprotein (Calo, Kaminski et al. 2010). In fact, the S-layer glycoprotein of *H.salinarum* contains three types of attached glycan, two N-linked and a third O-linked. Thus 15 Glc-Gal disaccharides are O-linked to a cluster of threonine residues at the C-terminus of the protein. In contrast one N-glycan is a repeating-unit pentasaccharide found only at one position (Asn-2 from the N terminus), while the other one is a sulphated oligosaccharide found at 10 positions scattered throughout the protein. The two N-glycan types are assembled on two different lipid carriers and rely on different linking sugars, GalNAc and GlcNAc respectively. Moreover, the repeating-unit glycan is only found on the S-layer glycoprotein; instead, the sulphated oligosaccharide can be also seen attached to archaeellins (Jarrell, Ding et al. 2014). The transfer of the repeating-unit to the protein is prevented by bacitracin, suggesting that the

lipid carrier required is Dol-PP (Sumper 1987). In contrast, the addition of the sulphated glycan is not inhibited by bacitracin, demonstrating the involvement of a different lipid carrier, the Dol-P (Wieland, Dompert et al. 1980), as mentioned above.

Considering another archaeal organism, *Haloferax volcanii*, a different and more complicated scenario has occurred. In fact, depending at which salt concentrations and in which laboratories the organism is cultured, three S-layer N-glycan structures have been described. When cultured in a medium with 3.5 M NaCl, Mengele and Sumper (Mengele and Sumper 1992) described S-layer N-glycans made by different sugars: Asn-13 and Asn-498 were described to be decorated with a linear polymer consisting of 10 β 1,4-linked glucose residues, while the oligosaccharide attached to Asn-274 and Asn-279 presented additional galactose and idose subunits. Recently, this initial description has been contradicted: in similar high-salinity conditions (3.4 M NaCl), a pentasaccharide composed of a hexose, two hexuronic acids, a methyl ester of a hexuronic acid and a terminal mannose (Man-MetHexA-HexA-HexA-Hex) decorate at least two Asn residues (Asn-13 and Asn-83) (Abu-Qarn, Yurist-Doutsch et al. 2007, Magidovich, Yurist-Doutsch et al. 2010). More recently, Guan and co-workers (Guan, Naparstek et al. 2012) have found a rhamnose Rha-Hex-Hex-sulphated Hex tetrasaccharide attached to Asn-498 of the S-layer glycoprotein and this position has not been modified at higher salinities. Furthermore, the pentasaccharide Man-MetHexA-HexA-HexA-Hex was still attached to Asn-13 and Asn-83 in cells cultured at lower salinity although to a smaller proportion of these residues comparing to growth at higher salt concentrations.

1.4 Mass spectrometry

Mass spectrometry is one of the most powerful of the analytical tools in structural biology and represents the current method of choice for the structural analysis of glycoconjugates, including glycoproteins, due to its ultra-high sensitivity, mass accuracy and ability to characterise individual components within complex, heterogeneous mixtures (Morris, Thompson et al. 1978). It involves the production of gas-phase ions from analytes, where the resulting ions are accelerated out of the ionisation source into a mass analyser, where they are separated according to their mass to charge ratios (m/z) and detected, using a charge amplification device as an electron- or photon-multiplier. In fact, the three key components of a mass spectrometer are the ionisation source, which, converts molecules into charged ions, a mass analyser that resolves the ions by their mass to charge ratio, and the ion detector that

detects and quantifies the resolved ions exiting the mass analyser. From the m/z of the ion, since the ion charge state is known, the mass of the original molecule can be calculated.

1.4.1 Historical background of mass spectrometry

In the early 20th century, the analytical technique of mass spectrometry was used only in physics laboratories to measure the masses of atoms and one of its first contributions to science was to prove the existence of isotopes. In the prestigious Cambridge's Cavendish laboratory, Sir Joseph John Thomson was the first person to demonstrate the mass to charge ratio m/e of the electron in 1906 using his "positive ray parabole" and thus becoming the first person to obtain a mass spectrum. Afterwards, together with his student Francis W. Aston, Thomson built what later would be recognised as the first mass spectrometer, able to measure the masses of charged atoms, leading to the discovery that many elements, such as neon and chlorine, are made of different isotopes (Griffiths 2008).

However, it was the importance of isotopes to the Manhattan Project and World War II that really promoted MS into prominence as a useful tool. In fact, mass spectrometry was applied to isolate and study the radioactive U-235 isotope for nuclear weapon development (Nier 1947) and at the same time, Alfred Nier was the one that pushed this technique to people outside the tight community of physicists, helping, for example, chemists and biologists by preparing ¹³C-enriched carbon, and geochemists in determining the age of the earth by measuring ²⁰⁷Pb/²⁰⁶Pb in the planet's crust, among other achievements (De Laeter 2006, Griffiths 2008).

Successively, well before the proteomics era, two fundamental developments occurred in the mass spectrometry field. First of all, the early ionisation method, named electric discharge by Thomson, was replaced by the Electron Impact Ionisation (later called the Electron Ionisation or EI) (see **section 1.4.2 A**) during the post-war period, which permitted the detection of molecular ions as well as fragment ions allowing the employment of mass spectrometry in general chemical analysis, including its application to the study of new dyestuffs being manufactured by the chemical industry to meet the needs of the booming cotton and woollen industries of North England. Secondly, a gas chromatograph-mass spectrometer (GC-MS) combined instrument was developed and it provided the coupling of mass spectrometry and gas chromatography allowing the separation of volatile materials online with subsequent mass spectrometric detection avoiding the necessity to purify a sample in advance (Eneroth 1964, Gohlke 1993). However, this device could only be applied to volatile samples, and

therefore was not directly applicable to bio-chemicals such as peptides. Consequently, it was necessary to overcome this obstacle via their derivatisation. Two different methodologies were developed: the poly-amino alcohol derivatisation method for peptides using GC-MS instruments by Klaus Biemann (Nau 1973) and the one developed by Howard Morris, which combines the permethylation reaction together with the N-acetylation of amino groups to create N-acetylated permethyl peptide derivatives (Geddes, Graham et al. 1969, Morris, Williams et al. 1971). This last approach demonstrated the first successful application of mass spectrometry to protein sequencing and for analysing mixtures of peptides. The latter ability was particularly important since at that time only the analysis of single pure peptides was feasible by classical Edman methods, and the purification of individual peptides was the rate-limiting step in protein sequencing.

By the early 1980s, a new ionisation technique, Fast Atom Bombardment (FAB) (see **section 1.4.2 C**), was developed by Michael Barber at the University of Manchester Institute of Science and Technology (Barber, Bordoli et al. 1981), which quickly revolutionised the field of MS-based biopolymer analysis, allowing the analysis of native oligosaccharides and peptides without derivative formation (Morris, Panico et al. 1981). The FAB ionisation method has been generally replaced by softer ionisation techniques, such as Electrospray Ionisation (ESI) (Fenn, Mann et al. 1989) and Matrix-Assisted Laser Desorption Ionisation (MALDI) (Karas 1988) which permit the analysis of large biological molecules without major degradation.

Mass spectrometry has become the method of choice for establishing biopolymer structures and both ESI and MALDI ionisation methods have become the routine methods for this kind of analysis. These two soft ionisation methods have therefore been extensively applied in the Glycoproteomic analysis described in this thesis.

1.4.2 Ionisation techniques

Different types of ionisation modes are available and used depending on the type of compound to be analysed and the specific information required. Moreover, in mass spectrometry it is possible to combine different mass analysers with a different range of ionisation techniques.

In the 1960s the analysis of low molecular weight volatile molecules was carried out to define glycan structures using mass spectrometry. In particular, at that time two ionisation techniques were the most popular: Electron Impact (EI), also known as electron

bombardment, and Chemical Ionisation (CI). The first one, EI, is a “hard” ionisation technique used to identify small molecules up to 1000 Da by extensive energy input and formation of fragment ions, resulting in being unsuitable for analysis of molecular ions (Ando, Kon et al. 1977). On the other hand, CI is a “soft” ionisation technique: the ionisation is less energetic upon interaction of the sample with the reagent gas and the ions produced are molecular ions with little fragmentation, even though it uses a similar ion source to EI. Both these two ionisation techniques require the samples to be in the gaseous phase before ionisation, but if the samples are sufficiently volatile, for example if the polar groups are derivatised, these methods are very powerful for structure analysis.

In the late 1960s, Beeky developed an exciting new ionisation technique, Field Desorption (FD) ionisation, which for the first time allowed the study of non-volatile compounds without derivative formation (Winkler and Beeky 1972). However, this technique proved difficult to use and was only employed in a few laboratories although it was essential to the solution of several important structural problems of the time, including the correction of the structure of the antibiotic Echinomycin (Dell 1975). In 1976 the plasma-desorption in-source (PD-MS) was introduced by Macfarlane and Torgerson (Macfarlane and Torgerson 1976), which permitted the ionisation of biomolecules of up to 20000 Da and was used for the mass determination of proteins. Nevertheless, PD-MS depended on fission products of radioactive californium for ionisation/desorption and was accompanied by poor resolution and peak broadening, due to the large kinetic energy distribution carried by the molecular ions, and thus poor mass accuracy, and for this reason was not widely used. Following the introduction of the High Field Magnet by Howard Morris and colleagues in this laboratory in 1977 (Morris 1981) and Fast Atom Bombardment mass spectrometry (FAB-MS) by Barber and coworkers in the 1980s (Barber, Bordoli et al. 1981), the structural characterisation of large involative glycoconjugates became possible (Morris 1981, Morris, Panico et al. 1981, Dell and Ballou 1983, Egge, Peter-Katalinić et al. 1983). FAB-MS was widely used for these analyses for almost twenty years, but then has been superseded by the more versatile technologies of MALDI (Karas 1988) and ESI (Fenn, Mann et al. 1989).

A Electron Impact (EI-MS)

In EI-MS, the gaseous volatile sample is introduced into the source from an online gas chromatograph or a probe (**Figure 1.4**) where electrons from a tungsten filament with energy

of 70eV, bombard it so as to form highly energetic radical molecular cations by the ejection of an unpaired electron and giving enough internal energy to fragment the molecular ion into stable fragments, producing a mass spectrum.

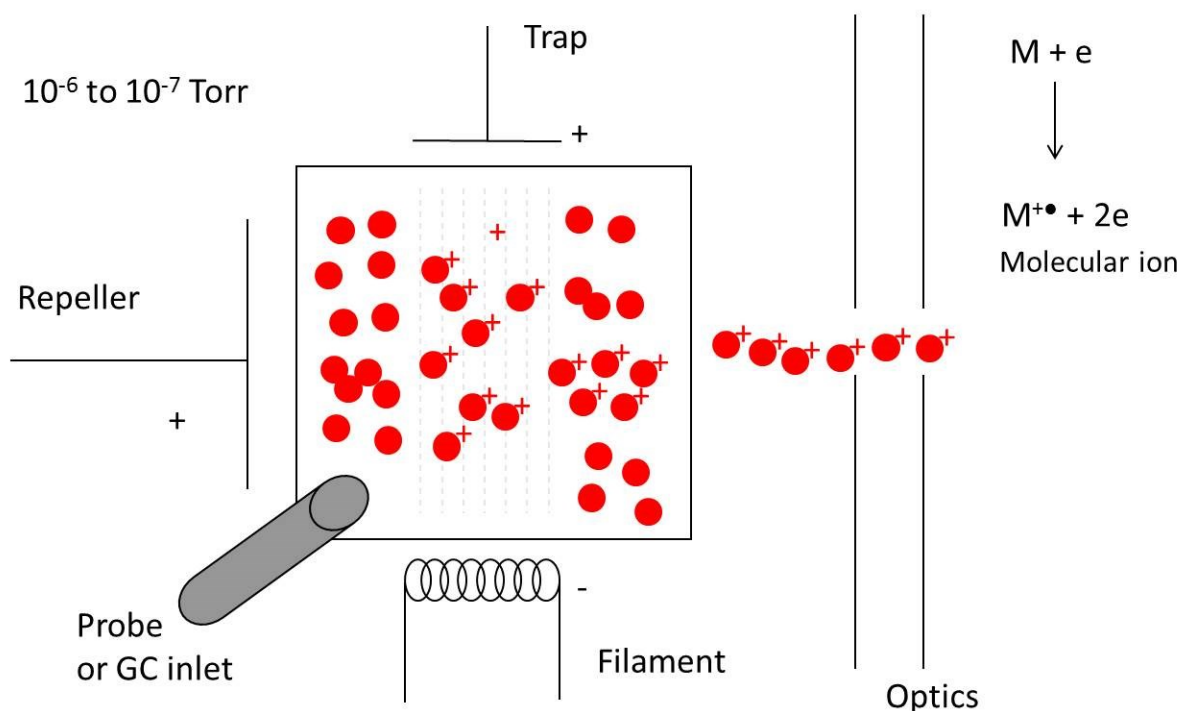


Figure 1.4 The diagram shows an Electron Impact mechanism where the red circles are the gaseous volatile sample and the red circles with a plus are the highly energetic radical molecular cations produced following the bombardment by the electrons coming from a tungsten filament. The product ion is a radical cation ($M^{+\bullet}$).

A useful application of EI-MS in glycoprotein analysis is for sugar composition and linkage determination by GC-MS (Sweet 1974).

Linkage analysis, also known as “methylation analysis”, is a well-established method whereby polysaccharides are permethylated, hydrolysed and then further acetylated for structural investigation (Björndal 1970). Its principle is to introduce a stable substituent onto each free hydroxyl group of the native glycan, which in **Figure 1.5** is an ether-linked methyl group. The glycosidic linkages, which are more labile than the ether-linked methyl groups, are then cleaved by acid hydrolysis, generating individual methylated monosaccharides with free hydroxyl groups at the positions that were previously involved in a linkage. The partially methylated monosaccharides are then derivatised to produce volatile molecules amenable to GC-MS analysis which carry an acetyl “marker” at linkage positions (Dell and Morris 2001).

The most common strategy comprises reduction of the monosaccharides to produce alcohols at C-1 by eliminating the formation of ring structures followed by derivatisation of the free hydroxyl groups, usually by acetylation. Each component of the mixture of partially methylated alditol acetates can then be identified by a combination of GC retention time and EI-MS fragmentation pattern. Unfortunately, this technique does not provide any information on the oligosaccharide sequence or anomeric configuration of the monosaccharide linkages (Varki 2009).

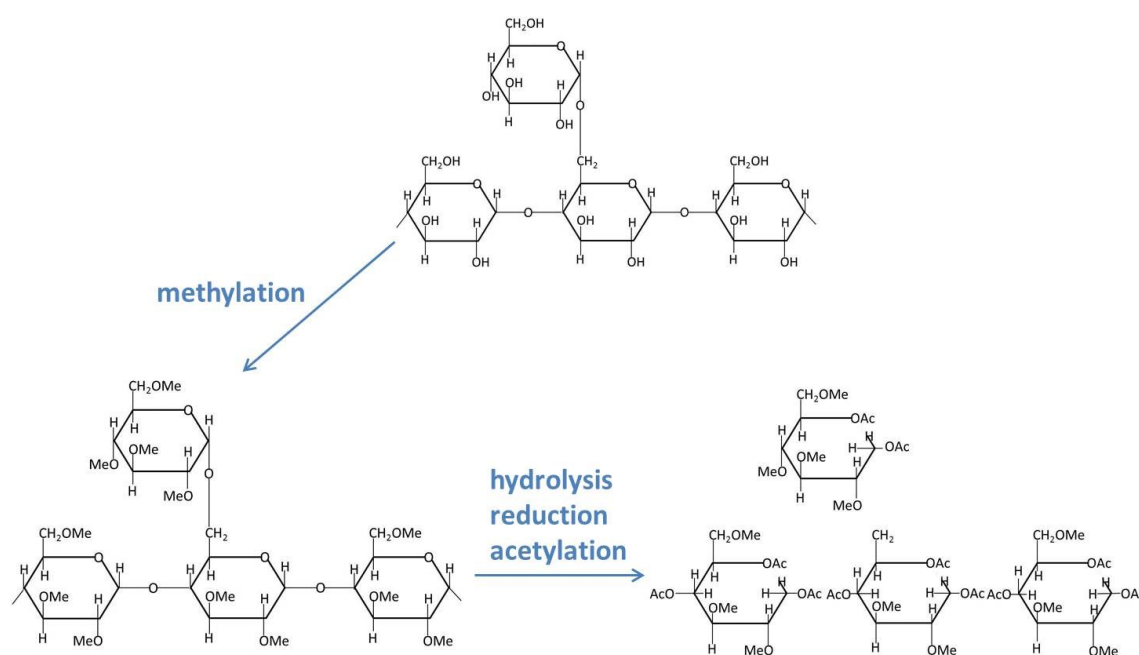
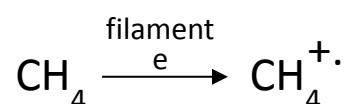


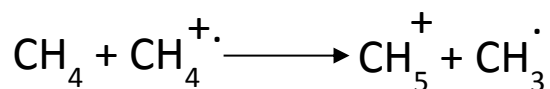
Figure 1.5 Example of linkage analysis. The chart displays the procedure of linkage analysis of a α 1,4-linked glucose branch. The terminal α 1,6 glucose, which is acetylated at C-1 and C-5, while the two α 1,4 glucose units on either side of the branching monosaccharide are acetylated at C-1, C-4 and C-5. The branching glucose is acetylated at C-1, C-4, C-5 and C-6. The schematic has been adapted from Varki and coworkers (Varki 2009).

B Chemical Ionisation (CI) and Field Desorption (FD)

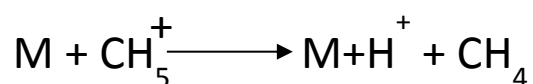
Chemical ionisation was the first of the “soft ionisation” methods, and is similar to electron impact and indeed uses an EI source to generate ions. The difference is that instead of being under high vacuum, the CI source is flooded with methane gas and the sample is a minor “contaminant” in that gas (**Figure 1.6**). This methane is first ionised, as in electron impact, but then because of the gas pressure many collisions can take place, and these lead to the formation of the chemical ionisation “reagent” CH_5^+ .



Main collision will be between $\text{CH}_4^{\cdot+}$ and CH_4



A collision between a sample molecule (which must also be in the gas phase – i.e. volatile) and a reagent ion then leads to the stronger base (for example an amino group on a peptide) abstracting the proton to become a “quasi-molecular ion” $\text{M}+\text{H}^+$.



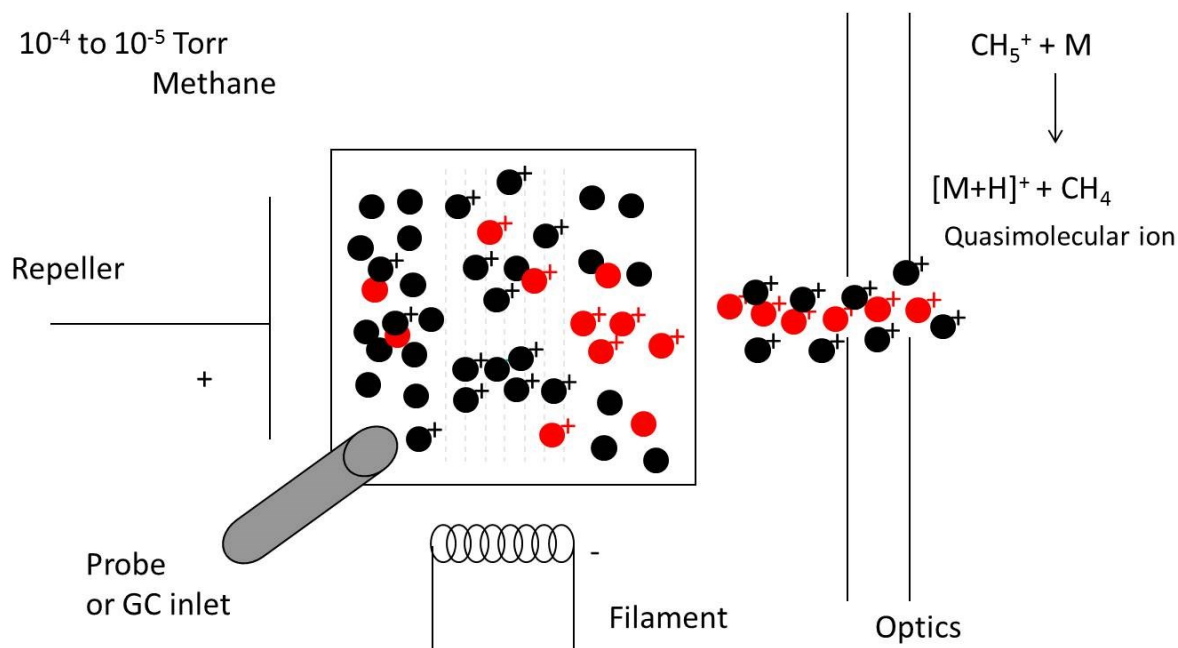
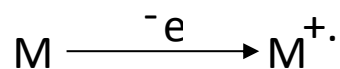


Figure 1.6 The diagram representation shows the Chemical Ionisation mechanism where the red circles are the gaseous volatile sample and the black circles are the reagent gas (methane). Following a collision between a sample molecule and a reagent ion, there is the formation of a quasi-molecular ion ($\text{M}+\text{H}^+$).

A Field Desorption (FD) ion source consists of an activated wire which is coated with the sample solution (and dried), and an extraction plate held a few millimetres away. A large potential difference is then applied usually up to 14 or 15 kV. The wire is pre-activated by growing polymeric “christmas trees” using benzonitrile before sample loading. The extremely high field strength, 12 kV over 2 mm on the tips of the “christmas tree” leads to electron tunneling from the sample on those tips into the wire.



This was the first method for directly ionising involatile biopolymers (1969) but is not now widely used because of technical difficulty.

C Fast Atom Bombardment Mass Spectrometry (FAB-MS)

Barber and coworkers introduced a new ionisation method, Fast Atom Bombardment (FAB) (Barber, Bordoli et al. 1981), which was applied for the analysis of biopolymers over the following decade, opening up the field to most areas of biomedical research. This method

was one of the first techniques to allow large intact carbohydrates to be ionised efficiently (Morris, Panico et al. 1981, Dell 1987, Dell and Morris 2001).

The sample is dissolved in a suitable solvent and an aliquot is added to a drop of a relatively non-volatile liquid matrix on the target. The most common FAB matrices are glycerol, thioglycerol or 3-nitrobenzyl alcohol. Once the probe is inserted into the mass spectrometer, the matrix is bombarded with a beam of high kinetic energy atoms, such as from the inert gas argon. Later, it was discovered in this laboratory that a beam of fast ions produce the same results (Morris, Panico et al. 1983) and the smaller ions guns can be placed nearer to the sample target giving higher sensitivity, and subsequent ion sources then used xenon or caesium guns for bombardment and ionisation. Thus, energy is transferred to underlying layers resulting in ionisation of the sample molecules, which are “sputtered” from the surface. All of this allows desorption and ionisation either through protonation/deprotonation and the coupling with cations or anions (adducts) in a positive or negative ion mode, and entrance into the gas phase. Therefore gas-phase ions are generated without prior volatisation of the sample thus allowing the analysis of polar, involatile and thermally labile compounds (**Figure 1.7**). Both positive and negative ions are produced during the sputtering process and either can be recorded by the appropriate choice of instrumental parameters. Molecules are ionised by the addition of a proton or cation, such as sodium, potassium, ammonium (positive ion formation) or by the loss of a proton or addition of an anion such as chloride (negative ion formation). During ionisation some internal energy is imparted to the molecule resulting in fragmentation of labile bonds and these molecular and fragment ions are accelerated and detected. For example in oligosaccharide analysis, FAB-MS can provide information on the degree of heterogeneity, on the type and sites of glycosylation and branching patterns of N- and O-linked glycans and, moreover, was used to identify polar molecules, such as peptides and oligosaccharides up to 6000 Da using a High Field Magnet double focusing ZAB instrument developed here at Imperial College London (Dell 1993).

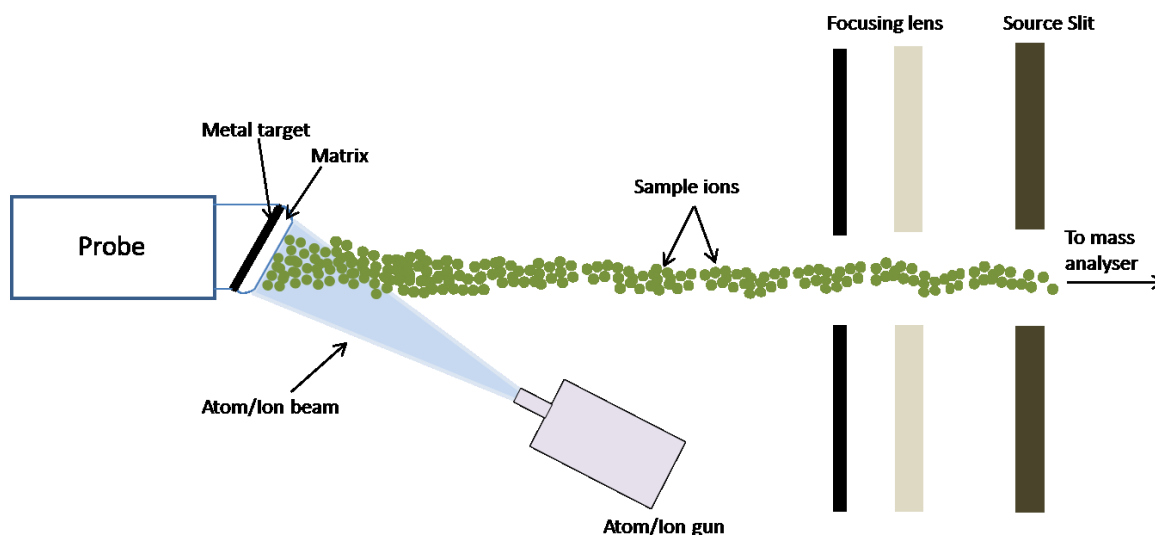


Figure 1.7 The diagram representation shows the Fast Atom Bombardment mechanism in which a beam of high energy atoms/ions strikes a surface to create sample ions. Once generated, the gas-phase ions can reach the mass analyser, which can be either a magnetic sector or a triple quadrupole.

D Electrospray Ionisation (ESI-MS)

ESI has become one of the most widely used ionisation techniques for proteomic and glycoproteomic analysis. It is one of the atmospheric pressure ionisation (API) techniques and was initially described in principle by Malcolm Dole in the late 1960s, but only in the late 1980s did it come to prominence in the work of John Fenn and collaborators, who demonstrated the use of electrospray for the ionisation of simple polar materials like arginine (Fenn, Mann et al. 1989).

In ESI-MS, a stream of liquid containing the sample of interest is directly injected into the atmospheric pressure ion source of a mass spectrometer. In the case of pure samples dissolved in the mobile phase, a direct injection “loop” can be used, but the method is especially useful when interfaced to chromatographic procedures like capillary zone electrophoresis (CZE), high performance liquid chromatography (HPLC) or “nanospray” ESI (see later). Once the sample is already dissolved in a mobile phase, the next step consists of pumping it through a fine stainless steel capillary and then introducing it into the source in solution through a narrow metal-tipped glass capillary. A potential difference of typically 3-6 kV is usually applied between the capillary and counter electrode situated at the entrance cone, and thus a spray of microdroplets is generated. These then pass through a series of “skimmers”, and undergo “drying” usually with the assistance of nitrogen, imposing charge accumulation on the liquid surface forming highly charged droplets. Upon evaporation, the

surface charge density is increased whilst droplet size decreases, resulting in the Rayleigh instability limit being reached, overcoming the surface tension to cause breakdown into even smaller droplets and ultimately the creation of a charged molecular species, devoid of solvent, ready for analysis either by a magnetic sector or more commonly a quadrupole analyser (**Figure 1.8**).

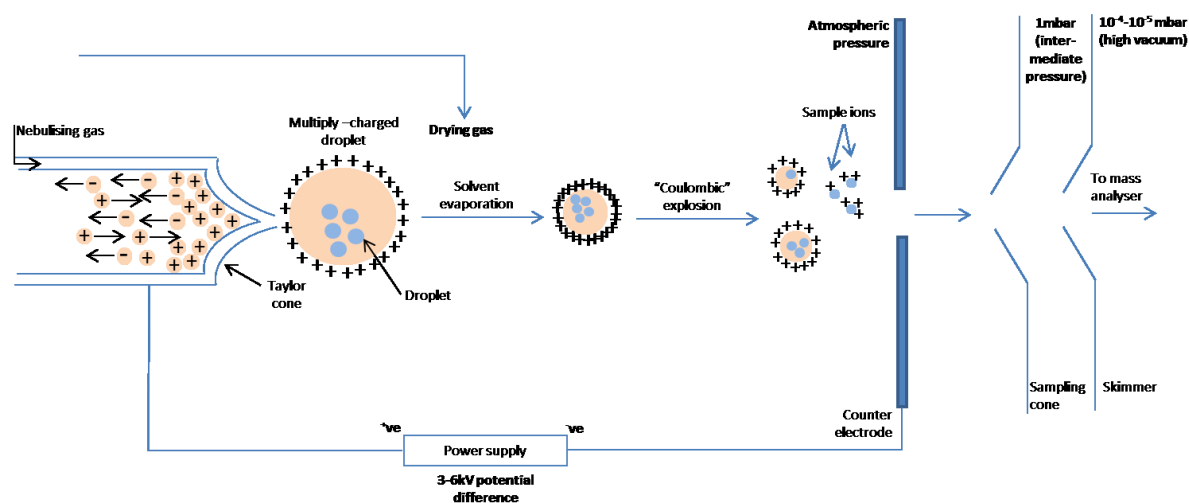


Figure 1.8 The diagram representation shows the Electrospray Ionisation mechanism. In positive ion mode, a positive potential is applied to the capillary tip with the following formation of positively charged droplets. Solvent evaporation of the charged droplets via a weak nitrogen countercurrent leads to gas-phase ions. The sample ions are driven by an electric potential and pressure difference and the potential difference over the cone orifice and downstream ion optical elements transports the sample ions into the high-vacuum region of the mass analyser chamber.

Large molecules, such as proteins and glycoproteins, usually tend to carry more than one charge and, as the mass spectrometer is an instrument which measures mass to charge ratio (m/z), even very large molecules exceeding 100 kDa can be amenable to ESI-MS analysis. A typical spectrum would be a distribution of signals carrying varying numbers of net positive or negative (depending on ion mode) charges on basic or acidic sites in the molecules (Dell 1993) as shown in **Figure 1.9**. In fact, ESI-MS is particularly useful as a method to study protein post-translation modification (PTM) by investigation of any mass difference between the observed signal and the calculated sum of the amino acids present in the sequence.

LC-MS reflects the compatibility between electrospray and liquid separation techniques, thus allowing protein and carbohydrate analysis and moreover, micro-separation methods suitable for glycan analysis have been developed in many laboratories worldwide (Morris, Paxton et al. 1996, Greer 1997, Novotny and Mechref 2005, Mechref and Novotny 2006). Early ESI sources ran at flow rates of a few $\mu\text{l}/\text{min}$, but nanospray sources operate at flow rates of 10-30

nl/min increasing sensitivity and the possibility for online nanoLC for complex mixtures (Wilm and Mann 1994).

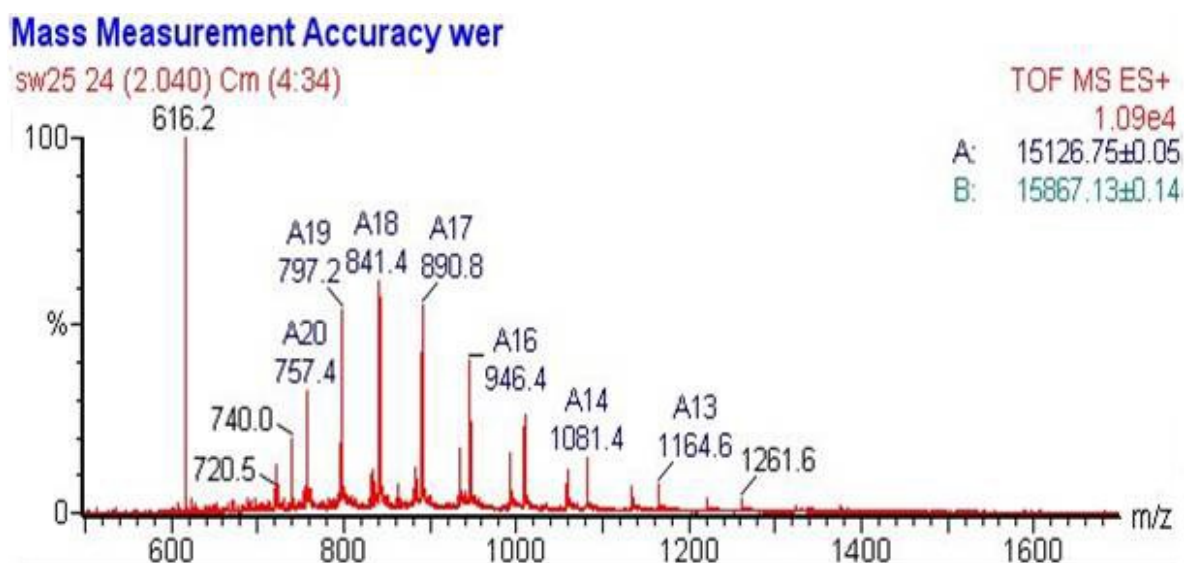


Figure 1.9 Typical ES-MS spectrum showing a distribution of signals carrying varying numbers of net positive charges. Specifically this MS spectrum shows two isoforms of haemoglobin, with the peak at m/z 616 being the heme group.

E Matrix-Assisted Laser Desorption Ionisation Mass Spectrometry (MALDI-MS)

In the late 1960s ionisation based on laser desorption was introduced and then it was developed and adapted for the analysis of biomolecules, but only in the 1980s were its limitations of poor ionisation efficiency and extensive fragmentation overcome with the use of a matrix to embed the sample leading to the matrix-assisted laser desorption ionisation (MALDI) technique. Karas and Hillenkamp showed that proteins and carbohydrates respectively can be ionised by MALDI (Karas 1988) and afterwards, Mock and colleagues applied it to N-linked glycans (Mock 1991). Since then it has become one of the most popular techniques for glycan analysis.

In contrast to FAB-MS, MALDI-MS uses a crystalline matrix, rather than a liquid one and a pulsed beam of photons, instead of a continuous beam of atoms or ions. The sample is embedded in a low molecular weight UV absorbing “crystalline” matrix. Moreover, and importantly, the matrix is chosen to have an absorption maximum near the wavelength of the pulsed laser that is used to ionise the sample. Nitrogen lasers with a wavelength of 337 nm are widely used. The matrix absorbs the pulsed laser energy, flash evaporates and then lifts

the sample into the gas phase from where the net positively charged molecules are extracted towards the TOF mass analyser as seen in **Figure 1.10**. Compared with FAB or ESI, MALDI is a “softer” ionisation technique and produces mainly singly charged quasi molecular ions of the form $[M+H]^+$ or $[M+Na]^+$, with little fragmentation, making this technique ideal for “fingerprinting” or mass profiling of complex glycan mixtures.

The choice of matrix is crucial so as to have an efficient ionisation of a given sample and many matrices have been developed, but one of the earliest, 2,5-dihydroxybenzoic acid (DHB), is the most popular for carbohydrate and glycoprotein analysis in this laboratory. Moreover, α -cyano-4-hydroxy cinnamic acid (HCCA) is used for peptides and glycopeptides, whereas sinapinic acid (SA) is the preferred one used for higher molecule weight proteins or glycoproteins. Analysis with diaminobenzophenone (DABP) and 2-(4'-hydroxylbenzeneazo) benzoic acid (HABA) are used for analysis of glycolipids.

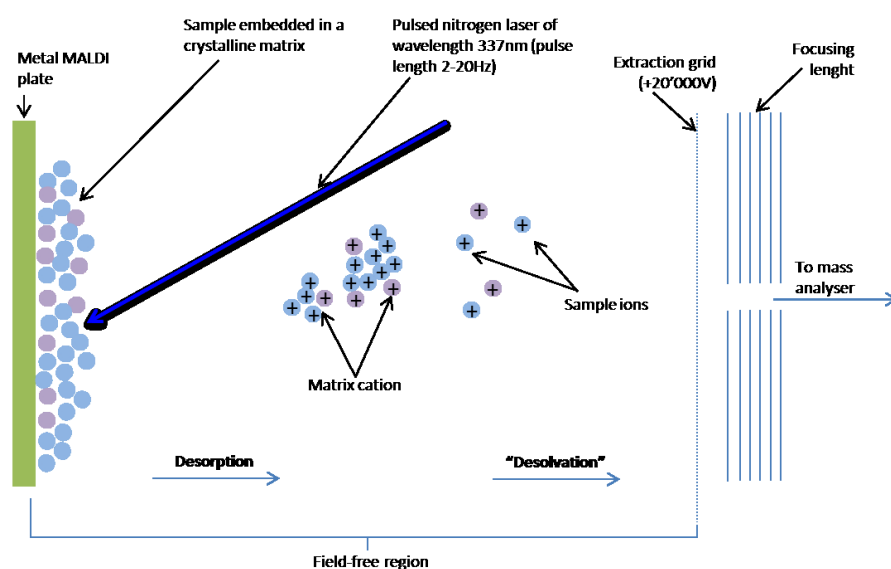


Figure 1.10 The diagram representation displays the Matrix Assisted Laser Desorption Ionisation mechanism in positive ion mode. The sample is mixed with a crystalline matrix which presents an absorption maximum near the wavelength of the pulsed laser used to ionise the sample. Once the matrix absorbs the pulsed laser energy, flash evaporated and push the sample into the gas phase. Then the gas-phase ions can fly towards the mass analyser.

1.4.3 Mass Analysers

Several mass analysers have been used in biopolymer mass spectrometry, including double focusing magnetic sector instruments, quadrupole mass analysers, time-of-flight (TOF) and

ion trap devices. All mass analysers in these instruments use the deflection behaviour of the charged particles moving through field regions, to effect their separation.

A Magnetic Sector

In a double-focusing Magnet Sector instrument (**Figure 1.11**) the ions leaving the source are first passed through an Electrostatic Analyser, which does not attempt to separate masses, but rather to produce a mono-energetic beam. This beam is then passed into the Magnet Field which deflects the ions according to their mass/charge ratio. The field strength is “scanned” overtime (eg 10 secs) in order to sequentially pass ions of differing mass through a slit to the detector, to produce the overall mass spectrum. The resolution of such an instrument is around 100,000 allowing relatively easy accurate mass measurement and therefore atomic composition assignment.

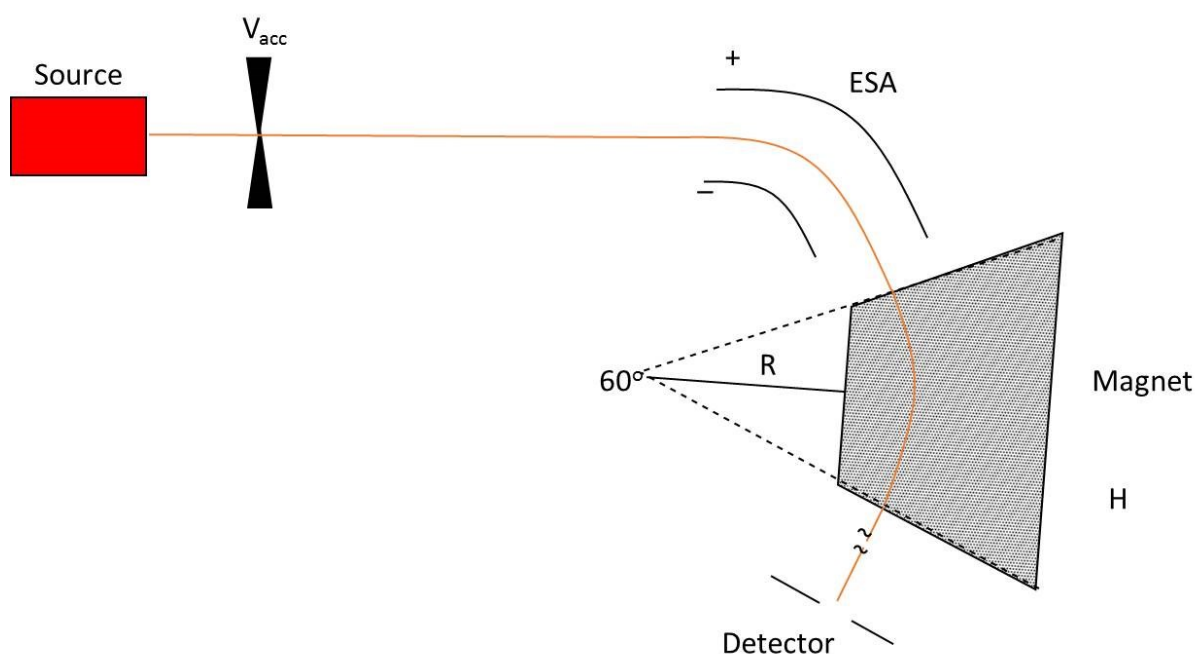


Figure 1.11 The diagram representation displays the Magnetic Sector mechanism. In a magnetic sector analyser the ions pass through an electrostatic analyser (ESA) which produce a mono-energetic beam (orange line) and once reached a magnetic field, is deflected to a circular motion of a unique radius in a direction perpendicular to the applied magnetic field.

B Ion Traps

Ion traps use oscillating electric fields generated by radiofrequency (RF) voltages applied to electrodes of opposite polarity to “trap” a range of m/z values. Scanning the RF or magnetic

field allows ions to be detected sequentially. For ion detection, the voltages are altered to destabilise ion motions, resulting in the ejection of the ions from the storage cell to the detector.

These instruments afford moderate resolving power and with fast analysis cycles are compatible with LC-MS analysis (**Figure 1.12**).

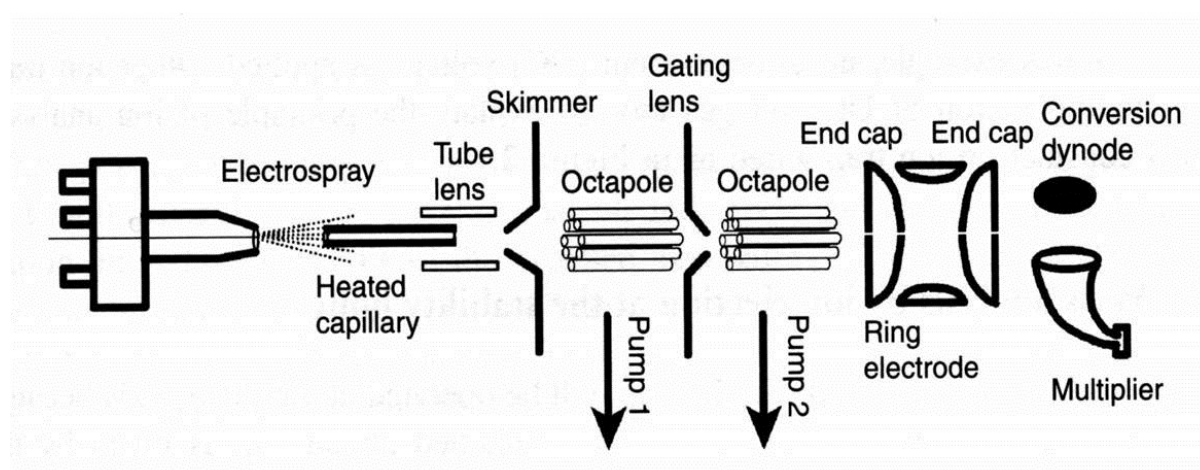


Figure 1.12 The diagram representation displays the Ion Trap mechanism.

C Quadrupole Analysers

Invented by Paul and Steinwedel in the mid 1950s (Paul and Steinwedel 1953) but commonly used from the 1970s, the quadrupole mass filter uses an electrical field comprising RF and DC components to separate the ions. It comprises four parallel rods with circular or hyperbolic cross-sections. The application of a fixed direct current component and an oscillating radiofrequency perpendicular to each other producing opposite current polarity in adjacent rods creates an electrical field. Changing the voltage on the fields permits only certain ions with the correct m/z value to pass through the filters without obstruction towards the detector, different from those with other m/z values which will collide with the rods and are lost (**Figure 1.13**). Therefore, in the quadrupole mass analyser, ion separation is based on the stability of the ion trajectory through the four parallel rods.

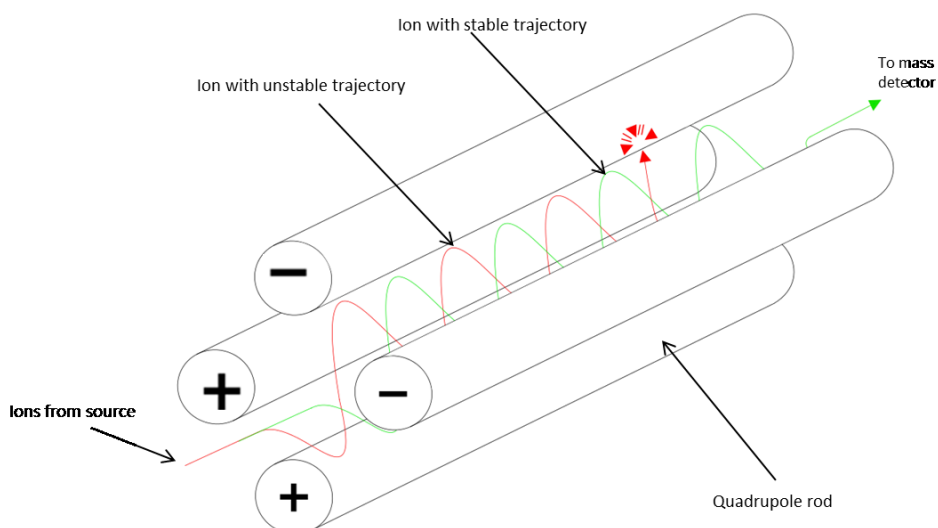


Figure 1.13 The diagram shows a quadrupole analyser where the red ion has the selected m/z value, therefore passing out of the quadrupole into the mass detector, while the path of the blue ion is intercepted by a quadrupole rod, which results in the ion being de-charged before reaching the mass detector and so not being detected.

These instruments, although they do not offer high resolution power, are rapid scanning and can operate at highly pressures and thus be interfaced with a wide variety of inlet systems, which is useful in combination with gas chromatography (GC-MS) and LC-ESI-MS. Quadrupole mass analysers are considered mass filters and are also now used as the first analyser in tandem mass spectrometry (see **section 1.4.4**).

D Time-of-Flight (TOF)

The time-of-flight analyser separates the ions based on their differences in velocities as they move after acceleration in a field-free “flight tube” towards the detector. The separation principle of TOF analysers is that the ions of different m/z values traverse the field-free region at different velocities and hence reach the detector at different times. To determine the flight time, all ions, regardless of their m/z values, have to enter the flight tube simultaneously. However, due to the spread of kinetic energies originally imparted to the ions during the ionisation process and time necessary to form the ions, ions of the same m/z value present a spread of velocities that contribute to a spread in the signal detected, resulting in reduced resolution. To improve the resolution, a “reflectron” device is employed.

The reflectron, or ion mirror, is a reflecting electric field and consists of a series of rings/grids with voltages that increase up to a value slightly greater than the voltage of the ion source, so

the more energetic ions penetrate further into the field to such an extent that their increased path length just compensates for their increased velocity, resulting in a reduction of the spread of initial kinetic energies of the ions produced and an improvement in resolution (**Figure 1.14**). High mass ions can't traverse the reflectron and so high resolution reflectron mode comes at the expense of mass range, due to the physical restrictions applied by the reflectron and this limits the mass range to about 5000 Da. All of this means that reflectrons are used for the analysis of low to medium molecular weight molecules, but deactivated in favour of the sensitivity of linear analysis to detect larger ions. The pulsed nature employed in MALDI-MS makes this technique perfectly suited to being associated with a TOF mass analyser, which requires one "packet" of ions to be analysed before the arrival of the next i.e. a dis-continuous ion source.

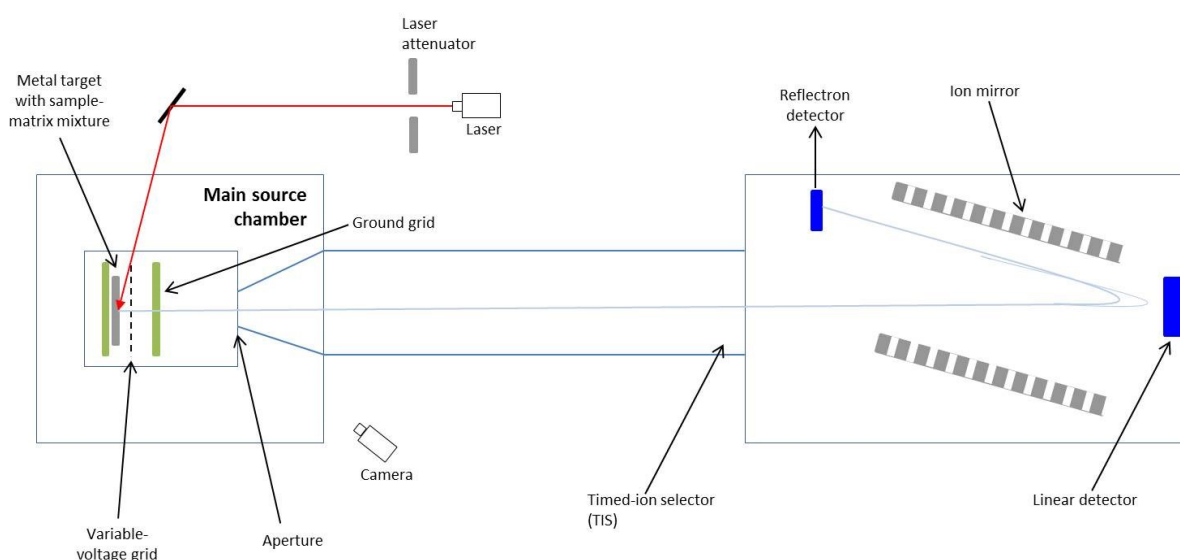


Figure 1.14 The above diagram depicts a Voyager-DE STR System with a Time-Of-Flight analyser. The TOF analyser separates the ions based on their differences in velocities therefore the ions reach the detector at different times. However, all ions have to enter the field-free flight tube at the same time. The ion mirror are necessary to improve the resolution.

1.4.4 Tandem mass spectrometry

In mass spectrometry, the molecular ion mass can bring information on the theoretical monosaccharide, amino acid or lipid compositions and confirm the presence or absence of modifying groups, such as sulphate or phosphate. However, fragmentation of the molecular

ion permits rigorous sequencing, comprising branch determination and site occupation in glycoproteomics, although in practice this can be very challenging. Both MALDI-MS and ESI-MS are soft ionisation techniques, thus very useful for mass profiling of glycan and glycopeptide mixtures, but for further sequence analysis, it is necessary to fragment the quasi-molecular ion and so to employ two mass analysers in tandem.

Several mass spectrometer designs now include two mass analysers in tandem, being the same or different types, and the term MS/MS is referring to such coupling when used to induce fragmentation. The majority of MS/MS experiments are carried out on instruments provided with triple quadrupole (Hunt, Giordani et al. 1982), hybrid quadrupole with time-of-flight (Q-TOF) (Morris, Paxton et al. 1996) and time-of-flight/time-of-flight (TOF/TOF) analysers (Vestal and Campbell 2005). In MS/MS experiments, the first mass analyser performs the mass selection of a precursor ion of interest from the mixture of ions produced in the source. Then, the selected ion is passed into a pressurised collision cell containing an inert gas, eg. argon or xenon, and undergoes Collision-Induced-Dissociation (CID), sometimes also called Collisionally-Activated-Decomposition (CAD). This collision promotes fragmentation of the precursor ion and these fragments are then passed into the second mass analyser, which separates the resulting fragment ions, and the overall spectrum is then finally detected. These fragment ions show a unique fingerprint pattern of the selected precursor ion. Tandem mass spectrometers are fundamental tools to afford detailed structural characterisation of individual components including peptide or glycan sequencing.

Recently a different method of fragmenting multiply-charged gaseous macromolecules in a mass spectrometer between the stages of tandem mass spectrometry has been developed and it has been named Electron-Transfer Dissociation (ETD) in which collision with an organic electron-donor molecule in the collision cell between the tandem analysers leads to a different and more gentle fragmentation process of benefit in some post-translational modification studies.

Much of the data acquired in this thesis has been created by exploiting the power of tandem mass spectrometry using the following state-of-the-art instruments: a quadrupole orthogonal acceleration time of flight (Q-STAR Pulsar) mass spectrometer, a Xevo G2 Q-TOF mass spectrometer, a Synapt G2-S Q-TOF mass spectrometer, a TripleTOF 5600 mass spectrometer and a MALDI-TOF-TOF mass spectrometer.

A Triple Quadrupole

The first successful tandem design, allowing MS/MS analysis, was the so-called Triple Quadrupole mass spectrometer, shown in **Figure 1.15**. Ions of interest (precursors or parents) are selected in Q1 for passage into a collision cell filled with inert gas for CID fragmentation (Q2). The product (or daughter) ions are then separated to produce a mass spectrum by scanning Q3. Note that Q2 is not in fact acting as a quadrupole, only as an ion guide around the gas cell.

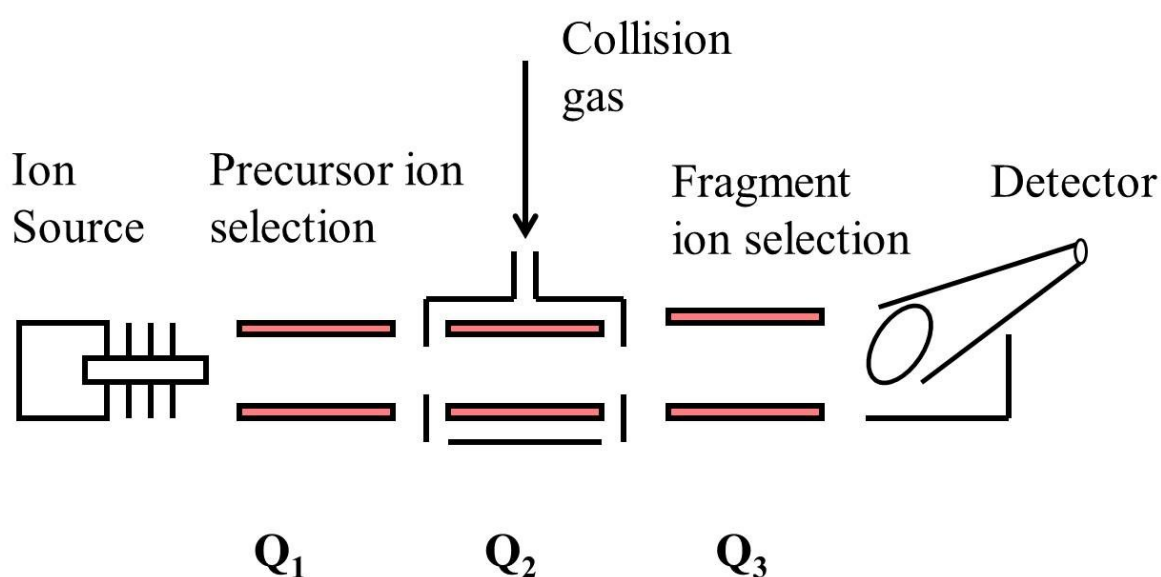


Figure 1.15 The above diagram depicts a Triple Quadrupole mass spectrometer. The Quadrupole labelled Q₂ is just an ion guide around the gas cell.

B Quadrupole Orthogonal Acceleration Time-of-Flight (Q-TOF)

In the early '90s, Howard Morris conceived a novel geometry instrument for high sensitivity unambiguous sequencing. He and the manufacturers (Morris, Paxton et al. 1996) subsequently developed the idea, a quadrupole/orthogonal-acceleration time-of-flight tandem mass spectrometer named Q-TOF, with the aim of overcoming the resolution and sensitivity deficiency of the triple quadrupole design, to give ultra high sensitivity MS/MS analysis together with the unambiguous mass assignment arising from good resolving power. This was achieved using non-scanning detection and reflectron TOF resolution, specifically

coupled with a quadrupole for ion selection and passage of the ions into the CID cell, ensuring low energy CID. The Q-TOF geometry therefore incorporates a quadrupole mass filter, a hexapole collision-gas cell and an orthogonal time-of-flight reflector analyser. Generally it is fitted with an ESI source, but it is also possible to find it coupled with MALDI ionisation. The performance characteristics of the first prototype Q-TOF and subsequent commercial instruments were routine femtomole (10^{-15} mole) sensitivity with good signal-to-noise ratios in MS/MS spectra, mass accuracies of 0.1 Da, and easy differentiation of singly-, doubly- and triply-charged precursor ions. A second generation Q-TOF mass spectrometer was born in the Q-STAR Pulsar as an example. The Q-STAR mass spectrometer has an optimum duty cycle, good transmission and sensitivity which together with a novel high pressure collision cell (LINAC) permitting high fragmentation efficiency and the production of good quality data for biopolymer sequencing. These instruments are preferably linked to a nanoLC column, enabling high sensitivity MS/MS experiments in an on-line fashion.

In MS mode, the collision cell acts only as an ion guide, whilst the TOF is the mass analyser; however, in MS/MS mode, the ion filter capability of the first quadrupole analyser is used to select the precursor ions of interest, which are then accelerated into the hexapole collision cell, where they undergo fragmentation via collision with nitrogen or argon gas, with the TOF used to mass analyse the resulting fragment ions before they reach the array detector (**Figure 1.16**).

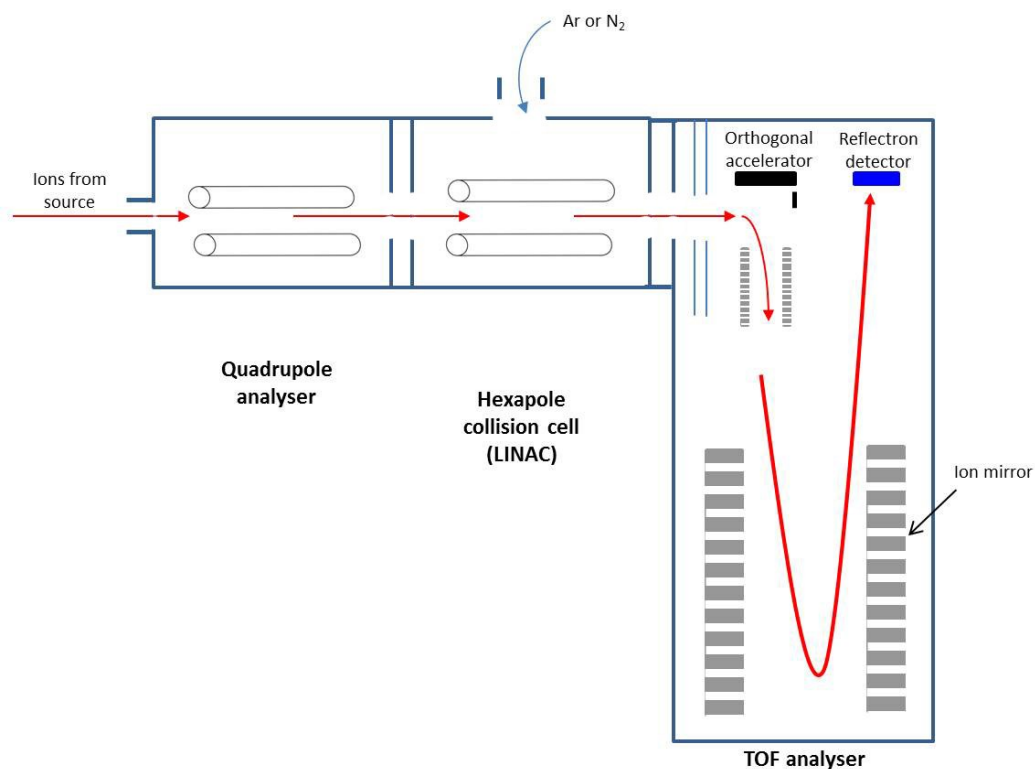


Figure 1.16 The diagram depicts a quadrupole analyser hexapole collision cell and orthogonal acceleration Time-Of-Flight (TOF). The hexapole collision cell makes the collision more efficient and the orthogonal geometry minimises the level of chemical noise coming from the collision cell.

C Tandem Time-of-Flight

The Q-TOF instrument design deliberately incorporates multiple low energy CID collisions to give simple easily interpreted spectra showing low-energy fragmentations; instead, on the other hand, for some purposes high energy CID is beneficial, therefore to provide additional information-rich cleavage products. MALDI-TOF/TOF instruments are able to produce CID fragments at both low and high energies, whilst preserving the resolution and sensitivity of the single TOF instrumentation (Vestal and Campbell 2005), their only drawback being the relatively low parent ion resolution compared with that available in a Q-TOF geometry. The MALDI-TOF/TOF instrument used in the production of some of the data described in this thesis was the 4800 MALDI-TOF/TOF (**Figure 1.17**).

MALDI-TOF/TOF instruments consist of a linear delayed extraction MALDI-TOF, a collision cell and a second TOF analyser. The ion source optics and electronics of the initially developed MALDI-TOF remain basically unchanged, but additional elements incorporated for the MS/MS mode include a timed-ion-selector (TIS). Considering the MS mode, ions are extracted and guided directly to the detector; however, in the MS/MS mode, the molecular

ions produced in the source pass through the first TOF and arrive at the TIS. This allows the selection of a precursor ion according to its velocity through the first TOF analyser, while other ions are deflected away. A time delay generator is programmed to open the TIS as the lightest mass of interest reaches the gate and to close it when the highest mass of interest has passed through, resulting in the retardation of selected ions by the deceleration lens and the entrance into the collision cell at a designated collision energy, defined by the potential energy between the source and the collision cell (i.e. 1kV, 2kV). Variation in the collision energy is accomplished by the adjusting of the ion source potential relative to that of the collision cell, where the pressure of the inert gas, like argon, is controlled to provide satisfactory fragmentation. Then, fragment ions produced in the collision cell travel with the same velocity as the decelerated precursor ions, until they are re-accelerated by a high voltage pulse into the second TOF analyser, reaching the detector via the reflectron and providing high resolution MS/MS data of the selected m/z parent ion.

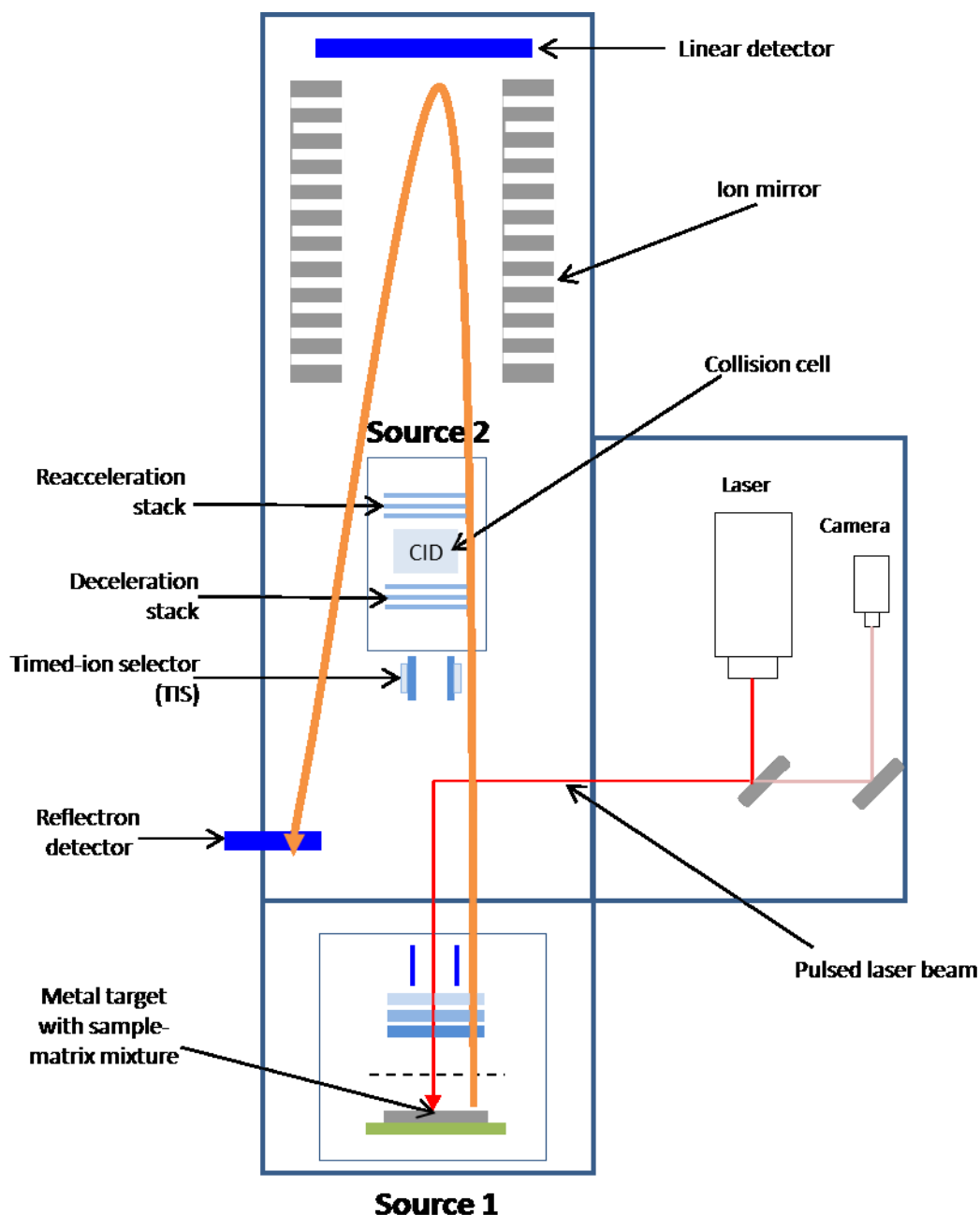


Figure 1.17 The MALDI TOF/TOF 4800 Analyser from Applied Biosystems is a double focusing instrument with linear and reflector flight path options, permitting the selection of individual precursor ions (TIS) for further fragmentation in the collision cell. Moreover fragmentation is facilitated by the presence of an inert gas, i.e. argon, in the collision cell (CID) and gas pressure is controlled to provide a satisfactory fragmentation.

1.4.5 Fragmentation & Interpretation

A Fragmentation of Peptides

When an ion is created by cationisation (protonation or sodiation) in ESI or MALDI the inherent stability of the oxonium (O-protonated) or ammonium (N-protonated) ion formed

allows for a “quasi-molecular ion” $[M+H]^+$ to be made that normally presents minimal natural fragmentation because of low internal-energy transfer. Initially, protonation is expected to happen at amine groups (the most basic sites) in amino acid side chains (Lys or Arg) or at the peptide amino-terminus itself ($R-NH_2 + H^+ \rightarrow R-NH_3^+$); then, the charge can be passed around the molecule due to its flexibility, including transfer to the peptide bonds where fragments of interest to sequence interpretation are then generated. In negative ion mode, instead, the functional groups that readily lose protons comprise carboxylic acids ($R-CO_2-H \rightarrow R-CO_2^- + H^+$) and alcohols ($R-OH \rightarrow R-O^- + H^+$) and these more susceptible groups form negatively charged $[M-H]^-$ quasi-molecular ions.

As stated previously, MALDI and ESI allow the investigation of intact molecules to provide molecular weight information, but for detailed structural studies it has been found necessary to apply digestion strategies to break the problem down into a series of solvable structural units, the peptides, glycopeptides and glycans in the case of glycoproteins. Therefore, protein digestion by various enzymes will create peptide fragments, such as in the case of trypsin digestion, which cleaves peptide chains at the carboxyl side of the amino acids Lys or Arg, except when either is followed by proline, and will form peptide fragments that possess a C-terminal NH_2 , which will be observed as either $[M+H]^+$ quasi-molecular ions in MALDI or $[M+2H]^{2+}$ in ESI in the positive ion mode.

Peptide fragments may be classified into two categories, coming from cleavage of bonds within the peptide backbone chain to give amino-terminal (N-) and carboxy-terminal (C-) fragment ions (Morris, Panico et al. 1981) and are annotated using the nomenclature described by Roepstorff, Fohlman and Biemann (Morris 1981, Roepstorff 1984, Johnson, Martin et al. 1987).

Applying appropriate low collision energy, two main types of cleavage can occur across the peptide bond. Considering singly charged species, peptide bond cleavage can produce two species, one charged and one neutral, with only the charged species being able to be accelerated and detected by the mass spectrometer. However, via different mechanisms of cleavage (**Figure 1.18**) the charge may be held either on the N-terminal or the C-terminal fragment, and thus both b and y ions are the most prominent fragments observed during MS/MS analysis on peptides.

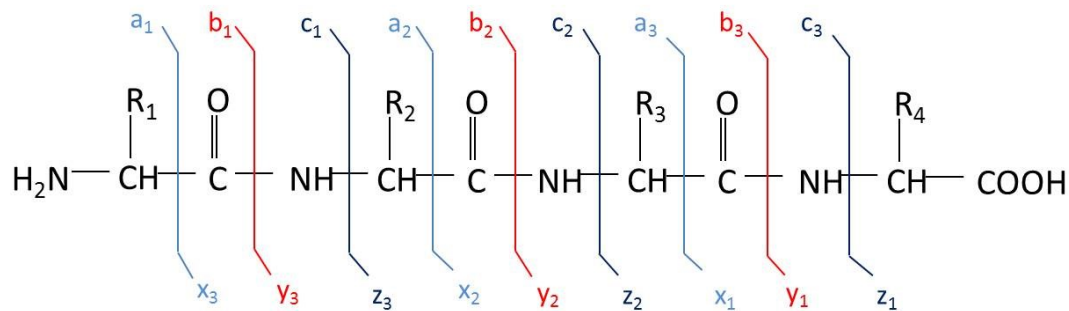


Figure 1.18 The peptide fragmentation nomenclature is shown. Protein backbone cleavages lead to two families of potential fragments, called a, b and c ions (if the charge is retained by the N-terminal fragment) and x, y and z ions (if the charge is on the C-terminal fragment). As the CO-NH bonds result in the most susceptible cleavage site, the b and/or y ions are generally the most abundant fragments observed during MS/MS experiments on peptides. Considering the ETD mechanism, the cleavage of N-C α bonds occurs and produces c- and z- type fragments. Nevertheless, the mass difference between two adjacent a/b/c or x/y/z ions provides the mass to identify the amino acid residue and thus allows sequencing of the selected peptide (Roepstorff 1984).

The mass difference between two adjacent b or y ions gives information on the mass and therefore identify of the amino acid residue, allowing sequencing of the peptide of interest.

Table 1.1 shows the most common amino acids including their composition, structure and monoisotopic masses. The two other types of fragments found in most spectra of doubly or higher charge states result from sub-cleavages of b and y ions to give aldimine (-28 Da from b ions) or immonium ions (-27 from amino acid residue masses).

1 letter code	3 letter code	Name and composition	Residue structure	Monoisotopic Mass (Da)
G	Gly	Glycine C ₂ H ₃ NO		57.02
A	Ala	Alanine C ₃ H ₅ NO		71.03
S	Ser	Serine C ₃ H ₅ NO ₂		87.03
P	Pro	Proline C ₅ H ₇ NO		97.05
V	Val	Valine C ₅ H ₉ NO		99.06
T	Thr	Threonine C ₄ H ₇ NO ₂		101.04
C	Cys	Cysteine C ₃ H ₅ NOS		103.00
I	Ile	Isoleucine C ₆ H ₁₁ NO		113.08
L	Leu	Leucine C ₆ H ₁₁ NO		113.08
N	Asn	Asparagine C ₄ H ₆ N ₂ O ₂		114.04
D	Asp	Aspartic acid C ₄ H ₅ NO ₃		115.02
Q	Gln	Glutamine C ₅ H ₈ N ₂ O ₂		128.05
K	Lys	Lysine C ₆ H ₁₂ N ₂ O		128.09
E	Glu	Glutamic acid C ₅ H ₇ NO ₃		129.04
M	Met	Methionine C ₅ H ₉ NOS		131.04
H	His	Histidine C ₆ H ₇ N ₃ O		137.05
F	Phe	Phenylalanine C ₉ H ₉ NO		147.06
R	Arg	Arginine C ₆ H ₁₂ N ₄ O		156.10
Y	Tyr	Tyrosine C ₉ H ₉ NO ₂		163.06
W	Trp	Tryptophan C ₁₁ H ₁₀ N ₂ O		186.07

Table 1.1 Amino acid residues, compositions, structures and monoisotopic masses.

B Fragmentation of Carbohydrates

Carbohydrate fragmentation pathways have been established using ESI-MS and FAB-MS instrumentation (Dell 1993) and are preserved, regardless of the ionisation methods used. In 1988 Domon and Costello introduced the accepted nomenclature for describing the fragmentation of carbohydrates (Domon 1988, Spina 2004, Stephens, Maslen et al. 2004). The simplest fragmentation at low collision energy results from cleavage of glycosidic bonds to produce reducing and non-reducing end fragment ions. Ions retaining the charge at the reducing end are designated X (cross-ring), Y and Z (glycosidic linkage), while ions with the charge at the non-reducing terminus are A (cross-ring), B and C (glycosidic linkage). Moreover, sugar rings are numbered from the non-reducing end for A, B and C ions and from the reducing end for the others (**Figure 1.19**) (Harvey 2010).

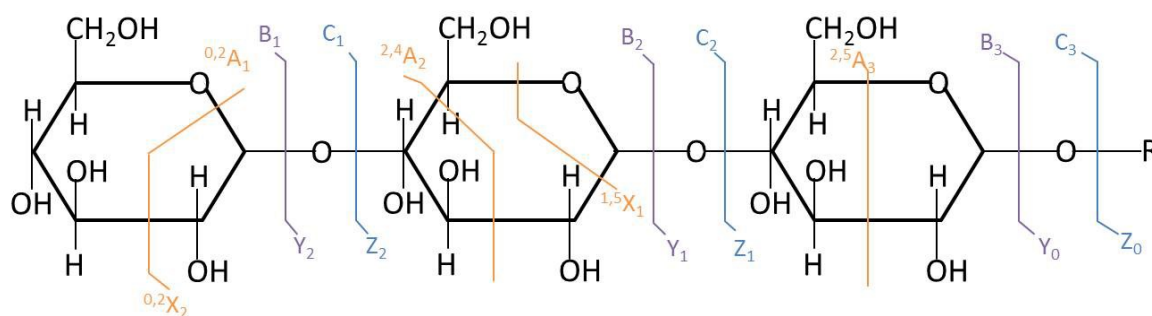


Figure 1.19 Systematic nomenclature of carbohydrate fragmentation. Cleavage of the oligosaccharide takes place by breaking the glycosidic bonds (B, C, Y, Z) and across rings (A and X). Fragment ions containing the non-reducing terminal are A, B and C types, while the reducing end fragments are X, Y and Z types. The subscripts of A and X denote the bonds broken in order to form the respective fragments.

In the positive ion mode, some native samples and most permethylated/peracetylated derivatives (see later) undergo A-type cleavages on the non-reducing side of glycosidic bonds to produce an oxonium ion, which occurs preferentially at amino sugar residues, such as GlcNAc or GalNAc. Instead, in the case of cleavage at the reducing side of a residue, a rearrangement is frequently observed via β -elimination, in which the loss of a water molecule and the formation of a double bond between C-1 and C-2 can be observed (Dell 1993). The second most common type of glycosidic cleavage, β -cleavage, takes place when the charge on the fragment ion is not located at the point of cleavage. Both positive and negative mode mass spectrometry can produce β -cleavages and the resulting fragment ions can be reducing or non-reducing, depending on which of the two bonds to the glycosidic oxygen was cleaved

and therefore providing sequence and branching information (see β -cleavage in **Figure 1.20**) (Dell 1987). At higher collision energy, it is possible to observe cross-ring cleavages. They arise from the sequential movement of electron pairs around the ring resulting in the breakage of single bonds and the formation of double bonds (see ring cleavages in **Figure 1.20**) (Dell and Ballou 1983). Cross-ring cleavages are useful to assign linkages and fragment ions can occur from two or more cleavage events in different parts of the molecule.

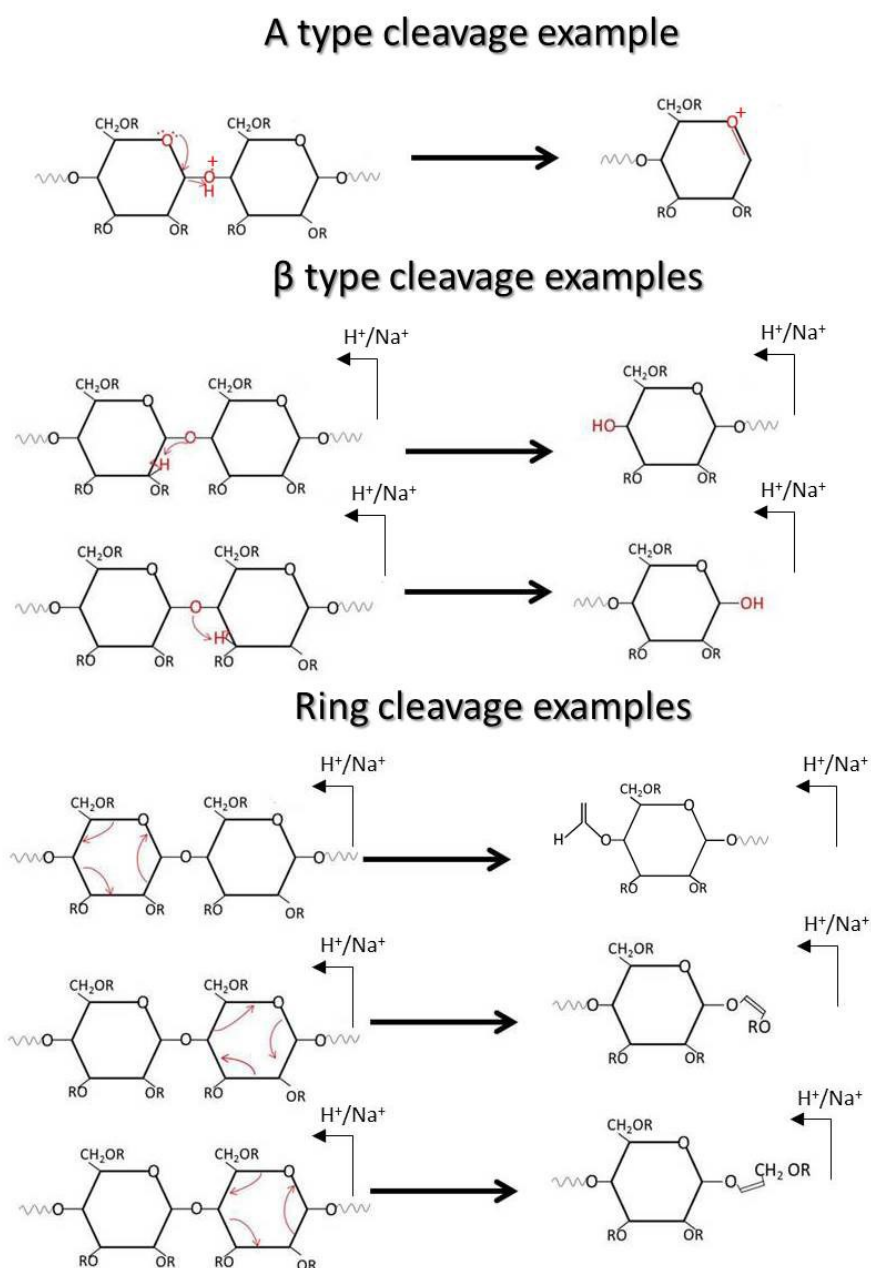


Figure 1.20 Fragmentation and cleavages of oligosaccharides (positive ion). A type cleavages form oxonium ions. Glycosidic cleavages with hydrogen transfer are known as β -cleavages, where the fragment ions can be either reducing (top panel) or non-reducing (lower panel). Examples of ring cleavages are shown in the bottom three panels.

Although, native carbohydrates can of course be analysed by mass spectrometry, due to their extensive inter-molecular hydrogen bonding and the lack of primary amino functions, they do not transfer into the gas phase as well as other biomolecules (Dell 1993, North, Huang et al. 2010). Thus, it is advantageous to derivatise glycans prior to analysis, to remove hydrogen bonding and allow high sensitivity structural analysis. Protection of functional hydroxyl and amide groups in carbohydrate characterisation is performed by permethylation or peracetylation (see **section 2.4.8**). Permethylation comprises the exchange of protons in such functional groups with hydrophobic methyl groups, which was first introduced by Hakomori and coworkers (Hakomori 1964), although not for the purpose of intact mass spectrometry analysis.

Moreover, the increased stability of labile bonds obtained after glycan derivatisation permits the formation of characteristic and predictable fragmentation patterns, which is of fundamental value in *de novo* sequencing of glycans. Some of the monosaccharides seen in the work presented in this thesis along with their composition, general ring structure and their underivatized or derivatised monoisotopic masses are shown in **Table 1.2**.


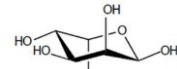
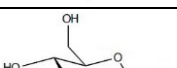
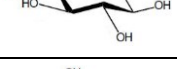
Symbol/code	Name/Composition/Example	Exemplar ring structure	Underivatized monoisotopic mass	Permethylated monoisotopic mass	Perdeuteromethylated monoisotopic mass
Pent	Pentose C ₅ H ₁₀ O ₅		132.04	160.07	166.11
dHex	DeoxyHexose C ₆ H ₁₂ O ₅		146.06	174.09	180.12
Hex	Hexose C ₆ H ₁₂ O ₆		162.05	204.10	213.16
HexNaC	Hexosamine C ₆ H ₁₃ NO ₅		203.08	245.12	254.18

Table 1.2 Monosaccharide residues, compositions, structures and masses.

Derivatisation of the reducing end by reductive amination is required for detection of carbohydrates by optical methods and is commonly used for mass spectrometric analysis for several purposes, such as enhancement of signals and modification of fragmentation patterns. Small aromatic amines, like 2-aminobenzamide (2-AB), 2-aminoacridone (AMAC) or benzylamine, are used for fluorescent detection and increase the proton affinity of the molecules, as $[M+H]^+$ ions are formed sometimes in preference to the more usual $[M+Na]^+$,

particularly with ESI where they have been described to enhance signal strength (Harvey 2010).

C Fragmentation of Glycopeptides

Glycopeptide fragmentation is more complex than individual peptide or oligosaccharide fragmentation because of the additional conjugation of the saccharide to the peptide backbone, and also because of differing fragmentation characteristics of the two polymer classes. Furthermore, a knowledge of the type of glycosylation would be useful since that would affect the strategy chosen for structural analysis. For example, in the case of an N-linked glycopeptide where the carbohydrate component could be relatively large, then N-deglycosylation with for example PNGase F would allow separation and individual study of the peptide and oligosaccharide components in a mammalian system. The site of attachment can in that case be inferred from the presence of the peptide consensus sequence Asn-X-Ser/Thr, where the Asn is converted to Asp by PNGase F. In contrast, in the case of an O-linked glycopeptide, glycan removal could be more problematic (no good comprehensive O-glycanases) and the strategy would probably be to analyse the conjugate in total. With respect to the CID MS/MS analyses of glycopeptides an important issue is the relative ease with which glycosidic bonds cleave compared to the amide bonds of the peptide. This can result in the “stripping off” of the glycan or glycan chain from the glycopeptide, and the loss of the site-of-attachment information. Whilst this problem can often be overcome by careful adjustment of collision energy parameters, an alternative new method is the application of Electron Transfer Dissociation (ETD). This has been shown to provide more comprehensive coverage of post-translational modifications (Sobott, McCammon et al. 2005). Particularly, this approach has the advantage of preserving unstable side-chain modifications during fragmentation (Mikesh, Ueberheide et al. 2006, Wiesner, Premsler et al. 2008). So far, several different instrumental approaches have been established to use electrons for fragmentation and recently a new generation Q-TOF instrument has been developed to incorporate this capability into the Synapt G2-S, which was used in this thesis work. This new technology incorporates ion drift technology which can be used to maximise proteomic coverage for protein identifications, and the supplementary use of ETD for relatively low internal energy fragmentation of the peptide backbone whilst retaining the labile carbohydrate attached. In fact, finding the site where the sugar is attached to the peptide remains one of the most difficult technical aspects of glycoprotein characterisation even when

using the power of modern MS. This kind of technology becomes fundamental to identify different O-linked structures, in particular, as they are problematic because of the lack of an amino acid consensus sequence to accurately predict possible substitution sites. The mechanism of ETD generates c/z-type ions rather than b/y-type ions (**Figure 1.16**), cleaves disulphide bonds, and retains structure features such as post-translational modifications which may be lost in other ionisation methods.

The most difficult glycopeptide analysis problems are those in which (a) small, even single sugar, attachments are present (and therefore are easily lost even in the ion source before CID), (b) situations in which completely novel structures are encountered and (c) there is a need to characterise large N-linked glycopeptides in complex mixtures, for which MALDI CID MS/MS can be particularly useful. This latter situation (c) was encountered during my training for this thesis research on first joining this laboratory, in work on the assignment of 24 N-linked glycosylation sites in virion-derived HIV-1 gp120 glycoprotein (Panico, Bouché et al. 2016) and the O-glycan training for this thesis work was also on the gp120 molecule (Stansell, Panico et al. 2015). Situations (a) and (b) above were both encountered during my thesis research and will be discussed in detail in chapters 3, 4 and 5.

1.5 Research Project Overview

The common theme linking the three projects illustrated in this thesis is the development and application of high sensitivity analytical techniques, as mass spectrometry, to challenging PTMs in both eukaryotic and prokaryotic organisms. Specifically, the research concentrates on the identification and structural characterisation of novel glycan structures in ADAMTS13 and in the bacterial pathogen *Clostridium difficile*.

1.5.1 Overview of ADAMTS13 Structural Biology

ADAMTS13 consists of a disintegrin and metalloprotease with a thrombospondin type 1 motif, (member 13). This large multi-domain protein regulates thrombogenesis by cleavage of an adhesive blood glycoprotein, von Willebrand factor (VWF) and so generating smaller and less thrombogenic fragments (Zheng 2013a). Originally, ADAMTS13 was studied for its association with a life threatening hematological disease, the Thrombotic Thrombocytopenic Purpura (TTP), which is directly linked to severe ADAMTS13 deficiency and characterised

by uncontrolled microvascular thrombosis that especially affects the cerebral and renal circulation (Moake 2002). Moreover, this plasma protease may also play a role in other hematologic pathologies, namely stroke and myocardial infarction (Zheng 2013b).

In addition to the ADAMTS13 metalloprotease and disintegrin domains, it also carries a thrombospondin type 1 repeat (TSR) just following those first two domains, then a Cysteine-rich domain and a spacer domain, terminating with additional TSRs within its protein sequence (Cal, Obaya et al. 2002). Specifically, ADAMTS13 includes an additional six TSRs and two CUB1 domains at its carboxyl end, and is the most divergent member of the family (Levy, Nichols et al. 2001, Soejima, Mimura et al. 2001, Zheng, Chung et al. 2001). Thrombospondin type 1 is itself a protein thought to play a role in cell adhesion and angiogenesis and TSRs include approximately 60 amino acids with conserved tryptophans and cysteines. In particular, a consensus sequence has been recognised as C-X-X-(S/T)-C-G which appears to act as a signal for O-fucosylation at S or T. Ricketts et al. have described the fucosylation at 6 of the consensus sites in recombinant ADAMTS13 but could show no evidence of fucosylation at TSR1. In radioactive labelling experiments these authors showed that the O-linked structures are disaccharides with a glucose- β 1,3-fucose structure, and they demonstrated that O-fucosylation of TSRs is by POFUT2, which is located in the ER and appears to play an important role in processing and secretion of ADAMTS13 (**Figure 1.21**) (Ricketts, Dlugosz et al. 2007).

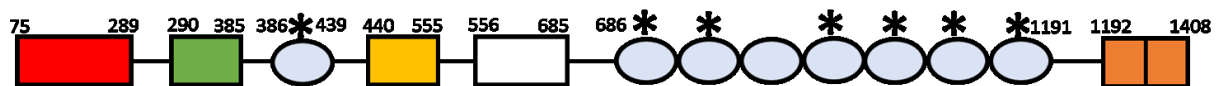


Figure 1.21 Domain structure of ADAMTS13: The domain structure of ADAMTS13 includes of a metalloprotease domain (red box), a disintegrin domain (green box), TSR1 (light blue oval), a cysteine rich domain (yellow box), a spacer domain (white box), seven additional TSRs and two CUB domains (orange boxes). Potential consensus sequences for O-fucosylation within the TSRs are indicated with *.

Moreover, it is known that there is a broad range of plasmatic ADAMTS13 activity in a healthy population (Mannucci, Capoferri et al. 2004). While the majority of investigative attention has been directed towards disease states in which ADAMTS13 activity is deficient, there is insufficient consideration for individuals whose ADAMTS13 plasma levels are higher than average and this may represent a compensatory mechanism against persistently

elevated levels of VWF in patients with venous thromboembolism (Mazetto, Orsi et al. 2012). In contrast, increased synthesis or enzymatic activity could also be caused from mutations or polymorphisms in the *ADAMTS13* gene, even though no known clinical disease or bleeding diathesis have been reported in association with excess ADAMTS13 activity so far. However, it has been seen that even minor decreases in the enzyme plasma level can be associated with increased likelihood of ischemic stroke. Therefore, these same genetic variants may be applied in the biotechnology industry with the aim to increase protein production while maintaining safety and efficacy.

There has been considerable general interest in the genetic factors controlling or influencing protein expression and/or secretion and/or activity, both within the molecular biology community and the biotechnology industry where the objective is the most efficient high-yield production of engineered proteins, including antibodies, cytokines and blood plasma proteins, such as ADAMTS13. To this end, a new field of research is examining codon usage during protein synthesis and why certain triplet codons are utilised to code for a particular amino acid rather than others, where a choice of up to six may exist. A single base change in the triplet codon leading to the same amino acid being coded for is called a silent or synonymous mutation. Many research works have now demonstrated that the conventional belief that synonymous mutations are universally “silent” is not correct and these polymorphisms can bear consequences on both protein expression and functionality in several cases through indirect means, such as modifying codon usage, and a growing body of these synonymous mutations have also been found in association with human disease (Gartner, Parker et al. 2013, Hunt, Simhadri et al. 2014, Supek, Miñana et al. 2014, Zheng, Kim et al. 2014). So far, almost 50 distinct diseases affecting various organ systems have been described in association with synonymous mutations (Sauna and Kimchi-Sarfaty 2011). In 2014, Supek et al. carried out comparative studies of 3,851 cancer exomes and over 400 genomes from 19 tumor types, which have revealed up to 1.3-fold enrichment of synonymous mutations in oncogenes, a phenomenon absent in tumor suppressor genes (Supek, Miñana et al. 2014). Because the protein product of a gene with a synonymous mutation nucleotide change is expected to possess an identical primary amino acid sequence to that of the WT gene, synonymous mutations may influence in more subtle ways than non-synonymous mutations, and for example, they may affect the efficiency of protein synthesis and/or disrupt protein structure via impaired translation kinetics (Komar 2007, Sauna and Kimchi-Sarfaty 2011) or affect certain post- or co-translational events. Although there is mounting evidence highlighting significant impacts by synonymous mutations on cellular function, many

biotechnology strategies for manufacturing protein-based drugs are grounded on the belief that such changes in the DNA sequence will be of no significance (Angov 2011, Supek, Miñana et al. 2014). A global gene optimisation strategy would consist of a much better understanding than currently available of the consequences of synonymous mutations on the structure, folding, secretion and function of target proteins, leading to the identification and characterisation of single synonymous substitutions that can substantially increase protein expression without carrying negative consequences for function.

Chapter 3 in this thesis describes a mass spectrometric study aimed at increasing that understanding for the case of a synonymous mutation in ADAMTS13 leading to increased secretion, and this work can be a template to aid in the development of predictive tools to locate “hot spots” within gene sequences amenable to optimisation and engender more precise and safe approaches for biotechnological gene design.

1.5.2 Overview of *Clostridium difficile*

Clostridium difficile is a Gram-positive, spore-forming, strictly anaerobic bacterium. It is the major cause of antibiotic-associated nosocomial diarrhoea in adults in Western countries and contamination by the resistant spores can provoke either asymptomatic carriage or clinical signs ranging from mild to severe diarrhoea to life-threatening pseudomembranous colitis (PMC) (Rupnik, Wilcox et al. 2009). In fact this pathogenic bacterium can exploit the disruption of the resident-intestinal flora, as a consequence of the administration of broad-spectrum antibiotics, both as prophylactics and to treat infection, to colonise and proliferate in the gut (Dethlefsen, Huse et al. 2008, Lawley, Clare et al. 2009).

The several known risk factors for *C.difficile* infections (CDI) are previous hospitalisation, underlying disease, advanced age (>65 years) and, most importantly, the use of antibiotics, even though newly affected populations include outpatients, like children and peri-partum women (Janoir 2016). All antibiotic classes can be related to CDI, but clindamycin, cephalosporins and fluoroquinolones are the most frequent mentioned (Slimings and Riley 2014). However, even low-risk antibiotics, i.e. trimethoprim and piperacillin-tazobactam, can predispose the patient to CDI, in particular where two or more courses of different antibiotics are prescribed and the cumulative damage to the intestinal microbiota could be sufficient to allow *C.difficile* to proliferate (Smits, Lyras et al. 2016).

Moreover, one of the key features of CDI is the high rate of recurrence, observed in 20%-30% of the cases: recurrences can either be relapses occurring with the initial strain or be caused by another strain (Rupnik, Wilcox et al. 2009).

The epidemiology of CDI has evolved over time and historically CDI was caused by several different strains, but since the early 2000s, the situation has changed as a consequence of the emergence of the so-called hypervirulent strains 027/NAP1/BI (where 027 refers to the PCR ribotype, BI refers to the restriction endonuclease group and NAP1 refers to the North American pulsotype) that spread from North America to Europe, becoming predominant in many Western countries (Janoir 2016). These strains are responsible for higher morbidity and mortality and nowadays CDI are caused by different multiple strains (Poxton 2013), including strains from emerging new lineages, such as the 078 PCR-ribotype strains found both in humans and in animals (Goorhuis, Bakker et al. 2008).

C.difficile pathogenesis consists in a three-step process starting with disruption of the gut microbiota (Smits, Lyras et al. 2016), followed by germination of indigenous or ingested spores, which starts the colonisation phase with the attachment of the bacterium to the host intestinal epithelium and its multiplication both at the surface and in the lumen. The final phase of virulence is the release of toxins and the onset of disease symptoms (Kirk, Banerji et al. 2016). The infection caused by this pathogen can also be facilitated by the release of two potent exotoxins, toxin A and toxin B, an enterotoxin and a cytotoxin respectively, produced by bacterial spores once germinated to form vegetative cells in the intestine (von Eichel-Streiber, Boquet et al. 1996, Calabi, Calabi et al. 2002). Furthermore, a key role in the mechanism of infection is occupied by the bacterial cell surface components, crucial in the interaction between the bacterium and the host, even though the molecular details of these interactions are still under investigations (Kirk, Banerji et al. 2016).

1.5.2.1 *Clostridium difficile* genome

In 2006, Sebahia and co-workers presented the first fully sequenced and annotated genome of *C.difficile*, the strain 630 RT012. This virulent and multidrug resistant strain was originally isolated in 1982 from a patient with PMC in Zurich, Switzerland. Its genome comprises a large circular chromosome of 4,290,252 bp (4.3 Mb), 3,776 putative protein-coding sequences (CDSs), a GC content of 29.06% and a plasmid, pCD630, of 7,881 bp containing 11 CDSs (Sebahia, Wren et al. 2006). Since then, many other genomes ranging in size from 4.1 to 4.3 Mbp have been entirely sequenced and annotated, such as B11 RT027 isolated in

the United States in 1988, R20291 RT027 isolated in the United Kingdom in 2006 and 2007855 RT027 isolated in France in 2007 (Stabler, He et al. 2009).

As in the case of strain 630, Whole-genome Sequencing (WGS) of these *C.difficile* strains have revealed much about the architecture of the *C.difficile* genome. In fact, *C.difficile* is characterised to have a highly dynamic and mosaic genome including a high proportion (~11% in strain 630) of mobile genetic elements. These comprise bacteriophages, group I introns, insertion sequences (IS), *sigK* intervening (skin) elements, clustered regularly interspersed short palindromic repeat (CRISPR)-*cas* elements, genomic islands and transposable and conjugative elements, together with an extensive range of accessory genes (Sebahia, Wren et al. 2006, Stabler, Valiente et al. 2010, Monot, Boursaux-Eude et al. 2011, Brouwer, Allan et al. 2012, Darling, Worden et al. 2014).

Several of the CDSs identified in the genome of *C.difficile* are related to adaptation and proliferation in the gastrointestinal tract (germination, adhesion and growth) and survival in challenging suboptimal environments (endospore formation) (Sebahia, Wren et al. 2006, Monot, Boursaux-Eude et al. 2011). All of these support the view that *C.difficile* occupies a highly dynamic niche and is able to coexist with its host for a long time (Sebahia, Wren et al. 2006).

The large and complex *C.difficile* genome, up to 42% larger than those of other closely related clostridial species, reflects its ability to survive even for long period within a diverse range of human, animal and abiotic environments (Knight, Elliott et al. 2015).

1.5.2.2 *Clostridium difficile* cell envelope and its components involved in host interactions

The colonisation step is characterised by the development of the bacteria in the colonic niche and comprises several features, such as bacterial evasion of host innate defences, adherence to epithelium mucosa and multiplication of vegetative cells (Janoir 2016). Bacterial cell surface components are crucial in the interaction between the bacterium and the host and *C.difficile* shows at its surface a large array of proteins (Wright, Wait et al. 2005). Recently, different studies have been conducted to better identify and characterise numerous cell wall polymers, as well as several surface proteins. The majority of these macromolecules are unique to *C.difficile*, so the cell envelope can be considered like an eligible target for the development of species-specific therapeutics (Kirk, Banerji et al. 2016).

A Peptidoglycan

Peptidoglycan (PG) is a key component of the cell wall with pleiotropic functions, such as maintenance of cell shape and integrity, and anchoring cell wall proteins (CWP). It presents a largely conserved structure, containing long glycan polymers cross-linked by short peptide chains. Polymers of the $\beta 1 \rightarrow 4$ linked disaccharide N-acetylglucosamine-N-acetylmuramic acid (GlcNAc-MurNAc) are the main constituents of the polysaccharide backbone and a short tetrapeptide branch, L-Ala-D-Glu-A₂pm-D-Ala (A₂pm: 2,6-diaminopimelic acid) is attached to the D-lactoyl group of MurNAc (Vollmer, Blanot et al. 2008, Peltier, Courtin et al. 2011). *C.difficile* shows a very high level of GlcNAc N-deacetylation (89-93%) comparing with that reported for other Gram-positive bacteria (Peltier, Courtin et al. 2011), while all MurNAc residues remain totally acetylated. Furthermore, the percentage of acetylated GlcNAc residues decreases more than twofold following the introduction of lysozyme, a crucial effector of the innate immune system that cleaves the PG backbone (Bevins and Salzman 2011, Ho, Williams et al. 2014).

The largely used antibiotic vancomycin inhibits cell wall synthesis through interaction with the terminal D-Ala-D-Ala of PG precursor (Johnson, Martin et al. 1987, Reynolds 1989), but resistance can be conferred via modification of these terminal residues (Kirk, Banerji et al. 2016).

B Secondary cell wall polysaccharides

C.difficile cell surface is characterised to have three anionic polymers. The first polysaccharide is PS-I, a branched penta-glycosylphosphate repeating unit that has originally been identified in a ribotype 027 strain and only found in a few strains (Ganeshapillai, Vinogradov et al. 2008). PS-II and PS-III consist in a polymer of hexaglycosylphosphate repeat units and a lipid bound glycosylphosphate polymer, respectively. They are more widely distributed compared to PS-I and have been discovered in all the strains investigated so far (Ganeshapillai, Vinogradov et al. 2008, Reid, Vinogradov et al. 2012). PS-I and PS-II have been illustrated as teichoic acid-like, even though they diverge from the simple glycerol phosphate or ribitol phosphate classic teichoic acids (Ganeshapillai, Vinogradov et al. 2008, Weidenmaier and Peschel 2008). Instead PS-III is a member of the extended lipoteichoic acid family (Percy 2014). Unfortunately, the biological significance of these polymers in *C.difficile* is still poorly understood, but PS-II has been confirmed to be the cell wall ligand that anchors members of the CWP family to the cell surface.

Thanks to their accessibility in the cell wall, these polymers may be considered potential vaccine targets. In fact, anti-PS-I antibodies can be seen in sera from healthy horses and chemically synthesised PS-I has been combined to a subunit of *C.difficile* toxin B to create a potential dual conjugate vaccine (Jiao, Ma et al. 2013), therefore demonstrating the potential of cell wall polysaccharides as vaccine candidates. Nevertheless, PS-II and PS-III may be more promising vaccine targets, since PS-I is not widespread between different *C.difficile* strains (Kirk, Banerji et al. 2016).

C *Clostridium difficile* S-layer

One structure involved in adhesion of *C.difficile* to enteric cells is the bacterial S-layer (surface layer), a proteinaceous two-dimensional paracrystalline array that completely coats the vegetative cell (Fagan, Albesa-Jove et al. 2009, Fagan and Fairweather 2014). *C.difficile* presents an S-layer with square-ordered lattice (Kawata, Takeoka et al. 1984), involving two distinct proteins, a high molecular weight (HMW) SLP (42-50 kDa) and a low molecular weight (LMW) SLP (22-38 kDa) (Cerquetti, Molinari et al. 2000) (**Figure 1.22**). These SLPs are generated by post-translational cleavage of a pre-protein (SlpA) encoded by *slpA* gene (Calabi, Ward et al. 2001), which is specifically located within a 36.6-kb cell wall protein (*cwp*) gene cluster (Sebahia, Wren et al. 2006). Furthermore, within the *cwp* gene cluster besides the *slpA* gene other cell wall proteins are encoded and in particular the *cwp84*, *cwp66*, *cwp2* and *secA2* constitute a 10 kb genetic region, denominated the S-layer cassette (SLC). The peculiarity is that different strains of *Clostridium difficile* present differences within the SLC cassette more than the rest of the *cwp* cluster (Dingle, Didelot et al. 2013).

Fagan and co-workers has established that SlpA comprises three identifiable subdomains: an N-terminal secretion signal, followed by the highly variable LMW region and finally the HMW region containing tandem cell wall binding 2 motifs (CWB2, PF04122) (Fagan and Fairweather 2011). The signal peptide conducts the translocation across the cell membrane via the accessory Sec system (Fagan and Fairweather 2011). Then, the SlpA protein is cleaved by the cell wall localised cysteine protease Cwp84 (Kirby, Ahern et al. 2009, Dang, Riva et al. 2010) to generate the two SLPs. The HMW SLP derives from the C-terminal region and it is relatively conserved among *C.difficile* strains; instead the LMW SLP derives from the N-terminal region and its sequence is more highly variable (Calabi, Calabi et al. 2002). HMW and LMW SLPs form a stable heterodimeric complex, also named H/L complex, via non-covalent interactions and they are able to self-assemble and so to constitute

the mature S-layer (Fagan, Albesa-Jove et al. 2009, Kirk, Banerji et al. 2016). The H/L complex shape, the positions of the interaction domains and the sequence variability of the LMW SLP suggest a model for orientation of the S-layer on the cell surface (Fagan, Albesa-Jove et al. 2009). The CWB2 motifs in the HMW SLP anchor the complex to the cell wall via an interaction with the PS-II (Willing, Candela et al. 2015) to display the LMW SLP on the cell surface. Unfortunately, the mechanism that leads to the development of the mature S-layer is still not clear and it is believed that S-layer self-assembly is a thermodynamically driven process (Chung, Shin et al. 2010) and some SLPs present a different crystallisation domain that resolves lateral interactions in the array (Smit, Jager et al. 2002).

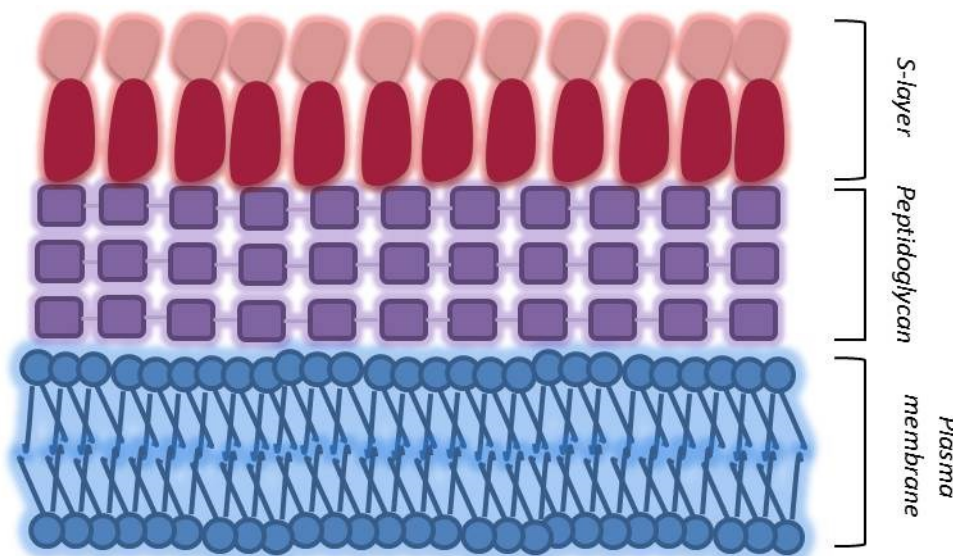


Figure 1.22 Scheme of cell envelopes of Gram-positive bacteria. The S-layer is attached to the peptidoglycan layer and is anchored to the the peptidoglycan via lipoteichoic acids, since they have a lipid component that can assist in anchoring as the lipid component is embedded in the plasma membrane.

The S-layer is implied in pathogen-host interactions crucial to pathogenesis and it has been proved that the SLPs promote *C.difficile* adhesion to culture cell lines and adhere to both gastrointestinal tissues and many extracellular matrix components (Takumi, Koga et al. 1991, Calabi, Calabi et al. 2002, Spigaglia, Barketi-Klai et al. 2013). The S-layer has also been involved in host immune activation via TLR4 (Ryan, Lynch et al. 2011).

Even though the SlpA amino acid sequence has been discovered to be highly conserved in some human *C.difficile* isolates of the same PCR-ribotype (Ní Eidhin, Ryan et al. 2006, Spigaglia, Galeotti et al. 2011), it has been recently demonstrated by Dingle and co-workers

that SlpA may be extremely variable between *C.difficile* strains (Dingle, Didelot et al. 2013) in fact, twelve highly divergent S-layer cassette variants have been found, investigating more than 800 clinical isolates from several hospitals in Oxford and Leeds, and these cassettes can randomly associate with different genotypes by homologous recombination events (S-layer switching). Interestingly, one of these SLCs, the cassette n° 11, was found in 13% of the isolates and presents an unusual genotype if compared with the strain 630: a shorter SlpA protein, a rearrangement of *cwp66* and *CD2790* and *cwp2* is missing. Unexpectedly, the SLC 11 contains a novel 23.8 kb, 19 open reading frames (ORFs) polysaccharide synthesis gene cluster, inserted in place of *cwp2* (**Figure 1.23**). It has been predicted that this 19 ORFs cluster is capable to encode a putative S-layer glycosylation mechanism as it presents a homology to another S-layer glycosylation cluster belonging to a non-pathogenic Gram-positive species, *Geobacillus stearothermophilus* (Dingle, Didelot et al. 2013). This research by Dingle and co-workers in Oxford constitutes the first description of putative S-layer glycosylation in a Gram-positive human pathogen or in any Clostridium species, in fact, to date, neither genetic nor phenotypic evidence of S-layer glycosylation had been found in *C.difficile* (Qazi, Hitchen et al. 2009).

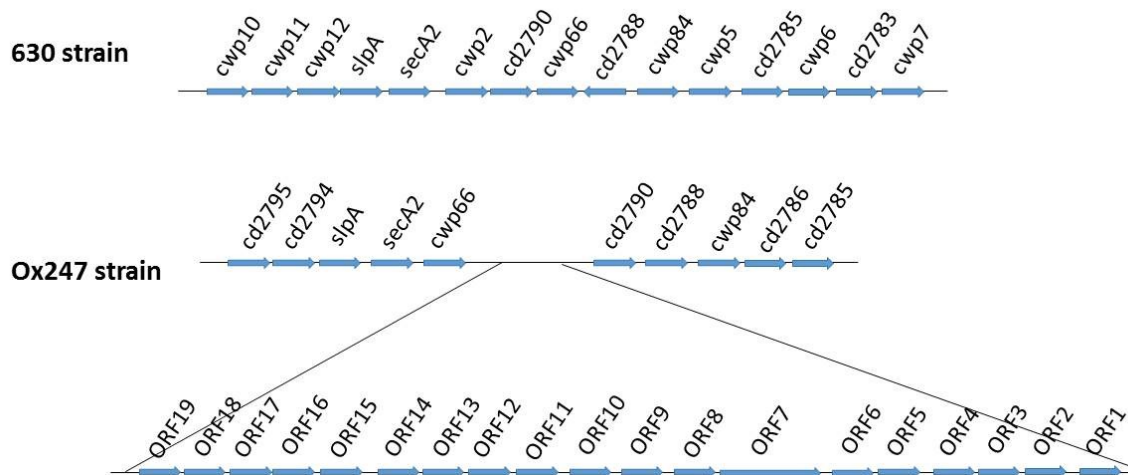


Figure 1.23 The cwp cluster of strains 630 and Ox247 are slightly different. The Ox247 strain lacks cwp2 and the cwp66 and cd2790 have been rearranged. This strain also includes a novel insertion of 23.8 kb containing 19 predicted open reading frames (ORFs) within one of the S-layer cassettes, SLC 11, between the cwp66 and the cd2790.

The gene cluster of this post-translational modification is usually located near the S-layer gene (Ristl, Steiner et al. 2011). Therefore, in *C.difficile* S-layer glycosylation may significantly modify the properties of the outer cell surface, with potentially crucial effects on

antigenicity, antibiotic permeability and virulence (Dingle, Didelot et al. 2013). Bioinformatic analysis suggests that this novel insertion encodes enzymes that could initiate, extend, membrane translocate and ligate a glycan chain to a substrate molecule. The first step, the initiation, comprises the transfer of a galactose from its nucleotide-activated form (UDP-Gal) to a lipid carrier (undecaprenyl phosphate). Then, several deoxysugars, specifically rhamnoses, are transferred from dTDP- β -L-Rha to the lipid-bound by action of four different rhamnosyltransferases, named WsaC through WsaF, which are responsible of the growing glycan chain. The termination of S-layer glycan chain elongation is by 2-O-methylation of the last sugar on the N-terminal. Finally, an ABC transporter system is responsible for binding of the 2-O-methylated glycan chain and then exports it through the plasma membrane (Steiner, Novotny et al. 2008). As some of the genes involved in this model are homologues to genes present in *Clostridium difficile*, including Ox247, this suggests a putative role of SLC-11 in S-layer glycosylation. In addition, to encoding all the components which characterise an S-layer glycosylation cluster, its 19 ORFs also encode a rhamnose biosynthesis pathway. However, with the exception of the putative OTase which has been shown by the Wren group at the London School of Hygiene and Tropical Medicine to express an active enzyme (personal communication), nothing was known about the other glycogenes or the S-layer glycan structures prior to the start of this thesis work.

D Cell wall protein family

C.difficile adheres to different cell lines such as Vero cells, Hela cells, Hep-2 cells, enterocyte-like Caco-2 cells, HT-29 and mucus-producing HT-29-MTX cells (Karjalainen, Barc et al. 1994, Naaber 1996, Drudy, apos et al. 2001) and this adhesion needs bacterial adhesins that are basically surface proteins. Many *C.difficile* adhesins have been characterised even though their role in the pathogenesis of CDI is still for some of them unclear (Janoir 2016). Among them, several belong to the Cwp family, a family of paralogous surface-associated proteins.

The *C.difficile* Cwp family comprises 29 members, all containing three tandem copies of the CWB2-anchored surface proteins (Kirk, Banerji et al. 2016). The similarities to the arrangement of CWB2 motifs within SlpA and the Cwps let speculate a similar mechanism responsible for anchoring these proteins to the anionic polymer PS-II in the *C.difficile* cell wall (Willing, Candela et al. 2015). In 2014, Fagan and Fairweather demonstrated that, beyond the presence of the CWB2-anchoring domain, many Cwps also include an additional

domain able to functionalise the S-layer (Fagan and Fairweather 2014). However, only a small number of these Cwps have been fully characterised so far, but several have been shown to cover critical roles in the interaction between *C.difficile* and the host (Kirk, Banerji et al. 2016).

1.5.2.3 *Clostridium difficile* flagella

C.difficile is motile through the presence of a flagellar apparatus. Until recently *C.difficile* flagella have chiefly been involved in the colonisation of the host (Tasteyre, Barc et al. 2001), although its contribution to the pathogenesis is still not fully understood due to its complexity. More than one scientific contribution highlights a direct role in virulence by flagellin, and not only because it provides force-driven motility towards nutrients available in the gut, but also for its ability to modulate toxin expression (Aubry, Hussack et al. 2012, Baban, Kuehne et al. 2013).

A Genetic organisation of *C.difficile* flagellar genes

C.difficile 630 and R20291 (BI/NAPI/027, hypervirulent ribotype) share several genes and proteins implicated in the structural integrity and motion of flagella in other bacterial species (Stevenson, Minton et al. 2015). Generally, activation of flagellar gene transcription is controlled by the master transcriptional regulator FlhDC (Anderson, Smith et al. 2010); however, many alternative master regulators have been found, like CtrA, VisNR, FleQ, FlrA, FlaK, LafK, SwrA and MogR (Smith and Hoover 2009). Genes involved in the motility are expressed at higher levels during the exponential growth phase and are negative controlled by the alternative sigma factor SigH (Soutourina and Bertin 2003, Derman, Söderholm et al. 2015), but a specific master regulator of *C.difficile* flagellar expression has not yet been characterised and this brings the possibility that flagellar gene activation in this pathogen might be governed by transcriptional regulators involved in other cellular processes (Stevenson, Minton et al. 2015).

The assembly of *C.difficile* 630 flagella is regulated by genes organised in three operons, named F1, F2 and F3 regulons (Stabler, He et al. 2009), very variable among lineages of *C.difficile*. The hierarchy of transcription of these flagellar genes begins with the early-stage genes located in the F3 region. The F2 region is located between F3 and F1 and in *C.difficile* 630 and strains in the ribotype 017 lineage comprises four flagellar-biosynthetic glycan genes

(Twine, Reid et al. 2009, Faulds-Pain, Twine et al. 2014). This regulon is responsible for post-translational modification of flagella, shown to be essential for functional flagella assembly and motility of *C.difficile* (Twine, Reid et al. 2009, Faulds-Pain, Twine et al. 2014). Finally, the late-stage flagellar genes are encoded in the F1 regulon and are orthologous to class III flagellar genes of other bacteria (Aldridge and Hughes 2002, Anderson, Smith et al. 2010). Many hypervirulent ribotype 027 strains present divergence of the F1 region (Stabler, Gerding et al. 2006, Stabler, He et al. 2009) and this diversity is believed to explain the differences in motility between *C.difficile* 630 and R20291 (Stabler, He et al. 2009).

B Flagellar glycosylation in *C.difficile*

In 2009, Twine and co-workers (Twine, Reid et al. 2009) demonstrated that the flagellin of *C.difficile* could be modified by O-linked glycan moieties. They highlighted that the flagellin glycans of *C.difficile* were genetically different not only from other species of Clostridia, i.e. *Clostridium botulinum*, but also among *C.difficile* strains. The flagellin of *C.difficile* 630 could be glycosylated with an N-acetyl hexosamine (HexNAc) residue at up to seven sites; in contrast, flagellin glycosylation in strains of the 027 ribotype (BI1 and BI7) was more complicated. Twine and co-workers suggested that flagellin from 027 ribotypes was modified via O-linkage to heterogenous glycans of up to five monosaccharide residues with masses of 204 (HexNAc), 146 (deoxyhexose), 160 (Methylated deoxyhexose) and 192 (Heptose) but an overall structure remained elusive, until the work carried out in this thesis. Moreover, further identification and characterisation of a specific glycosyltransferase gene (CD0240) showed that it was involved in the glycosylation process and its inactivation led to loss of the surface-associated flagellin protein rendering the strain non motile, even though the strain was still able to produce truncated polymerised flagella filaments (Twine, Reid et al. 2009, Faulds-Pain, Twine et al. 2014).

Following whole-genome comparative studies, it was shown that *C.difficile* strains 630, R20291 CD196 (027 ribotype) and M120 (078 ribotype) present genetic differences in the F2 region (Stabler, He et al. 2009), probably translating into differences in autoagglutination. Autoagglutination of flagellin results in changes in antigenic specificity due to differential glycosylation of flagellin (Guerry, Ewing et al. 2006) and the divergence of the F2 region together with the differences in agglutination highlighted in Stabler's study could be correlated with the fact that distinct spectra of flagellin glycosylation patterns occur between *C.difficile* 630 and 027 strains, as observed by Twine and coworkers (Twine, Reid et al.

2009). Nevertheless, in *C.difficile* changes in agglutination may not directly be linked to changes in flagellin glycosylation, as it is a multilevel process and likely comprises other surface proteins (Stabler, He et al. 2009).

Additional study has been focused on characterisation of the genes from the F2 region of *C.difficile* 630 and RT017 strains and individual mutants in three of the four PTM genes belonging to the F2 region (CD0241, CD0242 and Cd0244) result in loss of motility and display a sedimentation phenotype in *vitro* (Faulds-Pain, Twine et al. 2014) much more extreme than that presents in the flagellin (*fliC*) mutant. In agreement with what had been observed by Twine et al in 2009, mutants in these genes are able to produce flagellin, even though differences in flagellin molecular weight have been found for each one. These size differences have been thought to be due to changes in flagellin PTM originated by disruption of the genes (Faulds-Pain, Twine et al. 2014). Therefore, both the structural integrity and PTMs of flagella, including those studied here, are fundamental for the motility of *C.difficile* (Stevenson, Minton et al. 2015).

1.6 **Project aims**

The overall aim of this study is to apply and improve both current and novel mass spectrometric glycoproteomic strategies whilst characterising biologically important glycoproteins. The specific aims of each project are:

➤ **ADAMTS13**

To study the comparative ADAMTS13 O-glycomes of wild type, synonymous and non-synonymous mutants in order to help answer the question on causation of enhanced secretion seen in the synonymous mutant.

➤ ***Clostridium difficile* flagellin**

To characterise the unique flagellar glycan structure of the bacterial pathogen *Clostridium difficile* using various mass spectrometric techniques.

➤ ***Clostridium difficile* S-layer**

To characterise the S-layer glycosylation in *Clostridium difficile* Ox247 ribotyping 5 using a full battery of advanced MS techniques, including ETD.

Chapter 2:
Materials & Methods

2. Materials & Methods

2.1. Materials

- **Acros Organics** (Geel, Belgium): 3,4-diaminobenzophenone (DABP);
- **Alfa Aesar** (Lancashire, UK): acetic anhydride ((CH₃CO)₂O) and methyl iodide (CH₃I);
- **Applied Biosystems** (Warrington, UK): 4700 Mass Standards kit, comprising the peptide standards: bradykinin, fragment 1-8, angiotensin 1, adrenocorticotrophic hormone fragment (ACTH) 1-17, ACTH 18-39, ACTH 7-38 (**Table 2.1**);
- **BOC** (Guildford, UK): nitrogen and argon gases;
- **ELGA LabWater** (High Wycombe, UK): ultrapure 18 MΩ·cm³ distilled/deionised (Milli-Q) water from PURELAB Option-Q water purification system was used for all aqueous solutions;
- **Fluka** (Poole, UK): ammonium bicarbonate (CH₅NO₃), dithiothreitol (DTT), Dowex[®] resin, pyridine (C₅H₅N) and potassium hydroxide (KOH);
- **Invitrogen** (Paisley, UK): NuPAGE 4-12% Bis-Tris Gel 1.0 mm X 10 well, NuPAGE[®] MOPS SDS Running Buffer [20X], NuPAGE[®] Antioxidant, NuPAGE[®] LDS Sample Buffer [4X], NuPAGE[®] Sample Reducing Agent, BenchMark[™] Protein Ladder, Stainer A and Stainer B (Colloidal Blue Stain Kit);
- **Membrane Filtration Products, Inc.** (Texas, USA): CelluSep regenerated cellulose tubular membrane (MWCO: 12,000-14,000);
- **Pierce** (Rockford, USA): tri-Sil 'Z' Derivatizing agent (TMS), Snakeskin[®] dialysis tubing (7 kDa molecular weight cut off (MWCO));
- **ROMIL** (Waterbeach, UK): acetonitrile (ACN), ammonia (NH₃), acetic acid (CH₃COOH), chloroform (CH₃Cl), dimethylsulfoxide (DMSO), formic acid (HCO₂H) methanol (CH₃OH), propan-1-ol (C₃H₇OH) sodium hydroxide pellets (NAOH) and trifluoroacetic acid (TFA);
- **Sigma-Aldrich Corporation** (Poole, Dorset, UK): α-cyano-4-hydroxycinnamic acid (4-HCCA), hexanes (C₆H₁₄), hydrofluoric acid 48 wt. % in water (HF), iodoacetic acid (ICH₂CO₂H), iodomethane-d₃ (CD₃I), monosaccharide standards (arabinose, arabitol, fucose, galactose, glucose, inositol, mannose, rhamnose, xylose), m-nitrobenzyl alcohol

(m-NBA), ponceau-S red staining, potassium borohydrate (KBH₄), sodium borodeuteride (NaBD₄) and trypsin (porcine EC 3.4.21.4);

- **Waters** (Massachusetts, USA): Waters TOF GS2 Sample Kit-1, including the peptide standards [Glu¹]-fibrinopeptide B and Leucine Enkephalin.

Peptide fragment	Amino acid sequence	Predicted molecular weight (Da)
Bradykinin Clip 1-8	PPGFSPFR	903.46
Angiotensin 1	DRVYIHPFHL	1295.67
[Glu ¹]-fibrinopeptide B	EGVNDNEEGFFSAR	1569.66
Adrenocorticotrophic hormone (ACTH) Clip 1-17	SYSMEHFRWGKPVGKKR	2092.07
ACTH Clip 18-39	RPVKVYPNGAEDESAAEAFPLEF	2464.19
ACTH Clip 7-38	FRWGKPVGKKRRPVKYPNGAEDESAAEAFPLE	3656.92

Table 2.1 Masses and sequences of the peptides contained in the calibration peptide mix used for MALDI-MS and MS/MS.

2.2. Equipment & Consumables

Equipment:

- ABSciex 5600 mass spectrometer linked to an UltiMate 3000 Series autosampler;
- API Q-STARTM Hybrid LC-MS/MS system coupled to a Dionex/LC Packings LC nanocapillary column (15cm x 75µm C₁₈);
- Applied Biosystems 4800 MALDI-TOF-TOFTM mass spectrometer;
- Bruker SCIONTM SQ 456-GC mass spectrometer fitted with a 15 m 0.25 mmID 0.25µm bucker br-5ms column;
- Dionex, Sunnyvale, USA UltiMate 3000 with a probot spotter and fitted with a Pepmap analytical C₁₈ nanocapillary (75 µm ID x 15 cm length) for offline-nanoLC separation prior to MALDI-TOF/TOF analysis;
- Waters Synapt G2-S coupled with a nanoACQUITYTM LC;

- Xevo G2 Q-TOF LC-MS/MS mass spectrometer on-line to a Waters Acquity UPLC microbore reverse phase column (1 mm x 50 mm C₁₈).

Consumables:

- **Bennett Scientific** (Devon, UK): culture tube caps;
- **Fisher Scientific** (Loughborough, UK): 10 ml glass syringes;
- **Sigma-Aldrich** (Poole, UK): Lo-bind[®] eppendorf tubes;
- **Thermo Scientific** (Basingstoke, UK): Thermo Savant SPD121P Speed Vac connected to a RVT4104 refrigerated vapour trap and Edwards XDS10 pump for volume reduction and concentration, Thermo Savant ModuloD freeze dryer for lyophilisation, IEC Centra CL3 centrifuge for centrifugation;
- **VWR International Ltd** (Leicestershire, UK): automatic multi-tube vortexer used for shaking, screw cap glass culture tubes (13x100 mm corning, 7.5 ml), eppendorfs, glass Pasteur pipettes (150 mm and 230 mm);
- **Waters Ltd** (Hertfordshire, UK): reverse phase Classic C₁₈ (360 mg, 55-150 µm), Oasis[®] HLB, C₁₈ (Plus) Sep-Pak cartridges.

2.3. Sample preparation

All biological samples used in the projects discussed in this thesis were prepared by research collaborators, and specifically ADAMTS13 samples by United States Food and Drug Administration (FDA), *Clostridium difficile* flagellin samples by Professore Brendan Wren's group (LSHTD) and *Clostridium difficile* S-layer samples by Professor Neil Feirweather's group (Imperial College London).

2.4. Methods

2.4.1 Purification of *Clostridium difficile* S-layer proteins

For the removal of contaminants, such as culture media, reagents and salts, the *C.difficile* S-layer sample was dialysed against 2% acetic acid solution in water and then only against water using a regenerated cellulose tubular membrane (Cellu Sep MWCO 12,000-14,000).

2.4.2 SDS-PAGE electrophoresis of ADAMTS13 and *Clostridium difficile* S-layer proteins

ADAMTS13 and *Clostridium difficile* S-layer proteins were separated by SDS-PAGE. 4 µg and 20 µg of each sample respectively were diluted with NuPAGE[®] LDS Sample Buffer [4X], NuPAGE[®] sample reducing agent [10X] and deionised water. A pre-stained molecular weight protein marker (Invitrogen, Renfrew, Paisley, UK) was run in parallel. Samples were heated for 10 minutes at 70°C then loaded onto a NuPAGE 4-12% gel (Invitrogen, UK). Electrophoresis was carried out at 200 eV constant in a running buffer for 50 minutes (NuPAGE[®] MOPS SDS running buffer [20X] plus NuPAGE[®] Antioxidant). Protein detection was performed by colloidal blue stain kit (stainer A and stainer B).

2.4.3 In-gel digestion of protein bands and elution of peptides

Each gel band of interest (if not provided) was excised into approximately 1 mm x 1 mm sections. The gel pieces were de-stained by addition of 200 µl of ammonium bicarbonate (50 mM, pH 8.4) followed by 200 µl of acetonitrile (ACN) for 15 minutes. The supernatant was discarded and the gel pieces were dried on a heatblock. If cysteine residues in the amino acid sequence of the protein in question are likely to be present, disulphide bridges will generally need to be reduced (10 mM dithiothreitol at 56 °C for 30 minutes) and carboxymethylated (55 mM iodoacetic acid at room temperature for 30 minutes). The samples were dissolved in 15 µl of ammonium bicarbonate (50 mM, pH 8.4) and digested with a 25 ng/µl working solution of acidified grade modified trypsin (REF V5111, Promega) in Ammonium bicarbonate buffer (50 mM, pH 8.4) for 14-16 hr at 37 °C. After the reaction, the supernatant from the gel pieces was transferred to a 0.5 ml Lo-bind[®] eppendorf tube. 50 µl of 0.1% (v/v) trifluoroacetic acid (TFA) was added to the gel pieces and incubated at 37 °C for 10 minutes to halt the digestion, followed by the addition of 100 µl of ACN and incubation for further 15 minutes. The supernatant was pooled with the previously obtained supernatant. This process was repeated once and the (glyco)peptide extract was reduced in volume to approximately 10 µl and stored at 4 °C. Peptides were then solubilised in 30 µl of 0.1% formic acid to be processed with different mass spectrometric strategies.

2.4.4 In-solution digestion

An in-solution trypsin digestion was carried out on *Clostridium difficile* flagellin. The sample was incubated at 37 °C overnight in 50mM ammonium bicarbonate buffer at pH 8.4 at a protease:ratio of 1:20 (w/w). Trypsin hydrolyses the peptide bonds at the carboxyl side of lysine and arginine residues, unless they are followed by a proline residue.

2.4.5 Electrotransfer: tank transfer

The transferring of proteins from a polyacrylamide gel (SDS-PAGE) onto on Immobilon PVDF transfer membrane was achieved by using a tank transfer system. First of all sponges and filter paper had been immerse into transfer buffer (5% (v/v) Novex NuPAGE 20X, 1% methanol, 0.001% Antioxidant, 200 mg/l SDS). Then the filter paper has been placed onto the SDS gel and then the gel was turned over with the filter paper side down. Afterwards, the membrane was wet in methanol for 15 seconds, then it was soaked in Milli-Q water for 2 minutes and finally equilibrated in transfer buffer for at least 5 minutes. Later the transfer task was assembled: at the bottom of the cassette a foam pad was placed, one sheet of filter paper on top of it, then the gel was laid on the filter paper and on the top of the gel the membrane, a second sheet of filter paper and a second foam pad in this order. The cassette holder was located into the transfer tank covered with adequate buffer and the electrotransfer was carried out at 30 eV constant for 60 minutes. Protein detection was performed by ponceau S-red staining solution.

2.4.6 Release of O-glycans from *Clostridium difficile* S-layer glycopeptide

Various O-glycosidase, specific to certain core types, are available and capable to release some O-glycans, but unfortunately, these conserved linkages to the peptide backbone are not found in prokaryotic glycoproteins and so not applicable for the study of the O-glycosylation of the S-layer glycoprotein of *Clostridium difficile* Ox247. As a consequence of the specificity of the O-glycosidase enzyme, most O-glycans in eukaryotes are chemically released by way of alkaline β -elimination (Dell 1993). The following methods were employed in this thesis in order to release the glycans from the protein for further analysis.

2.4.6.1 Reductive elimination and purification of O-glycans

The β -elimination reaction is carried out under reducing condition to prevent glycan degradation by the “peeling” effect of reducing the terminal/linking residues to their alditol forms and so is usually referred to as reductive elimination. A lyophilised aliquote of *Clostridium difficile* Ox247 sample was dissolved in 400 μ l of 55 mg/ml potassium borohydride (KBH₄) in a 0.1 M potassium hydroxide (KOH) solution and incubate at 45°C for 20-24 h. The reaction was terminated by dropwise addition of glacial acetic acid.

2.4.6.2 Ion exchange clean-up

An ion-exchange Dowex[®] (50W-X8 (H⁺) 50-100 mesh) column was assembled to remove cationic salts, amino acids and peptides. The Dowex cation exchange column was built using a pasteur pipette plugged at the tapered end with a small amount of glass wool, where a piece of silicon tubing was placed with a flow-blocking adjustable clip. The Dowex conditioned with 20 ml of 5% (v/v) acetic acid and then the sample was loaded and eluted with 3 ml of 5% (v/v) acetic acid. The collected eluate was lyophilised and the excess of borates resulting from the reductive elimination were removed by co-evaporation with 4 x 500 μ l of a 10% (v/v) acetic acid in methanol under a stream of nitrogen.

2.4.7 Hydrofluoric acid (HF) hydrolysis

Hydrofluoric acid treatment was carried out to hydrolyse any phosphodiester linkages that might be present in the glycan chain decorating the *Clostridium difficile* Ox247 S-layer protein. The sample was transferred to Lo-bind[®] eppendorf tubes and the dried sample was incubated with 50 μ l of hydrofluoric acid (HF) on ice for 20 hrs at 4°C. The reaction was terminated by drying under a gentle stream of nitrogen and afterwards the glycoprotein was derivatised prior to analysis by mass spectrometry (see **section 2.4.8.1**).

2.4.8 Chemical derivatisation of carbohydrates for MS analysis

As discussed in Chapter 1, native glycans can be analysed directly by mass spectrometry, but do not ionise as efficiently due to their hydrophilic nature and are thus derivatised before

being analysed by mass spectrometry. Derivatisation methods can either employ a reducing end tag or protect most or all of the functional groups of the carbohydrates.

2.4.8.1 Permethylation

Permethylation includes the exchange of protons in hydroxyl and amide groups for hydrophobic methyl groups and involves the successive base-catalysed ionisation of these functional groups followed by methylation. This technique was initially introduced by Hakomori and colleagues (Chou 1979) who described the use of methylsulphinyl carbanion from dimethyl sulphoxide (DMSO) as the base and methyl iodide as the methyl donor and then the chemistry was optimised for high sensitivity application to peptides and carbohydrates by the discovery in isotope dilution MS experiments that the kinetics is on the second timescale rather than the hours and days originally used (Morris 1972). This led to all subsequent methods utilising permethylation times of minutes in order to minimise byproducts and maximise sensitivity. Permethylation is now commonly catalysed by weaker bases such as sodium hydroxide and has become the staple derivatisation step in many glycomic strategies (Dell 1990).

For specific oligosaccharide analysis, samples were methylated using the sodium hydroxide procedure. Five pellets of NaOH were ground with ~3 ml of DMSO in a pestle and mortar. 1 ml of the resulting slurry was added to the lyophilised samples, followed by 0.6 ml of methyl iodide. The reaction mixture was vigorously mixed on an automatic multi-tube vortexer for 40 minutes at room temperature. Afterwards, the reaction was quenched by dropwise addition of water, while constantly shaking the tube. The permethylated glycans were extracted in 1ml of chloroform and washed several times with 4ml of Milli-Q water (4x) and then the chloroform layer was finally dried under a gentle stream of nitrogen gas. The resulting glycans were finally purified by Sep-Pak[®] C₁₈ using the aqueous ACN system.

2.4.8.2 Deutero-permethylation

The sample was deuteropermethylated using the sodium hydroxide procedure. A slurry of NaOH-anhydrous DMSO (1 ml) was added to the sample, followed by 0.6 ml deuterated iodomethane-d₃. The reaction mixture was agitated on an automatic multi-tube vortexer for 40 minutes at room temperature and quenched by dropwise addition of water. The

deteuropermethylated glycans were extracted in 1 ml of chloroform, washing with 4 ml of water (4x) and dried under a gentle stream of nitrogen gas. The resulting glycans were then purified by Sep-Pak[®] C₁₈ using the aqueous ACN system as described in **section 2.4.8.3**.

2.4.8.3 Sep-Pak[®] C₁₈ purification of derivatised glycans

Purification of the permethylated sample was achieved by reverse-phase chromatography using C₁₈ Sep-Pak[®] cartridges. Sep-Pak[®] classic cartridges were attached to a 10 ml glass syringe (Waters corporation, UK) and conditioned successively with 5 ml of methanol, Milli-Q water, ACN and 15 ml of Milli-Q water. The sample was dissolved in 1:1 (v/v) methanol: Milli-Q water (200 µl) and loaded onto the cartridge. The sample was washed with 5 ml of Milli-Q water and then eluted stepwise with 3 ml of each 15%, 35%, 50%, 75% and 100% (v/v) aqueous ACN solution. The fraction volumes were reduced on the SpeedVac[®] and lyophilised.

2.4.8.4 Alditol acetates

For compositional analysis by alditol acetate, the non-permethylated samples were hydrolysed (2M TFA, 121°C, 2 hrs), dried under nitrogen and the monosaccharides reduced (10 mg/ml NaBD₄ in 2M aqueous NH₃, room temperature, 2 hrs). After borate removal with acidified methanol co-evaporation, they were re-acetylated (acetic anhydride, 100°C, 1 hr), purified by washing with chloroform and water (4x) and prepared for GC-MS by suspension in hexane.

2.4.8.5 Linkage analysis

Linkage information was achieved by the analysis of partially methylated partially acetylated alditol acetate derivatives. Samples were permethylated as described followed by hydrolysis (2M TFA, 121°C, 2 hrs) and reduction (10 mg/ml NaBD₄ in 2 M aqueous NH₃, room temperature, 2 hrs). After borate removal with acidified methanol co-evaporation, the monosaccharide samples were re-acetylated (acetic anhydride, 100°C, 1 hr), purified by washing with chloroform and water (4x) and prepared for GC-MS by suspension in hexane (Sweet 1974).

2.4.9 Mass spectrometric analysis

The mass spectrometric instrumentations employed for the screening analysis are ABSciex 5600 mass spectrometer linked to an UltiMate 3000 Series autosampler, API Q-STAR Pulsar coupled to a Dionex nano LC, a 4800 MALDI-TOF/TOFTM Analyser Applied Biosystems, Bruker SCIONTM SQ 456-GC mass spectrometer, Ultimate 3000 LC Packings Dionex, Xevo G2 Q-TOF LC-MS/MS and a new generation Q-TOF instrument, the Waters Synapt G2-S coupled with a nanoACQUITYTM LC.

2.4.9.1 GC-MS

GC-MS analysis was carried out on a Bruker SCIONTM SQ 456-GC instrument. The samples were dissolved in hexane and injected on a br-5ms column (15 m x 0.25 mm internal diameter). For analysis of the alditol acetate derivatives, the oven was held at 60°C for 1 min and increased to 190°C at 20°C min⁻¹, from where the temperature is increased to 230°C at 1°C min⁻¹. The final temperature increment is to 300°C raised at 25°C min⁻¹ and held for a total of 5 mins. The programme for the linkage analysis is started at 60°C for 1 min and increased to 300°C in a single ramp at 8 °C min⁻¹ and held for 5 minutes.

2.4.9.2 Nanospray MS and MS/MS with API Q-STAR ABSciex 5600 and Xevo G2 mass spectrometers

In nanospray mass spectrometry experiments, the sample, in approximately 2 µl of 5% acetic acid/30% acetonitrile, is loaded into a borosilicate needle which itself is coated in a conductive material, usually gold. The needle is placed in a holder which carries the spray-inducing voltage and placed opposite a counter-electrode with several thousand volts potential difference, at the entrance to the quadrupole analyser of the Q-TOF. This device is useful for creating good quality spectra from small sample quantities by acquiring data over a relatively long period of time (the spray time) of 15 to 30 minute compared to the few seconds to a minute of acquisition time which is normal for an on-line LC-MS run.

2.4.9.3 On-line LC-ES MS and MS/MS with an API Q-STAR ABSciex 5600 and Xevo G2 mass spectrometers

Tryptic digests of glycopeptides were analysed by online liquid chromatography and were resuspended in 25 μ l 0.1% (v/v) formic acid and analysed by nano-LC-ES-MS/MS using a reverse-phase nano-HPLC system (LC Packings, Dionex) connected to a quadrupole TOF mass spectrometer (API Q-STAR® Pulsar, Applied Biosystem/MDS Sciex, ABSciex 5600 and Xevo G2). Online peptide separation was achieved by a binary nano-HPLC gradient generated by an UltiMate pump fitted with a Famos autosampler and a Switchos microcolumn switching module (Dionex Ltd.) utilising a pre-microcolumn C₁₈ cartridge and an analytical C₁₈ nanocapillary (15cm x 75mm internal diameter, PepMap). The digests were first loaded onto the precolumns and eluted with 0.1% formic acid prior to the transferral onto the analytical nanocapillary HPLC column and eluted using a gradient of Solvent A (0.05% (v/v) formic acid in a 99.5% (v/v) ACN mixture). The instrument was pre-calibrated using 100 fmol/ μ l of [Glu¹]-fibrinopeptide B human in 0.1% TFA in 30% (v/v) aqueous acetonitrile solution. In the MS/MS mode, the collision gas used was argon and the pressure was maintained at 3.5×10^{-6} Torr. Data were acquired and processed by using the Analyst QS software with the automatic information-dependent acquisition (IDA) function. In Q-TOF experiments on both ABSciex 5600 and Xevo G2 similar procedures were used. Interpretation of data was made manually/visually. Preparation of the spectra annotated with structural assignments was carried out manually with CorelDraw X3 Graphics Suite Software.

2.4.9.4 Off-line LC-MALDI-TOF and TOF/TOF MS/MS

Clostridium difficile Ox247 peptides were analysed by *off-line* liquid chromatography/matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (*off-line* LC-MALDI-TOF MS). Peptides obtained were separated using an Ultimate 3000 LC system Dionex, fitted with an analytical Pepmap C₁₈ nanocapillary column (15 cm length, 75 mm internal diameter). The digests were loaded onto the column and eluted using solvent A (0.1% (v/v) trifluoroacetic acid, TFA, in 2% ACN) and solvent B (0.1% TFA in 90% ACN), in the following gradient: 0-60% solvent B (0-36 min), 60-90% solvent B (36-37 min), 90% solvent B (37-40 min) and 100% solvent A (40-41 min). Sample elutions were spotted directly to a

steel MALDI target plate using a Probot MALDI Spotter (LC-Packings, Dionex) and mixed with α -cyano-4-hydroxycinnamic acid matrix at a concentration of 4 mg/ml. MALDI TOF/TOF MS profiling was performed using the 4800 MALDI TOF/TOFTM Analyser Applied Biosystems mass spectrometer in the positive reflectron mode and set for delayed extraction, with major peaks selected using a data dependent acquisition method for collision induced dissociation (CID) and sequencing by tandem mass spectrometry. The instrument was calibrated using the 4700 calibration standard, calmix, as the external calibrant for the MS mode.

2.4.9.5 MALDI-TOF MS and TOF/TOF MS/MS analysis of glycans

Derivatised samples were dissolved in 10 μ l of methanol and 1 μ l aliquots were mixed with an equal volume of the sample specific matrix (10 mg/ml) in their appropriate solvents (typically DABP for glycans and HCCA for peptides) and was spotted onto a MALDI target plate, before drying under vacuum. MALDI-TOF MS and TOF/TOF MS/MS were performed using a 4800 MALDI TOF/TOFTM Analyser (Applied Biosystem, Damstadt, Germany) mass spectrometer. The instrument was calibrated using the 4700 calibration standard, calmix, as the external calibrant for the MS mode. For MALDI-TOF/TOF the molecular ions were selected and subjected to collision induced dissociation (CID), where the collision energy was set to 1kV and pressurised argon was used as the collision gas (3.5×10^{-6} Torr).

2.4.9.6 Waters Synapt G2-S

Clostridium difficile orf3::erm mutant, after trypsin digestion, was analysed by a new generation Q-TOF instrument: the Waters Synapt G2-S. The peptides obtained were separated using nanoACQUITYTM LC, fitted with a nanoACQUITY Ultra Performance LCTM C₁₈ column (15cm length, 75 μ m internal diameter). The instrument was calibrated using 100 fmol/ μ l of [Glu¹]-fibrinopeptide B human and 200 pmol/ μ l of leucine enkephalin in 0.1% formic acid in 50% (v/v) aqueous acetonitrile solution. As this new technology incorporates the supplementary use of Electron Transfer Dissociation (ETD), it was necessary using a glow discharge, and specifically 4-nitrotoluene, and also to supercharge the peptides before the mass spectrometric detection. The supercharge reagent adopted is the m-nitrobenzyl alcohol (m-NBA) (Sigma Aldrich) and a 0.2% solution of m-NBA in 50% ACN

and 0.1% FA was prepared. The presence of m-NBA in electrospray solutions of proteins between 1 and 20% involves a significant increase in the number of charges on the gas-phase analyte ions. Generating higher charge state ions by ESI is considered important for MS/MS studies, because the increasing charge enhances dissociation efficiency and information from fragmentation (Iavarone and Williams 2002). Jensen and co-workers (Kjeldsen 2007) demonstrated the effect of m-NBA to shift the overall charge-state distribution of several tryptic peptides, in particular in the case of BSA tryptic peptides which changed from 2⁺ to 3⁺. Furthermore they achieved high charge-state shifts without jeopardising the chromatographic performance or increasing the chemical noise level adding just 0.1% m-NBA to the two mobile-phase solutions.

2.4.10 Data processing and interpretation

The GC-MS data acquired were processed and analysed using Bruker MSWS 8 System Control software, whilst the MALDI-TOF MS and TOF/TOF MS/MS data were acquired using 4000 Series Explorer Instrument Control Software and were processed using Data Explorer[®] 4.9 software. The MS data were generally noise filtered (correction factor of 0.7). For data handling of ESI Q-STAR MS and MS/MS data and ABSciex 5600 MS and MS/MS data Analyst QS Software with an automatic information-dependent-acquisition (IDA) function was used. Finally, data obtained using Waters Xevo G2 and Synapt G2-S instrumentations were acquired and processed by using MassLynx V4.1 software.

To aid in the assignment of peptide molecular ion peaks MassLynx Mass Spectrometry Software by Waters was used for *in-silico* digestion of the proteins in question, allowing virtual digestion with various enzymes and predicting peptide molecular ions, missed cleavages, methionine oxidation and cysteine carboxymethylation. For correct identification of the protein being analysed Mascot searches were performed. Data from each of the MS/MS spectra were used to search “All entries” in the Swiss Prot database for peptide sequences consistent with the fragment ions using the Mascot Search Engine.

Interpretations of data were made manually/visually, as when dealing with the most complex interpretations, then it is generally recognised that the manual (eye/brain coordination) method of interpretation is the gold standard compared to software attempts to interpret such spectra. Software used for automated interpretations of mass spectra, including automated peptide and/or carbohydrate mapping and sequencing of conventional protein and glycoprotein digest spectra, are for example provided by most instrument manufactures for

application in proteomics for the identification of unknown proteins in various sample types. However, there is a major difference between that type of work and the ability to handle a wide range of possible PTMs, including unknowns, within the data set produced. Manual interpretation is based on a number of key pieces of “know-how” information which themselves are the result of the accumulation of empirical knowledge. One example of this would be the predictable range which manual interpretation might apply to the search for novel O-linked glycosylation of peptides with respect to retention times on reverse HPLC column. Experience shows that the key factor affecting the retention time of an O-linked glycopeptide on a reverse phase C₁₈ column for example is the structure of the peptide itself, with the carbohydrate playing a minimum role in its chromatography. It follows that in a mapping search of a protein tryptic digest for example, if we first screen for the free *in-silico* predicted tryptic peptide, then we can be reasonably confident that any potential respective glycopeptide up to and including typical 1-5 sugar structures will elute (and therefore can be searched for in the data set) within a +/- approximately 5 minute retention time of the free peptide. This is the result of the higher mass of the glycopeptide moving the retention time to a later point, but the hydrophilicity of the sugars moving the retention time to an earlier point, and the net result of being very close elution of the two products. There are of course some exceptions of general rules, which can include the presence of sialic acid for example, which may be influenced by the particular elution buffer used. Similarly, a wider retention time window may need to be screened for when searching for larger N-linked structures. Finally, preparation of the spectra annotated with structural assignments was carried out manually with CorelDraw X3 Graphics Suite Software and Adobe Illustrator CS4.

Chapter 3:
Characterisation of ADAMTS13 PTMs

3. Characterisation of ADAMTS13 PTMs

3.1 Introduction

ADAMTS13, as described in **section 1.5.1**, is a large 200 kDa multi-domain glycoprotein, specifically a disintegrin and metalloprotease with a thrombospondin type 1 motif, member 13, able to regulate thrombogenesis by cleavage of an adhesive blood glycoprotein, the von Willebrand factor (VWF) (Zheng 2013a, Zheng 2013b).

In the studies leading to our involvement in this work, our collaborators at the United States Food and Drug Administration (FDA), Dr. Chava Kimchi-Sarfaty and Dr. Ryan Hunt, had investigated the effects of six synonymous and six non-synonymous variants found in ADAMTS13 in the general population (Edwards, Hing et al. 2012), and a transient expression system in mammalian cells was used for studying the effects of these substitutions at the level of mRNA, protein expression and enzymatic activity. In initial experiments, a single synonymous variant at a proline residue within the metalloprotease domain, c. 354G>A [P118P] was found to lead to greater secretion of ADAMTS13 compared to WT. Our FDA colleagues were then faced with the question of why that would be the case, and whether the cause would be higher cellular expression or something related to the post-translational modification (PTM) of the protein perhaps enhancing secretion. One PTM event in particular has previously been reported to play a dramatic role in the secretion of recombinant ADAMTS13 in an HEK cell line, which is the O-glycosylation of Ser in a CXX(S/T)CG consensus sequence (where the Cys residues are conserved), within the thrombospondin repeat, by a fucosyl disaccharide (Ricketts, Dlugosz et al. 2007). Six substitutions of a possible seven consensus sites were reported by Ricketts et al. who also demonstrated that site directed mutagenesis of S→A at one site in particular led to a dramatic drop in ADAMTS13 secretion.

My role in this study, in collaboration with Biopharmaspec and the FDA, has been to try to compare the relevant O-glycome PTMs of WT ADAMTS13 with the synonymous P118P mutation ADAMTS13 glycoprotein, together with a non-synonymous P118F control which actually leads to a decreased glycoprotein secretion. A major challenge of this mass spectrometry study was the total sample availability of only 4 micrograms of this large glycoprotein corresponding to only 20 picomoles of material.

3.2 Experimental strategy

To profile the PTMs among WT, P118P and P118F ADAMTS13 relevant for secretion and protein folding by mass spectrometry, all of the available samples in each case were loaded onto a 4-12% SDS gel for final purification. The electrophoretogram obtained for the three samples is shown in **Figure 3.1** against a protein reference standard lane showing in each case a broad Coomassie-stained band in the region of 200 kDa.

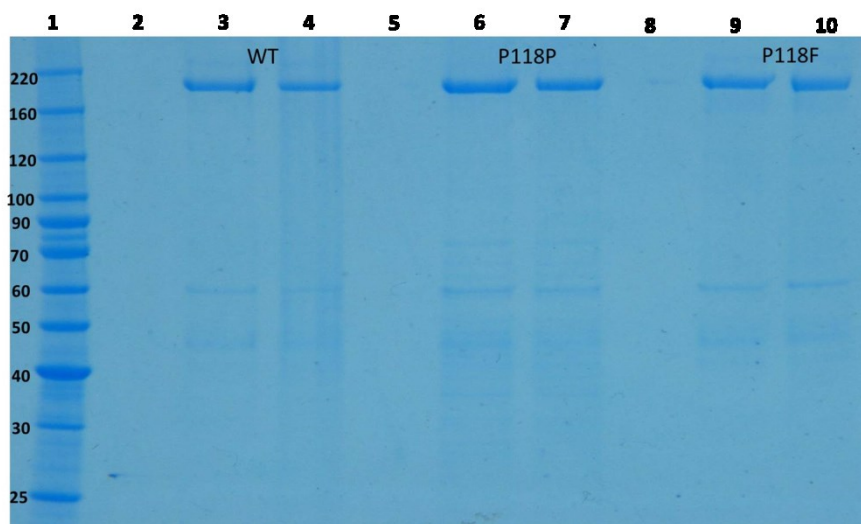


Figure 3.1 Separation of the ADAMTS13 glycoprotein samples by SDS-PAGE gel (Invitrogen NuPAGE 4-12% Bis Tris Gel) and visualised by Coomassie blue staining. Molecular size markers are indicated on the left. Three different kinds of samples were run, wild type (WT), synonymous P118P mutation ADAMTS13 glycoprotein and non-synonymous P118F control. Lane 1: bench mark (5 μ g); lanes 2-5-8: empty; lanes 3-4: WT; lanes 6-7: P118P; lanes 9-10: P118F.

Following excision from the gel and destaining, each band was treated by in-gel reduction and alkylation, followed by trypsin digestion. Strategically, because of the restricted amount of sample, each was divided to run 2/3 on the nanoLC-Q-STAR and 1/3 on the microLC-Q-TOF instruments in automated MS/MS mode, known as IDA and DDA respectively.

The MS total ion current traces (TIC) for the LC MS/MS runs of the ADAMTS13 samples were then generated, and the WT, P118P and P118F MS data sets were then screened manually against the theoretical doubly/triply/quadruply charged *in silico* tryptic digest data produced by MassLynx for the ADAMTS13 glycoprotein.

Once a peptide of interest was located, the MS/MS data sets in and around those locations were screened for possible PTMs.

A schematic summarising the key strategic steps in this protocol is shown in **Figure 3.2**.

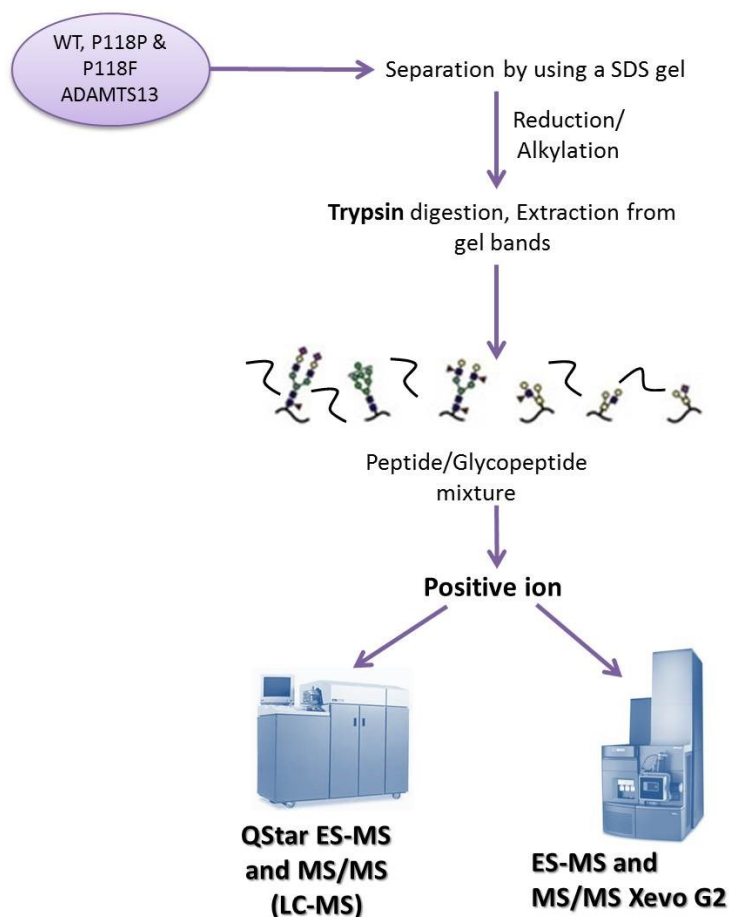


Figure 3.2 Experimental strategy employed in ADAMTS13 PTMs characterisation. Following trypsin in gel digestion of the samples, mass mapping was carried out. The mass spectrometric analysis was done in positive ion mode and using on-line LC-ES-MS and MS/MS.

3.3 Results & Interpretation

In the particular case of the ADAMTS13 research, presented in this chapter, the know-how described in **section 2.4.10** was particularly useful in locating anticipated PTMs which had been previously reported, but also in finding new ones, despite the minute amount of material available for study.

3.3.1 TSRs

The definitive study of factors potentially affecting the secretion of ADAMTS13 has been the work of Ricketts et al. (Ricketts, Dlugosz et al. 2007) described in the Introduction, in which they demonstrated the addition of a HexFuc disaccharide unit to six of the seven possible thrombospondin repeat consensus sequences (TSRs) in the molecule. They were unable to

find evidence for the PTM of the first, TSR1. In our collaboration with the FDA, it was necessary to make a comparative study between the three provided samples of wild type WT, the synonymous mutation sample P118P and the non-synonymous variant P118F, and the first step after the practical production of the tryptic digest LC-MS and MS/MS data was the generation of the *in-silico* tryptic map for the 1,427 residue ADAMTS13 protein shown in Table 3.1.

Untitled
Trypsin/K-IP /R-IP

Frag#	Res#	Sequence	Theor (Bo)	[M+H]	[M+2H]	[M+3H]	[M+4H]
T1	1-4	(-)MHQR (H)	570.27	571.28	286.14	191.10	143.58
T2	5-7	(R)HPR (A)	408.22	409.23	205.12	137.08	103.06
T3	8-9	(R)AR (J)	245.15	246.16	123.58	82.72	62.30
T4	10-57	(R)JPPLJVAGILAJGFLIG JWGPSHFQQSJLQALEPQAV SSYLSPGAPLK (G)	5246.46	5247.47	2624.24	1749.83	1312.62
T5	58-67	(K)GRPPSPGFQR (Q)	1097.57	1098.58	549.79	366.87	275.40
T6	68-69	(R)QR (Q)	302.17	303.18	152.09	101.73	76.55
T7	70-71	(R)QR (Q)	302.17	303.18	152.09	101.73	76.55
T8	72-73	(R)QR (R)	302.17	303.18	152.09	101.73	76.55
T9	74-74	(R)R (A)	174.11	175.12	88.06	59.05	44.54
T10	75-102	(R)AAGGILHLELLLVAVGPD VFQARQEDTER (Y)	2984.54	2985.54	1493.28	995.85	747.14
T11	103-116	(R)YVLTNLNIGAEILR (D)	1587.90	1588.91	794.96	530.31	397.98
T12	117-125	(R)DPSLGAQFR (V)	989.49	990.50	495.75	330.84	248.38
T13	126-130	(R)VHLVK (M)	594.39	595.39	298.20	199.14	149.60
T14	131-180	(K)MVLITPEEGAPNITANL TSSLLSVJGWSQTINPEDDT DPGHADLVLYITR (F)	5424.64	5425.65	2713.33	1809.22	1357.17
T15	181-190	(R)FDLELPDGNR (Q)	1174.56	1175.57	588.29	392.53	294.65
T16	191-193	(R)QVR (G)	401.24	402.25	201.63	134.75	101.32
T17	194-257	(R)GVTQLGGAJSPTWSJLI TEDTGFDLGVTIAHEIGHSF GLEHDGAPGSGJGPSGHVMA SDGAAPR (A)	6463.88	6464.89	3232.95	2155.63	1616.98
T18	258-267	(R)AGLAWSPJSR (R)	1104.50	1105.51	553.26	369.18	277.13
T19	268-268	(R)R (Q)	174.11	175.12	88.06	59.05	44.54
T20	269-278	(R)QLLSLLSAGR (A)	1056.63	1057.64	529.32	353.22	265.17
T21	279-280	(R)AR (J)	245.15	246.16	123.58	82.72	62.30
T22	281-312	(R)JVNDPPRPQPSAGHPP DAQPGLYYSANEQJR (V)	3608.58	3609.59	1805.30	1203.87	903.15
T23	313-318	(R)VAFGPK (A)	617.35	618.36	309.68	206.79	155.35
T24	319-326	(K)AVAJTFAR (E)	895.42	896.43	448.72	299.48	224.86
T25	327-349	(R)EHLDMJQALSJHTDPLD QSSJSR (L)	2749.07	2750.08	1375.54	917.37	688.28
T26	350-364	(R)LLVPLLDGTEJGVK (W)	1642.85	1643.86	822.43	548.62	411.72
T27	365-368	(K)WJSK (G)	580.23	581.24	291.12	194.42	146.07
T28	369-370	(K)GR (J)	231.13	232.14	116.57	78.05	58.79
T29	371-372	(R)JR (S)	335.13	336.13	168.57	112.72	84.79
T30	373-386	(R)SIVELTPIAAVHGR (W)	1461.83	1462.84	731.92	488.28	366.47
T31	387-393	(R)WSSWGER (S)	874.41	875.42	438.21	292.48	219.61
T32	394-398	(R)SPJSR (S)	606.24	607.25	304.13	203.09	152.57
T33	399-407	(R)SJGGGVVTR (R)	892.41	893.42	447.21	298.48	224.11
T34	408-408	(R)R (R)	174.11	175.12	88.06	59.05	44.54
T35	409-409	(R)R (Q)	174.11	175.12	88.06	59.05	44.54
T36	410-421	(R)QJNNPRPAFGGR (A)	1373.63	1374.63	687.82	458.88	344.41
T37	422-440	(R)AJVCGADLQAEMLNTQAJ EK (T)	2157.83	2158.84	1079.92	720.29	540.47

Untitled
Trypsin/K-IP /R-IP

Frag#	Res#	Sequence	Theor (Bo)	[M+H]	[M+2H]	[M+3H]	[M+4H]
T38	441-452	(K) TQLEEMSQQJAR (T)	1498.65	1499.66	750.34	500.56	375.67
T39	453-459	(R) TDGQPLR (S)	785.40	786.41	393.71	262.81	197.36
T40	460-484	(R) SSPGGASFYHWGAAPVH SQGDALJR (H)	2615.16	2616.17	1308.59	872.73	654.80
T41	485-488	(R) HBJR (A)	603.23	604.23	302.62	202.08	151.81
T42	489-497	(R) AIGSEFIMK (R)	994.52	995.52	498.27	332.51	249.64
T43	498-498	(K) R (G)	174.11	175.12	88.06	59.05	44.54
T44	499-507	(R) GDSFLDGTR (J)	966.44	967.45	484.23	323.15	242.62
T45	508-514	(R) JMPSGPR (E)	804.33	805.33	403.17	269.12	202.09
T46	515-528	(R) EDGTLSLJVSGSJR (T)	1541.63	1542.64	771.82	514.89	386.42
T47	529-535	(R) TFGJDR (M)	812.31	813.32	407.16	271.78	204.09
T48	536-544	(R) MDSQQVWDR (J)	1163.50	1164.51	582.76	388.84	291.88
T49	545-558	(R) JQVJGGDNSTJSR (K)	1599.56	1600.57	800.79	534.19	400.90
T50	559-559	(R) K (G)	146.11	147.11	74.06	49.71	37.53
T51	560-566	(K) GSFTAGR (A)	694.34	695.35	348.18	232.45	174.59
T52	567-568	(R) AR (E)	245.15	246.16	123.58	82.72	62.30
T53	569-598	(R) EYVTEFLTVTPNLTSVYI ANHRPLFTHLAVR (I)	3471.87	3472.87	1736.94	1158.30	868.97
T54	599-602	(R) IGGR (Y)	401.24	402.25	201.63	134.75	101.32
T55	603-608	(R) YVVAGK (M)	635.36	636.37	318.69	212.80	159.85
T56	609-625	(K) MSISENTTYPSSLEEDGR (V)	1879.90	1880.91	940.96	627.64	470.98
T57	626-629	(R) VEYR (V)	565.29	566.29	283.65	189.44	142.33
T58	630-636	(R) VALTEDR (L)	802.42	803.43	402.22	268.48	201.61
T59	637-639	(R) LPR (L)	384.25	385.26	193.13	129.09	97.07
T60	640-644	(R) LEEIR (I)	658.36	659.37	330.19	220.46	165.60
T61	645-659	(R) IWGFLQEDADIQVYR (R)	1801.90	1802.91	901.96	601.64	451.48
T62	660-660	(R) R (Y)	174.11	175.12	88.06	59.05	44.54
T63	661-683	(R) YGEEYGNLTREDITFTY FQPKPR (Q)	2791.36	2792.37	1396.69	931.46	698.85
T64	684-692	(R) QAWVWAAVR (G)	1085.58	1086.58	543.80	362.87	272.40
T65	693-704	(R) GPJSVSVJGAGLR (W)	1221.51	1222.52	611.76	408.18	306.39
T66	705-715	(R) WVNYSJLDQAR (K)	1411.62	1412.63	706.82	471.55	353.91
T67	716-716	(R) K (E)	146.11	147.11	74.06	49.71	37.53
T68	717-763	(K) ELVETVQJQGSQPPAW PEAJVLEPJPYPWAVGDFGP JSASJGGGLR (E)	5208.22	5209.23	2605.12	1737.08	1303.06
T69	764-768	(R) ERPVR (J)	655.38	656.38	328.70	219.47	164.85
T70	769-778	(R) JVEAQGSLK (T)	1104.55	1105.56	553.28	369.19	277.14
T71	779-784	(K) TLPEAR (J)	653.39	654.39	327.70	218.80	164.35
T72	785-786	(R) JR (A)	335.13	336.13	168.57	112.72	84.79
T73	787-807	(R) AGAQPPVALETJNPQP JPAR (W)	2237.02	2238.03	1119.52	746.68	560.26
T74	808-852	(R) WEVSEPPSSJTSAGGAGL ALENETJVPGDGLEAPVTE GPGSVDEK (L)	4517.99	4518.99	2260.00	1507.00	1130.50
T75	853-879	(K) LPAPERJVGMSJPPGNG HIDATSAGEK (A)	2822.24	2823.24	1412.13	941.75	706.57
T76	880-888	(K) APSPWGSIR (T)	969.50	970.51	485.76	324.18	243.38
T77	889-910	(R) TGAQAHHVWTPVAGSJS VSJGR (G)	2260.00	2261.01	1131.01	754.34	566.01
T78	911-916	(R) GLMELR (F)	717.38	718.39	359.70	240.14	180.35
T79	917-925	(R) FLJMDSALR (V)	1112.50	1113.51	557.26	371.84	279.13
T80	926-942	(R) VPVQEEIJGLASKPGSR	1826.92	1827.93	914.47	609.98	457.74
T81	943-943	(R) R (E)	174.11	175.12	88.06	59.05	44.54
T82	944-954	(R) EVJQAVPPEAR (W)	1287.56	1288.57	644.79	430.19	322.90
T83	955-958	(R) WQYK (L)	623.31	624.31	312.66	208.78	156.83
T84	959-968	(K) LAAJSVSVJGR (G)	1081.45	1082.46	541.73	361.49	271.37
T85	969-972	(R) GVVR (R)	429.27	430.28	215.64	144.10	108.33
T86	973-973	(R) R (I)	174.11	175.12	88.06	59.05	44.54
T87	974-979	(R) ILYJAR (A)	795.39	796.40	398.71	266.14	199.86
T88	980-1015	(R) AHGEDDGEELLLDTQJQ GLPRPEPQEAJSLEPJPFR (W)	4102.78	4103.79	2052.40	1368.60	1026.70

Untitled

Trypsin:K-IP /R-IP

Frag#	Res#	Sequence	Theor (Bo)	[M+H]	[M+2H]	[M+3H]	[M+4H]
T89	1016-1017	(R) WK (V)	332.18	333.19	167.10	111.74	84.05
T90	1018-1034	(K) VMSLGPJSASJGLGTAR (R)	1724.75	1725.76	863.38	575.93	432.20
T91	1035-1035	(R) R (S)	174.11	175.12	88.06	59.05	44.54
T92	1036-1075	(R) SVAJVLQDQGDVEVDE AAJAALVRPEASVPJLIADJ TYR (W)	4438.99	4440.00	2220.50	1480.67	1110.76
T93	1076-1094	(R) WHVGTWMEJSVVSJGDGI QR (R)	2265.92	2266.93	1133.97	756.32	567.49
T94	1095-1095	(R) R (R)	174.11	175.12	88.06	59.05	44.54
T95	1096-1096	(R) R (D)	174.11	175.12	88.06	59.05	44.54
T96	1097-1123	(R) DTJLGPQAQAPVPADFJ QHLPKPVTVR (G)	3003.46	3004.47	1502.74	1002.16	751.87
T97	1124-1149	(R) GJWAGPJVGQGTPLVLP HEEAAPGR (T)	2662.19	2663.20	1332.10	888.40	666.56
T98	1150-1165	(R) TTATPAGASLEWSQAR (G)	1645.81	1646.81	823.91	549.61	412.46
T99	1166-1176	(R) GLLFSPAPQPR (R)	1181.66	1182.66	591.84	394.89	296.42
T100	1177-1177	(R) R (L)	174.11	175.12	88.06	59.05	44.54
T101	1178-1194	(R) LLPGQENSVSQSSAJGR (Q)	1799.85	1800.86	900.93	600.96	450.97
T102	1195-1206	(R) QHLEPTGTIDMR (G)	1396.68	1397.68	699.35	466.57	350.18
T103	1207-1228	(R) GPGQADJAVAIIGRELGE VVTLR (V)	2236.16	2237.17	1119.09	746.40	560.05
T104	1229-1247	(R) VLESSLNJSAGDMLLLW GR (L)	2121.02	2122.03	1061.52	708.02	531.26
T105	1248-1251	(R) LTWR (K)	574.32	575.33	288.17	192.45	144.59
T106	1252-1252	(R) K (M)	146.11	147.11	74.06	49.71	37.53
T107	1253-1255	(K) MJR (K)	466.17	467.17	234.09	156.40	117.55
T108	1256-1256	(R) K (L)	146.11	147.11	74.06	49.71	37.53
T109	1257-1265	(K) LLDMTFSSK (T)	1040.52	1041.53	521.27	347.85	261.14
T110	1266-1272	(K) TNLVVR (Q)	801.47	802.48	401.74	268.16	201.38
T111	1273-1274	(R) QR (J)	302.17	303.18	152.09	101.73	76.55
T112	1275-1285	(R) JGRPGGVLLR (Y)	1141.60	1142.61	571.81	381.54	286.41
T113	1286-1297	(R) YGSQLAPEFYR (E)	1430.68	1431.69	716.35	477.90	358.68
T114	1298-1326	(R) EJDMQLFGPWGEIVSPS LSPATSNAGGJR (L)	3124.36	3125.37	1563.19	1042.46	782.10
T115	1327-1336	(R) LFINVAPHAR (I)	1136.65	1137.65	569.33	379.89	285.17
T116	1337-1361	(R) IAIHALATNMGATEGA NASYILIR (D)	2527.32	2528.33	1264.67	843.45	632.84
T117	1362-1367	(R) DTHSLR (T)	727.36	728.37	364.69	243.46	182.85
T118	1368-1396	(R) TTAFHGQQVLYWESESS QAEMEFSEGFLK (A)	3365.52	3366.53	1683.77	1122.85	842.39
T119	1397-1402	(K) AQASLR (G)	644.36	645.37	323.19	215.79	162.10
T120	1403-1422	(R) GQYWTLQSWVPEMQDPQ SNK (G)	2493.14	2494.15	1247.58	832.06	624.29
T121	1423-1424	(K) GK (E)	203.13	204.13	102.57	68.72	51.79
T122	1425-1427	(K) EGT (-)	305.12	306.13	153.57	102.72	77.29

Table 3.1 *in-silico* Tryptic Map for the 1,427 residue ADAMTS13 protein (J = Cmc = carboxymethylcysteine).

The total ion current (TIC) trace of the WT tryptic digest in the MS mode across the 90 minute nanoLC gradient is showed in **Figure 3.3** and its appearance immediately indicates that the digest has been successful with a variety of signals across the whole elution range, as expected for a macromolecule of this type.

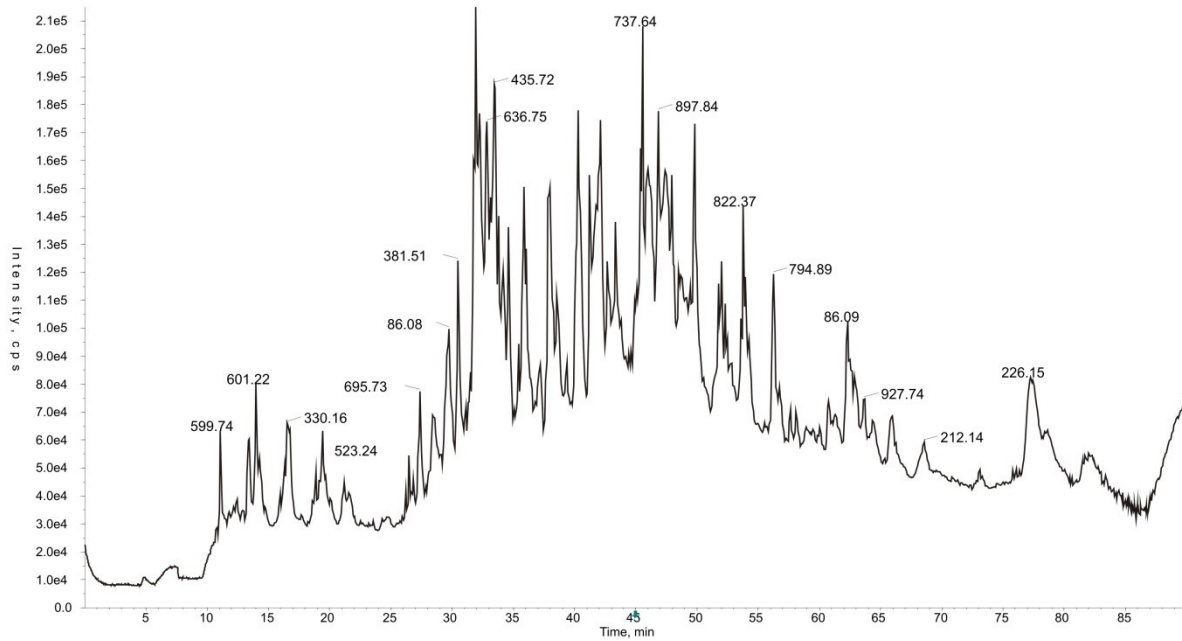


Figure 3.3 TIC (Total Ion Current) for the 90 minute separation of the ADAMTS13 WT digest *online*-nano-LC-ES-MS/MS experiment. The TIC represents the summed intensity across the entire range of masses being detected at every point during the run.

TSR1:

The first predicted thrombospondin consensus site, and the first consensus sequence in ADAMTS13, lies within the tryptic peptide SJGGGVVTR (residues 399-407), MW 892.41, predicted $[M + 2H]^{2+}$ 447.21. Screening for this mass in the MS trace of the WT sample in **Figure 3.3** finds a signal at 14.1 minute. **Figure 3.4** shows a screen shot of the MS LC profiles and MS and MS/MS spectra with the 447 signal highlighted, and its corresponding MS/MS spectrum shown in the bottom right of the picture. The fragmentation seen in this MS/MS spectrum clearly confirms the identity of the peptide of interest via signals at m/z 175 (y_1), 221 (a_2), 249 (b_2), 276 (y_2), 375 (y_3), 588 (y_6), 645 (y_7). The top right hand panel shows the MS spectrum before the instrument has automatically MS/MSed multiply charged signals within it, and this panel nicely illustrates the principle made previously regarding searching a predictable range surrounding the retention time of the free peptide, to look for O-linked carbohydrate PTMs. Within that panel there is a signal at m/z 601.22 which on expansion is seen to be doubly charged. The m/z difference to the m/z 447.21 free peptide is therefore $601.2 - 447.2 = 154.0$ and since these ions are doubly charged this corresponds to a 308 mass difference which could fit to a HexFuc disaccharide PTM.

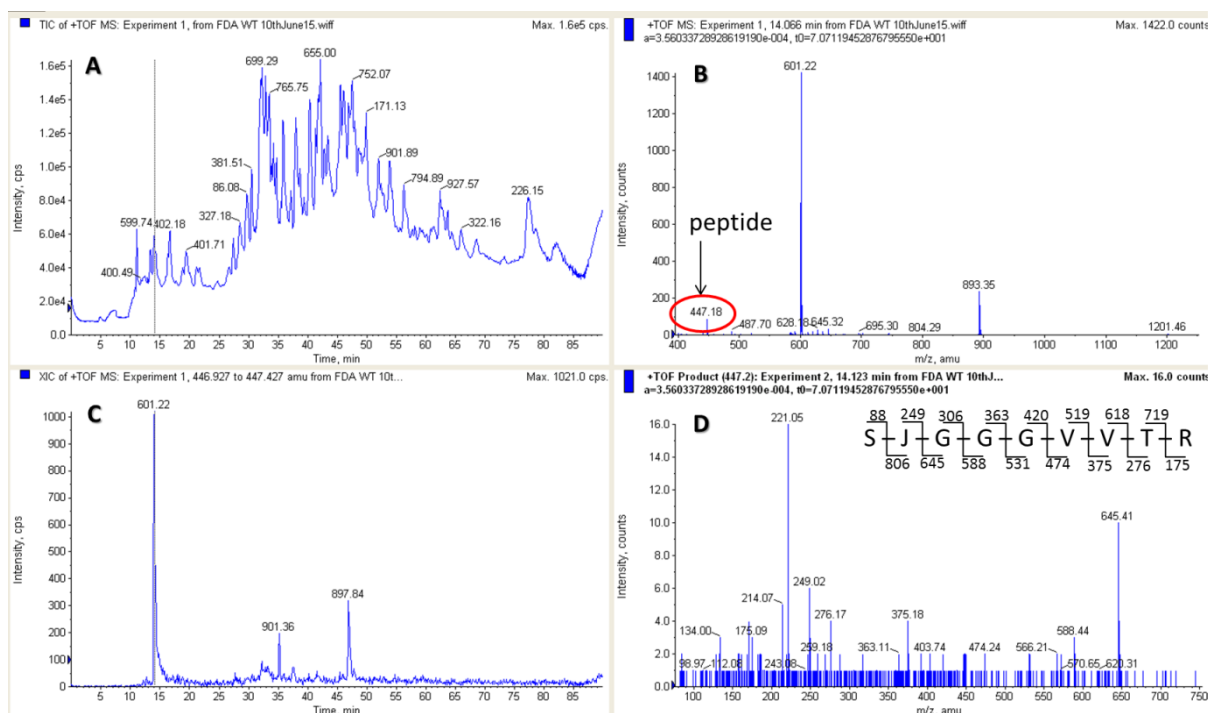


Figure 3.4 MS LC profiles and MS and MS/MS spectra with the 447 signal highlighted in red. The top left hand panel corresponds to the TIC; the top right hand panel shows the MS spectrum; the bottom left hand panel is the XIC (Extracting Ion Current) and the bottom right hand panel consists in the MS/MS spectrum of the m/z 447.

Looking now at the MS/MS of the m/z 601.2²⁺ ion which is shown in **Figure 3.5**, this absolutely proves the post-translational modification of peptide 447²⁺ because the precursor ion collapses immediately with the elimination of 308 Da to give the quasi-molecular ion of the 399-407 peptide seen at m/z 893.3, and the further b and y ions characteristic of this peptide (as found in **Figure 3.4**). This rapid loss of the PTM is absolutely characteristic of O-linked carbohydrate.

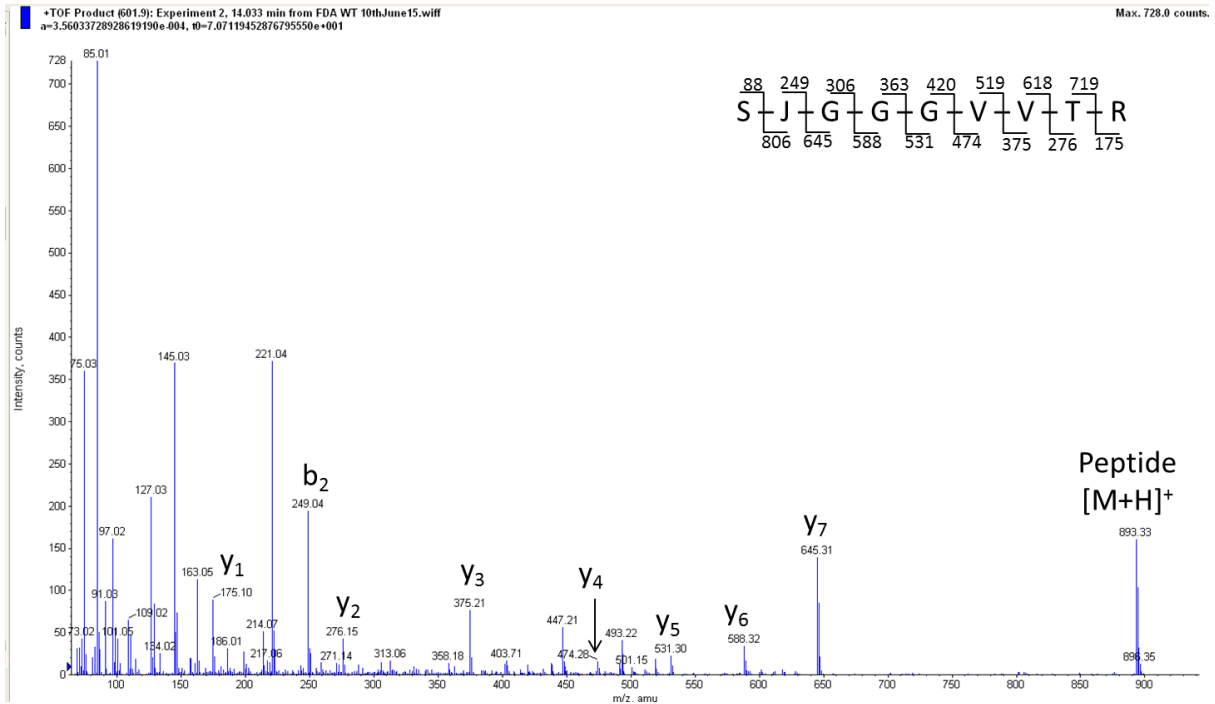


Figure 3.5 MS/MS spectrum of m/z 601²⁺ in WT ADAMTS13. Peptide fragmentation provides very strong evidence for the sequence SJGGGVVTR, all the y ions, except for y_1 , and some b ions have been found.

Comparing this data set now with the P118P and P118F samples, m/z 601.2²⁺ is also found in these data and the MS/MS data are shown for the respective samples in **Figure 3.6** and **3.7**.

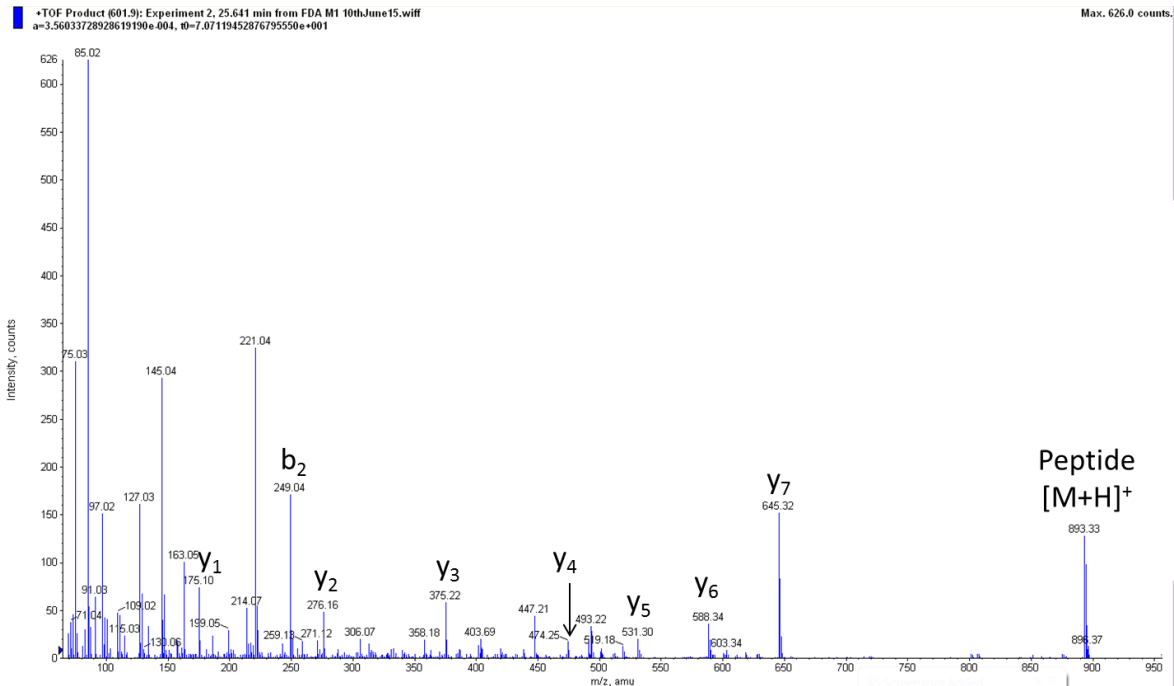


Figure 3.6 MS/MS spectrum of m/z 601²⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in **Figure 3.5**.

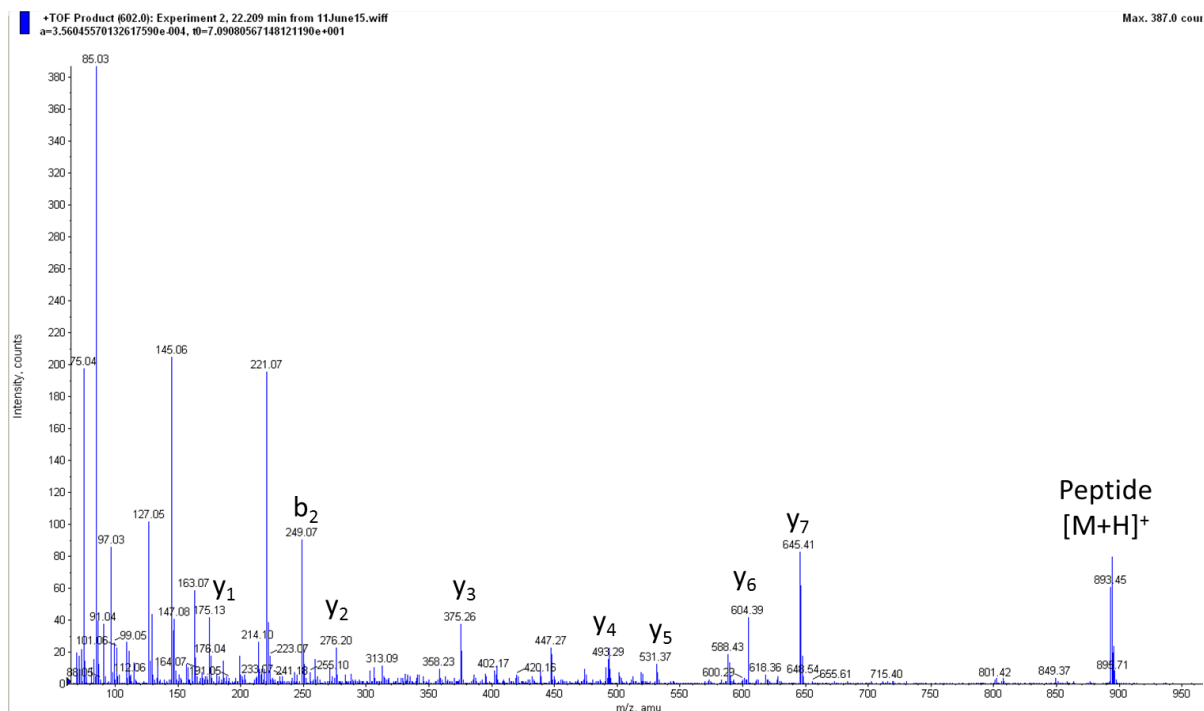


Figure 3.7 MS/MS spectrum of m/z 601²⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both **Figure 3.5** and **Figure 3.6**.

As can be seen with respect to the b and y ions and the quasi-molecular ion the data are very similar, again showing a facile loss of 308 Da corresponding to a HexFuc disaccharide.

Overall, in contrast to earlier studies in which no evidence for TSR1 substitution has been found (Ricketts, Dlugosz et al. 2007) the work presented here shows clear evidence for the TSR1 PTM disaccharide. Since the earlier studies link the TSR substitutions to secretion mechanisms of the protein (although there is no clear evidence to date of how that operates biochemically), then it would also be important to examine the degree of substitution of the TSRs, if it is possible. To do such work properly, isotopically labelled internal standards would be required and dose/response curves produced to define the amounts of the unglycosylated and glycosylated peptides in each sample, and this was not possible with the available material. However, it is possible to estimate approximate levels of substitution from ion current values of the glycopeptide discovered compared to ion currents of the free peptide and although the estimated level is slightly higher for the P118P sample, this could not be the basis of a quantitative conclusion from this work. Therefore, it would appear that differential substitution of the HexFuc disaccharide between WT and P118P will have to await further study as a possible explanation for increased secretion of P118P at least with respect to the newly found TSR1.

TSR2:

The second predicted thrombospondin consensus sequence in ADAMTS13 lies within the tryptic peptide GPJSVSJGAGLR (residues 693-704), MW 1221.51, predicted $[M + 2H]^{2+}$ 611.76²⁺. There is no significant signal for the free peptide in the MS chromatogram, however a strong related signal is found at m/z 765.8²⁺ ion and the MS/MS spectrum is shown in **Figure 3.8**. This proves the post-translational modification of peptide 611.76²⁺ with a HexFuc disaccharide because the precursor ion decomposes with the elimination of 308 Da to give the quasi-molecular ion of the GPJSVSJGAGLR peptide seen at m/z 1222.40, together with b and y ions which define this peptide sequence.

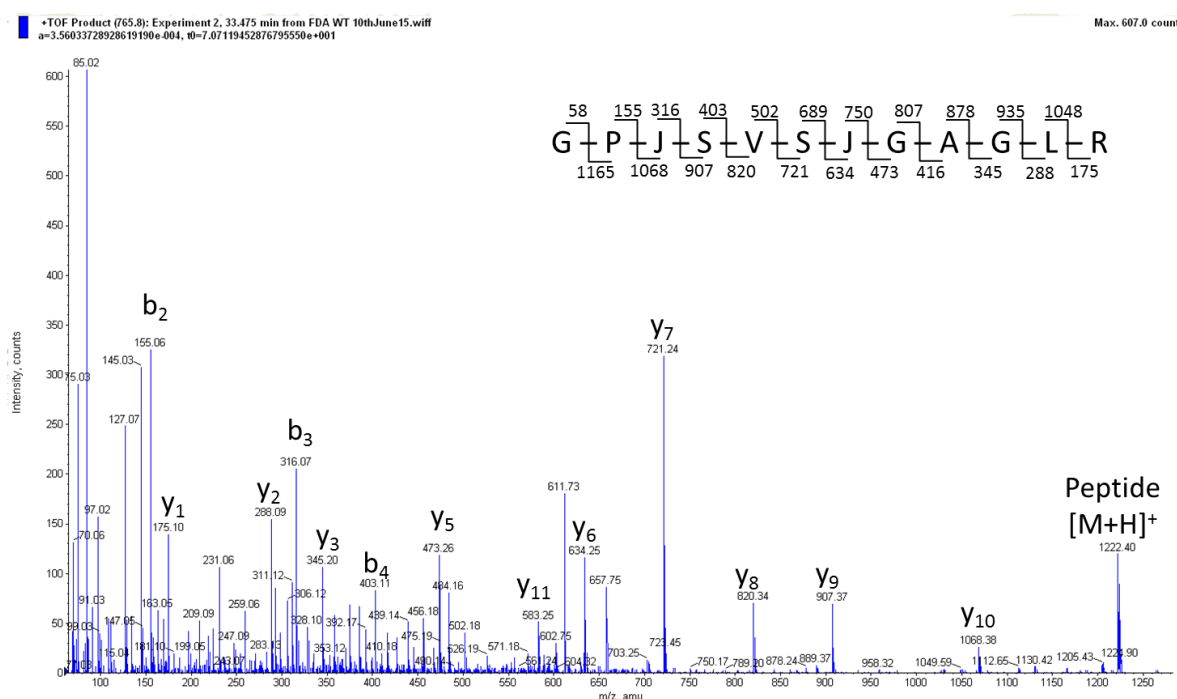


Figure 3.8 MS/MS spectrum of m/z 765.8²⁺ in WT ADAMTS13. Peptide fragmentation provides very strong evidence for the sequence GPJSVSJGAGLR, the majority of y ions and some b ions have been found.

Comparing these data now with the P118P and P118F samples, m/z 765.8²⁺ is also found in these data sets and the MS/MS spectra are shown for the respective samples in **Figure 3.9** and **3.10**. In terms of the b and y ions and the quasi-molecular ion generated, the glycopeptide assigned is equivalent to TSR2 in the WT data.

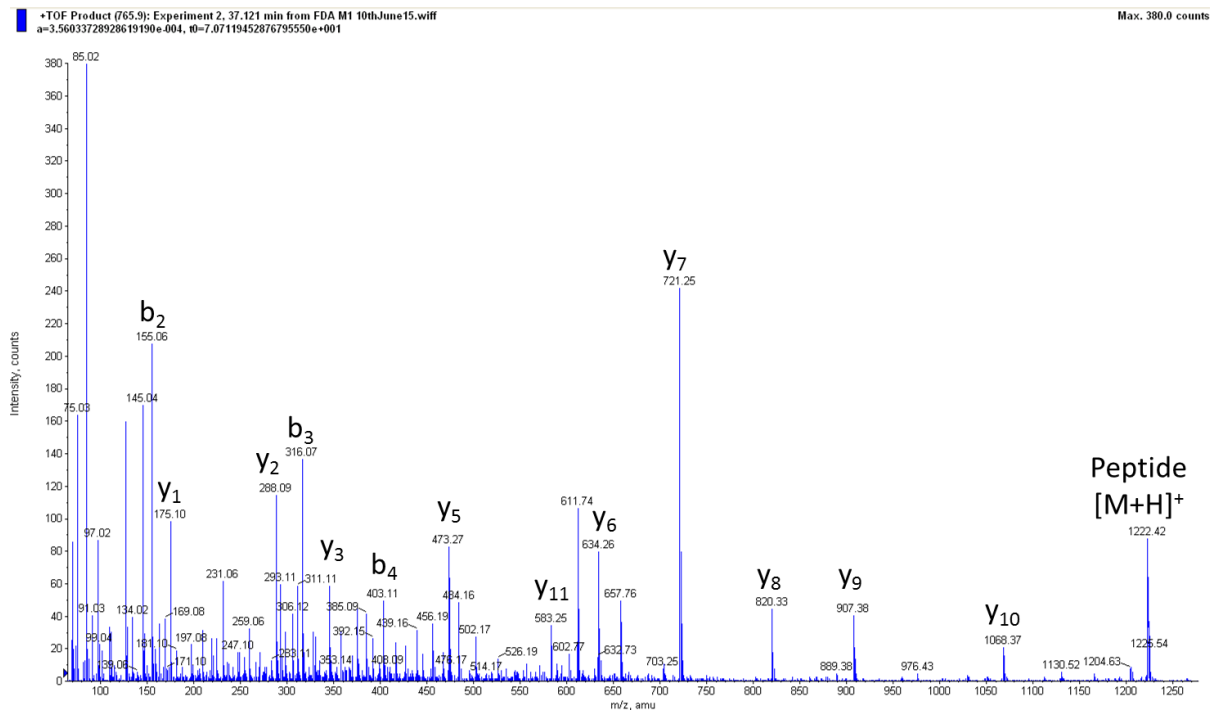


Figure 3.9 MS/MS spectrum of m/z 765.8²⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in **Figure 3.8**.

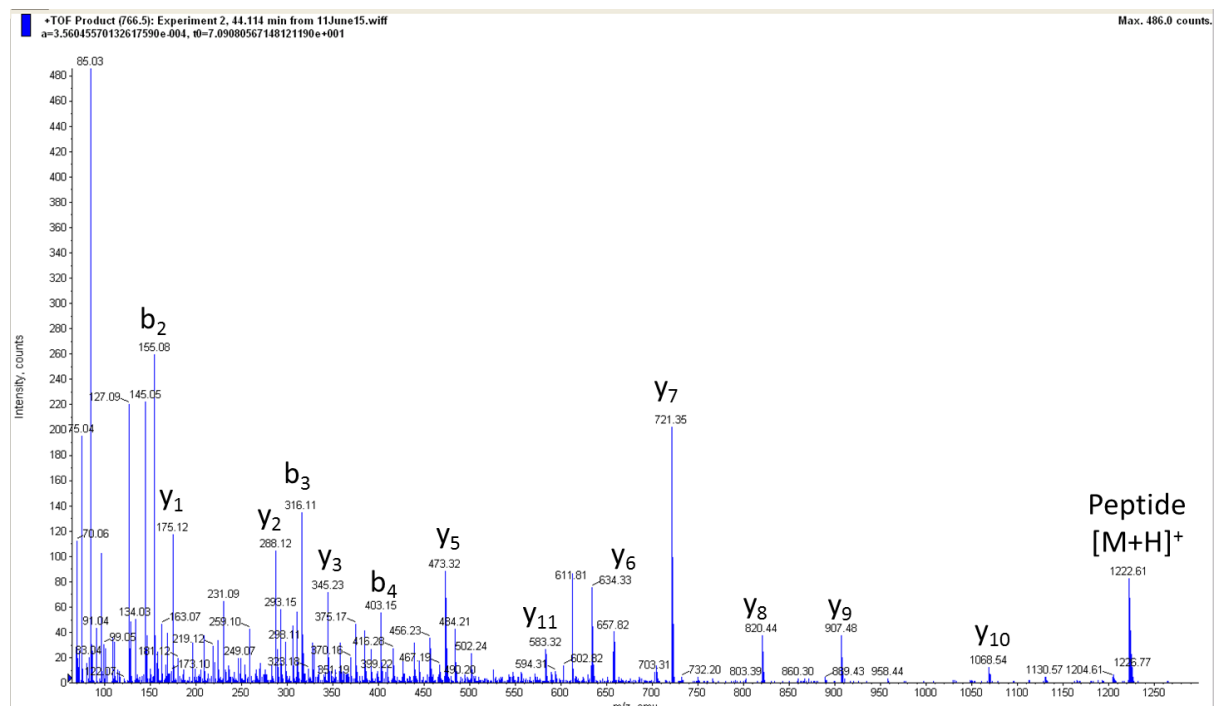


Figure 3.10 MS/MS spectrum of m/z 765.8²⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both **Figure 3.8** and **Figure 3.9**

TSR3:

The third predicted thrombospondin consensus sequence in ADAMTS13 lies within the tryptic peptide ELVETVQJQGSQQPPAWPEAJVLEPJPPYWAVGDFGPGJSASJGGGLR (residues 717-763), MW 5208.22, predicted $[M + 4H]^{4+}$ 1303.95⁴⁺. There is no significant signal for the free peptide in the MS chromatogram, however a related cluster of signals centred around m/z 1381 and seen to be fourthly charged is observed, and this corresponds to a mass difference of 77 Da, and being fourthly charged this is 308 Da corresponding to a HexFuc. The MS/MS spectrum is shown in **Figure 3.11**. This proves the post-translational modification of peptide 1303.95⁴⁺ with a HexosylFucose because the precursor ion decomposes to give a series of fragment ions which identify the tryptic peptide as residues 717-763 via the presence of a series of b and y ions and mid-chain proline fragments in the MS/MS spectrum in **Figure 3.11**, including y_5 , y_7 , y_8 , y_{11} , y_{12} and y_{13} both within and next to the consensus domain, seen at m/z 459, 707, 778, 1123, 1180 and 1327 respectively.

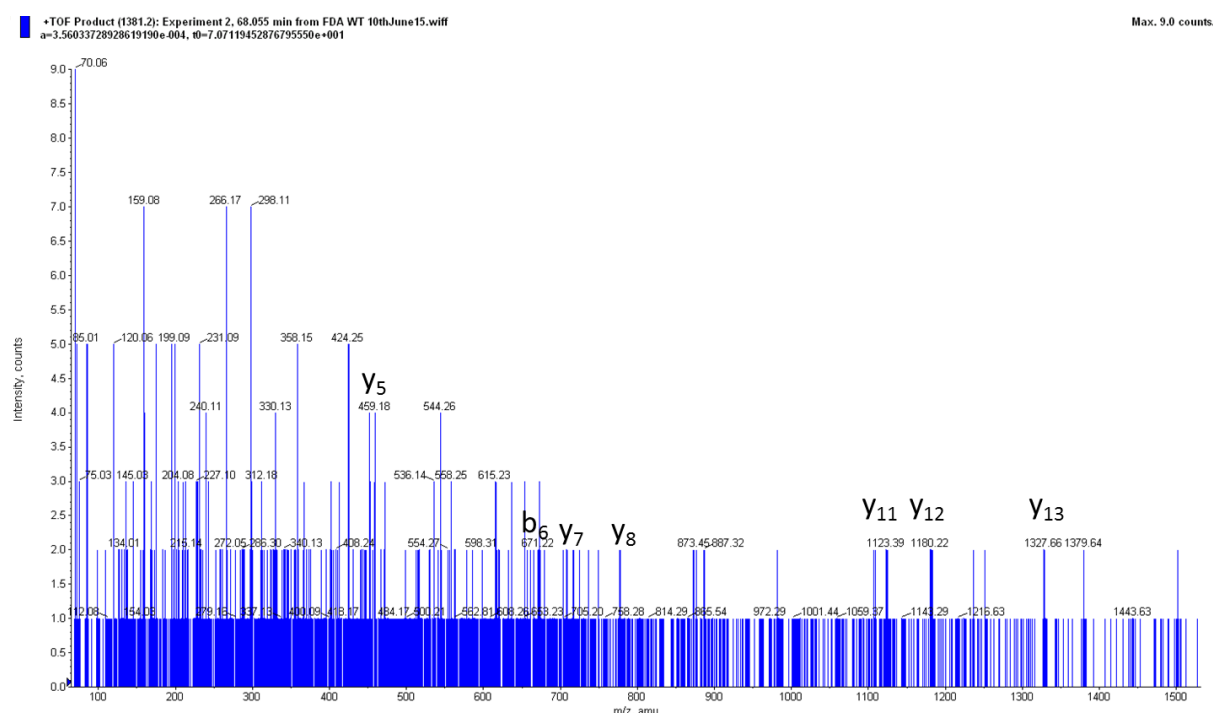


Figure 3.11 MS/MS spectrum of m/z 1381.2⁴⁺ in WT ADAMTS13. Peptide fragmentation provides evidence for the sequence ELVETVQJQGSQQPPAWPEAJVLEPJPPYWAVGDFGPGJSASJGGGLR.

Comparing these data now with the P118P and P118F samples, a signal grouping centred around m/z 1381⁴⁺ is also found in these data sets and the MS/MS spectra are shown for the

respective samples in **Figure 3.12** and **3.13**. In terms of the b and y ions and mid-chain fragments generated, the glycopeptide assigned is equivalent to TSR3 in the WT data.

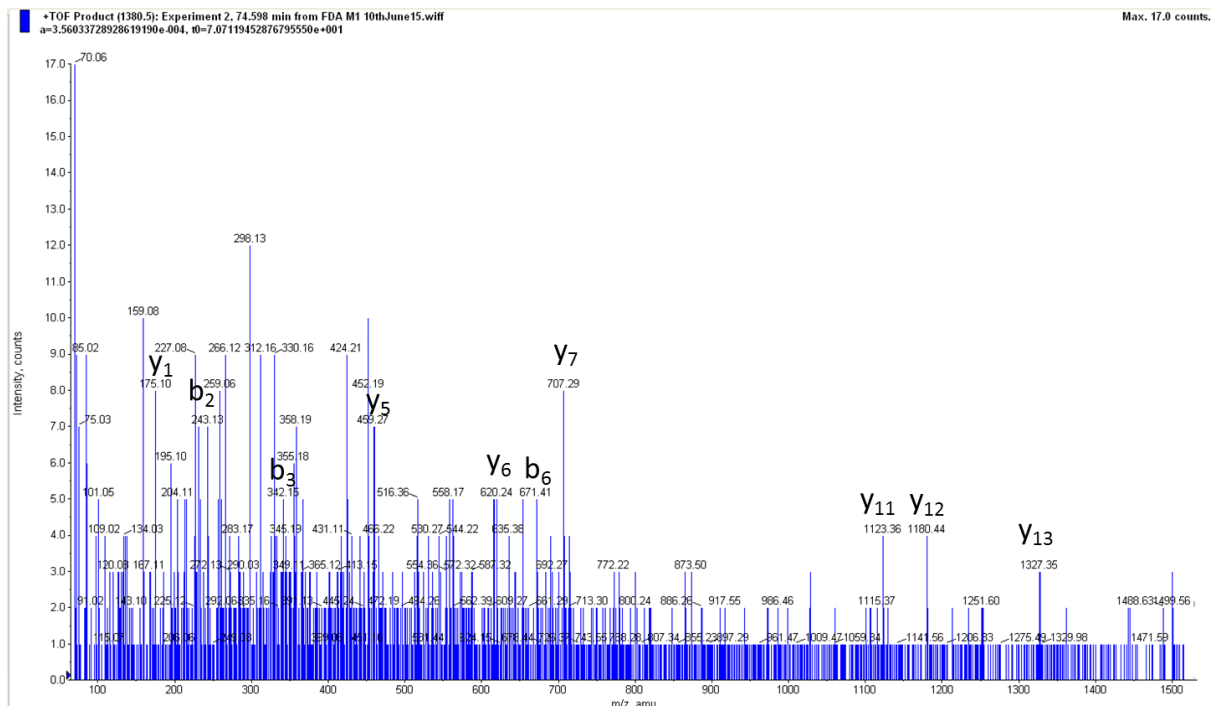


Figure 3.12 MS/MS spectrum of m/z 1380.5⁴⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in **Figure 3.11**.

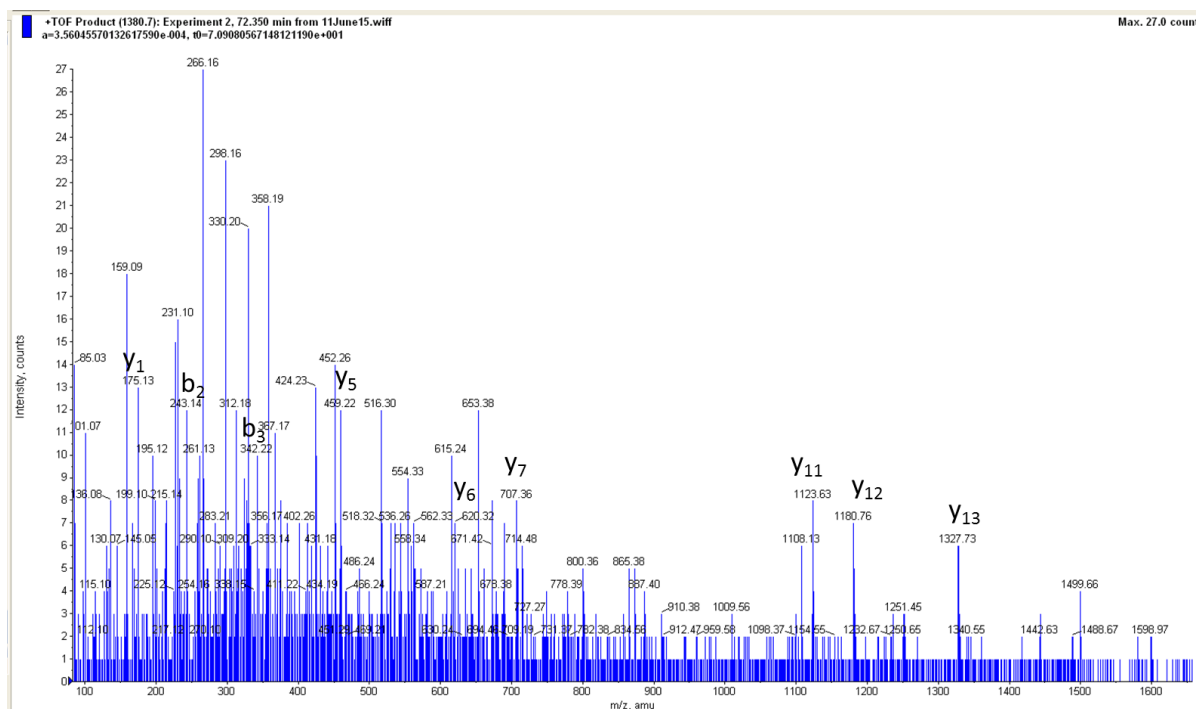


Figure 3.13 MS/MS spectrum of m/z 1380.65⁴⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both **Figure 3.11** and **Figure 3.12**.

TSR5:

The fourth predicted TSR was mis-classified in the original publication (TSR4) since it does not have the second conserved Cys residue. However, the ordering nomenclature is preserved here, and the fifth predicted thrombospondin glycosylation consensus sequence in ADAMTS13 lies within the tryptic peptide TGAQAAHVWTPVAGSJSVSJGR (residues 889-910), MW 2260.0, predicted $[M + 3H]^{3+}$ 754.30³⁺. There is a significant signal for the free peptide in the MS chromatogram at 37.7 minute and a related signal is found at m/z 857.0³⁺ ion and the MS/MS spectrum is shown in **Figure 3.14** which corresponds to a mass increment of 308 Da. This proves the post-translational modification of peptide 754.30³⁺ with a HexFuc because the precursor ion decomposes with the elimination of 308 Da to give the quasi-molecular ion of the TGAQAAHVWTPVAGSJSVSJGR peptide seen at m/z 1130.99²⁺ together with b and y ions which define this peptide sequence.

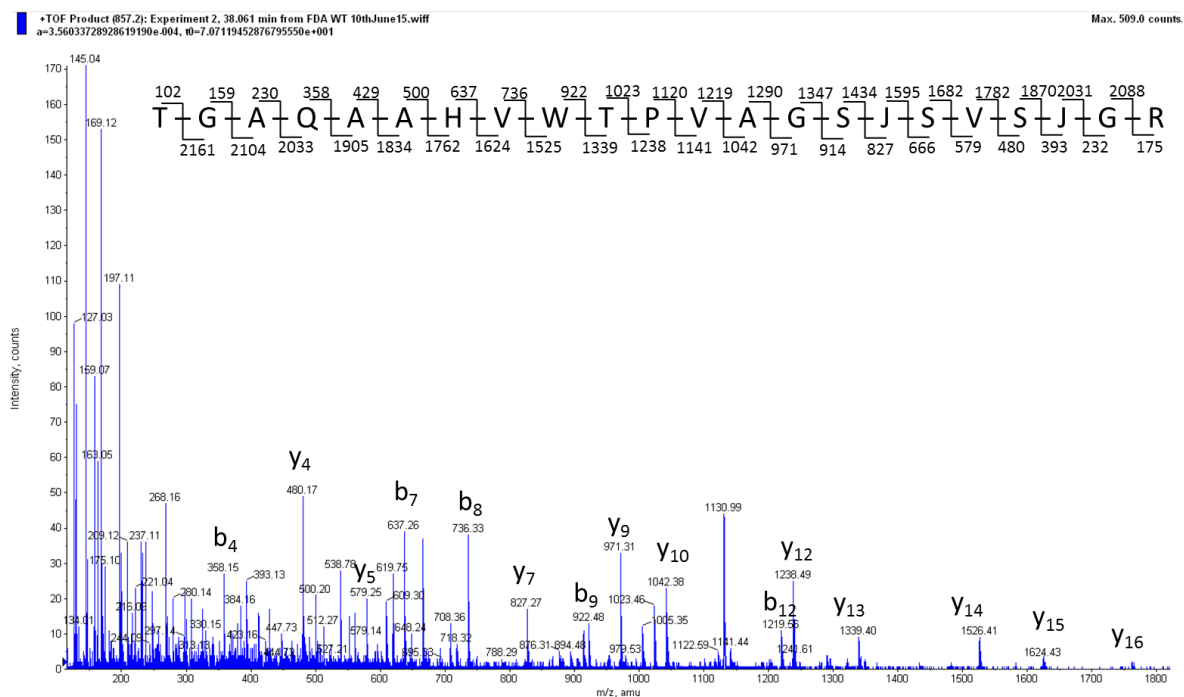


Figure 3.14 MS/MS spectrum of m/z 857.2³⁺ in WT ADAMTS13. Peptide fragmentation provides evidence for the sequence TGAQAAHVWTPVAGSJSVSJGR.

Comparing these data with the P118P and P118F samples, m/z 857.0³⁺ is also found in these data sets and the MS/MS spectra are shown for the respective samples in **Figure 3.15** and **3.16**. In terms of the b and y ions and the quasi-molecular ion generated, the glycopeptide assigned is equivalent to TSR5 in the WT data.

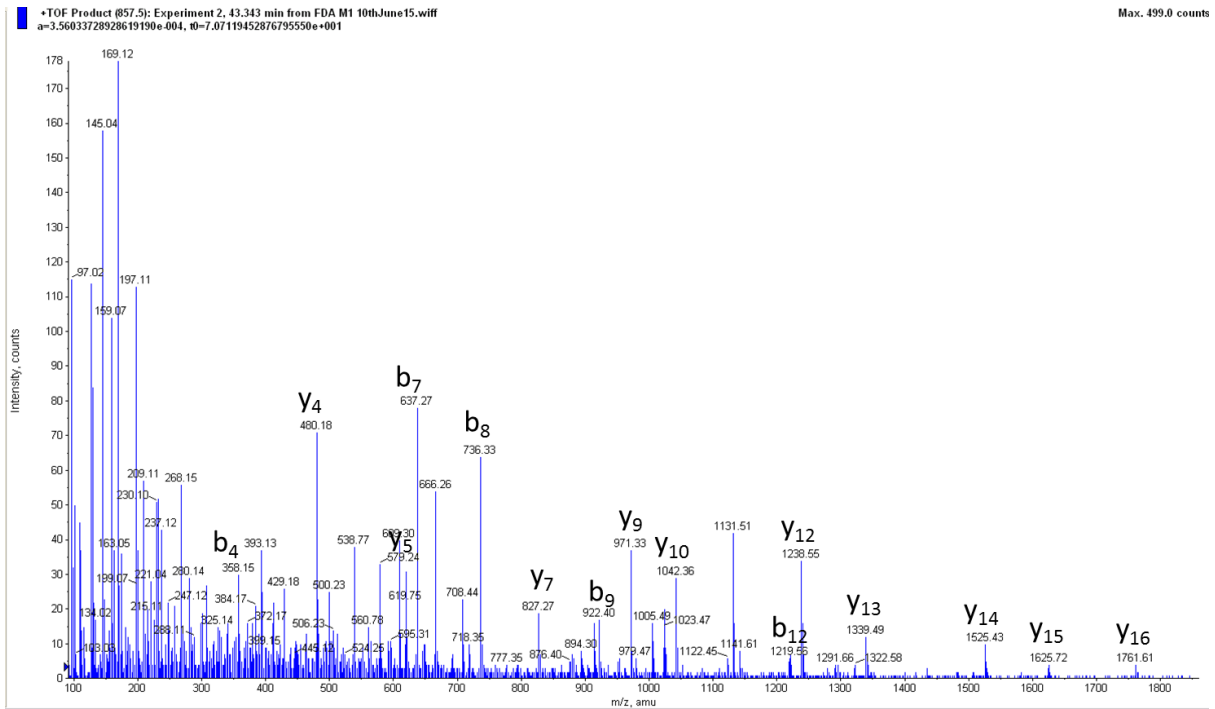


Figure 3.15 MS/MS spectrum of m/z 857.2³⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in **Figure 3.14**.

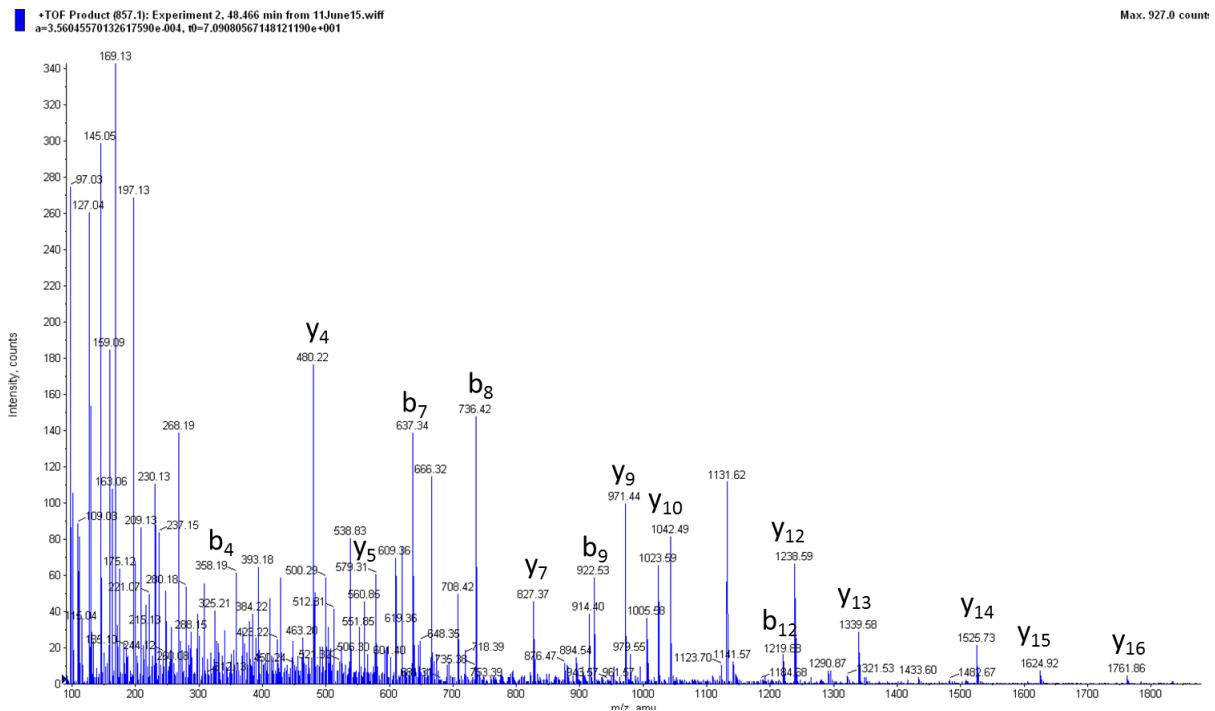


Figure 3.16 MS/MS spectrum of m/z 857.2³⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both **Figure 3.14** and **Figure 3.15**.

TSR6:

The sixth predicted thrombospondin consensus sequence in ADAMTS13 lies within the tryptic peptide LAAJSVSJGR (residues 959-968), MW 1081.45, predicted $[M + 2H]^{2+}$ 541.73²⁺. There is a significant signal for the free peptide in the MS chromatogram at 27.47 minute and a related signal is found at m/z 695.90²⁺ ion and the MS/MS spectrum is shown in **Figure 3.17**. This proves the post-translational modification of peptide 541.73²⁺ with a HexFuc because the precursor ion decomposes with the elimination of 308 Da to give the quasi-molecular ion of the LAAJSVSJGR peptide seen at m/z 1082.46, together with b and y ions which define this peptide sequence.

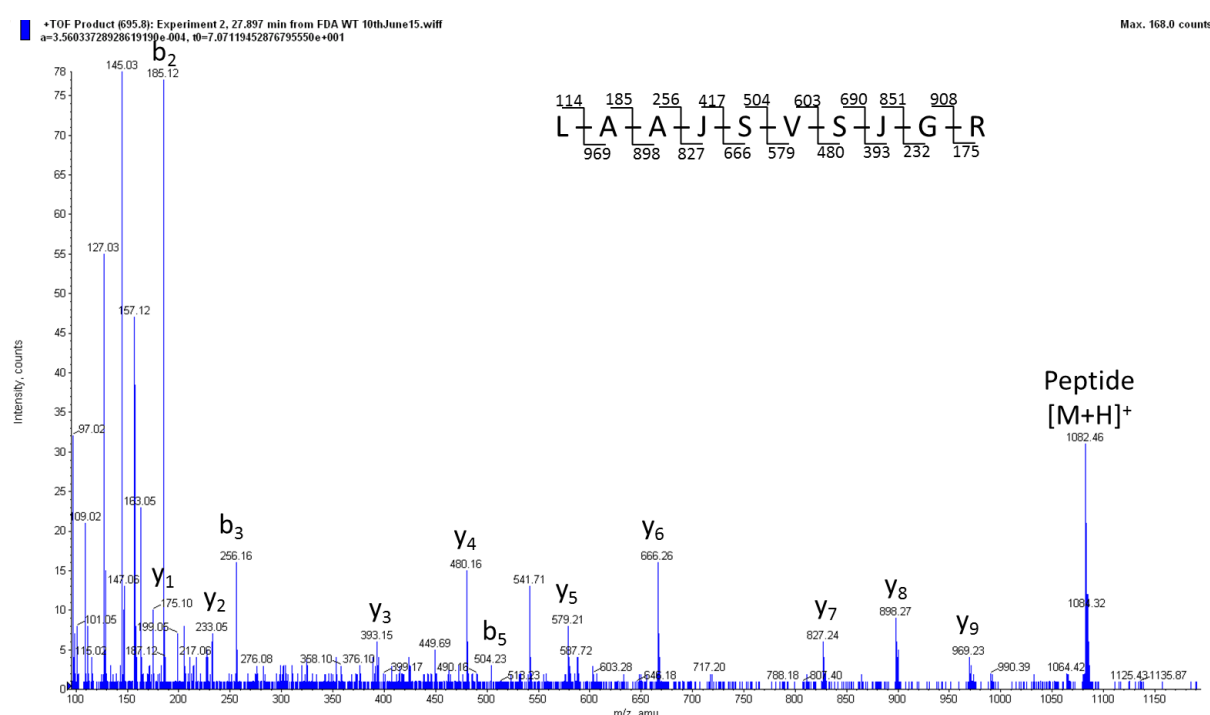


Figure 3.17 MS/MS spectrum of m/z 695.90²⁺ in WT ADAMTS13. Peptide fragmentation provides evidence for the sequence LAAJSVSJGR.

Comparing these data now with the P118P and P118F samples, m/z 695.90²⁺ is also found in these data sets and the MS/MS spectra are shown for the respective samples in **Figure 3.18** and **3.19**. In terms of the b and y ions and the quasi-molecular ion generated, the glycopeptide assigned is equivalent to TSR6 in the WT data.

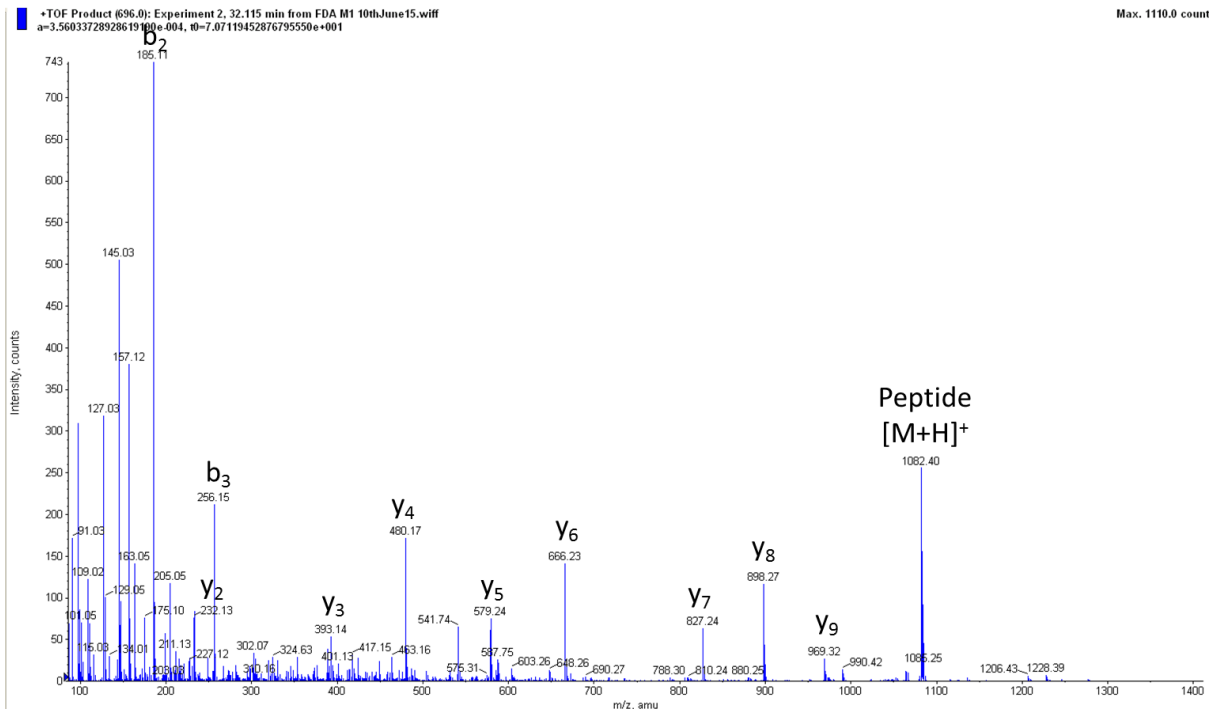


Figure 3.18 MS/MS spectrum of m/z 695.90²⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in **Figure 3.17**.

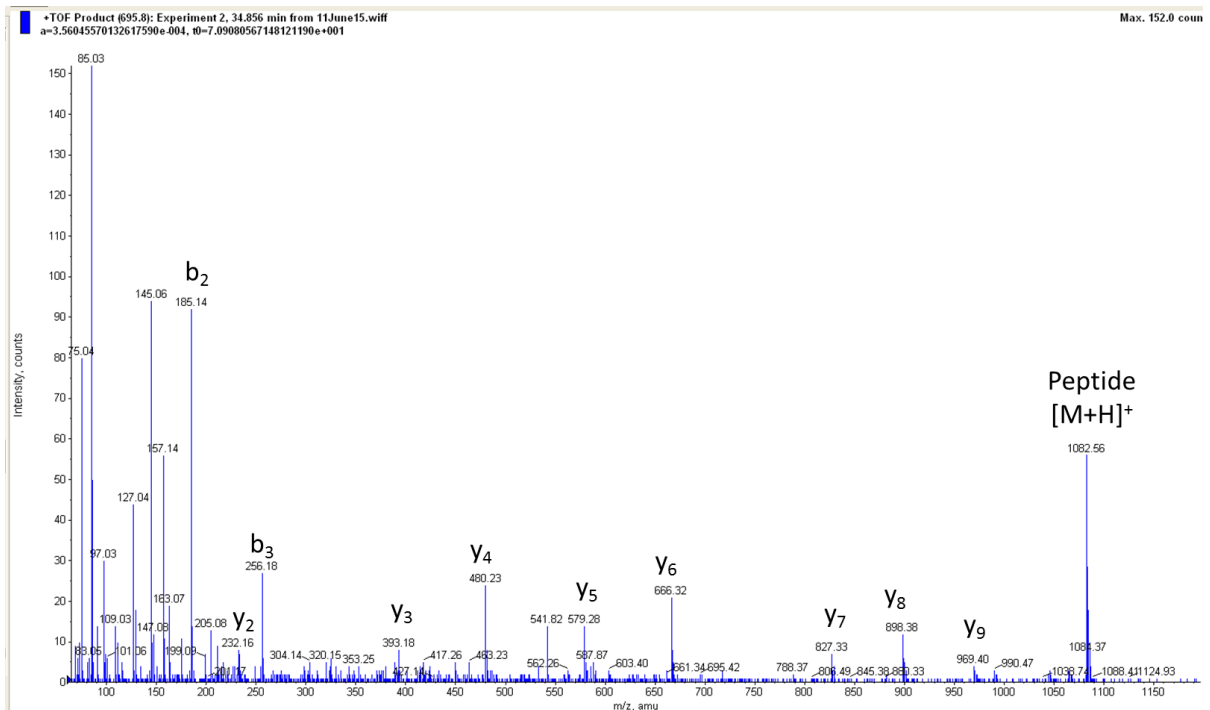


Figure 3.19 MS/MS spectrum of m/z 695.90²⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both **Figure 3.17** and **Figure 3.18**.

TSR7:

The seventh predicted thrombospondin glycosylation consensus sequence in ADAMTS13 lies within the tryptic peptide VMSLGPJSASJGLGTAR (residues 1018-1034), MW 1724.75, predicted $[M + 2H]^{2+}$ 863.38²⁺. Related signals are found in the case of this TSR at 936.4²⁺ and 1017.4²⁺ at different times in the LC-MS chromatogram (**Figure 3.20**), 42.8 and 42.0 minute respectively.

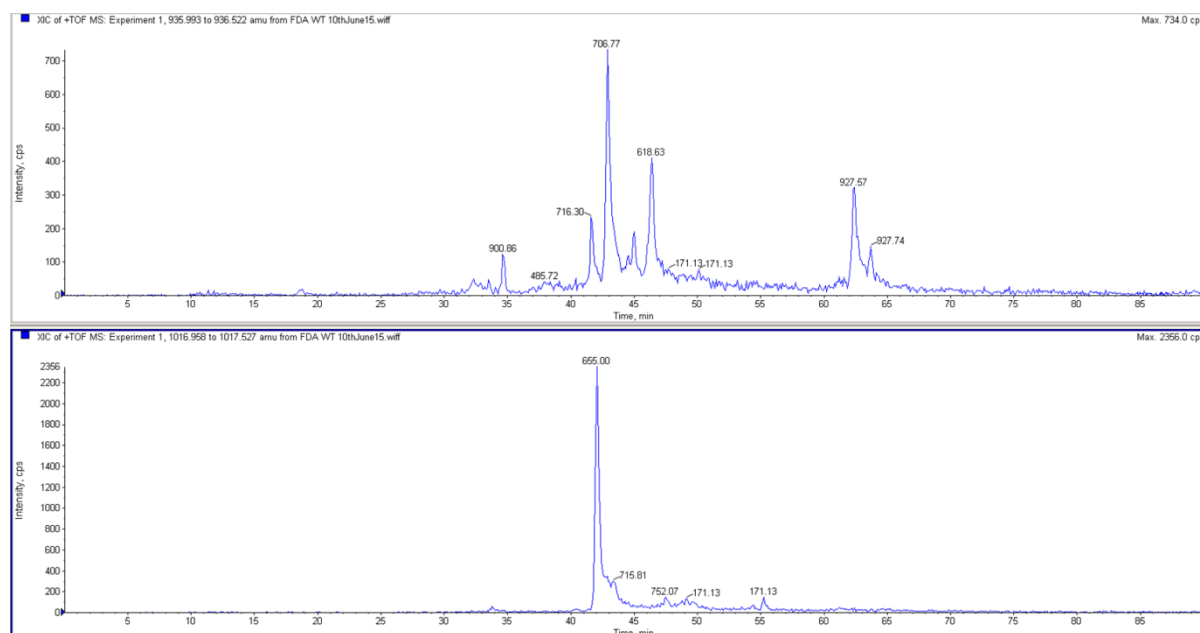


Figure 3.20 LC-MS chromatogram of 936.4²⁺ (top panel) and 1017.4²⁺ (bottom panel) in WT ADAMTS13.

The MS/MS spectra for these signals are shown in **Figure 3.21** and **3.22**. The MS/MS spectrum of 936.4²⁺, 1871.8 $[M+H]^+$, shows a clear decomposition to give the free peptide VMSLGPJSASJGLGTAR recognised by its b and y sequence ions for example (among others) at m/z 231 (b₂) and 246 (y₂). This proves that the 1871.8 $[M+H]^+$ has decomposed on the MS/MS to give 1725.7 $[M+H]^+$ and its subsequent fragments, and corresponds to a loss of 146 Da, the mass of fucose. This is the only one of the TSRs found in this study as a single fucosyl substitution rather than the HexFuc substitution found in earlier examples. The MS/MS of 1017.4²⁺ (**Figure 3.22**) again shows decomposition to give the VMSLGPJSASJGLGTAR peptide, but this time the loss is 308 mass units corresponding to a HexFuc unit, as seen in the other TSRs. The significance of the fucosyl peptide as a separate molecular entity is not understood since the other glycopeptides found in this study do not

show evidence of significant decomposition by loss of hexose, so sample handling and experimental method can probably be discounted.

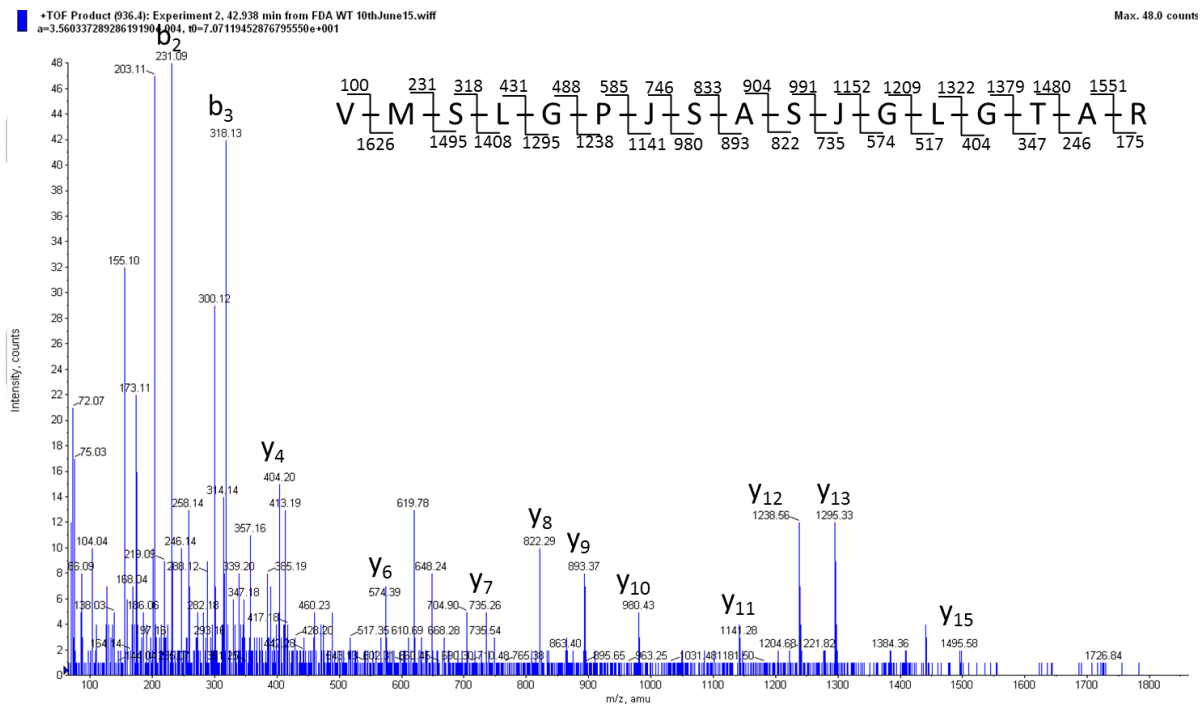


Figure 3.21 MS/MS spectrum of m/z 936.40²⁺ in WT ADAMTS13. Peptide fragmentation provides evidence for the sequence VMSLGPJSASJGLGTAR. The peak at m/z 1726 corresponds to a loss of 146 Da (fucose) from m/z 1871 [M+H]⁺.

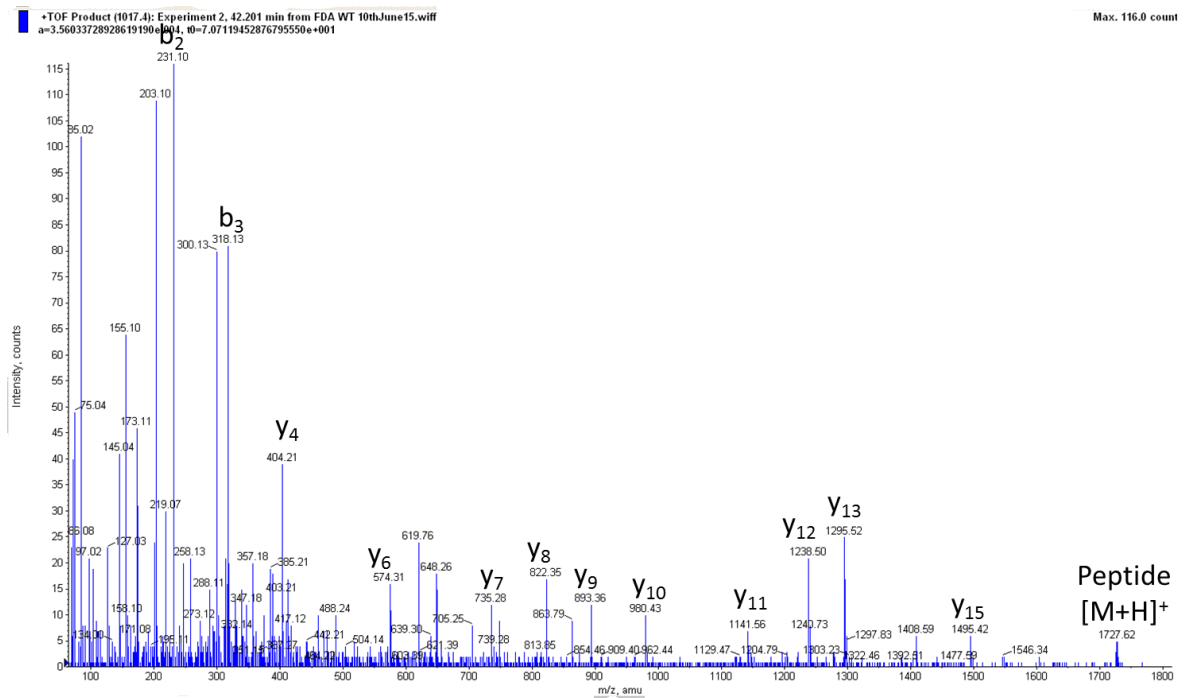


Figure 3.22 MS/MS spectrum of m/z 1017.4²⁺ in WT ADAMTS13. Peptide fragmentation provides evidence for the sequence VMSLGPJSASJGLGTAR.

Comparing these data with the P118P sample, m/z 936.4²⁺ and 1017.4²⁺ are also found in these data sets and the MS/MS spectra are shown for the respective samples in **Figure 3.23** and **3.24**. In terms of the b and y ions and the quasi-molecular ion generated, the fucosyl- and the HexFuc-glycopeptide are both found to be present equivalent to TSR7 in the WT data.

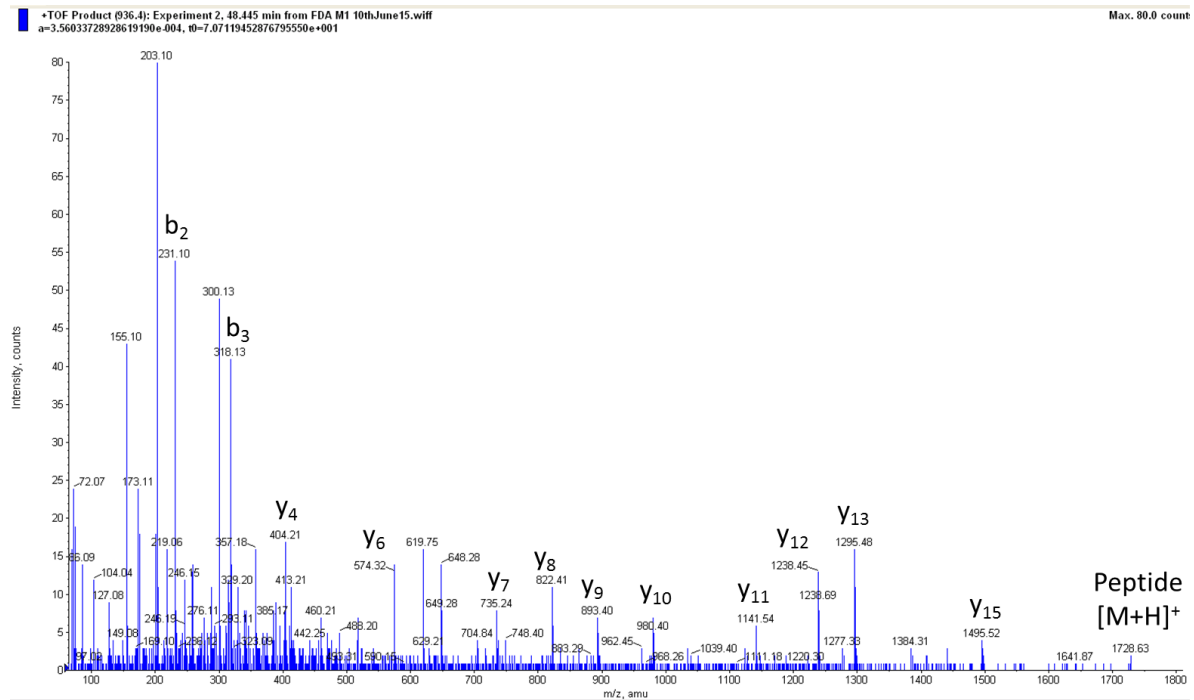


Figure 3.23 MS/MS spectrum of m/z 936.40²⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in **Figure 3.21**.

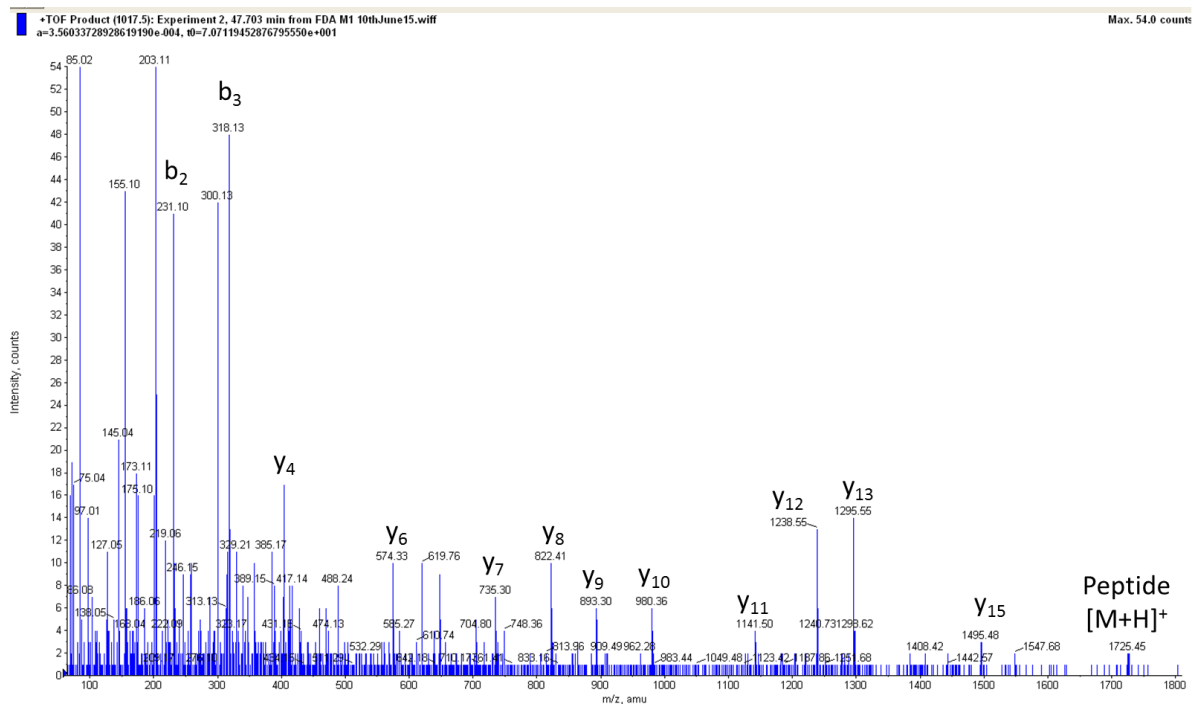


Figure 3.24 MS/MS spectrum of m/z 1017.4²⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in **Figure 3.22**.

The data for P118F sample also show the 936.4²⁺ and 1017.4²⁺ signals in the MS spectra, but in this case the 936.4²⁺ was too weak to be automatically chosen for MS/MS. The 1017.4²⁺ MS/MS data is shown in **Figure 3.25**, and overall although the P118F data are weaker, there is evidence that TSR7 site in both P118P and P118F is equivalent to the WT site in carrying both fucose and HexFuc disaccharide.

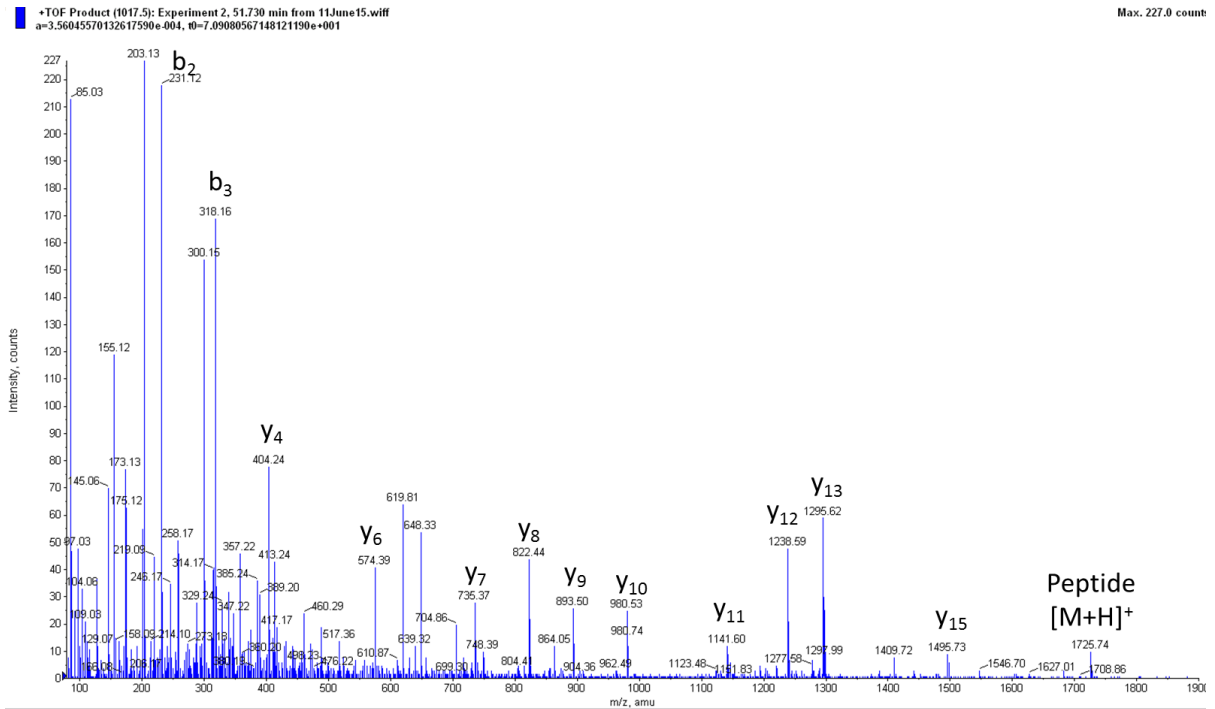


Figure 3.25 MS/MS spectrum of m/z 1017.4²⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both **Figure 3.22** and **Figure 3.24**.

TSR8:

The eighth predicted thrombospondin glycosylation consensus sequence in ADAMTS13 lies within the tryptic peptide WHVGTWMEJSVSJGDGIQR (residues 1076-1094), MW 2265.92, predicted $[M+3H]^{3+}$ 756.32³⁺. A related signal is found at m/z 859.0³⁺ (an increment of 308 Da) and the MS/MS spectrum is shown in **Figure 3.26**. This proves the post-translational modification of peptide 756.32³⁺ with a HexFuc because the precursor ion decomposes with the elimination of 308 Da to give the quasi-molecular ion of the WHVGTWMEJSVSJGDGIQR peptide seen at m/z 1133.9²⁺, together with b and y ions which define this peptide sequence.

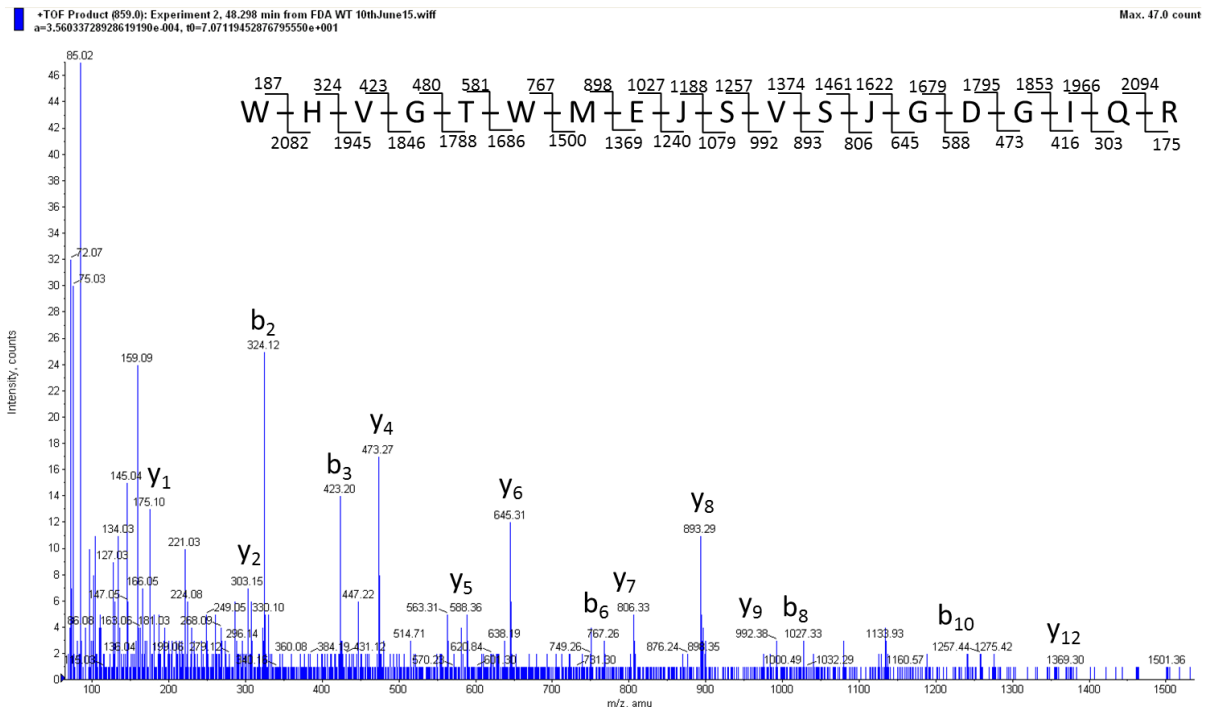


Figure 3.26 MS/MS spectrum of m/z 859.0³⁺ in WT ADAMTS13. Peptide fragmentation provides evidence for the sequence WHVGTWMEJSVSVJGDGIQR.

Comparing these data with the P118P and P118F samples, m/z 859.0³⁺ is also found in these data sets and the MS/MS spectra are shown for the respective samples in **Figure 3.27** and **3.28**. In terms of the b and y ions and the quasi-molecular ion generated, the glycopeptide assigned is equivalent to TSR8 in the WT data.

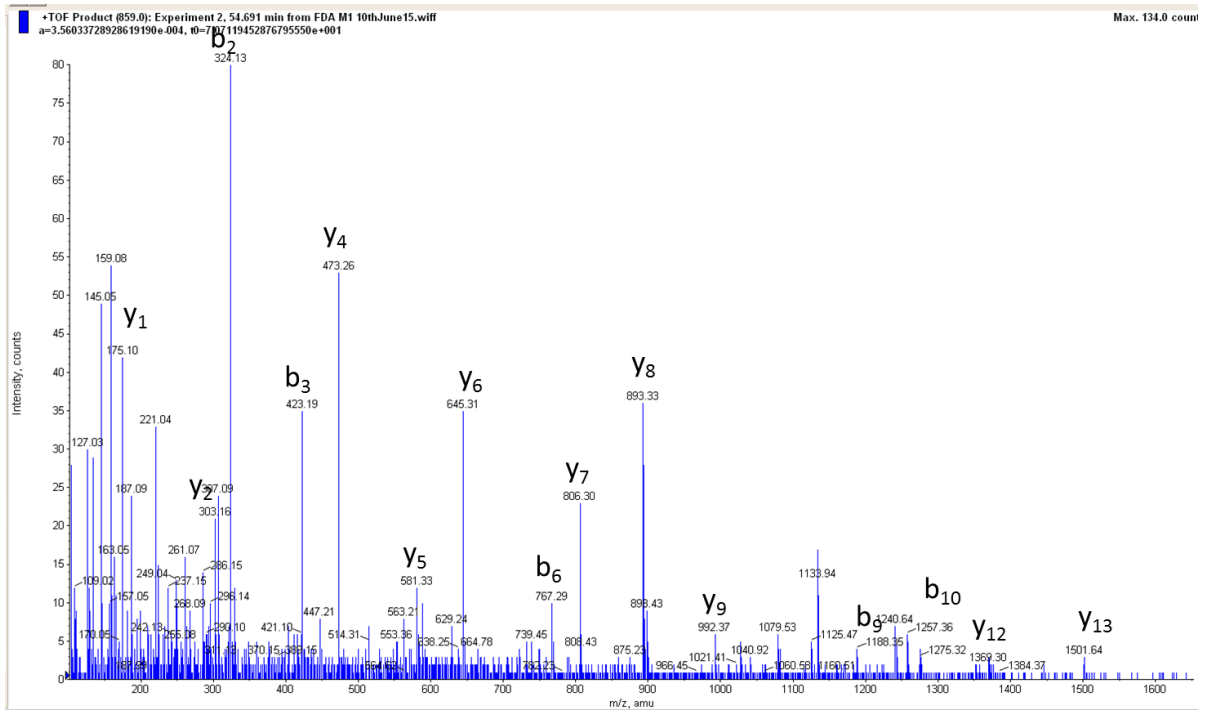


Figure 3.27 MS/MS spectrum of m/z 859.0³⁺ in P118P ADAMTS13. This MS/MS spectrum presents the same peptide fragmentation of the one shown in Figure 3.26.

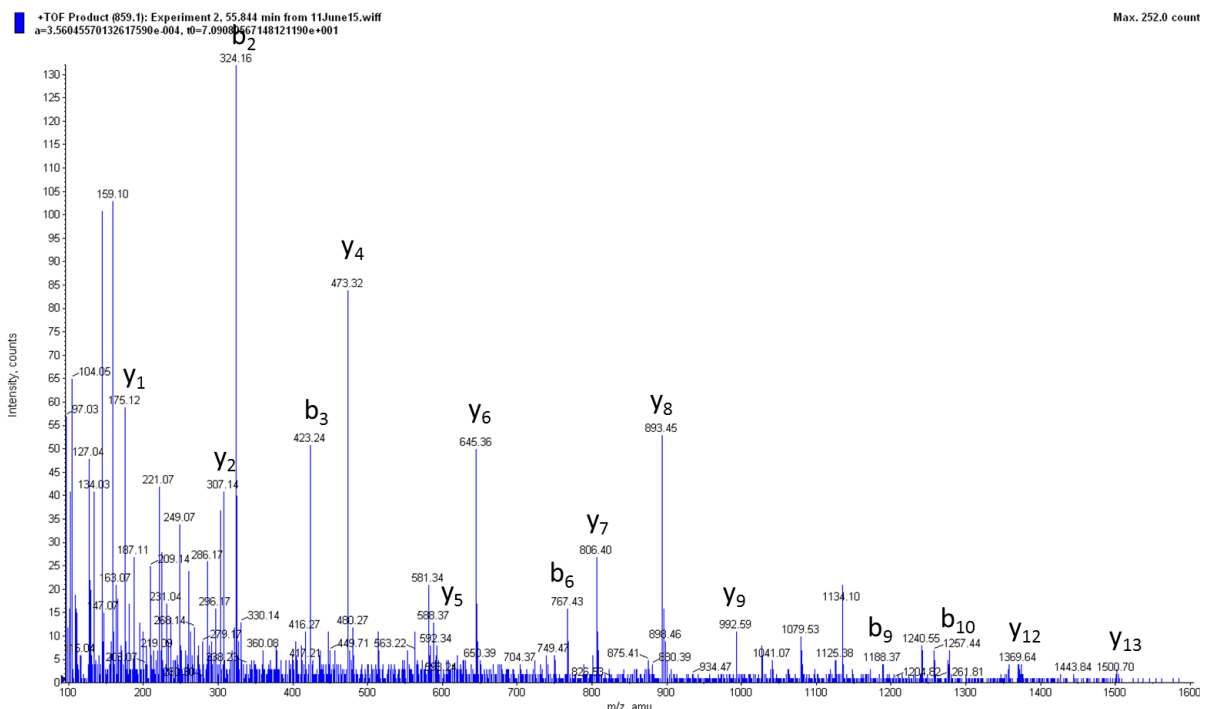


Figure 3.28 MS/MS spectrum of m/z 859.0³⁺ in P118F ADAMTS13. The MS/MS spectrum is very similar to the one shown in both Figure 3.26 and Figure 3.27.

3.3.2 Other Discoveries

A. Further mapping analysis of WT ADAMTS13 then led to the discovery of other undescribed PTM events at specific amino acid sites in the molecule. One of these was found at T31 (residues 387-393 in **Table 3.1**). Here a very peculiar loss was observed for a tryptophan containing peptide in a MS/MS spectrum of an ion at 519.2^{2+} , exhibiting a loss to 450.2^{2+} corresponding to 138 Da, which is not the mass of any amino acid or sugar. Detailed examination of the spectrum shown in **Figure 3.29** allows the manual identification of this peptide through a y ion series at m/z 175, 272, 329, 515, 602 and 689 corresponding to a partial sequence S-S-W-G-P-R assignable to residues 388-393 in the ADAMTS13 sequence. Interestingly these residues are predicted to be preceded in the sequence by a tryptophan residue, W-387, but inspection of the spectrum shows no evidence of an N-terminal tryptophan at m/z 187. However, the mass difference between the MS/MS signal of 519.2^{2+} equivalent to $M = 1036.4$ and the theoretical peptide 387-393 which is 874.4 is 162 Da which is the mass of hexose residue, and the sequence data analysis above strongly suggests that the N-terminal tryptophan residue of this peptide must be hexosylated. A literature search at this point for this tryptophan PTM found an X-ray study of a recombinant ADAMTS13 fragment, residues 287-685 (Akiyama, Nakayama et al. 2013) in which the electron density map suggested a C-mannosylation on the side chain of Trp 387, although the paper also states that in a parallel study of the WT fragment structure the electron density map was not clear enough to assign. A further search shows that various other studies on thrombospondins and smaller ADAMTS family members have suggested a hexosylation of either Trp 387 or Trp 390. Since the work reported here would be the first definitive demonstration of the hexosylation of intact ADAMTS13 residue Trp 387 it was therefore important to understand the strange MS/MS fragmentation observed for this molecule as seen in **Figure 3.29**. The spectrum can be rationalised by assuming two initial losses of water from the hexose ring, and a further collapse to give an acetylinic indole side chain on the first Trp. In total this corresponds to a loss of 138 Da and to the major doubly charged ion originally observed at m/z 450.2^{2+} . **Figure 3.29** is annotated to show how the other key signals in the spectrum eg m/z 156, 294, 744 together with the b_1 , b_2 and b_3 ions which are 24 Da higher than normal, fit in with this suggested fragmentation mechanism.

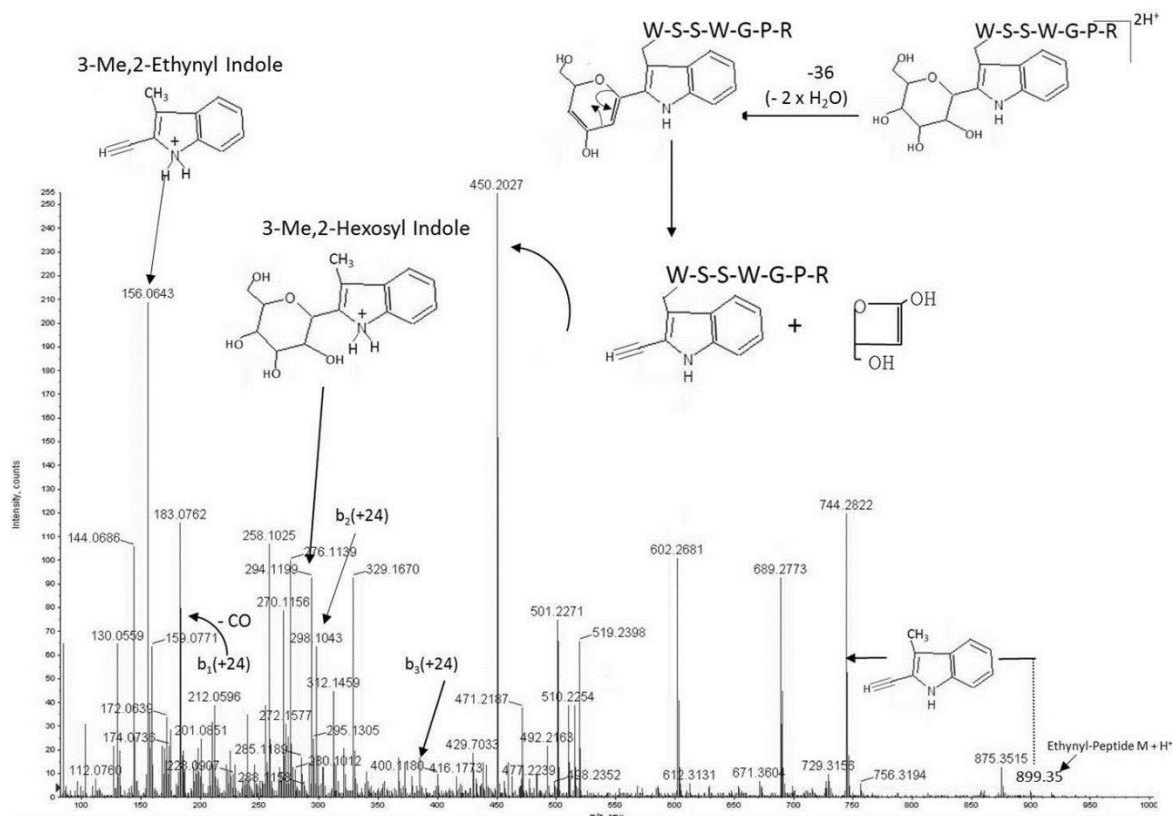


Figure 3.29 Glycosylation of W-387 in ADAMTS13. The quasi-molecular ion is 162 Da, a hexose unit, higher than the theoretical peptide mass, and the position of substitution of the hexose is shown to be on the first tryptophan, W-387, as proven by the interpretation of the fragment ions observed, and illustrated in schematic with suggested mechanisms. The principal fragment ions observed correspond to the novel formation of a 2-Ethynyl-Indole (Acetylenic substituent) on the side chain of W-387 resulting from loss of 138 Da caused by water loss and partial cleavage of the hexose ring, in preference to the normal β -elimination (162 Da) seen in O-linked glycosylation chemistry.

B. In addition, a previously unreported O-glycosylation of WT ADAMTS13 was found in this work at residue Ser-1170 which is present on the border of the TRS1-8 and CUB1 protein domains. This new finding was made by screening using the sugar reporter ions (low mass fragment ions) at m/z 163, 204, 292 and its water loss at 274. **Figure 3.30** provides the evidence that (a) the peptide portion of the molecule corresponds to residue 1166-1176 GLLFSPAPQPR $M+H^+$ 1182.68⁺ and (b) the peptide is O-glycosylated at Ser-1170 by DiSialyl Core-1 (NeuAc₂HexHexNAc) as interpreted from the signals at m/z 592.34 (doubly charged peptide), 693.38, 774.41, 839.41, 919.94 and 1065.50. At a slightly earlier retention time, a lesser amount of a NeuAc₂Hex₂HexNAc₂ glycoform is also present (m/z 832.3³⁺).

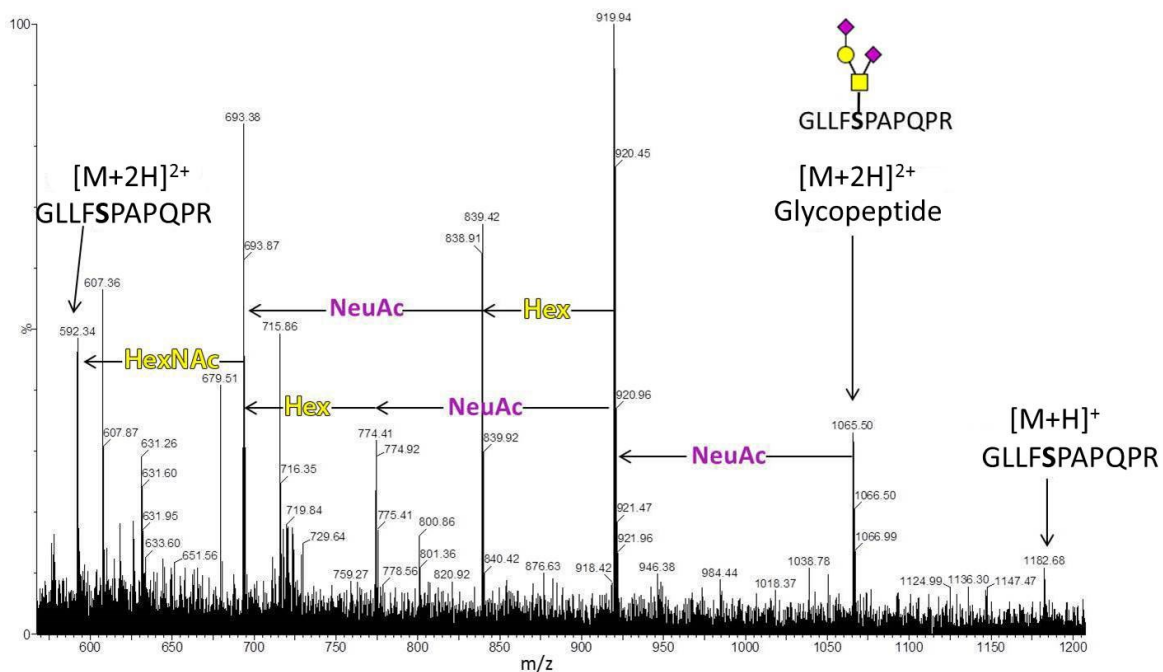


Figure 3.30 The partial MS spectrum (m/z 550-1200) at the corresponding LC-MS elution time where the MS/MS of a signal at 1065.5²⁺ gave b and yⁿ signals attributable to the sequence GLLFSPAPQPR in WT ADAMTS13. The spectrum indicates an ion at 1065.5²⁺ and a corresponding quasi-molecular ion for the peptide component of 1182.6 (591.8²⁺) which calculates for a NeuAc₂HexHexNAc unit attached to Ser-1170 of the peptide sequence assigned. From an understanding of mammalian biosynthetic pathways, this likely corresponds to a DiSialyl Core-1 structure.

This newly assigned ADAMTS13 glycopeptide substituting Ser-1170 is also observed in the P118P and P118F sample data sets (**Figure 3.31** and **3.32**).

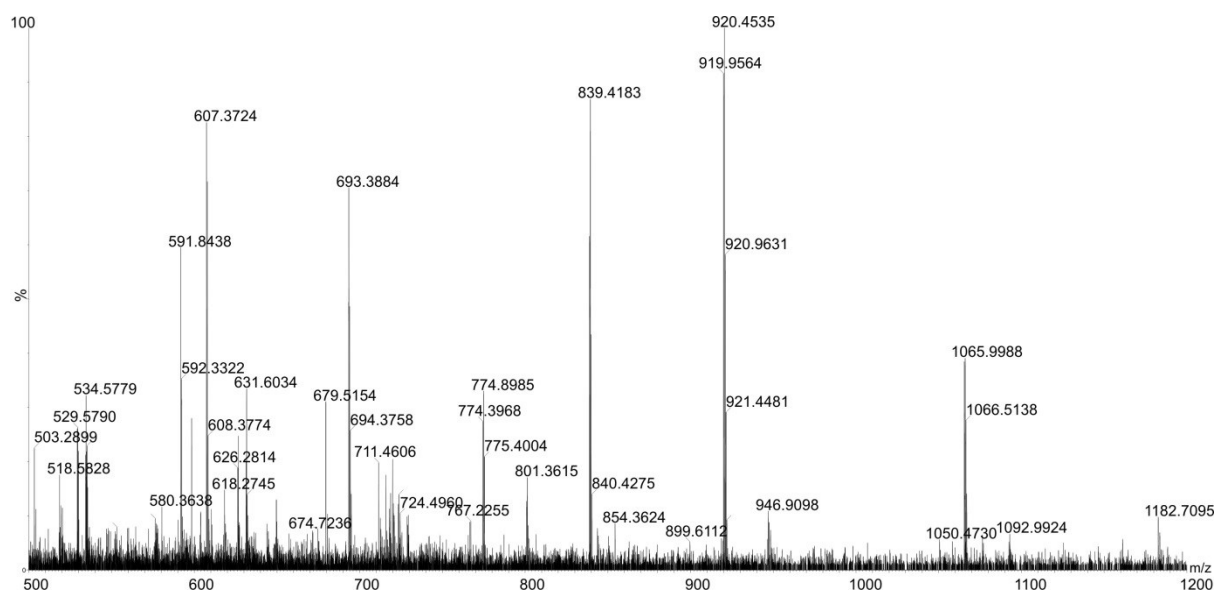


Figure 3.31 The partial MS spectrum (m/z 500-1200) where the MS/MS of a signal at 1065.5²⁺ gave b and yⁿ signals attributable to the sequence GLLFSPAPQPR in P118P ADAMTS13.

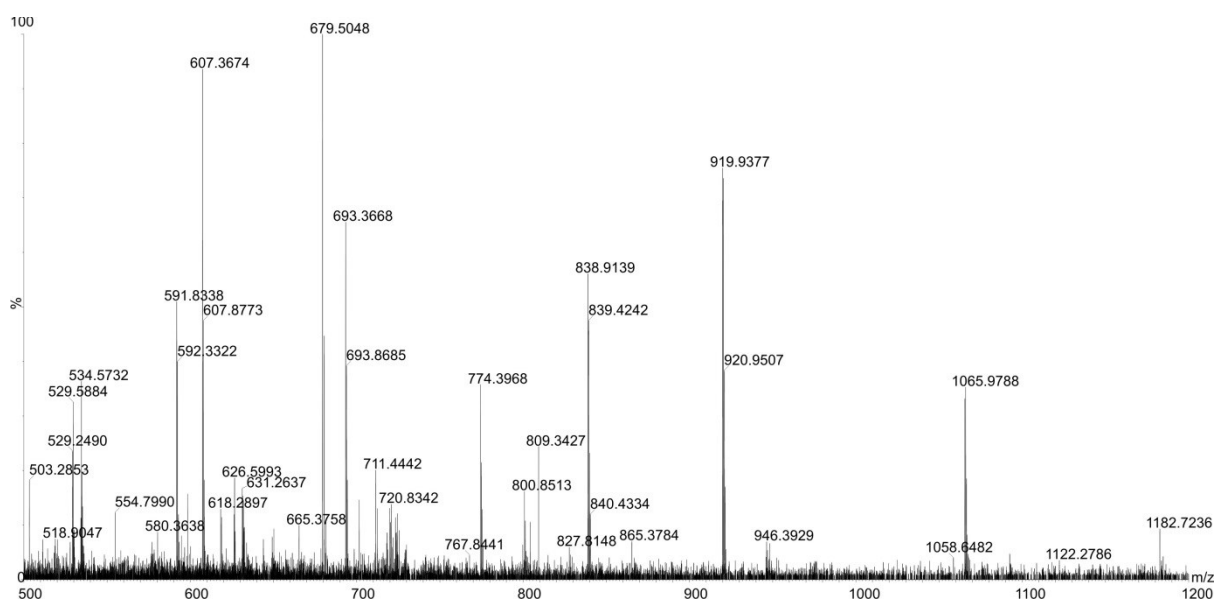


Figure 3.32 The partial MS spectrum (m/z 500-1200) where the MS/MS of a signal at 1065.5²⁺ gave b and yⁿ signals attributable to the sequence GLLFSPAPQPR in P118F ADAMTS13.

3.4 Discussion

Synonymous mutations have long been assumed to lack pathogenic relevance, and are often ignored owing to the belief that they have insignificant effects on protein expression and function. Consequently, there has been a concentration of research on structural and functional characteristics of non-synonymous mutations, while studies of synonymous mutations have been few in comparison. However, in recent years, it has become apparent that synonymous mutations can affect the expression, structure and functionality of proteins through indirect means, ranging from mRNA stability to the kinetics of protein translation (Hunt, Simhadri et al. 2014). This is reflected by an increasing amount of published work linking synonymous mutations to human disease (Sauna and Kimchi-Sarfaty 2011).

In the study initiated by our FDA collaborators, they focused on the characterisation of a single synonymous substitution, P118P in the *ADAMTS13* gene product. They compared the properties of this variant with WT and a non-synonymous control variant P118F. They chose a control variant that would demonstrate the sensitivity of the study methods due to the expectation that there would be some similarities in *ADAMTS13* properties of WT and the synonymous P118P variant. Remarkably, the P118P mutation produces an increase in

extracellular expression (secretion) of over 25% compared to its WT counterpart, with similar specific activity levels.

One protein characteristic, which has previously been shown to affect protein secretion, is a novel small O-linked modification on serine residues (O-fucosylation) present in certain sequence stretches found in a number of proteins, including the blood protein ADAMTS13, called Thrombospondin repeats (TSRs) (Hofsteenge, Huwiler et al. 2001). The possible enhancement or wider distribution of such a PTM could therefore have been responsible for the increased secretion seen in the synonymous mutant, or its origin could of course relate to an unknown determinant. The collaborative work in this chapter has been an initial attempt to study the comparative ADAMTS13 O-glycomes of WT, synonymous and non-synonymous mutants in order to help answer the above question on causation of enhanced secretion.

There is also clear interest in this field from both the FDA and the Biotechnology industry where increased yields of functional biopharmaceutical products would be an obvious benefit to the affordability of the final medicinal products. The glycoprotein analysis reported here reveals the finding of very similar PTMs in WT and both variants ADAMTS13. Seven O-fucosylated TSRs, six of which were known to be essential for ADAMTS13 secretion, were found. It is thought that the O-fucosylation of these domains is carried out by POFUT2 which requires prior folding of the protein in order to function. Therefore, this modification serves as a form of quality control (Ricketts, Dlugosz et al. 2007). As a result, it can be thought that all three secreted variants were properly folded as far as TSR domains are considered. It also should be noted that several novel PTMs not previously reported, namely O-glycosylation of TSR1, C-mannosylation of W387 and DiSialyl Core-1 O-glycosylation of S1170, were also discovered in all three variants. Although this structural discovery research cannot be regarded as quantitative, no very obvious differences were seen in the PTMs found in the three preparations, and this now suggests that the future work should concentrate on relative mRNA stability and/or other factors affecting protein translation kinetics, in order to explain differences in the secretion observed. The specific roles of these modifications and importantly, the presence of these PTMs decorating ADAMTS13 in other expression situations remain to be determined. It has been suggested, however, that sialic acid residues may inhibit serum protease activity. In fact, VWF has been found to be extensively glycosylated in this manner and the removal of these negatively charged sialic acid residues at the A1-A2 domain junction leads to increased VWF resistance against ADAMTS13 cleavage. Furthermore, it should also be noted that DiSialyl Core-1 is known to have a high affinity for

SIGLEC 7 (CD328), found in abundance on NK cells, which may relate to ADAMTS13 interaction with other components in blood (Canis, McKinnon et al. 2010).

The P118P ADAMTS13 variant characterised here is a unique example of a single synonymous mutation resulting in improved protein expression without conferring a detrimental impact on protein character. From a Biotechnological standpoint, the O-glycome findings presented in this chapter together with the FDA research support the idea that synonymous mutations, especially those that naturally occur within the human population, can be targetedly introduced into expression constructs for the purpose of increasing protein expression. While predicting the potential impact of uncharacterised synonymous mutations remains a challenging task, several additional synonymous variants exist in ADAMTS13 with similar codon usage metrics to P118P. Such variants could have additive or synergistic effects and are likely to be increasingly employed in the large scale production of therapeutic proteins or for the development of gene delivery technologies.

Chapter 4:
Type B Flagellin of Hypervirulent
***Clostridium difficile* Strains**

4. Type B Flagellin of Hypervirulent *Clostridium difficile* Strains

4.1 Historical Perspective on Type A and Type B flagellin

In previous work carried out by Twine and co-workers, a preliminary structure of flagellin glycosylation in the model strain 630 and some virulent strains had been determined by LC-MS and MS/MS analysis. The *C.difficile* flagellin is composed of a single flagellin protein, which is post-translationally modified by an O-linked sugar, essential for motility. In this original *C.difficile* sequenced strain, 630, it was shown to be modified at up to seven sites with an O-linked N-acetyl glucosamine (GlcNAc) linked to a methylated threonine via a phosphate group, termed Type A PTM in **Figure 4.1** (Twine, Reid et al. 2009, Faulds-Pain, Twine et al. 2014).

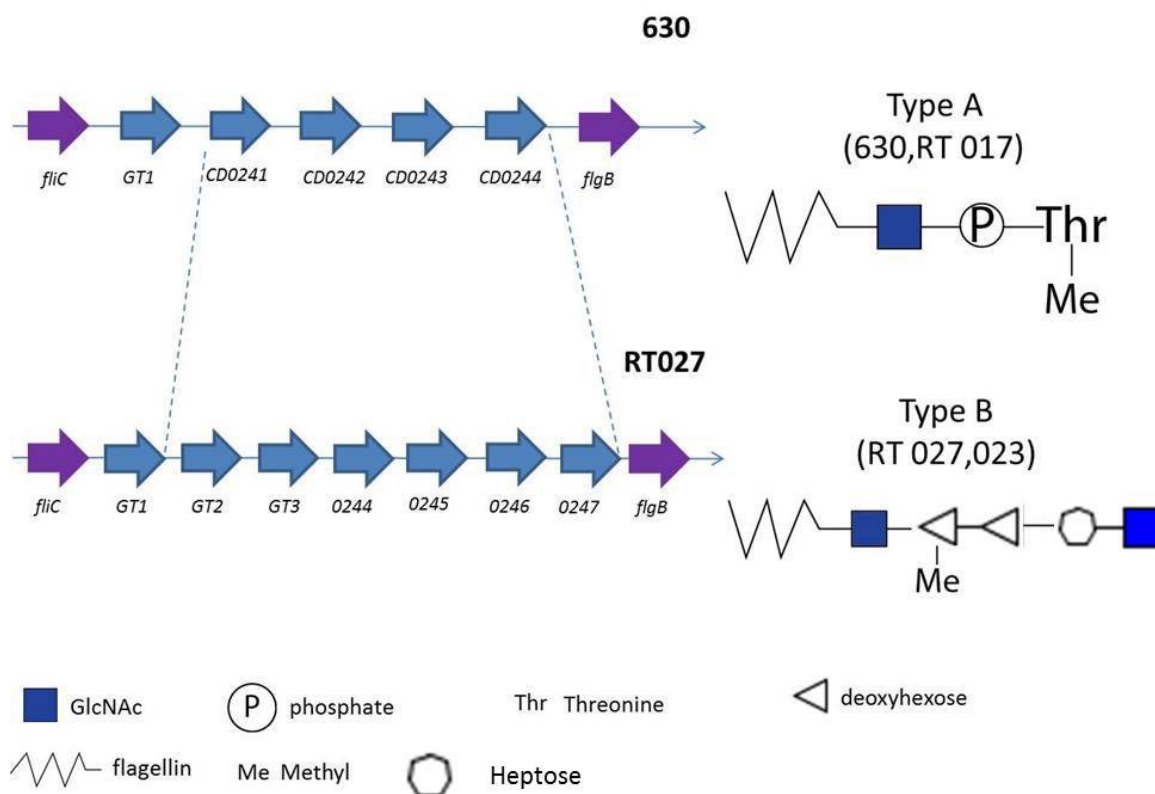


Figure 4.1 Type A and B flagellin in *C.difficile*. The *C.difficile* flagellin is formed of a single flagellin protein which is post-translationally modified by an O-linked sugar. Two types of flagellin PTM have been identified in *C.difficile*, named Type A and B.

The genes responsible for this modification (CD0240, CD0241, CD0242, CD0243 and CD0244) are encoded immediately downstream of the flagellin gene, *fliC*. Two types of flagellin PTM have now been identified in *C.difficile*, defined as Type A and more recently Type B (**Figure 4.1**) based largely on gene cassette analysis, but not rigorously defined (Twine, Reid et al. 2009, Hitchen, Twigger et al. 2010).

Previous genetic analysis of RT027 and RT023 strains, revealed that the gene lying downstream *fliC* in the Type B modification, codes for a glycosyltransferase 1 (GT1), and is highly similar to CD0240 of strain 630. Although the function of GT1 protein is yet to be confirmed, the first sugar of the Type B PTM is also an O-linked N-acetyl hexosamine, specifically a GlcNAc residue.

4.2 Experimental Strategy

As previously discussed in chapter 1, the structural characterisation of glycosylation in prokaryotes cannot rely on assistance from predictable biosynthetic pathways as is the case for eukaryotes, which makes the elucidation of unknown structures, such as presented by this project on the structural characterisation of the *Clostridium difficile* flagella, a much more challenging task.

The glycoproteomic experimental strategy, shown in **Figure 4.2** is a modification of that discussed in chapter 3, and this was initially applied to this project. Whilst this led to advances in our understanding of this PTM, other experimental strategies were needed to solve this structural problem, and these are described later.

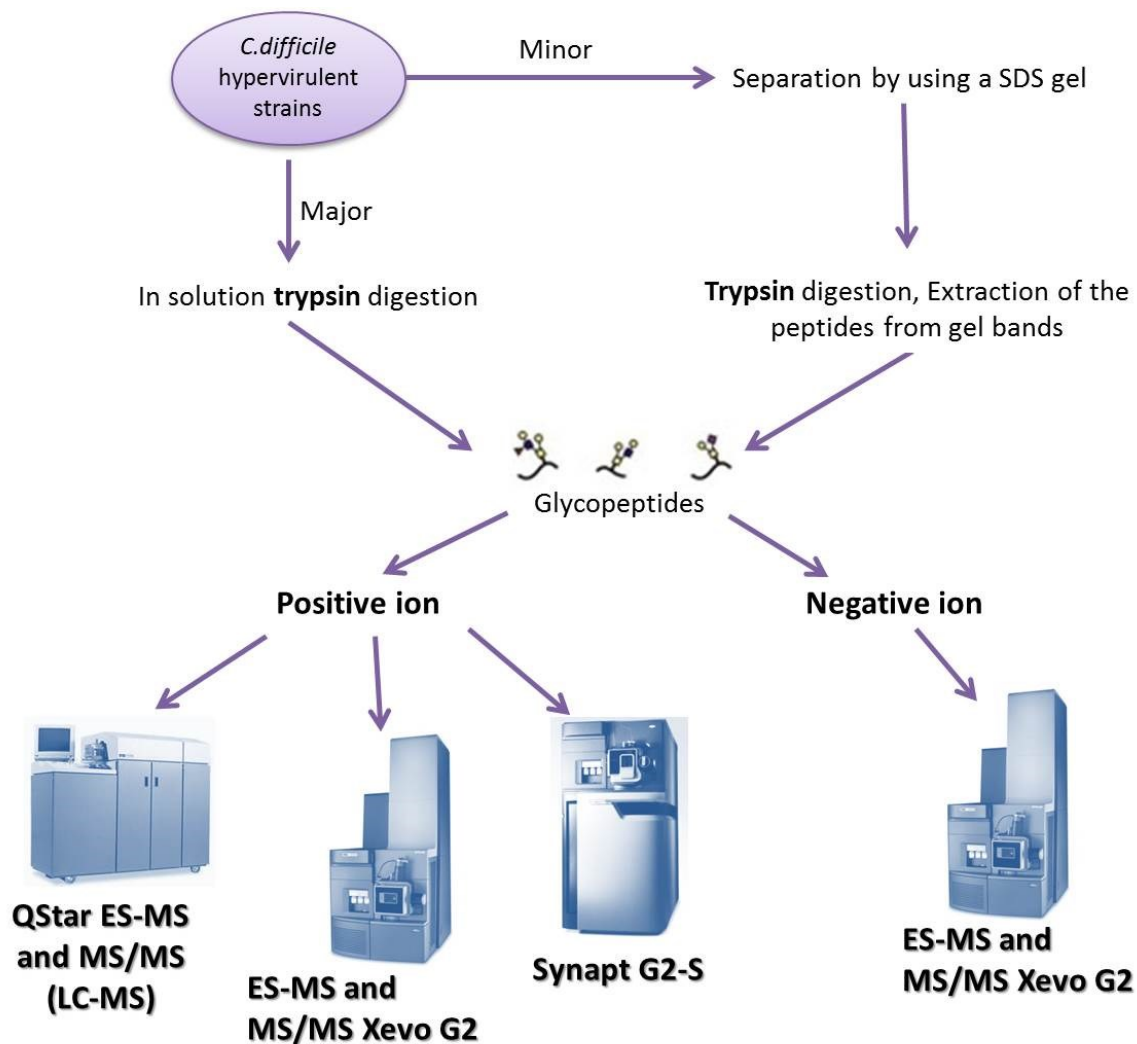


Figure 4.2 Experimental strategy employed in *Clostridium difficile* Type B flagellin structural characterisation. The sample was digested with trypsin, then mass mapping was carried out. The mass spectrometric analysis was done in both positive and negative ion modes using different on-line LC-ES-MS and MS/MS instrumentations.

Several hypervirulent clonal strains of *Clostridium difficile*, including RT027, RT023, RT106 and RT001, were studied in this work, and all of the bacteriology related to the samples listed in **Table 4.1** was carried out by our collaborators, Professor Brendan Wren and colleagues from the Department of Pathogen Molecular Biology at the London School of Hygiene and Tropical Medicine, UK. All flagellin preparations were initially provided as bands run on SDS-PAGE gels at LSHTM using our group protocols, and subsequently at our request as samples in solution.

Strains	Characteristics	Source
<i>Clostridium difficile</i>		
R20291	PCR ribotype 027, isolated from an outbreak in 2004-05	Stabler et al 2009
CD196	ribotype 027, isolated from a pseudomembranous colitis case	France, 1985
BI-16	PCR ribotype 027, isolated from an outbreak in 2004	Augusta, USA
CD305	PCR ribotype 023, isolated from an outbreak in 2011	Barts Hospital, UK
CD1426	PCR ribotype 023, isolated from an outbreak in 2010	Queens Hospital Remford, UK
CD1714	PCR ribotype 023, isolated from an outbreak in 2011	Whipp's Cross Hospital, UK
106-01	PCR ribotype 106, isolated from an outbreak in 2008	Fatal case, Glasgow, UK
001-01	PCR ribotype 001, isolated from an outbreak in 2008	Paisley, UK
001-07	PCR ribotype 001, isolated from an outbreak in 2008	Edinburgh, UK

Table 4.1 Strains of *C.difficile* studied in Professor Wren's laboratories.

A variety of structural analysis techniques were combined in the initial approach to the characterisation, including ES-MS strategies using Q-TOF technology in normal and nanospray ionisation modes, comprising positive- and negative-CID MS and MS/MS to produce unique and interpretable fragmentation patterns, together with high resolution accurate mass measurement to allow derivation of the atomic compositions of key ions of interest (**Figure 4.2**).

4.3 Results

A Tryptic Digest data

An *in silico* tryptic digest of the sequence shown in **Figure 4.3** for the Flic flagellin protein is given in **Table 4.2**. The highlighted peptides T27, T29 and T31, specifically LLDGSSTEIR, VALVNTSSIMSK and QMVSSLDVALK, are not found in the tryptic mass spectrometric map at the predicted doubly charged masses, but are created by the MS/MS fragmentation of higher mass ions, indicating that these peptides are post-translationally modified. Peptides T29 and T31 are seen to contain methionine and MS data on those peptides was variable between preparations and between runs due to variable oxidation of the methionine sulphur atom. This complicates the mass spectra, especially when looking for other unknown substitutions and therefore the work in this chapter describes the detailed analysis of peptide T27 LLDGSSTEIR only, since data showed that the other peptides carried the same modifications. Note that this peptide contains three hydroxy amino acids (two Ser and one Thr) which can be possible acceptor sites for the GlcNAc sugar described above. In the various preparations studied by ES LC-MS, this peptide was seen to be modified post-translationally to move the predicted doubly charged ion of 545.8 (**Table 4.2**) to higher masses, including signals which could be interpreted as additions of disaccharide combinations deoxyHex/MethyldeoxyHex, MethyldeoxyHex/deoxyHex and MethyldeoxyHex/MethyldeoxyHex in position 2 and 3 respectively where GlcNAc is position 1. At higher masses, signals are observed at m/z 991.0²⁺ or 998.1²⁺, and these incremental masses are much heavier and bear no relationship to the phosphomethyl threonine PTM found by Twine et al.

MSDYINEELIKKIKSRNDKDVNYTKEGKIMRVNTNVSALIANNQMGRNVNAQSKSMEKL
SSGVRIKRAADDAAGLAISEKMRAQIKGLDQAGRNVQDGISVVQTAEGALEETGNILQR
MRTL SVQSSNETNTAEERQKIADELLQLKDEVERISSSIEFNGKKLLDGSSTEIRLQVGANF
GTNVAGTTNNNNEIKVALVNTSSIMSKAGITSSTIASLNADGTSGTDAAKQMVSSLDVAL
KELNTRAKLGAQQNRLESTQNNLNNTIENVTAESRIRD TDVASEMVNLSKMNILVQA
SQSMLAQANQQPQGVLLG

Figure 4.3 *Clostridium difficile* flagellin sequence

Frag	Residues	Sequence	Theor	[M+H] ⁺	[M+2H] ²⁺	[M+3H] ³⁺
T1	1-11	(-)MSDYINEELIK(K)	1353.65	1354.66	677.83	452.22
T2	12-12	(K)K(I)	146.11	147.11	74.06	49.71
T3	13-14	(K)IK(S)	259.19	260.20	130.60	87.40
T4	15-16	(K)SR(N)	261.14	262.15	131.58	88.06
T5	17-19	(R)NDK(D)	375.18	376.18	188.60	126.07
T6	20-25	(K)DVNYTK(E)	738.35	739.36	370.19	247.13
T7	26-28	(K)EGK(I)	332.17	333.18	167.09	111.73
T8	29-31	(K)IMR(V)	418.24	419.24	210.13	140.42
T9	32-47	(R)VNTNVSALIANNQMGR(N)	1700.86	1701.87	851.44	567.96
T10	48-54	(R)NVNAQSK(S)	759.39	760.40	380.70	254.14
T11	55-58	(K)SMEK(L)	493.22	494.23	247.62	165.41
T12	59-64	(K)LSSGVR(I)	617.35	618.36	309.68	206.79
T13	65-66	(R)IK(R)	259.19	260.20	130.60	87.40
T14	67-67	(K)R(A)	174.11	175.12	88.06	59.05
T15	68-80	(R)AADDAAGLAISEK(M)	1230.61	1231.62	616.31	411.21
T16	81-82	(K)MR(A)	305.15	306.16	153.58	102.73
T17	83-86	(R)AQIK(G)	458.29	459.29	230.15	153.77
T18	87-93	(K)GLDQAGR(N)	715.36	716.37	358.69	239.46
T19	94-118	(R)NVQDGISVVQTAEGALEETGNILQR(M)	2640.34	2641.34	1321.18	881.12
T20	119-120	(R)MR(T)	305.15	306.16	153.58	102.73
T21	121-136	(R)TLVQSSNETNTAEER(Q)	1764.81	1765.82	883.41	589.28
T22	137-138	(R)QK(I)	274.16	275.17	138.09	92.40
T23	139-147	(K)IADELLQLK(D)	1041.61	1042.61	521.81	348.21
T24	148-152	(K)DEVER(I)	646.29	647.30	324.15	216.44
T25	153-162	(R)ISSIEFNGK(K)	1080.55	1081.55	541.28	361.19
T26	163-163	(K)K(L)	146.11	147.11	74.06	49.71
<u>T27</u>	<u>164-173</u>	<u>(K)LLDGSSTEIR(L)</u>	<u>1089.57</u>	<u>1090.57</u>	<u>545.79</u>	<u>364.20</u>
T28	174-195	(R)LQVGANFGTNAVAGTTNNNEIK(V)	2275.12	2276.13	1138.57	759.38
<u>T29</u>	<u>196-207</u>	<u>(K)VALVNTSSIMSK(A)</u>	<u>1248.67</u>	<u>1249.68</u>	<u>625.35</u>	<u>417.23</u>
T30	208-230	(K)AGITSSTIASLNADGTSGTDAAK(Q)	2108.02	2109.03	1055.02	703.68
<u>T31</u>	<u>231-241</u>	<u>(K)QMVSSLDVALK(E)</u>	<u>1189.64</u>	<u>1190.65</u>	<u>595.83</u>	<u>39.55</u>
T32	242-247	(K)ELNTR(A)	718.36	719.37	360.19	240.46
T33	248-249	(R)AK(L)	217.14	218.15	109.58	73.39
T34	250-256	(K)LGAQQNR(L)	785.41	786.42	393.72	262.81
T35	257-277	(R)LESTQNNLNNTIENVTAAESR(I)	2317.11	2318.12	1159.57	773.38
T36	278-279	(R)IR(D)	287.20	288.20	144.61	96.74
T37	280-292	(R)DTDVASEMVNLSK(M)	1407.66	1408.66	704.84	470.23
T38	293-319	(K)MNILVQASQSMLAQANQQPQGVLLG(-)	2879.50	2880.51	1440.76	960.84

Table 4.2 *C.difficile* flagellin sequence *in-silico* tryptic digest.

B The MS/MS of 991²⁺ and 998²⁺

Figures 4.4 and **Figure 4.5** show the LC-MS/MS spectra of m/z 991²⁺ from a hypervirulent RT023 strain and m/z 998²⁺ from a hypervirulent RT027 strain. An initial interpretation of both spectra clearly shows the first sugar linked to the peptide M+H⁺ ion at m/z 1090 to be the expected HexNAc (GlcNAc) at m/z 1293. Assuming simple glycosidic bond cleavages, in both cases the next sugars in the chain are assigned as MethyldeoxyHex to m/z 1453 and deoxyHex to m/z 1599. Beyond that mass, there are no clear signals up to the calculated M+H⁺ masses of 1980.9 and 1994.9 respectively. Turning to masses below m/z 1000, again there are common signals to both spectra at m/z 952 and 749 which are separated by a HexNAc mass difference. Looking at the possible peptide fragment ions, m/z 749 fits for the y₇^{''} ion (GSSTEIR) and therefore m/z 952 would correspond to the HexNAc substitution at one of the Ser or Thr residues. Other peptide fragment ions were seen in the spectra shown in **Figures 4.4** and **4.5** for example y₁^{''} at m/z 175 and y₂^{''} at m/z 288.

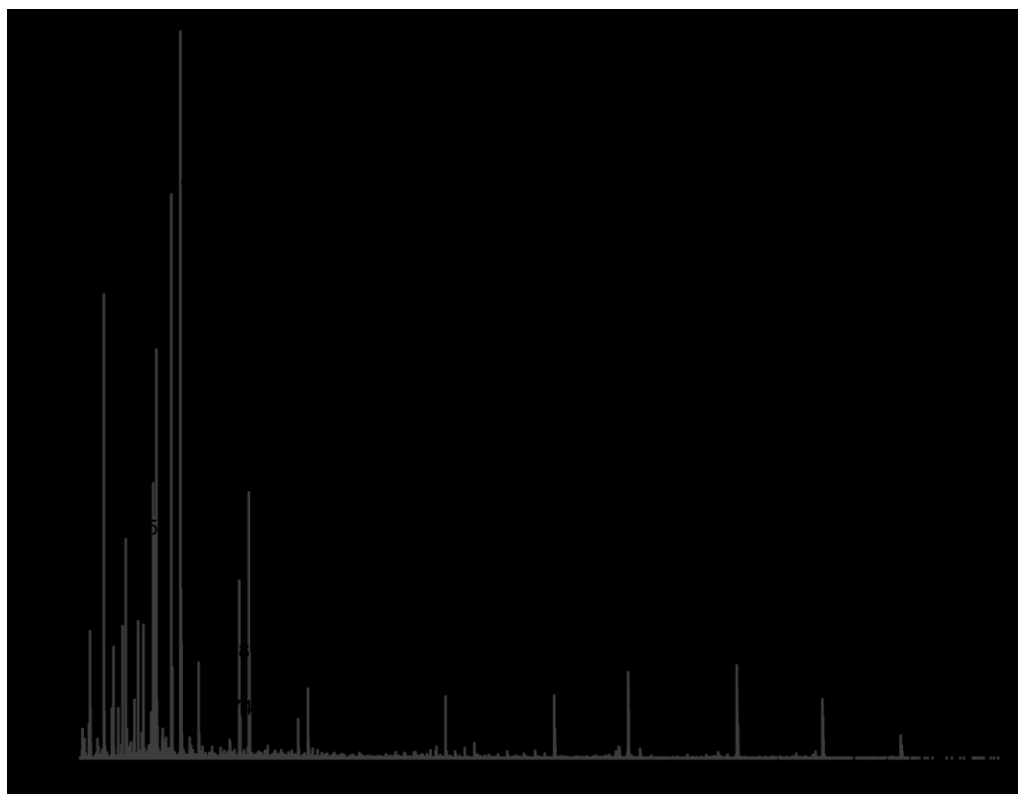


Figure 4.4 LC-MS/MS spectrum of m/z 991²⁺.

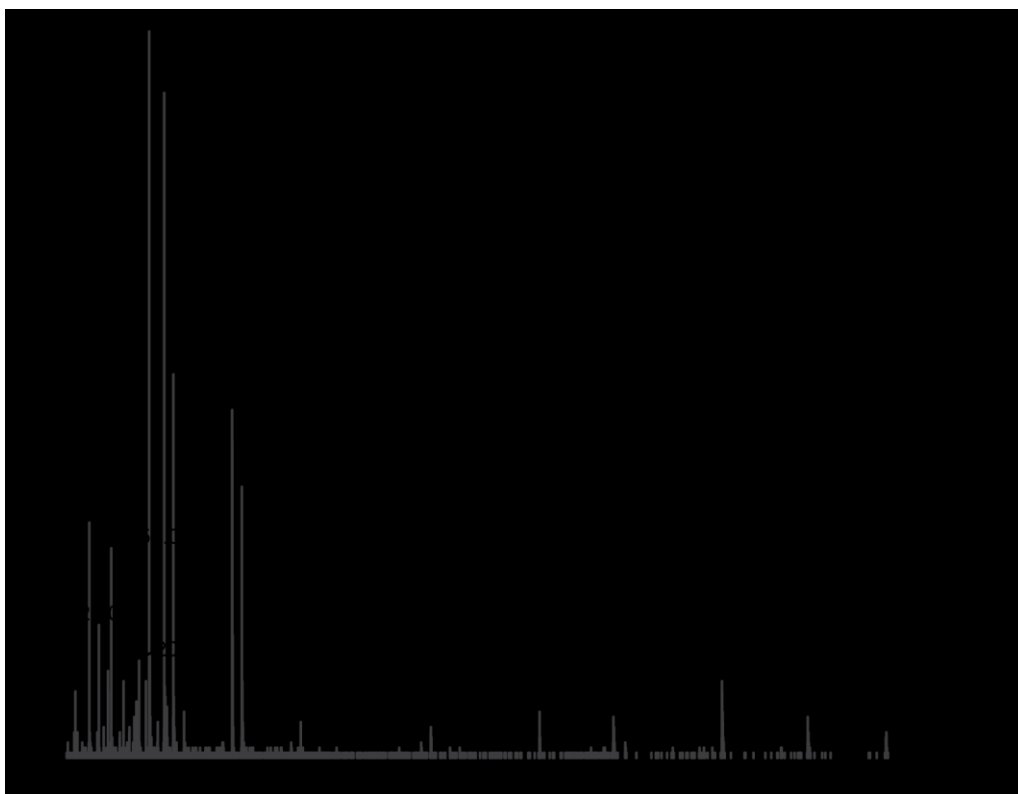


Figure 4.5 LC-MS/MS spectrum of m/z 998²⁺.

The clearly novel aspect of this glycopeptide is seen in the intense fragment ions respectively for m/z 991²⁺ at m/z 382, 364, 254, 237, 209 and 152, and for m/z 998²⁺ at m/z 396, 378, 268, 251, 223 and 152 which do not immediately correlate with any common structural entity previously observed in sugar or amino acid chemistry. Interestingly, in adding either the 396 or 382 fragment ions respectively to the m/z 1599 common fragment ion (assigned above as the peptide plus HexNAcMethyldeoxyHexdeoxyHex) gives the calculated masses of the two glycopeptides (using 991²⁺ and 998²⁺), meaning that the full structure is represented in the fragment ions observed, without any missing pieces.

C Ideas from Bioinformatics

At this point in dealing with the complete unknowns of the m/z 396 and 382 ions, it was thought that some clues could be obtained from the early bioinformatic analysis of the genes noted in **Figure 4.1**. The operon genes CD0245, CD0246 and CD0247 in bioinformatics analysis were predicted to be respectively coding for a carbamoyl phosphate synthetase (involved in the synthesis of pyrimidines and purines), a putative ornithine cyclodeaminase and a putative 3-amino-5-hydroxybenzoic acid synthase (personal communication, Esmeralda

Valiente). After studying several nucleoside analogues, interestingly, by *in-silico* coupling of 8-hydroxyadenosine to a deoxyHex it is possible to make a sensible putative fragment ion suggestion for m/z 396 (**Figure 4.6**) which could theoretically then decompose to m/z 268, 251 and ultimately 152.

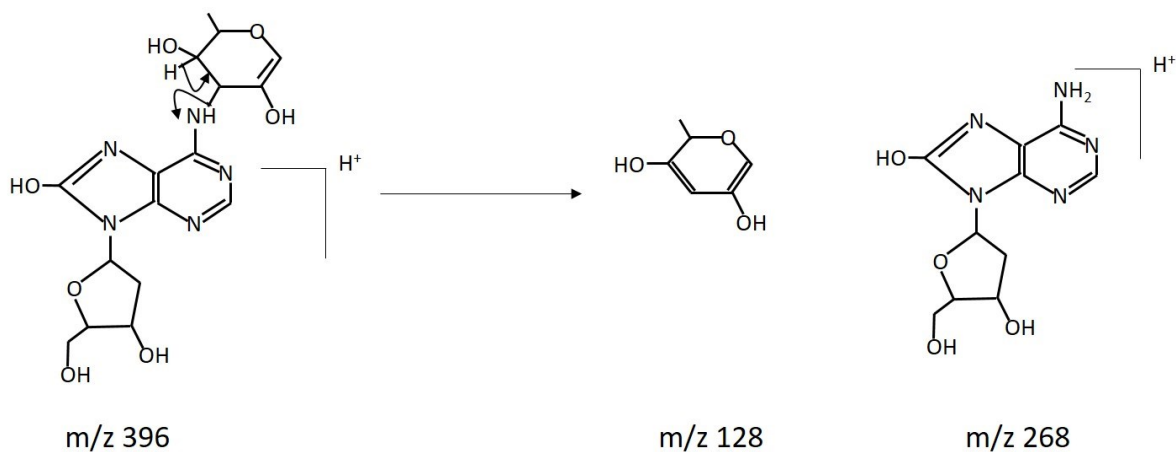
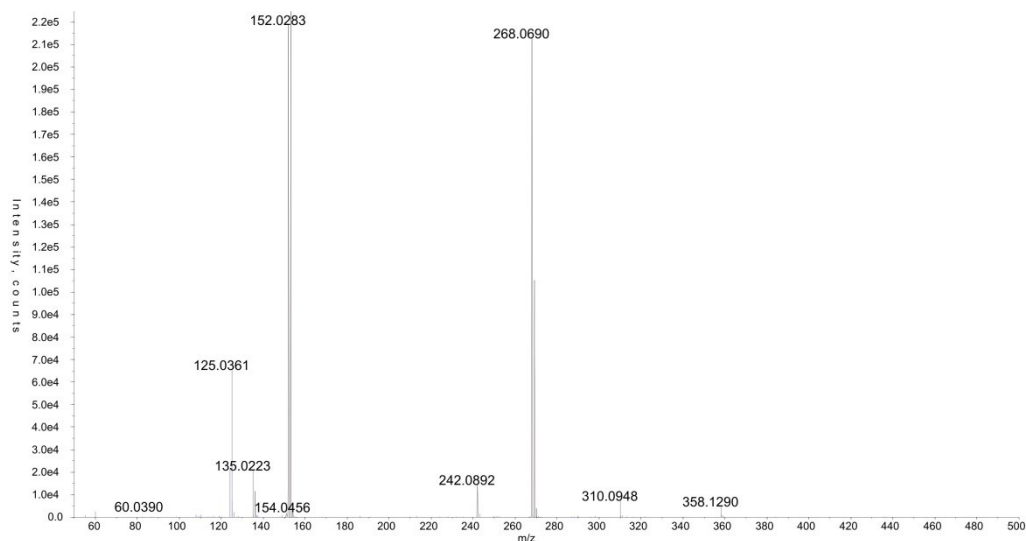
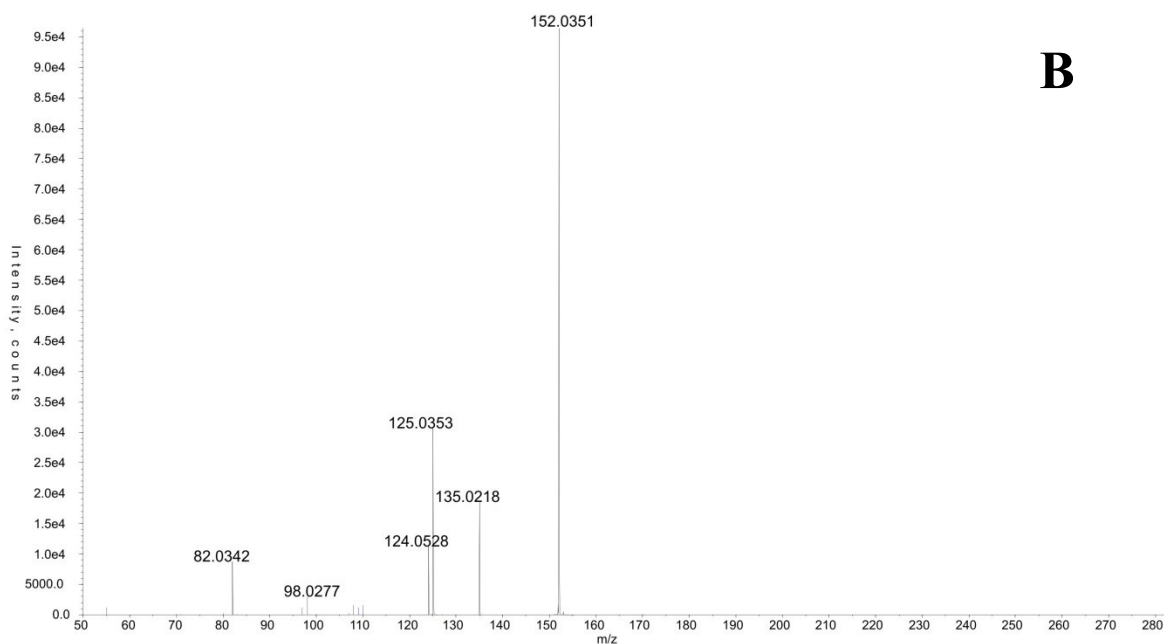


Figure 4.6 Fragment suggestion for the m/z 396 to give m/z 128 and m/z 268.

The availability of this commercial adenosine standard then allowed the study of its MS and MS/MS, and the nanospray spectrum of that material is shown in **Figure 4.7 A** (MS) and **4.7 B** (MS/MS). Although a standard for the actual predicted glycopeptide component structure drawn in **Figure 4.6**, including the deoxyHex sugar unit was not of course available, it would be expected that some common ions would exist for the sub-fragmentation of the adenosine unit. As seen in **Figure 4.7** for the adenosine analogue itself there are some apparently related mass numbers to the *C.difficile* natural glycopeptide product eg m/z 268 and m/z 152.



A



B

Figure 4.7 A-B: Nanospray of 8-hydroxyadenosine. The A spectrum is the MS spectrum of the commercial adenosine standard, whereas the B spectrum is the MS/MS of m/z 152.

However, the MS/MS of the particular ion of interest at m/z 152 shows no common fragment ions in either the 991²⁺ or 998²⁺ MS/MS spectra of the glycopeptide (**Figures 4.4 and 4.5**), i.e. the m/z 135 and 124/125 ions are not present in the glycopeptide MS/MS, showing that the similarity of mass numbers with the nucleoside standards is purely coincidental. Later work at High Resolution (see **section E**) also completely discounts this avenue of investigation.

D m/z 396 sub-fragment possibilities

Because the two glycopeptide quasi molecular ions are clearly related and 14 mass units apart (which High Resolution measurement shows to be a methyl group replacing a hydrogen atom), it was decided to concentrate on only the m/z 998²⁺ ion and its fragments in order to solve the unknown structure.

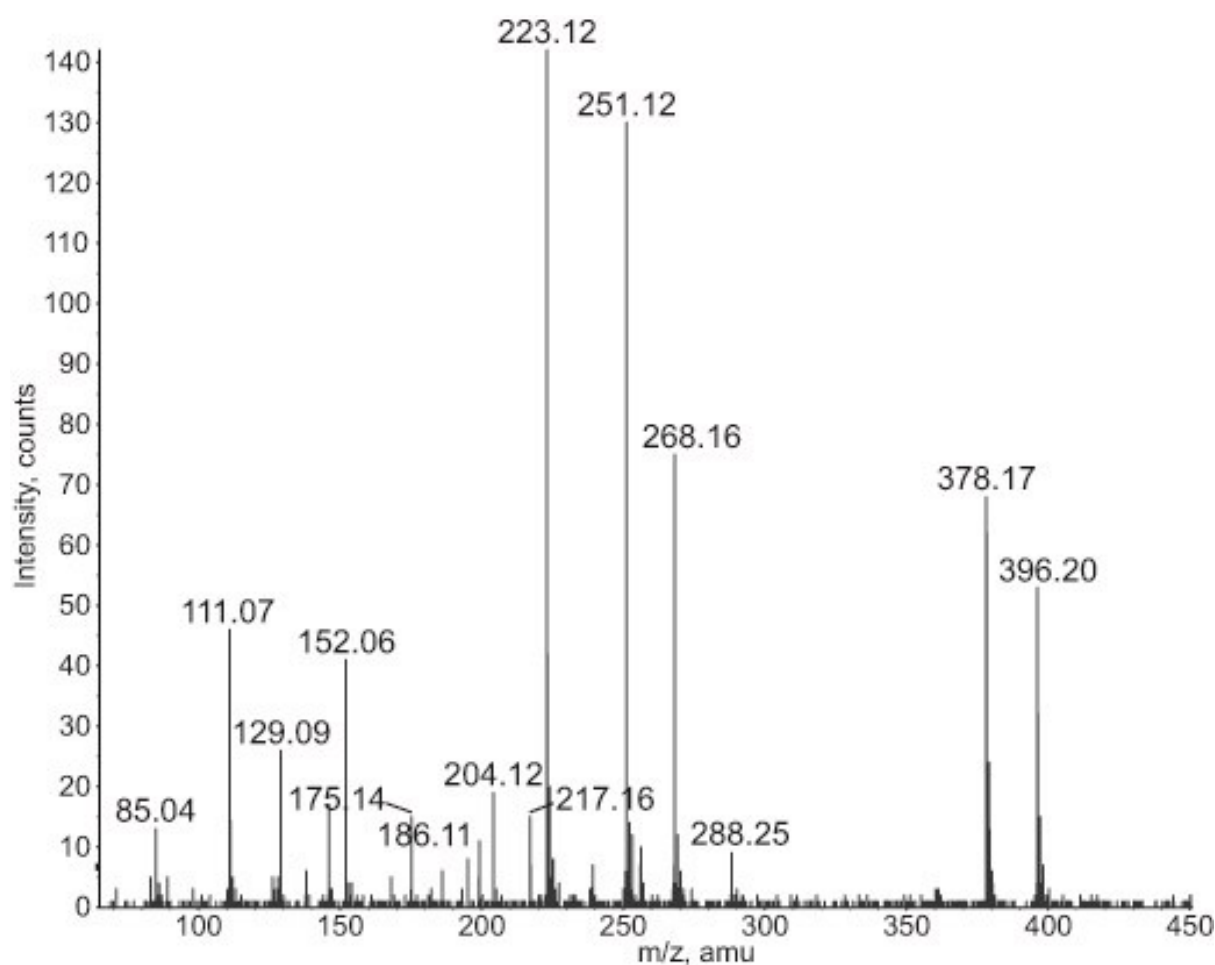


Figure 4.8 MS/MS spectrum of 998²⁺ from m/z 60 to 450.

Looking at a possible pathway of losses from m/z 396, as seen in **Figure 4.8** in the low mass range 60 to 450, the first principal loss is to m/z 268 followed by losses to m/z 251 and 223. The m/z 396 to 268 loss is 128 Da, and it seems reasonable to assume that since the sugars identified thus far up to m/z 1599 are deoxyhexoses and Methyldeoxyhexoses, then the m/z 268 ion could derive from the loss of another deoxyHex fragment unit, as suggested in the scheme shown below (**Figure 4.9**). That results from a β -elimination mechanism (red arrows), whereas a straight forward elimination by neutralising the positive charge on the

nitrogen (blue arrows) results in charge production on the sugar fragment giving m/z 129. The m/z 129 ion can then lose water to produce a postulated highly stable triply conjugated cyclic ion structure at m/z 111.

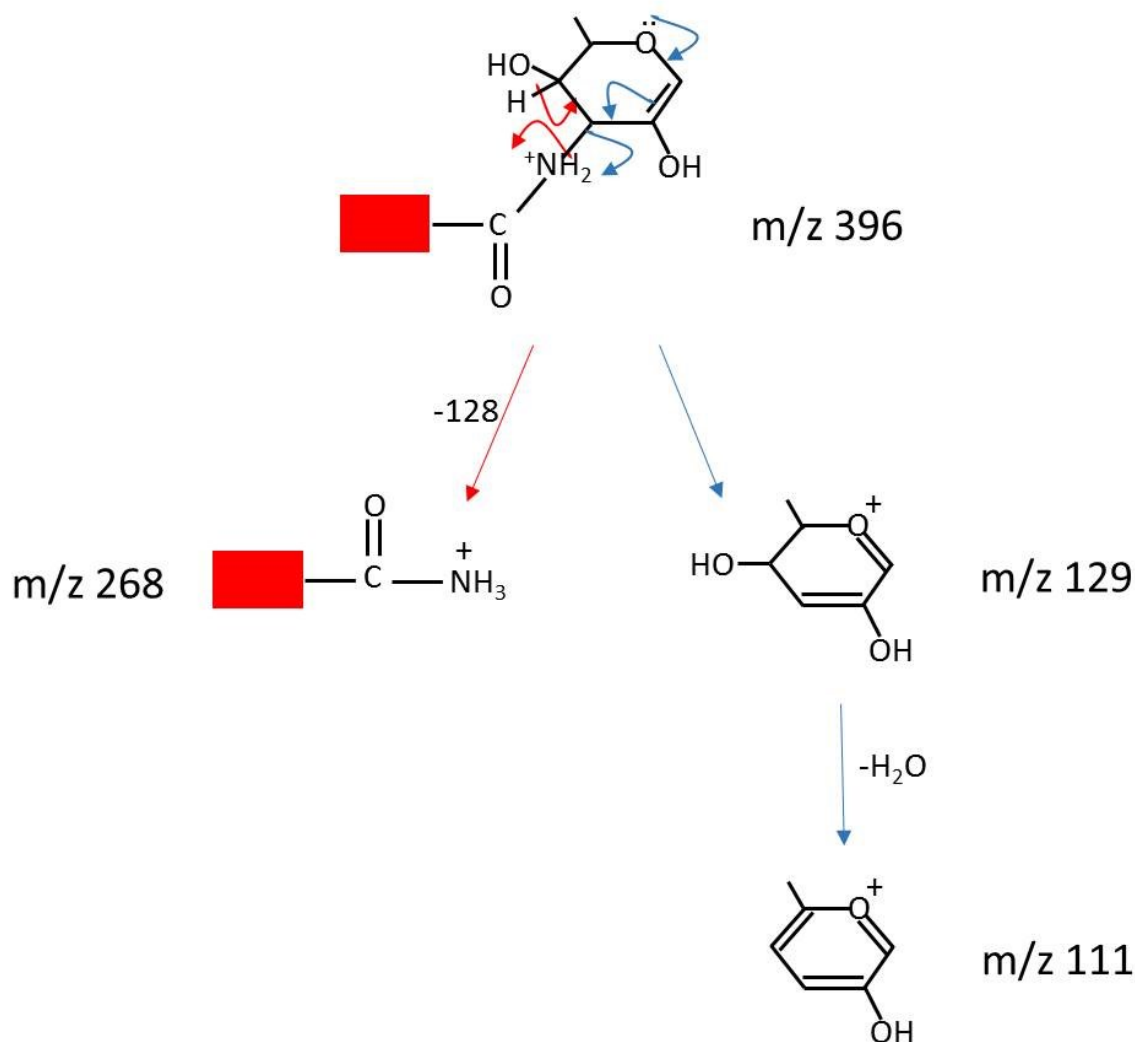


Figure 4.9 Possible β -elimination mechanism (red arrows) with consequently loss of 128 from 396 where the red box at this stage remains unknown, and an alternative elimination mechanism (blue arrows) with formation of highly stable triply conjugated cyclic ion structure at m/z 111, respectively.

Referring back to the **Figure 4.8**, the next principal losses are clearly 268 to 251 (-17 Da) and 251 to 223 (-28 Da). The 17 Da loss in ES can only be NH_3 , whereas a 28 Da could be CO or C_2H_4 . Since hydrogen and nitrogen are mass sufficient (greater than nominal mass) and oxygen is mass deficient (less than nominal mass) on a $C = 12.000000$ convention, then even at the basic Q-STAR resolution it is possible to see that the 268 to 251 loss is an NH_3

molecule, and 251 to 223 is CO not C₂H₄, as suggested in the fragmentation scheme shown below (**Figure 4.10**).

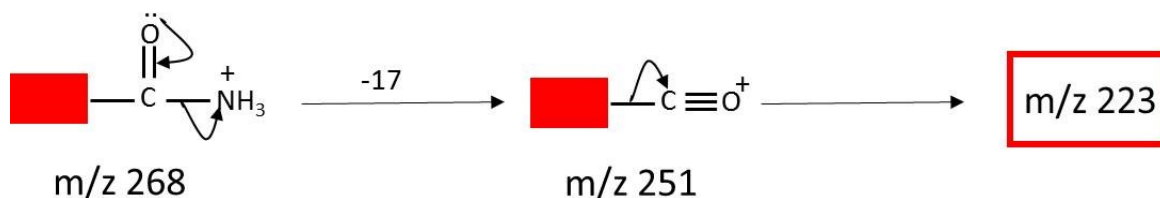


Figure 4.10 Possible fragmentation mechanism to lose NH₃ and CO from m/z 268.

E High Resolution Accurate Mass Measurement

At this point, it was clear that the increasing ambiguities related to possible structures of the lower mass ions including m/z 223 and 152 (red box above) could be removed or at least mitigated if a High Resolution spectrum of the 998²⁺ MS/MS data could be obtained. Such data, if achievable across the whole mass range in a “dynamic scan” on the nano LC-MS (with only a few seconds of acquisition time available) would also validate the interpretation of the higher mass ions in the spectrum. **Figure 4.11** shows the positive-ion online nano-LC MS/MS high resolution CID mass spectrum of m/z 998²⁺ across the entire mass range, giving the data obtained to four decimal places of mass accuracy.

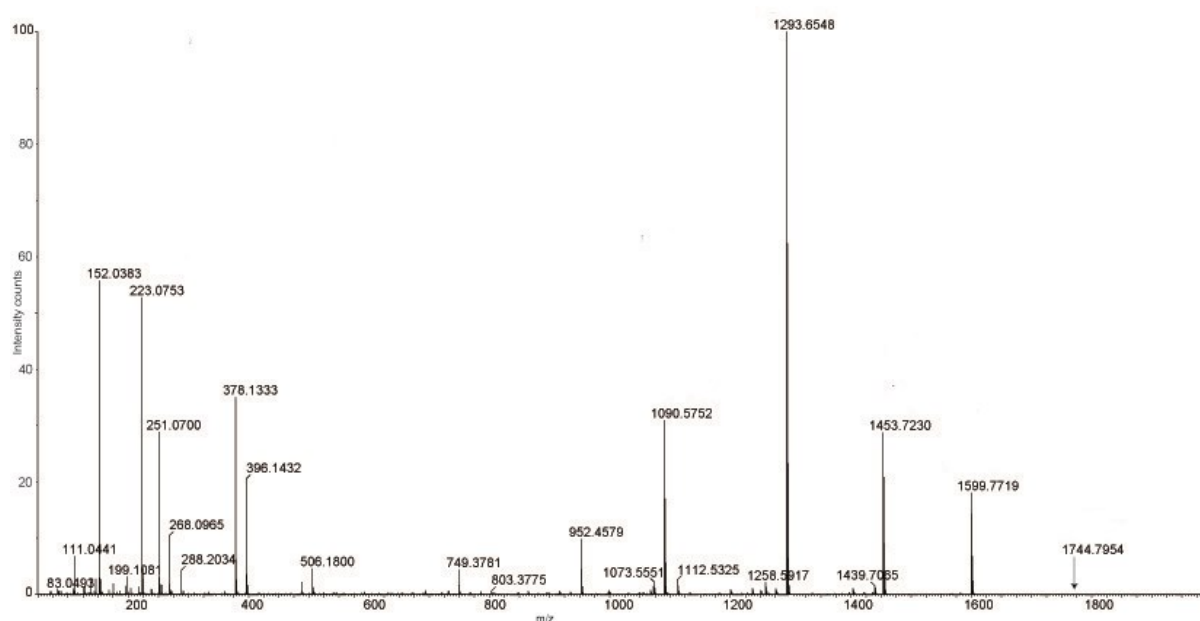


Figure 4.11 Positive ion online nano-LC MS/MS high resolution CID mass spectrum of m/z 998²⁺.

Zooming-in on the low mass region of the spectrum to find the less intense ions gives the data shown in **Figure 4.12**, **Figure 4.13** and **Figure 4.14**.

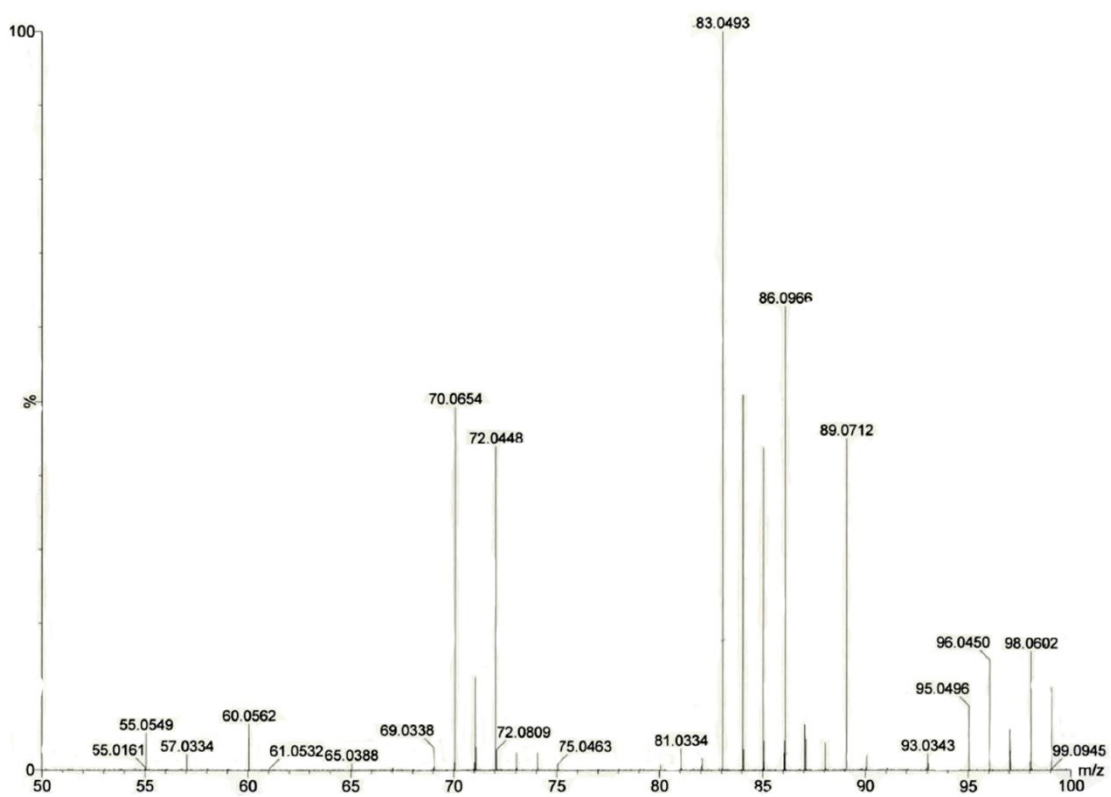


Figure 4.12 Positive ion online nano-LC MS/MS high resolution CID mass spectrum of m/z 998²⁺ from 50-100 m/z .

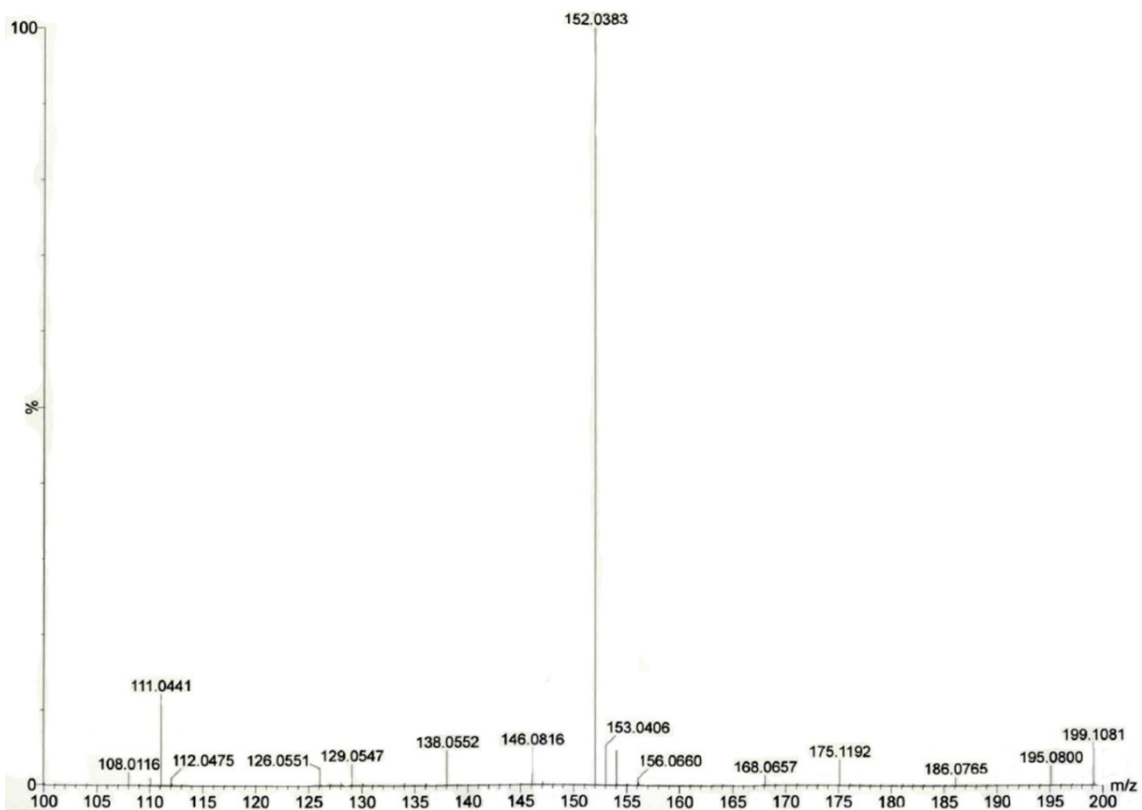


Figure 4.13 Positive ion online nano-LC MS/MS high resolution CID mass spectrum of m/z 998²⁺ from 100-200 m/z .

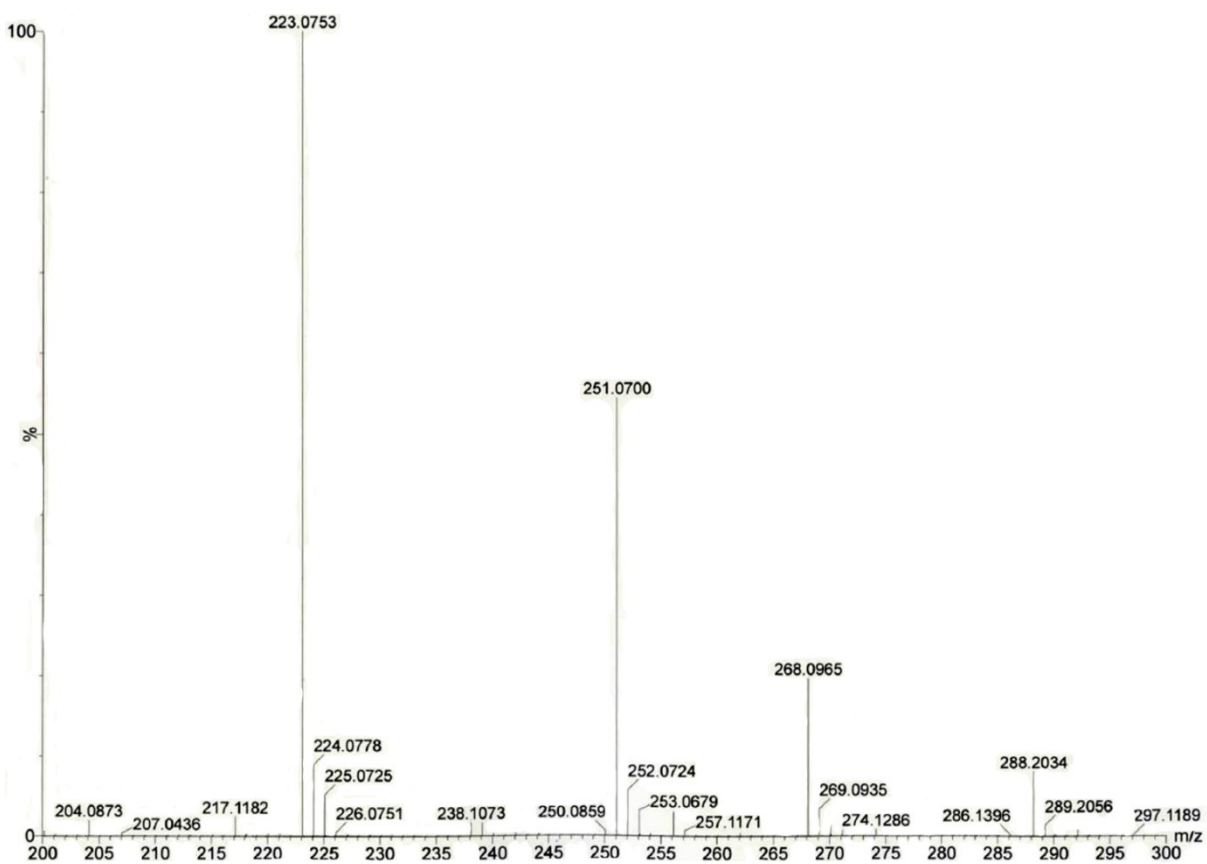


Figure 4.14 Positive ion online nano-LC MS/MS high resolution CID mass spectrum of m/z 998²⁺ from 200-300 m/z .

Since the instrument was tuned to a resolution $> 40,000$ for the run shown in **Figure 4.11**, then the anticipated mass accuracy obtainable is of the order of 10 parts per million (10 ppm) or less. Within the overall data set, it is possible to find fragment ions expected from the fragmentation of *known* portions of the glycopeptide, including of course the peptide sequence LLDGSSTEIR and the GlcNAc, residue 1, of the saccharide conjugate. These were therefore used as internal standards to check the mass accuracy of the data and hence to give a degree of confidence in the mass measurements observed in **Figures 4.11-4.14** for the *unknown* parts of structure. For example, the y_1 ” and y_2 ” ions of the protein-derived peptide at m/z 175 and 288 would be expected as fragments, as would the m/z 204 reporter ion for the GlcNAc sugar residue, and these are indeed present in **Figures 4.13-4.14** at m/z 175.1192, 204.0873 and 288.2034.

Computing the possible atomic compositions for these accurate masses, using C, H, N, O and S (being guided by the isotopic distributions observed in the spectrum) produces the possibilities shown in **Table 4.3**. In analysing these possibilities, using the m/z 288 data as an example, the high ppm errors were first deleted together with the sulphur-containing possibilities (since there is no +2 isotope for ^{34}S in the spectrum at m/z 290), leaving two possible atomic compositions of $\text{C}_{12}\text{H}_{26}\text{N}_5\text{O}_3$ and $\text{C}_8\text{H}_{22}\text{N}_{11}\text{O}$. The latter composition containing 11 nitrogens and 1 oxygen for only 8 carbons would be extremely difficult if not impossible to draw a chemical structure for, and in fact the $\text{C}_{12}\text{H}_{26}\text{N}_5\text{O}_3$ composition found does indeed correspond to the expected y_2 ” peptide ion, demonstrating an experimental mass accuracy of better than 1 ppm at that position in the mass scale. Similarly for m/z 204 and 175, the mass errors found for those atomic compositions are 0.5 ppm and 1.7 ppm respectively.

Mass	Calc. Mass	PPM	Formula
175.1192	175.1168	13.7	$\text{C}_2\text{H}_{11}\text{N}_{10}$
	175.1195	-1.7	$\text{C}_6\text{H}_{15}\text{N}_4\text{O}_2$
	175.1235	24.6	$\text{C}_{11}\text{H}_{15}\text{N}_2$
	175.1137	20.0	$\text{C}_9\text{H}_{19}\text{OS}$
204.0873	204.0847	12.7	$\text{C}_{12}\text{H}_{14}\text{NS}$
	204.0881	-3.9	$\text{C}_9\text{H}_{18}\text{NS}_2$
	204.0885	-5.9	$\text{C}_9\text{H}_{10}\text{N}_5\text{O}$
	204.0872	0.5	$\text{C}_8\text{H}_{14}\text{NO}_5$

	204.0919	22.5	C ₆ H ₁₄ N ₅ OS
	204.0845	13,7	C ₄ H ₁₀ N ₇ O ₃
	204.0832	-1.5	C ₃ H ₁₄ N ₃ O ₇
	204.0879	-2.9	CH ₁₄ N ₇ O ₃ S
288.2034	288.1997	12.8	C ₁₅ H ₃₀ NO ₂ S
	288.2076	-14.6	C ₁₇ H ₂₆ N ₃ O
	288.2036	-0.7	C₁₂H₂₆N₅O₃
	288.2043	-3.1	C ₅ H ₂₆ N ₁₁ OS
	288.2009	8.7	C ₈ H ₂₂ N ₁₁ O
	288.1995	13.5	C ₇ H ₂₆ N ₇ O ₅
	288.2081	-16.3	C ₂ H ₂₂ N ₁₅ O ₂

Table 4.3 Possible atomic compositions for the accurate masses observed: 175.1192, 204.0873 and 288.2034. The compositions highlighted in yellow are the one being accepted.

Having established that the mass accuracy of the High Resolution run is within the 10 ppm error expected, it was then possible to accept the reasonable atomic composition possibilities for the unknown ion species including m/z 111, 152, 223, 251, 268, 378 and 396 with a high degree of confidence, and these data are presented in **Table 4.4**.

Observed Mass (m/z)	Atomic Composition Assigned	Theoretical Mass (m/z)
111.0441	C ₆ H ₇ O ₂	111.0446
152.0383	C ₄ H ₁₀ NO ₃ S	152.0381
223.0753	C ₇ H ₁₅ N ₂ O ₄ S	223.0752
251.0700	C ₈ H ₁₅ N ₂ O ₅ S	251.0701
268.0965	C ₈ H ₁₈ N ₃ O ₅ S	268.0967
378.1333	C ₁₄ H ₂₄ N ₃ O ₇ S	378.1335
396.1432	C ₁₄ H ₂₆ N ₃ O ₈ S	396.1441

Table 4.4 Atomic compositions assigned for key signals in the High Resolution MS/MS data. This table shows the deduced atomic compositions for the experimental and measured masses observed for key signals in the High Resolution MS/MS data (**Figure 4.11**) in association with the theoretical masses of those compositions. Crucial discoveries from these data comprised the finding of sulphur in the m/z 152 and higher mass fragments, thus confirming a novel structural entity and allowing the interpretation of a clear fragmentation pathway between the m/z 396 and 152 ions.

The High Resolution mass experiment was critical to the structure elucidation of this glycopeptide unknown PTM in terms not only of the ions listed in **Table 4.4**, but also in confirming the other interpretation already postulated for ions seen in the MS/MS spectra in **Figures 4.4** and **4.5** including the ions at m/z 1090.5752, 1293.6548, 1453.7230, 1599.7719 and 1744.7954 in **Figure 4.11** where the latter fragment ion corresponds to the extension of the postulated Peptide-HexNAc-MethyldeoxyHex-deoxyHex chain by a further Amino-deoxyHex sugar unit (145 Da). The schematic of the detail at this present stage of the high mass interpretation is presented in **Figure 4.15**.

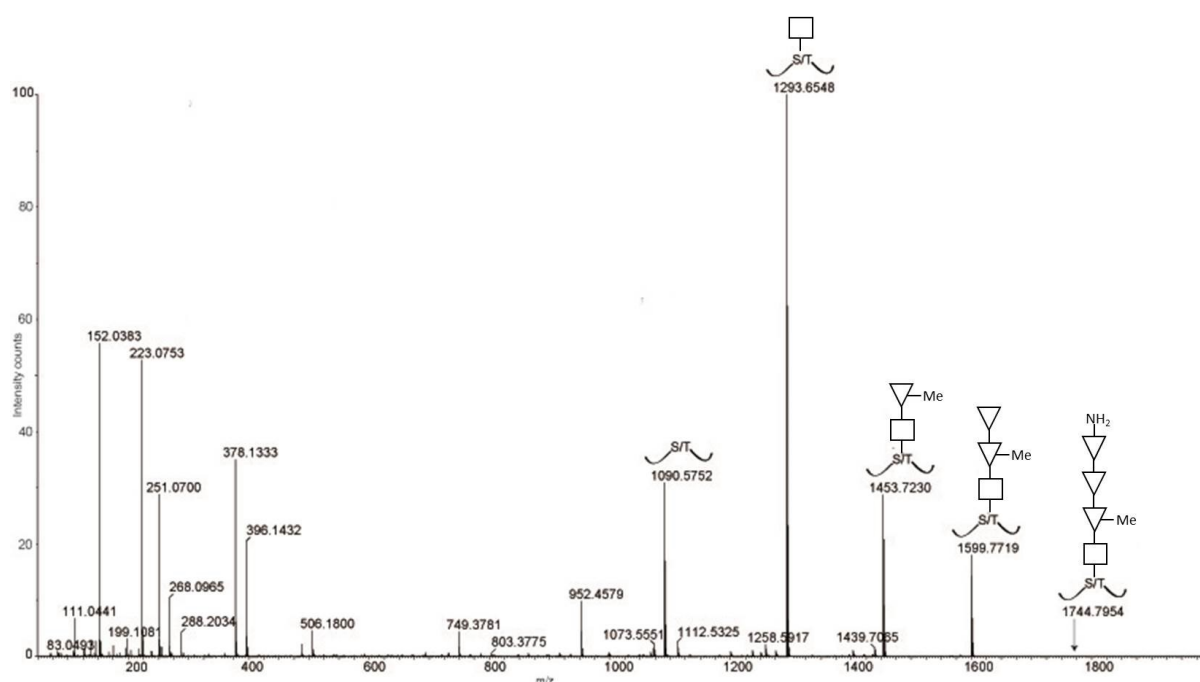


Figure 4.15 Positive ion online- nano-LC MS/MS high resolution CID mass spectrum of m/z 998²⁺. \square : peptide; \square : HexNAc; \square -Me: MethyldeoxyHex; \square : deoxyHex; \square -NH₂: Amino-dideoxyHex.

Reviewing **Figure 4.9** discussed earlier, the Amino-deoxyHex structural unit referred to above is the one postulated (itself as a fragment ion) at the top of this figure. That figure also postulates the production of an ion at m/z 111 and the High Resolution atomic composition determined in **Table 4.4** as C₆H₇O₂ now absolutely confirms this idea. m/z 111 could most easily arise to give the triple conjugation by assuming a 3-Amino substitution.

The next significant ion in the low mass region is m/z 152 and the observed accurate mass difference of 71.0370 Da to m/z 223 in **Figure 4.11**, as a result of 223.0753 minus 152.0383, might correspond to either alanine or its isomer N-methyl glycine in atomic composition,

since both compositions (C_3H_5NO) present a theoretical mass of 71.0371. Incorporating this idea into the schematic extends the postulated structure as shown **Figure 4.16**.

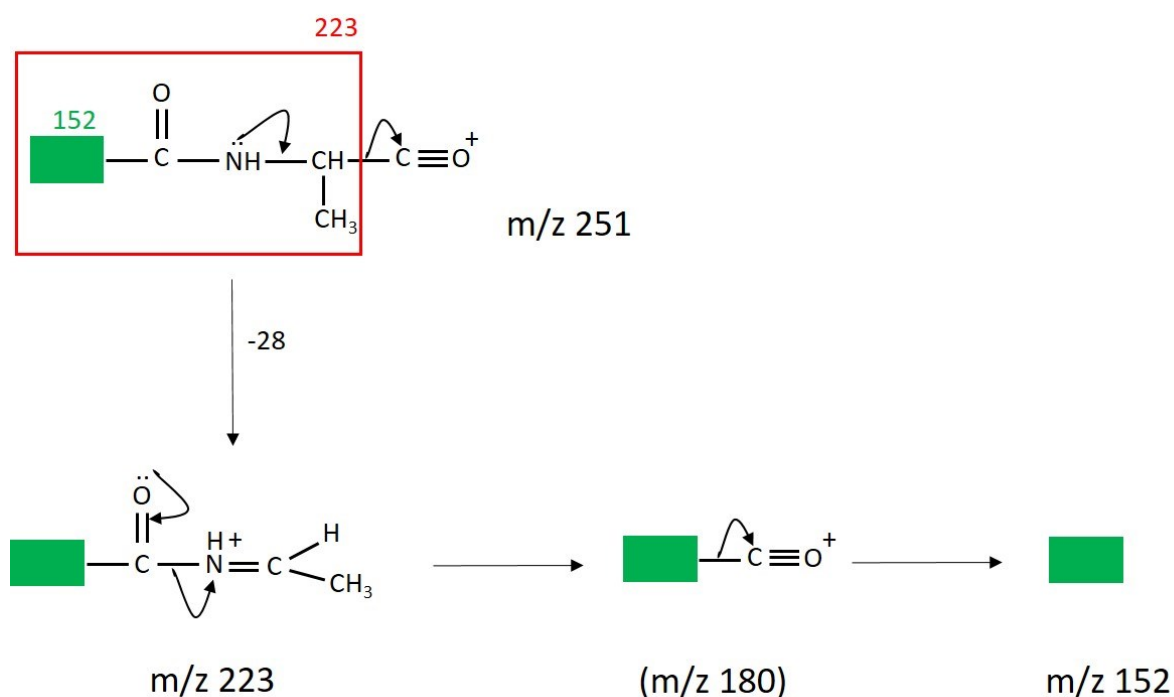
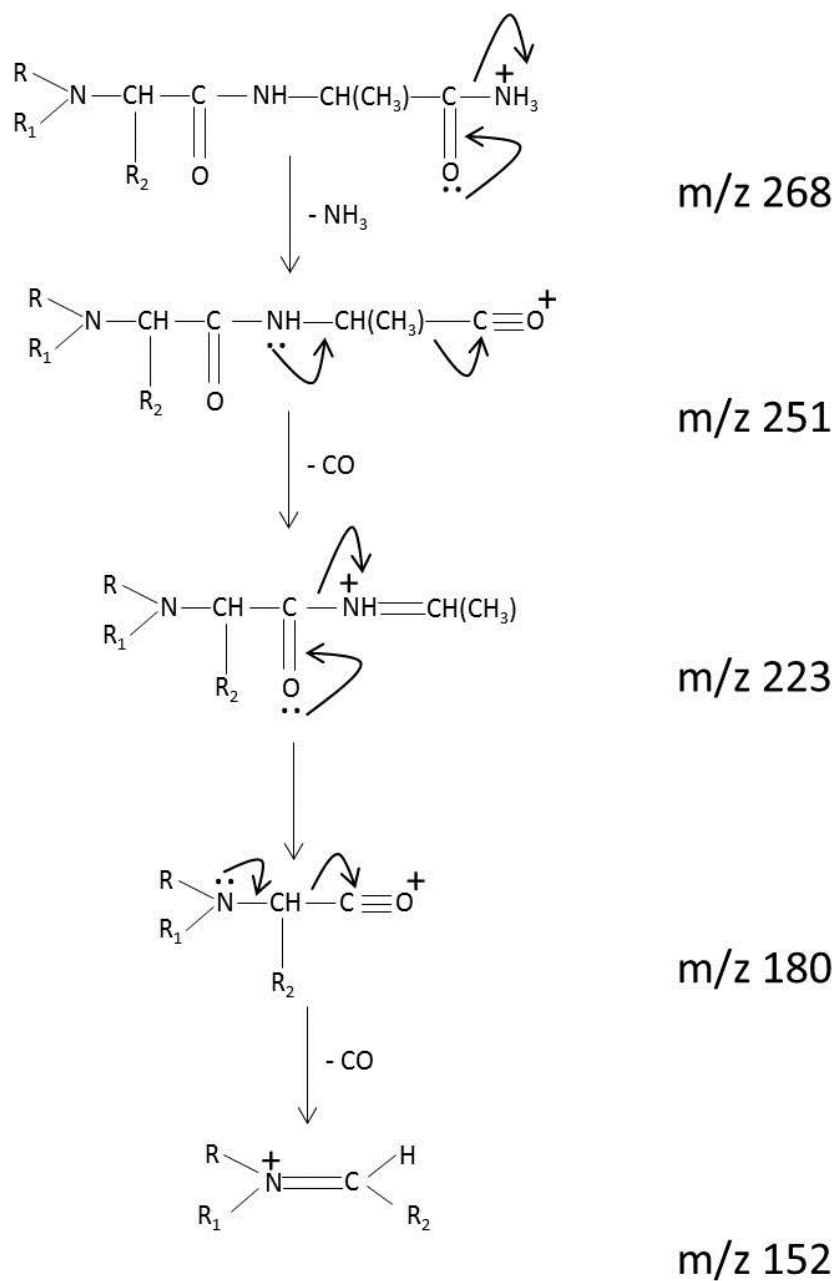


Figure 4.16 Possible fragmentation mechanism to form $m/z\ 152$.

In peptide mass spectrometry the carbon monoxide (CO) loss to give $m/z\ 152$ would be via an aldimine mechanism which would require a β -nitrogen within the 152 unit, leading to a postulated schematic from $m/z\ 268$ to 152 shown in **Scheme 1**.



Scheme 1 Summary of the aldimine mechanistic logic used to produce the m/z 152 ion.

It should be noted that the above **Scheme 1** could also apply equally well to the m/z 254, 237 and 209 ions seen in the MS/MS of 991²⁺ simply by replacing the postulated alanine shown (or methyl glycine) by glycine residue itself, NH-CH₂-CO, in the above structures.

F Additional Sub-fragmentation and Negative Ion Strategies

In the High Resolution data, the m/z 152 was found to possess an unusually mass-deficient accurate mass which suggests a sulphur-containing atomic composition of $C_4H_{10}NO_3S$ for this terminal fragment, which is not the formula of any previously reported protein- or carbohydrate-derived structural unit. In order to define this fragment in more detail, and to provide supporting evidence for the ideas in **Scheme 1** two further sets of experiments with new strategies were attempted on the remaining small quantities of material: (1) MS/MS analysis of several of the key fragment ions above to confirm mechanistically understandable breakdown products and (2) experiments in the negative ion MS and MS/MS modes to look for new and complementary fragment ion information.

- 1) CID MS/MS of the cone voltage induced m/z 152 from the 998^{2+} glycopeptide sample gives rise to two principal ion species seen in **Figure 4.17** at m/z 108 and m/z 70.

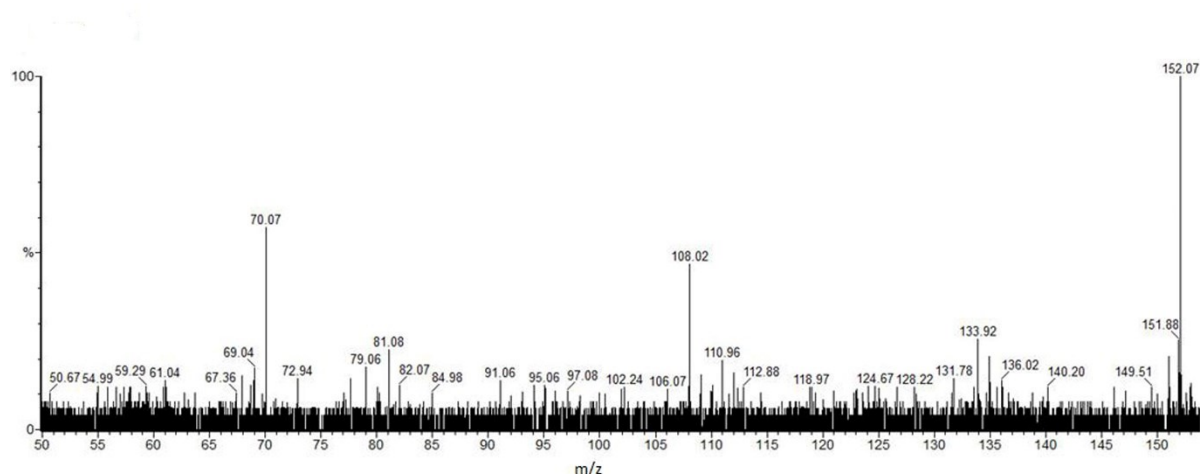


Figure 4.17 Positive ion nanospray CID MS/MS spectrum of m/z 152 produced via cone-voltage induced in-source fragmentation of 998^{2+} . Signals at m/z 70 and 108 correspond to losses of 82 and 44 mass units respectively.

These fragments are actually seen, but to a lesser extent of course, in the MS/MS High Resolution data in **Figures 4.12-4.13** and the atomic compositions are calculated to be $C_2H_6NO_2S$ for m/z 108 and C_4H_8N for m/z 70, and this suggests two overlapping component fragments, a sulphonic acid and an alkylamine, competitively derived from the m/z 152 ion ($C_4H_{10}NO_3S$).

2) The negative ion CID MS/MS spectrum obtained for m/z 996²⁻ is shown in **Figure 4.18**.

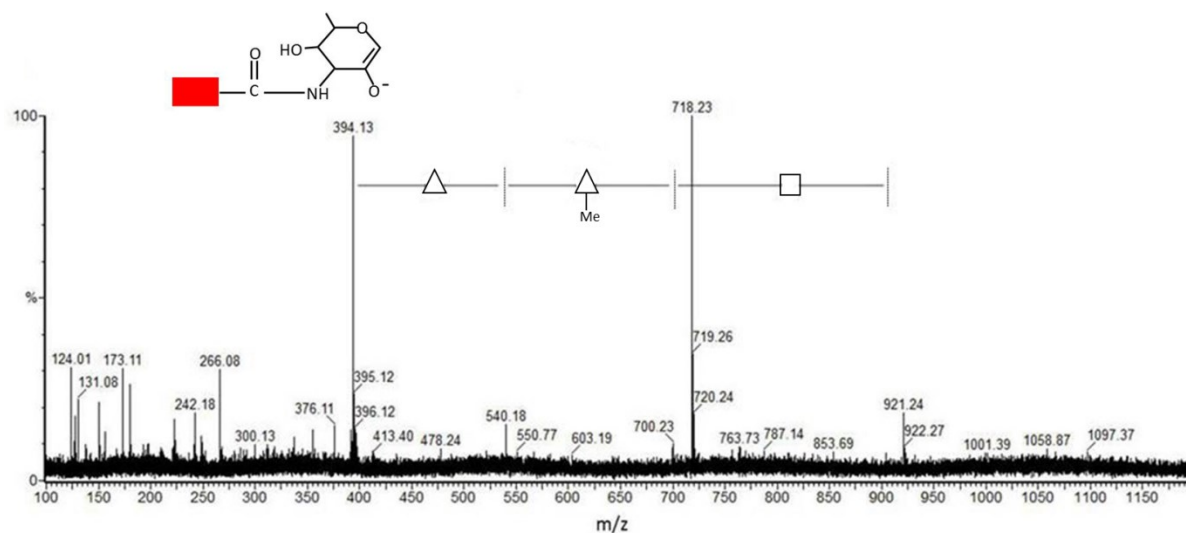


Figure 4.18 Negative ion nanospray CID MS/MS spectrum of m/z 996²⁻ (m/z 100-1200).

To begin with, these data confirm the basic structural features of the novel glycosylation inferred from the positive ion data in **Figure 4.15** and shown in **Scheme 1** regarding the expected principal fragments at m/z 921, 718, 540, 394, 266 and 150 (**Figure 4.19**). This figure, for comparative purposes, also shows the fragmentation leading to the main signals observed in the positive ion spectrum.

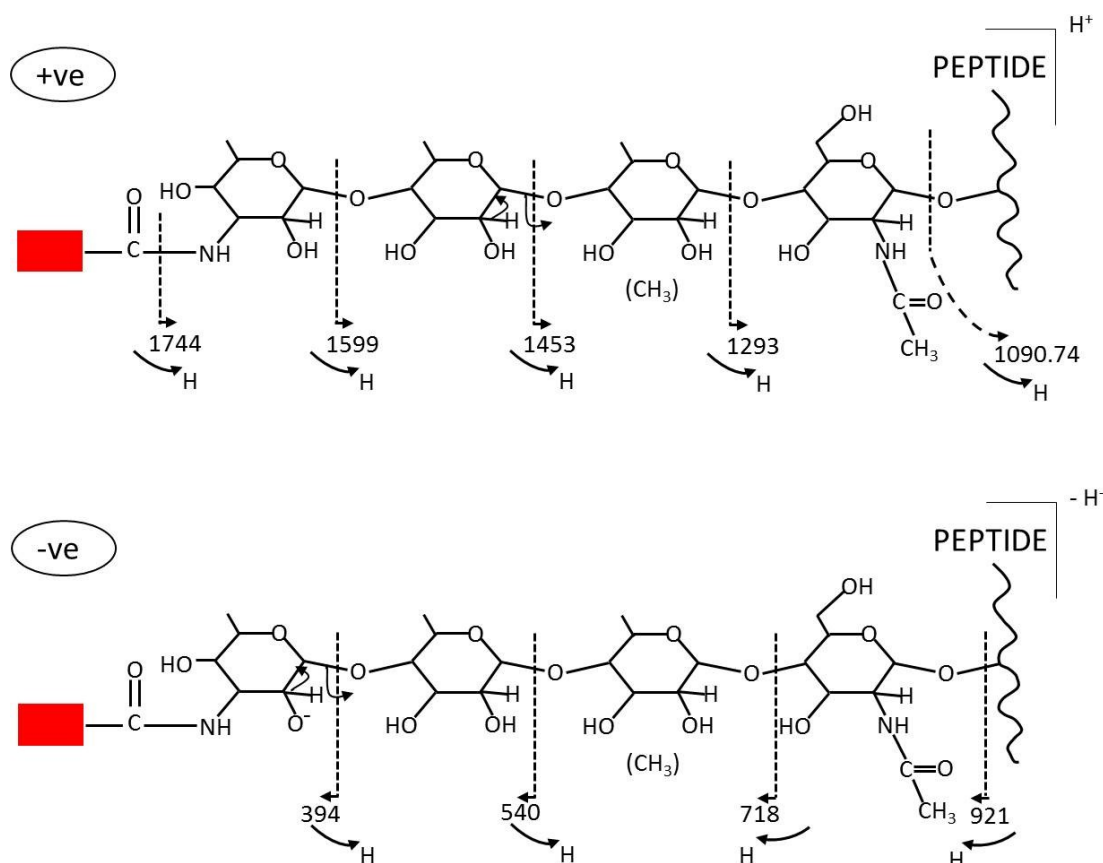


Figure 4.19 Suggested fragmentation mechanism to produce key signals in the positive and negative ion spectra of respectively the 998^{2+} and 996^{2-} glycopeptide quasi molecular ions.

Some other signals are derived from the carrier peptide LLDGSSTEIR, for example m/z 173 being attributable to the negative y_1 ion. Importantly, a new fragment ion with the equivalent not being observed in the positive ion MS/MS, is seen at m/z 124, and this is recognisable as corresponding to the mass of an Aminoethyl-Sulphonic acid, called Taurine ($C_2H_6NO_3S$, $[M-H]$).

The MS/MS of this 996^{2-} - derived signal (m/z 124) shown in **Figure 4.20** gives rise to an identical MS/MS spectrum to that observed for a synthetic sample of Taurine itself (not shown), being mainly a SO_3^- fragment ion base peak, thus providing strong evidence for this structural unit within the m/z 152 ion in the novel structure.

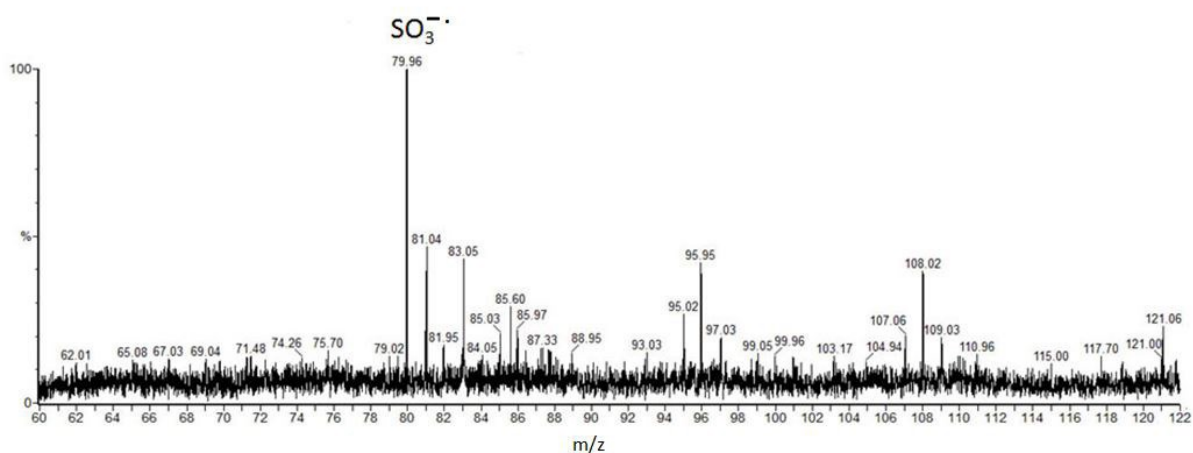


Figure 4.20 Negative ion nanospray CID MS/MS spectrum of m/z 124 produced via cone voltage induced in-source fragmentation of m/z 996²⁻.

Further experiments including subfragment MS/MS were carried out to confirm the structural conclusions from the mass spectrometric experiments. Nevertheless, as seen in the structural summary in **Scheme 1**, an ambiguity remains in the proposed novel m/z 268 unit whereby the amide-linked amino acid to the AminodeoxyHex could be either alanine or N-methyl glycine in the 998²⁺ structure, whilst it can only be glycine in the 991²⁺ structure. Moreover, there are several ways to arrange the R, R₁ and R₂ atoms in the unit incorporating the Aminoethyl Sulphonic acid-containing 152 (180) structure, comprising variations containing Taurine itself or cysteic acid with alkylations to produce the necessary accurate masses observed, and so satisfy the atomic compositions obtained from High Resolution mass measurement.

At this point, our NMR collaborators (Susan Logan, Evgenii Vinogradov and co-workers) based at the National Research Council, Ottawa, Canada were asked to attempt to resolve the above structural ambiguity, to confirm the findings from the MS glycoproteomic research described above, and to assign the configuration and linkages of the sugar residues found.

Following the isolation of sufficient material, our collaborators were able to digest the glycoprotein with proteinase K and isolate the 991²⁺ variant of the structure discovered in the MS experiments. Their subsequent NMR study then revealed the identity of the R, R₁ and R₂ groups in **Scheme 1** and the stereochemistry and linkages. Specifically, the NMR data permitted identification of the R₂ group as a methyl group, thus allowing an alanine assignment, with R being a hydrogen atom and R₁ being the ethyl sulphonic acid group giving rise to Taurine discovered in the MS/MS fragmentation. In addition, the glycan

portion of the structure, assigned mass spectrometrically as an Amino-deoxyHex-deoxyHex-MethyldeoxyHex-HexNAc unit in **Figure 4.15** has been then further assigned from the NMR data as α -Fuc3N-(1 \rightarrow 3)- α -Rha-(1 \rightarrow 2)- α -Rha3OMe-(1 \rightarrow 3)- β -GlcNAc-(1 \rightarrow)Ser.

The overall structure of the novel peptidyl-glycan found by combining the mass spectrometric and NMR data is shown in **Figure 4.21**, and this discovery has now been published (Bouché, Panico et al. 2016).

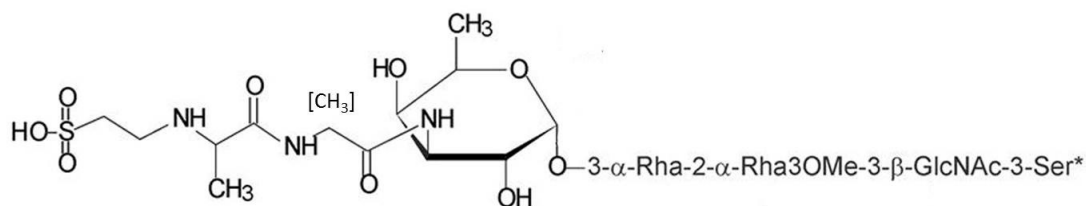


Figure 4.21 Chemical structure of the novel Sulphonated Peptidylamido-Glycan found in hypervirulent strains of *Clostridium difficile*. The structure drawn produces the tryptic glycopeptide m/z 991²⁺ variant and the [CH₃] above the glycine indicates the other variant observed at m/z 998²⁺ in the work described here.

4.4 Conclusion

Glycosylation is a key modification of both proteins and lipids, which often plays an important role in intermolecular and intercellular interactions. Bacterial protein glycosylation systems have come under enhanced study because of the increasing association with pathogenic species, in particular regarding their survival and pathogenesis. They have been described variously as being important in adhesion, motility, DNA uptake, biofilm formation, auto-aggregation, invasion, serum resistance, immune evasion and animal colonization (Szymanski, Burr et al. 2002, Larsen, Szymanski et al. 2004, Szymanski and Wren 2005, Guerry, Ewing et al. 2006, Asakura, Churin et al. 2010, Iwashkiw, Seper et al. 2012, Alemka, Nothaft et al. 2013). Recently, the structure and biological role of flagellar glycosylation in the enteric opportunistic pathogen *C.difficile* 630 has been described, by Twine et al. in 2009, and by Faulds-Pain et al. in 2014, the latter during the work described in this thesis. Investigations into the biological role of flagellar glycosylation in the emerging hypervirulent *C.difficile* strains, RT027 and RT023, in parallel work to the bio-analytical study described here, have now been reported (Valiente, Bouché et al. 2016). In these *C.difficile* studies, flagella post-translational modification is shown to be important in motility, aggregation and

adhesion to abiotic surfaces. In the case of *C.difficile* RT027 flagella glycosylation is also involved in Caco-2 cell adhesion.

The research presented here describes the discovery of a unique flagellar non-reducing-end Sulphonated Peptidylamido-glycan structure found on the glycoproteins from RT027 and also seen in other ribotypes, RT023, 001 and 106 (Bouché, Panico et al. 2016).

The mass spectrometric discovery of unique PTMs in the LC-MS and MS/MS data coming from tryptic digests of the FliC proteins was followed by the application of a full battery of advanced MS techniques (Hunt and Morris 1973, Morris, Paxton et al. 1996, Morris, Paxton et al. 1997, Billker, Lindo et al. 1998) to characterise the novel components to the extent possible on the small quantities of protein available. These led to the production of on-line High Resolution MS/MS data using a 40,000 resolving power Q-TOF geometry instrument. These experiments allowed the assignment of probable atomic compositions of all fragment ions, comprising those evidently not derived from normal glycopeptides or previously reported PTMs. For the first time the presence of sulphur-containing moieties were observed and, subsequent MS² and MS³ data obtained by the high sensitivity and mass accuracy of the Xevo Q-TOF geometry instrument (Morris, Paxton et al. 1996, Morris, Paxton et al. 1997) in both positive and negative ion mode have permitted the discovery of a Taurine (aminoethylsulphonic acid) unit in the breakdown fragments of several of the precursor ion species investigated. All these data sets allowed the finding of a novel structural unit presenting the terminal sulphonic acid group. This mass spectrometric work on hypervirulent strains has concentrated on what was found to be the most abundant LLDGSSTEIR tryptic glycopeptide, but preliminary mass spectrometric analyses show that similar modifications are present on at least other two flagellin tryptic peptides namely QMVSSLDVALK and VALVNTSSIMSK, with only minor amounts of the respective free non-glycosylated peptides found in the digest. In contrast to this very complex glycosylation of the flagella of *C.difficile* RT027, the wild type 630 strain is modified with single GlcNAc residue that is substituted with an N-methylated threonine linked via a phosphodiester bond (Twine, Reid et al. 2009, Faulds-Pain, Twine et al. 2014)(Twine, Reid et al. 2009, Faulds-Pain, Twine et al. 2014)(Twine, Reid et al. 2009, Faulds-Pain, Twine et al. 2014)(Twine, Reid et al. 2009, Faulds-Pain, Twine et al. 2014)(Twine, Reid et al. 2009, Faulds-Pain, Twine et al. 2014)(Twine, Reid et al. 2009, Faulds-Pain, Twine et al. 2014)(195,196)(Twine, Reid et al. 2009, Faulds-Pain, Twine et al. 2014). Despite the substantial differences in glycosylation, a common feature is the presence of a negatively charged functionality in the periphery of the PTM, namely a sulphonate in the hypervirulent strains and a phosphoester in the 630 strain. These charged groups are likely to

be involved in ionic interactions between the flagella and extracellular structures. This could explain the phenotype of *C.difficile* flagellar glycosylation knockouts where autoaggregation, biofilm formation and adhesion to Caco-2 cells are reduced (Faulds-Pain, Twine et al. 2014, Valiente, Bouché et al. 2016).

This study reveals a unique flagellar glycosylation structure in the bacterial pathogen *Clostridium difficile* RT027 strains which could provide this pathogenic organism with a novel strategy to escape the immune system and therefore become more virulent.

Chapter 5:
***Clostridium difficile* S-layer Research**

5. Clostridium difficile S-layer Research

5.1 Introduction

This chapter describes the research results on the work of profiling the S-layer glycoprotein of the Gram-positive bacterium *Clostridium difficile* Ox247, including investigations made into defining the glycosylation site occupancy as well as the glycan compositions. In addition, several *Orfs* of the S-layer cassette 11, discussed in the Introduction chapter 1, were functionally characterised.

The approaches employed in the experimental strategy are two-fold involving glycomics and glycoproteomics (summarised in **Figure 5.1**). The discovery of glycosylation of the S-layer was performed by comparing the S-layer proteins from several gene-deletions of the strain with the wild type *C.difficile* S-layer proteins, as purified by gel electrophoresis, enzymatically digested, separated by liquid chromatography and analysed using the Q-TOF methodology. To gain a further understanding of the overall structures discovered, full profiling of the S-layer glycoprotein using various techniques was employed, shedding light in a site-specific analysis strategy to define the site of attachment of the glycan chain on the S-layer, as well as the composition of the glycan moieties, and confirmation of the amino acid sequence involved in the glyco-conjugation. This required a range of analysis techniques including ETD, GC-MS, MALDI-MS and MS/MS, ES-MS and MS/MS as well as chemical derivatisation and hydrolysis.

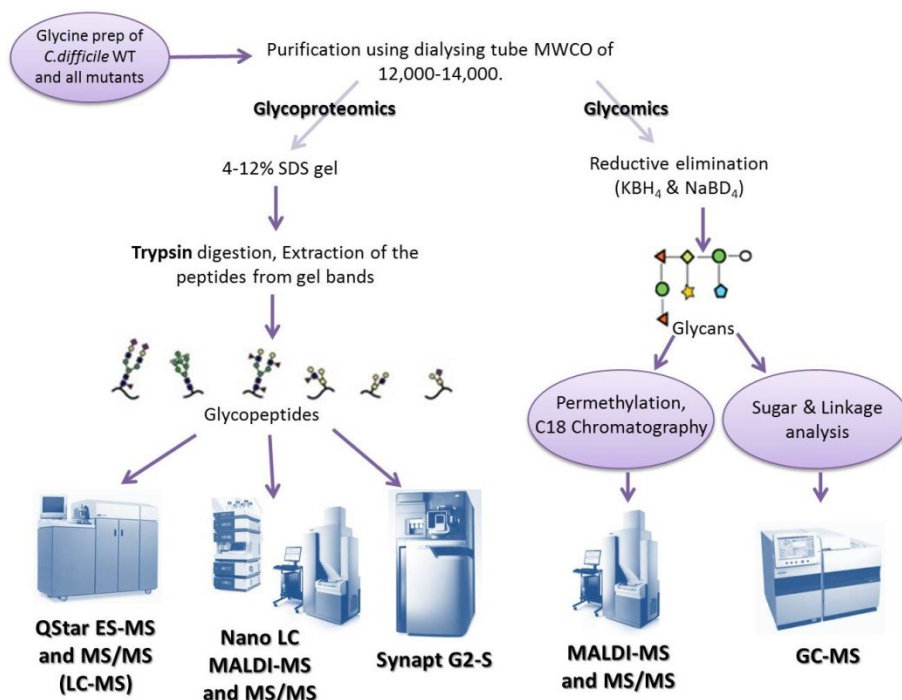


Figure 5.1 Simplified representation of the experimental strategy employed in this section which were two-fold involving glycomics and glycoproteomics. S-layer glycoprotein of *Clostridium difficile* Ox247 and the S-layer proteins from several gene-deletions of the strain under study were purified by gel electrophoresis, enzymatically digested, separated by liquid chromatography and analysed using Q-TOF methodology. Fragmentation data were acquired using multiple different complementary techniques including Collision Induced Dissociation (CID) and Electron Transfer Dissociation (ETD). Moreover, a range of analysis techniques comprising GC-MS, MALDI-MS and MS/MS, ES-MS and MS/MS as well as chemical derivatisation and hydrolysis were employed to gain a further understanding of the overall structures discovered.

Proteomics and glycoproteomics experiments were carried out on samples coming from the MRC Centre for Molecular Bacteriology and Infection at Imperial College London and in particular, this project has been carried out in collaboration with Prof. Neil Fairweather. Insertional mutants were created by his group in *Orf2*, *Orf3*, *Orf4*, *Orf7*, *Orf16* and *Orf19* using the Clostron technique, a targeted insertional mutagenesis system, followed by transformation of plasmids into *E. coli* and conjugation into *C. difficile*. **Figure 5.2** illustrates the putative functions of each gene within the S-layer cassette number 11 (SLC-11). *Orf2* is predicted to be a glycan biosynthesis initiation gene and so responsible for the transfer of a hexose from its nucleotide activated form to the lipid carrier undecaprenyl phosphate at the membrane; instead *Orf3* is an expected rhamnosyltransferase which may be responsible for the addition of the first deoxysugar onto the core hexose by an α 1,3 glycosidic bond; *Orf4* corresponds to a putative phosphoribose diphosphate decaprenyl-phosphate phosphoribosyltransferase, which catalyses transfer of a pentose sugar in the putative glycan structure; *Orf7* is the largest gene within the cluster and it is a putative multi-domain glycosyltransferase; *Orf16*, instead, is one of the five genes encoded which constitute the

rhamnose biosynthetic pathway (*Orfs 13, 14, 16, 17, 18*) and in particular it is predicted to encode the first enzyme, *RmlA*. Finally the *Orf19* is predicted to be an O-specific ligase for binding of glycan onto the protein acceptor (Giraud and Naismith 2000).

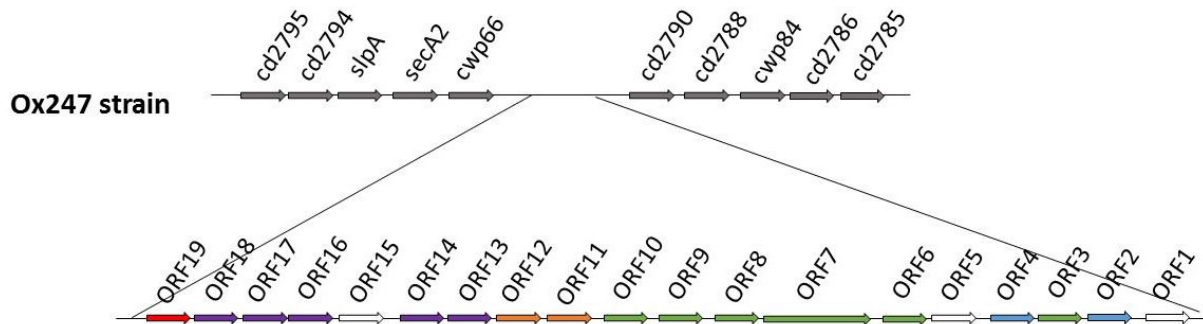


Figure 5.2 The putative 23.8 kb locus inserted between CD2790 and *cwp66* in *Clostridium difficile* Ox247 has been analysed by BLAST and the putative functions are indicated by different colour. White (1, 5, 15): unknown; Blue (2, 4): initiating glycosyltransferases; Green (3, 6- 10): glycosyltransferases; Orange (11, 12): ABC transporters; Purple (13, 14, 16-18): rhamnose biosynthetic genes; Red (19): O-specific ligase.

5.2 Glycoproteomic Results

One of the first indications that a protein may be glycosylated is through an aberrant migration on a gel and this has been shown by both the Fairweather and Wren laboratories to be the case for the small S-layer subunit (Hitchen, P.G. & Dell 1575-1580 2006). Five different mutants have been analysed during the work presented here and **Figure 5.3** shows how they separate on a 4-12% SDS gel run at 200 V for 60 minutes compared to the wild type (WT).

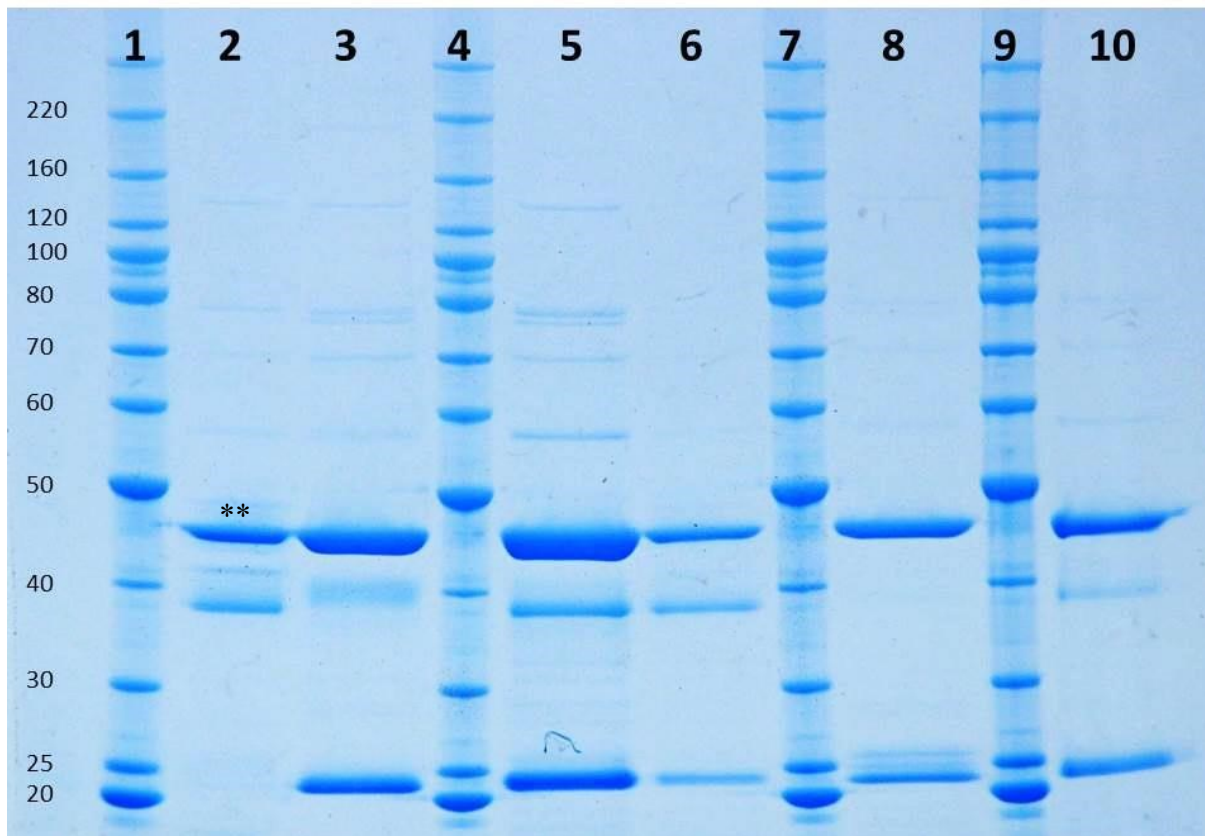


Figure 5.3 Separation of the *C.difficile* proteins by SDS-PAGE gel (Invitrogen NuPAGE 4-12% Bis Tris Gel) and visualized by Coomassie blue staining. Molecular size markers are indicated on the left. Six different kinds of samples were run, wild type (WT) and mutants of the SLC-11 strain Ox247 (*Orf2*, *Orf3*, *Orf4*, *Orf7* and *Orf16*). In the Ox247 strain HMW SLP and LMW SLP (defined in chapter 1) migrate between 50 and 40 kDa. Each mutant, instead, contains a band at ~20 kDa shown by mass spectrometry to be glycosylated LMW SLP. Lanes 1-4-7-9: bench mark (5 μ g); lane 2: wild type (WT); lane 3 $\Delta orf2$; lane 5 $\Delta orf3$; lane 6 $\Delta orf4$; lane 8 $\Delta orf7$ and lane 10 $\Delta orf16$; ** see text.

Excising the Coomassie-staining band at ~48 kDa from the WT Ox247 sample, indicated by ** in **Figure 5.3**, followed by destaining, digestion with trypsin and MS analysis leads to the discovery of a significant glycosylation event. The predicted amino acid sequence of *Clostridium difficile* Ox247 S-layer protein is shown in **Figure 5.4**. Tryptic digests of the various bands observed were analysed by *on-line* liquid chromatography using a reverse phase C₁₈ nano-LC connected to an electrospray quadrupole-TOF mass spectrometer (Q-STAR Pulsar). MS and MS/MS data were acquired using Analyst QS software with the automatic information-dependent MS/MS acquisition (IDA) function (see **Materials & Methods**, chapter 2).

MKKRNLAMAMA**AVTVV**GSAAP**VFA**AASDVISLQDGTNDKYTVSNTKASDLV**KDILAAQNLT**
TGAVILNKDKTVTFYDANEKDSSTPTGDKKVYSEQLTTANGNEDYVKTTLKNLDAGEYAIIDL
 TYNNAKTVEIKVVAASEKTVVSSDAKNSAKDIAEKYVFEDKDLENALKTINASDFSKTDSYYQ
 VVLYPKGKRLQGFSTYRATNYNEGTA YGNTPVILT LKSTSKSNLKTAVEELQKLNASYSNTTTLA
 GDDRIQTAIEISKEYYNDGEKSDHSADV KENVKNVVLVGANALVDGLVAAPLAAEKDAPLLL
 TSKDKLDSSVKSEIKRVLDLKTSTEVTKTVYIAGGVNSVSKEVVTELESMGLKVERFSGDDRY
 ETSLKIAD EIGLDNDKAYVVG GTGLADAMSIASVASTKLDGNGVVDRTNGHATPIVVVDGKA
 DKISDDLDSFLGSADVDIIGGFASVSEKMEEAISDATGKGVTRVKGDDRQDTNSEVIKTY YAND
 TEIAKA AVLDKDSGASSDAGVFNFYVAKDGSTKEDQLVDALAVGAVAGYKLAPVVLATDSLS
 SDQSV AISKVVG EKYSKDLTQVGGIANSVINKIKDLLDM

Figure 5.4 Amino acid sequence of *Clostridium difficile* Ox247 S-layer. Residues in red correspond to the signal peptide. Residues in black represent the LMW SLP “L” - 20 kDa and residues in blue indicates the HMW SLP “H” – 45 kDa. The tryptic glycopeptide DILAAQNLT**TGAVILNK** has been shown in bold and it is part of the LMW SLP sequence.

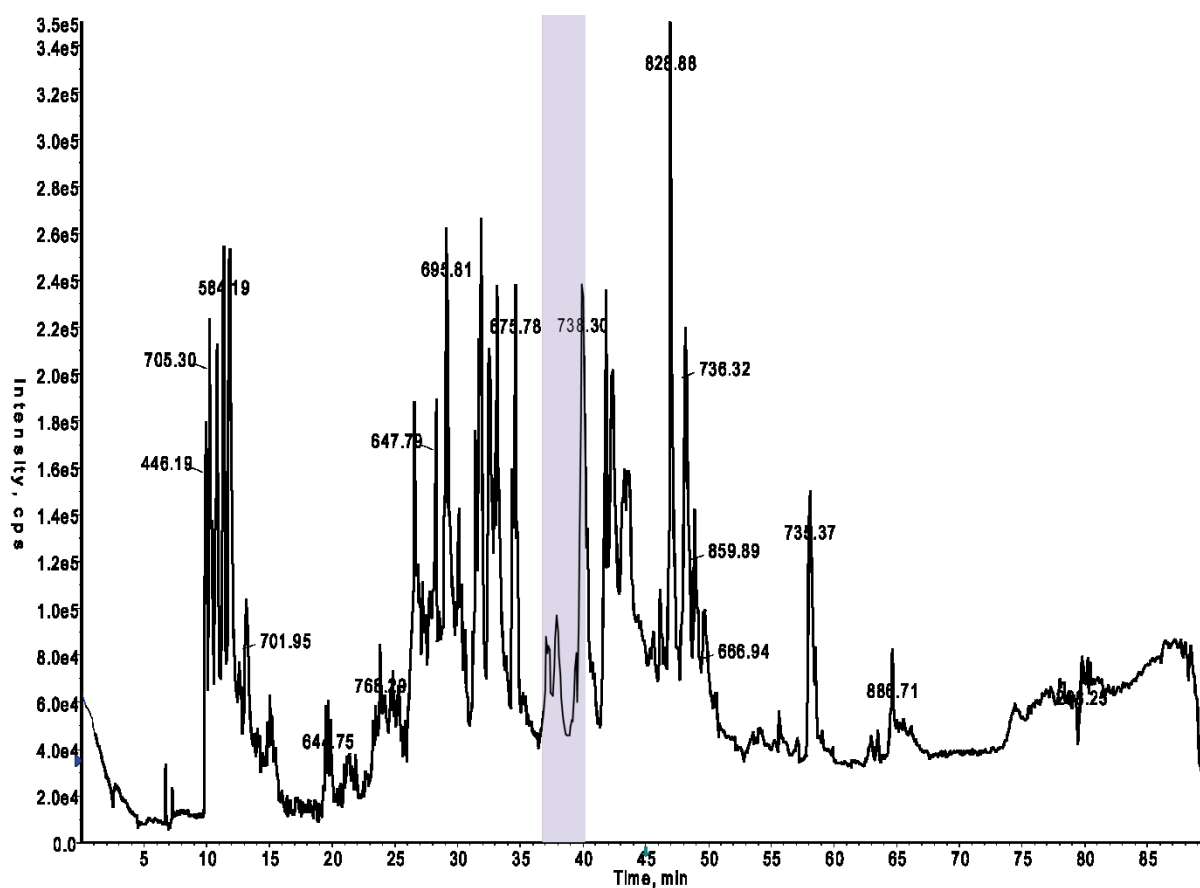


Figure 5.5 TIC (Total Ion Current) for a 90 minute separation of the Ox247 48 kDa (marked in ** in **Figure 5.3**) digest *online*-nano-LC-ES-MS/MS experiment. The shaded box highlights a retention time at approximately 37 minutes. MS spectra of components at this time reveal a pattern of doubly charged ions centred around m/z 1178 with cone-voltage-induced fragment intervals corresponding to deoxyHex, Hex or Pentose.

Looking at the TIC (Total Ion Current) trace (**Figure 5.5**) between 37 and 40 minutes of the tryptic digest of sample Ox247 48 kDa (sample marked in ** in **Figure 5.3**), it is possible to

see a pattern of doubly charged ions centred around m/z 1178 at intervals interpreted as Hex (81 Da), deoxyHex (73 Da) and Pent (66 Da). Thus, because of the presence of a co-chromatographing doubly charged peptide signal which can be assigned from MS/MS fragmentation as DILAAQNLTGAVILNK belonging to the LMW SLP region (**Figure 5.4**), these observations led to the discovery of glycosylation of this protein. Interpretation of this spectrum shown in **Figure 5.6** suggests a precursor molecule having a single structure of deoxyHex and Hex residues, branched with Pent, of at least twelve sugar units of the types illustrated in the schematic below.

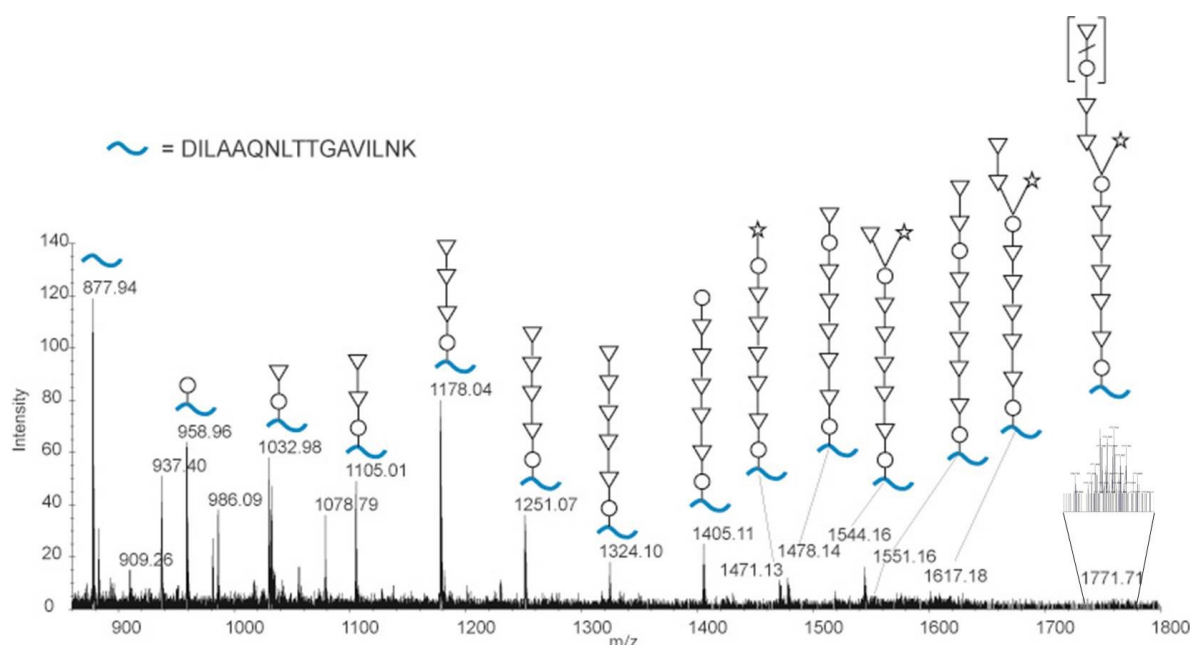


Figure 5.6 Pattern of doubly charged ions centred around m/z 1178 at intervals corresponding to deoxyhexose, hexose or pentose (Hex₃deoxyHex₈Pent₁). On the bottom right of the spectrum there is a magnification of the peak m/z 1771.71 showing that it is a real signal extending to PentHex₈Hex₃. : hexose; : deoxyhexose; : pentose.

Interestingly, the mass chromatograms of signals at for example 878, 959, 1032, 1105, 1178, 1251, 1324 and 1405 are all coincident at 37.9 minute. The important significance of this finding is that the fragments observed uniquely derive from one long structure and do not represent individual glycopeptide quasi-molecular ions of increasing mass. Furthermore, since there is an N-linked consensus sequence, the initial expectation could be linkage through the asparagine-59. However, the MS/MS fragmentation interpretation of 1178 (**Figure 5.7**) shows that this residue is not substituted, since fragment b_7 is a normal Asn cleavage, suggesting that the carbohydrate chain must be O-linked to one of the two threonine residues T-61 or T-62. The pentose branching shown at sugar residue 7 (hexose) in the schematic of **Figure 5.6** is evidenced by the mass difference between m/z 1471.13²⁺ and

1405.11²⁺ (66 Da), and the adjacent signal at m/z 1478.14²⁺ can then be interpreted as a dHex attachment to sugar residue 7 with the sensitive pentose having fallen off, (**Figure 5.8**). Such residues are sensitive to elimination.

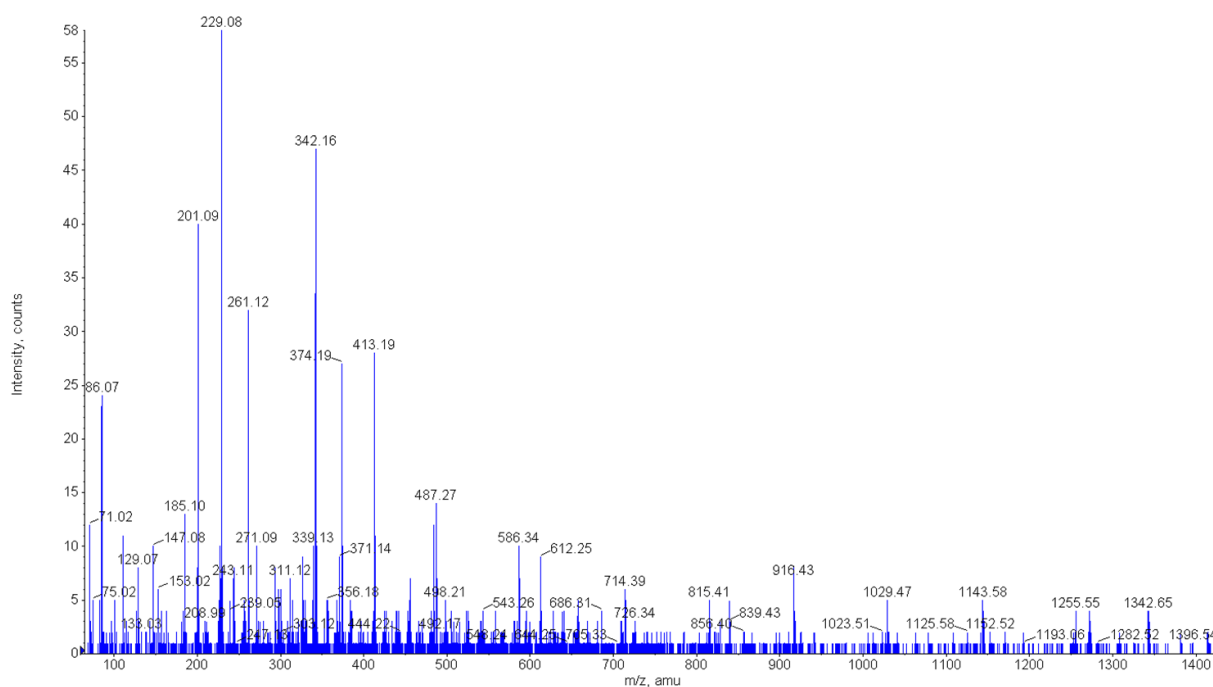
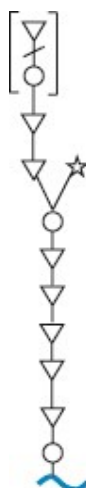


Figure 5.7 MS/MS spectrum of m/z 1178 confirming the DILAAQNLTGAVILNK peptide via signals at m/z 229 (b₂), 342 (b₃), 413 (b₄), 484 (b₅), 612 (b₆), 726 (b₇), 839 (b₈), 261 (y₂), 342 (b₃), 374 (y₃), 413 (b₄), 487 (y₅), 586 (y₆).

877.9²⁺: peptide DILAAQNLTGAVILNK

- 958.9694²⁺ : Hex
- 1032.9889²⁺ : dHexHex
- 1105.0173²⁺ : dHex₂Hex
- 1178.0446²⁺ : dHex₃Hex
- 1251.0721²⁺ : dHex₄Hex
- 1324.1037²⁺ : dHex₅Hex
- 1405.1169²⁺ : dHex₅Hex₂
- 1471.1368²⁺ : PentdHex₅Hex₂
- 1478.1482²⁺ : dHex₆Hex₂
- 1544.1632²⁺ : PentdHex₆Hex₂
- 1551.1695²⁺ : dHex₇Hex₂
- 1617.1808²⁺ : PentdHex₇Hex₂
- 1771.7197²⁺ : PentdHex₈Hex₃



Key

- = Hexose
- △ = deoxyHexose
- ☆ = Pentose

Figure 5.8 The list of the ions identified and related composition. Each number corresponds to the peptide plus the sugar/s attached and possible interpretation of the final structure schematic of the oligosaccharide attached at the DILAAQNLTGAVILNK peptide.

To better understand the biosynthesis of this unique type glycopeptide found in the WT sample, several mutants have been studied using our mass spectrometry methods to identify any changes in glycosylation (**Figure 5.3**). When compared to the WT strain, all mutants examined have shown the presence of a new band at 20 kDa and a different separation of the bands around 45 kDa. The mutants analysed are *Orf 2*, *Orf 3*, *Orf 4*, *Orf 7* and *Orf 16*. All the bands between 50 kDa and 20 kDa were excised, destained, digested with trypsin and analysed by MS.

Looking at the band at 20 kDa of the *Orf2::erm* mutant, the results (**Figure 5.9**) show that it corresponds to the unmodified LMW SLP, indicating that this particular mutant is glycosylation defective, confirming that *Orf2* is a glycan biosynthesis initiation gene and without it, glycosylation would not be present.

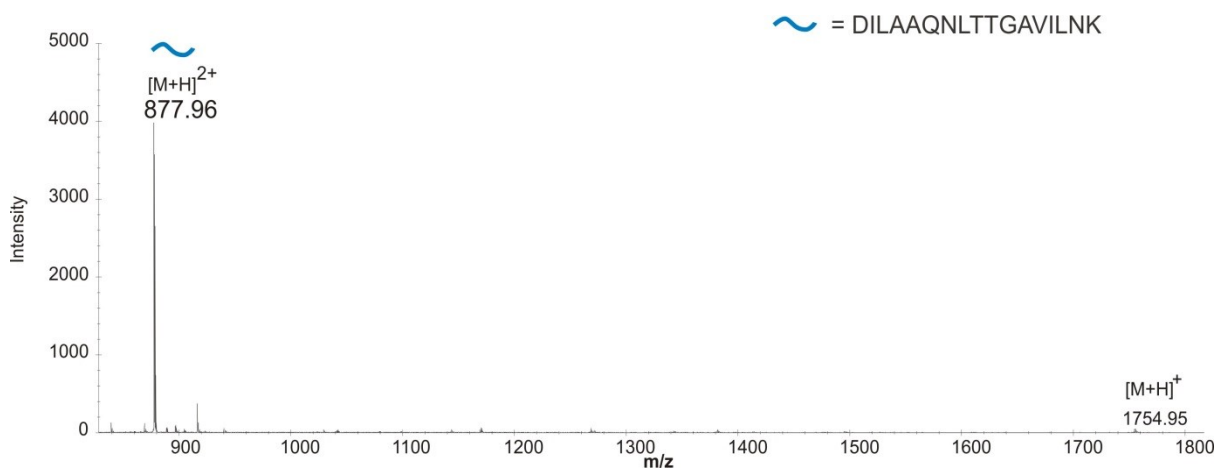


Figure 5.9 Summed MS data acquired at 39.5 min in the online nanoLC-ES-MS of a tryptic digest of *C.difficile Orf2::erm* mutant. The expanded middle mass region to show doubly-charged and singly-charged peaks that correspond to the DILAAQNLTGAVILNK peptide.

The *Orf3::erm* mutant contains a thick band between 20 and 25 kDa (**Figure 5.3**), even though the same sample concentration as the other mutants was loaded. Processing this band shows that the DILAAQNLTGAVILNK peptide is glycosylated by a very much shortened oligosaccharide unit. In fact, the data acquired from the Q-STAR show a deoxyHexHex attached to the LMW SLP (**Figure 5.10**). This enzyme is likely to be responsible for the addition of one or more of the deoxyhexoses but not the one attached to the peptide-linked hexose, since from the experimental results a deoxyHex is found to be attached to the Hex-1 on the peptide.

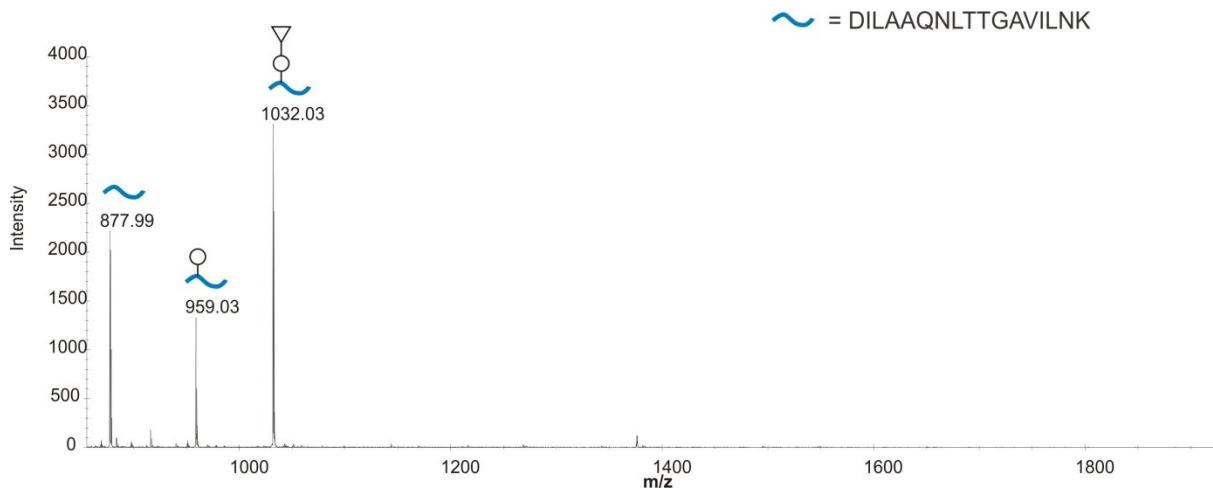


Figure 5.10 Summed MS data acquired at 44.0 min in the *online* nanoLC-ES-MS of a tryptic digest of *C.difficile* Orf3::erm mutant. Expanded middle mass region to show doubly-charged peaks that corresponds to glycans attached to the LMW SLP.

The *Orf4::erm* knock-out mutant S-layer was next analysed. *Orf4* is predicted to catalyse the transfer of a pentose sugar to the putative glycan structure (see **section 5.1**). The band at 20 kDa was processed and in analysing this band belonging to *Orf4::erm* in **Figure 5.3** by mass spectrometry, cone-voltage-induced fragments are observed corresponding to sugar compositions ranging from a single Hex up to deoxyHex₃Hex (**Figure 5.11**) attached to the peptide, which is consistent with the bioinformatics prediction of lack of pentose.

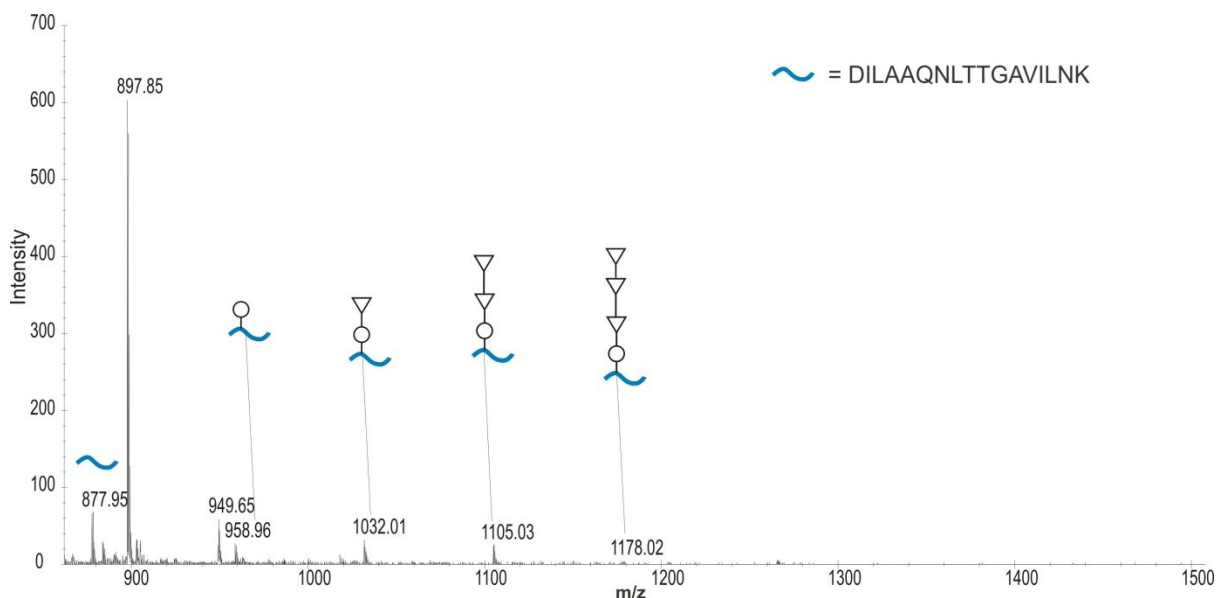


Figure 5.11 Summed MS data acquired at 40.2 min in the *online* nanoLC-ES-MS of a tryptic digest of the band at 20 kDa of *C.difficile* Orf4::erm mutant. Expanded middle mass region to show doubly-charged peaks that corresponds to glycan compositions ranging from a single hexose up to deoxyHex₃Hex. m/z 897.95 is from elsewhere in the digest.

The *Orf7::erm* mutant shows two extra bands above the 20 kDa LMW SLP. This different migration of the bands may be compatible with a putative function of *Orf7* as a multi-domain glycosyltransferase of unknown specificity, but the data obtained by ES-MS (LC/MS) are equivalent and the molecular ions observed in those three bands are the same. The glycan profile corresponds to dHex₅HexPent and the figure below represents just the summed MS data of the band at 25 kDa (**Figure 5.12**).

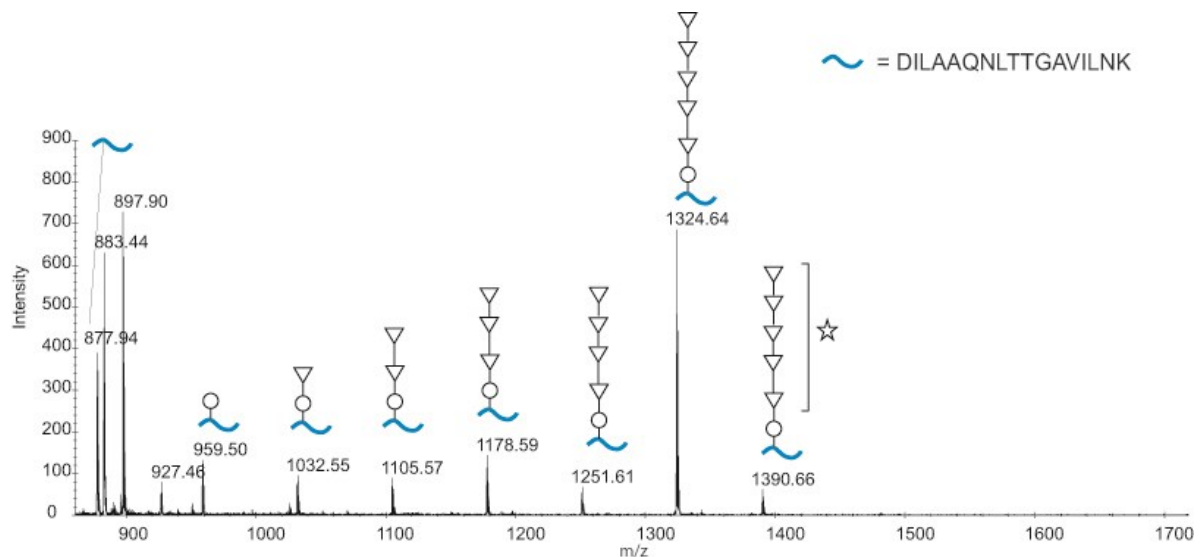


Figure 5.12 Summed MS data acquired at 40.6 min in the *online* nanoLC-ES-MS of a tryptic digest of the band at 25 kDa of *C.difficile Orf7::erm* mutant. Expanded middle mass region to show doubly-charged peaks that corresponds to glycan compositions ranging from a single hexose up to PentdeoxyHex₅Hex.

Interestingly these data show a pentose appearing earlier in the sugar chain than suggested by the WT fragmentation data summarised in **Figure 5.8**.

Finally, the band at 20 kDa of the *Orf16::erm* mutant was studied and found to produce a different pattern of doubly charged ions (**Figure 5.13**). As the *Orf16* is predicted to encode the first enzyme, *RmlA*, and the glycan found attached at the DILAAQNLTGAVILNK peptide is composed of one hexose and four deoxyhexoses. The data suggest that *Orf16* may be responsible for addition of the fifth deoxyhexose.

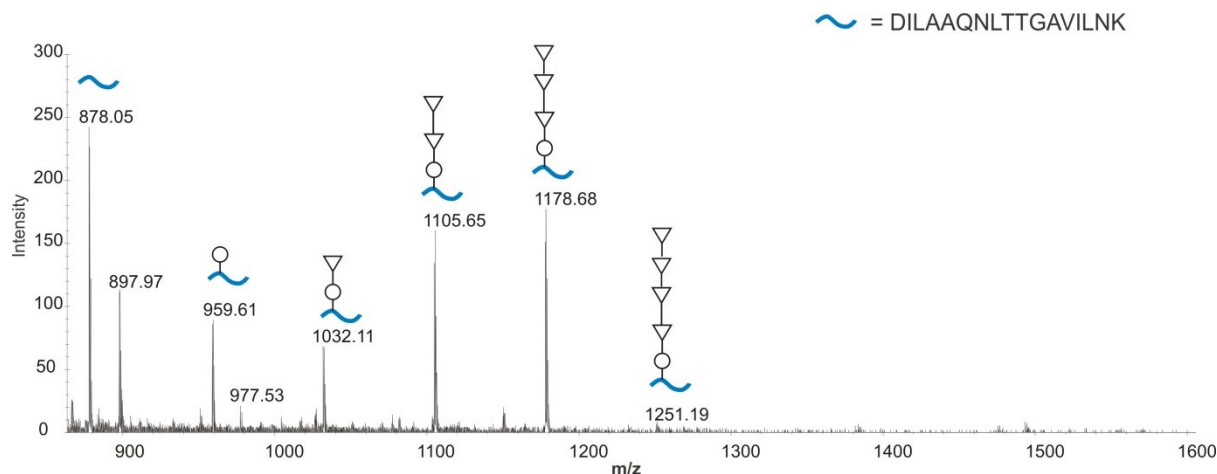


Figure 5.13 Summed MS data acquired at 42.9 min in the *online* nanoLC-ES-MS of a tryptic digest of the band at 20 kDa of *C.difficile Orf16::erm* mutant. Expanded middle mass region to show doubly-charged peaks that corresponds to glycans compositions ranging from a single hexose up to deoxyHex₄Hex.

Further experiments were done on two specific mutants, *Orf2::erm* and *Orf3::erm*, to confirm the data obtained by ES-MS (LC/MS). Following in-gel digestion with trypsin of the lower band at 20 kDa, the subsequent tryptic digests were analysed by *off-line* liquid chromatography/matrix-assisted laser desorption ionisation–time-of-flight/time-of-flight mass spectrometry (*off-line* nano-LC MALDI-TOF/TOF). The digests were first separated on a nano-capillary C₁₈ column and then eluted peptides sequentially spotted onto a MALDI plate. Both results (ES-MS and MALDI-MS) are confirmatory (**Figure 5.9**, **Figure 5.10**, **Figure 5.14** and **Figure 5.15**). **Figure 5.14** shows the MS of the *Orf2::erm* mutant and the major peak at m/z 1754.94⁺ corresponds to the DILAAQNLTGAVILNK peptide without any sugars attached.

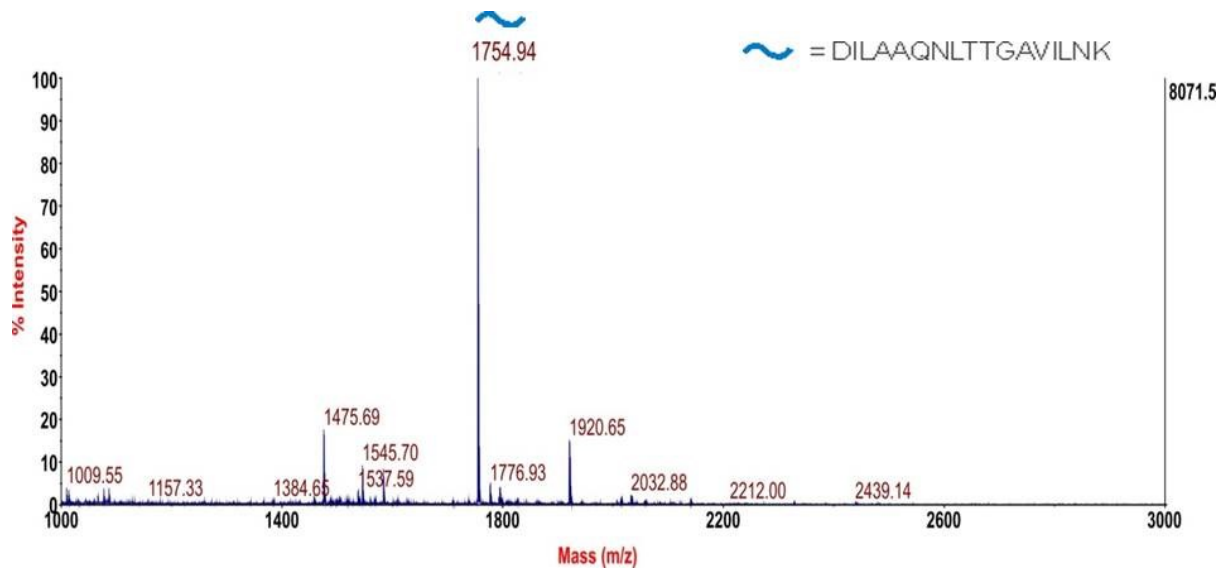


Figure 5.14 MALDI-TOF mass spectrum of *Clostridium difficile* Orf2::erm mutant. The peak at 1754.94 m/z is the DILAAQNLTGAVILNK peptide in accordance with the data obtained by ES-MS (LC/MS) (**Figure 5.7**).

Figure 5.15 illustrates the MS of the Orf3::erm mutant. The peak of interest is m/z 2063.14⁺. This singly-charged ion is consistent with the data shown in **Figure 5.10** because it coincides with the DILAAQNLTGAVILNK peptide plus two glycans attached (deoxyHexHex).

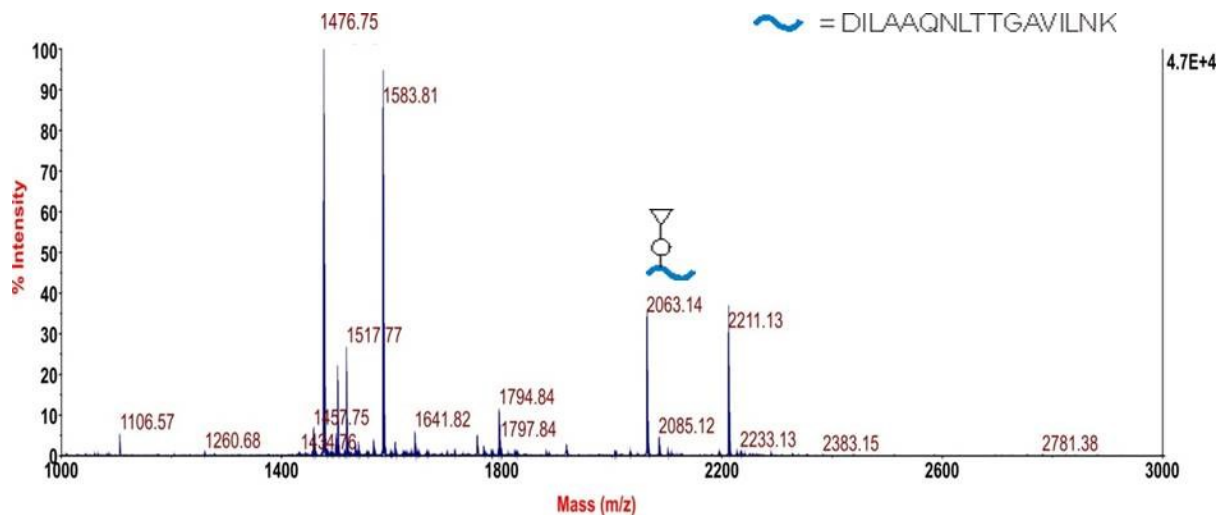


Figure 5.15 MALDI-TOF mass spectrum of *Clostridium difficile* Orf3::erm mutant. The peak at 2063.14 m/z is the DILAAQNLTGAVILNK peptide with an hexose and a deoxyhexose attached ($1754 + 162 + 146 = 2063$) in accordance with the data obtained by ES-MS (LC/MS) (**Figure 5.10**).

To confirm the presence of those two sugars on that specific peptide, and possibly to find the site of attachment, the MALDI TOF/TOF MS/MS spectrum of m/z 2063.14 was produced and can be seen in **Figure 5.16**.

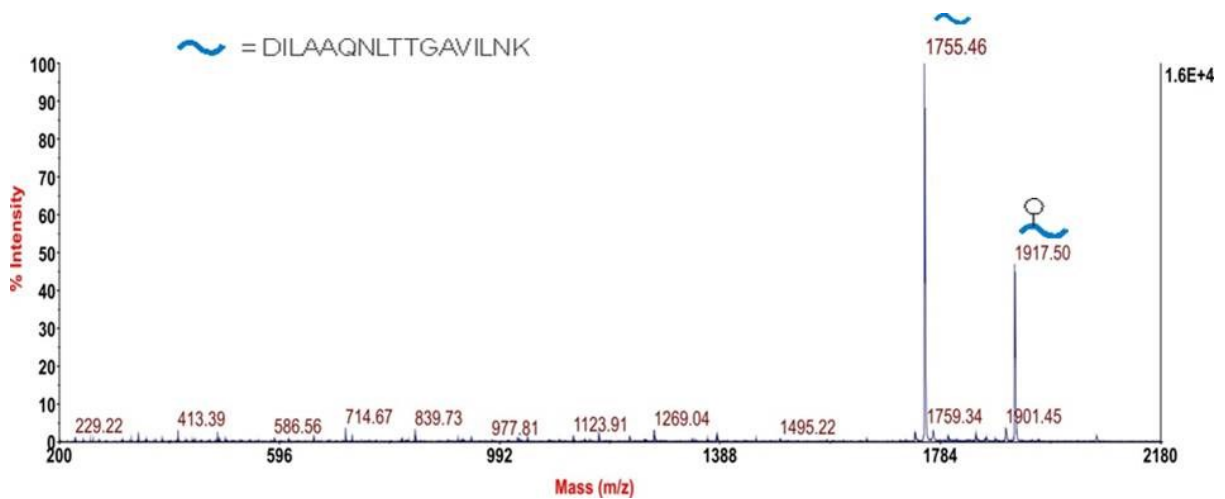
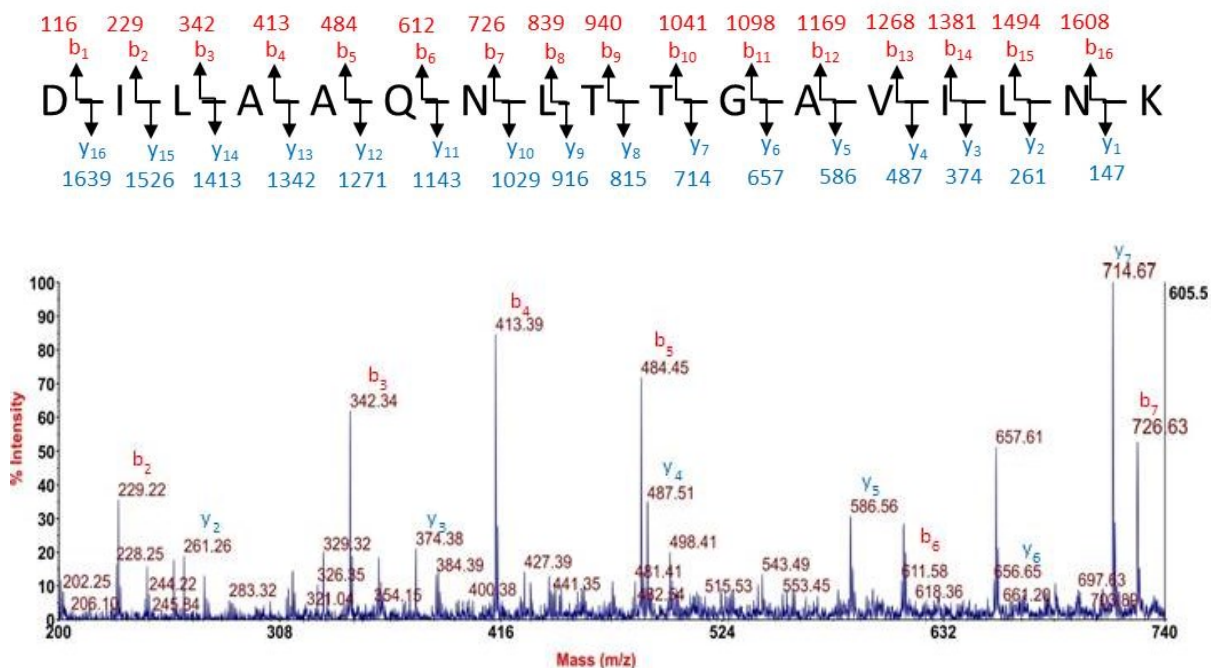


Figure 5.16 MALDI-TOF/TOF MS/MS spectra of m/z 2063.14 derived from the MS spectrum of *Clostridium difficile* Orf3::erm mutant. Assignment of the fragment ions are shown.

Looking at the MALDI TOF/TOF spectrum of m/z 2063.14, as expected the deoxyHex and Hex sugars are partially eliminated to give m/z 1917 and m/z 1755. The peptide sequence can then be unequivocally defined (**Figure 5.17**). All the b ions (except for b₁ and b₁₆) and most of the y ions have been found, validating the presence of a glycan chain on the DILAAQNLTGAVILNK peptide, belonging to the LMW SLP.



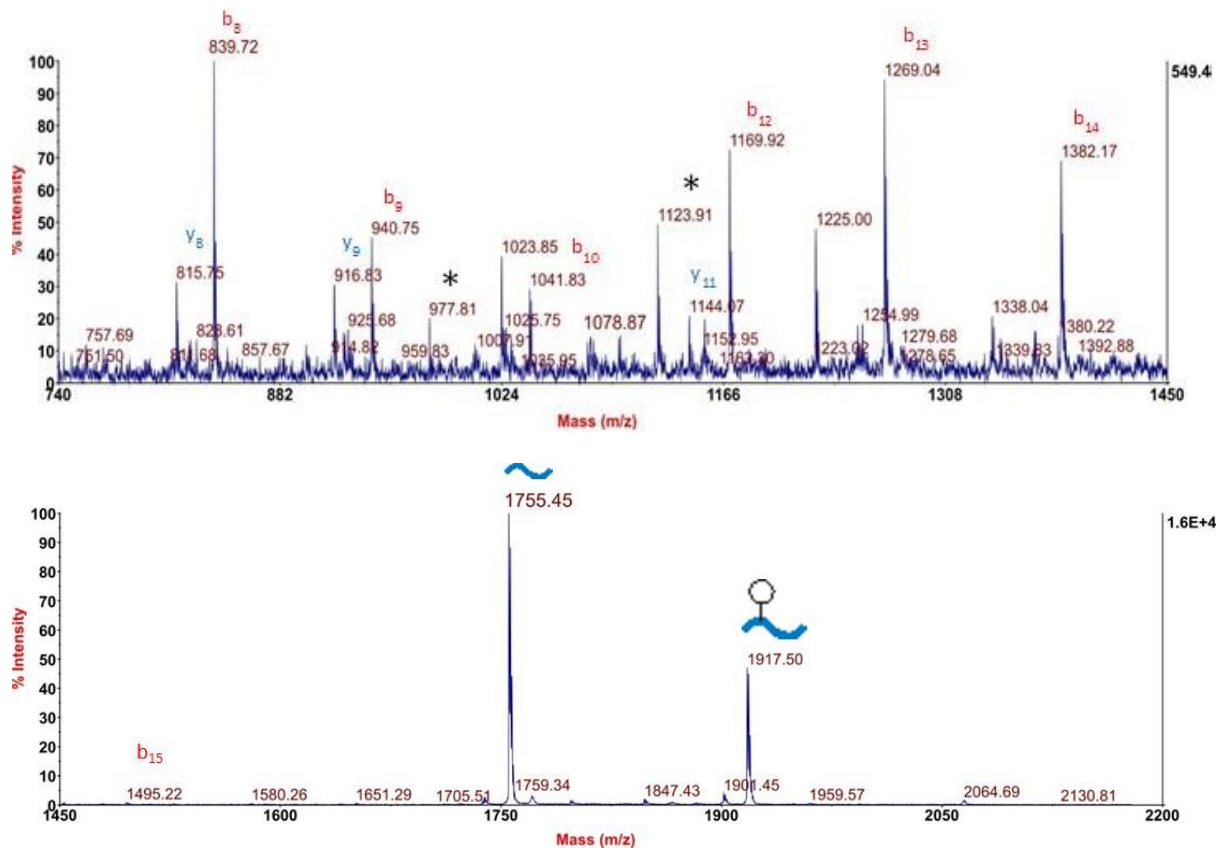


Figure 5.17 Expanded MALDI-TOF/TOF MS/MS spectra of m/z 2063.14 derived from the MS spectrum of *Clostridium difficile* *Orf3::erm* mutant. Peptide fragmentation provides very strong evidence for the sequence DILAAQNLTGAVILNK. b-ions are labelled in red and y-ions are labelled in blue. The peptide fragmentation is shown. Protein backbone cleavages involved lead to the formation of b ions (the charge is retained by the N-terminal fragment) and y ions (the charge is on the C-terminal fragment). The mass difference between two adjacent b or y ions provides the mass and identity of the amino acid residue and thus allows sequencing of the DILAAQNLTGAVILNK peptide. For site of attachment interpretation see text.

In fact, there are three possible candidate sites: DILAAQNLTGAVILNK, although Asn glycosylation is less likely than O-glycosylation because the OTase in *C.difficile* does not resemble other prokaryotic N-OTase. Examining the MS/MS data of the *Orf3::erm* mutant in **Figure 5.17**, there is the evidence that the short glycan chain (deoxyHexHex) is an O-linked glycan, and also signals indicating the site of attachment. The peptide sequence is characterised for having two adjacent threonine (T), but the MS/MS data show substitution of deoxyHexHex on the second threonine from the N-terminus of the sequence determined (T-62), supported by ions at m/z 815 (y_8), 977 ($\Delta 162$) and 1123 ($\Delta 146$) (glycopeptide marked by * in **Figure 5.17**). There is no equivalent evidence of species where only the first threonine (T-61) is substituted, because the “b” series signals are not there (m/z 940, 1102 and 1248). However, this argument is not infallible since b ion series often eliminate sugars more efficiently than y ion series.

5.3 ETD fragmentation

To confirm the glycosylation site data obtained by *off-line* nano-LC MALDI TOF/TOF, a further experiment was then carried out using ETD, incorporated into a new generation Q-TOF instrument, the Waters Synapt G2-S (see **section 1.2.5 C**). Following in-gel digestion with trypsin of the lower band at 20 kDa of the *Orf3::erm* mutant, and *on-line* liquid chromatography elution using a reverse phase C₁₈ nano-LC, the peptides were “supercharged” with a supercharge reagent, m-nitrobenzyl alcohol (m-NBA), as the increasing charge enhances dissociation efficiency and therefore the information from fragmentation (Iavarone and Williams 2002).

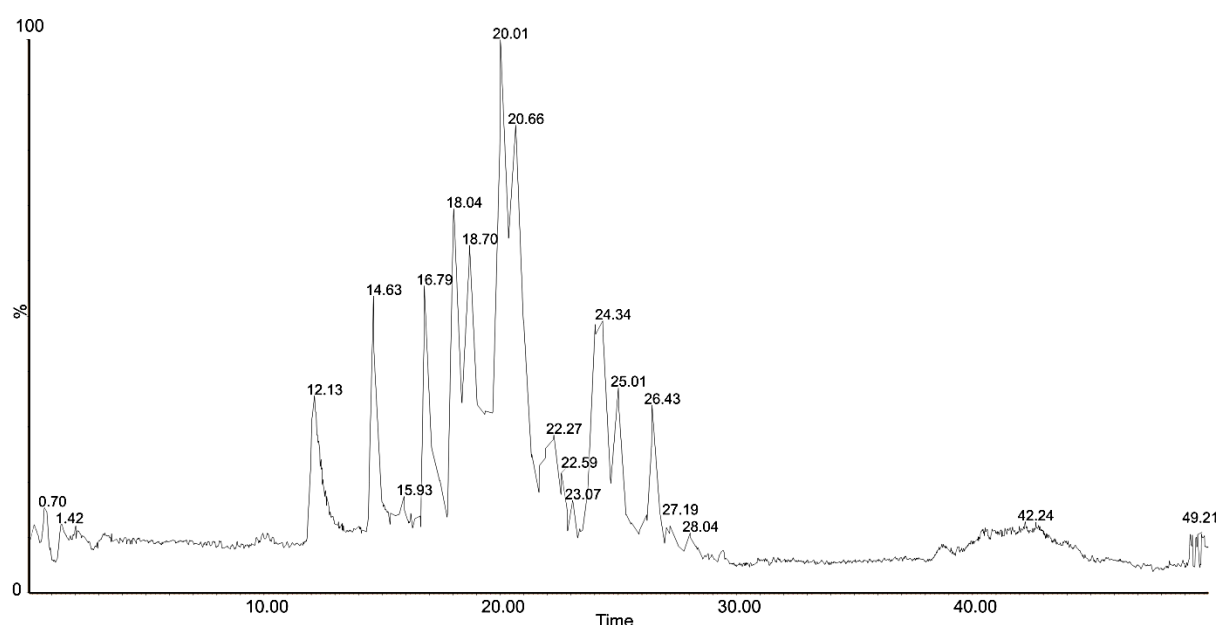
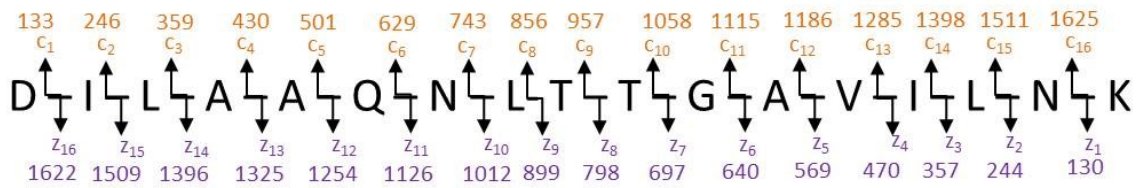


Figure 5.18 TIC for the 50 minute elution profile of *Orf3::erm* 20 kDa Tryptic digest chromatogram for an ETD experiment. The DILAAQNLTGAVILNK peptide is eluted at approximately 26 minute, whereas the glycopeptide is found at approximately 25 minute. Both are seen in a quadruply charged state.

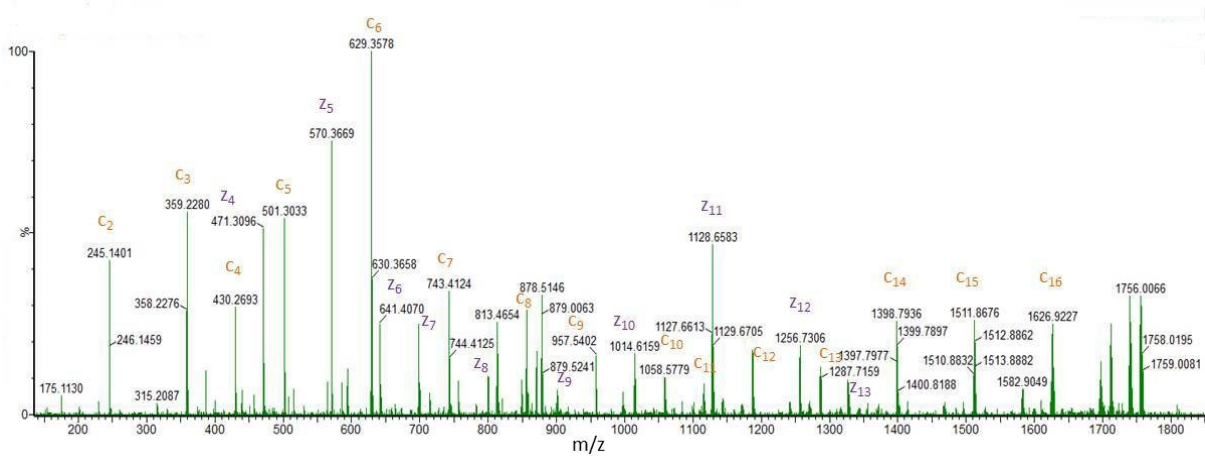
Interpreting the ETD spectrum of m/z 439.52⁴⁺ at 26 minutes, further evidence for this peptide backbone sequence is gained from the fragmentation pattern (**Figure 5.19 A & B**). All the c ions (except for c₁) and most of the z ions are found. Analysing the ETD spectrum of the glycopeptide $[M + 4H]^{4+}$ (m/z 516.55) at 25 minute, there is good evidence for the substitution of deoxyHexHex on the T-62 of the sequence determined, supported by several fragments. First of all considering the “c” series, there is no signal at m/z 1265 (c₉ +

deoxyHexHex), but m/z 957 free c_9 is strong. There is no significant free c_{10} signal at m/z 1058, but there is a strong peak at m/z 1367 that corresponds at $c_{10} + \text{deoxyHexHex}$; secondly coming from the C-terminus of the peptide, the free z_8 fragment ion is absent, but the glycosylated signal is at m/z 1107 ($z_8 + 1 + \text{deoxyHexHex}$) (**Figure 5.19 C**).

A



B



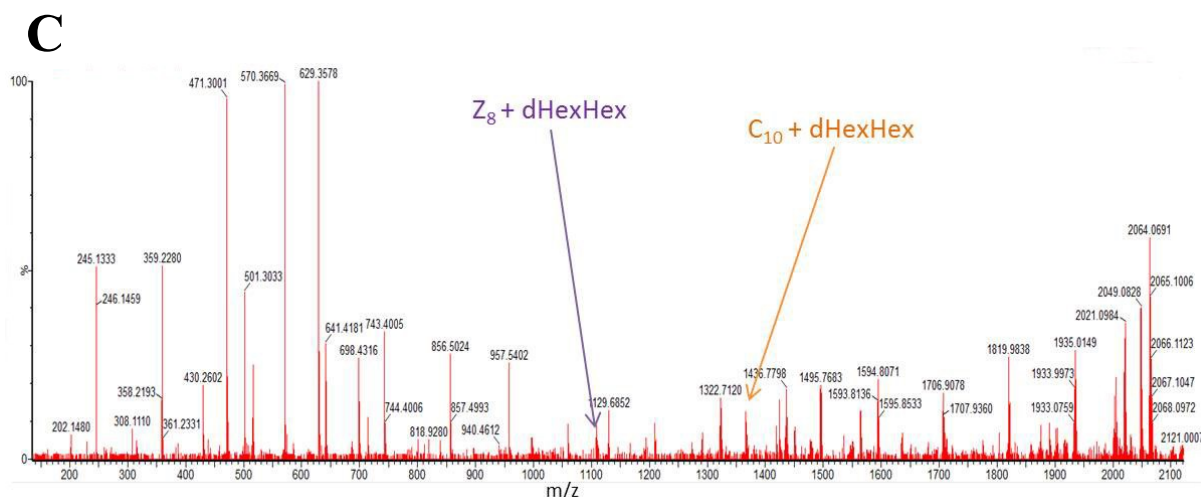


Figure 5.19 ETD spectra of m/z 439.52⁴⁺ and 516.55⁴⁺ derived from the *Clostridium difficile* Orf3::erm mutant. Peptide fragmentation provides very strong evidence for the sequence DILAAQLTTGAVILNK. c-ions are labelled in orange and z-ions are labelled in purple. **A:** The peptide fragmentation is shown. Considering the ETD mechanism, the cleavage of N-C α bonds occurs and produces c- and z- type fragments. The mass difference between two adjacent c or z ions provides the mass and identify the amino acid residue and thus allows sequencing of the selected peptide. **B:** ETD spectrum of the DILAAQLTTGAVILNK peptide: most of the peaks have been assigned at the “c” or “z” series. **C:** ETD spectrum of the glycopeptide.

These data therefore confirm the glycosylation site as Thr-62 but also supplement the MALDI MS/MS data by demonstrating unequivocally the absence of glycosylation at Thr-61.

5.4 β -elimination strategies

Having now discovered a novel long chain O-linked oligosaccharide on LMW SLP together with its precise site of attachment in the protein sequence, the objective was then to define the size of this unusual structure including its non-reducing end. For this objective, the oligosaccharide should ideally be removed from the protein backbone.

Several different procedures to release the O-glycan from the protein were applied, since in initial experiments the standard β -elimination procedure using hydroxide-borohydride was not successful. One method investigated was the process of transferring the glycoprotein from an SDS gel to an Immobilon PVDF transfer membrane using a tank transfer system. The WT strain was run on a 4-12% SDS gel and then the most intense bands were transferred to the PVDF membrane while maintaining their relative positions (**Figure 5.20**). The transfer method used was by electrotransfer.

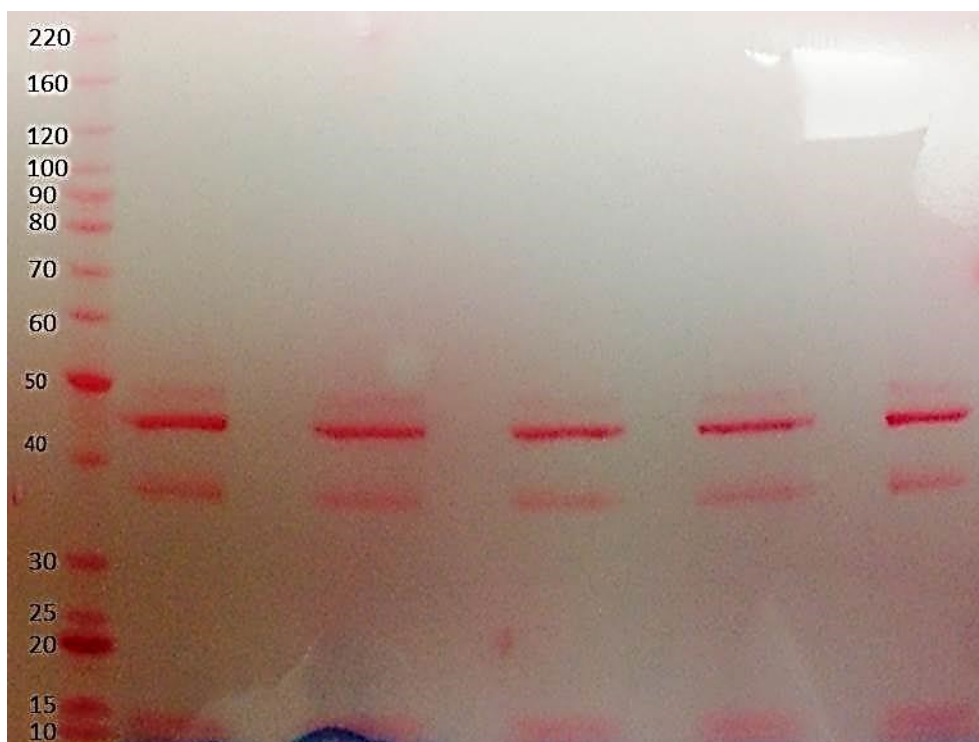


Figure 5.20 Separation of *C. difficile* Ox247 on an Immobilon PVDF transfer membrane and visualized after transfer by Ponceau-S red staining. Molecular size markers are indicated on the left.

The five bands at ~48 kDa were excised and three bands worth of material were analysed using the standard β -elimination strategy; two bands were directly derivatised by permethylation. In both cases the data obtained on MALDI TOF/TOF instrument were not of sufficient quality, as the presence of PEG contaminants seen as intense groups of signals with 44 Da spacing was too high.

Following those unsuccessful experiments, it was assumed that contaminants from the preparations of our microbiological collaborators might be responsible, and therefore another approach of dialysis of the WT sample against 5% acetic acid and then deionised water was taken to purify the sample. Further purification on SDS gels followed by digestion/extraction and β -elimination again failed and a final protocol was to directly β -eliminate the dialysed sample, and this procedure was then finally successful.

5.5 Analysis of the β -eliminated sample

Figure 5.21 illustrates the MS of the WT sample acquired using MALDI-TOF/TOF after permethylation of the sample. The two acetonitrile fractions (35% and 75%) were then analysed separately by MS.

Long glycan chains were found belonging to the S-layer protein of *C.difficile* Ox247 with the highest mass peak in the 35% fraction seen at m/z 3512 (Figure 5.21).

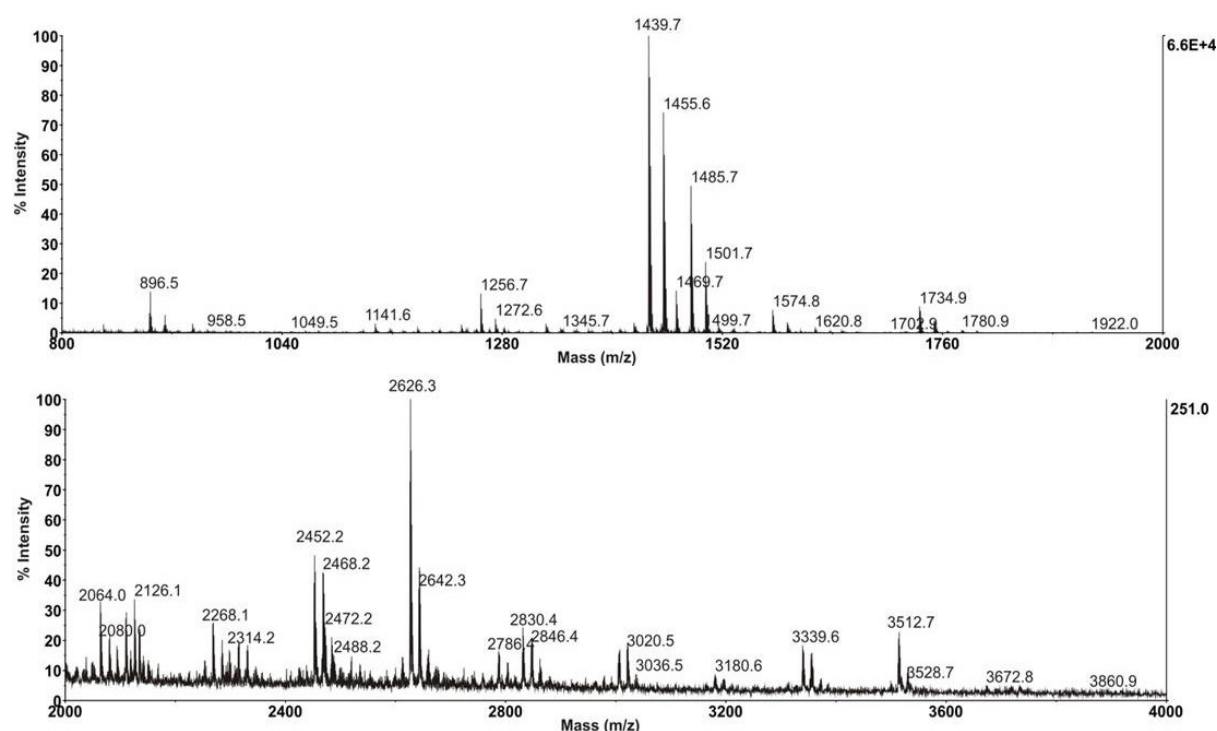


Figure 5.21 MS spectra of WT SLP *C.difficile* Ox247 from m/z 800 to 2000 and from m/z 2000 to 4000. Clearly, these spectra show the presence of a series of peaks corresponding to the mass differences of monosaccharide residues, specifically hexoses, deoxyhexoses and pentoses, belonging to the SLP of *C.difficile* Ox247.

The presence of many 204 and 174 mass differences indicates polymers related to the observation deduced earlier from the electrospray data and therefore, several peaks were then selected for MS/MS, such as m/z 2452 and 3512 (Figure 5.22). These spectra interestingly show patterns in the low end and middle masses, namely m/z 389, 563, 737, 767, 941/2, 1447/9, 1624, 1653, 1827/9, 2001, where it is possible to see a “phasing” of related structures separated by deoxyhexose or hexose, and specifically considering the MS/MS spectrum of m/z 3512 it is possible to see extended glycan chains formed by deoxyHex-Hex-deoxyHex-(deoxyHex,Pent)-deoxyHex-Hex-deoxyHex to get to m/z 2001 starting for example from m/z 563.

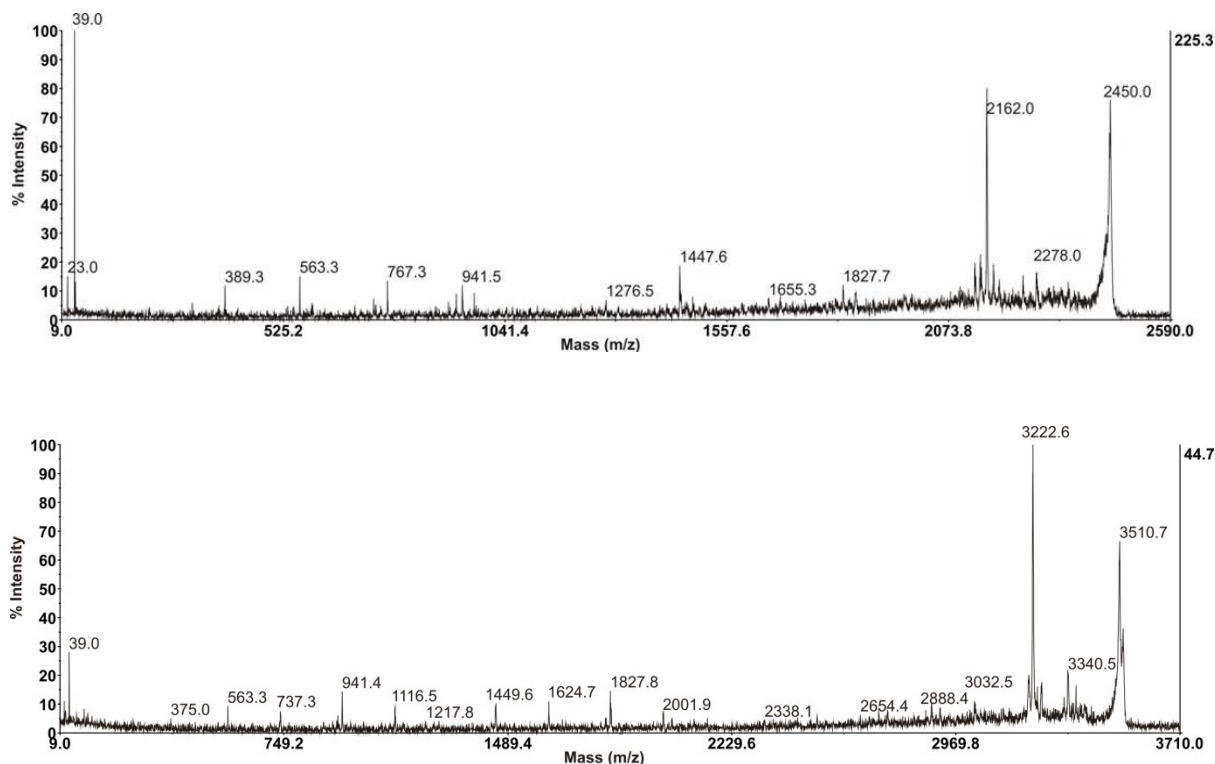


Figure 5.22 MS/MS spectra of m/z 2452 and 3512 belonging to WT SLP *C.difficile* Ox247.

Similarly, looking at the MS/MS spectrum of m/z 3339 (Figure 5.23) and starting from the peak at m/z 563, it is possible to see a glycan chain comprising Hex-deoxyHex-(deoxyHex,Pent,deoxyHex)-Hex-deoxyHex to get to m/z 1829, probably due to phased extensions.

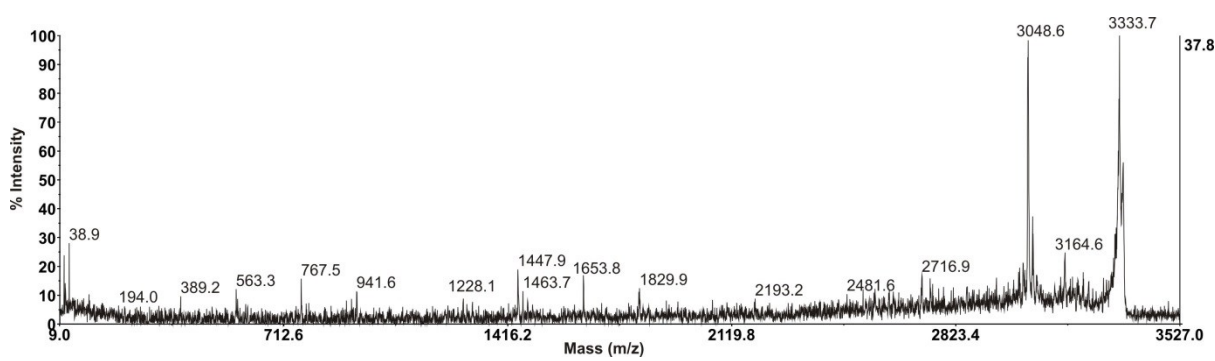


Figure 5.23 MS/MS spectrum m/z 3339.

Moreover, it is possible to infer a possible mechanism which explains the “starting number” at the non-reducing end via a sodiation of the glycosidic bond and β -elimination with

hydrogen transfer to the non-reducing end sugars to give mathematically 15 + sugar residue + 40 (OHNa). So postulating the glycan chain presents at the end one pentose and one deoxyhexose that means m/z $15+160+174+40 = 389$, then adding one more deoxyhexose it takes us to the above starting number 563. While some rationalisation of the spectra were possible in this way, it was not possible to extrapolate to a total structure encompassing the molecular ions using this logic and known sugar masses.

Interestingly, in the MS coming from the 75% ACN fraction of the WT sample acquired using MALDI TOF/TOF after permethylation, it is possible to see that the glycan chain decorating the S-layer protein of *C.difficile* Ox247 is much longer than the ES-MS data could indicate, possibly allowed by the very low internal energy transfer associated with MALDI ionisation. In fact, from the MALDI data in **Figure 5.24** there is evidence of a long glycan chain of almost fifty sugar residues with the highest reasonably intense peak at m/z 8520, with evidence that even that is not the limit of the structure. **Figure 5.24** consists of a zoom-in of the main spectrum to highlight the 204, 174 and 160 mass differences, which even at high mass values indicate the presence of polymers relating to the electrospray observations.

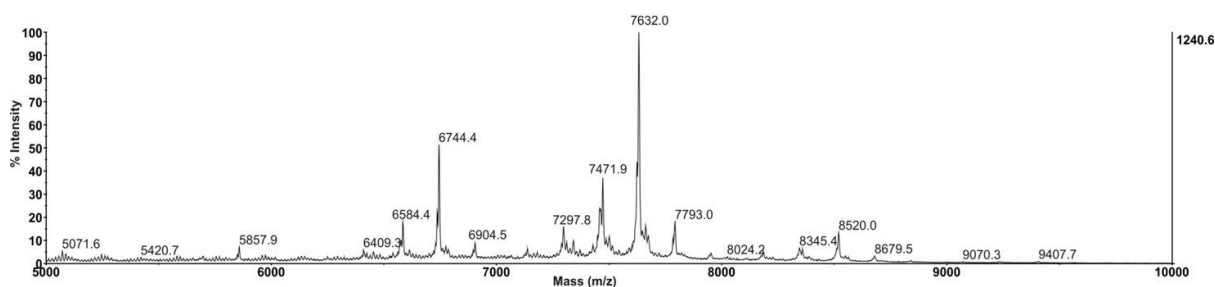


Figure 5.24 MS spectrum from m/z 5000 to 10000 of the WT strain provides very strong evidence for the presence of a long glycan chain of almost fifty sugars belonging to the S-layer protein of *C.difficile* Ox247.

It is also clear from the MS data in **Figure 5.24** that the phasing series discussed earlier now has peak maxima at some of the mass numbers observed which are 886 Da apart. This corresponds to a permethylated HexdHex₃Pent composition ($204+174 \times 3+160=886$). In other words, there is clear evidence for this repeating unit leading to different polymer lengths.

Several MS/MS experiments were then carried out by MALDI TOF/TOF, specifically the MS/MS of m/z 6744 and 7632 (**Figure 5.25**). In these MS/MS spectra, the presence of certain “starting numbers” are clear eg m/z 1509 which calculates for one reduced hexose (275), one hexose (204), five deoxyhexoses (5×174) and one pentose (160), which is equivalent to the (pre-reduced) oligosaccharide composition for m/z 1471^{2+} in **Figure 5.6**.

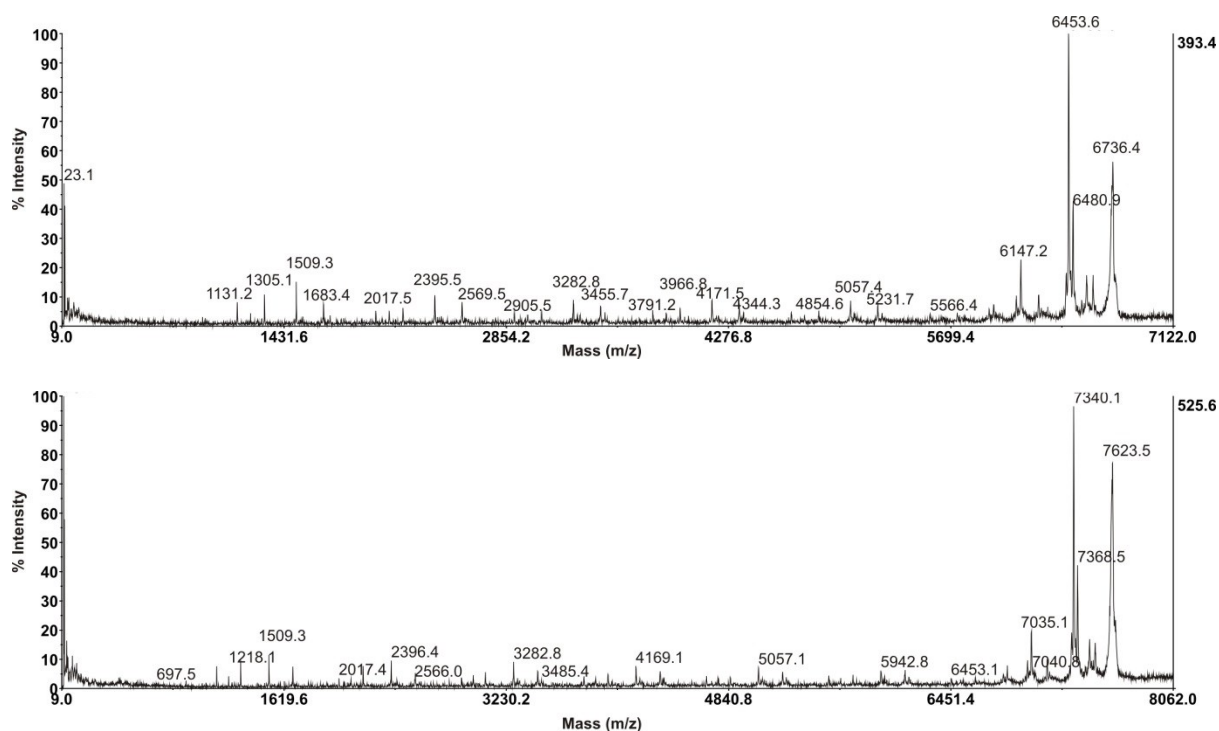


Figure 5.25 MS/MS spectra of respectively m/z 6744 and 7632 coming from the MS showed in Figure 5.24.

Starting from m/z 1509, a series of peaks is found corresponding to a difference of hexoses, deoxyhexoses and pentoses. To make the interpretation easier, the block unit of m/z 886 (eg 1509 to 2395) has been considered in the permethylated data. In fact, it appears that, the most favourable cleavage occurs between the deoxyhexose residue and the following hexose, likely as a consequence of the α or β configuration of the sugar anomeric carbon (Figure 5.26).

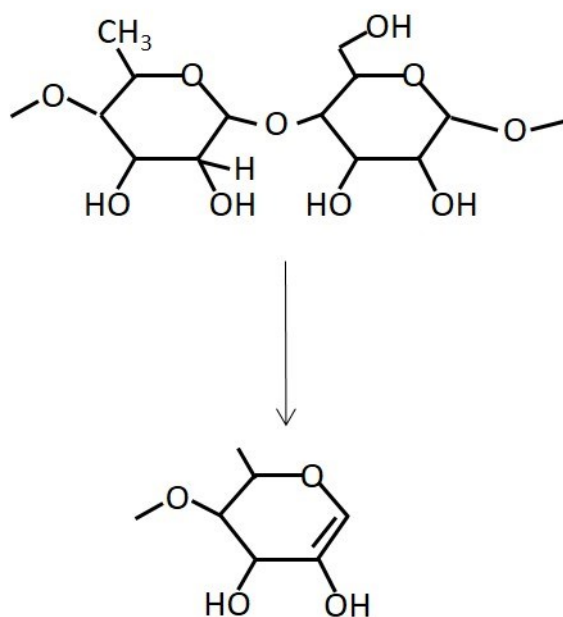


Figure 5.26 Cleavage between the Rha and the Glc residues within the glycan chain under study. This cleavage may be facilitated as a result of the α or β configuration of the anomeric carbon affecting the relative position in space of the oxygen and hydrogen atoms involved in the Hydrogen transfer.

Analysing the MS/MS spectrum of m/z 6744 shown in **Figure 5.25**, the most abundant peaks are the ones with a mass difference of 886 starting from “the core” (m/z 1509), such as the m/z 2395, 3281, 4167, 5053. Note since the reflectron resolution is lost going to higher masses (above around 3000) the data points become the unresolved average chemical masses which moves the computer-labelled mass to a higher position than the theoretical ^{12}C mass. Beyond this point and up to the quasi-molecular ion region (i.e. towards the non-reducing end of the structure) it is unfortunately not possible to fit those masses to simple Hex/dHex/Pent combinations and therefore the non-reducing end remains unidentified. There are however losses from the quasi-molecular ion regions of the MALDI spectra expanded in **Figures 5.27** and **5.28** which normally corresponds to losses from a non-reducing end with charge held on the remaining reducing end fragment. These losses were also not interpretable on the basis of the sugars so far found in the study.

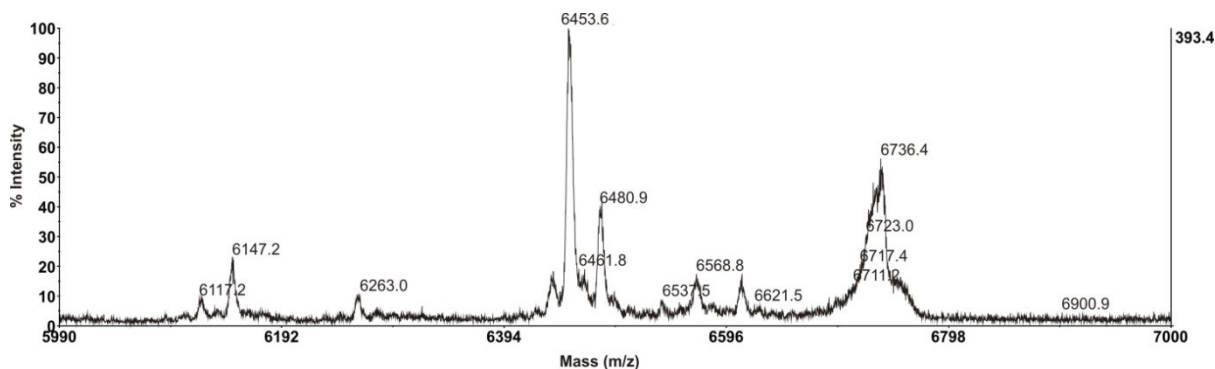


Figure 5.27 MS/MS spectrum of m/z 6744 at the high end on the mass spectrum and showing the unusual fragmentation pattern.

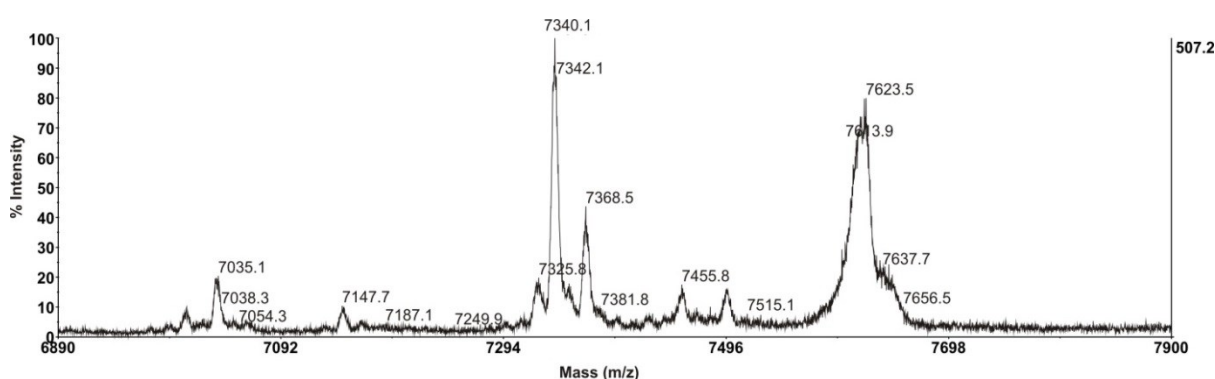


Figure 5.28 MS/MS spectrum of m/z 7632 at the high end on the mass spectrum and showing the unusual fragmentation pattern.

Another set of experiments was carried out in order to confirm and/or extend our knowledge of the structure gained from the MS and MS/MS data, consisting in the treatment of the *C.difficile* Ox247 S-layer with CD₃I to compare with the standard permethylation methodology (**Figures 5.24**). The deuteropermethyated *C.difficile* Ox247 WT SLP spectrum is shown in **Figure 5.29**, confirming the presence of long glycan chains of more than forty sugar residues decorating the SLP LMW of *C.difficile* Ox247, in which the quasi-molecular ions are showing the difference in chain lengths as HexdHex₃Pent via, for example, signals at m/z 6991 and 7911 (theoretical 919, equivalent to 886 in CH₃I permethylation). An easy loss of multiple pentoses (up to 4 or 5) is observed as 166 mass differences in molecular ion groupings.

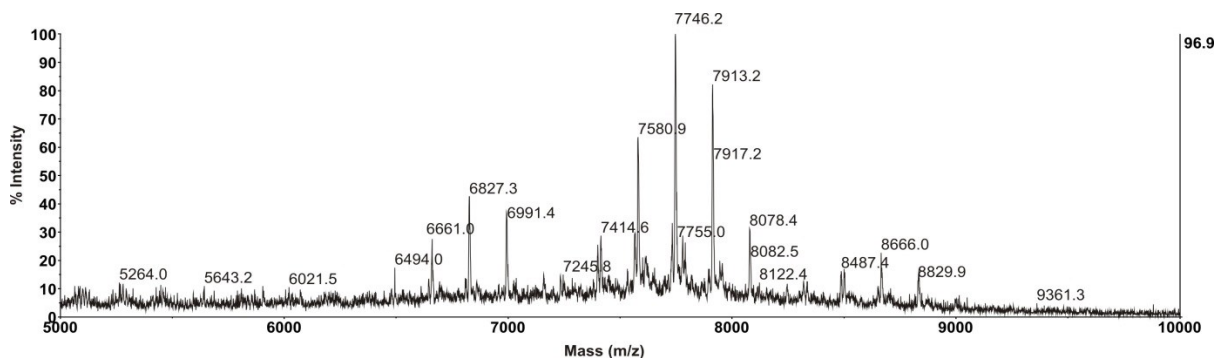


Figure 5.29 MS spectrum from m/z 5000 to 10000 of *C.difficile* WT SLP after β -elimination and deuteromethylation procedures have been performed. This MS spectrum of the WT strain confirms the presence of a long and fragile glycan chain of up to more than forty sugars belonging to the S-layer protein of *C.difficile* Ox247.

Several peaks were then MS/MSed, specifically m/z 6991 and 7746 (**Figure 5.30**) and analysing the MS/MS spectra confirms the interpretation of the non-deutero data.

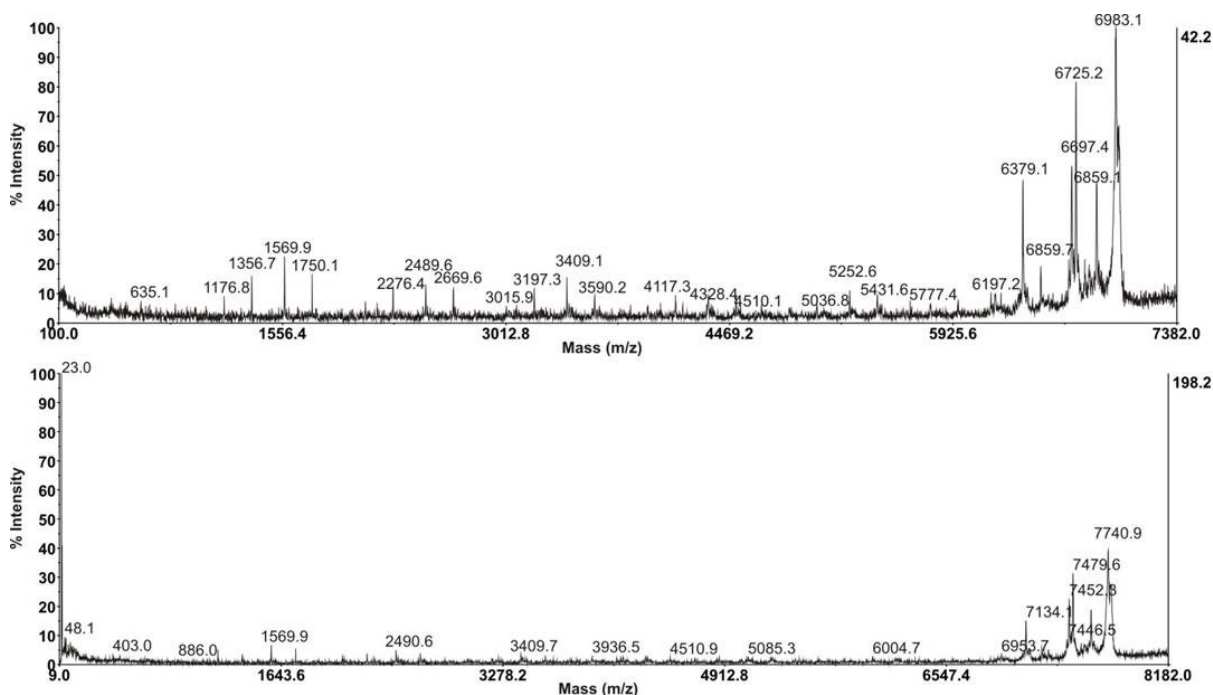


Figure 5.30 MS/MS spectra of m/z 6991 and 7746 of *C.difficile* WT SLP after β -elimination and deuteromethylation using CDI_3 procedures.

5.6 Composition data:

The next question to address in order to complete the sugar profiling of the novel long glycan chain decorating the LMW SLP of *C.difficile* Ox247 was the composition of the sugars

comprising in the glycan chain and if possible their linkages. The WT SLP *C.difficile* Ox247 sample, as extracted and purified by dialysis, was reductively eliminated and hydrolysed into monosaccharides. These were derivatised to form alditol acetates.

Alditol acetate derivatised samples were analysed by GC-MS and run alongside standards using the retention times (elution times) on the gas chromatogram as well as the specific mass spectrometric fingerprints of each monosaccharide to determine the monosaccharide composition of the glycans.

The GC-MS chromatograms for alditol acetate derivatives of the S-layer derived glycans together with standards are shown in **Figure 5.31**.

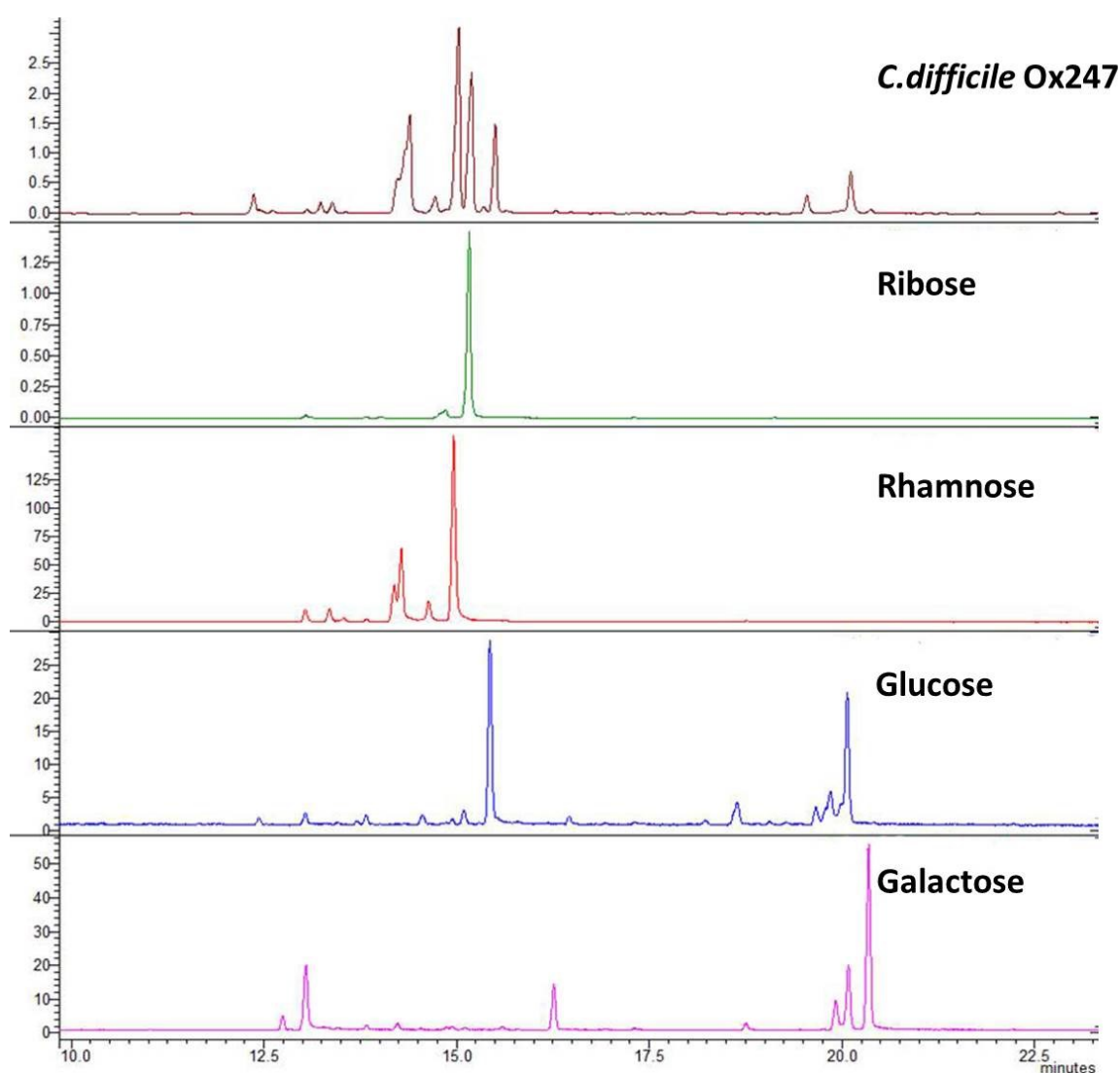


Figure 5.31 GC-MS chromatogram of alditol acetate derivatised glycans of *C.difficile* Ox247 WT SLP. The glycan composition of the *C.difficile* Ox247 S-layer (top panel) has been run alongside commercially obtained standards, starting from the top ribose, rhamnose, glucose and galactose respectively.

From the GC-MS retention times together with the MS data in **Figures 5.32, 5.33, 5.34** and **5.35** the O-glycan of the S-layer glycoprotein was found to contain rhamnose, ribose, galactose and glucose. For a complex natural product, such as that found here for the S-layer glycan, containing approximately fifty sugar residues or more, it was not thought reliable to try to make quantitative calculations of relative ratios. The GC-MS trace above also not unexpectedly contains signals other than standard sugars found, but these were not identifiable.

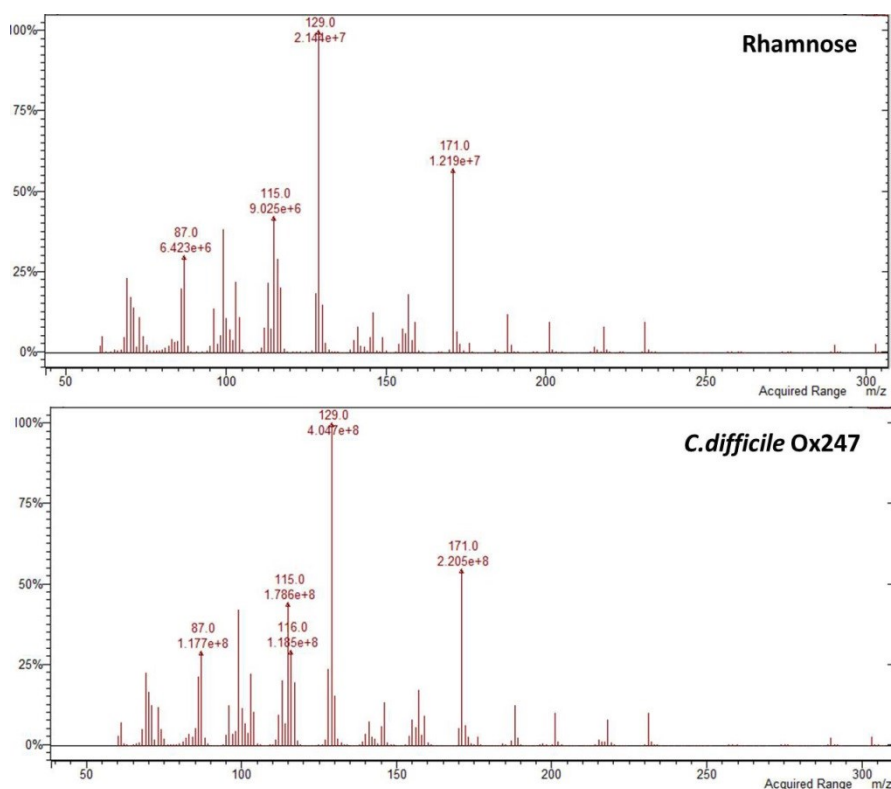


Figure 5.32 Mass spectrometric fingerprint of alditol acetate rhamnose. The fingerprint is from the GC-MS chromatogram of the alditol acetate commercially obtained standard for rhamnose moiety (top panel) and that of the *C.difficile* Ox247 WT SLP (bottom panel).

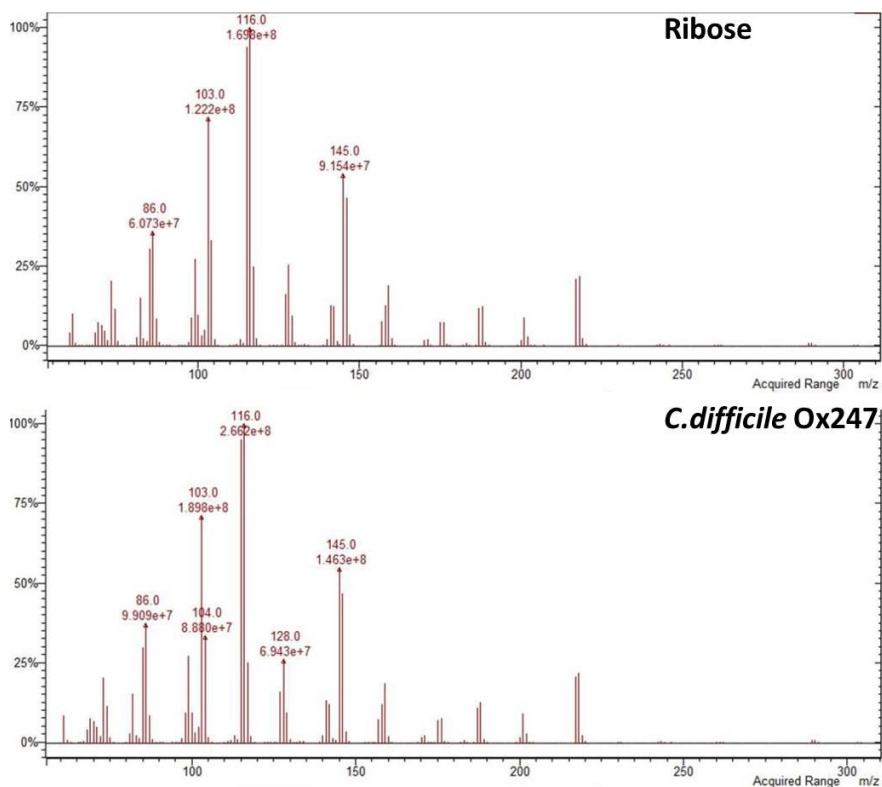


Figure 5.33 Mass spectrometric fingerprint of alditol acetate ribose. The fingerprint is from the GC-MS chromatogram of the alditol acetate commercially obtained standard for ribose moiety (top panel) and that of the *C.difficile* Ox247 WT SLP (bottom panel).

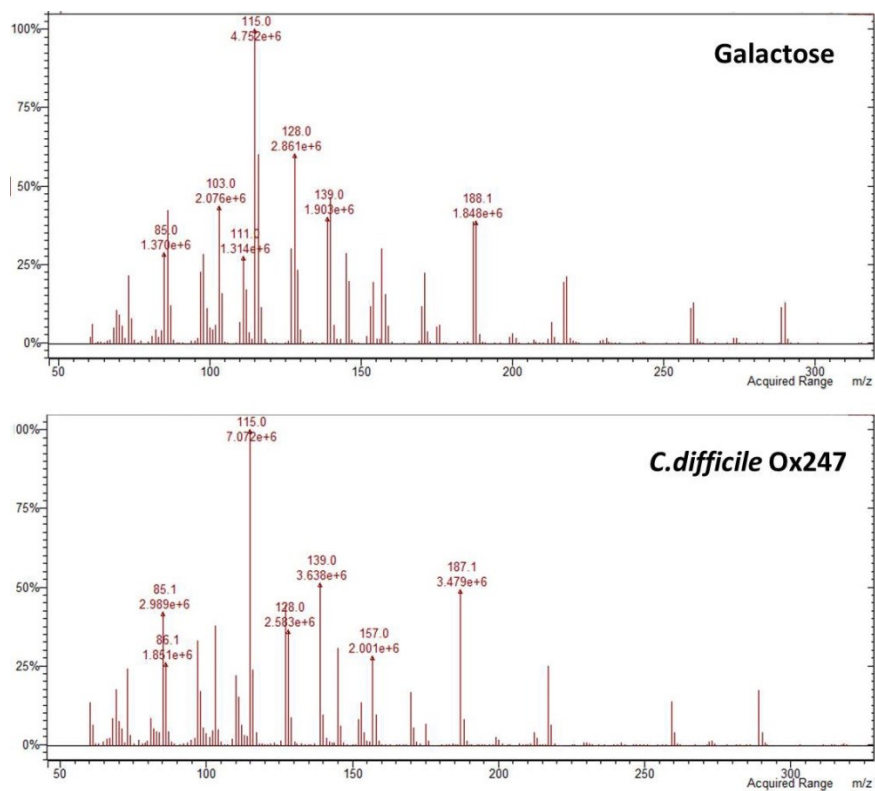


Figure 5.34 Mass spectrometric fingerprint of alditol acetate galactose. The fingerprint is from the GC-MS chromatogram of the alditol acetate commercially obtained standard for galactose moiety (top panel) and that of the *C.difficile* Ox247 WT SLP (bottom panel).

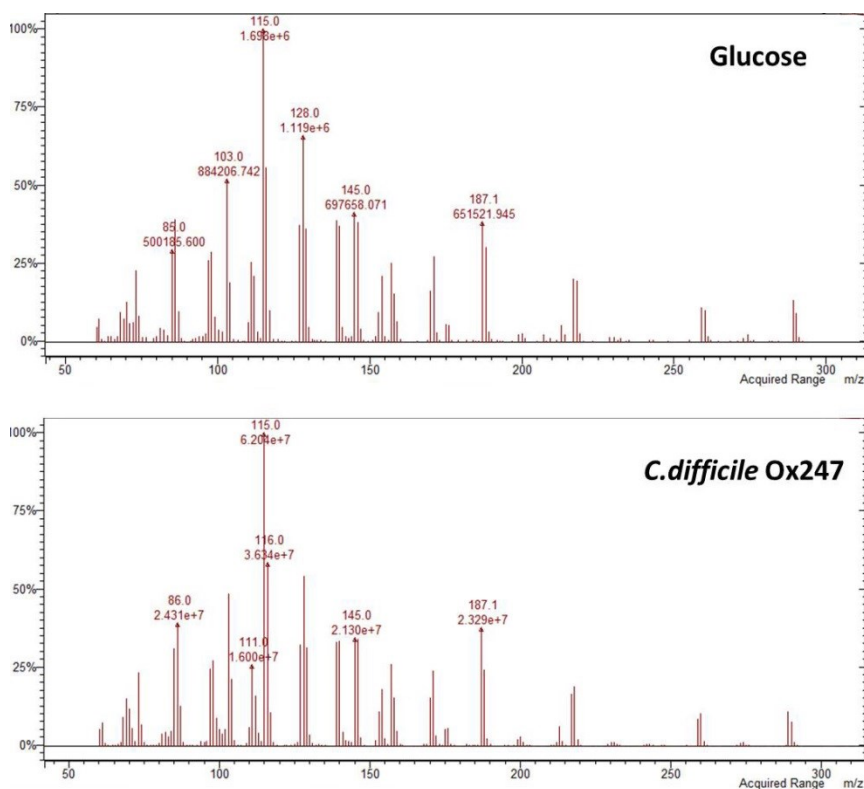


Figure 5.35 Mass spectrometric fingerprint of alditol acetate glucose. The fingerprint is from the GC-MS chromatogram of the alditol acetate commercially obtained standard for glucose moiety (top panel) and that of the *C. difficile* Ox247 WT SLP (bottom panel).

An attempt was made to carry out linkage analysis on the permethylated β -eliminated oligosaccharide by the hydrolysis and acetylation method to produce PMAAs (Sweet 1974). The GC-MS trace produced from the experiment is showed in **Figure 5.36**. Examination of the mass spectra of each peak eluted from the gas chromatography column permitted only some of them to be assigned by comparison with the database spectra publically available on the Complex Carbohydrate Research Centre website at Atlanta, Georgia (CCRC). Positive assignments were possible for terminal ribose, 3-linked rhamnose, 4-linked rhamnose, 3,4-linked rhamnose and 4-linked glucose (**Figure 5.37-5.41**).

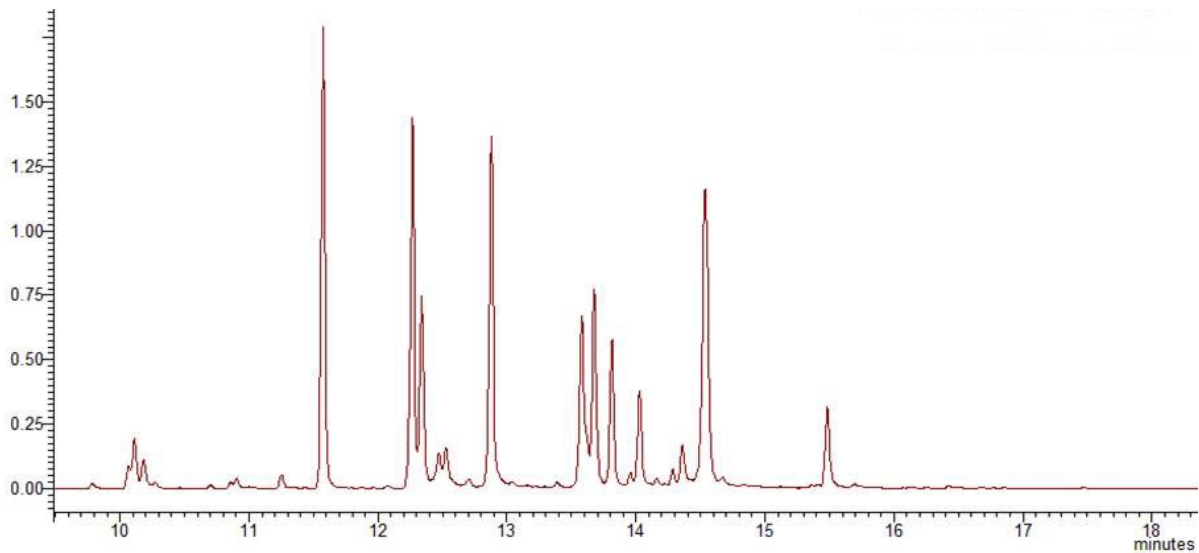


Figure 5.36 GC-MS linkage analysis of *C.difficile* Ox247 WT SLP. The samples were run on a Bruker SCION SQ 456-GC fitted with a br-5ms column. The annotation has been completed using the mass spectrometric fingerprints and fragmentation pathways of the monosaccharide structures.

Peak at 11.56 min: TRibose

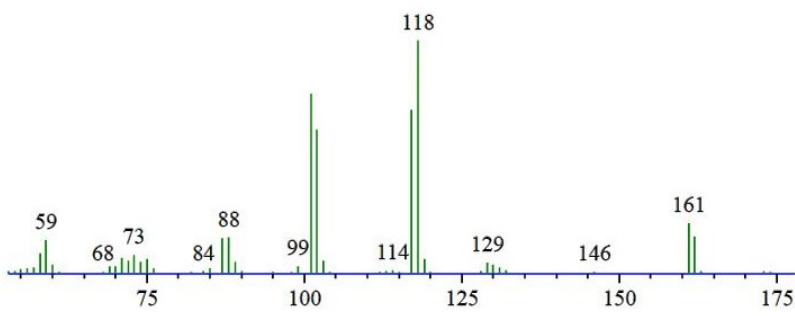
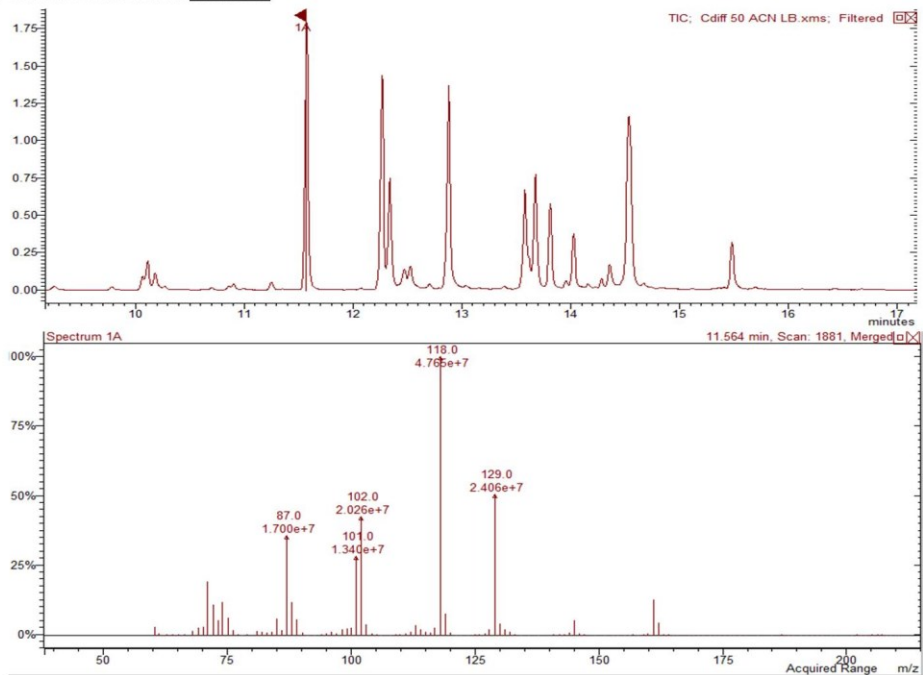


Figure 5.37 GC-MS chromatogram of linkage analysis of *C.difficile* Ox247 WT SLP (top panel), mass spectrometric fingerprint of a terminal ribose (middle panel) and mass spectrum of a terminal ribose available on the CCRC website (bottom panel).

Peak at 13.67 min: 4-linked Rhamnose

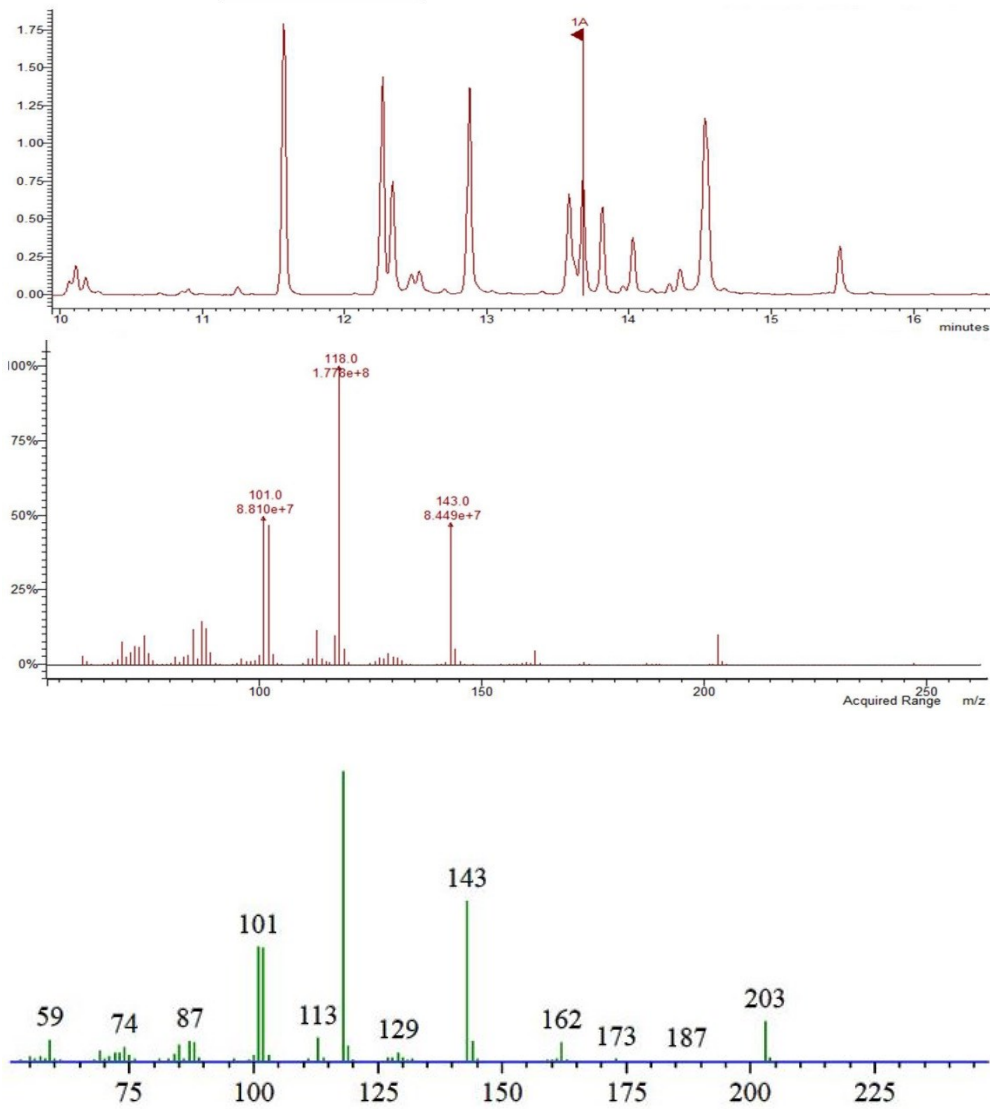


Figure 5.38 GC-MS chromatogram of linkage analysis of *C.difficile* Ox247 WT SLP (top panel), mass spectrometric fingerprint of a 4-linked rhamnose (middle panel) and mass spectrum of a 4-linked rhamnose available on the CCRC website (bottom panel).

Peak at 13.81 min: 3-linked Rhamnose

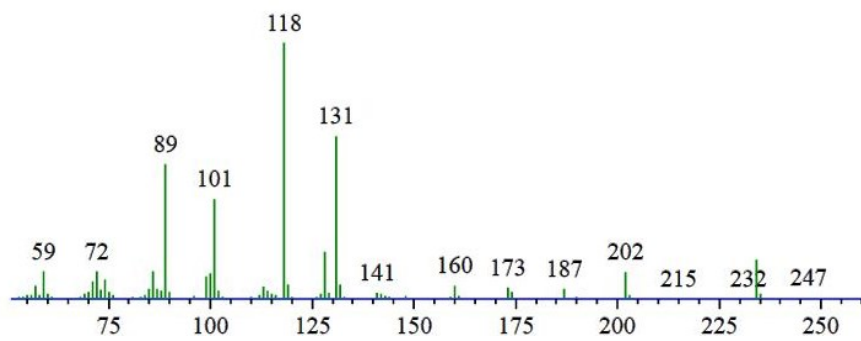
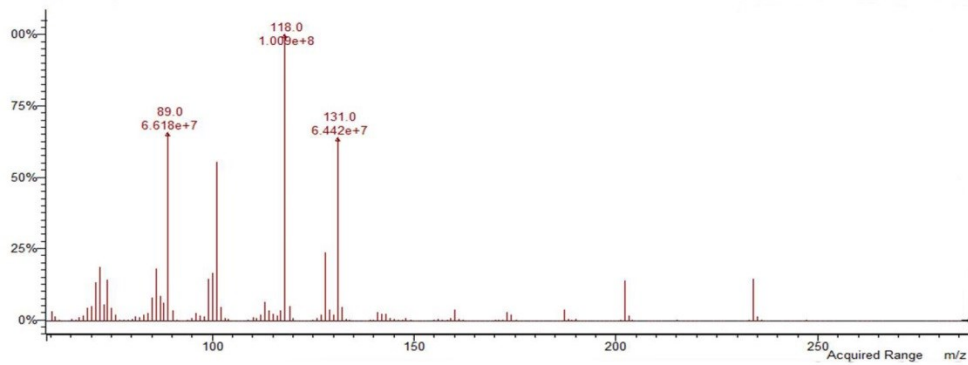
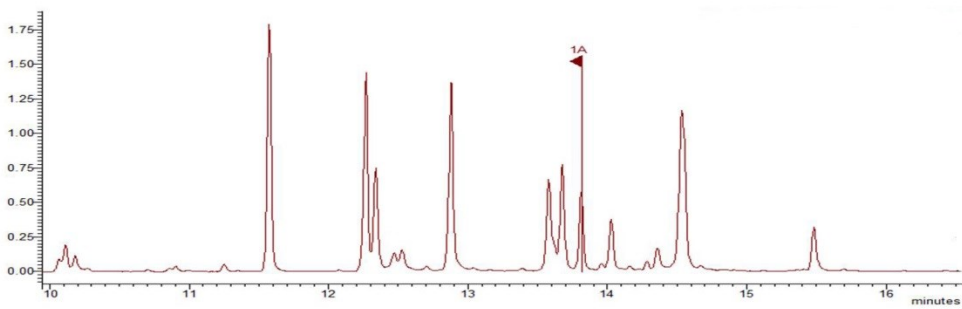


Figure 5.39 GC-MS chromatogram of linkage analysis of *C.difficile* Ox247 WT SLP (top panel), mass spectrometric fingerprint of a 3-linked rhamnose (middle panel) and mass spectrum of a 3-linked rhamnose available on the CCRC website (bottom panel).

Peak at 13.58 min: 3,4-linked Rhamnose

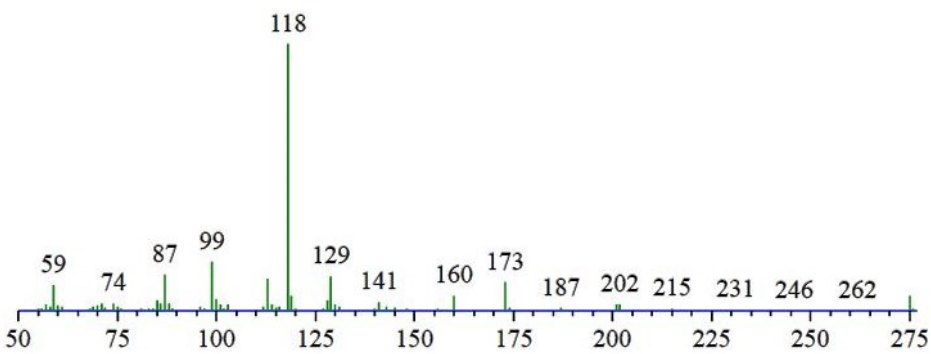
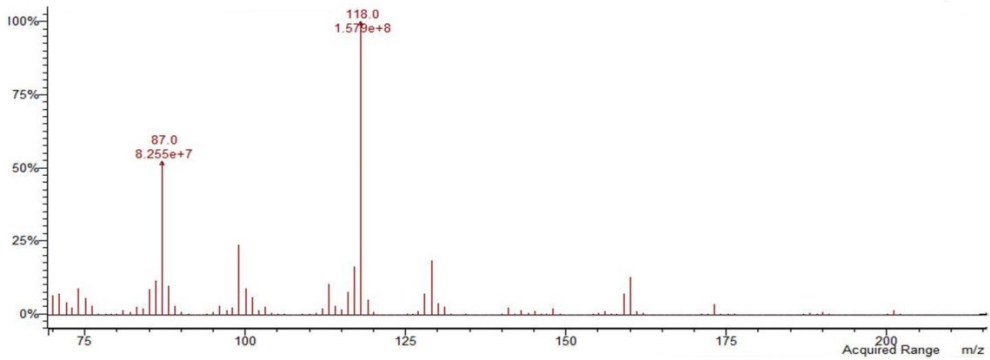
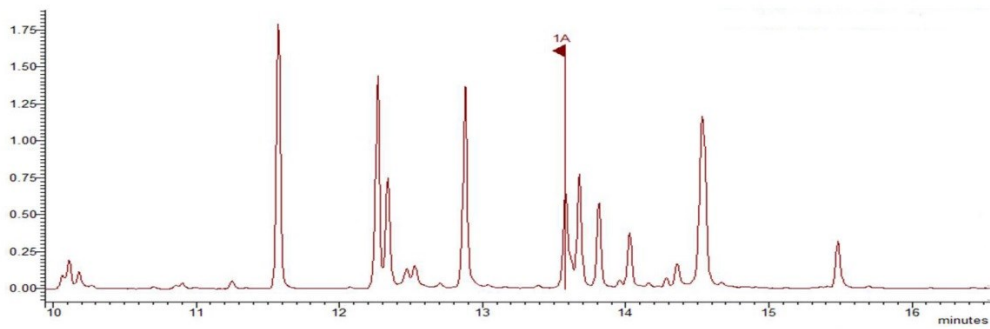


Figure 5.40 GC-MS chromatogram of linkage analysis of *C.difficile* Ox247 WT SLP (top panel), mass spectrometric fingerprint of a 3,4-linked rhamnose (middle panel) and mass spectrum of a 3,4-linked rhamnose available on the CCRC website (bottom panel).

Peak at 15.47 min: 4-linked Glucose

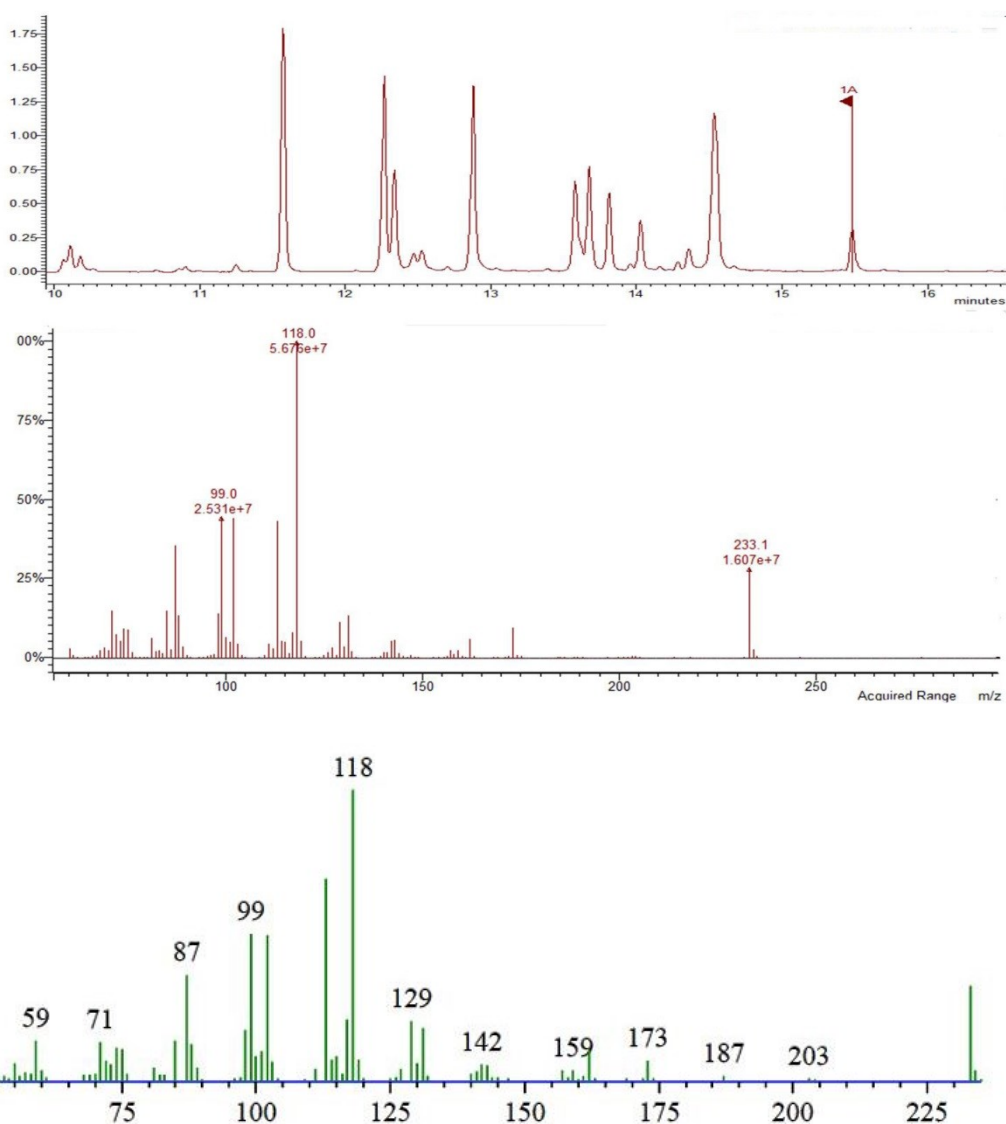


Figure 5.41 GC-MS chromatogram of linkage analysis of *C.difficile* Ox247 WT SLP (top panel), mass spectrometric fingerprint of a 4-linked glucose (middle panel) and mass spectrum of a 4-linked glucose available on the CCRC website (bottom panel).

Since the mass spectra of a number of peaks in the GC-MS trace could not be assigned as PMAAs, these data sets could not be relied upon to extend the structural analysis in terms of suggesting possibilities for the unknown portion at the non-reducing end of the molecule.

The data support the basic structural conclusions in the MS and MS/MS data sets regarding the presence of hexose, deoxyhexose and pentose units in the oligosaccharide, including the facile loss of terminal pentose branching units, but apart from the presence of some non-interpretible peaks, there were no further clues in the composition and linkage data as to the missing unit(s) needed to identify the non-reducing end structure.

5.7 NMR analysis

Further analyses were therefore now necessary, ideally via an orthogonal technique such as NMR, to complement the detailed MS studies and elucidate the final structure of this novel O-glycan chain decorating the LMW of *C.difficile* Ox247 S-layer. A summary of the O-glycan MS discoveries and structural conclusions, regarding the pentose branched oligosaccharide findings, was provided to our Canadian collaborators (Susan Logan and Evguenii Vinogradov), who specialise in the NMR analysis of bacterial oligosaccharides. Furthermore, a description of the tryptic peptide carrying the PTM on Thr-62 of peptide DILAAQNLTGAVILNK was given. This peptide component is unfortunately too large to consider for NMR of the intact glycopeptide discussed in this chapter, since the peptide resonances can overlap and confuse the NMR interpretation, and therefore our Canadian colleagues had to sub-digest using proteinase K, a very non-specific protease, which allows the digestion and removal of most of the peptide structure, sometimes leaving just a single amino acid residue attached to the oligosaccharide of interest. This, however, clearly simplifies the NMR spectra and aids interpretation.

Following the NMR analysis, Susan Logan, Evguenii Vinogradov and colleagues have suggested a structure for our novel oligosaccharide as seen in **Figure 5.42**. This suggestion is however not rigorous, particularly in the non-reducing capping residue assignment which is partly based on our own MS findings together with a negative ion spectrum produced by our Canadian colleagues. Furthermore, the positions of the postulated Phospho- and Acetyl substituents on the non-reducing terminal rhamnose ring in **Figure 5.42** are not absolutely defined in the NMR data, and nor is the peptide attachment at the reducing end of the molecule.

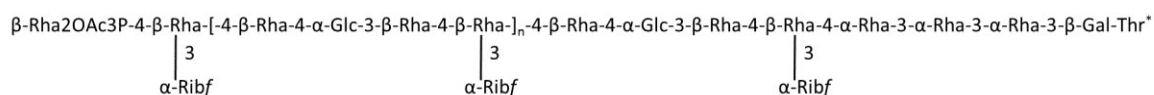


Figure 5.42 Suggested NMR structure describing sugar and linkage compositions of the O-glycan chain attached to the LMW SLP of *C.difficile* Ox247 of interest (Logan and Vinogradov personal communication). * identity of peptide unknown.

5.8 Overall Cross-correlation of the NMR and MS Data

The proteinase K digested sample used in the NMR study was then sent to our laboratory for cross-correlation experiments, including assistance in confirming the suggestions arising from the NMR study. This subsequent research is described below:

A Ribose attachment site:

The first apparent difference shown in the NMR structure above is the suggested site of attachment of the first ribose from the reducing-end of the oligosaccharide assigned in the MS experiments, described earlier in this chapter. Thus, the suggested attachment to rhamnose-4 contrasts with the structure shown in **Figure 5.6** and **5.8** where the MS fragmentation data suggest it is attached to the glucose from the reducing end. A deoxyhexose substitution position nearer to the peptide backbone would however, be consistent with the MS data on the *Orf7* mutant shown in **Figure 5.12**.

The fragmentation schematic shown in **Figure 5.43 A-B** shows the respective fragment ions predicted for the two alternative positions of substitution.

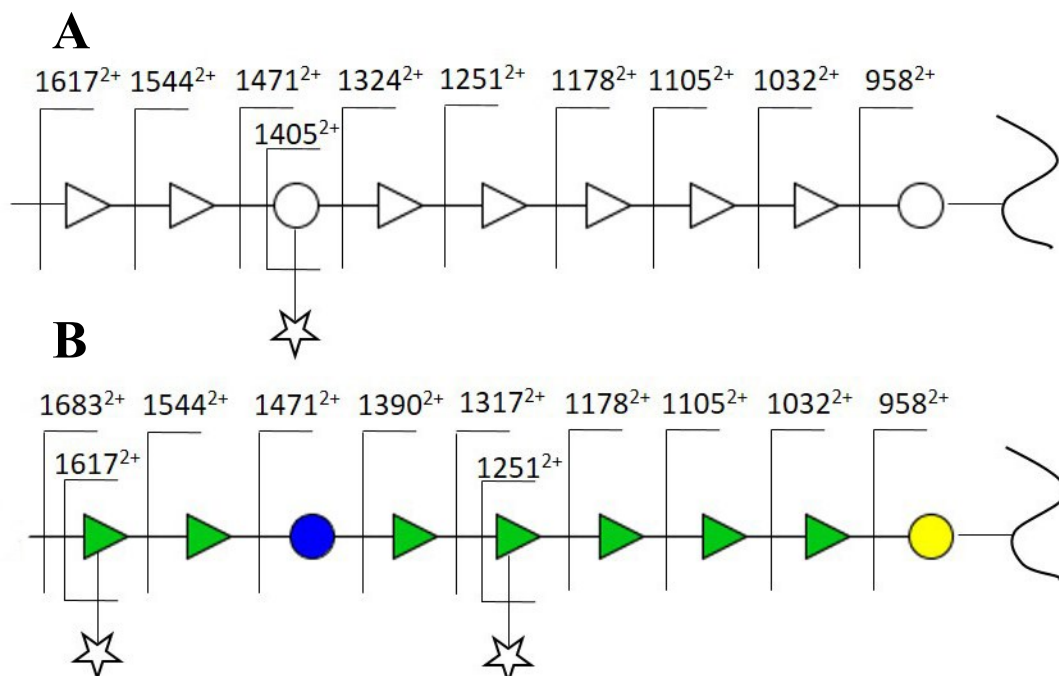


Figure 5.43 **A:** Suggested glycan structure from Q-STAR data shown in **Figure 5.6**. **B:** Suggested NMR structure. ● : Galactose; ▲ : Rhamnose; ● : Glucose; ○ : Hexose; ▷ : deoxyHexose; ☆ : Pentose.

Note that the 958^{2+} , 1032^{2+} , 1105^{2+} , 1178^{2+} and 1251^{2+} ions are common to both suggested structures. However, if the ribose is attached to Rha-4 then a doubly charged ion at m/z 1317^{2+} would be expected together with ions at m/z 1390^{2+} , 1471^{2+} , 1544^{2+} continuing along the chain to the hexose residue. However, if the labile ribose residue then falls off, fragment ions at m/z 1324^{2+} , 1405^{2+} and 1478^{2+} would be created, and these are indeed found in the spectrum in **Figure 5.6**. Also in that figure, the 1317^{2+} ion itself at the Rha-4 position is very weak compared to the m/z 1324^{2+} signal. This resulted in a suggested assignment of ribosylation at the second hexose position where the $1471^{2+}/1478^{2+}$ mass difference is obvious in the MS. From experience with other elimination mass losses, such as simple water loss from a threonine or serine residue in a peptide, the intensity of the loss can vary as fragmentation proceeds along the polymer chain, and the loss may still be observed or even be maximal a residue or two away from the elimination site. There is in fact therefore a rational explanation for the difference in interpretation, and the MS data also fit the rhamnose-4 suggestion from the NMR.

Finally, on this point, a spectrum of the glycopeptide run on a different instrument (the Xevo G2 Q-TOF) for the purpose of sample collection was re-examined to see whether a difference in ion source characteristics can affect this particular fragmentation and this spectrum is shown in **Figure 5.44**. Interestingly, in contrast to the Q-STAR spectrum in **Figure 5.6**, these data show a clear signal at m/z 1317^{2+} which would indeed correspond to ribosylation at Rha-4, and a further signal at m/z 1390^{2+} corresponding to cleavage at the next rhamnose which also then loses the ribose residue to give m/z 1324^{2+} . The following glucose residue seen at m/z 1471^{2+} and the ribose at position 4 (Rha-4) is then lost from that ion to give signal m/z 1405^{2+} .

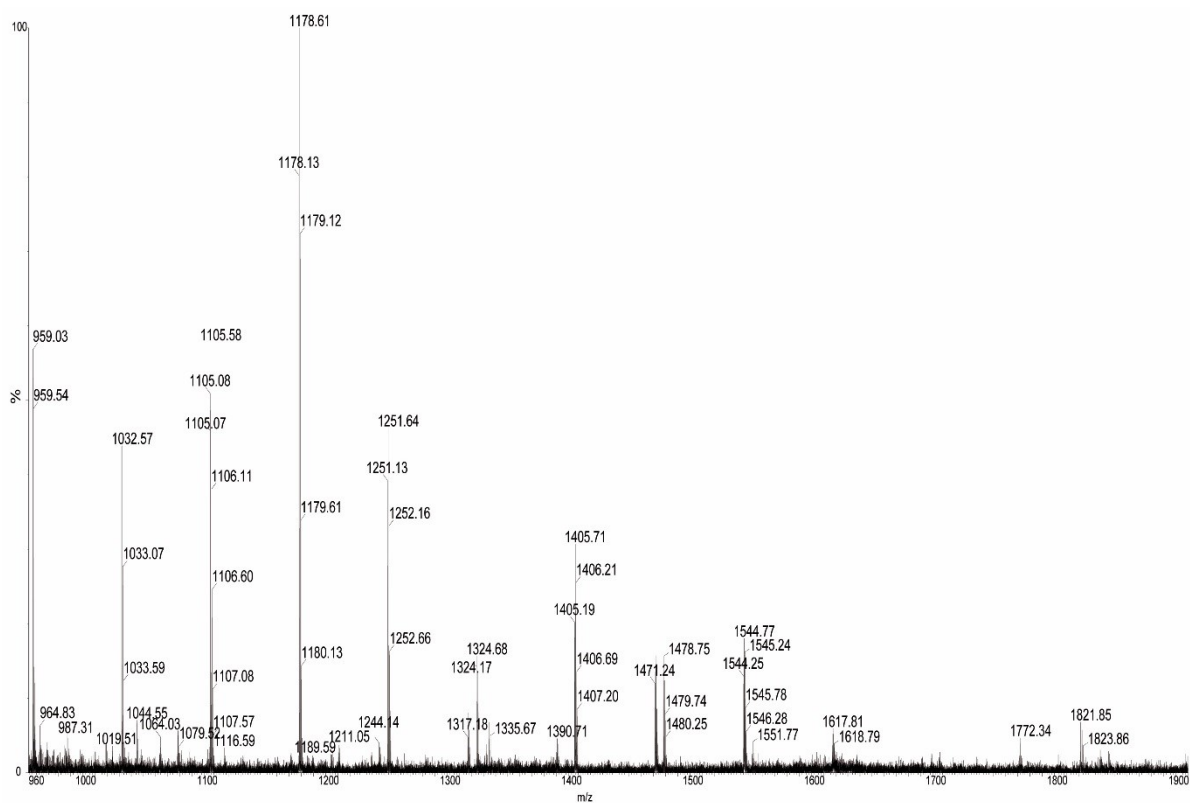


Figure 5.44 MS spectrum of the S-layer of *C.difficile* Ox247 obtained using Xevo G2 Q-TOF. This data set shows a clear signal at m/z 1317²⁺ corresponding to ribosylation at Rha-4, and at m/z 1390²⁺, corresponding to cleavage at the next rhamnose.

B Permethylated of the Glycopeptide:

The MS data produced by permethylation on the proteinase K digested sample are shown in **Figure 5.45** from the 35%-50% fraction. The intense signal at m/z 845 was chosen for further analysis by MS/MS and the spectrum produced is given in **Figure 5.46**.

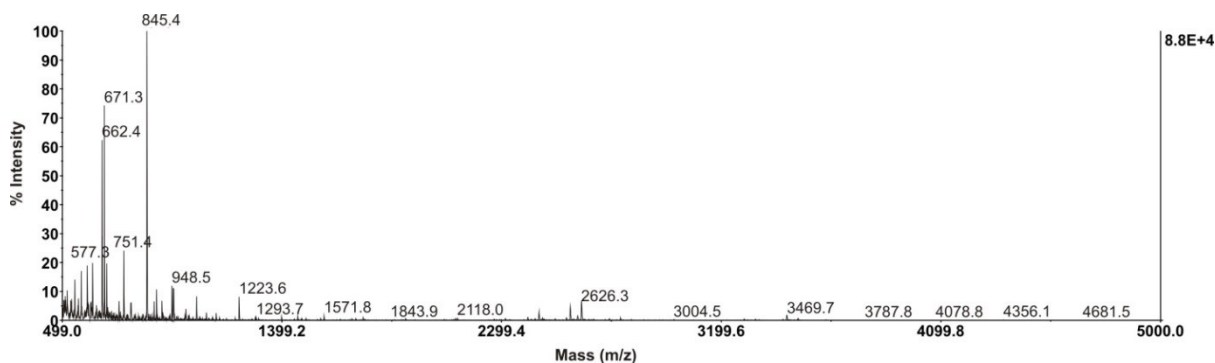


Figure 5.45 MS spectrum of the S-layer of *C.difficile* Ox247 sample after proteinase K digestion. This data set is consistent with the data shown in **Figure 5.24** and a series of peaks corresponding to the mass differences of monosaccharide residues, specifically hexoses, deoxyhexoses and pentoses, are seen all along the spectrum.

This spectrum provides good confirmation of the presence of a terminal Phospho-deoxyHex group in the molecule via fragment ions at m/z 719.4 and 563.3 which are interpreted as the loss of dimethyl phosphoric acid ($98+28=126$ Da) and loss of the terminal Phospho-deoxyHex unit via β -elimination with charge retention on the reducing end, respectively (**Figure 5.46**).

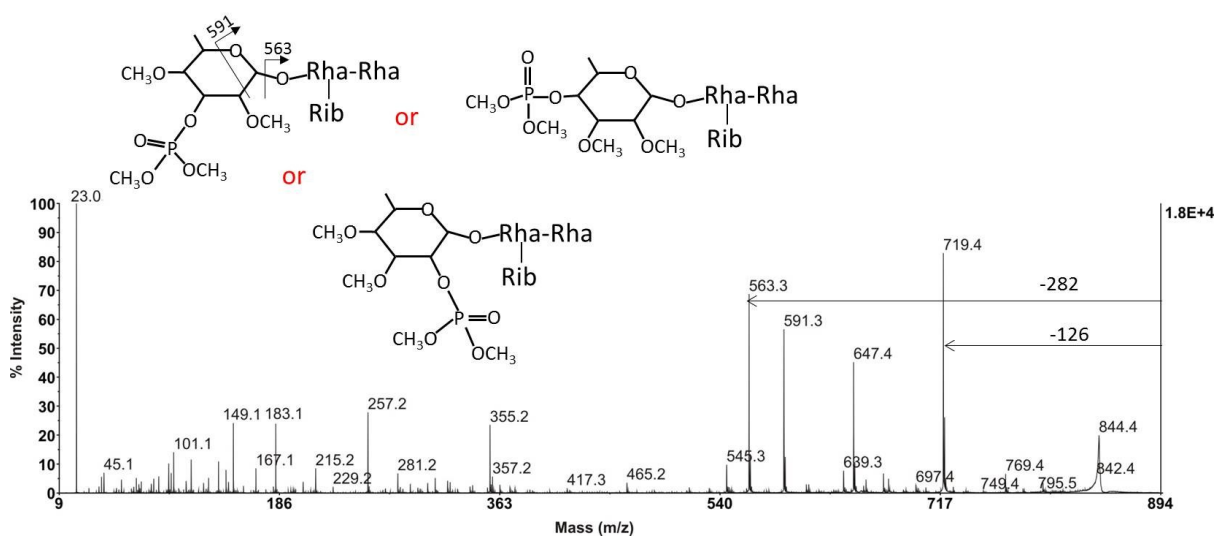


Figure 5.46 MS/MS spectrum of the m/z 845 peak seen in the MS spectrum shown in **Figure 5.44**. In this MS/MS, there is a clear loss of 126 from the molecular ion suggesting that the phosphate group has been dimethylated in the permethylation step. There are also further cross ring-type fragment ions in between the loss of 126 and the loss of 282, which corresponds to the loss of the complete capping residue.

The equivalent experiment using deuteriomethyl iodide (CD_3I) confirms (**Figure 5.47**) that the two methyl groups present on the phosphate moiety in **Figure 5.46** are from the derivatisation reaction and the native phosphate hydroxyls are free.

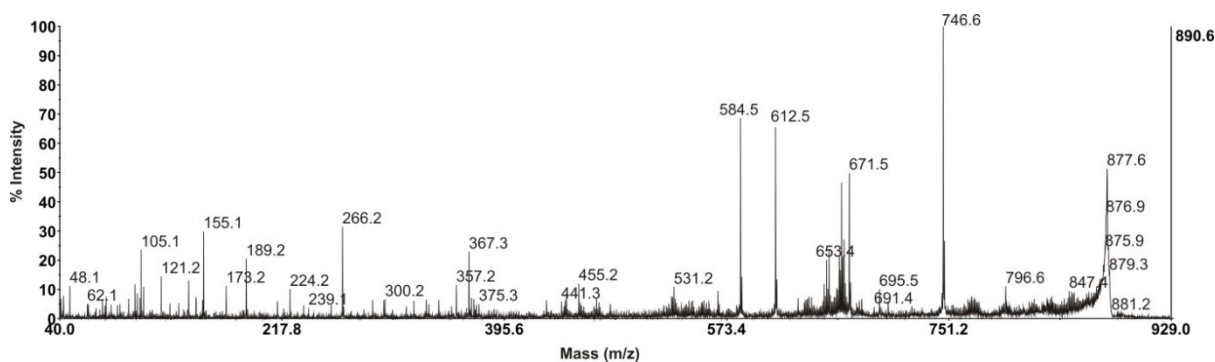


Figure 5.47 MS/MS spectrum of m/z 878, which can be correlated with the MS/MS at m/z 845 (**Figure 5.46**). There is a clear loss of 132 from the molecular ion ($878-132=746$) suggesting that the phosphate group is not naturally dimethylated or monomethylated, but the dimethyl groups have been added during the deuteropermethylation step.

NMR data are not definitive for the position of the phosphate substitution (personal communication from our Canadian colleagues), but considering the MS/MS spectrum in **Figure 5.46** can give some insights into that question by considering the deoxyHex ring structure following loss of the dimethyl phosphate fragment to produce m/z 719 and the subsequent ring cleavage to produce m/z 647. As shown in **Figure 5.48** and **5.49** for the various possible ring substitution positions for the phosphate, only position 2 or 3, and not 4, will allow the mechanistic formation of a m/z 647 ion.

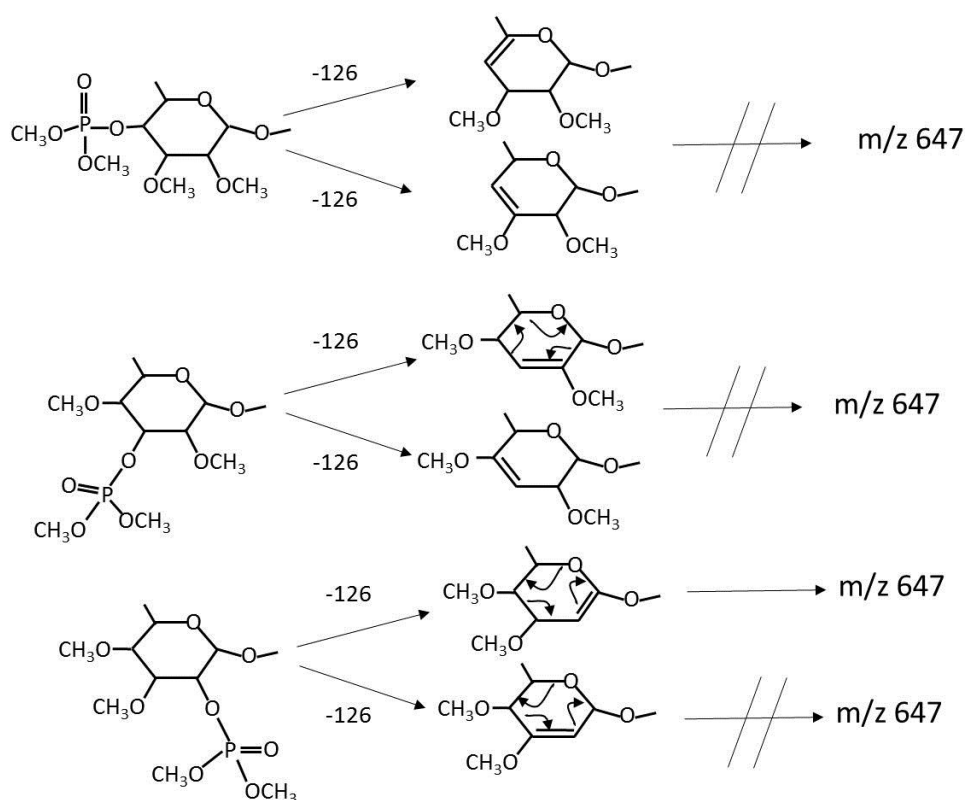


Figure 5.48 Various possible ring substitution positions for the dimethyl phosphate attached to the deoxyHex ring structure and the following loss of the dimethyl phosphate fragment (126) to produce m/z 719 and the subsequent ring cleavage to produce m/z 647. Considering the mechanism shown, only position 2 will allow the mechanistic formation of a m/z 647 ion.

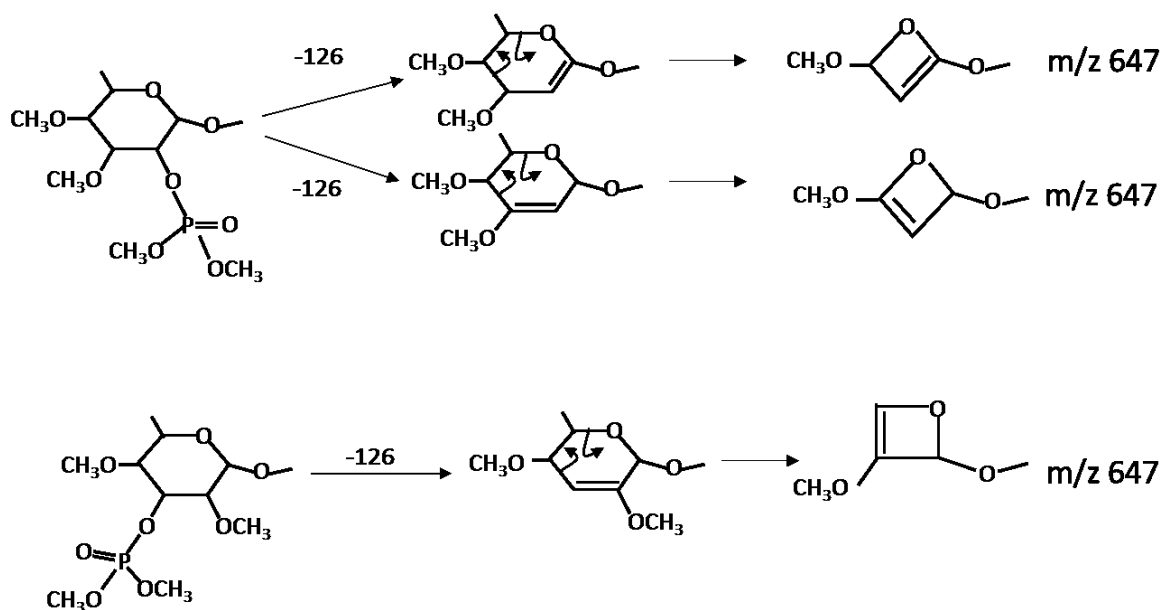


Figure 5.49 Considering a different mechanism, the dimethyl phosphate group can be attached to either carbon-2 or carbon-3 to allow the mechanistic formation of a m/z 647 ion.

Interestingly, the phosphate idea proposed in the NMR and confirmed in the above MS and MS/MS MALDI experiments now can allow a reasonable interpretation of the high mass MS/MS MALDI data of the reductively eliminated oligosaccharide described earlier in **Figure 5.27**. The main high mass signal centred at m/z 6453 then corresponds to a loss of the terminal Phospho-deoxyHex residue (282 Da) from the quasi-molecular ion, and the dimethyl Phosphoric acid loss can be seen at m/z 6609. The signal at m/z 6481 can be assigned to loss of the terminal unit less the ring-derived aldehyde group, equivalent to $283 - 29 = 254$. Tying the non-reducing and reducing ends of the oligosaccharide together using the permethylated reductively eliminated oligosaccharide MS data gives a theoretical mass (using a phosphatidyl-deoxyHex terminus, but without an acetyl group) of $6737.6 M + Na^+$ average chemical mass with a repeating unit of $n=5$ which compares well with the observed MS mass at m/z 6744, within the error of the experimental MALDI TOF method of $\pm 0.1\%$. The non-Acetylated mass has been used since it would have been expected that an O-Acetyl would be removed during the alkaline permethylation step.

C Identity of the Peptide Portion of the Proteinase K Glycopeptide:

Following the NMR analysis in Canada, Susan Logan's colleagues attempted to prove the presence of an acetyl group in the terminal Phospho-Rhamnose and produced the ES-MS spectrum shown in **Figure 5.50** by using the negative ion cone voltage fragmentation technique described in **Chapter 4**. My thesis work has included assisting in interpretation of this Canadian data.

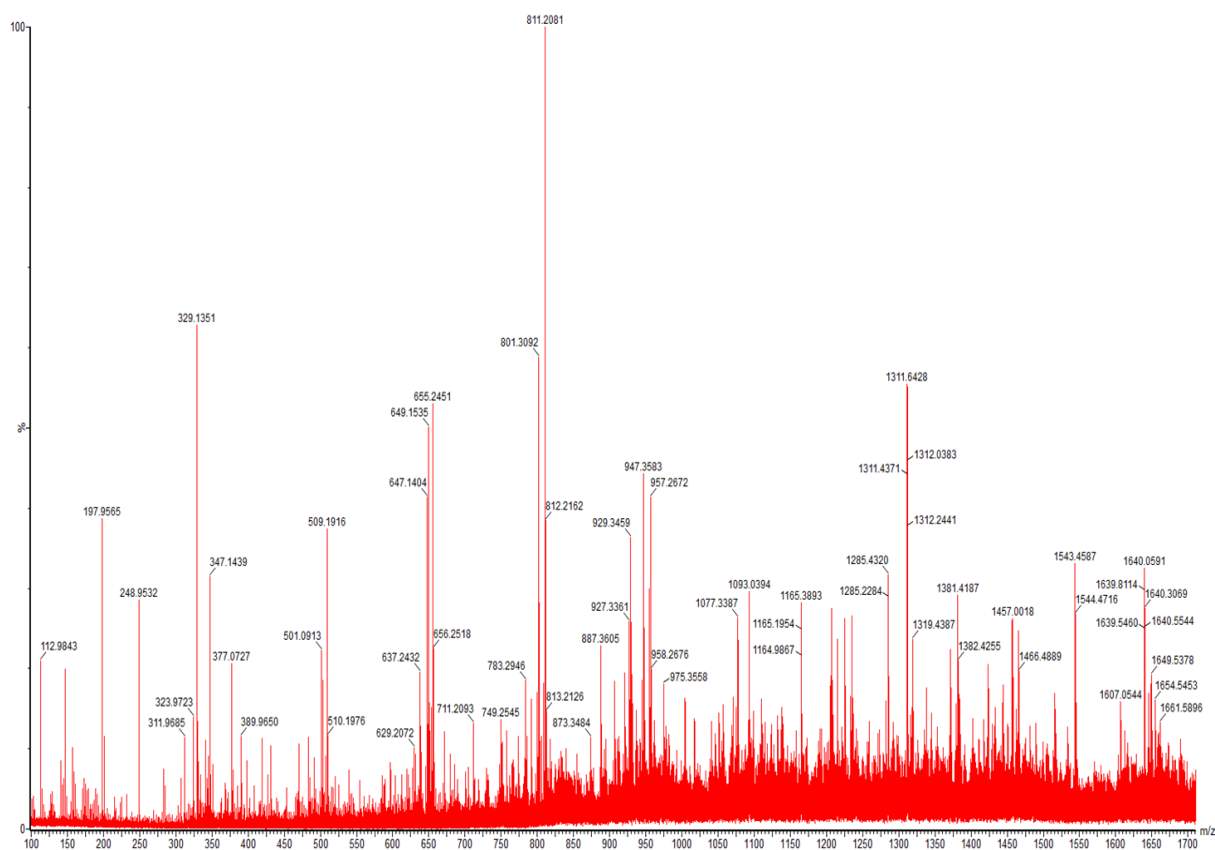
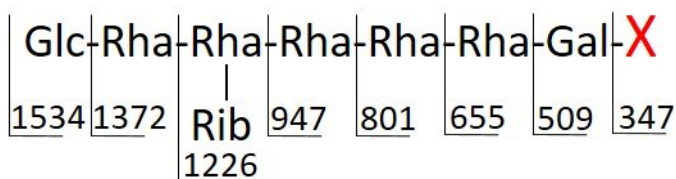


Figure 5.50 ES-MS spectrum in negative ion cone voltage fragmentation mode (spectrum from our Canadian collaborators).

My interpretation of this spectrum has revealed a series of source induced fragmentation ions namely m/z 347, 509, 655, 801, 947, 1226, 1372 and 1534 corresponding to:



These fragments can only be derived from the reducing end of the molecule because neither our earlier detailed fragmentation data nor the NMR structural conclusion finds five rhamnosides adjacent to each other, meaning that the reducing end is unique. Therefore, it follows that the m/z 347 negative ion, since it is not a sugar mass, should correspond to the terminal peptide portion of the molecule attached to the reducing end sugar. Considering the mechanism of negative ion formation this suggests that the threonine identified earlier in **section 5.2** will be present as threonine, and not de-hydro threonine in the m/z 347 ion. A screen of the DILAAQNLTGAVILNK sequence region previously shown by MS to contain the attachment site on the second threonine then allows a mass search using MassLynx which shows that only the peptide TTGA fits the 347 negative ion mass data. The glycopeptide produced by proteinase K digestion and studied by our colleagues in the Canadian NMR experiments was therefore containing this tetrapeptide, and not just the assigned threonine attachment.

D Final Structural Correlation & Conclusion:

Having now got a reasonable understanding of the reducing and non-reducing end structures, together with the repeating unit, within this novel oligosaccharide chain, it is possible to interpret the high mass positive and negative MALDI data on sample derived from the intact proteinase K digested S-layer glycopeptide.

HF Digested Proteinase K Digested Glycopeptide:

The data in **Figure 5.51** shows the results of HF digestion of the proteinase K digested S-layer glycopeptide followed by permethylation. The interpretation clearly shows the complete hydrolysis of the ribose side-chains and the loss of phosphate and acetyl together with other hydrolytic cleavages along the saccharide chain, including the peptide linkage. Signals such as that seen at m/z 6233 can be correlated to the residual $(Rha_3)-(GlcRha_3)_6-(GlcRha_5Gal)$ as the sodiated quasi-molecular ion, within the mass error of the MALDI TOF technique.

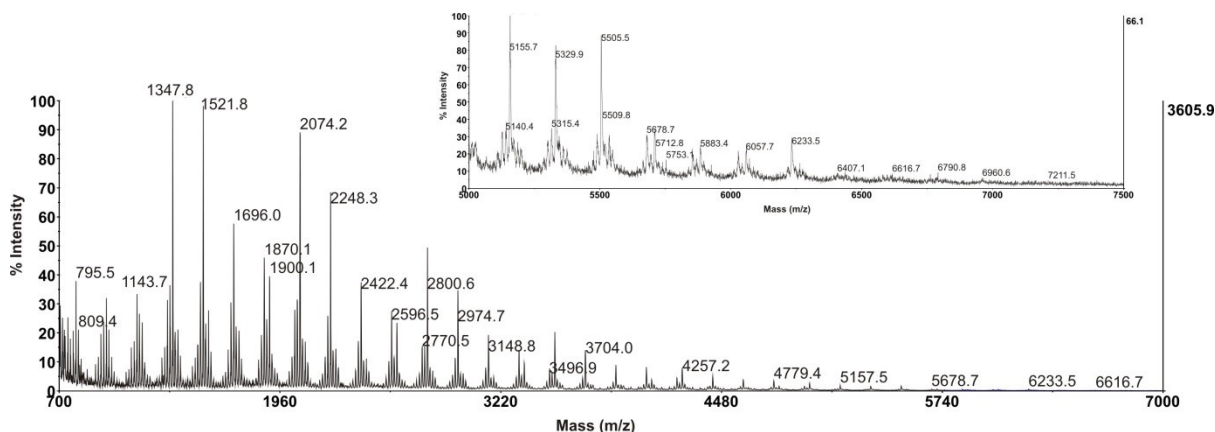


Figure 5.51 MALDI spectrum of HF digestion of the proteinase K digested S-layer glycopeptide followed by permethylation. The insert is a zooming-in of the main spectrum and specifically from m/z 5000 to 7500.

Unhydrolysed Proteinase K Digested Glycopeptide:

The negative ion MALDI TOF spectrum of the untreated glycopeptide itself is shown in **Figure 5.52** and the most significant high mass components observed are centred around m/z 6586, 6564, 6432, 5854, 5832 and 5698. In pattern recognition terms the 6586/6564 and 5854/5832 pairs probably correspond to $[M-H+Na-H]^-/[M-H]^-$ species.

The signal centred at m/z 6432 corresponds to the loss of ribose from $[M-H]^-$ 6564 which all the earlier work shown in this chapter confirms is the most sensitive sugar for elimination.

The final correlation in this molecular ion region of the spectrum is the mass difference between m/z 6564 and 5832, which is 732 Da corresponding to the repeat pentasaccharide unit (Glc-Rha-Rha/Rib-Rha) identified earlier (732.5 Da).

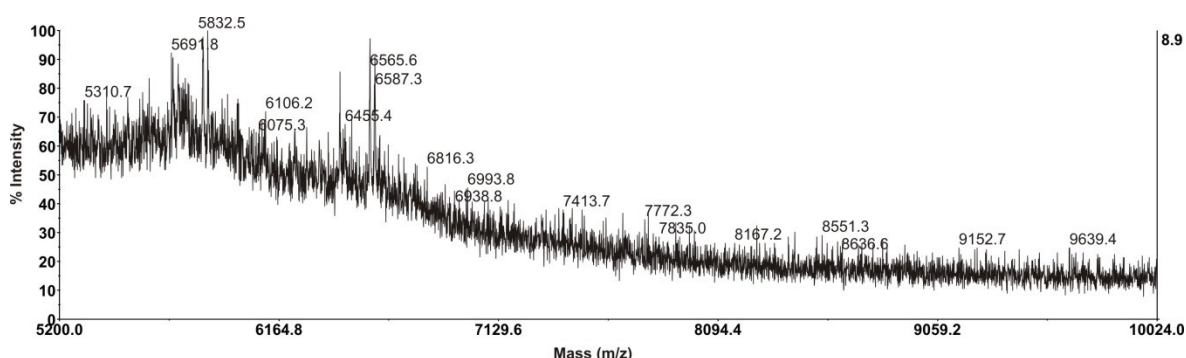


Figure 5.52 MALDI MS spectrum in negative ion mode from m/z 5000 to 10000.

The overall predicted structure of the glycopeptide combining all the MS experiments together with the NMR (without linkages or anomeric configurations) is depicted in **Figure 5.53**, where the proteinase K digested glycopeptide is highlighted in red.

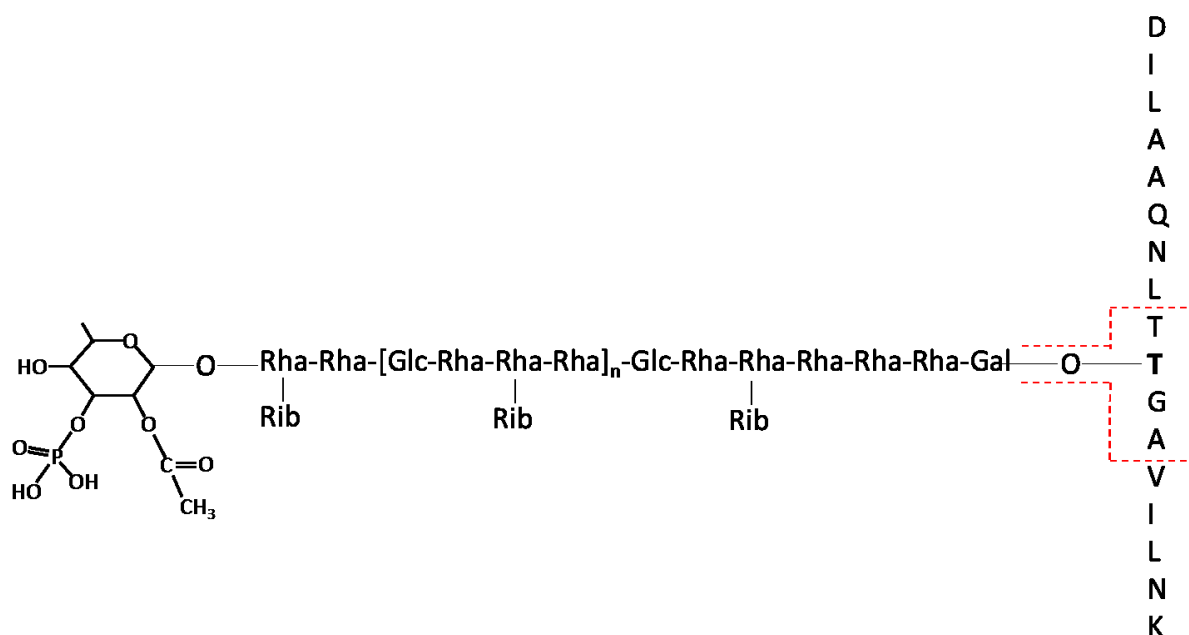


Figure 5.53 The overall predicted structure of the O-link glycopeptide decorating the S-layer of *Clostridium difficile* Ox247.

The theoretical average chemical mass of that unit calculates as 5891.6 for $n=5$ repeat unit and 6624.2 for $n=6$. The actual observed values for the presumed $[M-H]^-$ ions in **Figure 5.52** shows M 5832 and 6564 which are approximately 59 Da lower in mass than theoretical. This may be interpretable as loss of acetic acid (60 Da) by β -elimination from the predicted $[M-H]^-$ ions. Whilst this may be unexpected for low internal energy MALDI MS, without the availability of synthetic standards for the structure proposed, it cannot be discounted since perhaps the phosphate substitution is able to stabilise the new ring-double bond created by acetic acid elimination.

High Mass Q-STAR Tryptic Digestion Data:

Re-examining the Q-STAR spectrum in which the glycopeptide was initially discovered (**Figure 5.6**), and with the benefit of hindsight, now looking for much larger structures, it is possible to see very weak data around the 2000 position in the m/z scale, as shown in **Figure 5.54**.

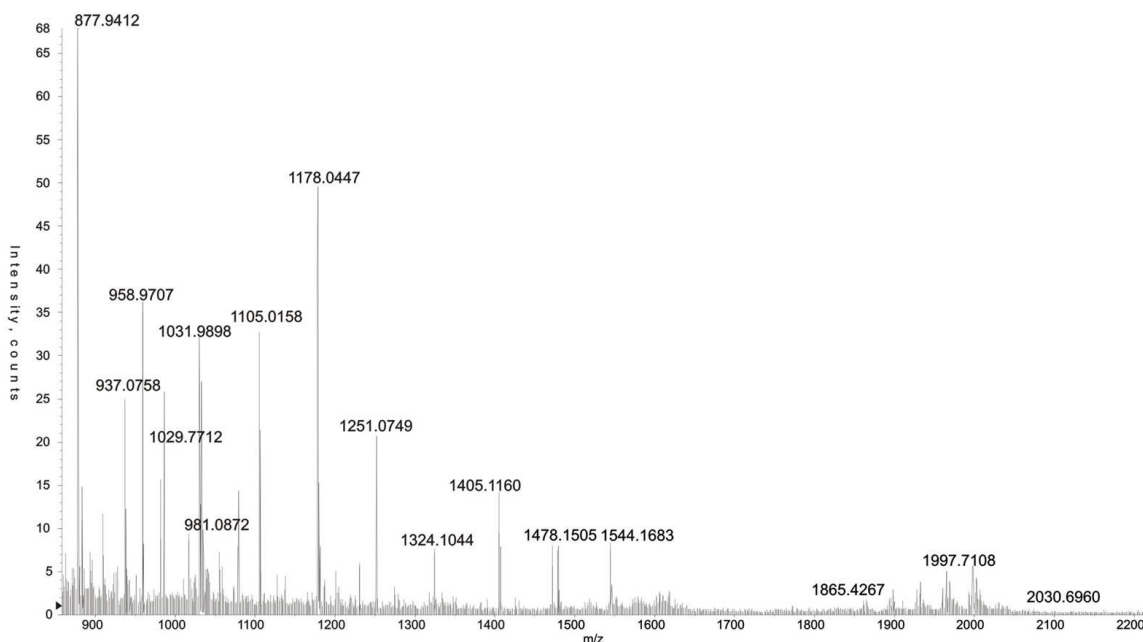


Figure 5.54 Q-STAR MS spectrum of the S-layer of *Clostridium difficile* Ox247 showing weak data around the 2000 position in the m/z scale.

These data are not strong enough to allow charge state determination by measuring the width between isotope signals, but by estimating the charge state and looking for the $z+1$ and $z-1$ variants, it is possible to conclude that the group of signals shown carry 4 charges. The principal signal group centred at m/z 1997.71 calculates for mass = 7986.84 Da. This figure is bigger than those seen in the MALDI data above, but if it is assumed to be $n = 8$, then $n=6$ would be 6520. The problem with the ES data in such a calculation is that there may have been enough internal energy in the ion to lose a number of riboses and/or phosphate or presumed acetate, making it less definitive in an overall mass calculation on such a novel glycopeptide.

Final Considerations:

Allowing for the above discrepancies between the MS and NMR data, which may be explicable as described, another observation of note is that the MALDI MS data sets on detailed analysis show weaker but nevertheless real satellite signals at higher masses than the principal “most intense” signals used in the data interpretation here.

These “satellites” would correspond to additions of 1, 2 and 3 ribose units on top of the structure concluded in **Figure 5.53**. The positions and extent of these substitutions are

currently unknown, and may be the subject of future work to more fully define the intricacies of this exciting novel glycopeptide structure.

This new structure is yet another example to add to the expanding list of S-layer glycoproteins found in this domain of life. In fact, so far approximately 40 different S-layer glycoprotein glycan structures have been fully or at least partially described and the glycan chain decorating the S-layer of *C.difficile* Ox247 presents similar aspects with these glycopeptide structures. Thus, only O-glycosidic linkages have been identified in both Gram-positive and Gram-negative bacterial species and these may involve serine, threonine or tyrosine residues (Zarschler 2010). The majority of the S-layer chains characterised are also linear or branched homo- or heterosaccharides, which comprise 20-50 identical repeating units. A minority, which include some lactic acid bacteria and the Gram-negative bacterium *Tannerella forsyntia*, have S-layer glycans which are short oligosaccharides without repeats similar to archaeal glycans. A wide spectrum of carbohydrates is involved in these glycan chains: neutral hexoses, 6-deoxyhexoses, amino sugars, uronic acid as well as rare sugars, i.e. α -D-Rhap, α -D-Fucp, α -D-FucpNAc, β -D-QuipNAc, D-glycero- β -D-manno-Hepp, β -D-ManpNAcA or Pse5Am7Gc and some of them are typical components of LPS O-antigens in Gram-negative bacteria (Messner, Schaffer & Kosma 2013). L-rhamnose is the most prevalent carbohydrate residue in the majority of glycans and in particular the L-rhamnose bound to a mannopyranosyl unit has been often observed as a component of linear and branched repeating units in glycans of *Thermoanaerobacter Thermohydrosulfuricus* strains (Bock et al 7137-7144 1994) and *Thermoanaerobacterium Thermosaccharolyticum* (Schaffer et al 5482-5492 2000), respectively. Frequently, non-carbohydrate substituents are used as capping of the terminal sugar residue at the non-reducing end of long S-layer glycan chain. These capping structures are believed to be implicated in chain-length determination of the polysaccharide during biosynthesis (Ristl 2011). For example, 2-O-methyl groups end the S-layer glycans of *G.stearothermophilus* (Schaffer 6230-6239 2002) or *Thb.Thermohydrosulfuricus* (Bock et al 7137-7144 1994). In fact, even in the case presented in this chapter, it is possible to believe that phospho- and acetyl substituents on the non-reducing terminal rhamnose ring of the glycopeptide attached to the S-layer of *C.difficile* Ox247 is implicated in the chain-length determination as the presence of different polymer lengths (see **Figures 5.25, 5.27 and 5.28**).

The highly variable S-layer glycoproteins, with the presence of possible terminal groups, can be seen as the Gram-positive equivalents of lipopolysaccharides, with the protein component substituting for the lipid A of Gram-negative bacteria (Raetz & Whitfield 635-700 2002).

So far most of the Gram-positive bacteria S-layer glycoproteins analysed share a common tripartite building plan, comprising the elongated glycan chain, an adapter region, a short oligosaccharide responsible for linking the entire S-layer glycan to the S-layer polypeptide backbone, and, finally, the glycosidic linkage to the S-layer proteins. Usually, S-layer proteins are multiply glycosylated and the linkage sugar is in the β -anomeric configuration in all the structures examined to date (Messner, Schaffer & Kosma 2013). In contrast, *C.difficile* Ox247 S-layer presents a single glycosylation site and specifically Thr-62 (see **Figure 5.19**). Moreover, based on the results obtained and shown in this chapter, it is now possible to correlate them with the bioinformatics predictions (see **section 5.1**). In fact, analysing several insertional mutants created by our collaborators (*Orf2*, *Orf3*, *Orf4*, *Orf7* and *Orf16*), it is possible to illustrate the putative functions of each gene within the S-layer cassette number 11. Specifically, *Orf2* has been confirmed to be a glycan biosynthesis initiation gene; *Orf3*, which has been predicted to be a rhamnosyltransferase responsible for the addition of the first deoxysugar onto the core hexose by an α 1,3 glycosidic bond, has been actually confirmed to be a rhamnosyltransferase but responsible for the addition of one or more deoxyhexoses but not the one attached to the peptide-linked galactose. The data obtained by the analysis of the *Orf4* mutant are consistent with the bioinformatics prediction that *Orf4* corresponds to a putative phosphoribose diphosphate decaprenyl-phosphate phosphoribosyltransferase. The absence of pentose in the observed glycans is consistent with *Orf4* being responsible for the transfer of a pentose sugar to the lipid-linked glycan precursor. The glycan profile obtained by the *Orf7* knock-out corresponds to dHex₅HexPent and this gene is likely to be the one responsible for the addition of a glucose to a rhamnose. Finally, *Orf16* is one of the five genes encoded which constitute the rhamnose biosynthetic pathway (*Orfs 13, 14, 16, 17, 18*) and was predicted to be the first enzyme, *RmlA*, responsible for the conversion of glucose-1-P to TDP-Glc. However, from the data obtained, *Orf16* looks likely to be a rhamnosyltransferase responsible for adding a rhamnose to another rhamnose and specifically via a β 3,4 glycosidic bond.

S-layer glycosylation is the best-studied example of glycosylated prokaryotic organisms, beside the fact it was the first glycoprotein detected in this domain of life (Mescher & Strominger 1976; Sleytr & Thorne 1976). Even though glycosylation is the major modification of S-layer proteins, not necessarily all prokaryotic S-layer proteins are glycosylated, but the presence of this specific PTM may be advantageous for the survival of the organisms in their natural competitive habitats, because it contributes to an enormous diversification of the prokaryotic cell surface (Messner, Schaffer & Kosma 2013) and

therefore it can be then considered as a possible target for colonisation-preventing vaccines (see chapter 6).

Chapter 6:
Conclusion

6. Conclusion

The goal of this PhD work was to address challenging structural questions related to several glycoproteins, belonging to both eukaryotes and prokaryotes, employing a wide spectrum of high sensitivity mass spectrometric methodologies in order to obtain further understanding of both mammalian and bacterial unusual glycosylation.

Appropriate structural analysis methods have been exploited for glycoproteomic problems with the focus on ADAMTS13, a disintegrin and metalloprotease multi-domain glycoprotein implicated in the regulation of thrombogenesis and the *Clostridium difficile* flagellin and S-layer glycoproteins which are involved in the host-pathogen interaction. The common theme linking these three projects consists in the novel structural discoveries obtained as a consequence of the progress in MS-based approaches. Thus, having the opportunity to use several high resolution mass spectrometric instrumentations with different and specific characteristics, it was possible to address biologically important questions which had been unresolved to date. For example, the Xevo G2 Q-TOF, with its capacity to collect the maximum amount of data from a single analysis and/or using picomoles of materials, was fundamental in the analysis of ADAMTS13 and its variants. Moreover, the Synapt G2-S mass spectrometer, characterised by presenting high-sensitivity and unrivalled selectivity and analytical peak capacity together with the powerful ETD capability, was crucial in allowing the discovery of a unique terminal Taurine (aminoethyl-sulphonic acid) peptidylamidoglycan unit and the determination of the single-site glycan attachment in *C.difficile* flagellin and S-layer, respectively. Another challenging aspect of this PhD work to overcome was the quality and/or quantity of materials available. In fact, different methods were applied and sometimes even developed in order to gain more structural insight into the nature of the glycan chains under study.

In particular, considering the *Clostridium difficile* S-layer project, further investigations would be necessary to better address the complexities of this novel and unusual glycopeptide structure. For example more experiments can be done to define the positions and extent of the “satellite ribose units” in the structure concluded in Chapter 5; moreover, the commissioning of new synthetic standards (which will require further funding) would be helpful to achieve a clear understanding of the presence/absence of the O-Acetyl substitution at the non-reducing end terminal rhamnose. Finally, the coupling of the mass spectrometric strategies described

here with the possible use of specific glucosidases or rhamnosidases able to target and isolate the terminal capping of this exciting novel glyco-structure would be very effective.

These novel O-glycome findings, shown in chapters 3, 4 and 5, may be introduced in future research in various ways. For example, one could be into expression constructs for the purpose to increase protein expression in order to achieve large scale production of therapeutic proteins or development of gene delivery technologies (ADAMTS13). Another way may be in helping understand the mechanisms of bacterial glycosylation and its role in pathogenesis underpinning advances in the design of new glycoconjugate vaccines, as in the case of the nosocomial pathogen *Clostridium difficile*. In fact, in the last 20 years the number of hospital stays associated with *C.difficile* in the USA has triples and close to 10% of these CDIs result in the death of the patient (Calo, Kaminski et al. 2010, Lucado 2012), in addition to the fact that *C.difficile* can persist indefinitely in a robust sporulated form, which can be easily transferred from person to person in the hospital environment and which is recalcitrant to many standard cleaning regimens (Gerding, Muto et al. 2008). As commonly used frontline antibiotics are less and less effective, alternative treatment options and particularly preventive measures such as vaccines are becoming more attractive (Kelly and LaMont 2008) to fight *C.difficile* associated disease. In fact, since protein glycosylation is a high-potential target for clinical intervention to help combat the continuing threat of infectious disease and also the clear interest from the biotechnology industry to develop a vaccine against this hospital-acquired pathogen, vaccination with surface antigens may inhibit colonisation, reduce the reservoir, and thus limit recurrence more efficiently than toxin-neutralising approaches. Therefore, *C.difficile* surface glycans are potential targets for colonisation-preventing vaccines, due to the well-known role of glycans in mucosal bacterial adhesion (Moxon 1990). In fact, it is possible that the use of native glycoproteins as subunit vaccines may reduce the cost and effort associated with future vaccine development, together with the development of antimicrobial alternatives to antibiotics to address the increasing threat of drug and multidrug resistant bacteria. Moreover, the presence of unique carbohydrate structures decorating the surface proteins of many pathogenic bacteria, including *C.difficile* as found in this research, will be exploited so to develop glycoprotein-based diagnostics and also, the increased knowledge of the diversity of bacterial glycoproteins has the potential to lead to strain-specific differentiation founded upon carbohydrate or protein-carbohydrate epitopes with a resulting decrease of the burden of bacterial diseases (Fulton, Smith et al. 2016).

Chapter 7:
References

Abu-Qarn, M., J. Eichler and N. Sharon (2008). "Not just for Eukarya anymore: protein glycosylation in Bacteria and Archaea." Current Opinion in Structural Biology **18**(5): 544-550.

Abu-Qarn, M., S. Yurist-Doutsch, A. Giordano, A. Trauner, H. R. Morris, P. Hitchen, O. Medalia, A. Dell and J. Eichler (2007). "Haloferax volcanii AglB and AglD are Involved in N-glycosylation of the S-layer Glycoprotein and Proper Assembly of the Surface Layer." Journal of Molecular Biology **374**(5): 1224-1236.

Akiyama, M., D. Nakayama, S. Takeda, K. Kokame, J. Takagi and T. Miyata (2013). "Crystal structure and enzymatic activity of an ADAMTS-13 mutant with the East Asian-specific P475S polymorphism." Journal of Thrombosis and Haemostasis **11**(7): 1399-1406.

Aldridge, P. and K. T. Hughes (2002). "Regulation of flagellar assembly." Current Opinion in Microbiology **5**(2): 160-165.

Alemka, A., H. Nothaft, J. Zheng and C. M. Szymanski (2013). "N-Glycosylation of *Campylobacter jejuni* Surface Proteins Promotes Bacterial Fitness." Infection and Immunity **81**(5): 1674-1682.

Anderson, J. K., T. G. Smith and T. R. Hoover (2010). "Sense and sensibility: flagellum-mediated gene regulation." Trends in microbiology **18**(1): 30.

Ando, S., K. Kon, Y. Nagai and T. Murata (1977). "Chemical Ionization and Electron Impact Mass Spectra of Oligosaccharides Derived from Sphingoglycolipids." The Journal of Biochemistry **82**(6): 1623-1631.

Angov, E. (2011). "Codon usage: Nature's roadmap to expression and folding of proteins." Biotechnology Journal **6**(6): 650-659.

Apweiler, R., H. Hermjakob and N. Sharon (1999). "On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database1." Biochimica et Biophysica Acta (BBA) - General Subjects **1473**(1): 4-8.

Asakura, H., Y. Churin, B. Bauer, J. P. Boettcher, S. Bartfeld, N. Hashii, N. Kawasaki, H. J. Mollenkopf, P. R. Jungblut, V. Brinkmann and T. F. Meyer (2010). "Helicobacter pylori HP0518 affects flagellin glycosylation to alter bacterial motility." Molecular Microbiology **78**(5): 1130-1144.

Aubry, A., G. Hussack, W. Chen, R. KuoLee, S. M. Twine, K. M. Fulton, S. Foote, C. D. Carrillo, J. Tanha and S. M. Logan (2012). "Modulation of Toxin Production by the Flagellar Regulon in *Clostridium difficile*." Infection and Immunity **80**(10): 3521-3532.

Baban, S. T., S. A. Kuehne, A. Barketi-Klai, S. T. Cartman, M. L. Kelly, K. R. Hardie, I. Kansau, A. Collignon and N. P. Minton (2013). "The Role of Flagella in *Clostridium difficile* Pathogenesis: Comparison between a Non-Epidemic and an Epidemic Strain." PLoS ONE **8**(9): e73026.

Barber, M., R. S. Bordoli, G. V. Garner, D. B. Gordon, R. D. Sedgwick, L. W. Tetler and A. N. Tyler (1981). "Fast-atom-bombardment mass spectra of enkephalins." Biochemical Journal **197**(2): 401-404.

Bell, S. D. and S. P. Jackson (2001). "Mechanism and regulation of transcription in archaea." Current Opinion in Microbiology **4**(2): 208-213.

Bevins, C. L. and N. H. Salzman (2011). "Paneth cells, antimicrobial peptides and maintenance of intestinal homeostasis." Nat Rev Micro **9**(5): 356-368.

Billker, O., V. Lindo, M. Panico, A. E. Etienne, T. Paxton, A. Dell, M. Rogers, R. E. Sinden and H. R. Morris (1998). "Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito." Nature **392**(6673): 289-292.

Bjoern, S., D. C. Foster, L. Thim, F. C. Wiberg, M. Christensen, Y. Komiyama, A. H. Pedersen and W. Kisiel (1991). "Human plasma and recombinant factor VII. Characterization of O-glycosylations at serine residues 52 and 60 and effects of site-directed mutagenesis of serine 52 to alanine." Journal of Biological Chemistry **266**(17): 11051-11057.

Björndal, H., Hellerqvist, C.G., Lindberg, B., Svensson, S. (1970). "Gas-liquid chromatography and mass spectrometry in methylation analysis of polysaccharides." Angew. Chem. Internat. **9**(8): 610-619.

Bouché, L., M. Panico, P. Hitchen, D. Binet, F. Sastre, A. Faulds-Pain, E. Valiente, E. Vinogradov, A. Aubry, K. Fulton, S. Twine, S. M. Logan, B. W. Wren, A. Dell and H. R. Morris (2016). "The Type B Flagellin of Hypervirulent *Clostridium difficile* Is Modified with Novel Sulfonated Peptidylamido-glycans." The Journal of Biological Chemistry **291**(49): 25439-25449.

Brouwer, M. S. M., E. Allan, P. Mullany and A. P. Roberts (2012). "Draft Genome Sequence of the Nontoxicogenic *Clostridium difficile* Strain CD37." Journal of Bacteriology **194**(8): 2125-2126.

Buko, A. M., E. J. Kentzer, A. Petros, G. Menon, E. R. Zuiderweg and V. K. Sarin (1991). "Characterization of a posttranslational fucosylation in the growth factor domain of urinary plasminogen activator." Proceedings of the National Academy of Sciences of the United States of America **88**(9): 3992-3996.

Burda, P. and M. Aebi (1999). "The dolichol pathway of N-linked glycosylation." Biochimica et Biophysica Acta (BBA) - General Subjects **1426**(2): 239-257.

Cal, S., A. J. Obaya, M. a. Llamazares, C. Garabaya, V. c. Quesada and C. López-Otín (2002). "Cloning, expression analysis, and structural characterization of seven novel human ADAMTSs, a family of metalloproteinases with disintegrin and thrombospondin-1 domains." Gene **283**(1-2): 49-62.

Calabi, E., F. Calabi, A. D. Phillips and N. F. Fairweather (2002). "Binding of *Clostridium difficile* surface layer proteins to gastrointestinal tissues." Infect Immun **70**(10): 5770-5778.

Calabi, E., S. Ward, B. Wren, T. Paxton, M. Panico, H. Morris, A. Dell, G. Dougan and N. Fairweather (2001). "Molecular characterization of the surface layer proteins from *Clostridium difficile*." Molecular Microbiology **40**(5): 1187-1199.

Calo, D., Z. Guan, S. Naparstek and J. Eichler (2011). "Different routes to the same ending: comparing the N-glycosylation processes of *Haloferax volcanii* and *Haloarcula marismortui*, two halophilic archaea from the Dead Sea." Molecular microbiology **81**(5): 1166-1177.

Calo, D., L. Kaminski and J. Eichler (2010). "Protein glycosylation in Archaea: Sweet and extremeD Calo et al." Glycobiology **20**(9): 1065-1076.

Castric, P. (1995). "pilO, a gene required for glycosylation of *Pseudomonas aeruginosa* 1244 pilin." Microbiology **141**(5): 1247-1254.

Castric, P., F. J. Cassels and R. W. Carlson (2001). "Structural Characterization of the *Pseudomonas aeruginosa* 1244 Pilin Glycan." Journal of Biological Chemistry **276**(28): 26479-26485.

Cerquetti, M., A. Molinari, A. Sebastianelli, M. Diociaiuti, R. Petruzzelli, C. Capo and P. Mastrantonio (2000). "Characterization of surface layer proteins from different *Clostridium difficile* clinical isolates." Microb Pathog **28**(6): 363-372.

Chen, M. M., K. J. Glover and B. Imperiali (2007). "From Peptide to Protein: Comparative Analysis of the Substrate Specificity of N-Linked Glycosylation in *C. jejuni*." Biochemistry **46**(18): 5579-5585.

Chou, M. W. Y., S.K. (1979). "Combined reversed-phase and normal-phase high-performance liquid chromatography in the purification and identification of 7,12-dimethylbenz[a]anthracene metabolites." J Chromatogr **20**(185): 635-654.

Chung, S., S.-H. Shin, C. R. Bertozzi and J. J. De Yoreo (2010). "Self-catalyzed growth of S layers via an amorphous-to-crystalline transition limited by folding kinetics." Proceedings of the National Academy of Sciences **107**(38): 16536-16541.

Dang, T. H. T., L. d. I. Riva, R. P. Fagan, E. M. Storck, W. P. Heal, C. Janoir, N. F. Fairweather and E. W. Tate (2010). "Chemical Probes of Surface Layer Biogenesis in *Clostridium difficile*." ACS Chemical Biology **5**(3): 279-285.

Darling, A. E., P. Worden, T. A. Chapman, P. R. Chowdhury, I. G. Charles and S. P. Djordjevic (2014). "The genome of *Clostridium difficile* 5.3." Gut Pathogens **6**: 4-4.

De Laeter, J., Kurz, M.D. (2006). "Alfred Nier and the sector field mass spectrometer." J Mass Spectrom **41**(7): 847-854.

Dell, A. (1987). "F.A.B.-mass spectrometry of carbohydrates." Adv Carbohydr Chem Biochem **45**: 19-72.

Dell, A. (1990). "[35] Preparation and desorption mass spectrometry of permethyl and peracetyl derivatives of oligosaccharides." Methods in Enzymology **193**: 647-660.

Dell, A. and C. E. Ballou (1983). "Fast-atom-bombardment, negative-ion mass spectrometry of the mycobacterial O-methyl-d-glucose polysaccharide and lipopolysaccharides." Carbohydrate Research **120**: 95-111.

Dell, A., A. Galadari, F. Sastre and P. Hitchen (2010). "Similarities and Differences in the Glycosylation Mechanisms in Prokaryotes and Eukaryotes." International Journal of Microbiology **2010**: 14.

Dell, A., Khoo, K., Panico M., McDowell, R.A., Etienne, A.T., Reason, A.J., Morris, H.R. (1993). FAB-MS and ES-MS of glycoproteins. Glycobiology: A practical approach. M. K. Fukuda, A. Oxford, Oxford University Press.

Dell, A. and H. R. Morris (2001). "Glycoprotein structure determination by mass spectrometry." Science **291**(5512): 2351-2356.

Dell, A., Williams, D.H., Morris, H.R., Smith, G.A., Feeney, J., Roberts, G.C. (1975). "Structure revision of the antibiotic echinomycin." J Am Chem Soc **97**(9): 2497-2502.

Derman, Y., H. Söderholm, M. Lindström and H. Korkeala (2015). "Role of csp genes in NaCl, pH, and ethanol stress response and motility in *Clostridium botulinum* ATCC 3502." Food Microbiology **46**: 463-470.

Dethlefsen, L., S. Huse, M. L. Sogin and D. A. Relman (2008). "The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing." PLoS Biology **6**(11): e280.

DiGiandomenico, A., M. J. Matewish, A. Bisailon, J. R. Stehle, J. S. Lam and P. Castric (2002). "Glycosylation of *Pseudomonas aeruginosa* 1244 pilin: glycan substrate specificity." Molecular Microbiology **46**(2): 519-530.

Dingle, K. E., X. Didelot, M. A. Ansari, D. W. Eyre, A. Vaughan, D. Griffiths, C. L. Ip, E. M. Batty, T. Golubchik, R. Bowden, K. A. Jolley, D. W. Hood, W. N. Fawley, A. S. Walker, T. E. Peto, M. H. Wilcox and D. W. Crook (2013). "Recombinational switching of the *Clostridium difficile* S-layer and a novel glycosylation gene cluster revealed by large-scale whole-genome sequencing." J Infect Dis **207**(4): 675-686.

Domon, B., Costello, C.E. (1988). "Structure elucidation of glycosphingolipids and gangliosides using high-performance tandem mass spectrometry." Biochemistry **27**(5): 1534-1543.

Drudy, D., apos, D. P. DONOGHUE, A. BAIRD, L. FENELON, apos and C. FARRELLY (2001). "Flow cytometric analysis of *Clostridium difficile* adherence to human intestinal epithelial cells." Journal of Medical Microbiology **50**(6): 526-534.

Edwards, N. C., Z. A. Hing, A. Perry, A. Blaisdell, D. B. Kopelman, R. Fathke, W. Plum, J. Newell, C. E. Allen, G. S. A. Shapiro, C. Okunji, I. Kosti, N. Shomron, V. Grigoryan, T. M. Przytycka, Z. E. Sauna, R. Salari, Y. Mandel-Gutfreund, A. A. Komar and C. Kimchi-Sarfaty (2012). "Characterization of Coding Synonymous and Non-Synonymous Variants in ADAMTS13 Using Ex Vivo and In Silico Approaches." PLoS ONE **7**(6): e38864.

Egge, H., J. Peter-Katalinić, J. Paz-Parente, G. Strecker, J. Montreuil and B. Fournet (1983). "Carbohydrate structures of hen ovomucoid." FEBS Letters **156**(2): 357-362.

Eichler, J. and M. W. W. Adams (2005). "Posttranslational Protein Modification in Archaea." Microbiology and Molecular Biology Reviews **69**(3): 393-425.

Elhammer, A. P., R. A. Poorman, E. Brown, L. L. Maggiora, J. G. Hoogerheide and F. J. Kézdy (1993). "The specificity of UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase as inferred from a database of in vivo substrates and from the in vitro glycosylation of proteins and peptides." Journal of Biological Chemistry **268**(14): 10029-10038.

Eneroth, O., Hellstroem, K., Ryhage, R. (1964). "Identification and quantification of neutral fecal steroids by gas-liquid chromatography and mass spectrometry: studies of human excretion during two dietary regimens." J Lipid Res **5**: 245-262.

Fagan, R. P., D. Albesa-Jove, O. Qazi, D. I. Svergun, K. A. Brown and N. F. Fairweather (2009). "Structural insights into the molecular organization of the S-layer from *Clostridium difficile*." Mol Microbiol **71**(5): 1308-1322.

Fagan, R. P. and N. F. Fairweather (2011). "Clostridium difficile Has Two Parallel and Essential Sec Secretion Systems." Journal of Biological Chemistry **286**(31): 27483-27493.

Fagan, R. P. and N. F. Fairweather (2014). "Biogenesis and functions of bacterial S-layers." Nat Rev Micro **12**(3): 211-222.

Faridmoayer, A., M. A. Fentabil, D. C. Mills, J. S. Klassen and M. F. Feldman (2007). "Functional Characterization of Bacterial Oligosaccharyltransferases Involved in O-Linked Protein Glycosylation." Journal of Bacteriology **189**(22): 8088-8098.

Faulds-Pain, A., S. M. Twine, E. Vinogradov, P. C. R. Strong, A. Dell, A. M. Buckley, G. R. Douce, E. Valiente, S. M. Logan and B. W. Wren (2014). "The post-translational modification of the *Clostridium difficile* flagellin affects motility, cell surface properties and virulence." Molecular Microbiology **94**(2): 272-289.

Fenn, J., M. Mann, C. Meng, S. Wong and C. Whitehouse (1989). "Electrospray ionization for mass spectrometry of large biomolecules." Science **246**(4926): 64-71.

Francis, C. A., J. M. Beman and M. M. M. Kuypers (2007). "New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation." ISME J **1**(1): 19-27.

Fulton, K. M., J. C. Smith and S. M. Twine (2016). "Clinical applications of bacterial glycoproteins." Expert Review of Proteomics **13**(4): 345-353.

Galperin, M. Y. (2007). "Using archaeal genomics to fight global warming and clostridia to fight cancer." Environmental Microbiology **9**(2): 279-286.

Ganeshapillai, J., E. Vinogradov, J. Rousseau, J. S. Weese and M. A. Monteiro (2008). "Clostridium difficile cell-surface polysaccharides composed of pentaglycosyl and hexaglycosyl phosphate repeating units." Carbohydrate Research **343**(4): 703-710.

Gartner, J. J., S. C. J. Parker, T. D. Prickett, K. Dutton-Regester, M. L. Stitzel, J. C. Lin, S. Davis, V. L. Simhadri, S. Jha, N. Katagiri, V. Gotea, J. K. Teer, X. Wei, M. A. Morken, U. K. Bhanot, N. C. S. Program, G. Chen, L. L. Elnitski, M. A. Davies, J. E. Gershenwald, H. Carter, R. Karchin, W. Robinson, S. Robinson, S. A. Rosenberg, F. S. Collins, G. Parmigiani, A. A. Komar, C. Kimchi-Sarfaty, N. K. Hayward, E. H. Margulies and Y. Samuels (2013). "Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma." Proceedings of the National Academy of Sciences **110**(33): 13481-13486.

Geddes, A. J., G. N. Graham, H. R. Morris, F. Lucas, M. Barber and W. A. Wolstenholme (1969). "Mass-spectrometric determination of the amino acid sequences in peptides isolated from protein silk fibroin of *Bombyx mori*." Biochemical Journal **114**(4): 695-702.

Gerding, D. N., C. A. Muto and J. R. C. Owens (2008). "Treatment of Clostridium difficile Infection." Clinical Infectious Diseases **46**(Supplement_1): S32-S42.

Giraud, M.-F. and J. H. Naismith (2000). "The rhamnose pathway." Current Opinion in Structural Biology **10**(6): 687-696.

Gohlke, R. S., McLafferty, F.W. (1993). "Early gas chromatography/mass spectrometry." J AM Soc Mass Spectrom. **4**(5): 367-371.

Goorhuis, A., D. Bakker, J. Corver, S. B. Debast, C. Harmanus, D. W. Notermans, A. A. Bergwerff, F. W. Dekker and E. J. Kuijper (2008). "Emergence of Clostridium difficile Infection Due to a New Hypervirulent Strain, Polymerase Chain Reaction Ribotype 078." Clinical Infectious Diseases **47**(9): 1162-1170.

Grass, S., A. Z. Buscher, W. E. Swords, M. A. Apicella, S. J. Barenkamp, N. Ozchlewski and J. W. St Geme (2003). "The Haemophilus influenzae HMW1 adhesin is glycosylated in a process that requires HMW1C and phosphoglucomutase, an enzyme involved in lipooligosaccharide biosynthesis." Molecular Microbiology **48**(3): 737-751.

Grass, S., C. F. Lichti, R. R. Townsend, J. Gross and J. W. St. Geme, III (2010). "The Haemophilus influenzae HMW1C Protein Is a Glycosyltransferase That Transfers Hexose Residues to Asparagine Sites in the HMW1 Adhesin." PLoS Pathog **6**(5): e1000919.

Greer, F. M., Morris, H.M. (1997). "Fast-atom bombardment and electrospray mass spectrometry of peptides, proteins and glycoproteins." Methods Mol Biol **64**: 147-163.

Griffiths, J. (2008). "A Brief History of Mass Spectrometry." Analytical Chemistry **80**(15): 5678-5683.

Gross, J., S. Grass, A. E. Davis, P. Gilmore-Erdmann, R. R. Townsend and J. W. St. Geme (2008). "The Haemophilus influenzae HMW1 Adhesin Is a Glycoprotein with an Unusual N-Linked Carbohydrate Modification." Journal of Biological Chemistry **283**(38): 26010-26015.

Guan, Z., S. Naparstek, D. Calo and J. Eichler (2012). "Protein glycosylation as an adaptive response in Archaea: growth at different salt concentrations leads to alterations in *Haloflex volcanii* S-layer glycoprotein N-glycosylation." Environmental microbiology **14**(3): 743-753.

Guerry, P. (2007). "Campylobacter flagella: not just for motility." Trends in Microbiology **15**(10): 456-461.

Guerry, P., C. P. Ewing, M. Schirm, M. Lorenzo, J. Kelly, D. Pattarini, G. Majam, P. Thibault and S. Logan (2006). "Changes in flagellin glycosylation affect Campylobacter autoagglutination and virulence." Molecular microbiology **60**(2): 299-311.

Hakomori, S.-I. (1964). "A Rapid Permethylation of Glycolipid, and Polysaccharide Catalyzed by Methylsulfinyl Carbanion in Dimethyl Sulfoxide." The Journal of Biochemistry **55**(2): 205-208.

Harris, R. J., V. T. Ling and M. W. Spellman (1992). "O-linked fucose is present in the first epidermal growth factor domain of factor XII but not protein C." Journal of Biological Chemistry **267**(8): 5102-5107.

Harris, R. J. S., M.W. (1993). "O-linked fucose and other posttranslational modifications unique to EGF modules." Glycobiology **3**: 219-224.

Harvey, D. J. (2010). "Analysis of carbohydrates and glycoconjugates by matrix-assisted laser desorption/ionization mass spectrometry: An update for 2009–2010." Mass Spectrometry Reviews **34**(3): 268-422.

Helenius, J. and M. Aebi (2002). "Transmembrane movement of dolichol linked carbohydrates during N-glycoprotein biosynthesis in the endoplasmic reticulum." Seminars in Cell & Developmental Biology **13**(3): 171-178.

Hendrixson, D. R. and V. J. DiRita (2004). "Identification of Campylobacter jejuni genes involved in commensal colonization of the chick gastrointestinal tract." Molecular Microbiology **52**(2): 471-484.

Hitchen, P. G. and A. Dell (2006). "Bacterial glycoproteomics." Microbiology **152**(6): 1575-1580.

Hitchen, Paul G., K. Twigger, E. Valiente, Rebecca H. Langdon, Brendan W. Wren and A. Dell (2010). "Glycoproteomics: a powerful tool for characterizing the diverse glycoforms of bacterial pilins and flagellins." Biochemical Society Transactions **38**(5): 1307.

Ho, T. D., K. B. Williams, Y. Chen, R. F. Helm, D. L. Popham and C. D. Ellermeier (2014). "Clostridium difficile Extracytoplasmic Function σ Factor $\sigma(V)$ Regulates Lysozyme Resistance and Is Necessary for Pathogenesis in the Hamster Model of Infection." Infection and Immunity **82**(6): 2345-2355.

Hofsteenge, J., K. G. Huwiler, B. Macek, D. Hess, J. Lawler, D. F. Mosher and J. Peter-Katalinic (2001). "C-Mannosylation and O-Fucosylation of the Thrombospondin Type 1 Module." Journal of Biological Chemistry **276**(9): 6485-6498.

- Hug, I. and M. F. Feldman (2011). "Analogies and homologies in lipopolysaccharide and glycoprotein biosynthesis in bacteria." Glycobiology **21**(2): 138-151.
- Hunt, D. F., A. B. Giordani, G. Rhodes and D. A. Herold (1982). "Mixture analysis by triple-quadrupole mass spectrometry: metabolic profiling of urinary carboxylic acids." Clinical Chemistry **28**(12): 2387-2392.
- Hunt, E. and H. R. Morris (1973). "Collagen cross-links. A mass-spectrometric and (1)H- and (13)C-nuclear-magnetic-resonance study." Biochemical Journal **135**(4): 833-843.
- Hunt, R. C., V. L. Simhadri, M. Iandoli, Z. E. Sauna and C. Kimchi-Sarfaty (2014). "Exposing synonymous mutations." Trends in Genetics **30**(7): 308-321.
- Iavarone, A. T. and E. R. Williams (2002). "Supercharging in electrospray ionization: effects on signal and charge." International Journal of Mass Spectrometry **219**(1): 63-72.
- Iwashkiw, J. A., A. Seper, B. S. Weber, N. E. Scott, E. Vinogradov, C. Stratilo, B. Reiz, S. J. Cordwell, R. Whittall, S. Schild and M. F. Feldman (2012). "Identification of a General O-linked Protein Glycosylation System in *Acinetobacter baumannii* and Its Role in Virulence and Biofilm Formation." PLoS Pathogens **8**(6): e1002758.
- Iwashkiw, J. A., N. F. Voza, R. L. Kinsella and M. F. Feldman (2013). "Pour some sugar on it: the expanding world of bacterial protein O-linked glycosylation." Molecular Microbiology **89**(1): 14-28.
- Janoir, C. (2016). "Virulence factors of *Clostridium difficile* and their role during infection." Anaerobe **37**: 13-24.
- Jarrell, K. F., Y. Ding, B. H. Meyer, S.-V. Albers, L. Kaminski and J. Eichler (2014). "N-Linked Glycosylation in Archaea: a Structural, Functional, and Genetic Analysis." Microbiology and Molecular Biology Reviews : **MMBR** **78**(2): 304-341.
- Jiao, Y., Z. Ma, D. Hodgins, B. Pequegnat, L. Bertolo, L. Arroyo and M. A. Monteiro (2013). "*Clostridium difficile* PSI polysaccharide: synthesis of pentasaccharide repeating block, conjugation to exotoxin B subunit, and detection of natural anti-PSI IgG antibodies in horse serum." Carbohydrate Research **378**: 15-25.
- Johnson, R. S., S. A. Martin, K. Biemann, J. T. Stults and J. T. Watson (1987). "Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine." Analytical Chemistry **59**(21): 2621-2625.
- Jones, M. A., K. L. Marston, C. A. Woodall, D. J. Maskell, D. Linton, A. V. Karlyshev, N. Dorrell, B. W. Wren and P. A. Barrow (2004). "Adaptation of *Campylobacter jejuni* NCTC11168 to High-Level Colonization of the Avian Gastrointestinal Tract." Infection and Immunity **72**(7): 3769-3776.
- Karas, M., Hillenkamo, F. (1988). "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons." Anal Chem **60**(20): 2299-2301.

Karjalainen, T., M. C. Barc, A. Collignon, S. Trollé, H. Boureau, J. Cotte-Laffitte and P. Bourlioux (1994). "Cloning of a genetic determinant from *Clostridium difficile* involved in adherence to tissue culture cells and mucus." *Infection and Immunity* **62**(10): 4347-4355.

Kawata, T., A. Takeoka, K. Takumi and K. Masuda (1984). "Demonstration and preliminary characterization of a regular array in the cell wall of *Clostridium difficile*." *FEMS Microbiology Letters* **24**(2-3): 323-328.

Kelly, C. P. and J. T. LaMont (2008). "*Clostridium difficile* — More Difficult Than Ever." *New England Journal of Medicine* **359**(18): 1932-1940.

Kirby, J. M., H. Ahern, A. K. Roberts, V. Kumar, Z. Freeman, K. R. Acharya and C. C. Shone (2009). "Cwp84, a Surface-associated Cysteine Protease, Plays a Role in the Maturation of the Surface Layer of *Clostridium difficile*." *The Journal of Biological Chemistry* **284**(50): 34666-34673.

Kirk, J. A., O. Banerji and R. P. Fagan (2016). "Characteristics of the *Clostridium difficile* cell envelope and its importance in therapeutics." *Microbial Biotechnology* **10**(1): 76-90.

Knight, D. R., B. Elliott, B. J. Chang, T. T. Perkins and T. V. Riley (2015). "Diversity and Evolution in the Genome of *Clostridium difficile*." *Clinical Microbiology Reviews* **28**(3): 721-741.

Komar, A. A. (2007). "SNPs, Silent But Not Invisible." *Science* **315**(5811): 466-467.

Koonin, E. V. and Y. I. Wolf (2008). "Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world." *Nucleic Acids Research* **36**(21): 6688-6719.

Kowarik, M., N. M. Young, S. Numao, B. L. Schulz, I. Hug, N. Callewaert, D. C. Mills, D. C. Watson, M. Hernandez, J. F. Kelly, M. Wacker and M. Aebi (2006). "Definition of the bacterial N-glycosylation site consensus sequence." *The EMBO Journal* **25**(9): 1957-1966.

Lampert, D. T. A., L. Katona and S. Roerig (1973). "Galactosylserine in extensin." *Biochemical Journal* **133**(1): 125-132.

Larsen, J. C., C. Szymanski and P. Guerry (2004). "N-Linked Protein Glycosylation Is Required for Full Competence in *Campylobacter jejuni* 81-176." *Journal of Bacteriology* **186**(19): 6508-6514.

Lawley, T. D., S. Clare, A. W. Walker, D. Goulding, R. A. Stabler, N. Croucher, P. Mastroeni, P. Scott, C. Raisen, L. Mottram, N. F. Fairweather, B. W. Wren, J. Parkhill and G. Dougan (2009). "Antibiotic Treatment of *Clostridium difficile* Carrier Mice Triggers a Supershedder State, Spore-Mediated Transmission, and Severe Disease in Immunocompromised Hosts." *Infection and Immunity* **77**(9): 3661-3669.

Levy, G. G., W. C. Nichols, E. C. Lian, T. Foroud, J. N. McClintick, B. M. McGee, A. Y. Yang, D. R. Siemieniak, K. R. Stark, R. Gruppo, R. Sarode, S. B. Shurin, V. Chandrasekaran, S. P. Stabler, H. Sabio, E. E. Bouhassira, J. D. Upshaw, D. Ginsburg and H.-M. Tsai (2001). "Mutations in a member of the ADAMTS gene family cause thrombotic thrombocytopenic purpura." *Nature* **413**(6855): 488-494.

Linton, D., E. Allan, A. V. Karlyshev, A. D. Cronshaw and B. W. Wren (2002). "Identification of N-acetylgalactosamine-containing glycoproteins PEB3 and CgpA in *Campylobacter jejuni*." Molecular Microbiology **43**(2): 497-508.

Linton, D., N. Dorrell, P. G. Hitchen, S. Amber, A. V. Karlyshev, H. R. Morris, A. Dell, M. A. Valvano, M. Aebi and B. W. Wren (2005). "Functional analysis of the *Campylobacter jejuni* N-linked protein glycosylation pathway." Molecular Microbiology **55**(6): 1695-1703.

Logan, S. M. (2006). "Flagellar glycosylation – a new component of the motility repertoire?" Microbiology **152**(5): 1249-1262.

Lucado, J., MPH, Gould, C., MD, MSCR, Elixhauser, A. (2012). *Clostridium difficile* Infections (CDI) in Hospital Stays, 2009, Rockville (MD): Agency for Healthcare Research and Quality (US).

Macfarlane, R. and D. Torgerson (1976). "Californium-252 plasma desorption mass spectroscopy." Science **191**(4230): 920-925.

Magidovich, H., S. Yurist-Doutsch, Z. Konrad, V. V. Ventura, A. Dell, P. G. Hitchen and J. Eichler (2010). "AglP is a S-adenosyl-L-methionine-dependent methyltransferase that participates in the N-glycosylation pathway of *Haloferax volcanii*." Molecular Microbiology **76**(1): 190-199.

Mannucci, P. M., C. Capoferri and M. T. Canciani (2004). "Plasma levels of von Willebrand factor regulate ADAMTS-13, its major cleaving protease." British Journal of Haematology **126**(2): 213-218.

Marshall (1974). "The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins." Biochem Soc Symp **40**: 17-26.

Mazetto, B. M., F. L. Orsi, A. Barnabé, É. V. De Paula, M. C. Flores-Nascimento and J. M. Annichino-Bizzacchi (2012). "Increased ADAMTS13 activity in patients with venous thromboembolism." Thrombosis Research **130**(6): 889-893.

Mechref, Y. and M. V. Novotny (2006). "Miniaturized separation techniques in glycomic investigations." Journal of Chromatography B **841**(1–2): 65-78.

Mengele, R. and M. Sumper (1992). "Drastic differences in glycosylation of related S-layer glycoproteins from moderate and extreme halophiles." Journal of Biological Chemistry **267**(12): 8182-8185.

Mescher, M. F. and J. L. Strominger (1976). "Purification and characterization of a prokaryotic glucoprotein from the cell envelope of *Halobacterium salinarium*." Journal of Biological Chemistry **251**(7): 2005-2014.

Messner, P. (2004). "Prokaryotic Glycoproteins: Unexplored but Important." Journal of Bacteriology **186**(9): 2517-2519.

Mikesh, L. M., B. Ueberheide, A. Chi, J. J. Coon, J. E. P. Syka, J. Shabanowitz and D. F. Hunt (2006). "The utility of ETD mass spectrometry in proteomic analysis." Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics **1764**(12): 1811-1822.

Moake, J. L. (2002). "Thrombotic Microangiopathies." New England Journal of Medicine **347**(8): 589-600.

Mock, K. K., Davey, M., Cottrell, J.S. (1991). "The analysis of underivatized oligosaccharides by matrix-assisted laser desorption mass spectrometry." Biochem Biophys Res Commun **177**(2): 644-651.

Monot, M., C. Boursaux-Eude, M. Thibonnier, D. Vallenet, I. Moszer, C. Medigue, I. Martin-Verstraete and B. Dupuy (2011). "Reannotation of the genome sequence of *Clostridium difficile* strain 630." Journal of Medical Microbiology **60**(8): 1193-1199.

Morris, H. R. (1972). "Studies towards the complete sequence determination of proteins by mass spectrometry; a rapid procedure for the successful permethylation of histidine containing peptides." FEBS Letters **22**(3): 257-260.

Morris, H. R., Dell, A., McDowell, R.A. (1981). "Extended performance using a high field magnet mass spectrometer." Biomed Mass Spectrom **8**(9): 463-473.

Morris, H. R., M. Panico, M. Barber, R. S. Bordoli, R. D. Sedgwick and A. Tyler (1981). "Fast atom bombardment: A new mass spectrometric method for peptide sequence analysis." Biochemical and Biophysical Research Communications **101**(2): 623-631.

Morris, H. R., M. Panico and G. W. Taylor (1983). "FAB-MAPPING of recombinant-DNA protein products." Biochemical and Biophysical Research Communications **117**(1): 299-305.

Morris, H. R., T. Paxton, A. Dell, J. Langhorne, M. Berg, R. S. Bordoli, J. Hoyes and R. H. Bateman (1996). "High Sensitivity Collisionally-activated Decomposition Tandem Mass Spectrometry on a Novel Quadrupole/Orthogonal-acceleration Time-of-flight Mass Spectrometer." Rapid Communications in Mass Spectrometry **10**(8): 889-896.

Morris, H. R., T. Paxton, M. Panico, R. McDowell and A. Dell (1997). "A Novel Geometry Mass Spectrometer, the Q-TOF, for Low-Femtomole/Attomole-Range Biopolymer Sequencing." Journal of Protein Chemistry **16**(5): 469-479.

Morris, H. R., M. R. Thompson, D. T. Osuga, A. I. Ahmed, S. M. Chan, J. R. Vandenheede and R. E. Feeney (1978). "Antifreeze glycoproteins from the blood of an antarctic fish. The structure of the proline-containing glycopeptides." Journal of Biological Chemistry **253**(14): 5155-5162.

Morris, H. R., D. H. Williams and R. P. Ambler (1971). "Determination of the sequences of protein-derived peptides and peptide mixtures by mass spectrometry." Biochemical Journal **125**(1): 189-201.

Moxon, E. R., Kroll, J.S. (1990). "The role of bacterial polysaccharide capsules as virulence factors." Curr Top Microbiol Immunol **150**: 65-85.

Muir, L. and Y. C. Lee (1970). "Glycopeptides from Earthworm Cuticle Collagen." Journal of Biological Chemistry **245**(3): 502-509.

Naaber, P., Lehto, E., Salminen, S., Mikelsaar, M. (1996). "Inhibition of adhesion of *Clostridium difficile* to Caco-2 cells." FEMS Immunol Med Microbiol **14**(4): 205-209.

Nau, H., Kelley J.A., Bienmann, K. (1973). "Determination of the amino acid sequence of the C-terminal cyanogen bromide fragment of actin by computer-assisted gas-chromatography mass spectrometry." J Am Chem Soc **95**(21): 7162-7164.

Ní Eidhin, D., A. W. Ryan, R. M. Doyle, J. B. Walsh and D. Kelleher (2006). "Sequence and phylogenetic analysis of the gene for surface layer protein, slpA, from 14 PCR ribotypes of *Clostridium difficile*." Journal of Medical Microbiology **55**(1): 69-83.

Nier, A. O. (1947). "A mass spectrometer for isotope and gas analyse." Rev Sci Instrum **18**(6): 398-411.

Nishimura, H., T. Takao, S. Hase, Y. Shimonishi and S. Iwanaga (1992). "Human factor IX has a tetrasaccharide O-glycosidically linked to serine 61 through the fucose residue." Journal of Biological Chemistry **267**(25): 17520-17525.

North, S. J., H.-H. Huang, S. Sundaram, J. Jang-Lee, A. T. Etienne, A. Trollope, S. Chalabi, A. Dell, P. Stanley and S. M. Haslam (2010). "Glycomics Profiling of Chinese Hamster Ovary Cell Glycosylation Mutants Reveals N-Glycans of a Novel Size and Complexity." Journal of Biological Chemistry **285**(8): 5759-5775.

Nothaft, H., X. Liu, D. J. McNally, J. Li and C. M. Szymanski (2009). "Study of free oligosaccharides derived from the bacterial N-glycosylation pathway." Proceedings of the National Academy of Sciences **106**(35): 15019-15024.

Nothaft, H. and C. M. Szymanski (2010). "Protein glycosylation in bacteria: sweeter than ever." Nat Rev Micro **8**(11): 765-778.

Novotny, M. V. and Y. Mechref (2005). "New hyphenated methodologies in high-sensitivity glycoprotein analysis." Journal of separation science **28**(15): 1956-1968.

O'Neill, M. A. and R. R. Selvendran (1980). "Glycoproteins from the cell wall of *Phaseolus coccineus*." Biochemical Journal **187**(1): 53-63.

Panico, M., L. Bouché, D. Binet, M.-J. O'Connor, D. Rahman, P.-C. Pang, K. Canis, S. J. North, R. C. Desrosiers, E. Chertova, B. F. Keele, J. W. Bess, J. D. Lifson, S. M. Haslam, A. Dell and H. R. Morris (2016). "Mapping the complete glycoproteome of virion-derived HIV-1 gp120 provides insights into broadly neutralizing antibody binding." Scientific Reports **6**: 32956.

Paul, W. and H. Steinwedel (1953). Notizen: Ein neues Massenspektrometer ohne Magnetfeld. Zeitschrift für Naturforschung A. **8**: 448.

Peltier, J., P. Courtin, I. El Meouche, L. Lemée, M.-P. Chapot-Chartier and J.-L. Pons (2011). "*Clostridium difficile* Has an Original Peptidoglycan Structure with a High Level of N-

Acetylglucosamine Deacetylation and Mainly 3-3 Cross-links." The Journal of Biological Chemistry **286**(33): 29053-29062.

Percy, M. G., Gründling, A. (2014). "Lipoteichoic acid synthesis and function in gram-positive bacteria." Annu Rev Microbiol **68**: 81-100.

Poxton, I. R. (2013). "The changing faces of *Clostridium difficile*: A personal reflection of the past 34 years." Anaerobe **24**: 124-127.

Qazi, O., P. Hitchen, B. Tissot, M. Panico, H. R. Morris, A. Dell and N. Fairweather (2009). "Mass spectrometric analysis of the S-layer proteins from *Clostridium difficile* demonstrates the absence of glycosylation." J Mass Spectrom **44**(3): 368-374.

Reid, C. W., E. Vinogradov, J. Li, H. C. Jarrell, S. M. Logan and J.-R. Brisson (2012). "Structural characterization of surface glycans from *Clostridium difficile*." Carbohydrate Research **354**: 65-73.

Reynolds, P. E. (1989). "Structure, biochemistry and mechanism of action of glycopeptide antibiotics." Eur J Clin Microbiol Infect Dis **8**(11): 943-950.

Ricketts, L. M., M. Dlugosz, K. B. Luther, R. S. Haltiwanger and E. M. Majerus (2007). "O-Fucosylation Is Required for ADAMTS13 Secretion." Journal of Biological Chemistry **282**(23): 17014-17023.

Ristl, R., K. Steiner, K. Zarschler, S. Zayni, P. Messner and C. Schaffer (2011). "The s-layer glycome-adding to the sugar coat of bacteria." Int J Microbiol **2011**.

Roepstorff, P., Fohlman, J. (1984). "Proposal for a common nomenclature for sequence ions in mass spectra of peptides." Biomed Mass Spectrom **11**(11): 601.

Rupnik, M., M. H. Wilcox and D. N. Gerding (2009). "*Clostridium difficile* infection: new developments in epidemiology and pathogenesis." Nat Rev Micro **7**(7): 526-536.

Ryan, A., M. Lynch, S. M. Smith, S. Amu, H. J. Nel, C. E. McCoy, J. K. Dowling, E. Draper, V. O'Reilly, C. McCarthy, J. O'Brien, D. Ní Eidhin, M. J. O'Connell, B. Keogh, C. O. Morton, T. R. Rogers, P. G. Fallon, L. A. O'Neill, D. Kelleher and C. E. Loscher (2011). "A Role for TLR4 in *Clostridium difficile* Infection and the Recognition of Surface Layer Proteins." PLoS Pathogens **7**(6): e1002076.

Sandman, K. R., J.N. (2000). "Structure and functional relationships of archaeal and eukaryal histones and nucleosomes." Arch Microbiol **173**(3): 165-169.

Sauna, Z. E. and C. Kimchi-Sarfaty (2011). "Understanding the contribution of synonymous mutations to human disease." Nat Rev Genet **12**(10): 683-691.

Schiller, B., A. Hykollari, S. Yan, K. Paschinger and I. B. H. Wilson (2012). "Complicated N-linked glycans in simple organisms." Biological chemistry **393**(8): 661-673.

Schirm, M., E. C. Soo, A. J. Aubry, J. Austin, P. Thibault and S. M. Logan (2003). "Structural, genetic and functional characterization of the flagellin glycosylation process in *Helicobacter pylori*." Molecular Microbiology **48**(6): 1579-1592.

Sebahia, M., B. W. Wren, P. Mullany, N. F. Fairweather, N. Minton, R. Stabler, N. R. Thomson, A. P. Roberts, A. M. Cerdeno-Tarraga, H. Wang, M. T. Holden, A. Wright, C. Churcher, M. A. Quail, S. Baker, N. Bason, K. Brooks, T. Chillingworth, A. Cronin, P. Davis, L. Dowd, A. Fraser, T. Feltwell, Z. Hance, S. Holroyd, K. Jagels, S. Moule, K. Mungall, C. Price, E. Rabbinowitsch, S. Sharp, M. Simmonds, K. Stevens, L. Unwin, S. Whithead, B. Dupuy, G. Dougan, B. Barrell and J. Parkhill (2006). "The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome." Nat Genet **38**(7): 779-786.

Silberstein, S. and R. Gilmore (1996). "Biochemistry, molecular biology, and genetics of the oligosaccharyltransferase." The FASEB Journal **10**(8): 849-858.

Sleytr, U. B. and K. J. Thorne (1976). "Chemical characterization of the regularly arranged surface layers of *Clostridium thermosaccharolyticum* and *Clostridium thermohydrosulfuricum*." Journal of Bacteriology **126**(1): 377-383.

Slimings, C. and T. V. Riley (2014). "Antibiotics and hospital-acquired *Clostridium difficile* infection: update of systematic review and meta-analysis." Journal of Antimicrobial Chemotherapy **69**(4): 881-891.

Smit, E., D. Jager, B. Martinez, F. J. Tielen and P. H. Pouwels (2002). "Structural and Functional Analysis of the S-layer Protein Crystallisation Domain of *Lactobacillus acidophilus* ATCC 4356: Evidence for Protein-Protein Interaction of two Subdomains." Journal of Molecular Biology **324**(5): 953-964.

Smith, T. G. and T. R. Hoover (2009). Chapter 8 Deciphering Bacterial Flagellar Gene Regulatory Networks in the Genomic Era. Advances in Applied Microbiology, Academic Press. **Volume 67**: 257-295.

Smits, W. K., D. Lyras, D. B. Lacy, M. H. Wilcox and E. J. Kuijper (2016). "*Clostridium difficile* infection." Nature Reviews Disease Primers **2**: 16020.

Sobott, F., M. G. McCammon, H. Hernández and C. V. Robinson (2005). "The flight of macromolecular complexes in a mass spectrometer." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **363**(1827): 379-391.

Soejima, K., N. Mimura, M. Hirashima, H. Maeda, T. Hamamoto, T. Nakagaki and C. Nozaki (2001). "A Novel Human Metalloprotease Synthesized in the Liver and Secreted into the Blood: Possibly, the von Willebrand Factor-Cleaving Protease?" The Journal of Biochemistry **130**(4): 475-480.

Soutourina, O. A. and P. N. Bertin (2003). "Regulation cascade of flagellar expression in Gram-negative bacteria." FEMS Microbiology Reviews **27**(4): 505-523.

Spigaglia, P., A. Barketi-Klai, A. Collignon, P. Mastrantonio, F. Barbanti, M. Rupnik, S. Janezic and I. Kansau (2013). "Surface-layer (S-layer) of human and animal *Clostridium difficile* strains and their behaviour in adherence to epithelial cells and intestinal colonization." Journal of Medical Microbiology **62**(9): 1386-1393.

Spigaglia, P., C. L. Galeotti, F. Barbanti, M. Scarselli, J. Van Broeck and P. Mastrantonio (2011). "The LMW surface-layer proteins of *Clostridium difficile* PCR ribotypes 027 and 001 share common immunogenic properties." Journal of Medical Microbiology **60**(8): 1168-1173.

Spina, E., Sturiale, L., Romeo, D., Impallomeni, G., Garozzo, D., Waidelich, D., Glueckmann, M. (2004). "New fragmentation mechanisms in matrix-assisted laser desorption/ionisation time-of-flight/time-of-flight tandem mass spectrometry of carbohydrates." Rapid Communications in Mass Spectrometry **18**(4): 392-398.

Spiro, R. G. (1973). "Glycoproteins." Adv Protein Chem **27**: 349-467.

Spiro, R. G. (2002). "Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds." Glycobiology **12**(4): 43-56.

Spiro, R. G. and V. D. Bhoyroo (1980). "Studies on the carbohydrate of collagens. Characterization of a glucuronic acid-mannose disaccharide unit from *Nereis* cuticle collagen." Journal of Biological Chemistry **255**(11): 5347-5354.

St Geme, J. W. (1994). "The HMW1 adhesin of nontypeable *Haemophilus influenzae* recognizes sialylated glycoprotein receptors on cultured human epithelial cells." Infection and Immunity **62**(9): 3881-3889.

Stabler, R. A., D. N. Gerding, J. G. Songer, D. Drudy, J. S. Brazier, H. T. Trinh, A. A. Witney, J. Hinds and B. W. Wren (2006). "Comparative phylogenomics of *Clostridium difficile* reveals clade specificity and microevolution of hypervirulent strains." J Bacteriol **188**.

Stabler, R. A., M. He, L. Dawson, M. Martin, E. Valiente, C. Corton, T. D. Lawley, M. Sebahia, M. A. Quail, G. Rose, D. N. Gerding, M. Gibert, M. R. Popoff, J. Parkhill, G. Dougan and B. W. Wren (2009). "Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium." Genome Biology **10**(9): R102.

Stabler, R. A., E. Valiente, L. F. Dawson, M. He, J. Parkhill and B. W. Wren (2010). "In-depth genetic analysis of *Clostridium difficile* PCR-ribotype 027 strains reveals high genome fluidity including point mutations and inversions." Gut Microbes **1**(4): 269-276.

Stansell, E., M. Panico, K. Canis, P.-C. Pang, L. Bouché, D. Binet, M.-J. O'Connor, E. Chertova, J. Bess, J. D. Lifson, S. M. Haslam, H. R. Morris, R. C. Desrosiers and A. Dell (2015). "Gp120 on HIV-1 Virions Lacks O-Linked Carbohydrate." PLOS ONE **10**(4): e0124784.

Steiner, K., R. Novotny, D. B. Werz, K. Zarschler, P. H. Seeberger, A. Hofinger, P. Kosma, C. Schäffer and P. Messner (2008). "Molecular Basis of S-layer Glycoprotein Glycan

Biosynthesis in *Geobacillus stearothermophilus*." Journal of Biological Chemistry **283**(30): 21120-21133.

Stephens, E., S. L. Maslen, L. G. Green and D. H. Williams (2004). "Fragmentation Characteristics of Neutral N-Linked Glycans Using a MALDI-TOF/TOF Tandem Mass Spectrometer." Analytical Chemistry **76**(8): 2343-2354.

Stevenson, E., N. P. Minton and S. A. Kuehne (2015). "The role of flagella in *Clostridium difficile* pathogenicity." Trends in Microbiology **23**(5): 275-282.

Stimson, E., M. Virji, K. Makepeace, A. Dell, H. R. Morris, G. Payne, J. R. Saunders, M. P. Jennings, S. Barker, M. Panico, I. Blench and E. R. Moxon (1995). "Meningococcal pilin: a glycoprotein substituted with digalactosyl 2,4-diacetamido-2,4,6-trideoxyhexose." Molecular Microbiology **17**(6): 1201-1214.

Sumper, M. (1987). "Halobacterial glycoprotein biosynthesis." Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes **906**(1): 69-79.

Supek, F., B. Miñana, J. Valcárcel, T. Gabaldón and B. Lehner (2014). "Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers." Cell **156**(6): 1324-1335.

Sweet, D. P., Shapiro, R.H., Albersheim, P. (1974). "The mass spectral fragmentation of partially ethylated alditol acetates, a derivative used in determining the glycosyl linkage composition of polysaccharides." Biomed Mass Spectrom **1**(4): 263-268.

Szymanski, C. M., D. H. Burr and P. Guerry (2002). "Campylobacter Protein Glycosylation Affects Host Cell Interactions." Infection and Immunity **70**(4): 2242-2244.

Szymanski, C. M., S. M. Logan, D. Linton and B. W. Wren (2003). "Campylobacter – a tale of two protein glycosylation systems." Trends in Microbiology **11**(5): 233-238.

Szymanski, C. M. and B. W. Wren (2005). "Protein glycosylation in bacterial mucosal pathogens." Nat Rev Micro **3**(3): 225-237.

Szymanski, C. M., R. Yao, C. P. Ewing, T. J. Trust and P. Guerry (1999). "Evidence for a system of general protein glycosylation in *Campylobacter jejuni*." Molecular Microbiology **32**(5): 1022-1030.

Takumi, K., T. Koga, T. Oka and Y. Endo (1991). "SELF-ASSEMBLY, ADHESION, AND CHEMICAL PROPERTIES OF TETRAGONARLY ARRAYED S-LAYER PROTEINS OF *CLOSTRIDIUM*." The Journal of General and Applied Microbiology **37**(6): 455-465.

Tasteyre, A., M.-C. Barc, A. Collignon, H. Boureau and T. Karjalainen (2001). "Role of FliC and FliD Flagellar Proteins of *Clostridium difficile* in Adherence and Gut Colonization." Infection and Immunity **69**(12): 7937-7940.

Taylor, M. E. D., K. (2011). Introduction to Glycobiology. Oxford University Press.

Twine, S. M., C. W. Reid, A. Aubry, D. R. McMullin, K. M. Fulton, J. Austin and S. M. Logan (2009). "Motility and Flagellar Glycosylation in *Clostridium difficile*." Journal of Bacteriology **191**(22): 7050-7062.

Valiente, E., L. Bouché, P. Hitchen, A. Faulds-Pain, M. Songane, L. F. Dawson, E. Donahue, R. A. Stabler, M. Panico, H. R. Morris, M. Bajaj-Elliott, S. M. Logan, A. Dell and B. W. Wren (2016). "Role of Glycosyltransferases Modifying Type B Flagellin of Emerging Hypervirulent *Clostridium difficile* Lineages and Their Impact on Motility and Biofilm Formation." The Journal of Biological Chemistry **291**(49): 25450-25461.

Varki, A., Cummings, R.D., Esko, J.D., Freeze, H.H., Stanley, P., Bertozzi, C.R., Hart, G.W. & Etzler, M.E. (2009). Essentials of Glycobiology.

Vestal, M. L. and J. M. Campbell (2005). Tandem Time-of-Flight Mass Spectrometry. Methods in Enzymology, Academic Press. **Volume 402**: 79-108.

Vollmer, W., D. Blanot and M. A. De Pedro (2008). "Peptidoglycan structure and architecture." FEMS Microbiology Reviews **32**(2): 149-167.

von Eichel-Streiber, C., P. Boquet, M. Sauerborn and M. Thelestam (1996). "Large clostridial cytotoxins — a family of glycosyltransferases modifying small GTP-binding proteins." Trends in Microbiology **4**(10): 375-382.

Wacker, M., D. Linton, P. G. Hitchen, M. Nita-Lazar, S. M. Haslam, S. J. North, M. Panico, H. R. Morris, A. Dell, B. W. Wren and M. Aebi (2002). "N-Linked Glycosylation in *Campylobacter jejuni* and Its Functional Transfer into *E. coli*." Science **298**(5599): 1790-1793.

Weerapana, E. and B. Imperiali (2006). "Asparagine-linked protein glycosylation: from eukaryotic to prokaryotic systems." Glycobiology **16**(6): 91R-101R.

Weidenmaier, C. and A. Peschel (2008). "Teichoic acids and related cell-wall glycopolymers in Gram-positive physiology and host interactions." Nat Rev Micro **6**(4): 276-287.

Wieland, F., W. Dompert, G. Bernhardt and M. Sumper (1980). "Halobacterial glycoprotein saccharides contain covalently linked sulphate." FEBS Letters **120**(1): 110-114.

Wiesner, J., T. Premisler and A. Sickmann (2008). "Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications." PROTEOMICS **8**(21): 4466-4483.

Willing, S. E., T. Candela, H. A. Shaw, Z. Seager, S. Mesnage, R. P. Fagan and N. F. Fairweather (2015). "C *lostridium difficile* surface proteins are anchored to the cell wall using CWB2 motifs that recognise the anionic polymer PSII." Molecular Microbiology **96**(3): 596-608.

Wilm, M. S. and M. Mann (1994). "Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last?" International Journal of Mass Spectrometry and Ion Processes **136**(2): 167-180.

Winkler, H. U. and H. D. Beckey (1972). "Field desorption mass spectrometry of peptides." Biochemical and Biophysical Research Communications **46**(2): 391-398.

Wright, A., R. Wait, S. Begum, B. Crossett, J. Nagy, K. Brown and N. Fairweather (2005). "Proteomic analysis of cell surface proteins from *Clostridium difficile*." PROTEOMICS **5**(9): 2443-2452.

Young, N. M., J.-R. Brisson, J. Kelly, D. C. Watson, L. Tessier, P. H. Lanthier, H. C. Jarrell, N. Cadotte, F. St. Michael, E. Aberg and C. M. Szymanski (2002). "Structure of the N-Linked Glycan Present on Multiple Glycoproteins in the Gram-negative Bacterium, *Campylobacter jejuni*." Journal of Biological Chemistry **277**(45): 42530-42539.

Zheng, S., H. Kim and Roel G. W. Verhaak (2014). "Silent Mutations Make Some Noise." Cell **156**(6): 1129-1131.

Zheng, X., D. Chung, T. K. Takayama, E. M. Majerus, J. E. Sadler and K. Fujikawa (2001). "Structure of von Willebrand Factor-cleaving Protease (ADAMTS13), a Metalloprotease Involved in Thrombotic Thrombocytopenic Purpura." Journal of Biological Chemistry **276**(44): 41059-41063.

Zheng, X. L. (2013a). "ADAMTS13, TTP and Beyond." Hereditary genetics : current research **2**(1): e104.

Zheng, X. L. (2013b). "Structure-function and regulation of ADAMTS13 protease." Journal of thrombosis and haemostasis : JTH **11**(0 1): 11-23.