Development of Top-Down Mass Spectrometry Approaches for the Analysis of Type IV Pili

Joseph Frederick Gault

THESE

pour obtenir le grade de

DOCTEUR DE L'ECOLE POLYTECHNIQUE

Spécialité : Chimie

par

JOSEPH FREDERICK GAULT

Development of Top-Down Mass Spectrometry Approaches for the Analysis of Type IV Pili

Directrice de thèse : Dr Julia Chamot-Rooke

Soutenue publiquement le 19 décembre 2013 à l'Institut Pasteur

devant la commission d'examen composée de:

Prof. Yves Méchulam	Président
Dr Daniel Lafitte	Rapporteur
Prof. Jérôme Lemoine	Rapporteur
Prof. Catherine E. Costello	Examinateur
Dr Guillaume Duménil	Examinateur
Dr Julia Chamot-Rooke	Directrice de Thèse

THESE

pour obtenir le grade de

DOCTEUR DE L'ECOLE POLYTECHNIQUE

Spécialité : Chimie

par

JOSEPH FREDERICK GAULT

Development of Top-Down Mass Spectrometry Approaches for the Analysis of Type IV Pili

Directrice de thèse : Dr Julia Chamot-Rooke

Soutenue publiquement le 19 décembre 2013 à l'Institut Pasteur

devant la commission d'examen composée de:

Prof. Yves Méchulam	Président
Dr Daniel Lafitte	Rapporteur
Prof. Jérôme Lemoine	Rapporteur
Prof. Catherine E. Costello	Examinateur
Dr Guillaume Duménil	Examinateur
Dr Julia Chamot-Rooke	Directrice de Thèse

Acknowledgements

Wow, it's over! The last four years have been an amazing experience over which I had had the joy to work with many groups in several labs both in and around Paris and Boston USA. These first couple of pages are devoted to the people I have met and friends I have made. You made the hard times more bearable and the happy ones truly fantastic. Thank you.

First of all I would like to thank the president Prof. Yves Méchulam and members of my jury Prof. Catherine Costello, Dr Guillaume Duménil, Dr Daniel Lafitte and Prof. Jérôme Lemoine for critical reading of my manuscript and for examination of my work.

It would not have been possible for me to undertake this thesis without the support of Gilles Ohanessian, director of my first lab, the DCMR at Polytechnique. Thank you for your support during my M2, my Monge application and throughout my time at l'X. Thank you also to all of the other members of the DCMR, and in particular the students with whom I shared my first couple of years, Magalie, Ophélie, Yasmine, Ahmad, Ashwani, Renjie, Jianqing and more recently Vanessa, Julien and Jana. To David and Amy for the late nights spent in the lab, for the raw vegetables, le vin rouge drunk from jam jars, French lessons and cheese. An introduction to French (Franco-American?) culture par excellence! David, I owe you much for the progress I made with my French, a special thanks for that. And to the other members of my group, to Edith for your patience and for helping me understand how to tame the Apex III, to Guillaume for the discussions, for sharing with me some of your vast knowledge and showing me that anything is possible with a bit of thought!

Much of the method development in the first part of my thesis relied experiments performed on mass spectrometers not present at the DCMR. Thanks to Philippe Maître for allowing me access to the ApeX Qe FT-ICR at Université Paris Sud (Orsay) and Vincent for your help making it work, and to Jean-Michel Camadro at the Institut Jacques Monod for access to the Orbitrap Velos and to Thibault and Manuel for all of your help.

My second home for much of this thesis has been the lab of Guillaume Duménil at the Paris Cardiovascular Research Centre, Hôpital Européen Georges-Pompidou. The sheer energy of the lab was a major driving force for me early on; I can't thank you enough for that, and the resulting scientific story rests on many of the biological experiments which you have performed. Thanks to Magali and Anne-Flore who have been there from the beginning and to Silke with whom I enjoyed working with so much at the end. A special thanks to Corrine, the vast majority of the 700 or so pili preps would not have been possible without you! To all the other members of the group,

Audrey, Keira, Hebert, Valentina, Paula, Flore, Tiffany, Arthur and to Eric and his group, thanks for welcoming me and for all of the good times!

Eight months of the last four years were also spent in Boston, MA in the lab of Prof. Catherine Costello. Living and working in Boston was an eye opening experience for me in many ways and I would like to thank all of the members of Cathy's lab for welcoming me so warmly. So far from France it was a joy to practise my French with Nancy, Roger and Sandrine. Thanks to Nancy for all the help with the Orbi, and for never making life dull, and to Sandrine for all your help with the FT. Thanks also to Cheng for spending so much time discussing with me and to Sean. A special thanks to Debora and Sabine for all of the fun times both in and out of the lab and particular those on the bike cycling around. Most of all thank you to Cathy, for accepting me into your lab and for your wise words, your guidance, support and for sharing your passion for science.

Back to France and at the beginning of my third year I moved to a new lab at the Institut Pasteur and a tiny apartment in the heart of Paris. Thanks to all of the brilliant scientists with whom I have had the privilege to collaborate (and talk about pili!): Deshmukh Gopaul (my tutor), Olivera Francetic, Gerard Pehau-Arnaudet, Rémi Fronzes, Michael Nilges, Mathias Ferber, Guillaume Bouvier and Raphaël Laurenceau. To my colleagues at Pasteur for introducing me to the strange new world of proteomics, especially those with whom I shared the last couple of months Sébastien, Mariette, Myriam, Cyril, Magalie (again!) Véronique, Thibault, Egor and Francis. To Debora for enriching the craziness of our office and being a great M2 student, and to "senior post-doc" Catherine, for being the best of friends and for all of our discussions.

Finally I would like to thank to thank the three people who have accompanied me on this journey from beginning to end. Guillaume, it has been a real pleasure working with you. Thank you for all of your time, for putting up with all of the questions and for helping me when things got tough. I will be forever grateful. To Julia, my supervisor. You have taught me a lot about science, about working with others and how to construct and communicate my ideas. It has been an incredible journey from Polytechnique to Pasteur and you have never stopped pushing forward, thank you. And to Christian, this thesis is in many ways as much yours as it is mine. You have shared the early mornings, the endless pili preps and have helped shape the ideas that have moulded both the work and me personally. You shall forever remain a friend.

Lastly I would like to thank Marianne, who has always been by my side, and the rest of my family. It is difficult to express how grateful I am to you all. You have always been there to support me, to guide me, and to offer sage advice. You have been my inspiration throughout and I have loved sharing my experiences with you.

Contents

Chapter 3

Chapter 4

Chapter 5

Investigation of PTM of PilE in Novel Clinical Isolates of *Neisseria meningitidis*

Abbreviations

2DE – two dimensional electrophoresis

AA – amino acid

AGC – automatic gain control

AI-ECD – activated ion electron capture dissociation

ASF – alternative splice form

BATDH – 2-acetamido 4-butyramido 2,4,6-trideoxy α-D-hexose

BIRD –Blackbody infrared radiative dissociation

CAD – collision-activated dissociation

CE – capillary electrophoresis

CFU – colony forming unit

CI – chemical ionisation

CID – collision induced dissociation

CLIO – "centre laser infrarouge d'Orsay"

CSD – charge state distribution

CSF – cerebrospinal fluid

CSP – competence stimulating peptide

Da – Dalton

DATDH – 2,4-diacetamido 2,4,6-trideoxy α-D-hexose

DMSO – dimethylsulfoxide

DNA – deoxyribonucleic acid

DT – dynamic trapping

DTT – dithiothreitol

ECD – electron capture dissociation

EI – electron impact

ESI – electrospray ionisation

ETD – electron transfer dissociation

ETnD – electron transfer no dissociation

ExD – electron activated dissociation

eV – electron Volt

FAB – fast atom bombardment

FD – Frequency domain

FDR – false discovery rate

FFT – Fast Fourier transform

FPLC – fast protein liquid chromatography

FT-ICR – Fourier transform ion cyclotron resonance

GATDH – 2-acetamido 4-glyceramido 2,4,6-trideoxy α-D-hexose

GC-MS – gas chromatography mass spectrometry

HCD – high energy C-trap dissociation

Hex – hexose

HILIC - hydrophilic interaction liquid chromatography

HPLC – high pressure liquid chromatography

HPP - human proteome project

ICR – ion cyclotron resonance

ID – identification

IEF – isoelectric focusing

IgG – immunoglobulin

IPTG – isopropyl β-D-1-thiogalactopyranoside

IR –infra-red

IRMPD –infra-red multi photon dissociation

ISD – in source dissociation

LC – liquid chromatography

LESA – liquid extraction surface analysis

LPS – lipopolysaccharide

LT – linear trap

LT Q – linear trap quadrupole

m – mass

MAIVI – matrix assisted ionisation vacuum ionisation

MALDI –matrix assisted laser desorption ionisation

M_{av} – average mass

m-NBA – meta-nitrobenzyl alcohol

m-RNA – messenger ribonucleic acid

MLST – multi locus sequence typing

M_{mono} – monoisotopic mass

MS – mass spectrometer / mass spectrometry

MS/MS – tandem mass spectrometry

MW – molecular weight

N*g* – *Neisseria gonorrhoeae*

Nm – *Neisseria meningitidis*

NSD – nozzle skimmer dissociation

ORF – open reading frame

PC – phosphocholine

PE – phosphoethanolamine

PFD – prefolding dissociation

PG – phosphoglycerol

pgl – protein glycosylation

pI – isoelectric point

PI – precursor ion

PMF – peptide mass fingerprinting

ppm – parts per million

pptA/pptB – phosphoglyceroltrasferase A/B

PTM – posttranslational modification

PSD – post source decay

Q – quadrupole

Q_f – quality factor

QqQ – triple quadrupole

Q-ToF – quadrupole time-of-flight

RF – radio frequency

RPLC – reverse phase liquid chromatography

S/N – signal-to-noise

SAP – single amino acid polymorphism

SCX – strong cation exchange

SDS-PAGE – sodium dodecyl sulfate poly acrylamide gel electrophoresis

SID – surface-induced dissociation

SEC – size exclusion chromatography

sIEF – solution isoelectric focusing

SIMS – secondary ion mass spectrometry

SNP – single nucleotide polymorphism

SORI-CAD – sustained off resonance irradiation for collisional activation

T4P – type IV pili

TEM – transmission electron microscopy

TCEP – tris(2-carboxyethyl)phosphine

TD –time domain

TDMS – top-down mass spectrometry

ToF – time of flight

UVPD – ultraviolet photon dissociation

WCX – weak cation exchange

WT – wild type

z – charge

General Introduction

Mass spectrometry has developed into an incredibly powerful tool for the identification and characterisation of biological molecules. Within the space of forty years the technique has evolved from measuring the mass of small, volatile organic compounds to characterising the composition of huge macromolecular complexes containing multiple proteins, lipids and nucleotides. As biology is becoming more concerned with understanding cellular processes on a molecular level, mass spectrometry provides a unique tool to study the protagonists in the cell that are the proteins themselves.

The protein complement of a cell is exceptionally dynamic, constantly changing in response to the cellular environment to ensure homeostasis and adapting over time to regulate cell division and senescence. The term proteome was coined to define this set of proteins, and contemporary proteomics (the study of the proteome) strives to identify and quantify all protein forms expressed by a cell, tissue or organism at a given moment in response to precisely defined stimuli, including their interactions with other biomolecules and cellular localisation.

In the post-genomic age, proteomics has evolved in order to meet the challenges of analysing such large and dynamic systems. Sophisticated high-throughput, bottom-up proteomics methods have been developed where entire proteomes are digested by proteolytic enzymes into large numbers of peptides, which are then characterised by mass spectrometry. Particularly in the last few years, bottom-up proteomics has been coupled with robust, high resolution separation strategies and extremely fast scanning, sensitive mass spectrometers, to simultaneously identify and quantify large numbers of peptides and, through them, large numbers of proteins. Rigorous bioinformatics approaches have also been developed to validate this data statistically.

The recently proposed Human Proteome Project (HPP), which comprises two audacious programs to map the protein based molecular architecture of the human body, is testament to how far proteomics has advanced. However, despite the progress, bottom-up proteomics has failed to live up to its promise in the important field of disease biomarker discovery. The success of the human genome project probably meant that the bar was set too high, too early, but the return on the huge in investment in proteomics in this domain has been undeniably low.

More and more evidence is being presented linking posttranslational modification (PTM) to disease and this perhaps explains why proteomics has underperformed in this area. Indeed bottom-up proteomics is fundamentally unable to address the complexity of the proteome imparted by sequence variants and PTMs. PTM is unique among the processes leading to proteome variation in that complete characterisation of posttranslationally modified proteins can only be achieved at the protein level. Usual approaches based on proteolytic digestion are poorly suited to this task, as they destroy the connectivity between peptides and their parent

proteoforms. New approaches to proteome analysis are therefore required, particularly those that go beyond the capabilities of even state-of-the-art bottom-up proteomics and that are specially adapted for characterising posttranslationally modified proteoforms. In this regard top-down mass spectrometry holds great promise, despite still being in its adolescence.

Top-down mass spectrometry is a holistic approach to protein analysis that involves the identification and characterisation of intact proteins. It is therefore most often concerned with proteoforms - the many different molecular forms in which the protein product of a single gene can be found. The approach is technically challenging and only in the past few years have the technological advances been made to conceive application of this technique to large scale studies or entire proteomes.

Bacterial proteomes are ideal models for application of the top-down methodology. They are large enough in size for studies developing top-down proteomics (large-scale proteome analysis based on top-down mass spectrometry) and are rich in the small to medium sized proteins that are more amenable to top-down characterisation. They are also fertile hunting grounds for new and unusual posttranslationally modified proteins. In addition, many bacteria are pathogenic and their fundamental biology is both incompletely understood and relevant to human health.

One such pathogen is the Gram negative bacterium *Neisseria meningitidis* (Nm), the etiological agent of cerebrospinal meningitis. Nm possesses a number of virulence factors, including type IV pili (T4P). T4P are extracellular organelles expressed by several bacterial species and are involved in multiple processes linked to colonisation and infection of the host. They are principally composed of a single protein, the major pilin, which is called PilE in *Neisseria* spp. In the small number of reference strains that have been partially characterised to date, this protein is found to be expressed in a number of heavily posttranslationally modified proteoforms. However, the biological function of many of these PTMs is currently unknown, as is the extent and variety of PTM present on PilE in the wider bacterial population.

In close collaboration with the group of Dr Guillaume Duménil, this thesis aimed to develop mass spectrometry methods and create a new top-down approach for the characterisation of proteoforms of PilE..

A reference strain of *Neisseria meningitidis* was chosen on which to set up the initial top-down experiment, with the aim of providing sufficient protein coverage for complete PTM localisation. The approach was then to be adapted for high-throughput analysis of PilE purified from more recent clinical isolates. Implementation of the methodology in a wider scale study, would allow the variety of PTM present on PilE in the wider bacterial population to be sampled. As an

additional but optional goal, it was desirable to transfer the developed technique from FT-ICR to Orbitrap mass spectrometers and from Nm to other pathogens for the characterisation of other pilins and different PTMs.

This manuscript is divided into eight chapters, covering the different aspects of the development and application of top-down mass spectrometry to pilin proteins. A materials and methods section is provided as an annex.

Chapter one is divided into two parts. The first provides an introduction to the methodology behind bottom-up proteomics and the fundamental limitations that render a top-down approach necessary for the complete characterisation mixtures of proteoforms and proteomes. The technological requirements of top-down mass spectrometry are outlined along with a brief discussion of how the top-down experiment is performed in both targeted and discovery modes, and adapted to the biological question at hand. This is illustrated by several examples. The second part of the chapter is dedicated to the bacterium *Neisseria meningitidis* and places the proteins under study in a biological context.

In chapter two, bottom-up mass spectrometry is used to completely characterise all proteoforms and map all PTMs expressed by PilE purified from the Nm 8013 reference strain. These results are presented in the form of a publication (*Journal of Mass Spectrometry*, 2013) and provide a basis for development of the top-down approach.

In chapter three, the refined bottom-up approach is applied to PilE expressed by a number of mutants of Nm 8013 in order to help elucidate the function of the phosphoglycerol modification present on PilE. The publication included in this chapter (*Science*, 2011) uses these results to detail a biological role for the phosphoglycerol modification. A molecular basis for our hypothesis is also outlined and the implications on the meningococcal lifecycle discussed.

Chapter four chronicles the development of a top-down MS strategy for complete PTM characterisation of PilE expressed by Nm 8013. The top-down experiment is constructed on both FT-ICR and Orbitrap platforms. The effect of several important experimental parameters is investigated using samples of PilE, and general trends in the fragmentation behaviour presented that can be used to guide future implementation of the top-down methodology. The technical challenges involved in implementing the top-down experiment on each instrument platform are reviewed.

Chapter five describes the characterisation of PilE expressed by previously uncharacterised clinical isolates of *Neisseria meningitidis.* These strains were isolated from cases of sepsis and meningitis treated at the Limoges university hospital during recent, sporadic outbreaks of

meningococcal disease in central France. Some of the issues dealing with clinical isolates are outlined and two case studies of complete top-down characterisation of all proteoforms of PilE are presented. The first, concerning the 278534D isolate is given in the form of an accepted invited manuscript (*Proteomics*, "Top-Down Proteomics" special issue scheduled 2014) and compares the bottom-up and top-down approaches for PilE proteoform characterisation. Analysis of this strain provides a concrete biological example of a case where the top-down methodology is required to completely characterise several proteoforms present in the same sample. The second study characterises multiple proteoforms of PilE from the 427707C strain which are expressed in a complex mixture and with similar masses. The suitability of the top-down methodology to handle such a situation is discussed and evidence is presented characterising a novel glycan expressed by this strain.

Chapter six places the characterisation of PilE from these clinical isolates of Nm in a wider context. All of the isolates belong to the class II group of Nm strains which express invariable pilin sequences. This is in contrast to the hypervariable primary structures of PilE expressed by class I strains. The results presented in this thesis map the PTMs of pilin from class II strains for the first time, and show that PilE from class II isolates consistently harbours an unprecedented level of glycosylation. Additional results presented in this chapter indicate that this elevated glycosylation level has little effect on the morphology of the pilus fibre but a substantial effect on its surface. It is also shown that the extent of glycosylation is not directed by the genetic background but rather by the pilin primary structure. For the first time a hypothesis is outlined detailing the role of glycosylation on PilE from class II strains as a means of antigenic variation.

Finally, in chapter seven, the top-down methodology developed during this thesis is used to characterise the major pilin from a novel transformation pilus expressed by the *Streptococcus pneumoniae*. The results are presented in the form of a publication (*PLOS Pathogens*, 2013) and represent both the first report of a *bona fide* type IV pilus in a Gram positive bacterium and the first characterisation of such a pilus by mass spectrometry.

A general conclusion will end this manuscript including perspectives for future development and application of the top-down mass spectrometry approach.

Chapter 1

Introduction

Part I - Mass Spectrometry

1. Protein Structure

Proteins are biological polymers constructed from covalently linked building blocks called amino acids. Free amino acids (AAs) are small, chiral molecules containing a central α-carbon atom to which is attached a primary amine, carboxylic acid and a functionalised side chain (Figure 1A).

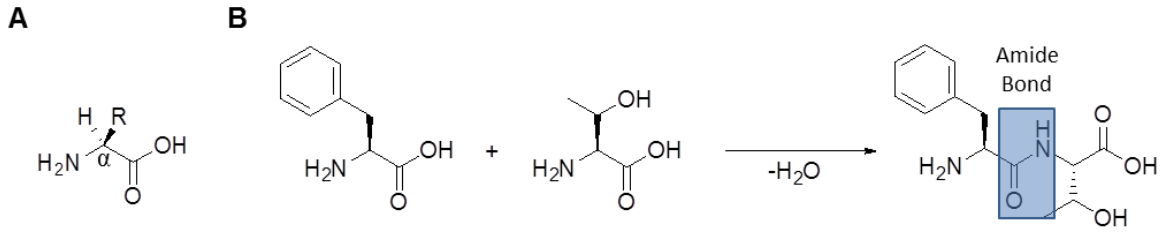


Figure 1 - A) Structure of a naturally occurring L-amino acid. Side chain represented by R B) Reaction of the amino acids phenylalanine and threonine to form the dipeptide Phe-Thr

A condensation reaction between the amine of one amino acid and the carboxylic acid of another results in the formation of an amide (or peptide) bond allowing several AAs to link together and form a non-branched covalently linked chain or peptide (Figure 1B). Peptides containing a few AAs (two to around twenty) are known as oligopeptides and longer continuous chains as polypeptides. Most proteins are long polypeptides greater than ≈ 50 amino acids in length whose number and the order in which they are arranged define the protein's primary structure.

Proteins and large peptides do not often exist as extended chains, but fold into distinct structural motifs such as alpha helices, beta sheets and loop regions (Figure 2 A-C). These regions of secondary structure are held together by extended hydrogen bonding networks between the backbone NH and C=O groups and their formation is governed by the conformational flexibility of neighbouring amino acid side chains. Additional structural stability may be impacted by metal ions or other ligands, and cysteine residues may form covalent –S-S– bonds which act as chemical cross links, anchoring the secondary structure into place and providing conformational rigidity. Further folding of these local structural elements gives the mature protein a tertiary structure (Figure 2 C). Tertiary structure ultimately determines a protein's function and even proteins with the same amino acid sequence may adopt different conformations that lead to drastically different activity. Misfolded proteins such as this are implicated in a number of important pathologies including cystic fibrosis, Alzheimers, Parkinsons and Gerstmann-Straussler-Scheinker disease[1]. The tertiary structure of a protein also determines its propensity to interact with other small

biomolecules and proteins. Indeed in many cases proteins do not function alone but interact with others in protein complexes. This defines a quaternary protein structure (Figure 2 D).

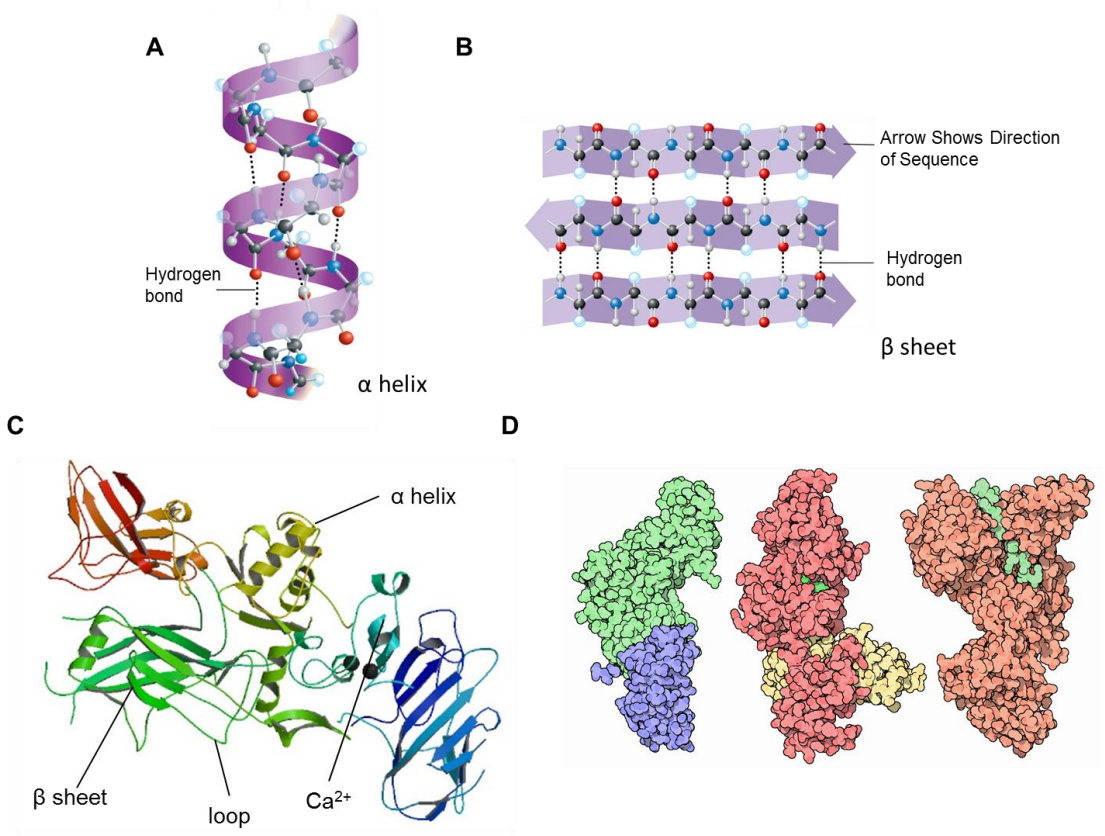


Figure 2 - A) & B) Schematic of protein secondary structure, α-helix and β-sheet respectively C) Protein tertiary structure - crystal structure of the PagA protective antigen, PDB 1ACC, exhibiting multiple secondary structural elements D) Protein quaternary structure showing interaction of 3 protein subunits PagA, Cya (PDB 1K90) and Lef (PDB 1JKY)

Organisms have evolved to express a distinct set of proteins dependent on their specific adaptations and requirements. The individual set of instructions that ultimately determines the amino acid sequence of expressed proteins is locked within the genetic code of an organism.

Protein Biogenesis

In both eukaryotes and prokaryotes (bacteria and archaea) genetic information is stored in the form of the polymeric biomolecule deoxyribonucleic acid (DNA). In the former, the DNA is housed in the nucleus of the cell as chromatin; tightly wound, linear, double helices of DNA coiled around protein complexes composed principally of histones. In the latter, DNA is mostly circular and double stranded, supercoiled around "histone like proteins" in the bacterial chromosome or as small, circular, individual sections known as plasmids. DNA is composed of 2'deoxyribonucleotide (or simply nucleotide) building blocks that are elaborated with four bases and joined together by

a phosphate di-ester linkage (Figure 3). Groups of three nucleotides code for a single amino acid and are referred to as codons. Since there are $4^3 = 64$ possible codon combinations we should perhaps expect 64 amino acid variants. However it turns out that there is considerable genetic degeneracy and several codons may code for the same amino acid. In addition codons are required to define the beginning (start codon) and the end (stop codon) of a protein. This results in base pair combinations coding for a total of 23 proteogenic amino acids of which the last three are more recent additions[2-7].

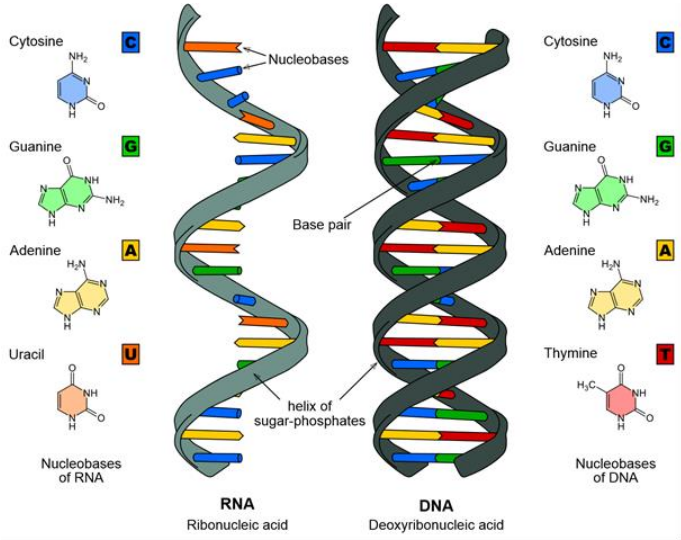


Figure 3 - Representation the molecular structure of the DNA double helix and single stranded RNA

The DNA of an organism contains many thousands of nucleotides. The human genome for example contains 3 billion base pairs spread over 46 linear pieces or chromosomes. This vast array of information is further increased since there are three possible ways to read the DNA (three reading frames) and some organisms can decode DNA in both the 5'-3' and 3'-5' directions. However not all of the DNA sequence codes for proteins. Indeed human DNA is now only thought to contain just over 20,000-protein coding genes or exons[8]. Protein coding genes are characterised by specific start and stop codons that demark each section of DNA and denote an open reading frame (ORF). In eukaryotes there is only a single start codon, although several alternate start codons are known in prokaryotes. In all domains of life there are multiple stop codons. The start and stop codons allow the creation of lists of hypothetical protein candidates from sequenced genomic data. In combination with other motifs such as the Shine-Dalgarno sequence in prokaryotes they also serve to direct the cellular machinery to where protein-coding genetic material begins and ends[9].

The way an organism translates this genetic information into proteins is rather complex and differs substantially between eukaryotes and bacteria, however in both cases it essentially involves two steps. In the first, the double stranded DNA helix is separated and each base in the anti-sense strand is transcribed to a corresponding base on an mRNA intermediate (Figure 4). In eukaryotes considerable processing of the mRNA then occurs and it is shuttled out of the nucleus to the cytoplasm.

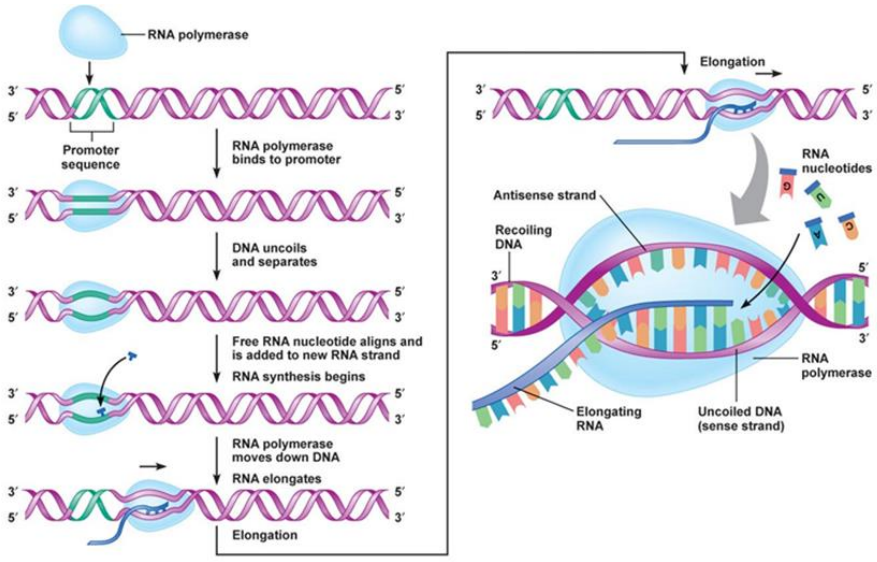


Figure 4 - Simplified description of the first stage in DNA transcription

In the second step, the mRNA is delivered to a very large set of protein complexes, the ribosome that translates the mRNA bases to their corresponding amino acids and polymerises them to form the nascent protein. Some detail of this process is shown in Figure 5.

2. Diversity Imparted from the Genome

Understanding the fundamental process of how stored genetic information is turned into an expressed protein allows one to appreciate the potential for gene product diversity. The situation described in the preceding section is an ideal one and in reality diversity may be imparted into the expressed proteome from a number of alternative outcomes of genetic processing (Figure 6). A common example of this is alternative splicing where combinations of exons from the same gene may be combined in different ways to form different gene products. These are known as alternative splice forms (ASFs). Errors in the copying process or other regulated processes can also give rise to single nucleotide polymorphisms (SNPs) in the gene which may (or may not) change the identity of an amino acid in the expressed gene product (single amino acid

polymorphism, SAP). Furthermore, there may be multiple possible start and stop codons for the same gene, which also result in distinct gene products.

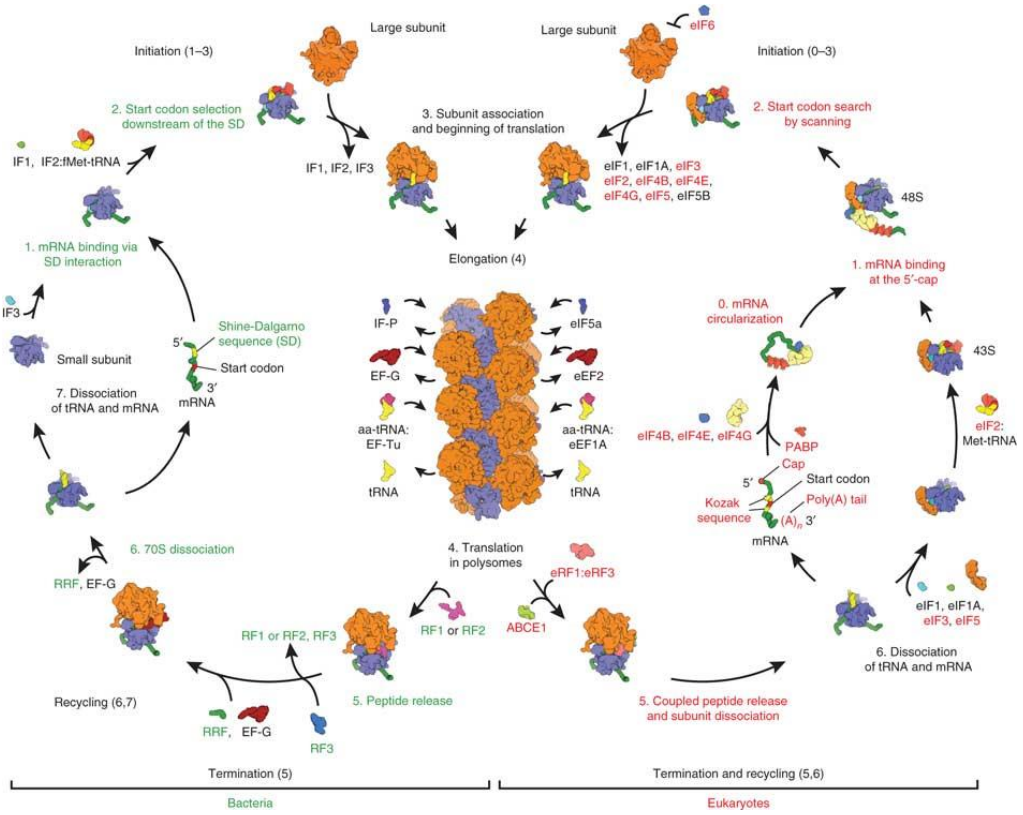


Figure 5 - Detail of translation events in both eukaryotes and prokaryotes showing the role of important cofactors and the dissociation and association of the ribosome during translation. From Melnikov *et al.*[10]

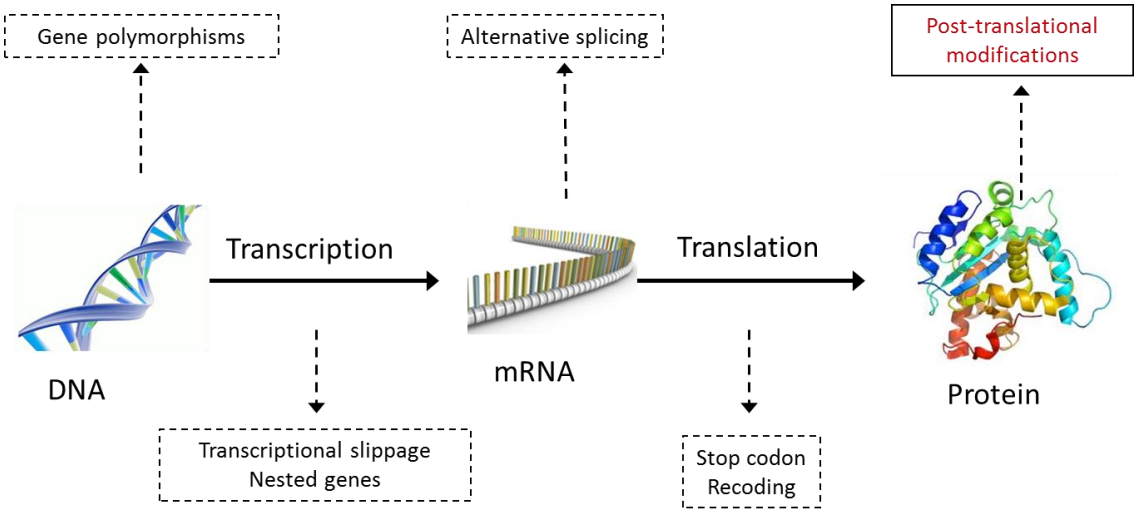


Figure 6 - Principal mechanisms leading to diversity in expressed gene products

3. Posttranslational Modification

Once the protein has been translated it may also become the target for further chemical elaboration by other proteins. This process is referred to as posttranslational modification (PTM) and encompasses a plethora of chemical changes including: proteolysis; modification of amino acid chirality (Pro cis→trans isomerism); modification of amino acid side chains or termini (deamidation, citrullination, pyroglutamate formation); the covalent addition of simple moieties such as hydroxyl, phosphate, methyl or acetyl groups and more complex molecules such as sugars (N-linked, O-linked); lipids and long aliphatic chains (myristolyation, palmitoylation); glycolipids (glypiation); polypeptides or even small proteins (SUMOlyation, ubiquitinination, neddylation, pupylation). The examples given here are by no means exhaustive but represent some of the most common modifications reported in the Swiss-Prot database[11]. Novel PTMs such as 3-phosphoglyceryl-lysine continue to be reported as understanding of cellular processes deepens[11]. Until relatively recently PTM was considered an exclusively eukaryotic process, however despite their smaller genomes many PTMs have since been discovered in bacteria and archaea, and perhaps because of their more promiscuous lifestyle they have become a rich fountain of new and interesting PTMs.

PTMs are increasingly being linked with regulation of protein function and as such are an essential piece of information to achieving total understanding of cellular processes. More and more evidence is being presented that many aspects of fundamental cell biology rest on pathways regulated by PTM; this includes transcription and gene expression directed by the histone code. The variety that PTM brings to gene products also helps explain the unexpectedly small number of proteins coded for by the human genome and others. Indeed it is being realised that the complexity that characterises biological processes occurring in higher organisms may not necessarily require a huge number of proteins but tight regulation and modulation of cell function may be imparted though PTM regulated process involving a much smaller set of modified proteins.

Several expressions have been employed to describe the different posttranslationally modified forms of a protein including, isoform, protein species, protein variant, protein form etc. Isoform has a strict IUPAC definition that restricts its use to gene products having a high sequence identity and issuing from the same gene family or polymorphisms. The term proteoform has recently been coined to encompass all variants of the same gene product including changes due to genetic variations, alternatively spliced RNA transcripts and post-translational modifications. It is this term that will be used throughout this thesis[12].

4. Identification of Proteins Using Bottom-Up Proteomics

Identifying the protein content of a biological sample has been an enviable challenge ever since the first studies on protein rich substances such as egg white and blood plasma. Despite these samples being highly abundant in protein (albumin and fibrinogen respectively) identification of the protein molecules presented a considerable challenge because their high molecular mass and diverse chemical reactivity rendered them difficult to characterisation by classical chemical methods. Clearly the most definitive way to identify a protein would be to determine its complete amino acid sequence.

In 1951 Frederick Sanger & Hans Tuppy were the first to achieve this remarkable feat, sequencing the 30 amino acid beta chain of the protein insulin[13, 14]. This was quickly followed by sequencing of the alpha chain[15, 16] and localisation on of the intra-chain and inter-chain cysteine bonds[17]. Sanger received the Nobel Prize for this work in 1958 and his basic methodology persists to this day.

Sanger pioneered a "divide and conquer" strategy to protein sequencing where the protein, too complex to analyse intact, was cleaved into smaller pieces using a combination of acid hydrolysis and proteolytic enzymes. The resultant peptides were then separated and individually sequenced though highly laborious steps of chromatographic separation, chemical derivatisation and further chromatographic separation. Since different proteolysis methods resulted in different peptides, the small peptide fragments often overlapped and could thus be pieced back together to eventually reveal the sequence of the entire protein. This divide and conquer strategy has remained essentially the same and forms the cornerstone of contemporary "bottom-up" proteomics.

4.1. Beginnings of Bottom-Up Proteomics

The techniques used for protein identification have however developed substantially since the 1950s. The introduction of sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) enabled proteins to be easily separated as a function of their molecular weight and the introduction of two dimensional electrophoresis (2DE) allowed more complex mixtures of proteins to be separated by both molecular weight (MW) and isoelectric point (pI), permitting hundreds to several thousand proteins to be separated and on large format polyacrylamide gels, focused into spots and visualised using dyes and fluorophores[18].

Whilst gel based techniques revolutionised protein separation, correlating an observed spot to a known protein remained challenging. Estimates of only MW and pI are clearly insufficient to provide definitive protein identification (protein ID). Initially proteins were extracted from gels

onto membranes and N-terminal sequences determined using Edman degradation. Databases of genomic data provided lists of putative proteins against which the N-terminal sequence could be compared, with a positive match providing a protein ID. However, extraction of the intact protein from the gel could be difficult, especially for high MW proteins. N-terminal sequencing was limited to the first 30-50 N-terminal residues and was also problematic for proteins with a high MW or those with a blocked N-terminus. The process was cyclic, identifying one amino acid at a time and therefore slow, taking several hours per protein even on automatic sequencers. It was also not particularly sensitive, requiring several micrograms of pure sample and protein identification proved impossible if the examined protein was not already present in a database or if the N-terminus of the protein had been modified posttranslationally. Internal posttranslational modification could also arrest sequencing prematurely. Using this method the differentiation of proteins with homologous N-termini was often simply impossible although integration of MW and pI from the 2D gel as additional parameters could narrow down the possibilities.

An interesting improvement was reported where extracted proteins were chemically digested using cyanogen bromide or skatole (4-methyl-2,3-benzopyrrole), to give a small number of large peptides. The digestion products were then sequenced by Edman degradation and matched against databases in an approach called mixed peptide sequencing[19]. However, a method to digest proteins into peptides within the gel matrix was developed which abrogated the need for protein extraction so long as the resulting peptides could be efficiently analysed[20]. In this regard the arrival of biological mass spectrometry proved a game changer, opening up the possibility for higher throughput large scale protein identification by two principal methodologies.

4.2. Peptide Mass Fingerprinting

Mass spectrometry had been used for many years for the structural examination of organic molecules. The organic compound of interest was often ionised *in vacuo* using high energy electrons (electron impact, EI) forming unstable radical cations which would undergo fragmentation or rearrangement and produce a number of daughter ions (fragment ions). Accelerating these ions through a magnetic field causes a deflection of their path, proportionate to their mass (m) and inversely proportional to their charge (z). Scanning though a range of field strengths to separate the ions followed by detection in a photomultiplier can be used to calculate the absolute mass of each fragment ion produced and structural information can be obtained from careful interpretation of the resultant mass spectrum.

Applying this procedure to peptides or other fragile biomolecules proved difficult due to their poor ionisation and thermolability. New ionisation techniques such as secondary ion mass spectrometry (SIMS), fast atom bombardment (FAB) and chemical ionisation (CI) were developed

that enabled peptide ionisation for the first time. The implementation of experimentally simpler, softer ionisation techniques such as matrix assisted laser desorption (MALDI)[21, 22] and electrospray ionisation (ESI)[23] bought in a new era of sensitive and routine analysis of underivatised, intact peptide and protein ions (Figure 7).

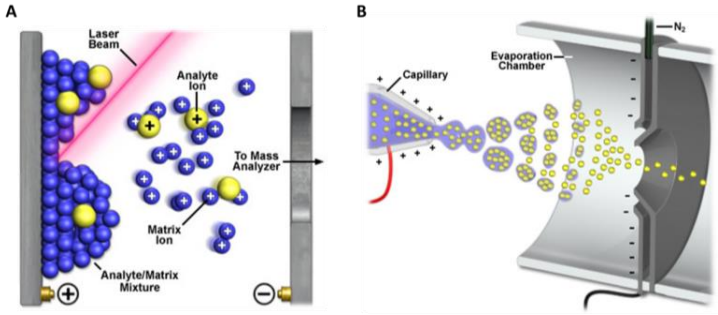


Figure 7 - Soft ionisation methods A) MALDI A mixture of analyte and chemical matrix crystallised onto a metal support is ionised *in vacuo* by a UV laser. The mixture is vaporised and a cation (usually proton) transferred from the matrix to the analyte. Positive ions are transferred into the mass spectrometer B) ESI Analytes dissolved in a volatile, usually acidic solution are sprayed from a positively charged needle or capillary Again the charge carrier is often a proton.

Refinement of ionisation techniques coupled with the development of new mass spectrometers equipped with time of flight (ToF) and quadrupole (Q) type mass analysers allowed the digest products of gel spots to be readily analysed in peptide profiling experiments. MALDI-time of flight mass spectrometry (MALDI ToF MS) became a popular choice since this type of analysis could be readily performed over a relatively large low mass range 400-4000 *m/z* and affords singly charged peptide ions. When proteins are digested with a proteolytic enzyme such as trypsin that cleaves at specific amino acid residues (Lys and Arg in the case of trypsin), mass analysis of the peptides formed gives a distinctive pattern for the protein of interest or peptide mass fingerprint (PMF).

With the revolution occurring in informatics and gene sequencing occurring at the same time it was quickly realised that PMF, or rather the masses of peptide signature ions, could be used in conjunction with databases of genomic information for the identification of proteins. If one had a list of all of the expected proteins in an organism, a list of expected peptides could be generated by taking into account the specificity of the enzyme used for digestion. Protein databases derived from genomic data could then be annotated with the masses of expected peptide ions, and comparison of experimental ions from the PMF profile and these *in silico* digestions used to identify the proteins present in 2D gels.

Thus the classical proteomics workflow was born. Complex protein samples were separated by 2DE, spots stained with dyes then excised, digested and PMF performed. Matching of PMF data with protein database finally furnished protein identification (Figure 8).

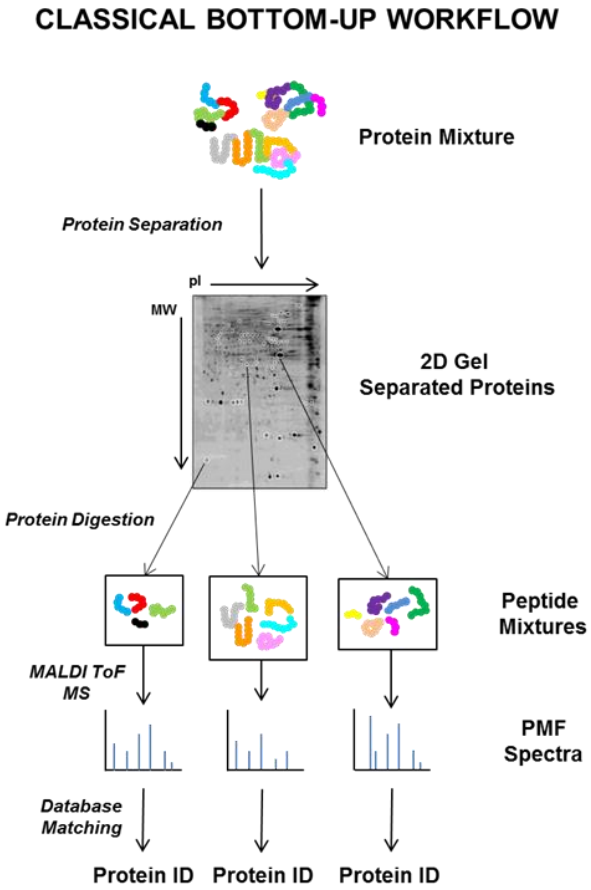


Figure 8 - Classical proteomics workflow for protein ID by PMF

PMF can be quite a successful approach to achieve protein identification especially when the high-resolution mass measurement is performed. This has been made possible by substantial technical improvements in ToF instruments over the last decade or so. The PMF approach does however suffer several drawbacks.

Firstly, unexpected peaks in the peptide digest are particularly problematic as they increase the risk of false positive protein ID. Miscleavages from inefficient proteolysis and some posttranslationally modified peptides can now be handled by appropriate database matching algorithms. However, complex mixtures of peptides from multiple proteins are still not well supported due to a lack of dynamic range in the MS step and increased risk of false positive protein ID. Secondly, merely examining peptide masses does not provide any amino acid sequence information and thus several peptide fragments are required to produce a confident protein match. Missing peptides due to large size, poor ionisation efficiency and ion suppression mean that in some cases PMF is only sufficient to narrow down the list of possible proteins rather than provide definitive protein identification. In addition highly posttranslationally modified proteins

and those not already in a database simply cannot be identified. If the observed peptide ions could be sequenced inside the mass spectrometer an additional level of information could be provided that would greatly facilitate protein identification.

4.3. Peptide Sequencing By Tandem MS

Early work by Biemann had shown that the fragments produced during electron ionisation mass spectrometry of individual, derivatised peptides could be used for amino acid sequencing[24] and once intact peptides could be successfully ionised it was quickly demonstrated that they could be fragmented inside the mass spectrometer in tandem MS (MS/MS) experiments. Various fragmentation techniques were developed for different instrumental platforms such as post source decay (PSD) in the free field region of the flight tube in MALDI ToF instruments[25, 26], collisionally activated dissociation (CAD) in specially designed collision cells in triple quadrupole mass spectrometers[27] and resonant excitation of the parent ion, in several forms of ion trap.

Fragmentation of peptides may either result in cleavage of the protein backbone or cleavage of the amino acid side chains. There are three principal places that cleavage can occur on the peptide backbone: between C_α and C=O of a particular amino acid; between C=O and NH (the amide bond) and between NH and C_α of the next amino acid. Roepstorff and Fohlman developed a nomenclature for peptide fragmentation[28] which was later extended by Biemann[29] that describes fragment ions issuing from cleavage of these three bonds. If the fragment ion contains the peptide N-terminus then it is termed *a*, *b* and *c* respectively. If the daughter ion is C-terminal then it is designated *x*, *y* and *z*. Ions are often followed by a subscript number indicating which peptide bond has been fragmented (counting from the respective terminus) (Figure 9).

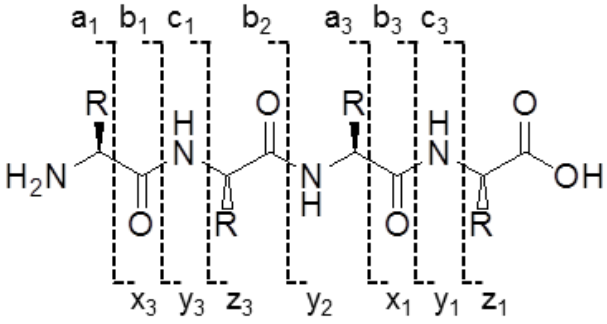


Figure 9 - Roepstorff nomenclature for peptide ion fragmentation[28]

Both PSD and CAD impart vibrational energy to the parent ion though successive low energy collisions with inert gasses such as N_2, He and Ar and produce characteristic *b* and *y* type fragment ions for which the activation barrier is generally the lowest. It was found that unlike Biemann's EI spectra, fragmentation spectra of peptides issuing from PSD and CAD were rarely dominated by one or two stable product ions, rather they contained many intense peaks. These peaks

correspond to series (or ladders) of ions built up from the N-terminus (*b* type ions) and or C-terminus (*y* type ions) of the peptide. The mass difference between each pair of ions in the series (rung of the ladder) corresponds to that of an individual amino acid. Often the two ion series meet in the middle or even overlap allowing the sequence of the entire peptide to be derived. If complete sequencing is achieved, uncertainty only remains for the isobaric amino acids leucine/isoleucine and near isobaric amino acids lysine/glutamine when experiments are performed with low resolution instruments.

At around the same time techniques such as pre-flight tube ion gating in MALDI ToF instruments, modulation of the rod potentials in hybrid mass spectrometers such as QqQ or Q-ToF instruments and modulation of the trapping potentials in ion trap mass spectrometers quickly became available to isolate, or separate, individual ions from mixtures such as PMF profiles inside the mass spectrometer. These ions could then be fragmented to provide sequencing information.

The advent of tandem mass spectrometry brought about two new methods for protein identification. The first was an augmentation of PMF identification where ions from a PMF profile would be selected and sequenced by MS/MS inside the mass spectrometer. The sequence tags produced helped to confirm protein identifications and brought added confidence to results. Instead of requiring many peptide fragments only one or two sequenced fragments were required to achieve a confident protein ID. Automation of this process would lead to high throughput identification of proteins from 2DE.

The second was the "direct sequence tag search" approach pioneered by Mann, where the PMF data could be discarded and the more specific sequence tags from MS/MS of peptide ions used directly to identify proteins from a database[30]. The amino acid sequences furnished though MS/MS could be searched directly against protein databases in order to identify the protein of interest.

This provoked a paradigm shift in protein identification and with the correct instrumental development one could now imagine peptides being constantly delivered to the mass spectrometer, selected and sequenced one after the other. Realisation of this idea relied on efficient ways to separate peptides and an efficient method to deliver them to the mass spectrometer.

4.4. LC-MS/MS for High-Throughput Bottom-Up Protein Identification

In contrast to MALDI, electrospray ion sources allow a free flowing liquid interface to be directly coupled to the mass spectrometer. To achieve efficient spray conditions several criteria must be met. In particular the analyte must be sprayed in a volatile solvent and be free of non-volatile salts

or particulates that may clog the electrospray needle. This requirement is even more stringent for smaller nanoESI interfaces that allow lower flow rates and thus greater sensitivity. Several such flow through methods such as capillary electrophoresis (CE) and liquid chromatography (LC) have been developed for the separation of both peptides and proteins from complex mixtures. CE involves separation as a function of differential mobility in an electric field and essentially depends on the charge carried by the analyte and its molecular mass. Whilst benefiting from incredibly high resolution, due to electroosmotic flow in the electrophoretic capillary, interfacing CE with mass spectrometry whilst maintaining high sensitivity has proved challenging and continues to be the subject of on-going work[31]. Liquid chromatography has enjoyed much more success.

LC involves separation of molecules based on their competing preference for stationary and mobile phases. Separation of peptides by this approach turns out to be much simpler than separation of proteins. For this reason the following description relates specifically to peptides, although the principles are essentially the same for proteins and the topic will be returned to later in this thesis. In the most commonly employed method, reverse phase LC (RPLC), a mixture of peptides is first adsorbed onto a hydrophobic solid phase. This usually is comprised of silica beads functionalised with long chain (often C_{18}) alkyl groups tightly packed into a column. The bound analyte is then eluted in an increasing gradient of hydrophobic solvent, such as the organic solvents methanol and acetonitrile. Hydrophilic peptides with little affinity for the C_{18} phase are eluted first followed by those of increasing hydrophobicity as the percentage of organic solvent in the mobile phase is increased. Peptide separation is thus is achieved as a function of analyte hydrophobicity.

With the addition of a volatile acid (such as formic or acetic acid) to aid ionisation, relatively minor modifications to the electrospray source and an appropriate flow rate, LC systems can be easily coupled directly to a mass spectrometer. MS analysis can thus be performed on peptides as soon as they elute from the column. If chromatography has been performed correctly they arrive to the mass spectrometer in concentrated, well separated bunches. There is generally ample time for both mass measurement and tandem MS/MS to be performed.

In reality several peptides may co-elute from the LC column and arrive at the MS together. In such cases automated parent ion selection, then fragmentation, enables consecutive MS/MS experiments to be performed without the need to make repeat MS measurements. MS/MS spectra and thus sequence information is linked to a specific parent ion in the acquisition data file. (This of course required significant advances in acquisition software, automation of instrument control and improvements in mass analyser resolution.)

Identification can now be achieved by matching peptide masses (from *m/z*) obtained in the MS spectrum against an appropriately digested protein database (with a tolerance related to analyser resolution). This first-pass filter narrows down the possible number of candidate peptides. Then masses derived from the MS/MS experiment can be matched with the theoretical fragment ions expected from the peptide candidates. Confidence metrics or scores for these peptide-spectrum matches (PSMs) have been invented in order to enumerate the quality of the match and confidence metrics imposed to discard matches under a certain significance threshold.

In this "shotgun" approach to protein identification, it is the peptides which are examined and identified experimentally. Only when a certain number of appropriate peptides have been identified at a given significance or confidence level. is the related protein also identified. The presence of a protein is directly inferred from peptide identification. Robust statistical tests such as decoy database searching and various types of false discovery rate (FDR) have been introduced in order to reduce the number of false peptide and protein identifications. Additional methods for validation of peptide and protein ID may also be used to add a further level of confidence. This modern LC MS/MS bottom-up proteomics workflow is illustrated in Figure 10.

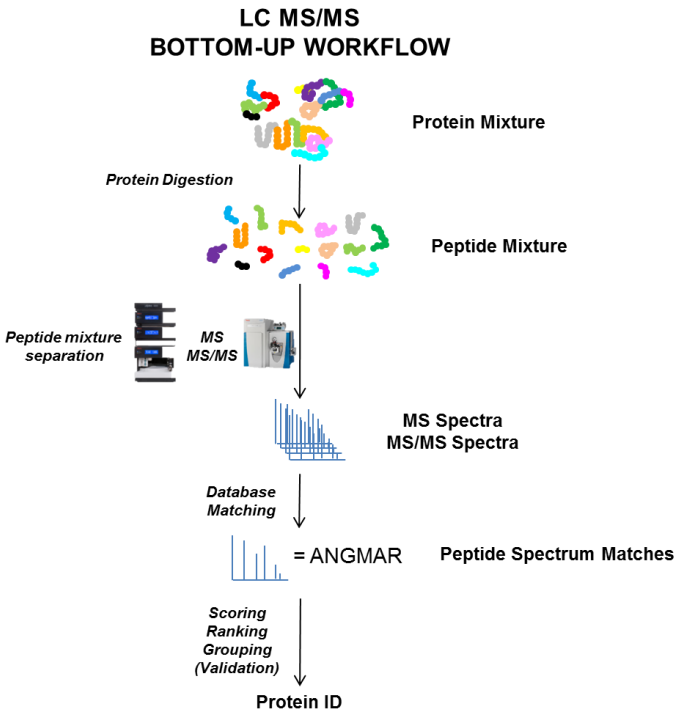


Figure 10 - Contemporary LC MS/MS bottom-up proteomics workflow

4.5. State-of-the-Art Bottom-Up Proteomics

Interestingly from its beginnings as a protein centric discipline, proteomics has become incredibly peptide centric. The arrival and adoption of robust LC-MS/MS methods represented a paradigm

shift for protein identification. No longer did proteins have to be painstakingly separated by 2DE then identified by MS but many proteins can be effectively identified after one pot digestion and LC-MS/MS. This idea is particularly compelling because it means that entire proteomes can be examined rather than isolated proteins and complete sample analysis achieved in much shorter time frame.

In order to divide very complex proteomes into smaller subsets or enrich for particular proteins or peptides of interest, many diverse fractionation techniques such as: strong and weak cation exchange (SCX, WCX); anion exchange; isoelectric focusing (IEF); hydrophilic interaction liquid chromatography (HILIC); size exclusion chromatography (SEC); gel based fractionation; and affinity purification strategies have been developed to complement separation by RPLC[32]. Specific protocols such as combined fractional diagonal chromatography (COFRADIC)[33] have also been developed for offline RPLC fractionation. MS workflows such as the MudPIT strategy pioneered by the Yates lab are then used for analysis of MS/MS data from these fractionated samples[34].

The quest to extend the dynamic range, to mine ever further into the proteome and to achieve higher and higher numbers of identified proteins has undoubtedly been an important driving factor in instrument development. Fast scanning, high resolution mass spectrometers with a large dynamic range and optimised peptide fragmentation such at the Thermo QExactive are producing very high quality, high confidence, high resolution MS/MS data. This means less and less is information is demanded before a positive peptide ID and therefore protein identification is accepted as fait accompli.

Instrument improvement and the introduction of ultra-high pressure LC systems is even allowing entire proteomes to be analysed in single LC runs on long nano-LC columns[35, 36]. As example of this performed during this thesis on the proteome of *Neisseria meningitidis* is shown in Figure 11. A recent report from the Mann group has identified over 4,000 proteins from Lys-C digested *Saccharomyces cerevisiae* lysates using this technique[37]. This is close to the total number of expected cellular proteins expressed under standard growth conditions. A very recent report from the Coon group has achieved similar results but with only a 1.3 hour acquisition time using a fast scanning Orbitrap Fusion mass spectrometer[38].

Increasingly accurate ways are being devised to quantify proteins from bottom up data including label-free and tag based techniques such as isotope-coded affinity tag (ICAT)[39], tandem mass tag (TMT)[40], isobaric tags for relative and absolute quantitation (iTRAQ)[41] and stable isotope labelling by/with amino acids in cell culture (SILAC)[42]. In the previous report by the Mann group

on the yeast proteome, over 3,500 proteins could be accurately quantified using a SILAC strategy with a similar method set up during a differential experiment.

A continually developing area of bottom-up proteomics is the identification of large numbers of post translational modifications at the proteome level. Since specific posttranslationally modified proteins often represent only a small fraction of the expressed proteome, enrichment of modified species at either the peptide or protein level is usually performed. Many PTM specific enrichment protocols have been devised such as immobilised metal affinity chromatography (IMAC) for phosphopeptides[43, 44], lectin enrichment in the case of glylcopeptides and immunoprecipitation of GG tagged peptides in the case of ubiquitination[45]. Specialised LC MS/MS strategies are then applied to increase coverage of modified peptides and identify PTM sites. For example a combination of SCX fractionation after protein digestion, IMAC phosphopeptide enrichment followed by RPLC MS/MS using CAD, ETD and a decision tree based MS method enabled the Coon group to identify 10,844 district phosphosites and 4,339 proteins in an analysis of the human embryonic cell phosphoproteome[46].

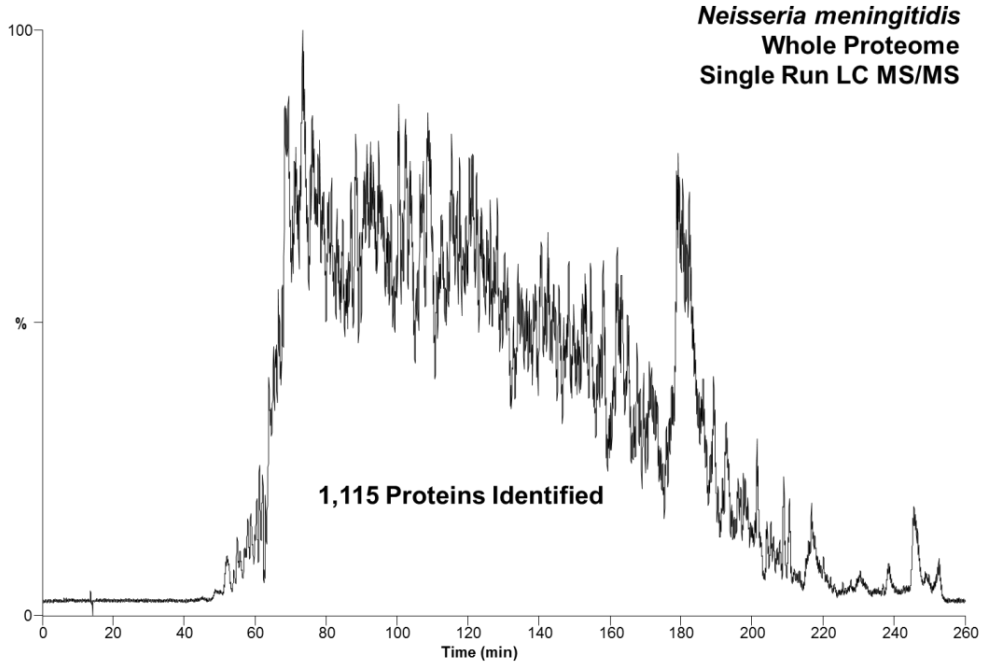


Figure 11 - Single run RPLC MS/MS on a cell lysate of *Neisseria meningitidis* using a 30cm nano-LC column and 4 hour chromatographic gradient. 1,115 proteins were identified at an FDR<1%. Data acquired by Miss Debora Pasquali

This thesis does not deal explicitly with bottom up-proteomics and the state of the art of this field including various methods for quantification, PTM identification, applications to interactomic and

systems biology and the need for good search engines and robust validation is presented elsewhere[47, 48]. What is of particular relevance is the divide and conquer bottom-up methodology used and the fundamental constraints that using this approach may impart for both the identification and characterisation of proteins.

5. Limitations of the Bottom-Up Approach for Proteoform Identification

The bottom-up approach is undoubtedly able to identify peptides very successfully. This is important as it partially fulfils the first goal of proteomics; confirming a protein encoded by the genome is actually expressed. Identification of a protein is assumed if one or more peptides has been positively matched against an appropriate genomic database within defined confidence limits. In effect a link is made between the peptide and the assumed gene product. The fundamental limitations of the bottom-up approach concern the validity of this assumption and stem from the digestion of proteins early on in the proteomics workflow.

The issues that arise from this assumption are outlined by the protein inference problem, described in detail by Nesvizhskii and Abersold for which there are two general cases[49]. In the simplest, the identified peptide may not be unique and potentially belong to several unrelated proteins present in the genomic database. This situation is attenuated in very large proteomes and has no finite solution; probabilistic methods must be used to rank the possible matches. Alternatively the peptide-protein link may be an oversimplification of reality and the assumed gene product may simply be incorrect. This situation is illustrated by two important examples.

The first is when the incorrect gene product has been assumed at the level of the protein primary sequence. A good example of this concerns ASFs. ASFs are common proteoforms and may represent a large proportion of the expressed proteome, for example a specially designed study on *Aspergillus flavus* has shown that ASFs may account for over 40% of identified proteins[50]. They also be extremely important factors in disease states, for example an ASF of the protein RON promotes cell motility in gastric carcinoma[51]. Protein databases do not generally contain all alternative splice gene products in an attempt to minimise redundancy and this makes the identification of ASFs in standard proteomics experiments unlikely. Even if databases were complete, many proteoforms of this kind would likely be missed simply due to poor sequence coverage because sampling peptides that cover all of the non-homologous regions is required for identification of the correct ASF. This problem is compounded in the case of single amino acid polymorphisms (SAPs). Like ASFs these are SAPs are ubiquitous and are often missed in proteomics studies as databases are not fully annotated[52]. Both ASFs and SAPs may require 100% protein sequence coverage in order to be confidently identified.

Incomplete sequence coverage is the first great weakness of the bottom-up methodology. Even in the most comprehensive proteomics studies, the average sequence coverage of identified proteins is below 50%. Rarely is a 100% sequence coverage achieved for any identified protein and in the case of low abundance proteins, where identification comes from only one or two peptides, sequence coverage of less than 10% is relatively normal. Inroads into this problem can be made by more extensive fractionation, the manufacture of faster mass spectrometers with greater sensitivity, greater dynamic range and more efficient fragmentation methods, that achieve more extensive protein coverage; but 100% sequence coverage from an entire proteome is simply an unrealistic expectation from bottom-up MS. Even for more abundant species some peptides will be too small to identify and will always remain underrepresented due to poor ionisation[53] and the proteotypic peptide phenomenon (even if multiple enzymes are used for digestion). Using a bottom-up methodology to achieve 100% sequence coverage of a majority of proteins in an entire proteome is an all but impossible goal and has even been described a surreal objective[54].

The second example concerns the case of PTM. Any protein identified with less than 100% sequence coverage could be potentially modified post translationally on an unsampled part of the sequence. One simply cannot be sure as the reference mass of the parent proteoform is unavailable due to the digestion step early on in the workflow. Furthermore, PTM often results in the expression of multiple proteoforms of the same gene product. Now identification may not concern one gene product but multiple proteoforms, each distinguished by the number and localisation of PTMs on the protein backbone.

Chait *et al*. and others have pointed out even if 100% sequence coverage could be achieved for one proteoform, bottom-up MS may still produce an incomplete picture because low abundance proteoforms may not be sampled in the conventional bottom-up experiment[55]. This is another manifestation of the 100% sequence coverage problem. There is however a more fundamental problem where PTM is concerned. PTMs rarely work alone and multiple PTMs are often present at different sites on different proteoforms of the same gene product. In certain cases a bottom-up methodology becomes fundamentally unable to completely describe populations of parent proteoforms even if 100% sequence coverage of each proteoform is achieved, including PTM identification and site assignment. The problem becomes a combinatorial one that becomes impossible to resolve using a bottom–up methodology[56].

6. Top-Down Mass Spectrometry

Top-down mass spectrometry (TDMS) has been proposed as alternative to the bottom-up approach. In TDMS protein identification and characterisation is performed entirely at the protein

level. Since no proteolytic digestion is involved there is no loss in connectivity during the analysis and experiments are performed on the expressed gene products themselves. TDMS therefore does not suffer from the drawbacks of the bottom-up methodology and offers the possibility of identifying both proteins and proteoforms.

It is helpful to straight away to distinguish two different types of TDMS experiment since whilst the fundamental pieces of information required from each may be similar, the experimental conditions may be quite different. The first may be described as "targeted mode" and is typified by the top-down identification and detailed deep characterisation of a single highly enriched protein or set of proteoforms. Often the putative identity of the protein *i.e.* the gene from which it has issued is a known piece of information. The second is the recently reported "discovery mode" where a complex sample is examined with the aim of identifying as many proteins and characterising as many proteoforms as possible. There is a greater emphasis placed on identification in this mode whilst retaining the deepest level of characterisation practically possible. General workflows fort the two modes are depicted in Figure 12.

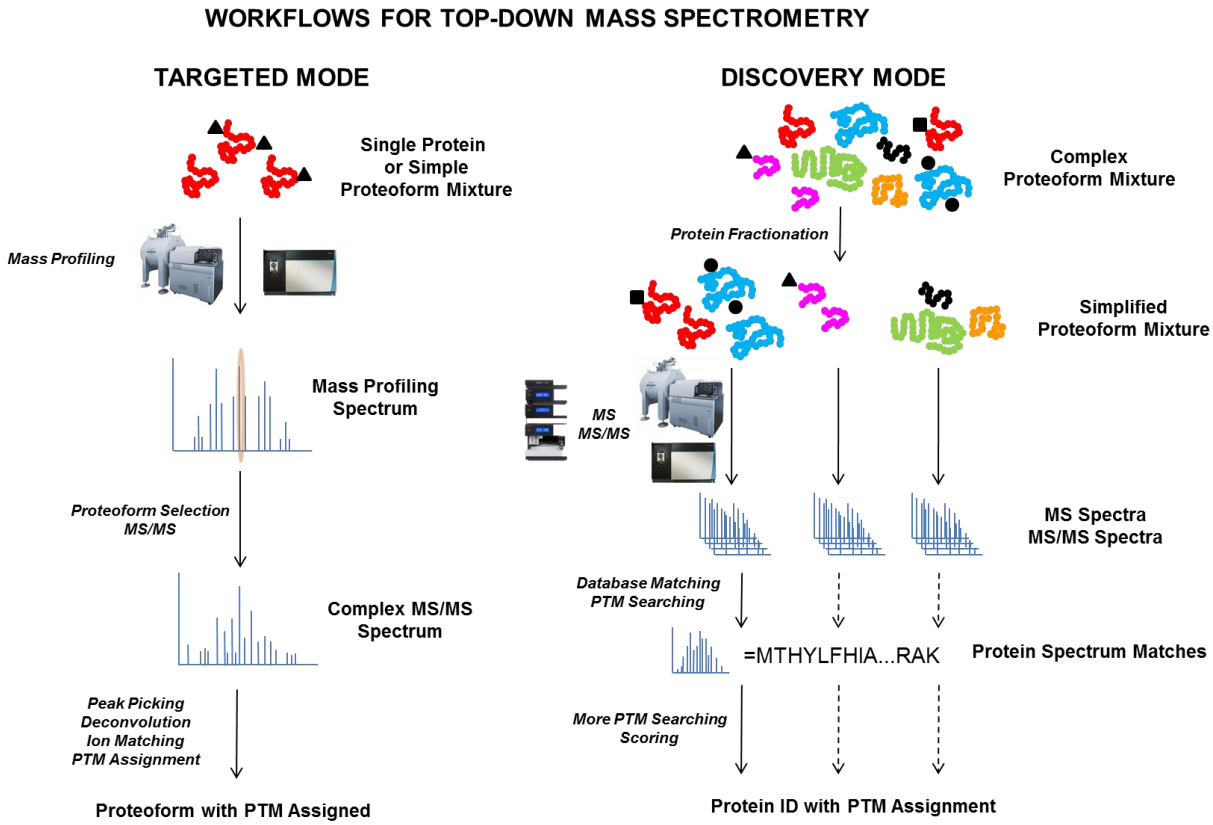


Figure 12 - General workflows for targeted and discovery mode top-down mass spectrometry

TDMS is technically demanding, hence why bottom-up has developed at a much faster rate. In all TDMS experiments there are some fundamental requirements that stem from the challenges that

come with working with proteins. Before it is demonstrated how a top-down experiment is performed in practice, these additional requirements will be outlined and briefly discussed.

6.1. Efficient Sample Ionisation

TDMS requires efficient methods to transfer intact proteins into the gas phase. Just as the soft ionisation techniques revolutionised peptide ionisation, they achieved a similar success for proteins. The two most commonly used ionisation methods in TDMS are MALDI and ESI, much for the same reasons as in bottom-up proteomics.

MALDI

Tanaka *et al*. presented his design for the laser desorption mass spectrometer in 1987 and with it the first entire protein mass spectra of lysozyme and carboxypeptidase-A. His results were formalised into a paper in 1988[21] and were followed closely by those of Hillenkamp who demonstrated an extended mass range up to 67 kDa for bovine albumin with his UVLD-ToF instrument[22]. MALDI enjoyed significant use in the early days of proteomics for mass measurement of intact proteins after extraction from 2D gels[57]. Whilst MALDI is particularly good for protein profiling, problems fragmenting singly charged protein ions have meant that other than a few isolated examples[58, 59], the alternative ESI is the ionisation method of choice for top-down experiments. New matrixes and the matrix assisted ionization vacuum ionisation (MAIVI)[60] approach are being developed to produce multiply charged ions and may change this in the future.

ESI and nano-ESI

Unlike MALDI, electrospray (ESI) produces highly charged proteins ions due to the nature of the desolvation process. Building upon earlier work on ion jets, electrospray was introduced by Yamashita and Fenn in 1984[61], interfaced with mass spectrometers and LC[62] and quickly applied to the analysis of large biomolecules including proteins ranging in size from 5.7 kDa (insulin) to a 133 kDa (BSA dimer)[23]. Electrosprayed proteins are almost always observed as a number of different peaks each corresponding to a different charge state of the protein at *m/z* values dependent on the protein mass and the number of ionising protons. This has been found to vary with electrospray solution composition, pH, temperature and instrumental parameters such as source temperature and pressure and these factors must be taken into account to ensure efficient ionisation.

Conveniently the observed charge state of a protein correlates loosely with protein size. This means that even very large electrosprayed proteins are usually visible in the *m/z* 200-4000 range available to most mass analysers. This can be seen most clearly in the ESI-Q-ToF spectrum of a

148 kDa immunoglobulin (IgG) acquired by M. Christian Malosse in our lab (resolution ≈10 000 at *m/z* 400).

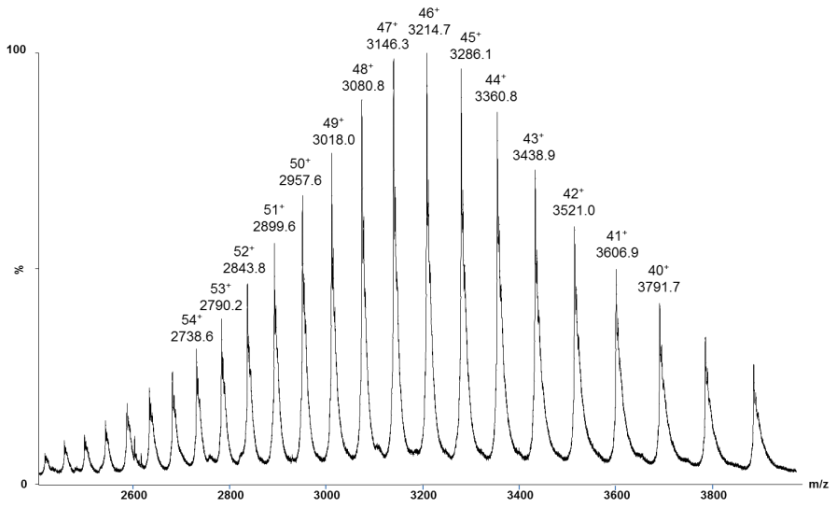


Figure 13 - ESI-Q-Tof Spectrum of an IgG (148 kDa)

There has been no step change in practical ESI since the introduction of the technique, only a refinement of the original process that has come with better understanding of the electrospray mechanism. This has focused on improving efficiency and sensitivity. Indeed these are very important parameters when performing top-down MS because ion loss at the front-end of the mass spectrometer is difficult to compensate for downstream and may drastically limit analysis options.

ESI has benefited from optimised source geometries and the introduction of sheath and nebulising gasses to increase desolvation efficiency. Theoretical treatment of the ESI process determined that smaller droplets provide more efficient ionisation. This led to miniaturisation of the standard electrospray source by Wilm and Mann who used pulled glass capillaries of 1-2 μm orifice diameter[63, 64]. With their interface they achieved flow rates in the low nL/min and sub picomolar sensitivity of proteins extracted from gels[65]. They also outlined the importance of salt removal from protein samples to ensure a stable continuous spray. Using a similar approach the McLafferty group achieved attomolar sensitivity and flow rates below 1 nL/min[66, 67]. For online nanospray as used in discovery mode, the pioneering work of Caprioli showed that flow rates of several hundred nL/min were possible and achieved zeptomolar sensitivity with peptides under highly optimised conditions[68]. This type of design forms the basis of modern LC-MS interfaces and performs very well with both peptides and proteins.

Another important parameter is the composition of the electrospray solvent. Two broad types of electrospray solution are used. The first type is a mix of water, organic solvent (such as acetonitrile or methanol) and a volatile acid to aid ionisation. This mixed phase spray solution is denaturing and destroys proteins solution phase tertiary structure. It is compatible with reverse phase chromatography, produces the best conditions for efficient ESI and is used in the majority of top-down studies. The second type is an aqueous buffer solution of a volatile salt such as ammonium bicarbonate. Spray solutions of this type are non-denaturing and allow the solution phase structure of proteins to be persevered. ESI in non-denaturing conditions is termed native spray and has been used in top-down studies that aim to probe tertiary structure[69-71].

Supercharging

The distribution of protein charge states and the maximum observed charge state can be an important factor in TDMS. Several methods have been developed to artificially increase the average number of charges carried by electrosprayed ions. Termed supercharging this usually involves mixing one of several additives such as sulfolene, dimethylsulfoxide (DMSO) or *m*-nitrobenzyl alcohol into the electrospray solution. The current hypothesis is that during electrospray these low vapour pressure, low basicity molecules make up an ever increasing proportion of liquid inside the evaporating droplet keeping the droplet intact whilst a greater proportion of available protons are sequestered by the analyte.

Other methods to shift the charge state envelope to higher charge states include complexation with trivalent metal ions[72] and heating of the capillary that is used to transfer ions from the source region of the mass spectrometer (electrothermal supercharging)[73]. Whilst the latter is an easy parameter to change and may be a practically useful solution, the addition of salt can have deleterious effects to signal intensity and complicate MS spectra. Supercharging additives can also cause pollution of the source region and ion optics. For these reasons, supercharging has found isolated uses in targeted mode TDMS but has not achieved widespread adoption.

6.2. Robust Sample Delivery Systems for Targeted TDMS

Targeted mode top-down experiments are mostly carried out on protein mixtures off-line (not coupled to LC). When sample quantity is high, ESI is used simply because it is more robust and easier to implement. nano-ESI is more appropriate for lower sample concentrations and smaller volumes. Pulled capillaries similar to those used by Wilm and Mann are still used to introduce the sample. These may take the form of metal needles, metal coated tapered glass capillaries or pulled glass capillaries where the circuit is completed by a metal wire placed inside the electrospray needle.

These needle types are all particularly laborious to prepare, position in front of the mass spectrometer and once in place are difficult to regulate. Particulates must be removed from the protein solution in order to prevent needle clogging and achieving a stable spray may be difficult with proteins that tend to aggregate or stick to the capillary orifice (Figure 14). Once a stable spray is established one must adopt a hit-and-hope attitude, since if the spray stops mid experiment, often due to needle clogging, there is no easy way to reinitiate the spray. This can be pretty irritating if it occurs half way through a 15-30 minute acquisition!

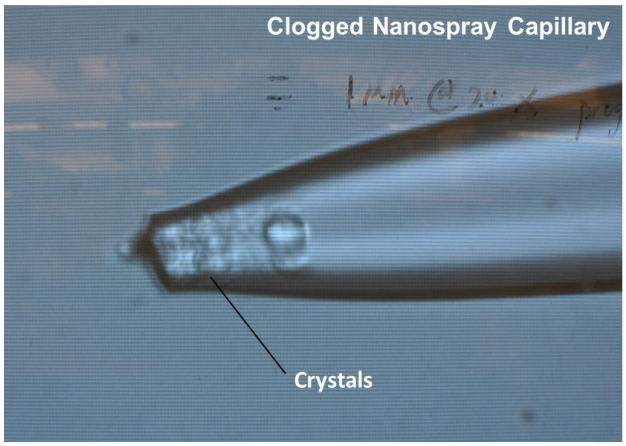


Figure 14 - A pulled borosilicate capillary with orifice diameter of around 3 µm that clogged mid run, presumably due to analyte precipitation or crystallisation, during experiments performed by the author

An automated solution, the Advion NanoMate, uses a robotic "mandrel" to aspirate sample from a 96 well plate (or other container or surface) and deliver it to a prefabricated conductive silicon wafer chip that is prepared with an array of nano-ESI nozzles. The chip sits in front of the MS orifice and acts as a nano-ESI source. The capability to automatically introduce sample, monitor the electrospray current during acquisition and change electrospray voltage, back pressure and nozzle mid run if required greatly facilitates the acquisition of top-down spectra especially over an extended time frame. The NanoMate is the method of choice for targeted mode TDMS and can also be coupled with LC.

An application illustrating the versatility of this device has been presented through top-down analysis of haemoglobin variants[74]. Here Edwards *et al.* take advantage of the liquid extraction surface analysis (LESA) capabilities of the NanoMate and haemoglobin is directly dissolved into the electrospray nozzle by automated soaking of dried blood spots in electrospray solution and immediate injection into the mass spectrometer (Figure 15).

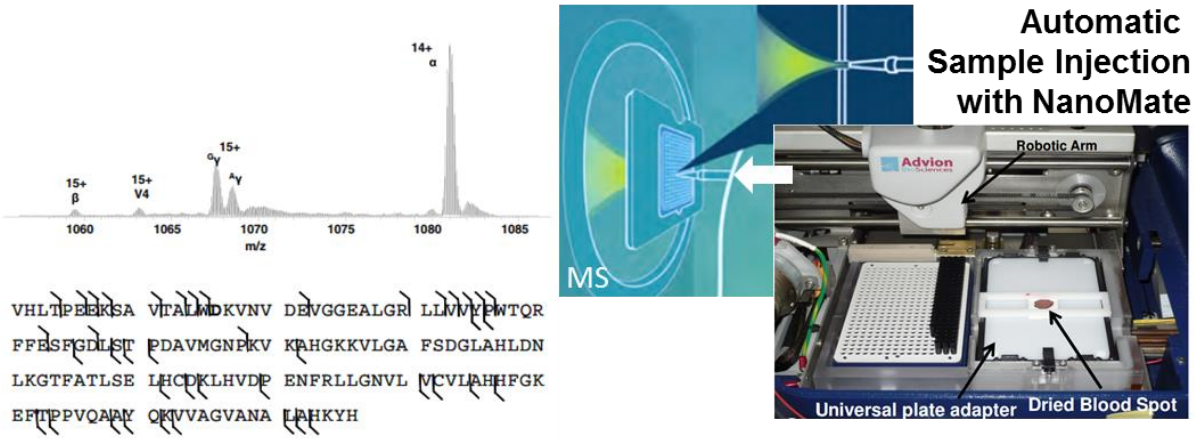


Figure 15 - Photograph of the LESA enabled NanoMate used for direct sampling of dried blood spots and cartoon of the electrospray process (right). Orbitrap mass profiling of a haemoglobin sample and coverage resulting from top-down CAD MS/MS of the unknown variant labelled V4 (left) Adapted from Edwards *et al.*[74]

6.3. High Resolution Mass Measurement

High Resolution for Protein Mass Measurement

Early mass spectra from electrosprayed proteins showed clear but broad peaks at multiple *m/z* values (much like Figure 13). Before a protein mass can be calculated from such spectra the individual charge state of the spectral peaks must be obtained. This is achieved by application of equation (1)[75] where z_n is the charge of a particular peak in the spectrum, m_n is the measured *m/z* value of that peak, m_{n-1} is a similar value from the peak in the charge state distribution with next smallest *m/z*, and m_a is the mass of the charge carrier, usually a proton. (There are many varieties of this equation that can be interchanged by simple substitution and rearrangement)

$$z_n = \frac{m_{n-1} - m_a}{m_n - m_{n-1}} \qquad (1)$$

Once the charge state of a peak has been determined simple multiplication of z_n and m_n minus the mass of the charge carriers, gives an approximation of the neutral protein mass M_r.

This can be scaled and averaged for all charge states to give a more accurate neutral M_r by application of equation (2) where n_0 the number of spectral peaks chosen for the calculation.

$$M_r = \frac{1}{n_0}\sum_n z_n(m_n - m_a) \qquad (2)$$

Reduction of the various charge states of a protein to their fundamental mode can be achieved by a number of mathematical treatments and is an example of spectral deconvolution. Deconvolution aims to collapse series of repeating peaks (charge states) into a single peak of nominal mass. There

are various ways to achieve this and the detail of the mathematical models are beyond the scope of this thesis. The maximum entropy method is particularly noteworthy and useful for transformation of spectral data into a distribution of singly charged or neutral isotope distributions. This can be especially useful for generating deconvoluted mass profiles of measured proteins.

Peptide ions are often visualised by MS as very narrow peaks whereas proteins give much broader peak shapes that increase in width with increasing mass. This is because the atoms that make up proteins naturally exist in different isotopic forms and therefore so do the proteins themselves. Even a very small protein such as ubiquitin (8.6 kDa) contains over a 1,200 atoms and thus a huge number of isotopologues with a range of distinct masses differing by approximately 1 Da.

On low resolution mass spectrometers, such as those used to recorded early protein spectra, even the charge states of small proteins are visualised as broad peaks. Only the average mass of the protein (M_{av}) can be determined from the peak apex. As the resolving power of the mass analyser is increased isotopic peaks are resolved. This is shown in Figure 16 for the theoretical $[M]^+$ ion of the protein ubiquitin.

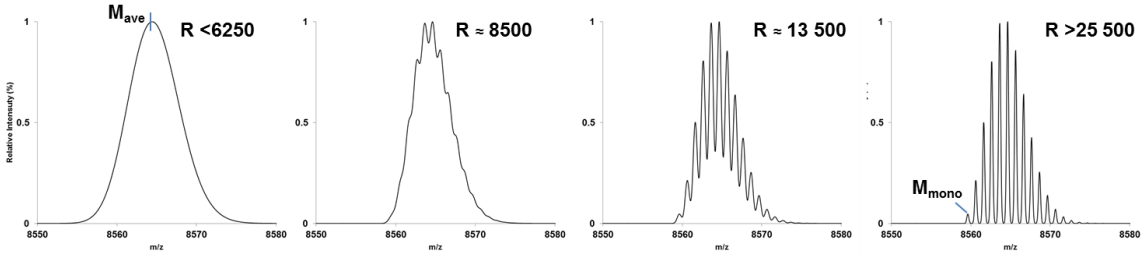


Figure 16 - Representation of the $[M]^+$ ion of the protein ubiquitin with increasing resolution, R, from unresolved isotopes (left) to baseline resolution (right). Peak shape is Gaussian. R is calculated using the $m/\Delta m_{50\%}$ definition

Once peaks are sampled at isotopic resolution, charge state determination is relatively straightforward. The mass difference between singly charged isotope peaks is approximately one mass unit, therefore the charge state of an isotope cluster is given by equation (3) where Δm_{iso} is the m/z difference between adjacent isotope peaks.

$$z \approx \frac{1}{\Delta m_{iso}} \tag{3}$$

Once the resolution is sufficient for the individual isotopes to be resolved, and if the signal to noise ratio (S/N) is high enough, a small peak will appear at the low mass end of the distribution. This monoisotopic peak corresponds to the protein exclusively formed of atoms in their lowest commonly observed isotopic state, the mass of which defines the monoisotopic protein mass (M_{mono}).

As the protein gets larger and the probability of forming the protein exclusively from atoms in their lowest isotopic mass decreases, the monoisotopic peak becomes even smaller with respect to the most intense peak in the isotopic envelope. This can easily be shown by considering the cases of a small and medium sized protein. For ubiquitin (8.6 kDa) the monoisotopic peak represents 4% of the intensity of the total envelope. For a typical medium sized protein (50 kDa) this drops to below 1×10^{-6} %. This means that for large proteins and low intensity signals, regardless of the resolving power of the analyser, the monoisotopic peak may never actually be distinguishable from the spectral baseline unless special measures are taken to reduce spectral noise either experimentally or by signal processing.

If data are sufficiently resolved to distinguish the isotopes but the S/N is too low for the monoisotopic peak to be visible, Senko, Beu and McLafferty proposed that a theoretical distribution of a hypothetical AA "averagine" could be used to calculate the position of the monoisotopic peak and thus obtain an accurate M_{mono}. Averagine has the formula $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$ and an average molecular mass of 111.1254 Da and is based on the natural abundance of the amino acids in the Protein Identification Resource database available at the time[76]. When trying to estimate the monoisotopic peak of a measured protein, the measured average mass is used to calculate non-integer number of averagines that would be required to achieve that mass. The atom numbers of the elements in this hypothetical "polyaveragine" are then rounded to integer values and the isotopic distribution of the averagine protein calculated and fitted to that of the observed distribution. This has been found to provide a very accurate estimation of the monoisotopic peak if the initial distribution is also of a good quality. This approach is also called "de-isotoping".

In targeted mode top-down experiments an accurate M_{mono} is extremely useful and can even be used to achieve protein ID without an MS/MS step. For example Robertson *et al.* used high resolution mass profiling of 18.6-18.7 kDa mouse urinary proteins (MUPs) with a 9.4T Apex III FT-ICR to identify four major urinary protein (MUP) components. Isotopic resolution of the 18.6-18.7 kDa parent ion peaks allowed calculation of an accurate M_{mono} and protein ID was simply achieved by comparing masses with reported MUP sequences. In one case where masses of the proteins were extremely close, careful examination of overlapping isotope clusters and examination of those predicted theoretically allowed the distinction of two forms that differed by only 0.985 Da (difference between peaks of 0.015 Da). This was later confirmed by LC separation.

An accurate M_{mono} can also greatly aid the assignment of PTMs and other proteoforms. When a high resolution spectrum is available, some PTMs such as cysteine oxidation may be investigated explicitly at the protein level without the need for fragmentation. Furthermore an accurate parent

protein mass will support high resolution fragmentation data and may prove decisive in PTM assignment. For example trimethylation and acetylation, both common modifications on histones, differ in mass by 0.036 Da. This small mass difference is only distinguishable at an elevated resolving power.

Limits to Achieving Isotopic Resolution of Proteins

Achieving isotopic resolution on even small proteins requires specialised mass analysers that permit high resolution mass measurement. The demands placed on this component depend both on the size of the protein and also the mode of sample ionisation. In the case of MALDI, where ionised proteins will be singly charged, the mass analyser requires a very large mass (*m/z*) range but a comparatively low resolving power, as the isotopologues of the protein will be separated by ≈1 Da. On the other hand, if ionisation is performed by electrospray, protein ions will be multiply charged. The *m/z* range required for protein mass measurement will therefore be smaller, but the resolution required is much higher, as isotopologues of the protein will now be separated by smaller *m/z* intervals that depend on the protein charge state (Equation 3).

High resolution ToF instruments provide a greater than standard resolution that is theoretically independent of *m/z*. However, the ion detectors traditionally used in these instruments decrease in sensitivity with the mass of the analyte. This limits the usefulness of ToF based MS for the analysis of large molecular weight species[77]. New technological developments such as the nanomembrane detector[78] and improvements to conversion dynodes are beginning to change this, however at present two analysers; the ion cyclotron resonance (ICR) cell, and more recently introduced Orbitrap[79, 80] are found the heart of most instrumental platforms used to perform TDMS[81].

The mode of operation of these two analysers is presented in greater detail in the Materials and Methods section and elsewhere[82, 83]. Briefly, both analysers trap ions in electrical (Orbitrap) or both electrical and magnetic fields (ICR). During detection, ions are caused to oscillate at a frequency that is inversely proportionate to their *m/z*. This oscillation is detected as an image current (transient) and a mass spectrum reconstructed from this signal by means of a Fourier transform.

The maximum theoretical resolution that can be achieved by these analysers depends on a number of parameters including *m/z* (decreases linearly with m/z for FT-ICR and with the square root of *m/z* for Orbitrap systems), magnetic field strength (in the case of FT-ICR), the potential applied to the outer electrodes (for Orbitrap), signal detection time, signal sampling rate (related to the Nyquist frequency) and the mathematical model by which the detected signal is processed.

The relationship that some these factors have to theoretical resolving power is summarised in Figure 17.

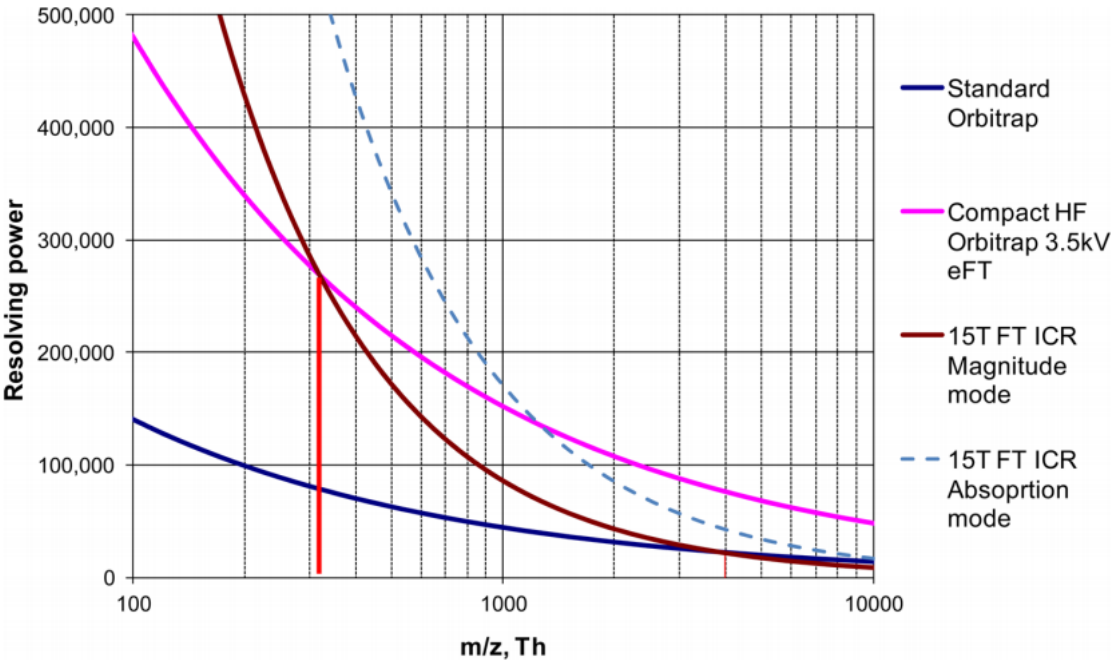


Figure 17 - Dependence of resolving power on *m/z* for the following analysers (all data are shown for a 0.76 s scan): Standard Orbitrap (magnitude mode, 3.5 kV on central electrode), Compact high-field trap (absorption mode, 3.5 kV on central electrode), FT-ICR (magnitude mode, 15 T), FT-ICR (absorption mode, 15 T). Taken from Zubarev *et al.*[79]

In practice, several additional factors decrease the maximum achievable resolution. Orbitrap systems are "resolving power limited" by the instrument software to 100,000 (Velos model), 240,000 (Velos Elite and Fusion models) at *m/z* 400. On both analyser types the most common experimental parameter that is changed to increase resolution is signal detection time. However, despite the very low pressures inside the ultra-high vacuum of the mass analyser ($\approx 1\times10^{-10}$ torr), the vacuum conditions are not ideal. Collisions with residual gas molecules dampen the detected signal over time and decrease the quality of the information that can be obtained from long transients. This is clearly shown in Figure 18 where a transient from the 33^{+} charge state of carbonic anhydrase has been recorded at standard and reduced pressures (pressure in the Orbitrap is determined by the pressure in HCD cell). Even at low pressures, signal dampening is clearly seen and this reduces the maximum achievable resolution.

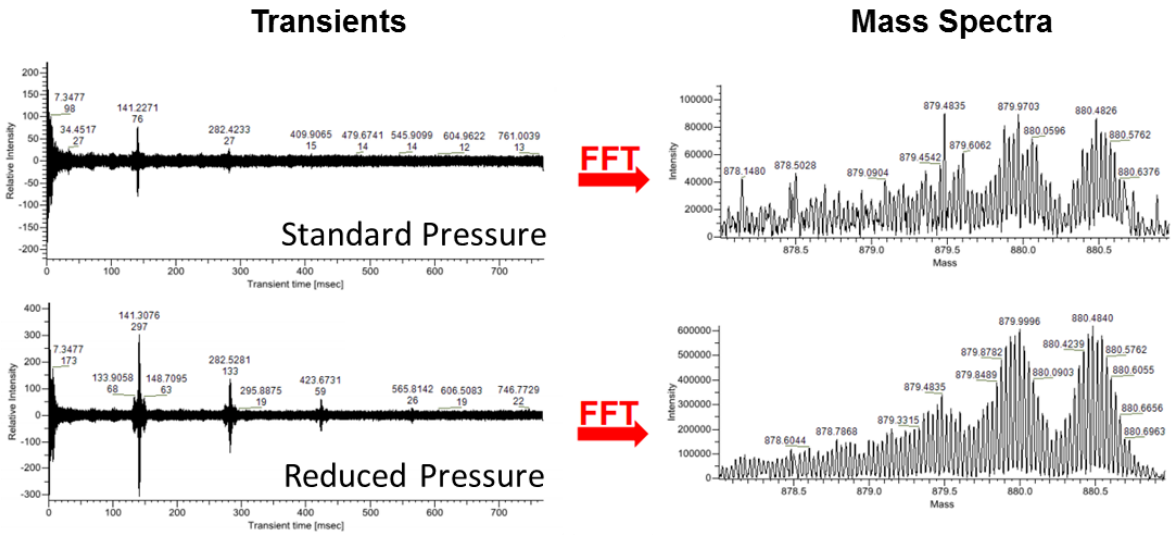


Figure 18 - Effect of pressure on recorded transient and spectral resolution for the 33^{+} charge state of carbonic anhydrase. Adapted from Thermo application note

In addition to long transient times, high resolution also requires high signal fidelity. High numbers of ions in the cell favour ion-ion interactions and increases space-charge effects. This results in non-ideal ion motion and leads to rapid signal loss, which in turn decreases the achievable resolution. Regulating the number of ions that are present the cell is essential to maintaining high resolution and high mass accuracy. This feature is present on Thermo instruments as they are equipped with an automatic gain control (AGC).

Another important parameter for signal survival over long periods of time is the quality of the electrostatic field in the Orbitrap (and the harmonicity of the DC trapping potential in the ICR cell). These both need to be as close to ideal as possible to ensure the homogeneity of the ion packets throughout detection. Resolution of more than one million has been achieved recently on an Orbitrap system in which this parameter, which mainly depends on the accuracy at which the electrodes are machined (nanometer range), had been carefully examined[84]. A new harmonised FT-ICR cell ("leaf cell") has been also recently introduced by Nikolaev *et al*. that has increased the experimentally achievable resolving power of FT-ICR instruments considerably (>20 million), even at low magnetic field strengths[85, 86].

Furthermore, large proteins are inherently expressed as a large number of isotopologues. Thus in a complete isotope envelope there are both a large number of ions, and many ions with very similar masses. These isotopologues will all have very similar precession frequencies. Such a scenario can greatly increase space charge effects and cause the precessing ion packets to mix together (phase locking) and their precession frequencies to coalesce. This phenomenon has been observed in both ICR cells and Orbitrap mass analysers[87], and results in mass shifts and a loss of

resolution. These problems are aggravated if the protein bears non-covalent adducts from salts or the electrospray solvent. Unfortunately both of these become more likely as the protein size increases.

Taken together these reasons help explain why unit mass resolution has only ever been reported for three proteins over 100 kDa[88-90]. The current unit-resolved mass record was set by Marshall's group in 2011[90]. They achieved baseline resolution of a 147 kDa monoclonal antibody on a home built 9.4T FT-ICR equipped with a "leaf type" ICR cell, at high spectral density (6-18 Mword) and with a transient acquisition time greater than 10s in duration.

High Resolution for Interpretation of MS/MS Data

Many different fragmentation techniques have been used in top-down experiments; two of which were used by Smith *et al.* to create the first assigned top-down spectra reported in the first issue of JASMS in 1990[91]. The protein melittin (2.8 kDa) was fragmented by in source dissociation (ISD, also called nozzle skimmer dissociation, NSD) on a QqQ mass spectrometer where applying a high voltage to their modified ESI inlet caused collision with gas molecules in the high pressure region of the ion source and CAD type fragmentation. In the same paper, CAD of cytochrome-C performed in the collision cell provided useful "protein fingerprints", but identifying the product ions from the spectra proved impossible due to "both the absence of product ion charge and the limited mass resolution". Later that year an article from the same group demonstrated more impressive top-down spectra from a more reasonably sized protein, RNAse A (13.8 kDa) using both ISD and CAD after cysteine bond reduction[92], however again assignment of the product ions proved difficult. At the end of this paper they conclude, "Ultrahigh resolution, a unique capability of FTMS, would permit unambiguous product-ion charge state determination, (by the observation of isotope peaks) and allow unknown sequences to be more easily interpretable", a clear nod to the requirement for high resolution to assign the fragments from protein fragmentation spectra.

MS/MS spectra of proteins are inherently complex. If a pair of complementary ions is considered, *b*/*y* for example, a small protein with 150 amino acids has the potential to produce 298 singly charged fragment ions. If multiply charged species have been fragmented (as is usually the case) the majority of fragment ions will be both multiply charged and present in a number of different charge states. In addition each fragment ion will have its own isotope cluster containing multiple peaks which become more numerous with increasing fragment ion mass. This greatly increases the total number of fragment ions present and results in a very crowded MS/MS spectrum, full of multiply charged isotopic distributions that often overlap Figure 19.

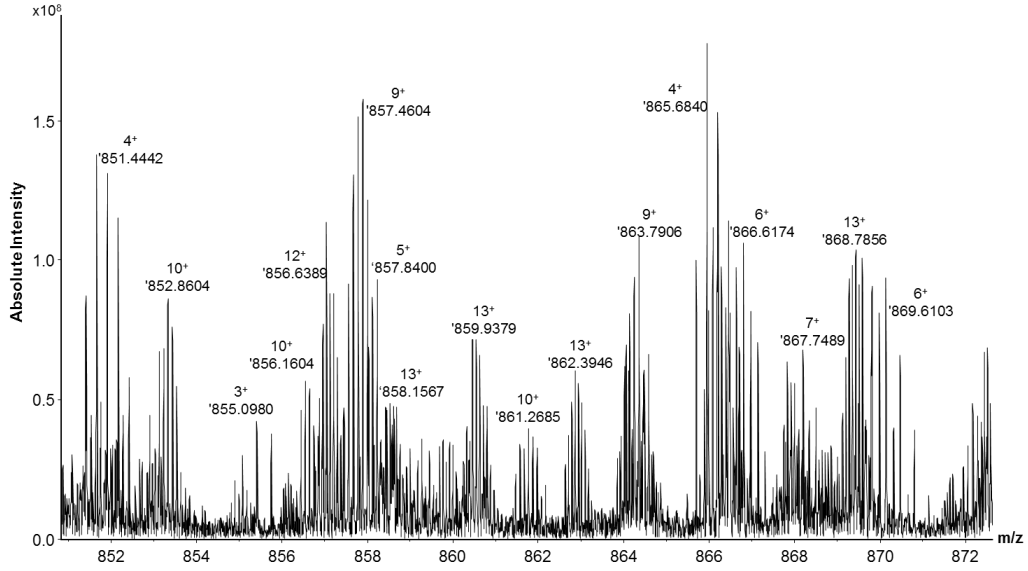


Figure 19 - Zoom on a 10 *m/z* window of a typical ECD MS/MS spectrum, in this case of the 20^+ charge state of the of the protein myoglobin (17 kDa), showing many multiply charged fragment ions. Ion masses are monoisotopic

Making sense of the "forest" of fragment ions produced is a considerable challenge that can be tackled in two possible ways. The first is to reduce the complexity of the spectrum by reducing the charge state of all fragment ions. This spreads the spectrum over a larger *m/z* range and reduces fragment ions to charge states that many mass spectrometers instruments can resolve. This strategy is achieved though ion/ion protein-transfer reactions and has been pioneered by McLuckey and co-workers in order to extend top-down to lower resolution trap instruments. The second is to develop robust algorithms to treat this complex data either manually or computationally. These approaches rely on calculating the charge state of the observed isotope clusters and therefore require highly resolved MS/MS data.

6.4. Efficient Fragmentation Methods Adapted for Proteins

As instruments have developed so have fragmentation techniques and modern mass spectrometers are equipped with an arsenal of different activation modes all of which can be used for protein fragmentation. Fragmentation techniques fall into two main camps, vibrational an electronic.

Vibrational Modes for Primary Structure Investigation and non-labile PTMs

Vibrational modes include ISD, CAD, HCD (high energy C-trap dissociation), BIRD (blackbody infrared dissociation), SORI-CAD (sustained off resonance induced collisionally activated dissociation), IRMPD (infrared multi photon dissociation) and SID (surface induced dissociation).

Vibrational activation can be performed on most mass spectrometers; either in a dedicated collision cell, ion trap or simply in the source region. HCD is limited to Orbitrap type instruments and BIRD, IRMPD and SORI-CAD are mostly associated with FT-ICR instruments.

In vibrational fragmentation modes energy is pumped into a molecule through collision with a surface (SID), inert gasses (ISD, CAD, HCD) or by absorption of low energy IR photons (BIRD, IRMPD). Since usually energy is deposited in small packets over several milliseconds this process is ergodic, energy is deposited and partitioned throughout the vibrational modes of the system until there is sufficient internal energy to overcome the activation barrier for bond cleavage. The fragmentation channels open to the molecule depend strongly on the quantity of energy involved, the manner in which it is deposited and on the molecule structure. Vibrational techniques produce mostly *b*/y type ions tend to be more efficient than electronic methods, with a greater percentage of parent ion(s) being converted into fragments (rather than charge reduced species). They have proved particularly useful for the characterisation of truncations and SAPs.

This is demonstrated in a report by Laitaoja *et al.* who used a combination of intact mass profiling and CAD performed in the external collision cell of a 12T Apex-Qe FT-ICR mass spectrometer to investigate sequence variants of the major bovine seminal plasma protein PDC-109 (13 kDa)[93]. Fragmentation was sufficient to identify four new variants including a proteoform truncated by 14 residues at the N-terminus, another variant with two point mutations; P10L and G14R, and two minor proteoforms that also appeared to be truncated at the N-terminus.

In another report performed on an LTQ-Orbitrap platform, sequence variants of the plasma proteins haemoglobin and transthyretin have been analysed by a multiple stage top-down methodology[94]. After mass profiling, protein ions of interest were selected and fragmented in a two-step MS^n strategy. ISD was used to generate large *b* and *y* type ions. The ions were examined for abundant fragments containing putative SAPs and these were selected in the LTQ and subjected to CAD to localise the modification site. In the case of transthyretin a G6S modification was elucidated. For haemoglobin the strategy was evaluated in detail and an E6V mutant confirmed on a sickle cell variant. This is one of the few examples of top-down MS on this older generation Orbitrap system.

Electronic Modes for Primary Structure Investigation and Labile PTMs

Fragmentation modes that involve electronic excitement include UVPD (ultraviolet photodissociation), ETD (electron transfer dissociation)[95], ECD (electron capture dissociation)[96] and higher energy forms[97]. ECD and ETD can be collectively referred to as ExD. UVPD is currently limited to spectrometers modified in-house of which several types have been reported. ECD (and its higher energy derivatives) is mainly limited to FT-ICR type instruments although some attempts have been made to couple it with other instrument types[98, 99] and a top-down ECD MS/MS of ubiquitin has been demonstrated in modified Q trap[100]. ETD is available on many different instrument types including Q-ToF, Q trap, LTQ, FT-ICR and others.

UVPD involves absorption of a photon that promotes a valence electron into a higher energy state. When performed with a 193 nm laser, ≈3 eV is deposited. This results in both fast and slow fragmentation, depending on the proportion of internal conversion during electronic relaxation that precedes fragmentation, and forms a mixture of ion types *a, b*, *c*, *x*, *y*, *z*, *v*, *w*, and *d*. UVPD is a fairly new technique for protein fragmentation, despite being at the origin of the discovery of ECD. It appears to be very efficient and enjoys high parent to daughter ion conversion. It is also extremely rapid requiring only several nanoseconds for fragmentation. Its utility when labile PTMs are present is yet to be evaluated however it seems incredibly promising for SAP identification and for localising non-labile PTMs such as oxidation[101].

ECD and ETD involve the capture or transfer respectively of low energy electrons <2 eV to a protein cation. This results in the formation of a radical cation and triggers a sequence of fast electronic transitions that ultimately results in fragmentation of the protein backbone. The detail of the mechanism is discussed in the Materials and Methods section. Fragmentation occurs very quickly after electron capture/transfer (it was originally reported to be non-ergodic although this has been refuted by theoretical calculation performed by Tureček and others) and produces primarily *c/z* type fragment ions.

Electronic modes find particular uses in the identification and characterisation of proteins harbouring labile PTMs[102, 103]. Modification such as glycosylation, phosphorylation, sulfonation and nitrosylation are easily lost during vibrational fragmentation due to facile cleavage of the protein-PTM bond. Conversely ECD and ETD preserve even the most labile PTMs when performed in the correct conditions and are particularly useful for PTM site localisation. They are however less efficient often producing abundant charge reduced molecular ions and requiring a longer duty cycle.

One of the most impressive uses of ECD and ETD has been for the characterisation of intact antibodies (>100 kDa). Tysbin's group analysed two intact antibodies, murine MOPC 21 IgG and human anti-Rhesus D IgG by ETD on a maXis UHR qTOF instrument achieving fairly impressive

22% and 16% sequence coverage respectively[104]. Using a different approach on an Orbtirap Velos Pro instrument they extended this coverage to around 33% on the Humira IgG1 kappa antibody (148 kDa) by combining up to 10,000 transients from multiple LC MS/MS runs and different ETD interaction times performed on different precursor charge states[105]. Marshall's group achieved a similar sequence coverage of 34% on the recombinant humanised IgG kappa antibody (148 kDa) by ECD fragmentation of all charge states up to *m/z* 3500 on a 9.4 T FT-ICR[106]. The phased MS/MS spectrum and associated sequence coverage is shown in Figure 20.

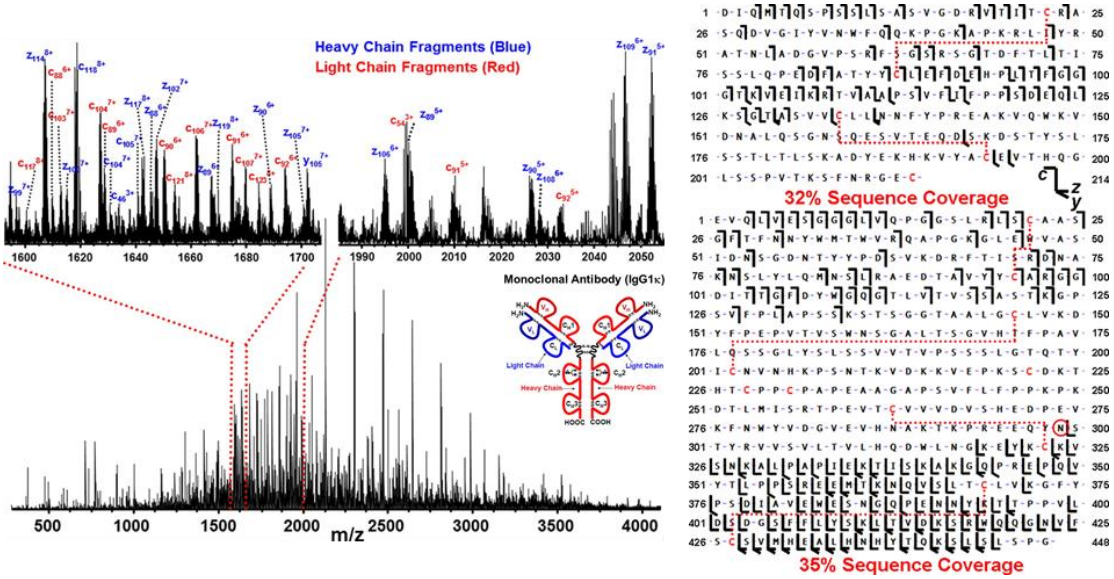


Figure 20 - ECD MS/MS of all charge states of an intact antibody. Coverage of the light (right top) and heavy chains (right bottom) indicates the localisation of the cysteine bonds and N-linked glycan. Adapted from Mao *et al.*[106]

Combinations of Vibrational and Electronic Fragmentation for Deep Characterisation

On polyvalent instruments the choice of fragmentation technique depends greatly on the aim of the top-down experiment and often the type of PTMs that one wishes to identify. The different strengths of the two types of fragmentation (vibrational and electronic) have often resulted in their complimentary use in targeted mode TDMS. Generally data from separate experiments is pooled together to produce a more complete picture of the proteoform of interest.

Peng *et al.* used a combination of ECD and CAD on a 7T LTQ-Ultra FT-ICR for the deep characterisation of human salivary α-amylase (HSAMY)[107]. This is a particularly challenging protein to work with because of its size (56 kDa) and the presence of 5 cysteine bonds. MS/MS of non-reduced native HSAMY identified cleavage of the 15 residue N-terminal signal peptide and formation of pyroglutamic acid. Unambiguous localisation of 3 cysteine bonds was also achieved in the native form, with two more overlapping cysteine bonds being localised through

fragmentation after partial reduction with tris (2-carboxyethyl) phosphine (TCEP). The monoisotopic mass furnished by high resolution mass profiling allowed confirmation of complete PTM characterisation (Figure 21).

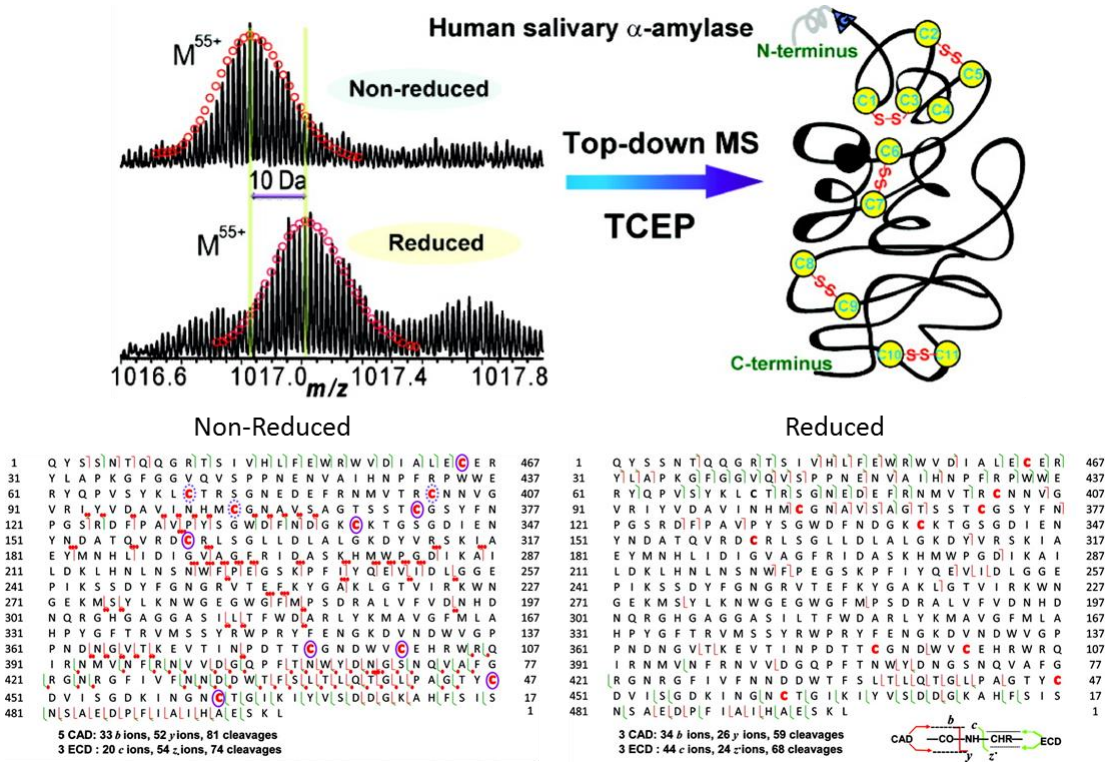


Figure 21 - High resolution MS spectra of non-reduced and completely reduced HSAMY, 10 Da corresponds to 5 intact disulfide bonds, shown in the cartoon of the protein. (Top) Combined sequence coverage from ECD and CAD top-down MS/MS of reduced and non reduced forms of HSAMY (bottom) one red dot on the fragment ion indicates +2 Da, 2 red dots +4 Da etc. Adapted from Peng *et al.*[107]

In a particularly interesting study Jia *et al.* used a polyvalent mass spectrometry approach for the *de novo* sequencing and PTM characterisation of crustacean hyperglycemic hormone (CHH)-family neuropeptides (8 kDa)[108]. Top down *de novo* sequencing using CAD, ECD and HCD performed on three different instrument platforms achieved 75% coverage of the primary sequence and the identification of two PTMs (pyroGlu and C-terminal amidation) on the *Callinectes sapidus* salivary gland CHH. This approach proved complementary to online TDMS and bottom-up in which 50% and 81% sequence coverage was achieved respectively and ultimately resulted in full sequence elucidation. This approach was successfully extended to other CHH peptides in *C. borealis* and MALDI imaging used to show their localisation in the sinus gland.

7. Performing Top-Down Mass Spectrometry

7.1. Targeted Mode Top-Down Mass Spectrometry for Deep Proteoform Characterisation

In a typical targeted mode TDMS experiment an enriched protein or simple set of proteoforms is prepared and analysed by tandem MS. If the sample is of low enough complexity the purification step can even be done away with, and components selected inside the mass spectrometer. Proteins of interest are ionised and injected into the mass spectrometer. Protein masses are then measured in a mass profiling experiment and may be compared to the expected mass from the genome in order to check for the presence of PTM. This is the first stage of TDMS. In the second stage protein ions of interest are manually selected, isolated and fragmented also inside the mass spectrometer. This produces complex fragmentation spectra. In the third step the raw data is processed either manually or by software tools to provide a list of monoisotopic fragment ion masses. In a fourth stage these lists are matched to expected fragment ions in the case of MS/MS and the fragmentation data visualised. Characterisation of PTMs may be performed in a fifth and final analysis stage.

Protein Separation

Targeted mode TDMS generally requires fairly large quantities of protein (at least tens of picomoles) especially if experimental parameters need to be refined. A good guideline is that the protein of interest needs to be by far the major component of the protein sample and gives a visible band when analysed by Coomassie stained SDS-PAGE. For this reason application based literature that uses deep characterisation mode TDMS is dominated by proteins of natural high abundance such as proteins from body fluids (haemoglobin, salivary proteins, urine proteins etc.), recombinant proteins, those purified though techniques amenable to larger quantities such as FPLC, tag based affinity approaches and more recently GelFree or highly optimised enrichment protocols.

Step 1 Mass Profiling

The first step of the TDMS experiment is mass profiling. The sample is injected using glass capillaries or a NanoMate system and the source parameters and transfer optics of the mass spectrometer tuned to ensure good transmission of the protein ions. If the sample signal is very low the ions may be accumulated in external quadrupole storage devices or ion traps in order to improve S/N. The ions are then transferred to the mass analyser and detected for the time required to achieve the appropriate (usually isotopic) resolution. Often several transients (scans) are accumulated in order to further improve S/N and improve the quality the observed isotopic

distributions. This provides a birds-eye view of the proteins present within a sample. Figure 22 shows a mass profile spectrum of a protein present in at least two major proteoforms.

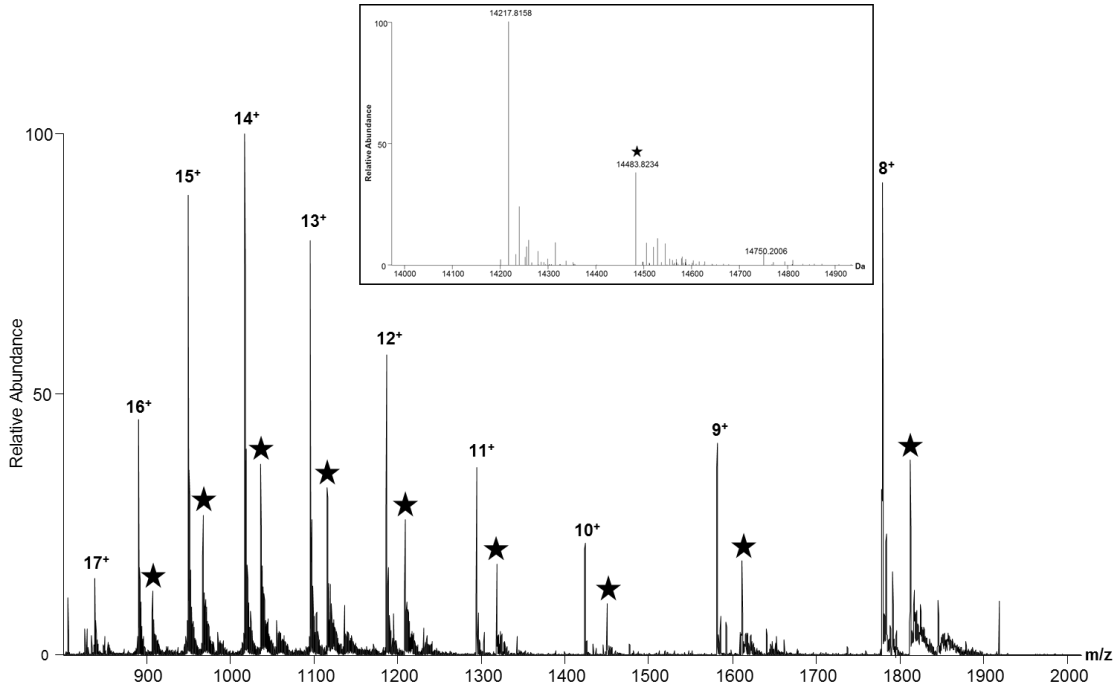


Figure 22 - Mass profiling of the small protein PpdD purified from *E. coli* and present in two major proteoforms

Once information has been obtained on the protein content of the sample, ions of interest may be identified and MS/MS performed. The experimental mass of an intact protein can also be compared the theoretical one (deduced from genomic data) and early information on the presence of PTMs, or unexpected truncations (such as Met removal) can be often easily deduced.

Step 2 Isolation and MS/MS

MS/MS may be performed on selected charge states or on the entire mass range. If a mixture of proteoforms is present (such as in Figure 22) selection of the desired proteoform is often performed, externally to the mass analyser in an appropriate quadrupole or ion trap. Ions are isolated and allowed to accumulate to the desired intensity. On FT-ICR platforms ion selection can also be performed within the ICR cell however this method of isolation is only practically used if a very narrow *m/z* window or multiple, separate *m/z* notches of the mass spectrum are required. On Orbitrap platforms the ion packet is then fragmented externally in the C-trap (HCD) or LTQ (ETD, CAD). In FT-ICR systems fragmentation may occur externally or inside the ICR cell in the case of ECD, IRMPD and SORI CAD. Fragmentation is followed by high resolution mass analysis of

fragment ions. An example of this is parent ion selection → fragmentation procedure is shown in Figure 23.

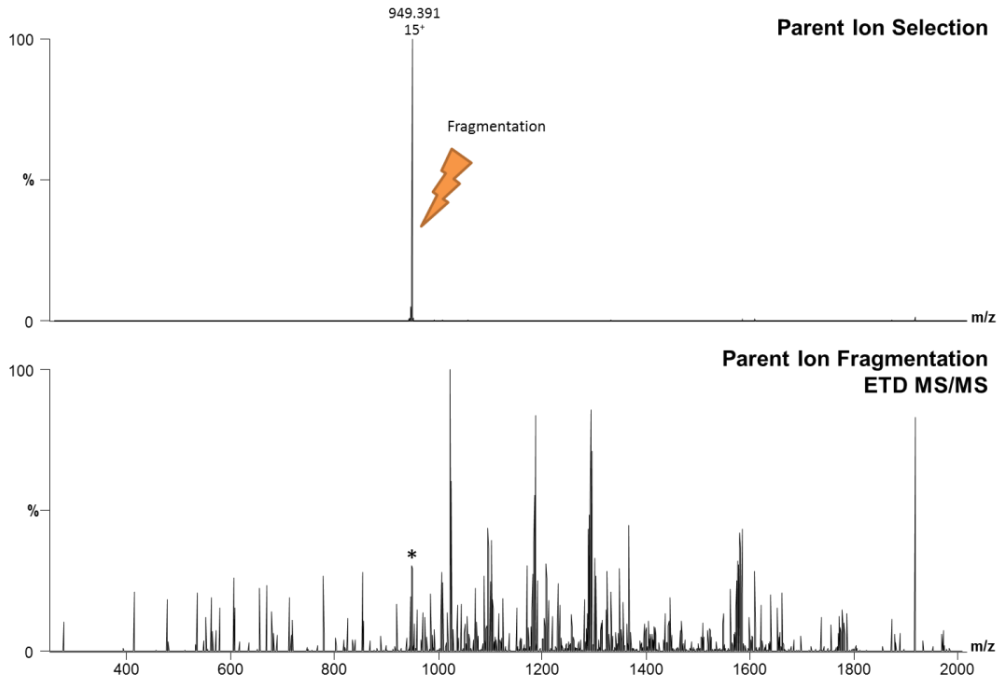


Figure 23 - Protein parent isolation (PpdD 15^+ charge state) and top-down ETD MS/MS performed on an Orbitrap Velos mass spectrometer

For proteins many possible fragmentation channels may be open and this dilutes the intensity of the fragment ions produced. The larger the protein the greater the number of possible fragmentation channels, the more complex the spectrum and the lower the average S/N of the individual fragment ion isotope distributions. Signal to noise has been shown to increase proportionately to the square root the number of scans recorded. Many repeat scans (often a few hundred but sometimes more than 1,000) are therefore usually taken and the transients summed to produce the final MS/MS spectrum. Acquisition time may last for several minutes up to several hours in order to achieve the required number of scans. The exact time depends on the duty cycle and this in turn depends strongly on three parameters; the parent ion accumulation time, the fragmentation mode chosen and the time required for ion detection (desired resolution).

Practical Considerations for Larger Proteins

The conservation of gas phase structure introduces some problems when fragmenting proteins of larger size. Unfortunately top-down MS/MS spectra of proteins >≈40 kDa are characterised by limited fragmentation, concentrated around the protein termini. Often no fragmentation at all is observed in the central region of the protein and in some cases even terminal regions may remain

impervious to examination. This is a major problem for extending TDMS to larger proteins in targeted mode and effectively limits the range of the proteome accessible in discovery mode. Several techniques have been developed in order to disrupt the residual structural elements and increase the mass range of proteins accessible to TDMS.

Supercharging has been used to induce unfolding but has limitations as previously discussed. Other approaches involve imparting vibrational energy to the molecule in an effort to disrupt tertiary structure prior or during the fragmentation event. Note that if activation is performed prior to fragmentation, care must be taken to ensure the protein does not have sufficient time to refold before the fragmentation step.

In a seminal work, McLafferty's group used various techniques to push the upper mass limit for TD MS/MS[109]. Strenuous IRPMD (long irradiation time) proved partially useful for fragmentation of β-Gal (116 kDa). More successful was a new approach termed "prefolding dissociation" (PFD). Here ions are subjected to low energy collisions in the high pressure (≈1×10^{-3} bar) capillary-source skimmer region of the mass spectrometer then rapidly accelerated in the relatively high pressure 1.3×10^{-6} bar post skimmer was also used. A cartoon fo this process is shown in Figure 24.

Initial application of this technique with the protein PurL (143.5 kDa) gave 50 assignable fragment ions and indicated N-terminal processing and removal of a Met residue. Combining spectra obtained under different experimental conditions (different capillary temperatures and pre and post skimmer voltages) gave ions representing 173 different cleavage sites.

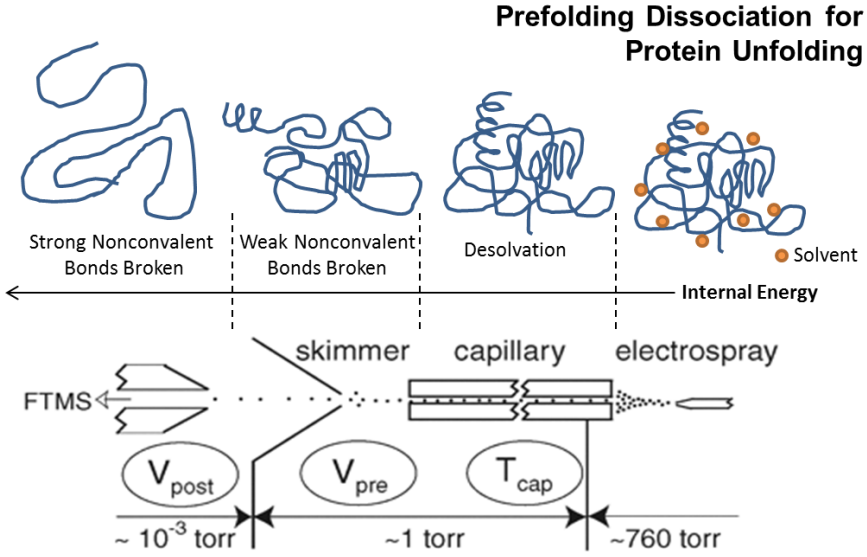


Figure 24 - PFD model for protein unfolding with the stages of pre-folding dissociation aligned with the appropriate region of the mass spectrometer. Adapted from Han *et al.*[109]

The addition of various ammonium salts to the electrospray solution changed the fragmentation pattern such that at total of 287 different cleavages were obtained representing 73% of the first 100 N- and C- terminal inter residue bonds. Use of PFD combining results from experiments performed with different parameters, yielded 87 cleavages for the 200 kDa human C4 glycoprotein. This is even more impressive given the high glycosylation state and number of cysteine bonds that tis protein carries.

Finally the combination of five PFD spectra obtained from the linear protein myco-cerosic aid synthase (229 kDa) gave 62 cleavages that extended over 100 AAs into the protein. This proof of concept study clearly showed that PFD enabled useful fragmentation spectra to be obtained for proteins over 200 kDa in size and sufficient fragmentation obtained to enable database identification. However for these larger proteins fragmentation was still limited to the termini and therefore is of limited use for detailed PTM characterisation. Moreover PFD would likely result in the loss of more labile PTMs as was observed here with human C4 glycoprotein. This work also underlines an important strategy of combining results obtained under different experimental conditions, in which different fragmentation channels may be favoured, in order to increase sequence coverage and protein characterisation.

Step 3 Spectral Processing – Peak Picking and Deconvolution

Once mass spectrometry data has been acquired spectral processing is performed. Both mass profiling and high-resolution MS/MS spectra are processed in similar ways either manually or using a variety of software tools.

In a manual approach isotope clusters are identified visually, their charge state calculated (by application of equation (3)), and an estimate of the M_{mono} made based on the expected number of peaks in the distribution for an ion of that size or an averagine based fitting algorithm. This is followed by deconvolution where M_{mono} is converted into a protein or fragment mass. M_{mono} determination is often subjective for large isotope clusters and can take over a week to treat a single top-down MS/MS spectrum!

Alternatively in an automatic approach several proprietary tools such as SNAP 2.0 (Bruker) or Xtract (Thermo) may be used on instrument specific data formats or freely available algorithms such as Thorough High Resolution Analysis of Spectra by Horn (THRASH) may be employed[110]. These types of tools perform peak picking, de-isotoping and deconvolution and output peak lists of the monoisotopic masses of spectral components. In most cases a score based on how well the isotope pattern matches the averagine based theoretical pattern is also provided (fit factor). This can be useful in evaluating confidence in the peak picked data. These tools do have limitations and

an empirical evidence suggest some perform better than others, particularly in extracting information from overlapping isotope clusters in MS/MS spectra.

The effectiveness of automated peak picking and de-isotoping tools of this type relates directly to the quality of the mass spectral data and the power of the post-acquisition processing. The ability to resolve low abundance peaks in the isotope distribution of fragment ions is directly dependent on ion intensity (S/N) and fragment ion size. The higher the S/N the more peaks in the isotope cluster will be visible and the greater probability that the cluster will be picked. The intensities of the visible ions will also have a direct impact on the calculated M_i since the averagine pattern will be shifted to ensure the best match with the observed peak pattern, no matter how poor it is. Low intensity signals must therefore reach a certain S/N threshold before they can be considered useful. For example one may be able to deduce three peaks from a large fragment ion from the baseline which is enough to deduce its charge state, but insufficient to estimate a meaningful M_{mono}. Using automatic tools for peak picking and deconvolution helps remove the subjectivity from data interpretation even if it may sometimes produce errors with low quality data.

Step 4 Ion Assignment & Data Visualisation

For mass profiling data, a deconvoluted mass profile is often reconstructed (Figure 22 inset). In targeted mode TDMS the putative sequence of the protein is often already known. The experimentally measured protein mass can therefore be compared to the expected mass, with any difference suggesting the presence of an ASF, SAP or PTM. The MS/MS data can then be used to confirm the modification status of the protein. The list of monoisotopic fragment masses produced from the peak picking and deconvolution step is used to match fragment ion masses against a reference list of predicted ions, calculated *in silico* from an inputted protein sequence. A mass tolerance either in Dalton or parts per million (ppm) is used at this point to avoid false assignments. The ion type searched for is obviously dependent on the fragmentation mode employed. Loss of neutrals (H_2O, NH_3 or AA) may complicate spectra but can also be used to confirm ion assignment.

This matching can either be performed manually or with several software packages such as Biotools (Bruker), BUPID Top-Down, BIG MASCOT[111], SEQUEST[112], OMSSA[113] and ProSightPTM[114] (sold as ProSightPC by Thermo). The result is usually visualised as a fragment map which depicts the protein sequence with fragment ions represented as blocks or lines. Again there is no universally accepted method for representing this data and groups have developed their own preferred formats.

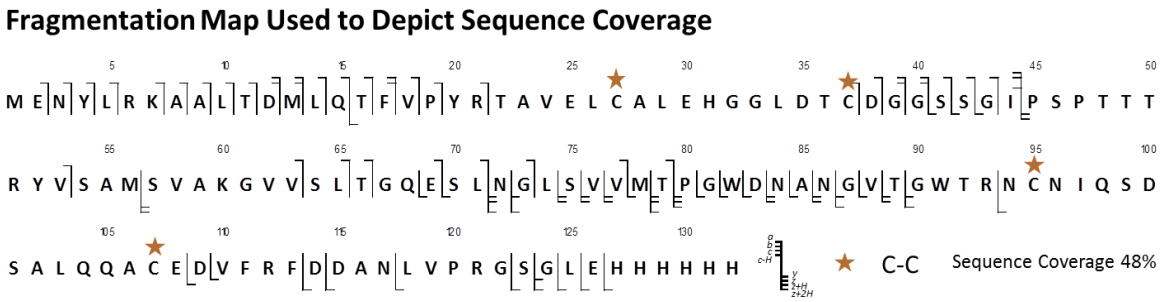


Figure 25 - Typical fragmentation map used to visualise top-down data

Step 5 PTM Assignment & Scoring

When PTMs or ASFs are suspected the procedure for ion assignment varies depending on whether they will be assigned manually or not. Often some prior biological knowledge coupled with a detailed examination of the mass profile can give some clues to the number and identity of PTMs present, and presents a good starting point. Definitive PTM identification and localisation is performed with the MS/MS data. In manual type assignment it is assumed that the spectrum should be sufficiently good to contain enough fragment ions. For any PTMs that are known to be present (chemical modifications such as carbamidomethylation of cysteines for example) the mass of the PTM is added in the appropriate place on the protein sequence and the theoretical ion masses recalculated.

A "first pass" at ion assignment is then performed and the data visualised. Often a sequence of fragment ions will be assigned and stop abruptly at a particular residue. This is a strong indication that this residue is modified. If no ions are assigned at all then this is a strong indication that the protein sequence is incorrect or the termini are modified. Once a putative site of modification has been identified previous biological knowledge can be used to test modifications on a particular site. The site is modified with the mass of the PTM, theoretical ions recalculated and assignment reattempted. If the ion sequence continues past this residue then one can reasonably assume that is modified. If not then other PTMs must be trialed. There are a number of manual techniques for PTM identification when the PTM is completely unknown or unexpected. This iterative process is continued until the mass of the modified protein matches that obtained from the mass profiling experiment.

Automatic Identification of PTMs in Targeted Mode TDMS

The process of PTM assignment can also be performed using specialised modes in a number of software tools. There are two types of tool available, the choice of which depends on how much information that is already known about the PTM (identity and localisation). The first type

requires the user to propose expected modification and is similar to the process performed during shotgun PTM identification in bottom-up proteomics. Tools that operate in this way include BIG MASCOT, SEQUEST, OMSSA and ProSightPTM.

The second type of tool will try to find completely unknown PTMs. There are several ways to do this and the details of the approach depend on the tool. In the "Delta m (Δm) Mode" of ProSightPTM the software calculates the difference between the observed and calculated molecular mass (Δm) at the protein level as a basis for localising PTM during ion matching.

A different approach to PTM localisation is provided by MS-TopDown[115] and its newer version MS-Align+[116]. This software finds spectral alignments between top down MS/MS spectra and truncated protein segments. With the precursor mass as an additional input and when data quality is high this tool can narrow down PTM containing regions and suggest PTM masses without any prior information. This approach is particularly powerful for *de novo* PTM identification.

Scoring of PTM Assignments

In targeted mode the protein ID is often known, therefore scoring generally only concerns different PTM assignments. When multiple PTMs are present there are often many ways to arrange the PTMs on the protein backbone and scoring systems can help assign confidence to each of these possibilities. Similarly to bottom-up proteomics when top-down experimental MS/MS data is processed automatically, some system, of scoring the protein spectrum match (PrSM) is applied to enumerate the confidence in the assignment. Different scoring metrics are used depending on the software chosen for data analysis.

Probably the most widely used scoring system is the Kelleher P-score as implemented in ProSightPTM[114]. The basis of the P score is shown in equation (4) where x is given by equation (5), M_a is the mass accuracy of the MS^2, n is the total number of fragment ions and f is the number of matching fragment ions.

$$P = P_{f,n} = \frac{(xf)^n \cdot e^{xf}}{n!} \tag{4}$$

$$x = \frac{1}{111.11} \cdot 2 \cdot M_a \cdot 2 \tag{5}$$

The P score (and related q score) essentially scores assignments based on the total number of fragment ion matches but there have been some problems raised with its appropriateness because of its relative simplicity. A newer "lambda score" is being worked on by Richard LeDuc (Indiana University) in collaboration with the Kelleher group. This new score is based on Bayesian probability, where prior information can be integrated into the scoring model. The function used

to calculate the score promises to take into account more parameters such as ion intensity and the propensity of cleavage between certain amino acids.

The construction of appropriate scoring functions remains an area of top-down proteomics that is in development and a topic of on-going research and intense debate.

7.2. Examples of Targeted Mode Top-Down Mass Spectrometry

Building on early studies on model proteins, TDMS is increasingly being applied to biological samples and has found particular utility in field of human health[117]. For example Ge's group has used targeted mode TDMS to study myocardial dysfunction.

In a 2011 report they published the first clinical application of top-down MS-based quantitative proteomics for biomarker discovery from tissues[118]. After affinity purification of Troponin I, a candidate biomarker for chronic heart failure, from both post-mortem and donor heart tissue, high resolution mass profiling was used to show that Troponin I was expressed in multiple proteoforms bearing different number of phosphate PTMs. They then quantitatively mapped the proportion of phosphorylation in samples taken from healthy hearts and those from patients suffering mild hypertrophy, sever hypertrophy, and congestive heart failure using (Figure 26).

Remarkably they found that the extent of phosphorylation correlated with disease progression and decreased as contractile dysfunction became increasingly severe. A similar decrease was shown when fresh tissue recovered after surgery from healthy donor hearts was compared with tissue from hearts exhibiting end-stage failure. Combining the mass profiling results from a number of patient samples in a large scale study enabled the proportions of non-modified, singly, doubly and triply modified phosphoforms to be statistical validated.

Using ECD fragmentation performed on a 7T LTQ-Ultra FT-ICR MS they also mapped the phosphorylation sites in these proteoforms and showed that during disease progression dephosphorylation occurred first at Ser^{22} then at Ser^{23}. These results demonstrate the power of TDMS to provide an overview of expressed proteoforms, provide label-free quantitative information, and then be used in a target approach for deep characterisation of a novel proteoform biomarker.

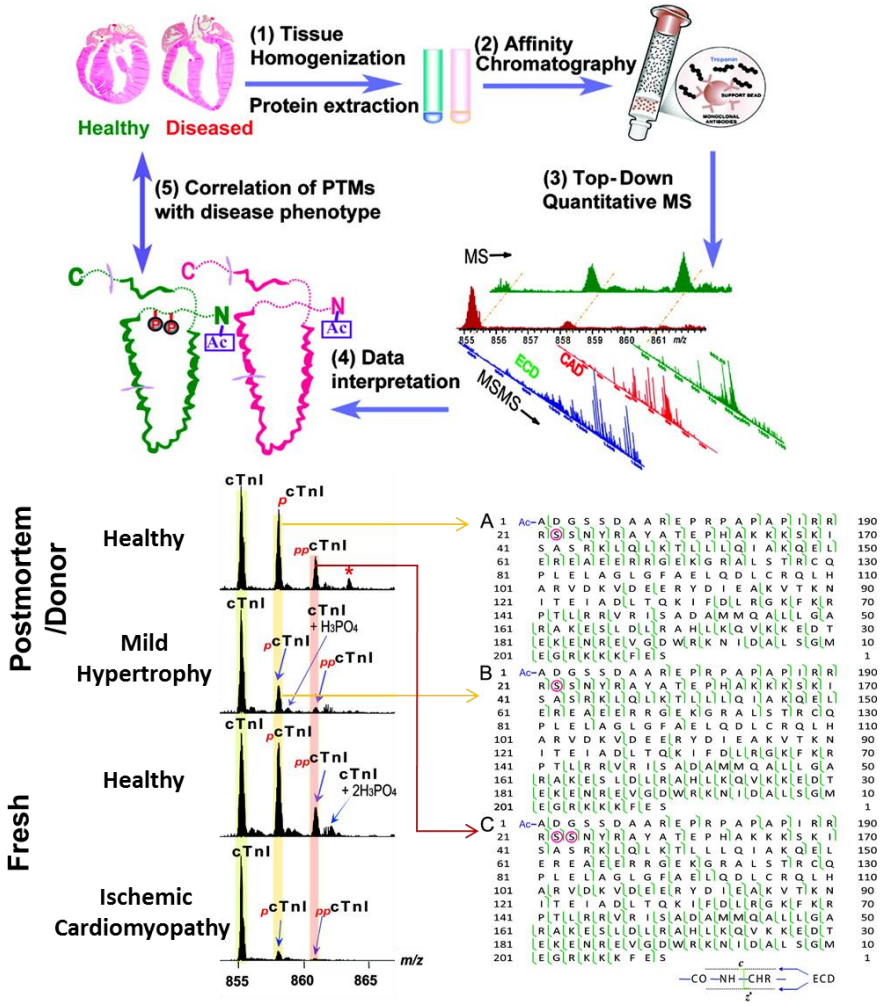


Figure 26 - Workflow used for Troponin I (cTnI) purification from heart tissue, quantitative MS for proteoform mapping and MS/MS for phosphorylation site identification. (Top) Mass profiles of the Troponin I proteoforms from healthy and diseased hearts. ECD MS/MS of the monophosphorylted proteoform from A) healthy & B) diseased hearts and C) the dephosphorylated form from a healthy heart. (Bottom) Adapted from Zhang et al.[118]

Another area where top-down mass spectrometry has been extensively applied is that of histone proteoform mapping[119-123]. Histones are a major part of the chromatin around which DNA is spooled and are known to carry multiple PTMs such as lysine and arginine methylation, lysine acetylation and serine or threonine phosphorylation. The combinations of these PTMs generate a histone code that directs chromatin related cellular processes. Core histones can be divided into four families [H4, H2B, H2A and H3]. In a recent report from the Paša-Tolíc group a total of 708 histone proteoforms were identified from a sample of purified HeLa cell core histones[124].

To separate histone forms prior to MS a metal free online 2D LC-MS/MS separation strategy was used where the major histone families were first separated by RPLC then proteoforms were further separated in a second dimension by WCX-HILIC coupled to a LTQ Orbitrap Velos Mass spectrometer (Figure 27). The system was as automated as possible to avoid time-consuming and labour intensive offline fractionation. Top-down MS/MS was performed with both ECD and CAD fragmentation and the results combined for maximum sequence coverage and precise PTM localisation. The results from MS/MS were processed using ProSightPTM, and even thought the identities of the core proteins are unknown, this report provides an example where automated PTM assignment and scoring becomes particularly useful if not essential. This study represents the largest investigation into histone modification performed to date and provides yet another example of the high level of diversity that proteoforms present bring to the proteome.

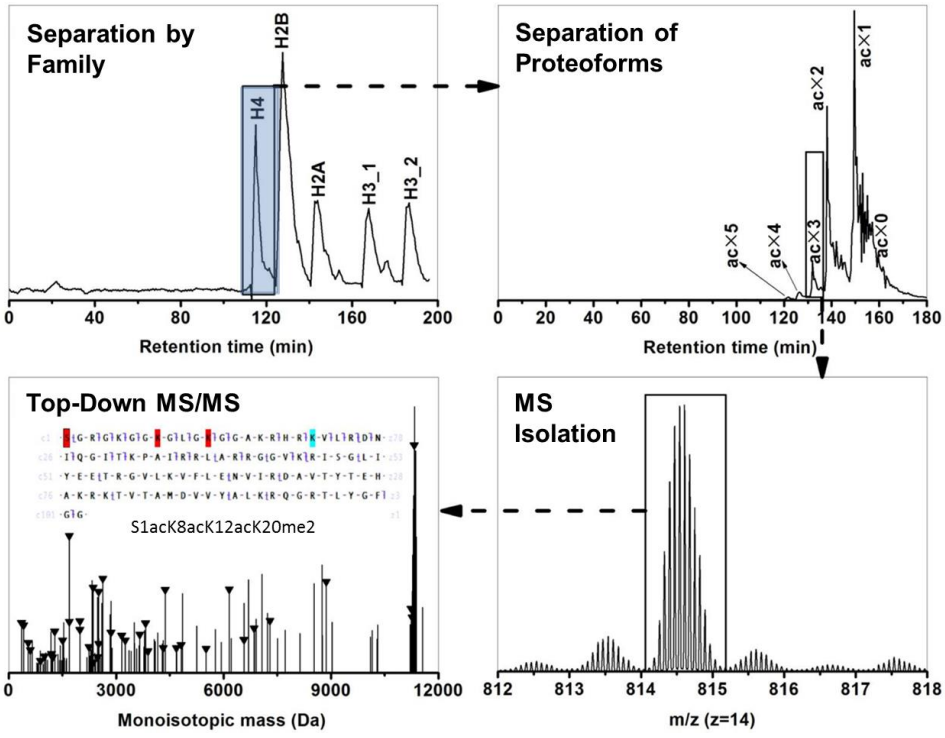


Figure 27 - 2D LC-MS/MS strategy for separation of core histones by family then further separation of proteoforms and top-down MS/MS. PTM assignment is shown for a component of the ac×3 fraction. Adapted from Tian *et al.*[124]

7.3. Discovery Mode Top-Down Proteomics

Discovery mode is a natural evolution of targeted mode TDMS where top-down experiment is performed in a high-throughput fashion on a more complex sample. The ultimate goal of discovery mode TDMS is the complete characterisation of all proteoforms present in an entire proteome. For this reason discovery mode is often simply referred to as top-down proteomics. The realisation of a robust top-down methodology has taken many years to perfect because further experimental demands are introduced by the increased number of analytes and a shorter experimental window imposed by the online MS/MS experiment. Some of these specific technical challenges and solutions are outlined below.

Pre-Fractionation & Sample Separation & Sample Injection

In discovery mode TDMS proteins are often separated by RPLC and delivered to the mass spectrometer in a fashion analogous to that found in the high-throughput bottom-up proteomics. Separation of proteins by RPLC results in a resolution lower than that achieved on peptides and so multidimensional fractionation methods are used to reduce the sample complexity to avoid too many co-eluting species[125].

Methods that separate proteins in solution are particularly desirable as they are amenable to multiplexing and facilitate integration into the top-down proteomics workflow. Important techniques that have been implemented into top-down workflows include, SEC and GelFree (molecular weight)[126], sIEF[127] & CE (pI)[127-129], cation and anion exchange chromatography (basic and acidic residues - related to pI), HILIC (hydrophilicity)[130, 131] and RP chromatography (hydrophobicity). Modifying these techniques for use with proteins however can be nontrivial and sometimes may require high concentrations of detergents, chaotropic agents or salt to maintain protein solubility. Dialysis for solvent exchange or precipitation is often required prior to sample injection in order to maintain compatibility with electrospray. These steps complicate the workflow and great care is often required to avoid sample loss.

Considerations During MS & MS/MS

The major difficulty in performing on-line top-down characterisation is the reduced time available for protein fragmentation on an LC time scale (<1-2 minutes). This limits the time available for scan acquisition. The factors that increase the duty cycle such as scan accumulation to improve S/N in both MS and MS/MS, electronic fragmentation modes (long activation time) over vibrational ones (short activation time), increased resolution though longer detection times are often performed at the expense of selection of other precursors and often a balance must be struck between deeper proteome coverage and more complete proteoform characterisation. This is a

particular problem on Bruker FT-ICR instruments in which the stages of the MS/MS experiment (accumulation, fragmentation, and detection) are performed in a strictly linear fashion. On newer Orbitrap systems these stages are overlapped for example while detection in one cycle is being performed accumulation is started in the next. This greatly increases scan speed. Since vibrational modes result in a greater percentage of parent to fragment ion conversion these are most often employed to achieve protein ID in top-down proteomics experiments. Electronic modes are often used to provide complementary fragmentation or identify labile PTMs.

Data Analysis, Protein Identification and PTM assignment in Discovery Mode

In discovery mode the approach to data analysis becomes much more akin to bottom-up proteomics. There is no question of performing manual data interpretation and thus appropriate software such as ProSightPTM, BIG MASCOT or MS-Align+ is required to perform both protein ID and PTM characterisation.

There are two basic strategies to achieve protein identification; the first uses the intact protein mass obtained from deconvolution of the MS scan plus fragmentation data to match, score and rank both non-modified and modified proteins. This is similar to the automatic approach described for targeted mode and thus BIG MASCOT, SEQUEST, OMSSA and ProSightPTM can be used for this type of identification. Extensive fragmentation is not required to establish confident protein ID with around 6-10 good quality fragment ions often being sufficient. This alleviates some of the experimental pressures required from extensive sequence coverage.

The second approach may not use the MS data but simply interrogates the data from MS/MS, looking for ladders of ions that correspond to putative sequence tags. These tags are then matched against the database creating a smaller set of proteins against which the full fragmentation data may be fully matched and scored. This option is also available in ProSightPTM. It is also used by the USTag tool which can perform protein ID but does not score matches[132]. MS-Align+ also relies on matching sequences of ions. At the very least this approach demands good sequence coverage in some localised areas of the protein. When deciding between multiple PTM matches an appropriate scoring system score becomes an incredibly important parameter as does the development of approaches to calculate an FDR. These are both topics of on-going research.

When deep characterisation of proteoforms is required and the search space is expanded to look for multiple combinations of multiple PTMs (often with a very large mass window) database searching becomes a highly computationally demanding procedure. This requires extension of software tools to support parallel computing on multi core architectures. Even then data treatment may take several days or even weeks to complete.

7.4. Examples of Discovery Mode Top-Down Proteomics

Large scale top-down proteomics studies have grown and in number and complexity since the first reports in the early 2000s and have now been performed on at least one proteome from each of the three domains of life, and the two model proteomes often used in bottom up proteomics; *Saccharomyces cerevisiae* (baker's yeast) and the human HeLa3 cell line. The results from major top-down proteomic studies published since 2003 are represented in (Figure 28).

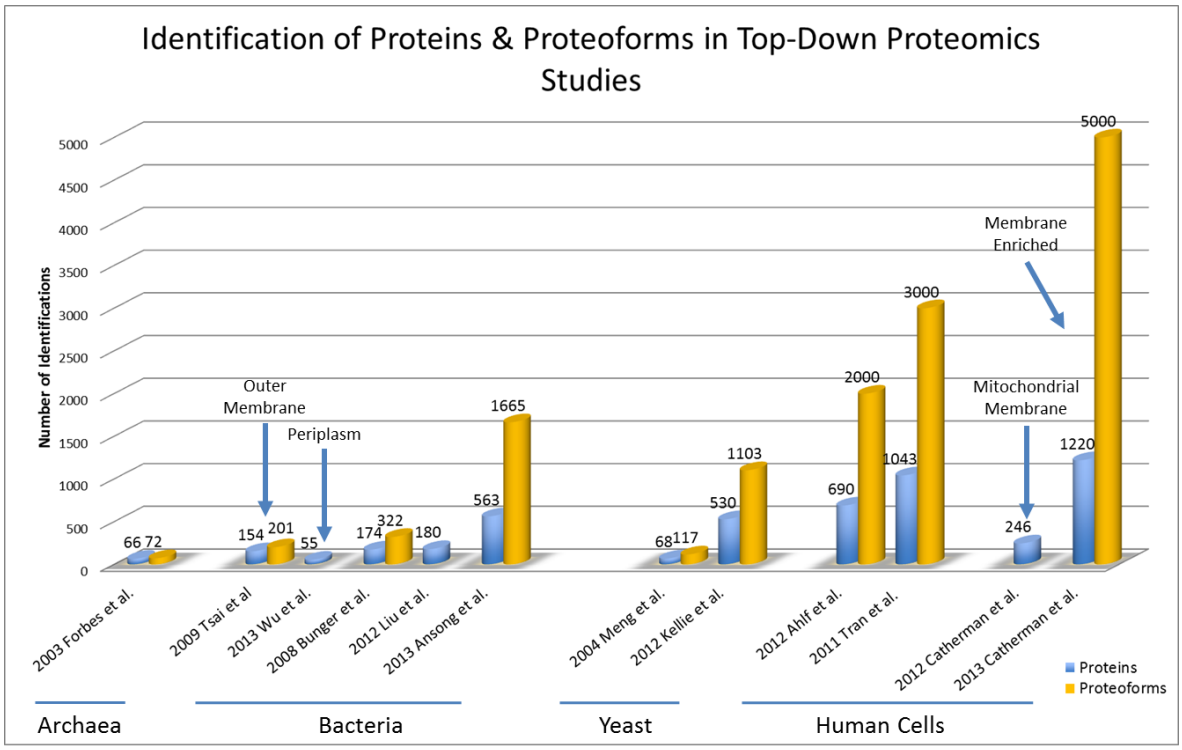


Figure 28 - Histogram of proteins (blue cylinders) and proteoforms (orange cylinders) identified in the major top-down proteomic studies published since 2003. Studies referred to are Forbes[133], Tsai[134], Wu[135], Bunger [136], Liu[116], Ansong[137], Meng[138], Kellie[139], Ahlf[140], Tran[141] and Catherman (2012)[142] (2013)[143]

There are two major trends that can be seen in the histogram. The first is the big improvement in both protein and proteoform ID since the introduction of the high–field Orbitrap in 2011. The second and the most striking feature is the number of proteoform IDs compared to the number of proteins. This is threefold higher in the studies from Ansong[137], Kellie[139], Ahlf[140] and Tran *et al.*[141] and nearly five time higher for the most recent study by the Kelleher group[143]. This provides a strong indication of the huge diversity that is imparted to all proteomes through proteoforms. Some of these studies are briefly presented in greater detail to outline some of the specific features of the top-down proteomics approach.

Until very recently the largest top-down proteomics study performed to date was published by Keheller group in 2011 and concentrates on the identification rather than characterisation of as many proteoforms as possible in HeLa S3 cells[141]. Fractions enriched in nuclear, cytosolic and mitochondrial membrane proteins were prepared from whole cell lysates. After multidimensional fractionation firstly by sIEF (four to five fractions) then GelFree (nine fractions) the proteome, which has now been separated into 36-45 fractions, was subjected to RPLC TDMS using a 12T LTQ FT Ultra FT-ICR mass spectrometer. This separation strategy is shown in Figure 29.

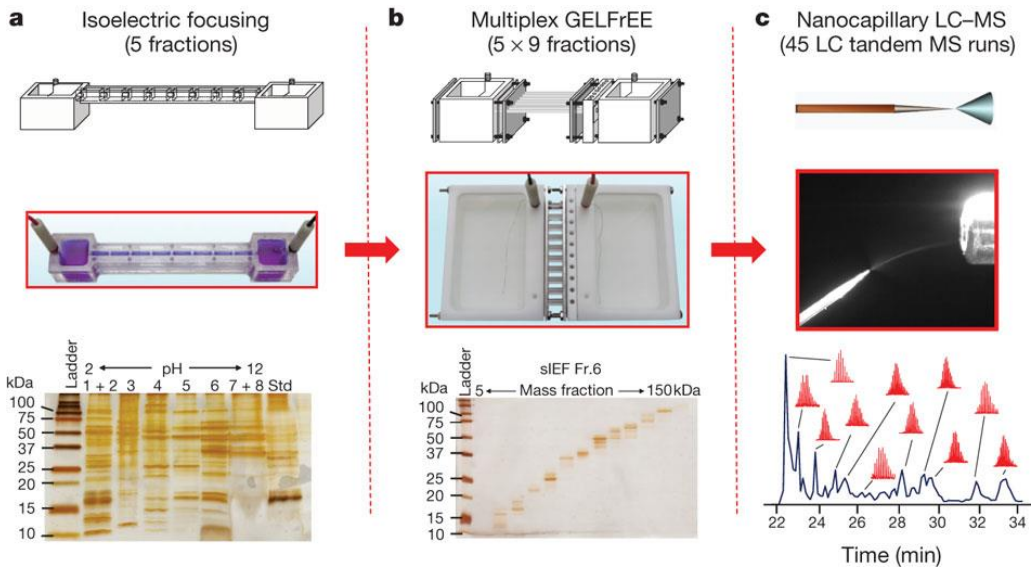


Figure 29 - Multidimensional separation strategy used for top-down proteomics of HeLa cells. Taken from Tran *et al.*[141]

A separate MS/MS method was used for proteins sparing three different mass ranges and both ISD and CAD modes were employed for protein fragmentation. Data were combined from three complete fractionation experiments performed on the nuclear extract, five on the cytosolic extract plus two GelFree RPLC-MS/MS runs on the mitochondrial membrane extracts totalling around 378 total LC-MS/MS runs. The resulting MS/MS data were first processed by several pieces of software for protein mass deconvolution then by a version of ProSightPTM for protein identification and PTM assignment that has been adapted to run on a cluster. 1,043 unique proteins and over 3,000 proteoforms were identified in this manner at an FDR of 5%.

Discovery mode TDMS has also been used by the same group for the identification and characterisation of mitochondrial membrane proteins. After mitochondrial separation from HeLa S3 cells and membrane purification, GelFree separation was followed by RPLC-MS/MS using a 12T LTQ-Velos-FT-Ultra MS. CID and ISD were employed as the fragmentation methods in a top 2 MS method. Pooling results from 30 GelFree fractions a total of 246 proteins were identified, 107

were mitochondrial, of which 83 were characterised as membrane proteins based on high confidence transmembrane helix prediction. Of these 83, 53 were identified with N-terminal Met cleavage and or acetylation. A myriad of other PTMs were found including myristoylation, trimethylation, ProGlu formation and an I→V SAP consistent with a reported SNP. Membrane proteins are notoriously difficult to handle by bottom-up methods and this study clearly shows the utility of the top-down approach for analysing proteins of this class.

A very recent study again by the same group has extended the study of membrane proteins to membrane enriched fractions of H1299 cells[143]. In this report 1,220 proteins were identified, 856 of which were designated as membrane associated by gene ontology. In addition over 5,000 proteoforms were observed. This sets a new bench-mark for the scale of top-down proteomics investigations.

In another report Ahlf *et al.* evaluated the new high field Orbitrap for TDMS proteome mapping using H1299 human lung cancer cells on a Velos Elite system[140]. A GelFree approach for initial sample fractionation was used followed by RPLC-MS/MS. A data dependent top 3 method was applied where HCD, ETD and CAD were performed sequentially on proteins with masses <25 kDa and ISD and HCD on proteins > 25kDa. Results were processed using ProSight (Figure 30).

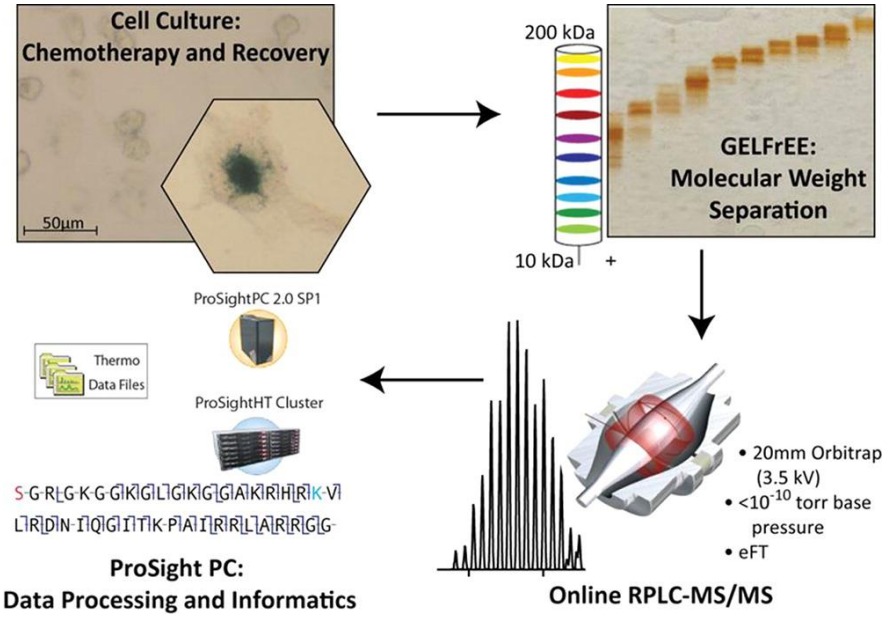


Figure 30 - Workflow used by Ahlf *et al.* for discovery mode TDMS of the proteome of H1299 human lung cancer cells. Taken from Ahlf *et al.*[140]

The complementarity of the fragmentation methods for protein ID was shown through a separate experiment on a single GelFree fraction. Results from 9 full experimental replicates were pooled leading to the identification of 690 unique gene products at an FDR of 5% and an impressive 2,366

proteoforms including 337 phosphorylations, 75 monomethylations, 58 dimethylations, 31 trimethylations and 892 acetylations. Note the high 3:1 proteoform/protein ratio in this study.

Moving to the bacterial proteome, the study by Ansong *et al*. is particularly important as it shows that discovery mode TDMS has developed sufficiently to be used in differential proteomics. The bacterium *Salmonella* Typhimurium was grown in both normal and minimal media representing basal and "infection-like" conditions. Without performing sample fractionation, RPLC-MS/MS was performed on a whole cell lysates of bacteria grown in these two conditions using a long nano-LC column (80 cm), long 250 min gradient and both HCD and ETD fragmentation on an Orbitrap Velos mass spectrometer. Data was processed using the MSAlign+ software tool resulting in the identification of 563 proteins and 1,665 proteoforms at a 5% FDR.

Of the many proteoforms identified, 25 exhibited S-thiolation. Thiolation was found to occur in two forms, glutathionylation or cysteinylation. Surprisingly the former appeared almost exclusively in the basal growth condition whilst the later was almost exclusive to the infection like condition. The PTM switch from glutathionylation to cysteinylation which occurred upon changing the growth conditions is evidenced for the YifE protein (Figure 31).

Many studies correlating proteomics to the disease state are differential and this work provides proof-of-principle that top-down proteomics can be applied to this area. When coupled with other top-down studies that have been performed on more relevant biological samples, such as that by Cabras *et al*. on the salivary proteome of patients with Down syndrome[144], it is clear that TDMS shows great promise to probe the roles that posttranslationally modified proteoforms play in disease.

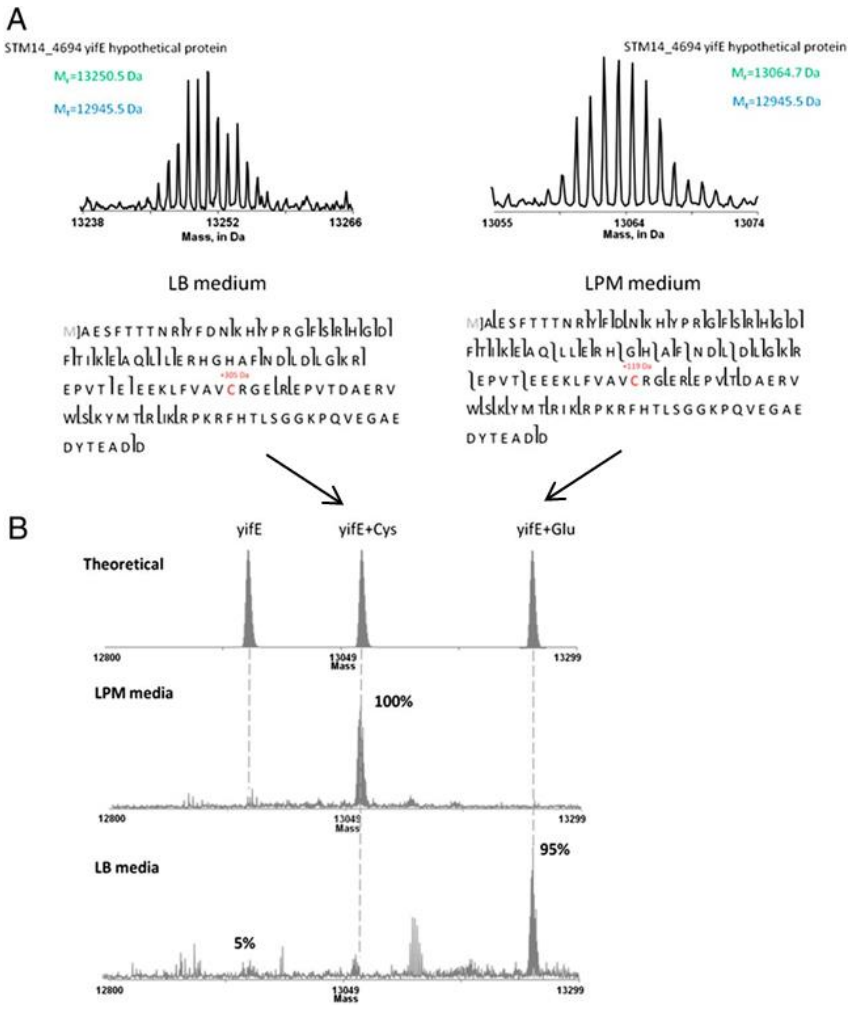


Figure 31 - A) Deconvoluted MS of YifE from basal and minimum media along with sequence coverage from top-down MS/MS showing the switch in S-thiolation forms. B) Quantification of the glutathionylated and cysteinylated proteoforms by mass profiling

Part II - Neisseria meningitidis - A Deadly Human Pathogen

1. *Neisseria meningitidis* and Meningococcal Disease

Neisseria meningitidis (Nm) is a Gram negative bacterium and commensal of the human nasopharynx. It is best known however as the etiological agent of cerebrospinal meningitis. First isolated by Weichselbaum in 1887, Nm exhibits a distinctive diplococcal shape, as shown when visualised by transmission electron microscopy (TEM) (Figure 32).

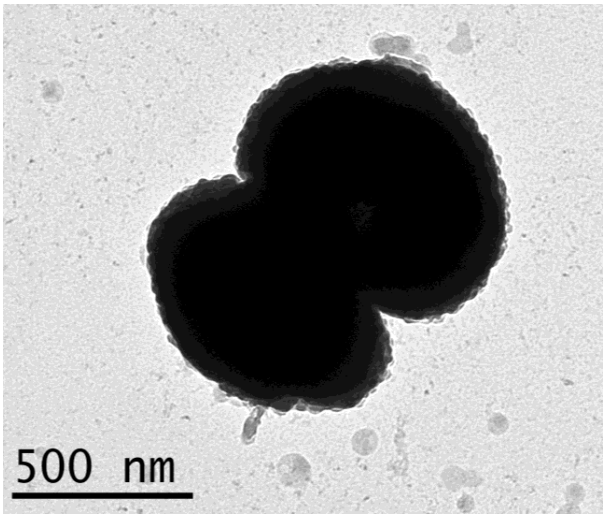


Figure 32 - Negative staining electron micrograph of a single Nm diplococcus of the 8013 strain

Pathogenic Nm is encapsulated by a polysaccharide layer that is lipid anchored into the bacterial outer membrane. Variation in sugars exposed on the surface of the capsule leads to different serological reactivity and the identification 13 distinct Nm serotypes (A, B, C, D, H, I, K, L, X, Y, Z, 29E and W135), of which five serogroups A, B, C, Y, and W135 are most associated with human disease[145]. Classification of the bacterium may also be performed in a more gene centric fashion by multi locus sequence typing (MLST)[146, 147]. In the case of Nm, this involves defining clonal complexes based on the sequence of seven housekeeping genes. This enables the genetic similarity of different strains to be evaluated and compared. The MLST approach is now widely adopted and has been used and to class commensal strains in carriage studies[148] and track hyper invasive lineages as they spread.

1.1. Epidemiology

Nm is an exclusively human commensal and is thought to be present in between 10-20% of the human population at any given time. Carriage rates have been reported to be significantly higher in closed or semi-closed populations such as university students and military recruits; presumably due to an increase in aerosol based transmission. The majority of the time Nm remains a benign commensal, but for as yet unknown reasons it may develop into a highly dangerous pathogen.

In western countries meningococcal disease is mostly caused by Nm serogroups B and C. Outbreaks are often geographically localised, sporadic and unpredictable resulting in a substantial public health burden[149]. Meningococcal disease is endemic in other parts of world, such as central Africa, where it is often attributed to serogroup A. The area of sub-Saharan Africa stretching from Senegal in the west, to Ethiopia in the east is particularly prone to frequent and deadly meningitis outbreaks and is therefore often termed the meningitidis belt. The number of cases in this region usually rises in the dry season and during large population migrations such as the Hajj pilgrimage where infection rates here have been known to reach 1,000 cases per 100,000 inhabitants (the WHO definition of an epidemic is > 100 cases per 100,000). Meningitis in Africa therefore constitutes an enormous global health burden and is the focus of international prevention efforts.

1.2. Meningococcal Disease

Meningococcal disease mostly manifests itself through early flu like symptoms. These can develop into purpural skin lesions, sensitivity to bright light, fever, vomiting, stiffness around the neck area and a severe headache. Meningococcal infection often develops very quickly and within a several hours may result in cerebrospinal meningitis and/or severe sepsis coupled with circulatory collapse (meningococcemia). Both of these conditions are potentially deadly and have a high rate of mortality that is around 50% if left untreated.

The rapid onset of the disease is exemplified by the recent, high profile death of a researcher in San Francisco, CA, USA who contracted meningitis after accidental exposure in the lab where he worked, only 17h after manifestation of the initial symptoms[150]. Diagnosis of meningococcal infection requires culture from blood or CSF (often obtained though lumbar puncture) and may be easily treated with appropriate antibiotics. This lowers the mortality rate to around 10-15%[151], but even if the patient survives the infection, subsequent disability and even limb loss are common outcomes. Rapid disease progression and seemingly benign initial symptoms often delay effective early treatment and make Nm an incredibly dangerous pathogen in countries with sophisticated medical care and devastating in areas where this is not the case.

1.3. Vaccination

Many cases of meningitis can be prevented though effective vaccination and both polysaccharide based and polysaccharide-protein conjugate vaccines have been developed against Nm serogroups A, C, Y, and W135. A highly effective quadrivalent vaccine catering for all of these serogroups was released in 2005. Recent programmes of widespread vaccination in countries throughout the meningitidis belt are having a particularly positive effect with the last major epidemic occurring in 2009[152].

Serogroup B strains (NmB) present a particular problem for vaccine development. In this serogroup the polysaccharide capsule is formed of the glycan (2→8)-α-*N*-acetylneuraminic acid which is similar to several human antigens. It has thus been deemed an unsuitable vaccine target for fear of autoimmune development[153]. Robbins *et al.* have however recently revaluated this position and argue that cross reactivity may not be as serious a problem as previously thought[154].

Other surface exposed molecules have also been used in outbreak specific meningococcal vaccines and are currently being trailed in several efforts to develop a broad based Nm vaccine that would be effective against NmB (Pfizer-Wyeth, Novartis)[155]. These newer vaccine candidates contain multiple components to increase strain coverage and increase efficacy, many of which are derived from surface proteins of the bacterium itself. In the wild type bacterial population these surface antigens are found to be hypervariable, and thus epitopes of emerging pathogenic strains may not be the same as those isolated previously[156, 157]. Vaccines are therefore expected to become less effective over time, even if they contain multiple components. A better understanding of the molecular mechanism of meningococcal disease may reveal alternative potential antigens and novel strategies to combat and prevent infection.

1.4. Escaping Immune Detection and Mechanisms of Disease

In addition to the polysaccharide capsule, *Neisseria meningitidis* harbours a number of surface antigens that influence bacterial interactions with the host (Figure 33)[156]. These include type IV pili, the lipopolysaccharide layer (LPS) and several ubiquitous surface proteins such as Opa and Opc (involved in adhesion to host cells), PorA (forms a cation selective pore) and NadA (an invasin promoting adhesion and invasion into host cells[158]). *Neisseria meningitidis* has evolved the ability to vary the structure of these surface exposed motifs in order to escape the host immune system. This is achieved using two main strategies; phase, and antigenic variation[159].

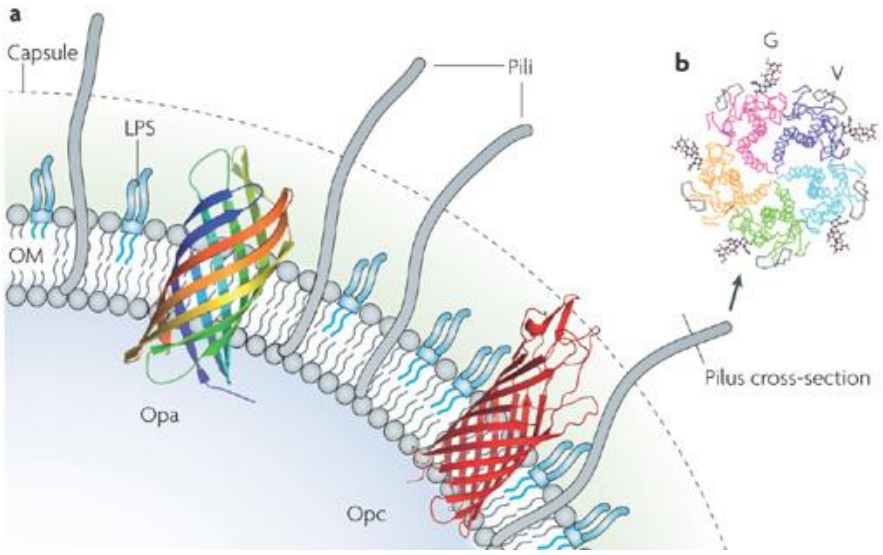


Figure 33 - Prominent outer-membrane (OM) components of *N. meningitidis* that influence bacterial interactions with host cells. Taken from Virji[156]

Phase variation is a mechanism for on/off switching of protein expression[160, 161]. Phase variable genes are characterised by repeat sequences or homopolymeric tracts in, or upstream of, the open reading frame (ORF) which increase the propensity for slipped strand mispairing during DNA duplication. This causes contraction and expansion of these repeat sequences, pushing the ORF in and out of frame and switching the relevant gene on or off. There are 24 known families of phase variable genes in *Neisseria* spp. and a further 25 strong candidates[162]. Over half of these encode for surface expressed proteins (such as Opc, Opa, PorA NadA), enzymes known to modify surface proteins (PglA, PglE, PglG, PglH) or are proteins involved in the synthesis of LPS (LgtA, LgtC, LgtD). The polysialyltransferase SiaD required for polysaccharide capsule expression is also phase variable[163].

Antigenic variation refers to the expression of multiple antigenically distinct forms of a single gene product within a clonal population[164]. As such, expression of proteoforms with different PTM complements and expression of proteoforms with different primary structures both constitute forms of antigenic variation. Antigenic variation of the protein primary structure may be the result of gene transfer which can be defined as unidirectional recombination of variant DNA into a homologous locus. This process occurs in Nm by a number of different molecular mechanisms. One of these mechanisms relating specifically to PilE, the major constituent of type IV pili is represented in Figure 36. This type of antigenic variation results in the expression of proteins which have hypervariable primary structures. Both OpA and PilE are known to undergo antigenic variation[159].

Employing both antigenic variation and phase variation to rapidly alter its surface structure aids *Neisseria meningitidis* in avoiding immune detection. In this regard it can therefore be consider

somewhat a “master of disguise”. This can pose particular issues for both vaccine design and bacterial clearage during infection.

Nm may become pathogenic after it has crossed the epithelial layer and gained access to the circulatory system. The trigger for this event is currently unknown. Once in the bloodstream the bacteria adhere to the blood vessels, multiply and aggregate. This colonisation of the vasculature triggers coagulation of the blood, inflammation and loss of vasculature integrity causing potentially life threatening septicaemia[165-167]. Proliferation in blood vessels also allows access to the blood brain barrier from where the meningococcus can gain access to the cerebral spinal fluid (CSF). For both initial colonisation of the nasopharyngeal mucosa and at each stage of this infection model, extracellular organelles called type IV pili (T4P) are found to be key protagonists mediating multiple interactions between bacteria and the host[168].

2. Type IV Pili

T4P are long, extracellular, filamentous organelles common to numerous bacterial species including *Pseudomonas aeruginosa*, *Vibrio cholera, Escherichia coli,* Nm and the related pathogen *Neisseria gonorrhoeae* (Ng) which colonises the human urogenital tract[169]. Nm and Ng exhibit some important differences, their preferred biological niche being just one example, but share significant similarities with respect to pilus biology. Whilst we are particularly concerned with Nm, evidence drawn from both Nm and Ng is useful to present a complete picture of these highly complex and versatile organelles.

2.1. T4P Function

T4P are implicated in multiple, diverse processes including manipulation of host cells, electron transfer, biofilm formation and bacterial motility[170]. In Nm type IV pili are required for four important life processes: host-cell adhesion, bacterial aggregation, signal induction and transformation[170].

Bacterial adhesion to both endothelial and epithelial cells has been shown to be severely impaired or completely abolished in non pilated variants of *Neisseria* spp. thus indicating that T4P are indispensable for host cell attachment[171-173].

Microscopy of bacterial colonies has shown that pili from neighbouring bacteria are often observed to aggregate or bundle. These inter-bacterial ties are believed to promote colony integrity and adhesiveness[174] and it has been shown *in vitro* that bundling partially helps bacterial colonies cope with shear stresses similar to that found in the vasculature[175, 176].

Pili have also been found to trigger several signalling processes in host cells. Attachment of T4P is required for the deformation of the host cell membrane, which also aids in coping with shear stress[176]. Attachment of T4P to human brain microvascular endothelial cells has been shown to recruit the Par3/Par6/PKCζ polarity complex that is required for intercellular junction formation. Recruitment of this complex weakens junctions and is thought to promote crossing of the blood brain barrier[177]. This recruitment is not found to occur for epithelial cells[178].

Lastly, T4P are required for natural transformation; a process that involves uptake of exogenous DNA and integration into the bacterial genome. Transformation is useful for promoting genetic variation and DNA repair. T4P are heavily implicated in the initial step of DNA binding which must take place before internalisation can occur[174].

2.2. Biogenesis & Structure

T4P are biological macropolymers formed of repeating protein subunits arranged in a helical fashion to create the long (several μm), thin (5-8 nm) and flexible fibre observed as the pilus (Figure 34)[179]. In *Neisseria* spp. the major component of T4P is the major pilin PilE, a 14-18 kDa protein coded by the *pilE* gene. Pilus assembly occurs in the periplasm. PilE is translated in the cytosol and transferred into the inner membrane ready for assembly. Here it is processed, polymerised in the periplasmic space and pushed out through a pore in the outer membrane to form long filaments observed as pili. Pili are dynamic organelles and may also retract back inside the bacterium. This is thought to occur through depolymerisation of the fibre and is necessary for transformation and a type of bacterial locomotion called twitching motility[180, 181].

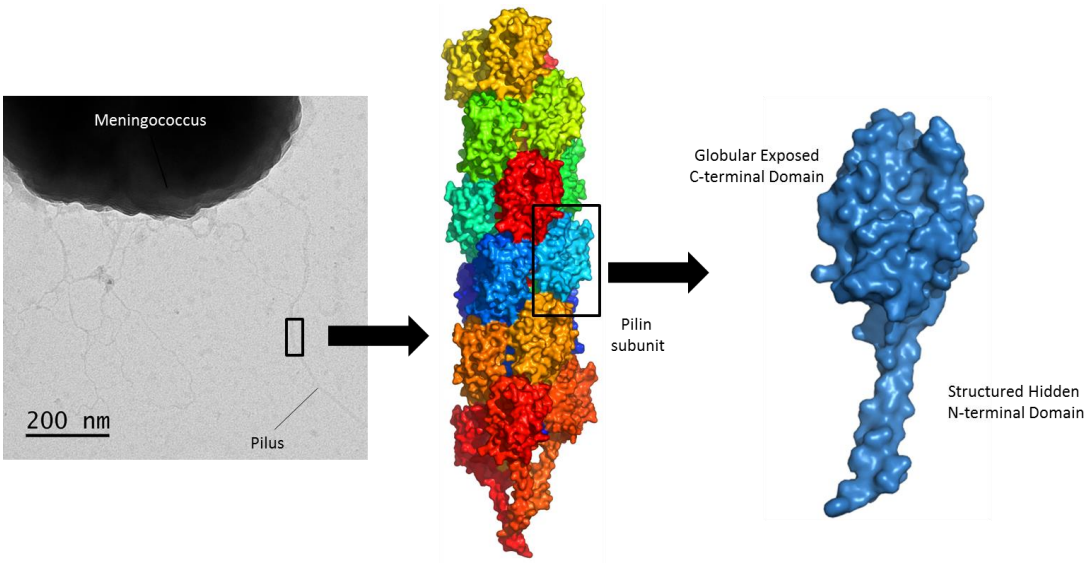


Figure 34 - Structure of the pilus showing helical arrangement of pilin subunits and the two major domains of the PilE monomer

Pilus expression in wild-type (WT) Nm requires 15 proteins including the major pilin PilE (PilC1/PilC2, PilD, PilE, PilF, PilG, PilH, PilI, PilJ, PilK, PilM, PilN, PilO, PilP, PilQ and PilW)[182, 183]. The specific role of many of these Pil proteins remains uncertain and the function of only a small group has been described in detail. PilD is a bifunctional enzyme that processes PilE into a mature form[184, 185]. PilF is a traffic ATPase thought to provide the chemical energy for pilus extension[186]. PilQ is the principal component of the outer membrane pore and together with PilM, PilN, PilO, PilP is thought to form a large periplasmic channel constituting a core transmembrane complex[187-189] in which PilG, PilH, PilI, PilJ, PilK and PilW are thought to play additional structural and stabilising roles[190]. PilC1 and PilC2 are involved in the fine tuning pilus adhesiveness and controlling host cell motility[191].

Seven additional proteins (ComP, PilT, PilT2, PilU, PilV, PilX and PilZ) are dispensable for pilus formation but play important roles in mediating the diverse array of pilus functions[192]. PilT is an ATPase that is responsible for pilus retraction. PilV, PilX and ComP (the minor pilins) share overall structural homology with PilE and are thought to be integrated within the pilus fibre since they co-purify with PilE after pilus purification. ComP has been linked to DNA binding[193], PilV to epithelial cell adherence[194] and the induction of signalling[176], and PilX has been found to promote bacterial aggregation[195]. PIlX is believed to counteract pilus retraction through a hook type D region that forms links with other PIlX subunits integrated into other bundled pili [194]. This increases friction between neighbouring fibres and impedes retraction.

PilE is by far the major component of the pilus fibre and biogenesis machinery and therefore has the greatest potential to mediate interactions with other biomolecules including those of the host immune system. The structure of PilE will therefore be examined in closer detail.

3. The Major Pilin - PilE

3.1. Structure of PilE

PilE is a ladle shaped protein with a long, hydrophobic, N-terminus that protrudes from a more globular C-terminal head (Figure 35)[196]. The hydrophobic alpha helical N-terminal "handle" plays a largely structural role promoting packing interactions on the inside of the pilus fibre. The globular "spoon" C-terminal domain of PilE is more hydrophilic with several important structural motifs that are exposed on the surface of the pilus. It is these surface exposed regions that mediate interactions with the biological milieu.

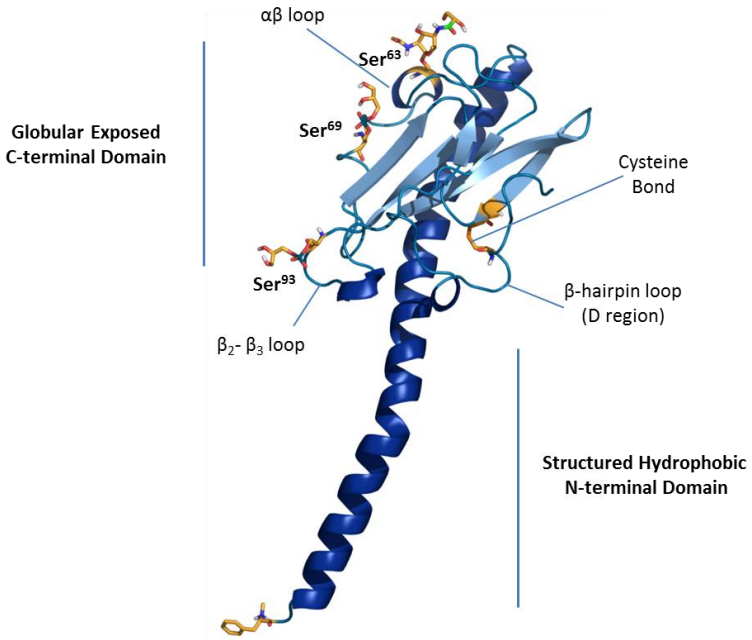


Figure 35 - Domains and structural features of the major pilin PilE

PilE is a small protein which varies in length from between 140 and 180 amino acids. The first 5-6 residues represent a short leader sequence that is cleaved posttranslationally by PilD before a conserved phenylalanine residue. The N-terminal region spanning the first 50 or so amino acids is virtually identical amongst all known PilE sequences. Outside of this region PilE may undergo considerable sequence variation. The middle proportion of PilE is described as mildly variable with local regions of hypervariability, whilst in the C-terminal D-region that is demarked by two cysteine residues, is particularly hypervariable and exhibits hardly any homology between strains. Figure 37 (top) shows an alignment of the "variable type" PilE sequences and is annotated with the protein secondary structure.

3.2. Sequence Variation

The hypervariability of the PilE primary sequence is another example of antigenic variation, used by Nm to promote immune escape. Indeed electron microscopy, X-ray crystallography and molecular modelling have shown that many of the sequence variable regions of PilE lie exposed on the surface of the pilus fibre. This is particularly apparent for the cystine bound D region loop.

The origins of this antigenic variation can be traced to homologous recombination of the *pilE* gene with several silent *pilS* cassettes that are clustered close to *pilE* in the bacterial genome[159]. *pilS* genes lack both the promoter and the initial 5' 150 bp sequence that corresponds to the N-terminus of PilE but share significant homology with the rest of the *pilE* gene. During recombination events DNA elements are transferred from *pilS* to pi*lE* in a unidirectional, RecA

dependent fashion and the original *pilE* gene is discarded. A carton of this process is shown in Figure 36.

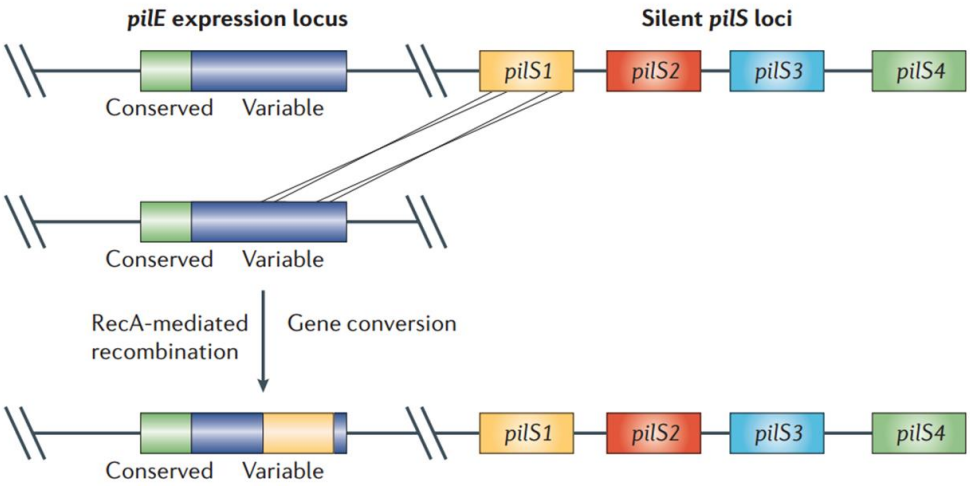


Figure 36 - Cartoon representing the homologous recombination of genetic material from the *pilS* cassettes into the *pilE* gene. This mechanism is responsible for antigenic variation of PilE. Taken from Davidsen *et al*.[159]

The molecular mechanism by which this exchange occurs is complex and a matter of current investigation[197-200] but the frequency of such gene conversion is remarkably high and has been estimated to occur at 1.6×10^{-3} events per colony forming unit per generation[201]. This method to escape the immune system has been a longstanding precept of pilus biology and is the principal reason why PilE was deemed unsuitable as a vaccine target in the late 1980s.

In 2010 a large bank of PilE sequences from clinical isolates of Nm was reported. These are shown in Figure 37. Unexpectedly, rather than being comprised entirely of the variable type, which is known to undergo antigenic variation, many strains showed significant and unprecedented conservation of their PilE sequence[202]*. The question therefore arises, "if antigenic variation of PilE promotes immune evasion, how do all of these strains that express an invariable PilE sequence escape immune detection?" The answer to this question is as yet unknown.

Note that in this paper and elsewhere PilE sequences are grouped into two classes based on reactivity with an SM1 monoclonal antibody. Whilst class I sequences are hypervariable and class II conserved, we find the term generates confusion and simply "variable" or "nonvariable" is an adequate description. The word "class" will be reserved exclusively for genetic organisation.

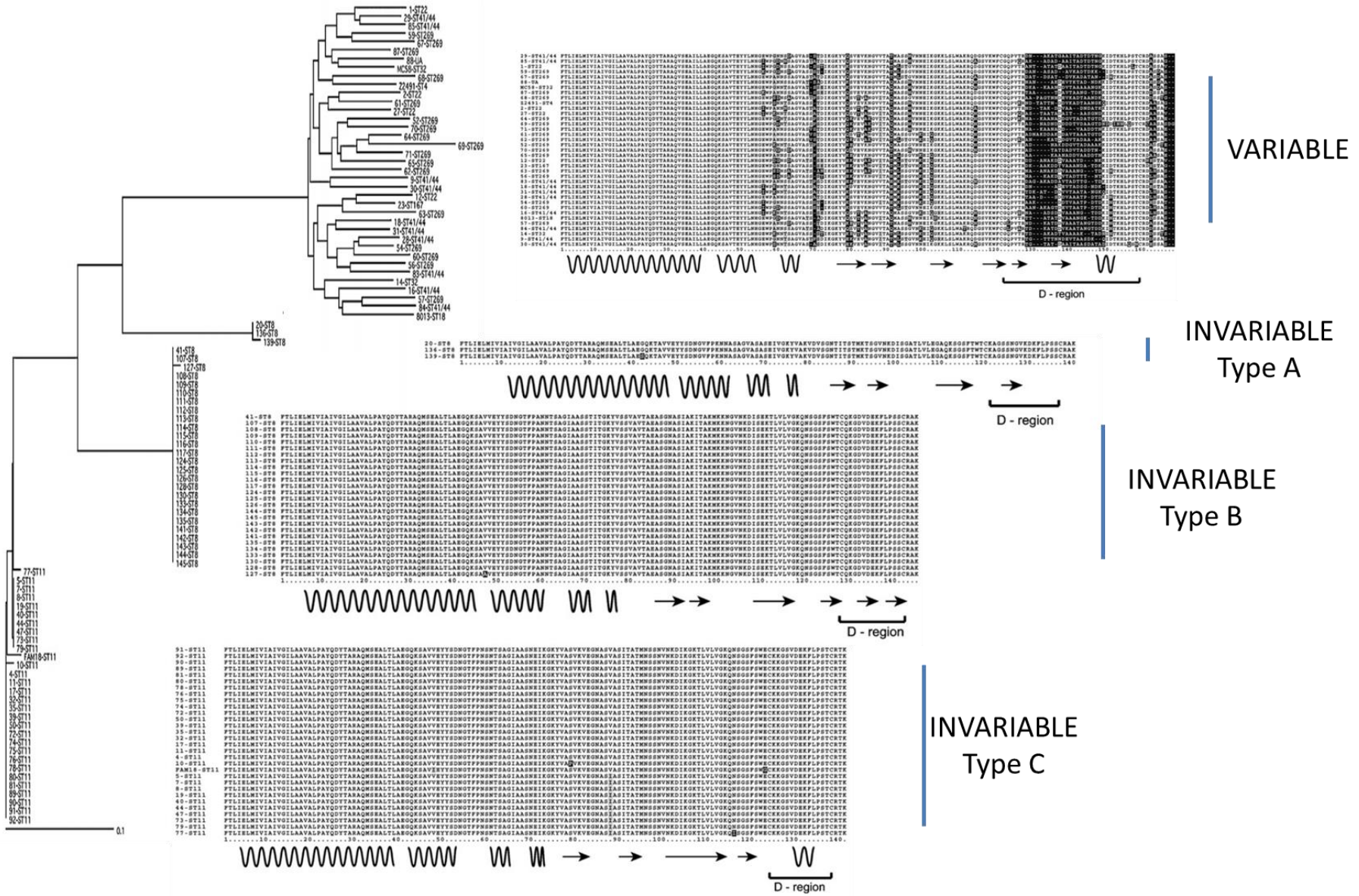


Figure 37 - Phylogram and sequence alignment of PilE from Nm isolates showing invariable and three groups of variable PilE sequences. Adapted from Cehovin *et al.*[202]

4. Posttranslational Modification of PilE

An additional means to promote structural diversity an antigenic variation is through posttranslational modification. PilE is a highly modified protein that has been found to harbour a number of unusual PTMs. After cleavage of the leader sequence, the bifunctional enzyme PilD methylates the N-terminusof PilE[185]. In addition to an intact disulfide bond between the cysteine residues enclosing the D region, these two PTMs are consistently present on all previously characterised pilins. The additional PTMs of PilE can be divided into two groups comprising *O*-linked glycans and phosphoforms. To date PTM of PilE has been investigated in only a limited number of Neisserial strains; two from Ng and three from Nm. Evidence from both Nm and Ng strains will therefore be grouped in order to present the most complete picture of pilin PTM.

4.1. Glycosylation

PilE may be glycosylated by two core glycans 2,4-diacetamido 2,4,6-trideoxy α-D-hexose (DATDH)[203] and 2-acetamido 4-glyceramido 2,4,6-trideoxy α-D-hexose (GATDH)[204] (Figure 38). These may be further elaborated by up to two additional hexose (Hex) subunits to form disaccharides and trisaccharides respectively. Pilin is always found to be glycosylated and in the strains examined to date Ser^{63} is consistently and exclusively reported to be the sole site of glycan attachment. The glycan is therefore located in the α-β loop region of PilE and exposed on the pilus surface. The evidence to support the identity and position of the glycan on the protein backbone comes from numerous studies, the conclusions of which have undergone refinement as experimental methods have developed. These reports are summarised in Table 1.

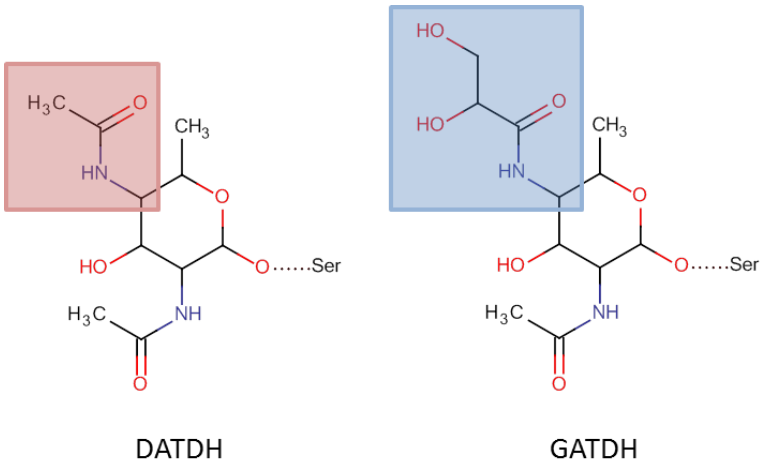


Figure 38 - Structure of the bacterial glycans DATDH and GATDH

Nm/ Ng	Strain	Parental Strain	Strain Derivatisation Details	Glycan			Phosphoforms			Comments
				Structure	Site	Method & Reference	Structure	Site	Method & Reference	
Ng	N400	MS11	Opa- mutant of MS11 (VD400) and then further modified with IPTG inducible *recA6*	GlcNAc-α1,3-Gal	Ser^{63}	X-ray [196]	P	Ser^{68}	X-ray [196]	Both glycan and P possibly misidentified
-	-	-	-	-	-	-	P/PG	Ser^{94} (minor)	X-ray [205]	Refinement of data in[196]
-	-	-	-	Ac-Hex-DATDH	Ser^{63}	MS [206]	PC/PE	Ser^{68}	MS [206]	PE changed to PC in *pilV* background & absent when pilV overexpressed
-	-	-	-	-	-	-	PE	Ser^{156}	MS [207]	Additional PE in ≈ 44% WT PilE. Further component not discussed in article
Ng	C30	MS11	non-piliated MS11 derived clone	α-D-galactopyranosyl-(1→3)-2,4-diacetamido-2,4-dideoxy-β-D-glucopyranoside (Gal-DADDGlc)	Ser^{63}	X-ray [179]	-	-	-	-
Nm	8013SB	8013	-	GlcNAc-α1,3-Gal	Ser^{63}	MS [208]	-	-	-	-
Nm	8013	8013	-	GATDH	Ser^{63}	MS [204]	-	-	-	First report of GATDH glycan
Nm	CS311#3	CS11	clonal variant[209]	Hex-Hex-DATDH	[45-73]	MS [203]	PC	Ser^{157} & Ser^{160}	-	First report of DATDH glycan
Nm	CS311#16	CS11	clonal variant[209]	Hex-Hex-DATDH	[45-73]	MS [203]	PG	Ser^{93}	-	
Nm	HTII125	NID280	DsiaB-D strain, *pilV*+	Hex-Hex-DATDH	Ser^{63}	MS [210]	PG	Ser^{94}	MS [210]	Residue mislabelled in article as Ser^{62}
-	-	-	-	DATDH	Ser^{63}	MS [210]	PC	Ser^{68}	MS [210]	-

Table 1 - Summary of all PTMs identified and characterised to date on PilE purified from *Neisseria* spp.

The *pgl* System, Glycan Structure & Phase Variation

When the DATDH glycan was first identifed on pili purified from *Nm*, discovery of glycosylation, and indeed PTM in general, was a novelty for bacteria. Since then several complete bacterial glycosylation systems have been described[211]. In *Neisseria* spp. the protein glycosylation (*pgl*) system is responsible for glycan synthesis and PTM. Hartley *et al.* have recently presented a detailed biochemical characterisation of this pathway including associated stereochemistry of the glycan intermediates (Figure 39)[212].

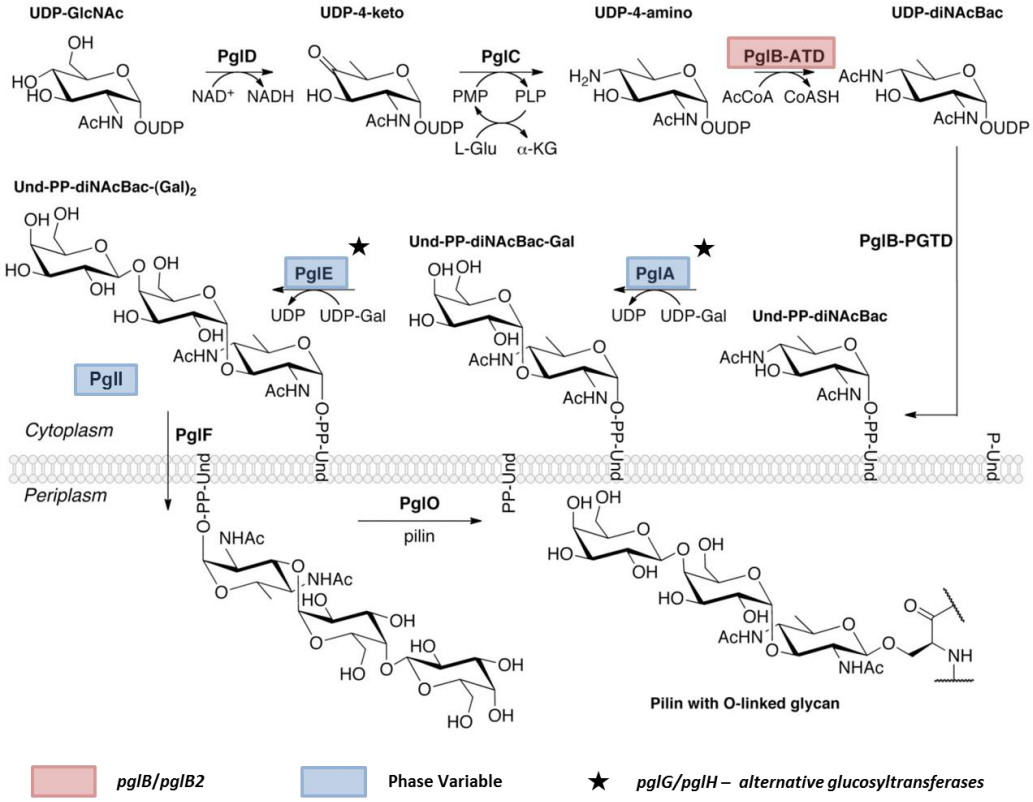


Figure 39 - *pgl* glycosylation system adapted from Hartley *et al.*[212] (Note that the homolog of PglO is PglL in Nm)

Synthesis of the glycan occurs in the cytoplasm where a succession of core enzymes (PglD, PglC, PglB) first acts on a core *O*-linked uridine phosphate (UDP) *N*-acetyl glucosamine (GlcNAc) sugar. The glycan core is constructed then transferred to a lipid anchored undecaprenyl phosphate (P-Und). The Und-PP linked sugar may then be further elaborated by PglE and PglA which attach successive galactose subunits. In some Nm strains PglH and PglI may fulfil this role with glucose as the substrate[213]. The glycan can be further *O*-acetylated, usually at only one site, by PglI[214]. Once completely formed, the sugar is translocated into the periplasm by the flippase PglF where

it is transferred en-bloc to its substrate by the oligosaccharyltransferase PglL (note that the homolog of PglO is PglL in Nm). Glycosylation of proteins therefore occurs in the periplasm.

Interestingly, there are several levels of variation present in the *pgl* pathway that result in different final glycan structures. The first occurs in the core *pgl* locus. Two alternative forms of the *pglB* gene, *pglB1* and *pglB2* code for proteins with different transferase domains: *pglB1* an acetyltransferase and *pglB2* a glycerotransferase. This results in the synthesis of two different core glycans, DATDH and GATDH respectively, that have alternate substituents at the 4 position (acetamido versus glyceramido) (Figure 38). *pglB2* has been shown to be present in approximately half of Nm isolates suggesting that half of clinically relevant strains express the GATDH glycan[204].

The second important level of variation concerns the other *pgl* genes, of which *pglA, pglE, pglG, pglH* and *pgI* are all found to be phase variable. This can result in the expression of up to six possible glycan structures for a given Nm strain[215]. Genotyping has been successfully used to predict the expressed glycan by examining the phase of the relevant genes.

The *pgl* system has recently been shown to be broad based in *Neisseria* spp. acting on substrates other than PilE[216-218]. The glycoproteins identified are all thought to be anchored or within the bacterial membranes and represent a wide spectrum of functions including a group involved in periplasmic electron transfer reactions (Ngo1769, AniA, CycB and CcoP) and another that form membrane fusion proteins in complex with efflux pumps (MtrC, MtrD, MacA). Glycosylation has also been found on a likely C-terminally truncated from of the PilQ outer membrane porin. These examples further underline the diversity that can be found in the bacterial proteome and highlight the fact that whilst overlooked for many years, bacteria too possess general systems for PTM.

4.2. Phosphoforms

In addition to glycosylation, PilE may also be modified by a variety of phosphoforms such as phosphate (P)[205], phosphoethanolamine (PE)[206, 207], phosphocholine (PC)[206, 207, 219] and phosphoglycerol (PG)[220] (Figure 40).

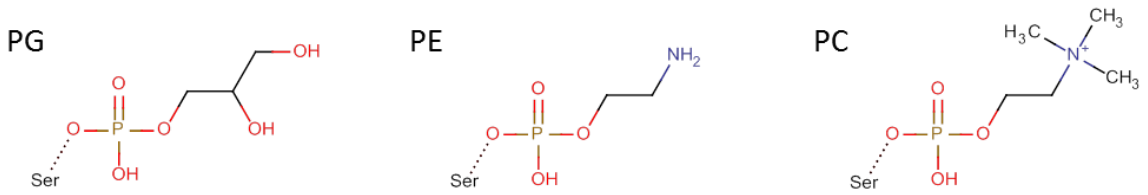


Figure 40 - Phosphoforms phosphoglycerol (PG), phosphoethanolamine (PE), phosphocholine (PC)

PilE is always found to be modified by at least one phosphoform and there is observable variation in the phosphoform structure. Phosphoform modification of pilins is found to be exclusively *O*-linked to serine residues and occurs at three distinct *loci*. The major modification site seems to be at $Ser^{68/69}$ with some pilin populations containing a minor component additionally modified at $Ser^{93/94}$. $Ser^{68/69}$ is located in the α-β loop and $Ser^{93/94}$ on the conserved β_2-β_3 loop feature connecting the second and third strand of the antiparallel, four strand beta sheet[168]. Both of these regions are exposed on the pilus surface. Similarly to glycosylation, the evidence to support these conclusions was obtained over a number of years through multiple studies. The relevant reports are also summarised in Table 1.

Enzymes Involved in Phosphoform Modification

In both Nm and Ng the enzyme PptA, coded by the phase variable *pptA* gene, is required for addition of PE and PC to PilE[221, 222]. Like glycosylation, phosphoform modification occurs in the periplasm. The preference for PE or PC seems to depend on a number of factors including expression of the minor pilin PilV, although the reasons for this are completely unknown[207]. PC and PE modification in Nm has recently been specifically attributed to a putative consensus sequence, XAS where X is a negatively charged amino acid[223]. The generality of this sequence has not yet been confirmed, however the PptA homologue in Ng does not seem to show this sequence specificity nor do the phosphoform modification sites reported in Nm HT1125. The mechanism by which PG is added to pilin was unknown at the beginning of this thesis.

5. Biological Role of PTM

5.1. Glycosylation

The surfaced exposed nature of the glycan, coupled with its high propensity for structural variation are thought to play roles in promoting antigenic variation. Convincing explanations for additional biological roles have remained elusive. Glycosylation is not required for pilus formation and Marceau *et al.* negated a role in pilus adhesiveness[208]. In a rather complex more recent report, Vik *et al*. study the effect of the glycan on a hexa-histidine tagged PilE mutant that is associated with a growth arrest phenotype[224]. They find that the growth arrest depends on the length of the glycan chain. Growth arrest is fully inhibited when the monosaccharide form is present (both DATDH and GATDH) but reduced in disaccharide constructs and negated when a trisaccharide is expressed. In conclusion they assert that the glycosylation status of PilE probably has some influence on pilin-subunit-subunit interactions and that is may be important for assembly and disassembly of the fibre in the periplasm. Whilst the study is interesting is validity in the wild type bacterium is unclear.

The glycan has also been linked to host cell receptor binding. An early study seemed to correlate putative pilin *N*-linked glycosylation status with bacterial adhesiveness, but since pilin glycosylation is now known to be *O*-linked the significance of this study is unclear[172]. Perhaps the point mutations made in the study induced adhesion related conformational changes in the pilus fibre? A more recent study in Ng suggested that that glycosylation was necessary for binding to the I domain of the CR3 receptor found on human cervical epithelial cells[225]. Since CR3 is a known signal transducer for phagocytic cells it was supposed that this may promote pathogen survival. Another report from the same group concerning Nm proposes that the glycan on the CS311#3 strain is involved in binding platelet activating factor[226].

Overall, other than its potential role in promoting immune escape through phase variation, the biological function of the glycan remains unclear.

5.2. Phosphoforms

The biological significance of the phosphoform modifications is also unclear and continues to be a subject of investigation. Similar to pilin glycosylation, phosphoform modification is always found in specific surface exposed regions of the globular domain. There is also some limited variation between phosphoforms and since they are surface accessible they have the potential to mediate interaction with other biomolecules and change the structure of the pilus surface. In a recent report from Jen *et al*. using the Nm CS311#3 strain, PC modification near the C-terminus has been shown to increase pilus-platelet activating factor binding affinity[226] however it must be noted that PC has not been previously reported in this region of PilE from any other strain.

Bibliography

[1] V. N. Uversky and A. L. Fink.Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochimica Et Biophysica Acta-Proteins and Proteomics*, **2004**, *1698*, 131.

[2] R. E. Thach, K. F. Dewey, J. C. Brown and P. Doty.Formylmethionine Codon AUG as an Initiator of Polypeptide Synthesis. *Science*, **1966**, *153*, 416.

[3] F. Zinoni, A. Birkmann, T. C. Stadtman and A. Böck.Nucleotide sequence and expression of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from Escherichia coli. *Proceedings of the National Academy of Sciences*, **1986**, *83*, 4650.

[4] I. Chambers, J. Frampton, P. Goldfarb, N. Affara, W. McBain and P. R. Harrison.The Structure Of The Mouse Glutathione-Peroxidase Gene - The Selenocysteine In The Active-Site Is Encoded By The Termination Codon. *Embo J.*, **1986**, *5*, 1221.

[5] J. Donovan and P. R. Copeland.The Efficiency of Selenocysteine Incorporation Is Regulated by Translation Initiation Factors. *Journal of Molecular Biology*, **2010**, *400*, 659.

[6] G. Srinivasan, C. M. James and J. A. Krzycki.Pyrrolysine Encoded by UAG in Archaea: Charging of a UAG-Decoding Specialized tRNA. *Science*, **2002**, *296*, 1459.

[7] B. Hao, W. Gong, T. K. Ferguson, C. M. James, J. A. Krzycki and M. K. Chan.A New UAG-Encoded Residue in the Structure of a Methanogen Methyltransferase. *Science*, **2002**, *296*, 1462.

[8] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. Manuel Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo and T. J. Hubbard.GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, **2012**, *22*, 1760.

[9] T. Nakamoto.Evolution and the universality of the mechanism of initiation of protein synthesis. *Gene*, **2009**, *432*, 1.

[10] S. Melnikov, A. Ben-Shem, N. G. de Loubresse, L. Jenner, G. Yusupova and M. Yusupov.One core, two shells: bacterial and eukaryotic ribosomes. *Nature Structural & Molecular Biology*, **2012**, *19*, 560.

[11] R. E. Moellering and B. F. Cravatt.Functional Lysine Modification by an Intrinsically Reactive Primary Glycolytic Metabolite. *Science*, **2013**, *341*, 549.

[12] L. M. Smith, N. L. Kelleher and P. Consortium Top Down.Proteoform: a single term describing protein complexity. *Nature Methods*, **2013**, *10*, 186.

[13] F. Sanger and H. Tuppy.The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *The Biochemical journal*, **1951**, *49*, 481.

[14] F. Sanger and H. Tuppy.The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *The Biochemical journal*, **1951**, *49*, 463.

[15] F. Sanger and E. O. P. Thompson.The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *The Biochemical journal*, **1953**, *53*, 366.

[16] F. Sanger and E. O. P. Thompson.The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *The Biochemical journal*, **1953**, *53*, 353.

[17] A. P. Ryle, F. Sanger, L. F. Smith and R. Kitai.The disulphide bonds of insulin. *The Biochemical journal*, **1955**, *60*, 541.

[18] T. H. Steinberg, in *Methods in Enzymology*, eds. R. B. Richard and P. D. Murray, Academic Press, 2009, vol. Volume 463, pp. 541.

[19] C. K. Damer, J. Partridge, W. R. Pearson and T. A. J. Haystead.Rapid identification of protein phosphatase 1-binding proteins by mixed peptide sequencing and data base searching - Characterization of a novel holoenzymic form of protein phosphatase 1. *Journal of Biological Chemistry*, **1998**, *273*, 24396.
[20] A. Shevchenko, M. Wilm, O. Vorm and M. Mann.Mass Spectrometric Sequencing of Proteins from Silver-Stained Polyacrylamide Gels. *Analytical Chemistry*, **1996**, *68*, 850.
[21] K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida and T. Matsuo.Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, **1988**, *2*, 151.
[22] M. Karas and F. Hillenkamp.Laser Desorption Ionization Of Proteins With Molecular Masses Exceeding 10000 Daltons. *Analytical Chemistry*, **1988**, *60*, 2299.
[23] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong and C. M. Whitehouse.Electrospray ionization for mass-spectrometry of large biomolecules. *Science*, **1989**, *246*, 64.
[24] K. Biemann, G. Gapp and J. Seibl.Application Of Mass Spectrometry To Structure Problems. I. Amino Acid Sequence In Peptides. *J. Am. Chem. Soc.*, **1959**, *81*, 2274.
[25] B. Spengler.Post-source decay analysis in matrix-assisted laser desorption/ionization mass spectrometry of biomolecules. *Journal of Mass Spectrometry*, **1997**, *32*, 1019.
[26] A. W. Purcell and J. J. Gorman.The use of post-source decay in matrix-assisted laser desorption/ionisation mass spectrometry to delineate T cell determinants. *Journal of Immunological Methods*, **2001**, *249*, 17.
[27] D. F. Hunt, J. R. Yates, J. Shabanowitz, S. Winston and C. R. Hauer.Protein Sequencing By Tandem Mass-Spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **1986**, *83*, 6233.
[28] P. Roepstorff and J. Fohlman.Letter to the editors. *Biological Mass Spectrometry*, **1984**, *11*, 601.
[29] K. Biemann.Contributions of mass-spectrometry to peptide and protein-structure. *Biomedical and Environmental Mass Spectrometry*, **1988**, *16*, 99.
[30] M. Mann and M. Wilm.Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, **1994**, *66*, 4390.
[31] D. C. Simpson and R. D. Smith.Combining capillary electrophoresis with mass spectrometry for applications in proteomics. *Electrophoresis*, **2005**, *26*, 1291.
[32] B. Manadas, V. M. Mendes, J. English and M. J. Dunn.Peptide fractionation in proteomics approaches. *Expert Review of Proteomics*, **2010**, *7*, 655.
[33] K. Gevaert, M. Goethals, L. Martens, J. Van Damme, A. Staes, G. R. Thomas and J. Vandekerckhove.Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nature Biotechnology*, **2003**, *21*, 566.
[34] A. J. Link, J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik and J. R. Yates.Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, **1999**, *17*, 676.
[35] S. S. Thakur, T. Geiger, B. Chatterjee, P. Bandilla, F. Froehlich, J. Cox and M. Mann.Deep and Highly Sensitive Proteome Coverage by LC-MS/MS Without Prefractionation. *Molecular & Cellular Proteomics*, **2011**, *10*.
[36] T. Koecher, P. Pichler, R. Swart and K. Mechtler.Analysis of protein mixtures from whole-cell extracts by single-run nanoLC-MS/MS using ultralong gradients. *Nature Protocols*, **2012**, *7*, 882.
[37] N. Nagaraj, N. A. Kulak, J. Cox, N. Neuhauser, K. Mayr, O. Hoerning, O. Vorm and M. Mann.System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap. *Molecular & Cellular Proteomics*, **2012**, *11*.
[38] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall and J. J. Coon.The One Hour Yeast Proteome. *Molecular & Cellular Proteomics*, **2013**.

[39] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb and R. Aebersold.Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, **1999**, *17*, 994.
[40] A. Thompson, J. Schafer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann and C. Hamon.Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, **2003**, *75*, 1895.
[41] P. L. Ross, Y. L. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson and D. J. Pappin.Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, **2004**, *3*, 1154.
[42] S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey and M. Mann.Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics*, **2002**, *1*, 376.
[43] L. Andersson and J. Porath.Isolation of phosphoproteins by immobilized metal (fe-3+) affinity-chromatography. *Analytical Biochemistry*, **1986**, *154*, 250.
[44] S. B. Ficarro, M. L. McCleland, P. T. Stukenberg, D. J. Burke, M. M. Ross, J. Shabanowitz, D. F. Hunt and F. M. White.Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nature Biotechnology*, **2002**, *20*, 301.
[45] K. A. Lee, L. P. Hammerle, P. S. Andrews, M. P. Stokes, T. Mustelin, J. C. Silva, R. A. Black and J. R. Doedens.Ubiquitin Ligase Substrate Identification through Quantitative Proteomics at Both the Protein and Peptide Levels. *Journal of Biological Chemistry*, **2011**, *286*, 41530.
[46] D. L. Swaney, C. D. Wenger, J. A. Thomson and J. J. Coon.Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **2009**, *106*, 995.
[47] A. Leitner and R. Aebersold.SnapShot: Mass Spectrometry for Protein and Proteome Analyses. *Cell*, **2013**, *154*.
[48] Y. Zhang, B. R. Fonslow, B. Shan, M.-C. Baek and J. R. Yates, III.Protein Analysis by Shotgun/Bottom-up Proteomics. *Chemical Reviews*, **2013**, *113*, 2343.
[49] A. I. Nesvizhskii and R. Aebersold.Interpretation of shotgun proteomic data - The protein inference problem. *Molecular & Cellular Proteomics*, **2005**, *4*, 1419.
[50] K.-Y. Chang, D. R. Georgianna, S. Heber, G. A. Payne and D. C. Muddiman.Detection of Alternative Splice Variants at the Proteome Level in Aspergillus flavus. *J. Proteome Res.*, **2010**, *9*, 1209.
[51] C. Ghigna, S. Giordano, H. H. Shen, F. Benvenuto, F. Castiglioni, P. M. Comoglio, M. R. Green, S. Riva and G. Biamonti.Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. *Molecular Cell*, **2005**, *20*, 881.
[52] P. Abraham, R. M. Adams, G. A. Tuskan and R. L. Hettich.Moving Away from the Reference Genome: Evaluating a Peptide Sequencing Tagging Approach for Single Amino Acid Polymorphism Identifications in the Genus Populus. *J. Proteome Res.*, **2013**, *12*, 3642.
[53] R. Aebersold and M. Mann.Mass spectrometry-based proteomics. *Nature*, **2003**, *422*, 198.
[54] B. Meyer, D. G. Papasotiriou and M. Karas.100% protein sequence coverage: a modern form of surrealism in proteomics. *Amino Acids*, **2011**, *41*, 291.
[55] B. T. Chait.Mass spectrometry: Bottom-up or top-down?, *Science*, **2006**, *314*, 65.
[56] F. Lanucara and C. E. Eyers.Top-down mass spectrometry for the analysis of combinatorial post-translational modifications. *Mass Spectrom. Rev.*, **2013**, *32*, 27.
[57] P. Jungblut and B. Thiede.Protein identification from 2-DE gels by MALDI mass spectrometry. *Mass Spectrom. Rev.*, **1997**, *16*, 145.
[58] D. Calligaris, C. Villard, L. Terras, D. Braguer, P. Verdier-Pinard and D. Lafitte.MALDI In-Source Decay of High Mass Protein lsoforms: Application to alpha- and beta-Tubulin Variants. *Analytical Chemistry*, **2010**, *82*, 6176.

[59] R. Ait-Belkacem, D. Calligaris, L. Sellami, C. Villard, S. Granjeaud, T. Schembri, C. Berenguer, L. H. Ouafik, D. Figarella-Branger, O. Chinot and D. Lafitte.Tubulin isoforms identified in the brain by MALDI in-source decay. *Journal of Proteomics*, **2013**, *79*, 172.
[60] E. D. Inutan and S. Trimpin.Matrix Assisted Ionization Vacuum (MAIV), a New Ionization Method for Biological Materials Analysis Using Mass Spectrometry. *Molecular & Cellular Proteomics*, **2013**, *12*, 792.
[61] M. Yamashita and J. B. Fenn.Electrospray ion-source - another variation on the free-jet theme. *Journal of Physical Chemistry*, **1984**, *88*, 4451.
[62] C. M. Whitehouse, R. N. Dreyer, M. Yamashita and J. B. Fenn.Electrospray interface for liquid chromatographs and mass spectrometers. *Analytical Chemistry*, **1985**, *57*, 675.
[63] M. S. Wilm and M. Mann.Electrospray And Taylor-Cone Theory, Doles Beam Of Macromolecules At Last. *International Journal of Mass Spectrometry*, **1994**, *136*, 167.
[64] M. Wilm and M. Mann.Analytical properties of the nanoelectrospray ion source. *Analytical Chemistry*, **1996**, *68*, 1.
[65] M. Wilm, A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer, T. Fotsis and M. Mann.Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature*, **1996**, *379*, 466.
[66] G. A. Valaskovic, N. L. Kelleher, D. P. Little, D. J. Aaserud and F. W. McLafferty.Attomole-Sensitivity Electrospray Source For Large-Molecule Mass-Spectrometry. *Analytical Chemistry*, **1995**, *67*, 3802.
[67] G. A. Valaskovic, N. L. Kelleher and F. W. McLafferty.Attomole protein characterization by capillary electrophoresis mass spectrometry. *Science*, **1996**, *273*, 1199.
[68] P. E. Andren, M. R. Emmett and R. M. Caprioli. Micro-Electrospray - Zeptomole-Attomole Per Microliter Sensitivity For Peptides. *Journal of the American Society for Mass Spectrometry*, **1994**, *5*, 867.
[69] H. Zhang, W. Cui, J. Wen, R. E. Blankenship and M. L. Gross.Native Electrospray and Electron-Capture Dissociation FTICR Mass Spectrometry for Top-Down Studies of Protein Assemblies. *Analytical Chemistry*, **2011**, *83*, 5598.
[70] M. T. Marty, H. Zhang, W. Cui, R. E. Blankenship, M. L. Gross and S. G. Sligar.Native Mass Spectrometry Characterization of Intact Nanodisc Lipoprotein Complexes. *Analytical Chemistry*, **2012**, *84*, 8957.
[71] J. Pan and C. H. Borchers.Top-down structural analysis of posttranslationally modified proteins by Fourier transform ion cyclotron resonance-MS with hydrogen/deuterium exchange and electron capture dissociation. *Proteomics*, **2013**, *13*, 974.
[72] T. G. Flick and E. R. Williams.Supercharging with Trivalent Metal Ions in Native Mass Spectrometry. *Journal of the American Society for Mass Spectrometry*, **2012**, *23*, 1885.
[73] C. A. Cassou, H. J. Sterling, A. C. Susa and E. R. Williams.Electrothermal Supercharging in Mass Spectrometry and Tandem Mass Spectrometry of Native Proteins. *Analytical Chemistry*, **2013**, *85*, 138.
[74] R. L. Edwards, P. Griffiths, J. Bunch and H. J. Cooper.Top-Down Proteomics and Direct Surface Sampling of Neonatal Dried Blood Spots: Diagnosis of Unknown Hemoglobin Variants. *Journal of the American Society for Mass Spectrometry*, **2012**, *23*, 1921.
[75] M. Mann, C. K. Meng and J. B. Fenn.Interpreting mass-spectra of multiply charged ions. *Analytical Chemistry*, **1989**, *61*, 1702.
[76] M. W. Senko, S. C. Beu and F. W. McLafferty.Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, **1995**, *6*, 229.
[77] X. Y. Chen, M. S. Westphall and L. M. Smith.Mass spectrometric analysis of DNA mixtures: Instrumental effects responsible for decreased sensitivity with increasing mass. *Analytical Chemistry*, **2003**, *75*, 5944.
[78] J. Park, H. Qin, M. Scalf, R. T. Hilger, M. S. Westphall, L. M. Smith and R. H. Blick.A Mechanical Nanomembrane Detector for Time-of-Flight Mass Spectrometry. *Nano Letters*, **2011**, *11*, 3681.

[79] R. A. Zubarev and A. Makarov.Orbitrap Mass Spectrometry. *Analytical Chemistry*, **2013**, *85*, 5288.
[80] Q. Z. Hu, R. J. Noll, H. Y. Li, A. Makarov, M. Hardman and R. G. Cooks.The Orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry*, **2005**, *40*, 430.
[81] A. G. Marshall and C. L. Hendrickson, in *Annual Review of Analytical Chemistry*, 2008, vol. 1, pp. 579.
[82] A. G. Marshall, C. L. Hendrickson and G. S. Jackson.Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrom. Rev.*, **1998**, *17*, 1.
[83] M. Scigelova, M. Hornshaw, A. Giannakopulos and A. Makarov.Fourier Transform Mass Spectrometry. *Molecular & Cellular Proteomics*, **2011**, *10*.
[84] E. Denisov, E. Damoc, O. Lange and A. Makarov.Orbitrap mass spectrometry with resolving powers above 1,000,000. *International Journal of Mass Spectrometry*, **2012**, *325*, 80.
[85] E. N. Nikolaev, I. A. Boldin, R. Jertz and G. Baykut.Initial Experimental Characterization of a New Ultra-High Resolution FTICR Cell with Dynamic Harmonization. *Journal of the American Society for Mass Spectrometry*, **2011**, *22*, 1125.
[86] E. N. Nikolaev, R. Jertz, A. Grigoryev and G. Baykut.Fine Structure in Isotopic Peak Distributions Measured Using a Dynamically Harmonized Fourier Transform Ion Cyclotron Resonance Cell at 7 T. *Analytical Chemistry*, **2012**, *84*, 2275.
[87] M. V. Gorshkov, L. Fornelli and Y. O. Tsybin.Observation of ion coalescence in Orbitrap Fourier transform mass spectrometry. *Rapid Communications in Mass Spectrometry*, **2012**, *26*, 1711.
[88] Y. Ge, I. N. Rybakova, Q. Xu and R. L. Moss.Top-down high-resolution mass spectrometry of cardiac myosin binding protein C revealed that truncation alters protein phosphorylation state. *Proceedings of the National Academy of Sciences of the United States of America*, **2009**, *106*, 12658.
[89] N. L. Kelleher, M. W. Senko, M. M. Siegel and F. W. McLafferty.Unit resolution mass spectra of 112 kDa molecules with 3 Da accuracy. *Journal of the American Society for Mass Spectrometry*, **1997**, *8*, 380.
[90] S. G. Valeja, N. K. Kaiser, F. Xian, C. L. Hendrickson, J. C. Rouse and A. G. Marshall.Unit Mass Baseline Resolution for an Intact 148 kDa Therapeutic Monoclonal Antibody by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Analytical Chemistry*, **2011**, *83*, 8391.
[91] R. D. Smith, J. A. Loo, C. J. Barinaga, C. G. Edmonds and H. R. Udseth.Collisional activation and collision-activated dissociation of large multiply charged polypeptides and proteins produced by electrospray ionization. *Journal of the American Society for Mass Spectrometry*, **1990**, *1*, 53.
[92] J. Loo, C. Edmonds and R. Smith.Primary sequence information from intact proteins by electrospray ionization tandem mass spectrometry. *Science*, **1990**, *248*, 201.
[93] M. Laitaoja, R. S. Sankhala, M. J. Swamy and J. Janis.Top-down mass spectrometry reveals new sequence variants of the major bovine seminal plasma protein PDC-109. *Journal of Mass Spectrometry*, **2012**, *47*, 853.
[94] R. Theberge, G. Infusini, W. Tong, M. E. McComb and C. E. Costello.Top-down analysis of small plasma proteins using an LTQ-Orbitrap. Potential for mass spectrometry-based clinical. assays for transthyretin and hemoglobin. *International Journal of Mass Spectrometry*, **2011**, *300*, 130.
[95] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz and D. F. Hunt.Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **2004**, *101*, 9528.
[96] R. A. Zubarev, N. L. Kelleher and F. W. McLafferty.Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.*, **1998**, *120*, 3265.

[97] F. Kjeldsen, K. F. Haselmann, B. A. Budnik, F. Jensen and R. A. Zubarev.Dissociative capture of hot (3-13 eV) electrons by polypeptide polycations: an efficient process accompanied by secondary fragmentation. *Chemical Physics Letters*, **2002**, *356*, 201.
[98] T. Baba, Y. Hashimoto, H. Hasegawa, A. Hirabayashi and I. Waki.Electron capture dissociation in a radio frequency ion trap. *Analytical Chemistry*, **2004**, *76*, 4263.
[99] O. A. Silivra, F. Kjeldsen, I. A. Ivonin and R. A. Zubarev.Electron capture dissociation of polypeptides in a three-dimensional quadrupole ion trap: Implementation and first results. *Journal of the American Society for Mass Spectrometry*, **2005**, *16*, 22.
[100] H. Satake, H. Hasegawa, A. Hirabayashi, Y. Hashimoto and T. Baba.Fast multiple electron capture dissociation in a linear radio frequency quadrupole ion trap. *Analytical Chemistry*, **2007**, *79*, 8755.
[101] J. B. Shaw, W. Li, D. D. Holden, Y. Zhang, J. Griep-Raming, R. T. Fellers, B. P. Early, P. M. Thomas, N. L. Kelleher and J. S. Brodbelt.Complete protein characterization using top-down mass spectrometry and ultraviolet photodissociation. *J. Am. Chem. Soc.*, **2013**, *135*, 12646.
[102] H. J. Cooper, K. Hakansson and A. G. Marshall.The role of electron capture dissociation in biomolecular analysis. *Mass Spectrom. Rev.*, **2005**, *24*, 201.
[103] J. Wiesner, T. Premsler and A. Sickmann.Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications. *Proteomics*, **2008**, *8*, 4466.
[104] Y. O. Tsybin, L. Fornelli, C. Stoermer, M. Luebeck, J. Parra, S. Nallet, F. M. Wurm and R. Hartmer.Structural Analysis of Intact Monoclonal Antibodies by Electron Transfer Dissociation Mass Spectrometry. *Analytical Chemistry*, **2011**, *83*, 8919.
[105] L. Fornelli, E. Damoc, P. M. Thomas, N. L. Kelleher, K. Aizikov, E. Denisov, A. Makarov and Y. O. Tsybin.Analysis of Intact Monoclonal Antibody IgG1 by Electron Transfer Dissociation Orbitrap FTMS. *Molecular & Cellular Proteomics*, **2012**, *11*, 1758.
[106] Y. Mao, S. G. Valeja, J. C. Rouse, C. L. Hendrickson and A. G. Marshall.Top-Down Structural Analysis of an Intact Monoclonal Antibody by Electron Capture Dissociation-Fourier Transform Ion Cyclotron Resonance-Mass Spectrometry. *Analytical Chemistry*, **2013**, *85*, 4239.
[107] Y. Peng, X. Chen, T. Sato, S. A. Rankin, R. F. Tsuji and Y. Ge.Purification and High-Resolution Top-Down Mass Spectrometric Characterization of Human Salivary alpha-Amylase. *Analytical Chemistry*, **2012**, *84*, 3339.
[108] C. Jia, L. Hui, W. Cao, C. B. Lietz, X. Jiang, R. Chen, A. D. Catherman, P. M. Thomas, Y. Ge, N. L. Kelleher and L. Li.High-definition De Novo Sequencing of Crustacean Hyperglycemic Hormone (CHH)-family Neuropeptides. *Molecular & Cellular Proteomics*, **2012**, *11*, 1951.
[109] X. Han, M. Jin, K. Breuker and F. W. McLafferty.Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science*, **2006**, *314*, 109.
[110] D. M. Horn, R. A. Zubarev and F. W. McLafferty.Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, **2000**, *11*, 320.
[111] N. M. Karabacak, L. Li, A. Tiwari, L. J. Hayward, P. Hong, M. L. Easterling and J. N. Agar.Sensitive and Specific Identification of Wild Type and Variant Proteins from 8 to 669 kDa Using Top-down Mass Spectrometry. *Molecular & Cellular Proteomics*, **2009**, *8*, 846.
[112] M. J. MacCoss, C. C. Wu and J. R. Yates.Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Analytical Chemistry*, **2002**, *74*, 5593.
[113] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Y. Yang, W. Y. Shi and S. H. Bryant.Open mass spectrometry search algorithm. *J. Proteome Res.*, **2004**, *3*, 958.
[114] L. Zamdborg, R. D. LeDuc, K. J. Glowacz, Y.-B. Kim, V. Viswanathan, I. T. Spaulding, B. P. Early, E. J. Bluhm, S. Babai and N. L. Kelleher.ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Research*, **2007**, *35*, W701.

[115] A. M. Frank, J. J. Pesavento, C. A. Mizzen, N. L. Kelleher and P. A. Pevzner.Interpreting top-down mass spectra using spectral alignment. *Analytical Chemistry*, **2008**, *80*, 2499.
[116] X. W. Liu, Y. Sirotkin, Y. F. Shen, G. Anderson, Y. S. Tsai, Y. S. Ting, D. R. Goodlett, R. D. Smith, V. Bafna and P. A. Pevzner.Protein Identification Using Top-Down. *Molecular & Cellular Proteomics*, **2012**, *11*.
[117] J. P. Savaryn, A. D. Catherman, P. M. Thomas, M. M. Abecassis and N. L. Kelleher.The emergence of top-down proteomics in clinical research. *Genome Medicine*, **2013**, *5*.
[118] J. Zhang, M. J. Guy, H. S. Norman, Y.-C. Chen, Q. Xu, X. Dong, H. Guner, S. Wang, T. Kohmoto, K. H. Young, R. L. Moss and Y. Ge.Top-Down Quantitative Proteomics Identified Phosphorylation of Cardiac Troponin I as a Candidate Biomarker for Chronic Heart Failure. *J. Proteome Res.*, **2011**, *10*, 4054.
[119] J. J. Pesavento, Y. B. Kim, G. K. Taylor and N. L. Kelleher.Shotgun annotation of histone modifications: A new approach for streamlined characterization of proteins by top down mass spectrometry. *J. Am. Chem. Soc.*, **2004**, *126*, 3386.
[120] M. Rolando, S. Sanulli, C. Rusniok, L. Gomez-Valero, C. Bertholet, T. Sahr, R. Margueron and C. Buchrieser.Legionella pneumophila effector RomA uniquely modifies host chromatin to repress gene expression and promote intracellular bacterial replication. *Cell host & microbe*, **2013**, *13*, 395.
[121] N. Siuti, M. J. Roth, C. A. Mizzen, N. L. Kelleher and J. J. Pesavento.Gene-specific characterization of human histone H2B by electron capture dissociation. *J. Proteome Res.*, **2006**, *5*, 233.
[122] M. T. Boyne, J. J. Pesavento, C. A. Mizzen and N. L. Kelleher.Precise characterization of human histories in the H2A gene family by top down mass spectrometry. *J. Proteome Res.*, **2006**, *5*, 248.
[123] L. Jiang, J. N. Smith, S. L. Anderson, P. Ma, C. A. Mizzen and N. L. Kelleher.Global assessment of combinatorial post-translational modification of core histones in yeast using contemporary mass spectrometry. *Journal of Biological Chemistry*, **2007**, *282*, 27923.
[124] Z. Tian, N. Tolic, R. Zhao, R. J. Moore, S. M. Hengel, E. W. Robinson, D. L. Stenoien, S. Wu, R. D. Smith and L. Pasa-Tolic.Enhanced top-down characterization of histone post-translational modifications. *Genome Biol*, **2012**, *13*.
[125] A. A. Doucette, J. C. Tran, M. J. Wall and S. Fitzsimmons.Intact proteome fractionation strategies compatible with mass spectrometry. *Expert Review of Proteomics*, **2011**, *8*, 787.
[126] J. C. Tran and A. A. Doucette.Gel-eluted liquid fraction entrapment electrophoresis: An electrophoretic method for broad molecular weight range proteome separation. *Analytical Chemistry*, **2008**, *80*, 1568.
[127] J. C. Tran and A. A. Doucette.Rapid and effective focusing in a carrier ampholyte solution lsoelectric focusing system: A Proteome prefractionation tool. *J. Proteome Res.*, **2008**, *7*, 1761.
[128] L. Sun, M. D. Knierman, G. Zhu and N. J. Dovichi.Fast Top-Down Intact Protein Characterization with Capillary Zone Electrophoresis-Electrospray Ionization Tandem Mass Spectrometry. *Analytical Chemistry*, **2013**, *85*, 5989.
[129] J. C. Tran and A. A. Doucette.Multiplexed Size Separation of Intact Proteins in Solution Phase for Mass Spectrometry. *Analytical Chemistry*, **2009**, *81*, 6201.
[130] J. J. Pesavento, C. A. Mizzen and N. L. Kelleher.Quantitative analysis of modified proteins and their positional isomers by tandem mass spectrometry: Human histone H4. *Analytical Chemistry*, **2006**, *78*, 4271.
[131] B. A. Garcia, J. J. Pesavento, C. A. Mizzen and N. L. Kelleher.Pervasive combinatorial modification of histone H3 in human cells. *Nature Methods*, **2007**, *4*, 487.
[132] Y. Shen, N. Tolic, K. K. Hixson, S. O. Purvine, G. A. Anderson and R. D. Smith.De novo sequencing of unique sequence tags for discovery of post-translational modifications of proteins. *Analytical Chemistry*, **2008**, *80*, 7742.

[133] C. Kelleher, S. Friel, G. Nolan and B. Forbes.Effect of social variation on the Irish diet. *Proceedings of the Nutrition Society*, **2002**, *61*, 527.
[134] Y. S. Tsai, A. Scherl, J. L. Shaw, C. L. MacKay, S. A. Shaffer, P. R. R. Langridge-Smith and D. R. Goodlett.Precursor Ion Independent Algorithm for Top-Down Shotgun Proteomics. *Journal of the American Society for Mass Spectrometry*, **2009**, *20*, 2154.
[135] S. Wu, R. N. Brown, S. H. Payne, D. Meng, R. Zhao, N. Tolic, L. Cao, A. Shukla, M. E. Monroe, R. J. Moore, M. S. Lipton and L. Pasa-Tolic.Top-Down Characterization of the Post-Translationally Modified Intact Periplasmic Proteome from the Bacterium Novosphingobium aromaticivorans. *International journal of proteomics*, **2013**, *2013*, 279590.
[136] M. K. Bunger, B. J. Cargile, A. Ngunjiri, J. L. Bundy and J. L. Stephenson.Automated proteomics of E-coli via top-down electron-transfer dissociation mass spectrometry. *Analytical Chemistry*, **2008**, *80*, 1459.
[137] C. Ansong, S. Wu, D. Meng, X. Liu, H. M. Brewer, B. L. Deatherage Kaiser, E. S. Nakayasu, J. R. Cort, P. Pevzner, R. D. Smith, F. Heffron, J. N. Adkins and L. Pasa-Tolic.Top-down proteomics reveals a unique protein S-thiolation switch in Salmonella Typhimurium in response to infection-like conditions. *Proceedings of the National Academy of Sciences of the United States of America*, **2013**, *110*, 10153.
[138] F. Y. Meng, Y. Du, L. M. Miller, S. M. Patrie, D. E. Robinson and N. L. Kelleher.Molecular-level description of proteins from Saccharomyces cerevisiae using quadrupole FT hybrid mass spectrometry for top down proteomics. *Analytical Chemistry*, **2004**, *76*, 2852.
[139] J. F. Kellie, A. D. Catherman, K. R. Durbin, J. C. Tran, J. D. Tipton, J. L. Norris, C. E. Witkowski, II, P. M. Thomas and N. L. Kelleher.Robust Analysis of the Yeast Proteome under 50 kDa by Molecular-Mass-Based Fractionation and Top-Down Mass Spectrometry. *Analytical Chemistry*, **2012**, *84*, 209.
[140] D. R. Ahlf, P. D. Compton, J. C. Tran, B. P. Early, P. M. Thomas and N. L. Kelleher.Evaluation of the Compact High-Field Orbitrap for Top-Down Proteomics of Human Cells. *J. Proteome Res.*, **2012**, *11*, 4308.
[141] J. C. Tran, L. Zamdborg, D. R. Ahlf, J. E. Lee, A. D. Catherman, K. R. Durbin, J. D. Tipton, A. Vellaichamy, J. F. Kellie, M. Li, C. Wu, S. M. M. Sweet, B. P. Early, N. Siuti, R. D. LeDuc, P. D. Compton, P. M. Thomas and N. L. Kelleher.Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature*, **2011**, *480*, 254.
[142] A. D. Catherman, M. Li, J. C. Tran, K. R. Durbin, P. D. Compton, B. P. Early, P. M. Thomas and N. L. Kelleher.Top Down Proteomics of Human Membrane Proteins from Enriched Mitochondrial Fractions. *Analytical Chemistry*, **2013**, *85*, 1880.
[143] A. D. Catherman, K. R. Durbin, D. R. Ahlf, B. P. Early, R. T. Fellers, J. C. Tran, P. M. Thomas and N. L. Kelleher.Large-scale top down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. *Molecular & Cellular Proteomics*, **2013**.
[144] T. Cabras, E. Pisano, C. Montaldo, M. R. Giuca, F. Iavarone, G. Zampino, M. Castagnola and I. Messana.Significant Modifications of the Salivary Proteome Potentially Associated with Complications of Down Syndrome Revealed by Top-down Proteomics. *Molecular & Cellular Proteomics*, **2013**, *12*, 1844.
[145] O. B. Harrison, H. Claus, Y. Jiang, J. S. Bennett, H. B. Bratcher, K. A. Jolley, C. Corton, R. Care, J. T. Poolman, W. D. Zollinger, C. E. Frasch, D. S. Stephens, I. Feavers, M. Frosch, J. Parkhill, U. Vogel, M. A. Quail, S. D. Bentley and M. C. J. Maiden.Description and nomenclature of Neisseria meningitidis capsule locus. *Emerg Infect Dis*, **2013**, *19*, 566.
[146] M. C. J. Maiden, in *Annual Review of Microbiology*, 2006, vol. 60, pp. 561.
[147] M. Pérez-Losada, P. Cabezas, E. Castro-Nallar and K. A. Crandall.Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infection, Genetics and Evolution*, **2013**, *16*, 38.
[148] D. A. Caugant, G. Tzanakaki and P. Kriz.Lessons from meningococcal carriage studies. *Fems Microbiol Rev*, **2007**, *31*, 52.

[149] A. Anonychuk, G. Woo, A. Vyse, N. Demarteau and A. C. Tricco.The Cost and Public Health Burden of Invasive Meningococcal Disease Outbreaks: A Systematic Review. *Pharmacoeconomics*, **2013**, *31*, 563.
[150] G. Miller.Death of California Researcher Spurs Investigation. *Science*, **2012**, *336*, 659.
[151] A. Sharip, F. Sorvillo, M. D. Redelings, L. Mascola, M. Wise and D. M. Nguyen.Population-based analysis of meningococcal disease mortality in the United States - 1990-2002. *Pediatric Infectious Disease Journal*, **2006**, *25*, 191.
[152] W. H. Organization.Meningococcal disease in countries of the African meningitis belt, 2012 – emerging needs and future perspectives. *Weekly Epidemiological Record*, **2013**, *88*, 129.
[153] J. Finne, M. Leinonen and P. H. Makela.Antigenic similarities between brain components and bacteria causing meningitis - implications for vaccine development and pathogenesis. *Lancet*, **1983**, *2*, 355.
[154] J. B. Robbins, R. Schneerson, G. Xie, L. Ake-Hanson and M. A. Miller.Capsular polysaccharide vaccine for Group B Neisseria meningitidis, Escherichia coli K1, and Pasteurella haemolytica A2. *Proceedings of the National Academy of Sciences of the United States of America*, **2011**, *108*, 17871.
[155] L. K. K. Tan, G. M. Carlone and R. Borrow.Advances in the Development of Vaccines against Neisseria meningitidis. *N. Engl. J. Med.*, **2010**, *362*, 1511.
[156] M. Virji.Pathogenic neisseriae: surface modulation, pathogenesis and infection control. *Nature Reviews Microbiology*, **2009**, *7*, 274.
[157] X. Bai and R. Borrow.Genetic shifts of Neisseria meningitidis serogroup B antigens and the quest for a broadly cross-protective vaccine. *Expert Review of Vaccines*, **2010**, *9*, 1203.
[158] B. Capecchi, J. Adu-Bobie, F. Di Marcello, L. Ciucchi, V. Masignani, A. Taddei, R. Rappuoli, M. Pizza and B. Arico.Neisseria meningitidis NadA is a new invasin which promotes bacterial adhesion to and penetration into human epithelial cells. *Molecular Microbiology*, **2005**, *55*, 687.
[159] T. Davidsen and T. Tonjum.Meningococcal genome dynamics. *Nature Reviews Microbiology*, **2006**, *4*, 11.
[160] B. Hallet.Playing Dr Jekyll and Mr Hyde: combined mechanisms of phase variation in bacteria. *Current Opinion in Microbiology*, **2001**, *4*, 570.
[161] Y. N. Srikhanta, K. L. Fox and M. P. Jennings.The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. *Nature Reviews Microbiology*, **2010**, *8*, 196.
[162] S. D. Bentley, G. S. Vernikos, L. A. S. Snyder, C. Churcher, C. Arrowsmith, T. Chillingworth, A. Cronin, P. H. Davis, N. E. Holroyd, K. Jagels, M. Maddison, S. Moule, E. Rabbinowitsch, S. Sharp, L. Unwin, S. Whitehead, M. A. Quail, M. Achtman, B. Barrell, N. J. Saunders and J. Parkhill.Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. *Plos Genetics*, **2007**, *3*, 230.
[163] S. Hammerschmidt, A. Muller, H. Sillmann, M. Muhlenhoff, R. Borrow, A. Fox, J. vanPutten, W. D. Zollinger, R. GerardySchahn and M. Frosch.Capsule phase variation in Neisseria meningitidis serogroup B by slipped-strand mispairing in the polysialyltransferase gene (siaD): Correlation with bacterial invasion and the outbreak of meningococcal disease. *Molecular Microbiology*, **1996**, *20*, 1211.
[164] K. A. Kline, A. K. Criss, A. Wallace and H. S. Seifert.Transposon mutagenesis identifies sites upstream of the Neisseria gonorrhoeae pilE gene that modulate pilin antigenic variation. *J Bacteriol*, **2007**, *189*, 3462.
[165] K. Trivedi, C. M. Tang and R. M. Exley.Mechanisms of meningococcal colonisation. *Trends Microbiol*, **2011**, *19*, 456.
[166] K. Melican and G. Dumenil.Vascular colonization by Neisseria meningitidis. *Current Opinion in Microbiology*, **2012**, *15*, 50.
[167] D. J. Hill, N. J. Griffiths, E. Borodina and M. Virji.Cellular and molecular biology of Neisseria meningitidis colonization and invasive disease. *Clinical Science*, **2010**, *118*, 547.

[168] L. Craig, M. E. Pique and J. A. Tainer.Type IV pilus structure and bacterial pathogenicity. *Nature Reviews Microbiology*, **2004**, *2*, 363.
[169] T. Proft and E. N. Baker.Pili in Gram-negative and Gram-positive bacteria - structure, assembly and their role in disease. *Cell. Mol. Life Sci.*, **2009**, *66*, 613.
[170] C. L. Giltner, Y. Nguyen and L. L. Burrows.Type IV Pilin Proteins: Versatile Molecular Modules. *Microbiology and Molecular Biology Reviews*, **2012**, *76*, 740.
[171] M. Virji, H. Kayhty, D. J. P. Ferguson, C. Alexandrescu, J. E. Heckels and E. R. Moxon.The role of pili in the interactions of pathogenic Neisseria with cultured human endothelial cells. *Molecular Microbiology*, **1991**, *5*, 1831.
[172] M. Virji, J. R. Saunders, G. Sims, K. Makepeace, D. Maskell and D. J. P. Ferguson. Pilus facilitated adherence of Neisseria meningitidis to human epithelial and endothelial cells modulation of adherence phenotype occurs concurrently with changes in primary amino-acid sequence and the glycosylation status of pilin. *Molecular Microbiology*, **1993**, *10*, 1013.
[173] X. Nassif, J. L. Beretti, J. Lowy, P. Stenberg, P. O'Gaora, J. Pfeifer, S. Normark and M. So.Roles of pilin and PilC in adhesion of Neisseria meningitidis to human epithelial and endothelial cells. *Proceedings of the National Academy of Sciences*, **1994**, *91*, 3769.
[174] M. Marceau, J.-L. Beretti and X. Nassif.High adhesiveness of encapsulated Neisseria meningitidis to epithelial cells is associated with the formation of bundles of pili. *Molecular Microbiology*, **1995**, *17*, 855.
[175] E. Mairey, A. Genovesio, E. Donnadieu, C. Bernard, F. Jaubert, E. Pinard, J. Seylaz, J. C. Olivo-Marin, X. Nassif and G. Dumenil.Cerebral microcirculation shear stress levels determine Neisseria meningitidis attachment sites along the blood-brain barrier. *Journal of Experimental Medicine*, **2006**, *203*, 1939.
[176] G. Mikaty, M. Soyer, E. Mairey, N. Henry, D. Dyer, K. T. Forest, P. Morand, S. Guadagnini, M. C. Prévost, X. Nassif and G. Duménil.Extracellular Bacterial Pathogen Induces Host Cell Surface Reorganization to Resist Shear Stress. *PLoS Pathog*, **2009**, *5*, e1000314.
[177] M. Coureuil, G. Mikaty, F. Miller, H. Lecuyer, C. Bernard, S. Bourdoulous, G. Dumenil, R. M. Mege, B. B. Weksler, I. A. Romero, P. O. Couraud and X. Nassif.Meningococcal Type IV Pili Recruit the Polarity Complex to Cross the Brain Endothelium. *Science*, **2009**, *325*, 83.
[178] H. Lecuyer, X. Nassif and M. Coureuil.Two Strikingly Different Signaling Pathways Are Induced by Meningococcal Type IV Pili on Endothelial and Epithelial Cells. *Infect Immun*, **2012**, *80*, 175.
[179] L. Craig, N. Volkmann, A. S. Arvai, M. E. Pique, M. Yeager, Edward H. Egelman and J. A. Tainer.Type IV Pilus Structure by Cryo-Electron Microscopy and Crystallography: Implications for Pilus Assembly and Functions. *Molecular Cell*, **2006**, *23*, 651.
[180] M. Wolfgang, P. Lauer, H. S. Park, L. Brossay, J. Hebert and M. Koomey.PilT mutations lead to simultaneous defects in competence for natural transformation and twitching motility in piliated Neisseria gonorrhoeae. *Molecular Microbiology*, **1998**, *29*, 321.
[181] A. J. Merz, M. So and M. P. Sheetz.Pilus retraction powers bacterial twitching motility. *Nature*, **2000**, *407*, 98.
[182] E. Carbonnelle, S. Helaine, L. Prouvensier, X. Nassif and V. Pelicic.Type IV pilus biogenesis in Neisseria meningitidis: PilW is involved in a step occurring after pilus assembly, essential for fibre stability and function. *Molecular Microbiology*, **2005**, *55*, 54.
[183] E. Carbonnelle, S. Helaine, X. Nassif and V. Pelicic.A systematic genetic analysis in Neisseria meningitidis defines the Pil proteins required for assembly, functionality, stabilization and export of type IV pili. *Molecular Microbiology*, **2006**, *61*, 1510.
[184] M. S. Strom, P. Bergman and S. Lory.Identification of Active-Site Cysteines in the Conserved Domain of PilD, the Bifunctional Type-IV Pilin Leader Peptidase N-Methyltransferase of Pseudomonas-Aeruginosa. *Journal of Biological Chemistry*, **1993**, *268*, 15788.
[185] M. S. Strom, D. N. Nunn and S. Lory.A Single Bifunctional Enzyme, PilD, Catalyzes Cleavage and N-Methylation of Proteins Belonging to the Type-IV Pilin Family.

Proceedings of the National Academy of Sciences of the United States of America, **1993**, *90*, 2404.
[186] N. E. Freitag, H. S. Seifert and M. Koomey.Characterization of the PilF-PilD pilus assembly locus of Neisseria gonorrhoeae. *Molecular Microbiology*, **1995**, *16*, 575.
[187] J.-L. Berry, M. M. Phelan, R. F. Collins, T. Adomavicius, T. Tønjum, S. A. Frye, L. Bird, R. Owens, R. C. Ford, L.-Y. Lian and J. P. Derrick.Structure and Assembly of a Trans-Periplasmic Channel for Type IV Pili in <italic>Neisseria meningitidis</italic>. *PLoS Pathog*, **2012**, *8*, e1002923.
[188] V. Pelicic.Type IV pili: e pluribus unum?, *Molecular Microbiology*, **2008**, *68*, 827.
[189] M. Georgiadou, M. Castagnini, G. Karimova, D. Ladant and V. Pelicic.Large-scale study of the interactions between proteins involved in type IV pilus biology in Neisseria meningitidis: characterization of a subcomplex involved in pilus assembly. *Molecular Microbiology*, **2012**, *84*, 857.
[190] R. F. Collins, M. Saleem and J. P. Derrick.Purification and three-dimensional electron Microscopy structure of the Neisseria meningitidis type IV pilus biogenesis protein PilG. *J Bacteriol*, **2007**, *189*, 6389.
[191] P. C. Morand, M. Drab, K. Rajalingam, X. Nassif and T. F. Meyer.Neisseria meningitidis differentially controls host cell motility through PilC1 and PilC2 components of type IV Pili. *PLoS One*, **2009**, *4*, e6834.
[192] D. Brown, S. Helaine, E. Carbonnelle and V. Pelicic.A systematic functional analysis reveals that a set of 7 genes is involved in fine tuning of the multiple functions mediated by type IV pili in Neisseria meningitidis. *Infect. Immun.*, **2010**, IAI.00099.
[193] A. Cehovin, P. J. Simpson, M. A. McDowell, D. R. Brown, R. Noschese, M. Pallett, J. Brady, G. S. Baldwin, S. M. Lea, S. J. Matthews and V. Pelicic.Specific DNA recognition mediated by a type IV pilin. *Proceedings of the National Academy of Sciences of the United States of America*, **2013**, *110*, 3065.
[194] H. C. Winther-Larsen, F. T. Hegge, M. Wolfgang, S. F. Hayes, J. P. M. van Putten and M. Koomey.Neisseria gonorrhoeae PilV, a type IV pilus-associated protein essential to human epithelial cell adherence. *Proceedings of the National Academy of Sciences of the United States of America*, **2001**, *98*, 15276.
[195] S. Helaine, D. H. Dyer, X. Nassif, V. Pelicic and K. T. Forest.3D structure/function analysis of PilX reveals how minor pilins can modulate the virulence properties of type IV pili. *Proceedings of the National Academy of Sciences of the United States of America*, **2007**, *104*, 15888.
[196] H. E. Parge, K. T. Forest, M. J. Hickey, D. A. Christensen, E. D. Getzoff and J. A. Tainer.Structure of the Fiber Forming Protein Pilin at 2.6-Angstrom Resolution. *Nature*, **1995**, *378*, 32.
[197] L. A. Cahoon, K. A. Manthei, E. Rotman, J. L. Keck and H. S. Seifert.Neisseria gonorrhoeae RecQ Helicase HRDC Domains Are Essential for Efficient Binding and Unwinding of the pilE Guanine Quartet Structure Required for Pilin Antigenic Variation. *J Bacteriol*, **2013**, *195*, 2255.
[198] L. A. Cahoon and H. S. Seifert.Transcription of a cis-acting, Noncoding, Small RNA Is Required for Pilin Antigenic Variation in Neisseria gonorrhoeae. *PLoS pathogens*, **2013**, *9*.
[199] L. A. Cahoon and H. S. Seifert.Focusing homologous recombination: pilin antigenic variation in the pathogenic Neisseria. *Molecular Microbiology*, **2011**, *81*, 1136.
[200] C. Vink, G. Rudenko and H. S. Seifert.Microbial antigenic variation mediated by homologous DNA recombination. *Fems Microbiol Rev*, **2012**, *36*, 917.
[201] R. A. Helm and H. S. Seifert.Frequency and Rate of Pilin Antigenic Variation of Neisseria meningitidis. *J Bacteriol*, **2010**, *192*, 3822.
[202] A. Cehovin, M. Winterbotham, J. Lucidarme, R. Borrow, C. M. Tang, R. M. Exley and V. Pelicic.Sequence conservation of pilus subunits in Neisseria meningitidis. *Vaccine*, **2010**, *28*, 4817.

[203] E. Stimson, 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.Meningococcal Pilin - A Glycoprotein Substituted With Digalactosyl 2,4-Diacetamido-2,4,6-Trideoxyhexose. *Molecular Microbiology*, **1995**, *17*, 1201.
[204] J. Chamot-Rooke, B. Rousseau, F. Lanternier, G. Mikaty, E. Mairey, C. Malosse, G. Bouchoux, V. Pelicic, L. Camoin, X. Nassif and G. Dumenil.Alternative Neisseria spp. type IV pilin glycosylation with a glyceramido acetamido trideoxyhexose residue. *Proceedings of the National Academy of Sciences of the United States of America*, **2007**, *104*, 14783.
[205] K. T. Forest, S. A. Dunham, M. Koomey and J. A. Tainer.Crystallographic structure reveals phosphorylated pilin from Neisseria: phosphoserine sites modify type IV pilus surface chemistry and fibre morphology. *Molecular Microbiology*, **1999**, *31*, 743.
[206] F. T. Hegge, P. G. Hitchen, F. E. Aas, H. Kristiansen, C. Lovold, W. Egge-Jacobsen, M. Panico, W. Y. Leong, V. Bull, M. Virji, H. R. Morris, A. Dell and M. Koomey.Unique modifications with phosphocholine and phosphoethanolamine define alternate antigenic forms of Neisseria gonorrhoeae type IV pili. *Proceedings of the National Academy of Sciences of the United States of America*, **2004**, *101*, 10798.
[207] F. E. Aas, W. Egge-Jacobsen, H. C. Winther-Larsen, C. Lovold, P. G. Hitchen, A. Dell and M. Koomey.Neisseria gonorrhoeae type IV pili undergo multisite, hierarchical modifications with phosphoethanolamine and phosphocholine requiring an enzyme structurally related to lipopolysaccharide phosphoethanolamine transferases. *Journal of Biological Chemistry*, **2006**, *281*, 27712.
[208] M. Marceau, K. Forest, J.-L. Béretti, J. Tainer and X. Nassif.Consequences of the loss of O-linked glycosylation of meningococcal type IV pilin on piliation and pilus-mediated adhesion. *Molecular Microbiology*, **1998**, *27*, 705.
[209] M. Virji, J. R. Saunders, G. Sims, K. Makepeace, D. Maskell and D. J. P. Ferguson.Pilus facilitated adherence of Neisseria meningitidis to human epithelial and endothelial cells modulation of adherence phenotype occurs concurrently with changes in primary amino acid sequence and the glycosylation status of pilin. *Molecular Microbiology*, **1993**, *10*, 1013.
[210] H. Takahashi, T. Yanagisawa, K. S. Kim, S. Yokoyama and M. Ohnishi.Meningococcal PilV Potentiates Neisseria meningitidis Type IV Pilus-Mediated Internalization into Human Endothelial and Epithelial Cells. *Infect Immun*, **2012**, *80*, 4154.
[211] M. Abu-Qarn, J. Eichler and N. Sharon.Not just for Eukarya anymore: protein glycosylation in Bacteria and Archaea. *Current Opinion in Structural Biology*, **2008**, *18*, 544.
[212] M. D. Hartley, M. J. Morrison, F. E. Aas, B. Borud, M. Koomey and B. Imperiali.Biochemical Characterization of the O-Linked Glycosylation Pathway in Neisseria gonorrhoeae Responsible for Biosynthesis of Protein Glycans Containing N,N '-Diacetylbacillosamine. *Biochemistry*, **2011**, *50*, 4936.
[213] B. Borud, R. Viburiene, M. D. Hartley, B. S. Paulsen, W. Egge-Jacobsen, B. Imperiali and M. Koomey.Genetic and molecular analyses reveal an evolutionary trajectory for glycan synthesis in a bacterial protein glycosylation system. *Proceedings of the National Academy of Sciences of the United States of America*, **2011**, *108*, 9643.
[214] M. J. Warren, L. F. Roddam, P. M. Power, T. D. Terry and M. P. Jennings.Analysis of the role of pglI in pilin glycosylation of Neisseria meningitidis. *FEMS Immunol. Med. Microbiol.*, **2004**, *41*, 43.
[215] P. M. Power, L. F. Roddam, K. Rutter, S. Z. Fitzpatrick, Y. N. Srikhanta and M. P. Jennings.Genetic characterization of pilin glycosylation and phase variation in Neisseria meningitidis. *Molecular Microbiology*, **2003**, *49*, 833.
[216] J. H. Anonsen, A. Vik, W. Egge-Jacobsen and M. Koomey.An Extended Spectrum of Target Proteins and Modification Sites in the General O-Linked Protein Glycosylation System in Neisseria gonorrhoeae. *J. Proteome Res.*, **2012**, *11*, 5781.
[217] A. Vik, F. E. Aas, J. H. Anonsen, S. Bilsborough, A. Schneider, W. Egge-Jacobsen and M. Koomey.Broad spectrum O-linked protein glycosylation in the human pathogen

Neisseria gonorrhoeae. *Proceedings of the National Academy of Sciences of the United States of America*, **2009**, *106*, 4447.
[218] S. C. Ku, B. L. Schulz, P. M. Power and M. P. Jennings.The pilin O-glycosylation pathway of pathogenic Neisseria is a general system that glycosylates AniA, an outer membrane nitrite reductase. *Biochem Bioph Res Co*, **2009**, *378*, 84.
[219] J. N. Weiser, J. B. Goldberg, N. Pan, L. Wilson and M. Virji.The phosphorylcholine epitope undergoes phase variation on a 43-kilodalton protein in Pseudomonas aeruginosa and on pili of Neisseria meningitidis and Neisseria gonorrhoeae. *Infect Immun*, **1998**, *66*, 4263.
[220] E. Stimson, M. Virji, S. Barker, M. Panico, I. Blench, J. Saunders, G. Payne, E. R. Moxon, A. Dell and H. R. Morris.Discovery of a novel protein modification: alpha-glycerophosphate is a substituent of meningococcal pilin. *Biochem. J.*, **1996**, *316*, 29.
[221] M. J. Warren and M. P. Jennings.Identification and characterization of pptA: a gene involved in the phase-variable expression of phosphorylcholine on pili of Neisseria meningitidis. *Infect Immun*, **2003**, *71*, 6892.
[222] C. L. Naessan, W. Egge-Jacobsen, R. W. Heiniger, M. C. Wolfgang, F. E. Aas, A. Rohr, H. C. Winther-Larsen and M. Koomey.Genetic and functional analyses of PptA, a phospho-form transferase targeting type IV pili in Neisseria gonorrhoeae. *J Bacteriol*, **2008**, *190*, 387.
[223] F. E. C. Jen, C. E. Jones, J. C. Wilson, B. L. Schulz and M. P. Jennings.Substrate recognition of a structure motif for phosphorylcholine post-translational modification in Neisseria meningitidis. *Biochem Bioph Res Co*, **2013**, *431*, 808.
[224] A. Vik, M. Aspholm, J. H. Anonsen, B. Borud, N. Roos and M. Koomey.Insights into type IV pilus biogenesis and dynamics from genetic analysis of a C-terminally tagged pilin: a role for O-linked glycosylation. *Molecular Microbiology*, **2012**, *85*, 1166.
[225] M. P. Jennings, F. E. C. Jen, L. F. Roddam, M. A. Apicella and J. L. Edwards.Neisseria gonorrhoeae pilin glycan contributes to CR3 activation during challenge of primary cervical epithelial cells. *Cellular Microbiology*, **2011**, *13*, 885.
[226] F. E. C. Jen, M. J. Warren, B. L. Schulz, P. M. Power, W. E. Swords, J. N. Weiser, M. A. Apicella, J. L. Edwards and M. P. Jennings.Dual Pili Post-translational Modifications Synergize to Mediate Meningococcal Adherence to Platelet Activating Factor Receptor on Human Airway Cells. *PLoS pathogens*, **2013**, *9*, e1003377.

Chapter 2

Development of Bottom-Up Mass Spectrometry for the Analysis of PilE from Neisseria meningitidis

Neisseria meningitis 8013 (Nm 8013) is a serogroup C strain isolated from the blood of a 57 year old male at the Institut Pasteur in 1989[1]. It belongs to the ST-18 clonal complex that has previously been associated with meningococcal disease in central Europe. Its genome has been fully sequenced and annotated and is freely available online via the NeMeSys portal[2]. This makes Nm 8013 both a very useful and relevant reference strain that has been used by many groups to study type IV pilus biology and other factors involved in host cell colonisation.

A clone of Nm 8013 (clone 12 Opa^-, Opc^-, $PilC1^+/PilC2^+$) has previously been used by our group in collaboration with that of Guillaume Duménil to investigate the posttranslational modification status of the major pilin PilE. In 2007 it was reported that PilE from this strain harboured the novel glycan substituent GATDH, rather than the DATDH sugar previously described for other Nm strains; a finding that could be generalised to almost half of clinical isolates[3]. This paper contains two important results that form the starting point for the work presented in this thesis.

The first concerns the identification and localisation of the GATDH glycan. In this paper, glycan identification and characterisation was achieved entirely by mass spectrometry. Positively charged ions of PilE were produced by nano electrospray ionisation, the glycan detached from these protein ions and characterised inside the mass spectrometer. In order to promote cleavage of the labile glycan protein bond without completely fragmenting the protein backbone, protein ions were subjected to moderate energy collisional activation in the high pressure source region. A schematic representation of this mild in-source dissociation (ISD) technique is shown in Figure 41. ISD produced very abundant glycan oxonium ions that appeared as intense peaks in the low mass region of the mass spectrum. Further selection and fragmentation of the GATDH oxonium ion yielded a characteristic glycan fingerprint. This was compared with a similar fingerprint produced by the DATDH oxonium ion and rationalisation of the differences between the fragmentation patterns enabled further structural elucidation of the GATDH glycan.

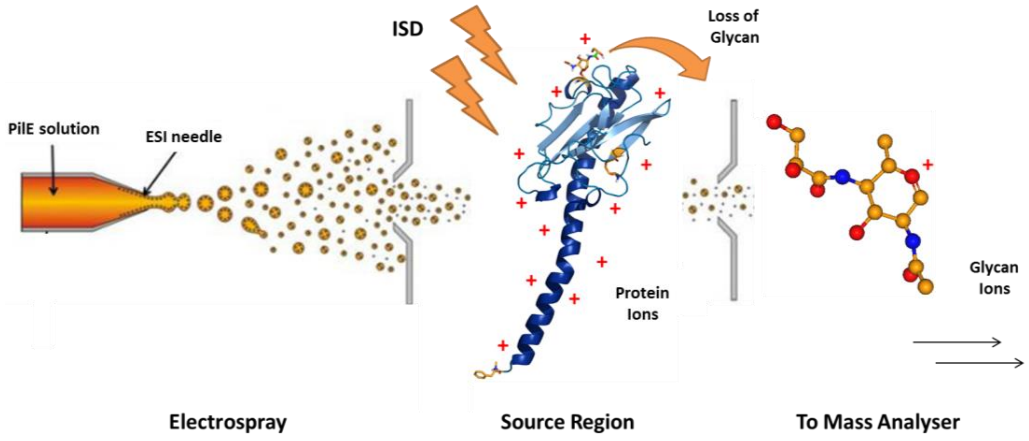


Figure 41 - Schematic representation of ISD showing the ESI of protein ions followed by application of collisional activation in the source region of the MS and loss of the glycan as an oxonium ion

The second point concerns the mass profiling results presented in the paper. High resolution FT-ICR mass profiling of wild type PilE revealed two peaks in an approximate 4:1 ratio (Figure 42). When the monoisotopic protein mass of the larger peak was compared to that predicted from the *pilE* gene, the experimental mass was higher than expected, even when taking into account known posttranslational modifications (N-terminal methylation, oxidised cysteines) and the additional mass imparted by the GATDH glycan. It was therefore supposed that PilE could be further posttranslational modified. In addition, a smaller peak at ≈17,645 Da was consistently present in mass profiles of PilE from multiple preparations. This suggested that perhaps PilE was expressed in multiple proteoforms which may harbour additional or different posttranslational modifications.

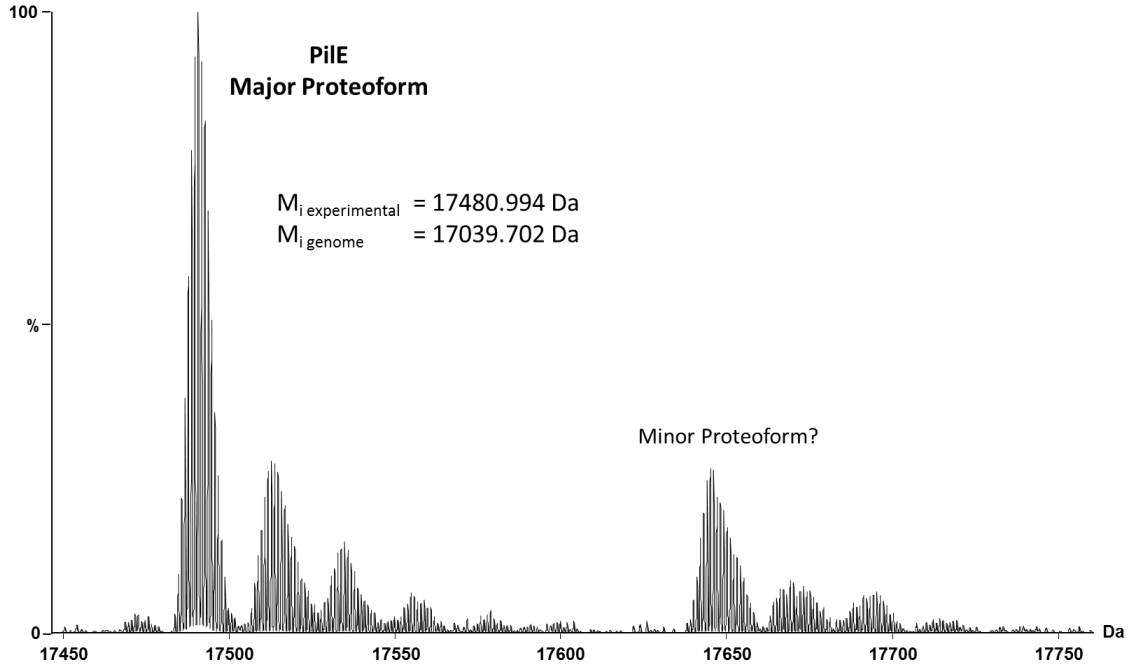


Figure 42 - nano-ESI FT-ICR high resolution mass profile of PilE-8013 deconvoluted across the entire mass range. Measured and predicted monoisotopic masses (M_i) of the major proteoform are indicated

It was, therefore, of great interest to develop a robust mass spectrometry based method which would allow the complete characterisation of PilE, including identification and site localisation of all PTMs, without any *a priori* information and without the time consuming creation of PilE sequence mutants. A combination of high resolution mass profiling and bottom-up mass spectrometry seemed ideally suited to this task. Given that the proteoform population appeared to contain only two components it also seemed suitable for the identification and characterisation of the putative second proteoform.

1. Development of a Bottom-Up MS Approach for the Characterisation of PilE from *Neisseria meningitidis* 8013 (PIlE-8013)

PilE can be purified from cultured bacteria using a procedure that has been modified from the quick pili preparation developed by Brinton *et al.*[4]. Bacteria are grown on agar plates and resuspended in ethanolamine buffer. Intact pili are separated from the bacterial body by mechanical shearing. Solubilisation of the pili fibres at high pH is followed by centrifugation to remove the bacterial debris and pili are then precipitated using ammonium sulphate. A final centrifuge step yields a pellet enriched in PilE that requires no further purification. All PilE preparations in this thesis were performed by the author and M. Christian Malosse from bacteria cultured in the lab of Dr Guillaume Duménil.

This protocol was implemented and re-optimised (due to lab relocation) to yield high quantities of PilE from the 8013 strain (PilE-8013). Only a small amount of high molecular weight contaminant was evidenced by SDS-PAGE. All future preparations presented were also controlled in this manner. After high resolution FT-ICR mass profiling, in-solution enzymatic digestion of PilE with trypsin yielded peptides that were characterised using off-line nano-ESI Q-ToF MS and examined manually for posttranslational modifications. The complete characterisation of PilE-8013 through the thorough analysis of this tryptic digest is presented in the following publication.

2. Published Article - "A combined mass spectrometry strategy for complete posttranslational modification mapping of Neisseria meningitidis major pilin"

Research article

Journal of MASS SPECTROMETRY

Received: 22 July 2013 Revised: 15 August 2013 Accepted: 15 August 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI 10.1002/jms.3262

A combined mass spectrometry strategy for complete posttranslational modification mapping of *Neisseria meningitidis* major pilin

Joseph Gault,[a,b] Christian Malosse,[a,b] Guillaume Duménil[c,d] and Julia Chamot-Rooke[a,b*]

Herein, we report a new approach, based on the combination of mass profiling and tandem mass spectrometry, to address the issue of localising all post-translational modifications (PTMs) on the major pilin protein PilE expressed by the pathogenic *Neisseria* species. PilE is the main component of type IV pili; filamentous organelles expressed at the surface of many bacterial pathogens and important virulence factors. Previous reports have shown that PilE can harbour various combinations of PTMs and have established strong links between PTM and pathogenesis. Complete PTM mapping of proteins involved in bacterial infection is therefore highly desirable. The methodology we propose here allowed us to fully characterise the PilE proteoforms of *Neisseria meningitidis* strain 8013, definitively identifying *all* PTMs present on *all* proteoforms and localising their position on the protein backbone. These modifications include a processed and methylated N-terminus, disulfide bridge, glycosylation and glycerophosphorylation at two different sites. A key element of our approach is high resolution, intact mass measurement of the proteoforms, a piece of information completely lacking in all classical bottom–up proteomics strategies used for PTM analysis and without which it is difficult to ensure complete PTM mapping. Copyright © 2013 John Wiley & Sons, Ltd.

Additional supporting information may be found in the online version of this article at the publisher's web site.

Keywords: Posttranslational Modification; pathogenic Neisseria; Accurate mass profiling; FT-ICR; type IV pili

Introduction

Neisseria meningitidis (Nm) and *Neisseria gonorrhoeae* (Ng) are gram-negative bacterial pathogens responsible for the eponymous human conditions meningitis and gonorrhea. Each inhabits a distinct biological niche; Ng is a commensal of the urogenital tract and Nm of the nasopharynx. To progress from benign commensal to potentially life threatening pathogen, Nm must first cross the epithelial layer and access the circulatory system. Here, it may proliferate causing septicaemia and cross the endothelial layer to proliferate in the cerebrospinal fluid causing inflammation and meningitis.[1,2] On the other hand, Ng mostly causes local inflammation but can also cross the epithelial barrier and spread within the host, causing systemic infections that may lead to arthritis, meningitis or pneumonia.[3] For both Ng and Nm bacterial adhesion is crucial for effective mucosal colonisation and relies on long, dynamic and filamentous organelles called type IV pili that protrude from the bacterial outer membrane.[2,4] Type IV pili are implicated in a variety of processes including DNA uptake, twitching motility, host cell motility, host cell adhesion and bacterial aggregation.[5] Importantly, they have been found to be essential for several phases of the bacterial life cycle that ultimately determine pathogenic efficacy.[6,7]

At the molecular level, Type IV pili are biological macropolymers composed of noncovalently bonded, pilin protein subunits that are arranged in a helical fashion to create long and flexible fibres. In *Neisseria*, the major component of the fibre (major pilin) is the pilin protein PilE. PilE is a 15–18 kDa protein coded by the *pilE* gene. It has been reported to harbour a number of post-translational modifications (PTMs) including a variety of phosphoforms such as phosphocholine,[8] phosphoethanolamine,[9,10] α-glycerophosphate or phosphoglycerol[11] and phosphate.[12] Several glycan motifs have also been described including 2-acetamido 4-glyceramido 2,4,6-trideoxy α-*D*-hexose (GATDH),[13] 2,4-diacetamido 2,4,6-trideoxy α-*D*-hexose (DATDH)[14] and variants arising from phase variation of the pilin glycosylation genes such as the disaccharide Hexose-DATDH (Hex-DATDH), the trisaccharide Hex-Hex-DATDH and their associated *O*-acetylated forms.[15]

The biological role of these PTMs remains a topic of intense current interest.[16] Although glycosylation is probably used by pathogens as a means to escape from the immune system[17,18] and has been linked to host cell adhesion in Ng,[19] its specific biological role in Nm remains unclear. A recent study has shown that *in vivo* modification of PilE with PG is a regulated PTM used by Nm to interfere with pilus–pilus interactions and promote bundle disaggregation; a prerequisite for bacterial

* *Correspondence to: Julia Chamot-Rooke, Structural Mass Spectrometry and Proteomics Unit, Institut Pasteur, CNRS UMR 3528, 26–28 Rue du Docteur Roux, 75724 Paris Cedex 15, France. E-mail: julia.chamot-rooke@pasteur.fr*

a *Département de Chimie, École Polytechnique, CNRS, Laboratoire des Mécanismes Réactionnels (DCMR), 91128, Palaiseau, France*

b *Structural Mass Spectrometry and Proteomics Unit, Institut Pasteur, CNRS UMR 3528, 26-28 Rue du Docteur Roux, 75724, Paris Cedex 15, France*

c *INSERM, U970, Paris Cardiovascular Research Center, Paris, France*

d *Faculté de Médecine Paris Descartes, Université Paris Descartes, Paris, France*

dissemination and invasion.[20] Because PTMs can be strongly related to pathogenesis, efficient analytical methodologies allowing not only their precise identification and localisation in a particular protein sequence, but also their complete and definite mapping across all protein forms (proteoforms) are required.[21] The accurate analysis of these proteoforms is the only way to determine the extent and nature of protein variation, which plays a central role in a wide variety of biological processes including infection and which is a critical piece of information often missing in contemporary proteomics.

The biophysical approaches that have been used to date to characterise pilin PTMs suffer a major drawback in that they only provide partial information about specific PTMs. There is no case where a complete picture of all PilE proteoforms and their associated PTMs has been presented for either Nm or Ng. Isoelectric point determination and monoclonal antibody recognition assays have been used to confirm the presence of PTMs on Ng and Nm pilins but offer no positional information for the modification on the protein backbone. X-ray crystallography experiments revealed the presence of the *O*-linked disaccharide *N*-acetyl glucosamine-α1,3-galactose (GlcNAc-α1,3-Gal)[22] and phosphate[12] on Ng PilE monomers but had to be associated with mass spectrometry (MS) to provide clear evidence of PTM localisation. In that particular case, a classical bottom–up workflow was used where the protein is digested enzymatically and the resulting peptides analysed by tandem mass spectrometry (MS/MS). This methodology was also utilised to identify the PG substituent on Nm pilin and coupled with chemical derivatisation to characterise the PG moiety itself.[11]

More recently, new MS-based approaches relying on the analysis and fragmentation of intact proteins in the gas phase (top–down approaches) have emerged, allowing multiple protein species to be visualised at the same time. This mass profiling strategy has been successfully employed in both Ng and Nm to identify the presence of several pilin proteoforms.[9,20] A top–down approach was also used on Nm PilE to characterise the GATDH subunit. In this study, glycan characterisation was achieved purely through gas phase fragmentation, without any chemical modification providing an excellent example of the large scale dimensionality that MS can offer, something that is often lacking in other techniques.[13]

It is now clear that PilE can be concurrently expressed as a number of different proteoforms, each harbouring its own set of PTMs. However, the characterisation of the entire proteoform population of a single neisserial strain, including complete PTM mapping, has as yet never been described. In this work, we propose a new and simple approach to address this issue, based on the combination of high-resolution mass profiling and MS/MS. This approach allowed us to completely map all PTMs on meningococcal PilE from strain 8013 giving a definitive account of the proteoform population.

Experimental

Pili preparation

Pili from Nm strain 8013 were prepared as described previously.[23] Briefly, the content of 10–12 petri dishes was harvested in 5 ml of 150 mM ethanolamine at pH 10.5. Pili were sheared by vortexing for 1 min. Bacteria were centrifuged at 4000 g for 30 min at 4 °C, and the resulting supernatant further centrifuged at 15 000 g for 30 min at ambient temperature. The supernatant was removed, pili precipitated by the addition of 10% vol. ammonium sulphate saturated in 150 mM ethanolamine pH 10.5 and allowed to stand for 1 h. The precipitate was pelleted by centrifugation at 4000 g for 1 h at 20 °C. Pellets were washed twice with phosphate-buffered saline and suspended in 100 μL distilled water. Protein solutions were desalted by C_4 ZipTip® (Millipore) and eluted in 10 μL 75:25:3 MeOH/water/HCOOH (v/v/v) before MS experiments.

Tryptic digestion

Samples of PilE in H_2O were reduced by one fifth volume of 10 mM dithiothretiol (Sigma Aldrich) in 0.1 M NH_4HCO_3 for 30 min at 56 °C under agitation and alkylated by 55 mM iodoacetamide in 0.1 M NH_4HCO_3. The carbamidomethylated protein was digested with trypsin (Roche) in a 1:35 trypsin: protein ratio overnight at 37 °C under agitation. Samples were taken directly from the digest and desalted by C_{18} ZipTip® (Millipore) before MS analysis in 50:50:0.1 acetonitrile/water/HCOOH (v/v/v).

Fourier transform Ion cyclotron resonance mass spectrometry

Whole protein MS profiling was carried out with a 7T APEX III FT-ICR MS (Bruker Daltonik, Bremen, Germany) equipped with a seven Tesla, actively shielded, superconducting magnet and infinity cell, and fitted with an Apollo™ one nano-electrospray source. Nano-electrospray glass capillaries (Proxeon) were filled with 2–5 μL of the protein solution and subsequently opened by breaking the tapered end of the tip under a microscope. A stable spray was obtained by applying a voltage of about −1 kV between the needle (grounded) and the entrance of the glass capillary used for ion transfer. A source temperature of 120 °C was used for all nano-electrospray experiments. The estimated flow rate was 20–50 nL/min. Ions were stored in the source region in a hexapole guide for 1 s and pulsed into the detection cell through a series of electrostatic lenses. Ions were finally trapped in the cell using SideKick™ and front and back trapping voltages of 0.9 V and 0.95 V. Mass spectra were acquired from m/z 600 to 3000 with 512 k data points. MS spectra were deconvoluted into neutral molecular masses species using DataAnalysis 4.0 (Bruker Daltonics) using the entire protein maximum entropy option. Peak picking was performed using the SNAP 2.0 algorithm. An external mass calibration using NaI (2 g/L in isopropanol/water) was performed weekly.

Quadrupole-time-of-flight mass spectrometry

Mass spectrometry and MS/MS experiments on tryptic peptides were performed on a Q-ToF-Premier™ (Waters Corp., Milford, MA, USA); a quadrupole, orthogonal acceleration time-of-flight mass spectrometer. The proteins were ionised using nano-electrospray ionisation in positive mode (ZSpray™). The source temperature was set to 80 °C. The capillary voltage was tuned manually between 2.1 and 2.6 kV and cone voltage set to 40 V. MS experiments were performed in wide pass quadrupole mode, with the time-of-flight data being collected between m/z 200–2000 with a low-collision energy of 5 eV. Argon was used as the collision gas. Scans were collected for 1 s and accumulated to increase the signal/noise ratio. MS/MS experiments were performed using a variable collision

Journal of MASS SPECTROMETRY

energy (20–32 eV), which was optimised for each precursor ion. Mass Lynx 4.1 was used both for acquisition and data processing. External calibration was performed daily with clusters of phosphoric acid (0.01 M in 50:50 acetonitrile/water). The mass range for the calibration was m/z 70–2000.

Results and discussion

The molecular mass of purified wild type PilE was measured experimentally by nanoESI-FT-ICR MS (Fig. 1A).

Two peaks in an approximate ratio of 4:1 were observed, together with smaller satellite peaks corresponding to sodium and potassium adducts. Monoisotopic neutral molecular masses of these two peaks were measured at 17 480.994 Da (major) and 17 634.919 Da (minor), respectively. We will subsequently refer to these forms as 'major' and 'minor'. This led us to the conclusion that in strain 8013 PilE predominately exists as two proteoforms in a 4:1 ratio. Note that this relative proportion does not take into account possible different ionisation efficiencies for the different proteoforms.

In order to evaluate the extent of PTM of PilE, we compared the measured protein masses to the theoretical mass obtained from the genome. Sequence information on the Nm 8013 strain is freely available through the online resource NeMeSys.[24] Direct translation of the *pilE* gene gives a 168 amino acid protein that begins with the leader sequence MNTLQKG. For surface expressed PilE, this seven amino acid sequence is cleaved from the prepillin by the bifunctional endoprotease PilD,[25] yielding a protein that has an *N*-terminal phenylalanine and a theoretical monoisotopic molecular mass of 17 039.702 Da. The sequence of mature PilE is shown in Fig. 1B.

Considering the measured mass of the major proteoform, these data indicate a +441.292 Da (17 480.994–17 039.702) experimental/theoretical difference. In the following sections, we outline how MS was used to fully explain this 441.292 Da discrepancy and completely describe the PTM of both PilE proteoforms.

Presence of a disulfide bond

Two conserved cysteine residues are omnipresent in PilE from different Nm strains, structurally forming a distinct topology enclosing the genetically hyper variable D region.[26] In the 8013 strain, they are present at residues 120 and 154 and are the only two cysteines in the protein sequence. Although some predictive software exists, unambiguous identification of intact disulfide bonds in proteins has classically involved enzymatic or chemical digestion before and after the protein has been subjected to reduction and cysteine derivatisation.[27] This is generally followed by direct examination of the digest products by MS. When a sufficiently high-resolution MS is available to obtain protein spectra at isotopic resolution, disulfide bond presence can be confidently assessed in a much shorter time scale by intact protein mass measurement before and after reduction with the appropriate reductant (mercaptoethanol, TCEP or DTT).[28]

After treatment of PilE with an excess of TCEP, the protein was desalted by C_4 ZipTip and its mass measured by FT-ICR MS (Supplementary Fig. 1). When compared to untreated wild type PilE a shift of the major isotope envelope by +1.998 Da was observed, consistent with reduction of a disulfide bond. A similar shift was observed for the minor proteoform in the spectrum and confirmed that both wild type PilE proteoforms exist with an

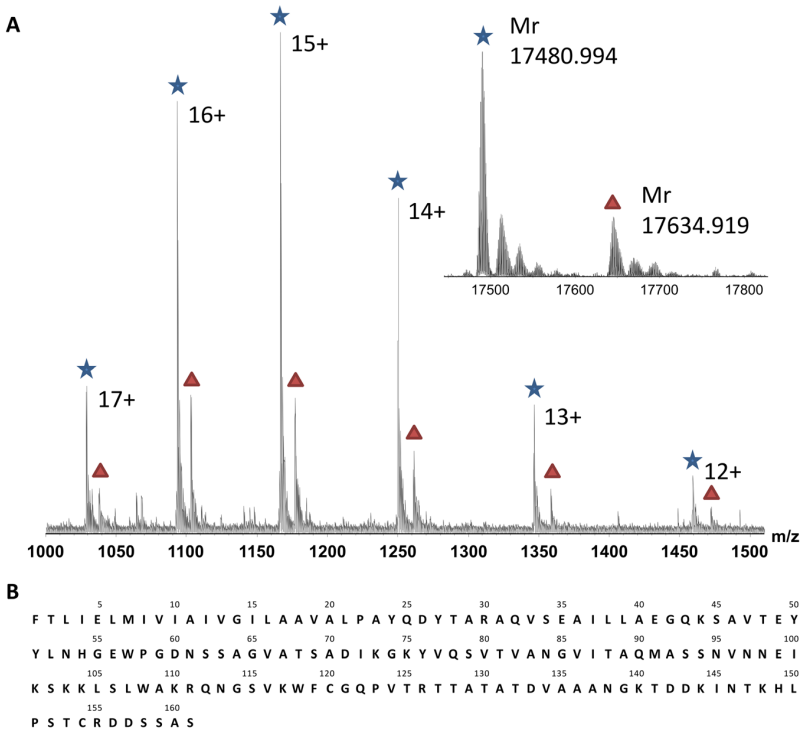


Figure 1. (A) Mass profiling of PilE. Peaks corresponding to major and minor proteoforms are highlighted with blue stars and red triangles respectively. A deconvoluted mass spectrum with monoisotopic neutral proteoform masses is shown in the inset. (B) Sequence of PilE.

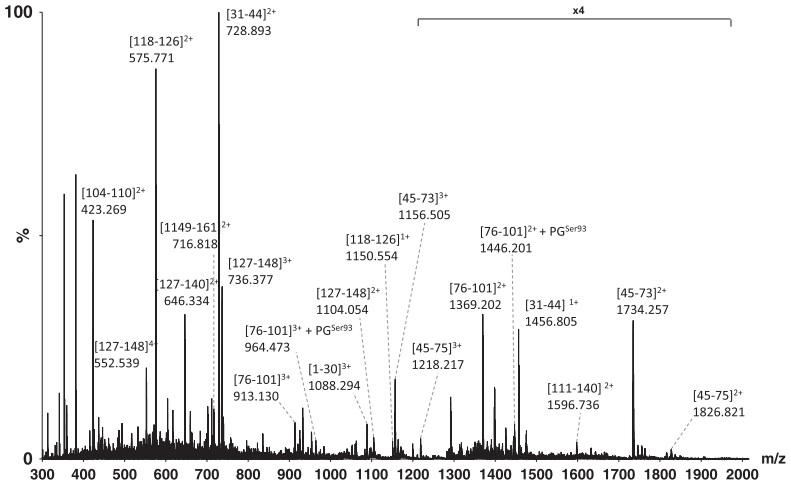


Figure 2. Peptide mass fingerprint of PilE tryptic digest products acquired by nanoESI-Q-ToF MS. [43–75] and [45–75] digest products always harbour GATDH and phosphoglycerol. A full list of ion masses is given in Table 1 with their associated error.

oxidised cysteine bridge. With respect to explaining the 441.292 Da theoretical/experimental mass difference, the presence of a disulfide bond effectively increases this mass by 2.015 Da to 443.307 Da.

N-terminal processing

After multiple repeats, it was established that the quick pilus preparation consistently produced a good yield of PilE with only very small quantities of high mass protein impurities. This is evidenced by the clean mass profiling experiments and was verified by SDS-PAGE (results not shown). Given the sample purity, crude PilE extracts were subjected to tryptic digestion without further clean-up and analysed directly by nanoESI-Q-ToF MS; that is without any chromatographic separation of the resulting digestion products (Fig. 2).

This approach allowed rapid examination of the digest and was amenable to MS/MS of even low-abundance peaks over an extended time frame. This was useful to improve the quality of the MS/MS spectra and increase both peptide sequence coverage and confidence whilst allowing manual tuning of collision energy in order to retain PTMs.

Peaks were assigned to digest products based on an *in silico* digestion of the PilE primary sequence (with no PTM) and comparison between experimental and theoretical monoisotopic masses (Peptide Mass Fingerprinting). An initial sequence coverage of 61% was achieved covering ranges [31–44], [76–101] and [104–161]; however, peptides spanning the [1–30] and [45–75] range were missing.

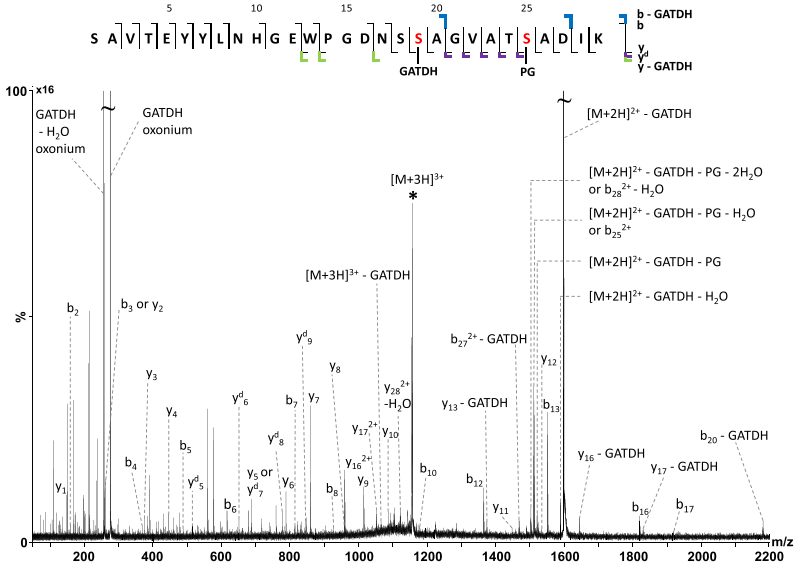


Figure 3. Tandem mass spectrum of m/z 1 156.558 (3+) corresponding to the [45–73] peptide.

Journal of MASS SPECTROMETRY

Tandem mass spectrometry was performed on all ions corresponding to predicted digest fragments and on as many unassigned peaks in the spectrum as possible. Of the unassigned peaks only the triply charged ion at m/z 1088.294 yielded fragments consistent with the peptide sequence FTLIELMIVIAIVGILAAVALPAYQDYTAR. This is the sequence of the N-terminal peptide [1–30], however, the absolute mass values of the parent ion and all *a*-type and *b*-type fragment ions were 14.02 Da greater than expected (Supplementary Fig. 2). This feature could easily be explained by N-terminal methylation of PilE. This modification has been previously reported and attributed to a secondary function of the enzyme PilD.[25] Indeed, when the phenylalanine is methylated, an a_1 ion at m/z 134.100 can also be assigned. This unusual product ion is expected because the methyl imparts additional stability to the phenylalanine immonium ion.[29] N-terminal methylation increases digest sequence coverage to 80% and reduces the unaccounted mass on the entire protein from 443.307 Da to 429.292 Da.

An unexpected deamidation

Interestingly, the 3+ ion at m/z 736.377 corresponding to the [127–148] digest product exhibited a severely distorted isotopic distribution (Supplementary Fig. 3). Quadrupole selection followed by collisionally activated dissociation of this ion confirmed almost complete deamidation of Asn^{138}. Deamidation is a common protein modification, which can either arise from a post-translational process or simply from sample ageing or degradation. It is known that asparagines vicinal to a glycine (which is the case here) are particularly prone to degradative deamidation.[30] Because different samples of PilE analysed by MS exhibited different proportions of deamidation and those analysed within a few hours of preparation exhibited very little, Asp^{138} was considered an artefact rather than a *bone fide* PTM. The presence of a deamidation reduces the unexplained mass by 0.984 Da to 428.308 Da.

Glycan and phosphoglycerol localisation

As reported previously, for this strain the PTMs GATDH and PG are always present on Ser^{63} and Ser^{69} respectively. Indeed, Ser^{63} is consistently and exclusively the sole reported site of glycan (GATDH, DATDH and DATH-Hex…) attachment for PilE. This is not the case for PilE phosphoform modification (PG, phosphocholine and phosphoethanolamine), which has mainly been reported on Ser^{68} or Ser^{69}, but also to a lesser extent on Ser^{93} or Ser^{94}.

Taking GATDH and PG modifications into account and making the necessary modifications to the theoretical digest, the [45–73] and [45–75] (one miscleavage) fragments bearing both modifications could be identified as doubly charged ions at m/z 1734.257 and m/z 1826.821 and triply charged ions at m/z 1156.505 and m/z 1218.217(Table 1).

To check the location of these modifications, both the 2+ and 3+ charge states of the more abundant [45–73] fragment were subjected to tandem MS. Fragmentation of the 2+ charge state, which required a fairly high-collision energy of around 30 eV, led to the production of many secondary fragments that complicated the spectrum, particularly in the low m/z range. MS/MS experiments on the 3+ charge state of the [45–73] fragment proved to be more informative (Fig. 3).

A striking feature of the resultant MS/MS spectrum is the oxonium ion of GATDH, which predominates at m/z 275.126 alongside its dehydrated form at m/z 257.115. These ions confirm the presence of GATDH on this peptide and can also be used in future experiments as reporter ions for this particular glycan. An intense, doubly charged ion at m/z 1597.197, corresponding to the loss of GATDH from the parent ion is also observed along

Table 1. Tryptic digest products of PilE from the associated mass spectrum in Fig. 2.

Digestion Fragment	Charge State	Measured m/z	Deconvoluted Monoisotopic Mass $[M+H]^+$ (Da)	Theoretical Monoisotopic Mass $[M+H]^+$ (Da)	Error (ppm)
[1-30]+Me	3	1088.294	3262.868	3262.852	4.90
[31-44]	1	1456.805	1456.805	1456.801	2.89
[45-73]+PG+GATDH	3	1156.505	3467.502	3467.516	-4.21
[45-73]+PG+GATDH	2	1734.257	3467.506	3467.516	-2.92
[45-75]+PG+GATDH	3	1218.217	3652.636	3652.633	1.03
[45-75]+PG+GATDH	2	1826.821	3652.635	3652.633	0.50
[76-101]	3	913.130	2737.377	2737.383	-2.15
[76-101]	2	1369.202	2737.396	2737.383	4.82
[76-101]+PG	3	964.473	2891.403	2891.386	5.99
[76-101]+PG	2	1446.201	2891.395	2891.386	3.28
[104-110]	2	423.269	845.531	845.524	7.78
[111-140]*	2	1596.736	3192.465	3192.597	-41.22
[118-126]	2	575.771	1150.535	1150.546	-9.65
[118-126]	1	1150.554	1150.554	1150.546	6.32
[127-140]	2	646.334	1291.660	1291.649	8.58
[127-140]	1	1291.647	1291.647	1291.649	-1.19
[127-148]	4	552.539	2207.135	2207.115	9.05
[127-148]	3	736.377	2207.117	2207.115	0.79
[127-148]	2	1104.054	2207.102	2207.115	-6.11
[149-161]	2	716.818	1432.629	1432.612	11.58

* The parts per million error is rather large for this fragment ion because it exhibits a very high proportion of deamidation. In consequence, the monoisotopic peak is extremely small and difficult to resolve from the baseline.

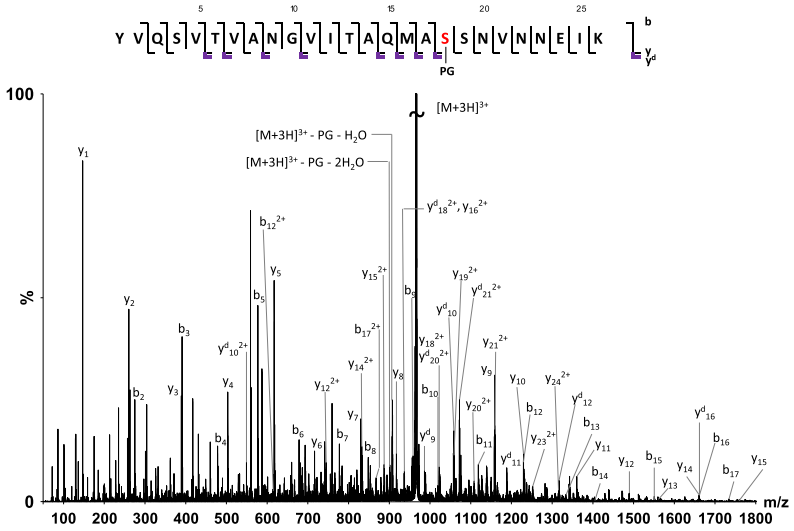


Figure 4. Tandem mass spectrum of m/z 964.472 (3+) corresponding to the [76–101]+PG peptide.

with additional molecular ion fragmentation products corresponding to further loss of PG (m/z 1520.121) concomitant with water (m/z 1511.186). This fragmentation pattern indicates that the glycosidic bond is much weaker than that between the PG and the protein backbone. Once the [45–73] fragment had been modified with GATDH and PG, a large series of *b*-type and *y*-type fragment ions could be assigned. Indeed, a complete *y* ion series y_1 to y_{12} is present and of particular importance are the y_{11} and y_{12} ions that confirm the modification of Ser^{63} with GATDH. *b* and *y* ions exhibiting loss of glycan and/or water are also present in some cases. These results confirm without ambiguity the localisation of GATDH and PG to Ser^{63} and Ser^{69}, respectively. The absence of ions in the digest corresponding to the [45–73] fragment without PG, or without GATDH or without both modifications, strongly suggests that PilE is always modified with these PTMs and thus confirms that the major peak in the mass profiling experiment corresponds to a single proteoform of PilE.

It is important to note that even if the identities of these modifications had been completely unknown, the extensive *b* and *y* ion coverage in the MS/MS spectrum would have allowed us to pinpoint two modifications 274 Da and 154 Da in mass on Ser^{63} and Ser^{69}, respectively. This would have enabled the *de novo* identification and localisation of both PTMs. Finally, the presence of a GATDH and a PG allows us to completely resolve the unexplained 428.308 Da experimental/theoretical mass difference for the major proteoform.

The difference of 153.925 Da between the major and minor proteoforms is very close to the theoretical 154.003 Da expected

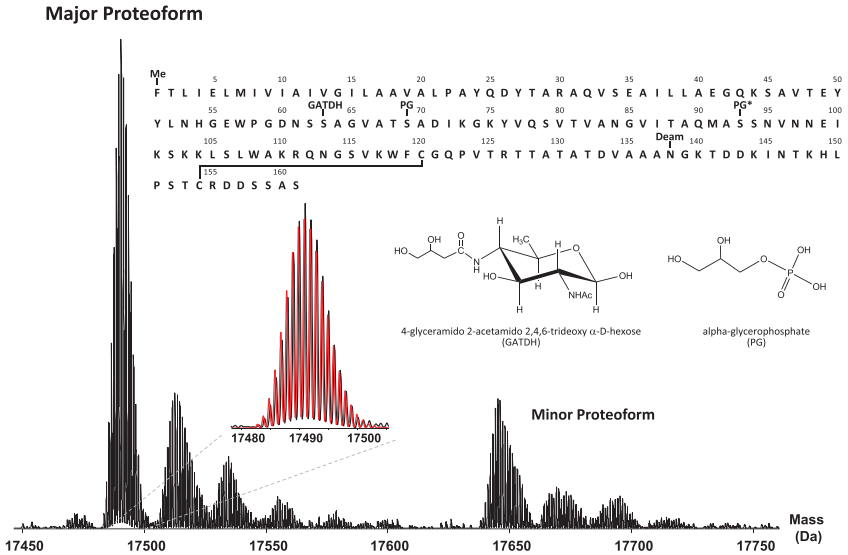


Figure 5. Fourier transform ion cyclotron resonance mass profile of PilE complete with protein sequence showing the location of all PTMs. PG* denotes modification only in the minor proteoform. Once all PTMs are taken into account the experimental spectrum (black) correlates very well with a theoretically generated isotope pattern (red).

Journal of MASS SPECTROMETRY

for phosphoglycerol. This confirms the presence of an additional PG modification in the minor form, which has previously been reported on Ser^{93}.[20] To check if the analysis of our digest would have been sufficient to obtain both the identification of an extra PG and its localisation without any *a priori* information, peaks in the digest spectrum were systematically checked for partner pairs at ≈+154 Da. The only partners found matching this criterion are the doubly charged ion at m/z 1446.201 and the triply charged ion at m/z 964.473 that both correspond to the [76–101] + PG fragment. In order to verify this assignment the ion at m/z 964.473 was subjected to MS/MS (Fig. 4).

The MS/MS spectrum is dominated by large series of *b* and *y* ions and also dehydroalanine type *y* ions (y^d) formed by loss of PG from Ser^{93}. Several triply charged ions of low abundance, corresponding to loss of PG from the molecular ion with one or several molecules of water are also visible in the spectrum.

The presence of the complete *y*-type ion sequence from y_1 (m/z 147.123) to y_{16}^{2+} (m/z 936.957) both confirms the presence of PG on Ser^{93} and also negates modification of Ser^{94}. Note that several *y* ions starting at y_9 (m/z 1158.527) are accompanied by dehydroalanine partners, for example y_9^d at m/z 986.501. The *b*-type ion series covers b_2 to b_{17}. It is particularly interesting to note that the 4:1 relative ion intensity ratio, which is observed for the [76–101] and [76–101] + PG fragment ions, matches the major:minor ratio observed for the intact proteoforms. This result indicates that the effect of the additional PG on ionisation efficiency is similar at both the peptide and protein level.

Complete post-translation modification mapping for both major and minor proteoforms

Taken in the context of the mass profiling experiment, which indicates the presence of two PilE proteoforms in a 4:1 ratio, our MS/MS data enables the complete characterisation of both major and minor proteoforms. The PTMs found for both proteoforms are: prepeptide cleavage, *N*-terminal methylation, GATDH on Ser^{63}, phosphoglycerol on Ser^{69} and a disulfide bridge. The minor proteoform harbours an additional phosphoglycerol on Ser^{93}. Both species are prone to degradative deamidation principally at Asn^{138}. Genetic point mutation of PTM sites has been used to confirm the exclusive modification of Ser^{69} with PG in the major PilE proteoform and exclusive modification of Ser^{69} and Ser^{93} with PG in the minor PilE proteoform.[20]

Once all PTMs are taken into account the sequence coverage obtained for the tryptic digest is extended to 99%. Furthermore, the mass discrepancy observed between theoretical and experimental PilE molecular masses in the mass profiling experiment can be completely resolved. For the major proteoform, a theoretical protein mass of 17 480.806 Da can be calculated, which when compared with the experimental 17 480.994 Da gives an associated error of +0.188 Da or 10.7 ppm. This is within the expected error range for mass measurement using an externally calibrated 7T FT-ICR instrument. As shown in Fig. 5, the overlap of theoretical and experimental isotopic patterns is excellent, confirming that our approach has led to a complete mapping of all PTMs. This point is important, as it confirms the exclusive presence of the described modifications and excludes any other PTMs.

Similar results are obtained for the minor proteoform: the calculated theoretical mass is 17 634.809 Da and the experimental mass 17 634.919 Da giving an associated error of +0.110 Da or 6.3 ppm. Therefore, all PTMs present on all PilE proteoforms have been accounted for and the proteoform population completely described for this Nm strain.

It is interesting to note that all of the identified PTMs are located in defined structural regions: α-β loop for the glycan at Ser^{63} and PG at Ser^{68}/Ser^{69}, β_2-β_3 loop for the PG at Ser^{93}/Ser^{94}. These structural features are located in the globular domain of PilE, which protrudes from the pilus surface (Supplementary Fig. 3).[31] This strongly suggests that the PTMs of PilE are directly involved in mediating interactions between the pilus and other molecules. In the case of PG modification at Ser^{93}, it has been shown that an increased modification level at this site changes the charge surface of the pilus fibre, making bundling unfavourable and disrupting pilus-pilus interactions.[20] It is unclear if the other modifications play similar roles but their conserved location in surface exposed regions suggest that it is probably the case.

Conclusion

In this paper, we describe a simple but efficient combined MS approach leading to complete PTM mapping of the PilE population of Nm. Using high-resolution mass profiling and MS/MS experiments, the proteoform population of PilE is completely and unambiguously characterised, including multiple PTM: prepeptide cleavage and *N*-terminal methylation, a disulfide bridge, glycerophosphorylation and glycosylation. Genetic point mutation has further validated our results.[20] The systematic approach, we propose here, is well suited to similar cases where there are multiple modifications leading to several proteoforms. An important feature is the measurement of accurate monoisotopic molecular masses of intact proteins, which can reveal the presence of modifications as subtle as deamidation. This approach allows us to obtain the 'the whole picture' (all proteoforms and PTMs) and not merely a partial view. This information is crucial for studies aiming to discover or study virulence factors of pathogenic bacteria; a key requirement in understanding pathogenesis and provides the basis for developing novel drugs and designing new vaccines.

References

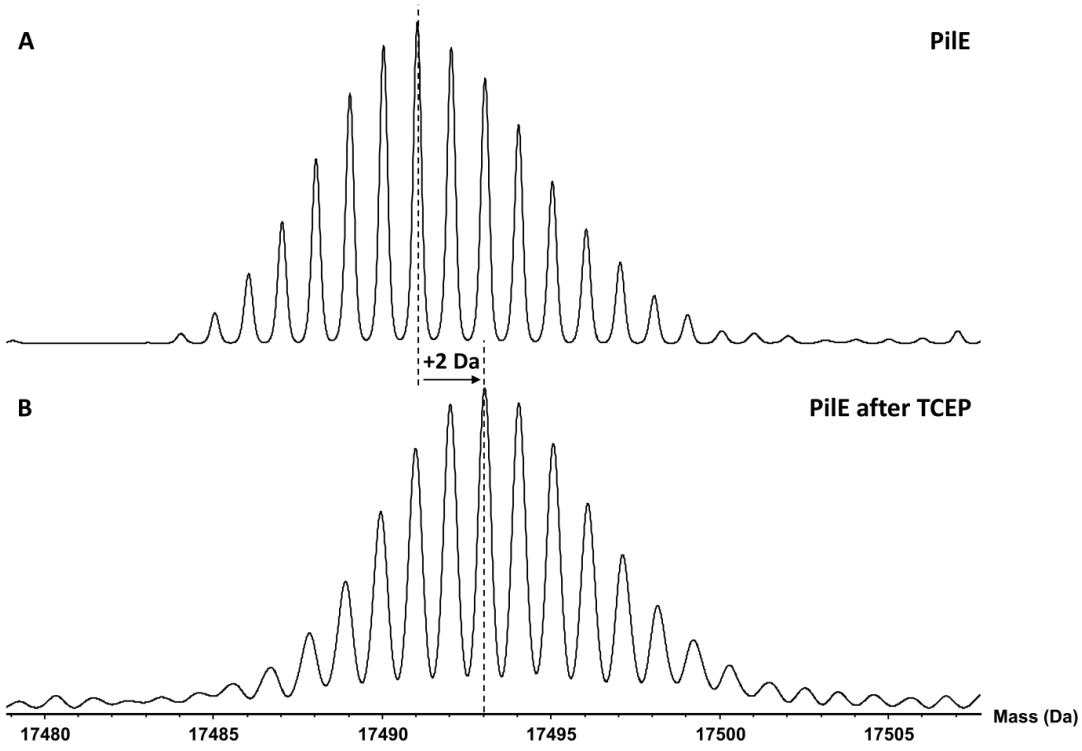
[1] K. Melican, G. Dumenil. Vascular colonization by *Neisseria meningitidis*. *Curr. Opin. Microbiol.* **2012**, *15*, 50.
[2] K. Trivedi, C. M. Tang, R. M. Exley.Mechanisms of meningococcal-colonisation. *Trends Microbiol.* **2011**, *19*, 456.
[3] J. L. Edwards, M. A. Apicella. The molecular mechanisms used by *Neisseria gonorrhoeae* to initiate infection differ between men and women. *Clin. Microbiol. Rev.* **2004**, *17*, 965.
[4] J. Swanson. Studies on gonococcus infection 4. Pili - their role in attachment of gonococci to tissue-culture cells. *J. Exp. Med.* **1973**, *137*, 571.
[5] C. L. Giltner, Y. Nguyen, L. L. Burrows. Type IV pilin proteins: versatile molecular modules. *Microbiol. Mol. Biol. Rev.* **2012**, *76*, 740.
[6] E. Mairey, A. Genovesio, E. Donnadieu, C. Bernard, F. Jaubert, E. Pinard, J. Seylaz, J. C. Olivo-Marin, X. Nassif, G. Dumenil. Cerebral microcirculation shear stress levels determine *Neisseria meningitidis* attachment sites along the blood–brain barrier. *J. Exp. Med.* **2006**, *203*, 1939.
[7] X. Nassif, S. Bourdoulous, E. Eugene, P. O. Couraud. How do extracellular pathogens cross the blood–brain barrier?, *Trends Microbiol.* **2002**, *10*, 227.
[8] J. N. Weiser, J. B. Goldberg, N. Pan, L. Wilson, M. Virji. The phosphorylcholine epitope undergoes phase variation on a 43-kilodalton protein in *Pseudomonas aeruginosa* and on pili of *Neisseria meningitidis* and *Neisseria gonorrhoeae*. *Infect. Immun.* **1998**, *66*, 4263.
[9] F. E. Aas, W. Egge-Jacobsen, H. C. Winther-Larsen, C. Lovold, P. G. Hitchen, A. Dell, M. Koomey. *Neisseria gonorrhoeae* type IV pili undergo

multisite, hierarchical modifications with phosphoethanolamine and phosphocholine requiring an enzyme structurally related to lipopolysaccharide phosphoethanolamine transferases. *J. Biol. Chem.* **2006**, *281*, 27712.
[10] F. T. Hegge, P. G. Hitchen, F. E. Aas, H. Kristiansen, C. Lovold, W. Egge-Jacobsen, M. Panico, W. Y. Leong, V. Bull, M. Virji, H. R. Morris, A. Dell, M. Koomey. Unique modifications with phosphocholine and phosphoethanolamine define alternate antigenic forms of *Neisseria gonorrhoeae* type IV pili. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 10798.
[11] E. Stimson, M. Virji, S. Barker, M. Panico, I. Blench, J. Saunders, G. Payne, E. R. Moxon, A. Dell and H. R. Morris. Discovery of a novel protein modification: alpha-glycerophosphate is a substituent of meningococcal pilin. *Biochem. J.* **1996**, *316*, 29.
[12] K. T. Forest, S. A. Dunham, M. Koomey, J. A. Tainer. Crystallographic structure reveals phosphorylated pilin from *Neisseria*: phosphoserine sites modify type IV pilus surface chemistry and fibre morphology. *Mol. Microbiol.* **1999**, *31*, 743.
[13] J. Chamot-Rooke, B. Rousseau, F. Lanternier, G. Mikaty, E. Mairey, C. Malosse, G. Bouchoux, V. Pelicic, L. Camoin, X. Nassif, G. Dumenil. Alternative Neisseria spp. type IV pilin glycosylation with a glyceramido acetamido trideoxyhexose residue. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 14783.
[14] E. Stimson, M. Virji, K. Makepeace, A. Dell, H. R. Morris, G. Payne, J. R. Saunders, M. P. Jennings, S. Barker, M. Panico, I. Blench, E. R. Moxon. Meningococcal Pilin - A Glycoprotein Substituted With Digalactosyl 2,4-Diacetamido-2,4,6-Trideoxyhexose. *Mol. Microbiol.* **1995**, *17*, 1201.
[15] B. Borud, F. E. Aas, A. Vik, H. C. Winther-Larsen, W. Egge-Jacobsen, M. Koomey. Genetic, structural, and antigenic analyses of glycan diversity in the O-linked protein glycosylation systems of human neisseria species. *J. Bacteriol.* **2010**, *192*, 2816.
[16] F. E. C. Jen, M. J. Warren, B. L. Schulz, P. M. Power, W. E. Swords, J. N. Weiser, M. A. Apicella, J. L. Edwards, M. P. Jennings. Dual pili post-translational modifications synergize to mediate meningococcal adherence to platelet activating factor receptor on human airway cells. *PLoS Pathog.* **2013**, *9*, e1003377.
[17] M. Virji. Pathogenic neisseriae: surface modulation, pathogenesis and infection control. *Nat. Rev. Microbiol.* **2009**, *7*, 274.
[18] D. J. Vigerust. Protein glycosylation in infectious disease pathobiology and treatment. *Cent. Eur. J. Biol.* **2011**, *6*, 802.
[19] M. P. Jennings, F. E. C. Jen, L. F. Roddam, M. A. Apicella, J. L. Edwards. *Neisseria gonorrhoeae* pilin glycan contributes to CR3 activation during challenge of primary cervical epithelial cells. *Cell. Microbiol.* **2011**, *13*, 885.
[20] J. Chamot-Rooke, G. Mikaty, C. Malosse, M. Soyer, A. Dumont, J. Gault, A. F. Imhaus, P. Martin, M. Trellet, G. Clary, P. Chafey, L. Camoin, M. Nilges, X. Nassif, G. Dumenil. Posttranslational modification of pili upon cell contact triggers *N. meningitidis* dissemination. *Science* **2011**, *331*, 778.
[21] L. M. Smith, N. L. Kelleher. Proteoform: a single term describing protein complexity. *Nat Meth* **2013**, *10*, 186.
[22] H. E. Parge, K. T. Forest, M. J. Hickey, D. A. Christensen, E. D. Getzoff, J. A. Tainer. Structure of the fiber forming protein pilin at 2.6-Angstrom resolution. *Nature*, **1995**, *378*, 32.
[23] E. Carbonnelle, S. Helaine, L. Prouvensier, X. Nassif, V. Pelicic. Type IV pilus biogenesis in *Neisseria meningitidis*: PilW is involved in a step occurring after pilus assembly, essential for fibre stability and function. *Mol. Microbiol.* **2005**, *55*, 54.
[24] C. Rusniok, D. Vallenet, S. Floquet, H. Ewles, C. Mouze-Soulama, D. Brown, A. Lajus, C. Buchrieser, C. Medigue, P. Glaser, V. Pelicic. NeMeSys: a biological resource for narrowing the gap between sequence and function in the human pathogen *Neisseria meningitidis*. *Genome Biol.*, **2009**, *10*, R110.
[25] M. S. Strom, D. N. Nunn, S. Lory. A single bifunctional enzyme, PilD, catalyzes cleavage and *N*-methylation of proteins belonging to the type-IV pilin family. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 2404.
[26] S. Hartung, A. S. Arvai, T. Wood, S. Kolappan, D. S. Shin, L. Craig, J. A. Tainer. Ultrahigh resolution and full-length pilin structures with insights for filament assembly, pathogenic functions, and vaccine potential. *J. Biol. Chem.* **2011**, *286*, 44254.
[27] J. J. Gorman, T. P. Wallis, J. J. Pitt. Protein disulfide bond determination by mass spectrometry. *Mass Spectrom. Rev.* **2002**, *21*, 183.
[28] M. Scigelova, P. S. Green, A. E. Giannakopulos, A. Rodger, D. H. G. Crout, P. J. Derrick. A practical protocol for the reduction of disulfide bonds in proteins prior to analysis by mass spectrometry. *Eur. J. Mass Spectrom.* **2001**, *7*, 29.
[29] J. L. Hsu, S. Y. Huang, J. T. Shiea, W. Y. Huang, S. H. Chen. Beyond quantitative proteomics: signal enhancement of the a(1) ion as a mass tag for peptide sequencing using dimethyl labeling. *J. Proteome Res.* **2005**, *4*, 101.
[30] O. V. Krokhin, M. Antonovici, W. Ens, J. A. Wilkins, K. G. Standing. Deamidation of -Asn-Gly- sequences during sample preparation for proteomics: consequences for MALDI and HPLC-MALDI analysis. *Anal. Chem.* **2006**, *78*, 6645.
[31] L. Craig, M. E. Pique, J. A. Tainer. Type IV pilus structure and bacterial pathogenicity. *Nat. Rev. Microbiol.* **2004**, *2*, 363.

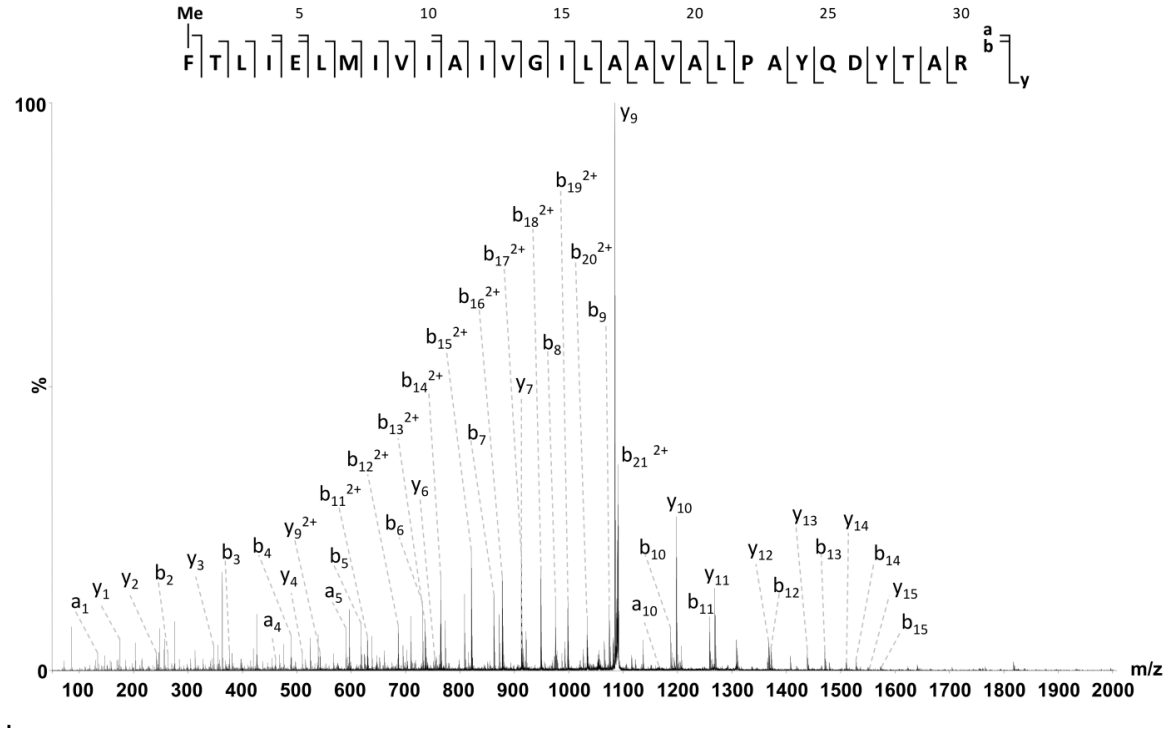
Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

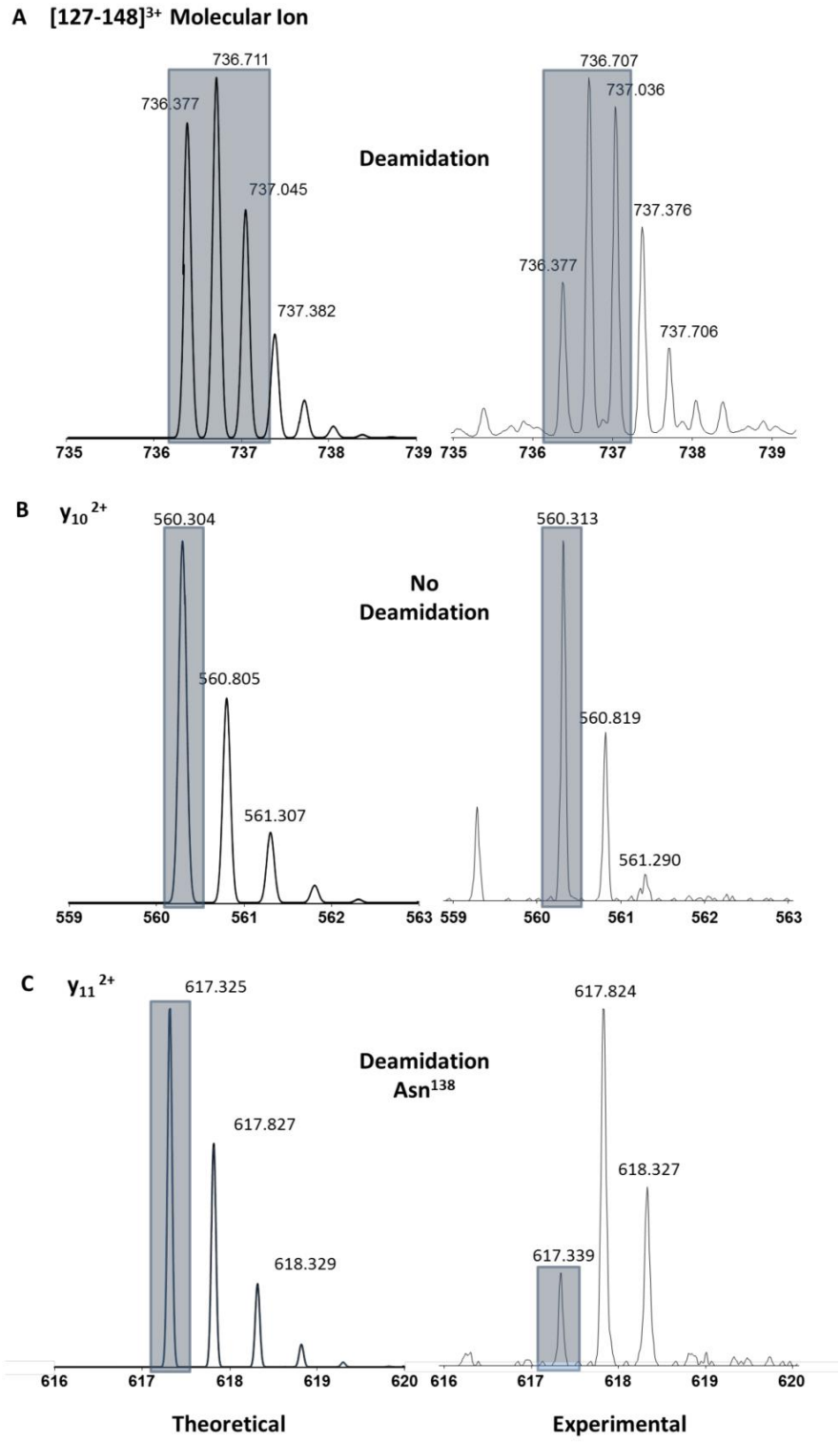
Supplementary Information



Supplementary Figure 1. High resolution FT-ICR MS measurement of PilE. Zoom on the isotope envelope of the major proteoform before (A) and after (B) treatment with TCEP showing a mass increase of ≈2 Da

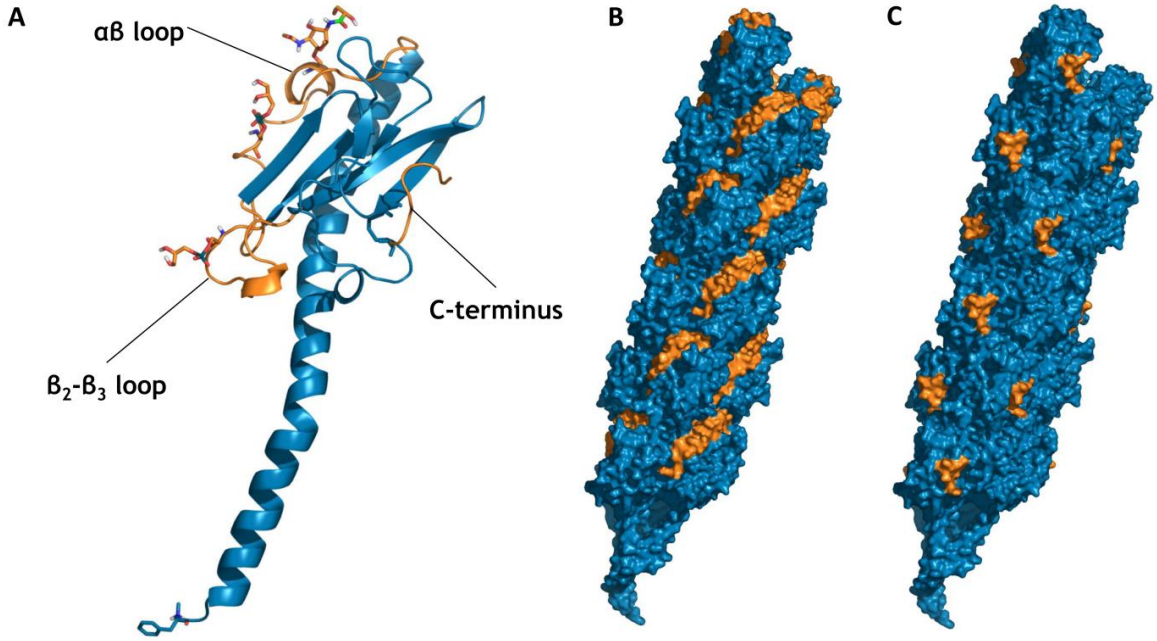


Supplementary Figure 2. MS/MS of the triply charged ion at m/z 1088.294. *b* and *y* fragment ions can be assigned to the N-terminal [1-30] peptide FTLIELMIVEAIVGILAAVAL and give a sequence coverage of 100%. The presence of the a_1 ion confirms methylation at the N-terminus.



Supplementary Figure 3. (A) Isotope envelope of the 3+ ion of the [127-148] peptide after tryptic digestion. Compared to the theoretical distribution (left) the second isotope is much larger than the first (right). This is indicative of a large proportion of peptide deamidation. (B) MS/MS of the [127-148] peptide and zoom on the y_{10} ion GKTDDKINTK. An excellent match with the theoretical distribution indicates no deamidation on this portion of the peptide sequence. (C)

MS/MS of the [127-148] peptide and zoom on the y_{11} ion NGKTDDKINTK. In the experimental distribution (right) the second isotope is of much higher intensity than the first. This does not fit with the predicted theoretical isotope pattern (left) and indicates that a large percentage of $Asn^{138} \rightarrow Asp^{138}$ conversion is responsible for the observed peptide deamidation.



Supplementary Figure 4. (A) PilE monomer with structural regions highlighted in orange. All posttranslational modifications are shown as sticks. (B) Intact pilus with αβ and β_2-β_3 loops highlighted in orange indicating that both of these regions are surface exposed. (C) Intact pilus with region between final cysteine residue and C-terminus highlighted in orange. This region is also exposed at the pilus surface.

3. Conclusions from Bottom-Up Characterisation of PilE-8013

3.1.Mass Spectrometry

The combination of high resolution mass profiling and bottom-up MS analysis proved successful for the complete characterisation of both proteoforms of PilE-8013. The off-line MS strategy used here is particularly adapted for the fragmentation of posttranslationally modified peptides over a long time scale and allows manual tuning of collision energies in order to limit labile PTM loss. It does however suffer from comparably low-sensitivity as all peptides issuing from the digest are visualised in a single spectrum and this compounds ion suppression effects with a limited dynamic range. Sensitivity could have been improved by performing nano-LC MS/MS and obviously ETD or ECD would have been more appropriate for PTM localisation. However, a nano-LC system was not available to us and the sensitivity required for examination and ECD MS/MS of ions present in the peptide digest could not be achieved using the Apex III FT-ICR instrument in the lab at that time.

Despite exhaustive manual MS/MS of ions obtained after digestion it could not be asserted with complete confidence that the other posttranslational modified peptides were not present in the digest. Genetic point mutation and further protein mass profiling were ultimately required to ensure the exclusivity of PTM modification sites. Nevertheless since both the identity of the PTMs and the modification sites have been defined for both proteoforms, the characterisation presented in this article provides a complete description of PilE expressed by wild type Nm 8013 onto which future methods can be developed and against which biological mutants can be compared.

3.2.Biological Relevance

This report represents the first complete characterisation of the entire proteoform population of PilE from a single Nm strain. In addition to a processed and methylated N-terminus, cysteine bridge and GATDH glycan at Ser^{63}, it was found that PilE was modified by a phosphoglycerol or glycerophosphate (PG) moiety at Ser^{69} in the major proteoform and both Ser^{69} and Ser^{93} in the minor proteoform.

PG is an very unusual PTM that was first discovered on meningococcal PilE purified from the C311#16 strain[5]. In this report tryptic digestion of purified PilE was followed by HPLC fractionation of the resultant peptides. Mass measurement of peptides present in one of these fractions by a combination of ESI QqQ MS and FAB two sector MS followed by N-terminal sequencing, led the authors to conclude that the Asn^{84}-Lys^{98} peptide was modified by an unknown PTM of mass ≈154 Da. They conjectured that this could be a PG group. Cleavage of the phosphate peptide bond under basic conditions to release the glycerophosphate moiety, followed by further

trimethylsilyl derivatisation and GC-MS analysis, confirmed the identity of the PTM as α-glycerophosphate. The chirality at the glycerol was not resolved.

Since glycerophosphate is a component of the bacterial inner membrane, it was originally proposed that a PG modification on PilE could perhaps act as a substrate for acetylation and the formation of a lipid anchor[6]. Apart from this extremely tentative hypothesis the function of the PG group remains unclear as does the pathway responsible for its addition to PilE. It was therefore of interest to rationalise the poorly described basis for PG modification and understand the biological function of this highly unusual PTM.

Bibliography

[1] X. Nassif, J. Lowy, P. Stenberg, P. Ogaora, A. Ganji and M. So.Antigenic Variation Of Pilin Regulates Adhesion Of Neisseria-Meningitidis To Human Epithelial-Cells. *Molecular Microbiology*, **1993**, *8*, 719.

[2] C. Rusniok, D. Vallenet, S. Floquet, H. Ewles, C. Mouze-Soulama, D. Brown, A. Lajus, C. Buchrieser, C. Medigue, P. Glaser and V. Pelicic.NeMeSys: a biological resource for narrowing the gap between sequence and function in the human pathogen Neisseria meningitidis. *Genome Biol*, **2009**, *10*, R110.

[3] J. Chamot-Rooke, B. Rousseau, F. Lanternier, G. Mikaty, E. Mairey, C. Malosse, G. Bouchoux, V. Pelicic, L. Camoin, X. Nassif and G. Dumenil.Alternative Neisseria spp. type IV pilin glycosylation with a glyceramido acetamido trideoxyhexose residue. *Proceedings of the National Academy of Sciences of the United States of America*, **2007**, *104*, 14783.

[4] C. C. Brinton, J. Bryan, J. Dillon, N. Guerina, L. J. Jacobson, A. Labik, S. Lee, A. Levine, S. Lim, J. McMichael, S. Polen, K. Rogers, A. C. C. To and S. C. M. To, Immunobiology of Neisseria gonorrhoeae : proceedings of a conference held in San Francisco, California, 18-20 January 1978, San Francisco, California, 1978.

[5] E. Stimson, M. Virji, S. Barker, M. Panico, I. Blench, J. Saunders, G. Payne, E. R. Moxon, A. Dell and H. R. Morris.Discovery of a novel protein modification: alpha-glycerophosphate is a substituent of meningococcal pilin. *Biochem. J.*, **1996**, *316*, 29.

[6] M. Virji.Post-translational modifications of meningococcal pili. Identification of common substituents: Glycans and alpha-glycerophosphate - A review. *Gene*, **1997**, *192*, 141.

Chapter 3

Deciphering the Role of Phosphoglycerol Modification of Type IV Pili in Neisseria meningitidis

At the time of this work no gene products of Nm had been annotated with phosphoglycerol transferase activity. Whilst investigating a group of 15 neisserial genes described in previous reports as being up regulated four hours after host cell contact[1-3], our collaborator Dr Guillaume Duménil discovered that one of these genes (*NMV_0885)* shared significant homology with one of another group of genes in *E. coli* (COG1368) titled phosphoglycerol transferase and related proteins. This raised the possibility that *NMV_0885* may code for a phosphoglycerol transferase in Nm and that this putative phosphoglycerol transferase may be responsible for the PG modification observed on PilE. The *NMV_0885* gene was given the putative designation phosphoglycerol transferase B (*pptB*).

In order to investigate this hypothesis a *pptB* deletion mutant (*ΔpptB*) was created. Mass profiling of purified PilE expressed by this mutant was devoid of PG (results in article inserted into this chapter). A further mutant was therefore created where the *pptB* gene was deleted, a copy inserted ectopically and placed under the control of an isopropyl β-D-1-thiogalactopyranoside (IPTG) inducible *lac* promoter ($pptB_{ind}$). This allowed the induction of various expression levels of *NMV_0885 in vitro* depending on the quantity of IPTG added to the culture medium. The appropriate concentration of IPTG was calculated to mimic the endogenous up-regulated *pptB* expression level that was reached several hours after host cell contact.

The effect of *pptB* up-regulation on pilin was determined by mass profiling of PilE purified from a Nm 8013 $pptB_{ind}^{+}$ construct ($^{+}$ indicates the addition of IPTG) and comparison with both the wild type and a $pptB_{ind}^{-}$ control (Figure 43). Note that these mass profiling experiments are performed using nano-ESI Q-ToF MS and the masses indicated are average protein masses (M_{av}) and have an associated error of ± 1 Da.

The mass profiles of PilE purified from the WT and $pptB_{ind}^{+}$ mutant are clearly different. The wild type profile shows two peaks separated by 154 Da. These correspond to the previously described major and minor proteoforms of PilE modified by one and two PG groups respectively. In the $pptB_{ind}^{+}$ mutant an additional peak representing a third proteoform appears at a mass corresponding to PilE+3PG. The PilE+2PG proteoform also appears to be much more abundant. This suggests that increased expression of PptB induces additional modification of PilE with PG. Furthermore when the $pptB_{ind}$ mutant is cultured in the absence of IPTG ($pptB_{ind}^{-}$) a peak corresponding to the mass of PilE devoid of PG modification appears. Together these observations suggest *that the NMV_0885* gene does indeed code for a phosphoglycerol transferase and that PptB is responsible for modification of PilE with PG.

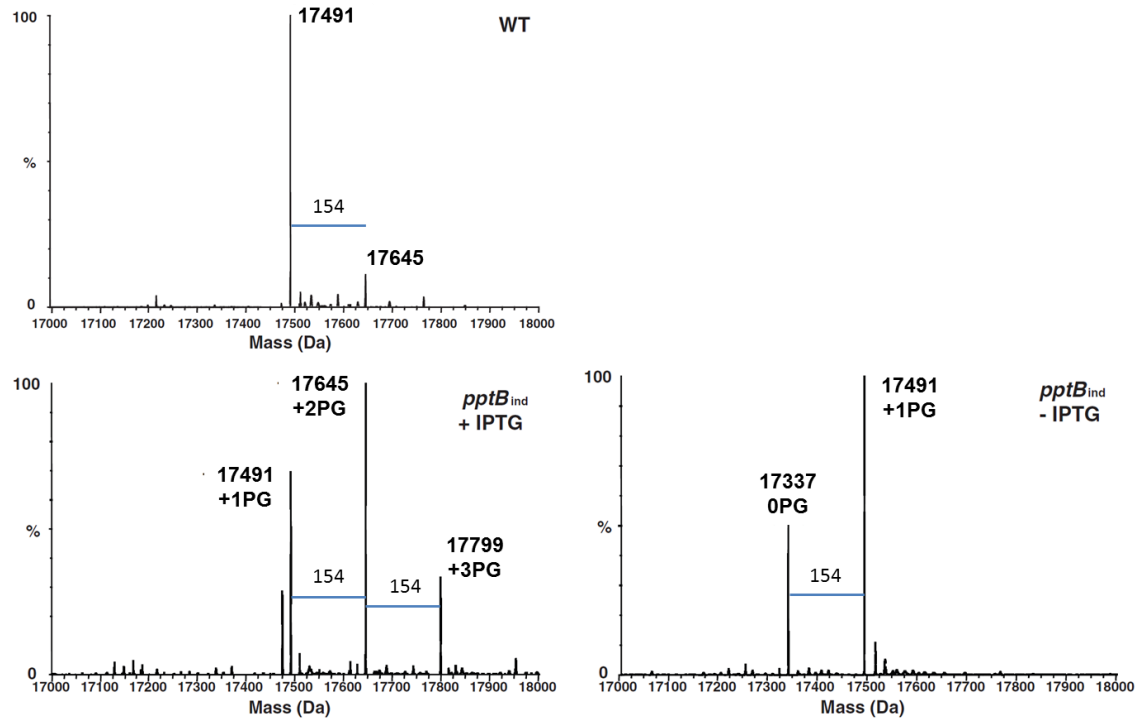


Figure 43 - nano-ESI Q-ToF mass profiling of PilE purified from wild type Nm 8013 (top) *pptB*$_{ind}^{+}$ (bottom left) and Δ*pptB*$_{ind}^{-}$ (bottom right). Spectra modified from Chamot-Rooke *et al.*[4]

1. Application of the Bottom-Up Approach to Map Phosphoglycerol Sites in the *pptB*$_{ind}^{+}$ Mutant

It was now desirable to completely characterise the three proteoforms expressed by the *pptB*$_{ind}^{+}$ construct. In a similar fashion to the characterisation of the wild type, the previously developed bottom-up methodology was applied. Tryptic digestion was performed on purified PilE and the digest spectrum manually examined for unexpected, multiply charged ions in an effort to identify all modified peptides present (Figure 44). Any such features were subjected to MS/MS for sequencing and PTM identification.

Localisation of the GATDH glycan on Ser^{63} and of one PG modification on Ser^{69} was easily achieved from fragmentation of the [45-73]+GATDH+PG digest product (results not shown). Similarly fragmentation of ions corresponding to [76-101]+PG enabled facile localisation of a PG group to Ser^{93}. However localisation of the third PG proved much more difficult to realise. Very careful examination of the digest revealed a triply charged ion at *m/z* 1269.616 corresponding to the [45-75] fragment plus one GATDH and two PG groups (Figure 44 inset). This very low intensity ion was selected and subjected to CAD over a long time frame (over 15 minutes) in order to sum enough scans to enable spectral interpretation.

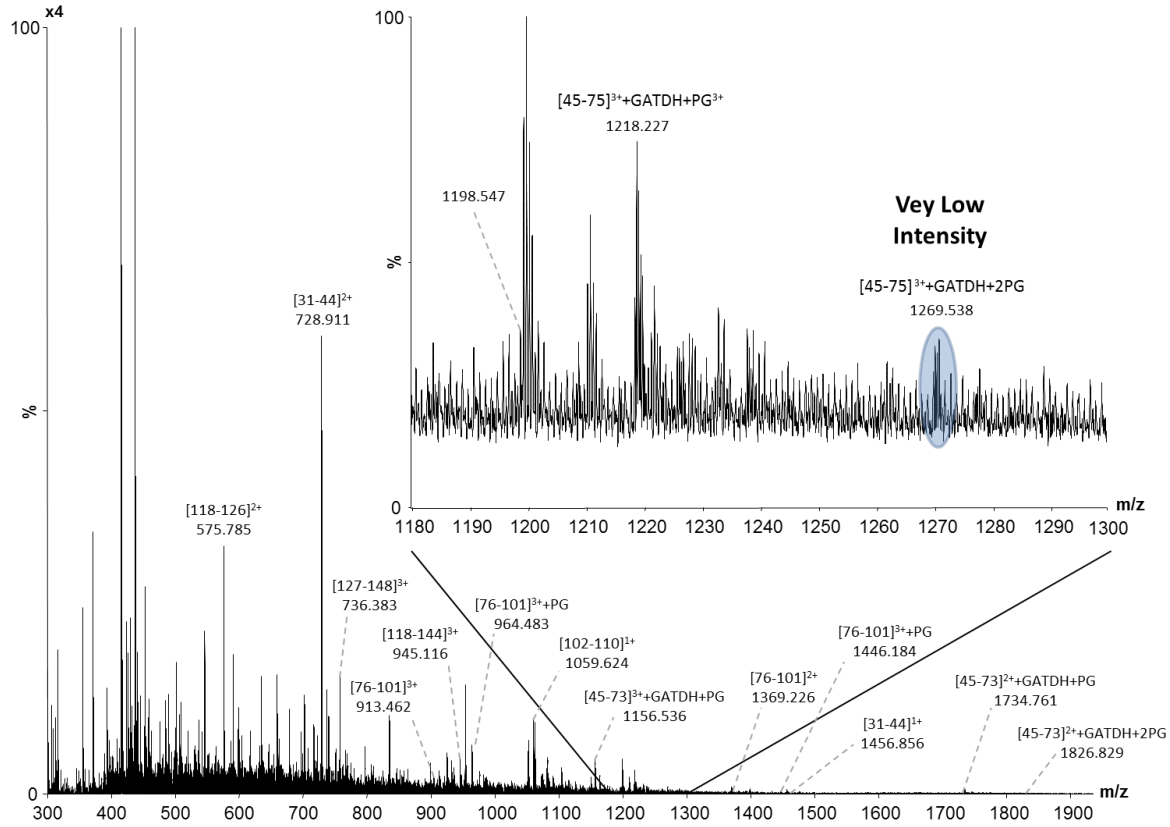


Figure 44 - nano-ESI Q-ToF MS of a tryptic digest of PilE from the *pptB*$_{ind}$$^{+}$ mutant. The very low intensity [45-75]$^{3+}$+GATDH+2PG ion is highlighted in the inset

The resultant MS/MS spectrum shown in Figure 45 is dominated by a large GATDH oxonium ion at *m/z* 275.132 and its dehydrated partner at *m/z* 257.122. In addition, triply charged ions corresponding to the molecular ion with loss of GATDH, 2PG-H_2O, GATDH-2PG-H_2O and GATDH-2PG-2H_2O are present at *m/z* 1178.712, 1160.897, 1069.536 and 1063.478, along with a similar doubly charged series corresponding to the molecular ion with loss of GATDH, GATDH-2PG-H_2O and GATDH-2PG-2H_2O at *m/z* 1766.818, 1603.766 and 1594.751. These ion series helped confirm the presence of two PG groups on this peptide.

Daughter ions from fragmentation of the backbone are rather low in intensity but nevertheless discernible from the baseline. An unmodified *y* type ion series can be assigned from y_2-y_6. Thereafter, both Ser^{69} and Thr^{68} must be modified with PG in order for additional *y* type ions y_8 1127.523, y_9 1198.490 and y_{11} 1354.624 to be assigned. This confirmed Thr^{68} as an additional PG modification site.

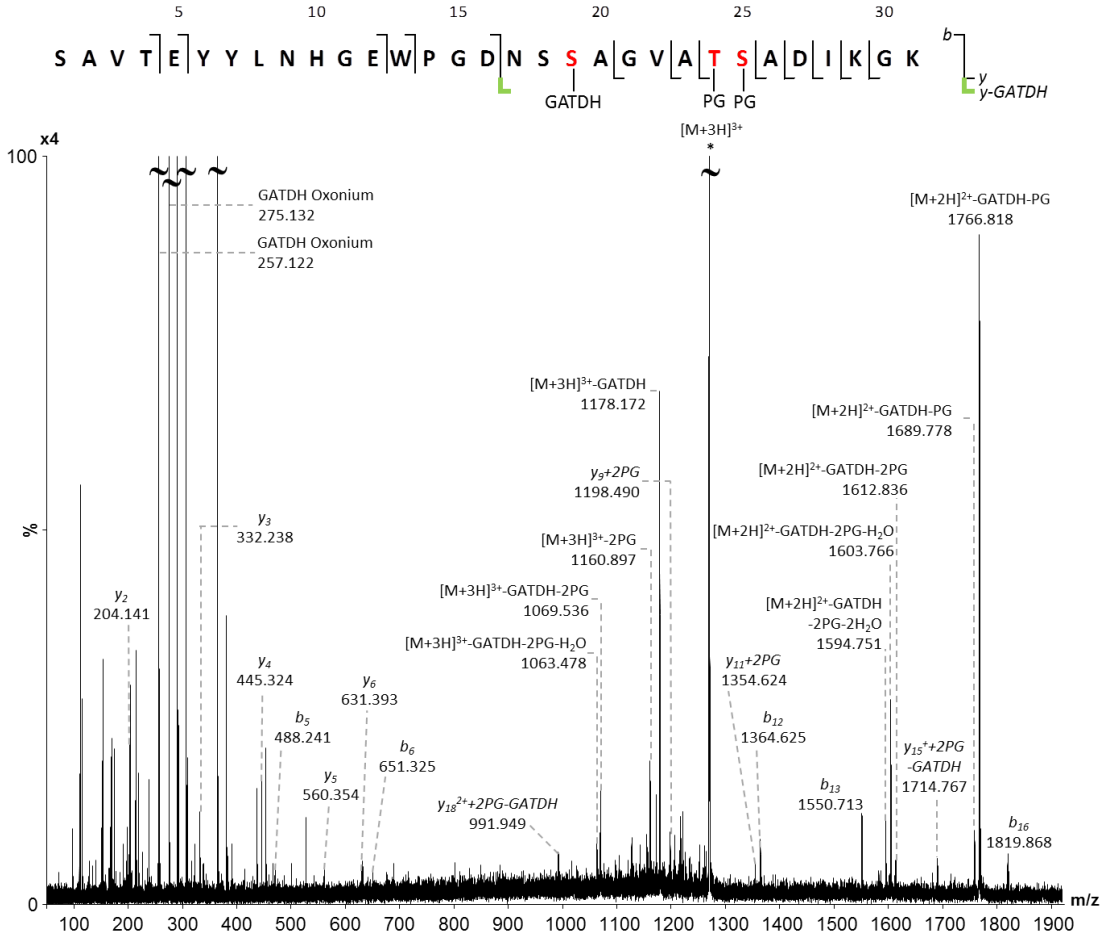


Figure 45 - nano-ESI Q-ToF MS/MS spectrum of the [45-75]$^{3+}$+GATDH+2PG peptide. The parent ion is marked with an asterisk and a fragmentation map is provided above the spectrum

The final step in the analysis requires the identified PG modifications to be related to the three proteoforms visualised in the mass profiling experiment. This can be related to having multiple identically shaped jigsaw puzzles (peptides) that need to be placed in the correct positions to reveal the correct picture (relating PTMs to parent proteoforms).

In total, three separate PG sites Ser^{69}, Ser^{93} and Thr^{68} were identified from examination of the tryptic digest products. As in the wild type the [76-101] peptide was identified principally in a naked (unmodified) form. It was also identified at lower abundance modified by one PG group. No ions corresponding to the mass of [76-101]+2PG were detected. Highly abundant ions corresponding to the [45-73] and [45-73] fragments both modified by one GATDH and one PG were easily identified. In contrast the [45-75]+2PG ion was extremely low in intensity and ions corresponding to [45-73]+2PG hardly visible at all. The digest spectrum was checked for other fragment ions with partners at ≈+154 Da but none were found.

Using the mass profiling data as a reference, only one combination of peptides is possible to account for the mass of the lowest observed PilE proteoform. This combination places a GATDH on Ser^{63} and PG on Ser^{69}. Similarly for the highest mass proteoform only one peptide combination is possible to reach the required mass, and assuming we have identified all modified peptides in the digest, this comprises of a GATDH on Ser^{63} and PG groups on Ser^{69}, Ser^{93} and Thr^{68}. However for the middle proteoform of PilE there are two different combinations of peptides that can be used to achieve the observed protein mass. In addition to the GATDH at Ser^{63} and PG at Ser^{69} the second PG site may therefore be modified by PG at either Thr^{68} or Ser^{93}. A simplified schematic of this combinatorial problem is given in Figure 46.

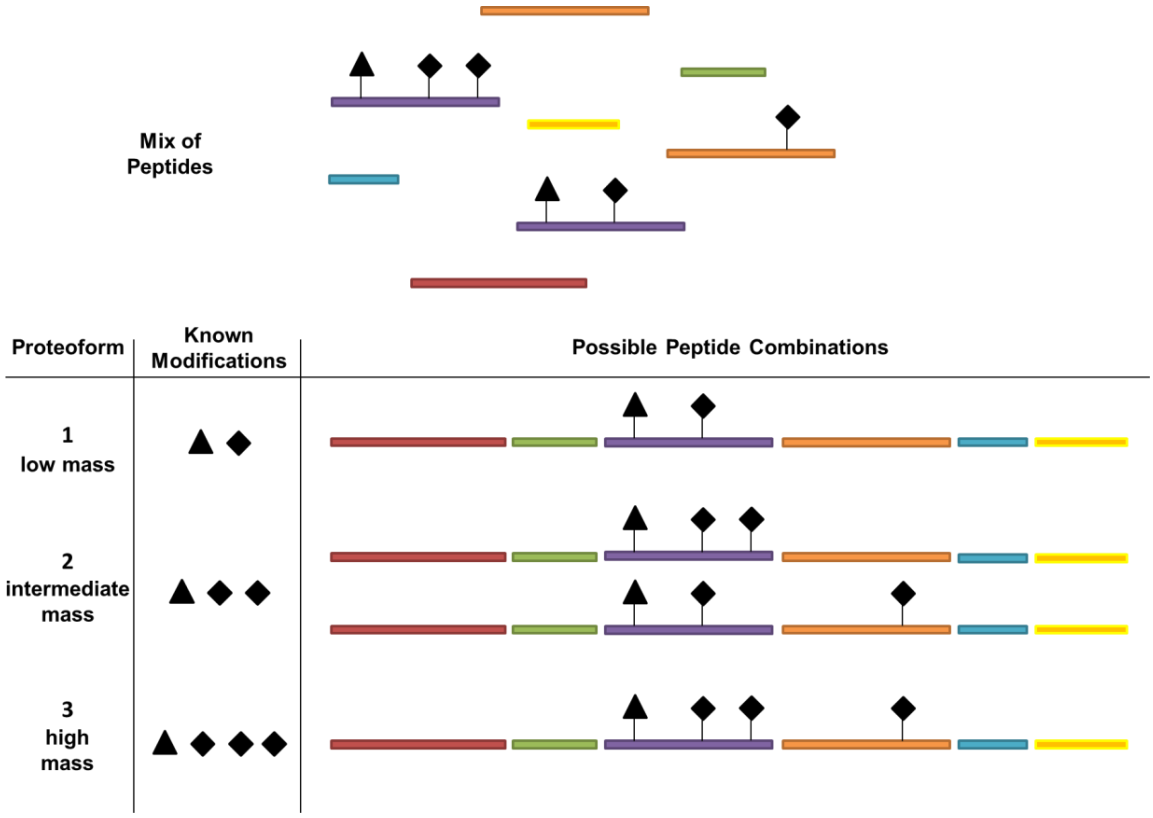


Figure 46 - Simplified schematic of peptide matching problem. For the intermediate mass proteoform there are two possible combinations of peptides that achieve the protein mass measured in the mass profiling experiment

In the case at hand we would be fairly confident in assigning the second PG group to Ser^{93} based on a combination of prior knowledge (Ser^{93} is the minor PG site in the WT) and the much greater ion abundance of the [76-101]+PG peptide compared with [45-75]+GATDH+2PG. It should be noted that reasoning based on this latter assumption may not always hold, as PTM is known to alter ionisation efficiency and therefore affect ion abundance. For this reason additional validation of the proposed modification sites was required.

Since Ser^{93} was thought to be the principal modification site in *pptB*$_{ind}^{+}$ a further derivative was created containing serine to alanine point mutation at Ser^{93} (*pptB*$_{ind}^{+}$ S39A). The mass profile of this mutant is shown in Figure 2 panel D of the article inserted later in this chapter. The intensities of the peaks in this S93A mutant confirm that Ser^{93} is the second site of PG modification. No point mutation was performed on Thr^{68}. These results place PG groups at Ser^{69} in the lowest mass proteoform, Ser^{69} and Ser^{93} in the intermediate mass form and Ser^{69}, Ser^{93} and Thr^{68} in the highest mass proteoform and completely resolve the PTM population of Nm 8013 *pptB*$_{ind}^{+}$.

2. Understanding the Biological Role of Phosphoglycerol Modification

To summarise the results thus far, it is supposed that the NMV_0885 gene codes for a phosphoglycerol transferase PptB and that this enzyme is responsible for modification of PilE with PG. Several hours after host contact PptB expression is increased and this results in modification of PilE with additional PG groups at Ser^{93} and to a lesser extent at Thr^{68}. In an effort to understand the biological function of this increased PG modification, the group of Guillaume Duménil performed various phenotypic assays with the *pptB*$_{ind}^{+}$ mutant. Intriguingly bacterial aggregation assays showed distinctly lower levels of aggregation when performed with *pptB*$_{ind}^{+}$ compared to the wild type or *pptB*$_{ind}^{-}$ mutant (results in the publication that follows).

Bacterial aggregation is known to be mediated by bundling interactions between pili. Examination of pili by negative staining electron microscopy showed a reduced degree of the bundling phenotype in the *pptB*$_{ind}^{+}$ mutant compared with the WT and *pptB*$_{ind}^{-}$ control. This suggested that when modified with PG at Ser^{93} the pilus bundling was severely impaired. Molecular modelling of the entire pilus fibres by the group of Prof. Michael Nilges at the Institut Pasteur provided a molecular basis for these observations.

Ser^{93} lies exposed on the pilus surface surrounded by a number of positively charged lysine residues. Upon modification of the Ser^{93} with PG, a negative charge is introduced in the centre of this patch. Calculations of the interaction potential between neighbouring fibres in a bundle demonstrate that once modified by PG at Ser^{93} the interaction energy becomes negligible and no longer favours pilus bundling. These results show that several hours after host cell contact PptB expression increases and this modifies PilE with increased levels of PG predominately at Ser^{93}. This weakens pilus-pilus interactions and promotes pilus unbundling allowing bacterial dissemination. This hypothesis is presented in full in the following paper.

3. Published article - "Posttranslational Modification of Pili upon Cell Contact Triggers *N. meningitidis* Dissemination"

M5313-derived and h-MCL induced higher IFN-β secretion than *L.g.*M− parasites or h-CL parasites (Fig. 2C). Furthermore, this expression was TLR3-TRIF dependent, with the MyD88 signaling pathway augmenting secretion (Fig. 2C).

Endosomal TLRs recognize nucleic acid motifs, with TLR7 and TLR3 recognizing single-stranded RNA (ssRNA) and double-stranded RNA (dsRNA), respectively (*23*). Our experimental evidence suggested that nucleic acid–derived motifs were involved in the host macrophage response to infection with metastasizing *L.g.* parasites. We observed increased production of CCL5, TNF-α, and IL-6 in macrophages exposed to single-stranded ribonuclease (ssRNAse)– and deoxyribonuclease (DNAse)–treated nucleic acids derived from *L.g.*M+ parasites, compared with *L.g.*M− and *L.major* LV39 (fig. S2). Although not statistically significant, these results suggested that the nucleic acid motif is resistant to ssRNAse and DNAse treatments and is likely to be dsRNA.

L.Viannia parasites, including *L.g.*M5313(M+) and *L. guyanensis* and *L. braziliensis* MCL human isolates, harbor the dsRNA *Leishmania RNA virus 1* (LRV1) (*24*–*26*). These viruses have a capsid coat protecting a 5.3-kb dsRNA genome (*27*). Metastasizing promastigotes had greater levels of LRV1 (*L.g.*M+ or h-MCL-LRV^{high}) than nonmetastasizing promastigotes (*L.g.*M− or h-CL-LRV^{low}) as shown by the presence of a ~5.3-kb, DNAse-insensitive, RNAse III–sensitive band in agarose gels, and LRV1 quantification by quantitative reverse transcriptase–polymerase chain reaction (qRT-PCR) (Fig. 3, A to C, and fig. S3A). We thus verified that macrophages treated with purified LRV1 dsRNA (fig. S3) induced a phenotype similar to that of macrophage infected with metastasizing parasites, and as shown by an increased expression of CXCL10, CCL5, TNF-α, IL-6, and IFN-β transcripts, this increase was TLR3 dependent (Fig. 3D). Because the *L.g.*M5313 M+ and M− parasites were not isogenic, we performed new experiments with parasites derived from the WHO reference strain *L.g.*M4147 that metastatizes in the hamster (*28*) and carries the LRV1-4 virus (*29*). Macrophage infection with *L.g.*M4147-LRV^{high} parasites produced significantly greater amounts of cytokines and chemokines than infection with its respective isogenic virus-free derivative *L.g.*M4147LRV^{neg}, in a TLR3-dependent manner (Fig. 3E and fig. S4) (*30*, *31*). Similar parasite burdens were observed for all parasites infected into the wild-type and the TLR-, TRIF-, and MyD88-deficient macrophages (table S1).

A role for TLR3 and LRV1 in leishmaniasis development was analyzed in vivo, with TLR3$^{-/-}$, TLR7$^{-/-}$, and WT mice that were infected in the footpad. A significant decrease in footpad swelling, and diminished parasite burden, were observed in TLR3$^{-/-}$ mice infected with *L.g.*M+LRV^{high} (M5313) or *L.g.*M4147−LRV^{high} parasites compared with wild-type mice (Fig. 4 and fig. S5). No consistent, significant decrease in disease pathology was observed between TLR3$^{-/-}$ and wild-type mice infected with *L.g.*M−LRV^{low} (Lg17) or *L.g.*M4147−LRV^{neg} or between TLR7$^{-/-}$ and wild-type infected mice with the different parasite isolates (Fig. 4 and Fig. S5). Further experimentation is required to elucidate the role of TLR7-dependent immune responses with respect to infection with LRV1-containing *Leishmania* parasites.

Our work showed that recognition of LRV1 within metastasizing *L.g.* parasites by the host promoted inflammation and subverted the immune response to infection to promote parasite persistence (*2*, *3*, *32*). Because recognition of LRV1 within the metastasizing *L.g.* parasites arises early after infection, we hypothesize that LRV1 dsRNA is released from dead parasites, unable to survive within the host macrophage. These results could open the door to better diagnosis of risk for MCL disease and facilitate the development of new and more efficient treatment regimes.

References and Notes

1. K. Weigle, N. G. Saravia, *Clin. Dermatol.* **14**, 433 (1996).
2. C. Vergel *et al.*, *J. Infect. Dis.* **194**, 503 (2006).
3. J. E. Martinez, L. Alba, *Trans. R. Soc. Trop. Med. Hyg.* **86**, 392 (1992).
4. A. Barral *et al.*, *Am. J. Pathol.* **147**, 947 (1995).
5. V. S. Amato, F. F. Tuon, H. A. Bacha, V. A. Neto, A. C. Nicodemo, *Acta Trop.* **105**, 1 (2008).
6. J. Arevalo *et al.*, *J. Infect. Dis.* **195**, 1846 (2007).
7. D. R. Faria *et al.*, *Infect. Immun.* **73**, 7853 (2005).
8. S. T. Gaze *et al.*, *Scand. J. Immunol.* **63**, 70 (2006).
9. C. Pirmez *et al.*, *J. Clin. Invest.* **91**, 1390 (1993).
10. D. A. Vargas-Inchaustegui *et al.*, *Infect. Immun.* **78**, 301 (2010).
11. J. M. Blackwell, *Parasitol. Today* **15**, 73 (1999).
12. L. Castellucci *et al.*, *J. Infect. Dis.* **194**, 519 (2006).
13. B. Travi, J. Rey-Ladino, N. G. Saravia, *J. Parasitol.* **74**, 1059 (1988).
14. J. E. Martínez, L. Valderrama, V. Gama, D. A. Leiby, N. G. Saravia, *J. Parasitol.* **86**, 792 (2000).
15. N. Acestor *et al.*, *J. Infect. Dis.* **194**, 1160 (2006).
16. Materials and methods are available as supporting material on *Science* Online.
17. C. Bogdan, M. Röllinghoff, A. Diefenbach, *Immunol. Rev.* **173**, 17 (2000).
18. F. H. Abou Fakher, N. Rachinel, M. Klimczak, J. Louis, N. Doyen, *J. Immunol.* **182**, 1386 (2009).
19. R. Ben-Othman, L. Guizani-Tabbane, K. Dellagi, *Mol. Immunol.* **45**, 3222 (2008).
20. K. A. Cavassani *et al.*, *J. Exp. Med.* **205**, 2609 (2008).
21. R. Le Goffic *et al.*, *PLoS Pathog.* **2**, e53 (2006).
22. K. S. Lang *et al.*, *J. Clin. Invest.* **116**, 2456 (2006).
23. S. L. Doyle, L. A. O'Neill, *Biochem. Pharmacol.* **72**, 1102 (2006).
24. L. Guilbride, P. J. Myler, K. Stuart, *Mol. Biochem. Parasitol.* **54**, 101 (1992).
25. G. Salinas, M. Zamora, K. Stuart, N. Saravia, *Am. J. Trop. Med. Hyg.* **54**, 425 (1996).
26. M. M. Ogg *et al.*, *Am. J. Trop. Med. Hyg.* **69**, 309 (2003).
27. T. L. Cadd, M. C. Keenan, J. L. Patterson, *J. Virol.* **67**, 5647 (1993).
28. J. A. Rey, B. L. Travi, A. Z. Valencia, N. G. Saravia, *Am. J. Trop. Med. Hyg.* **43**, 623 (1990).
29. G. Widmer, A. M. Comeau, D. B. Furlong, D. F. Wirth, J. L. Patterson, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 5979 (1989).
30. L. F. Lye *et al.*, *PLoS Pathog.* **6**, e1001161 (2010).
31. Y. T. Ro, S. M. Scheffter, J. L. Patterson, *J. Virol.* **71**, 8991 (1997).
32. A. Barral *et al.*, *Am. J. Trop. Med. Hyg.* **53**, 256 (1995).

33. We are grateful to N. Saravia (CIDEIM, Colombia) and Instituto Oswaldo Cruz, for *L. guyanensis* strains; S. Akira (Frontier Research center, Osaka University), P. Romero (LICR, Lausanne), and B. Ryffel (CNRS, Orléans) for knockout and mutant mice; M. Delorenzi (SIB, Lausanne) for bioinformatics expertise; F. Morgenthaler (Cellular Imaging Facility, Lausanne), S. Cawsey, and M.-A. Hartley for technical assistance; and J. Patterson and Y. T. Ro for the *L.g.*M4147 strains. This work was funded by FNRS grants 3100A0-116665/1 (N.F.) and 310030-120325 (P.L.), Foundation Pierre Mercier (S.M.), and NIH A129646 (S.M.B.). Microarray data are available within the Gene Expression Omnibus database (GSE21418) and at http://people.unil.ch/nicolasfasel/data-from-fasels-lab/.

Supporting Online Material

www.sciencemag.org/cgi/content/full/331/6018/775/DC1
Materials and Methods
Figs. S1 to S4
Table S1
References

20 October 2010; accepted 23 December 2010
10.1126/science.1199326

Posttranslational Modification of Pili upon Cell Contact Triggers *N. meningitidis* Dissemination

Julia Chamot-Rooke,[1,2] Guillain Mikaty,[3,4] Christian Malosse,[1,2] Magali Soyer,[4,5] Audrey Dumont,[4,5] Joseph Gault,[1,2] Anne-Flore Imhaus,[4,5] Patricia Martin,[3,4] Mikael Trellet,[6] Guilhem Clary,[4,7,8] Philippe Chafey,[4,7,8] Luc Camoin,[4,7,8] Michael Nilges,[6] Xavier Nassif,[3,4,9] Guillaume Duménil[4,5]*

The Gram-negative bacterium *Neisseria meningitidis* asymptomatically colonizes the throat of 10 to 30% of the human population, but throat colonization can also act as the port of entry to the blood (septicemia) and then the brain (meningitis). Colonization is mediated by filamentous organelles referred to as type IV pili, which allow the formation of bacterial aggregates associated with host cells. We found that proliferation of *N. meningitidis* in contact with host cells increased the transcription of a bacterial gene encoding a transferase that adds phosphoglycerol onto type IV pili. This unusual posttranslational modification specifically released type IV pili-dependent contacts between bacteria. In turn, this regulated detachment process allowed propagation of the bacterium to new colonization sites and also migration across the epithelium, a prerequisite for dissemination and invasive disease.

The Gram-negative bacterium *Neisseria meningitidis* is a leading cause of septicemia and meningitis in humans (*1*). Initially, individual bacteria adhere to the nasopharynx epithelium via their type IV pili, a filamentous organelle common to numerous pathogenic bac-

terial species (*2*). In the following hours, bacteria proliferate on the cellular surface in tight three-dimensional aggregates termed microcolonies. The formation of these aggregates results from homotypic, type IV pili–mediated contacts between the bacteria themselves and contacts between bacteria and the host cell plasma membrane. Contacts with host cells are enhanced by the formation of bacteria-induced plasma membrane protrusions (*3*). After this proliferation phase, individual bacteria are thought to detach from the microcolonies, leading to propagation to new hosts and dissemination throughout the body in case of invasive infection. Understanding the molecular mechanisms underlying the life cycle of *N. meningitidis* is a key step toward identification of prevention and treatment strategies of meningococcemia. The major component of *Neisseria* spp. type IV pili (PilE or pilin) is modified with phosphocholine (PC), phosphoethanolamine (PE), or phosphoglycerol (PG) (*4*–*6*). We wanted to determine the impact of these unusual posttranslational modifications (PTM) on the pathogenesis of *N. meningitidis*.

A whole-protein mass spectrometry approach was chosen to determine the phosphorylation state of type IV pili (*7*, *8*). Analysis of purified pili from the well-characterized 8013 strain (*9*) grown on solid medium (*10*) yielded a main peak with a mass of 17,491 daltons and a minor secondary peak with a mass of 17,645 daltons (Fig. 1A) corresponding to the addition of one PG (154 daltons, fig. S1A). Analysis of purified pili from strains carrying point mutations substituting conserved serine residues 69 and 93 of the PilE protein with alanines (Fig. 1, B and C) showed that all pilin subunits were modified with PG on Ser69 (17,491 daltons), whereas only about 15% of pilin subunits were also modified on Ser^{93} (17,645 dalton). The *NMV_0885* gene (ortholog of *NMA1705* and *NMB1508*) was a good candidate to carry out this activity because it is part of the cluster of orthologous group (COG) titled phosphoglycerol transferase and related proteins (COG1368, fig. S1C) (*11*). Analysis of type IV pili purified from a strain carrying a deletion in the *NMV_0885* gene revealed a single peak of 17,337 daltons that corresponds to pilin without any PG (Fig. 1D), demonstrating that this gene is responsible for the transfer of PG onto the pilin. We thus named the transferase PptB (pilin phosphotransferase B).

The *pptB* gene was previously described as a member of a group of 16 *N. meningitidis* genes containing a two-component system-regulated promoter referred to as CREN for contact regulatory element of *Neisseria* (*12*–*14*). Transcription of *pptB* increased two- to threefold over a period of 4 hours after adhesion to epithelial cells (*12*), suggesting that modification of type IV pili with PG could be triggered upon contact with host cells. To test this possibility, we expressed the *pptB* gene under the transcriptional control of the isopropyl-β-D-thiogalactopyranoside (IPTG)–inducible *lac* promoter to mimic the threefold induction found on host cells (Fig. 2A). In the presence of inducer, the peak corresponding to two PG modifications (17,645 daltons) became the most abundant form (Fig. 2, B and C, and fig. S2). Substitution of Ser^{93} into an alanine (S93A) (Fig. 2D) indicated that Ser^{93} is the main phosphorylation site upon induction of *pptB*, whereas Ser^{69} phosphorylation level remained constant. The expression level of the *pptB* gene thus determines the phosphorylation level of the pilin subunits on Ser^{93}.

We used two-dimensional gel electrophoresis followed by immunoblot to demonstrate increased modification of pilin with the negatively charged PG while bacteria were proliferating in contact with host cells. Analysis of PilE from bacteria growing in suspension displayed a major spot with an isoelectric point around 6 (fig. S3). Upon incubation with host cells, spots corresponding to acidic forms appeared after 2 and 4 hours. Pilin isoelectric point changes did not occur in the *pilES93A*-expressing strain. Thus, after *pptB* transcriptional increase upon contact with host cells, the isoelectric point of a significant proportion of the major pilin subunit became more acidic after modification of Ser^{93}.

To gain insight into the potential impact of this modification, we modeled the three-dimensional structure of the pilin from *N. meningitidis* with and without PG on Ser^{93} based on the known *N. gonorrhoeae* pilin structure and energy minimization in the context of the pilus fiber (Fig. 2E). Ser^{93} is surrounded by five lysine residues (Fig. 2F), a positively charged patch postulated to be important in type IV pili function (*15*). Modification of Ser^{93} with PG introduces a

[1]Ecole Polytechnique, Laboratoire des Mécanismes Réactionnels, Palaiseau F-91128, France. [2]CNRS, UMR7651, Palaiseau F-91128, France. [3]INSERM, U1002, Paris F-75015, France. [4]Université Paris Descartes, Faculté de Médecine Paris Descartes, Paris F-75006, France. [5]INSERM, U970, Paris Cardiovascular Research Center, Paris F-75015, France. [6]Unité de Bioinformatique Structurale, Unité de Recherche Associée CNRS 2185, Institut Pasteur, Paris F-75015, France. [7]Institut Cochin, CNRS (UMR 8104), Paris F-75014, France. [8]INSERM, U1016, Paris F-75014, France. [9]Assistance Publique–Hôpitaux de Paris, Hôpital Necker–Enfants Malades, Paris F-75015, France.

*To whom correspondance should be addressed. E-mail: guillaume.dumenil@inserm.fr

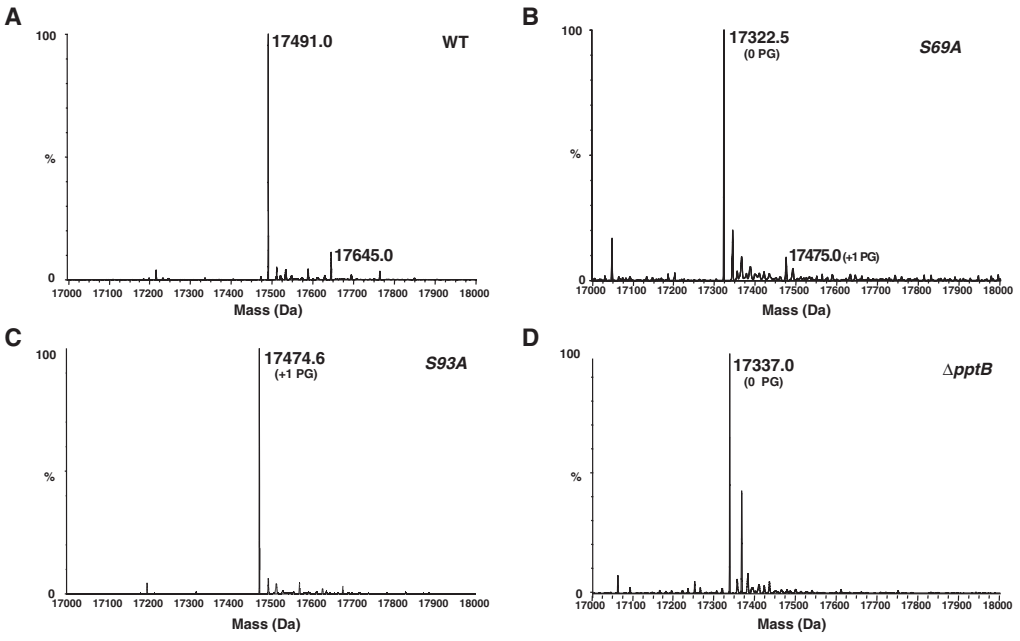


Fig. 1. Modification of the pilin subunit with PG by the PptB transferase. Whole-protein mass spectrometry analysis of type IV pili purified from the wild-type strain (WT) (**A**), from a strain harboring a S69A mutation in the *pilE* gene (**B**), from a strain harboring a S93A mutation in the *pilE* gene (**C**), and from a mutant in the *NMV_885* gene (Δ*pptB*) (**D**).

negative charge carried by the PG group protruding from the pilus structure (Fig. 2G and fig. S4). Modeling of bundles of four antiparallel pili fibers showed that the average interaction energy was largely favorable (60 kcal difference) for the pili with no modification on Ser^{93} when compared with the same structure with a PG on Ser^{93} (Fig. 3, A to C). Thus, modification of Ser^{93} with PG would strongly destabilize fiber interaction, suggesting that this PTM could have important consequences for type IV pili function.

We addressed the functional role of this PTM in key steps of *N. meningitidis* pathogenesis. Pilin modification with PG did not have any effect on the amount of type IV pili present on the bacterial surface (fig. S5, A and B). In contrast, as predicted by molecular modeling, ultrastructural analysis by negative staining showed that increased pilin modification with PG blocked the formation of bundles of pili (Fig. 3D). In the wild-type strain, pili bundles were commonly about 30 nm wide, thus containing several 6- to 8-nm-wide individual fibers (Fig. 3D). Upon transcriptional induction of the *pptB* gene, only thin fibers having the expected size of individual pili could be found (Fig. 3D and fig. S5C). Increased modification of pilin with PG thus blocks bundle formation.

Type IV pili bundle formation and *N. meningitidis* aggregation are linked (*16*). *pptB* gene deletion or the S93A substitution led to increased aggregate formation in suspension (Fig. 3E) and consistently increased transcription of the *pptB* gene abrogated bacterial aggregation (Fig. 3F). The effect of increased *pptB* transcription on aggregation was rescued by the S93A point mutation (Fig. 3G). Increased modification of Ser^{93} with PG thus strongly reduces type IV pili–dependent bacterial aggregation by introducing a negative charge at this site (fig. S6B). This anti-aggregative effect appeared to overcome the pro-aggregative activity of the minor pilin PilX (fig. S6, C and D).

The effect of pilin modification with PG on adhesion to epithelial cells was evaluated. The first step of adhesion, which is the contact of individual bacteria with the cell surface (Fig. 4A), occurred independently of the level of glycerophosphorylation (Fig. 4B). Bacterial microcolonies formed by the strains affected in pilin modification with PG did not appear morphologically different from the wild-type multilayered microcolonies (Fig. 4C). Furthermore, pilin glycerophosphorylation had little effect on the total number of cell-associated bacteria after 6 hours (Fig. 4D). To investigate the effect of

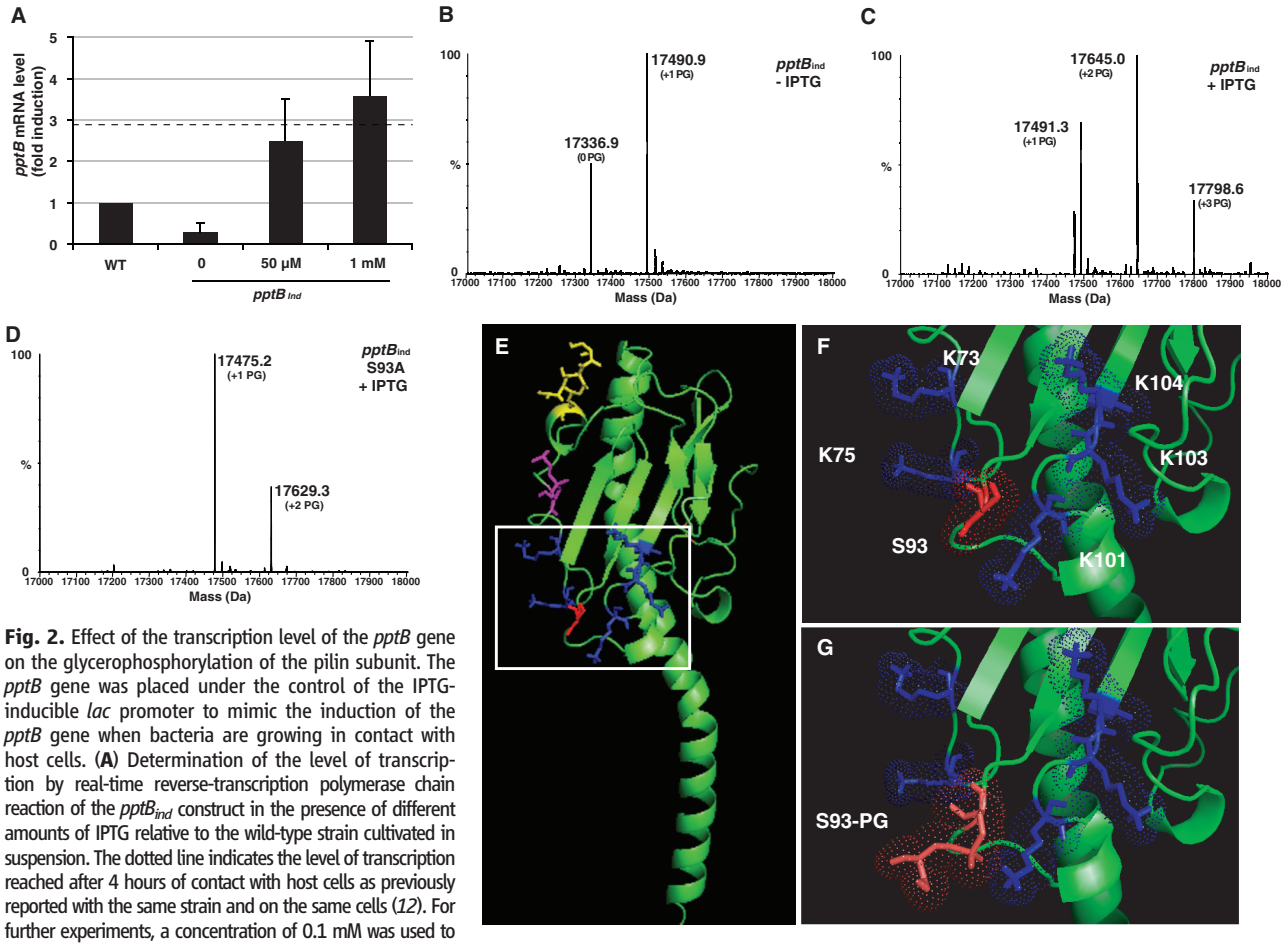


Fig. 2. Effect of the transcription level of the *pptB* gene on the glycerophosphorylation of the pilin subunit. The *pptB* gene was placed under the control of the IPTG-inducible *lac* promoter to mimic the induction of the *pptB* gene when bacteria are growing in contact with host cells. (**A**) Determination of the level of transcription by real-time reverse-transcription polymerase chain reaction of the $pptB_{ind}$ construct in the presence of different amounts of IPTG relative to the wild-type strain cultivated in suspension. The dotted line indicates the level of transcription reached after 4 hours of contact with host cells as previously reported with the same strain and on the same cells (*12*). For further experiments, a concentration of 0.1 mM was used to mimic the effect of cell contact. The graph represents the mean ± SEM of three independent experiments. (**B**) Whole-protein mass spectrometry analysis of the pilin subunit purified from the $pptB_{ind}$ strain in absence of IPTG and therefore with low amounts of PptB ($pptB_{ind}$ – IPTG). (**C**) Effect of increased *pptB* level after induction of the $pptB_{ind}$ construct with 0.1 mM IPTG on pilin mass ($pptB_{ind}$ + IPTG). (**D**) Analysis of the pilin subunit containing the S93A substitution in the $pptB_{ind}$ strain in presence of 0.1 mM IPTG ($pptB_{ind}$ S93A + IPTG). (**E**) Molecular modeling of the *N. meningitidis* pilin monomer with the PTMs found when bacteria are grown in suspension: GATDH (glyceramido tri-deoxy hexose) sugar modification on Ser^{63} (yellow) and glycerophosphate on Ser^{69} (magenta). The structure is displayed with the Pymol software (www.pymol.org). (**F**) Positively charged environment of Ser^{93} on the structure of PilE corresponding to the area outlined in (E). A patch of five lysine (K) residues indicated in blue surround Ser^{93} in red. (**G**) Predicted structure of the pilin modified with PG on Ser^{93}.

increased pilin glycerophosphorylation on detachment, we determined the number of bacteria disengaging from microcolonies over time. A laminar flow chamber was used as a tool to progressively collect detaching bacteria (Fig. 4E and fig. S7A). Whereas the number of wild-type bacteria released from the infected monolayer slowly increased with time after 3 hours of infection, detachment of the Δ*pptB* mutant was impaired (1.5×10^6 versus 3.4×10^5 bacteria per ml at 7 hours). About 20 to 30% of bacteria adhering at 6 hours detached in the following hours of infection. The increase in pilin glycerophosphorylation thus favors the release of a proportion of individual bacteria from the microcolonies that has little impact on the number of adhering bacteria.

Bacteria released from microcolonies are more likely to cross the epithelium (*17*), and we tested the ability of the Δ*pptB* mutant to transmigrate across an epithelial cell monolayer (Fig. 4F and fig. S7). After 6 hours, the Δ*pptB* strain had crossed the monolayer 20-fold less efficiently than the wild-type strain, indicating a defect as strong as the nonpiliated *pilE* strain. The strain expressing a point mutation on Ser^{93} of the pilin exhibited a similar defect, showing that the effect of the *pptB* deletion was mediated by the

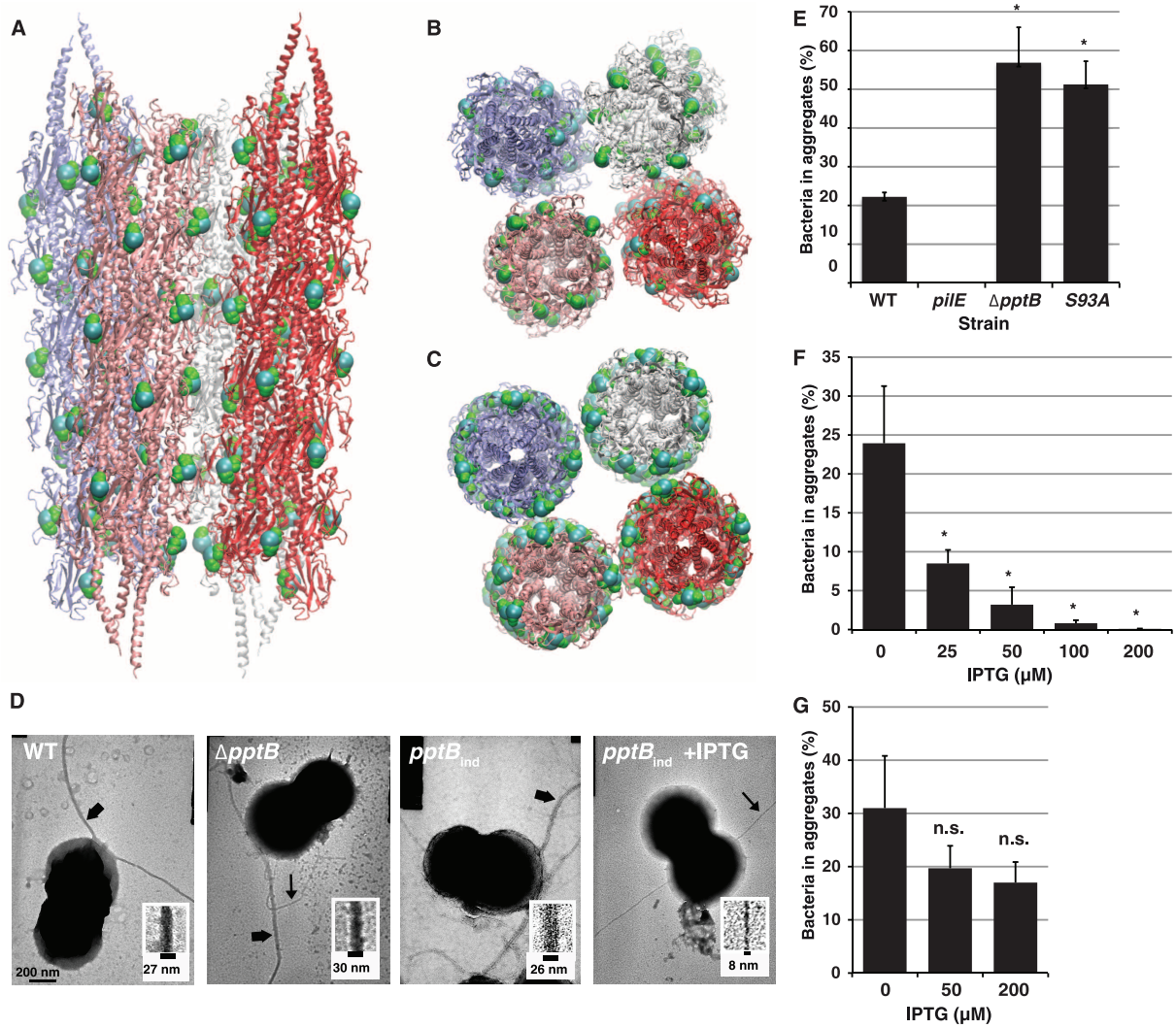


Fig. 3. Effect of pilin glycerophosphorylation on the aggregative functions of type IV pili. (**A** to **C**) Molecular modeling of pili fibers interacting with each other in a bundle. We modeled a regular antiparallel bundle involving four pili fibers by a combination of systematic and random search in the distance between pili, the rotation, and the tilt of the pilus from the overall bundle axis and analyzed them in terms of their interaction energy and the geometry. Each pilus fiber contains 20 monomers of the pilin subunit. (A) Side view of the bundle formed by pilin subunits displaying a PG only on Ser^{69}. The PG modifications are displayed as spheres, phosphate atoms in blue and carbon atoms in green. (B) Top view of the structure shown in (A), showing the close contacts between the fibers. (C) Top view of the unstable bundle that would be formed by the pilin modified with PG on both Ser^{69} and Ser^{93}. (**D**) Effect of increased pilin phosphorylation on the organization of pilus fibers analyzed by transmission electron microscopy for the WT strain, the Δ*pptB* strain, and the $pptB_{ind}$ strain with and without 0.1 mM IPTG. Wide arrows indicate thick bundles of fibers, and narrow arrows indicate thin fibers likely to be individual pili (6 to 8 nm). Insets represent a higher magnification of pili fibers. (**E** to **G**) The ability of bacteria to form aggregates was determined by a technique based on direct observation by fluorescence microscopy. The percentage of bacteria engaged in an aggregate relative to the total amount of bacteria in the same volume is indicated. Four independent experiments each done in triplicate were performed; error bars denote mean ± SEM. Statistical analyses were done with Student's *t* test; n.s., nonsignificant; $*P < 0.05$. (E) Aggregation was compared between the WT strain, the nonpiliated *pilE* strain, the Δ*pptB* strain, and the pilES93A strain. (F) Aggregation was determined in the $pptB_{ind}$ strain in the presence of different amounts of IPTG. (G) Effect on aggregation of increased *pptB* expression in a mutant expressing the PilE protein with an alanine on position 93 ($pptB_{ind}$*PilES93A*).

www.sciencemag.org/cgi/content/full/331/6018/778/DC1

Supporting Online Material for

Posttranslational Modification of Pili upon Cell Contact Triggers *N. meningitidis* Dissemination

Julia Chamot-Rooke, Guillain Mikaty, Christian Malosse, Magali Soyer, Audrey Dumont, Joseph Gault, Anne-Flore Imhaus, Patricia Martin, Mikael Trellet, Guilhem Clary, Philippe Chafey, Luc Camoin, Michael Nilges, Xavier Nassif, Guillaume Duménil*

*To whom correspondence should be addressed. E-mail: guillaume.dumenil@inserm.fr

Published 11 February 2011, *Science* **331**, 778 (2011)
DOI: 10.1126/science.1200729

This PDF file includes:

Supporting online material for

Cell-contact induced posttranslational modification of type IV pilin triggers *Neisseria meningitidis* dissemination

Julia Chamot-Rooke[1,2], Guillain Mikaty[3,4], Christian Malosse[1,2], Magali Soyer[4, 6], Audrey Dumont[4, 6], Joseph Gault[1,2], Anne-Flore Imhaus[4, 6], Patricia Martin[3,4], Mikael Trellet[5], Guilhem Clary[4, 7, 8], Philippe Chafey[4, 7, 8], Luc Camoin[4, 7, 8], Michael Nilges[5], Xavier Nassif[3, 4, 9], and Guillaume Duménil[4, 6, *]

[1] Ecole Polytechnique, Laboratoire des Mécanismes Réactionnels, Palaiseau, F-91128, France

[2] CNRS, UMR7651, Palaiseau, F-91128, France

[3] INSERM, U1002, Paris, F-75015 France

[4] Université Paris Descartes, Faculté de Médecine Paris Descartes, Paris, F-75006, France

[5] Unité de Bioinformatique Structurale, URA CNRS 2185, Institut Pasteur, Paris, F-75015 France

[6] INSERM, U970, Paris Cardiovascular Research Center, Paris, F-75015, France

[7] Institut Cochin, CNRS (UMR 8104), Paris, F-75014, France

[8] INSERM, U1016, Paris, F-75014, France.

[9] AP-HP, Hôpital Necker-enfants malades, Paris, F-75015, France

* To whom correspondence should be addressed. Email: guillaume.dumenil@inserm.fr.

This file includes:

Materials and Methods

Bacterial strains and mutagenesis

All *N. meningitidis* strains described in this study were derived from the recently sequenced 8013 serogroup C strain [http://www.genoscope.cns.fr/agc/nemesys] (*1*). *N. meningitidis* strains were grown on GCB agar plates (Difco) containing Kellogg's supplements, in a moist atmosphere containing 5% CO_2 at 37°C. GFP was expressed by introducing the pAM239 plasmid by conjugation (*2*). The non-adhesive *pilE* and *pilC1* strains are described elsewhere (*3, 4*). The *pptA* mutant (8013 *pptA::* mini-*Himar1*) used in this study was derived from a library of transposition mutants described elsewhere (*1, 5*). In-frame deletion of the *pptB* gene was introduced into the *N. meningitidis* chromosome by allelic exchange using the spectinomycin cassette from the pT1Ω1 plasmid (*6*). To complement Δ*pptB* mutants, the WT *pptB* ORF was cloned in the pGCC4 vector, adjacent to *lac*IOP regulatory sequences (*7*) and introduced into the chromosome by homologous recombination. To generate point mutations in the *pilE* gene we took advantage of the *pilE::kan* transcriptional fusion described elsewhere which allows introduction of the chosen *pilE* allele at the endogenous site under the control of its own promoter (*8*). Point mutations in the *pilE* gene were introduced with the Quickchange mutagenesis kit (Stratagene) according to the manufacturer's instructions.

Q-ToF

Pili were prepared as previously described (*9*). Top down analysis of PilE was performed on a Q-TOF-Premier™- (Waters Corp., Milford, MA, USA). The source temperature was set to 80°C. The capillary and cone voltages were set to 2500 and 40 V. The

Q-TOF Premier instrument was operated in wide pass quadrupole mode, for MS experiments, with the TOF data being collected between *m/z* 400–2000 with a low collision energy of 10 eV. Argon was used as the collision gas. Scans were collected for 1 s and accumulated to increase the signal/noise ratio. The MS/MS experiments were performed using a variable collision energy (10–30 eV), which was optimized for each precursor ion. Mass Lynx 4.1 was used both for acquisition and data processing. Deconvolution of multiply charged ions into neutral species was performed using MaxEnt1 in the mass range [10 – 25 kDa] with a resolution of 0.01 Da/channel. An external calibration in MS was done with clusters of phosphoric acid (0.01M in 50:50 Acetonitrile:H_2O v:v). The mass range for the calibration was *m/z* 70 - 2000.

Cell culture

Cells were grown at 37°C in a humidified incubator under 5% CO_2. The human endometrial cell line HEC-1B (HTB113) and human intestinal epithelial cell line Caco-2 were purchased from the American Type Culture Collection (Rockville, Md., USA) and maintained in DMEM medium supplemented with 10% fetal bovine serum (FBS; PAA Laboratories). The HEC-1B cell line was selected because it is extensively used in the field of *Neisseria* infections and it was used to demonstrate the induction of the CREN promoters when bacteria are in contact with host cells (*10*). The Caco-2 cell line is an intestinal cell line that was chosen because it efficiently forms tight junctions and generates transepithelial electrical resistance.

Bacterial aggregation assay

Bacteria grown on GCB agar plates were adjusted to OD_{600}=0.05 and then incubated for 2 hours at 37°C in pre-warmed RPMI supplemented with 10 % FBS with gentle agitation. The bacterial suspension was concentrated to OD_{600}=0.6 or 0.3 by a 1 min centrifugation at 15000 g followed by resuspension in medium containing 0.5 µg/ml of DAPI. Bacterial

suspensions were briefly vortexed and transferred in a glass-bottom 96-well plate (Nunc, Rochester, USA). After 30 min incubation, aggregates were observed microscopically with a 4x lens and size and number determined with the ImageJ software (*11*). Two images were captured per well, corresponding to the surface of most of the well. Using high magnification images each bacterium was estimated to occupy 4.6 μm^3. This value was used to determine the number of bacteria per aggregate based on their volume. Bacterial aggregates smaller than 6 μm in diameter were not considered (about 50 individual bacteria).

Bacterial adhesion, detachment and transmigration assays

Initial adhesion assay. Experiments using the laminar flow chamber were done essentially as described (*2*). HEC1B epithelial cells growing on disposable flow chambers were used (Ibidi GmbH, München, Germany). Experiments using the flow chamber were performed in DMEM supplemented with 2% serum and maintained at 37°C. The bacterial culture was diluted to $7.5x10^7$ bacteria/ml and was introduced into the chamber using a syringe pump (Harvard Apparatus). Adhesion of individual bacteria was recorded using a Nikon Eclipse Ti-E/B inverted microscope with a 20x objective and a Hamamatsu ORCA03 CCD camera.

Adhesion and proliferation in static conditions. For bacterial adhesion to epithelial cells, 24 well plates were seeded with 10^5 HEC-1B cells per well and the monolayers were infected with 10^7 bacteria (MOI=100). After 1h of contact, unbound bacteria were removed by three washes and the infection was continued for 5 h. Adherent bacteria, recovered by scraping the wells, were counted by plating appropriate dilutions on GCB agar plates.

Bacterial detachment assay. Epithelial cells were grown in disposable flow chambers. Bacteria grown on GCB agar plates were adjusted to OD_{600}=0.02 in prewarmed RPMI medium containing 10% fetal bovine serum and cultivated for 2h at 37°. Cells were infected with 10^6 bacteria (MOI=100), adhesion allowed to proceed for 30 min, unbound bacteria removed by three extensive washes and infection continued for 2h in an incubator. Infected

cells were then placed directly in a 0.15 dynes/cm^2 flow. DMEM supplemented with 10% FBS was maintained at 37°C and introduced into the chamber using a syringe pump. Every hour, samples coming out of the flow chamber were collected, serial dilutions performed and a fraction was plated on GCB agar plates.

Bacterial transmigration assay. Caco-2 cells were grown on 12 mm diameter culture plate insert with 3 μm pores (Millipore, Cork, Ireland) for a period of 6 days to reach a trans-epithelial resistance of 600-1000 ohms/cm^2. The upper compartment was infected at an MOI of 100, infection allowed to proceed for 4 hours, inserts transferred to a new well and bacteria were collected in the lower compartments after 90 min. Results were normalized with passage across well without cells to minimize potential interstrain differences and results presented as a percentage relative to the wild type strain.

Electron microscopy

For negative staining transmission electron microscopy, a drop of bacterial suspension in PBS (OD_{600}=1) was placed on a Formvar-coated grid for 10 min. Bacteria were fixed for 5 min with 10 mM cacodylate buffer (pH7.5) containing 2.5% glutaraldehyde. Grids were then washed twice with water and stained for 10 min with 1% phosphotungstic acid, air-dried and viewed using a JEOL JEM-100CX microscope operated at 80 kV.

For scanning electron microscopy, infected cells were fixed with 2.5% glutaraldehyde in 0.1 M cacodylate buffer (pH 7.2) 1h at room temperature. Samples were washed three times for 5 min in 0.2 M cacodylate buffer (pH 7.2), fixed for 1 h in 1% (wt/vol) osmium tetroxide in 0.2 M cacodylate buffer (pH 7.2), and then rinsed with distilled water. Samples were dehydrated through a graded series of 25, 50, 75 and 95% ethanol solution (5 min each step). Samples were then dehydrated for 10 min in 100% ethanol followed by critical point drying with CO_2. Dried specimens were sputtered with 10 nm gold palladium, with a GATAN Ion Beam Coater and were examined and photographed with a JEOL JSM 6700F field

emission scanning electron microscope operating at 5 Kv. Images were acquired with the upper SE detector (SEI).

Modeling of the N. meningitidis pilin and pilus structure

Due to the high sequence identity (77 %) the sequences of *N. gonorrhoeae* and *N. meningitidis* pilins could be simply aligned by eye. The basis of the modeling was the *N. gonorrhoeae* pilin structure as deposited in the model of the pilus (PDB code 2HIL). Missing backbone and side chains were added and optimized for packing within the context of the pilus, in a multi-stage procedure that we implemented in the program CNS (*12*). We assumed that the overall helical parameters of the pilus are the same for *N. gonorrhoeae* and *N. meningitidis* (rise 10.5 Å, angle 105.5). The symmetry of the pilus was enforced throughout the modeling of the pilus using the NCS STRICT command in CNS. In this way, only one single protomer is modeled explicitly, while all the neighbors are treated as images that are created on the fly to calculate the non-bonded interactions. 20 neighbors of the pilin were included in the calculation. We used a modified version of the CHARMM19 force field for all modeling.

The first stage is a quick optimization of the geometry and the packing, with a simplified non-bonded interaction (repulsive Van der Waals only). During this stage, positional restraints were used on those residues that were strictly identical to the residues in the *N. meningitidis* pilin. The second stage is a refinement *in vacuo*, using adapted non-bonded parameters (a distance-dependent dielectric, a switching function between 2 and 9 Å, and a non-bonded cut-off of 10 Å). The third stage is a short refinement in water, similar to the one used in NMR structure determination (*13*). We used a water layer of 10 Å thickness and a non-bonded cutoff of 12 Å. During this stage, the harmonic positional restraints were slowly switched off. During all three stages, the initial structures were maintained in a flexible and adaptive way using log-harmonic distance restraints and automated weighting (*14*).

The CHARMM19 force field was extended for the serine modifications (*15*). Topology and parameter files for these modifications were obtained with the help of the PRODRG2 server (*16*). The atom types were as far as possible mapped onto those of the CHARMM19 force field, or, if not possible, onto those of the CHARMM11 force field (for example, for the glycerophosphate group).

Bundles of pili were generated as symmetric antiparallel tetramers by randomly varying the distance, the rotation angle around the long axis of a pilus, and the crossing angle between pili. The energetic analysis was performed with the ACE generalized Born model implemented in CNS for symmetric systems (*17*). The binding energy was estimated as the difference between the electrostatic, van der Waals and generalized Born contributions to the total energy calculated in the complex and in an isolated pilus. We used 6 for the internal dielectric and 80 for the external dielectric.

2D gel electrophoresis

The isoelectric point of the major pilin subunit in different conditions was determined by 2D gel electrophoresis followed by immunoblot and detection of PilE with specific antiserum. Infection of an epithelial monolayer growing in a 6-well plate was initiated for a period of 30 min at an MOI of 400, cells were washed, infection was allowed to proceed for 2-4 hours as indicated, rinsed with PBS and loading buffer added directly in the wells (8 M urea, 2 M thiourea, 4% (w/v) CHAPS). All samples were treated with 2D Clean-Up kit (GE Healthcare) according to the manufacturer's instructions and the resultant dry pellets were resuspended in loading buffer. Two-dimensional gel electrophoresis was performed as described previously (*18*) and proteins were blotted onto nitrocellulose membrane by standard western blotting procedures (*19*). The PilE protein was detected with a polyclonal antiserum directed against the PilE protein (diluted 1/1000), followed by horseradish peroxydase-linked

anti-IgG (Jackson Immunoresearch Laboratories, diluted 1/10000) and ECL Plus luminescence kit (Amersham Biosciences).

Supporting figures:

Figure S1: Pilin modification with PG, structure and gene involved.

A, Chemical representation of the modification of a serine with phosphoglycerol. **B**, Type IV pili were purified from a mutant in the *pptA*, a gene responsible for the transfer of PE and PC onto *N. gonorrhoeae* pilin and the molecular mass of the major pilin subunit was determined by mass spectrometry. **C,** Phylogenetic analysis of members of COG1368. To generate the phylogenetic tree, multiple alignments were done using the Muscle software package (*20*) and the phylogeny was done with the PhyML software (*21*). We took advantage of the on-line integrated tools found at the following address: http://www.phylogeny.fr/ (*22*). To generate and edit the tree we used the iTol (interactive tree of life) on-line tool (*23*). The genes were colored according to their association with one of the four groups that appear from this analysis. Genes with known function are indicated in a darker color and their function is indicated. Red dots indicate the genes in the COG that present significant homology with PptB and its counterparts in the different *N. meningitidis* strains. Asterisks indicate the genes reported to be regulated at the transcriptional level by changes in the environment.

Figure S2: Relationship between pptB transcription level and pilin modification with PG

The WT and *pptBind* strains were cultivated in the presence of different concentrations of IPTG, pili purified and analyzed by mass spectrometry. Occupancy of serine 93 with PG in

the *pptB*$_{ind}$ strain was plotted as the relative to the total amount of pilin based on the analysis of spectra from three samples for each point (circles). The experiment was done in parallel for the wild type strain (squares). Error bars denote mean of three experiments +/- SEM.

Figure S3: The isoelectric point of the pilin becomes more acidic during growth on the cellular surface.

A, A total bacterial lysate was analyzed by 2D gel electrophoresis, transferred onto a nitrocellulose membrane and proteins were visualized using Ponceau S. **B**, The isoelectric point of PilE from a WT strain was compared when bacteria are proliferating in suspension prior to host cell infection (0) with bacteria proliferating on host cells for a period of 2 hours (2h) and 4 hours (4h). As a control, cells were infected with the *pptB*$_{ind}$ strain in the absence of IPTG and with the *pptB*$_{ind}$*pilES93A* for 4 hours. In each case the same amount of protein was loaded. In the case of the time point 0 the bacterial lysate was mixed with a cellular lysate prepared separately. Open arrowheads indicate the position of the more basic forms, full arrowhead indicates the main form of the pilin and the arrows indicate the acidic forms that accumulated after contact with host cells. The corresponding area analyzed by western blot on panel B is indicated as a rectangle on the Ponceau S stain on panel A.

Figure S4: Impact of serine 93 modification with PG on pilus electrostatic surface

A, An explicit model of a twentimer was generated by repeated duplication, translation and rotation of the primary pilin and represented using the Pymol software. The pilin backbone is in green and the phosphoglycerol on serine 93 in space filling model. **B**, **C,** Modeling of electrostatic surface distribution indicates introduction of a diffuse negative charge around serine 93 when PG is present. Electrostatic calculations were done with APBS (*24*), with an

interior dielectric of 4 and an exterior dielectric of 80, a cube length of 300 Å and grid dimensions of 129 * 129 * 129 points. Defaults were used for all other parameters. The electrostatic potential was mapped onto the surface with -5 and 5 kT as extreme values for red and blue. The surface of the pilus was displayed with VMD (*25*). **B**, Electrostatic surface around serine 93. Blue indicates positive charge and red negative. A positively charged cavity is surrounded by an oval **C**, Electrostatic surface around serine 93 when modified with phosphoglycerol, the cavity surrounded by the oval has lost its positive charge and is physically filled by the modification.

Figure S5: Impact of PG modification on the amount of pili and their ability to form bundles

The amount of pili expressed by different bacteria was determined on whole bacteria preparations by ELISA assay essentially as previously (*26*). **A**, The wild type strain (WT) was compared to the *pptB* deletion strain (Δ*pptB*) and the non-piliated *pilE* mutant (*pilE*). Results are presented relative to the wild type strain. Statistical analyses were done with Student's *t* test. n.s. not significant. **B**, The $pptB_{ind}$ strain in the presence of inducer, 0.1 mM of IPTG. Results are presented relative to the non-induced conditions. **C**, Ability of different strains to form pili bundles was determined. Bacteria displaying pili were classified in two categories: (*i*) with individual fibers only; (*ii*) with bundles (including bacteria with bundles and individual fibers). One hundred bacteria were scored per strain and condition. Three independent experiments each done in triplicate were performed, error bars denote mean of three experiments +/- SEM.

Figure S6: Aggregation

A, The graph represents the level of aggregation as a function of the level of pilin modification with PG. This plot was obtained by combining: (*i*) the ability of the $pptB_{ind}$ strain to form aggregates as a function of IPTG concentration (Fig. 3f) and; (*ii*) the level of modification of pilin with PG also as a function of IPTG concentration in the same strain (Fig. S2). **B**, To demonstrate that the effect of serine 93 modification was mostly due to introduction of a charge at this site, serine 93 was mutated into different amino acids with different properties: negative charge, aspartic acid (S93D) and glutamic acid (S93E); bulky, tyrosine (S93Y). Three independent experiments each done in triplicate were performed, error bars denote mean of three experiments +/- SEM. Statistical analyses were done with Student's *t* test. n.s. not significant, $*p < 0.05$. **C**, The level of insertion of the minor pilin PilX into pili was determined. Pili were purified from the indicated strains, the amount of the major pilin was normalized in the different samples and the amount of PilX was determined by western blot. **D**, Impact of PilX mutation on pilin modification with PG as determined by mass spectrometry. PilE isolated from the *pilX* mutant displays the same pattern as the wild type strain.

Figure S7: Detachment and invasion

A, Detachment of bacteria from aggregates proliferating on the cellular surface after 7h. Results from three independent experiments each done in duplicate were pooled on this graph for the 7h time point. **B**, Caco-2 cells were grown on 12 mm diameter culture plate inserts with 3 µm pores (Millipore, Cork, Ireland) for a period of 6 days to reach a trans-epithelial resistance of 600-1000 ohms/cm^2. Samples from the top and bottom compartments of the inserts were collected at the different time points and the results are indicated as a ratio of the amount of bacteria that transmigrated relative to the amount of bacteria on the top of the insert. Filled circles correspond to the wild type strain, open circles to the Δ*pptB* strain and

squares to the non-piliated *pilE* strain. **C**, Impact of PG modification on the invasion process. Invasion of Caco-2 cells by *N. meningitidis* was tested using a gentamicin resistance assay. The number of gentamicin resistant colony-forming units were normalized by the number of total cell associated bacteria. Results are presented relative to the wild type strain. Three independent experiments each done in triplicate were performed, error bars denote mean of three experiments +/- SEM. Statistical analyes were done with Student's *t* test, $*p < 0.05$.

Figure S8: Model of epithelium colonization

Proposed sequence of events taking place during tissue colonization by *Neisseria meningitidis*: *adhesion*, individual bacteria in suspension attach to the cellular surface, transcription of the *pptB* gene is induced to reach full expression after 2-4 hours; *proliferation*, bacteria multiply in tight aggregates involving bacteria/cell and bacteria/bacteria contacts, pilin is increasingly modified with PG; *detachment*, pilin modification with PG disengages the interaction between pili thus allowing the detachment of individual bacteria from the surface of the microcolony; *propagation*, detached bacteria in suspension can colonize new sites in the same host or in another host. The level of *pptB* goes back to basal level and the level of pilin phosphorylation decreases back to initial levels allowing a new cycle to take place and; *dissemination*, bacteria disengaging from the microcolony but remaining in contact with host cells are in a position to cross the epithelium and disseminate throughout the body.

Supporting references:

1. C. Rusniok *et al.*, *Genome Biol* **10**, R110 (Oct 9, 2009).
2. E. Mairey *et al.*, *J Exp Med* **203**, 1939 (Aug 7, 2006).

3. C. Pujol, E. Eugene, M. Marceau, X. Nassif, *Proc Natl Acad Sci U S A* **96**, 4017 (Mar 30, 1999).
4. P. C. Morand, P. Tattevin, E. Eugene, J. L. Beretti, X. Nassif, *Mol Microbiol* **40**, 846 (May, 2001).
5. M. C. Geoffroy, S. Floquet, A. Metais, X. Nassif, V. Pelicic, *Genome Res* **13**, 391 (Mar, 2003).
6. S. R. Klee *et al.*, *Infect Immun* **68**, 2082 (Apr, 2000).
7. I. J. Mehr, C. D. Long, C. D. Serkin, H. S. Seifert, *Genetics* **154**, 523 (Feb, 2000).
8. X. Nassif *et al.*, *Mol Microbiol* **8**, 719 (May, 1993).
9. J. Chamot-Rooke *et al.*, *Proc Natl Acad Sci U S A* **104**, 14783 (Sep 11, 2007).
10. S. Morelle, E. Carbonnelle, X. Nassif, *J Bacteriol* **185**, 2618 (Apr, 2003).
11. M. D. Abramoff, P. J. Magelhaes, S. J. Ram, *Biophotonics International* **11**, 36 (2004).
12. A. T. Brunger *et al.*, *Acta Crystallogr D Biol Crystallogr* **54**, 905 (Sep 1, 1998).
13. J. P. Linge, M. A. Williams, C. A. Spronk, A. M. Bonvin, M. Nilges, *Proteins* **50**, 496 (Feb 15, 2003).
14. M. Nilges *et al.*, *Structure* **16**, 1305 (Sep 10, 2008).
15. B. R. Brooks *et al.*, *J Comp Chem* **4**, 187 (1983).
16. A. W. Schuttelkopf, D. M. van Aalten, *Acta Crystallogr D Biol Crystallogr* **60**, 1355 (Aug, 2004).
17. L. Moulinier, D. A. Case, T. Simonson, *Acta Crystallogr D Biol Crystallogr* **59**, 2094 (Dec, 2003).
18. A. Gorg *et al.*, *Electrophoresis* **21**, 1037 (Apr, 2000).
19. H. Towbin, T. Staehelin, J. Gordon, *Proc Natl Acad Sci U S A* **76**, 4350 (Sep, 1979).
20. R. C. Edgar, *Nucleic Acids Res* **32**, 1792 (2004).

21. S. Guindon, O. Gascuel, *Syst Biol* **52**, 696 (Oct, 2003).

22. A. Dereeper *et al.*, *Nucleic Acids Res*, (May 6, 2008).

23. I. Letunic, P. Bork, *Bioinformatics* **23**, 127 (Jan 1, 2007).

24. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, *Proc Natl Acad Sci U S A* **98**, 10037 (Aug 28, 2001).

25. W. Humphrey, A. Dalke, K. Schulten, *J Mol Graph* **14**, 33 (Feb, 1996).

26. S. Helaine *et al.*, *Mol Microbiol* **55**, 65 (Jan, 2005).

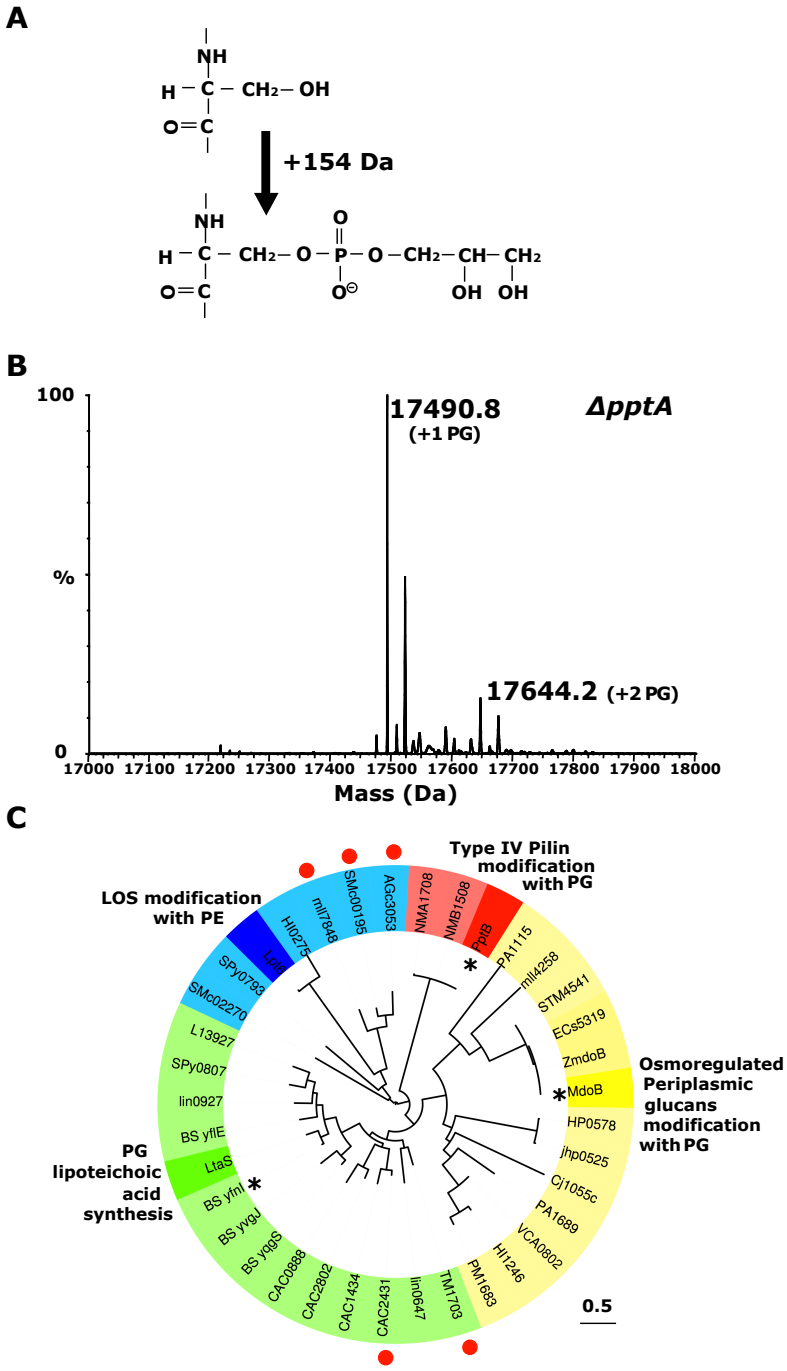

A
+154 Da
B
ΔpptA
17490.8
(+1 PG)
17644.2 (+2 PG)
100
%
0
17000 17100 17200 17300 17400 17500 17600 17700 17800 17900 18000
Mass (Da)
C
Type IV Pilin modification with PG
LOS modification with PE
Osmoregulated Periplasmic glucans modification with PG
PG lipoteichoic acid synthesis
NMA1708
NMB1508
PptB
PA1115
nll4258
STM4541
ECs5319
ZmdoB
MdoB
HP0578
jhp0525
Cj1055c
PA1689
VCA0802
HI1246
PM1693
TM1703
lin0647
CAC2431
CAC1434
CAC2802
CAC0888
BS ykoS
BS yvgJ
BS yfnI
LtaS
BS yflE
lin0927
SPy0807
L13927
SMc02270
SPy0793
LptC
HI0275
nll7848
SMc00195
AGc3053
0.5

Figure S1

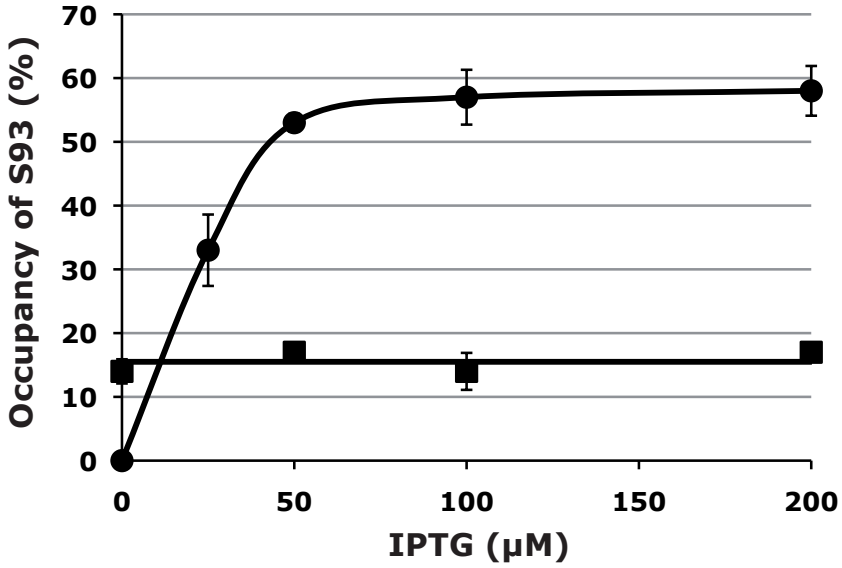


Figure S2

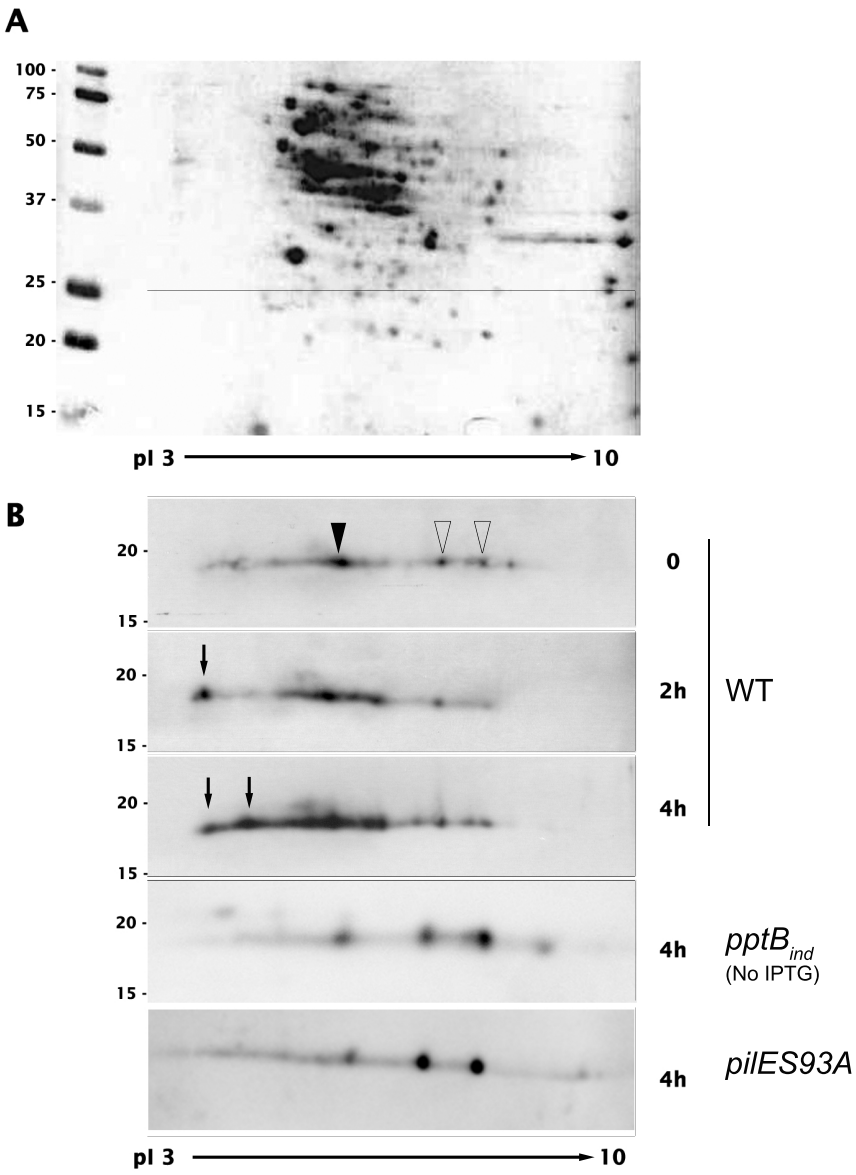

A
100
75
50
37
25
20
15
pI 3
10
B
20
15
0
2h
WT
4h
4h
pptB ind
(No IPTG)
4h
pilES93A
pI 3
10

Figure S3

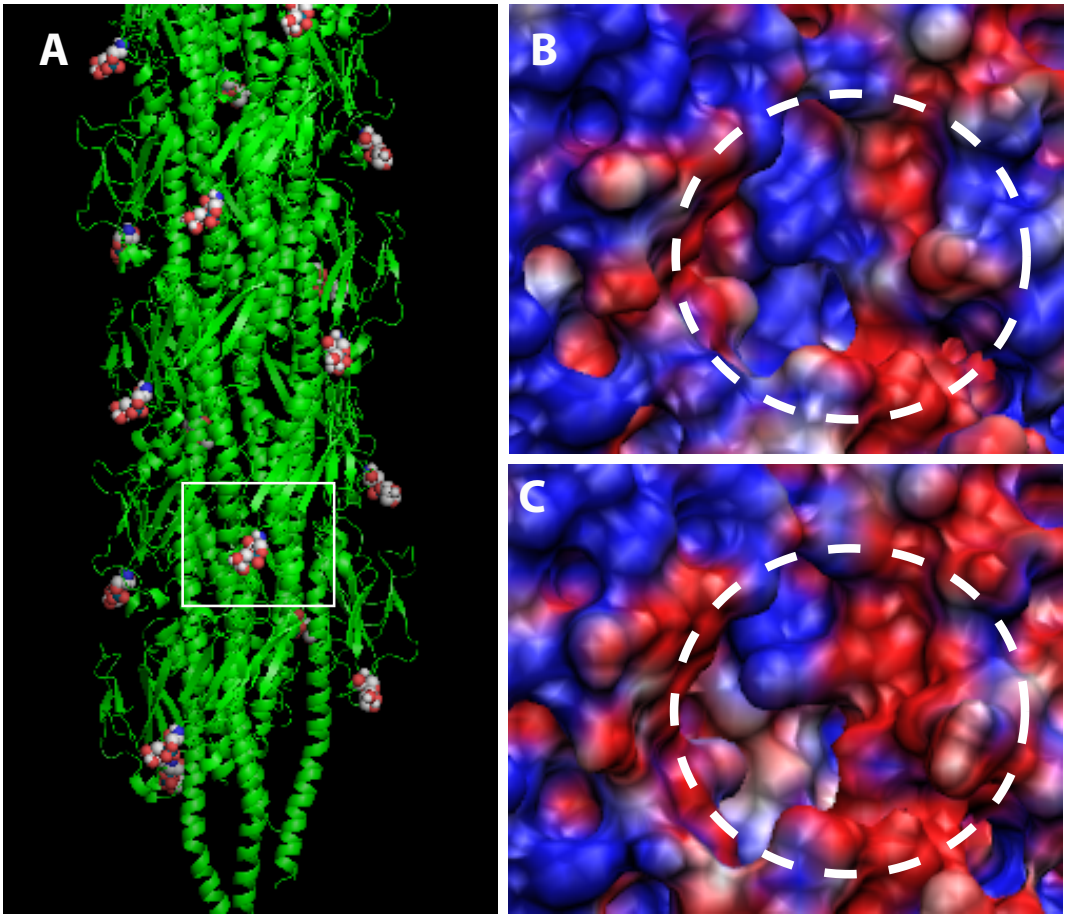
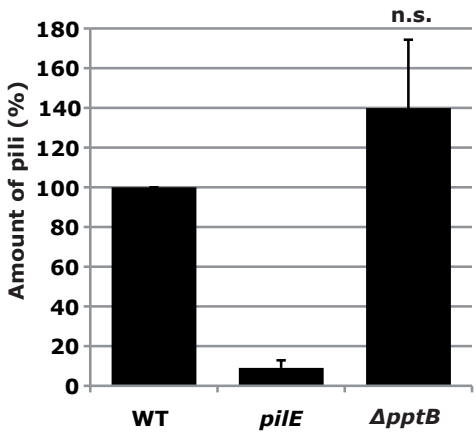
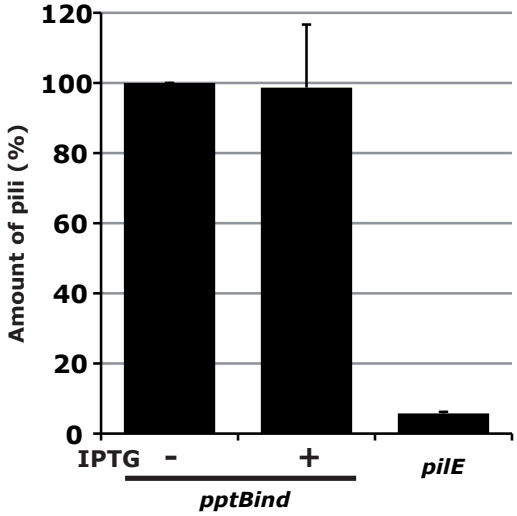


Figure S4

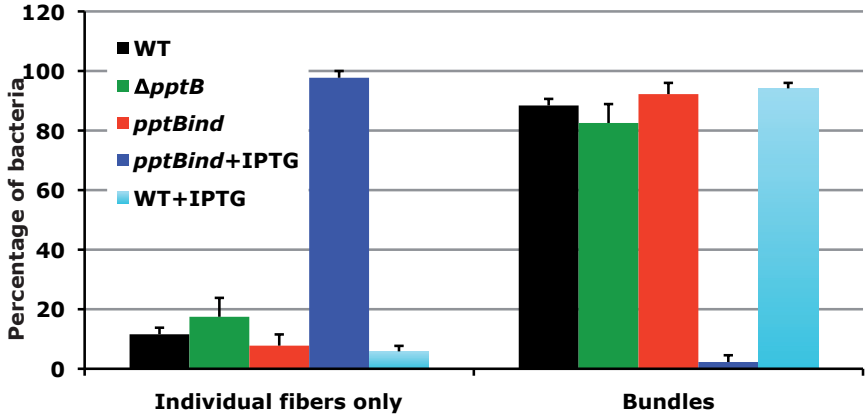
A



B



C



Supplemental figure S5

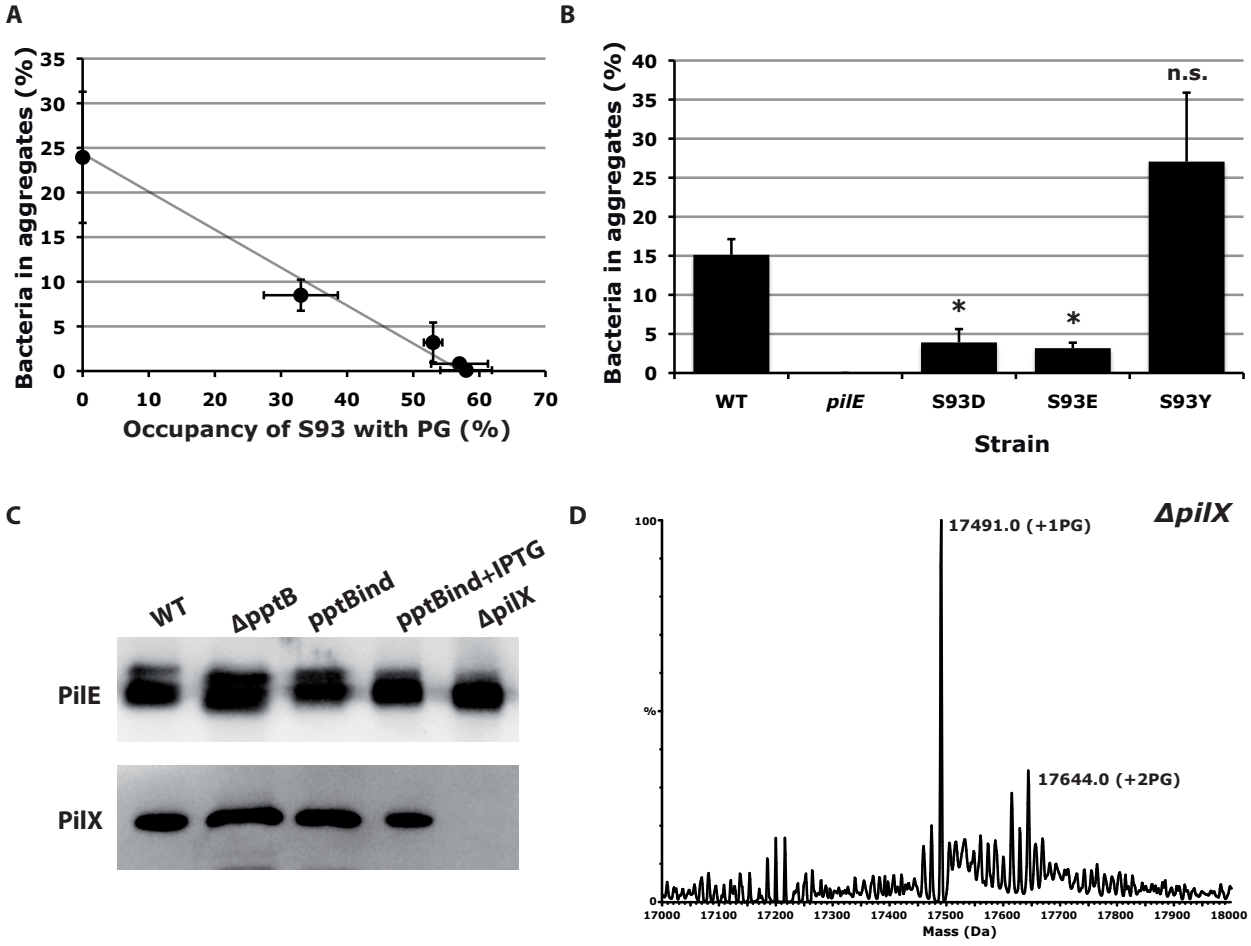

A
Bacteria in aggregates (%)
35
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
Occupancy of S93 with PG (%)
B
Bacteria in aggregates (%)
40
35
30
25
20
15
10
5
0
n.s.
*
*
WT
pilE
S93D
S93E
S93Y
Strain
C
WT
ΔpptB
pptBind
pptBind+IPTG
ΔpilX
PilE
PilX
D
ΔpilX
100
%
0
17491.0 (+1PG)
17644.0 (+2PG)
17000
17100
17200
17300
17400
17500
17600
17700
17800
17900
18000
Mass (Da)

Figure S6

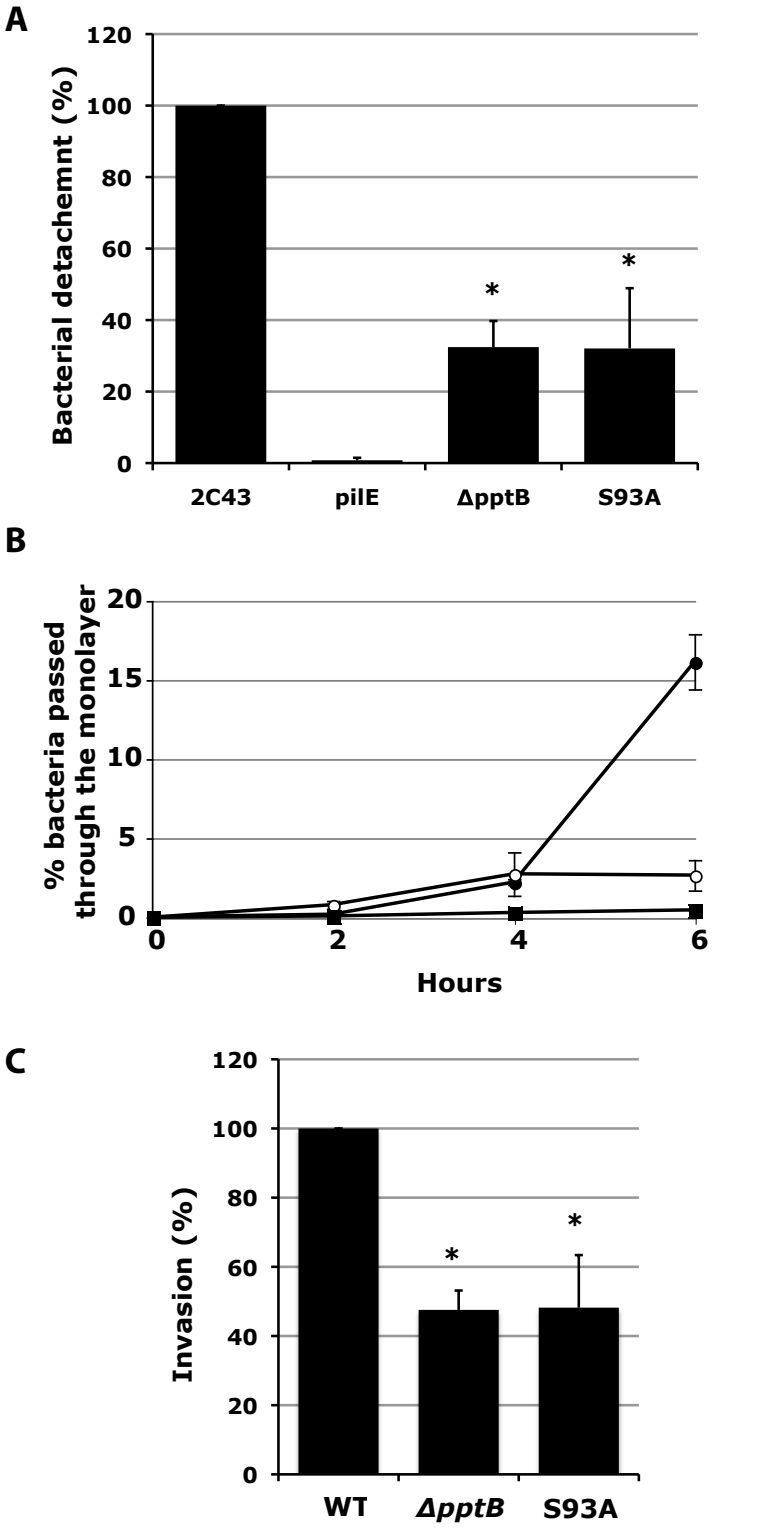

A
Bacterial detachemnt (%)
120
100
80
60
40
20
0
*
*
2C43
pilE
ΔpptB
S93A
B
% bacteria passed through the monolayer
20
15
10
5
0
0
2
4
6
Hours
C
Invasion (%)
120
100
80
60
40
20
0
*
*
WT
ΔpptB
S93A

Figure S7

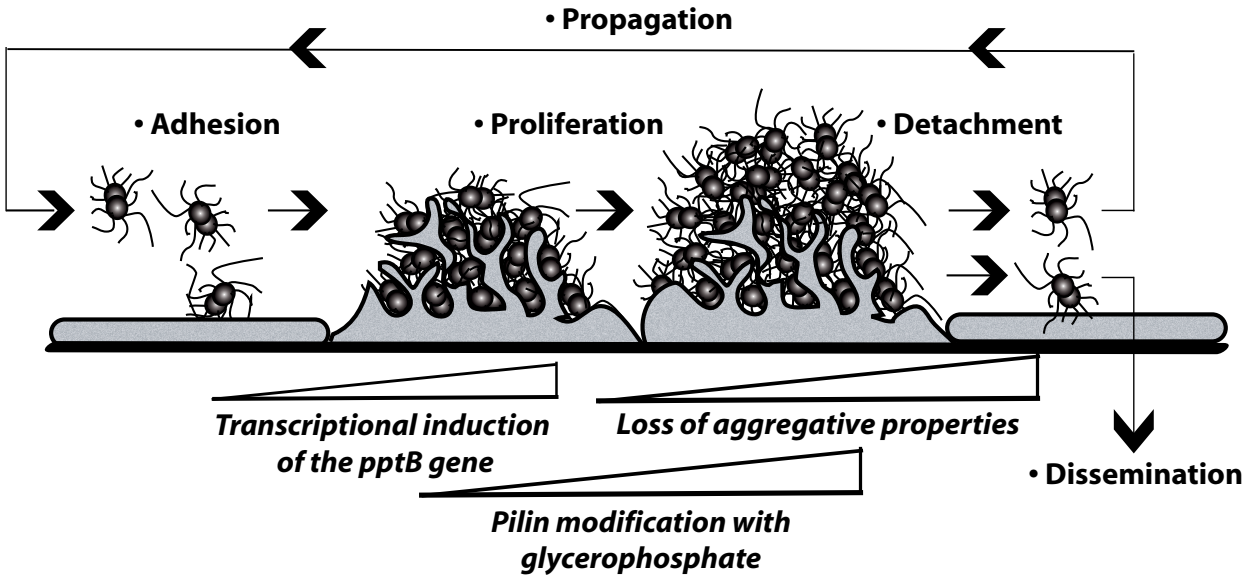

• Propagation
• Adhesion
• Proliferation
• Detachment
Transcriptional induction of the pptB gene
Loss of aggregative properties
• Dissemination
Pilin modification with glycerophosphate

Figure S8

4. Additional Results for "Posttranslational Modification of Pili Upon Cell Contact Triggers *N. meningitidis* Dissemination"

In order to further support our conclusions, during the preparation of this article additional mutants were created and their respective PilE populations completely characterised including mass profiling, PTM identification and site localisation. This was performed using the using the previously described mass profiling plus bottom-up MS methodology. In some cases in-gel digestion was also utilised after separation of PilE by SDS PAGE as this gave cleaner digests that facilitated the identification of modified peptides. For brevity, only the mass profiling spectra are shown (Figure 47) and the results of the bottom-up experiments summarised rather than each MS/MS spectrum being presented and described in detail (Table 2).

Firstly, PilE from the double point mutant S93A S94A and its *pptB* inducible counterpart $pptB_{ind^+}$S93A S94A were prepared and analysed in order to provide further evidence that Ser^{94} was not the third site of PG modification. This was particularly important since double modification at neighbouring Ser^{93} and Ser^{94} would introduce a double negative charge into this region and would affect the results of the molecular modelling. Mass profiling of the S93A S94A mutant (Figure 47 top) gave the expected protein mass, and site localisation confirmed that only Ser^{69} is modified with PG. The $pptB_{ind^+}$ S93A mutant gave the expected profile showing two peaks of similar abundance with masses corresponding to PilE+PG and PilE+2PG. In this mutant Ser^{69} and Thr^{68} were identified as the sites of PG modification.

In two additional experiments, mass profiling (Figure 47 bottom left) and PTM characterisation of the $pptB_{ind^+}$ S63A mutant were performed in order to investigate whether the presence of GATDH had any effect on the phosphoglycerolation state. Loss of glycan had no discernible effect on the mass profile but site localisation data suggested that in the absence of GATDH at Ser^{63}, Ser^{94} was modified with PG rather than Thr^{68}. This result was not explored further at the time and is perhaps worthy of future investigation. Finally a *pilV* deletion mutant, Δ*pilV*, deficient for expression for the minor pilin PilV, was characterised as a complement to the Δ*pilX* mutant presented in the supplementary information of the article. Both mass profiling (Figure 47 bottom right) and site localisation gave results similar to the wild type.

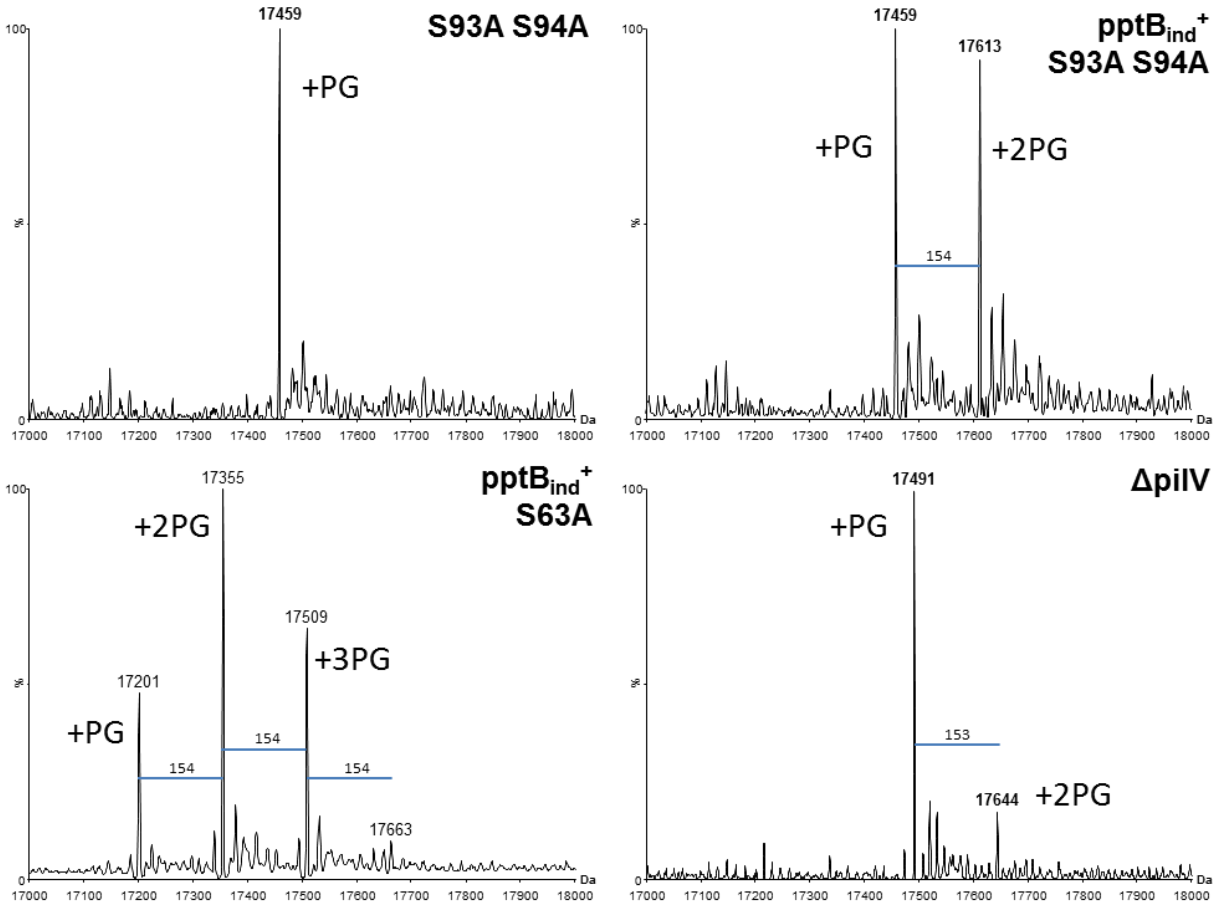


Figure 47 - nano-ESI Q-ToF mass profiling of PilE purified from S93A S94A, *pptB$_{ind}^{+}$* S93A S94A (top left and right respectively) *pptB$_{ind}^{+}$* S63A (bottom left) and Δ*pilV* mutants (bottom right). Mass differences between the peaks are highlighted with coloured bars

Mutant	GATDH Localisation	PG Modification & Localisation			PG Site Occupancy Ratio
		1	2	3	
WT	Ser^{63}	Ser^{69}	Ser^{93}	-	5:1
Δ*pilV*	Ser^{63}	Ser^{69}	Ser^{93}	-	5:1
pptB$_{ind}^{+}$	Ser^{63}	Ser^{69}	Ser^{93}	Thr^{68}	2:2:1
***pptB$_{ind}^{+}$* S63A**	-	Ser^{69}	Ser^{93}	Ser^{94}	7:5:2
***pptB$_{ind}^{+}$* S93A**	Ser^{63}	Ser^{69}	Thr^{68}	-	5:2
***pptB$_{ind}^{+}$* S93A S94A**	Ser^{63}	Ser^{69}	Thr^{68}	-	2:1

Table 2 - Summary of PTM characterisation performed on additional Nm 8013 mutants. Occupancy ratio is calculated from the mass profiling experiments

5. Conclusions from the Investigation Into the Biological Function of PG

5.1. Biological Relevance

The hypothesis presented in this work suggests that modification of PilE with PG is induced several hours after host cell contact and is a prerequisite for pilus unbundling, bacterial dissemination and transmigration of the epithelial barrier. These are key steps in both the colonisation and infection model of Nm as depicted in the cartoon in Figure 48.

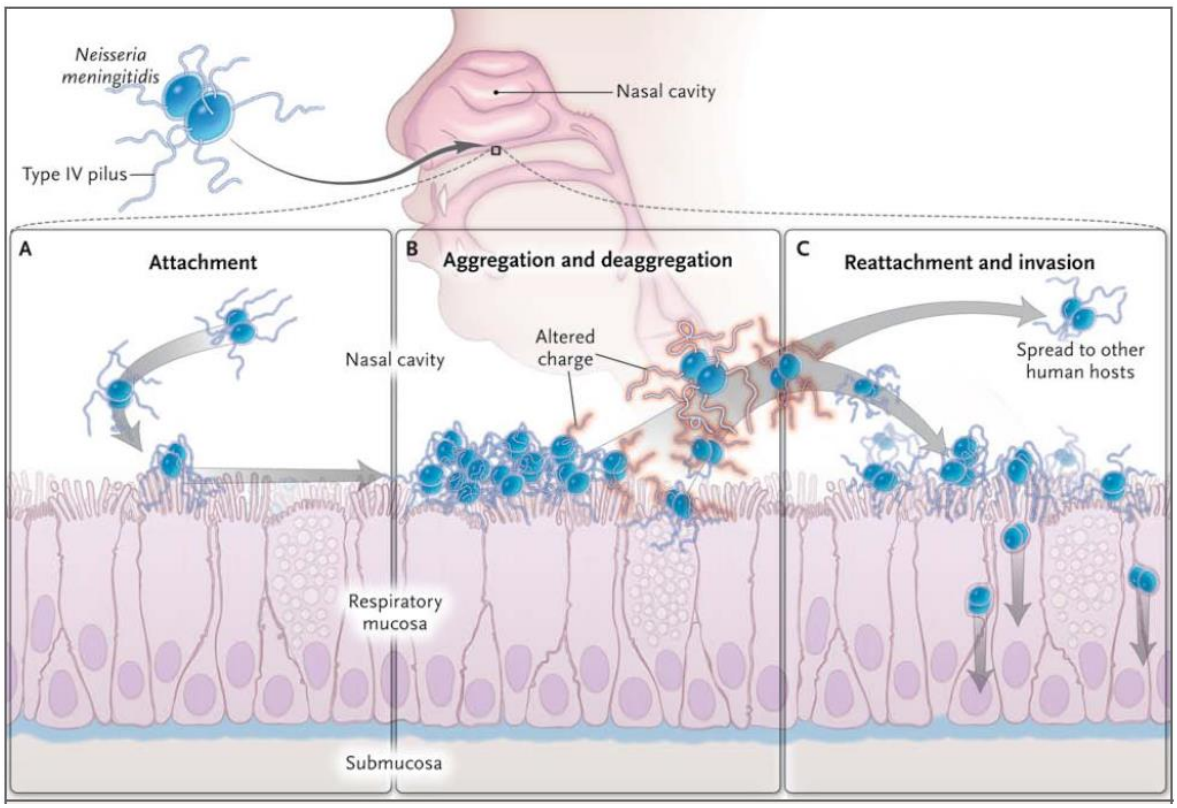


Figure 48 - Role of pili in Nm colonisation and the effect of PG modification on bacterial aggregation and crossing of the epithelium. Taken from Quagliarello *et al.*[5]

Pili are required for initial attachment of bacteria to the epithelial cells of the nasopharyngeal mucosa (panel A). During colonisation, bundling of pili from neighbouring bacteria promotes colony integrity through bacterial aggregation (panel B). Several hours after initial host cell contact *pptB* is up-regulated and PptB expression levels increase. PilE is modified with additional PG groups mainly at Ser^{93}. This changes the surface charge of the pilus fibres and promotes pili unbundling and bacterial disaggregation (panel B). Dissemination of colony members can then occur. This can be followed by further colonisation or invasion of the mucosal epithelium (panel C).

In addition to providing a biological explanation for the role of the PG modification, this work also raises several questions regarding the molecular trigger, signalling pathway and the generality of this unbundling mechanism for other strains of Nm[5-7]. The mechanics involved in protein modification also remain to be unravelled since modification of bundled pili with PG would require separation of fibres from the bundle or entire bundle reaction. These are all subjects for future research.

5.2. Mass Spectrometry

Mass profiling of PilE was an integral aspect of this work to describe the function of the PG modification. The characterisation of the $pptB_{ind^+}$ mutant which expressed three proteoforms of PilE did however highlight several weaknesses of the bottom-up methodology. As explained in the introduction, when a bottom-up approach is performed on a protein population in which multiple proteoforms are present, the tryptic digestion step results in a loss of connectivity between the PTMs and their parent proteoforms. Up until now this limitation has been compensated for by using a combination of intact mass profiling and genetic point mutation to confirm the exclusive occupancy of modification sites. A method that would abrogate this time consuming step would obviously be particularly useful. A top-down MS approach would fulfil this requirement and the development of this methodology is chronicled in the next chapter.

Bibliography

[1] X. Nassif, J. Lowy, P. Stenberg, P. Ogaora, A. Ganji and M. So.Antigenic Variation Of Pilin Regulates Adhesion Of Neisseria-Meningitidis To Human Epithelial-Cells. *Molecular Microbiology*, **1993**, *8*, 719.

[2] C. Rusniok, D. Vallenet, S. Floquet, H. Ewles, C. Mouze-Soulama, D. Brown, A. Lajus, C. Buchrieser, C. Medigue, P. Glaser and V. Pelicic.NeMeSys: a biological resource for narrowing the gap between sequence and function in the human pathogen Neisseria meningitidis. *Genome Biol*, **2009**, *10*, R110.

[3] J. Chamot-Rooke, B. Rousseau, F. Lanternier, G. Mikaty, E. Mairey, C. Malosse, G. Bouchoux, V. Pelicic, L. Camoin, X. Nassif and G. Dumenil.Alternative Neisseria spp. type IV pilin glycosylation with a glyceramido acetamido trideoxyhexose residue. *Proceedings of the National Academy of Sciences of the United States of America*, **2007**, *104*, 14783.

[4] C. C. Brinton, J. Bryan, J. Dillon, N. Guerina, L. J. Jacobson, A. Labik, S. Lee, A. Levine, S. Lim, J. McMichael, S. Polen, K. Rogers, A. C. C. To and S. C. M. To, Immunobiology of Neisseria gonorrhoeae : proceedings of a conference held in San Francisco, California, 18-20 January 1978, San Francisco, California, 1978.

[5] E. Stimson, M. Virji, S. Barker, M. Panico, I. Blench, J. Saunders, G. Payne, E. R. Moxon, A. Dell and H. R. Morris.Discovery of a novel protein modification: alpha-glycerophosphate is a substituent of meningococcal pilin. *Biochem. J.*, **1996**, *316*, 29.

[6] M. Virji.Post-translational modifications of meningococcal pili. Identification of common substituents: Glycans and alpha-glycerophosphate - A review. *Gene*, **1997**, *192*, 141.

Chapter 4

Development of Top-Down Mass Spectrometry of PilE for Complete Posttranslational Modification Characterisation

Characterisation of PilE by a top-down MS approach has the potential to allow fast and complete PTM mapping of individual proteoforms inside the mass spectrometer. Since no enzymatic digestion step is involved, the connectivity between PTMs and their parent proteoforms is retained and the PTM complement of each individual proteoform can be explicitly defined. All PTMs of all proteoforms of PilE from the Nm 8013 reference strain have already been defined through a combination of bottom-up MS and genetic point mutation. PilE-8013 therefore provides an excellent base onto which a top-down method can be built.

The goals of the top-down experiment are twofold. The first is simply to achieve sufficiently extensive backbone fragmentation to allow PTM localisation. Once this goal has been achieved, the second is to maximise the overall sequence coverage. The progress made towards the second goal will enable the approach to be critically evaluated for use on pilins expressed by other meningococcal strains and even different organisms, where the PTMs may not be located in the same regions of the protein. As the development of the top-down experiment progresses, results will therefore be judged against each of these two objectives.

The top-down experiment is fairly complex with many of the experimental and instrumental parameters being protein dependant. This ranges from the voltages and RF frequencies used on the components inside the mass spectrometer, to the specific parameters used for electronic and vibrational fragmentation and finally to spectral treatment and data analysis. Each requires extensive manual optimisation to reveal the potential and the limitations of the top-down approach. For this reason very few, if any, reports of complete parameter optimisation at the protein level exist. The most complete account uses the peptide standard substance P presumably because both protein and biological samples are often low in concentration and difficult to handle[1]. There is therefore no standardised approach to experimental design.

The following chapter details the construction and optimisation of a robust top-down methodology for the analysis of PilE-8013 using both FT-ICR and Orbitrap platforms. Important experimental parameters are identified and investigated as the experiment is developed.

1. Top-Down Analysis of PilE-8013 on a 7 Tesla Bruker Apex III FT-ICR Mass Spectrometer

Initial development of the top-down experiment began on the Bruker 7T Apex III FT-ICR mass spectrometer installed in the DCMR lab at Ecole Polytechnique in 2003. A schematic and picture of this instrument are shown below in Figure 49.

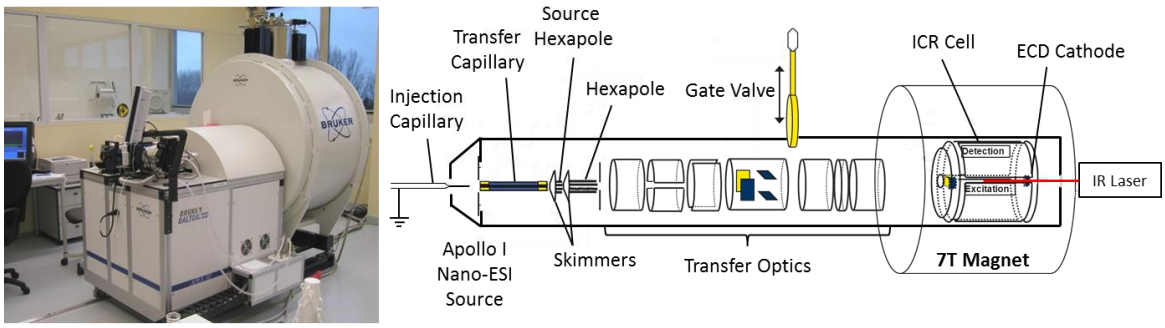


Figure 49 - Schematic (right) and photograph (left) of the Bruker Apex III FT-ICR mass spectrometer installed in the DCMR lab at Ecole Polytechnique

This generation of FT-ICR mass spectrometer is fairly basic compared to current state-of-the-art instrumentation. Sample injection in nano-electrospray (nanospray) mode is performed using the classical glass capillary method. Ions are introduced into the mass spectrometer by spraying them on axis (directly at the capillary entrance). After desolvation in the heated transfer capillary ions are guided into the low pressure region of the mass spectrometer by two simple skimmers. The pre-hexapole in the source region acts primarily as an ion guide and the hexapole as a mass filter for ion selection. There is no multipole specifically designed for ion accumulation or fragmentation. As with all Bruker FT-ICR instruments, the voltages and RF frequencies on these components must be manually tuned for each new analyte as they are extremely protein dependant. Once inside the high vacuum region of the MS, ions are guided by a series of transfer optics through the fringe field of the magnet into the ICR cell where they are trapped and detected. Parameters for the high vacuum ion transfer optics and ICR cell parameters do not need to be drastically altered for most experiments. Trapping of the ions may require optimisation of cell parameters depending on the technique used.

1.1. Apex III Performance Overview

In order to provide a first estimate to the experimental parameters, the protein standard equine myoglobin was used. This was a good model protein for which transmission parameters could be optimised since it has a similar molecular weight to PilE-8013, exhibits a similar charge state profile and can be easily prepared at high sample concentrations. For nanospray, a metal coated,

tapered glass capillary was filled with a solution of myoglobin and was opened by breaking a small piece off the tapered end using a home-built stage mounted under a microscope. The needle was installed and nano-electrospray initiated. Lens voltages and RF potentials of the source components were iteratively modified to maximise signal intensity.

Once reasonable ion transmission had been achieved a sample of PilE-8013 was loaded and the optimisation process fine-tuned to further increase signal intensity as much as possible. A full MS spectrum of PilE (Figure 50 top) was obtained by acquiring 15 scans and summing the transients before apodization, zero filling and Fourier-transform (this process is handled by the acquisition software).

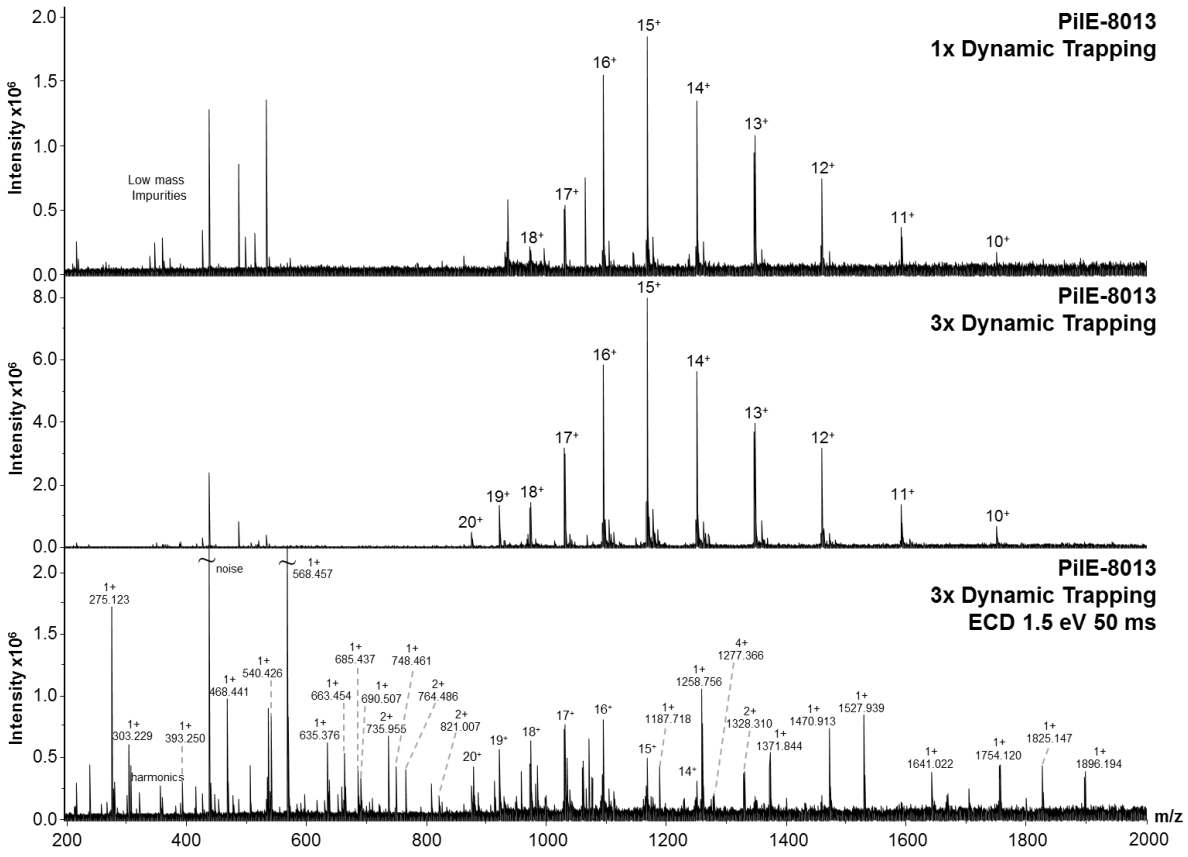


Figure 50 - FT-ICR mass spectrum of PilE after optimisation of experimental parameters (top) three rounds of dynamic trapping (middle) and ECD fragmentation over the entire mass range (bottom). All spectra are the result of summing 15 transients

Unfortunately, even after extensive parameter observation and spectral accumulation, the signal intensity achieved in the MS experiment was rather low (Figure 50 top). Good signal intensity is a critical prerequisite to generate informative MS/MS spectra because the ECD process is inefficient and often produces abundant, charged reduced molecular ions that are not useful for protein characterisation. A clear single-scan precursor ion intensity threshold of around 1×10^6 exists on

this instrument, below which ECD fragmentation is non-optimal and even ineffectual (this number has been derived empirically from previous experiments on protein standards). More intense precursor ions are therefore required to produce exploitable MS/MS information.

1.2. Ion Accumulation to Improve Signal Intensity

Signal intensity can be improved by accumulating the stream of ions produced by the electrospray source into larger ion packets in different regions of the mass spectrometer. Firstly the time delay values associated with pulsing ions in and out of the hexapoles were raised. This produces a trapping effect that can be exploited to increase ion intensity but is limited in its effectiveness as these components are not specially designed for ion trapping. In addition, the dynamic trapping technique (DT), which involves repeatedly charging the ICR cell with ions in order to boost the signal intensity, was also employed[2, 3]. This process requires pulsing inert gas into the cell in order to efficiently trap the ions before the next packet is introduced. Once "full" the cell must be pumped out to generate sufficiently low pressure for excitation and detection and this has the drawback of introducing a very long duty cycle (several seconds). Three rounds of DT were found to improve signal intensity almost fourfold (Figure 50 middle) and now that a reasonable precursor ion intensity had been achieved MS/MS could be trialled.

1.3. Initial Attempts at ECD MS/MS

As has been previously explained, there are many possible fragmentation modes that can be utilised to perform top-down MS. Since the aim of the top-down experiment here is to localise labile PTMs, ECD is the fragmentation mode of choice. There are two additional experimental parameters that must be selected when performing ECD: the electron energy and the electron irradiation time. These will collectively be referred to as the ECD parameters and for this initial attempt at ECD MS/MS were fixed at 1.5 eV and 50 ms based on previous experience of fragmentation of proteins standards. The signal intensity achieved in the full MS experiment was insufficient for isolation and fragmentation of single charge states of PilE-8013, so three rounds of DT were used and ECD MS/MS was performed on the entire mass range. Fifteen transients were summed and the resulting spectrum is shown in Figure 50 (bottom).

Now that an ECD MS/MS spectrum had been obtained it must be analysed. Interpretation of top-down spectra is complex and there is as yet no universally accepted workflow. Individual groups therefore either treat data manually, or develop their own algorithms for peak picking, de-isotoping and deconvolution, and-in house tools for fragment ion assignment. For this initial top-down ECD MS/MS spectrum of PIlE-8013, two methods of spectral analysis were trialled and compared.

1.4. An Automated Approach to Peak Picking, Deconvolution and Ion Assignment

The automated approach uses the manufacturer's software. The MS/MS spectrum was imported into Bruker Data Analysis 4.0 SP5 for peak picking and deconvolution. Within Data Analysis the SNAP 2.0 algorithm can be used to identify isotopic envelopes and extract a list of neutral mono-isotopic masses. Several user-defined parameters must be entered that determine the stringency of peak picking. Four of these parameters are particularly important and define cut-off values below which candidate spectral features are discarded. The first two are intensity thresholds (absolute, I_a and relative, I_r), the third a signal to noise (S/N) threshold and the fourth a "quality factor" (Q_f). During peak picking SNAP will try to match each candidate isotope pattern to an appropriate theoretical isotope pattern generated from averagine. The fit is represented by a score or quality factor, which ranges from 1.0, a perfect match, to 0.0 indicating an incredibly poor fit. Setting this value to 0.5 for example, will discard any picked patterns with $Q_f < 0.5$.

It was found from preliminary investigations with fragmentation data acquired from the model protein myoglobin, that setting the intensity thresholds to non-limiting values ($I_a = 0$, $I_r = 1.0\times10^{-5}$) and the S/N factor to 2 consistently resulted in "sensible" peaks being picked (*i.e.* nothing that was clearly part of the baseline). Using these values and an initial $Q_f = 0.9$, 17 patterns were picked from the ECD spectrum of PilE-8013.

Now the spectrum had been processed the deconvoluted masses of the 17 patterns were exported as a list to the Bruker BioTools software for assignment. This stage concerns matching the deconvoluted masses against theoretical ions expected from PilE-8013. Of the 17 patterns 15 represented distinct masses, the other two being additional charge states. Using 20 ppm peak picking tolerance and the standard ion types defined in BioTools for ECD fragmentation (*c*, *z*, $z^{+\cdot}$), no assignment was achieved until the N-terminus was modified with a methyl group. Even then only 3 of the 15 masses could be matched to sequence fragments of PilE (c_4-c_6). Clearly there are more than 3 fragments present in the MS/MS spectrum (Figure 50 bottom) and thus this standard peak picking and ion assignment protocol required further refinement.

Widening the Search to Other Fragment Ion Types

ECD commonly produces *c* and $z^{+\cdot}$ type ions, however the formation of other ion types is well documented in the literature; alternative ECD pathways have been shown to form *a* and *y* type ions[4, 5] and *b* type ions have been observed in some ECD spectra[6, 7]. In addition hydrogen transfer reactions may occur during fragmentation producing *c-H* and *z+H* ions (often called *c-1* and *z+1*)[8, 9]. Increasing the types of ions searched for to *a*, *b*, *c*, *c-H*, *y*, $z^{+\cdot}$, *z+H*, *z+2H* ions allowed 7 of the 15 masses to be assigned. Interestingly these additional ions were all *b* type b_{11}-b_{14}. This was an

* Note that BioTools defines a *z* ion in the classical manner *i.e.* an even electron carbocation from

improvement but visual inspection of the spectrum clearly indicates that there are more features that remained unpicked and therefore unassigned.

Investigating the Effect of SNAP Q_f

The initial peak picking parameters were re-examined and the effect of the Q_f on the number of patterns picked and masses assigned was investigated (Figure 51).

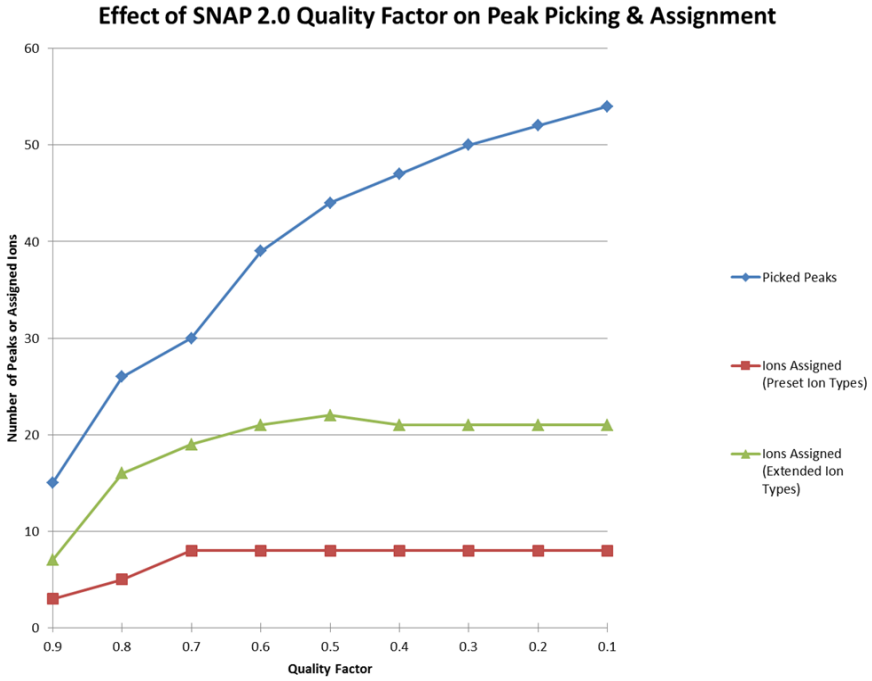


Figure 51 - Plot of number of picked peaks against SNAP 2.0 Q_f values (blue curve) and number of assigned masses with the standard *c*, *z*, $z^{+\cdot}$ (red curve) and extended *a*, *b*, *c*, *c-H*, *y*, $z^{+\cdot}$, *z+H*, *z+2H* ion series (green curve)

It was found that decreasing the quality factor increased both the number of picked peaks, from 15 at Q_f = 0.9 to 54 at Q_f = 0.1 and the number of assigned masses, from 7 at Q_f = 0.9 to 20 at Q_f = 0.1 The biggest jump in assigned peaks occurred between Q_f = 0.9 and 0.7 and remained static after that. Of the new peaks assigned most were singly charged ions that were not previously identified, seemly because their ^{13}C peak was very low in intensity and obscured by the baseline. Several multiple charged ions with distorted isotope patterns were also assigned. A good example of a real, multiply charged ion with a distorted isotope envelope, that is not identified with the standard ion series or with higher Q_f values, is the $z+H_{49}$ ion at *m/z* 1277.366 and Q_f = 0.614 (Figure 52).

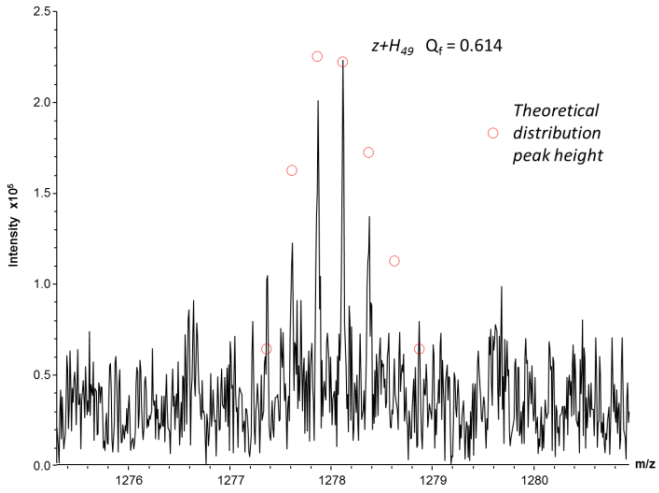


Figure 52 - $z+H_{49}$ ion at m/z 1277.366 exhibiting a highly distorted isotope pattern

Choosing Peak Picking Parameters

After much consideration it was decided to set the Q_f value to 0.3 for future analyses. Thus a large number of patterns would be picked and, if needs be, more stringent criteria could be used during ion assignment. This was based on the belief that the SNAP 2.0 algorithm was fairly robust at distinguishing real features from both the baseline and from high intensity noise (such as shot noise). In addition even at low Q_f values few (if any) peaks were picked that would not have been identified manually.

Automatic peak picking and ion assignment were now performed with these finalised parameters and resulted in the identification of 21 distinct ions (c_3-c_6, c_{18}, c_{25}, c_{31}, b_9-b_{18}, b_{20}, z_{161}, $z+H_{49}$, $z+2H_{49}$). These fragments correspond to 18 inter-residue cleavages and give total sequence coverage of 12% (Figure 53). Note that the sequence coverage in brackets in the fragmentation map is the coverage obtained if the region between Cys[120] and Cys[154] is not considered in the calculation. The z_{161} ion is a false assignment formed by neutral loss from the precursor. It has been included here and is shaded in grey in the fragmentation map but ions of this type will be removed from future analyses.

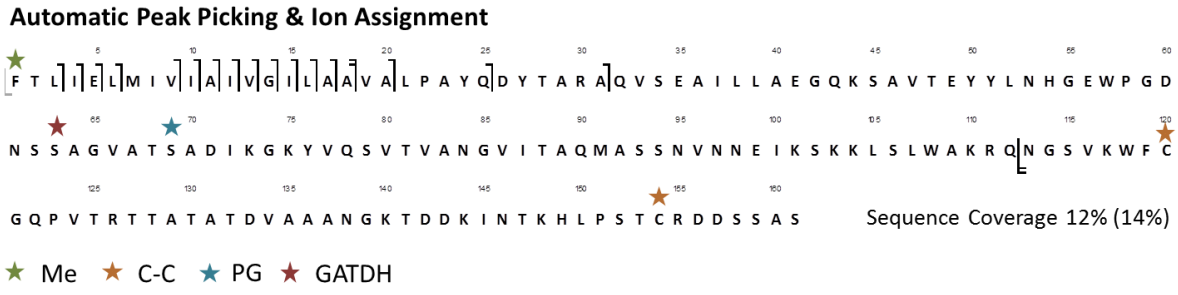


Figure 53 - Fragmentation maps of ions assigned to PilE-8013 after automatic peak picking and ion assignment

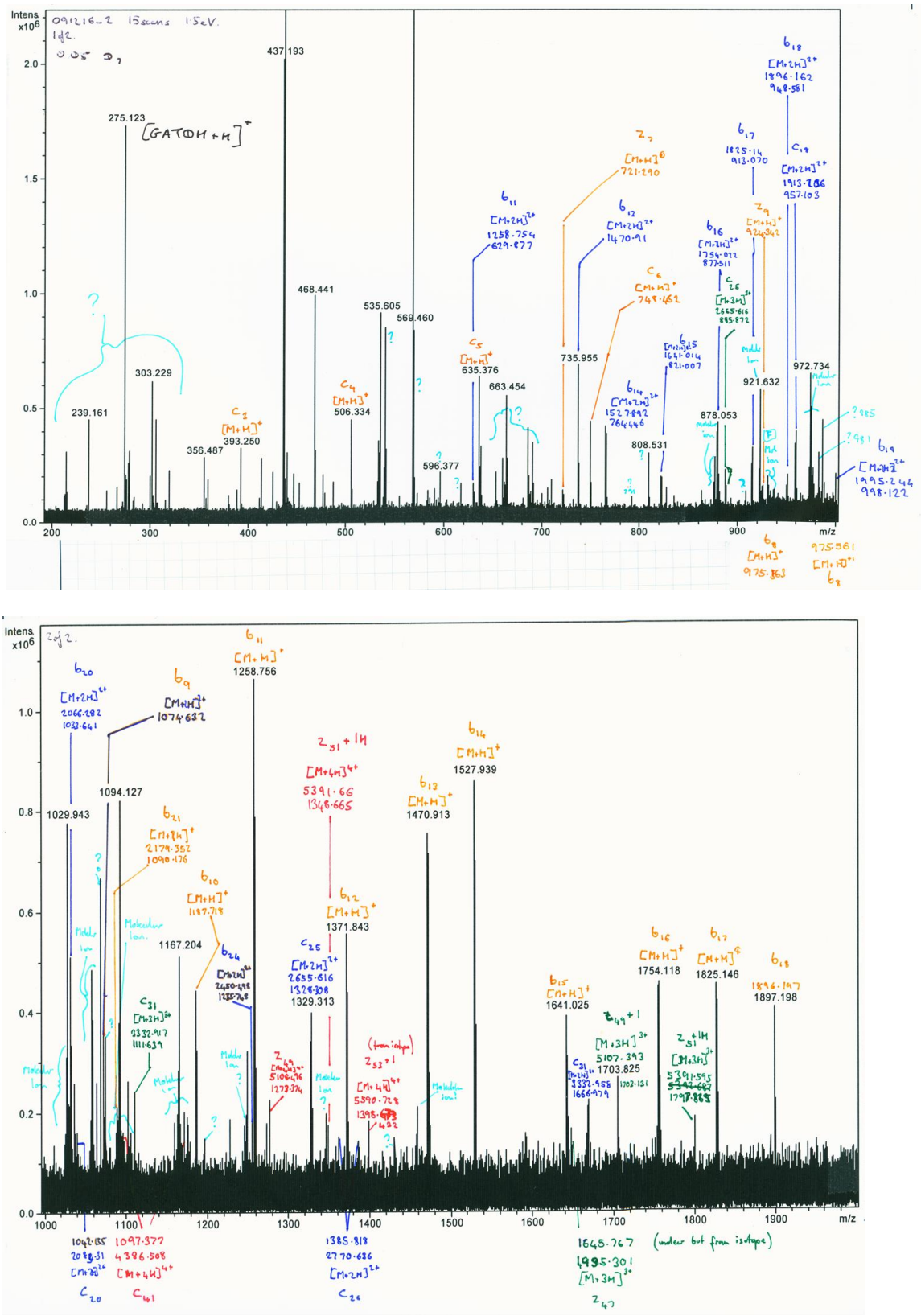


Figure 54 - Manually assigned ECD MS/MS spectrum of PilE-8013 acquired on the Apex III FT-ICR

1.5. Manual Peak Picking and Ion Assignment

Once an automatic method had been developed and tested, a manual approach to peak picking deconvolution and ion assignment was trialled. For the spectral features classed as peaks, charge states were determined by hand and deconvoluted masses calculated. These were then cross referenced against a table of expected fragments for PilE-8013 generated using the online software tool Protein Prospector (http://prospector.ucsf.edu/prospector/mshome.htm). Matched peaks were then manually annotated onto the spectrum (Figure 54) and onto an empty fragmentation map, a digitised version of which is shown in Figure 55.

Figure 55 - Fragmentation maps of ions assigned to PilE-8013 after manual assignment

Manual picking and assignment identified 43 isotopic patterns of which 31 could be assigned to fragment ions representing 29 inter residue cleavages. This resulted in an overall sequence coverage of 18%.

1.6. Comparison of Manual and Automatic Data Analysis

This initial ECD MS/MS spectrum of PilE-8013 provided a relatively simple example upon which the manual and automatic methods of spectral treatment can be compared. Firstly, considering the ions picked by both methods, at $Q_f = 0.3$ the SNAP algorithm picked 50 features as opposed to the 43 picked using the manual method. Most of the extra patterns picked by the automatic method were charge reduced molecular ions or molecular ions with neutral loss that were ignored when peak picking was performed manually.

All of the automatically identified fragment ion peaks were picked by the manual method and thus were manually validated. The manual method did however identify several peaks that were not detected by the automated method. Some of these features, such as the peak at *m/z* 720.283 corresponding to the z_7 ion (Figure 56), are clearly real ions and have several visible isotope peaks. They are most likely ignored by SNAP because the z_7 and $z+1_7$ ions are overlapping and this

distorts the isotope pattern sufficiently for this feature to be discarded. This nicely illustrates a limitation of the automatic approach and highlights an area for algorithm improvement.

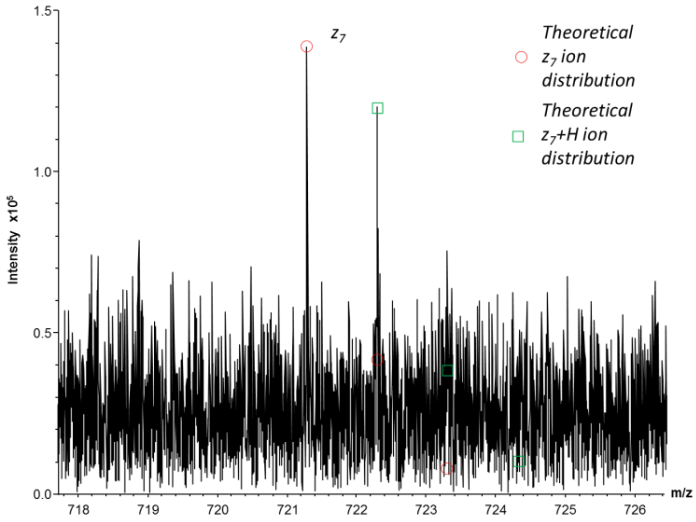


Figure 56 - Overlapping z_7 and $z+1_7$ ions at m/z 721.290 consistently ignored by the SNAP 2.0 algorithm

Even if it may fail to identify some small fragment ions, automatic peak picking and assignment does however offer some additional advantages over its manual counterpart. Manual picking took several hours even with this rather simple spectrum whereas automatic assignment only took a few minutes. It is therefore significantly quicker and is arguably more accurate when the data is complex and accurate monoisotopic mass calculation from large isotope clusters becomes difficult to perform manually. It is also completely objective, with the same criteria applied for each spectral feature no matter its *m/z*, intensity or the complexity of the isotope cluster. Properties such as the quality factor and mass error can also easily be computed and are useful pieces of information to include in scoring systems.

A combination of automatic and manual assignment is therefore advocated for future experiments. In a first step, automatic assignment will be used with low intensity threshold values and a low SNAP threshold. This provides a first approximation of spectral assignment that suffices when the aim of the experiment is simply parameter optimisation. If a higher level of confidence or more complete coverage is required the spectrum should be examined manually for peaks missed by SNAP, especially those that are singly charged, and any important peaks crucial for the assignment of PTM should be further manually validated. It must be highlighted at this point that no matter if the peak picking and ion assignment step is performed manually or automatically, generating the fragmentation maps which are required for data visualisation were especially time consuming as BioTools does not contain a feature for automatic fragment map generation.

1.7. Conclusions from ECD MS/MS of PilE-8013 using a 7T Bruker Apex III FT-ICR MS

The ECD MS/MS experiment performed on the Apex III instrument results in minimal fragmentation of PilE-8013 (sequence coverage 18%) and the production of small fragment ions mostly originating from cleavages close to the N-terminus. No significant fragmentation was observed in other regions of the protein except for several isolated ions corresponding to fragments from either side of the cysteine residues. No ions specific to the minor proteoform of PilE where Ser^{93} is modified with PG could be assigned. Clearly ECD MS/MS on the Apex III does not provide sufficient sequence coverage for PTM localisation and requires further optimisation.

The single scan signal intensity of PilE-8013 was always rather low on this instrument and was consistently below the 1×10^6 threshold required for optimal ECD. Such low intensity was unexpected since the PilE preparation produced protein concentrations around 10 pmol/μL as estimated by SDS-PAGE and the same sample analysed on a Q-ToF mass spectrometer gave a much more intense signal with higher S/N. It was therefore concluded that the major obstacle to achieving more informative ECD MS/MS spectra was insufficient precursor ion intensity, and that this was caused by a lack of instrument sensitivity. Options to overcome this problem are rather limited on the Apex III FT-ICR and, since the DT experiment has a very long duty cycle and a stable spray was difficult to maintain for prolonged periods, poor precursor ion intensity is difficult to compensate for by summing the transients from a large number of experiments. Consequently it was thought that a more sophisticated, later generation FT-ICR instrument may provide better sensitivity and allow more confident evaluation of the ECD MS/MS experiment.

2. Top-Down Analysis of PilE-8013 on a 7 Tesla Bruker Apex Qe FT-ICR Mass Spectrometer

Experimental development was therefore continued on a modified Apex Qe FT-ICR located in the CLIO facility at Université Paris-Sud 11, Orsay, nearby to Ecole Polytechnique (Figure 57). This instrument features several improvements over the Apex III. Notably it has an orthogonal source region and ion funnels for much improved ion transmission and sensitivity. It also has a quadrupole for ion selection and accumulation, and a dedicated collision cell for CID MS/MS. Furthermore the instrument control software has been augmented by a software tool that permits monitoring of the spray current during spectral acquisition.

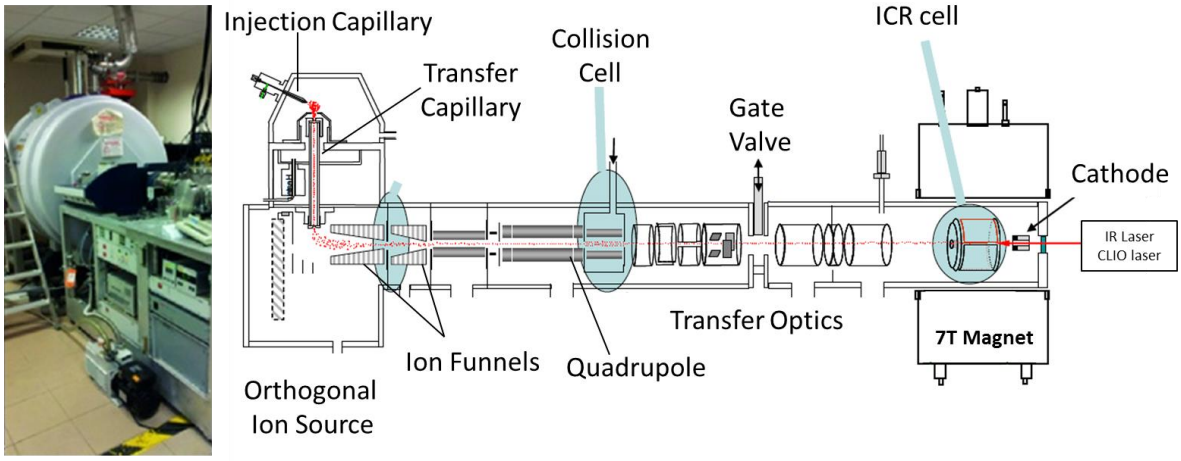


Figure 57 - Photograph of the modified Bruker Apex Qe FT-ICR mass spectrometer at Université Paris-Sud (left) and instrument schematic (right)

2.1. Apex Qe Performance Overview

Sample injection on this instrument is performed using metal-coated glass capillaries in a similar way to the Apex III. However with this updated Apollo II source the spray is performed almost perpendicularly to the transfer capillary entrance. After initial manual optimisation of the source parameters and transfer optics, first with the model protein myoglobin, then with PilE-8013, a MS spectrum of PilE was acquired. The signal intensity on this instrument was approximately tenfold higher than that of the Apex III with only a modest 0.2 s quadrupole accumulation time. ECD MS/MS was therefore trialled on the full spectral range.

After rapidly examining the effect of accumulating multiple transients on protein sequence coverage (both with PilE-8013 and myoglobin), it was concluded that summing of several hundred transients was necessary to produce good quality MS/MS spectra. Acquisition time for the full experiment therefore lasted anything between 5 and 20 minutes. A stable spray is consistently required throughout that time as variation in the spray can drastically affect fragmentation. The spray should therefore be surveyed throughout the acquisition and the current monitoring software was invaluable in this regard.

2.2. Empirical Observations from Variation of Experimental Parameters

Initial results from MS/MS experiments were very encouraging with many more ECD fragments being produced than in analogous experiments on the Apex III FT-ICR. It was therefore decided to explore the effect of different experimental parameters on the precursor ion intensities and ECD fragmentation pattern. Some important general trends can be garnered from visual inspection of the acquired spectra which are summarised in Figure 58. Each panel shows the effect of a different experimental parameter. All ECD MS/MS experiments were performed directly on PilE-8013.

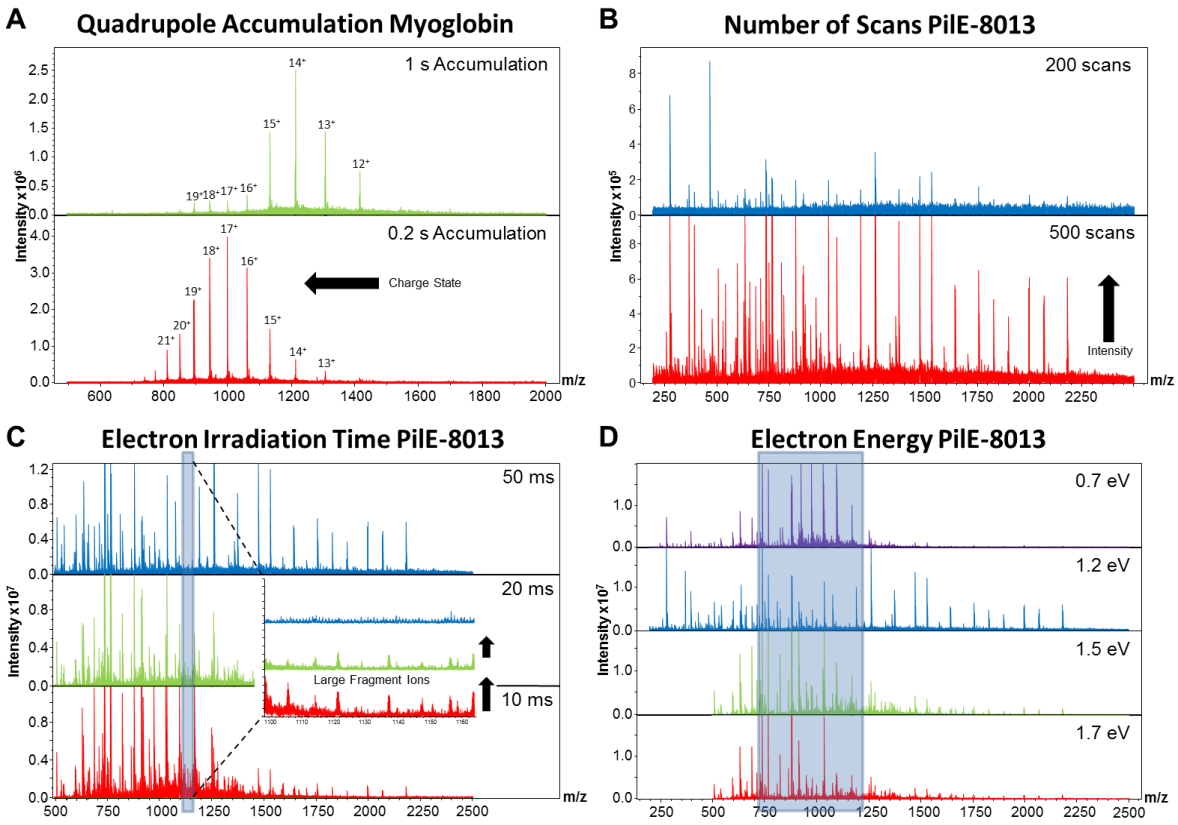


Figure 58 - Visually evident trends in ECD MS/MS spectra performed under different experimental conditions. In panel D the region of interest is highlighted in blue

Panel A shows the effect of an increased quadrupole accumulation time on the charge state envelope of myoglobin. Whilst increasing the accumulation time does increase the signal intensity, it appears to skew the charge state envelope to higher *m/z* values and lower charge states. This charge discrimination is a known phenomenon[10] and resembles that reported by Belov *et al*. for longer accumulation times with low quadrupole pressures[11]. This shifting effect may negate any gain in signal intensity and reduce the overall sequence coverage due to less efficient ECD capture. For this reason MS/MS experiments where precursor ions were accumulated in the external quadrupole for longer times are not presented for PilE-8013.

Panel B clearly shows the effect of transient accumulation, where two ECD MS/MS spectra of PilE-8013 acquired with the same ECD parameters (electron energy 1.2 eV, irradiation time 10 ms) but a different number of summed transients are compared. The first is the result of the accumulation of 200 transients and the second of 500. The same y axis scale is used in both panels. Clearly more ions are visible when more scans have been accumulated (bottom versus top). To extract the maximum information from the ECD spectrum it was considered from this and other experiments that the accumulation of more than 300 transients was therefore necessary.

Panel C illustrates the effect of decreasing the electron irradiation time from 50 to 20, then to 10 ms on fragmentation. At lower irradiation times many high mass, high charge state ions begin to appear from the baseline (see inset panel). These peaks represent large ions issuing from the central region of the protein. Conversely the abundance of small fragments decreases. This observation suggests that with longer irradiation times multiple electron capture events may occur resulting in further fragmentation of previously formed fragment ions. This has the overall effect of favouring the formation of smaller fragment ions that mainly provide coverage of the terminal regions of the protein. This is a key observation. If information is required from the more central regions of the protein, as is the case for PilE-8013, then a short irradiation time may be required in order to prevent the spectral depletion of larger, information rich fragment ions.

Finally in panel D, the ECD fragmentation pattern of PilE-8013 is compared at constant irradiation time (10ms) but with increasing electron voltage from 0.7 eV to 1.7 eV. The spectral differences are more subtle here than for the other three experimental parameters but it seems that the number of peaks in the *m/z* 700-1200 spectral range increases with increasing electron energy (highlighted in blue in the figure). This suggests that using higher electron energies either increased parent to daughter conversion or perhaps favoured the formation of a greater percentage of larger, more highly charged fragment ions. The spectrum acquired at 1.2 eV is interesting as it produced a large series of particularly abundant *b* type ions.

2.3. Effect of Experimental Parameters on Fragmentation

Now these general trends have been outlined, the fragmentation data obtained from PilE-8013 was used to investigate the effect of the ECD parameters; (irradiation time and electron energy) with the aim of finding optimal conditions that produced sufficiently extensive sequence coverage for PTM localisation. ECD MS/MS was performed on the full spectral range after 0.2 s precursor ion accumulation time, at three different electron irradiation times (5, 10 and 50 ms) and with three different electron energies (1.2, 1.5 and 1.7 eV). Out of a possible nine different experimental conditions, six were sampled (see points on Figures 59 &60).

Data analysis was performed using the previously optimised automatic method. Spectra acquired at different irradiation times and electron energies were processed in Data Analysis 4.0 using the previously defined SNAP 2.0 thresholds (I_a = 0, I_r = 1.0×10^{-5}) and fragment ion lists generated. These were then matched against the extended list of *a*, *b*, *c*, *c-H*, *y*, $z^{+\cdot}$, *z+H* and *z+2H* fragment ions in BioTools using a 20 ppm error tolerance. Fragmentation maps were prepared by hand and the number of fragment ions in the central region of the protein between Tyr^{51} and Arg^{111} were counted. Particular attention was paid to this 50 amino acid portion of the sequence because it is known to contain the majority of PTMs and it is fragmentation in this region that will bring us

closer to achieving full PTM characterisation. Analysing the data in this way provided a quantitative method to evaluate this aspect. Overall sequence coverage was also calculated.

The results are presented in the form of 3D contour plots as this enables simultaneous visualisation of the effect of both irradiation time and electron energy. For missing experimental conditions dummy points marked by a ⓧ symbol have been introduced to aid interpretation. The effect of the ECD parameters on sequence coverage in the central region of the protein is shown first in Figure 59 along with a fragmentation map issuing from the experiment that provided the maximum coverage in this region (the central Tyr^{51}-Arg^{111} region is highlighted in blue).

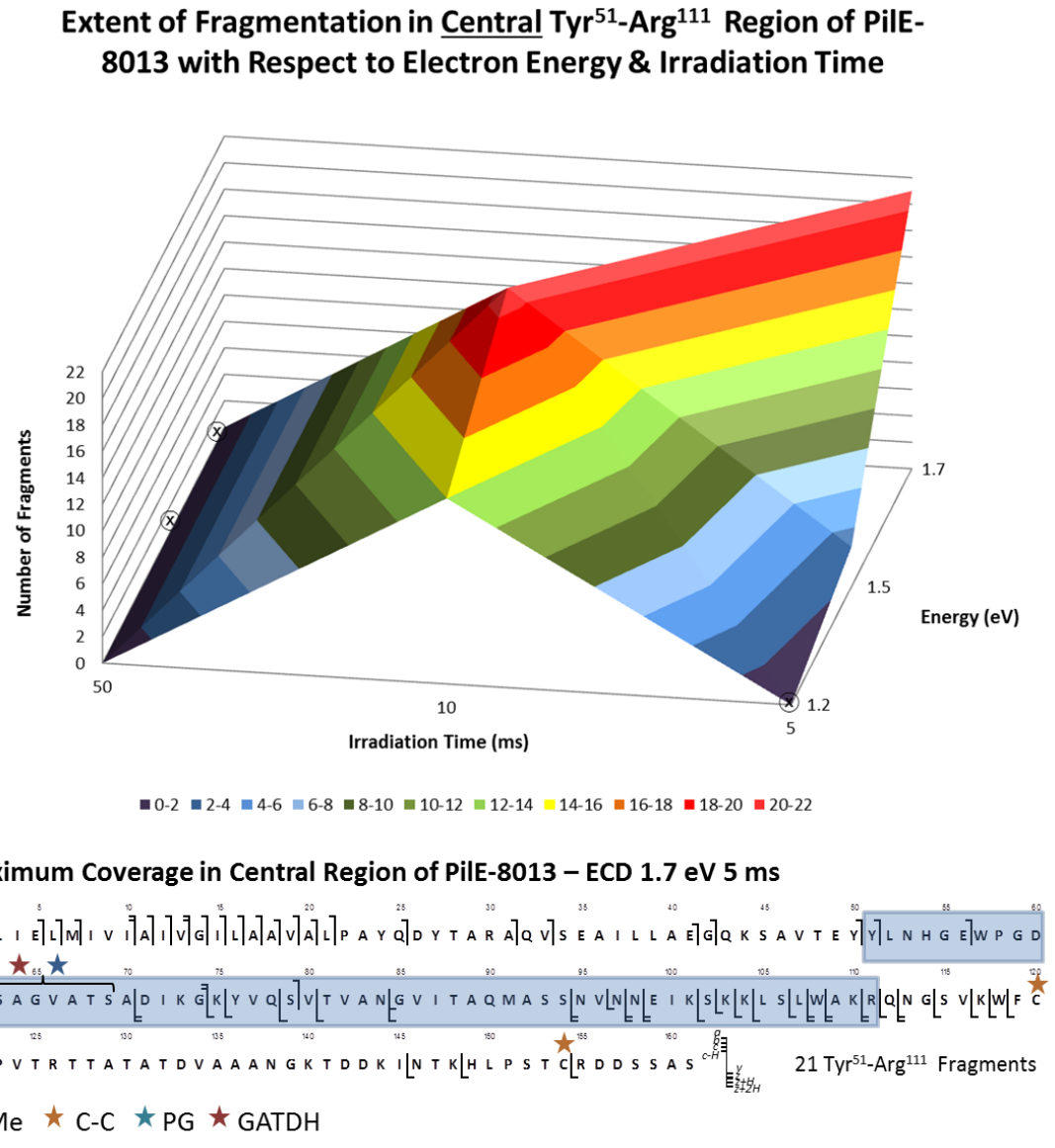


Figure 59 - 3D contour plot of fragmentation in the central region Tyr^{51}-Arg^{111} of PilE as a function of electron irradiation time and energy (top) Fragmentation map resulting from one of the two ECD MS/MS experiments giving the maximum Tyr^{51}-Arg^{111} coverage (bottom)

The contour plot shows that increased sequence coverage in the central region of the protein is clearly favoured by higher electron energies and shorter irradiation times. Two conditions, 5 ms irradiation with 1.7 eV electrons and 10 ms irradiation with 1.5 eV electrons provided a maximum of 21 inter-residue cleavages between Tyr^{51}-Arg^{111}. This tallies with the empirical observations presented previously and supports the idea that a shorter irradiation time produces larger fragment ions. The same data set analysed with respect to the overall sequence coverage is shown in Figure 60.

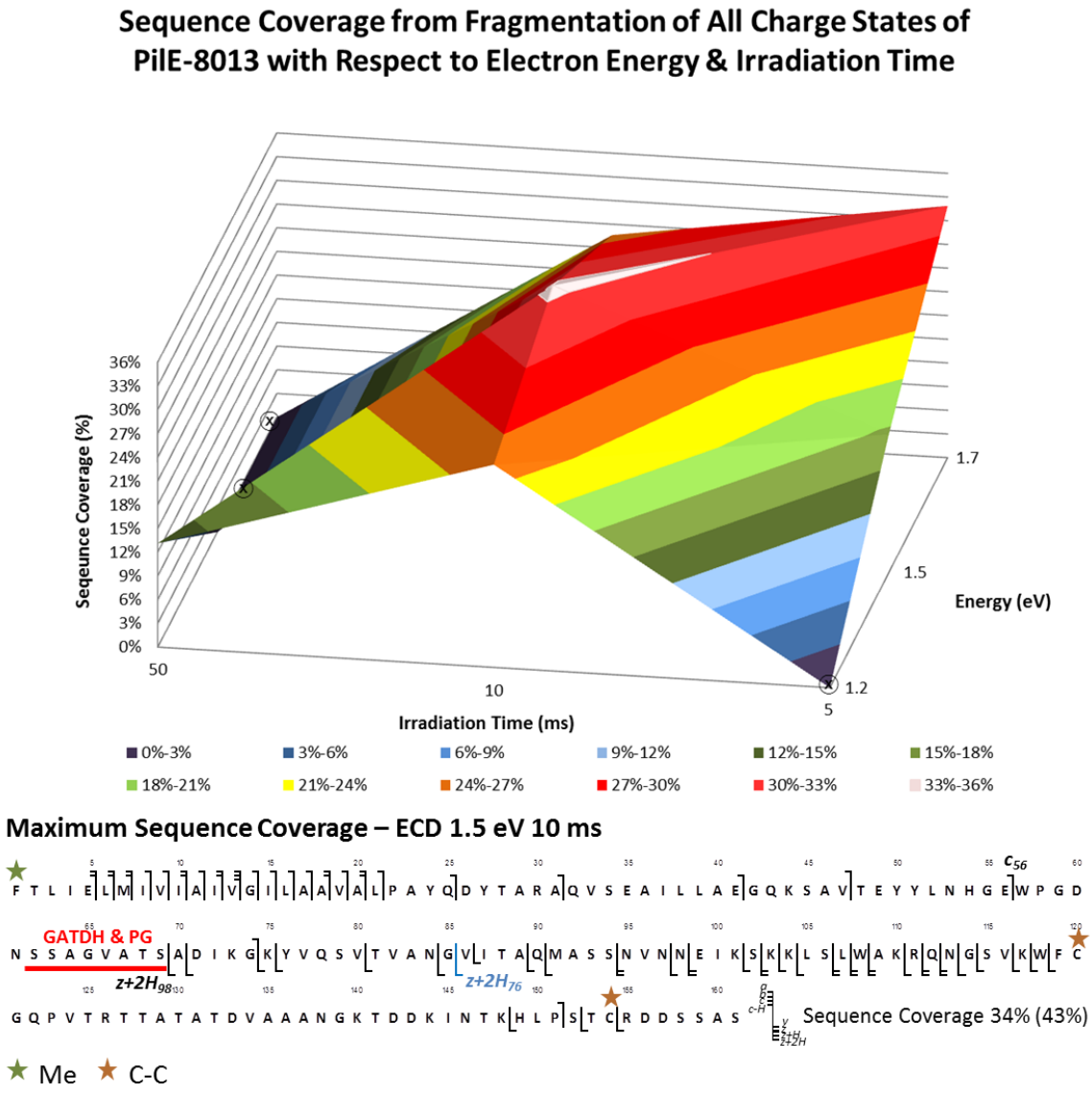


Figure 60 - 3D contour plot of overall sequence coverage of PilE as a function of electron irradiation time and energy (top) along with fragmentation map of the ECD MS/MS experiment giving the maximum sequence coverage (bottom)

Interestingly in this case the maximum coverage of 34% was achieved by 10 ms irradiation with 1.5 eV electrons, although this was extremely close to the 33% achieved at 5 ms and 1.7 eV. It is conceivable that a slightly longer irradiation time results in a greater parent-to-fragment ion conversion and this results in greater overall coverage, although this cannot be verified without the full MS as a reference in each case.

When matching deconvoluted fragment ion masses to theoretical ions of PilE-8013, very little coverage was obtained until the N-terminus was modified by a methyl group and both cysteines "oxidised". Large *c* and *z* type ions series could then be assigned reaching c_{56} and $z+2H_{98}$. Interestingly the *z* ion series is *z+H* type up until Gly^{114} at which point it changes predominately to *z+2H* type. This may be due to deamidation of Asn^{113}. These *c* and *z* type ions series localise the GATDH and PG modifications to the region between Tyr^{57} and Ser^{69} but the lack of fragmentation between these residues precludes more precise identification of the modification sites. When this region is modified by PG and GATDH moieties one additional *c* type ion can be assigned at Gly^{74} and one *b* type ion at Pro^{151}.

2.4. Conclusions from Fragmentation Performed on Apex Qe FT-ICR

Top down ECD MS/MS of PilE-8013 on the Apex Qe FT-ICR provided much more extensive sequence coverage than that achieved on the Apex III (34% versus 18%). Studying the effect of experimental parameters on the fragmentation of this biological sample, including those specifically related to ECD, enabled formulation of several empirical rules for top-down fragmentation to be drawn. These general rules form the basis of our understanding of the top-down experiment and provide a guide for further experimental optimisation.

The fragmentation achieved on all charge states and proteoforms of PilE-8013 was sufficiently extensive to identify the N-terminal methylation and clearly indicated the presence of an intact cysteine bond. The c_{51}, $z+2H_{97}$ and $z+2H_{98}$ ions also enabled the region of the sequence which is modified by PG and GATDH to be narrowed down. Indeed the only residues capable of bearing these *O*-linked modifications are Ser^{62}, Ser^{63}, Thr^{68} and Ser^{69}. However the fragmentation in the central region of the protein is still insufficient for full PTM characterisation. Furthermore only a single ion was found corresponding to the minor proteoform of PilE-8013 bearing an additional PG group at Ser^{93}. This $z+2H_{76}$ ion is highlighted in blue on the fragmentation map in Figure 60.

Since the experiments here were performed on the full spectral range, they are incompatible with the characterisation of individual proteoforms. In addition, close examination of the ECD MS/MS spectra, particularly those resulting in more extensive sequence coverage, revealed many features close to the baseline that were definitely present but not sufficiently intense to be confidently picked and deconvoluted. It was thought that a combination of increased sensitivity and increased

resolution may enable experiments to be performed on individual charge states, and would aid in resolving more highly charged fragments both from the baseline and from overlapping patterns in crowded regions of the spectrum. Both of these features are available on the latest generation of Bruker FT-ICR instrument.

3. Top-Down Analysis of PilE-8013 on a 12 Tesla Bruker solariX FT-ICR Mass Spectrometer

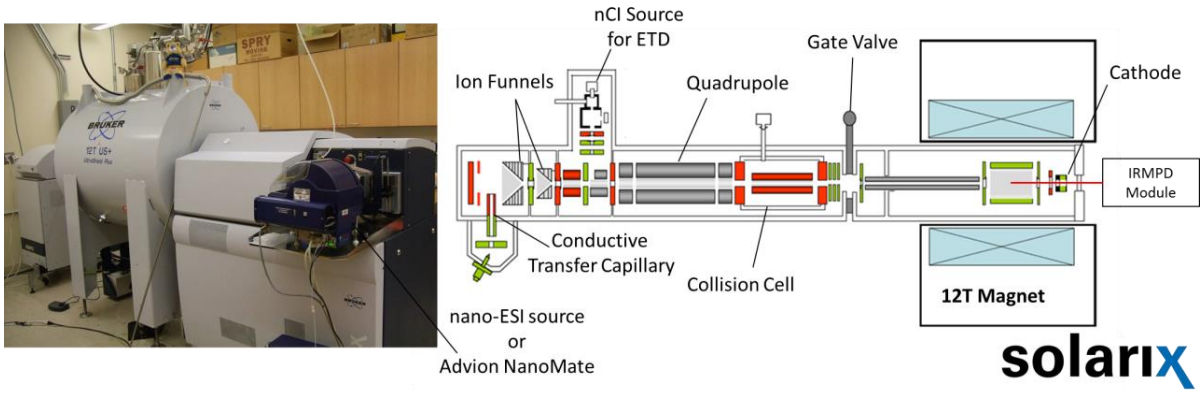


Figure 61 - Schematic of Bruker solariX FT-ICR (right) and photograph of instrument equipped with NanoMate in the Costello lab at Boston University (left)

Experimental optimisation was therefore continued on a 12 Tesla solariX FT-ICR instrument installed in the lab of Prof. Catherine Costello at Boston University, Boston, MA, USA. The solariX platform represents the state-of-the-art Bruker FT-ICR and contains major upgrades to many instrument features including the electronics and acquisition management software. Importantly the nano-ESI source has been overhauled and the ion source transfer optics have again been improved. The latest version includes a conductive transfer capillary. A NanoMate device was available to be coupled with the FT-ICR MS in order to facilitate sample injection. In addition the 12 Tesla magnet provides an almost twofold resolution improvement for the same transient length compared to the Apex Qe and Apex III instruments used in previous experiments and increased sensitivity. This instrument also benefits from a commercial IRMPD module and ETD capability which increase the number of fragmentation modes that can be used for PTM localisation.

3.1. solariX Performance Overview

Sample injection was performed using the NanoMate device or in some instances with the improved nano-ESI source using pulled borosilicate capillaries. As before, preliminary experiments were carried out on myoglobin and the full range of instrument parameters

optimised to maximise ion transfer. From these initial experiments it was clear that the solariX platform provides a much higher signal than the Apex Qe instrument for the same concentration of electrosprayed analyte. Absolute signal intensities were between ten and a hundred fold higher. ECD fragmentation was then trialled. Initial ECD MS/MS experiments performed on myoglobin both on the full spectral range and also on individual charge states were particularly successful, and produced extensive backbone cleavage. Before experiments were commenced on PilE-8013 the higher performance of this instrument enabled some of the fundamental experimental parameters to be re-investigated in greater depth.

3.2. Creation of Software Tool for Ion Assignment and Fragment Map Generation

Given that these experiments required interpretation of many top-down MS/MS spectra and the time-limiting step for data analysis had been identified as data visualisation, there was a clear need to automate this laborious stage of data treatment. Several software tools, both freely available (BUPID top-down) and commercial (ProSightPTM), partially fulfil this requirement but are limited in the type of fragment ions considered and cannot be customised to output specific pieces of statistical information.

To fulfil the desire for these features, a new software tool was created allowing automatic ion assignment and high quality fragment map creation. It was also programmed to output various pieces of statistical data such as sequence coverage and graphs of fragment mass error. A very brief overview of the fragment map creation function is given in Figure 62. This tool proved invaluable for visualisation and analysis of the data generated in the following experiments and has been used to generate all of the fragmentation maps presented in this thesis.

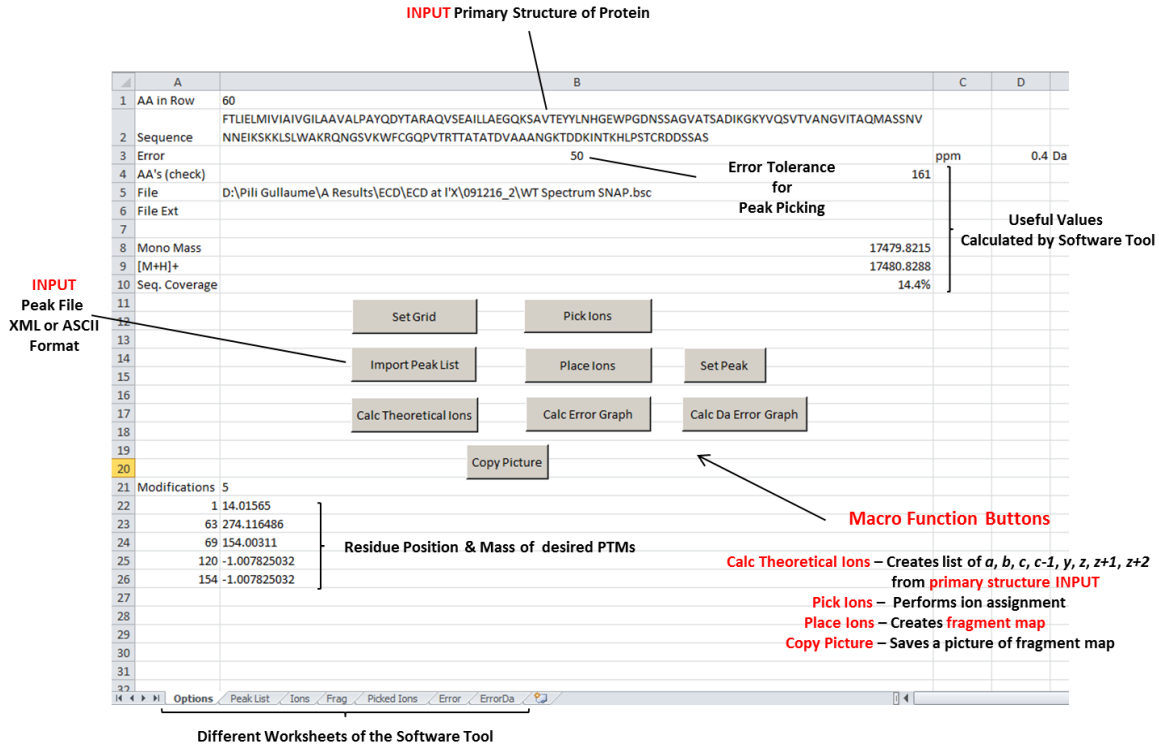


Figure 62 - Screenshot of the main console of a new software tool created for automatic ion assignment and fragment map generation

3.3. Effect of Precursor Ion Intensity on Sequence Coverage

It has already been shown that increased fragment ion intensities can be achieved by increasing the number of summed transients. Increasing the PI intensity should in theory have a similar effect and low abundance fragment ions should therefore be more readily distinguishable from the baseline. It is unclear what exactly this effect will have on sequence coverage and so it is worthy of investigation.

Whilst increasing sample concentration will obviously increase the signal intensity, accumulation of ions in the quadrupole can also be used to artificially increase the abundance of the PI. It has been observed that when accumulation is performed on multiple charge states, charge discrimination may complicate the accumulation of parent ions; however this phenomenon is not evident if a single charge state is selected and accumulated in the quadrupole. With the improved signal on the solariX FT-ICR MS, performing experiments on single charge states of single proteoforms becomes feasible for the first time and one may now ask the questions, "will working with a higher intensity precursor ion or increasing the number of summed transients really maximise sequence coverage? And if so, which will provide the greatest effect?"

To answer these questions a series of ECD MS/MS experiments were performed on the 17^{+} charge state of myoglobin. This charge state was chosen as it was both very abundant and close to the

maximum charge state observed for PilE-8013. Electron energy was fixed at 1.5 eV and irradiation time at 5 ms. The quadrupole accumulation time was altered to achieve a range of precursor ion intensities from 5×10^7 to 8×10^8 and multiple spectra were acquired resulting from different numbers of summed transients. The data were processed automatically in Data Analysis 4.0 with stringent peak picking parameters ($Q_f = 0.3$ and S/N = 2) and the resultant peak lists exported to Bio Tools where *a*, *b*, *c*, *c-H*, $z^{+\cdot}$, *z+H*, *z+2H* ions were assigned using a low error tolerance of 3 ppm. The sequence coverage was calculated manually and plotted against the number of scans at different precursor intensities (Figure 63).

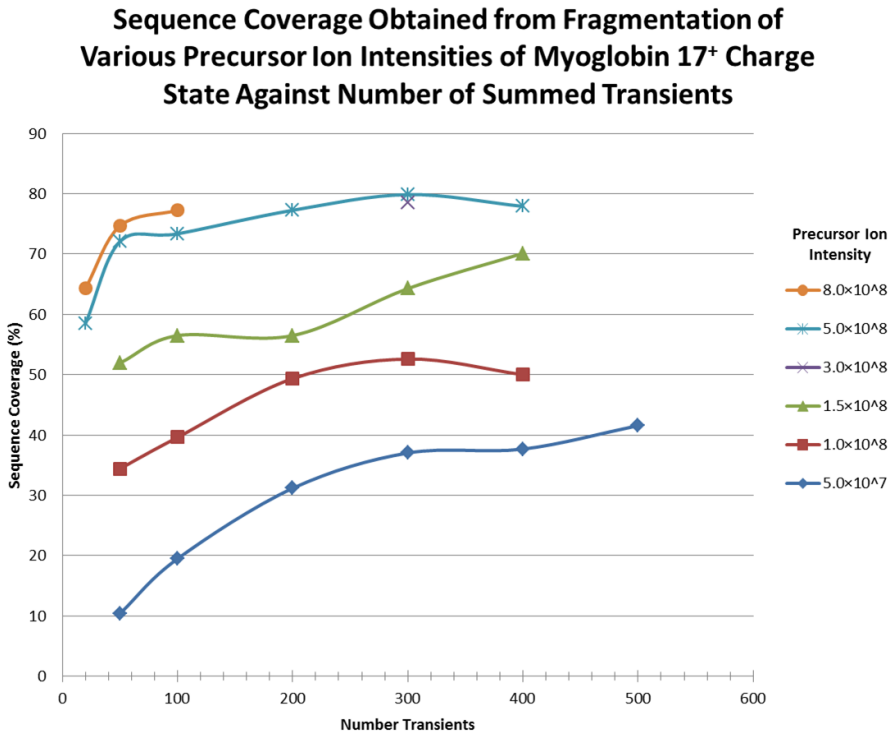


Figure 63 - Plot of sequence coverage obtained from ECD FT-ICR MS/MS of the myoglobin 17^+ charge state at different numbers of summed transients and different precursor ion intensities

As expected, the curves in Figure 63 indicate that for all PI intensities, summing transients always has a positive effect on sequence coverage. The effect is greater for lower intensity precursor ions and the majority of the improvement occurs between 50 and 200 transients, whereafter the benefit of summing scans is less pronounced. On the other hand the initial PI intensity has an extremely strong effect on sequence coverage. The increase in sequence coverage upon increased PI intensity is almost exponential (e^{-1}) regardless of the number of summed scans. Accordingly, when only 50 transients are summed, simply increasing the PI intensity by a factor of ten from 5×10^7 to 5×10^8 increases the sequence coverage 500% from 10.4% to 58.4%. There is little difference between sequence coverage obtained from 5×10^8 and 8×10^8 PI intensity.

These results suggest that the benefits of increasing the precursor ion intensity through quadrupole accumulation far outweigh the summing of multiple transients. Three hundred transients from a PI of 5×10^7 provide a similar sequence coverage to 350 transients from a 1×10^8 PI at a fraction of the total experiment time. A simple explanation for this observation is that with a higher number of precursor ions, the overlap in the ICR cell between the ion cloud and the electron beam is better and thus leads to an increase both in the number of fragments and in their abundance.

Furthermore, extrapolating the curves in Figure 63 suggests that, even if a large number of transients are summed (>1000), sequence coverage will only increase very slowly, or may never increase above a certain level. What is more, this level appears to depend on the initial PI intensity. The results also show that, whatever the PI intensity, summing more than 300 scans has only a modest effect on sequence coverage. The decrease between 300 and 400 scans for PIs of 1×10^8 and 5×10^8 is probably due to fluctuations in the spray throughout the experiment.

When performing experiments on biological samples such as PilE-8013, the concentration is often fairly low and there are limited options for sample concentration outside of the mass spectrometer. This limitation became more significant for samples in which multiple proteoforms are expressed. This means that, upon electrospray, protein ions of interest may not be particularly abundant. In addition the sample volume is not usually very large and achieving a stable spray over an extended timeframe can be difficult. This limits the time window in which the top-down experiment can be performed. Finding the most effective way to improve sequence coverage over the shortest possible timeframe is therefore extremely important.

Signal intensity is the first parameter to choose. If the ion of interest has low abundance, one faces the choice of accumulating the PI then performing MS/MS or accepting whatever the PI value may be and accumulating a large number of transients. More often than not, a compromise between the two is required since increasing both parameters results in too long a duty cycle. The results on myoglobin suggest that for a finite experiment time it is most useful to maximise PI intensity as much as possible and aim to acquire around 300 scans.

3.4. Investigation of ECD Parameters and Charge State for Optimal MS/MS of PilE-8013

The next set of parameters that must be decided are the precursor ion charge state on which the experiment should be performed and the ECD parameters. These are expected to be more protein dependent than the precursor ion intensity and so were investigated using PilE-8013 directly. In a experiment similar to that performed on the Apex Qe, the effects that different ECD parameters have on fragmentation were investigated, but again in more depth and on individual charge states rather than the whole mass range. To show the true effect of the ECD parameters, all other experimental variables including the precursor ion intensity must be kept constant. This required careful experimental design.

Since different charge states naturally have different abundances, a target intensity value that was same for all experiments needed to be chosen and, because Bruker FT-ICR instruments do not benefit from an automatic gain control system (AGC), this meant that quadrupole accumulation time had to be adjusted appropriately. This target intensity needs to be achievable in a reasonable

time frame for all charge states investigated. Given the poor efficiency of quadrupole accumulation on this instrument it therefore cannot be too high, otherwise it will be unattainable for low abundance charge states. However, it must be high enough to achieve sufficient sequence coverage for evaluation of ECD parameters. Therefore 1×10^8 was considered an appropriate compromise that would be achievable for the 16^+-18^+ charge state range of interest and would yield acceptable sequence coverage.

To ensure comparable data between experimental conditions the electrospray must remain stable throughout the course of the entire top-down experiment. Any fluctuations may affect the precursor ion intensity and this has already been shown to have a strong effect on fragmentation. Monitoring of the electrospray current throughout the run is therefore crucial. Thankfully this capability is provided by the NanoMate. As an additional check, measurement of the precursor ion intensity was made before and after the main experimen. Between 5 and 10 separate scans of the parent ion were taken before and 5 after the experiment and the intensities averaged. If variation away from the 1×10^8 target value was greater than ±15% the experiment was repeated (This tolerance also helped compensate for the fact that the absolute intensity value displayed is in fact dependent on charge).

Three hundred transients were summed for each experiment, as this seemed a reasonable compromise between experiment duration and accumulation sufficient to obtain high quality spectra. All acquired spectra were this time internally calibrated using the large series of *b* and *c* type ions that conveniently span the entire *m/z* range. Peak picking parameters were fixed at Q_f = 0.3, S/N = 2 and picked ions were matched against theoretical peaks with a tolerance of 5 ppm and using the home built software tool. Overall sequence coverage and coverage in the central Tyr^{51}-Arg^{111} region was calculated. In some cases repetitions of the same experimental condition were performed. If only a single repeat had been acquired and the resulting sequence coverage was very close, the higher of the two was retained. If repeats totalled three or more, outliers were discarded and the median sequence coverage retained.

The results of 28 experiments performed on different charge states of PilE-8013 under different ECD MS/MS conditions are shown in Figure 64 for Tyr^{51}-Arg^{111} and Figure 65 for overall coverage. In both cases a single experimental condition produced the maximum sequence coverage and the associated fragmentation map is shown in Figure 66.

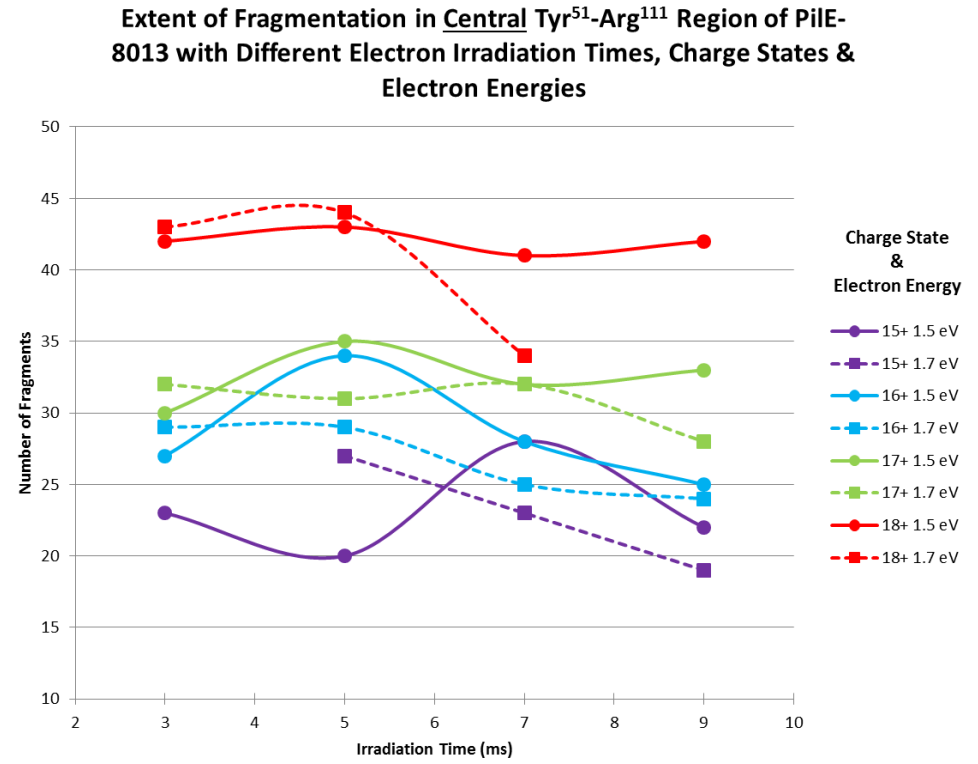


Figure 64 - Plot of coverage obtained from central region of PilE-8013 from ECD FT-ICR MS/MS performed with different electron irradiation times and energies on different charge states

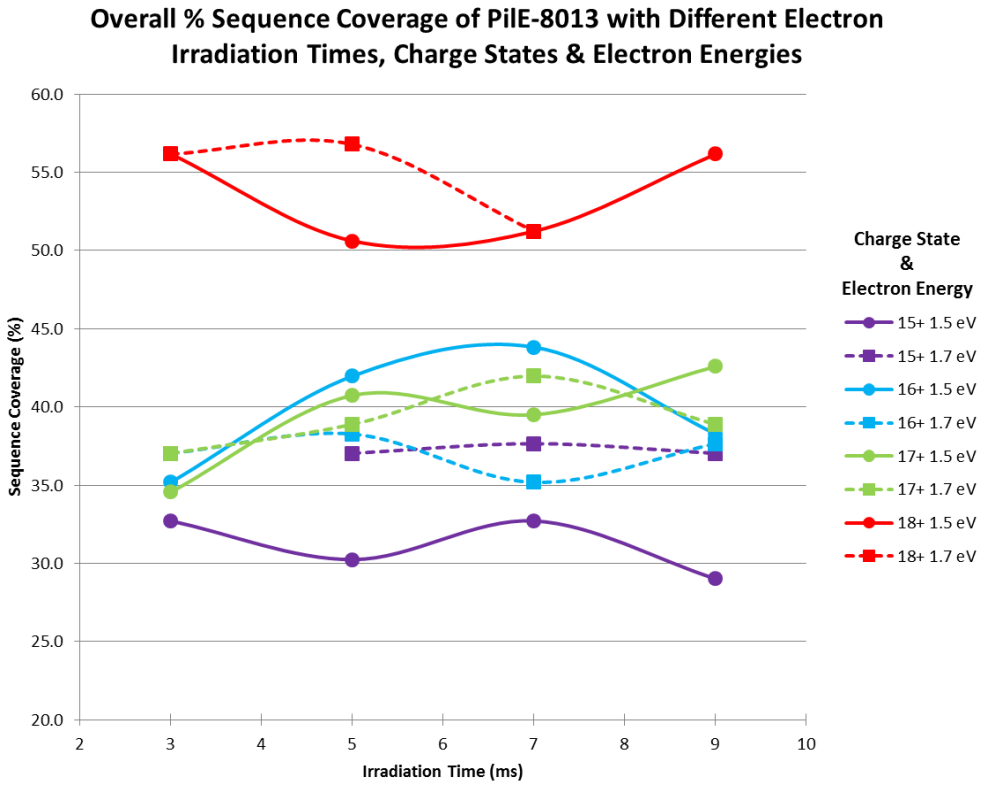


Figure 65 - Plot of overall sequence coverage obtained from ECD FT-ICR MS/MS of PilE-8013 performed with different electron irradiation times and energies on different charge states

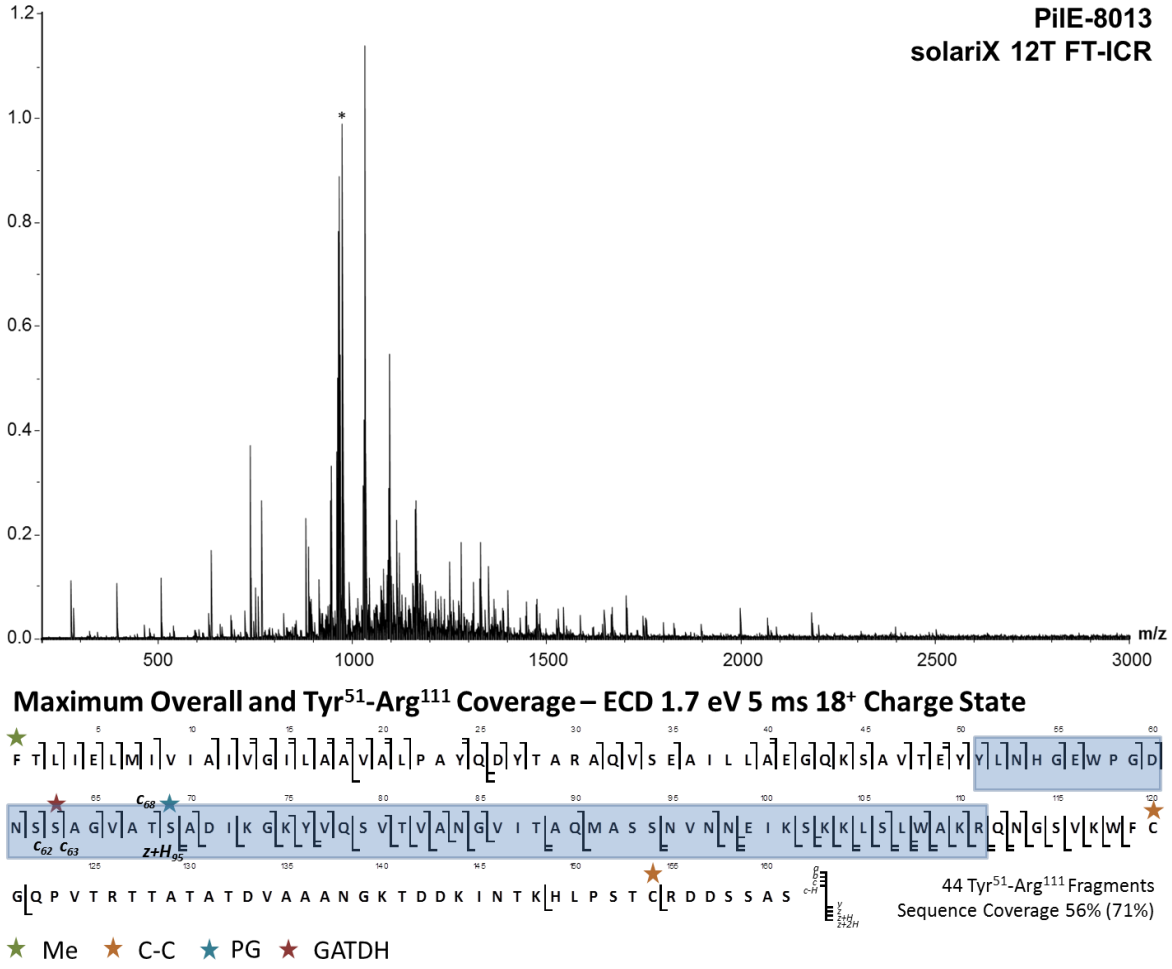


Figure 66 - Fragmentation map of the ECD MS/MS experiment giving both the maximum overall sequence coverage and maximum coverage in the Tyr^{51}-Arg^{111} region

The most important trend that can be taken from the two data sets is that higher charge states provide more extensive sequence coverage. The correlation between sequence coverage, irradiation time and electron energy is less clear and suggests that these parameters have a more subtle effect on fragmentation that is more easily distorted by other experimental variables.

Considering together both the overall sequence coverage and fragmentation in the central Tyr^{51}-Arg^{111} region of PilE-8013, the maximum sequence coverage in this data set was obtained by fragmentation of the 18^+ charge state with 1.7 eV electrons at 5 ms irradiation time. Despite being performed on only a single charge state the sequence coverage is much higher at 56% than that achieved on the Apex Qe with similar ECD parameters. The coverage of the central region of PilE-8013 is also more extensive. The 42 fragments formed here are double that obtained previously. A significant number of *c/z* partner pairs are also now present (double cleavages), and this further increases the confidence of the ion assignment.

Again *b* and *c* type ions are produced from fragmentation of the N-terminus and C-terminal fragmentation is extensive after Cys^{120}. In contrast to the best results obtained the Apex Qe, in this experiment the *c* type ions at c_{62} and c_{63} enable localisation of the GATDH subunit to Ser^{63}. In addition the c_{68}, $z+H_{95}$ and $z+2H_{95}$ ions allow the PG moiety to be placed on Ser^{69}. This assignment is further validated by the c_{65} and c_{66} ions between residues Ser^{63} and Ser^{69}. This set of experimental conditions therefore enables complete characterisation of the major proteoform of PilE-8013 in a single "one-shot" experiment.

There are several additional approaches to furnish different and potentially improved fragmentation when performing top-down MS including supercharging, a "brute force approach" and utilisation of different fragmentation modes. The development of a top-down method for the investigation of PIlE-8013 would not be complete without considering some of these alternatives. The results from these experiments are not included here, but are presented in the materials and methods section, since they did not significantly improve on the results already outlined. The brute force strategy did increase overall sequence coverage to 62% but only after similar experimental optimisation to that described in this chapter.

At the same time that these FT-ICR MS experiments were being performed, the top-down experiment was also being developed on an Orbitrap platform. This will be described in the following section.

4. Top-Down Analysis of PilE-8013 on an Orbitrap Velos Mass Spectrometer

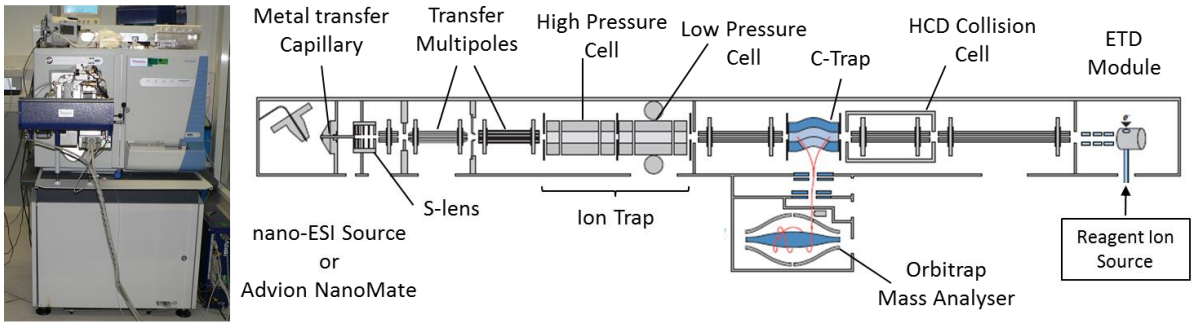


Figure 67 - Schematic of Thermo Orbitrap Velos mass spectrometer (right) and photograph of instrument equipped with NanoMate at the Institut Pasteur (left)

Orbitrap systems are high resolution FTMS instruments that have also shown promising results for top-down characterisation of small proteins. In addition to the FT-ICR approach described in the previous section, the top-down experiment for the analysis of PilE-8013 was concurrently

developed on an Orbitrap platform in order to evaluate any differences between the two technologies and assess their complementarity. Initial experiments were performed on an LTQ-Orbitrap Velos mass spectrometer equipped with an ETD module that is installed in the lab of Prof. Jean-Michel Camadro at the Institut Jacques Monod, Paris. These were later extended on an Orbitrap Velos at the Institut Pasteur, Paris.

4.1. Orbitrap Velos Performance Overview

The Velos system is a hybrid LTQ-Orbitrap mass spectrometer[12] and therefore the top-down experiment is conceptually quite similar to that performed on the Apex Qe or solariX FT-ICR. Following nano-electrospray of the analyte from a metal coated class glass capillary (or NanoMate), ions are then transferred through the S lens and two multipole ion guides before being accumulated in the ion trap (LTQ). Accumulation in the LTQ is directed by an AGC which automatically adjusts the number of trapped charges to a user defined level. The ion packet is then pulsed into the C-trap where they are electrodynamically squeezed into a tight ion packet and transferred to the Orbitrap for detection.

After suitable automatic tuning of the transfer optics, first with myoglobin then with PilE-8013, a well-defined mass profile of PilE was obtained and acceptable signal achieved to evaluate ETD fragmentation. ETD on the Orbitrap Velos is performed in the LTQ. Both analyte ions and ETD reagent (fluoranthene anions) are accumulated in separate regions of the segmented trap. They are then left to react for a certain "interaction" time. This is the major user-defined variable in the ETD experiment as the electron energy is dictated by the chemical properties of the ETD reagent used. In a similar set of experiments to those performed on the Apex Qe the effect of the interaction time on sequence coverage was trialled on different charge states of the major proteoform of PilE-8013. Similar trends in fragmentation were observed as with ExD on the FT-ICR instrument. Sequence coverage is favoured by short (5-15 ms) but not very short (<5 ms) interaction times and higher charge states.

Orbitrap platforms are very different to FT-ICR instruments as they are primarily designed forproteomics applications and hence have a more "user orientated" interface that simplifies routine operation. Settings for many of the component voltages and more advanced features useful for top-down MS are therefore not particularly accessible or only partially supported. After much time exploring the top-down experiment on this instrument, several key prerequisites to achieving high quality top-down data have been discovered and are listed below.

- Precursor ion intensity must be in the 10^5 range (profile) and can be improved by a combination of manual tuning of the source optics, AGC target value and maximum injection time

- Equally the intensity of the reagent ion can be improved by manual tuning of the transfer optics and should be in the 10^6 intensity range. To ensure MS/MS reproducibility the fluoranthene ETD reagent pressure must also be very stable

- Reducing the gas pressure in the HCD cell manually by $0.2\text{-}0.4\times10^{-5}$ torr can improve the intensity of large fragment ions, especially for larger proteins

- If scan averaging is used in LTQ Tune the number of scans should be set to maximum

- Several microscans should be accumulated to improve S/N (this is totally different to scan averaging) and to circumvent the limitations of the software that can only store a limited number of spectra in memory

- The full profile FT mode should be used to improve peak picking of large multi-charged ions that are close to the baseline and small 1^+ and 2^+ fragments where the ^{13}C isotope has low intensity

Once these changes had been implemented the top-down experiment was repeated in order to evaluate the full potential of the Orbitrap Velos for top-down characterisation of PilE-8013. The 17^+ charge state of PilE-8013 was fragmented with 8 ms interaction time and 27 scans, comprising of 50 microscans each, were acquired at a resolution of 60,000 at *m/z* 400. Scans 24-27 were averaged and peaks picked and deconvoluted using the Xtract tool with an S/N cutoff of 2 and a fit factor of 35% (remainder 25%). Ions were assigned using the previously described tool with a peak picking error of 10 ppm. This larger error tolerance was used as it appears charge-space effects shift the masses of the ions in central regions of the spectrum to a greater extent than in an ICR cell and internal calibration was more difficult to perform due to the absence of many singly charged N-terminal ions. The resulting ETD spectrum and fragmentation map are given in Figure 68.

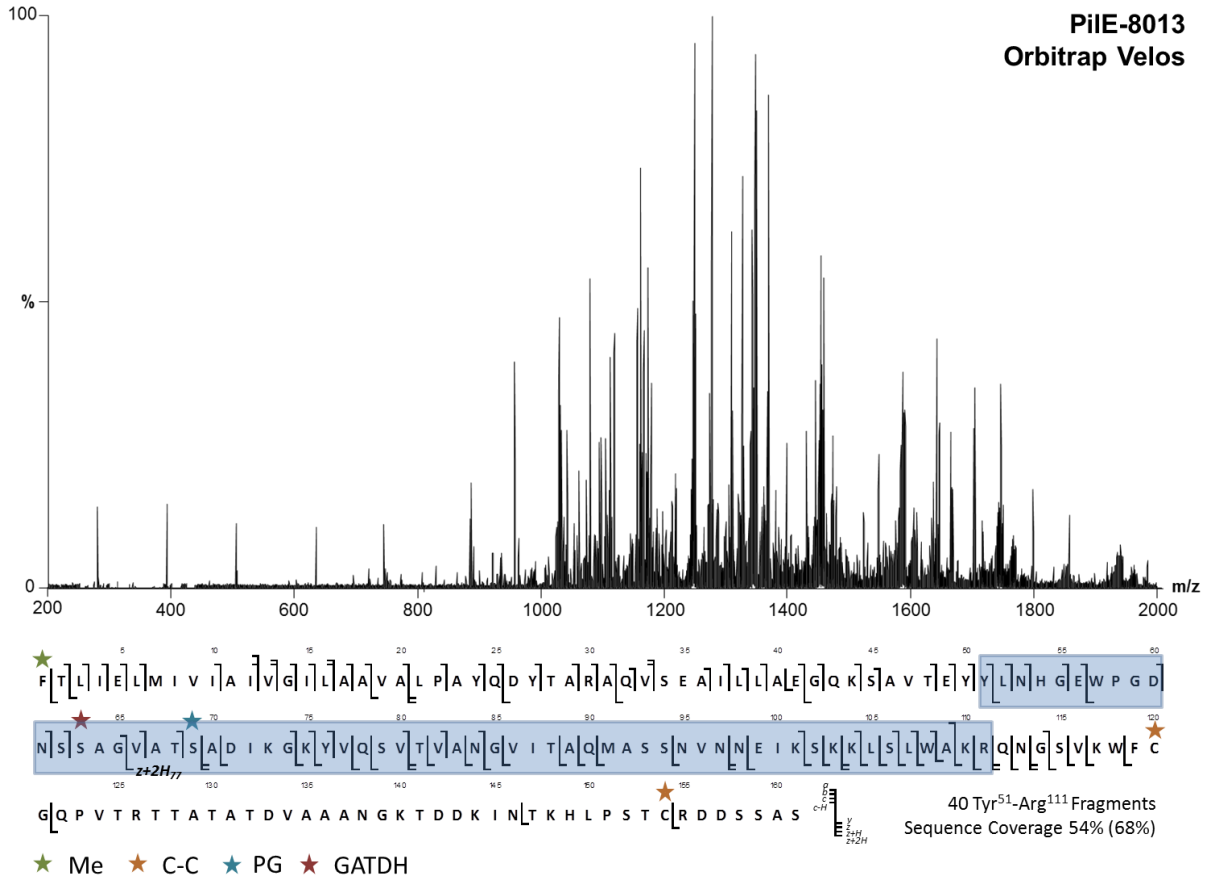


Figure 68 - ETD MS/MS of 17^{+} charge state of PilE-8013 acquired on an Orbitrap Velos. The summed MS/MS spectrum is shown along with a fragmentation map depicting sequence coverage

Once appropriate parameters had been chosen, fragmentation of PilE-8013 was extensive and similar to results from ECD experiments performed on the solariX FT-ICR. Coverage in the central region is slightly less expansive at 40 fragments, as is overall sequence coverage at 54%; however the c_{63}, $c\text{-}H_{65}$, c_{66}, c_{68}, $z{+}2H_{77}$, $z{+}H_{73}$ and $z{+}2H_{73}$ ions allow localisation of the GATDH moiety to be to Ser^{63} and PG group to Ser^{69}. The Orbitrap Velos can therefore also be useful for investigation of PTMs of PilE and, due to the superior ion selection efficiency of the LTQ, this instrument may be particularly useful for characterisation of minor proteoforms.

5. Conclusions from the development of the top-down MS/MS Methodology

5.1. Fragmentation of PilE-8013

Fragmentation of PilE by both ECD and ETD produces extensive backbone cleavage and there are several interesting features of the fragmentation patterns produced. The most striking is probably the absence of fragmentation between the two cysteines Cys^{120} and Cys^{159}. Whilst ECD has been

documented to cause the rupture of disulfide bonds this clearly does not seem to be the case here. This is perhaps due to the gas phase tertiary structure of the PilE ions. Indeed it has been suggested by Simons *et al.* that there must be special proximity between the S-S and C=O-H^+ before electron transfer to S-S and cysteine cleavage can occur[13]. In the case of PilE-8013 the intact S-S does not appear to be problematic as, once parameters have been optimised, sequence coverage over the rest of the protein backbone is very high.

Indeed the intact disulfide bond effectively prevents fragmentation in a 39 amino acid section of the protein that is not of interest for PTM localisation. This vastly reduces the number channels from which fragment ions may be produced. Since the total average fragment ion abundance is proportionate to the number of open channels, keeping the cysteines oxidised is expected to reduce dilution of the MS/MS signal and improve fragment ion intensity. This facilitates the identification of ions in the regions of interest. Cystine reduction would only be necessary if one was interested in exploring the fragmentation of a less structured form of the protein.

Another interesting feature is that fragmentation at the N-terminus often produces an abundant series of *b* and *c* type ions that run until Pro^{22} even when overall fragmentation is not very extensive. The origin of the abundant *b* type ion series is enigmatic. One possibility is that it could be linked to the absence of basic residues in this region (apart from the N-terminus). Indeed Liu and Håkansson have shown that ECD fragmentation of peptides without basic amino acids produces significant numbers of *b* type ions[14]. Alternatively, vibrational activation of PilE ions either during or post ECD could lead to *b* ion formation. Loss of the glycan and the presence of the oxonium ion in the spectrum is a good barometer of the internal energy of PilE. In the ECD spectra GATDH oxonium ions are clearly visible in the low mass range. In the ETD spectra, performed on the Orbitrap neither the GATDH oxonium ion nor the large *b* type N-terminal ion series are visible. This supports the case for vibrational excitation.

Only the regions between residues Thr^{28}-Thr^{48}, Ser^{63}-Gly^{74} and Met^{91}-Lys^{101} seem more impervious to fragmentation. The solution phase structure of PilE suggests that the Thr^{28}-Thr^{48} is both alpha helical and protected by the globular domain. If the gas phase structure is similar this may explain its resistance to fragmentation. The Ser^{63}-Gly^{74} region contains two PTMs which may affect the ExD fragmentation in this region and produce a masking effect. The resilience of the Met^{91}-Lys^{101} region is more difficult to explain. It is part of the β_2-β_3 loop and residues Asn^{97}-Lys^{103} may be involved in a short helix or structured loop which may make this region more difficult to fragment, alternatively electron capture at the sulphur may direct fragmentation to channels that do not involve backbone cleavage, such as loss of $SH^{\cdot}$.

5.2. Mass Spectrometry and Biological Relevance

Building on the initial experiments performed with the Apex III instrument and the general trends elucidated from ECD MS/MS on all charge states and proteoforms using the Apex Qe FT-ICR, the potential of the top-down approach has been finally realised using the state-of-the-art solariX platform. Extensive sequence coverage has also been obtained using an Orbitrap Velos platform.

Development of the top-down MS/MS experiment has concentrated on the major proteoform of PilE using ECD as the fragmentation method on the FT-ICR and ETD on the Orbitrap Velos. Other techniques have been trialled (see Materials and Methods) but not optimised as extensively. AI-ECD requires more optimisation of experimental parameters and as does ETD on the FT-ICR for more extensive sequence coverage. Collisional activation shows interesting fragmentation complementarity and may prove very useful for identifying the glycan present on the protein backbone through either CAD or ISD experiments and could perhaps be used for phosphoform localisation.

A top-down ExD MS/MS experiment has been successfully developed that can be performed on individual charge states and single proteoforms of PilE-8013. For FT-ICR MS/MS, experimental parameters such as the number of transients to accumulate, precursor ion intensity and ECD parameters were carefully optimised enabling extensive fragmentation of the protein backbone. The sequence coverage achieved was more than sufficient for complete PTM localisation of the major proteoform of PiLE-8013. Similar results were obtained though optimisation of experimental parameters on the Orbitrap Velos.

The detailed investigation of ECD parameters at the same precursor ion intensity and comparison with a brute force MS/MS approach (see Materials and Methods) reveals that, whatever the proteoform abundance and precursor ion intensity, appropriate ECD parameter selection is key to the success of the top-down experiment. When the best parameters are selected, sequence coverage in the central region of PilE-8013 similar to that provided by a “brute force” strategy can be achieved with three to four times less precursor intensity on the appropriate charge state. Results obtained on the FT-ICR platform could likely be improved using additional fragmentation modes such as AI-ECD and spectral phasing would definitely improve both resolution and S/N. An improvement in quadrupole efficiency is also necessary if this instrument is to be used for the analysis of low abundance species.

On the Orbitrap Velos, optimisation of the ETD reaction time, performing experiments at sufficiently high precursor ion intensity and increasing the number of acquired microscans are equally important for experimental success. Results are expected to be even better on the more recent Orbitrap models such as the Orbitrap Elite or newly released Fusion since unlike the Velos

they contain a high-field Orbitrap, greater ion capacity for MS/MS, an improved ETD source, spectral phasing as standard, a faster scan speed, much higher resolution and are therefore more adapted to top-down MS/MS[15, 16].

Nevertheless, for both ETD on the Orbitrap Velos and especially ECD on the solariX FT-ICR, the completeness of fragmentation coverage that has been achieved after experimental optimisation is impressive, with almost single amino acid resolution obtained for the majority of the PilE sequence outside of the cysteine bridged region. This bodes well for application of the approach to other pilins, especially those that may be even more highly posttranslationally modified.

Bibliography

[1] T. W. D. Chan and W. H. H. Ip.Optimization of experimental parameters for electron capture dissociation of peptides in a Fourier transform mass spectrometer. *Journal of the American Society for Mass Spectrometry*, **2002**, *13*, 1396.

[2] M. V. Gorshkov, S. H. Guan and A. G. Marshall.Dynamic ion trapping for Fourier transform ion-cyclotron resonance mass spectrometry simultaneous positive-ion and negative-ion detection. *Rapid Communications in Mass Spectrometry*, **1992**, *6*, 166.

[3] Y. O. Tsybin, M. Witt, G. Baykut and P. Hakansson.Electron capture dissociation Fourier transform ion cyclotron resonance mass spectrometry in the electron energy range 0-50 eV. *Rapid Communications in Mass Spectrometry*, **2004**, *18*, 1607.

[4] R. A. Zubarev, N. A. Kruger, E. K. Fridriksson, M. A. Lewis, D. M. Horn, B. K. Carpenter and F. W. McLafferty.Electron capture dissociation of gaseous multiply-charged proteins is favored at disulfide bonds and other sites of high hydrogen atom affinity. *J. Am. Chem. Soc.*, **1999**, *121*, 2857.

[5] R. A. Zubarev, K. F. Haselmann, B. Budnik, F. Kjeldsen and F. Jensen.Towards an understanding of the mechanism of electron-capture dissociation: a historical perspective and modern ideas. *Eur. J. Mass Spectrom.*, **2002**, *8*, 337.

[6] H. J. Cooper, R. R. Hudgins, K. Hakansson and A. G. Marshall.Secondary fragmentation of linear peptides in electron capture dissociation. *International Journal of Mass Spectrometry*, **2003**, *228*, 723.

[7] H. J. Cooper.Investigation of the presence of b ions in electron capture dissociation mass spectra. *Journal of the American Society for Mass Spectrometry*, **2005**, *16*, 1932.

[8] P. B. O'Connor, C. Lin, J. J. Cournoyer, J. L. Pittman, M. Belyayev and B. A. Budnik.Long-lived electron capture dissociation product ions experience radical migration via hydrogen abstraction. *Journal of the American Society for Mass Spectrometry*, **2006**, *17*, 576.

[9] B. J. Bythell.To Jump or Not To Jump? C-alpha Hydrogen Atom Transfer in Post-cleavage Radical-Cation Complexes. *Journal of Physical Chemistry A*, **2013**, *117*, 1189.

[10] M. E. Belov, E. N. Nikolaev, R. Harkewicz, C. D. Masselon, K. Alving and R. D. Smith.Ion discrimination during ion accumulation in a quadrupole interface external to a Fourier transform ion cyclotron resonance mass spectrometer. *International Journal of Mass Spectrometry*, **2001**, *208*, 205.

[11] M. E. Belov, M. V. Gorshkov, K. Alving and R. D. Smith.Optimal pressure conditions for unbiased external ion accumulation in a two-dimensional radio-frequency quadrupole for Fourier transform ion cyclotron resonance mass spectrometry. *Rapid Communications in Mass Spectrometry*, **2001**, *15*, 1988.

[12] J. V. Olsen, J. C. Schwartz, J. Griep-Raming, M. L. Nielsen, E. Damoc, E. Denisov, O. Lange, P. Remes, D. Taylor, M. Splendore, E. R. Wouters, M. Senko, A. Makarov, M. Mann and S. Horning.A Dual Pressure Linear Ion Trap Orbitrap Instrument with Very High Sequencing Speed. *Molecular & Cellular Proteomics*, **2009**, *8*, 2759.

[13] J. Simons.Mechanisms for S-S and N-C-alpha bond cleavage in peptide ECD and ETD mass spectrometry. *Chemical Physics Letters*, **2010**, *484*, 81.

[14] H. Liu and K. Hakansson.Abundant b-type ions produced in electron capture dissociation of peptides without basic amino acid residues. *Journal of the American Society for Mass Spectrometry*, **2007**, *18*, 2007.

[15] D. R. Ahlf, P. D. Compton, J. C. Tran, B. P. Early, P. M. Thomas and N. L. Kelleher.Evaluation of the Compact High-Field Orbitrap for Top-Down Proteomics of Human Cells. *J. Proteome Res.*, **2012**, *11*, 4308.

[16] R. A. Zubarev and A. Makarov.Orbitrap Mass Spectrometry. *Analytical Chemistry*, **2013**, *85*, 5288.

Chapter 5

Investigation of PTM of PilE in Novel Clinical Isolates of Neisseria meningitidis

Full characterisation of the PTM population of PilE has so far been achieved only once for the 8013 strain: these results were presented in Chapters 2 and 3 of this thesis. At the time this work was performed only partial characterisation had previously been achieved on two other Nm strains (8013SB and CS311). Since then the HTII125 isolate has also been partially characterised. Strains 8013SB and CS311 have become classical laboratory reference strains and have been repeatedly cultured in the lab environment over a number of years. It was therefore of great interest to us to examine the PilE PTM population of a wider variety of Nm strains. Recent clinical isolates were an ideal pool from which such pathogenic strains could be selected.

1. Selection of Strains

In collaboration with Dr Marie-Cecile Ploy at the Limoges university hospital, a set of around 50 isolates was obtained from patients treated for sepsis and meningitis during sporadic outbreaks of meningococcal disease in central France. Samples had been collected from a number of medical services including local doctors and various hospital departments, over an 8 year period from 2003 to 2011 and from patients representing a wide age range from 1 day to 96 years old. Interestingly, for some patients multiple samples had been taken and cultured from different locations in the body, most frequently the throat, blood and cerebrospinal fluid (CSF).

Fifteen samples were chosen for further study spanning a time period of 4 years from 2003-2007 (Table 3). Strains belonging to serogroup B were avoided for safety reasons but selected isolates spanned a range of different serotypes and body fluids in order to sample a large variety of clinically relevant strains. In two cases, isolates that had been cultured from the throat, blood and CSF of the same patient were chosen, as it was interesting to determine whether PilE purified from these bacteria would have the same PTM profile.

2. Mass Profiling of PilE from Previously Uncharacterised Clinical Isolates

Dealing with clinically isolated bacteria can present several challenges and the first is to ensure they can be grown effectively in the laboratory environment. After several attempts, three of the fifteen initial isolates (100643, 184148 and 244412) could not be cultured. From the remaining twelve isolates, PilE purification was attempted using the usual protocol. Initial results were not particularly encouraging, however, and acceptable yields of PilE could only be achieved from 427707.

Isolate Number	Serotype	Sample From	Service	Patient DOB	Sample Date
100643	C	Blood	Paediatrics	14/06/1990	10/02/2007
455712	C	Blood	Paediatrics	17/07/1995	23/06/2005
424675	C	Blood	Clinical Haematology	28/02/2000	20/07/2003
256949	C	Blood	Emergency Room	13/08/1935	26/01/2005
219900	C	CSF	Emergency Room	12/01/1958	06/03/2003
427708	C	Throat	Paediatrics	10/04/2002	02/10/2003
427709	C	Blood			
427707	C	CSF			
278534	A	Throat	Emergency Room	27/09/1929	07/02/2006
278536	A	Blood			
278533	A	CSF			
446377	W135	Blood	Paediatrics	19/10/2004	10/12/2004
184148	W135	CSF	Paediatrics	28/08/2005	21/03/2006
218240	Y	Blood	Emergency Room	24/07/1937	20/12/2007
244412	Y	Blood	Emergency Room	09/06/1981	07/06/2004

Table 3 – Initial selection of clinical isolates of *N. meningitidis*. Coloured groups indicate sets of samples taken from different locations of the same patient

In an effort to try to understand why only one of the preparations produced useful amounts of PilE, even after several attempts, any differences in the aspect of the cultures were noted and the size and compactness of pellets at different stages of the protocol were recorded. It was quickly realised that the quantity of cultured bacteria varied considerably among strains. In an initial step to try to omptimise the preparation, the number of bacteria per plate was normalised empirically. Note that despite this measure, the amount of pili produced from different isolates is expected to vary naturally as certain meningococcal strains are known to express more pili than others.

For strains that produced more diffuse pellets, the centrifugation time was increased in order to improve sample purity, and for strains that produced only a small quantity of PilE, the results of several preparations were pooled. After considerable effort, enough PilE to perform MS analysis was obtained from six of the initial fifteen isolates.

2.1.Initial Mass Profiling of Clinical isolates

Mass profiling of PilE was therefore performed by nano-ESI Q-ToF MS (Figure 69). Note that the deconvoluted protein masses are average masses (M_{av}) and are subject to a ±1 Da error. The mass profiling spectra are all fairly complex with multiple peaks representing multiple forms of PilE (Figure 69). In many cases, examination of the raw data suggested that many minor proteoforms were also present, but their ions were too low in abundance to be deconvoluted correctly. In some cases, 219900 for example, the signal was very low and deconvolution proved difficult. In other

cases multiple satellite peaks at ±12-14 Da appeared after deconvolution making it difficult to choose the correct peak in each proteoform cluster and complicating the analysis.

Importantly there appeared to be between three and six major PilE proteoforms present for each clinical isolate. Closer examination of the mass profiles revealed that the approximate mass differences between the major peaks seemed to correspond to different glycan moieties (see Table 3 for masses); DATDH for 278534, GATDH for 219900 and GATDH-Hex for the other profiled strains. In some spectra abundant proteoforms exhibiting a mass difference of ≈112 Da were also present.

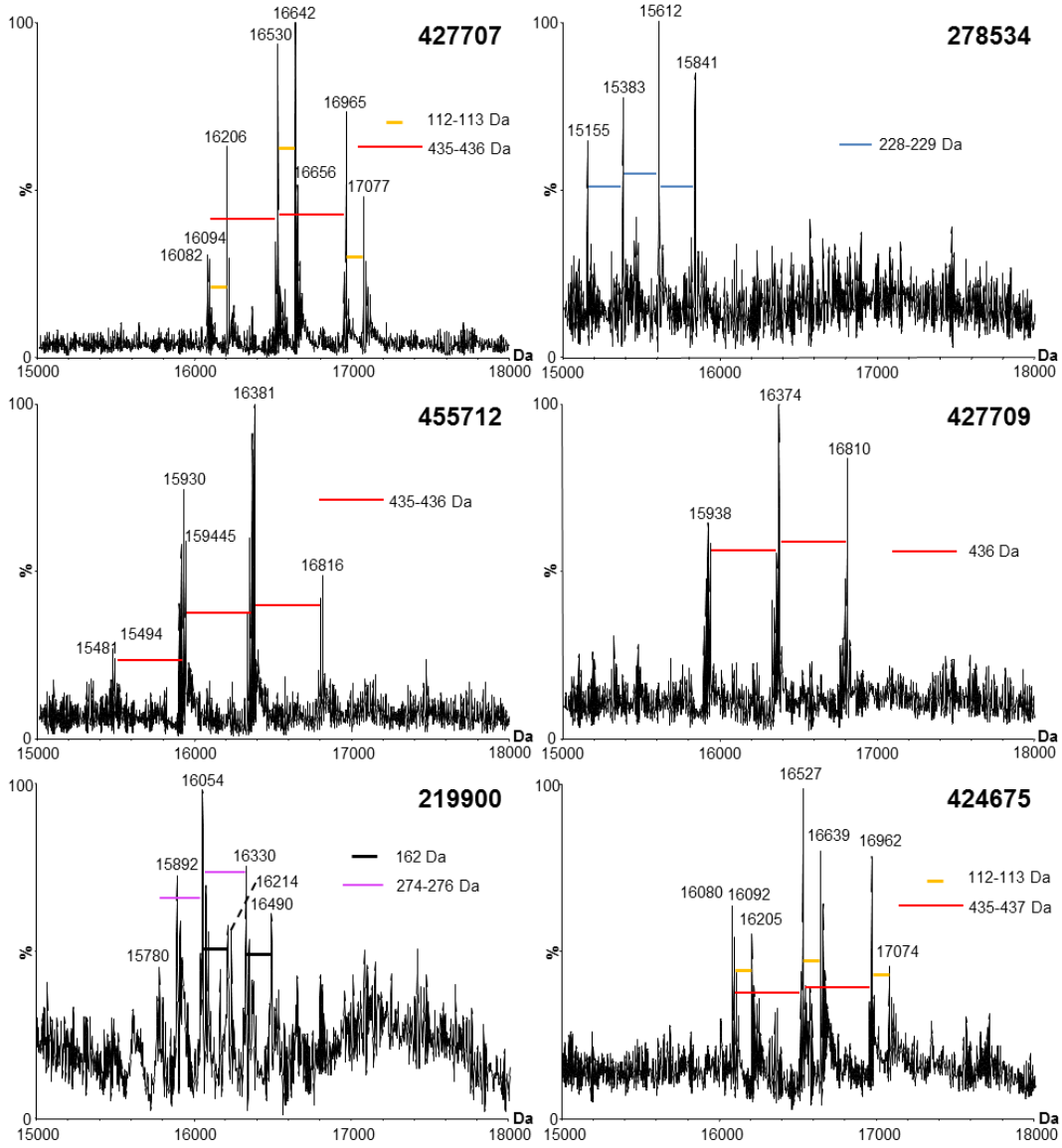


Figure 69 - nano-ESI Q-ToF Mass profiling of PilE purified from clinical isolates

PilE purified from *Neisseria* spp. has only ever been reported with one glycan subunit and never in multiple glycoforms. Since there were many proteoforms present in the sample, it was initially

supposed that sample heterogeneity could at least in part be the origin of the complex mass profiling patterns. Indeed, it is possible that these clinically obtained samples may contain more than one strain of Nm, each harbouring its own complement of PilE proteoforms due to the previously described antigenic variation.

2.2.Isolation and Mass Profiling of Clones of Clinical Isolates

In order to test this hypothesis four distinct colony forming units (CFUs) were selected and cultured from the 427707 isolate; which presented a particularly complex mass profile. Selecting CFUs in this way eliminates any doubt that the purified PilE comes from a single bacterium and its progeny and not a mixture of several strains. These will henceforth be referred to simply as clones and the original isolate the "mother".

Four clones of the 427707 strain were prepared and PilE purified from each. PilE was produced in large quantities and of high purity at the first attempt and gave the mass profiles shown in Figure 70.

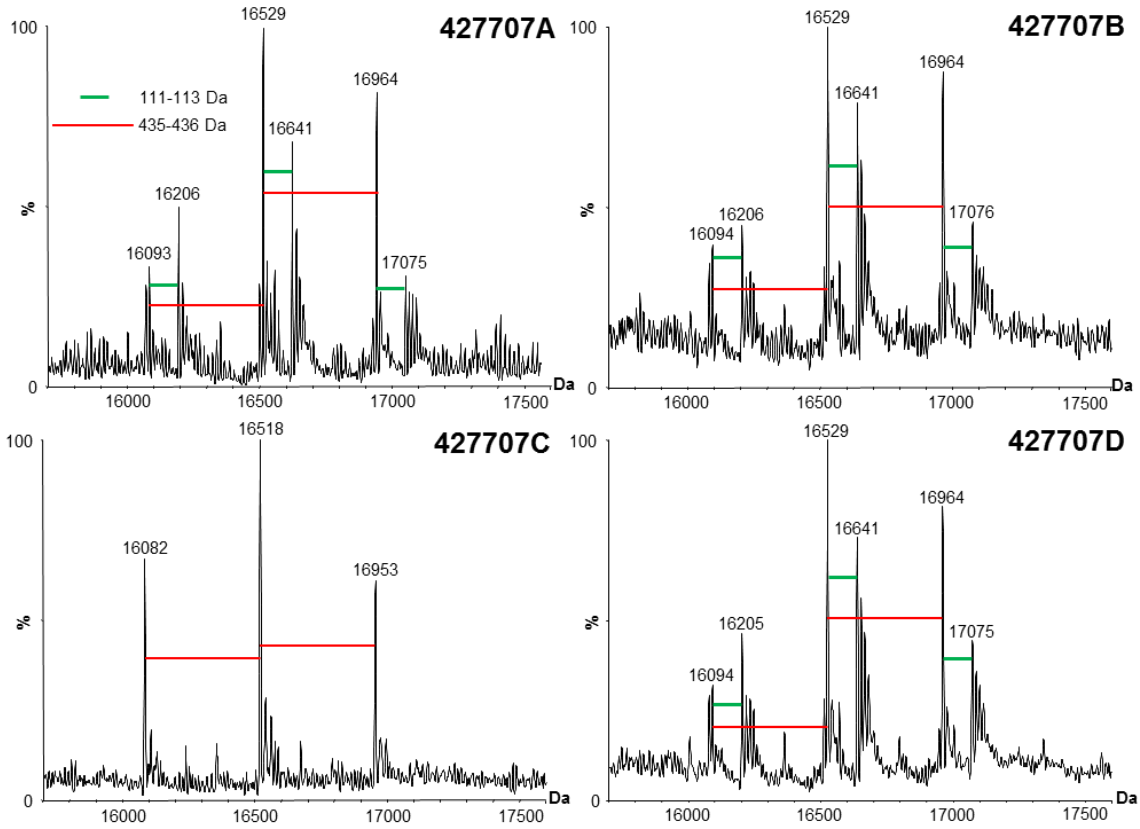


Figure 70 - nano-ESI Q-ToF mass profiles of PilE extracted from clones A-D of isolate 427707

PilE from clones A, B and D produced similar mass profiles greatly resembling that obtained from the mother. However PilE from clone C produced a markedly different profile with only three major peaks rather than the six found in the other clones. Given this apparent heterogeneity, it

was of interest to determine whether similar differences were observed between clones cultured from other isolates.

Clones were therefore prepared, PilE purified and mass profiling performed for isolates 455712A-D (Figure 71), 427709A-D (Figure 72), 424675A-D (Figure 73) and 278534A-D

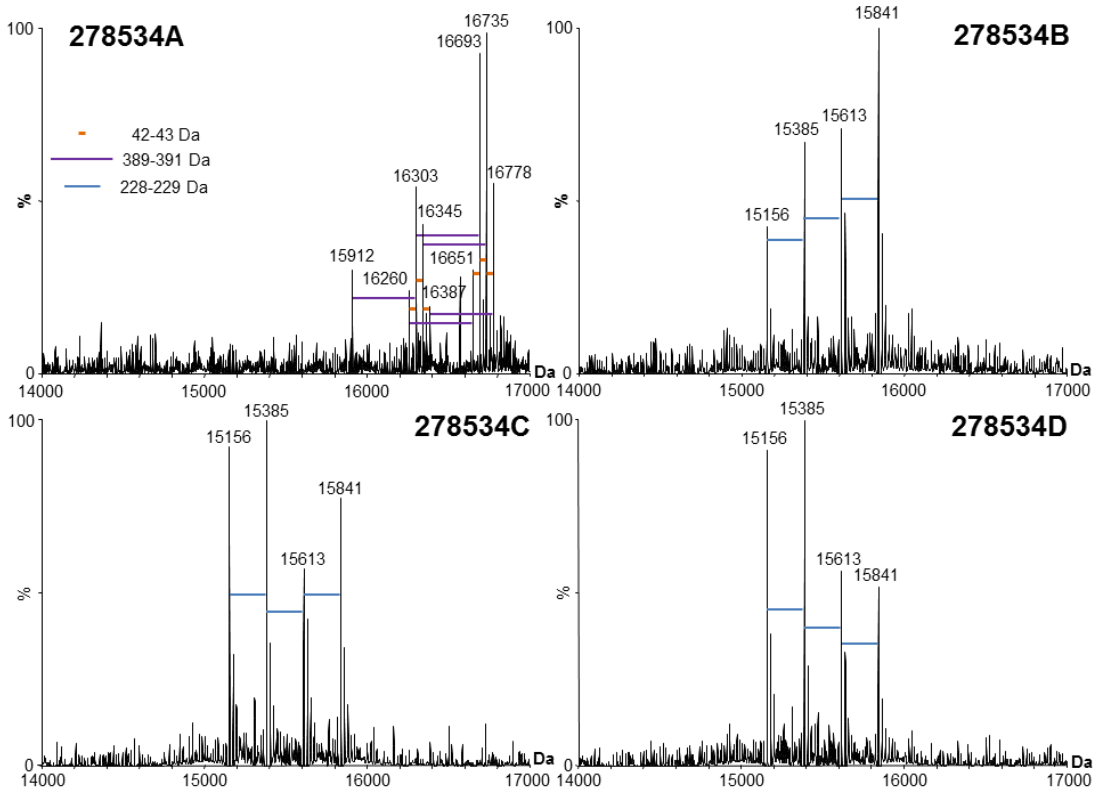


Figure 74). Clones of 219900 were prepared and mass profiling performed; however the protein signal was too weak to provide exploitable deconvoluted spectra and this isolate will be removed from the subsequent analysis.

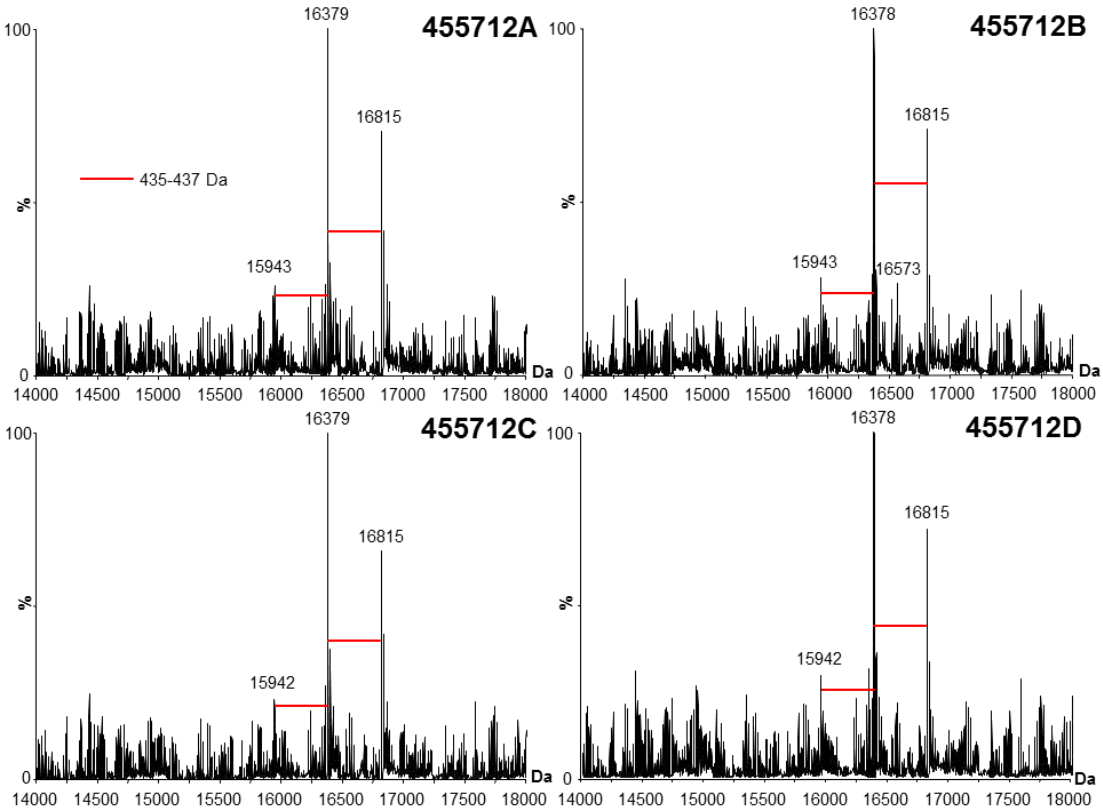


Figure 71 - nano-ESI Q-ToF mass profiles of PilE extracted from clones A-D of isolate 455712

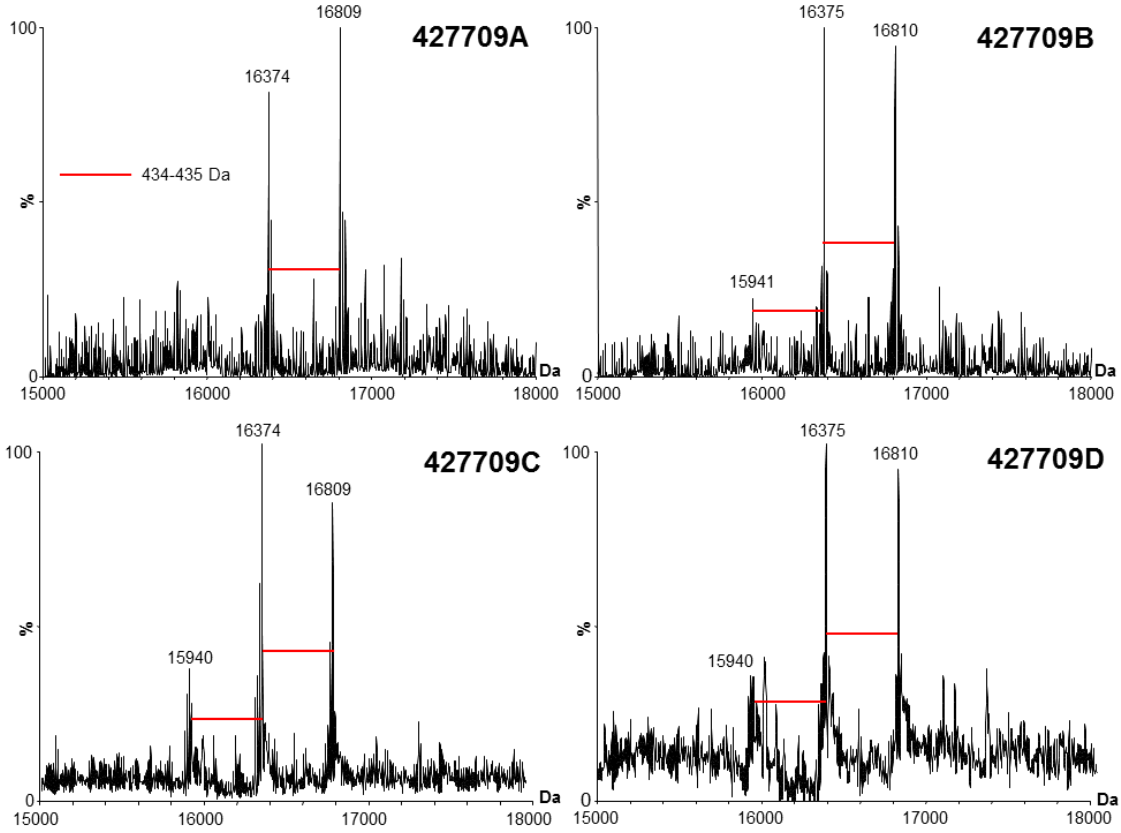


Figure 72 - nano-ESI Q-ToF mass profiles of PilE extracted from clones A-D of isolate 427709

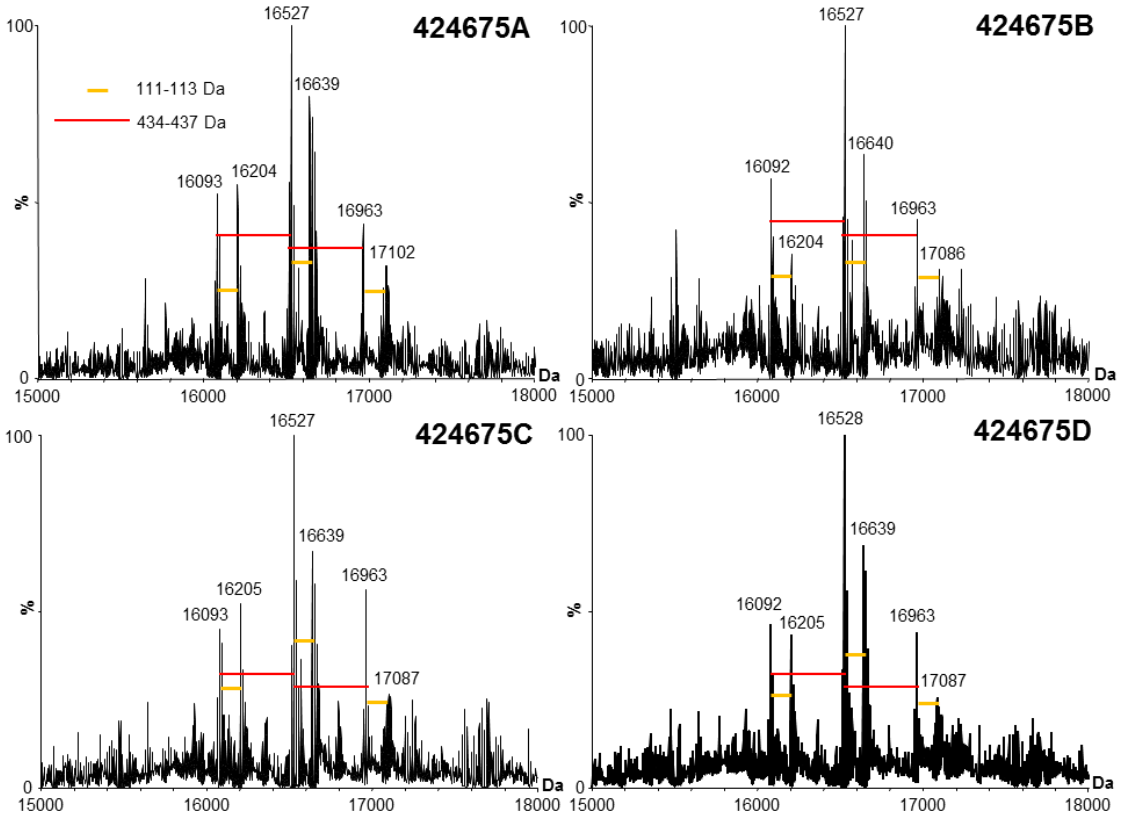


Figure 73 - nano-ESI Q-ToF mass profiles of PilE extracted from clones A-D of isolate 424675

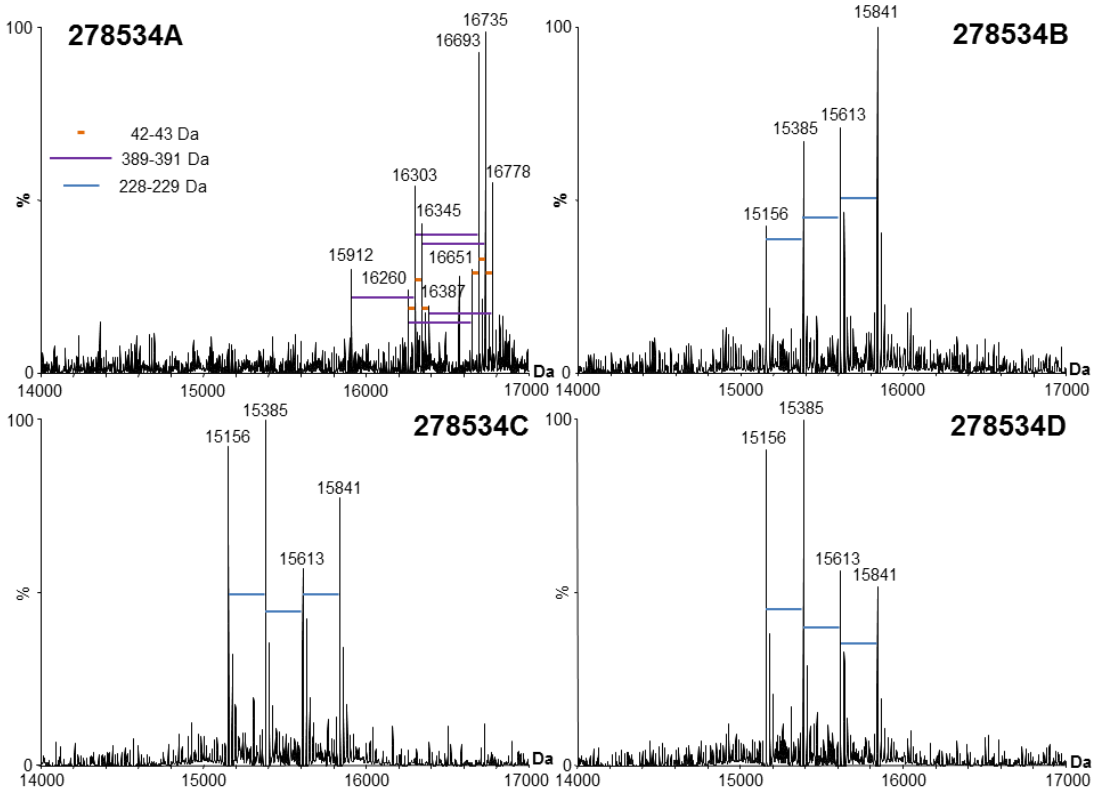


Figure 74 - nano-ESI Q-ToF mass profiles of PilE extracted from clones A-D of isolate 278534

For isolates 455712, 424675 and 427709 mass profiling of all four clones produced spectra that were similar to their respective mother and no intra-strain variation was detected. For 278534 clones B, C and D resembled the mother whereas clone A exhibited a vastly different profile. Of the five remaining isolates, only PilE purified from 278534 and 427707 exhibited different intra-strain mass profiles.

3. Sequencing of the *pilE* Gene and Analysis of Mass Profiles from Clones

In order to investigate this observed variation further and determine whether it was linked to previously described phase variation of the genes responsible for PTM, the primary structure of PilE was required. The *pilE* gene was therefore sequenced for all "A clones" where the mass profiles are similar, both A and C clones for 427707 and both A and D clones for 278534 (Table 4). In each case no intra-strain variation in *pilE* sequence was observed. This is important as it confirmed that the different mass profiles for clones prepared from the same mother were not simply the result of variation of the PilE primary structure.

What is more, only two different PilE were expressed by all seven isolates (including 219900) (Table 4). This strongly suggested that the different mass profiles observed for strains with the same PilE sequence (compare 278534 to 424675 for example), are due to differences in the PTM status of PilE.

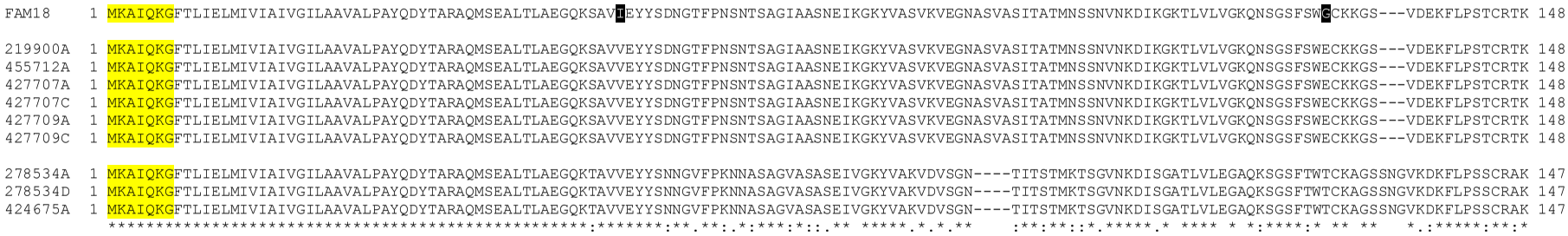


Table 4 - Sequence of PilE expressed by FAM20 and clustal type alignment[1, 2] of the *pilE* gene amplified from clinical strains of *Neisseria meningitidis*. Leader sequence highlighted in yellow

PTM	M_{av} Increase (Da)	M_{mono} Increase (Da)	Oxonium Mass (Da)	Elemental Composition
DATDH	**228**	228.111	229.252	$C_{10}H_{16}N_2O_4$
DATDH-*O*Ac	**270**	270.122	271.129	$C_{12}H_{18}N_2O_5$
DATDH-Hex	**390**	390.164	391.171	$C_{16}H_{26}N_2O_9$
DATDH-Hex-*O*Ac	**432**	432.174	433.182	$C_{18}H_{28}N_2O_{10}$
DATDH-Hex-Hex	**552**	552.217	553.224	$C_{22}H_{36}N_2O_{14}$
DATDH-Hex-Hex-*O*Ac	**594**	594.227	595.234	$C_{24}H_{38}N_2O_{15}$
GATDH	**274**	274.116	275.124	$C_{11}H_{18}N_2O_6$
GATDH-*O*Ac	**316**	316.127	317.134	$C_{13}H_{20}N_2O_7$
GATDH-Hex	**436**	436.169	437.177	$C_{17}H_{28}N_2O_{11}$
GATDH-Hex-*O*Ac	**478**	478.180	479.187	$C_{19}H_{30}N_2O_{12}$
GATDH-Hex-Hex	**598**	598.222	599.229	$C_{23}H_{38}N_2O_{16}$
GATDH-Hex-Hex-*O*Ac	**640**	640.233	641.240	$C_{25}H_{40}N_2O_{17}$
Phosphate	**80**	79.966	-	HO_3P
Phosphoglycerol	**154**	154.003	-	$C_3H_7O_5P$
Phosphoethanolamine	**123-124**	123.009 (124.016)*	-	$C_2H_6NO_3P$ (+H)*
Phosphocholine	**166**	166.063	-	$C_5H_{13}NO_3P$

Table 5 - Table of reported PTMs of PilE with expected monoisotopic mass addition upon modification and elemental composition. Oxonium ion masses are given for sugars. *Phosphoethanolamine carries a primary amine which may be protonated. The mass of the protonated from is given in parentheses

4. PTM of PilE from Clinical Isolates

A great deal of information on the possible PTM state of PilE can be obtained from the initial mass profiling data. First of all, the measured M_{av} of PilE can be compared with that expected from the genome. In every case the difference is very large, much greater than previously observed for PilE and suggests extensive PTM even for the lowest mass proteoform (Table 6).

Secondly it appears that PilE is expressed in multiple proteoforms. The mass difference between observed proteoform peaks can be calculated and cross referenced against the mass differences expected from reported PTMs of PilE to identify the PTMs on these higher mass proteoforms. A summary of this is given in Table 6. Whilst this approach does not completely define the PTM complement of each proteoform, it can give us some clues as to the type of modifications present.

Isolate	Mass Difference $M_{av(exp)} - M_{av(genome)}$ (Da)	Δ_{mass} Between Proteoforms (Da)	Proposed Glycan Modification
219900A	888	275	GATDH
455712A	1050	436	GATDH-Hex
427707A	1202	435	GATDH-Hex?
427707C	1190	435	GATDH-Hex?
427709A	1048	435	GATDH-Hex?
278534A	1379	390, 432	(*O*-Ac)DATDH-Hex
278534D	623	228	DATDH
424675A	1547	435	GATDH-Hex

Table 6 - Selected isolates and predicted glycan modifcation from mass profiling. Mass difference is for lowest mass proteoform

219900 is likely to be modified with GATDH or GATDH-Hex. Strains 427707A-D, 427709A-D, 424675A-D and 455712A-D exhibit differences of around 435 between major peaks and are likely to be modified with GATDH-Hex. Differences of 112 Da between major and minor peaks in these strains are difficult to explain (other than CH_2 x 8) but as mentioned previously they could possibly be 124 Da and this is the expected mass of PE. 278534B,C,D is probably modified by DATDH and 278534A with either DATDH-Hex or DATDH-*O*Ac. For 278534A the peaks exhibiting mass differences of around -42 Da could be due to sample degradation and loss of *O*-acetylation in the form of acetic acid. All clones from the same mother expressed the same core glycan (DATDH or GATDH); however phase variation appears to be present between some clones, 278534A, B, C compared to 278534D for example.

Taken together these observations present strong evidence that PilE from these isolates is heavily modified by multiple glycans. However, mass profiling of the isolates by Q-ToF MS only produces low resolution mass data. This is insufficient to accurately calculate the mass difference between the observed and theoretical protein masses or the exact mass difference between observed proteoforms. High resolution mass data is required to enable further conclusions to be drawn about the PTM status of these isolates, including the large discrepancy between theoretical and observed mass values.

Several of the strains were therefore subjected to high resolution mass measurement and top-down MS/MS in order to provide an accurate protein mass and comprehensive PTM characterisation, including site localisation for each of the observed major proteoforms. This was first performed on the 278534D strain where the top-down and bottom-up methodologies were compared.

5. Deep Characterisation of PilE-278534D

5.1. Accepted Article – "Complete Post-Translational Modification Mapping of Pathogenic *N. meningitidis* Pilins Requires Top-Down Mass Spectrometry"

RESEARCH ARTICLE

Complete posttranslational modification mapping of pathogenic *Neisseria meningitidis* pilins requires top-down mass spectrometry

Joseph Gault[1,2], Christian Malosse[1], Silke Machata[3,4], Corinne Millien[3,4], Isabelle Podglajen[5], Marie-Cécile Ploy[6], Catherine E. Costello[7], Guillaume Duménil[3,4] and Julia Chamot-Rooke[1]

[1] Structural Mass Spectrometry and Proteomics Unit, Institut Pasteur, CNRS UMR 3528, Paris, France
[2] Laboratoire des Mécanismes Réactionnels (DCMR), Département de Chimie, École Polytechnique, CNRS, Palaiseau, France
[3] INSERM U970, Paris Cardiovascular Research Center, Paris, France
[4] Université Paris Descartes, Faculté de Médecine Paris Descartes, Paris, France
[5] Service de Microbiologie, Assistance Publique-Hôpitaux de Paris, Hôpital Européen Georges-Pompidou, Paris, France
[6] INSERM UMR1092, Faculté de Médecine, Université de Limoges, Limoges, France
[7] Mass Spectrometry Resource, Department of Biochemistry, Boston University School of Medicine, Boston, MA, USA

Received: September 4, 2013
Revised: September 4, 2013
Accepted: October 11, 2013

In pathogenic bacteria, posttranslationally modified proteins have been found to promote bacterial survival, replication, and evasion from the host immune system. In the human pathogen *Neisseria meningitidis*, the protein PilE (15–18 kDa) is the major building block of type IV pili, extracellular filamentous organelles that play a major role in mediating pathogenesis. Previous reports have shown that PilE can be expressed as a number of different proteoforms, each harboring its own set of PTMs and that specific proteoforms are key in promoting bacterial virulence. Efficient tools that allow complete PTM mapping of proteins involved in bacterial infection are therefore strongly needed. As we show in this study, a simple combination of mass profiling and bottom-up proteomics is fundamentally unable to achieve this goal when more than two proteoforms are present simultaneously. In a *N. meningitidis* strain isolated from a patient with meningitis, mass profiling revealed the presence of four major proteoforms of PilE, in a 1:1:1:1 ratio. Due to the complexity of the sample, a top-down approach was required to achieve complete PTM mapping for all four proteoforms, highlighting an unprecedented extent of glycosylation. Top-down MS therefore appears to be a promising tool for the analysis of highly posttranslationally modified proteins involved in bacterial virulence.

Keywords:
Neisseria meningitidis / Pili / Proteoforms / PTM / Technology / Top-down MS

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

PTM increases the functional diversity of proteins by covalent addition of functional groups, modification of amino acid side chains, and proteolysis. PTMs are implicated in almost all aspects of normal cell biology and pathogenesis. In viral or bacterial infection, pathogens often use PTMs to manipulate pathways in the host cell in order to promote

Correspondence: Dr. Julia Chamot-Rooke, Structural Mass Spectrometry and Proteomics Unit, Institut Pasteur, CNRS UMR 3528, 26–28 Rue du Docteur Roux, 75724 Paris Cedex 15, France
E-mail: julia.chamot-rooke@pasteur.fr
Fax: +33-0-169334803

Abbreviations: DATDH, 2,4-diacetamido 2,4,6-trideoxy α-D-hexose; **ECD**, electron capture dissociation; **ETD**, electron transfer dissociation; **HCD**, higher energy collisional dissociation; **PG**, phosphoglycerol

Colour Online: See the article online to view Figs. 2 and 4 in colour.

their own survival, replication, and evasion from the host immune system [1, 2]. For a long-time PTM was considered an exclusively eukaryotic process but it is now widely accepted to also occur in bacteria and archaea. Recent evidence supports the hypothesis that acetylation broadly impacts bacterial physiology [3]. Highly phosphorylated bacterial proteins have been described as being potential intermediates of degradative pathways [4, 5] and sulfated proteins have been shown to trigger the host immune system and bacterial cell–cell communication [6]. Bacterial surface structures such as flagella (*Pseudomonas aeruginosa* and *Campylobacter jejuni*) and pili (*Neisseria* spp. and *P. aeruginosa*) are all found to be particularly rich in posttranslationally modified proteins [7]. Indeed studies on these organelles have led to the description of several complete microbial glycosylation models [8]. As many of the proposed bacterial glycoproteins are surface-exposed, these modified proteins have been postulated to play important roles in pathogenicity and antigenicity.

Type IV pili (T4P) of pathogenic *Neisseria* are hair-like structures that protrude from the bacterial surface and are implicated in a wide variety of processes including bacterial motility and DNA uptake [9]. Since T4P are also required for host-cell adhesion, and thus play a crucial role in colonization of the host, they are considered as a major bacterial virulence factor. T4P are protein macropolymers predominantly composed of a single protein subunit, the major pilin. This pilin protein is arranged in a helical fashion to create the long and flexible pilus fibre. In *Neisseria* the major pilin is the protein PilE, which is highly posttranslationally modified. It is always N-terminally processed and methylated and carries a pair of oxidised cysteines close to the C-terminus. It is glycosylated by the unusual glycan 2,4-diacetamido 2,4,6-trideoxy α-D-hexose (DATDH) [10] or 2-acetamido 4-glyceramido 2,4,6-trideoxy-α-D-hexose [11], which can be further elaborated by up to two galactose or glucose subunits and may be *O*-acetylated. In addition, this protein may also harbor a number of phosphoforms such as phosphate (P) [12], phosphoethanolamine [13, 14], phosphocholine [13–15], and phosphoglycerol (PG) [16].

PilE has been reported to be concurrently expressed as a number of different proteoforms each carrying an array of different PTMs. (The term proteoform has recently been proposed to describe the different molecular forms in which the protein product of a single gene can be found, including changes due to genetic variations, alternatively spliced RNA transcripts, and posttranslational modification [17]). This was first suggested from X-ray crystallography data that showed a weak electron density for phosphate around Ser^{94} in *N. gonorrhoeae* strain MS11, in addition to the phosphate on Ser^{68} [12]. This indicated that a small proportion of the PilE population harbored an additional phosphate group. More recently proteoforms of PilE from both *Neisseria meningitidis* and *Neisseria gonorrhoeae* have been evidenced by MS in intact mass profiling experiments. In *N. gonorrhoeae*, pilins modified with both phosphoethanolamine and phosphocholine have been reported and the ratio of the different phosphoforms expressed has been shown to be dependent on the presence of the glycan and of the minor pilin PilV [13]. In *N. meningitidis*, proteoform variation also centers around the phosphoform [18]. Mass profiling of the 8013 strain has shown PilE to be expressed as two proteoforms in a 4:1 major/minor ratio [11]. The minor proteoform carries an extra PG on Ser^{93}. In this strain PG has been found to be a regulated PTM that mediates bacterial pathogenesis. Increased modification of PilE with PG at Ser^{93}, several hours after host cell contact, changes the electrostatic surface of the pilus fibre and disrupts pilus–pilus interactions. This is a prerequisite for crossing the epithelial layer and a key step in pathogenesis [19]. The extent and variation of pilin PTMs in highly pathogenic *N. meningitidis* strains is therefore of particular interest, as is understanding their role in bacterial virulence [20].

The tools available at present to achieve complete PTM mapping of all expressed proteoforms of a protein of interest are extremely limited. MS-based proteomics has proven to be powerful approach for the identification of individual PTMs, leading in some cases to the global identification of hundreds to thousands of PTMs within a sample [21]. However, peptide-based, bottom-up proteomics fails to provide a complete picture of PTM since the connectivity between peptides and their parent proteoforms is lost. This is particularly important when two or more modifications work together on a single protein. Moreover, bottom-up strategies do not provide proteoform level information, i.e. the explicit identities of the proteoforms present in the sample and their relative abundance, which is crucial information to understand proteoform function and in vivo regulation. Top-down MS, which is based on the analysis of intact proteins, preserves the proteoforms and thus facilitates their full characterization including PTMs [22, 23]. Top-down approaches have been successfully applied to the characterization of various protein PTMs [23–25] and recent improvements in intact protein chromatographic separations and high performance FTMS instrumentation have greatly expanded the observable range of proteoforms in complex samples [26, 27]. However, the analytical requirements of top-down MS remain particularly challenging due to the size of the systems under study (proteins and not peptides). Top-down MS is currently limited to high-resolution instrumental platforms (FT-ICR, LTQ-Orbitrap, or high resolution ToF instruments) and separation of intact proteins can be a difficult task, as can efficient protein fragmentation on an LC time scale. Finally, the software options available for top-down MS data analysis are rather limited, although recent developments in the field are helping to overcome this [28, 29].

To date the PTM complement of PilE has only been investigated in two different strains of *N. meningitidis* (8013 and C311) [10, 11, 16, 30] and the diversity of PTMs present in the population of *N. meningitidis* strains found in human patients remains unknown. Because of the link between PTMs and virulence, the analysis of novel strains is of great interest, but it is limited by the lack of approaches that are sufficiently rapid and amenable to high throughput analysis. To

determine the feasibility of typing clinical strains in terms of pilin PTMs, we isolated a strain from a patient hospitalized with meningitis and characterized its pilin PTMs by different approaches. Mass profiling revealed the presence of four major proteoforms of PilE, in a 1:1:1:1 ratio. Initially, a bottom-up strategy, which had proven to be efficient for the analysis of the other strains was used, but it led to an incomplete PTM mapping. We therefore employed top-down MS to select and fragment the intact proteoforms individually. Strengths and weaknesses of both approaches for the analysis of these challenging proteins will be discussed. We will show that when more than two proteoforms of a single protein are present, bottom-up alone is fundamentally unable to map all PTMs and top-down is required to achieve this goal. Top-down mass spectrometry therefore appears a promising tool for the analysis of highly posttranslationally modified proteins involved in bacterial virulence.

2 Materials and methods

2.2 Bacterial strain

The strain of *Neisseria meningitidis* used in this study was isolated from a patient with meningitis at the Limoges hospital in the Haute Vienne County of France in 2006 (strain number 278534). It is a serogroup A capsular serotype and multilocus sequence typing revealed that it is part of the ST-5 complex/subgroup III clonal complex. *N. meningitidis* was grown on solid GCB Agar (Laboratorios Conda, Spain) containing Kellogg's supplements [31]. The major pilin gene (*pilE*) was PCR amplified and sequenced using oligos NG1705 (GTCAAACCCGGTCATTGTCC) and NG1706 (CAGGAGTCATCCAAATGAAAGC) [32].

2.3 PilE Preparation

Pili were prepared as described previously [33]. Briefly, the content of 10–12 Petri dishes was harvested in 5 mL of 150 mM ethanolamine at pH 10.5. Pili were sheared by vortexing for 1 min. Bacteria were centrifuged at 4000 × *g*, 30 min, 4°C and the resulting supernatant further centrifuged at 15 000 × *g*, 30 min, ambient temperature. The supernatant was removed, pili precipitated by the addition of 10% vol. ammonium sulfate saturated in 150 mM ethanolamine pH 10.5 and allowed to stand for 1 h. The precipitate was pelleted by centrifugation at 4000 × *g*, 1 h, 20°C. Pellets were washed twice with PBS and suspended in 100 μL distilled water.

2.4 Bottom-up MS

Ten microliters of crude PilE preparation was suspended in Laemmli buffer and separated by SDS-PAGE. After removal of the Bio-Safe Coomassie stain (Biorad) used for visualization, the band corresponding to PilE was excised and digested in-gel as described elsewhere [34]. Briefly gel pieces were reduced, alkylated, and digested overnight with trypsin (Promega) at 37°C. After desalting by C_{18} Ziptip® samples were eluted into 10 μL spray solution of ACN:H_2O:HCOOH (50:50:0.1) for MS. The resulting proteolytic peptides were examined in positive ion mode by direct infusion nano-ESI, using a TriVersa Nanomate (Advion Biosciences, Harlow, UK) on an Orbitrap Velos mass spectrometer, equipped with an electron transfer dissociation (ETD) module (Thermo Fisher Scientific, Bremen, Germany). A full set of automated positive ion calibrations was performed immediately prior to mass measurement, as were the calibrations for reagent ion transfer. All spectra were acquired in full profile mode. For MS experiments ions were accumulated in the ion trap and then transferred to the Orbitrap for high-resolution mass measurement. For MS/MS experiments, ions were selected with an appropriate mass window and higher energy collision dissociation (HCD) was performed at normalized collision energies of 15–25%, with other activation parameters left as default. For ETD the reagent gas was fluoranthene and the interaction time tuned to maximize sequence coverage. Supplemental activation was also applied. The FT automatic gain control was set at 1×10^6 for MS and 2×10^5 for MS^n experiments. Spectra were acquired in the FTMS over several minutes with between one and three microscans and a resolution of 60 000 for MS and 30 000 at m/z 400 for MS/MS before being processed with Thermo Xcalibur 2.2. Peak picking was performed manually for all spectra and fragmentation maps were generated using a home built package.

2.5 Top-down MS

Top-down experiments were performed on a solariX 12T hybrid Qh-FT-ICR (Bruker Daltonics, Billerica, MA) equipped with a hollow dispenser cathode. Crude protein extracts were desalted by C_4 Ziptip® (Millipore) and eluted into 10 μL electrospray solution MeOH:H_2O:HCOOH (75:25:3). Protein was introduced into the mass spectrometer through pulled borosilicate capillaries or by using a TriVersa NanoMate® (Advion Ithaca, NY). The NanoMate was the injection method of choice over pulled capillaries with wider tips, even for samples prone to aggregation and needle (nozzle) clogging. For mass profiling experiments ions were accumulated in the hexapole for 0.1–1 s before being transferred to the Infinity™ cell for detection. Spectra were accumulated for 50–200 scans. For MS/MS spectra, since the concentration of sample, and therefore single-scan intensity, was highly preparation dependent, ions were accumulated for $\leq$ 4 s in the hexapole in order to reach a threshold precursor ion intensity. Ions were then transferred to the Infinity™ cell where electron capture dissociation (ECD) was performed with a pulse length of 5–10 ms and electron energy of 1.0–1.7 eV. For each experiment, 300–450 scans were accumulated. The number of data points (1 Mega points) was chosen to have near baseline

resolution without detrimentally decreasing scan speed. Calibration was performed monthly with clusters of NaI.

2.6 Top-down data analysis

Data processing was performed using DataAnalysis 4.0 SP5 (Bruker Daltonics, Billerica, MA). For MS experiments spectra were deconvoluted using the maximum entropy option. This gives much cleaner deconvoluted spectra than the other options available. For MS/MS experiments, acquired spectra were internally calibrated and peak picked using the SNAP 2.0 algorithm; quality factor 0.1, S/N 2, relative intensity 1×10^{-5} (%), absolute intensity 0 and a maximum charge state altered to just above the maximum observed charge state. Peak picking results were saved as an XML file and peak assignment was performed by importing this data into a home built package for ion assignment and automated fragmentation map creation. PTM assignment was performed manually with this software on combined peak lists from the 14+ and 15+ charge states.

3 Results and discussion

3.1 Mass profiling

Purification of PilE from the 278534 strain produced a large amount of protein in high purity (see SDS-PAGE in Supporting Information Fig. 1). When measured by FT-ICR MS the crude sample gave a complex MS spectrum exhibiting multiple charge state envelopes. Deconvolution of the raw data gave four major peaks with monoisotopic neutral molecular masses (M_r) of 15 146.7058, 15 374.8325, 15 602.9369, and 15 831.0584 in an approximate 1:1:1:1 ratio (Fig. 1A).

To provide a reference point for further investigation the *pilE* gene from the *N. meningitidis* 278534 strain was sequenced. Surface expressed PilE is known to be posttranslationally processed by the endoprotease *pilD* that cleaves a short N-terminal leader sequence or prepilin before a conserved phenylalanine residue [35]. Making this modification to the initial 147 amino acid sequence furnished a 140 amino acid protein with the theoretical M_r of 14,524.47 (sequence depicted in Fig. 1B). Even when compared to the lowest mass major peak observed in the MS profile this represents a difference of over 620 Da and indicated that PilE from this particular clinical strain could be highly posttranslationally modified.

3.2 Bottom-up analysis

A MS strategy based on the combination of accurate high resolution intact mass measurement of proteoforms and MS/MS experiments on peptides (bottom-up approach) had previously proven useful in identifying PTMs on PilE from *N. meningitidis* strain 8013 [19, 36]. Therefore a similar approach was initially employed here. A sample of PilE was subjected to SDS-PAGE followed by in-gel tryptic digestion. The digest was then analyzed by nano-ESI-FTMS on an Orbitrap mass spectrometer (Fig. 2).

Comparison of the experimental masses measured for this tryptic digest and theoretical ones calculated in silico from the PilE sequence revealed the presence of nonmodified (naked) peptides spanning the ranges [31–59], [76–112], and [122–138]. A [1–30]+14.016 Da peptide was also observed, confirming the N-terminal methylation of PilE and leading to a sequence coverage of over 80% (Table 1).

All peptide ions were isolated and subjected to HCD in order to confirm their identity (data not shown). Despite the high sequence coverage obtained with these ions, numerous abundant, multiply charged peaks present in the MS spectrum could not be attributed. In addition, no peptide spanning the [60–75] region could be assigned. Since this region of PilE is almost always posttranslationally modified, its absence encouraged us to investigate these multiply charged, nonassigned peaks in the hunt for additional PTM-bearing peptides.

When ions observed at m/z 790.8960, 815.4392, 651.2957, and 739.7317 were subjected to HCD, three very abundant fragment ions appeared at m/z 229.118, 211.108, and 169.097 in the resulting MS/MS spectra (Fig. 3A, C, E, G).

These ions correspond to the oxonium ion of the DATDH glycan, its dehydrated partner and a fragment ion characteristic of the glycan core and indicate that all fragmented peptides contain this sugar moiety. DATDH is a previously described PTM for PilE and these reporter ions have previously been used to identify glycosylated peptides [37, 38]. The analysis of the other fragment ions in the spectra (y/b ions) confirmed the sequence for the four peptides ions as [80–92], [99–112], [113–121], and [93–112], respectively, but could not be used to localize the sites of glycosylation. All four precursor ions were therefore subjected to ETD in order to confidently localize the glycosylation sites (Fig. 3B, D, F, H). For the [80–92] peptide, the glycan could be easily localized on Ser^{83}, for [99–112] and [93–112] it was localized exclusively on Ser^{101} and finally for [113–121], Ser^{113} was found to be modified with DATDH. In none of the ETD spectra was any trace of the reporter ions detected at m/z 229.118, 211.108, or 169.097 (see Supporting Information Tables for all MS/MS data).

In addition, when fragmented by HCD, two other doubly charged ions, at m/z 928.934 and 1042.999, produced intense reporter ion signals for DATDH. In both cases the [60–75] tryptic peptide was identified suggesting multiple forms to be present, each modified by DATDH. ETD data identified a DATDH on Ser^{63} for the precursor m/z 928.934, on both Ser^{63} and Ser^{68} for m/z 1042.99 and showed that both forms were further modified by PG on Ser^{70} (Fig. 3I, J).

Upon inclusion of these modified peptides, the sequence coverage from the digest was extended to 98% with only the two last amino acids of the sequence, AK, unaccounted for. These results were corroborated but not improved upon

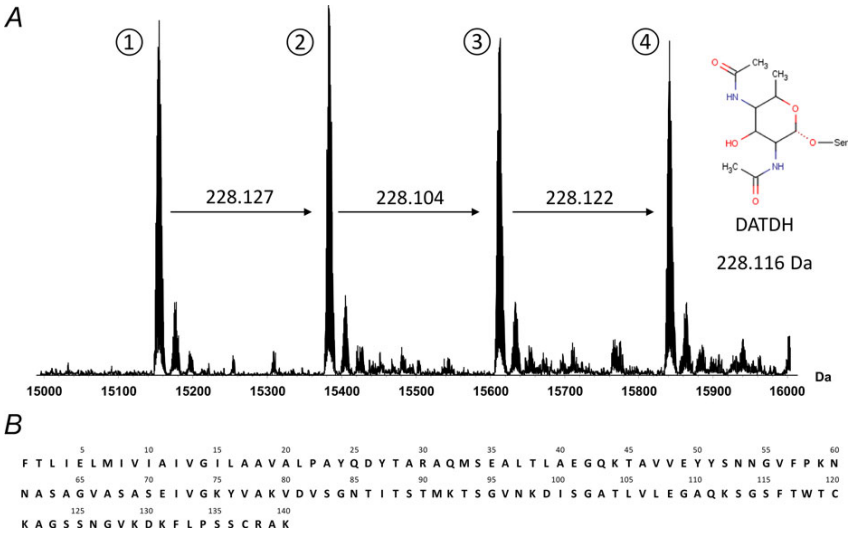


Figure 1. Mass spectrum of PilE extracted from strain 278534 and deconvoluted across the entire mass range. Four major peaks are observed with associated salt adduct peaks. Many lower-intensity features are clearly distinguishable close to the baseline. 190 × 142 mm (300 × 300 DPI).

using a LC-MS approach. Interestingly, the presence of several peptides in both nonmodified and modified forms confirmed that PilE was expressed as multiple proteoforms. Furthermore it appeared that these proteoforms differ in the number of DATDH glycan units. This hypothesis correlated perfectly with the pattern observed by mass profiling, where the mass difference between the four major peaks is approximately 228.1 Da—the mass expected from addition of DATDH. PilE thus appeared to be present in multiple glycoforms each bearing a different number of DATDH subunits.

Armed with mass profiling data and the results from the bottom-up experiments, one can begin to assign peptide combinations, and thus PTMs, to specific proteoforms. Since all peptides are found in nonmodified forms apart from [113–121] and [60–75], the lowest mass forms of these two peptides [113–121] + DATDH and [60–75] + PG + DATDH plus the exclusively nonmodified peptides, may naturally be attributed to the lowest mass proteoform. This assignment seems satisfactory since it results in the correct 15 146.7058 Da protein mass with an experimental theoretical mass error of only 0.03 ppm. The combination of peptides giving the protein with the heaviest mass; [60–75] + PG + 2DATDH, [80–92] + DATDH, [99–112] + DATDH, [113–121] + DATDH could similarly be assigned to the highest mass proteoform. This also seems

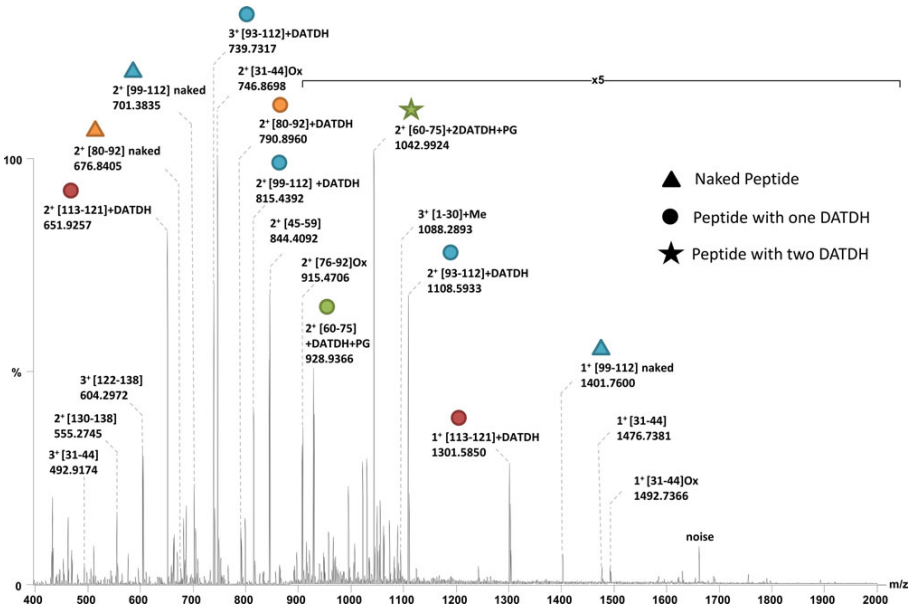


Figure 2. Nano-ESI-FTMS analysis of the PilE digest. For peptides that exist in multiple modification states the naked form is marked with a triangle, a circle denotes modification with one DATDH and a star with two DATDH. Peptides spanning the same modification site are represented by shapes of the same color. A full list of peptide masses is given in Table 1. 190 × 142 mm (300 × 300 DPI).

Table 1. Digest product ions of PilE from 278534 strain (from Fig. 2)

Digestion product	Measured *m/z*	Charge	Measured monoisotopic mass $[M+H]^+$	Theoretical monoisotopic mass $[M+H]^+$	Error (ppm)
[1–30]+Me	1088.2893	3	3262.8533	3262.8524	0.290
[31–44]	492.9174	3	1476.7376	1476.7363	0.912
[31–44]	738.8725	2	1476.7377	1476.7363	0.964
[31–44]	1476.7381	1	1476.7381	1476.7363	1.219
[31–44]Ox	746.8698	2	1492.7323	1492.7312	0.753
[31–44]Ox	1492.7366	1	1492.7366	1492.7312	3.618
[45–59]	844.4092	2	1687.8111	1687.8326	12.724[a)]
[60–75]+DATDH+PG	928.9366	2	1856.8659	1856.8637	1.197
[60–75]+2DATDH+PG	1042.9924	2	2084.9775	2084.9747	1.354
[76–92]Ox	915.4706	2	1829.9339	1829.9313	1.434
[80–92]	676.8405	2	1352.6737	1352.6726	0.831
[80–92]+DATDH	790.8960	2	1580.7847	1580.7836	0.711
[93–112]+DATDH	739.7317	3	2217.1805	2216.1769	[b)]
[93–112]+DATDH	1108.5933	2	2216.1793	2216.1769	1.094
[99–112]	701.3835	2	1401.7597	1401.7584	0.944
[99–112]	1401.7600	1	1401.7600	1401.7584	1.141
[99–112]+DATDH	815.4392	2	1629.8711	1629.8694	1.057
[113–121]+DATDH	651.2957	2	1301.5841	1301.5831	0.786
[113–121]+DATDH	1301.5850	1	1301.5850	1301.5831	1.460
[122–138]	604.2972	3	1810.8770	1809.8912	[a)]
[130–138]	555.2745	2	1109.5417	1109.5408	0.832

a) Errors for these ions are greater than expected since these digest products contain asparagine vicinal to glycine residues and are therefore subject to facile deamidation. For the ion at *m/z* 844.4092 deamidation is incomplete but the monoisotopic peak is very close to the baseline. For the ion at *m/z* 604.2972 complete deamidation increases the observed mass by approximately 1 Da.
b) The monoisotopic peak for this ion is obscured by the doubly charged [31–44] at *m/z* 738.8725.

correct and is the only peptide combination possible to reach a total mass of 15 831.0584 Da. However, a problem arises when considering the two intermediate proteoforms, since various combinations of peptides are possible to achieve the observed protein masses. Here, even with the mass profile as a reference, a bottom-up methodology is intrinsically unable to relate modified peptides to their parent proteoforms without making several important assumptions. One may think that examining the ion intensities of modified peptides and relating them to abundances may help solve this problem, but this approach is unrealistic since PTM has been well documented to drastically affect peptide ionization efficiency [39].

To achieve complete proteoform mapping, each proteoform must be investigated separately. Off-line fractionation methods require prior knowledge of the PTMs present on each proteoforms and are often difficult to implement at the protein level. A top-down MS approach allows each proteoform to be easily isolated and fragmented separately in the mass spectrometer. This top-down strategy was therefore applied to PilE purified from Nm 278534D and tested on both an LTQ-Orbitrap Velos mass spectrometer, employing ETD as the fragmentation method, and a 12T solariX FT-ICR mass spectrometer, using ECD for protein fragmentation. Higher overall sequence coverage was obtained from ECD MS/MS on the FT-ICR instrument and therefore this approach was chosen for all top-down experiments.

3.3 Top-down analysis

PilE from *N. meningitidis* 8013 strain has previously been investigated by top-down ECD MS/MS using an FT-ICR mass spectrometer (unpublished data) and experimental parameters optimized in that study were used as a starting point here (irradiation time, energy of the electrons etc.). The PilE sample examined in this study did however exhibit some differences compared to that purified from the 8013-reference strain (PilE-8013). Clogging of nanoelectrospray needles during spectral acquisition was much more acute; despite centrifugation prior to sample injection and the use of pulled capillaries with a wider tip. A more stable spray was obtained using a Triversa NanoMate and most importantly this injection method allowed the spray signal to be monitored during spectral acquisition and the nozzle to be quickly changed if necessary. The charge state envelope was also different for the 278534 strain. It was shifted to lower charge states, with the maximum observed here 16+ compared to 19+ for the 8013 strain. This is possibly due to the difference in protein size or the different nature and number of PTMs carried by the two proteins. The envelope was also much more complex, the presence of four major proteoforms causing some proteoforms of one charge state to overlap with other proteoforms from another. Proteoforms of the highest exploitable charge states (15+ and 14+) were isolated in the hexapole of the

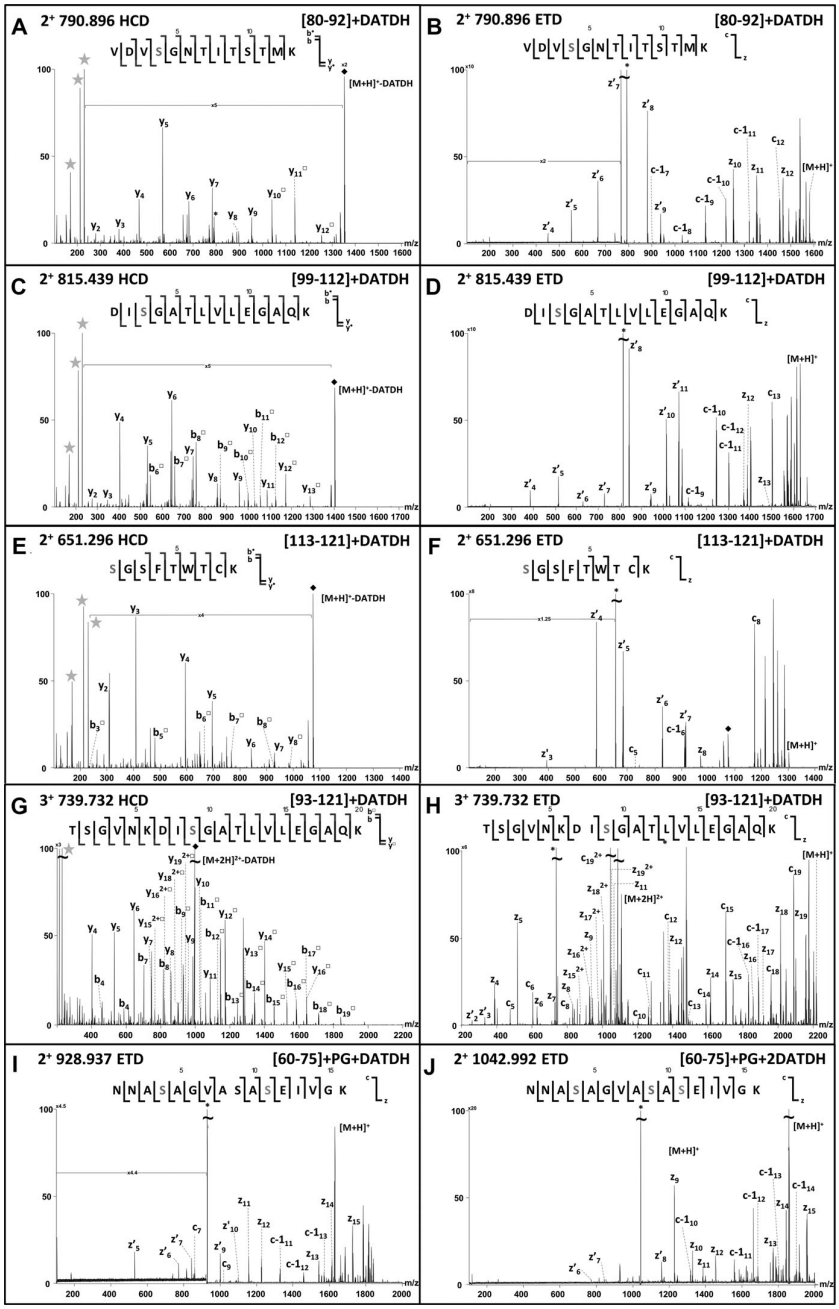


Figure 3. HCD and ETD spectra of modified peptides. Molecular ions are marked with a star *. Charge-reduced molecular ions exhibiting loss of DATDH are marked with a diamond ♦, other ions with loss of DATDH marked with a square □. Green stars ⋆ indicate DATDH reporter ions. In the fragmentation maps residues modified with DATDH are colored red and those modified with PG are colored green. In the ETD spectra ions marked z′ correspond to z + H and in the fragmentation maps both c and c-H are represented by a single bar as are z and z′. In panel H several abundant ions are not labeled and originate from overlap of the parent ion with [31–44]2+ at *m/z* 738.8725. 190 × 142 mm (300 × 300 DPI).

12T FT-ICR mass spectrometer and single charge states were subjected to top-down ECD MS/MS.

In general the 15+ ions afforded greater sequence coverage for the same precursor intensity but the 14+ charge state was more abundant and, for some proteoforms, furnished even greater sequence coverage when precursor ions were allowed to accumulate. Extensive fragmentation was observed in the top-down spectra especially in the case of the highest mass proteoform (15 831.0584 Da) (Fig. 4D). Cleavage maps are also shown for the other three proteoforms (Fig. 4A–C). Interestingly, the intact cysteine bridge between Cys^{120}–Cys^{137} appeared to strongly inhibit fragmentation in this region for the 14+ charge state but not for the 15+ charge state.

For proteoform 1, fairly extensive N- and C-terminal fragmentation enabled straightforward identification of the methylated N-terminus and a DATDH glycan on Ser^{113}. The

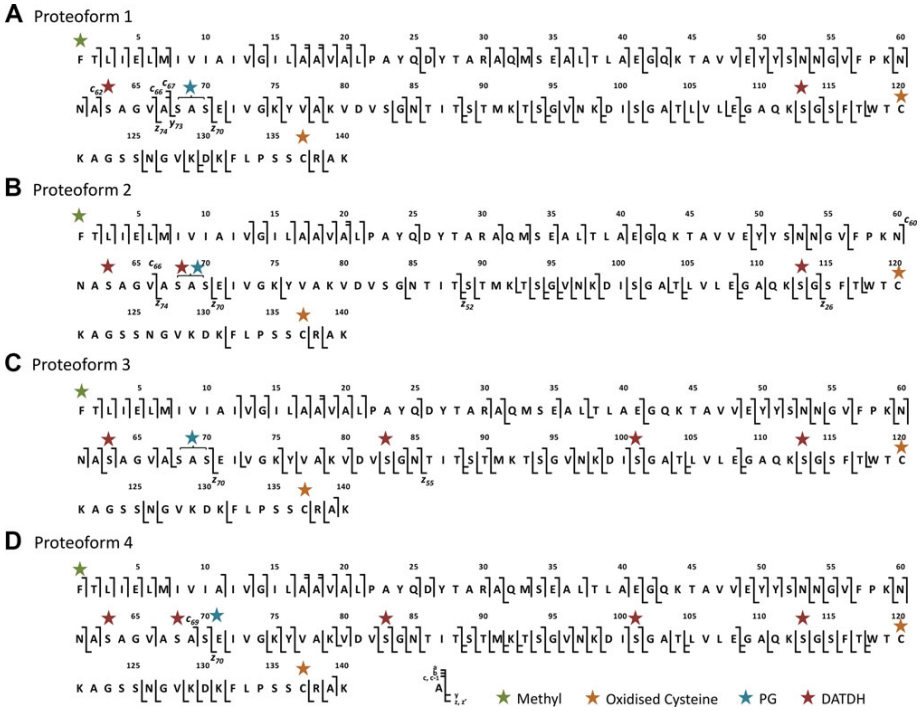


Figure 4. Fragmentation maps of proteoforms 1–4 resulting from ECD fragmentation of 14+ and 15+ charge states. Proteoform 2 is from the 15+ charge state only. Colored stars denote different PTMs. z and z′ (z+H) ions are represented by a single bar as are c and c-H ions. Particular ions of interest are labeled. 190 × 142 mm (300 × 300 DPI).

c_{62} ion at m/z 1115.4284 (6+), 1339.3085 (5+), c_{66} ion at m/z 1205.8037 (6^+), 1446.7672 (5^+), and c_{67} ion at m/z 1217.6449 (6^+) enabled localization of a second DATDH on Ser^{63}. The presence of the same c ions in spectra obtained from both the 14+ and 15+ protein forms increases confidence in this assignment, as does the presence of multiple charge states for c_{66} in each spectrum. In addition the C-terminal fragment ions z_{70}, z_{74}, and y_{73} enabled identification of a phosphoglycerol group on Ser^{68} or Ser^{70} but the absence of ions between these residues precluded definitive localization. This was however provided by the bottom-up data that showed Ser^{70} to be exclusively modified with phosphoglycerol. For proteoform 1 this gave a PTM complement of two DATDH glycans at Ser^{63} and Ser^{113} and a phosphoglycerol group at Ser^{70} in addition to the expected N-terminal methylation and cysteine bond. The theoretical protein mass for this assignment of 15 146.7017 Da correlated exceptionally well with the 15 146.7058 Da experimental value, giving a + 0.03 ppm error.

For proteoform 2, two DATDH subunits were also easily identified on Ser^{63} and Ser^{113}; the former by the c_{60} and c_{66} ions and the latter by a large series of z ions from z_{26} to z_{52}. Despite poor fragmentation in the central region of the protein, the sequence coverage was sufficient to indicate that no PTM was present between residues 71 and 112. The difference in mass between z_{70} and z_{74} (698.2519 Da) indicated the presence of both a PG and a DATDH between Ala^{67} and Ser^{70} but as for proteoform 1 the absence of fragments between these residues prevented definitive location. Taking into account the bottom-up data a PG group was assigned on Ser^{70} and therefore a DATDH assigned to Ser^{68}. Again the measured mass of 15 374.8325 Da corresponded excellently to the theoretical mass for this PTM assignment (15 374.8122 Da) with an error of +1.3 ppm. In comparison to proteoform 1, proteoform 2 exhibited an additional DATDH subunit at Ser^{68}.

For proteoform 3, a better sequence coverage in the regions of interest allowed an easy assignment of the four DATDH groups to Ser^{63}, Ser^{83}, Ser^{101}, and Ser^{113}. Again a PG could be identified either on Ser^{68} or Ser^{70} and bottom-up data were used to confirm the latter position. As the number of modifications increases so does the number of possibilities for site localization. Deciding upon the correct PTM assignment therefore becomes increasingly difficult, especially when potential modification sites are close together in the protein sequence. In the case of proteoform 3, the large series of z type ions from z_{55} to z_{70} allowed confident assignment of DATDH on Ser^{83} rather than Ser^{68}. Our results indicated that, in this proteoform, two previously unmodified serines are now glycosylated and one that was occupied in proteoform 2 is now nonmodified. Similarly to proteoforms 1 and 2, the measured mass of 15 602.9369 Da corresponds excellently to the theoretical mass for this PTM assignment (15 602.9226 Da) with an error of +0.9 ppm.

Finally, for the highest mass proteoform (proteoform 4) the assignment of four DATDH was found to be similar to proteoform 3 and an additional glycan was localized on Ser^{68}, as already observed for proteoform 2. This proteoform led to

the highest sequence coverage in ECD (62%) and the presence of the c_{69}, c_{70}, and complementary z_{70} ions mean that for this proteoform no bottom-up data are required to assign PG to Ser^{70}. The measured mass of 15 831.0584 Da again corresponds very well to the theoretical mass (15 831.0331 Da) with an error of +1.5 ppm.

These results indicated that in most cases spectra were of sufficient quality and fragmentation sufficiently extensive to allow PTMs to be unambiguously localized on the protein backbone. The only exception was for the location of the PG group that needed information from bottom-up experiments to provide the exact modification site for three of the four proteoforms. This result may be explained by a lower efficiency of electron capture at sites close to PTMs, or even capture of the electron by the modifications themselves. This would likely modify the overall reactivity of the ECD process.

Taken in the context of the mass profiling experiment, the correlation between experimental and theoretical masses for all PTM assignments is excellent, with errors at the protein level consistently below 2 ppm. The complement of posttranslational modifications has been explicitly defined for each proteoform and the bottom-up and top-down data are in perfect agreement. All proteoforms are modified with a PG at Ser^{70} and DATDH at Ser^{63} and Ser^{113}. The extra sites of glycosylation are Ser^{68}, Ser^{83}, and Ser^{101} but our results also revealed that the glycosylation process does not appear to be completely successive. This was not expected since the glycerotransferase PglO is known to transfer the glycan *en-bloc* to the PilE substrate and we might therefore expect sequential glycan addition based on decreasing affinity for sites on the protein backbone. Our results suggest that modification at one site may alter the affinity for the others, however the specificity of this enzyme is currently unknown. Most importantly, this level of glycosylation is a hitherto unreported phenomenon for pilin of *Neisseria* spp.

3.4 Strengths and weaknesses of both approaches

First of all, it is important to point out that, when aiming to characterize all PTM present on a protein or its different proteoforms, the mass profiling experiment is a key piece of information. In the case of the strain studied here, mass profiling of PilE indicated the presence of four major proteoforms, each expressed in similar abundance and each modified with phosphoglycerol plus two, three, four, and five DATDH subunits, respectively. This acted as primary reference for all future experiments.

The bottom-up approach was powerful for identifying posttranslationally modified peptides and when coupled with an appropriate fragmentation technique such as ETD, for localizing the sites of posttranslational modification themselves. However, in order to achieve complete characterization of a mixture of proteoforms and thus explicitly define the PTM content of each one, it is necessary to link these pieces of information together and to relate modified peptides to their parent proteoforms. Given the homogeneity of proteoform abundance, the strain characterized here was an ideal case to explore whether a bottom-up proteomics approach is able to achieve this goal, or not. Our results clearly show that although it is possible to completely map all PTMs on the lightest and heaviest forms of PilE, information obtained from the bottom-up approach is not sufficient for mapping of the middle forms.

On the other hand, by selecting individual proteoforms and subjecting them to top-down ECD MS/MS the connectivity that was lost by the bottom-up approach was retained thus allowing the complete assignment of all glycosylation sites. This is expected to be the case in other mixtures of more than two proteoforms where each proteoform is modified by different numbers of a particular PTM. One must also point out that an important weakness of the top-down approach remains the analysis of data; from peak picking, to PTM assignment and scoring. Although currently available bioinformatics tools are sufficient for proteins with one or two known modifications, the analysis of highly modified proteins remains a challenge and manual interpretation is often needed. Improvement in this field is a requirement for more confident PTM assignment and high throughput analysis.

4 Concluding remarks

In this study, the protein PilE, purified from a previously uncharacterized strain of *Neisseria meningitidis* isolated from a patient hospitalized with meningitis, has been analyzed by different mass spectrometric approaches. Mass profiling showed that PilE exists as four major proteoforms, differing only by the number of DATDH glycans. A bottom-up strategy was initially chosen to map all PTMs of the four proteoforms. ETD on glycopeptides proved very useful for reliable assignment of glycosylation sites, however bottom-up only proved capable in characterizing the PTM content of the lightest and heaviest proteoforms but not those of intermediate mass. Individual proteoforms were therefore selected and subjected to top-down ECD MS/MS on a 12T FT-ICR mass spectrometer. The sequence coverage obtained in each case allowed unambiguous identification of an increasing number of glycan subunits. Combining the top-down and bottom-up data allowed complete PTM characterization of all proteoforms. The lightest form was found to be N-terminally processed and methylated, to carry a disulfide bridge close to the C-terminus, one phosphoglycerol on Ser^{70} and DATDH at Ser^{63} and Ser^{113}. The additional glycosylation sites for the other proteoforms are Ser^{68}, Ser^{83}, and Ser^{101}. Such an extent of glycosylation and indeed of PTM has never before been described for PilE and may be linked to increased pathogenicity or antigenicity. In general our results show that, in proteoform mixtures with more than two components, where multiple modifications of the same mass are present, a top-down mass spectrometry approach is necessary for complete proteoform mapping. In

this study a 12T FT-ICR mass spectrometer was chosen for the top-down approach but it is probable that in the near future these experiments will be possible on a routine basis on other instrumental platforms such as later generation Orbitrap systems.

This work was supported by INSERM (ATIP-Avenir starting grant) and by the European Research Council (starting grant) (G.D.), by the CNRS and Institut Pasteur (J.C.R.) and NIH grants P41 RR10888/GM104603 and S10 RR025082 (C.E.C.). J.C.R. and J.G. gratefully acknowledge the Monge Ph.D scholarship from Ecole Polytechnique that has funded the research placement of J.G. in Prof Costello's group.

The authors have declared no conflict of interest.

5 References

[1] Broberg, C. A., Orth, K., Tipping the balance by manipulating post-translational modifications. *Curr. Opin. Microbiol.* 2010, *13*, 34–40.

[2] Ribet, D., Cossart, P., Post-translational modifications in host cells during bacterial infection. *Febs. Lett.* 2010, *584*, 2748–2758.

[3] Hu, L. I., Lima, B. P., Wolfe, A. J., Bacterial protein acetylation: the dawning of a new age. *Mol. Microbiol.* 2010, *77*, 15–21.

[4] Rosen, R., Becher, D., Buttner, K., Biran, D. et al., Highly phosphorylated bacterial proteins. *Proteomics* 2004, *4*, 3068–3077.

[5] Mijakovic, I., Protein phosphorylation in bacteria. *Febs. J.* 2010, *277*, 20–21.

[6] Han, S.-W., Lee, S.-W., Bahar, O., Schwessinger, B. et al., Tyrosine sulfation in a Gram-negative bacterium. *Nat. Commun.* 2012, *3*, 1153.

[7] Iwashkiw, J. A., Vozza, N. F., Kinsella, R. L., Feldman, M. F., Pour some sugar on it: the expanding world of bacterial protein *O*-linked glycosylation. *Mol. Microbiol.* 2013, *89*, 14–28.

[8] Nothaft, H., Szymanski, C. M., Protein glycosylation in bacteria: sweeter than ever. *Nat. Rev. Microbiol.* 2010, *8*, 765–778.

[9] Giltner, C. L., Nguyen, Y., Burrows, L. L., Type IV pilin proteins: versatile molecular modules. *Microbiol. Mol. Biol. Rev.* 2012, *76*, 740–772.

[10] Stimson, E., Virji, M., Makepeace, K., Dell, A. et al., Meningococcal pilin—a glycoprotein substituted with digalactosyl 2,4-diacetamido-2,4,6-trideoxyhexose. *Mol. Microbiol.* 1995, *17*, 1201–1214.

[11] Chamot-Rooke, J., Rousseau, B., Lanternier, F., Mikaty, G. et al., Alternative *Neisseria* spp. type IV pilin glycosylation with a glyceramido acetamido trideoxyhexose residue. *Proc. Natl. Acad. Sci. USA* 2007, *104*, 14783–14788.

[12] Forest, K. T., Dunham, S. A., Koomey, M., Tainer, J. A., Crystallographic structure reveals phosphorylated pilin from *Neisseria*: phosphoserine sites modify type IV pilus surface chemistry and fibre morphology. *Mol. Microbiol.* 1999, *31*, 743–752.

[13] Hegge, F. T., Hitchen, P. G., Aas, F. E., Kristiansen, H. et al., Unique modifications with phosphocholine and phosphoethanolamine define alternate antigenic forms of *Neisseria gonorrhoeae* type IV pili. *Proc. Natl. Acad. Sci. USA* 2004, *101*, 10798–10803.

[14] Aas, F. E., Egge-Jacobsen, W., Winther-Larsen, H. C., Lovold, C. et al., Neisseria gonorrhoeae type IV pili undergo multisite, hierarchical modifications with phosphoethanolamine and phosphocholine requiring an enzyme structurally related to lipopolysaccharide phosphoethanolamine transferases. *J. Biol. Chem.* 2006, *281*, 27712–27723.

[15] Weiser, J. N., Goldberg, J. B., Pan, N., Wilson, L., Virji, M., The phosphorylcholine epitope undergoes phase variation on a 43-kilodalton protein in *Pseudomonas aeruginosa* and on pili of *Neisseria meningitidis* and *Neisseria gonorrhoeae*. *Infect Immun.* 1998, *66*, 4263–4267.

[16] Stimson, E., Virji, M., Barker, S., Panico, M. et al., Discovery of a novel protein modification: alpha-glycerophosphate is a substituent of meningococcal pilin. *Biochem. J.* 1996, *316*, 29–33.

[17] Smith, L. M., Kelleher, N. L., Consortium Top Down Proteomics, Proteoform: a single term describing protein complexity. *Nat. Methods* 2013, *10*, 186–187.

[18] Jen, F. E. C., Warren, M. J., Schulz, B. L., Power, P. M. et al., Dual pili post-translational modifications synergize to mediate meningococcal adherence to platelet activating factor receptor on human airway cells. *PLoS Pathog.* 2013, *9*, e1003377–e1003377.

[19] Chamot-Rooke, J., Mikaty, G., Malosse, C., Soyer, M. et al., Posttranslational modification of pili upon cell contact triggers *N. meningitidis* dissemination. *Science* 2011, *331*, 778–782.

[20] Quagliarello, V., Dissemination of *Neisseria meningitidis*. *N. Engl. J. Med.* 2011, *364*, 1573–1575.

[21] Mann, M., Jensen, O. N., Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* 2003, *21*, 255–261.

[22] Lanucara, F., Eyers, C. E., Top-down mass spectrometry for the analysis of combinatorial post-translational modifications. *Mass Spectrom. Rev.* 2013, *32*, 27–42.

[23] Ansong, C., Wu, S., Meng, D., Liu, X. et al., Top-down proteomics reveals a unique protein S-thiolation switch in *Salmonella typhimurium* in response to infection-like conditions. *Proc. Natl. Acad. Sci. USA* 2013, *110*, 10153–10158.

[24] Zhang, H., Ge, Y., Comprehensive analysis of protein modifications by top-down mass spectrometry. *Circ. Cardiovasc. Genet.* 2011, *4*, 711.

[25] Siuti, N., Kelleher, N. L., Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* 2007, *4*, 817–821.

[26] Tran, J. C., Zamdborg, L., Ahlf, D. R., Lee, J. E. et al., Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 2011, *480*, 254–U141.

[27] Catherman, A. D., Durbin, K. R., Ahlf, D. R., Early, B. P. et al., Large-scale top down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. *Mol. Cell. Proteomics* 2013, *12*, 3465–3473.

[28] Liu, X. W., Sirotkin, Y., Shen, Y. F., Anderson, G. et al., Protein Identification Using Top-Down. *Mol. Cell. Proteomics* 2012, *11*, M111.008524.

[29] Zamdborg, L., LeDuc, R. D., Glowacz, K. J., Kim, Y.-B. et al., ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* 2007, *35*, W701–W706.

[30] Marceau, M., Forest, K., Béretti, J.-L., Tainer, J., Nassif, X., Consequences of the loss of O-linked glycosylation of meningococcal type IV pilin on piliation and pilus-mediated adhesion. *Mol. Microbiol.* 1998, *27*, 705–715.

[31] Kellogg, D. S., Jr., Cohen, I. R., Norins, L. C., Schroeter, A. L., Reising, G., Neisseria gonorrhoeae. II. Colonial variation and pathogenicity during 35 months in vitro. *J. Bacteriol* 1968, *96*, 596–605.

[32] Kahler, C. M., Martin, L. E., Tzeng, Y. L., Miller, Y. K. et al., Polymorphisms in pilin glycosylation locus of *Neisseria meningitidis* expressing class II pili. *Infect Immun.* 2001, *69*, 3597–3604.

[33] Carbonnelle, E., Helaine, S., Prouvensier, L., Nassif, X., Pelicic, V., Type IV pilus biogenesis in *Neisseria meningitidis*: PilW is involved in a step occurring after pilus assembly, essential for fibre stability and function. *Mol. Microbiol.* 2005, *55*, 54–64.

[34] Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., Mann, M., In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protocols* 2006, *1*, 2856–2860.

[35] Strom, M. S., Nunn, D. N., Lory, S., A single bifunctional enzyme, PilD, catalyzes cleavage and N-methylation of proteins belonging to the type-IV pilin family. *Proc. Natl. Acad. Sci. USA* 1993, *90*, 2404–2408.

[36] Gault, J. M. C., Duménil, G., Chamot-Rooke, J., A combined mass spectrometry strategy for complete posttranslational modification mapping of *N. meningitidis* major pilin. *J. Mass Spectrom.* 2013, *48*, 1096–9888.

[37] Vik, A., Aas, F. E., Anonsen, J. H., Bilsborough, S. et al., Broad spectrum O-linked protein glycosylation in the human pathogen *Neisseria gonorrhoeae*. *Proc. Natl. Acad. Sci. USA* 2009, *106*, 4447–4452.

[38] Anonsen, J. H., Vik, A., Egge-Jacobsen, W., Koomey, M., An extended spectrum of target proteins and modification sites in the general O-linked protein glycosylation system in *Neisseria gonorrhoeae*. *J. Proteome Res.* 2012, *11*, 5781–5793.

[39] Gao, Y., Wang, Y., A method to determine the ionization efficiency change of peptides caused by phosphorylation. *J. Am. Soc. Mass Spectrom.* 2007, *18*, 1973–1976.

Correct Figures

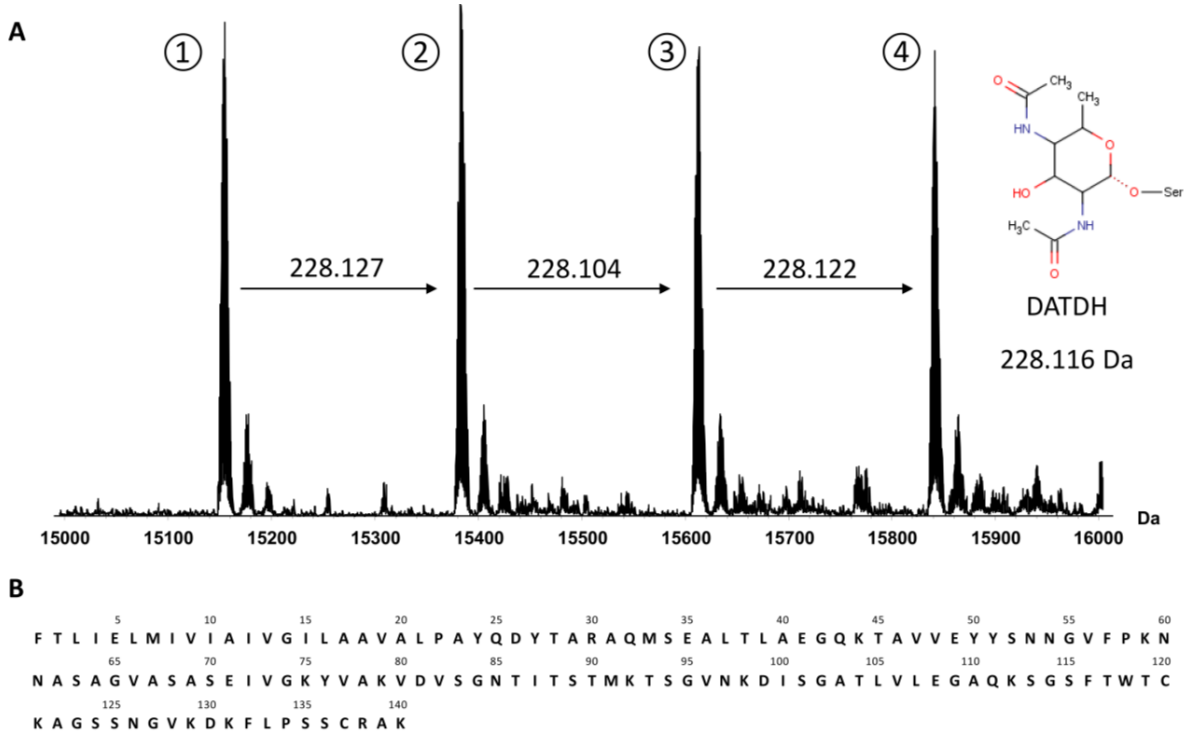


Figure 1 – Mass spectrum of PilE extracted from strain 278534 and deconvoluted across the entire mass range. Four major peaks are observed with associated salt adduct peaks. Many smaller features are clearly distinguishable close to the baseline.

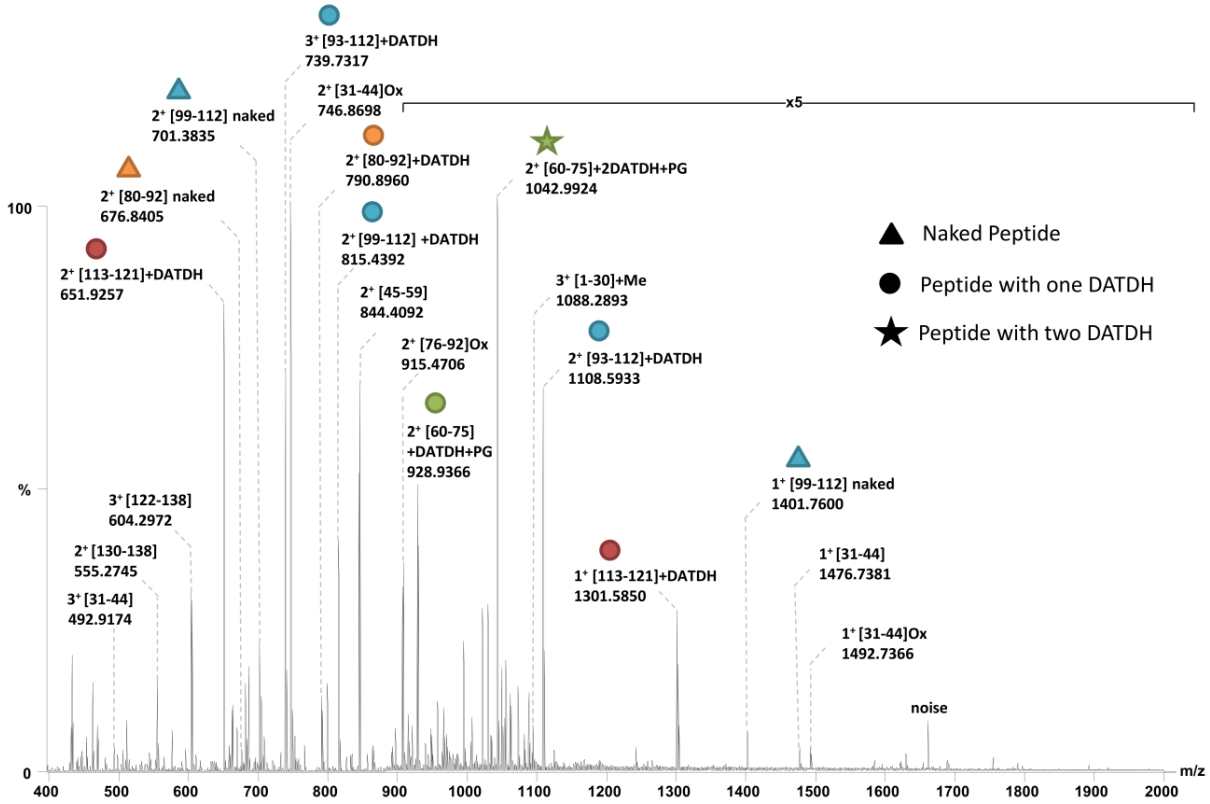


Figure 2 – nanoESI-FTMS analysis of the PilE digest. For peptides that exist in multiple modification states the naked form is marked with a triangle, a circle denotes modification with one DATDH and a star with two DATDH. Peptides spanning the same modification site are represented by shapes of the same colour. A full list of peptide masses is given in table 1.

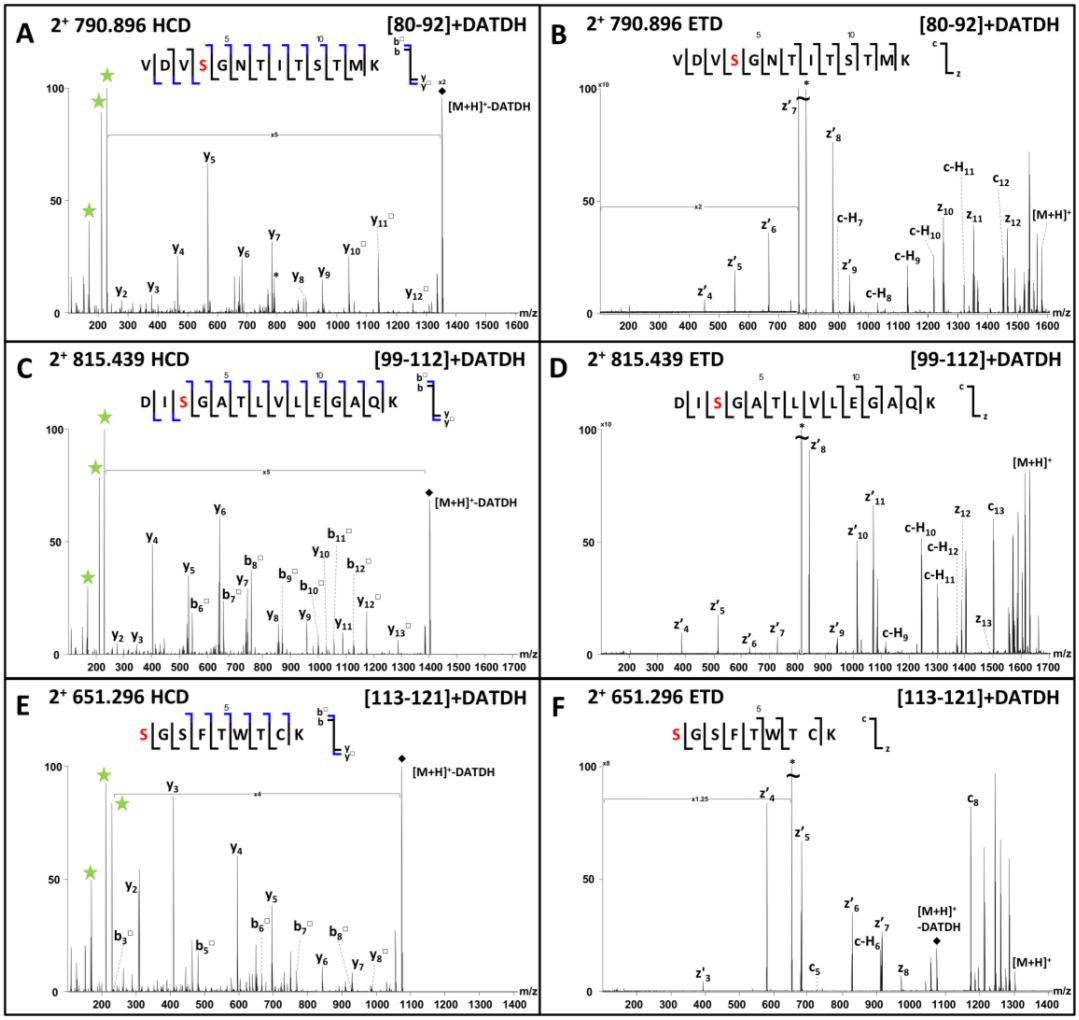

A
2+ 790.896 HCD
[80-92]+DATDH
V D V S G N T I T S T M K
[M+H]+-DATDH
B
2+ 790.896 ETD
[80-92]+DATDH
V D V S G N T I T S T M K
[M+H]+
C
2+ 815.439 HCD
[99-112]+DATDH
D I S G A T L V L E G A Q K
[M+H]+-DATDH
D
2+ 815.439 ETD
[99-112]+DATDH
D I S G A T L V L E G A Q K
[M+H]+
E
2+ 651.296 HCD
[113-121]+DATDH
S G S F T W T C K
[M+H]+-DATDH
F
2+ 651.296 ETD
[113-121]+DATDH
S G S F T W T C K
[M+H]+
-DATDH
[M+H]+
G
3+ 739.732 HCD
[93-112]+DATDH
T S G V N K D I S G A T L V L E G A Q K
[M+2H]2+-DATDH
H
3+ 739.732 ETD
[93-112]+DATDH
T S G V N K D I S G A T L V L E G A Q K
[M+2H]2+
[M+H]+
I
2+ 928.937 ETD
[60-75]+PG+DATDH
N N A S A G V A S A S E I V G K
[M+H]+-DATDH
J
2+ 1042.992 ETD
[60-75]+PG+2DATDH
N N A S A G V A S A S E I V G K
[M+H]+
-DATDH

Figure 3 – HCD and ETD spectra of modified peptides. Molecular ions are marked with a star *. Charge-reduced molecular ions exhibiting loss of DATDH are marked with a diamond ♦, other ions with loss of DATDH marked with a square □. Green stars ★ indicate DATDH reporter ions. In the fragmentation maps residues modified with DATDH are coloured red and those modified with PG are coloured green. In the ETD spectra ions marked *z'* correspond to *z+H* and in the fragmentation maps both *c* and *c-H* are represented by a single bar as are *z* and *z'*. In panel H several abundant ions are not labelled and originate from overlap of the parent ion with $[31\text{-}44]^{2+}$ at *m/z* 738.8725.

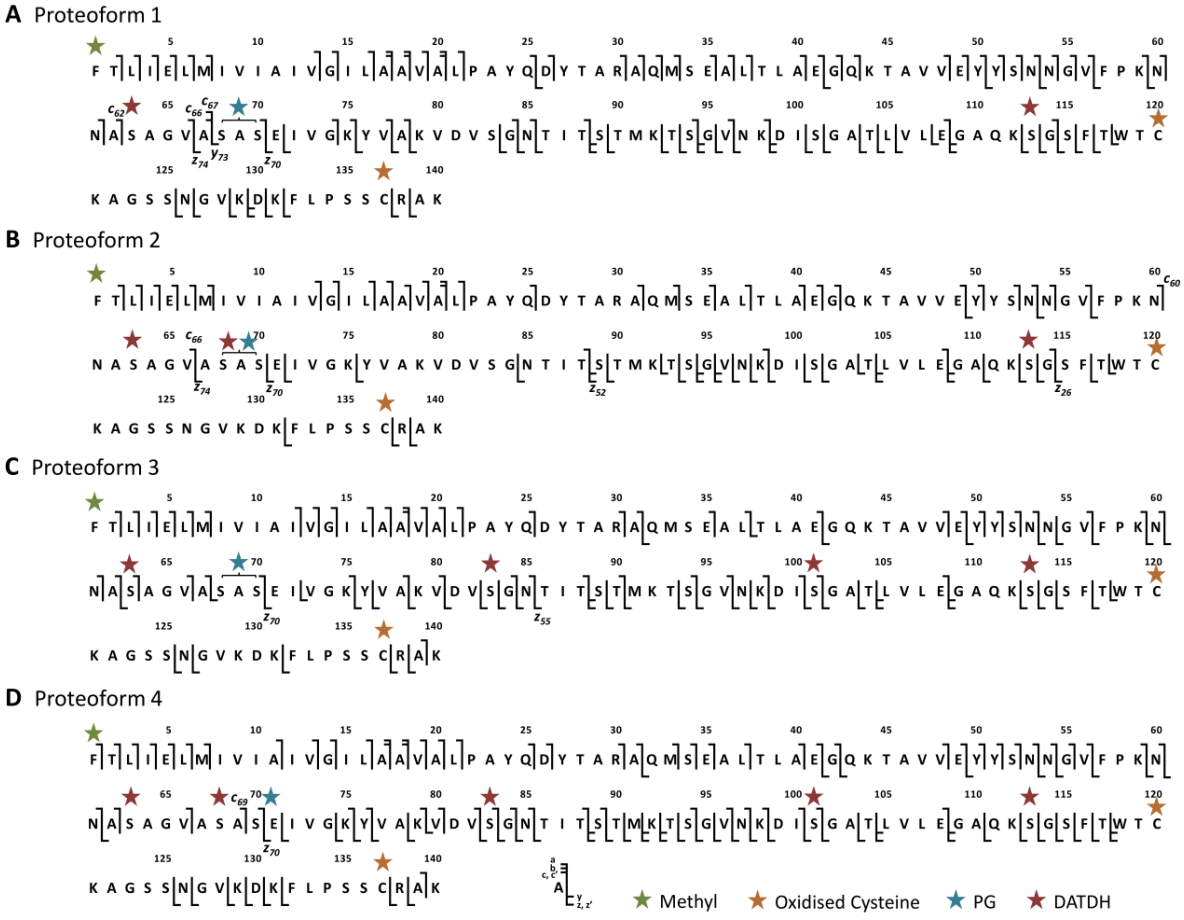
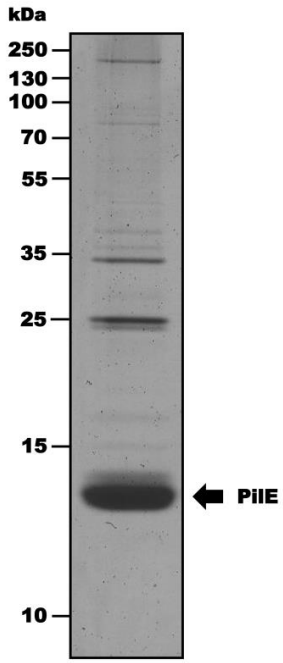


Figure 4 – Fragmentation maps of proteoforms 1 to 4 resulting from ECD fragmentation of 14^+ and 15^+ charge states. Proteoform 2 is from the 15^+ charge state only. Coloured stars denote different PTM's. *z* and *z' (z+H)* ions are represented by a single bar as are *c* and *c-H* ions. Particular ions of interest are labelled.



Supplementary Figure 1 - Coomassie blue stained SDS PAGE of crude PilE preparation showing an intense band for PilE in the expected mass range.

6. Conclusions from Deep Characterisation of PilE-278534D

The analysis of PilE-278534D strain provides several important results concerning the PTM status of PilE and the transferability and utility of the top-down MS methodology.

6.1. Mass Spectrometry

Firstly it is clear that the top-down methodology developed on PilE-8013 can be successfully applied to other strains expressing PilE sequences of the invariable type. Experimental parameters transferred relatively well despite the difference in the protein size and the number of PTMs. Good overall sequence coverage is obtained, although it is not as extensive as that obtained for PilE-8013 and is improved when the results from several charge states are combined.

When multiple putative modification sites are located very close together on the protein backbone, as is the case here for Ser^{68} and Ser^{70}, it is quite difficult to achieve single amino acid resolution and definitively localise the PTM site. Indeed this was only achieved for the highest mass proteoform of PilE-278534D where the MS/MS had been more extensively optimised in an effort to characterise the greatest number of PTMs.

As discussed in the manuscript, the reasons for this are difficult to pinpoint and are perhaps linked to the presence of the PTM themselves; indeed it is conceivable that in proteoform 4 new fragmentation channels were opened by the presence of PG on Ser^{68}. Alternatively the fact that ETD, performed at the peptide level, succeeded in providing amino acid resolution around this site, may suggest that a more compact gas phase protein structure played a role in inhibiting fragmentation through extensive H bonding. In addition, the local proton positions may have been different between the two experiments.

In highly modified proteins with regions of high PTM density such as this, a combination of bottom-up and top-down methods can prove useful particularly if one plays to the strength of each and combines the results to provide the most complete picture of PTM. It is important to stress that in cases where greater than two proteoforms, modified by different numbers of the same PTM are simultaneously expressed, top-down MS becomes essential to map individual PTMs to their parent proteoforms and achieve complete PTM characterisation. PilE-278534D provides a simple and concrete example of such a case.

6.2. Biological Relevance

PilE from 278534D is found to be heavily posttranslationally modified and is expressed in multiple forms each harbouring an unprecedented number of DATDH glycan subunits. To confirm the glycosylation state of PilE-278534D mutants were created by the group of Dr Guillaume Duménil where either the *pglD* or *pglC* genes were deleted (Δ*pglD* or Δ*pglC*). Both of these mutants were expected to be glycan deficient as these core *pgl* genes are necessary for glycan synthesis. PilE from these mutants was purified and profiled by nano-ESI Orbitrap MS (Figure 75).

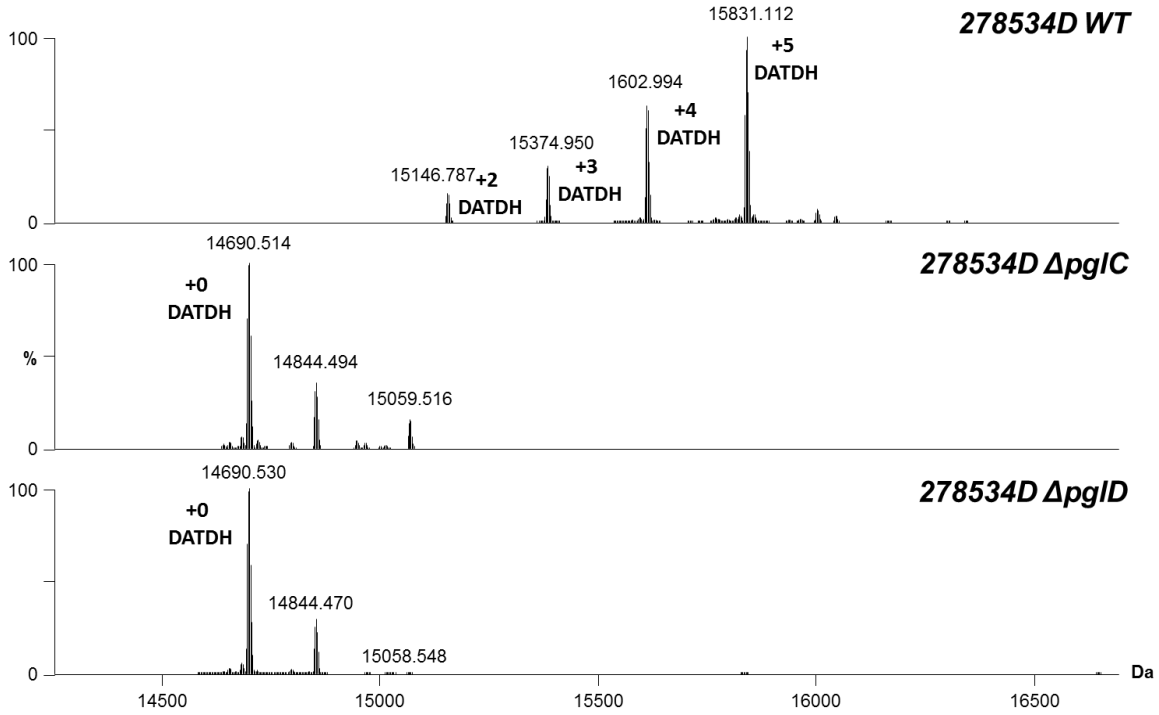


Figure 75 - nano-ESI Orbitrap MS of 278534D WT, Δ*pglC* and Δ*pglD* mutants. Spectra show the full deconvoluted isotope pattern and masses are neutral M_{mono}

The mass profiling spectra of PilE purified from the Δ*pglC* and Δ*pglD* mutants were quite different from that of the WT. They are shifted to lower mass and exhibit two or three major proteoforms as opposed to the four proteoforms present in the WT. The mass of the major proteoform of PilE-278534Δ*pglC* is 14690.514 Da and of PilE-278534Δ*pglD* is 14690.530 Da. This represents a difference of 154.037 and 154.053 Da compared to the mass predicted from the *pilE* gene (14536.477 Da, including oxidised cysteines and a methylated N-terminus). Importantly the measured protein masses are 456.273 and 456.257 Da lower than the lowest mass proteoform in the WT. This corresponds well with the 456.221 expected from two DATDH subunits. Together these data confirm that the major proteoform in both *pgl* mutants is modified with a single PG subunit and is DATDH deficient and support the results presented in the accepted article.

Interestingly in both Δ*pglC* and Δ*pglD* mutants, two additional proteoforms were present at masses corresponding to the addition of one ≈+154 Da and two ≈+308-309 Da PG groups respectively. These high abundance proteoforms were not observed in the WT indicating that loss of DATDH results in an increased global level of PG modification. Whilst we do not know if glycosylation and phosphoform modification occur sequentially, or if both the glycotransferase PglL and phosphotransferases PptA and PptB compete for the same sites, this result may provide evidence for the potential interplay between phosphoform and glycan modifications.

PilE-278534D is therefore posttranslationally modified to a far greater extent than any other pilin previously reported. Expressed PilE seems to always be modified by at least two DATDH subunits and one PG. It carries an unprecedented number of DATDH glycan subunits in all proteoforms. To our knowledge it is also the first invariable type pilin to be characterised to date. This raised the possibility that the other clinical isolates, that also exhibited large experimental theoretical mass deviations and multiple peaks in their spectra, could also be heavily posttranslationally modified. Complete characterisation was therefore performed on another strain 427707C.

7. Deep Characterisation of PilE-427707C

PilE from the 427707C strain was purified and mass profiled by solariX FT-ICR MS. Deconvolution of the MS data gave three major peaks at 16072.003, 16506.1823 and 16941.373 Da when monoisotopic protein masses were calculated using the manufacturer's software (Figure 76). These will henceforth be referred to as the low, intermediate and high mass forms and the mass profile is in good agreement with that obtained by Q-ToF MS.

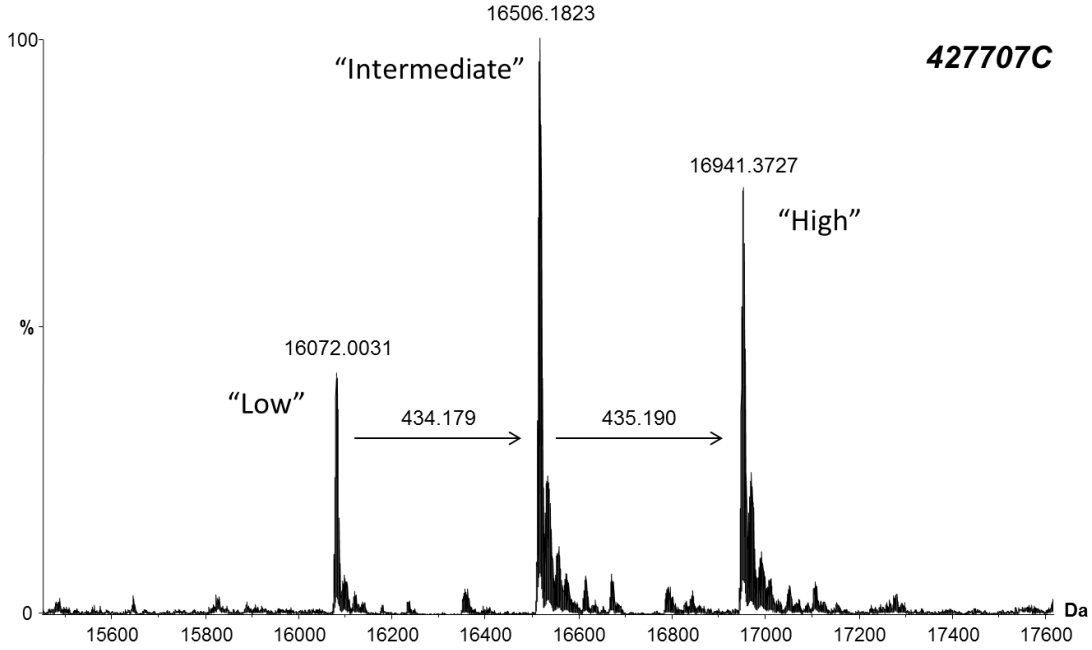


Figure 76 - FT-ICR mass profiling of PilE-427707C showing three major distributions. M_{mono} have been calculated by the SNAP 2.0 algorithm

7.1. PTM of Intermediate Mass Form - Part I

Top-down ECD MS/MS was initially performed on "intermediate" mass form since it was the most abundant. The spectrum was internally calibrated, and peak lists obtained from fragmentation of the 16^+, 17^+ and 18^+ charge states were combined and manual PTM assignment attempted. This process was started at the C-terminal end of the protein. Once the two cysteine residues had been oxidised a large series of *z* type ions could be assigned up to z_{24} with only a single ion $z+H_{28}$ assigned after that (Figure 77A). The sequence coverage stopped abruptly at Ser^{117} and since serine is a commonly modified amino acid in PilE, various PTMs were trialled on this site.

Surprisingly no previously reported PTM for PilE, including the various phosphoforms and glycans, resulted in greatly improved sequence coverage. The sequence coverage when Ser^{117} is modified with GATDH-Hex is shown in Figure 77B. It is known that hydrogen transfer reactions

may take place in ECD resulting in ions that are 1 Da larger or smaller than expected. To examine whether such processes were responsible for the abrupt end of the series of assigned fragment ions, the number of ion types was expanded. (Practically this can be achieved by increasing the peak picking error tolerance and looking for ion assignments at ±1 Da). Different known PTMs were placed on Ser^{117} and the number of ions searched for widened to include those at ±1 Da.

Interestingly when Ser^{117} was modified with GATDH-Hex and once ions at ±1 Da were allowed in the search, a large series of *z* type ions z_{26}-z_{39}, z_{41}, z_{43}, z_{44} was suddenly assigned (Figure 77C, new ions highlighted in red). The mass error for the *z* ion sequence varied around the baseline until z_{25} but then jumped close to -1 Da from z_{26} onwards. This distinctive "mass jump" in the assigned *z* ion sequence can be most easily visualised in Figure 78 and is indicative of an unassigned or mis-assigned PTM on the protein backbone.

At first glance this may seem rather innocuous and perhaps attributed to mis-assignment of the monoisotopic peaks for these ions by the automatic peak picking software or loss of H during fragmentation. However, the mass difference is neither that of a hydrogen atom (1.0078) nor a $^{13}C \rightarrow ^{12}C$ shift; rather it is consistently smaller (Table 7). When all newly assigned *z* ions are considered, an average mass shift of -0.966 Da can be calculated.

Ser^{117} was therefore modified by a hypothetical modification of 435.203 Da (the mass of GATDH-Hex minus 0.966 Da) and ion assignment performed again at 5 ppm tolerance (Figure 77D). This strategy of modifying Ser^{117} with a hypothetical PTM may seem rather artificial but without some sort of software tool, iterative mass modification based on trends in observed ion series and sequence coverage is the only way to approach manual *de novo* PTM assignment. This time the *z* type ion series was continued to z_{44} (highlighted in red).

Out of interest, ions at ±1 Da were again searched and ion assignment performed (Figure 77E). Now two series of new ions could be assigned and are highlighted in red. The first were *z+H* type $z+H_{27\text{-}39}$, $z+H_{41}$, $z+H_{43}$, $z+H_{44}$ and the second consisted of many more matches for the *z* type ions $z_{27\text{-}29}$, $z_{31\text{-}43}$. The combined average mass difference for these newly assigned ions is -1.006 Da.

In an effort to explain this rather odd behaviour, Ser^{117} was again modified by a further -1.006 Da, bringing the total hypothetical modification mass to 434.197 Da and ion reassignment was performed once more at 5 ppm tolerance (Figure 77F). This time a large number of both *z* and *z+H* type ions were assigned (highlighted in red). Searching for additional ions at ±1 Da did not result in a significant number of new ion matches, therefore Ser^{117} was considered to be modified by a PTM approximately 434.197 Da in mass.

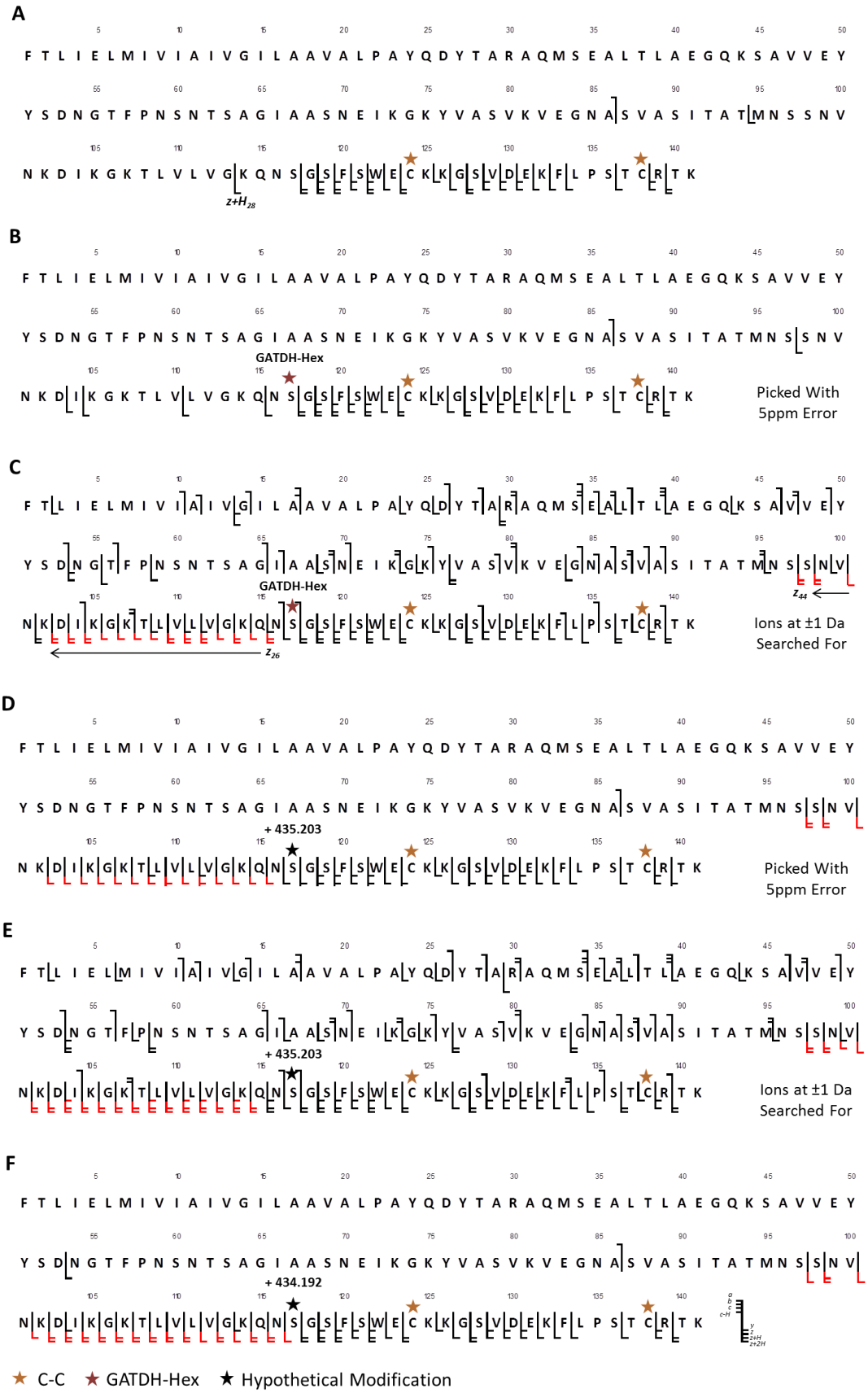


Figure 77 - Fragmentation maps of PilE-427707C produced during the iterative assignment of PTM to Ser^{117}

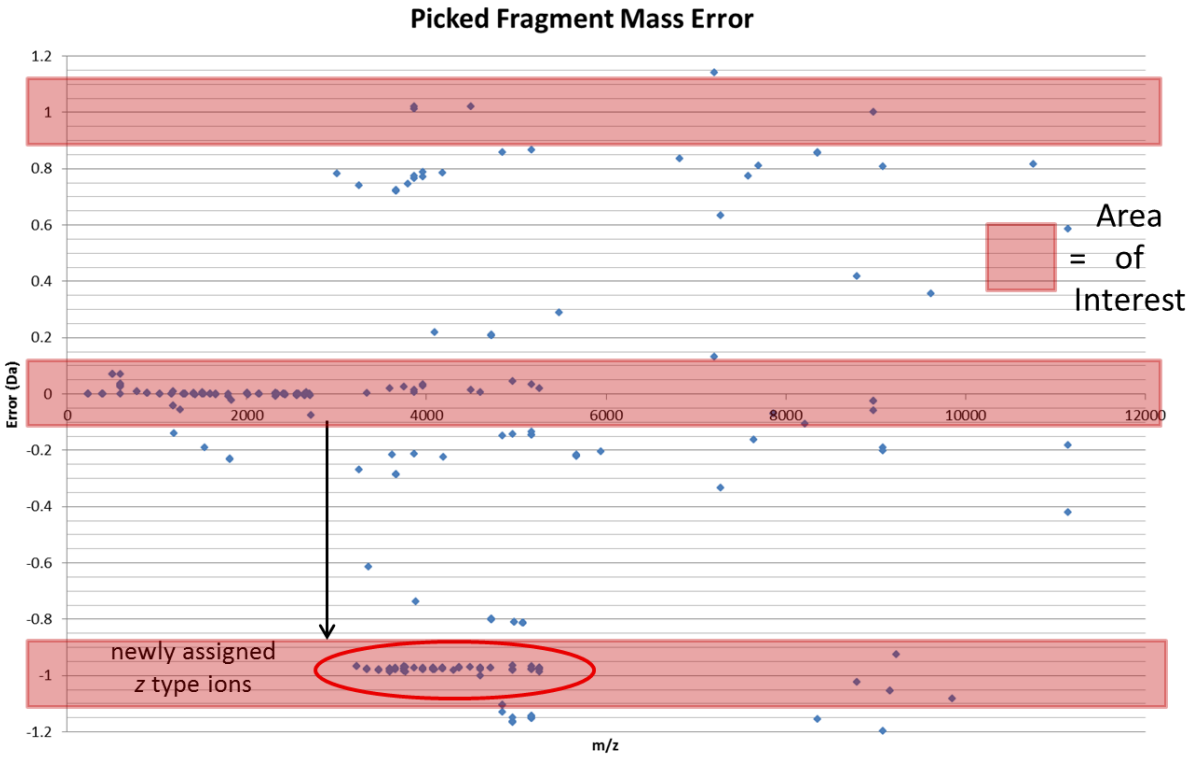


Figure 78 - "Mass jump" in *z* type ion assignment error for the intermediate mass form of PilE-427707C when Ser^{117} is modified with GATDH-Hex and the peak picking tolerance is widened to 300 ppm

Ion	Picked Peak [M+H]+	Theoretical Mass	Error (ppm)	Error (Da)
z_{26}	3338.5493	3339.5270	-292.7518	-0.9777
z_{27}	3466.6057	3467.5856	-282.5635	-0.9798
z_{28}	3594.7057	3595.6805	-271.1042	-0.9748
$z+H_{28}$	3595.7018	3596.6883	-274.2972	-0.9866
z_{29}	3651.7248	3652.7020	-267.5162	-0.9772
z_{30}	3750.7950	3751.7704	-259.9871	-0.9754
$z+H_{30}$	3751.7952	3752.7782	-261.9338	-0.9830
z_{31}	3863.8809	3864.8545	-251.8963	-0.9735
z_{32}	3962.9474	3963.9229	-246.0945	-0.9755
$z+H_{32}$	3963.9513	3964.9307	-247.0063	-0.9794
$z+H_{32}$	3963.9576	3964.9307	-245.4184	-0.9731
z_{33}	4076.0345	4077.0069	-238.5108	-0.9724
z_{34}	4177.0804	4178.0546	-233.1734	-0.9742
z_{35}	4305.1686	4306.1496	-227.8093	-0.9810
z_{36}	4362.1977	4363.1710	-223.0866	-0.9734
z_{37}	4490.2948	4491.2660	-216.2364	-0.9712
z_{38}	4603.3787	4604.3501	-210.9609	-0.9713
$z+H_{38}$	4604.3558	4605.3579	-217.5952	-1.0021
z_{39}	4718.4033	4719.3770	-206.3213	-0.9737
$z+H_{39}$	4719.5848	4720.3848	-169.4949	-0.8001
z_{41}	4960.5363	4961.5149	-197.2422	-0.9786
$z+H_{41}$	4961.5594	4962.5227	-194.1262	-0.9634
z_{43}	5173.6480	5174.6262	-189.0543	-0.9783
$z+H_{43}$	5174.6595	5175.6341	-188.3060	-0.9746
z_{44}	5260.6834	5261.6583	-185.2782	-0.9749
$z+H_{44}$	5261.6790	5262.6661	-187.5702	-0.9871

Table 7 - *z* and *z+H* ions picked at mass difference ≈-1Da in this table duplicates have been removed and the median value picked in each case to reduce the list whilst giving a representative list

Of all of the known PTMs for PilE the 434.192 Da mass for this proposed PTM is closest to that of GATDH-Hex. The low mass region of the ECD spectrum was therefore examined to check whether some oxonium ions had been formed from vibrational activation during the ECD experiment. Surprisingly two singly charged ions were present at *m/z* 435.197313 and 437.176602 in an approximate 3:1 ratio (Figure 79). The mass of the ion at m/z 437.176602 corresponds exceptionally well to the accurate mass expected from the GATDH-Hex oxonium ion. The associated error and elemental composition are shown in the figure inset. However, the ion at *m/z* 435.197313 could not be explained by any theoretical fragment ion of PilE.

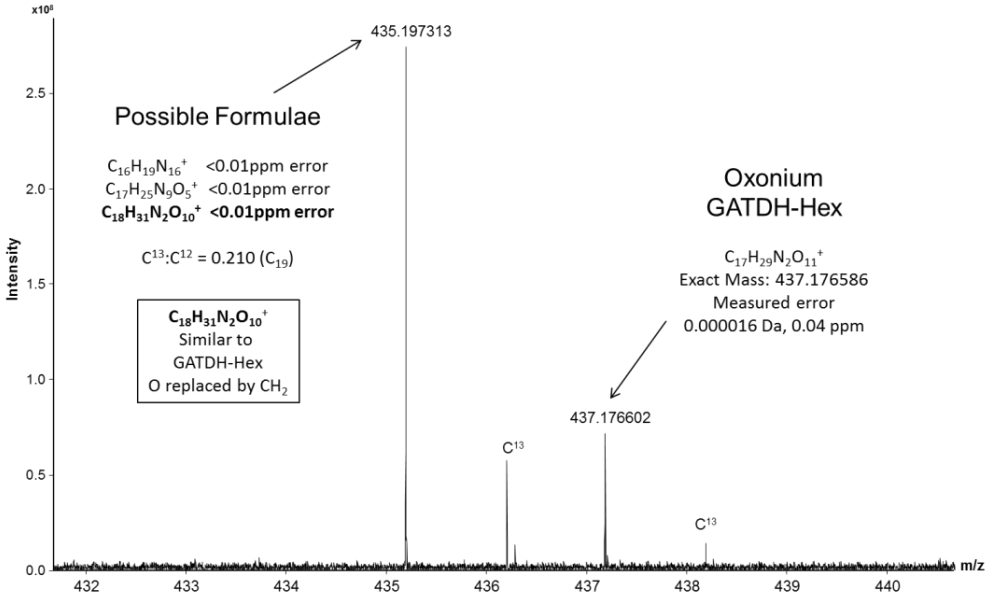


Figure 79 - Zoom of the *m/z* 432-440 mass range of a nano-ESI ECD FT-ICR MS/MS spectrum of the lowest proteoform of PilE-427707C

The accurate mass for this ion was used to generate possible molecular formulae. Three formulae were possible at less than 0.01 ppm error. The intensity ratio of the ^{13}C to ^{12}C isotope peaks was also used to estimate the number of carbon atoms in the molecule giving a value of 19. When combined these results suggested that a formula of $C_{18}H_{31}N_2O_{10}^+$ provided the best fit with the observed mass. Interestingly this formula is similar to oxonium ion for GATDH-Hex but with a CH_2 group in place of an oxygen atom.

A molecule with such a molecular formula would result in a PTM mass of +434.190 Da which also fits well with the hypothetical PTM mass of 434.197 Da estimated during the ion assignment on Ser^{117}. Before additional PTM assignment is continued on the intermediate mass form of PilE-427707C, it was decided to perform some additional experiments in order to check whether the ion observed in the low mass region of the ECD spectrum at m/z 435.197 originated from a *bona fide* PTM.

7.2. Characterisation and Identification of the 434 Da PTM

In-source fragmentation has previously been used to examine the glycosylation state of PilE[3]. Moderate energy ISD is known to break the protein-glycan bond and form abundant glycan oxonium ions without causing extensive fragmentation of the protein backbone. These reporter ions are characteristic of particular glycan structures and provide a fingerprint to identify any attached glycan. ISD was therefore performed on PilE-427707C in an LTQ Orbitrap mass spectrometer (Figure 80). Note that the Orbitrap platform was chosen over FT-ICR for these experiments as the MS^n experiments that follow are much more easily performed in the LTQ ion trap than the ICR cell.

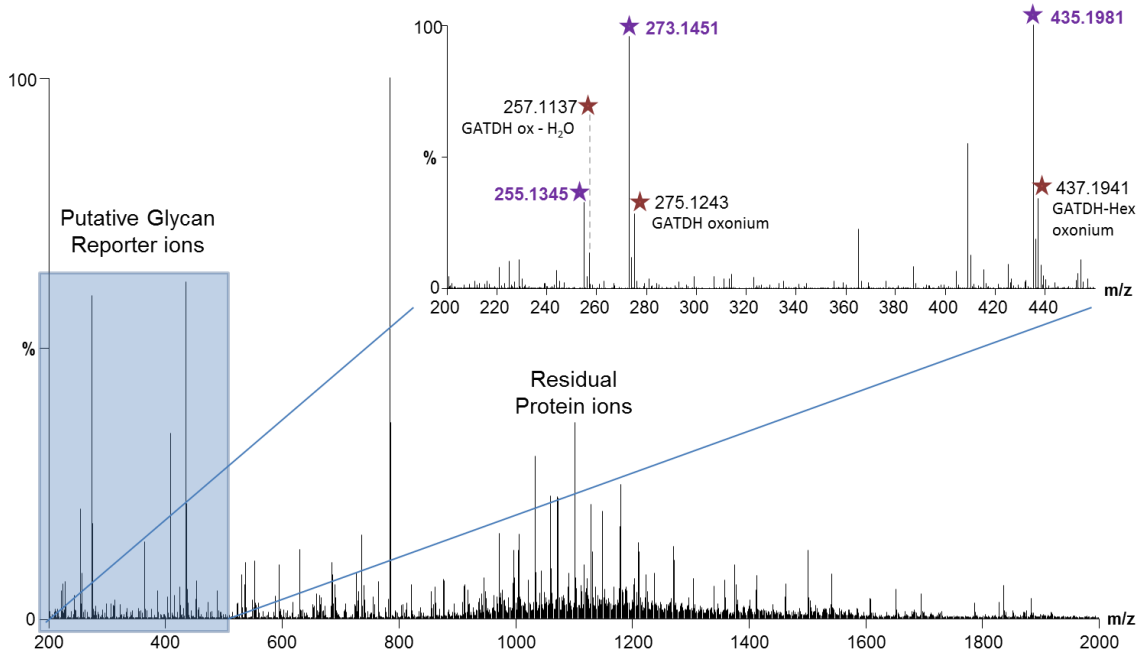


Figure 80 - nano-ESI ISD Orbitrap spectrum of PilE-427707C. Glycan ions are shown in the inset. ★Red stars represent ions related to GATDH-Hex and ★purple stars and purple text represent ions related to an unknown precursor

The resulting ISD spectrum showed several high abundance ions in the expected mass range for glycan reporter ions (Figure 80). The single charged ions marked by red stars in the inset are all known reporter ions for the GATDH-Hex glycan. The ions at m/z 437.1941, 275.1243 and 257.1137 are the GATDH-Hex oxonium ion, the GATDH oxonium ion (formed through facile loss of a hexose from GATDH-Hex) and the GATDH oxonium ion with loss of H_2O respectively. Curiously these GATDH type ions were always accompanied by ions at -1.979 m/z that were approximately three times more abundant. What is more, the ion pair at *m/z* 437 and 435 is extremely similar in both mass and intensity ratio to that observed in the ECD spectrum. Taken

together this strongly suggests that these unknown ions highlighted in purple may be reporter ions for a *bona fide* PTM.

In order to further probe the structure of the molecule responsible for the ion at *m/z* 435, this ion was individually selected in the LTQ using a small window of ≈± 1 Da and subjected to CAD. The resultant spectrum is shown in the upper panel of Figure 81. A similar spectrum of the 437 ion, whose fragmentation pattern confirms it to be the oxonium ion of GATDH-Hex, is also presented in the lower panel. Upon CAD the m/z 435 ion forms two main daughter two ions at m/z 273 and 255. This bears a remarkable similarity to fragmentation of the GATDH-Hex ion at 437 *m/z* and suggests that the 435 also loses a hexose residue (≈162 Da) resulting in a core ion that is prone to loss of H_2O. It also provides the first piece of evidence that the ion at 435 may have a similar structure to GATDH-Hex.

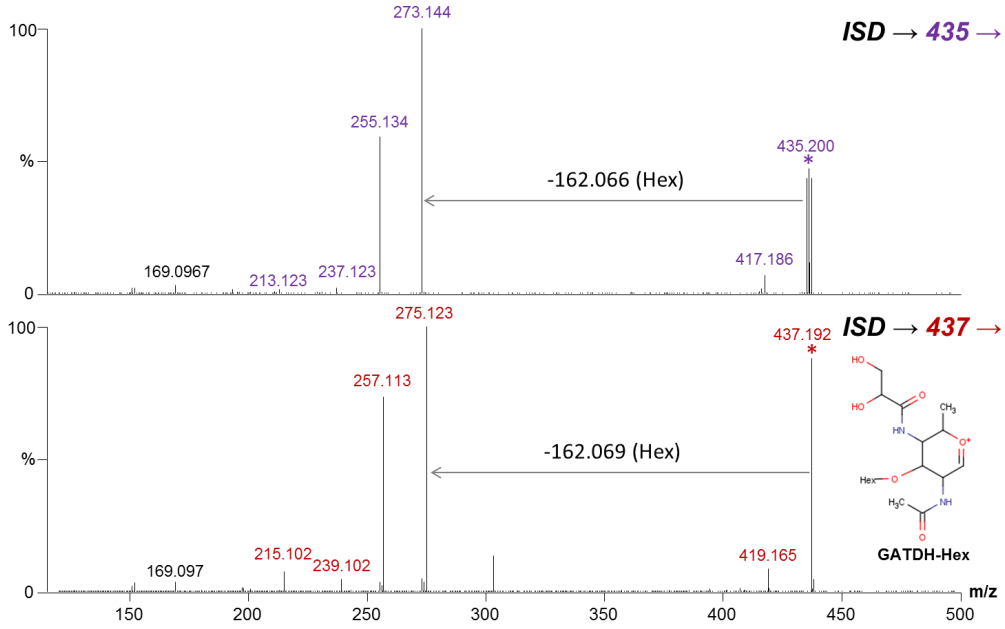


Figure 81 - MS^3 spectrum of the ions at 435 (upper panel) and 437 (lower panel). The structure of the GATGH-Hex oxonium ion is given for reference

The daughter ion at m/z 255 was subjected to a further round of MS (MS^4) resulting the spectrum shown in the upper panel of Figure 82. The fragmentation spectrum of the m/z 257 ion which issued from m/z 437 is provided for reference in the lower panel. Again there are remarkable similarities between the spectra issuing from the ions at 235 and 237 *m/z*, particularly in the low mass range where the daughter ions at *m/z* 169, 152, 151, 127 and 110 are all at exactly the same mass. These ions are formed through loss of side chains (see Figure 82, lower panel) and are characteristic of the glycan core of both DATDH and GATDH[3]. The presence of these ions in the

spectrum of 255 thus supports the hypothesis that the central structure of the parent ion at m/z 435 is that of a DATDH/GATDH like glycan (it will therefore be referred to as unknown glycan).

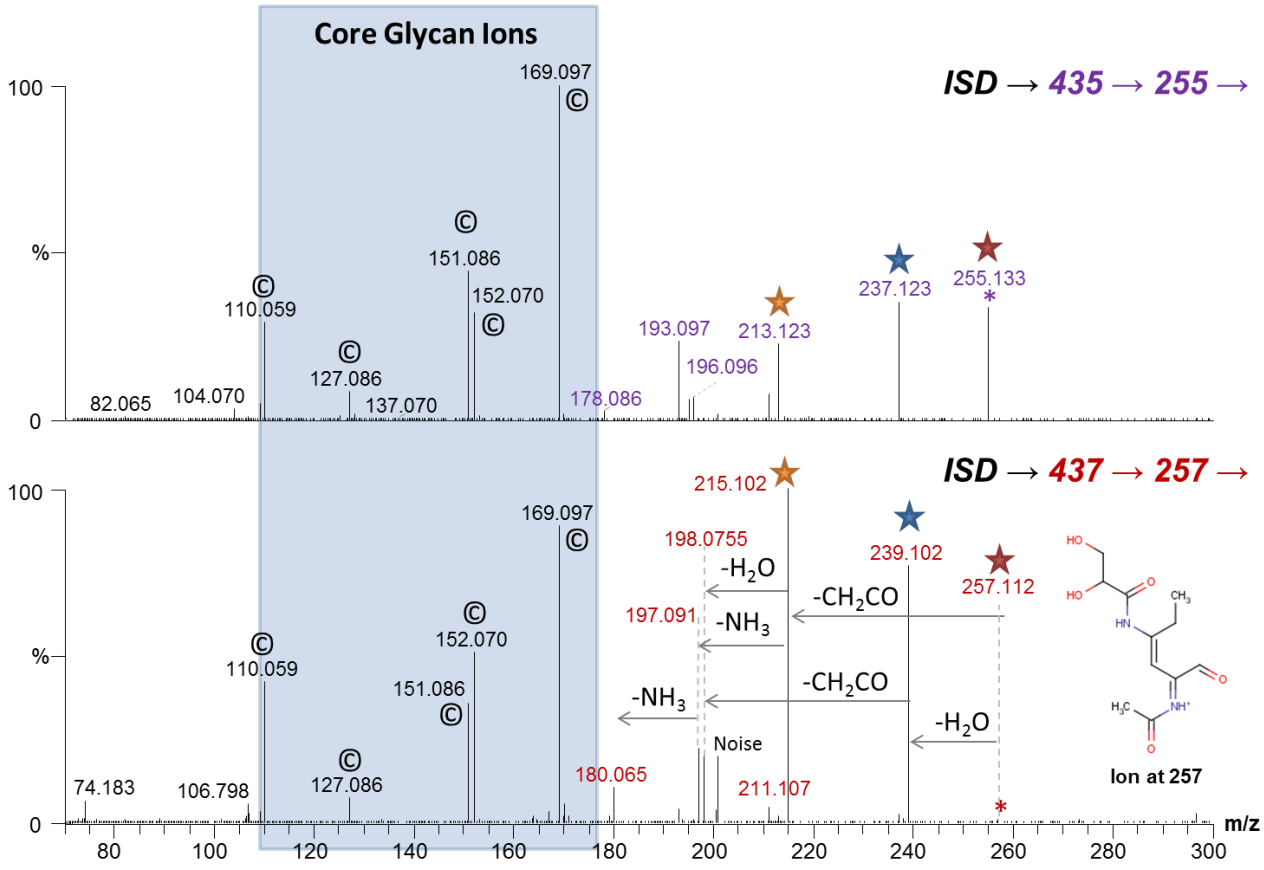


Figure 82 - MS^4 spectrum of the ions at 255 (upper panel) and 257 (lower panel)

Further examination the upper mass range of the 255 spectrum (Figure 82) also reveals a surprisingly similar pattern of ions at -1.98 Da to those observed in the reference spectrum. This suggests that the unknown glycan is functionalised in a similar fashion to GATDH. In particular the ion at m/z 213 confirms the presence of at least one N-acetyl group (213 is probably formed through loss of CH_2CO from 255, just as 215 is from 257). In the case of GATDH the corresponding ion at m/z 215 can further fragment losing the glycerol group (88 Da) and forming the core glycan ion at m/z 127. It was therefore of interest to see if a similar pattern is observed for the 213 ion of the unknown glycan.

The fragmentation of *m/z* 213 was performed in a MS^6 experiment (Figure 83). Note that the 255 ion in this experiment is produced thorough CAD of the 273 ion via the complete ISD→437→275→257→215 pathway. At this level of MS both the *m/z* values and the ion intensities are very low and thus detection is performed in the LTQ (hence the mass shift and lower mass accuracy) rather than the Orbitrap, in order to prevent signal loss during transmission.

Fragmentation of the 215 ion performed in a similar fashion is shown as a reference in the lower panel.

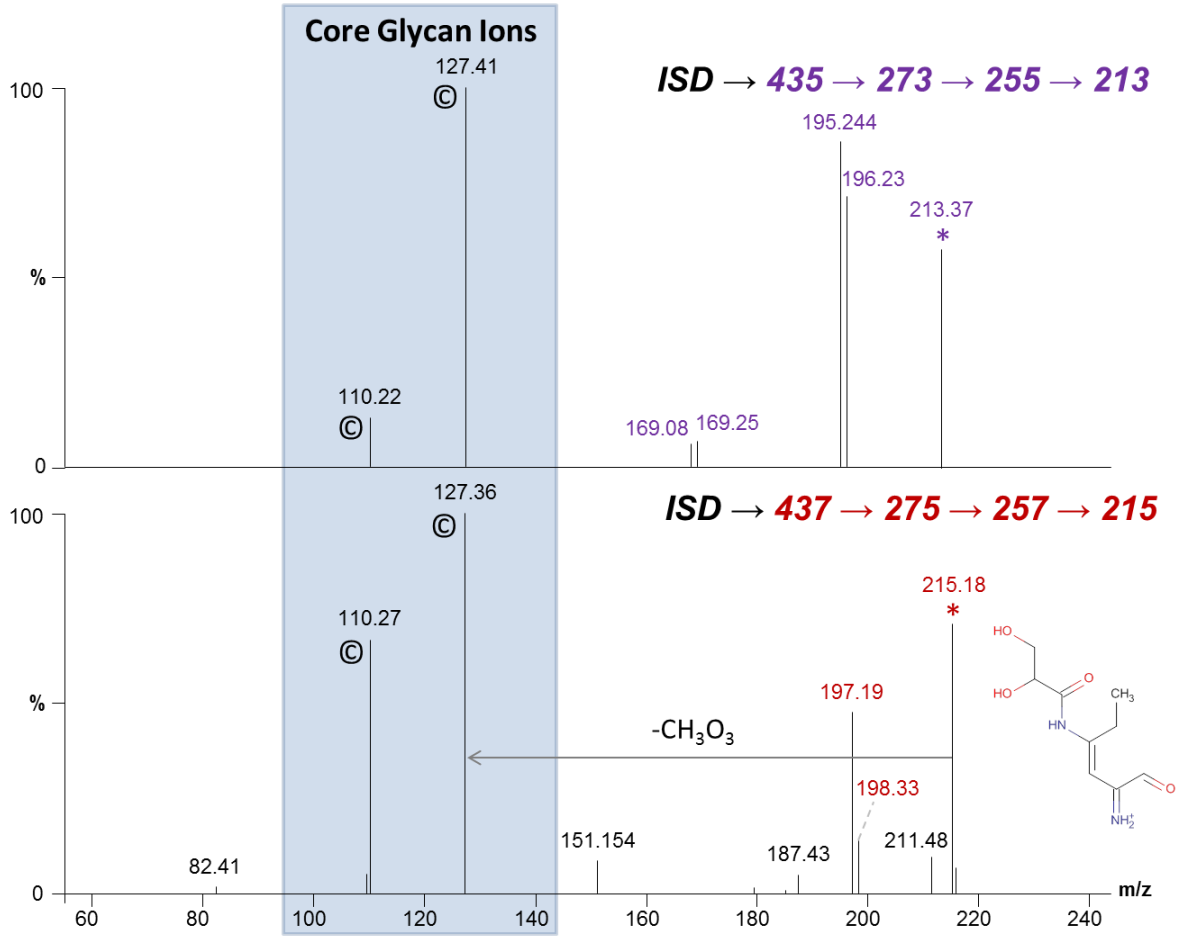


Figure 83 - MS^6 The upper panel is a sum of the 213 isolated spectrum as in the CAD spectrum the 213 ion is of extremely low intensity

Upon fragmentation of the ion at m/z 213, two core glycan daughter ions are produced at m/z 127 and 110. Fragmentation of the 215 ion from GATDH forms the same ions. This provides further evidence supporting the fact that the core structure of the unknown glycan is the same as GATDH. Furthermore it localises the mass difference between GATDH-Hex and the unknown glycan to the functionality at the either the 2 or 4 position.

Combining this fragmentation data with the elemental composition of the parent ion at m/z 437, the functionalised amido group at the 2 or 4 positions is proposed to have the formula C_3H_7O rather than $C_2H_5O_2$ in GATDH. Several structural isomers are possible for this elemental composition that can be grouped into alcohols (A) and ethers (B) (Figure 84).

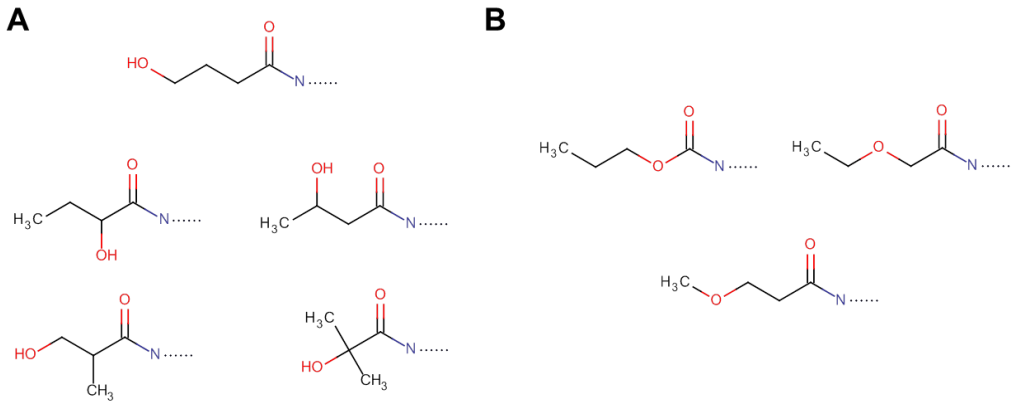


Figure 84 - Possible isomers for the amido functionality at position 2 or 4. A) alcohols B) ethers

Further characterisation of this functionality was attempted by *O*-linked glycan release from a tryptic digest of PilE followed by permethylation using a relatively recent spin column protocol[4, 5]. It was then planned to characterise the permethyalted glycan by GC-MS. An ammonium:borane based elimination protocol was used successfully to detach the sugar from PilE-427707C and the spin column permethylation protocol was also successful when trialled on a maltodextrin standard. However, when the two were combined on PilE-427707C results were inconclusive.

Without a definitive structure for the amido group at position 2 or 4 it is difficult to name this sugar. Since the core glycan is believed to be similar to GATDH and the amido group is supposed to have four carbon atoms, the name butyramido 2-acetamido 4-butyramido 2,4,6-trideoxy α-D-hexose or BATDH is tentatively proposed (Figure 85). Thus the 434 Da modification will be called BATDH-Hex which has an elemental composition of $C_{18}H_{31}N_2O_{11}$ and a monoisotopic modification mass of 434. 1900 Da.

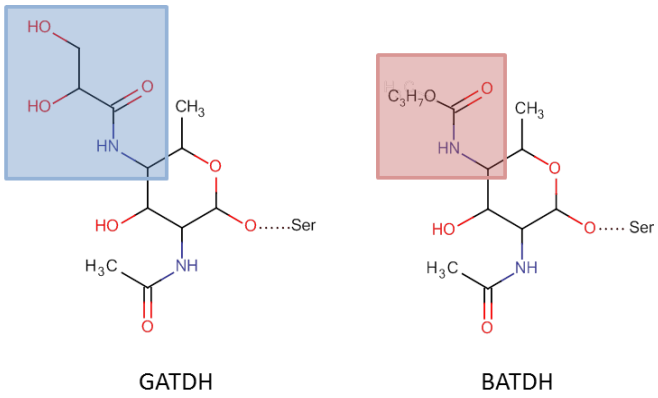


Figure 85 - Proposed structure of the BATDH glycan and comparison to GATDH (difference in functionalisation at the 4 position is highlighted)

7.3. PTM of Intermediate Mass Form – Part II

Now the identity of the 434 Da PTM has been resolved, PTM assignment can be continued for the intermediate mass form of PilE-427707C. Modification of Ser^{117} with BATDH-Hex increased the *z* ion coverage from the C-terminus to z_{44}. Applying additional BATDH-Hex modification to Ser^{90} and Ser^{87} and PG modification to $Ser^{97}$and Ser^{69} increased the sequence coverage to 88% (Figure 86).

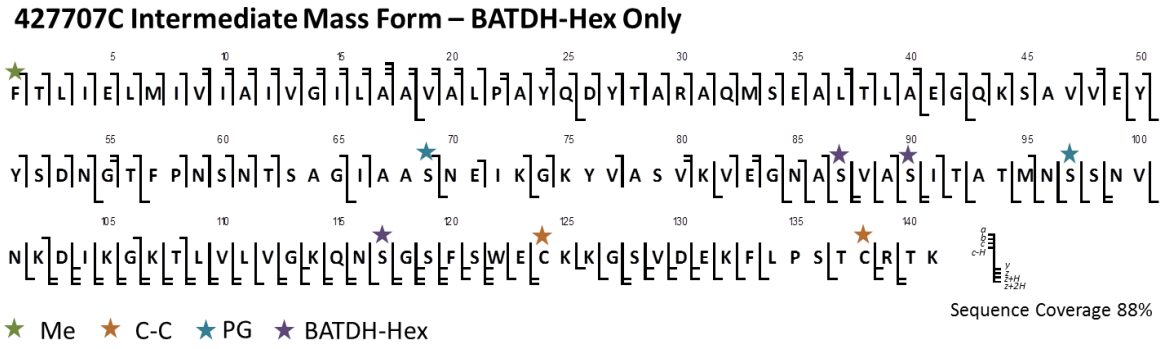


Figure 86 - PTM assignment for intermediate mass form of PilE-427707C. PTMs are three BATDH-Hex and two PG groups

Despite the fact that this PTM assignment produces an excellent sequence coverage at a low peak picking error threshold (5 ppm), the molecular mass with this assignment is 16505.203 Da compared to 16506.182 Da measured in the high resolution mass profiling experiment. The difference between these values is -0.979 Da and is thus difficult to explain by deamidation; which would result in an increase in mass of 0.984 Da not a decrease, or misappropriation of an H atom (1.008 Da). In addition, the ISD spectra suggested that a GATDH-Hex glycan should probably be amongst the PTM complement, and the PTM assignment in Figure 86 does not contain GATDH-Hex.

More combinations of PTMs were therefore trialled and it was discovered that two alternative PTM assignments both with one GATDH-Hex and two BATDH-Hex subunits also produced good sequence coverage (Figure 87). The difference between the two lies in the position of the GATDH-Hex: in panel A it is on Ser^{63}, in panel B it is on Ser^{90} (residues highlighted in the figure).

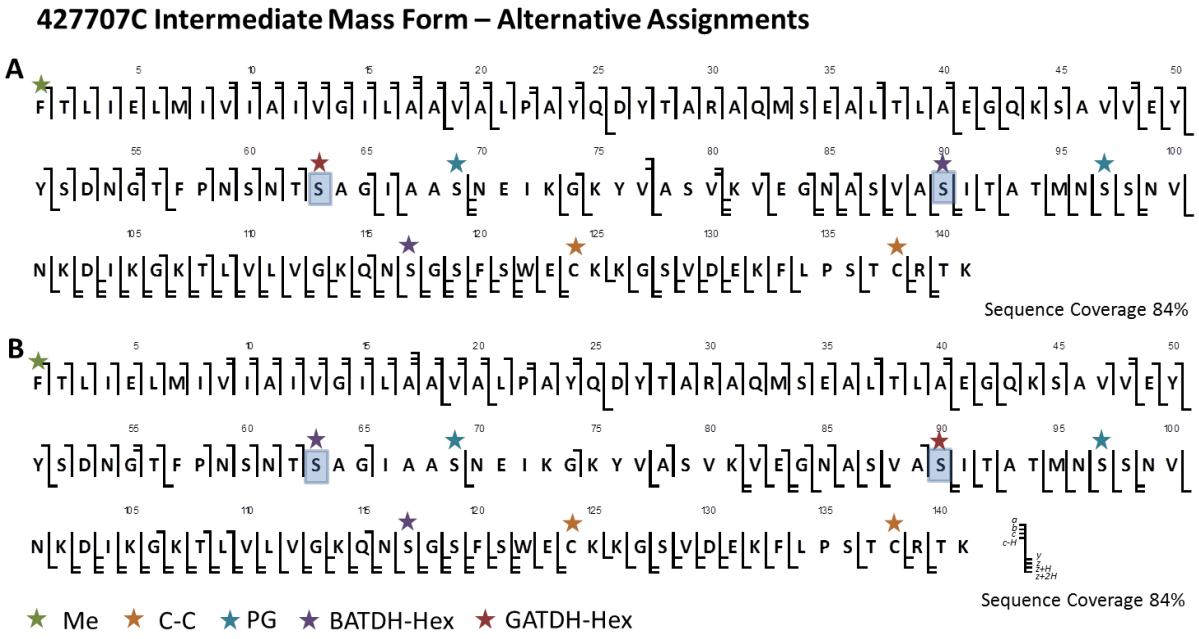


Figure 87 - Two alternative PTM assignments for intermediate mass form of PilE-427707C. In both cases PTMs are one GATDH-Hex, two BATDH-Hex and two PG groups

The mass of PilE with this PTM complement is 16507.182 Da and is therefore 1.000 Da greater than the 16506.182 Da measured in the mass profiling experiment. This was extremely puzzling and a great deal of effort was expended checking the spectral acquisition and data analysis for problems that would result in ≈1 Da mass shifts. No experimental exlaination for the ≈1 Da mass shift was found.

Once all experimental factors had been ruled out, attention was turned to the isotopic patterns of the protein. Upon close examination of the experimental distributions making up the three major peaks in the mass profiling spectrum, each seemed slightly broader than one would expect from an averagine type distribution. Concentrating on the intermediate mass form, theoretical distributions were created for each of the two possible PTM assignments and fitted against the experimental distribution (Figure 88). This was achieved by creating a custom tool in Microsoft Excel (using isotope abundance data from the Bruker tool Compass IsotopePattern) which modelled peaks as ideal Lorenzian distributions at any desired resolution and enabled fitting to an experimental distribution.

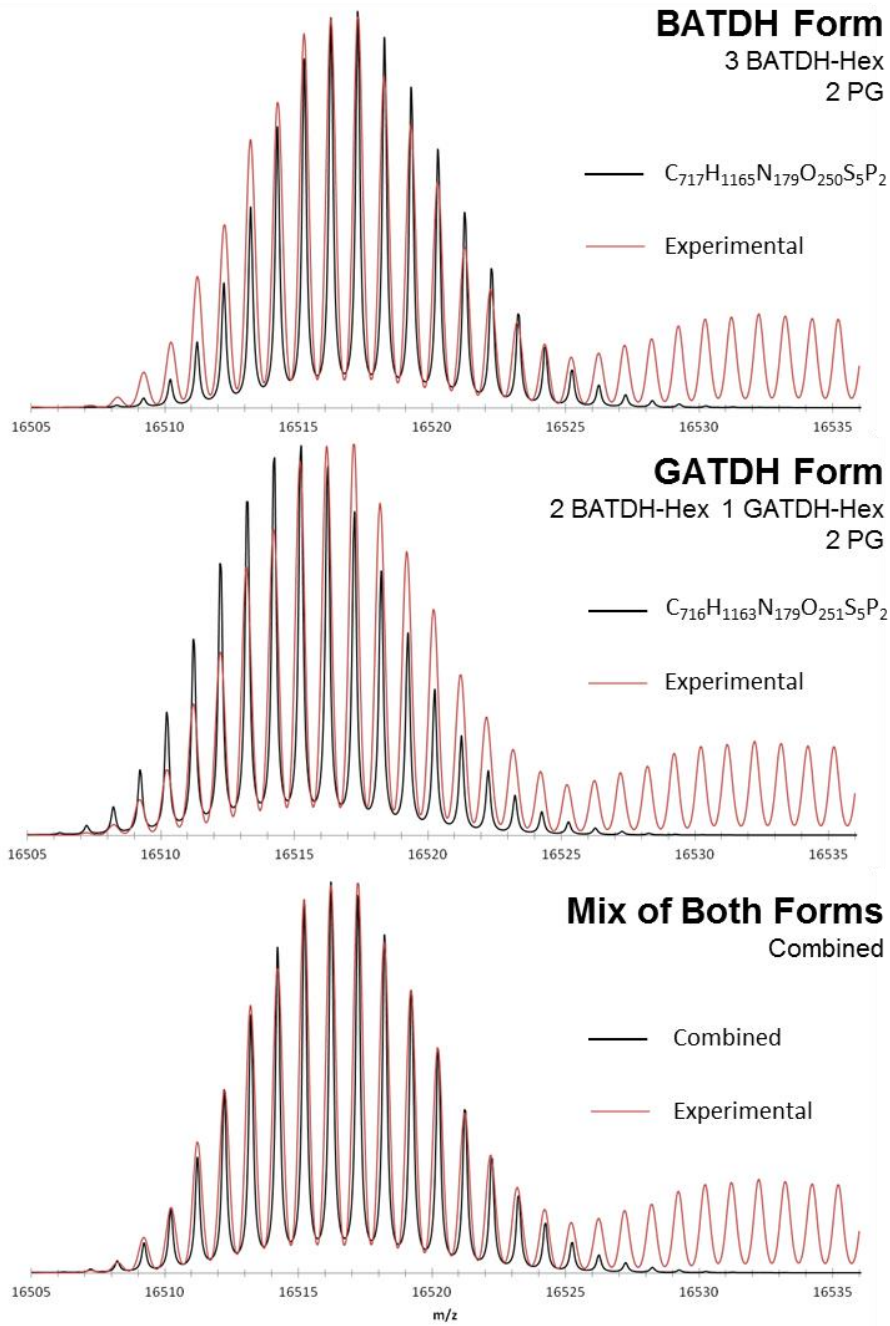


Figure 88 - Fitting of theoretical isotope patterns to the experimental isotope pattern of the intermediate mass form of PilE-427707C. Upper panel GATDH form, middle panel BATDH form, lower panel a mix of the two forms in a 35:65 ratio

The first theoretical pattern with three BATDH-Hex and two PG PTMs ("BATDH form") was shifted to the right of the experimental data by approximately 1 Da (Figure 88 upper panel) whereas the second with one GATDH-Hex, two BATDH-Hex and two PG groups ("GATDH form") was shifted approximately 1 Da to the left (Figure 88 middle panel). When these patterns were combined in a 35:65 ratio they produced a much better fit to the experimental data (Figure 88 lower panel). As

deconvolution may have distorted the protein isotope pattern somewhat, this matching process was also performed on the 15^+ and 16^+ charge states with similar results.

This led to the conclusion that the experimentally observed isotope pattern from this intermediate mass form was in reality a mix of two patterns from two different proteoforms of PilE-427707C, one with three BATDH glycans and the other with two BATDH-Hex and one GATDH-Hex glycan. The apparent mass difference between these assignments and the measured protein mass arose because the SNAP 2.0 peak picking algorithm used by the Bruker software found the best fitting isotope pattern to be exactly in-between those of the two proteoforms.

7.4. PTM of Low and High Mass Forms of PilE-427707C

Low Mass Form

MS/MS was performed on the 16^+ charge state of the low mass from of PilE-427707C and the fragment mass lists combined from three separate experiments. Upon PTM assignment modification of Ser^{117} with BATDH-Hex was required to increase sequence coverage past this residue, however fragmentation in the central region of the protein was less extensive than for the intermediate mass proteoforms and this made additional PTM assignment more rather difficult.

Similar to the explaination that involves the intermediate mass form, two different sets of PTMs could be proposed. The "BATDH form" has two BATDH-Hex on Ser^{117} and Ser^{90}, and two PG groups on Ser^{69} and Ser^{97} or Ser^{98} (Figure 89A). The "GATDH form" contains one GATDH-Hex, one BATDH-Hex and two PG groups but again there are two possible assignments that give very similar sequence coverage. Either one GATDH-Hex at Ser^{63}, one BATDH-Hex at Ser^{117} and PG groups at Ser^{69} and Ser^{97} or Ser^{98} (Figure 89B) can be placed on the protein backbone or a BATDH-Hex at Ser^{63}, a GATDH-Hex at Ser^{117} and PG groups at Ser^{69} and Ser^{97} or Ser^{98} can be assigned(Figure 89C). These two assignments differ only in the position of the GATDH-Hex glycan (residues are highlighted in blue in the figure).

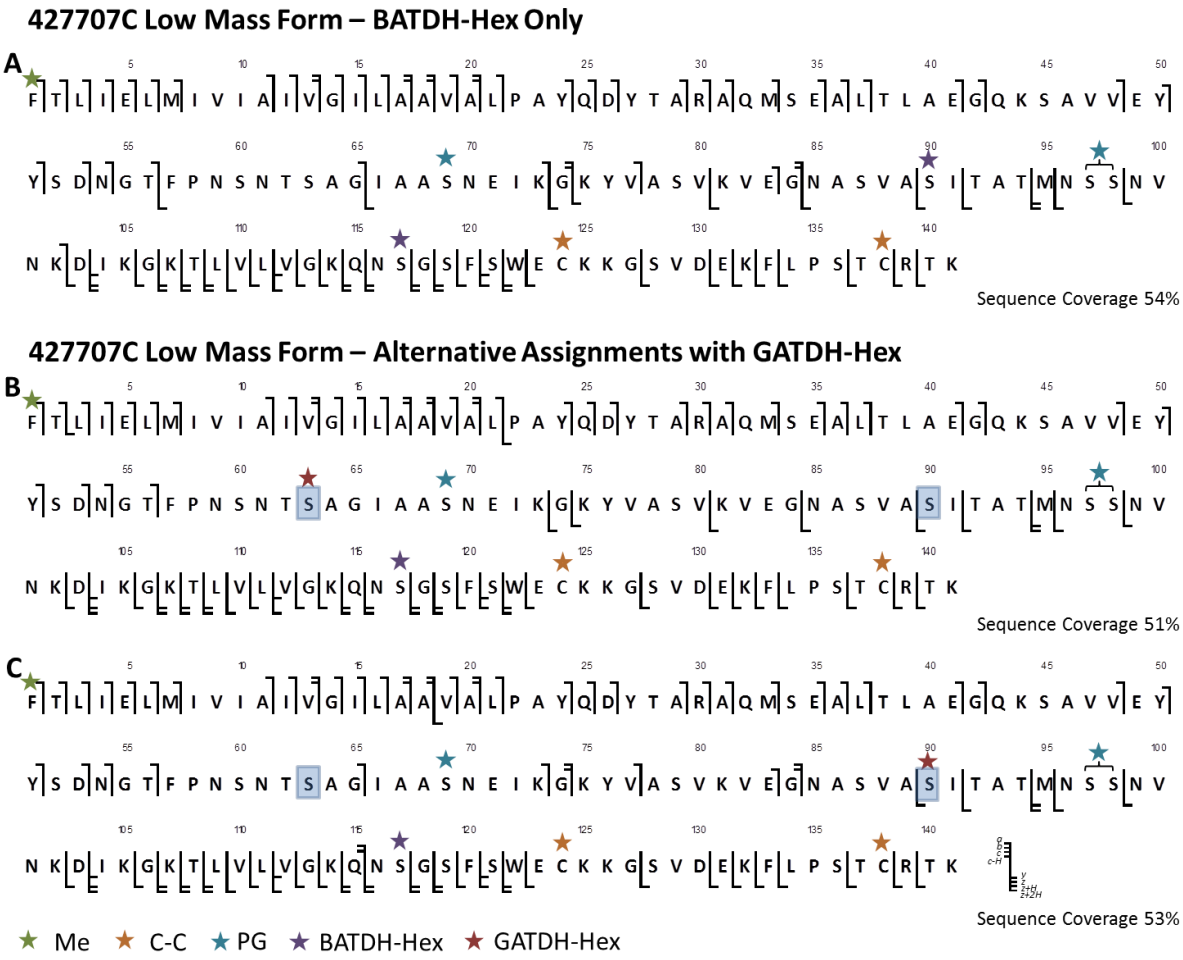


Figure 89 - PTM assignments for low mass form of PilE-427707C. A) BATDH-Hex form. B & C) Two alternative assignments with one GATDH-Hex

The mass of the "BATDH form" is 16071.013 Da and the mass of the "GATDH form" is 16072.992 Da. Neither matches the experimental protein mass of 16072.003 Da and neither theoretical distribution exactly matches the measured experimental distribution; either at the whole protein level or the 15^{+} and 16^{+} charge states. Similar to the intermediate mass form a mix of the two theoretical patterns provided a very good fit, this time in a 55:45 ratio. It appeared that the peak picking software had again found the best fitting isotope pattern to be exactly in-between the two distributions.

High Mass Form

MS/MS was performed on the 17^{+} charge state of the high mass form of PilE-427707C. The results presented here are from a single experiment. As for the low and intermediate mass forms, two general sets of PTMs can be assigned. The "BATDH form" has four BATDH-Hex glycans on Ser^{117}, Ser^{90}, $Ser^{87}$and $Ser^{63}$and two PG groups on Ser^{69} and Ser^{97}. For the GATDH form there were again

two assignments that provided very similar sequence coverage: the first has the GATDH-Hex on Ser^{63} and BATDH-Hex glycans on Ser^{117}, Ser^{90} and Ser^{87} with two PG groups on Ser^{69} and Ser^{97}, and the second has GATDH on Ser^{90} with BATDH-Hex glycans on Ser^{117}, Ser^{87} and Ser^{63} with two PG groups on Ser^{69} and Ser^{97}. The residues that differ are again highlighted in blue.

This time the mass of the "GATDH form" 16941.3722 Da compares well with the experimental mass of 16941.3727 Da, but on closer inspection the isotopic patterns were broader than the "GATDH form" pattern alone and a 30:70 mix of the "BATDH form" and "GATDH form" at 16939.3929 Da was required in order to fit a theoretical distributions to the experimental pattern. In this case it appears that the software had picked the experimental peak of the high mass form.

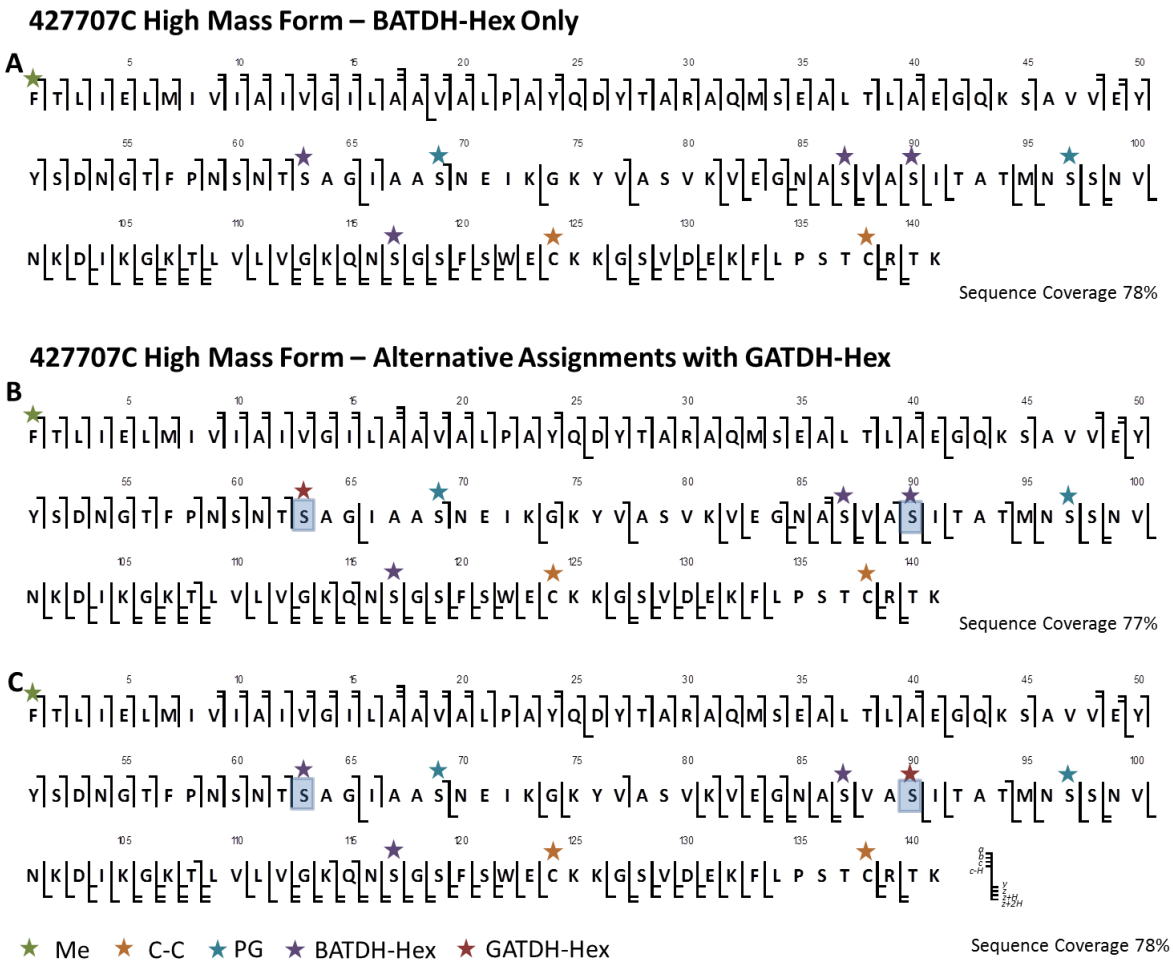


Figure 90 - PTM assignments for high mass form of PilE-427707C. A) BATDH-Hex form. B & C) Two alternative assignments with one GATDH-Hex

7.5. Summary of PTM Assignment of PilE-427707C

PilE purified from the 427707C strain appears to be expressed in at least six proteoforms that vary in the number of disaccharide substituents. These proteoforms may be divided into two groups. In the first, PilE is modified with two PG moieties and between two and four BATDH-Hex subunits. In the second, PilE is modified with two PG groups, one GATDH-Hex subunit and between one and three additional BATDH-Hex glycans. The location of the second GATDH-Hex subunit is uncertain and could either be at Ser^{90} or Ser^{63} position. Previous biological knowledge would suggest the Ser^{63} position is more likely, however sequence coverage is slightly better when GATDH-Hex is placed on Ser^{90} in both the low and high mass forms. These assignments are summarised in Figure 91.

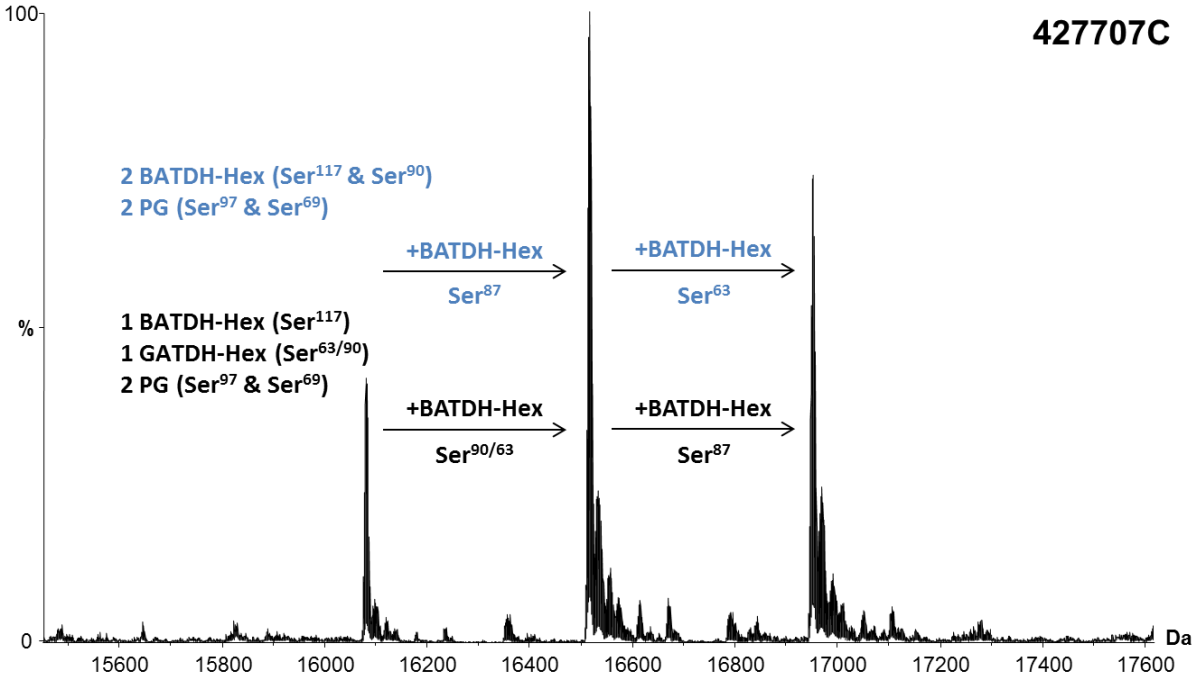


Figure 91 - Summary of PTMs present on PilE-427707C

8. Conclusions from Deep Characterisation of PilE-427707C

8.1. Mass Spectrometry

PilE from strain 427707C has proved to be a good test for the top-down methodology. From a rather simple looking mass profiling spectrum six potential proteoforms and a new glycan modification have been identified. Without isotopic resolution mass profiling and highly accurate fragmentation data it would have been almost impossible to identify the two BATDH-Hex glycans *de novo* as was the case here.

Sequence coverage was good (78%) even when fragmentation was performed on a single charge state (high mass form). The combination of fragmentation data from three separate charge states proved especially useful for improving sequence coverage, without which the GATDH form would have been difficult to identify, and helped localise the PG modification onto Ser^{97} rather than Ser^{98}. This was not possible for the low mass proteoforms.

Ion assignment was complicated by the presence of both the GATDH and BATDH forms but fortuitously the ≈2 Da mass difference between these alternative glycan PTMs meant that there was no overlap of the *z* and *z+H* ions from either form. For this reason there can be a reasonable degree of confidence in the ion assignment. It must be mentioned that artificial deamidation has previously been reported on PilE and Asn^{54} & Asn^{85} and to a lesser extent Asn^{99} & Asn^{91} may be prone to degradative deamidation in PilE-427707C. The extent of deamidation was difficult to judge here because of the complexity of the proteoform population.

For the GATDH form two assignments have been proposed for the low, intermediate and high mass proteoforms. It was very difficult to support one or the other based on the fragmentation data available. This highlights both the need for good scoring systems and automatic data processing that could quickly tests and scores all possible PTM forms. It also underlines the requirement for more complex scoring systems than those solely based on the number of assigned fragments (such as the P score). It would be interesting to see how a software tool handled *a priori* identification of the BATDH-Hex modification and whether it would have been able to predict the correct modification mass or not. It may also be possible to separate the intact proteoforms either biochemically or by ion mobility. If successful this would greatly simply PTM assignment.

8.2. Biological Relevance

It is clear that the PilE from the 427707C strain is also highly glycosylated. To confirm this, 427707CΔ*pglC* and 427707CΔ*pglD* deletion mutants were prepared. However, after more than five separate attempts at PilE preparations, no band for PilE could be visualised by Coomassie stained SDS-PAGE and no proteins were observed in the appropriate mass range by intact mass

profiling. It was initially thought that in this strain the lack of glycan may cause a change in solubility that resulted in loss of PilE during the preparation. Using a polyclonal antibody raised against PilE-8013, PilE-427707C was followed during each stage of the purification by Western blot. Reactivity with the antibody was however rather poor and these results were inconclusive.

Bacterial aggregation assays showed a normal phenotype for the Δ*pglC* and Δ*pglD* strains, strongly suggesting that pili were expressed by these strains and pili were visualised on the 427707CΔ*pglD* strain by transission electron microscopy (Figure 92). The inability to purify pili from the *pglC* and *pglD* mutants precluded the use of MS to validate the top-down MS results.

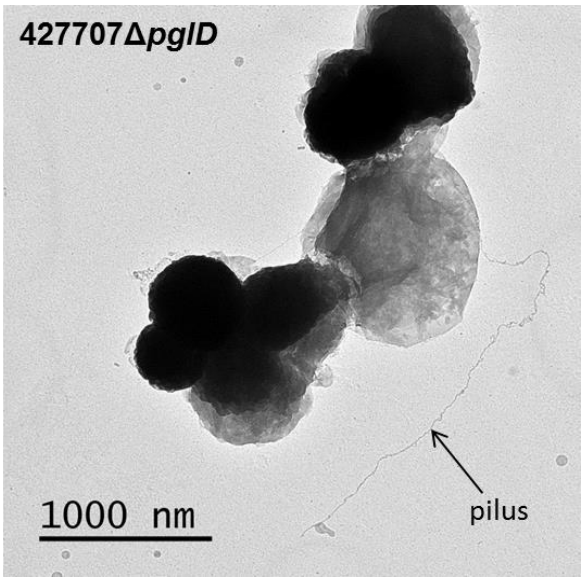


Figure 92 - Transmission electron micrograph of bacteria from the 427707C *pglD* mutant confirming the expression of pili

The identification of the new glycan BATDH is a rather unexpected result. The *pglB* gene is responsible for defining the glycan functionality at the 4 position[6] and codes for a protein having two catalytic domains, an N-terminal phosphoglycosyl transferase domain and C-terminal acetyl transferase domain (this transfers an acetyl group to the UDP-4-amino precursor). Alternatively, the acetyl domain may be replaced by ATP-grasp domain that functionalises the 4 position of the glycan with a glycerol moiety. In this case the protein is coded for by the *pglB2* gene[3]. Half of clinical isolates are known to carry the *pglB1* allele and thus express DATDH derived glycans, and the other half a *pglB2* allele and a GATDH core glycan.

The presence of the BATDH glycan suggests that there is an additional factor at play here. This could conceivably come from some unknown function encoded by the *pglB* gene. Indeed the genetic organisation of the *pglB2* region is rather complex with four distinct topologies, three of

which have been described in *Neisseria* spp.[7] The significance of the different *pgl* gene structures is currently unknown and sequencing data has shown the 427707C strain to express the more complex overlapping ORF (results not shown). Alternatively there may be other enzymes that act on the GATDH glycan downstream or alternative substrates for the ATP grasp domain. The presence of BATDH-Hex remains confounding and warrants further investigation.

What is clear is that the clinical strains analysed to date exhibit hitherto unreported levels of glycosylation. This is in stark contrast to the previously characterised 8013 and MS11 strains. It is therefore of interest to see if this phenomenon is common to all strains bearing invariable type PilE sequences.

Bibliography

[1] C. Notredame, D. G. Higgins and J. Heringa.T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **2000**, *302*, 205.

[2] P. Di Tommaso, S. Moretti, I. Xenarios, M. Orobitg, A. Montanyola, J.-M. Chang, J.-F. Taly and C. Notredame.T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Research*, **2011**, *39*, W13.

[3] J. Chamot-Rooke, B. Rousseau, F. Lanternier, G. Mikaty, E. Mairey, C. Malosse, G. Bouchoux, V. Pelicic, L. Camoin, X. Nassif and G. Dumenil.Alternative Neisseria spp. type IV pilin glycosylation with a glyceramido acetamido trideoxyhexose residue. *Proceedings of the National Academy of Sciences of the United States of America*, **2007**, *104*, 14783.

[4] P. Kang, Y. Mechref and M. V. Novotny.High-throughput solid-phase permethylation of glycans prior to mass spectrometry. *Rapid Communications in Mass Spectrometry*, **2008**, *22*, 721.

[5] P. Kang, Y. Mechref, I. Klouckova and M. V. Novotny.Solid-phase permethylation of glycans for mass spectrometric analysis. *Rapid Communications in Mass Spectrometry*, **2005**, *19*, 3421.

[6] M. D. Hartley, M. J. Morrison, F. E. Aas, B. Borud, M. Koomey and B. Imperiali.Biochemical Characterization of the O-Linked Glycosylation Pathway in Neisseria gonorrhoeae Responsible for Biosynthesis of Protein Glycans Containing N,N '-Diacetylbacillosamine. *Biochemistry*, **2011**, *50*, 4936.

[7] R. Viburiene, A. Vik, M. Koomey and B. Borud.Allelic Variation in a Simple Sequence Repeat Element of Neisserial pglB2 and Its Consequences for Protein Expression and Protein Glycosylation. *J Bacteriol*, **2013**, *195*, 3476.

Chapter 6

Large Scale Analysis of the Glycosylation Pattern of PilE Expressed by Uncharacterised Clinical Isolates of Neisseria meningitidis

1. Genomic Analysis of PilE Primary Structure

Sequencing of the *pilE* gene from seven strains isolated at the Limoges university hospital has shown that many of the strains express the same PilE sequence. The regions of the genome flanking the *pilE* gene are different in class I and class II strains and thus amplification is achieved by class specific primers. In all seven strains sequencing confirmed that these isolates belong to the group of Nm strains that have a class II genetic organisation. This class expresses PilE with invariable primary sequences. The archetypal class II reference strain is FAM18 which was isolated in the 1980s in North Carolina, USA from the CSF of a young patient suffering with meningitis. It is a member of the ET-37/ST-11 clonal complex that has been associated with disease worldwide[1]. We had at our disposal a closely related, nalidixic acid resistant clone of FAM18 named FAM20. FAM20 was also included in the analysis as a class II reference strain.

In order to place the PilE sequences from the Limoges isolates in the context of previously reported invariable and variable type PilE sequences, a phylogenic tree was created of the PilE sequences from the Limoges isolates, the three groups of invariable PilE sequences and a selection of well characterised variable type sequences (Figure 93). The PilE sequences from the relevant groups were also aligned (Figure 94).

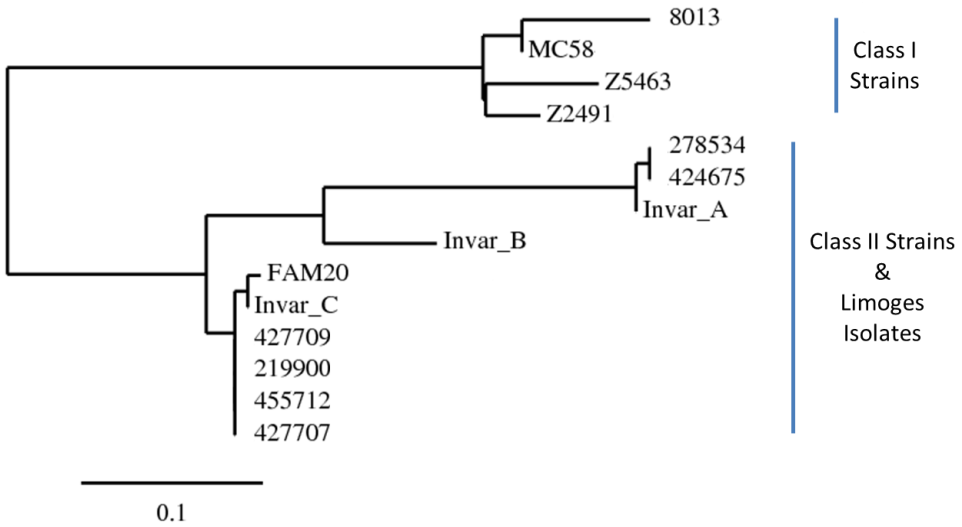


Figure 93 - Phylogenic tree of the PilE sequences from the Limoges isolates studied in this work, the three groups of invariable PilE sequences and a selection of well characterised variable type sequences. The tree was created using the Phlogeny.fr online resource[2]

```
8013    FTLIELMIVIAIVGILAAVALPAYQDYTARAQVSEAILLAEGQKSAVTEYYLNHGEWPGDNSSAGVATSADIKGKYVQSVTVANGVITAQMASSNVNNEIKSKKLSLWAKRQNGSVKWFCGQPVTRTTATA---TDVAAANGKTDDKINTKHLPSTCRDDSSAS
MC58    FTLIELMIVIAIVGILAAVALPAYQDYTARAQVSEAILLAEGQKSAVTEYYLNHGEWPGNNISAGVATSSEIKGKYVKSVEVKNGVVTAQMASSNVNNEIKGKKLSLWAKRQNGSVKWFCGQPVTRDKAKAANDDVTAAAAANGKKIDTKHLPSTCRDASIAS
Z5463   FTLIELMIVIAIVGILAAVALPAYQDYTARAQVSEAILLAEGQKSAVTEYYLNHGTWPGDNNSAGVASSSKIKGKYVKEVEVKNGVVTATMLSSGVNNEIKGKKLSLWAKRQDGSVKWFCGQPVKRNDTA---TDVAADSG--NDKIDTKHLPSTCRDAASAG
Z2491   FTLIELMIVIAIVGILAAVALPAYQDYTARAQVSEAILLAEGQKSAVTEYYLNHGEWPSNNISAGVASSTDIKGKYVQSVEVKNGVVTATMASSNVNNEIKGKKLSLWAKRQDGSVKWFCGQPVKRNDTATTNDDVKADTAANGKQIDTKHLPSTCRDAASAG
        ************************************************** **.:*.*****:*:.******:.* * ***:** * **.******.**********:**********.*  :    ** *  .  ..:*:********** :.*.


278534  FTLIELMIVIAIVGILAAVALPAYQDYTARAQMSEALTLAEGQKTAVVEYYSNNGVFPKNNASAGVASASEIVGKYVAKVDVSGNTITSTMKTSGVNKDISGATLVLEGAQKSGSFTWTCKAGSSNGVKDKFLPSSCRAK
InvarA  FTLIELMIVIAIVGILAAVALPAYQDYTARAQMSEALTLAEGQKTAVVEYYSDNGVFPKNNASAGVASASEIVGKYVAKVDVSGNTITSTMKTSGVNKDISGATLVLEGAQKSGSFTWTCKAGSSNGVKDKFLPSSCRAK
        **************************************************:**********************************************************************************************


FAM20   FTLIELMIVIAIVGILAAVALPAYQDYTARAQMSEALTLAEGQKSAVIEYYSDNGTFPNSNTSAGIAASNEIKGKYVASVKVEGNASVASITATMNSSNVNKDIKGKTLVLVGKQNSGSFSWGCKKGSVDEKFLPSTCRTK
InvarC  FTLIELMIVIAIVGILAAVALPAYQDYTARAQMSEALTLAEGQKSAVIEYYSDNGTFPNSNTSAGIAASNEIKGKYVASVKVEGNASVASITATMNSSNVNKDIKGKTLVLVGKQNSGSFSWECKKGSVDEKFLPSTCRTK
427709  FTLIELMIVIAIVGILAAVALPAYQDYTARAQMSEALTLAEGQKSAVVEYYSDNGTFPNSNTSAGIAASNEIKGKYVASVKVEGNASVASITATMNSSNVNKDIKGKTLVLVGKQNSGSFSWECKKGSVDEKFLPSTCRTK
219900  FTLIELMIVIAIVGILAAVALPAYQDYTARAQMSEALTLAEGQKSAVVEYYSDNGTFPNSNTSAGIAASNEIKGKYVASVKVEGNASVASITATMNSSNVNKDIKGKTLVLVGKQNSGSFSWECKKGSVDEKFLPSTCRTK
455712  FTLIELMIVIAIVGILAAVALPAYQDYTARAQMSEALTLAEGQKSAVVEYYSDNGTFPNSNTSAGIAASNEIKGKYVASVKVEGNASVASITATMNSSNVNKDIKGKTLVLVGKQNSGSFSWECKKGSVDEKFLPSTCRTK
427707  FTLIELMIVIAIVGILAAVALPAYQDYTARAQMSEALTLAEGQKSAVVEYYSDNGTFPNSNTSAGIAASNEIKGKYVASVKVEGNASVASITATMNSSNVNKDIKGKTLVLVGKQNSGSFSWECKKGSVDEKFLPSTCRTK
        **********************************************:********************************************************************** ******************
```

Figure 94 - Clustal alignment of PilE expressed by Limoges isolates and other class I and class II strains

The tree in Figure 93 is divided into two discrete sections, the upper branch containing variable type PilE sequences and the lower branch the invariable type. The sequence expressed by 278534 and 424675 are closely related to that expressed by the invariable type A group and indeed the sequence alignment data in Figure 94 shows that they are almost identical. The sequence of PilE expressed by the 427707, 427709, 219900 and 455712 strains is similarly identical to that expressed by the invariable type C group and FAM20. The Limoges isolates therefore all express invariable PilE sequences that are typical of class II strains.

2. PilE from Class II Strains is Expressed in Multiple Proteoforms that Exhibit High Levels of Glycosylation

Full PTM characterisation using bottom-up and top-down mass spectrometry has shown that PilE expressed by the 278534D and 427707C strains is highly glycosylated. Without similar deep characterisation PilE purified from the other Limoges isolates is also proposed to harbour high glycosylation levels. This is based on the large mass difference between the protein mass, as measured by Q-ToF MS, and that predicted from the genome.

Mass profiling experiments indicate that PilE from all isolates is expressed in multiple proteoforms and the identity of the expressed glycan has been proposed by measuring the mass difference between the major proteoform peaks. However, even with the measured mass and proposed glycan identity, the precision provided by Q-ToF mass profiling is not sufficient to decipher the full PTM complement for each of the profiled strains and confirm the glycosylation status. High resolution mass profiling, which furnishes more accurate mass was therefore performed on PilE purified from FAM20 and selected clones from all of the Limoges isolates (Figure 95).

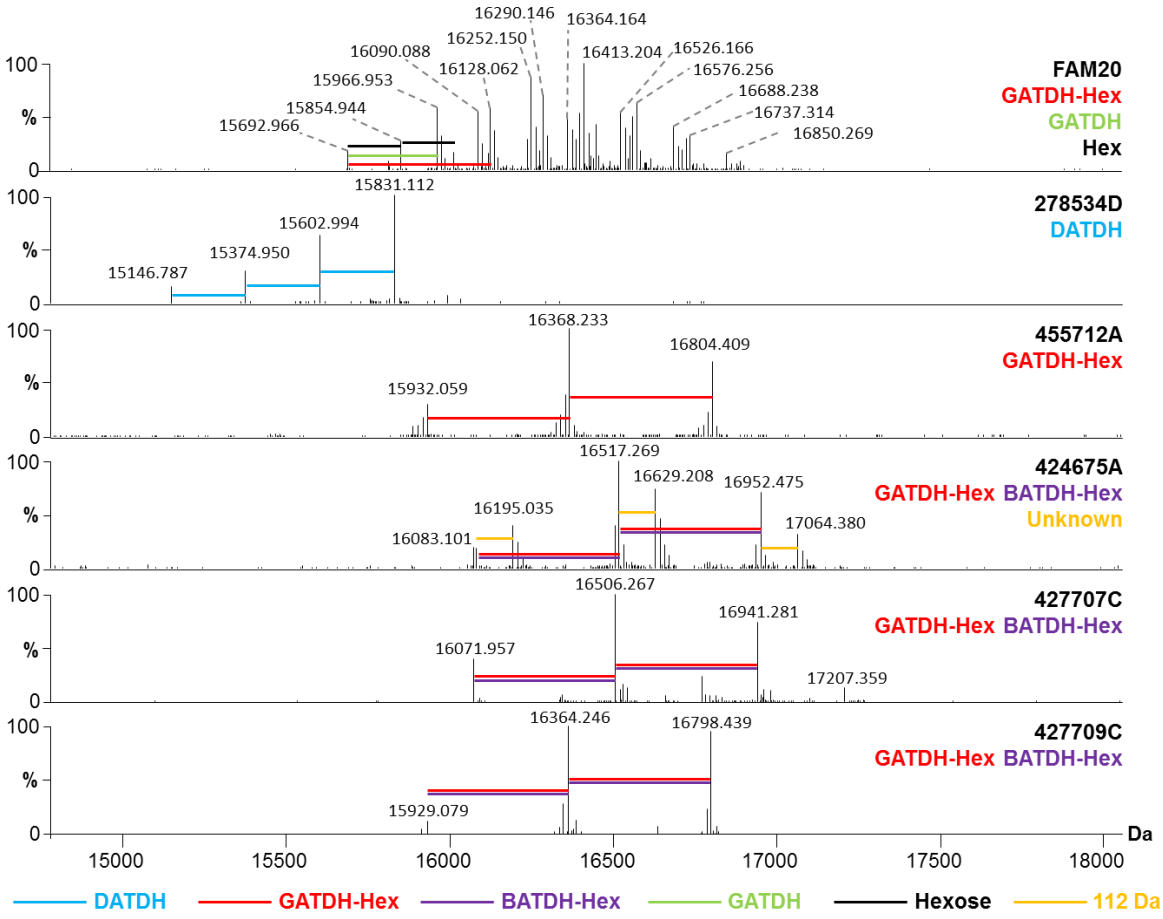


Figure 95 - High resolution Orbitrap mass profiling of FAM20 and selected clones of the Limoges isolates. The mass difference between major proteoforms is reported on the right of the spectrum

For all isolates the high resolution mass profiling gave spectra similar to those obtained previously, however the higher resolution of the instrument enabled the monoisotopic protein mass of each proteoform to be obtained and the difference between the major proteoforms to be calculated with much greater precision. This in turn enables putative modifications to be suggested with much higher confidence.

The proteoforms of 455712 are separated by a mass of 436.17 Da consistent with GATDH-Hex. The origins of the additional peaks at +111.9 Da are currently unknown. The three major peaks of 424675A and 427709C are separated by different two masses of 434.2 and 435.2 Da. This is almost exactly the same as the 434.3 and 435.2 Da observed for PilE-427707C and suggests that PilE from these strains may also harbour both the GATDH-Hex and BATDH-Hex glycans. FAM20 exhibits an extremely complex proteoform pattern with differences between peaks corresponding to that of GATDH, GATDH-Hex and Hexose. Clones of this strain were prepared, PilE purified and

mass profiled to see if any expressed a simplified PilE proteoform population, but all resembled the parent.

To provide further confirmation of glycan expressed by each strain "glycotyping" can be performed by ISD of the entire protein population or HCD on isolated proteoforms (Figure 96).

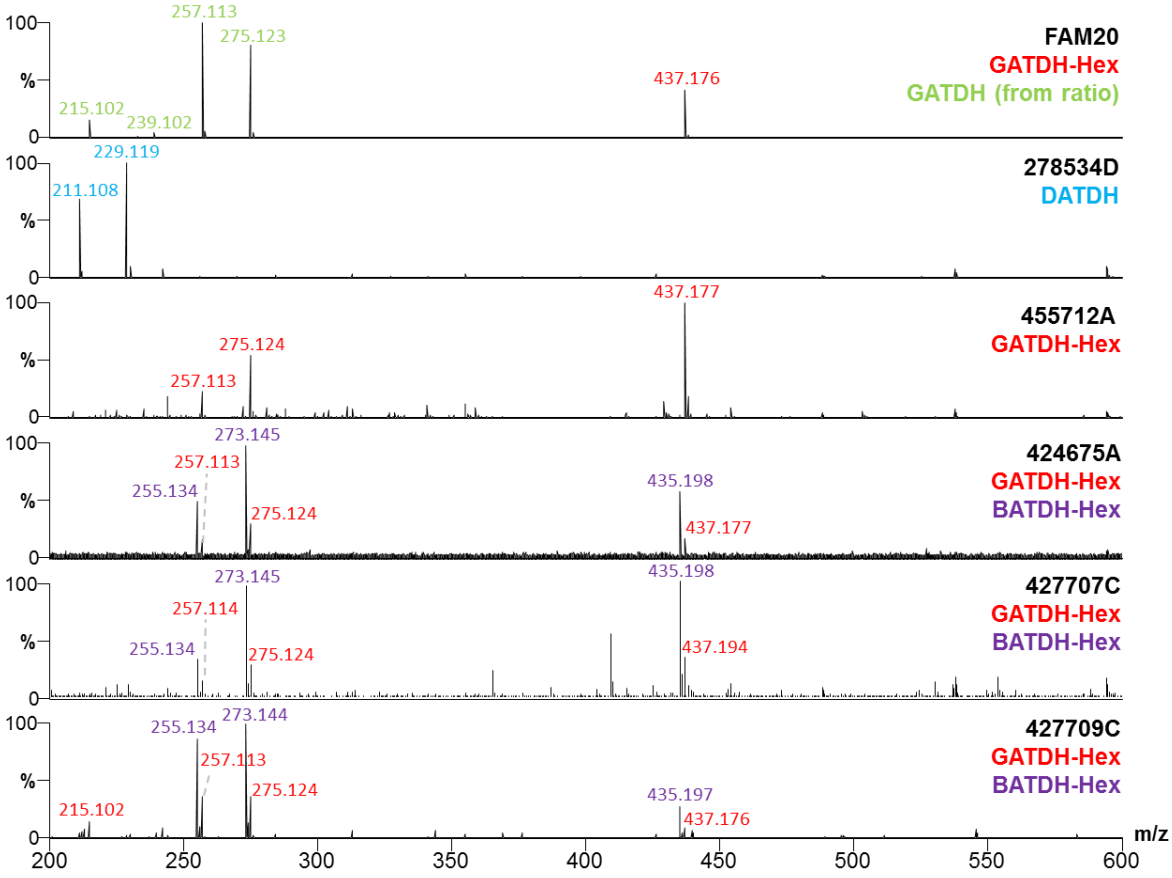


Figure 96 - Glycotyping of PilE purified from clones of the Limoges isolates and FAM20 using ISD or HCD

As expected the oxonium ion fingerprint of DATDH is obtained from 278534D and a fingerprint characteristic of GATDH-Hex is produced from 455712A. A pattern corresponding to co-expression of both the GATDH-Hex and BATDH-Hex glycans is also observed from 427707C and 427709C. This is to be expected as these strains are isolated from the blood and CSF of the same patient. Interestingly 4244675A also expressed the same pattern confirming that this strain also harbours BATDH-Hex. This isolate was obtained separately to 427707C and 427709C and thus indicates that the expression of the novel BATH-Hex glycan appears not to be restricted to a single meningococcal strain.

FAM20 is an especially complex case. Glycotyping clearly shows the presence of the GATDH-Hex glycan, however LC MS/MS experiments on a tryptic digest of PilE-FAM20 have shown that both

GATDH and GATDH-Hex glycans are present on a myriad of backbone sites. Mapping of these sites is a subject of current investigation.

Combining the high resolution mass profiling and glycotyping data enables full PTM complements to be proposed for FAM20 and the remaining isolates that have not been previously completely characterised (Table 8). In some cases there is a difference of one or two Dalton between the modified mass and that observed experimentally. When both the BATDH-Hex and GATDH-Hex glycans are co-expressed, this is to be expected due to the error involved in peak picking that was discussed previously in chapter 5. In addition the presence of PC on the protein backbone may introduce a 1 Da mass error as it is unclear how the software will handle deconvolution of this fixed charge species.

Strain	Predicted M_{mono} from *pilE* (Da)	Measured M_{mono} (Da)	Modified M_{mono} (Da)	Proposed PTMs† (Lowest Mass Proteoform)	Number Proteo-forms
FAM20	14824.621	**15692.966**	**15692.860**	2GATDH, 2PG	>10
278534D	14524.478	**15146.787**	**15146.702**	2DATDH-Hex, PG	4
455712A	14882.627	**15932.059**	**15933.029**	2GATDH-Hex, PC	6
424675A	14524.478	**16083.101**	**16085.065**	3BATDH-Hex, 2PE	3-6
427707C	14882.627	**16071.957**	**16072.992**	GATDH-Hex, BATDH-Hex, 2PG	6
427709A	14882.627	**15929.079**	**15929.070**	2BATDH-Hex, PC	3-6

Table 8 - Comparison of the of the experimentally measured mass of the lowest proteoform of PilE and that predicted from the genome ($M_{i\ genome}$) for FAM20 and several Limoges isolates †Modifications proposed are in addition to a N-terminal processing, methylations and cysteine reduction

When taken together the data in Table 8 indicated that in all strains expressing invariable type PilE sequences, PilE is both heavily glycosylated and expressed in multiple proteoforms each harbouring between two and six glycan subunits.

3. High Levels of Glycosylation are Directed by the Primary Structure of PilE

It has been established that in contrast to class I strains, class II strains express highly glycosylated PilE. The question now arises, "What is special about class II strains that results in this unexpectedly high glycosylation level?" The most obvious candidate for examination is the PglL glycerotransferase that is responsible for transfer of the glycan onto PilE in the periplasm. There does not appear to be any difference in the genetic organisation of the *pglL* locus between class II (FAM18) and class I strains (8013, MC58). Furthermore the primary structure of PglL is virtually identical in both cases and sequence alignment indicated no point mutations in the reported periplasmic exposed loop domains[3].

Little is known about the substrate specificity of this enzyme other than its preference for low complexity regions rich in serine, alanine and proline[4] and certain structural homologies[3]. Since variable and invariable PilE sequences share little homology other than the conserved N-terminus (Figure 94) and showed vastly different glycosylation patterns, it was thought that perhaps the primary sequence of PilE played a role in determining the extent of glycosylation. In order to test this hypothesis an invariable type primary sequence of PilE must be expressed in the genetic background from a class I strain. A mutant of Nm 8013 was therefore created by Corinne Millien in the group of Guillaume Duménil where the endogenous *pilE* gene was deleted and replaced by the *pilE* gene from Nm 427707C. This mutant will be referred to as 8013*pilE*-427707C.

The 8013*pilE*-427707C mutant expressed a normal aggregation phenotype confirming it expressed functional pili and PilE-8013*pilE*-427707C could be purified in good yield as evidenced by SDS-PAGE (results not shown). High resolution mass profiling and ISD glycotyping was therefore performed by Orbitrap MS (Figure 97).

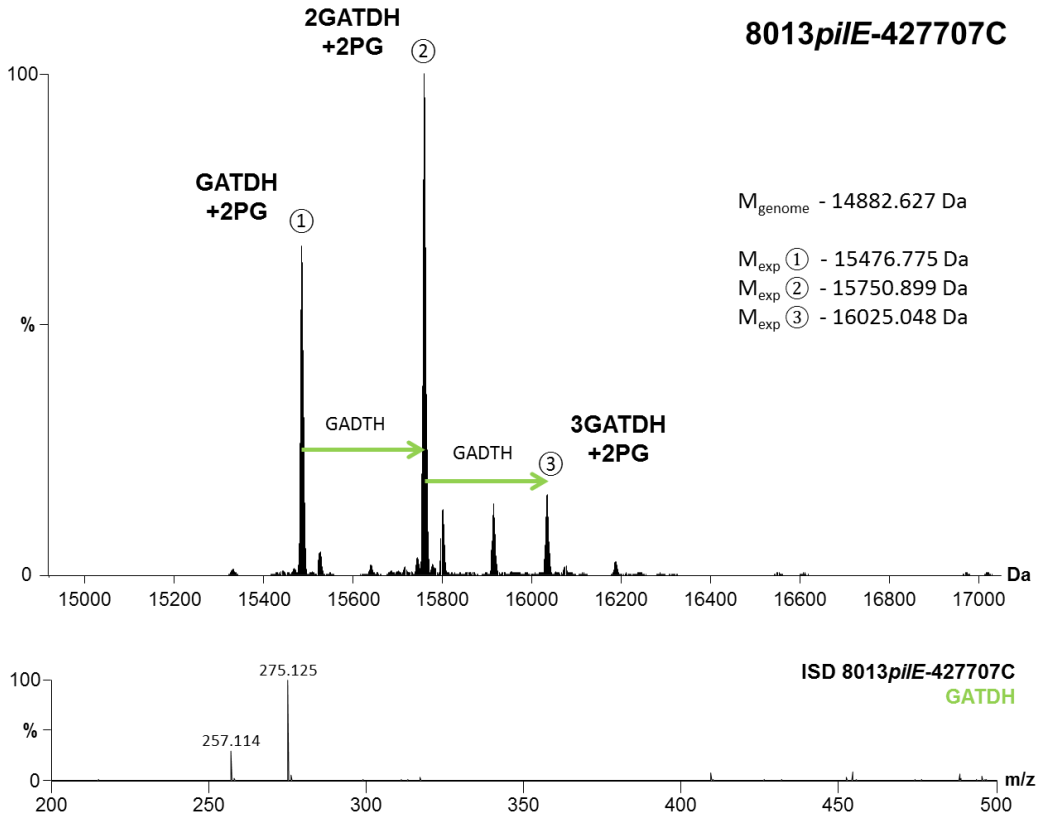


Figure 97 - High resolution Orbitrap mass profile of PilE expressed by the 8013*pilE*-427707C mutant (top). Peaks are labelled with suggested PTMs and monoisotopic masses of the 3 proteoforms are shown on the right of the figure. The result of an ISD glycotyping experiment is also presented (bottom)

Mass profiling of PilE from the 8013*pilE*-427707C mutant displayed three major proteoforms and a remarkably similar profile to wild type Nm 427707C. This is in contrast to wild type Nm 8013 which expresses only two proteoforms, each harbouring a single GATDH unit. Note that the identity of the expressed sugar is now exclusively GATDH as evidenced by glycotyping (Figure 97) not a mixture of GATDH-Hex and BATDH-Hex as in the 427707C wild type. This is expected since Nm 8013 expresses a different set of *pgl* genes and it is this that directs the identity of the glycan.

When the masses of the three proteoforms expressed by 8013*pilE*-427707C were compared to that expected from the genome, it appeared that each was heavily post transitionally modified, bearing two PG groups and one, two and three GATDH subunits respectively. In order to map the PTM sites, top-down characterisation of the three proteoforms was therefore performed by ETD Orbitrap MS/MS. The resulting PTM assignments are given in Figure 98.

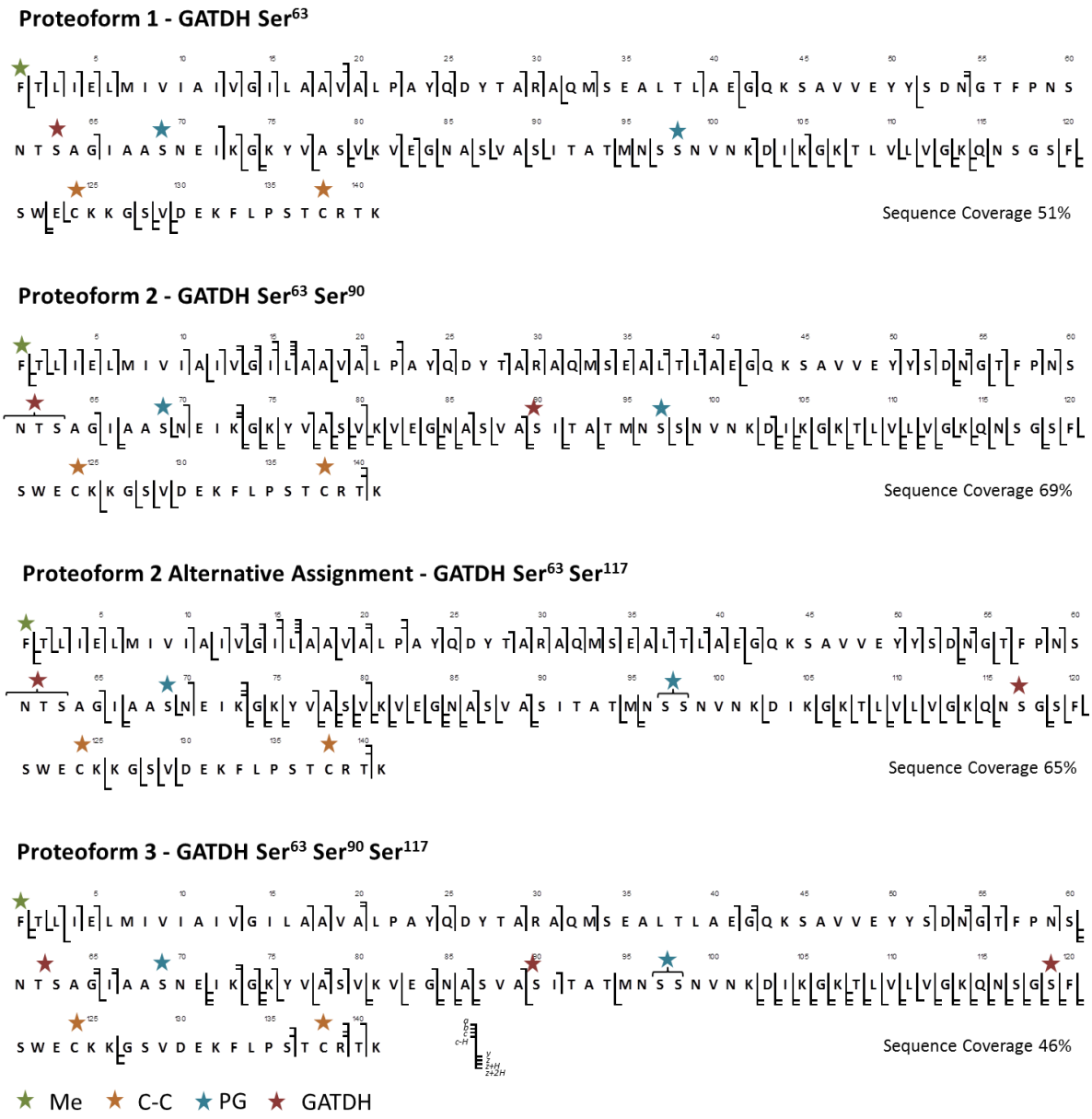


Figure 98 - Top-down Orbitrap ETD MS/MS of three major proteoforms of PilE-8013*pilE*-427707C

ETD fragmentation of the lowest mass proteoform showed that it harboured a GATDH glycan on Ser^{60} or Ser^{63} and PG groups on Ser^{69} and either Ser^{97} or Ser^{98}. A similar experiment performed on the intermediate mass proteoform gave two possible assignments yielding similar overall sequence coverage. In the first, PG groups could be localised to Ser^{69} and Ser^{97} and GADTH subunits were found on Ser^{90} and on either Ser^{60} or Ser^{63}. In the second, PG groups are found on Ser^{69} and on either Ser^{97} or Ser^{98} and the one GATDH subunit is clearly easily located on Ser^{117}. The other GATDH is located on a on either Ser^{63} or Ser^{60}. There is no evidence that both Ser^{90} and Ser^{117} are modified with GATDH in this form at the expense of glycosylation at Ser^{63}. In the highest mass proteoform Ser^{69} and either Ser^{96} or Ser^{97} are modified with PG and Ser^{63} Ser^{90} and Ser^{117} harbour GATDH. These results are consistent with modification of PilE-8013*pilE*-427707C with

PG at Ser^{69} and Ser^{97} and GATDH at Ser^{63} Ser^{90} and Ser^{117}. These are the same modification sites as identified in wild type 427707C.

These results show that when an invariable PilE sequence is expressed in a class I genetic background multiglycosylation of PilE is observed. This points to a situation where the **identity** of the sugar is determined by the *pgl* genes but the **number and location** of glycan modifications is directed by the PilE primary sequence. Invariable type PilE sequences are therefore always modified with a large number of glycan subunits.

4. High Levels of Glycosylation Do Not Impact Pilus Fibre Morphology

Now that multisite glycosylation has been established as a feature of invariable sequence type PilE, it was of interest to examine the effect that such high glycosylation levels had on the morphology of the pilus fibre. Pili prepared from wild type 278534D and a Δ*pglD* mutant that does not express the glycan, were therefore examined by negative staining transmission electron microscopy (TEM). Images were acquired by M. Gérard Pehau-Arnaudet in the Ultrastructural Microscopy Platform at the Institut Pasteur, and data were treated by the author with help from Dr Rémi Fronzes. An overview of pili produced from a typical preparation is given in Figure 99 (top) along with a zoomed image of bundles and individual pilus fibres in both the 278534D wild type and Δ*pglD* mutant (bottom left and right respectively).

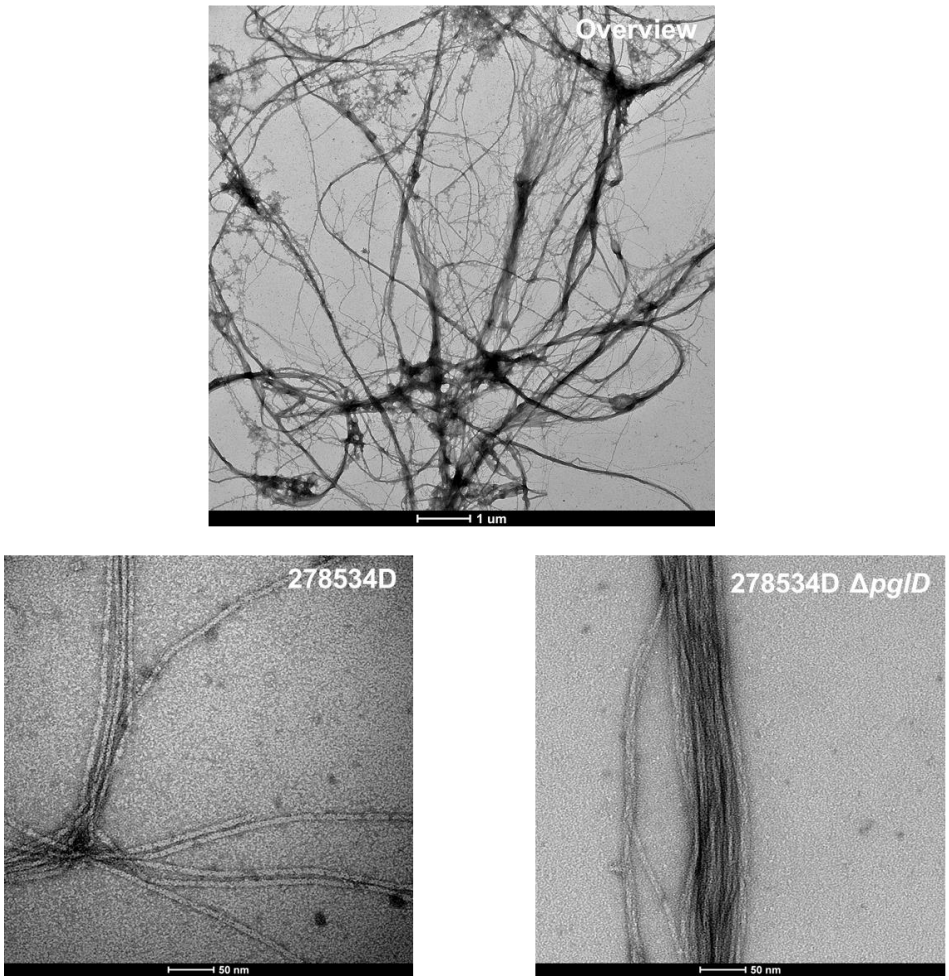


Figure 99 - Negative stain TEM images of a pilus preparation (top) and zoomed on individual fibres from the 278534D WT (bottom left) and Δ*pglD* mutant (bottom right)

The lack of glycan in the Δ*pglD* mutant did not affect either the bundling phenotype or overall morphology of the pilus fibre. Measurement of the pilus width was also attempted in order to ascertain if it affected the size of the pilus diameter. Accurately measuring the pilus width from these single images using an electronic ruler was incredibly difficult, even at high magnifications. A more rigorous approach was therefore pursued similar to that of Hilleringmann *et al*.[5] After calibration of the microscope using images of tobacco mosaic virus, multiple images of pili were taken under the same conditions with identical magnification and defocus. Several hundred slices of pili were selected from these images using the boxer tool in the EMAN2[6].The pilus particles were then processed, aligned against a solid rectangle, summed and clustered into a single particle in the IMAGIC software[7]. This produced a much more defined image whose profile was measured using the image processing tool Image-J. The width profiles obtained from 278534D WT and Δ*pglD* mutant are shown in Figure 100 along with the respective summed particles (Figure 100).

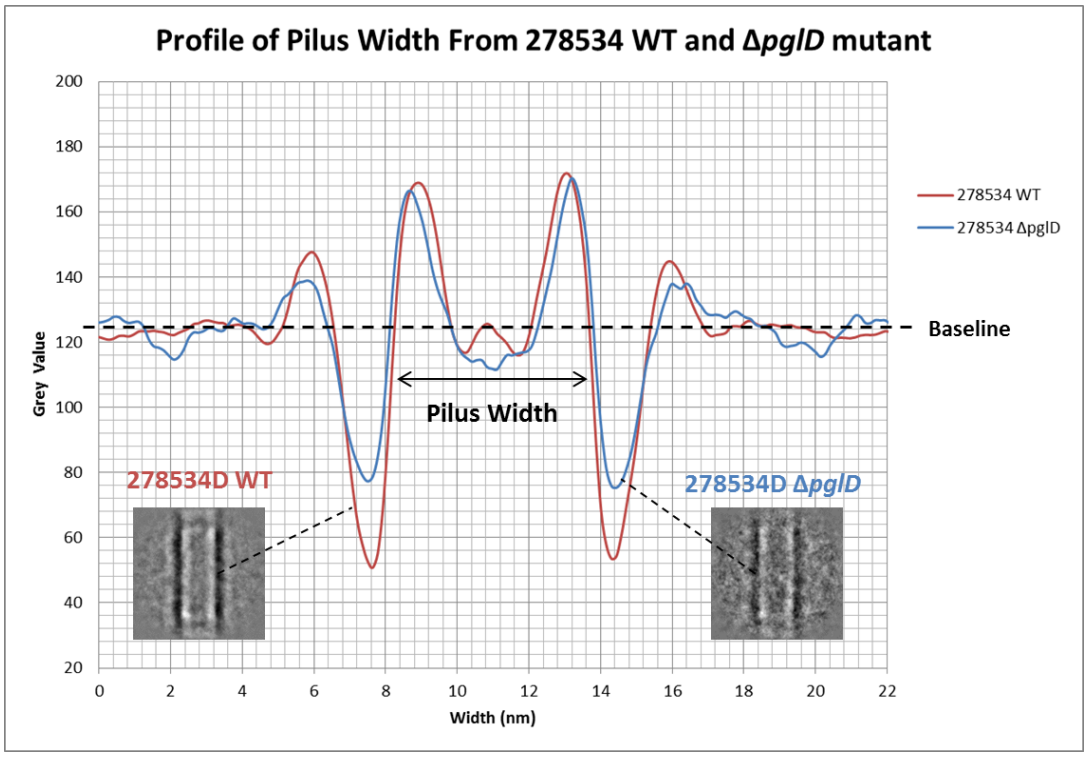


Figure 100 - Pilus width measured from summed and clustered particles of pili from 278534 WT and Δ*pglD* mutant

When the two pilus profiles are aligned against the baseline they are almost identical. This indicated that glycosylation did not affect the pilus width to an appreciable extent and suggests that high levels of glycosylation do not affect pilus morphology; at least under the conditions in which TEM was performed (*in vacuo*, no solvent). A similar experiment was attempted on the 427707C strain, however, PilE could only be purified in minute quantities in the respective Δ*pglD* mutant and experiments on this isolate are on-going.

5. Molecular Modelling Reveals that High Levels of Glycosylation Strongly Affect the Pilus Surface

It has previously been shown that PTM of pili can have a strong effect on the surface of the pilus fibre[8] and the impact of high glycosylation levels on the pilus surface was therefore investigated. Molecular modelling of the pilus fibres from the 427707C and 278534D strains was performed by our collaborators Dr Mathias Ferber and Dr Guillaume Bouvier in the group of Prof. Michael Nilges at the Institut Pasteur. The modelling procedure was adapted from that performed in previous work[8]. In particular the software package Modeller was used for homology modelling the correct

sugar stereochemistries from the recently published data by Hartley *et al*.[9] were imposed by the manual optimisation of force-field parameters.

The PTM site localisation data provided by top-down mass spectrometry was invaluable to the process and enabled the appropriate PTM to be placed in the correct locus on the protein backbone ensuring representative molecular modelling. Models of the pilus fibres from the class I 8013 strain and class II 278514D and 427707C strains are compared against an unmodified fibre in Figure 101.

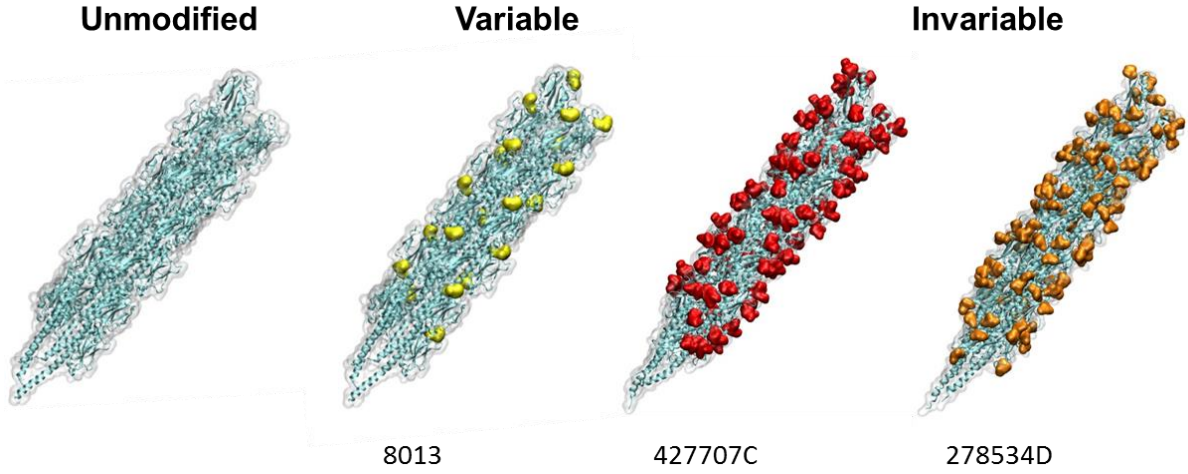


Figure 101 - Molecular modelling of unmodified and wild type pilus fibres from the 8013, 427707C (five GATDH-Hex glycans) and 278534D (five DATDH glycans). Glycan modifications are highlighted by coloured surface

When compared to the unmodified and 8013 fibre, models of the pilus from both Nm 278534D and Nm 427707C strains showed that glycosylation has a profound effect on the pilus surface. All mapped glycosylation sites are found to be surface exposed and when modified with DATDH in Nm 278534D and GATDH-Gal in Nm 427707C sugar moieties are seen to completely coat the pilus fibre. Solvent accessibility calculations show that this results a 20.64% (SD=2.29) reduction in the exposed surface area in 278534D and a 22.7% reduction in 427707C.

6. Conclusions from Investigation into Glycosylation of Type IV Pili in Strains of *N. meningitidis* with Invariable Primary Sequences

High resolution mass profiling experiments have consistently shown that class II strains of Nm harbour invariable type PilE sequences and express PilE in multiple, high abundance proteoforms. These proteoforms have much larger masses than expected from the genome. A combination of top-down and bottom-up mass spectrometry has been used to demonstrate that PilE from these strains is highly glycosylated. This is confirmed by comparative mass profiling of wild type and

glycan deficient strains. The unique capabilities of top-down mass spectrometry have been used to completely characterise the proteoform population of PilE expressed by Nm 427707C and 27834D. Using the PTM localisation data provided by this technique, molecular modelling of pilus fibres has been performed and shows that glycosylation sites are always surfaces exposed. In contrast to pili from class I strains the high glycosylation level of PilE in class II strains result in a fibre that is completely covered by glycan subunits and that reduces solvent surface accessibility by around 20%.

Solvent accessibility is one standard measure, but in this case visualisation of the pilus fibre with respect to another molecule such as an IgG provides a much more informative comparison (Figure 102).

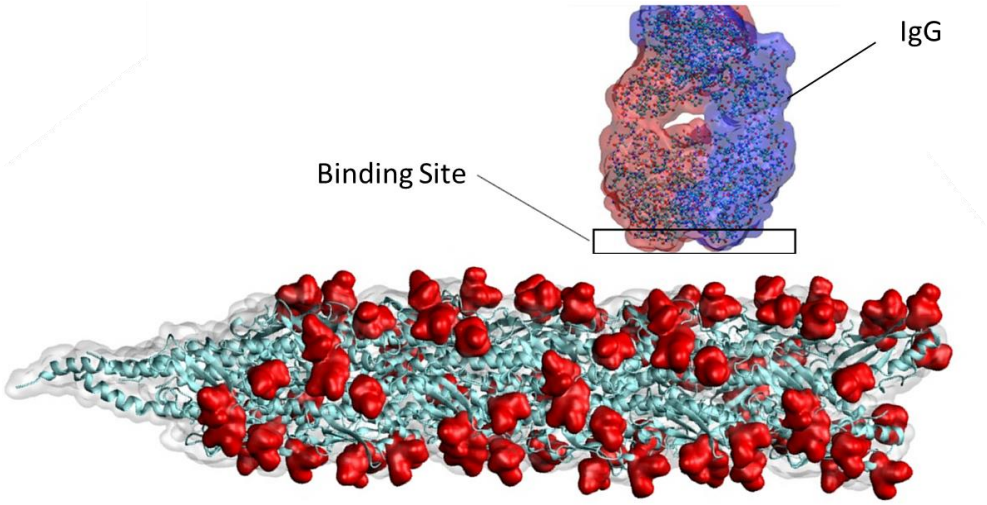


Figure 102 - Same scale model of IgG binding site and pilus fibre from the 427707C strain. The interaction area of the IgG binding site with respect to the glycosylated pilus surface is highlighted

Aligning the IgG alongside the pilus fibre at the same scale shows the clear effect the GATDH-Hex modifications would have on antibody challenge. The pilus fibre is effectively shielded from recognition by a glycan layer.

7. Biological Role of Glycosylation in Strains of *Neisseria meningitidis* Expressing Invariable Pilin Sequences

In keeping with its usually commensal lifestyle, *Neisseria meningitidis* has evolved several mechanisms to promote antigenic variation and escape from the human immune system. This includes on/off expression of surface antigens through phase variation, modulation of the primary structure of surface exposed proteins through gene transfer, and posttranslational modification[10]. PilE is an abundant surface protein that in class I strains exhibits primary

sequence hypervariability. This is the result of homologous recombination of the *pilE* gene with several silent *pilS* cassettes and is thought to be primary mechanism by which PilE evades immune detection.

In stark contrast, class II strains express a high degree of PilE primary sequence conservation, consistent with the absence of homologous recombination or other mechanism of gene conversion[11]. This idea is supported by the fact PilE from FAM18 has shown not to undergo any significant antigenic variation[12] and that the G quartet forming sequence required for pilus antigenic variation is degenerate in FAM18[13]. Class II strains nevertheless represent a significant proportion of pathogenic isolates[11, 14] and all of the Limoges isolates described in this work appear to be members of this class. How PilE expressed by class II strains escapes immune detection is currently unknown.

In this thesis it has been shown that invariable type pilin sequences from class II isolates consistently express multiple proteoforms of PilE all with high levels of glycosylation. It has also been shown that extensive glycosylation is directed specifically by the primary structure of PilE. Characterisation of pilus fibres formed of invariable type pilins by TEM and molecular modelling has demonstrated that glycosylation has little impact on pilus morphology but a significant impact on the surface of the pilus fibre. Indeed glycan modification sites are always surface exposed. It is known that the structure of the glycan can vary due to hard coded differences in the *pglB* gene and phase variation of other *pgl* genes *pglA*, *pglE*, *pglG*, *pglH* and *pglI*. Further diversity has been evidenced in this work by the novel BATDH glycan core.

It is therefore proposed that in class II strains of *Neisseria meningitidis* antigenic variation is promoted though extensive glycosylation and modulation of the pilus surface rather than mutation of the PilE primary sequence. This is illustrated by the cartoon in Figure 103.

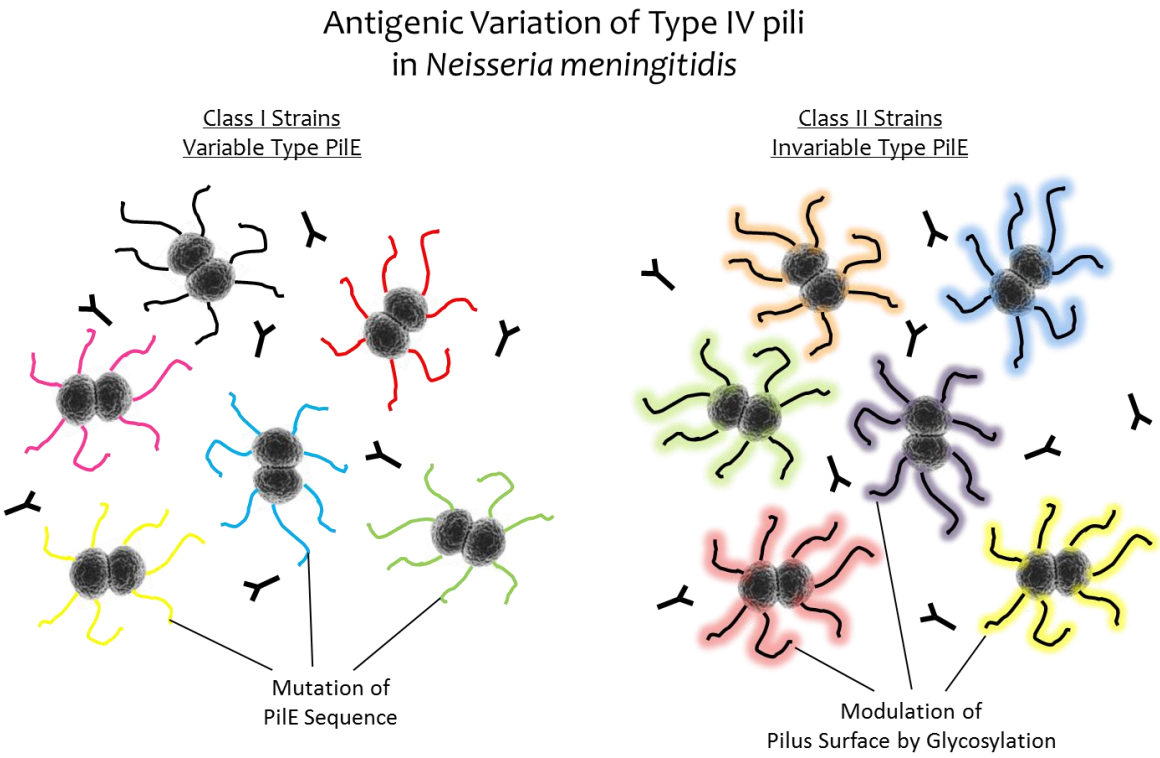


Figure 103 - Proposed mechanism for antigenic variation of PilE from clas II strains of *Neisseria meningitids*

Evidence from top-down mass spectrometry has shown that PilE from class II strains can be both expressed in multiple glycoforms, (different numbers of glycan subunits per pilin monomer) and that each within each glycoform there may be further heterogeneity with different protein backbone sites being modified. Coupled with the previous described modulation of the glycan structure due to the phase variability of the *pgl* genes and the additionally diversity imparted by the phosphoform modification that also decorates the pilus, this represents a dynamic and diverse protective coating for the pilus fibre in class II strains that is believed to protect the PilE primary sequence from opsonisation.

To this author's knowledge this represents the first proposed explanation for the immune evasion of pilin in class II strains of *Neisseria meningitidis*.

Bibliography

[1] J. F. Wang, D. A. Caugant, G. Morelli, B. Koumare and M. Achtman.Antigenic and epidemiologic properties of the ET-37 complex of Neisseria meningitidis. *Journal of Infectious Diseases*, **1993**, *167*, 1320.

[2] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J. F. Dufayard, S. Guindon, V. Lefort, M. Lescot, J. M. Claverie and O. Gascuel.Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research*, **2008**, *36*, W465.

[3] B. L. Schulz, F. E. C. Jen, P. M. Power, C. E. Jones, K. L. Fox, S. C. Ku, J. T. Blanchfield and M. P. Jennings.Identification of Bacterial Protein O-Oligosaccharyltransferases and Their Glycoprotein Substrates. *PLoS One*, **2013**, *8*.

[4] J. H. Anonsen, A. Vik, W. Egge-Jacobsen and M. Koomey.An Extended Spectrum of Target Proteins and Modification Sites in the General O-Linked Protein Glycosylation System in Neisseria gonorrhoeae. *J. Proteome Res.*, **2012**, *11*, 5781.

[5] M. Hilleringmann, F. Giusti, B. C. Baudner, V. Masignani, A. Covacci, R. Rappuoli, M. A. Barocchi and I. Ferlenghi.Pneumococcal pili are composed of protofilaments exposing adhesive clusters of Rrg A. *PLoS pathogens*, **2008**, *4*.

[6] G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees and S. J. Ludtke.EMAN2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology*, **2007**, *157*, 38.

[7] M. vanHeel, G. Harauz, E. V. Orlova, R. Schmidt and M. Schatz.A new generation of the IMAGIC image processing system. *Journal of Structural Biology*, **1996**, *116*, 17.

[8] J. Chamot-Rooke, G. Mikaty, C. Malosse, M. Soyer, A. Dumont, J. Gault, A. F. Imhaus, P. Martin, M. Trellet, G. Clary, P. Chafey, L. Camoin, M. Nilges, X. Nassif and G. Dumenil.Posttranslational Modification of Pili upon Cell Contact Triggers N. meningitidis Dissemination. *Science*, **2011**, *331*, 778.

[9] M. D. Hartley, M. J. Morrison, F. E. Aas, B. Borud, M. Koomey and B. Imperiali.Biochemical Characterization of the O-Linked Glycosylation Pathway in Neisseria gonorrhoeae Responsible for Biosynthesis of Protein Glycans Containing N,N '-Diacetylbacillosamine. *Biochemistry*, **2011**, *50*, 4936.

[10] T. Davidsen and T. Tonjum.Meningococcal genome dynamics. *Nature Reviews Microbiology*, **2006**, *4*, 11.

[11] A. Cehovin, M. Winterbotham, J. Lucidarme, R. Borrow, C. M. Tang, R. M. Exley and V. Pelicic.Sequence conservation of pilus subunits in Neisseria meningitidis. *Vaccine*, **2010**, *28*, 4817.

[12] R. A. Helm and H. S. Seifert.Frequency and Rate of Pilin Antigenic Variation of Neisseria meningitidis. *J Bacteriol*, **2010**, *192*, 3822.

[13] L. A. Cahoon and H. S. Seifert.An Alternative DNA Structure Is Necessary for Pilin Antigenic Variation in Neisseria gonorrhoeae. *Science*, **2009**, *325*, 764.

[14] X. Sun, H. Zhou, L. Xu, H. Yang, Y. Gao, B. Zhu and Z. Shao.Prevalence and genetic diversity of two adhesion-related genes, pilE and nadA, in Neisseria meningitidis in China. *Epidemiology and Infection*, **2013**, *141*, 2163.

Chapter 7

Top-down MS Characterisation of Pilins Expressed by Other Pathogens

Preceding chapters have presented extensive refinement of the top-down experiment performed both FT-ICR and Orbitrap mass spectrometers for the analysis of the major pilin PilE expressed by *Neisseria meningitidis*. During the course of this thesis this methodology has also been applied to pilins expressed by other pathogens. This has included characterisation of the PulG and PpdD pilins from *Escherichia coli* in collaboration with Dr Olivera Francetic at the Institut Pasteur, Paris and has recently been used to probe the composition of the *Streptococcus pneumoniae* transformation pilus in collaboration with the group of Dr Rémi Fronzes, also at the Institut Pasteur.

1. Identification and Characterisation of a Type IV Pilus in the Gram Positive Bacterium *Streptococcus pneumoniae*

First isolated by Louis Pasteur in 1881 and even named "*Micrococcus pasteuri*" by his contemporary Sternberg, *S. pneumoniae* is a Gram positive bacterium and transient commensal of the human upper respiratory tract, which sporadically develops virulent strains through mutation of its capsid or surface proteins. *S. pneumoniae* is best known today as the most common bacterial cause of acute respiratory infection and *otitis media*. However at the beginning of the 20th century this bacterium killed more people worldwide than any other pathogen and still remains a major cause of child death in less developed countries.

S. pneumoniae is known to become briefly, naturally competent at the start of pre-exponential growth. Competence is stimulated by a peptide pheromone called competence-stimulating peptide (CSP). Several genes are known to be up-regulated upon induction of competence. This includes the *comG* operon which codes for a set of proteins comprising ComGA, a putative ATPase, ComGB and ComGD putative membrane proteins, and the pilin proteins ComGC, ComGE, ComGF and ComGG. The genes coding for these proteins share significant homology with those required for type IV pili expression in Gram negative bacteria. However whilst several types of pili are known to be expressed by *S. pneumoniae*, no type IV pilus has ever previously been visualised or characterised.

By working on a particular strain of *S. pneumoniae,* developed by Dr Jean-Pierre Claverys (Université Paul Sabatier, Toulouse) on which competence can be easily induced, the group of Rémi Fronzes has been able to confirm the existence of a competence pilus and visualise it by transmission electron microscopy (TEM). Questions then arose concerning the molecular composition of this pilus and more particularly the identity of the major pilin. Through analogy to type IV pili, ComGC seemed a likely candidate and a FLAG tagged version of the *comGC* gene was inserted ectopically into a different competence up-regulated locus in order to probe the level of ComGC in the pilus fibre.

After induction of competence, anti-FLAG immunogold labelling and TEM experiments were performed. As expected the gold nanoparticles localised to pilus but only a few particles were observed over the whole length of the fibre, rather than the 50% coverage expected. This puzzling observation was continually the case after several repeats and bought into doubt the hypothesis that the pilin was primarily composed of GomGC subunits.

A sample of purified pili from this strain was therefore subject to characterisation by top-down mass spectrometry using an Orbitrap Velos, in order to identify and the major constituent of the pilus. The results of this work are presented as part of the following publication.

2. Published Article – "A Type IV Pilus Mediates DNA Binding during Natural Transformation in *Streptococcus pneumoniae*"

OPEN ACCESS Freely available online

PLOS | PATHOGENS

A Type IV Pilus Mediates DNA Binding during Natural Transformation in *Streptococcus pneumoniae*

Raphaël Laurenceau[1,2], Gérard Péhau-Arnaudet[2], Sonia Baconnais[3], Joseph Gault[2,4], Christian Malosse[2,4], Annick Dujeancourt[1,2], Nathalie Campo[5,6], Julia Chamot-Rooke[2,4], Eric Le Cam[3], Jean-Pierre Claverys[5,6], Rémi Fronzes[1,2]*

1 Institut Pasteur, Groupe Biologie Structurale de la Sécrétion Bactérienne, Paris, France, **2** CNRS, UMR3528, Paris, France, **3** Maintenance des Génomes et Microscopies Moléculaire, UMR 8126 CNRS-Université Paris Sud, Institut Gustave Roussy, Villejuif, France, **4** Institut Pasteur, Unité de Spectrométrie de Masse Structurale et Protéomique, Paris, France, **5** CNRS, UMR5100, Toulouse, France, **6** Université de Toulouse, UPS, Laboratoire de Microbiologie et Génétique Moléculaires, Toulouse, France

Abstract

Natural genetic transformation is widely distributed in bacteria and generally occurs during a genetically programmed differentiated state called competence. This process promotes genome plasticity and adaptability in Gram-negative and Gram-positive bacteria. Transformation requires the binding and internalization of exogenous DNA, the mechanisms of which are unclear. Here, we report the discovery of a transformation pilus at the surface of competent *Streptococcus pneumoniae* cells. This Type IV-like pilus, which is primarily composed of the ComGC pilin, is required for transformation. We provide evidence that it directly binds DNA and propose that the transformation pilus is the primary DNA receptor on the bacterial cell during transformation in *S. pneumoniae*. Being a central component of the transformation apparatus, the transformation pilus enables *S. pneumoniae*, a major Gram-positive human pathogen, to acquire resistance to antibiotics and to escape vaccines through the binding and incorporation of new genetic material.

Citation: Laurenceau R, Péhau-Arnaudet G, Baconnais S, Gault J, Malosse C, et al. (2013) A Type IV Pilus Mediates DNA Binding during Natural Transformation in *Streptococcus pneumoniae*. PLoS Pathog 9(6): e1003473. doi:10.1371/journal.ppat.1003473

Editor: Carlos Javier Orihuela, The University of Texas Health Science Center at San Antonio, United States of America

Received April 23, 2013; **Accepted** May 17, 2013; **Published** June 27, 2013

Funding: The Agence Nationale pour la Recherche, Institut Pasteur and the Centre National de la Recherche Scientifique have supported this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: remi.fronzes@pasteur.fr

Introduction

Natural transformation, first discovered in *Streptococcus pneumoniae* [1], is observed in many Gram-negative and Gram-positive bacteria [2]. It increases bacterial adaptability by promoting genome plasticity through intra- and inter-species genetic exchange [3]. In *S. pneumoniae*, a major human pathogen responsible for severe diseases such as pneumonia, meningitis and septicemia, transformation is presumably responsible for capsular serotype switching and could therefore reduce the efficiency of capsule-based vaccines after a short period [4]. In this species, it occurs during a genetically programmed and differentiated state called competence that is briefly induced at the beginning of exponential growth. During this competent state, pneumococci secrete a peptide pheromone called Competence-Stimulating-Peptide (CSP) [5], which spreads competence in the pneumococcal population. Interestingly, in *S. pneumoniae*, some antibiotics and DNA-damaging agents induce competence, which would act as an alternative SOS response and ultimately increases bacterial resistance to external stresses [6].

During transformation, environmental DNA is bound at the surface of competent cells and transported through the cell envelope to the cytosolic compartment. This process has been mostly studied in the Gram-positive bacterium *Bacillus subtilis* with additional information coming from studies in *S. pneumoniae* [7,8]. In both species, a DNA translocation apparatus mediates the transfer of DNA through the cellular membrane. In *S. pneumoniae*, it is composed of ComEA, EndA, ComEC and ComFA. Incoming double-stranded DNA would bind the membrane receptor ComEA. One DNA strand crosses the membrane through ComEC while the endonuclease EndA degrades the other strand. On the cytoplasmic side, ComFA, an ATPase that contains a helicase-like domain, would facilitate DNA internalization through ComEC. Once inside the bacterium, single-stranded DNA is either integrated into the chromosome by RecA-mediated homologous recombination or entirely degraded.

Strikingly, all transformable Gram-positive bacteria also carry a *comG* operon that resembles operons encoding Type IV pili and Type II secretion pseudopili in Gram-negative bacteria, as well as a gene encoding a prepilin peptidase homolog, *pilD* [7]. In *B. subtilis* and *S. pneumoniae*, *comG* and *pilD* genes are exclusively expressed in competent cells and are essential for transformation [9,10,11]. In *S. pneumoniae*, the *comG* operon encodes a putative ATPase (ComGA), a polytopic membrane protein (ComGB) and five prepilin candidates named ComGC, ComGD, ComGE, ComGF and ComGG (Figure 1A and B and table S1). By homology with Type IV pili, it is generally proposed that these proteins could be involved in the assembly of a transformation pseudo-pilus at the surface of competent cells [7,8,12]. So far, two studies show that a large macromolecular complex containing ComGC can be found at the surface of competent *B. subtilis* cells [9,12]. In this complex, ComGC subunits appear to be linked together by disulfide bridges [9]. All the other ComG proteins and the PilD homolog, ComC, are necessary for the formation of this

Author Summary

Natural genetic transformation, first discovered in *Streptococcus pneumoniae* by Griffith in 1928, is observed in many Gram-negative and Gram-positive bacteria. This process promotes genome plasticity and adaptability. In particular, it enables many human pathogens such as *Streptococcus pneumoniae, Staphylococcus aureus* or *Neisseria gonorrhoeae* to acquire resistance to antibiotics and/or to escape vaccines through the binding and incorporation of new genetic material. While it is well established that this process requires the binding and internalization of external DNA, the molecular details of these steps are unknown. In this study, we discovered a new appendage at the surface of *S. pneumoniae* cells. We show that this appendage is similar in morphology and composition to appendages called Type IV pili commonly found in Gram-negative bacteria. We demonstrate that this new pneumococcal pilus is essential for transformation and that it directly binds DNA. We propose that the transformation pilus is an essential piece of the transformation apparatus by capturing exogenous DNA at the bacterial cell surface.

complex [9]. It was proposed that this complex could correspond to the transformation pseudo-pilus.

Despite these first clues, no transformation appendage could be directly visualized at the surface of any competent Gram-positive bacterium. In addition, the function of the ComG proteins during transformation remains unclear. Mutations in the cytosolic ComGA protein abolish DNA binding at the surface of both *B. subtilis* and *S. pneumoniae* [13,14,15]. This strongly suggests that the ComGC-containing macromolecular complex detected at the surface of competent *B. subtilis* cells could bind DNA. However, it was recently shown that ComGA is the only ComG protein essential to the initial DNA binding at the surface of competent *B. subtilis* cells [14]. This protein would interact with an unknown DNA receptor at the surface of competent cells while the other ComG proteins would only be required at a later stage during transformation.

In this study, we provide the first direct evidence for the existence of a transformation pilus in a Gram-positive bacterium. We discovered a new appendage at the surface of competent pneumococci that we could visualize using immuno-fluorescence and electron microscopy. Competent cells harbor one or a few appendages that are morphologically similar to Type IV pili found in Gram-negative bacteria. We were able to purify this pilus and showed that it is essentially composed of the ComGC pilin. We also demonstrate that pilus assembly is required for transformation. As we provide direct evidence that the transformation pilus binds extracellular DNA, we propose it is the primary DNA receptor at the surface of competent pneumococci.

Results

An extracellular appendage containing ComGC is detected at the surface of competent *S. pneumoniae*

Mechanical shearing is frequently used to release bacterial surface appendages and to study their protein composition [16]. To see if ComGC was part of a macromolecular complex at the surface of competent *S. pneumoniae*, we adapted the method to this bacterium and raised antibodies against the purified soluble domain of ComGC. Using this antibody, we showed that ComGC could be detected by immunoblotting in the sheared fraction of competent bacteria (Figure 2A). While ComGC level in the cell fraction was not affected, no ComGC could be found in the sheared fraction in a *comGA* knockout mutant (Figure 2A). These data strongly suggest that ComGC is part of an extra-cellular appendage and that ComGA is necessary to its assembly.

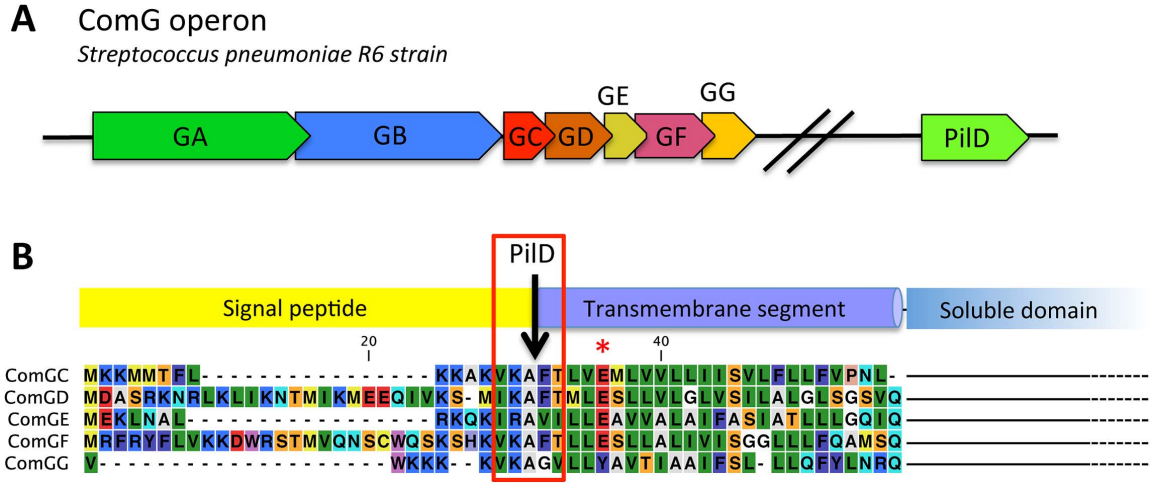


Figure 1. Genes potentially essential for transformation pilus assembly in *S. pneumonia and* prepilin sequences. (**A**) Genes that are potentially essential in transformation pilus assembly. The *comG* operon contains 7 genes named *comGA* to *comGG*. A gene that encodes a pre-pilin peptidase homolog, *pilD*, is found elsewhere on the chromosome. The name used to designate the *comG* and *pilD* genes varies in different pneumoccocal strains. For clarity, we refer to the comG nomenclature used in B. subtilis. Names of the corresponding genes in different S. pneumoniae strains are found in the Table S1. (**B**) Alignment of the N-terminal extremity of the five possible prepilins show that ComGC, ComGD, ComGE and comGF contain a canonical prepilin cleavage motif (red rectangle) [20]. All four proteins also contain a conserved glutamic acid residue in position 5 after the cleavage site (red asterisk), a conserved feature of all Type IV pilins [43]. ComGG has a degenerated peptidase motif and does not contain a conserved glutamic acid in position 5. Secondary structure predictions for the five proteins propose a N-terminal hydrophobic helix and a C-terminal soluble domain of variable size (not shown in the figure), the typical topology found in Type IV pilins.
doi:10.1371/journal.ppat.1003473.g001

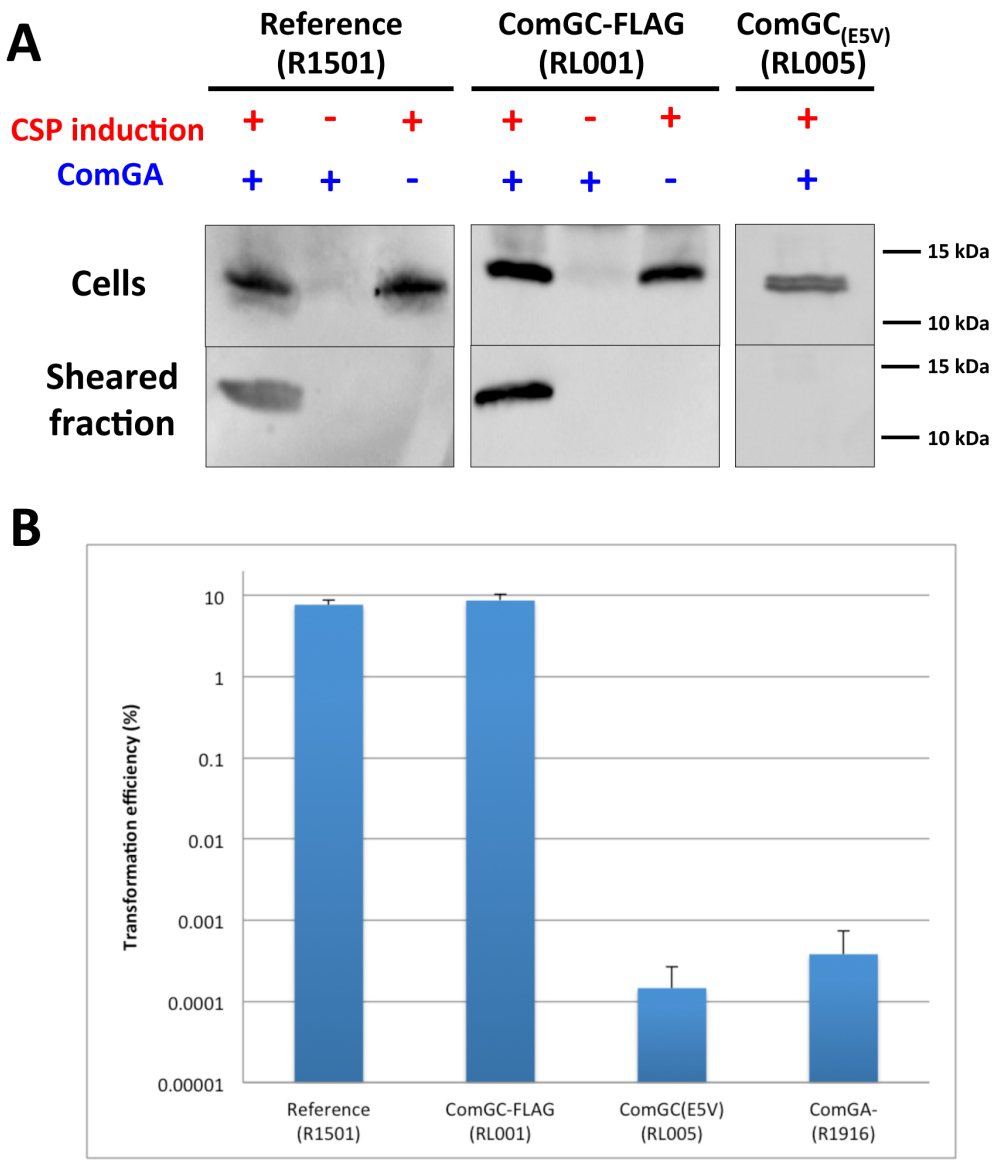


Figure 2. Competence-induced appendages assembly and transformation efficiency. (**A**) Detection of ComGC in the sheared and cellular fractions by immunoblot. ComGC was detected in the sheared fraction of reference (R1501) and ComGC-FLAG (RL001) competent strains. ComGC was not detected in the sheared fraction of cultures that were not competence-induced. Deletion of *comGA* completely abolished detection of the ComGC pilin in the sheared fraction in both reference and ComGC-FLAG strains. Substitution of the conserved glutamic acid in position 5 of the mature ComGC pilin (Figure 1) by alanine ($ComGC_{(E5V)}$) also completely abolished detection of ComGC in the sheared fraction. For the reference and $ComGC_{(E5V)}$ strains, detection of ComGC was performed with a polyclonal rabbit antibody raised against the soluble domain of ComGC. With the ComGC-FLAG strain, the FLAG-tagged pilin was detected by an anti-FLAG monoclonal antibody. (**B**) Transformation assay using the reference strain (R1501), comGC-FLAG strain (RL001), $comGC_{(E5V)}$ mutant (RL005) and *comGA* deletion mutant (R1916). Both $comGC_{(E5V)}$ and *comGA* mutants were defective for transformation.
doi:10.1371/journal.ppat.1003473.g002

A ComGC-containing appendage is visualized at the surface of competent pneumococci

We inserted a FLAG tag at the C-terminus of ComGC to directly visualize the competence-induced appendages by immuno-fluorescence. It was not possible to insert the sequence encoding the tag at the *comGC* locus on the chromosome because *comGC* and *comGD* genes overlap in the *comG* operon. Therefore, a copy of *comGC* encoding a C-terminally FLAG-tagged ComGC (ComGC-FLAG) was integrated ectopically into the chromosome of *S. pneumoniae* under the control of a competence-induced promoter [17]. The transformation efficiency was not affected in this strain (Figure 2B). Using anti-FLAG antibodies, we could show by immuno-fluorescence that almost all the cells appeared to harbour one or a few ComGC foci or distinct fluorescent appendages (Figure 3A and B; Figure S1A). Due to sample

preparation, many broken appendages were also found in the background. No preferential location of the foci/appendages at the cell surface was observed. They are absent in *comGA* knockout cells (Figure 3A). Note that anti-ComGC antibodies were not able to label the competent cells. They probably recognize epitopes that are masked when ComGC is included in the appendages.

Competence-induced appendage morphology and composition

Using electron microscopy, we observed filaments attached to the cell surface of negatively stained competent pneumococci (Figure 4A and B). These flexible filaments are 5–6 nm in diameter. Their length could reach up to 2–3 micrometers (Figure 4A). A maximum of 2–3 filaments per cell could be observed. Their average length was difficult to assess because they break easily into smaller fragments during sample preparation. Using the ComGC-FLAG expressing strain, we confirmed by immunogold-labelling that they contain ComGC (Figure 4C).

Appendages were then purified using anti-FLAG affinity chromatography after mechanical shearing. Appendage fragments of between 50 and 500 nm in length were observed by electron microscopy (Figure 5A), showing that these filamentous structures do not disassemble during purification. SDS-PAGE analysis of the purified fraction showed that ComGC is the major component of the appendages (Figure 5B). Using whole protein mass profiling by high-resolution mass spectrometry [18], we could only detect ComGC and ComGC-FLAG in the purified material (Figure 5C), confirming that ComGC is the major constituent of these appendages. Monoisotopic mass measurements of intact proteins and top-down fragmentation using a variety of activation techniques confirmed that the ComGC prepilin is cleaved after the alanine residue in position 15 and that the first amino acid of the mature protein is methylated, presumably by PilD (Figure S2). Indeed, PilD homologs in Gram-negative bacteria catalyze this post-translational modification of the Type IV pilins [19]. No other post-translational modification was detected in ComGC. Other proteins, including other ComG proteins, were not detected in the purified material by the methods used in this study. This suggests that these proteins are either absent, present in very low amount within the appendage or weakly bound to it and lost during sample preparation. These morphological and biochemical features are typical of Gram-negative Type IV pili. Therefore, we propose that the competence-induced appendage observed in *S. pneumoniae* belongs to the Type IV pilus family.

The competence-induced pili are required for transformation

It was important to determine whether these competence-induced pili were involved in transformation. Indeed, it was previously shown that *S. pneumoniae* and *B. subtilis comGA* knockout could not be transformed (Figure 2B) [9] [13]. In this study, we were able to show in *S. pneumoniae* that *comGA* mutant cells lack pili (Figure 2A and 3A). It was enticing to conclude that competence-induced pili assembly is essential for transformation. However, it was recently shown that a *comGA* mutation could have a pleiotropic effect on transformation in *B. subtilis* [14]. Therefore, we generated a *comGC* mutant in *S. pneumoniae* in which the conserved glutamic acid in position 5 was substituted by an alanine (Figure 1B). Such a substitution was shown to impair Type IV pilus assembly in Gram-negative bacteria [20]. ComGC cellular level was not affected by this point mutation (Figure 2A). Our results show that this mutant strain could not assemble any pilus and that it was defective for transformation (Figure 2A and B). Therefore we conclude from the analysis of both *comGA* and $comGC_{(E5A)}$ mutants that the assembly of the competence-induced pilus is required for transformation.

Transformation pili bind extracellular DNA

The nature of the primary DNA receptor at the surface of transformable Gram-positive bacteria is not known. It is generally proposed that the transformation pseudopilus would bind extracellular DNA at the surface of competent Gram-positive bacteria [8,21]. However, this hypothesis has never been confirmed experimentally. Using affinity purification, we show that DNA naturally released in the culture medium co-fractionates with the purified pili. No DNA could be found in the purified fraction in absence of the pilus (Figure 6A). These data were a first hint suggesting that DNA present in the environment could bind to the transformation pilus. However, it was not clear if this binding was related to the transformation process or fortuitous. By using specific electron microscopy methods [22], we visualized DNA directly bound to the transformation appendage after adding linear double stranded DNA (dsDNA) to competent bacteria. Long stretches of dsDNA interacting with the transformation pilus were observed with clearly visible multiple contact points (Figure 6 B–E). Interestingly, it was extremely difficult to see DNA bound on the pilus in the reference bacteria (R1501 strain), which are known to internalize exogenous DNA quickly [23]. On the other hand, in *ComEC* and *comFA* mutants, we could easily observe bound DNA on transformation pili. These strains are defective for DNA uptake and accumulate bound DNA at their surface [13]. Given that the dsDNA was added in large excess, no difference between the reference and mutant strains should be observed if DNA binding on the pili was a coincidental event. The fact that the uncoupling of DNA binding and uptake processes facilitates the observation of the DNA/pilus interaction is a strong indication that DNA binding on the transformation pilus is related to the transformation process.

Discussion

The *comG* operon is conserved in all transformable Gram-positive bacteria. This operon encodes proteins that are homologous to proteins involved in Type IV pilus assembly in Gram-negative bacteria. Therefore it has been proposed that a pilus (or pseudopilus) could be assembled at the surface of competent Gram-positive bacteria. Since all *comG* genes are essential for transformation, this pilus could be directly involved in transformation. The first biochemical clues for the existence of a transformation pilus were found in *B. subtilis* although decisive observational support was lacking. In addition, it was not clear if the ComGC-containing macromolecular complex found in *B. subtilis* was a common feature of competent Gram-positive bacteria or specific to this species. Finally, the function of this putative transformation pilus, and in general of the ComG proteins, was unclear.

Discovery of a new pneumococcal appendage

The pneumococcal transformation pilus represents a newly discovered pneumococcal surface structure. For a long time, no external appendage could be found at the surface of *S. pneumoniae* cells while many electron microscopy images were published in the literature. Recently, sortase-mediated pili have been discovered in some pathogenic *S. pneumoniae* strains [24]. To our knowledge, no specific ultrastructural study of competent *S. pneumoniae* has ever been described. Here, we analysed a laboratory strain that is

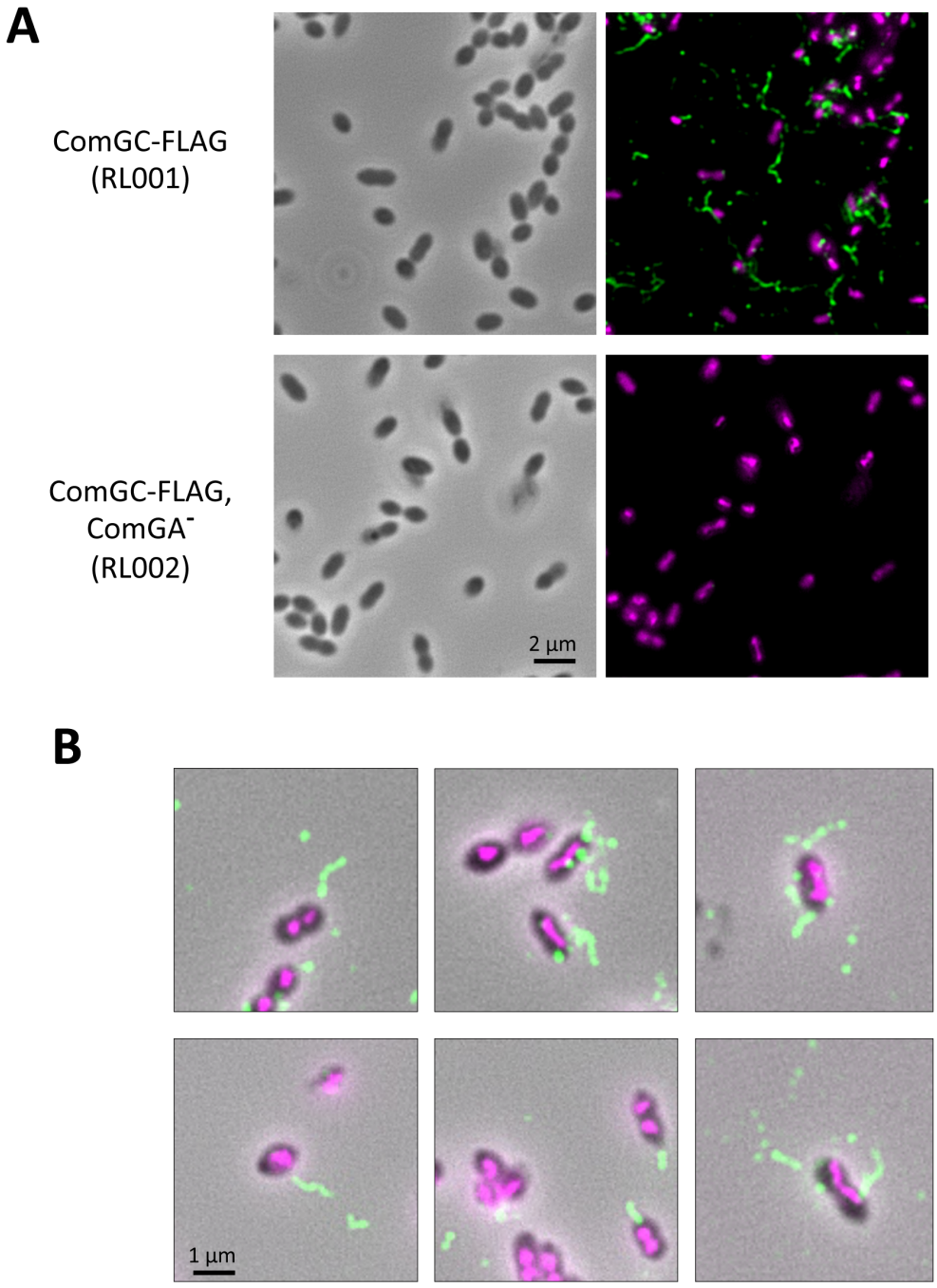


Figure 3. Direct visualization of competence-induced appendages by Immuno-fluorescence. (**A**) Immuno-fluorescence microscopy showing intact competent cells expressing a FLAG-tagged ComGC pilin in presence (top row) or absence of ComGA (bottom row). Left column correspond to bright field image, right column to overlay between anti-FLAG antibody fluorescence (green) and DAPI fluorescence (magenta). (**B**) Zoom on several bacterial cells visualized by immuno-fluorescence. Overlay of bright filed image, anti-FLAG antibody fluorescence (green) and DAPI fluorescence (magenta). Distinct appendages are visible on competent cells.
doi:10.1371/journal.ppat.1003473.g003

commonly used to study the transformation process in *S. pneumoniae* [10] [13]. In this strain, competence can be induced in a rapid and synchronous manner upon addition of synthetic CSP in the medium of an exponentially growing culture [5,25]. To make sure that the appearance of the transformation pilus is a common feature of competent pneumococci and not a mere one-

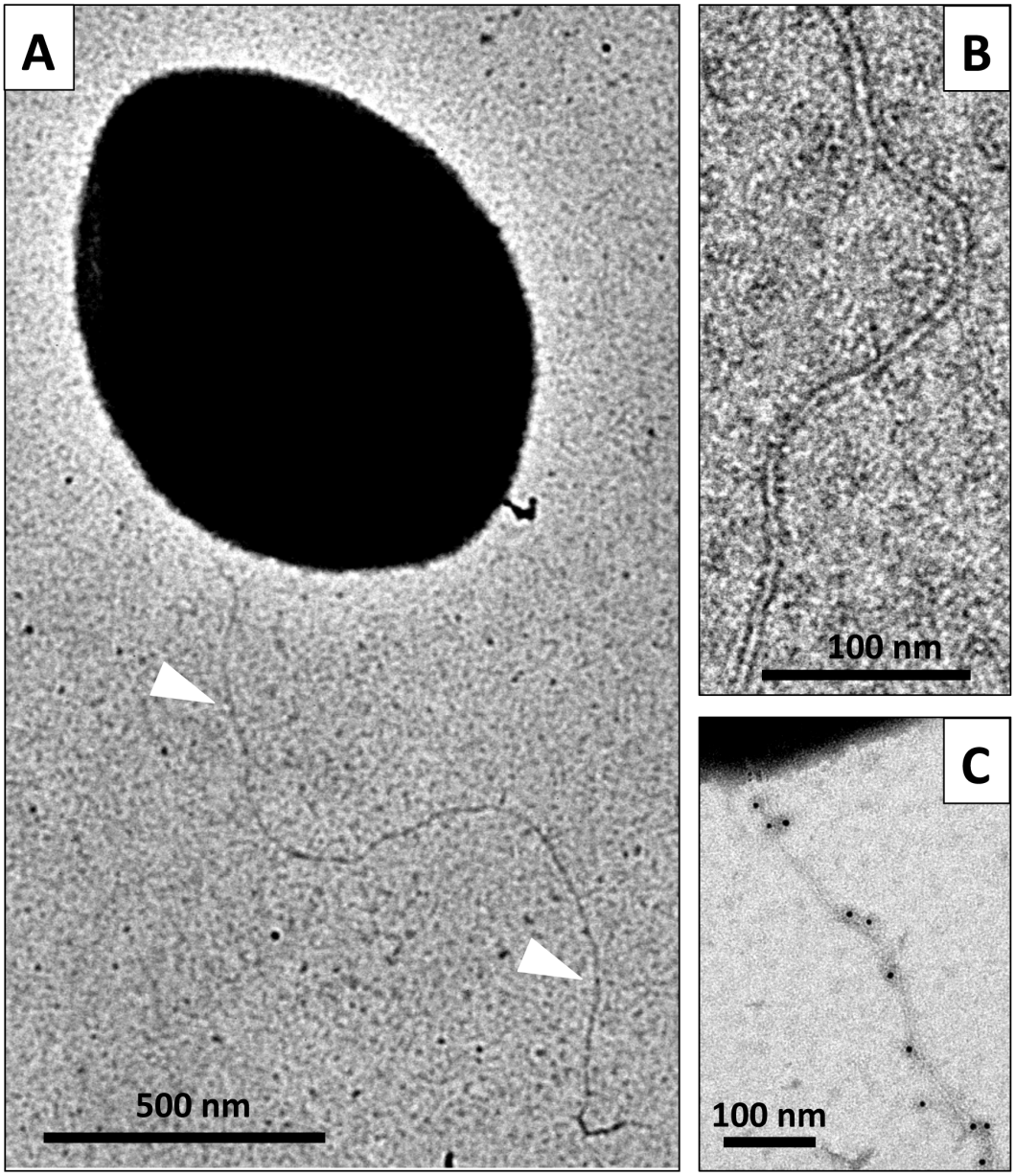


Figure 4. Direct visualization of the competence-induced appendage. Competent reference bacteria observed by transmission electron microscopy. Single appendages, 5–6 nm wide, were observed at the surface most of the competent cells in the culture. Many long filaments were observed, reaching up to several micrometers in length. (**A**) a competent *S. pneumoniae* cell with a long pilus (white triangle). (**B**) closer view of a transformation pilus. (**C**) A pilus observed by transmission electron microscopy after immunogold labeling with anti-FLAG antibody (5 nm gold beads) using the ComGC-FLAG strain. ComGC-FLAG proteins are detected within the appendages.
doi:10.1371/journal.ppat.1003473.g004

off property of our reference strain, we observed negatively stained G54 and CP strains by electron microscopy. The G54 strain is a wild-type clinical strain. The CP strain is a laboratory strain that has a different genetic background than our reference strain [26]. In both cases, transformation pili were observed at the surface of competent cells (Figure S3). Therefore, we think that transformation pili are found at the surface of most, if not all, pneumoccocal strains, including clinical strains.

The transformation pilus is a Type IV pilus

The pneumococcal transformation pilus is morphologically very similar to Type IV pili found in many Gram-negative bacteria. Its major component, the ComGC pilin, is cleaved and probably methylated by a PilD homolog. We therefore propose that the transformation pilus is a *bona fide* Type IV pilus. Since its length can reach up to 2–3 μm, we think that the "pseudo-pilus" appellation does not apply to the pneumococcal transformation

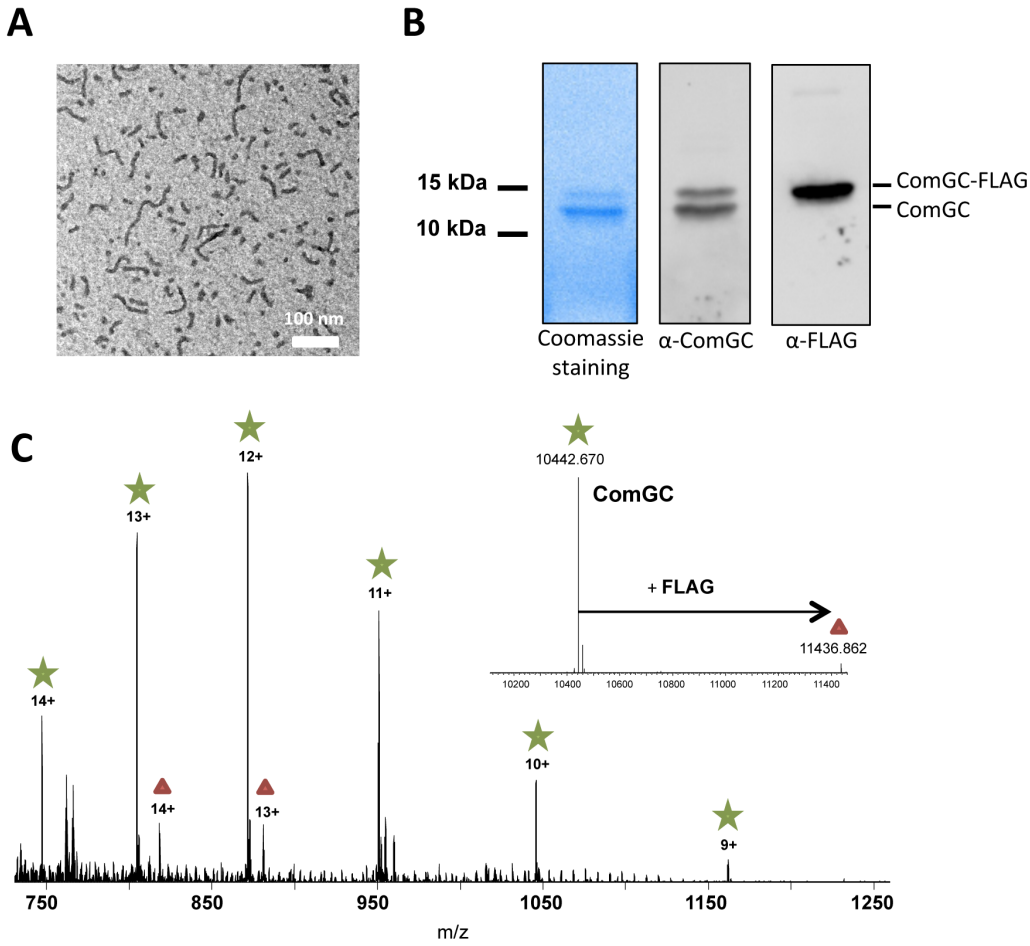


Figure 5. Nature of the transformation pilus. (**A**) Purified pili visualized by negative stain electron microscopy. Short pilus fragments ranging from 50 to 500 nm were observed. (**B**) SDS-PAGE analysis of the purified fraction. Left lane, Coommassie blue staining. Middle lane, Western blot using anti-ComGC antibody. Right lane, Western blot analysis using anti-FLAG analysis. (**C**) Nano-ESI FT-MS spectrum of purified pili. Peaks corresponding to ComGC are labelled with a green star those corresponding to ComGC-FLAG with a red triangle. ComGC-FLAG is present at approximately 6+/−2% abundance of the native form. The deconvoluted spectrum showing monoisotopic masses of the neutral protein forms is presented in the inset. The measured mass of methylated comGC (10,442.670 Da) compares very well to the calculated theoretical mass (10,442.636 Da) with an error of +3 ppm.
doi:10.1371/journal.ppat.1003473.g005

appendage. By comparison, the type II secretion pseudo-pilus is just 50–100 nm long [27]. The transformation pilus is the first Type IV pilus clearly observed in a Gram-positive bacterium. So far, Type IV pilus-dependent gliding motility had been described in *Clostridium* species [28]. However, no clear picture of this pilus was provided. A recent genomic study show the existence of numerous and diverse Type IV pilus-like operons in a wide range of Gram-positive bacteria [29]. This suggests that many other Type IV-like pili remain to be discovered in these bacteria. The conservation of *comG* operons argues in favor of the presence of a transformation pilus in all naturally transformable Gram-positive bacteria. However, species-specific variations in pilus length can be anticipated because of variations in thickness of the capsule and/or the cell wall.

Pilus function. We envision the transformation pilus to act as a ''DNA-trap'' to capture DNA in the environment. In Gram-negative bacteria, Type IV pilus assembly is also essential for natural transformation [30]. It is proposed that this pilus interacts directly with DNA but the molecular details of this interaction remains enigmatic [31,32,33]. No DNA binding protein could be identified in the pilus [34]. Recently a minor pilin, ComP, has been identified as a DNA receptor specific of genus-specific DNA uptake sequence (DUS) motifs in *Neisseria meningitidis* [35]. Our data explicitly show that the pneumoccocal transformation pilus binds DNA. At this stage, it is not clear if this DNA binding ability is due to the physicochemical properties of the pilus and/or due to a yet undetected minor pilin. Interestingly, only ComGA homolog was found indispensable for initial DNA binding at the surface of *B. subtilis* [14]. These bacteria could assemble only short transformation pili that are not sufficient to bind DNA at the surface of the competent cells. An unknown DNA receptor that interacts with ComGA ATPase could be required for efficient DNA binding in this species. It is possible that the mechanism of initial DNA binding at the surface of competent Gram-positive bacteria vary.

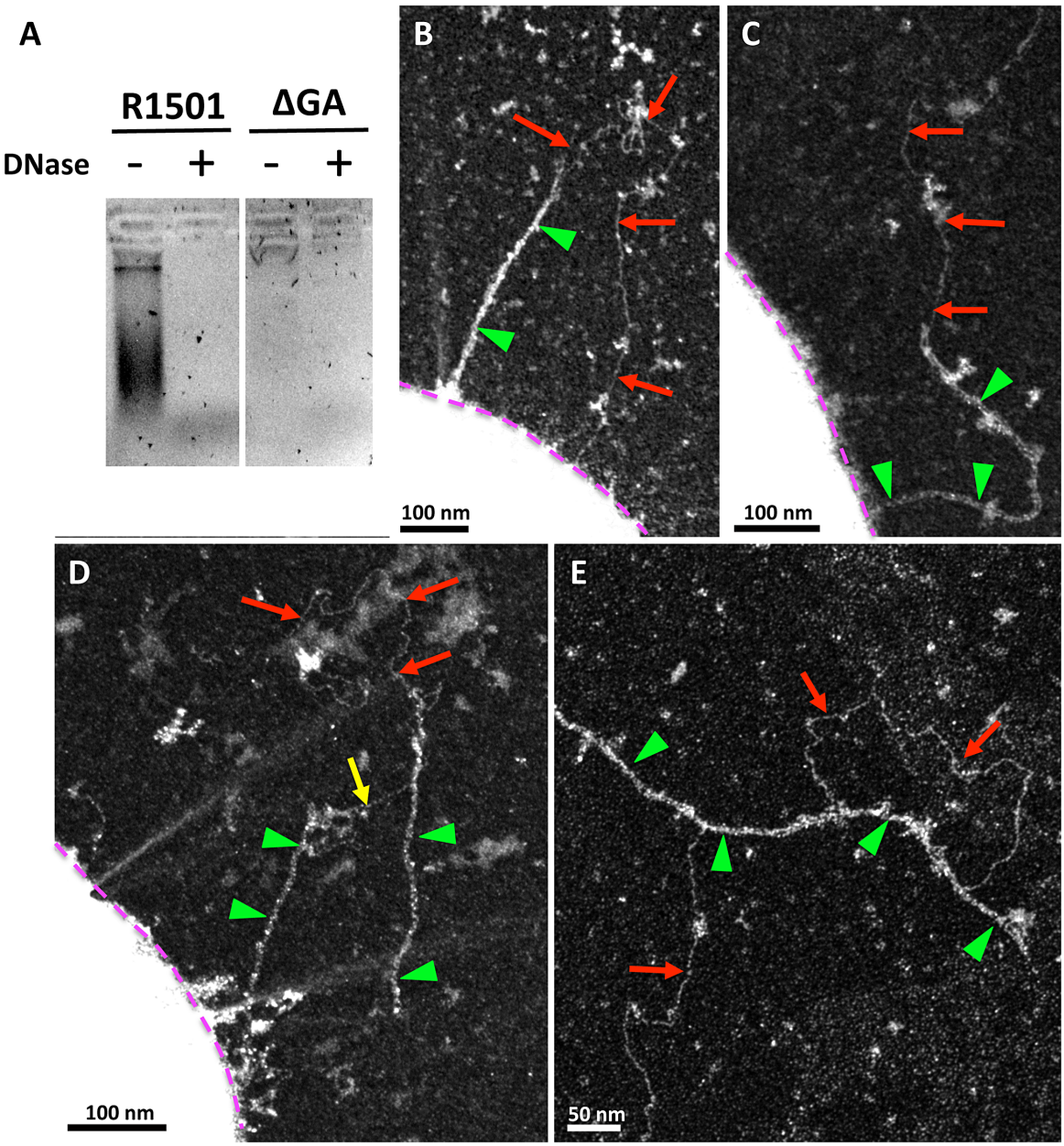


Figure 6. Interactions between transformation pili and DNA. (**A**) DNA co-purified with the pili by affinity chromatography was revealed after migration on agarose gel and Gel Green staining. DNA was present in the ComGC-FLAG strain. In the absence of ComGA, DNA was not detected. DNase treatment unambiguously showed that the bands detected on the gel are DNA. Linear DNA (red arrow) bound to the pilus (green triangle) in *ΔcomFA* (**B**) and *ΔcomEC* (**C**), (**D**) and (**E**) strains. (**B**) and (**C**), linear DNA interacting with a single pilus. The DNA molecule contacts the pilus at several points. **D**, a DNA molecule maintains a broken pilus close to the bacterium (yellow arrow), demonstrating the existence of several DNA binding sites on the pilus. (**E**) Detached pilus found in the medium with bound DNA. In (**B**), (**C**) and (**D**), bacterial cell envelope is highlighted by a purple dotted line.
doi:10.1371/journal.ppat.1003473.g006

Pilus assembly and retraction. In Gram-negative bacteria, 12 to 15 proteins are necessary for Type IV pilus biogenesis [36,37]. The assembly of the pneumococcal transformation pilus could require the simplest Type IV pilus assembly apparatus discovered so far in bacteria. Indeed, only 7 ComG proteins might be sufficient to assemble this pilus. Strikingly, Type IV pili are retractile appendages in Gram-negative bacteria [38]. It is therefore appealing to propose that the transformation pilus could retract to guide bound DNA through the cell wall and the polysaccharide capsule that is often present in bacteria (as in *S. pneumoniae*), and, ultimately deliver it to the ComEA/EndA complex located in the membrane [7]. It is unlikely that this

DNA would find its way along several hundreds of nanometers through the cell wall and capsule without pilus retraction. However, a dedicated ATPase called PilT powers pilus retraction in Gram-negative bacteria and *S. pneumoniae* genome does not encode any PilT homolog. All the *comG* operons that have been identified so far in Gram-positive bacteria encode only one ATPase, ComGA, that is required for the pilus assembly. Therefore, the possible retraction of the transformation pilus and role of ComGA in this process should be assessed.

Our data clearly establish the existence and function of a transformation pilus at the surface of competent pneumococci. As an essential component of the transformation apparatus, it enables this major human Gram-positive pathogen to acquire resistance to antibiotics and to escape vaccines through the binding and incorporation of new genetic material. Future work should establish whether transformation pili exist and play similar roles in other transformable pathogens. Intriguingly, ComG proteins also appear to play an important and direct role in phagosomal escape and virulence in *Listeria monocytogenes* [39]. It would be interesting to determine whether they also assemble into a pilus to play this role.

Materials and Methods

Strain construction and growth

Cells were grown at 37°C under anaerobic condition, without agitation, in a Casamino Acid Tryptone medium (CAT) up to $OD_{600} = 0.3$ for stock cultures [40]. After addition of 15% glycerol, stocks were kept frozen at −80°C. For competence induction, cells were grown in CAT supplemented with BSA (2 g/L), calcium chloride (1 mM) and adjusted to pH = 7.8. Competence was triggered by incubating cells with the Competence Stimulating Peptide (CSP) at $OD_{600} = 0.1$ for 12 min as described previously [40]. For transformation, DNA was then added and transformants were selected on CAT agar plates [17]. Competence was induced following the same protocol in G54 and TCP1251 strains.

For transformation efficiency assays, 100 µL of competent bacteria were transformed by the addition of 100 ng of *S. pneumoniae* R304 genomic DNA (contains the streptomycin resistance gene *str41*). Bacteria were plated in presence and absence of streptomycin (100 µg/mL final concentration) and incubated at 37°C overnight before colony counting.

The annotated names of the *comG* genes in different strains of *S. pneumoniae* are listed in Table S1. The *S. pneumoniae* strains used derived from the non-capsulated R6 strain and are listed in Table S2. The *comGC-FLAG* gene was cloned by PCR using genomic DNA of pneumococcal R6 strain (ATCC BAA-255) as template. The resulting fragment was digested with *NcoI* and *BamHI* and inserted into the same sites of the pCEPx vector [17]. RL001 strain was constructed by transformation of R1501 cells with the pCEPx plasmid containing *comGC-FLAG*, followed by selection with kanamycin (Kan). RL002 was obtained by transformation of RL001 with R1062 chromosomal DNA and selection with spectinomycin (Spc). For RL003, a 2 kb genomic fragment of R6 genome containing *comGC* in the middle was amplified, and the codon 20 was changed from GAG to GTG by cross-over PCR. R1501 was transformed with this modified genomic fragment, and clones were screened by sequencing the *comGC* gene.

Chemically competent *Escherichia coli* BL21 Star (Life Technologies) were used for heterologous production of ComGC soluble domain. The corresponding DNA sequence was amplified from genomic DNA of strain R800 and cloned into pET15b expression vector (Novagen), using *NdeI/XhoI*. The protein was purified from the soluble fraction using IMAC affinity and gel filtration in 50 mM Tris/HCl pH = 8, 200 mM NaCl. The anti-ComGC were raised against the purified protein (Eurogentec).

Detection and purification of cell surface appendages

Shearing experiments were adapted from Sauvonnet et al. [41]. Competence was induced exactly as described above in a 50 mL culture. Cells were harvested by centrifugation 15 min at 4,500 g, 4°C. The pellet was suspended in 1 mL LB and immediately vortexed for 1 min to apply mechanical pressure. The suspension was then centrifuged twice at 13,000 g for 5 min to separate the bacteria (pellet fraction) from the pilus-enriched supernatant (sheared fraction). The supernatant was then precipitated with 10% trichloroacetic acid for 30 min on ice. Both fractions were loaded on SDS 15% polyacrylamide gels and subjected to electrophoresis and immunoblot with rabbit polyclonal antibodies raised against ComGC soluble domain (38–108) or anti-FLAG M2 antibody (Sigma-Aldrich F1804).

The pili containing ComGC-FLAG were purified from the sheared fraction of a 1 L culture. Shearing was performed in 2 mL Tris Buffered Saline (TBS, Tris pH 7.6 0.05 M, NaCl 0.15 M, protease inhibitor cocktail Roche 11873580001) and incubated overnight on a rotating wheel at 4°C with ANTI-FLAG M2 affinity resin (Sigma-Aldrich A2220). After washing with TBS, the pili were eluted by adding 3×FLAG-peptide at 100 µg/mL (Sigma Aldrich F4799) 30 min at room temperature under agitation.

To prevent DNA aspecific binding on the ANTI-FLAG M2 affinity resin, the resin was saturated 2 h at 4°C with a 1.5 kb PCR fragment (20 ng/µL). For DNA detection, 20 µL of the eluted pili were run on a 1% agarose gel and stained with SYBR safe (Life technologies S33102).

Visualization of pili and immunogold labelling

Competence was induced exactly as described above in a 10 mL culture. Cells were harvested by centrifugation 15 min at 4,500 g, 4°C. The pellet was suspended in 60 µL phosphate-buffered saline (PBS) (Sigma-Aldrich P4417). A drop of this suspension was placed on a glow discharged carbon coated grid (EMS, USA) for 1 min. The grid was then placed on a drop of PBS-3% formaldehyde, 0.2% glutaraldehyde for 10 min, and washed on drops of distilled water. The grids were then treated with 2% uranyl acetate in water. Specimens were examined using a Philips CM12 transmission electron microscope operated at 120 kV. Pictures were recorded using a camera KeenView (SIS, Germany) and ITEM software. For immunogold labelling, additional steps were applied after fixation: 3 washes with PBS, PBS–50 mM NH_4Cl (10 min), 3 washes with PBS, PBS with 1% BSA (5 min), 1 hour incubation with ANTI-FLAG M2 antibody (Sigma-Aldrich F1804) diluted 1/100 in PBS with 1% BSA, 3 washes with PBS-BSA 1% (5 min), 1 hour incubation with goat anti-mouse antibody (5 nm gold particles, BritishBioCell, UK) diluted 1/25 in PBS containing 1% BSA.

Fluorescence microscopy

S. pneumoniae cells were grown in the same conditions as above for visualization by electron microscopy. Cells were harvested by centrifugation for 15 min at 4,500 g, 4°C. The pellet was suspended in 500 µL PBS and directly immobilized on poly-L-lysine-coated coverslips. Samples were fixed for 30 min with 3.7% formaldehyde, washed 3 times with PBS containing 1% BSA and incubated on a 100 µL drop of anti-FLAG antibodies (1:300) and secondary Alexa Fluor 488- coupled anti-mouse IgG (Invitrogen). Samples were examined with an Axio Imager.A2 microscope

(Zeiss). Images were taken with AxioVision (Zeiss) and processed in ImageJ [42].

Mass spectrometry

Protein samples were desalted and eluted directly into a 10 µL spray solution of methanol:water:formic acid (75:25:3). Approximately 4 µL was loaded into a coated, medium sized, nano-ESI capillary (Proxeon) and introduced into an Orbitrap Velos mass spectrometer, equipped with ETD module (Thermo Fisher Scientific, Bremen, Germany) using the off-line nanospray source in positive ion mode. A full set of automated positive ion calibrations was performed immediately prior to mass measurement. The transfer capillary temperature was lowered to 100°C, sheath and axillary gasses switched off and source transfer parameters optimised using the auto tune feature. Helium was used as the collision gas in the linear ion trap. For MSn experiments, ions were selected with a 3 Da window and both CID and HCD were performed at normalised collision energies of 15–25%, with the appropriate HCD charge state set and other activation parameters left as default. For ETD the reagent gas was fluoranthene and the interaction time 10 ms. Supplemental activation was used as noted. The FT automatic gain control (AGC) was set at 1×10^6 for MS and 2×10^5 for MSn experiments. Spectra were acquired in the FTMS over several minutes with one microscan and a resolution of 60,000 @ m/z 400 before being summed using Qualbrowser in Thermo Xcalibur 2.1. Summed spectra were then deconvoluted using Xtract and a, b, c−1, y, z, z+1 ions assigned using in house software at a tolerance of 5 ppm. N-terminal ions were verified manually.

Positive staining electron microscopy

Five microliters of bacterial culture (wild-type, *ΔcomFA* or *ΔcomEC*) were diluted in 45 µL of Tris 10 mM, pH 8, NaCl 150 mM. Bacteriophage lambda DNA (0,1 mg/ml final) was then added to bacteria. Five µL of mix were immediately adsorbed onto a 600 mesh copper grid coated with a thin carbon film, activated by glow-discharge. After 1 min, grids were washed with 0,02% (w/vol) uranyl acetate solution (Merck, France) and then dried with filter paper. TEM observations were carried out with a Zeiss 912AB transmission electron microscope in filtered crystallographic dark field mode. Electron micrographs were obtained using a ProScan 1024 HSC digital camera and Soft Imaging Software system.

Supporting Information

Figure S1 Visualization of competence-induced appendages by Immuno-fluorescence. Same picture as in figure 3A, in high resolution. Left column correspond to bright field image, right column to overlay between anti-FLAG antibody fluorescence (green) and DAPI fluorescence (magenta).
(PDF)

Figure S2 Mass spectrometry analysis of the major pilus component. Fragment map of GomGC generated from several top-down mass spectrometry experiments. Sequence coverage is 74%. MS/MS spectra formed through different fragmentation techniques were deconvoluted and de-isotoped in Xtract and the resulting peak lists combined. Fragment peaks were picked and assigned from this combined list using in house software at a tolerance of 5 ppm. Individual experimental conditions were as follows; ETD 14+ charge state 7 ms activation time; 13+ charge state 10 ms activation time, 5 ms activation time with and without supplementary activation; HCD 30 eV collision energy, 13 eV collision energy; CAD 20 eV collision energy.
(PDF)

Figure S3 Transformation pili are observed in other pneumococcal strains. Competent G54 and TCP1251 *S. pneumoniae* cells were observed by transmission electron microscopy. The same appendages were detected in these strains.
(PDF)

Table S1 *ComG* and *pilD* genes in different pneumococcal strains. The name used to designate the comG and pilD genes varies in different pneumoccocal strains. For clarity, we refer to the comG nomenclature used in B. subtilis. Names of the corresponding genes in different S. pneumoniae strains are found in the table.
(DOCX)

Table S2 Strains and plasmids. The strains and plasmids used in this study are listed in this table.
(DOCX)

Acknowledgments

We thank Olivera Francetic and David Cisneros for their help to set up the shearing assay. We thank all the members of the BSSB group for stimulating discussions.

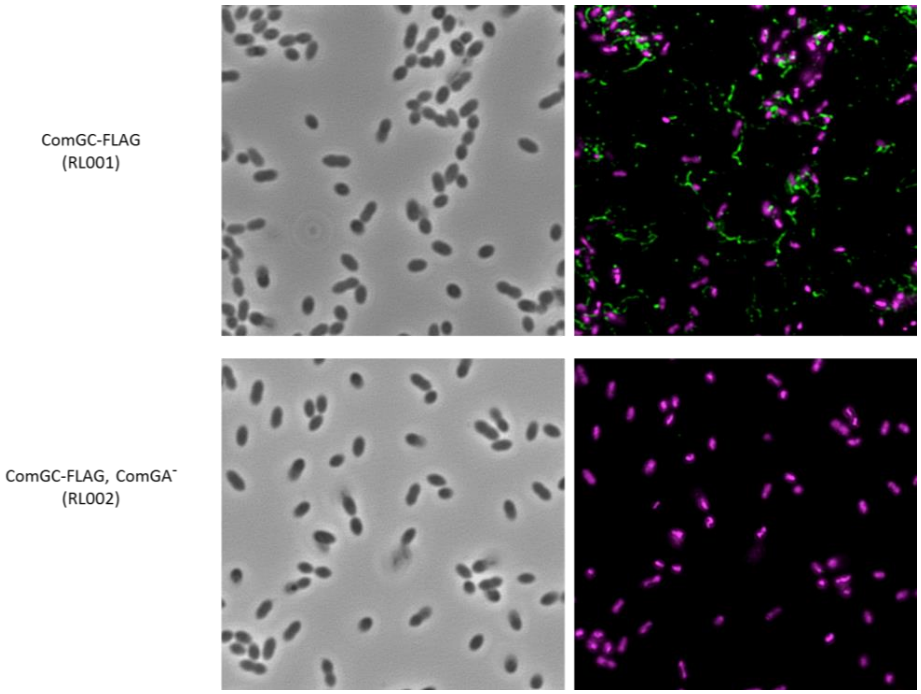
Author Contributions

Conceived and designed the experiments: JCR ELC JPC RF. Performed the experiments: RL GPA SB JG CM AD NC. Analyzed the data: JCR ELC JPC RF. Wrote the paper: RL JG NC JCR JPC RF.

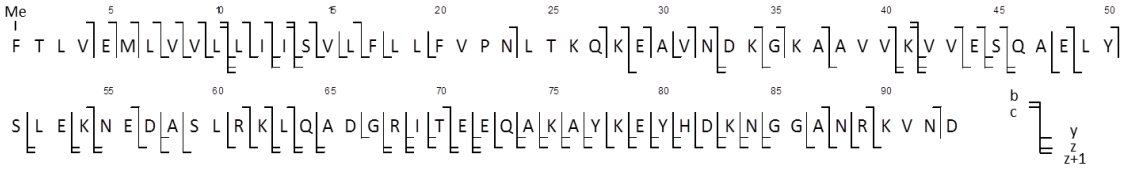
References

1. Griffith F (1928) The Significance of Pneumococcal Types. The Journal of hygiene 27: 113–159.
2. Johnsborg O, Eldholm V, Havarstein LS (2007) Natural genetic transformation: prevalence, mechanisms and function. Research in microbiology 158: 767–778.
3. Popa O, Dagan T (2011) Trends and barriers to lateral gene transfer in prokaryotes. Current opinion in microbiology 14: 615–623.
4. Hiller NL, Ahmed A, Powell E, Martin DP, Eutsey R, et al. (2010) Generation of genic diversity among Streptococcus pneumoniae strains via horizontal gene transfer during a chronic polyclonal pediatric infection. PLoS pathogens 6: e1001108.
5. Havarstein LS, Coomaraswamy G, Morrison DA (1995) An unmodified heptadecapeptide pheromone induces competence for genetic transformation in Streptococcus pneumoniae. Proceedings of the National Academy of Sciences of the United States of America 92: 11140–11144.
6. Prudhomme M, Attaiech L, Sanchez G, Martin B, Claverys JP (2006) Antibiotic stress induces genetic transformability in the human pathogen Streptococcus pneumoniae. Science 313: 89–92.
7. Chen I, Dubnau D (2004) DNA uptake during bacterial transformation. Nature reviews Microbiology 2: 241–249.
8. Claverys JP, Martin B, Polard P (2009) The genetic transformation machinery: composition, localization, and mechanism. FEMS microbiology reviews 33: 643–656.
9. Chen I, Provvedi R, Dubnau D (2006) A macromolecular complex formed by a pilin-like protein in competent Bacillus subtilis. The Journal of biological chemistry 281: 21720–21727.
10. Dagkessamanskaia A, Moscoso M, Henard V, Guiral S, Overweg K, et al. (2004) Interconnection of competence, stress and CiaR regulons in Streptococcus pneumoniae: competence triggers stationary phase autolysis of ciaR mutant cells. Molecular microbiology 51: 1071–1086.
11. Peterson SN, Sung CK, Cline R, Desai BV, Snesrud EC, et al. (2004) Identification of competence pheromone responsive genes in Streptococcus pneumoniae by use of DNA microarrays. Molecular microbiology 51: 1051–1070.
12. Kaufenstein M, van der Laan M, Graumann PL (2011) The three-layered DNA uptake machinery at the cell pole in competent Bacillus subtilis cells is a stable complex. Journal of bacteriology 193: 1633–1642.
13. Berge M, Moscoso M, Prudhomme M, Martin B, Claverys JP (2002) Uptake of transforming DNA in Gram-positive bacteria: a view from Streptococcus pneumoniae. Molecular microbiology 45: 411–421.

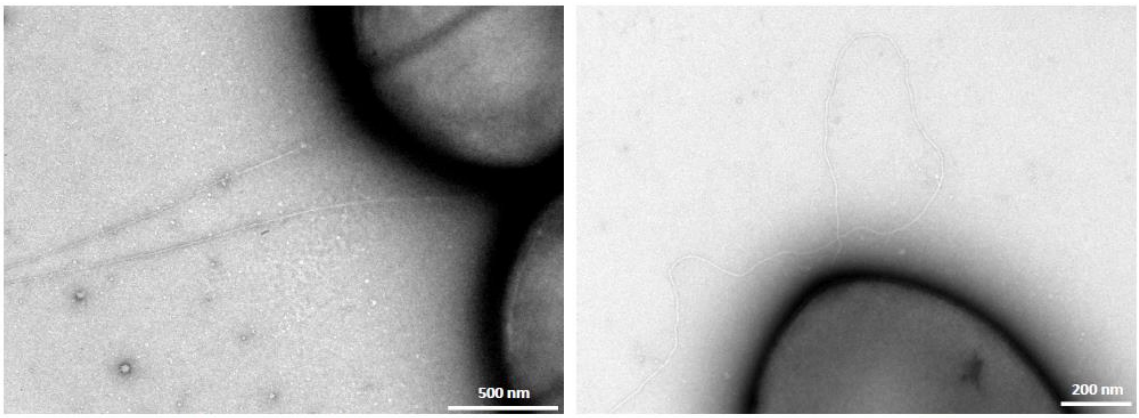
14. Briley K, Jr., Dorsey-Oresto A, Prepiak P, Dias MJ, Mann JM, et al. (2011) The secretion ATPase ComGA is required for the binding and transport of transforming DNA. Molecular microbiology 81: 818–830.
15. Chung YS, Dubnau D (1998) All seven comG open reading frames are required for DNA binding during transformation of competent Bacillus subtilis. Journal of bacteriology 180: 41–45.
16. Nunn D, Bergman S, Lory S (1990) Products of three accessory genes, pilB, pilC, and pilD, are required for biogenesis of Pseudomonas aeruginosa pili. Journal of bacteriology 172: 2911–2919.
17. Martin B, Granadel C, Campo N, Henard V, Prudhomme M, et al. (2010) Expression and maintenance of ComD-ComE, the two-component signal-transduction system that controls competence of Streptococcus pneumoniae. Molecular microbiology 75: 1513–1528.
18. Chamot-Rooke J, Mikaty G, Malosse C, Soyer M, Dumont A, et al. (2011) Posttranslational modification of pili upon cell contact triggers N. meningitidis dissemination. Science 331: 778–782.
19. Strom MS, Nunn DN, Lory S (1993) A single bifunctional enzyme, PilD, catalyzes cleavage and N-methylation of proteins belonging to the type IV pilin family. Proceedings of the National Academy of Sciences of the United States of America 90: 2404–2408.
20. Strom MS, Lory S (1991) Amino acid substitutions in pilin of Pseudomonas aeruginosa. Effect on leader peptide cleavage, amino-terminal methylation, and pilus assembly. The Journal of biological chemistry 266: 1656–1664.
21. Burton B, Dubnau D (2010) Membrane-associated DNA transport machines. Cold Spring Harbor perspectives in biology 2: a000406.
22. Dupaigne P, Le Breton C, Fabre F, Gangloff S, Le Cam E, et al. (2008) The Srs2 helicase activity is stimulated by Rad51 filaments on dsDNA: implications for crossover incidence during mitotic recombination. Molecular cell 29: 243–254.
23. Mejean V, Claverys JP (1993) DNA processing during entry in transformation of Streptococcus pneumoniae. The Journal of biological chemistry 268: 5594–5599.
24. Barocchi MA, Ries J, Zogaj X, Hemsley C, Albiger B, et al. (2006) A pneumococcal pilus influences virulence and host inflammatory responses. Proceedings of the National Academy of Sciences of the United States of America 103: 2857–2862.
25. Alloing G, Martin B, Granadel C, Claverys JP (1998) Development of competence in Streptococcus pneumonaie: pheromone autoinduction and control of quorum sensing by the oligopeptide permease. Molecular microbiology 29: 75–83.
26. Pestova EV, Havarstein LS, Morrison DA (1996) Regulation of competence for genetic transformation in Streptococcus pneumoniae by an auto-induced peptide pheromone and a two-component regulatory system. Molecular microbiology 21: 853–862.
27. Campos M, Cisneros DA, Nivaskumar M, Francetic O (2013) The type II secretion system - a dynamic fiber assembly nanomachine. Research in microbiology [Epub ahead of print] doi: 10.1016/j.resmic.2013.03.013.
28. Varga JJ, Nguyen V, O'Brien DK, Rodgers K, Walker RA, et al. (2006) Type IV pili-dependent gliding motility in the Gram-positive pathogen Clostridium perfringens and other Clostridia. Molecular microbiology 62: 680–694.
29. Imam S, Chen Z, Roos DS, Pohlschroder M (2011) Identification of surprisingly diverse type IV pili, across a broad range of gram-positive bacteria. PloS one 6: e28919.
30. Craig L, Pique ME, Tainer JA (2004) Type IV pilus structure and bacterial pathogenicity. Nature reviews Microbiology 2: 363–378.
31. Biswas GD, Sox T, Blackman E, Sparling PF (1977) Factors affecting genetic transformation of Neisseria gonorrhoeae. Journal of bacteriology 129: 983–992.
32. Dougherty TJ, Asmus A, Tomasz A (1979) Specificity of DNA uptake in genetic transformation of gonococci. Biochemical and biophysical research communications 86: 97–104.
33. van Schaik EJ, Giltner CL, Audette GF, Keizer DW, Bautista DL, et al. (2005) DNA binding: a novel function of Pseudomonas aeruginosa type IV pili. Journal of bacteriology 187: 1455–1464.
34. Lang E, Haugen K, Fleckenstein B, Homberset H, Frye SA, et al. (2009) Identification of neisserial DNA binding components. Microbiology 155: 852–862.
35. Cehovin A, Simpson PJ, McDowell MA, Brown DR, Noschese R, et al. (2013) Specific DNA recognition mediated by a type IV pilin. Proceedings of the National Academy of Sciences of the United States of America 110: 3065–3070.
36. Carbonnelle E, Helaine S, Nassif X, Pelicic V (2006) A systematic genetic analysis in Neisseria meningitidis defines the Pil proteins required for assembly, functionality, stabilization and export of type IV pili. Molecular microbiology 61: 1510–1522.
37. Georgiadou M, Castagnini M, Karimova G, Ladant D, Pelicic V (2012) Large-scale study of the interactions between proteins involved in type IV pilus biology in Neisseria meningitidis: characterization of a subcomplex involved in pilus assembly. Molecular microbiology 84: 857–873.
38. Merz AJ, So M, Sheetz MP (2000) Pilus retraction powers bacterial twitching motility. Nature 407: 98–102.
39. Rabinovich L, Sigal N, Borovok I, Nir-Paz R, Herskovits AA (2012) Prophage excision activates Listeria competence genes that promote phagosomal escape and virulence. Cell 150: 792–802.
40. Martin B, Prudhomme M, Alloing G, Granadel C, Claverys JP (2000) Cross-regulation of competence pheromone production and export in the early control of transformation in Streptococcus pneumoniae. Molecular microbiology 38: 867–878.
41. Sauvonnet N, Vignon G, Pugsley AP, Gounon P (2000) Pilus formation and protein secretion by the same machinery in Escherichia coli. The EMBO journal 19: 2221–2228.
42. Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. Nature methods 9: 671–675.
43. Hansen JK, Forest KT (2006) Type IV pilin structures: insights on shared architecture, fiber assembly, receptor binding and type II secretion. Journal of molecular microbiology and biotechnology 11: 192–207.



Supplementary Figure 1



Supplementary Figure 2



S. Pneumoniae G54 strain
(capsulated)

S. Pneumoniae TCP1251 strain

Supplementary Figure 3

Strain	*comGA*	*comGB*	*comGC*	*comGD*	*comGE*	*comGF*	*comGG*	*pilD*
R6	*spr1864 /cglA*	*spr1863 /cglB*	*spr_1862 /cglC*	*spr_1861 /cglD*	*Not annotated*	*spr_1859*	*spr_1858*	*pilD/spr1628*
D39	*SPD_1863 /cglA*	*SPD_1862 /cglB*	*SPD_1861 /cglC*	*SPD_1860 /cglD*	*SPD_1859*	*SPD_1858*	*SPD_1857*	*SPD_1593*
G54	*SPG_1968*	*SPG_1967*	*SPG_1966*	*SPG_1965*	*SPG_1964*	*SPG_1963*	*SPG_1962*	*SPG_1704*
CP	*Not Sequenced*	*Not Sequenced*	*Not Sequenced*	*Not Sequenced*	*Not Sequenced*	*Not Sequenced*	*Not Sequenced*	*Not Sequenced*
TIGR4	*SP_2053*	*SP_2052*	*SP_2051*	*SP_2050*	*SP_2049*	*SP_2048*	*SP_2047*	*SP_1808*

Table S1

Strain number	Genotype/relevant feature[a]	Reference
R800	R6 derivative	[1]
G54	Clinical isolate of serotype 19F	[2]
TCP1251	Rx derivative but *malM511, rpsL1, bgl1; Sm^R*	[3]
R1501	R800 but Δ*comC*	[4]
R304	R800 derivative, *nov1, rif23, str41; Nov^R, Rif^R, Sm^R*	[5]
R1916	R1501 but *ssbB::luc (ssbB$^+$), comGA::kan; Cm^R, Kan^R*	Claverys' strain collection
R998	R1501 but *comEC::spc; Cm^R, Spc^R*	Claverys' strain collection
R1063	R1501 but *comFA::spc; Cm^R, Spc^R*	Claverys' strain collection
RL001	R1501, but CEPx-*comGC-FLAG* (from plasmid pCEPx-*comGC-FLAG*); *Kan^R*	This study
RL002	RL001 , but *comGA::spc3^C* (from strain R1062); *Kan^R, Spc^R*	This study
RL003	R1501 but *comGC* E20A (point mutation of ComGC pilin)	This study

Plasmids		
pCEPx	ColE1 (pBR322) derivative containing the ComX-dependent promoter, P_X, and the RBS of ssbB; *Kan^R*	[6]

^{R}, resistance.

Table S2

1. Lefevre JC, Claverys JP, Sicard AM (1979) Donor deoxyribonucleic acid length and marker effect in pneumococcal transformation. Journal of bacteriology 138: 80-86.

2. Dopazo J, Mendoza A, Herrero J, Caldara F, Humbert Y, et al. (2001) Annotated draft genomic sequence from a Streptococcus pneumoniae type 19F clinical isolate. Microbial drug resistance 7: 99-125.

3. Pestova EV, Havarstein LS, Morrison DA (1996) Regulation of competence for genetic transformation in Streptococcus pneumoniae by an auto-induced peptide pheromone and a two-component regulatory system. Molecular microbiology 21: 853-862.

4. Dagkessamanskaia A, Moscoso M, Henard V, Guiral S, Overweg K, et al. (2004) Interconnection of competence, stress and CiaR regulons in Streptococcus pneumoniae: competence triggers stationary phase autolysis of ciaR mutant cells. Molecular microbiology 51: 1071-1086.

5. Mortier-Barriere I, de Saizieu A, Claverys JP, Martin B (1998) Competence-specific induction of recA is required for full recombination proficiency during transformation in Streptococcus pneumoniae. Molecular microbiology 27: 159-170.

6. Martin B, Granadel C, Campo N, Henard V, Prudhomme M, et al. (2010) Expression and maintenance of ComD-ComE, the two-component signal-transduction system that controls competence of Streptococcus pneumoniae. Molecular microbiology 75: 1513-1528.

3. Conclusions from Analysis of ComGC from *Streptococcus pneumoniae*

Mass Spectrometry and Biological Relevance

Mass profiling by Orbitrap MS confirmed that ComGC was the major component of the competence pilus and provided an explanation for the puzzling immunogold labelling results. Integration of only a small proportion of ComGC-FLAG into the pilus fibre was found to be the origin of the lower than expected level of immune gold labelling. The ratio of ComGC-FLAG to ComGC calculated from the mass profiling spectra correlated well with the ratio calculated from immunogold labelling experiments. This provides further evidence that mass spectrometry performed on the protein level can be at least semi-quantitative and is a good method to probe the relative abundance of co-expressed proteoforms.

MS/MS experiments performed on ComGC identified a truncated and methylated N-terminus and ruled out the presence of other posttranslational modifications on the pilus fibre. This included the presence of isopeptide bonds which have been observed in other pili from Gram positive bacteria[1]. N-terminal processing confirmed the action of the PilD homolog in *S. pneumoniae*.

The HCD and CAD fragmentation modes were responsible for the b_4-b_{16} and y_{76}-y_{86} ion series formed from fragmentation of the N-terminal region of ComGC (supplementary figure 2 of preceding article). This fragmentation pattern is similar to that from meningococcal PilE. Extensive sequence coverage was achieved by combining the results of both electronic and vibrational fragmentation methods and confirms that with minor adjustments, the experimental parameters developed for fragmentation of PilE are useful for the identification of pilins from other bacterial pathogens. This probably occurs because of their similar size and the fact that they likely share a similar gas phase structure.

All of the mass spectrometry evidence detailed in this work supports the conclusion that ComGC is the major pilin of a *bona fide* type IV pilus. This is the first time such an appendage has been identified in a Gram positive bacterium.

Bibliography

[1] H. J. Kang, F. Coulibaly, F. Clow, T. Proft and E. N. Baker.Stabilizing isopeptide bonds revealed in Gram-positive bacterial pilus structure. *Science*, **2007**, *318*, 1625.

General Conclusion

The principal goal of this thesis was to develop a top-down mass spectrometry approach for the identification of posttranslational modifications (PTMs) on the protein PilE: the primary constituent of type IV pili expressed by the human pathogen *Neisseria meningitidis* (Nm). This approach was to be developed on a reference strain and then extended to novel, previously uncharacterised clinical isolates with the goal of mapping the diversity of PTM in the wider bacterial population. If successful, it was hoped to apply the methodology to pili expressed by other pathogens that may harbour as yet unknown PTMs.

To provide a base upon which to construct the top-down experiment, mass profiling coupled with a bottom-up mass spectrometry approach was first used to map all PTMs on all proteoforms of PilE expressed by the Nm 8013 reference strain (PilE-8013). This approach was then applied to PilE from several mutants of the 8013 strain and the PTM localisation data used to help elucidate the biological function of the phosphoglycerol modification. The role of this unusual PTM was completely unknown at the beginning of this thesis.

Building on these results a top-down methodology was then developed for PTM characterisation of the major proteoform of PilE-8013. This was performed in parallel on both FT-ICR and Orbitrap platforms. Since top-down mass spectrometry can be both technical demanding and protein dependant, several generations of FT-ICR mass spectrometer were evaluated and many parameters were investigated on both instrument types, through large scale experiments performed directly on PilE. This allowed conditions for ECD fragmentation on the solariX FT-ICR and ETD on the Orbitrap Velos to be highly optimised. When employed for top down MS/MS, extensive protein coverage was achieved that was more than sufficient for PTM localisation. This allowed all PTMs of the major proteoform of PilE to be mapped in a single shot and resulted in high confidence PTM characterisation in a much shorter time frame than with the bottom-up approach.

Several clinical isolates were then selected from a large group of Nm strains obtained from patients treated for sepsis and meningitis at the Limoges university hospital during recent sporadic outbreaks of meningitis. Clones of some of these strains were isolated and PilE examined by mass profiling experiments. It was soon clear that these isolates expressed PilE in a greater number of proteoforms than ever previously described and harboured extensive posttranslational modification.

The top-down approach developed for PilE-8013 analysis on the solariX FT-ICR mass spectrometer was then used to fully characterise PilE from two of these strains. In the first case the top-down and bottom-up methodologies were compared for analysis of PilE-278534D. The top-down experiment developed on PilE-8013 transferred extremely well and when refined,

provided sufficient sequence coverage for complete PTM localisation, including a much higher level of glycosylation than previously reported. Whilst the bottom-up approach excelled at identifying the PTM sites, top-down MS was necessary to make the link between the PTMs and their parent proteoforms. This is a real biological example of a general case where bottom-up proteomics is fundamentally unable to achieve complete PTM characterisation and a top-down methodology is required to map PTMs to their parent proteoforms.

Top-down characterisation of PilE expressed by a second strain, 427707C, was also performed. A combination of high resolution mass profiling and top-down MS/MS revealed the presence of at least six proteoforms, compared with the three visualised by low resolution mass profiling, and also indicated the presence of an unknown PTM. This PTM was characterised entirely by mass spectrometry revealing its identity as a previously unreported glycan, with a similar structure to GATDH. Similarly to PilE-278534D, top-down MS/MS showed that PilE expressed by the 427707C strain was highly posttranslationally modified with multiple glycan moieties.

Sequencing of the *pilE* gene indicated that 427707C, 278534D, and indeed all of these previously uncharacterised clinical isolates had a class II genetic organisation and expressed PilE with highly conserved primary structures. High resolution mass profiling was used to provide further evidence that in all of the clinical isolates PilE was modified with an unprecedented level of glycosylation. This was supported by glycotyping experiments which definitively identified the glycan on all strains, including those that had not been fully characterised by top down MS/MS. This represents both the first large scale study on PTM of PilE and also the first complete PTM characterisation of PilE expressed by class II strains.

Armed with this information it was now of interest to explore the origins of this hyper-glycosylated phenotype. Using a mutant expressing an invariable type PilE sequence from a class I strain in a class II genetic background, it was shown that the glycosylation status of PilE was directed by the PilE primary structure. To characterise the effect of glycosylation on the pilus fibre additional techniques were employed. Transmission electron microscopy indicated that glycosylation had little effect on fibre morphology; however molecular modelling clearly indicated the glycosylation had a profound effect on the fibre's surface. Pili from the class II strains are covered in glycan subunits that significantly reduced the surface accessibility of the pilus.

Placing these results in a wider biological context suggests that strains with a class II genetic organisation express PilE with invariable primary sequences and high levels of glycosylation. This is stark contrast to PilE expressed by class I strains which have highly variable primary sequences but exhibit lower glycosylation levels. This observation may offer a response to the open question posed in the introduction to this thesis, "if antigenic variation PilE promotes immune evasion, how

do all of these strains that express an invariable PilE sequence escape immune detection?" We believe that in class II strains, high levels of glycosylation coupled with variation provided by the *pgl* system provide a different means of antigenic variation to the principle mechanism employed in class I strains. To our knowledge this is the first time such an explanation has been proposed.

Finally, the top-down method has been transferred and applied to type IV pili expressed by other pathogens such as *E. coli* and *S. pneumoniae*. In the latter it has been used to characterise a type IV pilus expressed by a Gram positive bacterium for the first time.

The initial goals set out at the beginning of the thesis have been achieved. The top-down and bottom-up approaches have been compared on both FT-ICR and Orbitrap platforms and the strengths and weaknesses of each platform have been explored. The top-down mass spectrometry methods developed have been applied to pilin proteins purified from both reference strains and previously uncharacterised clinical isolates of *Neisseria meningitidis*, and also other pathogenic bacteria. The unique ability of mass spectrometry to characterise proteoforms on the molecular level has at each stage been used to raise, then answer, biological questions. Once one question had been resolved, the methodology was developed in order answer another. Whilst the experiments themselves were tailored to PilE, techniques and knowledge that are highly relevant to the general top-down approach have also been learnt and these will be invaluable in extending the methodology to more complex samples and entire proteomes.

The work presented in this thesis provides a clear example of the diversity that is present in the bacterial proteome and the relevance that posttranslationally modified proteins have in mediating host-pathogen interactions. Little is currently known about the extent of global posttranslational modification in bacteria, let alone those relevant to human health. Whilst many studies concentrate on the changes in the proteome of the host cell upon infection, very few have attempted to differentially map changes in the bacterial proteome. Top-down proteomics has now developed sufficiently to be set to this task and the results presented in this thesis confirm that this area is ripe for new discoveries.

The work presented herein also provides an example of the synergic discovery process that will be key to the development of top-down proteomics, particularly in field of human health. Top-down proteomics not only has the potential to test biological hypotheses but has the unique ability to drive and direct biological thought in a more complete way than bottom-up proteomics can ever hope to achieve. It is becoming clear that cellular processes are governed not by protein but by proteoform expression, and top-down proteomics represents a unique tool to provide a quantitative, detailed overview of entire proteomes in all their complexity.

Employing top-down proteomics to characterise the proteoform complement of specific cell types would exponentially increase our understanding of human biology and allow cellular pathways to be mapped at the proteoform level for the first time. This would have profound implications for the way we approach human health and disease. Correlation of basal proteoform populations to a diseased state would allow top-down proteomics to transcend the monitoring of known biomarkers and drive the discovery of new ones. It may even enable the creation of precise, proteoform targeted therapies and permit personalised, molecular level disease diagnosis that far surpasses the capabilities of modern medicine. Whatever future applications top-down proteomics may find, it is clear that it will have a major impact on our comprehension of fundamental biology.

If the human genome project provides a dictionary to the language of life, top-down proteomics surely promises to teach us how it is spoken.

Annex

Materials & Methods

1. Pili Preparation

For details for pili preparation please see one of the articles inserted into this thesis.

2. Mass Spectrometry

Several mass spectrometers were used during this thesis. Details regarding instrument set up are found in the articles inserted into chapters 2 and 3 for the Q-Tof Premier (Waters) and Apex III FT-ICR[1, 2] and the accepted manuscript inserted into chapter 4 for the Orbitrap Velos and solariX FT-ICR. All additional details are included in the text.

2.1. Basic Principles of ICR and Orbitrap Analysers

ICR

Ion cyclotron resonance exploits the fact that a charged particle travelling perpendicularly to a magnetic field will experience a force perpendicular to that field (Lorenz force). Lawrence and Livingston demonstrated experimentally in 1932 that this resulted in a circular procession of the particle at a characteristic unperturbed cyclotron frequency ω_c defined by equation (6) where B_0 is the magnetic field strength, q the charge carried by the particle and m the particle mass[3].

$$\omega_c = \frac{qB_0}{m} \qquad (6)$$

This important relationship suggests that if an ion could be trapped in such a field, its rotational frequency would be directionally proportionate to its *m/z*. When the magnetic field strength is known this frequency to *m/z* conversion is given by equation (7) following substitution for q. The typical rotational frequency of an ion at *m/z* 800 in a 12T field is thus 230 kHz and falls in the low radio frequency (RF) range.

$$v_c = \frac{\omega_c}{2\pi} = \frac{1.535611 \times 10^7 \, B_0}{m/z} \qquad (7)$$

Trapping of such charged particles is achieved in a Penning trap or ICR cell by application of an axial electric field perpendicular to that of the magnetic field using two metal plates or end caps. Ion motion is now complicated by coupling of the cyclotron frequency with the radial component of the trapping potential and the relevant equations of motion now have quadratic form, giving rise to two solutions: ω_+ the reduced cyclotron frequency and ω_- the magnetron frequency. First approximations for these experimentally important frequencies are given in equation (8) where ω_c is defined by equation (6) and ω_z, the axial oscillation frequency, by equation (9) where q is the elementary charge, V_{trap} the voltage applied to the trapping plates, a the trap length in meters and α a constant that depends on the trap geometry. Trapped ions thus have a stable, but complex, motion that is described in detail elsewhere.

$$\omega_+ = \frac{\omega_c}{2} + \sqrt{\left(\frac{\omega_c}{2}\right)^2 - \frac{\omega_z^2}{2}} \qquad \omega_- = \frac{\omega_c}{2} - \sqrt{\left(\frac{\omega_c}{2}\right)^2 - \frac{\omega_z^2}{2}} \qquad (8)$$

$$\omega_z = \sqrt{\frac{2qV_{trap}\alpha}{ma^2}} \qquad (9)$$

Understanding the basic principles of ion motion in an ICR cell is important as they are key to the success or failure of the top-down experiment. Even in hybrid FT-ICR instruments ion trapping, fragmentation and mass measurement may all occur within the ICR cell and the ion packet must be suitably controlled during each of the three stages.

When ions are first injected into an ICR cell such as that shown in Figure 104, they are squeezed radially by the magnetic field and axially by the end plate trapping voltage and tend to reside near the centre of the cell. Ions may be slightly spread out in the *xy* plane due to magnetron expansion[4]. Often there will be a slight pause of several ms in the experiment here to allow "relaxation" of the ion cloud.

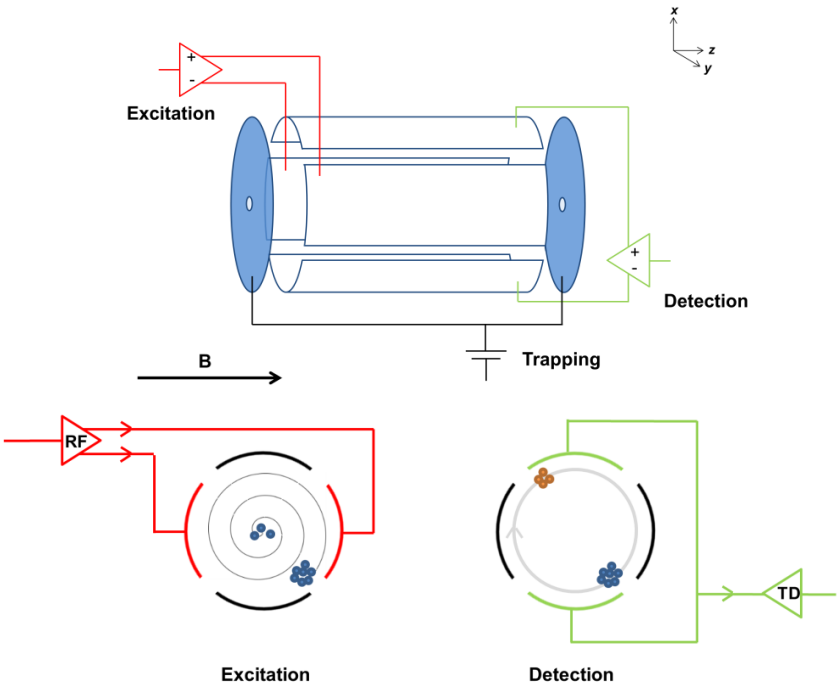


Figure 104 - Schematic of a cylindircal ICR cell (Top) and Ion motion during excitation and detection (Bottom)

Whilst ions of the same *m/z* will all precess with identical frequency, at this point in the experiment they will be out of phase due to their different arrival times in the cell and slightly different kinetic energy. In order to be detected they must be excited into a larger orbit close to

the two detection plates at the limits of the cell. This can be achieved by application of a broadband RF sweep (chirp) to the excitation plates that excites all ions nearly simultaneously to a radius which is independent of *m/z* and is given by equation (10). Other excitation modes are reviewed elsewhere.

$$r = \frac{\beta_{dipolar} V_{p-p} \sqrt{\frac{1}{Sweep\ Rate}}}{2dB_0} \qquad (10)$$

During excitation ions with the same *m/z* ratio also begin to align in phase producing comet type clouds and eventually forming well defined ion packets that process at the reduced cyclotron frequency close to cell limit (Figure 104 bottom). Each of these rotating ion packets will induce an alternating image current on the detector plates according to Faraday's law with a frequency proportionate to their *m/z* ratio and an intensity proportionate to the number of ions and their charge state. Since each ion packet induces its own current at its own particular frequency, the measured current is a superposition of all of the individual sinusoidal ion currents and thus an incredibly complex waveform.

In the ideal scenario of a perfect vacuum and with no ion-ion interactions, after excitation ions would continue to rotate unhindered at their reduced cyclotron frequency allowing an unlimited time for current detection, however even at the very high vacuum conditions of $<10^{-9}$-10^{-11} torr in modern instruments, collisions with residual gas particles and space charge effects cause dampening of the signal over time. Detection is therefore usually performed for several milliseconds up to several seconds depending on the resolution required. The resolution at low pressure is given by equation (11).

$$R = \frac{m}{\Delta m} = \frac{1.274 \times 10^7\ B_0 T_{aq}}{m/z} \qquad (11)$$

Equation (11) demonstrates why ICR achieves such high resolution. If one considers an ion at *m/z* 800 with a detection time of 1 s and modest field strength of 7 T the resolving power is an impressive 110,000. Several important relationships should also be noted from this equation namely that resolution decreases linearly with *m/z* and is directly proportional to both the magnetic field strength and acquisition time.

Extracting exploitable information from the recorded time domain signal is achieved by application of a Fourier transform (FT) to the outputted time domain (TD) waveform. This converts the signal into a frequency domain (FD) spectrum with both real and imaginary parts. The real part contains the individual frequencies of measured ions and is called the absorption mode spectrum. The imaginary part contains the dispersion mode spectrum. The peak shape,

intensity and the mass of these spectra depend on an additional property called the phase of the ions. Historically phase has been difficult to calculate since it depends on *m/z* and various instrumental parameters and, in order to overcome this problem, the magnitude mode spectrum which combines both real and imaginary parts of the FT has been used. The relationship between the different frequency modes is given by equation (12).

$$\omega_{mag} = \sqrt{\omega_{real}^2 + \omega_{imag}^2} \qquad (12)$$

The use of magnitude mode has the unfortunate effect of decreasing the resolution by approximately factor of 2 but abrogates the need for phase calculation. Recently, a robust and easy to use tool for the calculation of magnitude mode FT-ICR spectra (phasing) has been published[5, 6].

Further peak processing such apodisation is often used to improve peak shape and is also reviewed elsewhere.

The practical conclusions that should be taken from the above treatment are summarised below.

- *m/z* is easily determined from the frequency of orbiting trapped ions
- Signal amplitude translates to peak intensity and depends on the number of ions and their charge state
- Resolution is often very high and theoretically depends on *m/z*, magnetic field strength, detection time, and the form of the FT spectrum used (magnitude versus absorption)
- Mass precision depends on the sampling rate and therefore m/z (Not discussed above)
- Mass accuracy is determined by effective calibration across entire mass range and the form of the calibration curve applied (Not discussed above)

In the above discussion more complex factors such as ion-ion spatial charge effects, non-ideal trapping and the form of excitation and detection frequencies have not been developed. For economy of space many fascinating techniques and operation modes have also been omitted. For an introduction to these, Marshall's indispensable primer on FT-ICR is a fantastic starting point[7] or alternatively the less technical more historical review which followed[8]. A more recent mini-review compactly summarises more recent developments[9].

Orbitrap

The Orbitrap is a relatively new type of mass analyser only introduced commercially in 2005 but with roots that can be traced back to the 1920s and the Kingdon trap[10]. In contrast to the ICR cell there is no magnetic field and no applied RF voltage for excitation or detection. The Orbitrap analyser consists simply of a central spindle surrounded by two electrically isolated cup shaped outer electrodes, Figure 105. Ions are injected into the trap through a small specially machined hole in the side of one electrode and at the same time the voltage on the spindle ramped to ensure efficient capture of the incoming ions. This is known as electrodynamic squeezing. The electric field generated between the spindle and outer electrodes coupled with the shape of the electrodes causes the incoming ions to be pushed away from the trap extremities towards the fattest part of the spindle, setting up a harmonic oscillation about the *z* axis given by equation (13), where k is a constant.

$$\omega_z = \sqrt{\frac{k}{m/z}} \qquad (13)$$

A detailed treatment of the equations of motion[11] shows that the frequency of oscillation in the z axis is dependent only on the *m/z* ratio of the trapped ions and also that ions of different *m/z* oscillate back and forth around the spindle as well defined rings. Detection of the oscillating ion packets (rings) is performed without resonant excitation by simply detecting the image current generated on the two outer electrodes. This is much the same as the detection step in FT-ICR and the signal experiences similar dampening, which limits the exploitable detection time. Signal processing is also performed in a similar way through application of a Fourier transform.

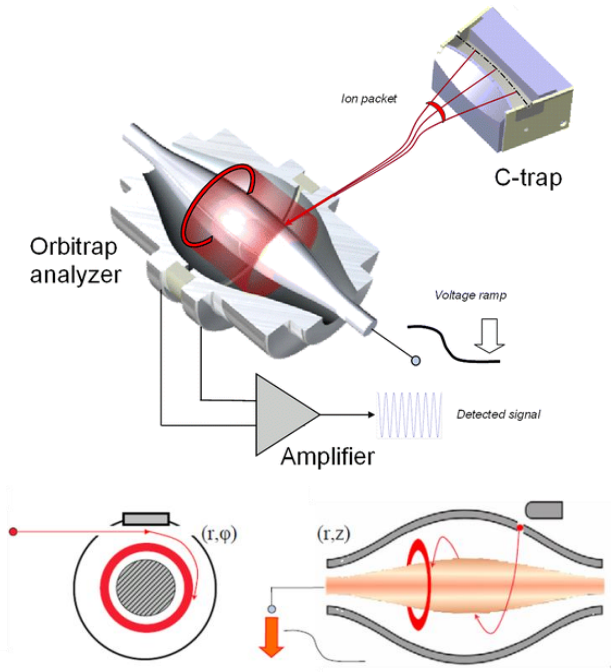


Figure 105 - Schematic of ion injection into an Orbitrap mass anlayser, precession of ions aound the inner spindle and detection of an image current on the outer spindle electrodes. (top) On-axis and side view (bottom)

There have been several generations of commercial Orbitrap platforms containing two generations of Orbitrap analyser (standard and high field). An overview can be found in the recent review by Zubarev and Makarov[12].

2.2. Fragmentation Modes

Vibrational Activation

In vibrational fragmentation modes energy is pumped into a molecule through collision with a solid body (SID), inert gasses (ISD, CAD, HCD) or by low energy IR photons (BIRD, IRMPD). Since this process is ergodic, energy is deposited and partitioned throughout the vibrational modes of the system until there is sufficient internal energy to overcome the activation barrier for bond cleavage. The fragmentation channels open to the molecule depend strongly on the quantity of energy involved and the manner in which it is deposited and on the molecule structure.

In practice this means that different fragmentation modes lead to different types of product ions. Low energy CID (1-100 eV) tends to yield *b* and *y* ions whereas high energy CID (>1 keV) leads also to side chain fragmentation products (*d*, *w*, *v* type ions). The most useful type of fragmentation often involves gradual accumulation of energy within the system rather than depositing it in a single shot. Often this is achieved through multiple low energy collisions with

inert gasses (usually N_2 or Ar in collision cells and He in ion traps) which results in fragmentation of both protein and peptide ions into predominantly *b* and *y* type ions. Another important factor that has been found to influence fragmentation products is the charge state and number of basic amino acids.

Protonation of proteins and peptides occurs on side chains of basic amino acids (K, R H), the N-terminus and the nitrogen or oxygen of the backbone carbonyl in decreasing order of preference. Fragmentation has been found to be charge directed[13] and enhanced cleavage around basic sites has been be explained by detailed studies into the mechanism of CAD. These studies culminate in the mobile proton model outlined by Wysocki[13] and then elaborated and reviewed by Paizs[14]. The mechanism for the formation of *b* and *y* ions is shown in Figure 106 along with one possible pathway for the generation of *a* ions, which may also be found in CAD spectra.

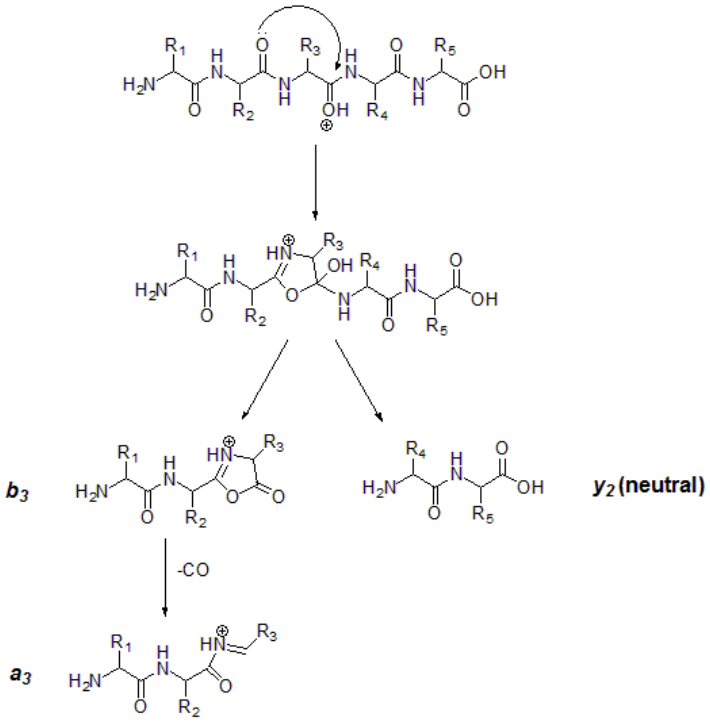


Figure 106 - Formation of *b* and *y* type ions via an oxazolone intermediate

Electronic Activation

In ECD performed on an FT-ICR instrument electrons are produced by applying a voltage to an heated metal alloy electron gun located to the rear of the ICR cell. After trapping, plus a brief relaxation of analyte ions, the gun is fired and electrons guided into the cell by positive potentials on the trapping plate and the lens, which can also act as an energy gate. The radius of electron beam is squeezed by the magnetic field to produce a smaller collimated beam of electrons which

will interact with the trapped positive ion cloud. In most instruments the gun is annular rather than a solid disk in order to make space for a laser that can fire through the electrode. The key experimental parameters are the electron energy and the length of time for which the electron pulse is applied (see following section for details).

In the case of ETD performed on an Orbitrap Velos system, as was mostly the case during this thesis, the ETD reaction occurs in the segmented linear ion trap (LTQ). Analyte ions are first injected, accumulated and "parked" in the front end of the LTQ. The ETD reagent fluoranthene anion is formed at the rear of the instrument after vaporisation of fluoranthene and electron capture. The anion is transferred through the instrument and accumulated in the rear of the LTQ. The potentials in the trap are switched allowing interaction of the analyte and fluoranthene, for a set period of time before excess reagent is ejected from the trap. The nature of the LTQ-Orbitrap limits the number of user tuneable parameters for the ETD reaction. The most important experimental parameters are the reagent ion AGC, that is the quantity of fluoranthene ions accumulated in trap, and the interaction time.

Both ETD and ECD are thought to proceed via similar reaction pathways for which there have been two proposed mechanisms, the Cornell (Figure 107) and Utah-Washington mechanism (Figure 108). Understanding the chemistry involved in ion formation provides insight into experimentally observed phenomena and helps explain why some proteins may fragment differently to others.

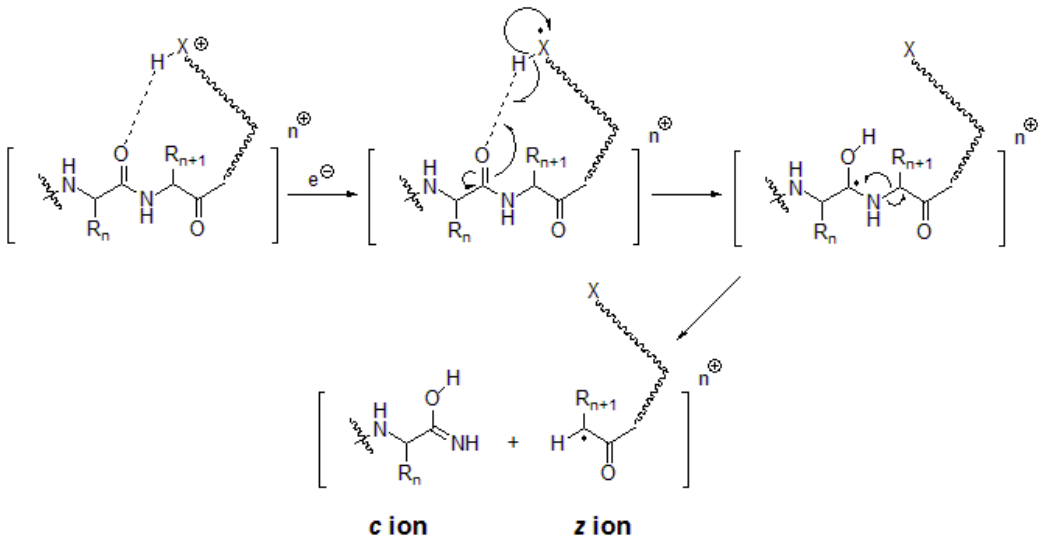


Figure 107 - Cornell mechanism

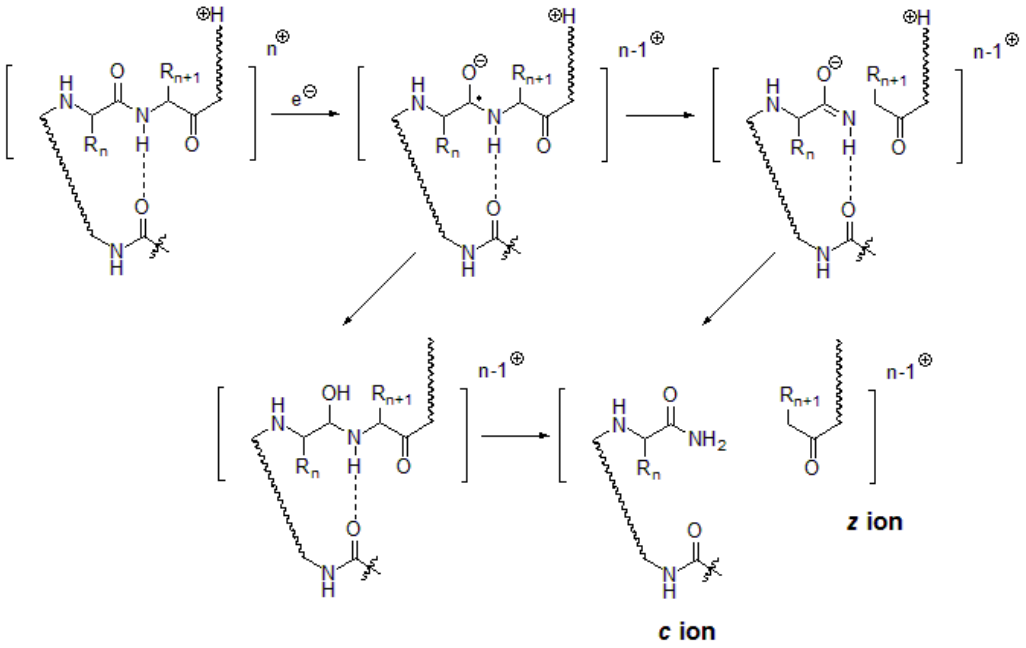


Figure 108 - Utah-Washington mechanism

In the Cornell mechanism this electron is captured by a high energy Rydberg orbital that centres on a positively charged site. Rapid radiative or non-radiative decay to a lower energy orbital then occurs before dissociative recombination and the generation of a free hydrogen atom which may attack nucleophilic sites such as the backbone carbonyl. This is followed by N-C_α bond cleavage which is now more thermodynamically favourable producing *c* and *z* type ions.

In the alternative Utah-Washington mechanism the electron is either quickly transferred to, or captured directly by, a C=O π* orbital forming an amide superbase. Hydrogen transfer is not required in this mechanism and helps explain bond cleavage in the protein backbone far from protonated sites or where H transfer has been artificially blocked. Direct N-C_α bond cleavage then proton transfer, or proton transfer followed by N-C_α bond cleavage also results in the formation of c and z type ions.

2.3. Strategies for Improving Coverage in TDMS

Several strategies can be used to improve sequence overage in TDMS.

1. **Increase Parent Ion Conversion & Fragmentation Efficiency**

The first and simplest strategy is to increase conversion of the parent ion into fragment ions. In collisional type activation this is most often achieved by increasing the collision energy rather than increasing the activation time. Conversely in ECD it is usually achieved by increasing the electron pulse length rather than electron flux and in ETD it is achieved by increasing the time the reagent gas interacts with the analyte ion packet rather than reagent ion density. Care must be

taken when attempting to convert the entire parent however as at higher collision energies or longer interaction times multiple fragmentation events can occur (fragments being fragmented again) and the effects of this may be difficult to predict. In addition some fragmentation products may not be very useful such as charge reduced species in ECD and ETD and immonium ions in CAD. Multiple capture or transfer of electrons can also lead to neutralisation, which is not optimal.

2. Increase Useful Fragmentation

The second technique is related to the first and involves improving fragmentation yield by increasing the efficiency of the fragmentation process. Some precision is required in the terminology used here. Fragmentation efficiency concerns the likelihood that each collision or electron capture results in formation of an exploitable fragment ion; as opposed to charge reduction or an increase in internal energy. If collisions are more likely to give product ions the fragmentation process is evidently more efficient.

For large ions with complex and extensive hydrogen bonding networks electron capture in ETD and ECD can be non-dissociative. This may simply result in charge reduction without bond cleavage. For ETD this has been referred to as "ET no D" (ETnD) and this is a common feature of ECD and ETD spectra. Alternatively it may involve N-C_α cleavage without separation of the daughter ions. It has been found that vibrational activation of ions at the same time or very quickly after electron transfer/capture can improve fragmentation yields by ensuring that if fragmentation occurs, the daughter ion-ion complex is dissociated. For ETD performed in an ion trap such as an LTQ this strategy is widely encountered as the supplementary activation option. In ECD irradiation with low energy IR photons performed during or post ECD can accomplish the same effect and in some studies it is used as a matter of course. For the viral proyl-4-hydroxylase (26 kDa), for example, mass profiling was used to monitor self-oxidation at several time points during a 12 h incubation[15]. Four oxidation sites were visualised by mass profiling. Each were separately isolated accumulated in an external quadrupole before being identified by a combination of IRMPD and in beam AIECD.

3. Use Experimental Conditions Expected to Provide More Fragmentation Channels

The third strategy involves sampling experimental conditions that are expected to produce different fragmentation patterns. It is important to realise that the fragmentation pattern, that is to say the type of ions produced and their relative abundances is something altogether different from fragmentation efficiency. Indeed one could consider a hypothetical case where 100% parent to daughter ion conversion is achieved through a 100% efficient process but only one type of fragment ion is produced. Fragmentation may be complete and very efficient but not particularly

useful! This distinction is important because many experimental conditions result in both a concurrent change in fragmentation efficiency and a change in fragment ion products. The two are often difficult to delineate as in increase in fragmentation efficiency can result in the identification of new fragment ion due purely to improvements in S/N. An often used technique to improve sequence coverage and a good example of this "increased efficiency"-"different fragmentation pattern" duality is fragmentation of elevated protein charge states.

4. Exploit the Different Fragmentation Behaviour of Different Charge States

In the case of ECD and ETD, fragmentation efficiency has been shown to increase with the square of charge state. A similar effect is observed in CAD. This occurs because more highly charged ions are more unfolded due to greater internal coulomb repulsion. They will therefore have a greater collisional/electron capture cross sectional area. For CAD more collisions will occur in a given time period and thus the energy required to fragment the ion from each collision is lower. For ECD and ETD electron capture is more likely and an increase in recombination energy concomitant with an increase in charge state also seems to increase the probability for electron capture to produce fragments rather than merely result in charge reduction. Increasing the charge state of ubiquitin (11 kDa) from 5+ to 13+ has been found to increases the sequence coverage obtained with ECD under the same conditions from 0 to 72%[16].

The origin of this phenomenon is usually explained by the different number of protons present on the protein backbone. Both CAD and ECD proceed through charge directed mechanisms. The location of the protons directly determines the open fragmentation channels. Breuker and McLaffery have spearheaded research into this area at the protein level and have shown that ECD of ubiquitin can essentially be decomposed into constituent charge site spectra that issue from protonation at specific residues on the protein backbone[17].

Increasing charge state does not necessarily mean increased sequence coverage however, as fragmentation channels may be both opened and closed. Rožman and Gaskell showed that for CAD of ubiquitin in a HCT Ultra PTM Discovery ion trap the 9^+ and 10^+ charge states provided the most sequence coverage from a distribution of 4^+ to 13^+[18]. In addition they performed ETD followed by PTR on a number of different proteins from 2.3 kDa to 29 kDa in size. They showed that for most small proteins with one or two charges sequence coverage was greatly improved on the higher charge state, although for coticotopin (4.5 kDa)the intermediate 5^+ provided the most coverage. For larger proteins an increase was observed that seemed to plateau at intermediate charge states and for carbonic anhydrase (29 kDa sequence coverage was similar for charge states 25^+-44^+.

This phenomenon occurs because charge state signatures do not guarantee protein conformations (see Fig 3. In Skinner *et al.*[17]) and protein conformation is a second key factor that determines the available fragmentation channels and thus observed fragmentation ion pattern. Evidence from computational studies of ECD reactions in peptides has found that hydrogen bonding played an important stabilising role during the reaction steps (refs).r group and others have shown that the solvation of charges though hydrogen bonds affects both the predominance of one ECD mechanism over the other (Cornell favoured >80% of the time), the fragmentation channels that are available and the fragmentation ion pattern. However, whilst it is widely accepted that proteins have a gas phase structure and can undergo gas phase folding (and unfolding)[19, 20], the interplay between conformation and proton number and localisation, and how that determines fragmentation at the protein level is currently a topic of debate[21].

5. Change the Gas Phase Conformation

An alternative to exploiting the fragmentation behaviour of different charge state is to change the gas phase conformation of a single charge state and to direct different fragment patterns. Often one refers to unfolding of the protein before fragmentation. Whilst perhaps an oversimplification, since any change in conformation may be concomitant with a change in proton localisation as well as fragmentation efficiency, it has proved useful in obtaining increased sequence coverage. The most commonly encountered experiment of this type is activated-ion ECD (AI-ECD). AI-ECD most commonly combines sub-dissociative IR irradiation for protein unfolding (preactivation) and ECD for fragmentation and is performed in the ICR cell of an FT-ICR instrument. It has proved useful in many studies and is strongly advocated by several groups. There are several varieties of AI-ECD that depend in the geometry of the electron gun and laser. The implementation of this technique can however be rather tricky, requiring optimal alignment of the laser and precise timing to ensure the same packet of ions that is thermally activated will be fragmented by ECD.

6. Mix it up

The sixth option is to simply combine as many sensible experimental conditions, charge states and fragmentation modes as possible to take advantage of the largest number of fragmentation channels.

7. Go Middle-Down

An alternative to the top-down strategy is the so-called middle down approach. Here proteins are digested chemically or with enzymes expected to produce very large fragments, such as the OmpT enzyme which cleaves between two consecutive basic amino acid residues such as KK or RK [22]. This method has the advantage of decreasing the size for very large proteins, which are difficult

to analyse by top-down MS, whilst retaining so of the beneficial features of the top-down approach.

2.4. Application of Strategies to Improve Sequence Coverage on PilE-8013

The solariX instrument is equipped with multiple additional fragmentation modes. Several of these were trialled to see if they could provide any improvement or complementary to the fragmentation already obtained using ECD. The first technique that was investigated was a brute force alternative to the previous ECD experiment.

A Brute Force ECD Approach

The initial experiments performed on myoglobin in chapter 4 revealed a strong correlation between sequence coverage and precursor ion intensity. An alternative strategy for achieving high sequence coverage of PilE-8013 without lengthy parameter optimisation may simply be to choose the most abundant charge state, increase the PI intensity to the maximum that the instrument can handle and perform ECD MS/MS on this very intense parent ion.

This "brute force" approach was tested on the 17^{+} charge state of PilE-8013 which was accumulated to a parent ion intensity of between $3\text{-}4\times10^{8}$ and subjected to ECD MS/MS. Fragmentation was performed with various electron irradiation times and electron energies. The spray stability was monitored throughout acquisition. Spectra were recalibrated internally and ions picked and assigned with the same parameters as the large experiment investigating ECD parameters on the experiment 1×10^{8} intensity PIs. When repeat experiments were performed with the same conditions the result with the highest sequence coverage was retained. The effect of ECD parameters on the central region of the protein between Tyr^{51} and Arg^{111} is shown using a 3D contour plot (Figure 109) and on overall sequence coverage (Figure 110).

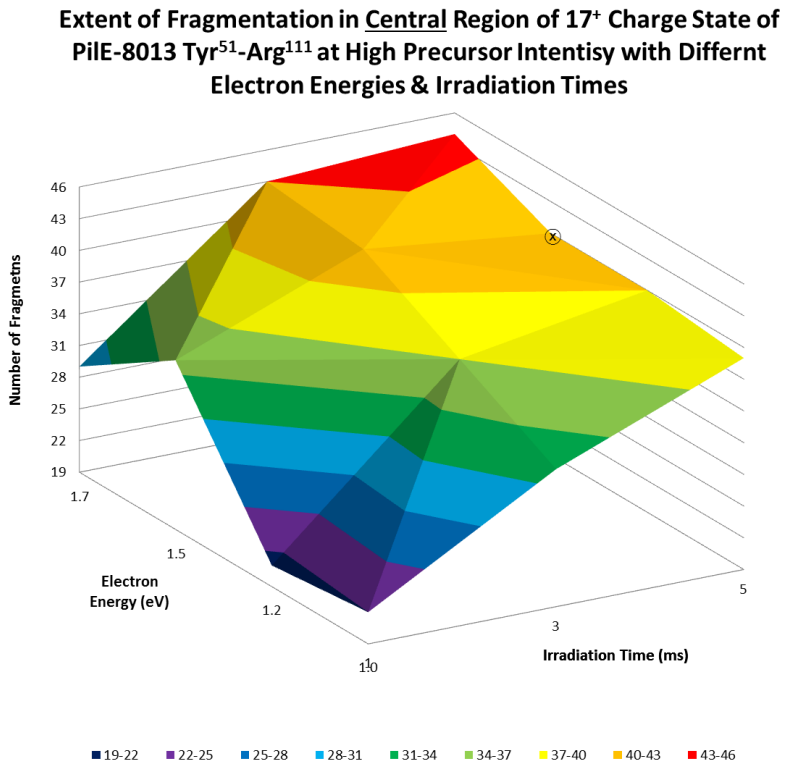


Figure 109 - 3D contour plot of coverage in the central Tyr51-Arg111 region of PilE-8013 as a function of electron irradiation time and energy after ECD fragmentation of the very intense 17+ charge state. ⓧ indicates a dummy point required due to a missing experiment

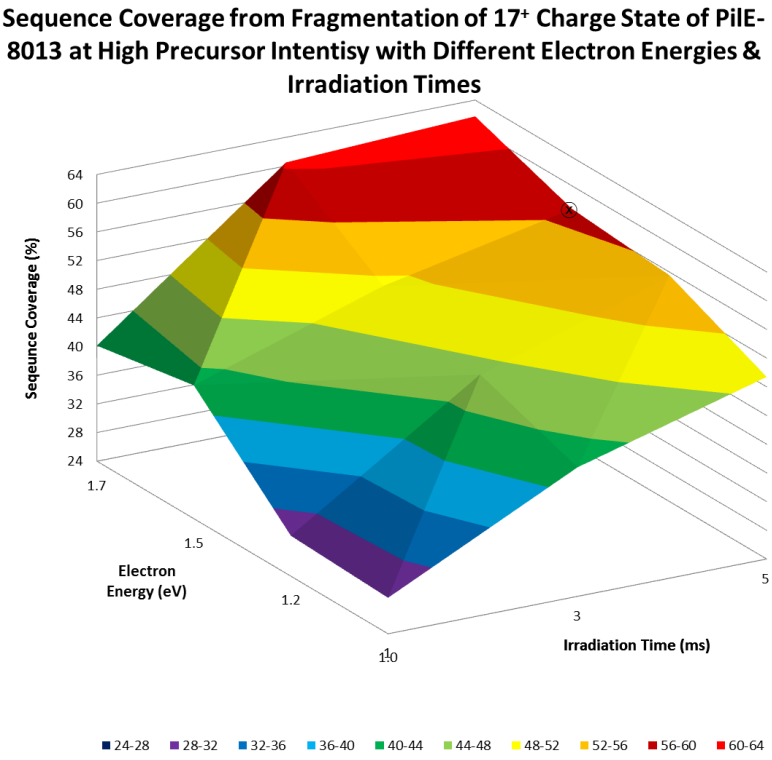


Figure 110 - 3D contour plot of total sequence coverage of PilE-8013 as a function of electron irradiation time and energy after ECD fragmentation of the very intense 17+ charge state. ⓧ indicates a dummy point required due to a missing experiment.

Similar profiles are obtained for this data set when it is treated with respect to coverage in the Tyr^{51}-Arg^{111} region or overall sequence coverage. Similar trends to previous experiments are also observed. Short irradiation times provided favoured increased sequence coverage however very short pulses of 3 ms and 1 ms resulted in less extensive coverage. The 1 ms electron pulse, which is at the limits of the instrument electronics, exhibited far lower levels of parent to daughter ion conversion, presumably due to lower level of electron capture.

In this experiment a greater number of electron energies were trialled and a clear improvement in sequence coverage of PilE-8013 is observed upon increasing the electron energy. The most extensive sequence coverage both overall and in the Tyr^{51}-Arg^{111} region was again provided by 5 ms irradiation with 1.7 eV electrons and a fragmentation map is given in Figure 111.

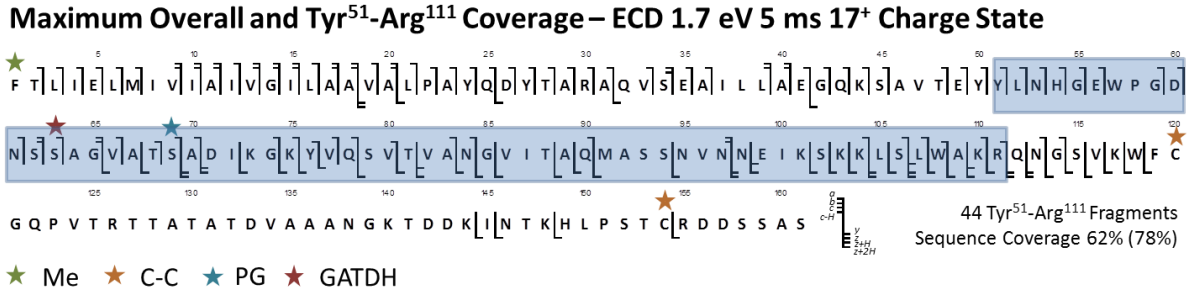


Figure 111 - Fragmentation map of the ECD MS/MS experiment giving both the maximum overall sequence coverage and maximum coverage in the Tyr51-Arg111 region from the brute force approach

Fragmentation of the 17^{+} charge state of PilE-8013 using this brute force approach produces a higher maximum sequence coverage than that achieved in the previous ECD parameter optimisation experiment on the 1×10^{8} intensity, 18^{+} charge state (62% vs 56%). This improvement comes mainly through increased coverage in the Thr^{28}-Thr^{48} region. The number of the fragments issuing from in the central region is the same at 44 and coverage of the areas of poor fragmentation Ser^{63}-Gly^{74} and Met^{91}-Lys^{101} is broadly similar. The "brute force" approach is therefore capable of providing high levels of informative sequence coverage of PilE-8013 but importantly this is highly dependent on optimised fragmentation conditions being used.

ETD

Unlike previous generations of Bruker FT-ICR instrument ETD fragmentation is available on the solariX platform and is performed in the collision cell. The analyte is accumulated in this region for a user defined time. Then the polarity of some of the transfer optics is switched and negatively charged ETD reagent transferred from the CI source into the collision cell. In these experiments fluoranthene was used as the ETD reagent. Once in the collision cell fluoranthene anions are left

to react with the protein analyte ions for a certain time. Both reagent accumulation time and reagent-analyte interaction time are user defined parameters.

After quick parameter optimisation with myoglobin the 18^+ charge state of PilE-8013 was accumulated to a precursor intensity of 1×10^8 and subjected to ETD. Reagent ions were accumulated for 0.185 s and 300 transients summed to produce the final spectrum. Two interaction times of 10 and 15 ms were trialled. The resultant fragmentation patterns were very similar but the longer 15 ms interaction time gave marginally better sequence coverage (Figure 112). Note that peak picking was performed with the same parameters as before although the error tolerance was extended to 7 ppm as internal calibration could not be achieved due to the absence of many of the N-terminal ions.

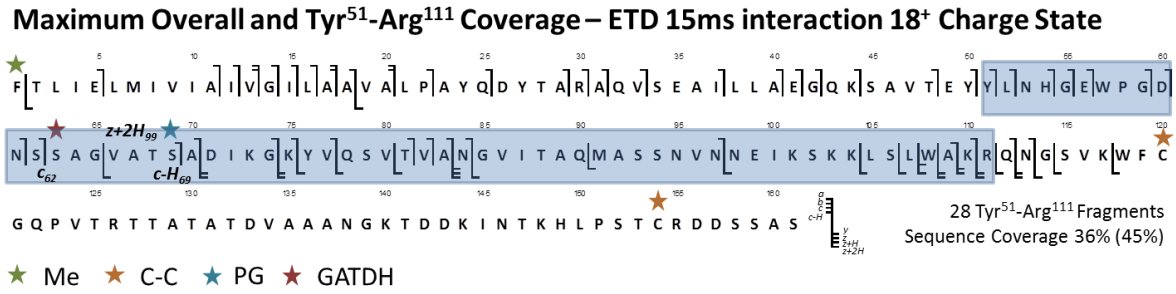


Figure 112 - Fragmentation map from 15 ms ETD FT-ICR MS/MS of the 17+ charge state of the major proteoform of PilE

ETD fragmentation of the 18^+ charge state of PilE-8013 produces a similar pattern of fragments to that obtained in ECD. The coverage was less extensive, particularly in the N-terminal region as was fragmentation between Tyr^{51} and Arg^{111}, however ETD conditions had not been extensively optimised. The c_{62} and $c\text{-}H_{69}$ ions narrow down the area of GATDH and PG modification to the Ser^{63}-Ser^{69} region and the $z+2H_{99}$ ion enables localisation of the glycan to Ser^{63} and of PG to either Thr^{68} or Ser^{69}. These results suggest that with further optimisation ETD could also prove very useful for characterisation of PilE-8013. Experimental conditions were not refined further as fragmentation seemed very similar to ECD which had already been highly optimised and provided confident PTM localisation.

CAD

Collision based fragmentation modes are not as useful for PTM localisation as electronic ones however CAD was performed on PilE-8013 to see if collisional type activation bought any fragmentation complementarity. MS/MS was performed in the collision cell on the 17^+ charge state at voltages ranging from 5-20 eV. The sequence coverage from 20 accumulated transients of 20 eV CAD is shown in Figure 113.

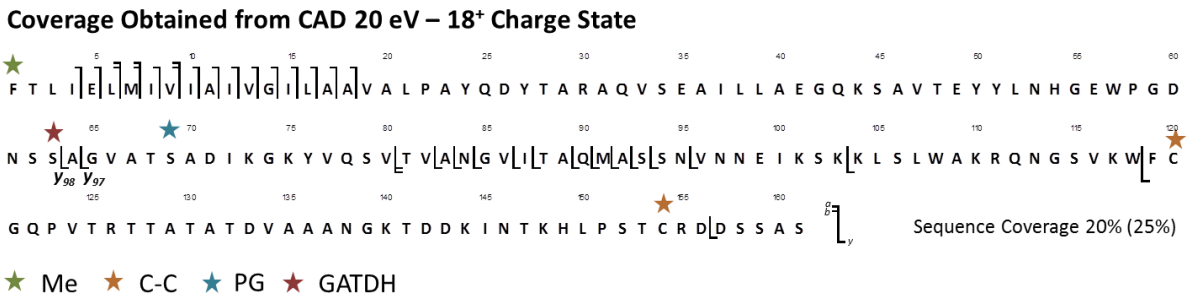


Figure 113 - Fragmentation map from 20 eV CAD FT-ICR MS/MS of the 17+ charge state of PilE-8013

CAD fragmentation of PilE produced many fragments, many of which were found to be large internal ions. Cleavage producing non-internal ions centred around two distinct local regions. A large series of *b* type ions was produced from fragmentation of the N-terminus up to Val^{19} and fragmentation in the central region of the protein between Val^{80}-An^{95} produced a large series of *y* type ions. Fragmentation occurs in regions of the protein that are devoid of basic amino acid residues and in the case of the Val^{80}-An^{95} region, proved incredibly complementary to that provided by ECD. This mirrors similar CAD/ExD complementarity made by the Zubarev group on the preference of inter residue cleavage[23, 24] and on membrane proteins by both the Kelleher[25] and Whitelegge groups[26].

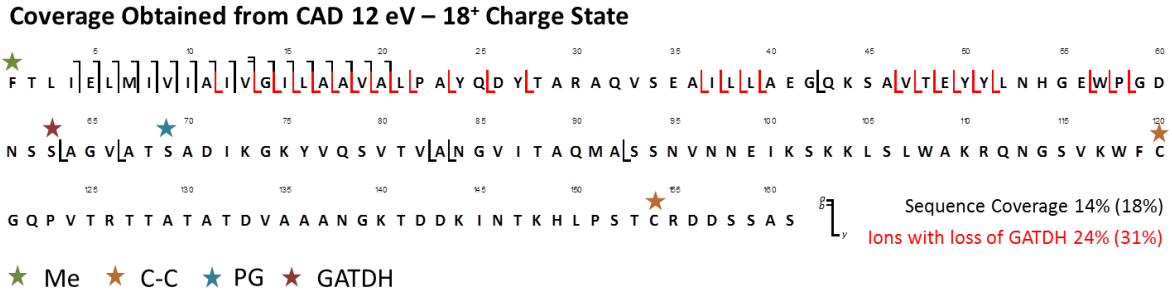


Figure 114 - Fragmentation map from 12 eV CAD FT-ICR MS/MS of the 17+ charge state of PilE-8013. Ions exhibiting loss of GATDH are highlighted in red

The utility of this method for PTM localisation in PilE-8013 is definitely rather limited in the case of the glycan, however in previous CAD MS/MS experiments performed on a Q-ToF instrument the PG group has proved to be more resilient to collisional fragmentation. This may explain the presence of the y_{97} and y_{98} observed here. Examining the spectral assignment from CAD performed at 12 eV (Figure 114) provides clear evidence that the PG group may be retained on the protein backbone. When assignment is performed without the GATDH modification (this situation mirrors glycan loss though collision activation) a large series of *y* type ions that still bear the PG group may still be assigned and are highlighted in red. This suggests that CAD fragmentation at

12 eV is high enough in energy to cause cleavage of the protein-glycan bond but not trigger the McLafferty type β-elimination that leads to PG loss. Together these observations suggest that CAD provides complementary fragmentation of PilE-8013 to ECD and ETD that it may be used to increase sequence coverage and in particular to identify the position of PG groups.

The next two strategies to increase sequence coverage were both investigated using the model protein myoglobin.

Activated ion ECD (AI-ECD)

AI-ECD has been previously used to improve the ECD spectra of large proteins by unfolding prior to ECD fragmentation[27]. It can also be used on smaller proteins such as PilE and even peptides to provide different fragmentation patterns and sometimes improved ExD coverage[28]. There are several ways to activate ions prior to ECD and a popular method is using infrared (IR) photons produced by a CO_2 laser. In a slight modification to the usual ECD MS/MS experiment ions are trapped in the ICR cell, irradiated with IR photons and then ECD is performed on this activated ion packet.

The experiment proceeds in distinct stages, depicted in Figure 115. Ions are injected into the ICR cell at time t_0 and after a short relaxation time (T_{Delay1}) they are subjected to IR activation for time T_{IR}. There is then another delay of time T_{Delay2} after which ECD is performed with an irradiation time of T_{ECD}. There is a final short delay before the excitation-detection phase is started at t_5. The reasons for this IR-delay-ECD pulse sequence concern the alignment of the laser, the electron gun and the motion of the ions within the cell.

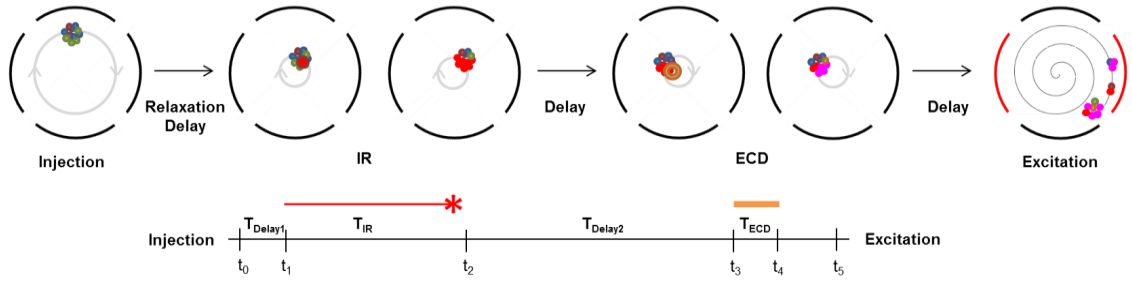


Figure 115 - Schematic of the stages in the Ai-ECD experiment

The solariX instrument is equipped with a hollow dispenser ECD cathode which is annular in shape. This produces a ring of electrons which are squeezed by the magnetic field into a beam (or smaller ring) as they are guided into the ICR cell by the ECD lens and positive potential on the back trapping plate. The position of the cathode is fixed and cannot be altered. The shape of the cathode means that an IR laser can be placed “on axis” and shot directly through the cathode centre. The

laser beam must be positioned to ensure good overlap of the beam and the ions trapped in the ICR cell.

This arrangement often means that the ECD gun and IR laser shoot in different places, both slightly off-centre of the ICR cell (Figure 116).

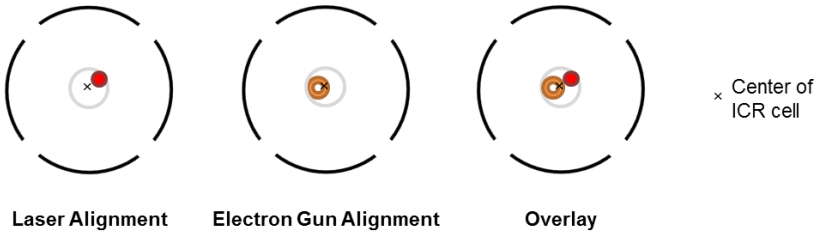


Figure 116 - Hypothetical position of the electron and IR laser beams with respect to ICR cell

The trapped analyte ions are not static. They are oscillating backwards and forwards in along the axis of the ICR cell (Into and out of the page) and they experience a magnetron motion that leads to radial procession. Because of this and the different positions of the laser and cathode T_{Delay2} must be properly tuned. For efficient AI-ECD the same packet of processing ions must be activated by IR and fragmented by ECD.

In the AI-ECD experiment the first parameter to optimise is laser power and pulse duration. Ideally the pulse must be as short as possible to minimise the duty cycle and impart as much energy as possible without triggering dissociation. For the laser attached to this instrument 80-90% power was used and only the pulse length optimised. All of these optimisation experiments were performed with myoglobin. After laser alignment, the effect of T_{IR} on precursor fragmentation was investigated. The 19^+ ion of myoglobin was subjected to IR activation for different values of T_{IR}. The precursor ion intensity at T_{IR} = 0 was monitored periodically throughout the experiment to ensure that it remained stable. For each value of T_{IR} the percentage fragmentation was calculated by plotting the observed parent ion intensity against the parent ion intensity at T_{IR} = 0 (Figure 117). Each intensity value represents the sum of 15 transients.

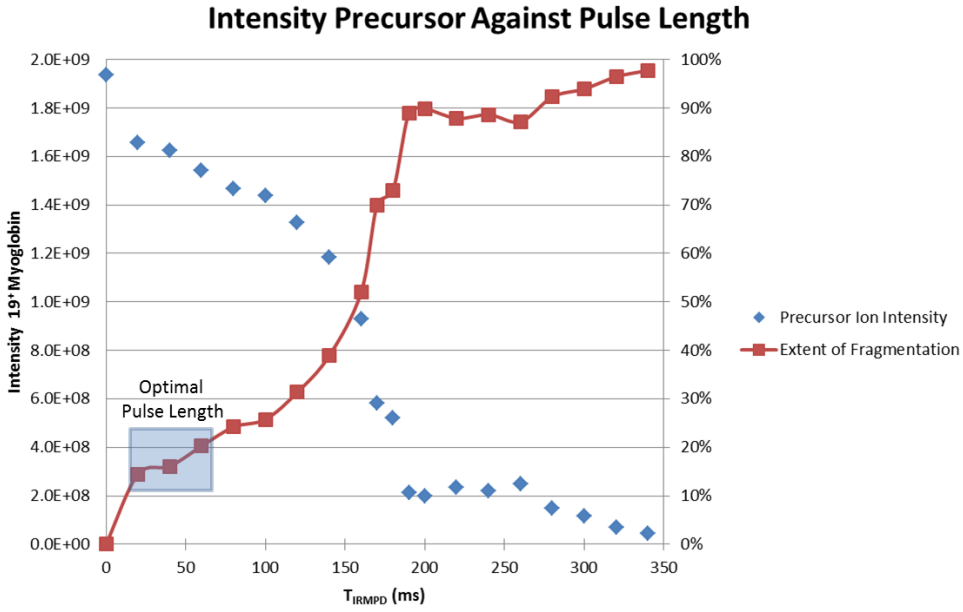


Figure 117 - Precursor ion conversion against IR pulse length

After a large increase in fragmentation between 0 and 10 ms irradiation, the extent of fragmentation increased slowly and linearly until 120s at which point it again increased more rapidly. Almost complete fragmentation of the precursor was achieved at 180 ms irradiation. A pulse length of between 40 and 60 ms was therefore considered short enough to increase internal energy of ions but not long enough to cause deleterious extensive precursor ion fragmentation.

The next parameter to optimise is the delay time between the IR pulse and ECD, T_{Delay2} The general approach here is iterative modification of T_{Delay2} for a given value of T_{ECD}, T_{Delay1} and T_{IR} until ECD fragmentation is maximised. T_{Delay2} must be then re-optimised if either T_{Delay1} or T_{IR} are changed. Trapped ions have previously been shown to undergo m/z independent magnetron motion that causes a radial precession about the ICR cell axis[29]. Correct timing of IR activation and ECD is therefore crucial to the success of the AI-ECD experiment[30]. It was thought that calculation of the magnetron motion period might enable a more rational approach for pulse length optimisation to be developed.

The magnetron motion period was calculated by measuring the intensity of the 19^+ precursor after a short ECD pulse (T_{ECD} = 10ms) and varying T_{Delay1} with no IR irradiation or Delay2 (T_{Delay2} = 0 and T_{IR} = 0). Lower precursor ion intensity indicates more extensive fragmentation and is indicative of overlap of the electron beam and the protein ion packet. The results of this experiment are shown in Figure 118 where the intensity is normalised to the maximum intensity achieved when ECD was performed as soon as the ions had entered the cell (T_{Delay1} = 0).

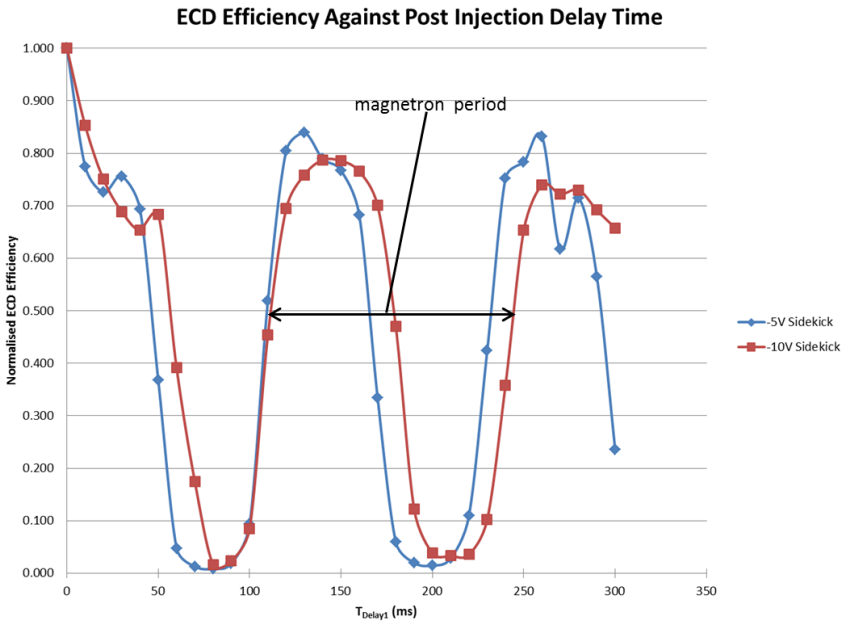


Figure 118 - Extent of ECD fragmentation of 19+ parent ion of myoglobin as a function of TDelay1. The effect of magnetron motion is clearly visible

The oscillation of percentage precursor ion conversion with respect to time is shown at two different sidekick voltages. The oscillation period can be estimated at ≈130 ms and provides a estimation to the magnetron procession frequency. This value correlates very well with the data acquired from varying the IR pulse length. In that experiment the sharp increase in IR induced dissociation between 120 and 180 ms could be due to further activation of the initial ion packet once it had completed its first period and was subject to reactivation.

Now a reasonable pulse length and the magnetron period had been calculated the approach was ready for extensive optimisation on myoglobin. It seemed that 60 ms IR could be followed by ECD in the 100 to 200 ms window. This would mean T_{Delay2} should be in the 40-140 ms time range. The success of this approach would hinge on the assumption that the same region of the ion packet is activated by both IR and ECD.

Investigations into the AI-ECD experiment were halted at this point due to the time involved in experimental optimisation and the fact that fragmentation using ECD provided sufficient coverage for PTM localisation, however all of the necessary preliminary experiments have been performed if the technique is required in future investigations. Indeed the AI-ECD approach merits further investigation if other pilins do not provide such extensive ECD fragmentation. Other AI-ECD techniques such as simultaneous and post-AI-ECD could also be trialled.

Supercharging to Access Higher Charge States

Supercharging is a method of increasing the charge of electrosprayed ions by adding low basicity compounds to the electrospray solution. This often results in the production of a higher maximum charge state and concomitant skewing of the charge state distribution (CSD) towards more highly charged ions (lower *m/z*). It was thought that increasing the abundance of highly charged PilE ions might facilitate top-down data acquisition by shortening experiment times. It may also provide access to more highly charged precursors which could produce different fragmentation patterns.

Implementation of the supercharging experiment was first attempted the model protein myoglobin and the supercharging reagent sulfolane at various concentrations from 1-100 mM. Whilst concentrations of 10 and 100 mM created new maximum charge states and significantly narrowed the observed CSD, an unfortunate side effect of supercharging using sulfolane was significant source pollution even at low concentrations. This required deep cleaning of the source region and transfer lenses before passing the machine to the next user. The approach was therefore not pursued any further. It may however be worth retrying with or other supercharging reagents such as *m*-nitrobenzyl alcohol (*m*-NBA)[31, 32] although this was not attempted for fear of similar source pollution from this similarly low volatility compounds.

3. Electron Microscopy

Electron microscopy was performed as detailed in the published article inserted into chapter 7. For pili preparations no fixation step was performed and pili were directly treated with uranyl acetate. Further details are given in the main text of the manuscript.

Bibliography

[1] J. Chamot-Rooke, G. Mikaty, C. Malosse, M. Soyer, A. Dumont, J. Gault, A. F. Imhaus, P. Martin, M. Trellet, G. Clary, P. Chafey, L. Camoin, M. Nilges, X. Nassif and G. Dumenil.Posttranslational Modification of Pili upon Cell Contact Triggers N. meningitidis Dissemination. *Science*, **2011**, *331*, 778.

[2] M. C. Gault J., Duménil G., Chamot-Rooke J.,A Combined Mass Spectrometry Strategy for Complete Posttranslational Modification Mapping of *N. meningitidis* Major Pilin. *Journal of Mass Spectrometry*, **2013**, Accepted.

[3] E. O. Lawrence and M. S. Livingston.The Production of High Speed Light Ions Without the Use of High Voltages. *Physical Review*, **1932**, *40*, 19.

[4] I. J. Amster.Fourier transform mass spectrometry. *Journal of Mass Spectrometry*, **1996**, *31*, 1325.

[5] D. P. A. Kilgour, R. Wills, Y. Qi and P. B. O'Connor.Autophaser: An Algorithm for Automated Generation of Absorption Mode Spectra for FT-ICR MS. *Analytical Chemistry*, **2013**, *85*, 3903.

[6] D. P. A. Kilgour, M. J. Neal, A. J. Soulby and P. B. O'Connor.Improved optimization of the Fourier transform ion cyclotron resonance mass spectrometry phase correction function using a genetic algorithm. *Rapid Communications in Mass Spectrometry*, **2013**, *27*, 1977.

[7] A. G. Marshall, C. L. Hendrickson and G. S. Jackson.Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrom. Rev.*, **1998**, *17*, 1.

[8] A. G. Marshall.Milestones in Fourier transform ion cyclotron resonance mass spectrometry technique development. *International Journal of Mass Spectrometry*, **2000**, *200*, 331.

[9] A. G. Marshall, C. L. Hendrickson, M. R. Ernmetta, R. P. Rodgers, G. T. Blakney and C. L. Nilsson.Fourier transform ion cyclotron resonance: state of the art. *European Journal of Mass Spectrometry*, **2007**, *13*, 57.

[10] K. H. Kingdon.A Method for the Neutralization of Electron Space Charge by Positive Ionization at Very Low Gas Pressures. *Physical Review*, **1923**, *21*, 408.

[11] A. Makarov.Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*, **2000**, *72*, 1156.

[12] R. A. Zubarev and A. Makarov.Orbitrap Mass Spectrometry. *Analytical Chemistry*, **2013**, *85*, 5288.

[13] V. H. Wysocki, G. Tsaprailis, L. L. Smith and L. A. Breci.Special feature: Commentary - Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry*, **2000**, *35*, 1399.

[14] B. Paizs and S. Suhai.Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.*, **2005**, *24*, 508.

[15] Y. Ge, B. G. Lawhorn, M. Elnaggar, S. K. Sze, T. P. Begley and F. W. McLafferty.Detection of four oxidation sites in viral prolyl-4-hydroxylase by top-down mass spectrometry. *Protein Science*, **2003**, *12*, 2320.

[16] K. Breuker, H. B. Oh, D. M. Horn, B. A. Cerda and F. W. McLafferty.Detailed unfolding and folding of gaseous ubiquitin ions characterized by electron capture dissociation. *J. Am. Chem. Soc.*, **2002**, *124*, 6407.

[17] O. S. Skinner, K. Breuker and F. W. McLafferty.Charge Site Mass Spectra: Conformation-Sensitive Components of the Electron Capture Dissociation Spectrum of a Protein. *Journal of the American Society for Mass Spectrometry*, **2013**, *24*, 807.

[18] M. Rozman and S. J. Gaskell.Charge state dependent top-down characterisation using electron transfer dissociation. *Rapid Communications in Mass Spectrometry*, **2012**, *26*, 282.

[19] K. Breuker and F. W. McLafferty.Stepwise evolution of protein native structure with electrospray into the gas phase, 10(-12) to 10(2) S. *Proceedings of the National Academy of Sciences of the United States of America*, **2008**, *105*, 18145.

[20] O. S. Skinner, F. W. McLafferty and K. Breuker.How Ubiquitin Unfolds after Transfer into the Gas Phase. *Journal of the American Society for Mass Spectrometry*, **2012**, *23*, 1011.

[21] Z. Hall and C. V. Robinson.Do Charge State Signatures Guarantee Protein Conformations?, *Journal of the American Society for Mass Spectrometry*, **2012**, *23*, 1161.
[22] C. Wu, J. C. Tran, L. Zamdborg, K. R. Durbin, M. Li, D. R. Ahlf, B. P. Early, P. M. Thomas, J. V. Sweedler and N. L. Kelleher.A protease for 'middle-down' proteomics. *Nature Methods*, **2012**, *9*, 822.
[23] M. M. Savitski, F. Kjeldsen, M. L. Nielsen and R. A. Zubarev.Complementary sequence preferences of electron-capture dissociation and vibrational excitation in fragmentation of polypeptide polycations. *Angewandte Chemie-International Edition*, **2006**, *45*, 5301.
[24] R. A. Zubarev, A. R. Zubarev and M. M. Savitski.Electron capture/transfer versus collisionally activated/induced dissociations: Solo or duet?, *Journal of the American Society for Mass Spectrometry*, **2008**, *19*, 753.
[25] A. D. Catherman, M. Li, J. C. Tran, K. R. Durbin, P. D. Compton, B. P. Early, P. M. Thomas and N. L. Kelleher.Top Down Proteomics of Human Membrane Proteins from Enriched Mitochondrial Fractions. *Analytical Chemistry*, **2013**, *85*, 1880.
[26] V. Zabrouskov and J. P. Whitelegge.Increased coverage in the transmembrane domain with activated-ion electron capture dissociation for top-down Fourier-transform mass spectrometry of integral membrane proteins. *J. Proteome Res.*, **2007**, *6*, 2205.
[27] D. M. Horn, Y. Ge and F. W. McLafferty.Activated ion electron capture dissociation for mass spectral sequencing of larger (42 kDa) proteins. *Analytical Chemistry*, **2000**, *72*, 4778.
[28] A. R. Ledvina, N. A. Beauchene, G. C. McAlister, J. E. P. Syka, J. C. Schwartz, J. Griep-Raming, M. S. Westphall and J. J. Coon.Activated-Ion Electron Transfer Dissociation Improves the Ability of Electron Transfer Dissociation to Identify Peptides in a Complex Mixture. *Analytical Chemistry*, **2010**, *82*, 10068.
[29] Y. O. Tsybin, C. L. Hendrickson, S. C. Beu and A. G. Marshall.Impact of ion magnetron motion on electron capture dissociation Fourier transform ion cyclotron resonance mass spectrometry. *International Journal of Mass Spectrometry*, **2006**, *255*, 144.
[30] V. A. Mikhailov and H. J. Cooper.Activated Ion Electron Capture Dissociation (AI ECD) of Proteins: Synchronization of Infrared and Electron Irradiation with Ion Magnetron Motion. *Journal of the American Society for Mass Spectrometry*, **2009**, *20*, 763.
[31] S. H. Lomeli, I. X. Peng, S. Yin, R. R. O. Loo and J. A. Loo.New Reagents for Increasing ESI Multiple Charging of Proteins and Protein Complexes. *Journal of the American Society for Mass Spectrometry*, **2010**, *21*, 127.
[32] S. G. Valeja, J. D. Tipton, M. R. Emmett and A. G. Marshall.New Reagents for Enhanced Liquid Chromatographic Separation and Charging of Intact Protein Ions for Electrospray Ionization Mass Spectrometry. *Analytical Chemistry*, **2010**, *82*, 7515.

Abstract

Top-down mass spectrometry (TDMS) is an alternative protein characterisation strategy to the more widespread bottom-up (BU) approach. TDMS has the unique ability to fully characterise the variety of protein products expressed by the cell (proteoforms), including those bearing posttranslational modifications (PTMs). In this thesis TDMS has been developed on both FT-ICR and Orbitrap mass spectrometers for the analysis of type IV pili (T4P). This includes the first T4P to be visualised in a Gram positive bacterium (*Streptococcus pneumoniae*).

T4P are filamentous, extracellular organelles primarily composed of a single protein subunit or major pilin that can be highly posttranslationally modified. For the major pilin, PilE, of the human pathogen *Neisseria meningitidis* (Nm), TDMS was extensively optimised and the first complete characterisation of all proteoforms of PilE from a single Nm strain performed. A biological role has been proposed for the enigmatic phosphoglycerol PTM.

The approach was extended and applied in the first large scale study of PTMs on PilE from uncharacterised, pathogenic strains of Nm. Comparison of the TD and BU methodologies revealed both their complementarity and the inherent weakness of the BU approach for full proteoform characterisation. TDMS was combined with other structural techniques to reveal that pilins from the previously unstudied class II isolates of Nm are extensively glycosylated and that glycosylation is both driven by the primary structure of PilE and has a profound effect on pilus surface topology. These observations have been used to offer the first explanation of how T4P expressed by class II isolates of Nm avoid immune detection.

Résumé

La caractérisation de protéines par spectrométrie de masse «top-down» (TDMS) possède plusieurs avantages sur l'approche «bottom-up» (BU), notamment pour la caractérisation des protéoformes; c'est-à-dire l'ensemble des protéines exprimées par la cellule, y compris celles portant les modifications post-traductionnelles (MPT). Dans cette thèse la TDMS a été développée pour l'analyse des pili de type IV (T4P) à la fois sur Orbitrap et sur FT-ICR.

Les T4P sont des organelles extracellulaires composées principalement d'une protéine, la piline majeure, qui peut être fortement décorée par des MPT. Pour la piline majeure PilE, du pathogène humain Neisseria meningitidis (Nm), la TDMS a été optimisée pour obtenir la première caractérisation de toutes les protéoformes de PilE exprimées par une souche de référence et un rôle biologique a été proposé pour la MPT glycérophosphate.

De plus la TDMS a été appliquée pour une étude à plus grande échelle des MPT des pili provenant des souches cliniques de Nm non-caractérisées. La comparaison des méthodologies TD et BU a révélé à la fois leur complémentarité et la faiblesse inhérente à l'approche BU pour la caractérisation complète de protéoformes. En combinaison avec d'autres techniques structurales, il a été montré que les pili exprimés par les souches de Nm de classe II sont fortement glycosylés, que la glycosylation est dirigée par la séquence de PilE, et que ces sucres modifient fortement la surface des fibres de pili. Ces observations ont amené à une hypothèse nouvelle sur le mécanisme d'évasion immunologique des T4P de classe II chez Nm. De plus la première caractérisation d'une T4P exprimée par une bactérie Gram positif a été réalisée.