NOVEL GLYCAN-TARGETED EXTRACELLULAR PROTEASES FROM DIVERGENT MUCOSAL MICROBES

Didier A. NDEH

A Thesis Submitted for the Degree of Doctor of Philosophy

2010 – 2013

Institute for Cell and Molecular Biosciences

Newcastle University

Acknowledgements

I would like to express my sincere gratitude to my supervisors Professor Robert Hirt and Dr David Bolam, for their unwavering support throughout the course of my PhD studies. I have benefited a lot from your knowledge, experience, generosity and all the words of encouragement. You both have been very instrumental to my success and there are so many wonderful attributes about you worth emulating. Life just made it possible that I should meet the right people, with the right mind-set, in the right university, at the right time. What more could I have asked for?

I would also like to thank Professor Martin Embley, Professor Harry Gilbert and Dr Timothy Cheek for every role they played and more importantly their desire to support me even after my PhD studies. This reinforced my confidence and gave me a huge sense of security and optimism for the future.

The six hour discussion with my assessors Professor Colin and Dr Natalie was very enlightening as they both showed great insight in my research area. For this I wish to offer a big thank you to them and to say I found their suggestions for corrections to my thesis and ideas for future work very useful.

I cannot think of any other practical way I would have completed my dissertation without the financial support I received from the joint effort of the Cameroonian government, the UK department for international development (DFID) and Newcastle University through the UK Commonwealth scholarships. This support has been phenomenal and I will for ever be grateful for this life changing opportunity.

The positive spirit of my parents, siblings and external relations has never been overlooked. It is worth thanking God for the wonderful relationship we share as a family and I would like to dedicate this to us all. So this if for you; Frunwi Mugri, Nchang Fri, Enni Akwa, Bih Gladys Taboh Divine, Ma Tekum Debora and my friend and Dad Mr Ndeh Joseph.

Finally, I have had a wonderful time with all current and past members of Labs M2035 and M2041 to whom am very grateful for all their contributions. These include (in any order) Sirintra Nakjang, Matthew Collison, Paul Dean, ME Geggie, Alina Goldberg Cavalleri, Ekaterina Kozhevnikova, Kacper Sendra, Valentina Margarita, Andrew Watson, Tom Williams, Sarah Heaps, Svetlana Cherlin, Peter Major, Elisabeth Lowe, Sarah Shapiro, Jose Munoz, Jonathon Briggs, Max Temple, Ana Luis, Lucy Crouch, Fiona Cuskin, Artur Rogowski, Adam Jackson, Xiaoyang Zhang, Alan Cartmell, Kate Cameron, Maria Francesca and Immacolata Venditto.

Abstract

Trillions of microorganisms inhabit mucosal surfaces of the human body. Despite increasing evidence of their impact on human health, many of the molecular mechanisms underlying host-microbial interactions (HMI) are poorly understood. To contribute to our understanding of HMI at mucosal surfaces, we investigated the novel family of M60-like/PF13402 domain-containing proteins and their putative functional partners.

M60-like domains are shared by proteins from several mucosal microbes including two important human mucosal microbes; the bacterial mutualist *Bacteroides thetaiotaomicron* and the protist pathogen *Trichomonas vaginalis*, suggesting these proteins are important for interaction with the mucosal layer. We initially tested our hypothesis that these are glycoprotein-targeted metal dependent proteases in both these organisms. The three M60-like domains of *B. thetaiotaomicron* proteins (BT4244, BT3015 and BT4272) exhibited mucin protease activity. This proteolytic activity was shown to be inhibited in a mutant version of the protein (BT4244-FL-E575D) as well as in the presence of Ethylenediaminetetraacetic acid (EDTA), implying BT4244 and its relatives are metal dependent proteases. All M60-like proteins from *B. thetaiotaomicron* contained a carbohydrate binding module (CBM) from family 32 and these were shown to be capable of binding galacto-configured sugars that are common to mucin glycans, while in contrast the putative carbohydrate binding PA14 domain of the *T. vaginalis* TVAG339720 M60-like protein interacted with heparin and its sulphated derivatives. Mucins are glycoproteins and prominent components of the mucus secreted at mucosal surfaces while heparin is a close relative of heparan sulphate which typically exists as part of proteoglycans in the glycocalyx of mucosal epithelia. Although the actual target of the M60-like domain of TVAG339720 and its relatives in *T. vaginalis* are not currently known, the interaction of the TVAG339720 PA14 domain with heparin suggests that these may be proteases targeting proteoglycans and play a role in adhesion of the pathogen to the epithelial layer, a key initial step in pathogenesis.

M60-like domain-containing proteins of *B. thetaiotaomicron* are also components of Sus-like systems. Sus-like systems are Bacteroidetes specific machinery that comprise a suite of cell-envelope located carbohydrate-active enzymes and sugar binding proteins that target complex glycans, with each Sus-like system tuned to the degradation of a specific glycan. The Sus-like system containing the BT4244 enzyme (BT4240-50), encoded by the polysaccharide locus (PUL) PULBT_4240-50 was characterised in this study. The results demonstrated that BT4244 is a surface protein and that its proteolytic activity is part of a concerted action of BT4240-50 components to utilise complex mucin glycoproteins containing the T (Galβ1-3GalNAc) and F (GalNAcα1-3GalNAc) antigens. Gene deletion studies revealed that PULBT_4240-50 provides a competitive advantage to the organism when grown on mucins, probably through its possession of the N-acetylgalactosamine (GalNAc) kinase BT4240, which was shown to be crucial for GalNAc utilisation. Finally, although variably conserved in closely related Bacteroides, the high frequency of PULBT_4240-50 components in this group of organisms suggests it may be an important evolutionary adaptation for survival at mucosal surfaces. Our findings not only set the stage for future functional studies on the novel M60-like/PF13402 family of proteins and their functional partners, but also further our understanding of host-microbial interactions at mucosal surfaces.

Content

List of Tables

List of Figures

Abbreviations

ΔBT_3015	*B. thetaiotaomicron* mutant with a deletion to the BT_3015 gene
ΔBT_4240	*B. thetaiotaomicron* mutant with a deletion to the BT_4240 gene
ΔBT_4242	*B. thetaiotaomicron* mutant with a deletion to the BT_4242 gene
ΔBT_4244	*B. thetaiotaomicron* mutant with a deletion to the BT_4244 gene
ΔBT_4272	*B. thetaiotaomicron* mutant with a deletion to the BT_4272 gene
ΔBT_3015ΔBT_4244ΔBT_4272	*B. thetaiotaomicron* mutant with a deletion to the BT_3015, BT_4244 and BT_4272 genes
ΔPULBT_4240-50	*B. thetaiotaomicron* mutant containing a deletion to all genes encoding the BT4240-50 Sus –like system
BT4240-50	Sus-like system encoded by genes starting from BT_4240 chronologically to BT_4250
A280nm	Absorbance at a wavelength of 280 nm
A600nm	Absorbance at a wavelength of 600 nm
A340nm	Absorbance at a wavelength of 340 nm
A420nm	Absorbance at a wavelength of 420 nm
AGE	Agarose gel electrophoresis
Asn	Asparagine
BACON	Bacteroidetes-Associated Carbohydrate-binding Often N-terminal CBM
BLAST	Basic Local Alignment Search Tool
BSM	Mucin from bovine submaxillary glands (Type I-S)
CAZy	Carbohydrate Active Enzyme server
CBM	Carbohydrate-binding module
CBM32	Family 32 carbohydrate-binding modules
CFE	Cell-free extract
CS	Chondroitin sulphate
DEPC	Diethyl pyrocarbonate
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded DNA
DUF	Domain of unknown function
EDTA	Ethylenediaminetetraacetic acid
FT	Flow-through sample during IMAC
Fuc	Fucose
Fus	Fructan utilization system
Gal	Galactose
GalNAc	N-acetyl-D-galactosamine
GalNAcα1-3GalNAcα	Forsmann disaccharide (F antigen)
GalNAcα-Ser/Thr	Tn antigen
Galβ1-3GalNAcα	Galacto-N-biose (T antigen)
Galβ1-3GalNAcα-Ser/Thr	Galactosyl Tn antigen
Galβ1-3GlcNAcα	Lacto-N-biose (LNB)
GBDL	Galactose-binding domain-like
GH	Glycoside hydrolases
Glc	D-Glucose
GlcNAc	N-acetyl-D-glucosamine
HAA	*Helix aspersa* (garden snail) agglutinin
HEPES	4-(2-Hydroxyethyl) piperazine-1-ethanesulfonic acid
His_6-tag	Polyhistidine tag
HMO	Human milk oligosaccharide
HPLC	High-performance liquid chromatography
IEC	Ion-exchange chromatography
IgA	Immunoglobulin A
IgA	Immunoglobulin A

IMAC	Immobilized metal ion affinity chromatography
IPTG	Isopropyl-β-D-thiogalactopyranoside
ITC	Isothermal titration calorimetry
KEGG	Kyoto Encyclopedia of Genes and Genomes
KOD	*Thermococcus kodakaraensis* DNA polymerase enzyme
MDM	Modified diamond's medium
MEROPS	On-line database for peptidases
MUC	Denotes human mucin gene
muc	Denotes mouse mucin gene
NAD+	Nicotinamide adenine dinucleotide (oxidized)
NADH	Nicotinamide adenine dinucleotide (reduced)
NC	Nitrocellulose membrane
NeuAc	N-Acetylneuraminic acid
Nus Tag™	Tag containing amino acid sequence of *E.coli* NusA protein
OD600nm	Optical Density at 600nm
PA14	CBM named after a ~14 kDa domain in the anthrax Protective Antigen
PAS	Periodic acid-Schiff
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PFAM	On-line database of protein families
PGMII	Porcine gastric mucin type II (bound sialic acid, ~1%)
PGMIII	Porcine gastric mucin type III (bound sialic acid, 0.5-1.5%)
*p*NP	Paranitrophenol
PTS – region/domain	Region/domain containing high amounts of Proline, Serine and Threonine amino acids
PUL	Polysaccharide utilization locus
PVDF	Polyvinyl difluoride membrane
qPCR	Quantitative polymerase chain reaction
SAGE	SDS-agarose gel electrophoresis
SDGC	Sucrose density gradient centrifugation
SDS-PAGE	Sodium dodecyl sulfate – polyacrylamide gel electrophoresis
SEED alignment	Alignment with a set of representative members a protein family
Ser `	Serine
Sialylα2-6GalNAcα-Ser/Thr	Sialyl Tn antigen
SP	Signal peptide
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
Sus	Starch utilization system
TAE	Tris-acetate-EDTA
TBE	Tris-borate-EDTA
TEMED	N,N,N',N'-Tetramethylethylenediamine
TLC	Thin layer chromatography
TR	Tandem repeat
TYG	Trypticase yeast extract-glucose
UDP-GalNAc	Uridine diphosphate - N-acetyl-D-galactosamine
VNTR	Variable number tandem repeat
vWF	Von Willebrand factor

Journal Articles

The article below contains work conducted by the author

Nakjang, S., Ndeh, D. A., Wipat, A., Bolam, D. N., & Hirt, R. P. (2012). A novel extracellular metallopeptidase domain shared by animal host-associated mutualistic and pathogenic microbes. *PloS One, 7*(1), e30287. doi:10.1371/journal.pone.0030287; 10.1371/journal.pone.0030287

Other articles containing work conducted by the author during the course of this investigation are still in preparation.

CHAPTER I

General Introduction

I.1 Introduction

An understanding of the genetic and molecular mechanisms underlying host microbial interactions in disease and health remains crucial to the development of novel and innovative strategies for the management and improvement of health care. The importance of studies aimed at revealing new insights into some of these processes thus cannot be overemphasized especially now that demand for novel solutions to many of our world's health problems is also on the rise.

Recent findings that the human microbiota especially our resident gut microbiota does not only potentially encode a more extensive proteome than the human genome but can also negatively influence our health under certain physiological circumstances, have drawn a great deal of attention from the research and medical community (Qin *et al.*, 2010, Sekirov *et al.*, 2010) . Our resident gut microbes have now been implicated in many important public health hazards including obesity, diabetes, cancer, allergy, autism (Turnbaugh *et al.*, 2006, Sekirov *et al.*, 2010, Clemente *et al.*, 2012), a situation that is only further worsened by our currently limited knowledge of the molecular and genetic underpinnings of most of the processes involved. Even so, a lot about the mechanisms by which they positively impact on our health, something they are generally known for, still remains obscure. Foreign invading pathogens on the other hand have been the subject of intensive research for so many years and their threat to human health is well-known and documented. Although significant progress has been made over the years with this group of microbes, just like their resident counterparts, a lot about the mechanisms by which they influence human health also remains unclear.

To date, two major approaches have been used to study the human microbiota and interactions with its host. The first is largely sequencing – based and currently adopted by large-scale metagenomic projects including the Human Microbiome Project (HMP) (http://commonfund.nih.gov/hmp/). Data from this and other metagenomic projects have so far revealed vital insights into the content, diversity and functioning of human-associated microbial communities (Yatsunenko *et al.*, 2012, Ling *et al.*, 2013, Martínez *et al.*, 2013). There is also the function-driven approach involving the molecular and biochemical characterisation of microbial encoded factors. This rather painstaking approach remains one of the most reliable ways of predicting the activities of our resident microbiota (including foreign invading microbes) and the various mechanisms by which they interact with us to influence our health.

The current study is in line with the latter and focuses on host-microbial interactions at mucosal surfaces involving two important human gut and urogenital tract microbes; *Bacteroides thetaiotaomicron* and *Trichomonas vaginalis*, respectively. Both mucosal microbes are known to share several genetic features including genes encoding the novel, putative surface-exposed M60-like domain-containing family of proteins (PF13402) (Nakjang *et al.*, 2012), which alongside their functional partners in *B. thetaiotaomicron* were of major interest in this study.

I.2 Mucosal surfaces in the human body

Mucosae represent layers of tissue, contiguous with the skin and lining areas of the body that come in contact with air. Mucosal tissue lines internal organs and important tracts in the human body such as the respiratory, gastrointestinal and the urogenital tracts (RT, GIT and UGT respectively) (Figure I.1), forming a protective barrier in these areas. The surface of mucosal tissue (mucosal surfaces) is home to a wide range of secretions, the most prominent being the characteristic mucus of the gastrointestinal and respiratory tracts (Atuma *et al.*, 2001, Nataro *et al.*, 2005). Mucus, secretory IgA (SIgA), defensins, proteolytic enzymes, epithelial cell surface glycocalyx components and resident microflora all constitute important elements of the mucosal immune system (Macdonald, 2003, Kaiserlian *et al.*, 2005).

I.2.1 Mucus

Mucus is a thick, viscoelastic, sticky substance, synthesized, stored and secreted by goblet cells in mucosal tissue and some glands of the body. Components of mucus include water (95%), glycoproteins (mucins) (1-10%), antibodies, electrolytes and nucleic acids (Hollander, 1963, Allen and Snary, 1972, MacFarlane *et al.*, 2005). Studies on mucus in the rat stomach and small intestine reveal that secreted mucus actually forms two continuous layers over mucosal surfaces where it is secreted. These include an upper loosely adherent, movable gel layer overlying a less movable adherent layer in direct contact with epithelial cells (Atuma *et al.*, 2001, Johansson *et al.*, 2011). The stomach and colon contain well-defined and continuous layers of each type of mucus while the small intestine rather shows a discontinuous mucus organisation without well-defined layers (Figure I.1). It has also been reported that unlike the lower less movable layer, the upper movable layer in the mouse colon is home to many gut microbes (Johansson *et al.*, 2008).

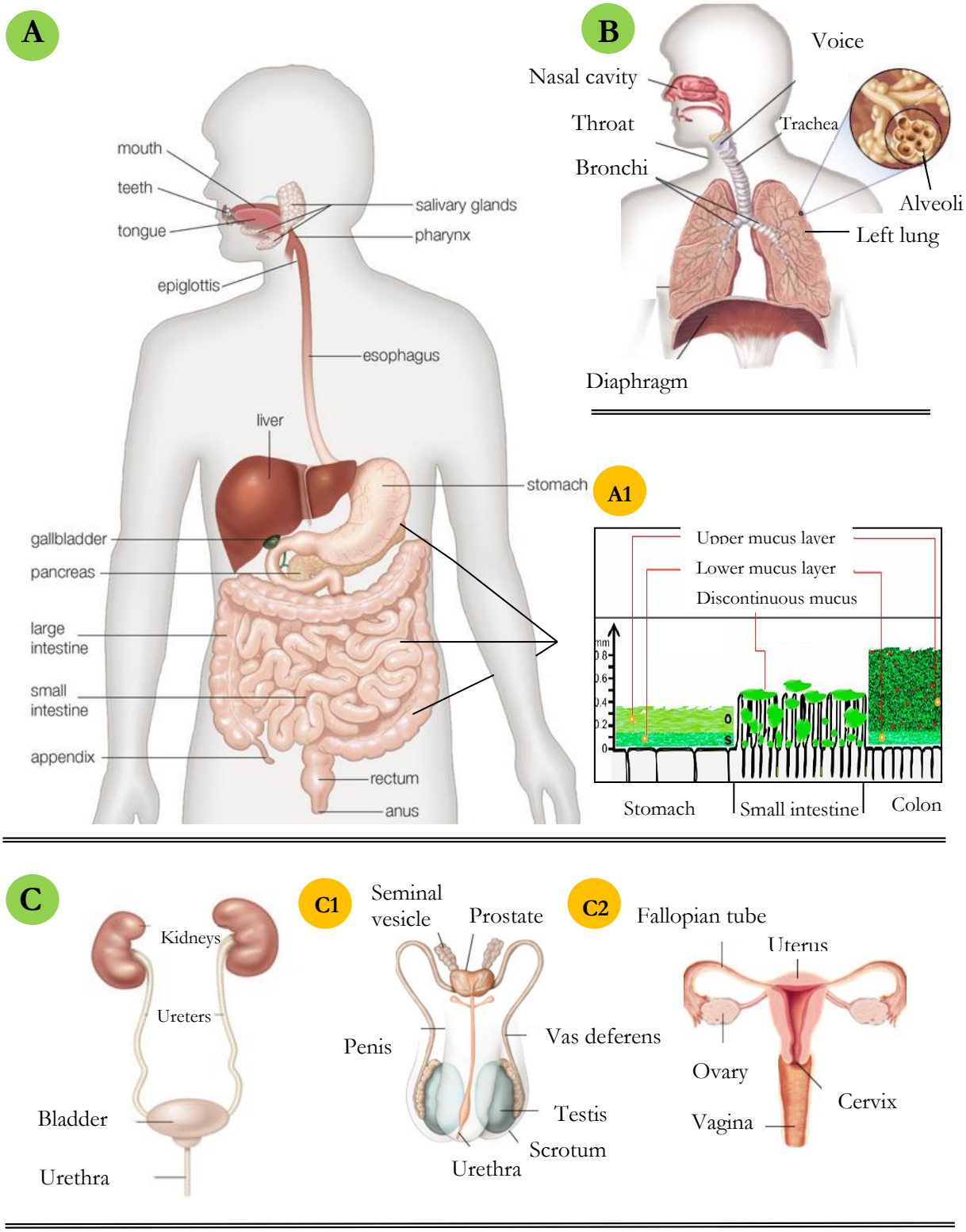


Figure I.1 – Examples of systems in the human body containing mucosal tissue and surfaces. **A**: Gastrointestinal system, **A1**: Cross-sectional view of gastrointestinal tissue showing mucus layers **B**: Respiratory system and **C:** Urogenital system in both males and females. **C1 & C2:** Rest of urogenital system in males and females respectively. Images were modified from Johansson *et al.*, (2011) and Encyclopaedia Britannica online (Written by Harrison R J)

Some important functions of the mucus bilayer include lubrication (e.g. during digestion) and physical protection against invading pathogens, toxins and other potentially harmful environmental substances (Macfarlane *et al.*, 2005, Thornton and Sheehan, 2004). There is indeed evidence of up regulation of mucus components such as mucins following exposure to bacterial lipopolysaccharides (Dohrman *et al.*, 1998). Paradoxically, mucus is a potentially rich source of carbon and nitrogen for gut microbes, some of which encode mucin degrading enzymes (Hernandez-Gutierrez *et al.*, 2004, Grys *et al.*, 2005, Gutierrez-Jimenez *et al.*, 2008, Szabady *et al.*, 2011, Ruiz-Perez *et al.*, 2011)

I.2.1.1 Mucins

Mucins are high molecular weight glycoproteins (>1000 KDa) and important structural components contributing significantly to the lubricative and viscoelastic properties of mucus (Gerken, 1993). The general makeup of mucins is a peptide core to which are attached several sugar side chains giving it a 'bottle brush' appearance (Bansil and Turner, 2006).

Mucins can be either secreted (gel forming and extracellular) or membrane bound (cell surface anchored or membrane tethered) (Figure I.2) and both share a lot of similarities (Perez-Vilar and Hill, 1999, Hattrup and Gendler, 2008, Dharmani *et al.*, 2009). An important distinguishing feature between them is the presence of a membrane-spanning domain and cytoplasmic tail in membrane tethered mucins (Hattrup and Gendler, 2008).

Mucin gene expression is tissue-specific and to date over 21 human mucin (MUC) genes have been identified (Table I.1, Figure I.3). A characteristic feature of almost all mucin types is the presence of variable number tandem repeats (VNTR) of amino acid motifs (Vinall *et al.*, 1998, Jiang *et al.*, 2000, Perez-Vilar and Hill, 1999). Repeats in different mucin types which may be identical or degenerate often contain a high number of proline, threonine and serine residues (PTS repeats) (Vinall *et al.*, 1998, Perez-Vilar and Hill, 1999) with serine and threonine representing potential glycosylation sites in the structure (Hattrup and Gendler, 2008). Repeat domains are often centrally located in the mucin peptide and flanked by several other types of domains typically cysteine rich domains some of which show similarity to C-terminal cystine knot domains and von Willebrand factor (vWF) C and D domains (Section I.2.1.1.3, Figure I.8). Terminal cysteine rich domains are essential for polymerization of mucins via disulphide bonds (Bansil and Turner, 2006, Sheehan *et al.*, 2004).

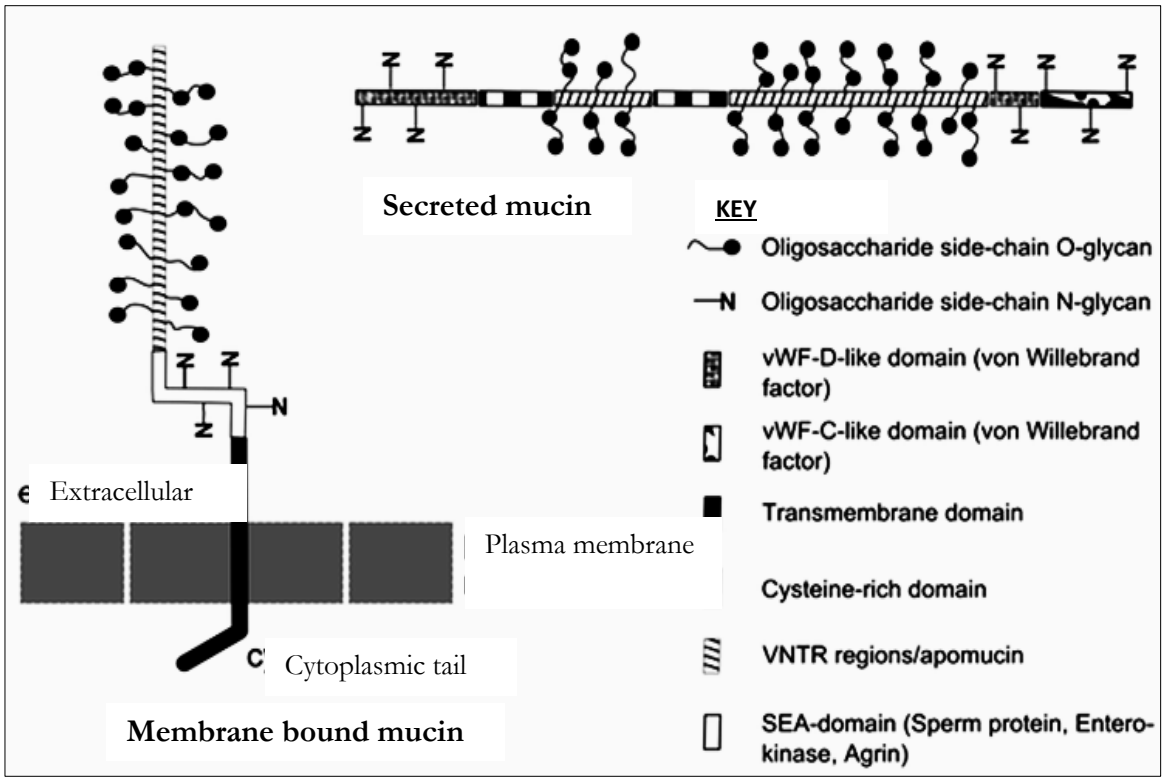


Figure I.2 - Typical structural organization of secreted and membrane bound mucins. Modified from Sipaul *et al.*, (2011).

Mucin gene	Form	TR/cysteine	Species
MUC1	Membrane bound	TR	H, R, M
MUC2	Secreted	Cysteine rich	H, R, M
MUC3A	Membrane bound	TR	H, R, M
MUC3B	Membrane bound	TR	H, R, M
MUC4	Membrane bound	TR	H, R, M
MUC5AC	Secreted	Cysteine rich	H, R, M
MUC5B	Secreted	Cysteine rich	
MUC6	Secreted	Cysteine rich	H, R, M
MUC7	Secreted	Cysteine rich	H, R, M
MUC8	Secreted	Cysteine poor	H, R, M
MUC9	Secreted	Cysteine poor	H, R, M
MUC10	Membrane bound	TR	R, M
MUC11	Membrane bound	TR	H, R, M
MUC12	Membrane bound	TR	H, R, M
MUC13	Membrane bound	TR	H, R, M
MUC14	Membrane bound	TR	H, R, M
MUC15	Membrane bound	TR	H, R, M
MUC16	Membrane bound	TR	H, R, M
MUC17	Membrane bound	TR	H, R, M
MUC18	Membrane bound	None	H, R, M
MUC19	Secreted	Cysteine rich	H, R, M
MUC20	Membrane bound	TR	H, R, M
MUC21	Membrane bound	TR	H, M

Table I.1 - Diversity and properties of human mucins. TR: Tandem repeat, H: Human, R: Rat M: Mouse (Dharmani *et al.*, 2009)

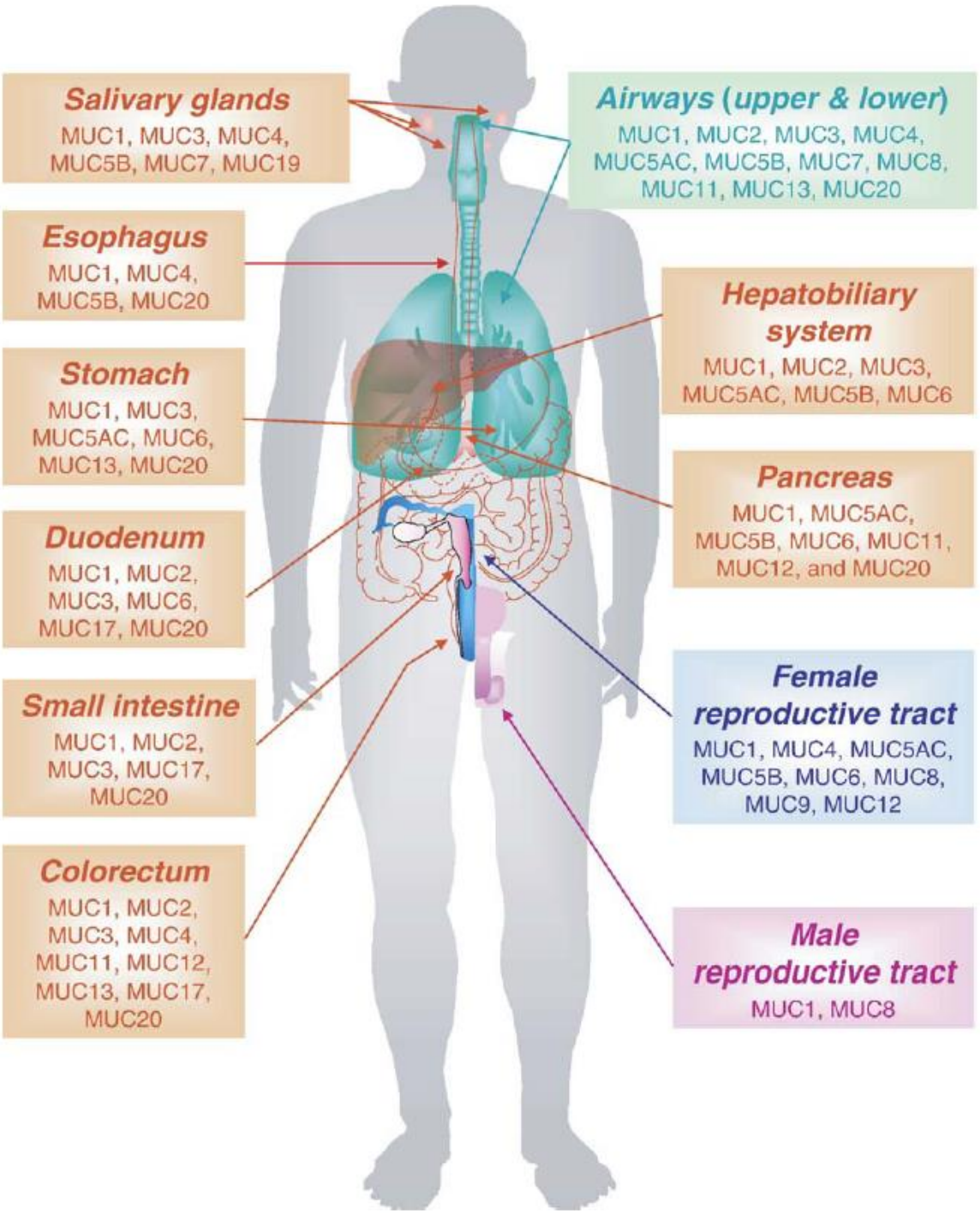


Figure I.3 - Mucin gene expression in different areas of the human body. Mucins indicated in annotation boxes are colour coded to match the colour of the various organs (in the image) in the human body expressing them. MUC21 which is not indicated in the image above is known to be present in the lungs, large intestines, thymus and testes (Dharmani *et al.*, 2009). This image was adapted from Andrianifahanana *et al.*, 2006.

I.2.1.1.1 Mucin glycosylation

A bulk of the molecular weight of mucins (about 90%) is contributed by the attached sugars (Perez-Vilar and Hill, 1999). Two major types of mucin glycosylation are known. These include the mucin type - O and N-glycosylations, named after the type of linkage present between the attached sugars and the protein backbone. Glycosylation generally has a marked impact on the structure and physicochemical properties of proteins. In the case of mucins, it is important for their stability, protection, folding, solubility and rheological properties (Thornton *et al.*, 2008).

I.2.1.1.1.1 Mucin O- glycosylation

O-glycosylation is the predominant form of mucin glycosylation. Mucin O-glycosylation occurs in the Golgi complex and typically commences with the attachment of the acetylated sugar α-N-acetylgalactosamine (αGalNAc) to the peptide backbone of the immature mucin. This is catalyzed by a large family of enzymes termed polypeptide *N*-acetylgalactosaminyltransferases (ppGalNAc-Ts) capable of linking αGalNAc (from UDP-GalNAc) with the hydroxyl groups of serine or threonine in the tandem repeats of the immature mucin polypeptide backbone (Tran and Hagen 2013, Nakayama *et al.*, 2013, Jensen *et al.*, 2010). The resulting bond formed between GalNAc and serine or threonine in the peptide backbone is termed an O-glycosidic bond and can involve other sugars such as xylose, fucose, mannose, galactose, and glucose in place of GalNAc in other glycoproteins (Nakayama *et al.*, 2013, Rose and Voynow, 2006). The core structure that is formed following the addition of GalNAc to serine or threonine is termed the Tn or tumor antigen (αGalNAc-Ser/Thr). Extension of this core by other glycosyl transeferase enzymes through the addition of a combination of sugars such as Gal and GlcNAc or GalNAc, results in the formation of at least 8 different types of mucin-type O-core structures (Figure I.4). Different types of mucins contain different types and amounts of these core structures which can be further enzymatically extended into more complex glycan structures as observed with human MUC5B and MUC2 (Thomsson *et al.,* 2002, Larsson *et al.,* 2009, Jensen *et al.*, 2010, Moran *et al.*, 2011). A list of some O-glycan structures obtained by mass spectrometry from different areas of the intestines of two human subjects is given below in Table I.2. Very extended or complex mucin glycans are typically divided in to three units namely the core, backbone and peripheral units as shown in the model structure of a complex mucin in Figure I.5. Also see Figure I.4 for a summary of the steps involved in the synthesis of various mucin type-O glycan core structures in mammals. More details about the process are given in Nakayama *et al.*, 2013.

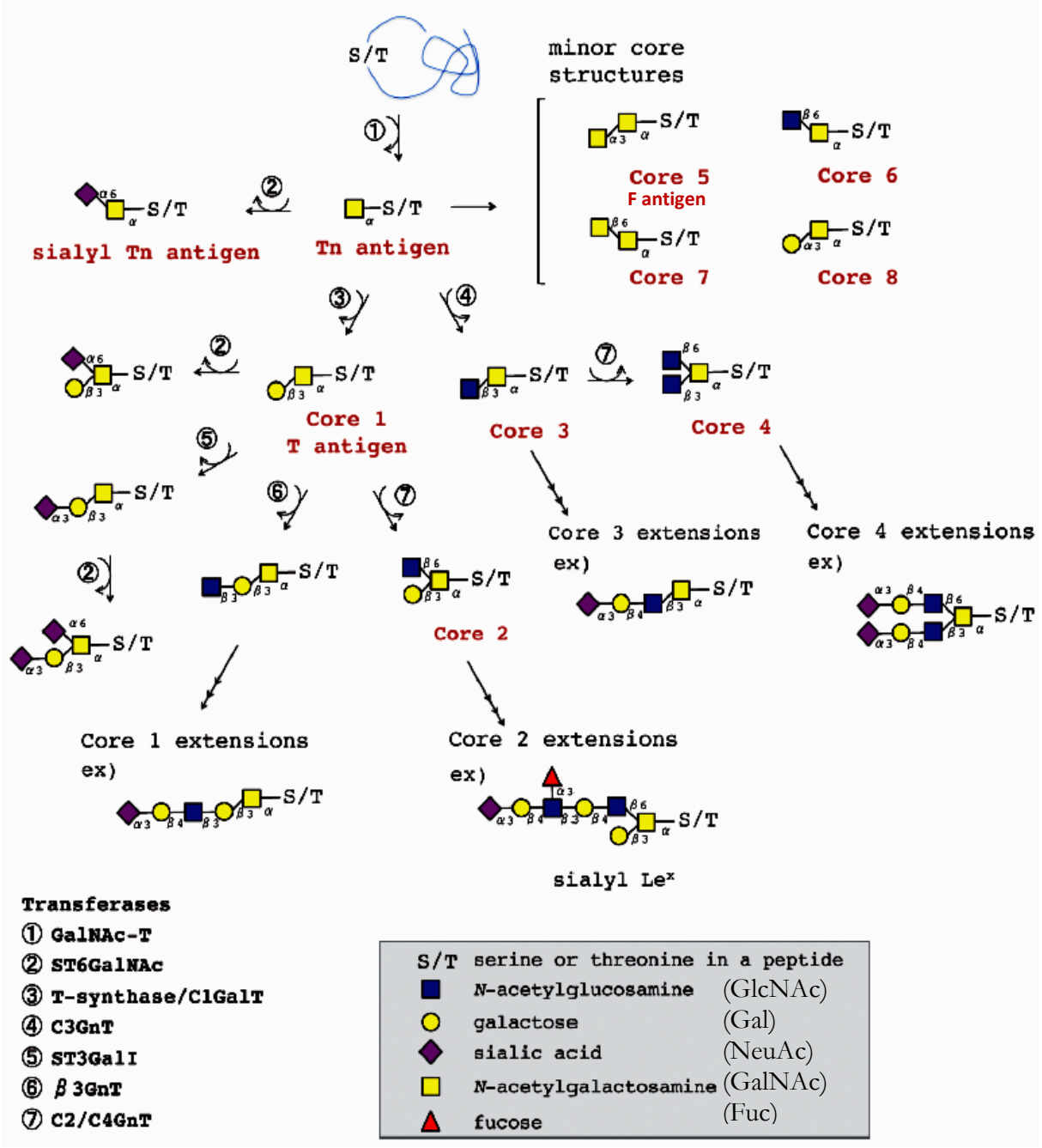


Figure I.4 – Biosynthesis of mucin type – O glycans. Mucin biosynthesis commences with the addition of N-acetylgalactosamine to a serine or threonine residue in the apomucin protein by the enzyme *N*-acetylgalactosaminyltransferase (GalNAc-T) forming the Tn antigen (GalNAcα1-O-Ser/Thr). Enzymatic extension of the Tn antigen by other specific transferases results in the formation of a variety of mucin core structures which can be further extended into more complex glycan structures. See a more detailed description of the process in Nakayama *et al.*, 2013 from where this image was adapted.

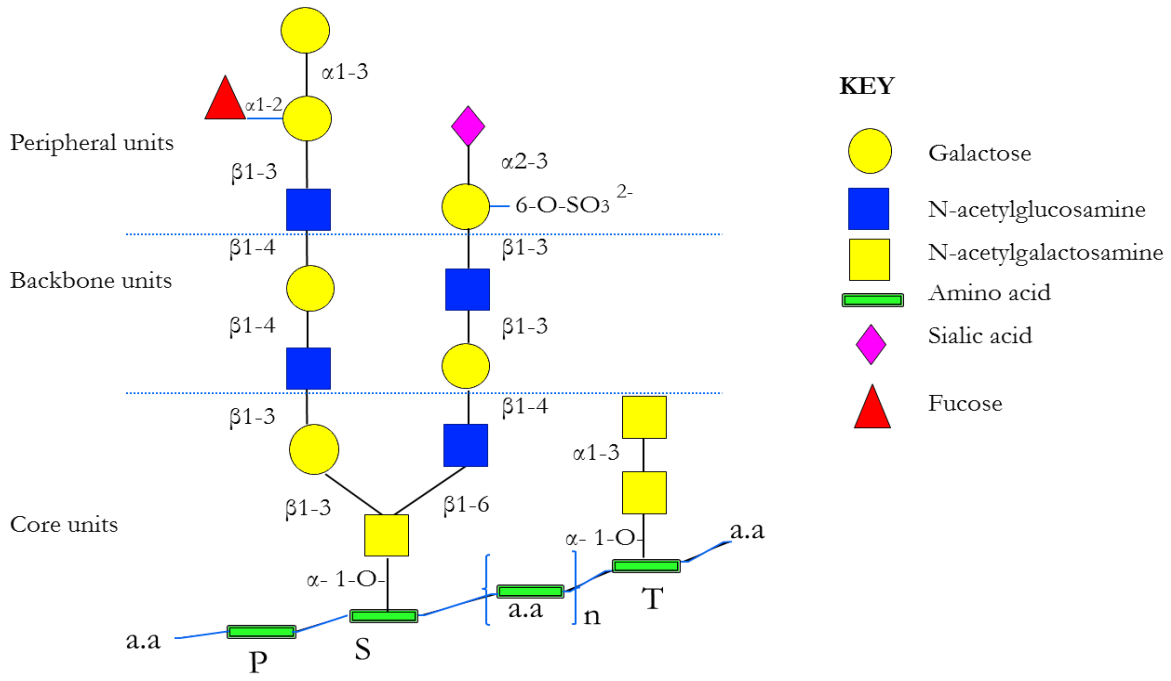


Figure I.5 - Model structure of a complex mucin showing various units. This image was redesigned from Wiggins *et al.*, (2001), with information from various other sources including Martens *et al.*, (2008), Moran *et al.*, (2011), Jensen *et al.*, (2010) and Nakayam *et al.*, (2013).

Sequence/composition of oligosaccharide alditols	I		C		T		S		R	
	a	b	a	b	a	b	a	b	a	b
GlcNAc→3GalNAc-ol	+	+	+	+	+	+	+	+	+	+
GalNAc→3GalNAc-ol	+	+	+	+	+	+	+	+	+	+
Gal→3GlcNAc→3GalNAc-ol	+	+	+	+	+	+	+	+	+	+
Gal→3(GlcNAc→6)GalNAc-ol	+	+	+	+	+	+	+	+	+	+
GlcNAc→3Gal→3GalNAc-ol	+	+	+	+	+	+	+	+	+	+
Fuc→2Gal→3GalNAc-ol	−	−	−	−	−	−	+	+	−	−
Gal→3(Fuc→4)GlcNAc→3GalNAc-ol	+	+	+	+	+	+	+	+	+	+
Gal→3(Gal→4GlcNAc→6)GalNAc-ol	+	+	+	+	−	−	+	+	−	+
Gal→4GlcNAc→3Gal→3→GalNAc-ol	+	+	+	+	−	−	−	−	−	−
HexNAc→Gal→3GlcNAc→3GalNAc-ol	+	+	+	+	+	+	+	+	+	+
HexNAc→3Gal→3(GlcNAc→6)GalNAc-ol	+	+	+	+	+	+	+	+	+	+
Gal→4GlcNAc→3(GlcNAc→6)GalNAc-ol	−	+	+	−	−	−	+	−	+	+
(Fuc→2)Gal→3(Fuc→4)GlcNAc→3GalNAc-ol	−	−	+	+	+	+	−	−	−	−
(Fuc→)GlcNAc→3(Fuc→2)Gal→3GalNAc-ol	−	−	−	+	−	−	−	−	−	−
Gal→4GlcNAc→3(Fuc→2)Gal→3GalNAc-ol	+	+	−	−	−	−	−	−	−	−
(Fuc→2)Gal→3(Gal→4GlcNAc→6)GalNAc-ol	+	+	−	−	−	−	−	−	−	−
2 Gal, GlcNAc, Fuc, GalNAc-ol	+	+	+	+	+	−	−	−	+	+
HexNAc→3Gal→3(Fuc→4)GlcNAc→3GalNAc-ol	+	+	+	+	−	−	−	−	−	−
Gal, 2HexNAc, Fuc, GalNAc-ol	+	+	+	+	+	+	−	−	−	−
Gal, 2GlcNAc, Fuc, GalNAc-ol (core 2)	+	−	−	−	−	−	−	−	−	−

Sequence/composition of oligosaccharide alditols	I		C		T		S		R	
	a	b	a	b	a	b	a	b	a	b
Gal, 2GlcNAc, Fuc, GalNAc-ol (core 4)	+	+	+	−	−	−	−	−	−	−
2 Gal, 2 HexNAc, GalNAc-ol	+	+	+	+	+	+	+	+	+	+
HexNAc→3Gal→4GlcNAc→3(GlcNAc→6)GalNAc-ol	−	−	−	−	−	−	−	−	−	+
Gal→3[(Fuc→2)Gal→3(Fuc→4)GlcNAc→6]GalNAc-ol	+	−	−	−	−	−	−	−	−	−
(Fuc→2)Gal→3[(Fuc→2)Gal→4GlcNAc→6]GalNAc-ol	+	−	−	−	−	−	−	−	−	−
2 Gal, HexNAc, 2 Fuc, GalNAc-ol	+	−	+	−	−	−	−	−	−	−
HexNAc→3(Fuc→2)Gal→3(Fuc→4)GlcNAc→3GalNAc-ol	+	+	+	+	−	−	−	−	−	−
2 Gal, 2 HexNAc, Fuc, GalNAc-ol	+	+	+	+	−	−	−	−	−	+
Gal, 3 HexNAc, Fuc, GalNAc-ol	−	−	−	−	−	−	−	−	−	+
HexNAc→3Gal→3(Fuc→4)GlcNAc→3(GlcNAc→6)GalNAc-ol	+	+	−	−	−	−	−	−	−	−
2 Gal, 3 HexNAc, GalNAc-ol	−	−	−	−	−	−	+	+	+	+
2 Gal, HexNAc, 3 Fuc, GalNAc-ol	+	−	−	−	−	−	−	−	−	−
2 Gal,2 HexNAc, 2 Fuc, GalNAc-ol	+	+	+	+	+	+	−	−	−	−
(Fuc→2)Gal→3(Fuc→4)GlcNAc→3(Gal→4GlcNAc→6)GalNAc-ol	+	+	−	−	−	−	−	−	−	−
3 Gal, 2 HexNAc, Fuc, GalNAc-ol	+	+	−	−	+	+	−	−	−	−
HexNAc→3(Fuc→2)Gal→3(Fuc→4)GlcNAc→3(GlcNAc→6)GalNAc-ol	+	+	−	−	−	−	−	−	−	−
2 Gal, 3 HexNAc, 1 Fuc, GalNAc-ol	+	+	−	−	−	−	−	−	−	−
2 Gal, 2 HexNAc, 3 Fuc, GalNAc-ol	+	+	−	−	−	−	−	−	−	−
3 Gal, 2 HexNAc, 2 Fuc, GalNAc-ol	+	+	−	−	−	−	−	−	−	−
2 Gal, 3 HexNAc, 2 Fuc, GalNAc-ol	+	+	−	−	−	−	−	−	−	−
2 Gal, 4 HexNAc, 1 Fuc, GalNAc-ol	+	+	−	−	−	−	−	−	−	−
3 Gal, 2 HexNAc, 3 Fuc, GalNAc-ol	+	+	−	−	−	−	−	−	−	−
2 Gal, 3 HexNAc, 3 Fuc, GalNAc-ol	+	+	−	−	−	−	−	−	−	−
2 Gal, 4 HexNAc, 2 Fuc, GalNAc-ol	+	+	−	−	−	−	−	−	−	−
3 Gal, 3 HexNAc, 3 Fuc, GalNAc-ol	+	+	−	−	−	−	−	−	−	−
2 Gal, 4 HexNAc, 3 Fuc, GalNAc-ol	−	+	−	−	−	−	−	−	−	−
Oligosaccharides with one NeuAc residue										
NeuAc→6GalNAc-ol	+	+	+	+	+	+	+	+	+	+
Gal→3(NeuAc→6)GalNAc-ol	+	+	+	+	+	+	−	−	+	+
(NeuAc→3)Gal→3GalNAc-ol	−	−	−	−	−	−	+	+	−	−
GalNAc→3(NeuAc→6)GalNAc-ol	+	+	+	+	+	+	+	+	+	+
GlcNAc→3(NeuAc→6)GalNAc-ol	+	+	+	+	+	+	+	+	+	+
(Fuc→2)Gal→3(NeuAc→6)GalNAc-ol	+	−	+	−	−	−	−	−	−	−
Gal→3GlcNAc→3(NeuAc→6)GalNAc-ol	+	+	+	+	+	+	+	+	+	+
(NeuAc→3)Gal→3(GlcNAc→6)GalNAc-ol	−	−	−	−	−	−	−	+	−	−
(NeuAc→3)Gal→4GlcNAc→3GalNAc-ol	−	−	−	−	−	−	−	+	−	−
GlcNAc→3Gal→3(NeuAc→6)GalNAc-ol	+	−	+	−	+	+	+	+	+	+
Gal→3(Fuc→4)GlcNAc→3(NeuAc→6)GalNAc-ol	+	+	+	+	+	+	+	+	+	+
Gal→4(Fuc→3)GlcNAc→3(NeuAc→6)GalNAc-ol	+	+	+	+	+	+	+	+	+	+
(NeuAc→3)Gal→4(Fuc→3)GlcNAc→3GalNAc-ol	−	−	−	−	−	−	−	+	−	−
HexNAc→3Gal→4GlcNAc→3(NeuAc→6)GalNAc-ol	−	+	−	+	−	−	−	−	−	−
GalNAc→4(NeuAc→3)Gal→4GlcNAc→3GalNAc-ol	−	−	−	−	+	+	+	+	+	+
GalNAc→4(NeuAc→3)Gal→3GlcNAc→3GalNAc-ol	−	−	−	−	+	+	+	+	+	+
(Fuc→2)Gal→3(Fuc→4)GlcNAc→3(NeuAc→6)GalNAc-ol	+	+	+	+	+	−	−	−	−	−
HexNAc→3Gal→3(Fuc→4)GlcNAc→3(NeuAc→6)GalNAc-ol	+	+	+	+	+	−	−	−	−	−
HexNAc→3(Fuc→2)Gal→3GlcNAc→3(NeuAc→6)GalNAc-ol	+	−	+	−	−	−	−	−	−	−
(NeuAc→3)Gal→4GlcNAc→3Gal→GlcNAc→3GalNAc-ol	−	−	−	−	−	−	+	+	−	−
HexNAc→3(Fuc→2)Gal→3(Fuc→4)GlcNAc→3(NeuAc→6)GalNAc-ol	+	+	+	+	−	−	−	−	−	−
NeuAc→3Gal→4(Fuc→3)GlcNAc→3Gal→3[Gal→4(Fuc→3)GlcNAc→6]GalNAc-ol	−	−	−	−	−	−	+	+	+	+

Sequence/composition of oligosaccharide alditols	I a	I b	C a	C b	T a	T b	S a	S b	R a	R b
Oligosaccharides wih one sulphate residue										
(SO3−)3Gal→4GlcNAc→3GalNAc-ol	−	−	−	+	−	−	+	+	+	+
Gal→4(SO3−)6GlcNAc→3GalNAc-ol	+	−	−	+	−	−	+	+	+	+
(Fuc→2)Gal→4(SO3−)6GlcNAc→3GalNAc-ol	+	−	−	+	−	−	−	−	−	+
(SO3−)3Gal→4(Fuc→3)GlcNAc→3GalNAc-ol	−	−	+	+	+	+	+	+	+	+
Gal→3[(SO3−)3Gal→4GlcNAc→6]GalNAc-ol	−	−	−	−	+	−	+	+	+	+
Gal→3[Gal→4(SO3−)6GlcNAc→6]GalNAc-ol	−	−	−	−	+	−	−	−	+	+
(SO3−)3Gal→4GlcNAc→3Gal→3GalNAc-ol	−	−	+	+	+	−	−	−	−	−
Gal→3[(SO3−)3Gal→4(Fuc→3)GlcNAc→6]GalNAc-ol	−	−	−	+	+	+	−	+	+	+
(SO3−)3Gal→4(Fuc→3)GlcNAc→3Gal→3GalNAc-ol	−	−	−	−	+	−	+	+	+	+
(SO3−)3Gal→4(Fuc→3)GlcNAc→3Gal→3[Gal→4(Fuc→3)GlcNAc→6]GalNAc-ol	−	−	−	−	−	−	−	+	−	−
Oligosaccharides with two acidic residues										
(SO3−)3Gal→4GlcNAc→3(NeuAc→6)GalNAc-ol	−	−	−	−	+	+	+	+	+	+
(NeuAc→3)Gal→3(NeuAc→6)GalNAc-ol	+	+	+	+	+	+	+	+	−	−
(SO3−)3Gal→3[(SO3−)3Gal→4(Fuc→3)GlcNAc→6]GalNAc-ol	−	−	−	−	−	−	−	+	−	−
(SO3−)3Gal→4(Fuc→3)GlcNAc→3(NeuAc→6)GalNAc-ol	+	+	+	+	+	+	+	+	+	+
(NeuAc→3)Gal→3[(SO3−)3Gal→4GlcNAc→6]GalNAc-ol	−	−	−	−	−	−	+	−	+	+
(NeuAc→3)Gal→4GlcNAc→3(NeuAc→6)GalNAc-ol	+	+	−	+	+	+	+	+	+	+
(SO3−)3Gal→4(Fuc→3)GlcNAc→3Gal→3(NeuAc→6)GalNAc-ol	−	−	−	−	+	+	−	−	−	−
(NeuAc→3)Gal→4(Fuc→3)GlcNAc→3(NeuAc→6)GalNAc-ol	−	−	+	+	+	+	+	+	+	+
(SO3−)3Gal→4GlcNAc→3Gal→4GlcNAc→3(NeuAc→6)GalNAc-ol	−	−	−	+	−	−	−	+	−	−
GalNAc→4(NeuAc→3)Gal→4GlcNAc→3(NeuAc→6)GalNAc-ol	−	−	−	+	+	+	+	+	+	+
GalNAc→4(NeuAc→3)Gal→3GlcNAc→3(NeuAc→6)GalNAc-ol	−	−	−	+	+	+	+	+	+	+
(SO3−)3Gal→4GlcNAc→3Gal→3[(SO3−)3Gal→4(Fuc→3)GlcNAc→6]GalNAc-ol	−	−	−	−	−	−	−	+	−	−
(SO3−)3Gal→4(Fuc→3)GlcNAc→3Gal→4GlcNAc→3(NeuAc→6)GalNAc-ol	−	−	+	+	−	−	+	+	+	−
3 Gal, 2 HexNac, NeuAc, SO3−, GalNAc-ol	−	−	−	−	−	−	−	−	+	−
(SO3−)3Gal→4(Fuc→3)GlcNAc→3Gal→3[(SO3−)3Gal→4(Fuc→3)GlcNAc→6]GalNAc-ol	−	−	−	+	−	−	+	+	+	+
(SO3-)3Gal→4(Fuc→3)GlcNAc→3Gal→4(Fuc→3)GlcNAc-3(NeuAc→6)GalNAc-ol	−	−	+	+	+	+	+	+	+	+
2 Gal, 2 HexNAc, Fuc, 2 NeuAc, GalNAc-ol	−	−	+	−	−	−	−	−	−	−
2Gal, 2 HexNAc, 2 Fuc, 2 NeuAc, GalNAc-ol	−	−	−	−	−	−	−	−	+	−

Table I.2 - Glycan structures detected in mucins from different areas of the intestines of two human subjects (a and b) with blood group ALeb (A-Lewisb). I = Ileum, C= cecum, T = transverse, S = sigmoid colon, R= rectum. All data presented in this table was adapted from Robbe *et al.*, 2004.

I.2.1.1.1.2 Mucin N-glycosylation

N-glycosylation of mucins is also possible (Hattrup and Gendler, 2008, Moran *et al.*, 2011), targeting asparagine in a unique tri-peptide consensus sequence; Asn-Xaa-Ser/Thr (Xaa = any amino acid apart from proline) when present in the nascent mucin polypeptide (Parry *et al.*, 2006, Moran *et al.*, 2011). It is much less predominant compared to O-glycosylation and has equally received less research attention. N-glycosylation of proteins in mammals unlike O-glycosylation occurs in the ER, commencing with the addition of a pre-assembled glycolipid precursor designated; $Glc_3Man_9GlcNAc_2$-P-P-Dol to asparagine (in the Asn-Xaa-Ser/Thr tri-peptide consensus) in the immature mucin protein. The preassembled structure is prepared in a series of reactions involving several transferase enzymes encoded by a family of genes termed ALGs (asparagine-linked glycosylation) (Figure I.6) while the transfer of the structure to the protein backbone of the immature mucin is catalysed by a large enzyme complex termed OST (oligosaccharyltransferase)(Potapenko *et al.*, 2010). The resulting glycoprotein structure is later on modified by a series of enzymes including glycoside hydrolases and glycosyltransferases some in the golgi appartaus to produce a variety of N-glycan structures. Depending on their content and complexity they are classified as either high mannose, complex or hybrid N-glycans. A summary of the process adapted from Potapenko *et al.*, (2010) is given below in Figure I.6 with more details of the process in the same publication.

I.2.1.1.2 MUC1 – an example of a predominantly membrane anchored mucin

Mucin 1 (MUC1), variously called CD227, episialin, or epithelial membrane antigen (EMA) is a product of the human MUC1 gene and a prominent urogenital tract mucin of males and females. It is highly expressed in the uterus, cervix, vagina, prostate, ovaries, kidneys and has also been reported in many others areas of the body including the breast, lungs, cornea, salivary glands, oesophagus, stomach, pancreas, small and large intestine (Figure I.3 and Dharmani *et al.*, 2009).

MUC1 is a high molecular weight glycoprotein consisting of a peptide backbone (with an estimated molecular weight of 125-220KDa), to which are attached several glycan side chains. The inclusion of attached glycans brings the overall molecular mass of the mature glycoprotein to about 250-500KDa (Lancaster *et al.*, 1990, Brayman *et al.*, 2004, Lagow *et al.*, 1999). Although shed and secreted versions of MUC1 have been reported (Boshell *et al.*, 1992, Engelmann *et al.*, 2005, Baruch *et al.*, 1999, Hanisch *et al.*, 2000) It is often found tethered to the apical membrane of epithelial cells that produce it, thanks to the presence of a trans-membrane domain in the structure of some variants e.g MUC1/REP and MUC1/Y (Baruch *et al.*, 1999).

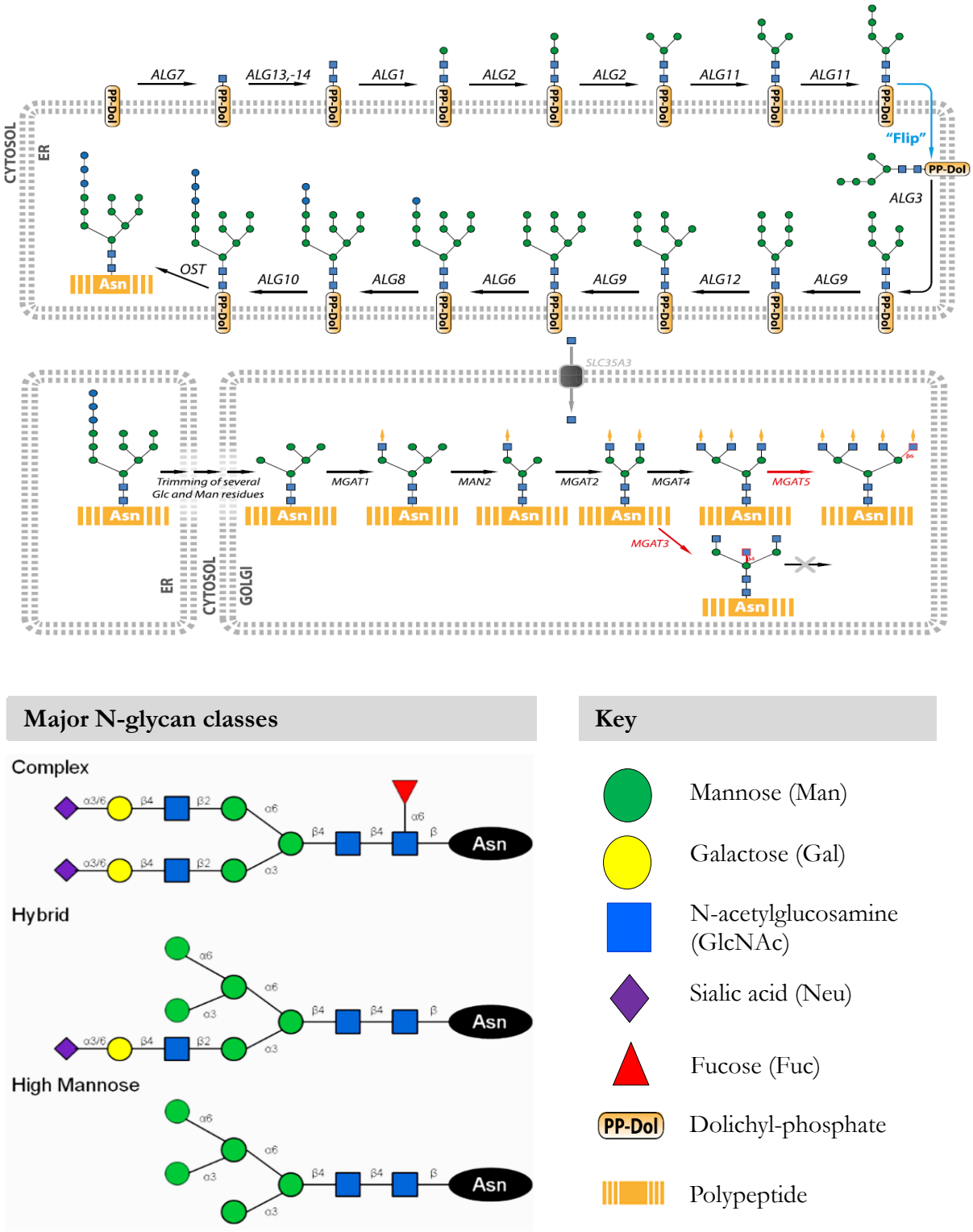


Figure I.6 - N-glycan biosynthesis in humans. Enzymes in the endoplasmic reticulum (ER) encoded by the ALG (asparagine-linked glycosylation) gene family catalyze the formation of $Glc_3Man_9GlcNAc_2$-P-P-Dol which is then transferred by the enzyme OST (oligosaccharyltransferase) to asparagine (Asn) in a unique tripeptide consensus of the nascent mucin polypeptide. Subsequent modifications of the glycoprotein by glycoside hydrolase and glycosyltransferase enzymes some in the golgi apparatus [mainly mannosidase (MAN) and mannosyl N-acetylglucosaminyltransferases (MGAT)], lead to the formation of a variety of N-glycans classified as either complex, hybrid or high mannose N-glycans. N-glycans like their O-glycans can also be sialylated or fucosylated. Images modified from Potapenko *et al.*, (2010), and Nettleship, (2012)

The well characterized MUC1/REP also contains two other important domains namely the cytoplasmic and extracellular domains. Following translation, the protein is processed into two cleavage products, one containing the extracellular domain alone and the other, the transmembrane and cytoplasmic domains. This autoproteolysis is thought to occur in a conserved region of the protein termed the SEA domain (Levitin *et al.*, 2005). When anchored to the epithelial membrane, both cleavage products are held together by non-covalent SDS sensitive bonds (Ligtenberg *et al.*, 1992, Baruch *et al.*, 1999).

The short cytoplasmic domain is in contact with the interior or cytoplasm of the epithelial cell while the extracellular domain enables the extension of MUC1 towards the outside of the cell e.g into the lumen or glandular ducts as in the uterus (Figure I.7). This is often to an extent (about 200 nM) known to be further than many extracellular surface proteins including E-cadherin (Wesseling *et al.*, 1996). The extracellular domain also contains degenerate variable tandem repeat sequences rich in serine and threonine residues, with some serving as sites for the attachment of O-linked glycans (Roy and Baek, 2002, Hanisch *et al.*, 2000).

MUC1 O-glycosylation is typically Core 2 - based (Galβ3(Galβ3/4GlcNAcβ6)GalNAc-S/T) (Figure I.7). MUC1 expression and glycosylation is also thought to change in diseases such as breast and colon cancer (Schroeder *et al.*, 2004, Backstrom, *et al.*, 2009, Taylor-Papadimitriou *et al*, 1999). With regards to glycosylation, there is decreased production of extended Core 2 - based glycans and a corresponding increase in shorter T antigen (Core 1) - based glycans (Figure I.7 and Taylor-Papadimitriou *et al*, 1999). This has been attributed to the lack of expression of the enzyme Core 2 p6-GlcNAc-transferase (Galβ1–3GalNAc/β-6-N-acetylglucosaminyltransferase) (Brockhausen *et al.*, 1995, Muller *et al.*, 1999) important for the addition of β1-6 linked GlcNAc to GalNAc in the T antigen chain of the growing glycoprotein. There is also a corresponding increase in α3-sialyltransferase activity in the cancer cell lines leading to the increased sialylation of T antigen and Tn antigen, thus explaining the predominance of sialylated forms of these sugars in breast cancers (Brockhausen *et al.*, 1995, Figure I.7). The nature of O-glycans in prostate cancer is still a subject of debate. While Arai *et al.*, 2005 have reported increased levels of sialylated MUC1 in biopsies from prostate cancer patients at different stages of the disease, the same observation has not been made for MUC1 from C42B prostate cancer cell lines, whose O-glycan structures are not known to be significantly different from normal (Premaratne *et al.*, 2011, Backstrom *et al.*, 2009).

N-linked glycans are also present on MUC1, close to the trans-membrane domain attached to the epithelial cell surface (Figure I.7A). A study comparing secreted and membrane anchored MUC1 N-glycans in human milk showed that secreted MUC1 contained more high mannose N-glycans as opposed to complex type N-glycans in their membrane anchored counterparts (Parry *et al.*, 2006).

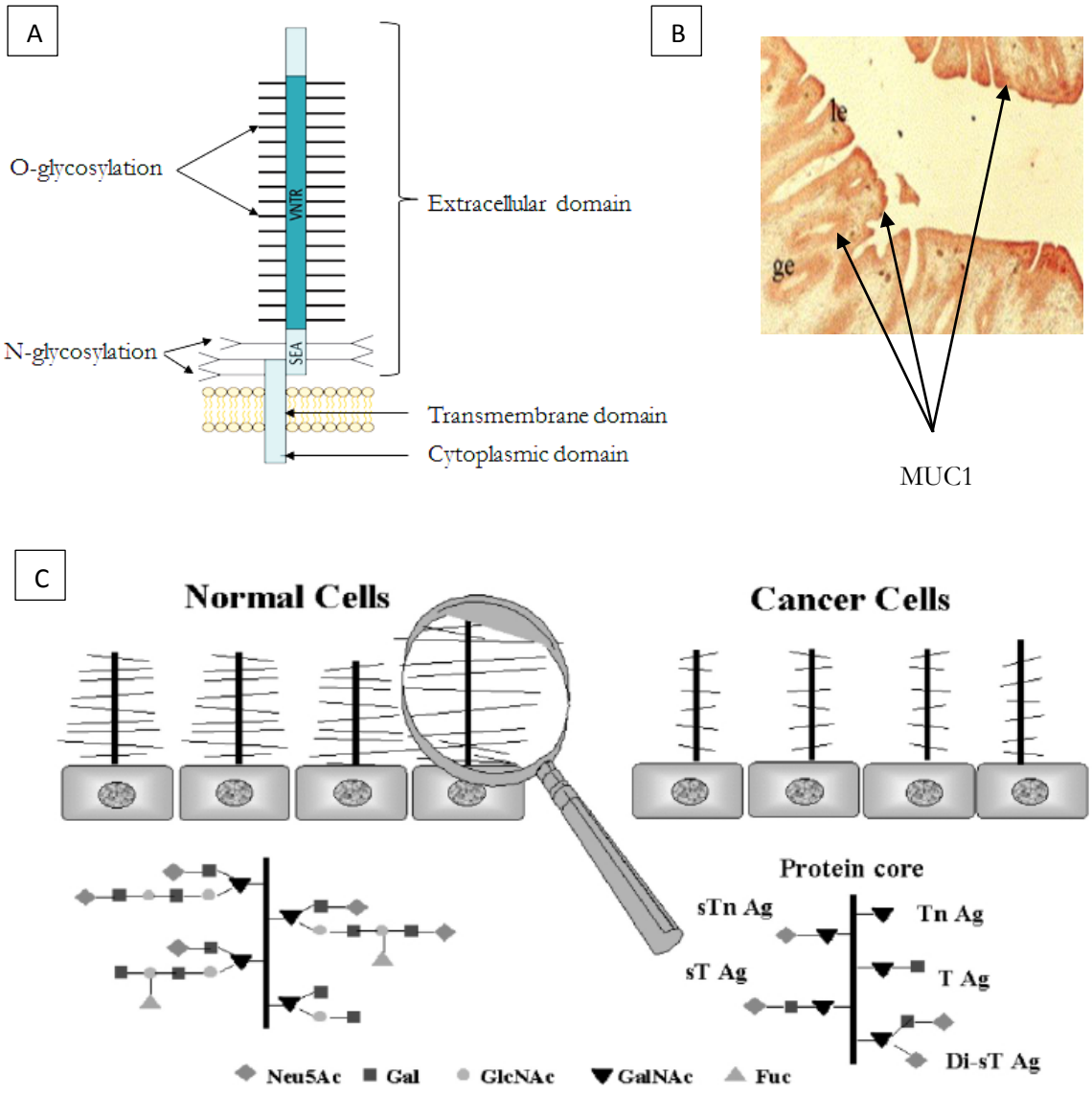


Figure I.7 – Properties of human mucin-1 (MUC1). A: Schematic diagram showing structural features of membrane tethered MUC1 (McGuckin *et al.*, 2011). **B:** Staining of MUC1 on luminal (le) and glandular epithelia (ge) of canine uterus (Ishiguro *et al.*, 2007). **C:** Differences in MUC1 O-glycosylation in normal and cancer breast cells Short chained core-1 based glycans predominate in the cancer cells (Roy and Baek, 2002).

I.2.1.1.2.1 Importance of MUC1

The observation that null mice unable to produce MUC1 were more prone to infection and inflammation of their lower reproductive tract tissues suggests that MUC1 like many other mucins are important in conferring resistance against infection (DeSouza *et al.*, 1999, Carson *et al.*, 1998). MUC1 is also thought to play a role in reproduction in mice by being repressed under the influence of reproductive hormones to ease embryo implantation (Carson *et al.*, 1998, Brayman *et al* 2004). Phosphorylation of the cytoplasmic domain of MUC1 is also possible, an indication that the molecule may also be involved in signal transduction across cells (Zrihan-Licht *et al.*, 1994, Hanisch and Muller., 2000).

I.2.1.1.3 MUC2 – an example of a predominantly secreted mucin

Mucin 2 is a secreted gel-forming mucin produced by goblet cells in respiratory and gastrointestinal tracts (Andrianifahanana *et al.*, 2006, McGuckin *et al.*, 2011 and Figure I.3). The encoded protein is about 5100 amino acids long (Allen *et al.*, 1998) and just like MUC1 contains several domains including glycosylated PTS containing VNTR domains (McGuckin *et al.*, 2011). It is however structurally more complex than MUC1, e.g in terms of glycosylation and the formation of secondary structures (Larsson *et al.*, 2009). The molecular weight of a MUC2 polymer as result can get as high as 2.5 MDa (Larsson *et al.*, 2009).

Secreted MUC2 contains about 11 protein domains, 7 of which are Von Willebrand factor (vWF)-like domains (Figure I.8). vWF is a large multimeric glycoprotein present in blood plasma and known to play an important role in blood clotting by binding to blood clotting proteins such as factor VIII and mediating the adhesion of blood platelets to wound sites (Sadler, 1998, Ruggeri and Ware 1993). A deficiency of the protein in humans indeed leads to a bleeding disorder termed von willebrand disease (VWD). The functions of the various vWF -like domains in MUC2 are however unclear but terminal vWF like domains are likely involved in the formation of complex MUC2 structures through the formation of C-terminal disulphide-held dimers as is the case with the vWF glycoprotein (Ruggeri and Ware, 1993, Asker *et al.*, 1998). These together with disulphide linked N-terminal trimers significantly contribute to the complex MUC2 gel network of the colon (Godl *et al.*, 2002, Johansson *et al.*, 2011). Disulphide linkages contributed by cysteine rich CK and CysD (Cys 1 and Cys 2) domains (Figure I.8) also play an important role in the formation of high order/complex polymeric MUC2 structures (Ambort *et al.*, 2011, Bell *et al.*, 2001, 2003).

Once secreted, in the colon, MUC2 forms two major gel layers an outer loosely adherent and an inner tightly packed layer (Figure I.1). The outer layer is home to resident or commensal microflora while the inner layer is generally thought to be sterile or void of microbes (Johansson *et al.*, 2011). MUC2 thus provides important protection for underlying epithelia against resident and foreign bacteria.

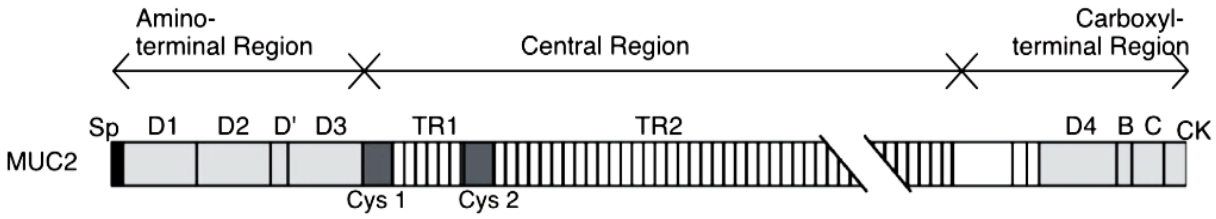


Figure I.8 - Structural features of human MUC2 apomucin. SP: Signal peptide B, C, D1, D2, D3, and D4, D': Von Willebrand factor-like domains TR1 and TR 2: Tandem repeat domains, CK: cystine knot domain. Cys1, Cys2: cysteine rich domains (Rousseau *et al.*, 2004, Ambort *et al.*, 2011)

A list of MUC2 glycans MUC2 O-glycans of the sigmoid colon detected by mass spectrometry is provided in Appendix Table D.1. The data reveal a prominence of Core-3 (GlcNAcβ1-3GalNAc) and Core-5 of F antigen (GalNAcα1-3GalNAc) - based structures. MUC2 expression and O-glycosylation is altered in disease conditions such as ulcerative colitis and colorectal cancers (Aksoy *et al.*, 2000, Larsson *et al.*, 2011, Brockhausen, 2006). MUC2 indeed is known to be a major carrier of the colorectal cancer associated sialyl-Tn (Neu5Acα2-6GalNAc-O-Ser/Thr) and Tn (GalNAc-O-Ser/Thr) antigens; (Conze *et al.*, 2010, Larsson *et al.*, 2011, Brockhausen, 2006). The Sialyl-Tn antigen is equally highly expressed in most gastric, ovarian, breast and pancreatic carcinomas (Conze *et al.*, 2010).

N-linked glycans on the other hand occur in low abundance in MUC2 compared to O- linked glycans. Although they have received lesser attention, they are thought to be important in mucin folding, dimerization, maturation and signaling (Asker *et al.*, 1998, Bell *et al.*, 2003).

I.2.1.1.3.1 Importance of MUC2

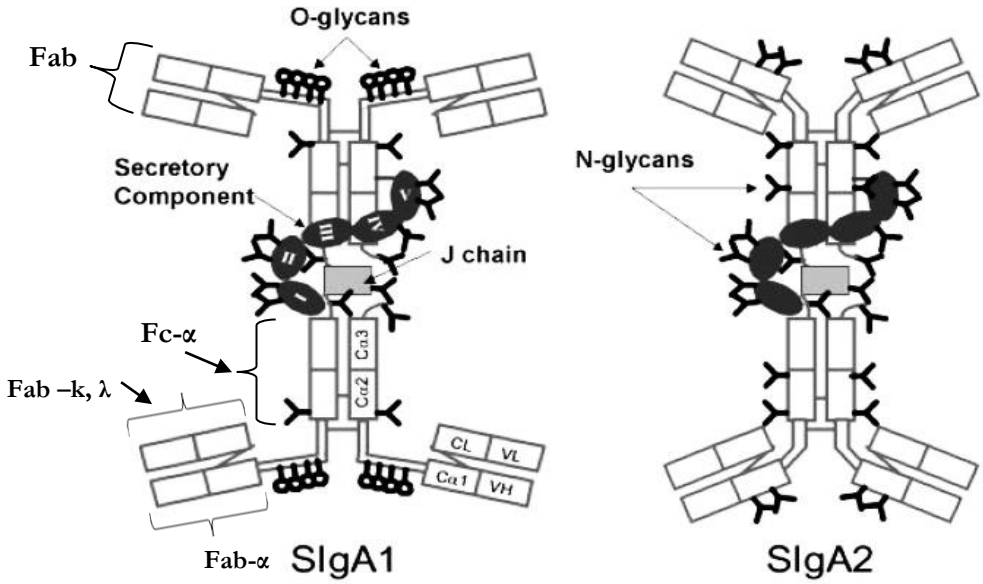
A major role for MUC2 is the physical protection of underlying epithelia, be it from bacteria, viruses or other harmful substances ingested into the gut lumen. The importance of MUC2 in this regard is also demonstrated in studies suggesting increased risks of colitis and colorectal cancers in muc2 deficient mice (Velcich *et al.*, 2002, Van der Sluis *et al.*, 2006). Secondly, MUC2 may serve an important role as a source of nutrients and an attachment site for both resident and foreign bacteria (Johansson *et al.*, 2011). As a result, it may also contribute in shaping the human gut microbiota composition (Koropatkin *et al.*, 2012).

I.2.1.2 Immunoglobulin A

IgA is a prominent component of mucous secretions and represents the predominant class of secreted immunoglobulins at mucosal surfaces (Underdown and Schiff, 1986, Morton *et al.*, 1993). IgA can exist in both monomeric (serum IgA) and dimeric forms [secretory IgA (SIgA)] with dimeric or secretory IgA containing additional components including the J-chain and secretory component (Figure I.9, Kazeeva and Shevelev, 2007). SIgA is the predominant form in mucous secretions playing a significant role in the protection of mucosa surfaces from invasion by mucosal microbes and other harmful environmental agents (Hurlimann and Darling, 1971).

There exists two major isotypes of the IgA molecule, namely IgA1 and IgA2 with varying distribution in various human tissues (Crago *et al.*, 1984, Morton *et al.*, 1993). The percentage of IgA1 in human serum can be as high as 90% while IgA2 levels may reach 50% in secretions (Delacroix *et al.*, 1982, Yoo and Morrison, 2005). IgA2 further exists in three different allotypic forms namely IgA2m(1), IgA2m(2) and IgA2m(n) (Torano and Putnam, 1978, Kazeeva and Shevelev, 2007). Both IgA1 and 2 isoforms however, show significant sequence identity to each other with the main difference being an additional mucin-like insertion sequence at the hinge region of IgA1 (Torano and Putnam, 1978) containing type-O glycosylations (Figure I.9, Yoo and Morrison, 2005).

A



B

Ser/Thr -*O*- GalNAc

Ser/Thr -*O*- GalNAc -*β1,3*- Gal

Ser/Thr -*O*- GalNAc -*β1,3*- Gal
α2,3
Sialic acid

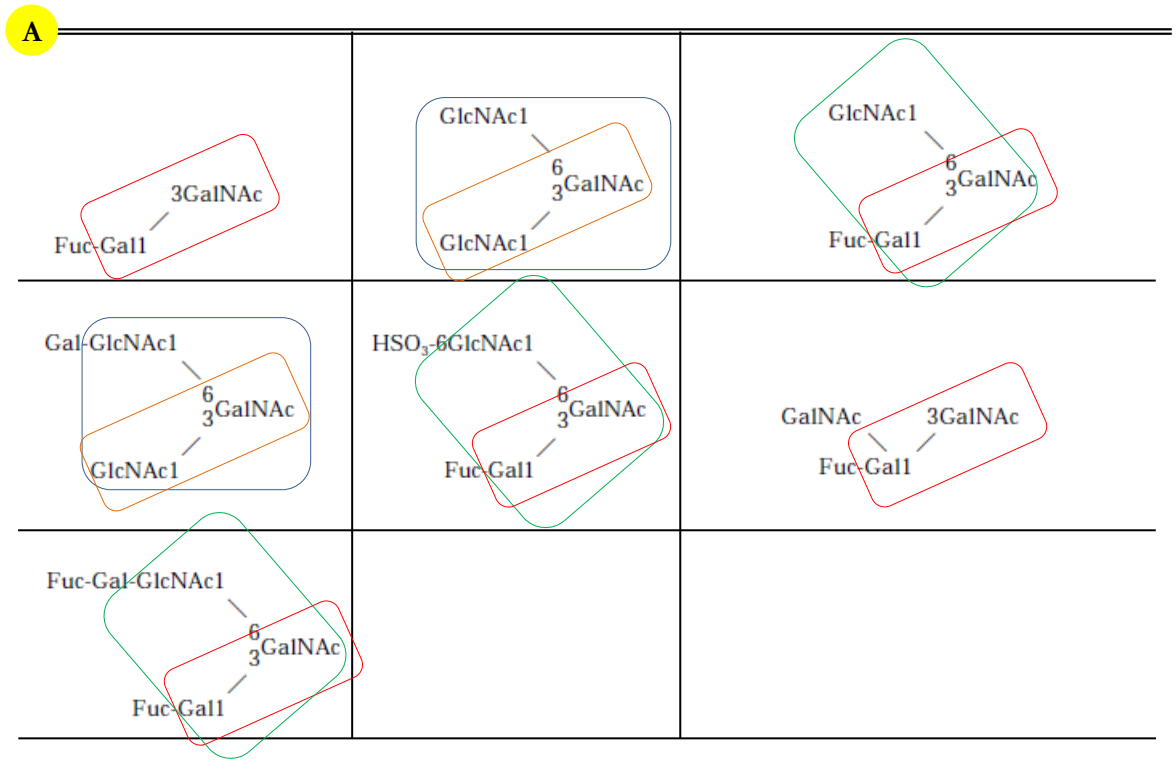
Ser/Thr -*O*- GalNAc -*β1,3*- Gal
α2,6 *α2,3*
Sialic acid **Sialic acid**

Figure I.9 - Structural features of human secretory IgA. A: Dimeric structures of SIgA1 and SIgA2. The IgA molecule consists of two identical chains termed heavy and light chains. The heavy chain is composed of four major domains namely the VH, Cα1, Cα2, and Cα3, while the light chain consists of two domains namely the VL and CL domains. V stands for domains within the variable region (amino acids in this region are variable in different variants of the immunoglobulin) of the molecule where the antigen binding (ab) site is located while C stands for domains in the constant region (identical in antibodies of same isotype) of the molecule. H and α indicates domains are present within the heavy chain (α being the greek denotation for the heavy chain in IgA) while L are for domains within the light chain. Lamda (λ) and kappa (k) denotes the various types of light chains in mammalian immunoglobulins. Following digestion of immunoglobulins with the enzyme papain, two major fragments (F) are produced one containing the antigen binding site (Fab) composed of the entire light chain (Fab-k,λ) and the variable VH and constant Cα1 region of the heavy chain (Fab-α) and the other containing the remaining two constant regions (Cα2 and Cα3) of the heavy chain (Fc-α) (Wang and Fudenberg, 1972). The secretory component is composed of five domains termed I, II, III, IV, and V which are N-glycosylated in IgA1 and IgA2 while the extended hinge insertion region of IgA1 absent in IgA2 is O-glycosylated. **B:** Four major O-linked glycan structures in human IgA1 (Modified from Royle *et al.*, 2003 and Novak *et al.*, 2008).

I.2.1.3 Porcine gastric and bovine submaxillary mucins

Substantial research has been carried out on porcine gastric mucins (PGM) in recent years as a model for human gastric mucins. This is partly due to anatomical and physiological similarities between the human and pig stomachs and PGM sequence similarity to human mucins such as MUC2 and MUC5AC (Celli *et al.*, 2005, Turner *et al.*, 1999). PGM is also a low cost substitute for human mucins capable of forming complex high molecular weight (~$9x10^6$ g/mol) polymeric structures similar to those of human MUC2 (Fiebrig *et al.*1995, Bansil and Tuner., 2006). A substantial amount of the PGM protein sequence contains VNTR domains with PTS repeats which are potential O-glycosylation sites (Turner *et al.*, 1999, Bansil and Tuner., 2006). Mass spectrometry data suggests that PGM O-linked glycans vary in length and complexity and are predominantly Core 1 and Core 2 - based with a few Core 3 and 4 structures (Figure I.10, Karlsson *et al.*, 1997, 2002).

Bovine submaxillary mucins (BSM) on the other hand are much less complex and contain shorter O-linked glycan structures. Mass spectrometry data on BSM glycans shows the predominance of sialylated O-linked GalNAc (Figure I.10, Tsuji and osawa 1986). The protein contains a few non-uniform repeat sequences and is phylogenetically related to the human and porcine MUC19 (Chen *et al.*, 2004).



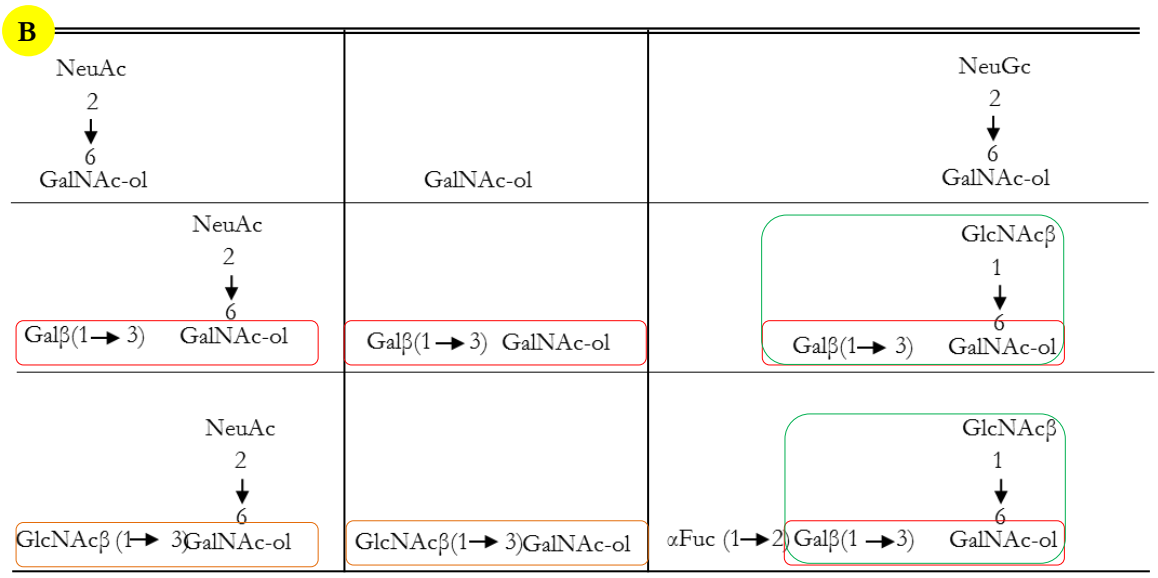
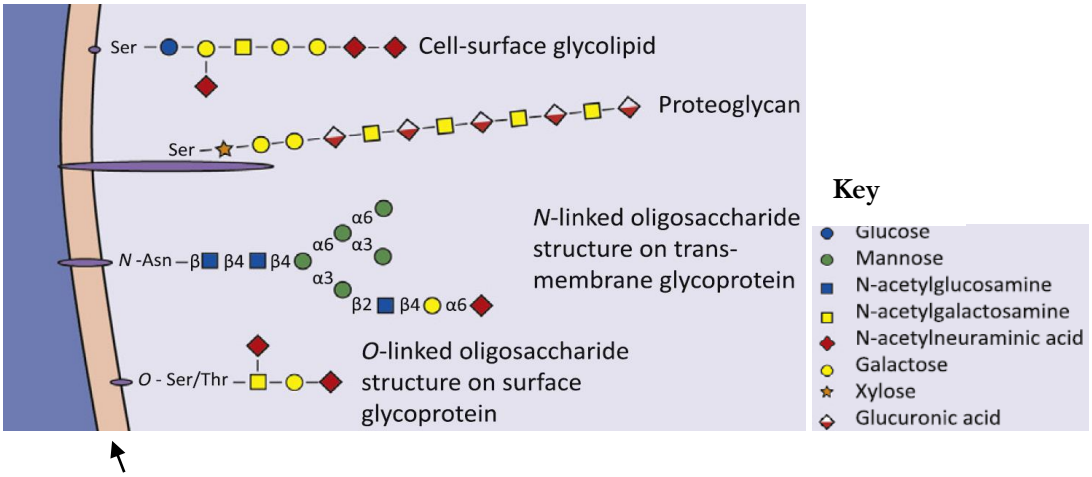


Figure I.10 – Prominent O-linked glycans of porcine gastric and bovine sub-maxillary mucins. A: Porcine gastric mucins (PGM) **B**: Bovine submaxillary mucins (BSM). Core 1 is indicated by red rectangles; core 2 by green rectangles, core 3 by brown rectangles and core 4 by blue rectangles (also see Figure I.4 for various core structures). Various structures were obtained from Tsuji and osawa 1986 and Karlsson *et al.*, 2002. See a list of other structures detected in mucins from different areas of the pig stomach in Karlsson *et al.*, 1997.

I.2.2 Epithelial cell surface glycocalyx

Transmembrane mucins such as MUC1 (Section I.2.1.1.2) alongside other epithelial cell surface anchored glycoproteins, glycolipids and proteoglycans constitute what is referred to as a glycocalyx (Figure I.11, Ouwerkerk *et al.*, 2013). These components are all glycosylated but are differentiated from each other depending on the nature of the glycans and what they are attached to (Varki and Sharon, 2009, Moran *et al.*, 2011). Glycolipids or glycosphingolipids contain sugars that are attached to a lipid ceramide backbone through glucose or galactose while proteoglycans like glycoproteins contain sugars attached to a protein backbone (Varki and Sharon, 2009, Moran *et al.*, 2011). Proteoglycans are distinguished from glycoproteins by the presence of repeating disaccharide units of acidic sugars termed glycosaminoglycans e.g. chondroitin sulfate/dermatan sulfate and heparan sulfate/heparin (Varki and Sharon, 2009, Moran *et al* 2011, Dick *et al.*, 2012).

A



Epithelial cell membrane

B

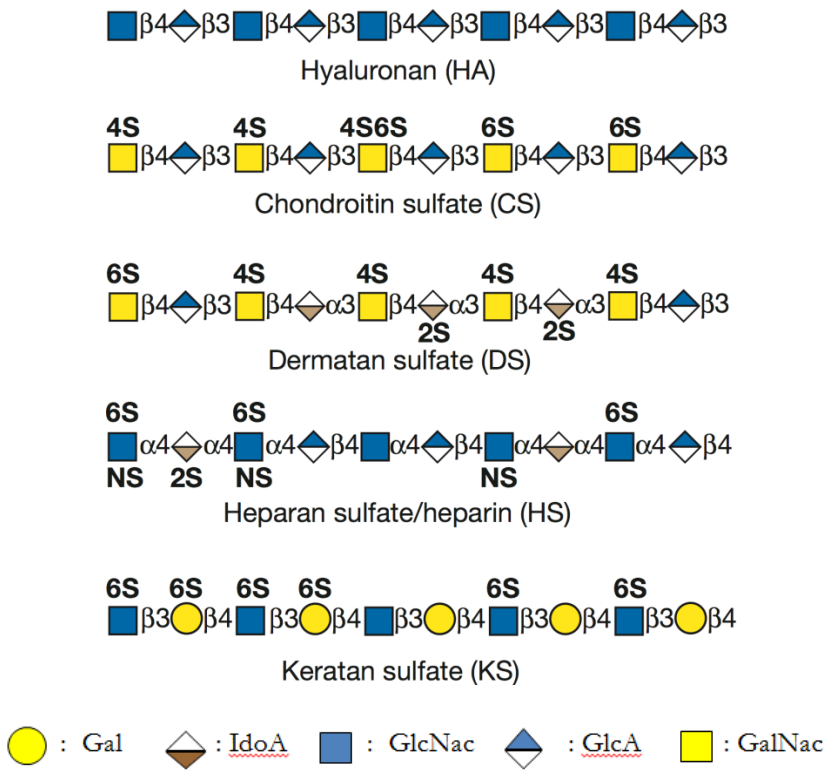


Figure I.11 – Major glycoconjugate classes within the epithelial cell surface glycocalyx and glycosaminoglycan diversity. A: Representation of the glycocalyx above a colonic epithelial cell (modified from Ouwerkerk *et al.*, 2013). **B:** Various glycosaminoglycan structures produced by mammalian cells. Except for hyaluronan, glycosaminoglycans are generally acidic or highly negatively charged due to sulphation (Esko *et al*, 2009, Varki and Sharon, 2009)

Glycocalyx components on epithelial cells can act as receptors for microbial adhesion, while the entire structure serves as a protective barrier against invading pathogens (Barnich *et al.*, 2007, Rousset *et al.*, 1998, Egberts *et al.*, 1984, Ouwerkerk *et al.*, 2013). The functions of the glycocalyx also include those performed by MUC1 (Section I.2.1.1.2.1) which is part of the glycocalyx. The glycocalyx thus represents an important second line of defense after the secreted mucus layer at mucosal surfaces (Varki and Sharon, 2009, Ouwerkerk *et al.*, 2013, Moran *et al.*, 2011).

I.3 Mucosal surfaces and Host-microbial interactions (HMI)

Mucosal surfaces play vital physiological roles in the human body including digestion, nutrient transport, homeostasis, reproduction and defense (Nataro *et al.*, 2005). They are also key players in host-microbial interactions, serving not only as initial entry points of microbes into the human body but also as the major point of contact between the human host and colonizing microbes.
Several aspects of mucosal surfaces including their exposure to the external environment, large surface area, the presence of adherent and sticky mucus and other protective secretions, nutrient availability, pH, temperature and other physiological factors (Johansson *et al.*, 2011, Sansonetti, 2004, Macpherson *et al.*, 2001, Hooper, 2009) make them particularly favorable to resident and foreign microbes. A significant number of colonizing microbes have also evolved strategies to enable them adapt and flourish at mucosal surfaces in otherwise unfavorable circumstances. As a consequence mucosal surfaces are host to a huge amount of diverse and complex microbial communities, with over 1,000 bacterial species (Qin *et al.*, 2010) thriving at the level of the gastrointestinal tract alone. Indeed over 100 trillion microbes, 10 times the total number of human somatic and germ cells put together [with a gene pool that exceeds the human gene pool by factor of 150 (Qin *et al.*, 2010)] inhabit the nutrient rich human intestinal tract (Savage, 1977, Ley *et al.*, 2006). This number is likely to be higher when microbes inhabiting other mucosal surfaces in the body such as the urogenital and respiratory tracts are taken into consideration. (Larsen and Monif, 2001, Zhou *et al.*, 2004, Konno *et al.*, 2006). Of particular interest however is the fact that there can be very serious health consequences as a result of our interactions with such large numbers of colonizing microbes. While the outcome of our interactions with mutualistic mucosal microbes is largely beneficial, HMI involving parasitic or pathogenic mucosal microbes are largely detrimental to the host as discussed in the subsequent section.

I.3.1 Overview of the importance of Host-microbial interactions to the human host

I.3.1.1 Mutualistic mucosal microbes

I.3.1.1.1 Host nutrition

The human gut microflora is thought to contribute about 10-15% of the daily calorie intake of its host (McNeil, 1984, Bergman, 1990). It is composed of several bacterial species capable of digesting otherwise indigestible dietary polysaccharides to short chain fatty acids that are easily absorbable by the host (Hooper *et al.*, 2002). Short chain fatty acids such as acetate, butyrate and propionate are known to enhance colonic health (Scheppach, 1994). Evidence also suggests that gut bacteria are capable of stimulating or enhancing the host's capacity to process some nutrients and xenobiotics (Claus *et al.*, 2011). Probiotic bacteria (live microbes that when consumed confer health benefits to the host) including members of the genus *Lactobacillus* often found in fermented milk products such as yoghurt have long been used to enhance lactose digestibility and assimilation in sufferers of lactose intolerance (Lin *et al.*, 1998). Finally, of recent, there have been reports of an association between obesity and gut microbiota composition (Ley *et al.*, 2005, Turnbaugh *et al.*, 2006, Zhang *et al.*, 2009), further demonstrating the importance of the gut microbiota in human nutrition and health.

I.3.1.1.2 Host immunity

Several lines of evidence suggest that mucosal microflora also play key roles in the development and modulation of host immunity. Studies involving the colonization of germ free mice with gut microflora for example revealed that these microbes including the *B. thetaiotaomicron* contribute to the development of the host immune system and angiogenesis, a process leading to the formation of new blood vessels from old ones (Hooper *et al.*, 2003). *B. thetaiotaomicron* is also capable of stimulating the secretion of SIgA in the murine intestine (Yanagibashi *et al.*, 2009), thus enhancing immunity to pathogens. The modulation of host immunity is important for the development of tolerance to frequently encountered foodstuffs, harmless environmental antigens and normal microflora (Braun-Fahrlander *et al.*, 2002). The tight association between an altered microbiota, allergies and chronic inflammatory bowel disease is evidence of their involvement in immunomodulation (Noverr and Huffnagle, 2005, Macfarlane *et al.*, 2009). Resident microflora may also enhance host immunity against harmful microflora through colonization resistance (Cebra *et al.*, 1999). Mutualistic bacteria including members of the *Bacteroidetes* and *Firmicutes* divisions constitute about 90% of the total human intestinal microflora (Eckburg *et al.*, 2005, Dethlefsen *et al.*, 2007, Tremaroli and Backhed, 2012) and such abundance may contribute to

restrict the proliferation of other bacteria including harmful bacteria. A detailed review into the role of the gut microbiota in host immunity is provided in Kamada *et al.*, 2013.
At the level of the female urogenital tract, mucosal microbes such as *L. acidophilus* play a role in the maintenance of a relatively low vaginal pH through the production of lactic acid (pH < 4.5), a condition that is thought to be hostile to many pathogenic bacteria (Boris *et al.*, 1998, Juarez Tomas *et al.*, 2003, Ronnqvist *et al.*, 2007). Some lactic acid bacteria also produce antimicrobial substances such bacteriocins and hydrogen peroxide (H_2O_2) which are equally important for the control of the vaginal ecosystem (Tomas *et al.*, 2004)

I.3.1.2 Pathogenic mucosal microbes

The outcomes of host microbial interactions involving pathogenic mucosal microbes unlike commensals are often detrimental to the host. This is not only by virtue of the infections these organisms cause but also, in some cases their ability to compromise the host immune system to other invading microbes and toxic substances.

Some common infections caused by mucosal microbes include amoebaic dysentery, caused by *Entamoeba histolytica* (a eukaryotic parasite), Trichomoniasis caused by *Trichomonas vaginalis* (also a eukaryotic parasite), diarrheal diseases caused by the *Escherichia coli* O157:H7 (a bacterial pathogen), Cholera infections caused by *Vibrio cholera* (a bacterial pathogen), and Typhoiditis caused by *Salmonella typhi* (a bacterial pathogen). These pathogens often possess unique capabilities and virulence factors that enable them cause this variety of infections.

It is also worth noting that although mutualistic relationships have been vastly looked upon as being beneficial, the quality of these relationships can be greatly affected by other factors including antibiotics, one's diet or health status (Maslowski and Mackay 2011). Any of these could lead to microbial dysbiosis which in turn can lead to serious pathophysiologies (Maslowski and Mackay 2011). Commensal microflora under certain conditions may also cause serious opportunistic infections e.g abdominal infections associated with members of the *Bacteroides fragilis* group (Goldstein *et al.*, 1996) and endocarditis caused by *Streptococcus sanguis* (Meddens *et al.*, 1982).

The above discussion is only a highlight of some of the important roles played by mucosal microflora in human health. Yet, despite the growing body of evidences, a lot about the mechanisms by which these organisms influence our health whether positively or negatively still remains unclear.

I.4 Membrane proteins of bacterial and eukaryotic mucosal microbes

Membrane proteins are broadly classified as either integral, peripheral or lipid anchored proteins (Singer, 1974, Reithmeier, 2001) depending on the nature of their association with the bacterial or eukaryotic cell membrane.

I.4.1 Lipid anchored Proteins

Lipid anchored proteins are widespread amongst gram positive and gram negative bacteria. A typical lipid anchor is the diacylglycerol group, often linked to a conserved N-terminal cysteine residue present within the protein. This group of lipid anchored proteins is termed lipoproteins. In most gram negative bacteria, the cysteine residue is part of what is referred to as the lipobox [LVI][ASTVI][GAS]C where it serves as the anchor point for the diacylglycerol group linking the entire protein to the lipid bilayer of the cell (Figure I.12A, Babu *et al.*, 2006, Kovac-simon *et al.*, 2011). Lipid anchored membrane proteins may face inwards into periplasm or out into the external environment (Figure I.12B)

Some proteins are attached to the lipid bilayer through a glycosyl-phosphatidylinositol molecule (GPI – anchored proteins). They represent a common group of lipid anchored proteins occurring with higher frequencies in eukaryotic species (Eisenhaber *et al.*, 2001, Brown and Wanneck, 1992, McConville *et al.*, 1993, Gerber *et al.*, 1992).

I.4.2 Integral membrane proteins

Integral membrane proteins form structures that traverse the lipid bilayer and hence contain a transmembrane region (Figure I.12C). The transmembrane segment often contains a high number of hydrophobic amino acids important for the proteins interaction with the hydrophobic lipid bilayer of the cell (Elofsson and von Heijne, 2007). It can take the form of α- helices or β-sheets/barrels (Ramasarma and Joshi, 2001).

I.4.3 Peripheral membrane proteins

Peripheral unlike integral membrane proteins are not directly linked to the membrane but rather interact with integral membrane proteins through polar or ionic interactions (Reithmeier, 2001). Just like lipoproteins, they could face into the periplasm or outward into the extracellular environment (Figure I.12C)

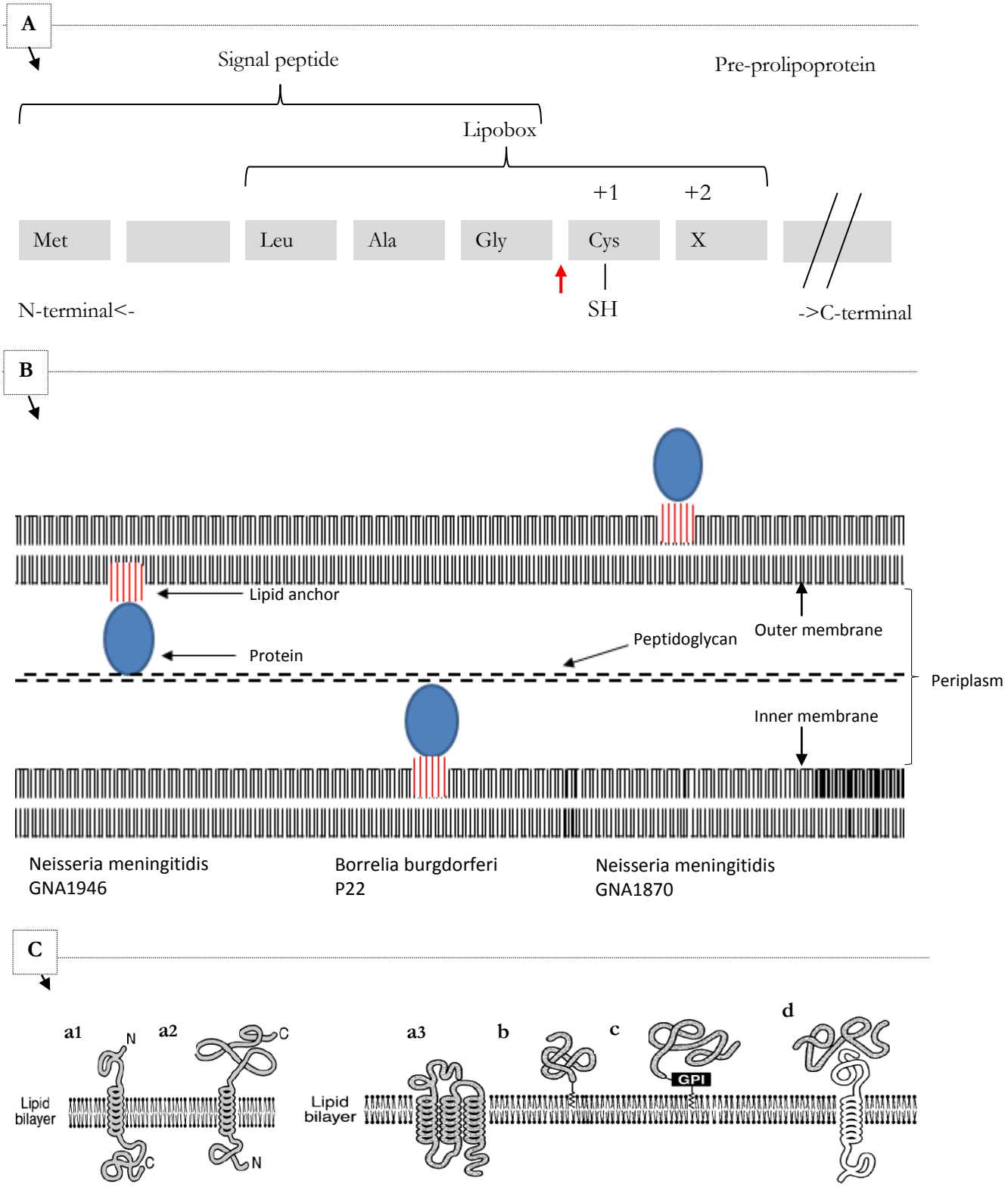


Figure I.12 - Membrane proteins of bacterial and lower eukaryotic microbes. A: Typical N-terminal features of bacterial lipoproteins showing the lipobox and the cysteine lipidation target. +1 and +2 indicate the position of amino acids relative to cleavage site of a signal peptidase enzyme (red arrow) **B:** Distribution of lipoproteins at the gram negative bacterial cell envelop. **C:** Summary of various classes of membrane proteins; **a1** and **a2**: Single pass transmembrane proteins, **a3**: multi-pass transmembrane protein, **b**: Lipid-chain anchored protein **c:** GPI-anchored membrane protein **d**: Peripheral membrane protein (Images modified from Kovac-simon *et al.*, 2011 and Chou and Elrod, 1999)

I.5 Surface exposed proteins of mucosal microbes

For the purpose of clarity, surface proteins are defined here as proteins emanating from the microbe into the extracellular environment. In mucosal microbes, this group of proteins represents an important interface in host microbial interactions, functioning as adhesins, metabolic enzymes, receptors or virulence factors.

As adhesins of mucosal microbes, surface proteins can mediate contact with mucosal surfaces by binding to mucus components such as mucins. An example is the 22-kDa surface-exposed Spr1345 protein of the respiratory tract pathogen *Streptococcus pneumoniae,* which has been shown to bind porcine gastric mucins and the surface of lung cancer cell lines (A549) *in-vitro* (Du *et al.*, 2011). This binding is thought to be mediated by a domain in the protein termed a mucin-binding domain (MucBD) which interestingly has also been identified (through homology searches) in surface proteins from other mucosal microbes such as *Lactobacillus lactis* and *Lactobacillus reuteri* (Du *et al.*, 2011). Spr1345 is thought to contribute towards microbial adherence and subsequent colonization of mucosal surfaces. Mucin binding surface proteins have also been reported in other members of the *Lactobacillus* genus that colonize the gut (Rojas *et al.*, 2002, Roos and Jonsson, 2002).

Surface proteins of gut mucosal microbes are also important tools for nutrient acquisition. They may function as metabolic enzymes, environmental sensing, signal transduction and transporter elements. They may operate as individual stand-alone elements in these roles or as parts of complex co-regulated systems, prominent examples of which are the polysaccharide utilization systems of the gut commensal *B. thetaiotaomicron* and its relatives (Martens *et al.*, 2009a, Sonnenburg 2010, Pope *et al.*, 2012, Mackenzie *et al.*, 2012).

Surface proteins have often been exploited as targets in vaccine development for their role as virulence factors of many pathogenic microbes including mucosal microbes (Kovacs-Simon *et al.*, 2011, Tai *et al.*, 2006, Jedrzejas *et al.*, 2001). As virulence factors, they may function as toxins, immunodulators, adhesins, and metabolic enzymes of the pathogens producing them. Examples of surface proteins with a role in virulence include the SpeB cell surface cysteine protease (also a secreted exotoxin) of *S. pyogenes* (Hytonen *et al.*, 2001), GP63 and CP65 surface proteases of *T. vaginalis* (Ma *et al.*, 2011, Alvarez-sanchez 2000) and a variety of other surface proteins of *S. pneumoniae* (Figure I.13).

Surface proteins thus represent important elements in host-microbial interactions involving both pathogenic and mutualistic mucosal microbes (Lebeer *et al.*, 2010)

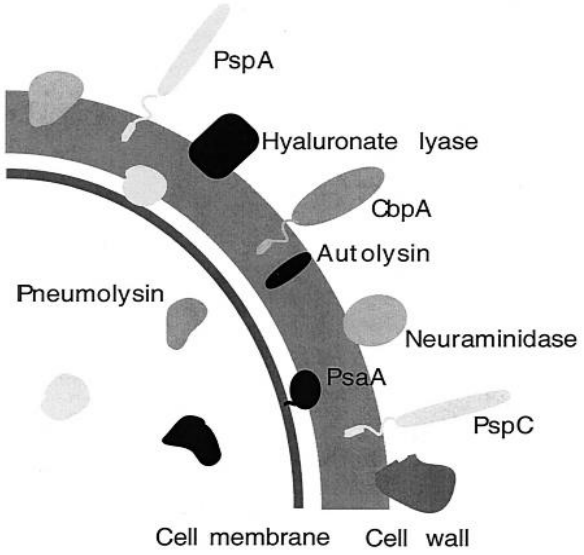


Figure I.13 Virulence factors of *S. pneumoniae* and their cellular localization. *S. pneumoniae* is a mucosal microbe and causative agent of pneumonia infections. As is the case with many other mucosal pathogens, several important virulence factors of *S. pneumoniae* are cell surface localized (Jedrzejas *et al.*, 2001)

1.6 *Bacteroides thetaiotaomicron*

Extensive studies on the prominent human gut commensal; *B. thetaiotaomicron* have greatly helped with our understanding of the molecular mechanisms underlying host-microbial interactions at mucosal surfaces. *B. thetaiotaomicron*, is an anaerobic non-sporulating Gram-negative bacterium and a member of the *Bacteroidetes* division, which in addition to the Firmicutes constitute over >90% of the resident human gut micro flora (Eckburg *et al.*, 2005, Dethlefsen *et al.*, 2007). It is known to colonize the colon and distal section of the small intestine and has been detected in human faecal samples (Hooper and Gordon 2001, Carson *et al.*, 2005). It is rod-shaped in nature (Figure I.14) and capable of producing an extracellular polysaccharide capsule (Martens *et al.*, 2009b).Close relatives of the organism which are equally predominant members of the GIT microbiota include *Bacteroides vulgatus, Bacteroides caccae* and *Bacteroides fragilis* (See scientific classification of *Bacteroides* species in Table I.3).

The entire 6.25MB genome has been sequenced and contains about 4779 predicted proteins in its proteome (Xu *et al.*, 2003). In-silico and gene transcriptional data suggest that the organism has

dedicated a significant portion of its genome to carbohydrate acquisition and utilization, capsular polysaccharide biosynthesis and DNA mobilization (Martens *et al.*, 2008, Xu *et al.*, 2003)

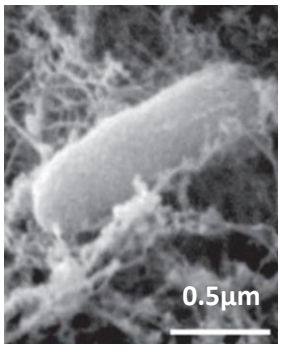


Figure I.14 - Scanning electron micrograph of *Bacteroides thetaiotaomicron* embedded in mucus. Adapted from Sonnenburg *et al.,* (2005)

Kingdom:	Bacteria
Phylum:	Bacteroidetes
Class:	Bacteroidetes
Order:	Bacteroidales
Family:	Bacteroidaceae
Genus:	*Bacteroides*
Species:	*Bacteroides thetaiotaomicron*

Table I. 3 Scientific classification of *B. thetaiotaomicron*

B. thetaiotaomicron is capable of metabolizing a variety of complex dietary polysaccharides some of which are otherwise indigestible by the host. Examples of these include starch, heparin, chondroitin sulphates and pectins (Koropatkin *et al.*, 2012). Studies also suggest that under conditions of limited dietary glycan supply, the organism can switch to the utilization of host derived glycans such as mucins, an ability which is thought to provide it with a significant competitive and survival advantage in the gastrointestinal tract (Martens *et al.*, 2008). Its ability to consume mucin glycans has been demonstrated in *in-vitro* growth experiments with porcine gastric mucins (PGMIII) (Benjdia *et al.*, 2011, Martens *et al.*, 2008). *B. thetaiotaomicron* can also influence nutrient acquisition and the production of defense components by the human host as highlighted in Section I.3.1.1.2 making it an important member of the gut microflora contributing to both host nutrition and immunity (Hooper *et al.*, 2003, Wexler, 2007, Yanagibashi *et al.*, 2009). Its contribution to mucosal homeostasis was recently demonstrated by Wrzosek *et al.*, (2013) who showed that it is capable of promoting globet cell differentiation and the production of mucus and sialylated mucins, effects which are attenuated by another prominent gut commensal *Faecalibacterium prausnitzii* (Miquel *et al.*, 2013), leading to the maintenance of colonic epithelial homeostasis.

B. thetaiotaomicron may cause opportunistic infections such as intra-abdominal sepsis and bacteremia in immune-compromised individuals (Goldstein, 1996, Redondo *et al.*, 1995). It is capable of developing resistance to β-lactams antibiotics such as Clindamycin and Cefotexin (Teng *et al.*, 2002, Edwards, 1997).

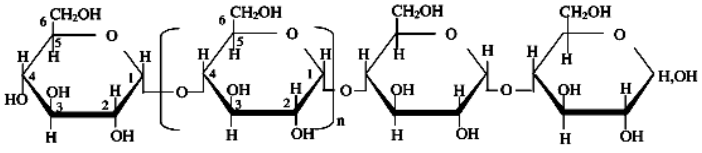
I.6.1 Surface proteins and polysaccharide utilisation loci (PULs) of *B. thetaiotaomicron*

Like many other mucosal microbes, several *B. thetaiotaomicron* surface proteins are involved in mucosal colonization and survival processes such as cell adhesion, (Rogemond and Guinet, 1986), signaling and nutrient acquisition (Xu, *et al.*, 2003). A good number of proteins encoded by *B. thetaiotaomicron* are putative surface proteins and elements of various polysaccharide utilization loci (PULs) (Xu, *et al.*, 2003, Martens *et al.*, 2009a). PULs are gene clusters in the genome of *B. thetaiotaomicron* encoding cell-envelop associated multi-protein complexes dedicated to carbohydrate utilization. Currently over 88 different polysaccharide utilization loci have been identified in the genome of *B. thetaiotaomicron* although detailed functional data is only available for two of them. These include the starch and fructan utilization loci encoding proteins of the starch utilization and fructan utilization systems of *B. thetaiotaomicron* (Martens *et al.*, 2009a, Sonnenburg 2010).

I.6.1.1 The starch utilization system (Sus) of *B. thetaiotaomicron*

Starch is a high-energy nutrient and an important component of the human diet. It is present in significant amounts in many staple foods including rice, wheat, corn, cassava and potatoes, consumed by human populations of both the developed and developing world. It consists of two alpha (α)-D glucose polymers termed amylose and amylopectin (Figure I.15). Amylopectin is usually in higher amounts compared to amylose and consists of α1-4 linked glucopyranosyl units with branching α1-6 bonds linking the backbone to other glucose units (Buleon *et al.*, 1998). Amylose on the other hand is a less complex linear polymer of alpha 1-4 linked glucopyranosyl units, with only occasional α1-6 branch points (Tester *et al.*, 2004, Buleon *et al.*, 1998).

Amylose



Amylopectin

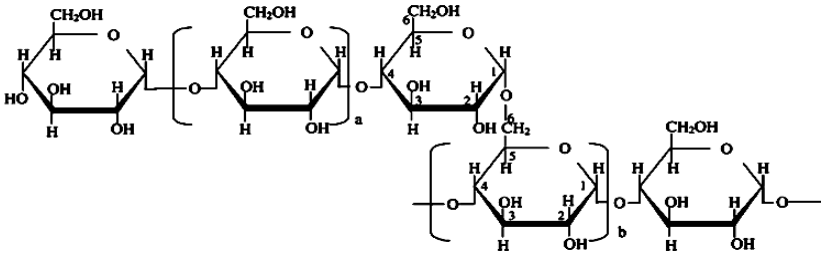
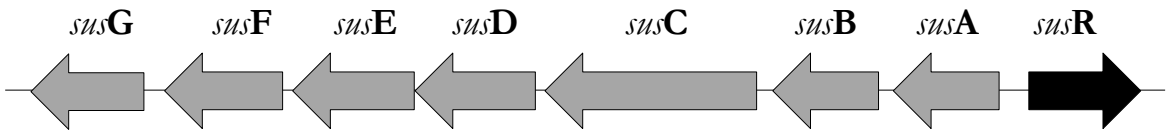


Figure I.15 - Structure of starch glycans. Adapted from Tester *et al.*, (2004)

The digestion of dietary starch in the human gut is made possible thanks to salivary, pancreatic and intestinal enzymes. Salivary and pancreatic starch degrading enzymes are mainly α-endo-amylases (E.C. 3.2.1.1) whose action on starch yields short linear and branched oligosaccharide structures that are later on degraded in the small intestine by brush border exo-acting maltase-glucoamylase (E.C. 3.2.1.20 and 3.2.1.3) and sucrase-isomaltase (E.C.3.2.1.48 and 3.2.1.10) enzymes (Ao *et al.*, 2007, Swallow, 2003).

Starch degradation up till the small intestine often does not go to completion and some escape into the large intestine or colon. These are referred to as "resistant starch" (Englyst and Mcfarlane, 1986, Vonk *et al.*, 2000, Carciofi, 2012) and alongside other undigested plant cell wall polysaccharides such as celluloses and pectins can be fermented by colonic mucosal microbes including *B. thetaiotaomicron* to produce short chain fatty acids such as butyrate, acetate and propionate (Cummings, 1981). Short chain fatty acids are easily absorbable by the human host and confer important health benefits to the host (Scheppach, 1994)

The starch utilisation locus contains a total of eight genes encoding various components of the starch utilisation system in *B. thetaiotaomicron* (Martens *et al.*, 2009a). A list of the PUL genes (*sus*R, A, B, C, D, E, F and G) and their corresponding locus tags is provided in the key of Figure I.16. It has a GC content of about 45% and occupies a 15.821 kbp region in the genome of the organism. The genes are organized into two transcriptional units (the first containing the *sus*A gene alone and the second from *sus*B to G, all under the regulation of *sus*R (Cho *et al.*, 2001). All the *sus* structural genes are unidirectionally transcribed.



Key

Gene name	Gene locus	Annotation of encoded protein
*sus*G	BT_3698	Alpha-amylase (Glycoside Hydrolase Family 13)(SusG)
*sus*F	BT_3699	Outer membrane protein(SusF)
*sus*E	BT_3700	Outer membrane protein(SusE)
*sus*D	BT_3701	SusD
*sus*C	BT_3702	SusC
*sus*B	BT_3703	Glycoside Hydrolase Family 97(SusB)
*sus*A	BT_3704	Alpha-amylase (Glycoside Hydrolase Family 13) (SusA)
*sus*R	BT_3705	Transcriptional regulator(SusR)

Figure I.16 - Genes of the starch utilisation locus of *B. thetaiotaomicron*. Top: Genomic organisation of PUL genes (i.e. genes encoding various components of the Sus system). Bottom: List of Sus genes and their annotations as in Martens *et al*., 2008. See list of glycoside hydrolase families in section I.9 and the activities of various classes on the CAZy database at http://www.cazy.org/ (Cantarel *et al.,* 2009)

The SusR protein contains two transmembrane domains enabling it to span the inner membrane from the periplasmic to the cytoplasmic side (Figure I.18). In the presence of starch or maltose, SusR activates the expression of various Sus proteins to enable the acquisition and utilisation of the target substrate (Cho *et al*., 2001).

SusD, E, F, and G all contain a type II signal peptide and are located to the outer membrane of the cell (Reeves *et al*., 1997, Cameron *et al*., 2012). SusC is also located to the outer membrane of the cell and is a member of the Ton-B dependent receptor family of proteins (Reeves *et al*., 1996, 1997, Koropatkin 2012, Ferguson and Deisenhofer, 2002). SusD, E and F have been shown to bind starch oligosaccharides such as maltohepatose and cyclodextrins (Koropatkin 2008, Cameron *et al*., 2012). SusE and F are thought to achieve this with the aid of distinct carbohydrate binding modules in their structures as shown in Figure I.17. Together, the concerted action of the SusC, D, E, and F enables the organism to bind starch and starch oligosaccharides to its surface (Reeves *et al*., 1997).

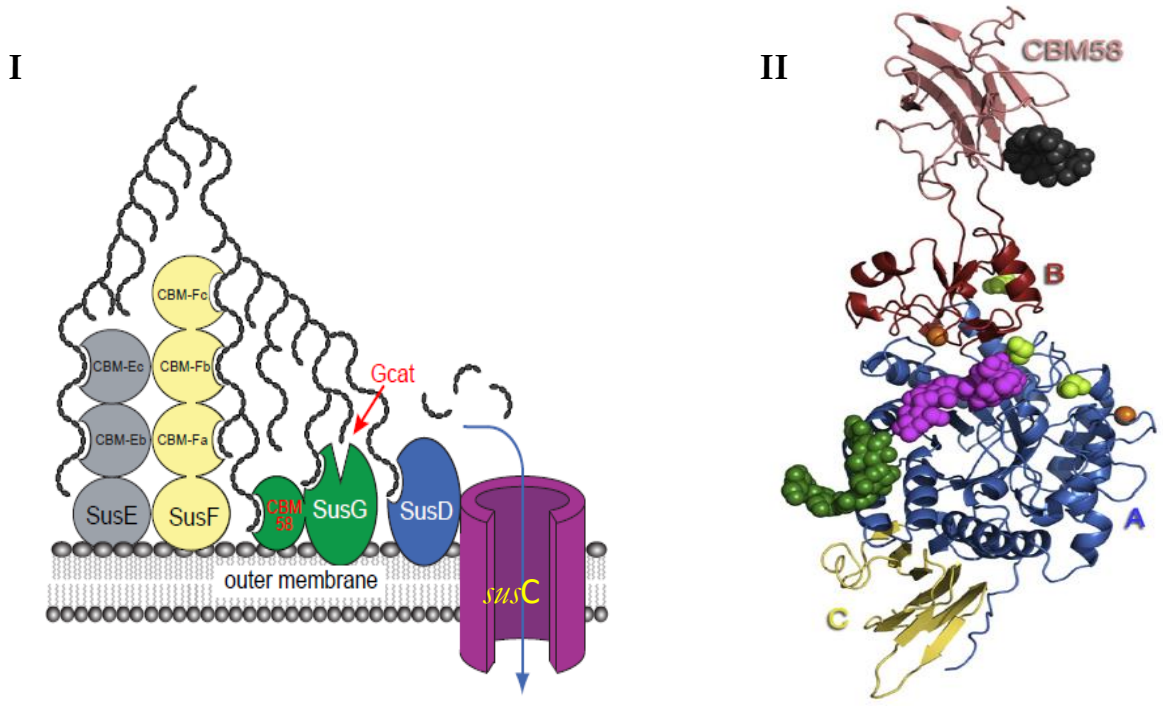


Figure I.17 – The surface carbohydrate binding apparatus of the Sus system of *B. thetaiotaomicron*. **I:** Schematic representation of the outer surface carbohydrate binding apparatus of the Sus system. Starch binding is a concerted action of outer membrane proteins SusE, F, G, D and C. SusE, F and G achieve this through defined carbohydrate binding domains in their structures. Unlike SusG, SusC, D, E and F are not currently known to possess any enzymatic activities. SusG is an endo-acting amylase enzyme with a CBM58 domain and a catalytic site (Gcat). Sugar import into the cell is achieved through the TonB-dependent SusC transporter **II:** Structure of the SusG protein showing constituent domains (domain A - blue, domain B - red, domain C - yellow) including the CBM58 domain bound to molecules of maltoheptaose (gray, mauve and green spheres). Images adapted from Cameron *et al.*, (2012) and Koropatkin *et al.*, (2010).

SusG on the other hand not only contains a carbohydrate binding module in its structure (CBM58) but also displays endo- alpha amylase activity (Shipman *et al.*, 1999 and Martens 2009). It is capable of hydrolysing glucose α1-4 linkages in the bound starch polysaccharides yielding shorter starch oligosaccharides [although predicted to be at least longer than maltotriose (Martens *et al.*, 2009)] that are later imported into the cell through the TonB-dependent SusC porin (Figure I.17 and I.18). TonB-dependent proteins are typically outer membrane spanning β-barrels that couple energy from the proton motive force and an inner membrane TonB-ExbBD complex to the transport of solutes and macromolecules through the outer membrane of the cell (Ferguson and Deisenhofer, 2002). SusC likely employs a similar mechanism to import bound starch oligosaccharides into the periplasmic space (Koropatkin *et al.*, 2012).

SusA and B proteins both contain a type I signal peptide in their structure and hence predicted to be periplasmic (Figure I.18). SusA like SusG also belongs to the GH13 family of glycoside hydrolases; although unlike SusG is an exo-acting neopullulanase (D'Elia and Salyers, 1996, Martens 2009a). In conjunction with SusB (a GH97 enzyme with α-glucosidase activity) both further degrade the SusG - processed oligosaccharides imported into the periplasmic space through SusC (D'Elia and Salyers, 1996, Koropatkin *et al.*, 2012, Martens 2009a).

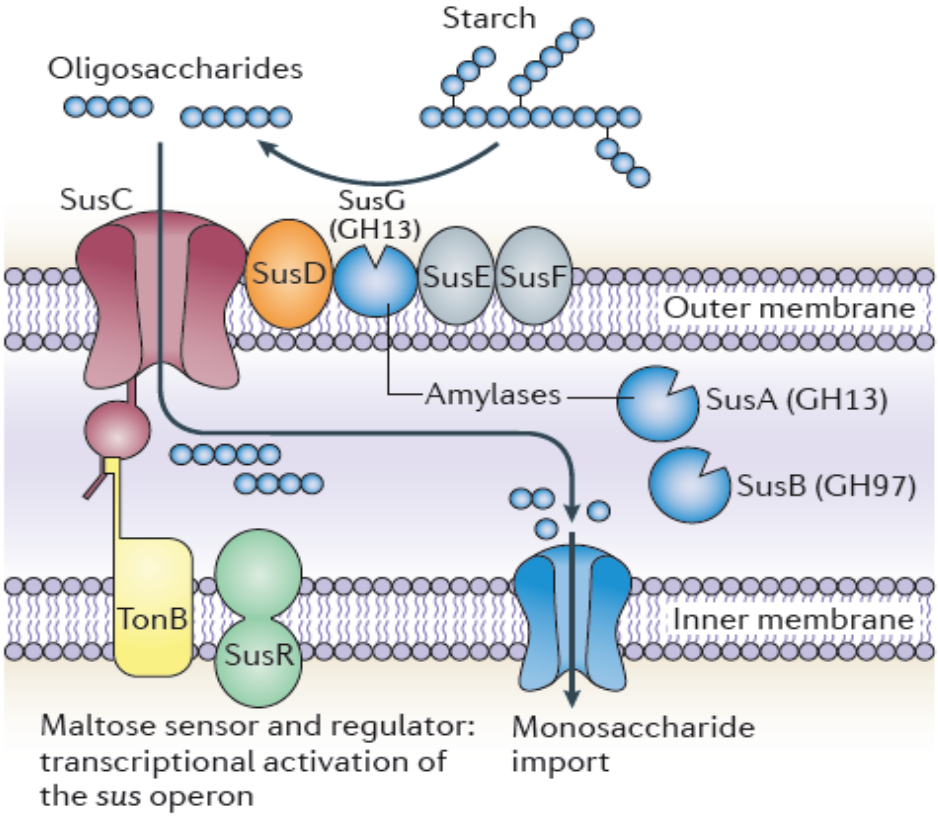


Figure I.18 - Schematic representation of the Sus system of *B. thetaiotaomicron*. Bound starch (Figure I.17) is initially processed by the outer membrane SusG endo-acting amylase and transported into the periplasmic space with the aid of the SusCD complex. Periplasmic exo-acting enzymes SusA and B further act on imported oligosaccharides yielding monosaccharides that are imported into the cytoplasmic space through an inner membrane transporter (adapted from Koropatkin *et al.*, 2012). A detailed insight into the mechanism of the Sus system is also provided in Martens *et al.*, 2009.

I.6.1.2 The fructan utilization system (Fus) of *B. thetaiotaomicron*

Fructans are polymers of fructose, derived from plant carbohydrates such as Inulin and Levan. Inulin and levan contain polymers of β2-1 and β2-6 fructose units respectively. These are known to be resistant to host digestive enzymes in the upper gastrointestinal tract and hence constitute an important source of dietary glycans for gut bacteria in the colon (Sonnenburg *et al.*, 2010).

A fructan utilization locus has been identified (Figure I.19, Sonnenburg *et al.*, 2010) in *B. thetaiotaomicron* and other *Bacteroides* species containing a host of genes encoding enzymes, carbohydrate recognition and regulatory proteins that seem to coordinate the acquisition and utilization of fructans in a manner analogous to the mechanism of the Sus system (Figure I.18).

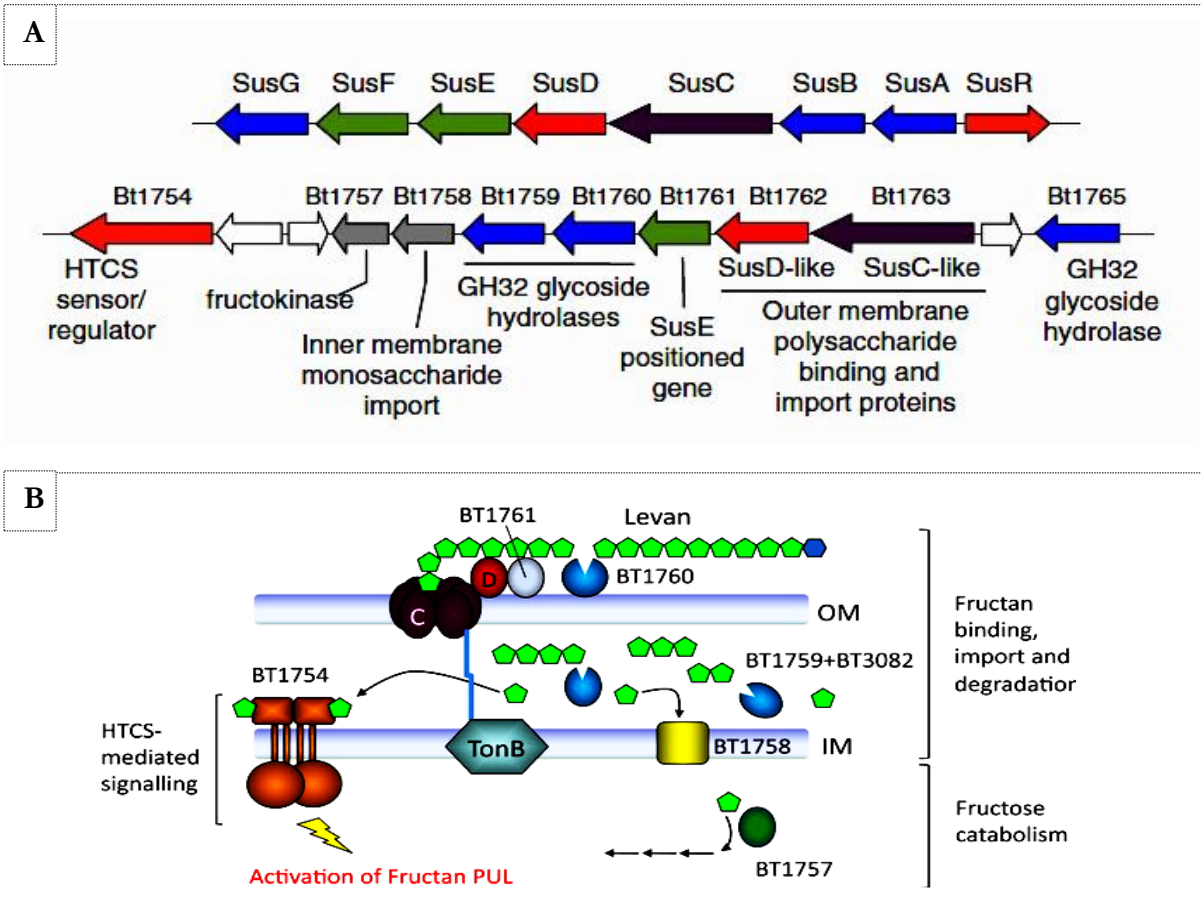


Figure I.19 - The fructan utilization system (Fus) of *B. thetaiotaomicron*. **A:** Comparison of the Sus and Fus loci. Genes are colour coded to indicate broad functional similarities **B** - Schematic representation to show cellular localisation of Fus elements and their role in fructan utilisation. In the mechanism that follows, bound fructan on the surface of the cell is endolytically cleaved by the BT1760 endo-fructanase. This is later imported through the susC homologue (BT1763) into the periplasmc space containing GH32 enzymes that make further cuts in the molecules. While some of the released fructose end products are imported into the cell cytoplasm, some are used to further activate the PUL through binding to the BT1754 inner membrane hybrid two-component system (HTCS) regulator protein (All images were taken from Sonnenburg *et al.*, 2010)

I.7 *T. vaginalis*

T. vaginalis is an anaerobic eukaryotic protozoan and causative agent of human Trichomoniasis, a sexually transmitted disease (STD) of the urogenital tract that annually affects millions of individuals worldwide (Gerbase *et al.*, 1998, Schwebke and Burgess., 2004). The active, motile feeding stage of the organism (trohpozoite) is pyriform or pear-shaped in nature but can also become amoeboid following cytoadherence (Petrin *et al.,* 1998, Harp and Chowdhury, 2011). The organism is flagellated and contains a large nucleus and other internal organelles such as the axostyle, costa and pelta (Figure I.20). *T. vaginalis* acquires nutrients from vaginal secretions, phagocytosed bacteria, vaginal epithelial cells (VECs) and erythrocytes (Juliano *et al.*, 1991, Rendon-Maldonado *et al.,* 1998, Seema and Arti, 2008). It lacks mitochondria and peroxisomes and instead contains unusual double membrane bound energy-producing organelles known as hydrogenosomes (Schneider *et al.,* 2011). Hydrogenosomes are involved in carbohydrate metabolism, producing energy (ATP), acetate, carbon dioxide and hydrogen as end products from pyruvate and malate substrates (Muller *et al.*, 1993). They are found in diverse anaerobic eukaryotic microbes and are a typical feature of the parabasalid lineage to which the *T.vaginalis* belongs (Embley *et al.*, 2003, Malik *et al.*, 2011) (see scientific classification of *T. vaginalis* in Table I.4).

The life cycle of the organism consists of two stages, mainly the infective and diagnostic stages (Figure I.20). During the diagnostic stage, *T. vaginalis* is detectable in vaginal, prostatic secretions and urine. Following longitudinal binary fission, more trophozoites are produced and concentrate in the vagina or orifice of the urethra before transmission by sexual intercourse (Harp and Chowdhury, 2011). In women, *T. vaginalis* infections can cause severe vaginal inflammation and irritation and has also been linked to cervical cancer (Petrin *et al.*, 1998). In males, the infection is generally asymptomatic and can lead to urethritis, prostatitis and prostate cancer (Kuberski, 1981, Abdolrasouli *et al.*, 2007, Sutcliffe *et al.*, 2010). In both men and women, *T. vaginalis* infection can lead to infertility and predisposition to human immunodeficiency virus (HIV) infection (Sorvillo and Kerndt, 1998).

Kingdom:	Protista
Phylum:	Metamonada (Parabasala)
Class:	Parabasalia
Order:	Trichomonadida
Family:	Trichomonadidae
Genus:	*Trichomonas*
Species:	*Trichomonas vaginalis*

Table I.4 - Scientific classification of *Trichomonas vaginalis*

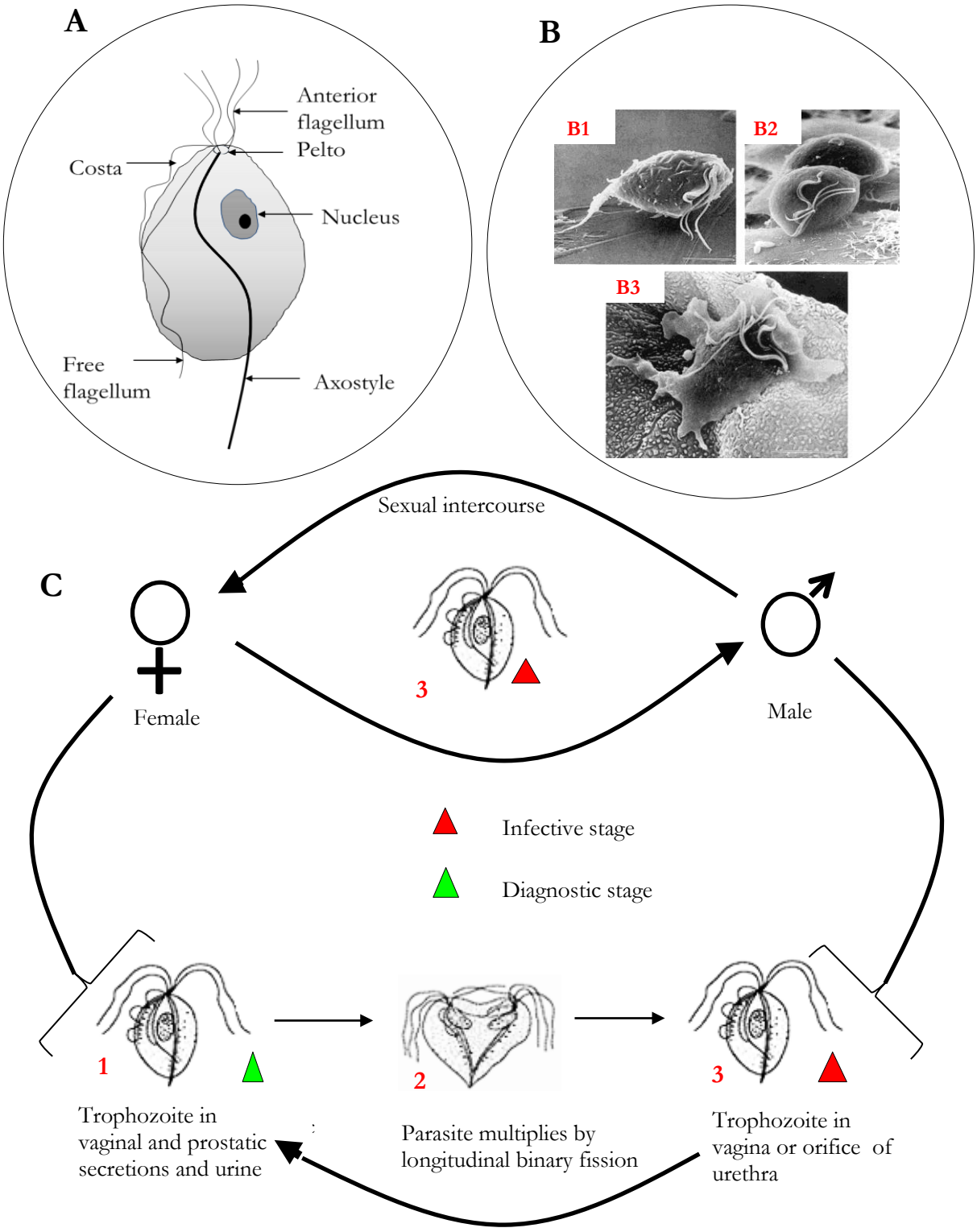


Figure I.20 - Overview of *T. vaginalis* characteristics. **A:** Morphological features **B:** Electron micrography images. B1**:** *T. vaginalis* in broth culture. **B2:** *T. vaginalis* in contact with vaginal epithelial cells. **B3:** Example of an amoeboid form of *T. vaginalis*. **C:** Life cycle of *T. vaginalis* (Images modified from Petrin *et al.*, 1998, Harp and Chowdhury, 2011 and the DPDx/CDC America website)

I.7.1 Surface proteins of *Trichomonas vaginalis*

The entire 160 Mb genome of *T. vaginalis* strain G3 has been sequenced and contains ~ 60 000 protein coding genes according to data from the GiardiaDB and TrichDB databases (GiardiaDB and TrichDB, Carlton *et al.*, 2007, Harp and Chowdhury, 2011). About 65% of the genome is repetitive composed of virus, transposon, retrotransposon and unclassified repeat DNA sequences. The genome contains a large repertoire of genes consistent with carbohydrate and amino acid metabolism, defense against oxidative stress, transport and pathogenesis (Mendoza-lopez *et al.*, 2000, Carlton *et al.*, 2007, Hirt *et al.*, 2007, de Miguel *et al.*, 2010). A total of 3000 candidate surface-exposed proteins from ten different protein families (major cartegories of which were the BspA-like, GP63-like proteins and adhesins) have been identified from the sequenced genome and some have already been biochemically characterized (Hirt *et al.*, 2007, Harp and Chowdhury, 2011).

As a pathogenic mucosal microbe, *T. vaginalis* surface proteins with adherence and metabolic functions represent important virulence factors. An example is the 30 kDa CP30 surface cysteine proteinase produced by the organism which has been shown to be capable of degrading female urogenital tract proteins as well as adhering to HeLa cervical carcinoma cell lines (Mendoza-lopez *et al.*, 2000). The recently characterized TvGP63 protease is also another example of a surface localized virulence factor of *T. vaginalis* (Ma *et al.*, 2011). TvGP63 whose protease activity is inhibited by the cysteine proteinase inhibitor; 1, 10 phenathroline is thought to play a role in epithelial cell destruction during *T. vaginalis* infection (Ma *et al.*, 2011). Surface proteins alongside the lipophosphoglycans in the glycocalyx of the *T. vaginalis* cell play a crucial role in mediating *T. vaginalis* cytoadherence (Bastida-Corcuera., 2005, Harp and Chowdhury, 2011).

I.7.2 Diagnosis and Treatment of Trichomoniasis

T. vaginalis is detectable using unsophisticated microscopic techniques due to its size. Microscopic evaluation is however usually less sensitive compared to the wide range of tests that have recently been developed for *T. vaginalis* diagnosis (see comparison in Harp and Chowdhury, 2011). The rapid antigen test that detects *T. vaginalis* membrane proteins and a variety of nucleic acid based tests are examples of some novel diagnostic strategies that offer high sensitivity and specificity (Harp and Chowdhury, 2011). Current treatment for Trichomoniasis is done with the 5-nitroimidazole compounds; Metronidazole and Tinidazole. The limited number of drug treatments for the disease coupled with evidence of resistance to some drugs (Schmid *et al.*, 2001) has prompted research for alternative therapeutic strategies.

I.8 M60-like domain-containing proteins

Published genome sequences of important mutualistic and parasitic mucosal microbes such as *B. thetaiotaomicron* (Xu *et al.*, 2003), and *T. vaginalis* (Carlton *et al.*, 2007) respectively, provide exciting opportunities to study the role of surface proteins in host-microbial interactions. Following comparative genomic studies, these organisms were found to share a number of gene families including candidate surface metabolic enzymes (Nakjang *et al.*, 2012). One set of such genes characterized through bioinformatic analyses was identified as candidate zinc metalloproteases sharing a novel protein domain termed "M60-like (PF13402) domains" (Nakjang *et al.*, 2012). They were so named by virtue of their profile similarity to an existing protein family termed family M60 [MEROPS database, (Rawlings *et al.,* 2012)], epitomized by the insect baculovirus (*Lymantria dispar nucleopolyhedrovirus)* enhancin protease also known to be capable of degrading insect intestinal mucins (Wang and Granados, 1997, Nakjang *et al.*, 2012).

M60-like domain-containing proteins show a broad taxonomic distribution as they are shared by a large number of bacterial and eukaryotic microbes [PFAM database, (Punta *et al.,* 2012)]. Curiously, a significant proportion of M60-like positive organisms are important mucosal microbes including in addition to *B. thetaiotaomicron* and *T. vaginalis*, other human mucosal microbes such as *Bacteroides fragilis, Bacteroides caccae, Bacillus anthracis, Clostridium perfringens, Vibrio cholera, Entamoeba histolytica, Cryptosporidium species.* suggesting that these proteins might play an important role in the biology of mucosal microbes as well as in host microbial interactions.

I.8.1 M60-like domain-containing proteins are putative gluzincin family proteases

I.8.1.1 Overview of proteases and classification

Proteases are enzymes capable of hydrolyzing peptide bonds in protein substrates. They are generally classified as serine, threonine, cysteine, aspartate, glutamate or metalloproteases as on the MEROPS online database (http://merops.sanger.ac.uk/) for proteolytic enzymes (Rawlings *et al.*, 2012). The MEROPS database also includes mixed and unknown classes. Protease classes are divided into clans or superfamilies which can be further divided into families (MEROPS). Their classification takes into consideration amongst other factors, evolutionary relationships, protein structure and sequence homology, active site amino acids, and their cofactor requirements (Mansfeld *et al.*, 2007, Gomis-Rüth, 2003). Proteases can also be endo- or exo-acting depending on the position of the peptide bond in the protein that they target (Figure I.21). Endo-acting proteases cleave internal peptide bonds while exo-acting proteases cleave terminally located peptide bonds (Dos Santos, 2011).

Proteases perform vital physiological and pathological functions in microbial systems (Figure I.21), including serving as nutritional and virulence factors (Miyoshi and Shinoda, 2000, Ruiz-Perez *et al.*, 2011, Szabady *et al.*, 2011, Mendoza-lopez *et al.*, 2000, Franco *et al.*, 2005, Dos Santos, 2011) .

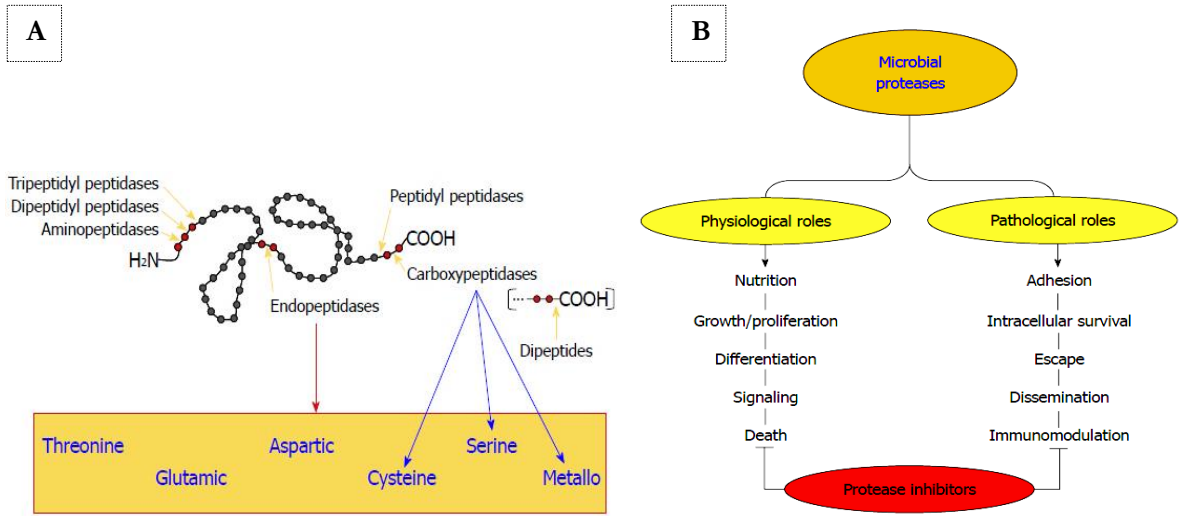


Figure I.21 – General classification of proteases (A) and an overview of their role in microbial systems (B). Images were adapted from Dos Santos, (2011).

I.8.1.1.2 Metalloproteases

As the name implies these are protease requiring metallic cofactors for their activity. A popular co-factor for this group is zinc and metalloproteases requiring zinc for their activity are generally termed zinc metalloproteases. A significant number of the known zinc metalloproteases contain the HEXXH motif or consensus sequence (X= any amino acid) in their catalytic sites (Figure I.22) and belong to the zincin superfamily or clan *(*Miyoshi and Shinoda, 2000, Gomis-Ruth, 2003, Levine, 2011, Coleman, 1998). The two histidine residues in the consensus perform the function of coordinating the catalytic zinc ion (Zn^{2+}) in the active site of the enzyme while the glutamic acid (E) residue serves as the catalytically active amino acid (Fukasawa *et al.*, 2011, Ramos *et al.*, 2001). A third Zn^{2+} binding ligand differentiates the zincin clan further into subclans e.g gluczincins (E), aspzincins (D) or metzincins (H/D) (Gomis-Rüth, 2003, Levine, 2011).

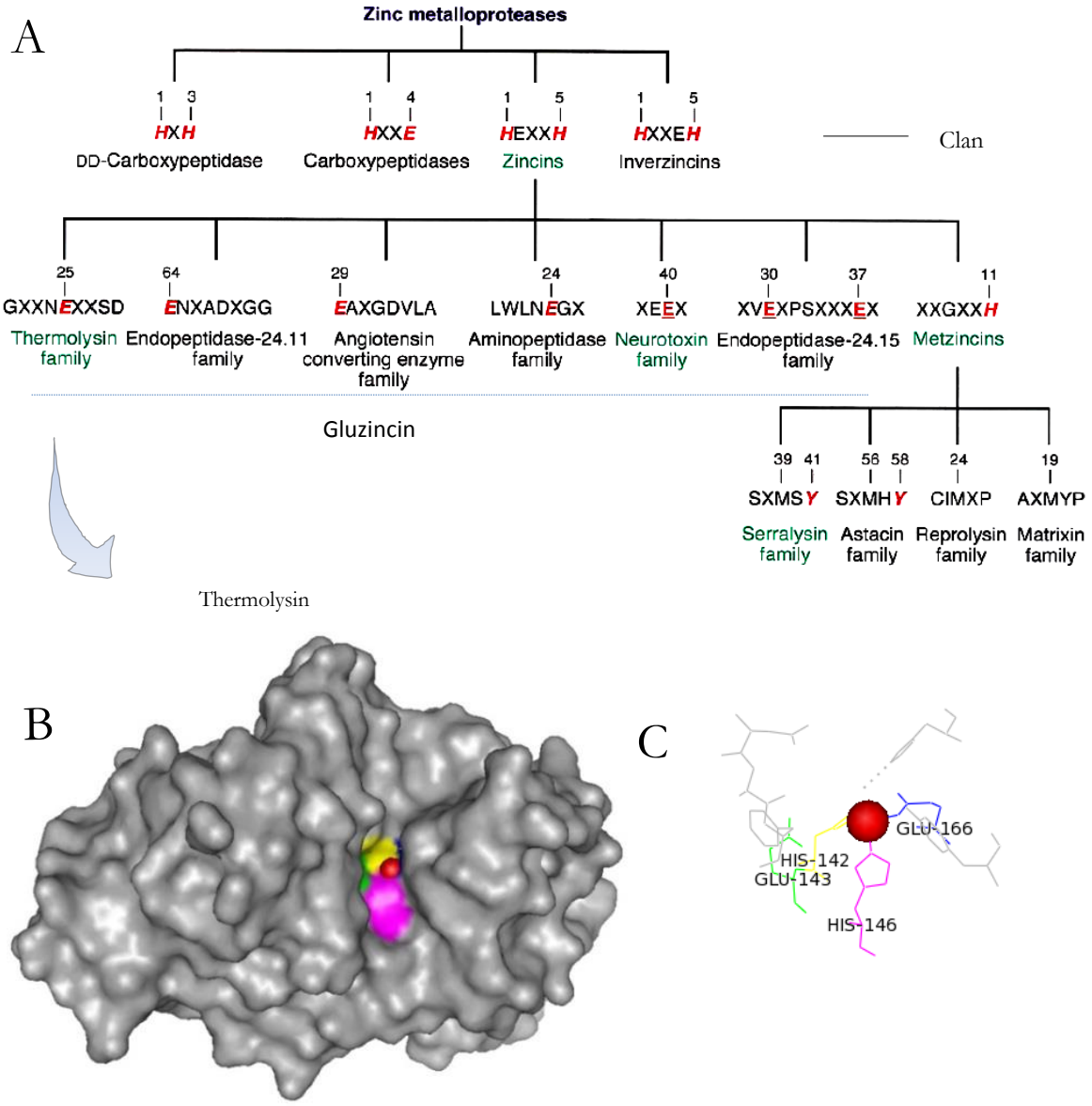


Figure I.22 – Overview of zinc metalloproteases. A: Classification of zinc metalloproteases based on sequence features. Residues in red are various metal binding ligands considered during the classification of zinc metalloproteases into various clans and families (Miyoshi and Shinoda, 2000) **B:** Structure of the *B. thermoproteolyticus* thermolysin protease (PDB 2A7G.pdb) showing components of its zincin motif (HELTH). The green shading indicates the position of the catalytic glutamic acid (Glu143), the yellow and magenta shadings are the histidine residues coordinating the zinc ion (Zn^{2+}) at the active site, while the blue shading represents the third metal binding ligand. All Zn^{2+} ion coordinating ligands are also shown in Figure I.22C in stick representation with the red sphere representing the ion.

According to the PFAM database (Punta *et al.,* 2012), M60-like domains have currently been identified in over 817 sequences from both prokaryotic and eukaryotic origins. In both groups of organisms, a seed alignment (alignment that contains a small set of representative members of the family) shows a highly conserved HEXXH motif followed by a series of other conserved residues including glutamic acid (E) residue (downstream of the HEXXH motif) (Figure I.23). On account of these features, M60-like proteases are thus classified as putative zincin proteases under the MAE subclan of metalloproteases also termed gluzincins on the MEROPS database (Rawlings *et al*., 2012, Figure I.22).

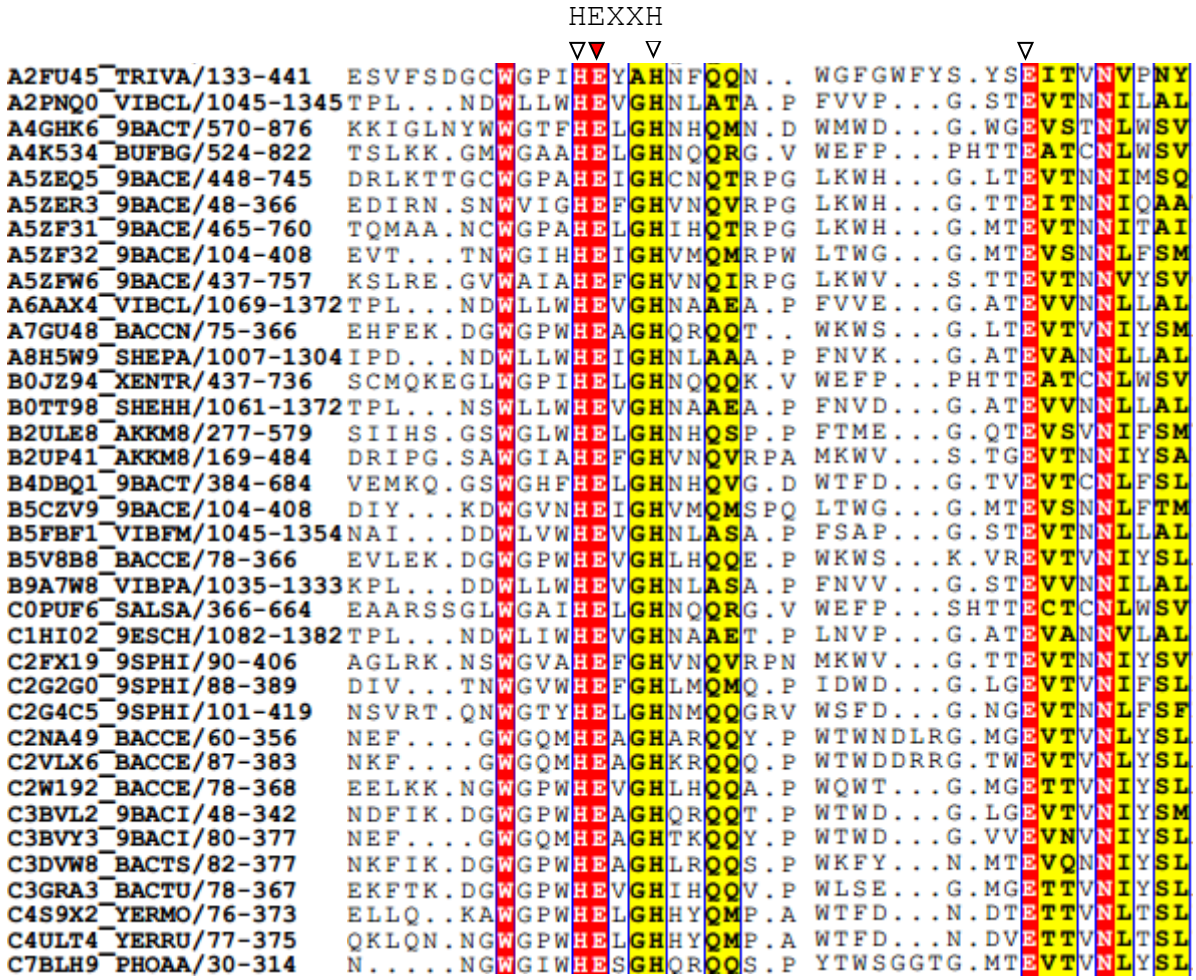


Figure I.23 - Seed alignment of M60-like domain-containing sequences showing a highly conserved zincin motif (HEXXH). The red inverted triangle points to the putative catalytic glutamic acid residue (E) while open inverted triangles point to various conserved metal binding residues. Aligned sequences (from both prokaryotic and eukaryotic origins) were all retrieved from the PFAM database (Punta *et al.,* 2012) (Uniprot IDs indicated on the left). Sequence alignments were viewed using the ESPript 2.2 utility at http://espript.ibcp.fr/ESPript/ESPript/ (Gouet *et al*., 1999) with a global similarity score threshold of 0.7. Red highlights are for amino acid residues showing 100% conservation while yellow highlights are for residues showing less than 100% conservation but above the global score threshold.

I.8.1.1.2.1 General proteolytic mechanism of zincins

A prototypic member of the zincin clan is the *B. thermoproteolyticus* thermolysin protease. As shown in Figure I.22, the protein contains the HEXXH motif (HELTH) and a third glutamic acid (E) metal binding ligand. In addition to water, Zn^{2+} ion is coordinated by 4 ligands (His 142, His 146, Glu166 and water) in the active site of the native enzyme. In the presence of a substrate at the active site, the carboxyl group of the catalytic glutamic acid residue (Glu143) forms hydrogen bonds with the water molecule coordinated to the electrophilic Zn^{2+} ion (Figure I.24). This causes a nucleophlic attack by the water molecule on the carbonyl atom close to the scissile peptide bond in the substrate (Matthews *et al.*, 1988, Coleman, 1998, Gomis-Rüth, 2003). The result is a tetrahedral intermediate stabilized by other residues in the enzyme structure. The distortion of the peptide bond in the structure occurs when Glu143 transfers protons acquired during the formation of the intermediate to the nitrogen of the scissile bond as shown below.

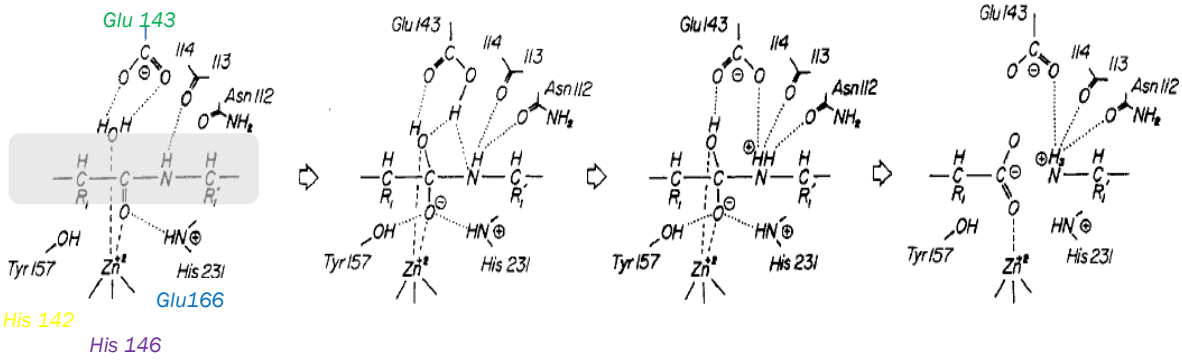


Figure I.24 - Sequence of events leading to the cleavage of a scissile peptide bond within a protein substrate by zincin proteases. The gray highlighted sequence represents a hypothetical protein substrate targeted for cleavage by a zincin protease (in this case shown for the *B. thermoproteolyticus* thermolysin protease*)*. See Figure I.22 for the location of various residues in the HEXXH consensus and the third metal binding ligand (Modified from Matthews, 1988 and Fernandez *et al.*, 2001).

I.8.2 Carbohydrate binding modules of M60-like domain-containing proteins

Carbohydrate active enzymes (CAZymes) are defined as enzymes capable of degrading, modifying, or creating glycosidic bonds. The online CAZy database (Cantarel *et al.,* 2009) includes glycoside hydrolases (GH), glycosyltransferases (GT) polysaccharide lyases (PL) and carbohydrate esterases (CE) in this group. CAZymes often contain one or more non-catalytic domains in their structure termed carbohydrate binding modules or CBMs. As the name implies, they generally serve to bind specific carbohydrate components within a target substrate. By doing so, they may potentiate the activity of the parent enzyme through proximity or targeting effects (Boraston *et al.*,2004, Shoseyov *et al.*, 2006, Herve *et al.*, 2010), or in some cases cause the disruption of the target substrate (Din *et al.*, 1991).

Initially identified as cellulose binding domains (Gilkes 1998, Boraston *et al* 2004), the number of CBMs with varying ligand specificities has risen rapidly over the past years. Indeed CBMs have now been classified into over 67 different families (with more yet to be classified) on the CAZy database (Cantarel *et al.,* 2009), targeting a wide range of substrates from monosaccharides to complex glycans. The classification into families is based on sequence similarities and various CBM families are identified by a number written in front of the CBM abbreviation e.g CBM32 to denote a family 32 carbohydrate binding module as seen on the CAZy database (Cantarel *et al.,* 2009). It is however worth noting that there are other putative and biochemically characterized carbohydrate binding domains that are not named following this convention e.g the BACON and PA14 carbohydrate binding domains respectively that are only available from the PFAM database (Punta *et al.,* 2012). BACON domains are a new family of protein domains and stand for Bacteriodetes-Associated Carbohydrate-binding Often N-terminal (Mello *et al.,* 2010). Their carbohydrate binding properties are yet to be characterized. PA14 domains are named based on their similarity to a ~14kDa domain in the anthrax protective antigen (Rigden *et al.,* 2004). Some PA14 domains detected in surface exposed epithelial cell adhesins (Epa proteins) of *Candida albicans* are capable of binding galacto-configured sugars (Maestre-Reyna *et al.*, 2012, Zupancic *et al.*, 2008).

CBMs are also classified into different types based on structural and functional similarities (Boraston *et al.*, 2004). A detailed insight into this classification scheme is provided in Boraston *et al.*, 2004 and Guillen *et al.*, 2010. In brief, CBMs are classified into Types A, B and C. Type A or 'surface-binding' CBMs refers to CBMs with a flat or platform like binding site. It is the

orientation of aromatic side chains at the binding site of the CBM that provides the planar platform typical of this group (Figure I.25). CBM1, 2 and 3 are examples of CBM families belonging to this group. The aromatic amino acid side chains of Type B CBMs on the other hand are oriented in a manner that results in the formation of a sandwich or twisted binding platform (Figure I.25).. This in addition to the extended nature of Type B binding sites (>15Å) is consistent with their ability to accommodate long glycan chains (Boraston *et al.*, 2004). In contrast, Type C CBMs or "small sugar CBMs" which lack extended binding site grooves or clefts. They contain smaller binding pockets and as a result are adapted for binding smaller sugars such as mono-, di- and trisaccharides. Examples are members of CBM families 9, 8 and 32 (Boraston *et al.*, 2004). The authors also discussed the classification of CBMs based on fold families, although it is deemed to be less useful in terms of predicting CBM function compared to type classifications (Boraston *et al.*, 2004).

Some enzymes are known to contain more than one CBM domains (multivalent) (Ficko-Blean and Boraston., 2009, Guillen *et al.*, 2010, Ficko-blean *et al.*, 2012) with different or similar substrate specificities and this may enhance their affinity for the target substrate through an avidity effect. Some CBMs also possess more than one binding site for similar reasons (Boraston *et al.*, 2004).

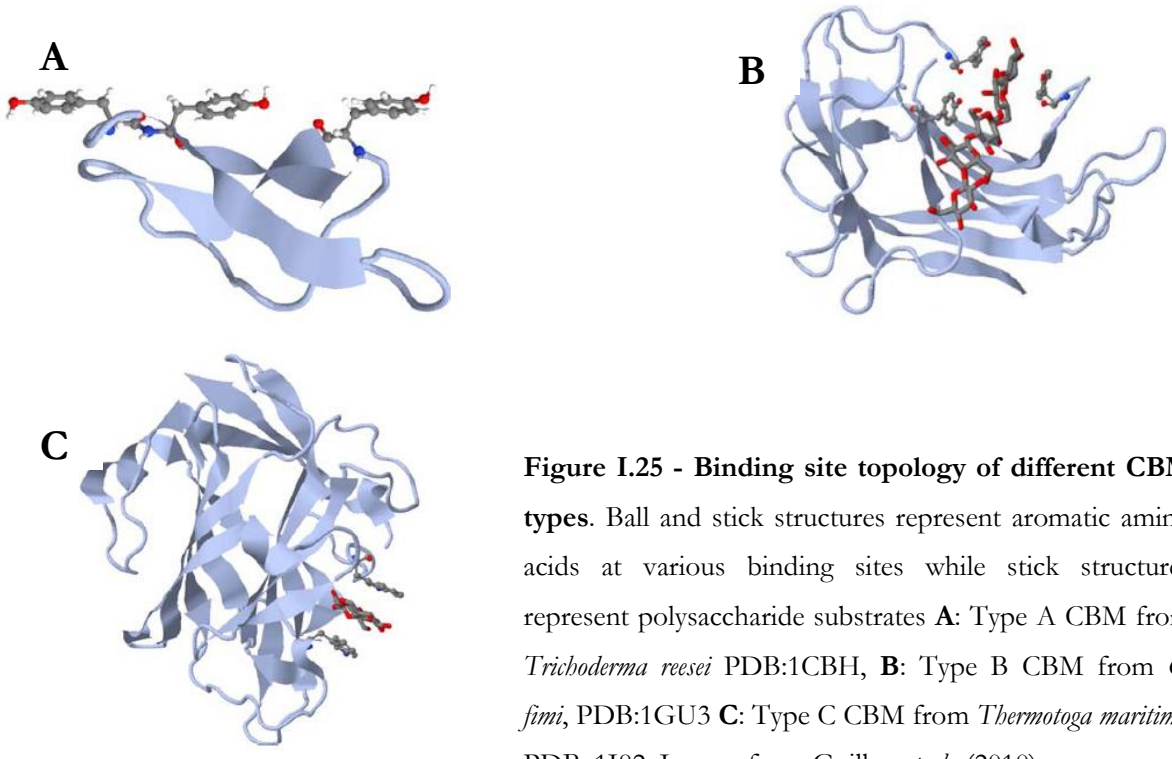


Figure I.25 - Binding site topology of different CBM types. Ball and stick structures represent aromatic amino acids at various binding sites while stick structures represent polysaccharide substrates **A**: Type A CBM from *Trichoderma reesei* PDB:1CBH, **B**: Type B CBM from *C. fimi*, PDB:1GU3 **C**: Type C CBM from *Thermotoga maritime*, PDB: 1I82. Images from Guillen *et al.,* (2010).

Putative CBMs were detected in many M60-like domain-containing proteins (Nakjang *et al.*, 2012). Examples include CBM32, CBM5, CBM12, CBM52, BACON and PA14 domains. Of these, the highest frequency of occurrence was observed for the family 32 (CBM32) carbohydrate binding modules as shown in Figure I.26.

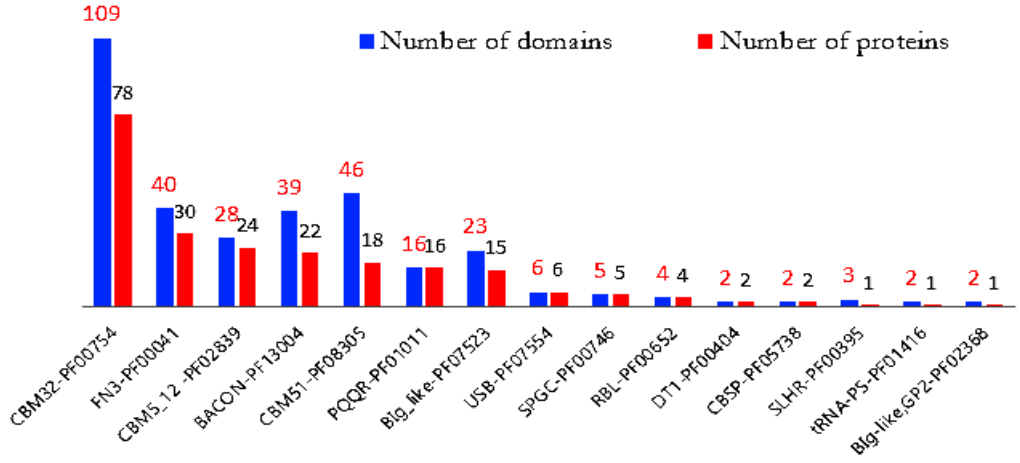


Figure I.26- Frequency of various domains associated with M60-like domain-containing proteins. This graph was generated using data from Sirintra Nakjang's thesis, 2012. CBM32 modules are a prominent feature of M60-like domain-containing proteins. More details of various domains can be obtained from the PFAM database (Punta *et al.,* 2012) using the PFAM ID's provided on the graph.

I.8.3 Evidence for extracellular localization of M60-like domain-containing proteins

A significant number (~70%) of M60-like/PF13402-containing proteins also possess additional features consistent with extracellular or cell surface localisation. Most either contain a putative signal peptide (SP) sequence or one or more transmembrane domains (TMDs) (PFAM, Nakjang *et al.*, 2012). Genes encoding M60-like domain-containing proteins also exist as members of different polysaccharide utilisation loci (PULs) in the genome of *B. thetaiotaomicron* (Table I.5). As indicated in Section I.6, gene products from these clusters are often cell envelope associated, some of them extracellularly (Sections I.6.1.1 and I.6.1.2). Their association with PULs also suggests that M60-like domain-containing proteins may functionally interact with other PUL components including amongst others, prominent members such as glycoside hydrolases (Martens *et al.*, 2008). Examples of putative glycoside hydrolase families encoded by M60-like PULs include families GH109, GH2, GH43, and GH35 (Table I.5). See subsequent section (Section I.9) for an overview of glycoside hydrolases.

PUL 5	GENES	Annotation
	BT_0262	hypothetical protein
	BT_0263	hypothetical protein
	BT_0264	Glycoside Hydrolase Family 43
	BT_0265	Glycoside Hydrolase Family 43
	BT_0266	hypothetical protein
	BT_0267	Hybrid two-component system regulator
	BT_0268	susC-like
	BT_0269	susD-like
	BT_0270	hypothetical protein
	BT_0271	hypothetical protein
	BT_0272	susC-like
	BT_0273	susD-like
	BT_0274	hypothetical protein
	BT_0275	hypothetical protein
	BT_0276	hypothetical protein
	BT_0277	hypothetical protein
	BT_0278	hypothetical protein
	BT_0279	hypothetical protein
	BT_0280	transposase for insertion sequence element ISRM3
	BT_0284	putative peptidoglycan binding protein (LPXTG motif)
	BT_0285	putative tolQ-type transport protein
	BT_0286	hypothetical protein
	BT_0287	putative biopolymer transmembrane protein
	BT_0288	hypothetical protein
	BT_0290	Glycoside Hydrolase Family 35
	BT_0291	Integrase

PUL 45	GENES	Annotation
	BT_3010	ECF-type sigma factor
	BT_3011	anti-sigma factor
	BT_3012	susC-like
	BT_3013	susD-like
	BT_3014	putative chitobiase
	BT_3015	hypothetical protein
	BT_3016	TonB-dependent receptor
	BT_3017	acid phosphatase

PUL78	GENES	Annotation
	BT_4240	Conserved hypothetical protein, with a phosphotransferase enzyme family domain
	BT_4241	Glycoside hydrolase family 2
	BT_4242	Putative transporter
	BT_4243	Putative oxidoreductase (putative secreted protein)
	BT_4244	Hypothetical protein
	BT_4245	Hypothetical protein
	BT_4246	Susd-like
	BT_4247	Susc-like
	BT_4248	Anti-sigma factor'
	BT_4249	Anti-sigma factor
	BT_4250	ECF-type sigma factor

PUL 79	GENES	Annotation
	BT_4266	Hypothetical protein
	BT_4267	Susc-like
	BT_4268	Susd-like
	BT_4269	Hypothetical protein
	BT_4270	Hypothetical protein
	BT_4271	Hypothetical protein
	BT_4272	Hypothetical protein

Table I.5 - PULs containing M60-like domain-containing proteins in *B. thetaiotaomicron*. The M60-like protein of each PUL is highlighted in red (See full list of PULs in Martens *et al.*, 2008).

I.9 Glycoside hydrolases

Glycoside hydrolases (EC 3.2.1.-) occur in essentially all domains of life and catalyse the hydrolysis of glycosidic bonds within polysaccharide or between polysaccharide and non-polysaccharide structures. There are several levels of their classification, for example they can be classified as endo or exo-acting, inverting or retaining enzymes or into various GH families.

I.9.1 Endo or exo-acting

As is the case with proteases, the classification of GHs as endo or exo-acting enzymes is dependent on whether they cleave internal or terminally located glycosidic linkages in their target substrates. This is also discussed in Sections I.6.1.1 and I.6.1.2 in relation to GHs of the Sus system.

I.9.2 Inverting and retaining enzymes

GHs can be classified as inverting or retaining depending on changes to the stereochemistry at the anomeric carbon following hydrolysis (Rye and Withers, 2000, Vuong and Wilson, 2010). Both classes of glycoside hydrolases require amino acids with carboxylic acid groups at their active sites to achieve hydrolysis. In inverting enzymes, one of the carboxyl amino acids serves as catalytic acid residue and the other, a catalytic base residue while retaining glycosidases contain a nucleophile and a general acid/base residue (Figure I.27).

Inverting enzymes act through a single step mechanism involving the donation of a proton to the anomeric carbon by the catalytic acid residue followed by a nucleophilic attack of the anomeric carbon by water earlier activated by the catalytic base. This leads to a change in the stereochemistry at the anomeric carbon. Retaining enzymes catalyse the hydrolysis of the glycosidic bond by a double displacement mechanism. In the reaction that follows, the glycosyl oxygen atom (O-glycosidic bond oxygen) is protonated initially by the general acid/base residue while a nucleophilic attack of the anomeric carbon by the nucleophile leads to the formation of a substrate glycosyl-enzyme intermediate (Figure I.27). In the second step, the previously deprotonated general acid/base residue then acts as a base to activate a water molecule that nucleophilically attacks the anomeric carbon of the glycosyl-enzyme intermediate. This two-step procedure ensures the retention of the stereochemistry at the anomeric carbon (Vuong and Wilson, 2010)

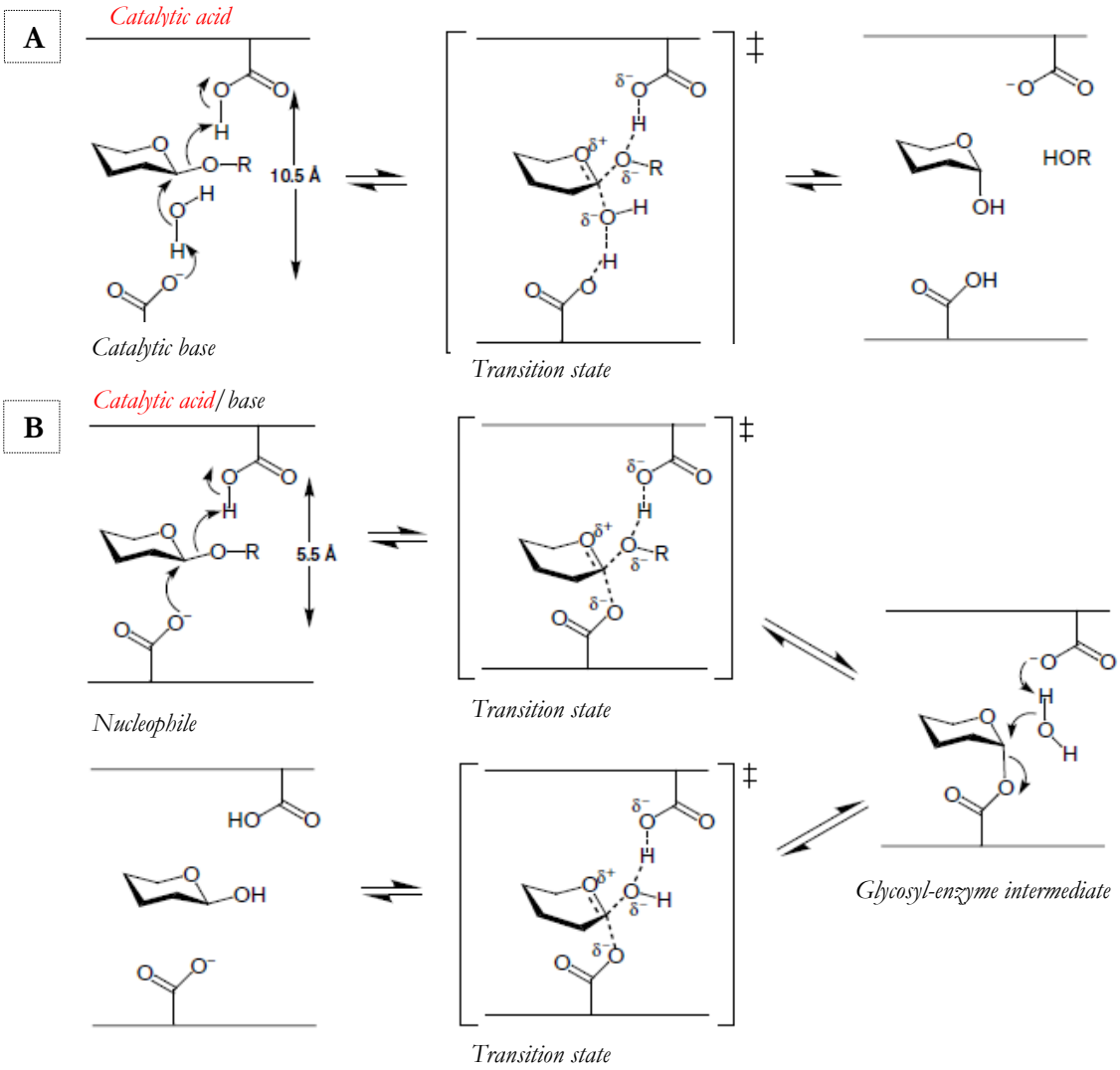


Figure I.27 - Mechanism of inverting and retaining glycoside hydrolases. A: Inverting enzyme mechanism **B:** Retaining enzyme mechanism (modified from Rye and Withers, 2000)

I.9.3 GH Clans and families

GH enzymes can also be classified into families based on sequence similarities [CAZY database (Punta *et al.*, 2012), Davies and Sinnot, 2008). Families are further grouped into clans based on similarities in the three dimensional structure, catalytic residues, mechanism of glycosidic bond cleavage and evolutionary ancestry (Naumoff, 2011). A list of various GH families and the mechanisms of hydrolysis is given below in Table I.6.

Clan	Families (GH)	Anomeric configuration	Three-dimensional structure
GH-A	1, 2, 5, 10, 17, 26, 30, 35, 39, 42, 50, 51, 53, 59, 72, 79, 86, 113	retained (eq.)	$(\beta/\alpha)_8$-barrel
GH-B	7, 16	retained (eq.)	β-sandwich
GH-C	11, 12	retained (eq.)	β-sandwich
GH-D	27, 31, 36	retained (ax.)	$(\beta/\alpha)_8$-barrel
GH-E	33, 34, 83, 93	retained (eq.)	6-bladed β-propeller
GH-F	43, 62	inverted (eq.)	5-bladed β-propeller
GH-G	37, 63	inverted (ax.)	$(\alpha/\alpha)_6$-barrel
GH-H	13, 70, 77	retained (ax.)	$(\beta/\alpha)_8$-barrel
GH-I	24, 46, 80	inverted (eq.)	α + β-lysozyme
GH-J	32, 68	retained (β-furanoside)	5-bladed β-propeller
GH-K	18, 20, 85	retained (eq.)	$(\beta/\alpha)_8$-barrel
GH-L	15, 65, 125	inverted (ax.)	$(\alpha/\alpha)_6$-barrel
GH-M	8, 48	inverted (eq.)	$(\alpha/\alpha)_6$-barrel
GH-N	28, 49	inverted (ax.)	$(\beta)_3$-solenoid

Table I.6 - GH families and clans. Members of the same clan for the most part employ similar catalytic strategies [i.e. retaining or inverting mechanisms (ax.: axial and eq: equatorial)] and possess similar three dimensional structures (Naumoff 2011).

Other GH families that have not been allocated to clans in the CAZy database include, GH129, GH109 and GH4 families which also happen to use an unusual mechanism involving NAD+ as cofactor [CAZy (Punta *et al.*, 2012), Naumoff 2011, Liu *et al.*, 2007]. In the PFAM database, GH4 (PF02056) and GH109 belong to clan CL0063 (PFAM, Naumoff 2011).

I.10 Objectives of this study

The overall aim of this study was to gain new insights into host microbial interactions at mucosal surfaces in disease and health by studying the novel family of M60-like domain-containing proteins and their functional partners in *B. thetaiotaomicron* and *T. vaginalis.*

The specific objectives were as follows;

Characterize M60-like domain-containing proteins from *B. thetaiotaomicron* and *T. vaginalis* using *in-silico* and biochemical approaches.

Analyze the functional context of a PUL associated M60-like domain containing protein in *B. thetaiotaomicron*.

Evaluate the contribution of a *B. thetaiotaomicron* M60-like PUL to the organisms' fitness and survival on mucins *in-vitro*.

CHAPTER II

Materials and methods

II.1 Molecular biology and Biochemistry

II.1.1 Bacterial strains

Various *Escherichia coli (E. coli)* strains used during the course of this study are listed below in Table II.1.

Strain	Genotype	Use	Reference
BL21(DE3)	F- *ompT hsdS*$_B$(r$_B$- m$_B$-) *gal dcm* (DE3)	Protein Expression	Studier and Moffatt, 1986
One Shot™ TOP10	F- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) ±80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *end*A1 *ara*D139 Δ(*ara, leu*)7697 *gal*U *gal*K Δλ- *rps*L *nup*G *ton*A *hsd*R	DNA cloning (plasmid propagation)	Invitrogen
CC118 λ *-pir*	*⊿(ara-leu) araD ⊿lacX74 galE galK phoA20 thi-1 rpsE rpoB argE (Am) recA1 λpir*	Gene deletion (plasmid propagation)	Herrero *et al.*, 1990
S17.1 λ*-pir*	*hsd*R *recA pro* RP4-2 (Tc::Mu; Km::Tn*7*)(λ *pir*)	Gene deletion	Skorupski and Taylor, 1996

Table II. 1 - Bacterial strains used in this study.

II.1.2 Plasmids

Cloning and DNA deletion plasmids used during the course of this study are listed below in Table II.2

Plasmids	~Size (kbp)	Phenotype/ Genotype	Reference
pET28a	5.4	Kan^r, T7, *lac, laclq*	Novagen
pET43.1a	7.3	Amp^r, T7, *lac*, *laclq*	Novagen
minipRSET-A	2.9	Amp^r, N-His	Invitrogen
*p*Exchange-tdk	4.2	Amp^r	Koropatkin *et al.*, 2008

Table II. 2 - Plasmids used in this study. Please see maps for various plasmids in appendix C.

II.1.3 Growth Media

The preparation of various growth media used during the course of this study are described below in Table II.3.

Medium	Composition	Amount per litre
Luria-Bertani (LB) medium	Bacto®tryptone	10 g
	Bacto®yeast extract	5 g
	NaCl	10 g
	After dissolving components in about 900 ml of water, the pH of the solution was adjusted to 7.4 with NaOH and the final volume taken to a litre prior to sterilisation (Section II.1.5)	
LB – agar	2 g of agar (AGAR NO.1, OXOID) was added to 100 ml of LB medium (2%) and autoclaved at 121 ^{o}C for 20 min. See below for preparation of selective media	
TYG medium	Tryptone Peptone	10 g
	Bacto Yeast Extract	5 g
	Glucose	2 g
	Cysteine (free base)	0.5 g
	1 M KPO_4 pH 7.2	100 ml
	Vitamin K solution, 1 mg/ml	1 ml
	TYG salts	40 ml
	0.8% $CaCl_2$	1 ml
	$FeSO_4$, 0.4 mg/ml	1 ml
	Resazurin, 0.25 mg/ml	4 ml
	The final volume of the mixture was taken up to 1L, mixed properly and 5ml transferred into glass test tubes. The test tubes containing media were then clogged with cotton and later autoclaved at 121 ^{o}C for 20 min. After autocalving media were allowed to cool to room temperature and 5 µl of His-Hem solution (Section II.1.3.1.1) added prior to use.	
Minimal medium	NH_4 SO_4	1 g
	Na_2CO_3	1 g
	cysteine, free base	0.5 g
	1 M KPO_4 pH 7.2	100 ml
	Vitamin K solution, 1 mg/ml	1 ml
	$FeSO_4$, 0.4 mg/ml	10 ml
	Resazurin, 0.25 mg/ml	4 ml

Medium	Composition	Amount per litre
Minimal medium (continued)	Vitamin B_{12}, 0.01 mg/ml	0.5 ml
	Mineral Salts for defined medium	50 ml
	After preparation of the minimal medium, subsequent steps were as above for the TYG medium except for the fact that before inoculation with bacteria, the solution of a desired substrate (in about 100-200 µl) is added to the medium. For some substrates such as PGM that are not affected by autoclaving, the appropriate amount was weight and autoclaved together with medium	
Modified Diamond's Medium (MDM)	Trypticase peptone	20 g
	Yeast extract	10 g
	Maltose	5 g
	Ascorbic acid	1 g
	Iron II sulphate heptahydrate	0.1 g
	KCl	1 g
	KH_2PO_4	1 g
	K_2HPO_4	0.5 g
	$KHCO_3$	1 g
	Media contents were mixed and 200 ml volumes transferred to into 5 blue-topped 200 ml bottles and taken for autoclaving. After autoclaving 20 ml of horse serum (from -20 ^oC) was added per 200 ml of medium followed by 2 ml of penicillin – streptomycin to obtain a complete MDM medium. The MDM medium was then shared into ~50 ml fractions in 50 ml conical tubes and tightly corked. They were also further sealed with parafilm. Tubes containing MDM were stored at 37^oC for about a week.	
Brain heart infusion (BHI)	3.75 g in of BHI and 2 g of agar were dissolved in 100 ml of distilled water and autoclaved. The sample after autoclaving was allowed to cool and when desired, other ingredients such as antibiotics were added before pouring into petri dishes	

Table II.3 - Preparation of growth media

II.1.3.1 Other media components

II.1.3.1.1 His-Hem solution (0.2 M Histidine pH 8.0)

4.2 g of Histidine - HCl monohydrate (Sigma cat.; H7875) was dissolved in 80 ml distilled water and pH adjusted to 8 with 10N NaOH. The final volume was brought up to 100 ml with distilled water. 12 mg of Hematin (Sigma cat.; H3281) was mixed with 10 ml of 0.2 M Histidine pH 8.0 and dissolved by end-over-end rotation or vigorous shaking for several hours. The sample was later filter-sterilized using a 0.2 μm filter and stored at 4 °C.

II.1.4 Selective media

The preparation of stock antibiotic solutions was as described below in Table II.4. After autoclaving, the LB-agar mixture (Table II.3) was allowed to cool to about 50 °C and the necessary antibiotics added. The mixture was later poured into 90 mm sterile petri dishes (~25 ml per dish) and allowed on a working bench to solidify before storage at 4 °C.

Antibiotic	Stock concentration	Final antibiotic concentration	Storage
Ampicillin	50 mg/ml in water	50 μg/ml	4 °C for < 5 days
Kanamycin	10 mg/ml in water	10 μg/ml	4 °C for < 5 days

Table II.4 - Preparation of antibiotics

II.1.5 Sterilisation

Unless otherwise stated, media, glassware and solutions were sterilized by autoclaving using a portable steam sterilizer or autoclave (Prestige Medical) at 121 °C, 32 lb / inch $^{-2}$ for 20 min. For solutions that could not be sterilized by this means, filter sterilization was carried out using a sterile syringe (Plastipak®,Becton Dickinson) and an appropriate pore-sized (0.22-1 μm) Millipore filter discs (Supor® Acrodisc®).

II.1.6 Storage of DNA and bacteria

For long term storage, glycerol stocks (25 % glycerol) of bacterial strains were kept at -80 °C in cryovials. *E. coli* colonies on agar plates were only stored at 4 °C for a maximum of 4 days before re-use. Plasmids were stored at -20 °C in EB buffer (10 mM Tris/HCl buffer, pH 8.5).

II.1.7 Plating bacteria

To plate bacterial cells on LB-agar, a glass spreader was immersed in 100 % ethanol and passed through a burnsen flame to enable sterilization as the ethanol burned off the spreader. The spreader was allowed to cool for about a minute and used to evenly spread 100 µl of bacterial suspension over the surface of the agar. Plates were incubated in an inverted position at 37 °C for overnight growth in an incubator (LEEC Ltd).

II.1.8 Growth of *B. thetaiotaomicron*

TYG was used as the culture medium for *B. thetaiotaomicron* in this study. 5 ml - 10 ml of TYG and MM media were inoculated with about 50 µl - 100 µl of *B. thetaiotaomicron* from a glycerol stock or directly from culture medium. Anaerobic conditions were achieved using pyrogallol and sodium bicarbonate. In brief, tubes containing media were clogged with the cotton wool used during autoclaving and the cotton burned under a hood to partially extract oxygen from within the tubes. Burned cottons were then pushed half way down each tube and soaked with 200 µl of 35% pyrogallol followed by same volume of 10% $NaHCO_3$. Tubes were immediately corked with plastic stoppers and taken to a 37 °C incubator for growth.

II.1.9 Growth of *T. vaginalis*

MDM medium (Table II.3) was used as culture medium for *T. vaginalis* in this study. Frozen *T. vaginalis* cells from liquid nitrogen were quickly defrosted in warm tap water (~40 °C) and 2 ml inoculated into tubes containing about 50 ml MDM in a sterile laminar flow hood. Tube caps were then tightly applied and further sealed with parafilm before incubating at 37 °C for about 2 days.

II.1.10 Centrifugation

Bacterial cells were often harvested from culture (100-1000 ml) in 500 ml centrifuge pots (Nalgene) using low speed centrifugation at 5000 × g for 10 min at 4 °C. The 500 ml centrifuge pots are adapted for use with the JA-10 rotor of a Beckman J2-21 centrifuge (Beckman Coulter, Inc.). Bacterial cells (usually the *E.coli* TOP 10 cells) from below 10 ml culture volumes were

harvested in sterile 25 ml Sterilin tubes by centrifugation at 5000 × g for 5 min using a fixed angle Hettich Zentrifugen bench centrifuge (Hettich Lab technology). For culture volumes of 1-2 ml, appropriate Eppendorf tubes were used with a HERAEUS, PICO17 bench top centrifuge (Thermo Scientific).

II.1.11 Chemically competent *E. coli*

Chemically competent cells used for cloning and expression experiments were prepared by a modification of the protocol described by Cohen *et al.* (Cohen *et al*, 1972). A single colony of *E.coli* cells was used to inoculate 5 ml of LB for overnight growth at 37 °C while shaking at 180 rpm. 1 ml of the overnight culture was used to inoculate 100 ml of LB in a non-baffled 1 L flask and cells allowed to grow to an optical density (OD600nm) of about 0.4. The flask containing cells was later placed on ice for 20 min after which cells were collected into 4 x 25 ml sterilin tubes and centrifuged at 280 x *g* for 5 min at room temperature. The supernatant was discarded in each case and cells resuspended in 3 ml of ice cold 0.1 M $CaCl_2$. The previous step was repeated and the cells resuspended 1 ml sterile ice-cold 0.1 M $MgCl_2$. Cells were finally made competent by allowing on ice for 2 h. For long term use, competent cells were stored as 100 µl aliquots in 1.5 ml Eppendorf tubes with 25 % (v/v) glycerol at -80°C.

II.1.12 Genomic DNA extraction

II.1.12.1 *B. thetaiotaomicron*

B. thetaiotaomicron genomic extraction was carried out using the Sigma GenElute™ Bacterial Genomic DNA Kit (Sigma cat.; NA2100) according to the manufacturer's instructions.

II.1.12.2 *T. vaginalis*

II.1.12.2.1 Cell lyses and preparation for Phenol chloroform extraction

T. vaginalis genomic DNA was purified by phenol chloroform extraction (Bewsey *et al.*, 1991). *T. vaginalis* cells (75-100ml culture volume) were harvested at mid-log phase by centrifugation at 900 g for 5 min at 4 °C and washed twice in 20 ml PBS at 4 °C. The cells were centrifuged again, re-suspended in 1ml PBS and later placed on ice for about 5 min. 20 µl of DEPC (Diethylpyrocarbonate) was then added to the cell suspension in a fume hood and allowed to sediment on ice for 5 min. 40 ml of freshly prepared DEPC- Triton X-100-RBS [40µl of DEPC plus 100 ml of cold Triton X-100-RBS (10 mM NaCl, 10mM Tris-HCl, 5mM $MgCl_2$)] was then added to mixture and centrifuged for 3 min at 900 g at 4 °C. The previous step was repeated and the final cell pellet resuspended in 5 ml proteinase K buffer (1.5% SDS, 1% Proteinase K). The

mixture was incubated at 55 °C for 2 h with shaking every 15 min to increase the rate at which the pellet dissolved. Proteinase K was later-on heat inactivated after by incubating the mixture at 65 °C for 10 min.

II.1.12.2.2 Phenol/chloroform extraction

At this stage of the process only polypropylene tubes which are resistant to phenol/chloroform were used. Equal volumes of Sigma phenol: chloroform: isoamyl alcohol solution (Sigma cat.; P2069) and the nucleic acid solution containing inactivated proteinase K above were mixed and agitated briefly until an emulsion formed. The solution was then centrifuged at 12,000 x g for 3-5 min at room temperature. The aqueous phase of the centrifuged mixture was pipetted into a new tube while the organic phase and interface between the organic and aqueous phases were both discarded. The previous step was repeated on the aqueous phase until no protein was visible at the interface between the aqueous and organic phases. An equal volume of chloroform was then mixed with the aqueous phase and centrifuge at 12,000 x g for 3-5 min at room temperature to remove residual phenol. The aqueous phase was then pipetted into a new tube for ethanol precipitation.

II.1.12.2.3 Ethanol precipitation

3 M Sodium Acetate (pH 5.2) was added to the aqueous phase from above at a rate of 1/10th the volume of the aqueous phase followed by two volumes of 100% ice cold ethanol. The tube contents were mixed by inverting several times and stored at -20 °C for at least 1 h. After this step the solution containing the DNA was centrifuged at 13 200 rpm for 20 min using a HERAEUS, PICO17 bench top centrifuge (Thermo Scientific) and the pellet washed with ~ 500 µl of 70% ice cold ethanol by further centrifugation at 13 200 rpm for 5 min in a 1.5 ml Eppendorf tube. The pellet was then vacuum-dried after removal of the supernatant and reconstituted in about 200 µl of TE buffer (10mM Tris/1mM EDTA). The DNA was quantified using a Nanodrop spectrophotometer (Section II.1.13).

II.1.13 Determination of DNA and protein concentration

The concentration of DNA or protein in samples was routinely estimated by absorbance at 260 nm and 280 nm (A260nm, A280nm) respectively using a NanoDrop 2000 UV-Vis spectrophotometer (Thermo Fisher Scientific Inc, USA) and the beer lamberts equation:

$$A=\varepsilon CI$$

Where A = absorbance 280 or 260nm, ε = molar extinction coefficient, I = length of light path (cm), and C = molar concentration of sample.

II.1.14 Primers

All primers used in this study were synthesised by Sigma (Sigma Adrich, UK). Primer parameters were estimated using the online Oligonucleotide Properties Calculator tool at http://www.basic.northwestern.edu/biotools/OligoCalc.html. Primer lengths were usually greater than or equal to 18 bp and melting temperatures (T_m) greater than or equal to 50 °C. For cloning experiments, restriction sites were added to the 5'- ends of various primers alongside a CCGG or GCGC spacer to enable cleavage by restriction enzymes and susbsequent ligation in similarly cut vectors. Primers were often synthesized dry and resuspended in highly distilled water to the desired concentration upon receipt.

$$T_m = 64.9 + 41*(yG + zC - 16.4)/ (wA + xT + yG + zC)$$

Where w, x, y, z are the number of the bases A, T, G, C in the sequence, respectively.

II.1.15 Polymerase chain reaction (PCR)

DNA amplification and site directed mutagenesis was routinely performed by PCR (Mullis & Faloona, 1987) using the Novagen Hot start PCR kit (Novagen). A typical PCR reaction set-up is given in Table II.5. Amplification reactions were performed using a PHC-3 thermocylcer (Biorad) using the standard program below (Table II.6) unless otherwise indicated.

Components and concentrations	Volume
Autoclaved distilled water	19 µl
10 x KOD buffer minus Mg^{2+} (10 x)	5 µl
dNTP's (2 mM)	5 µl
Q-solution (DMSO)	5 µl
$MgSO_4$ (25 mM)	4 µl
Template DNA (~70 ng/µl)	1 µl
Novagen® KOD DNA Polymerase (2.5 U/µl)	1 µl
Forward oligonucleotide (5 µM)	5 µl
Reverse oligonucleotide primer (5 µM)	5 µl
Total volume	50 µl

Table II.5 - Typical PCR reaction set-up

Program name	Event	Temperature	Duration	Number of Cycles
Program 1	Denaturation	95 °C	1 min	1
Program 2	Denaturation	95 °C	1 min	
	Annealing	50 °C	1 min	30
	Extension	68 °C	atleast 1min/1kbp fragment size	
Program 3	Polishing	68 °C	10 min	1
Program 4	Storage	10 °C	≤ 24hr	1

Table II.6 - Typical PCR reaction program

II.1.16 Quantitative/Real time Polymerase chain reaction (qPCR)

Please see Section II.1.46.5.

II.1.17 Site-directed mutagenesis

Single amino acid mutations were introduced into recombinant proteins using a modified version of the protocol described for the QuickChangeTM Site-Directed Mutagenesis Kit (Stratagene). The main modification was the use of high fidelity Novagen ® KOD DNA Polymerase in place of the PfuTurbo DNA polymerase enzyme recommended for the kit. Mutations were targeted at DNA fragments already cloned into plasmids. Primers used in the PCR reaction were designed to contain the desired mutation. A typical site-directed mutagenesis PCR set-up and program resembled that described for standard PCR reactions in Section II.1.15 except that extension times were often higher due to the need to amplify the full recombinant plasmid. Also only about 18 cycles of Program 2 were performed. After completion of the amplification reactions, the enzyme DpnI was added to the reaction products at a rate of 30 units per 50 µl of the PCR reaction (for 1 h at 37 °C) to enable digestion of the methylated and unmutated template dsDNA. This step leaves behind only unmethylated but mutated PCR amplicons. 5 µl of the digestion reaction was then used for transformation of chemically competent One Shot™ TOP10 *E.coli* cells (Section II.1.10).

II.1.18 Analyses of PCR results by agarose gel electrophoresis

The results of PCR reactions were analysed by agarose gel electrophoresis (AGE) (Meyers *et al*, 1976). AGE allows the detection and separation of amplified DNA fragments according to size. Electrophoresis equipments used were the HU10 Mini Plus Horizontal Gel Unit (SCIE-PLAS Ltd) and a BDH horizontal gel mould connected to a Bio-rad Mini-PROTEAN® Tetra Cell power supply (Bio-rad). To run an AGE experiment, 0.5 g of low grade (MELFORD Ltd) or SeaKem® Gold Agarose (Lonza) was dissolved in 50 ml of TBE buffer (biroad) or TAE (Amersham) in a 200 ml conical flask to obtain a 1 % agarose solution. The solution was stirred and heated in a microwave oven at 450 watts of power for at least 1.5 min. This was later allowed to cool to about 60 ºC followed by addition of 5 µl (0.5-1 mg/ml) of an ethidium bromide solution. The solution was mixed by gentle swirling and poured into a gel casting mould set-up according to the manufacturers' instructions. As the gel solidified, 5 µl of DNA loading buffer [0.25% (w/v) bromophenol blue, 50% (v/v) Glycerol, 10 x TBE buffer (8.9 mM Tris base, 8.9 mM Boric acid, 2 mM EDTA pH 8.0)] were added to 5 µl of the solution containing the PCR products to be analysed. After setting, the gel was submerged in 50 ml of TBE buffer followed by application of the PCR samples [~10 µl/sample alongside standards (7 µl of Bioline HyperLadder™ I markers)]. Electrophoresis was run at a constant voltage (70 V for Biorad machine and 100 V for the Amersham machine) for about 1 h. Results were visualised in the UV range using Bio-Rad Gel Doc 1000 system (Bio-Rad). An example of data obtained following agarose gel electrophoresis of PCR amplified products is shown below in Figure II.1.

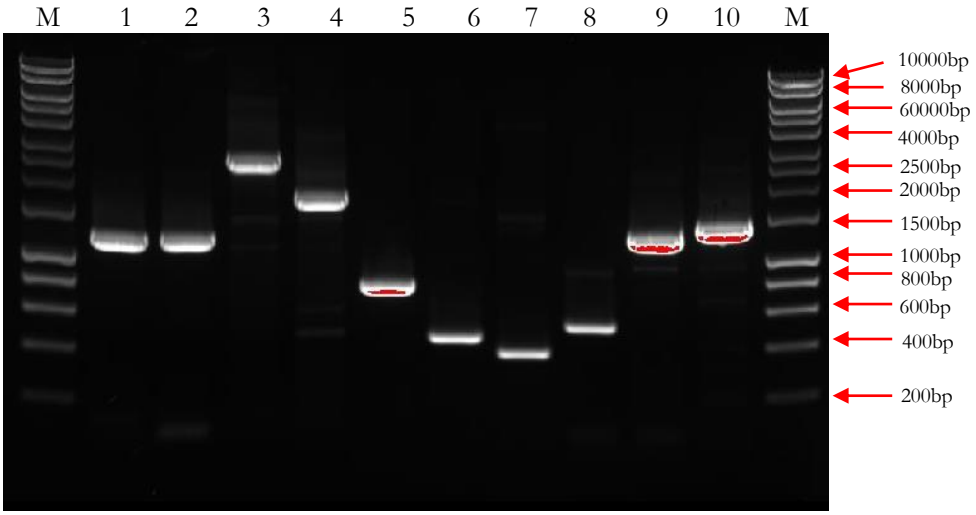


Figure II.1 - Example of data obtained following agarose gel electrophoresis of PCR products. PCR products and standards were analysed on a 1% agarose gel in TAE buffer. Bands in lanes 1-10 are amplified products from different DNA fragments from the genome of *B. thetaiotaomicron*. Lanes M: Bioline HyperLadder™ I standards (7 µl), Lanes 1-10: AGE results from 5 µl of different PCR reactions.

II.1.19 Purification of PCR products

PCR products were purified using the Qiagen QIAquick PCR Purification Kit (Qiagen) as described in manufacturer's instructions.

II.1.20 DNA digestion with restriction enzymes

Digestion of amplified DNA and plasmids containing cleavage sites for restriction enzymes was carried out prior to ligation reactions (Section II.1.22). Restriction enzymes and buffers used during the process were ordered from fermentas (MBI Fermentas, UK). Digestion reactions were set-up according to the manufacturers' instructions. A typical set – up is shown below in Table II.7.

Components	***Volume***
Distilled water	1-2 µl
DNA fragment / plasmid (~0.1 – 0.5 µg)	50 µl
Restriction enzyme buffer (10 x)	6 µl
Restriction enzyme (10 U)	2-3 µl
Total	~ 60 µl

Table II.7 - Set-up of a typical restriction enzyme digestion reaction. One unit of enzyme is defined as the amount of enzyme required to cleave 1 µg of DNA in 1 h at 37 °C. Digestion reactions were often incubated in a water bath at 37 °C for at least 1 h.

II.1.21 DNA extraction from agarose gels

After DNA digestion with restriction enzymes, agarose gel electrophoresis (with high quality seakem agarose (Seakem)] was performed on samples followed by gel extraction to purify the digested DNA. This was achieved using the Qiagen QIAquick Gel Extraction Kit (Qiagen) according to the manufacturers' instructions.

II.1.22 Ligation reactions

Following the purification of digested PCR products (insert) and plasmids from agarose gels, the concentration of DNA in each sample was determined and the data used in setting up ligation reactions. Ligation reactions were performed using the Novagen rapid ligation kit (Novagen) according to the manufacturers' instructions. Ligation reactions often contained a plasmid to insert concentration ratio of 1:3 (Table II.8).

Components	***Volume***
Vector (10 ng/µl)	2 µl
Insert DNA (10 ng/µl)	6 µl
5x Ligase buffer	4 µl
T_4 DNA Ligase (4 U/µl)	1 µl
H_2O (Nuclease free water)	7 µl
Total volume	**20 µl**

Table II.8 - Example of a ligation reaction set-up. Ligation reactions were allowed to run at 37 °C for at least 1 h.

II.1.23 Transformation and growth of competent *E. coli*

Competent cells from -80°C were allowed to thaw on ice for about 5-10 min followed by addition of 2-5 µl of plasmid or the ligation mix above. The mixture was allowed on ice for a further 1 h an hour before heat-shocking by incubation in a Techne Dri-Block™ DB-2A at 42 °C for 2 min. After heat shocking cells were immediately returned on ice for ~ 3 min. Transformed cells were then plated on agar containing an appropriate selection antibiotic (Sections II.1.3 and II.1.7). In the case of One Shot™ TOP10 cells, 250 µl of sterile LB medium was added to the heat-shocked cells and allowed to grow at 37 °C for 1 hour in a rotating incubator (180 rpm) before plating. In each case, plated cells were allowed to grow overnight at 37 °C.

II.1.24 Plasmid DNA purification

Following overnight growth, single colonies of transformed *E.coli* cells were separately subcultured into 5 ml or 100 ml of LB (with appropriate antibiotic) overnight at 37 °C while shaking (180 rpm). These were then used for plasmid extraction. For small scale plasmid DNA purications (5 ml-10 ml culture volumes) the QIAprep ® Spin Miniprep Kit (Qiagen)was used, while large scale purifications (>10ml) were carried out using the Plasmid Midi Kit (Qiagen). In both cases, the protocol was as described in the manufacturers' instructions

II.1.25 Automated DNA sequencing

Automated DNA sequencing was performed using the MWG Value Read service (MWG Biotech AG, Ebersberg, Munich, Germany) to check for mutations in cloned DNA sequences. 5 µl of plasmid DNAs were dried by vacuum lyophilization at room temperature in a 1.5 ml Eppendorf tube and labelled with pre-ordered sequencing labels before posting to MWG. Sequencing primers used were either custom - designed or standard sequencing primers available from the MWG website such as the T7 - forward (TAATACGACTCACTATAGGG) and reverse primers (CTAGTTATTGCTCAGCGGT) complementary to regions within most of the plasmids used in this study. Sequencing data received from MWG were analysed by alignment with the original KEGG database DNA sequence using a multiple sequence alignment tool such as Multalign (http://multalin.toulouse.inra.fr/multalin/).

II.1.26 Over-expression and purification of recombinant proteins in *E. coli*

II.1.26.1 Induction of protein expression and cell lysis

E.coli BL21 (DE3) host cells were used for recombinant protein expression. The cells were transformed with sequenced recombinant plasmids and later plated on appropriate selective media for overnight growth (Section II.1.23). The next day, a loop-ful of colonies harbouring plamsids were scraped and inoculated into 100 ml volumes of LB containing antibiotic in 200 ml baffled/unbaffled conical flasks. Cells were grown at 37 °C with aeration (180 rpm) until an OD600nm of ~0.6. Depending on the protein, at this stage, cells were either directly induced at 37 °C with 1 mM IPTG or cooled under running tap water to about 16 °C before induction with same concentration of IPTG. Cells induced at 37 °C were allowed to grow for just 4-5 hrs at the same temperature with aeration (180rpm) before protein purification while those at 16 °C were grown overnight before protein purification. For protein purification, cells were harvested by centrifugation at 5000 x g for 10 min, the supernatant discarded and the pellet resuspended in 5

ml of Talon buffer (20 mM Tris/HCl pH 8.0 plus 100 mM NaCl) per 100 ml of original culture volume. Cells were then lysed by sonication for 1-2 min (0.5 second cycling) on ice using a B. Braun Labsonic U sonicator (B. Braun, Melsungen, Germany) set at low intensity (~45 watts). Lysed cells were transferred into 50 ml centrifuge tubes (Nalgene) and centrifuged at 15 000 rpm [using a JA25.5 rotor in a Beckman J2-21 centrifuge (Beckman Coulter, Inc.)] for 20-30 min at 4 °C. The resulting supernatant or cell free extract (CFE or soluble fraction) was collected for protein purification while pellet fractions (insoluble fractions) were resuspended in 10 ml of Talon buffer and saved on ice for later analyses.

II.1.26.2 Immobilised metal affinity chromatography (IMAC)

The purification of recombinant N- or C- terminal 6 x His-tagged proteins from the CFE was achieved by immobilised metal affinity chromatography (IMAC). This was carried out using TALON Metal Affinity Resins (Clontech Laboratories Inc) containing bound Cobalt (Co^{2+}) capable of binding to the histidine tag of the recombinant proteins. Briefly, HisTALON Gravity Columns (Clontech Laboratories Inc) were filled with a 2.5 ml bed volume of Talon resin and equilibrated with at least 10 ml of Talon buffer (20 mM Tris/HCl pH 8.0 plus 100 mM NaCl). The CFE solution was then applied onto the resin bed in the column and allowed to drain by gravity. The flow through (FT) was collected and saved for later analyses. The resin was washed with at least 20 ml of Talon buffer followed by a stringent wash with 5 ml of Talon buffer containing 10 mM imidazole. Elution of the bound protein from the resin was achieved by sequential application of 5 ml volumes of Talon buffer containing 100 mM imidazole. All eluted fractions were collected and saved for subsequent analyses by SDS PAGE (Section II.1.27) or for further purification by Ion-exchange or gel filtration chromatography.

II.1.26.3 Ion-exchange and gel filtration chromatography

Proteins with very low purity after IMAC purification were further purified by ion-exchange chromatography (IEC). High purity is also desired for crystallisation and protein structure studies. Ion exchange chromatography was performed using the Bio-Rad BioLogic DuoFlow™ System connected to a UNO™ Q12 anion exchange column (Bio-Rad). The flow rate of applied samples was 1 ml/min and purified samples were collected using a Bio-Rad BioFrac™ fraction collector (Bio-rad). Protein samples were initially dialysed overnight into 10 mM Tris/HCl pH 8.0 (Buffer A) and then loaded onto the column (equilibrated with ~200 ml of buffer A) through a 4 ml loop. The elution buffer (Buffer B) contained 10 mM Tris/HCl pH 8.0 and 500 mM

NaCl. Gel filtration was done using the same system but with a 16/60 Superdex™ 200 gel filtration column. The equilibration and elution buffer was 10 mM Tris/HCl pH 8.0 containing 150 mM NaCl. 1.5 ml fractions of purified samples were collected and analysed by SDS PAGE. At the end, desired fractions were pooled together and concentrated into a buffer of choice.

II.1.27 Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

Protein expression was analysed by SDS-PAGE as described by Laemmli (Laemmli, 1970). Samples were routinely analysed on 12.5 % polyacrylamide gels (Acrylogel 3; BDH Electran®) using the Bio-rad Mini-PROTEAN® Tetra Cell system (Bio-rad) according to the manufacturer's instructions. Information on the preparation of solutions and buffers for SDS-PAGE experiments is given below in Table II.9. 5 µl of SDS loading buffer was added to 5 µl of the pellet fraction and 10 µl of all other fractions (CFE, FT, 10 mM, and 100 mM fractions). Except otherwise stated, samples were often boiled at 98 °C for 2-3 min in a boiling water bath after which they were cooled to room temperature and applied to SDS-PAGE gels. Samples were applied alongside protein standards (Figure II.2) to enable the estimation of protein molecular weights after staining. SDS PAGE was run at 150 volts and gels after electrophoresis were stained with InstantBlue™ stain (Expedeon) for at least 15 min, after which they were washed in excess distilled water overnight followed by image acquisition using a Canon PowerShoot A75 camera (Canon).

Component	Volume/Amount
Resolving gel (12.5 %)	~For 4 gels
0.75 M Tris/HCl buffer, pH 8.8 with 0.2 % SDS	9.4 ml
40 % Acrylamide (BDH Electran acrylamide, 3 % (w/v) bisacrylamide)	5.8 ml
d.d. H_2O	3.5 ml
10 % (w/v) Ammonium persulphate	90 µl
TEMED	30 µl
Stacking gel	
0.25 M Tris/HCl buffer, pH 8.8 with 0.2 % SDS	3.75 ml
40 % Acrylamide (BDH Electran acrylamide, 3 % (w/v) bisacrylamide)	0.75 ml
d.d. H_2O	3.0 ml
10 % (w/v) Ammonium persulphate	60 µl
TEMED	20 µl
Sample/Loading buffer	
SDS	10 % (w/v)
0.25 M Tris/HCl buffer, pH 8.8 with 0.2 % SDS	5 ml
Glycerol	25 % (w/v)
β-mercaptoethanol	2.5 ml
Bromophenol blue dye	0.1 %
Running buffer	
32 mM Tris/190 mM glycine, pH 8.3	350 ml
SDS	0.1 %

Table II.9 - Preparation of SDS PAGE gels and buffers

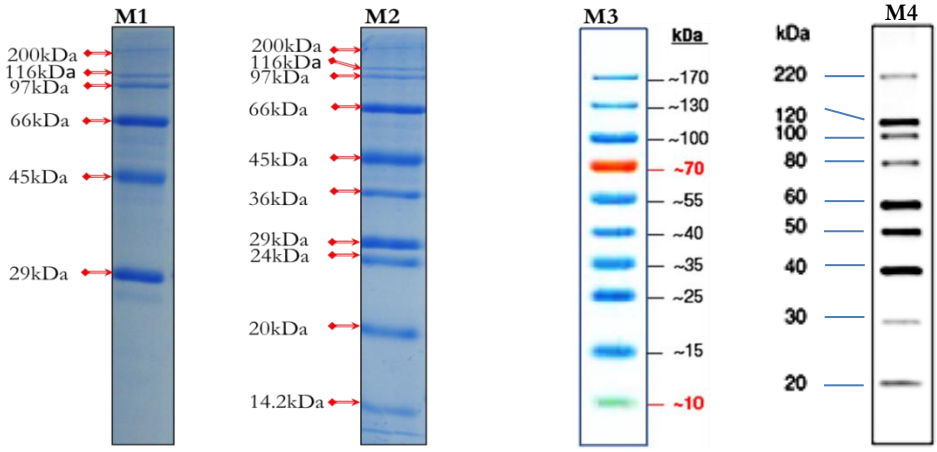


Figure II.2 - Band profile of protein markers used in this study. M1: Sigma high molecular weight markers (Sigma, Cat. SDS6H2) on 12.5% SDS-PAGE gels. M2: Sigma low molecular weight markers (Sigma Cat. SDS7) 12% SDS-PAGE gels. M3: PageRuler Prestained Protein markers (Thermo Pierce Cat. SM0671) on 12% SDS-PAGE M4: MagicMark# XP Western Protein Standard (20-220 kDa) (Life technologies) on Western blot.

II.1.28 SDS-agarose gel electrophoresis (SAGE)

SAGE was used for the electrophoresis of very high molecular weight molecules (>200kDa)such as mucin glycoproteins (Szabady *et al.*, 2011). SAGE gels were prepared by mixing 1% agarose and 0.1% SDS in a total volume of 20 ml TAE buffer (40 mM Tris acetate, pH 8.3, containing 1 mM EDTA). The mixture was heated in a microwave oven (400 watts) for about 2 min and allowed to cool to about 60 °C before casting in an HU10 Mini Plus Horizontal Gel Unit (SCIE-PLAS Ltd) agarose gel mould. Electrophoresis using SAGE gels was routinely performed at 100 V in TAE buffer.

II.1.29 Western blotting

The transfer of electrophoresed proteins and glycoproteins onto Nitrocellulose (NC) or polyvinylidene difluoride (PVDF) membranes was achieved by Western blotting using the Biorad Trans-Blot Turbo Transfer System (Bio-rad). PVDF membranes cut to an appropriate size were initially soaked in methanol for 30 s and immediately rinsed in distilled water for 30 s. The membranes were then submerged in transfer buffer for another minute or two before use in blotting. NC membranes on the other hand were ready for use by simply soaking in transfer buffer for 2 min. A transfer sandwich containing two filter papers (thoroughly soaked in transfer buffer) at the bottom, followed by the membrane (PVDF or NC), the gel and another set of two

filter papers was prepared and introduced in the blotting equipment according to the manufacture's instructions. Blotting was carried out at 25 V for 30 min.

II.1.30 Buffer exchanging and concentrating proteins

Vivaspin™ centrifugal filter concentrators (VivaScience) were routinely used to concentrate and buffer exchange proteins. Briefly, protein samples to be concentrated were transferred into concentrators with the appropriate molecular weight (5, 10 or 30 kDa) cut-off filter, followed by centrifugation at ~3500 × g for about an hour using a swing bucket type - MSE Mistral 3000i bench centrifuge (MSE, UK). This step could be repeated several times to exclude more of the unwanted buffer thus concentrating the protein solution. When buffer exchanging, the concentrated proteins were diluted in a buffer of choice or water (for crystallography experiments) and the process repeated about three more times.

II.1.31 Protein crystallization screen

Crystallisation screens were performed using the sitting drop vapour diffusion method (Figure II.3) with commercially available screen solutions [JCSG+, PACT, and STRUCTURE – (Qiagen)]. The concentration of proteins used for crystal screens ranged from 10 to 20 mg/ml. Proteins were added to 96-well plates containing screening solutions using a mosquito™ (TTP Labtech) nanolitre pipetting robot at a rate of 1+1, 2+1 (1ul of protein plus 1µl of screen solution and 2 µl protein plus 1 µl of screen solution). In some cases, a known ligand for the protein to be crystallized was added to the protein (10 mM of ligand) before mixing with screen solutions. Crystal structures were viewed using a Leica MZ-6 crystallization microscope (Leica MICROSYSTEMS)

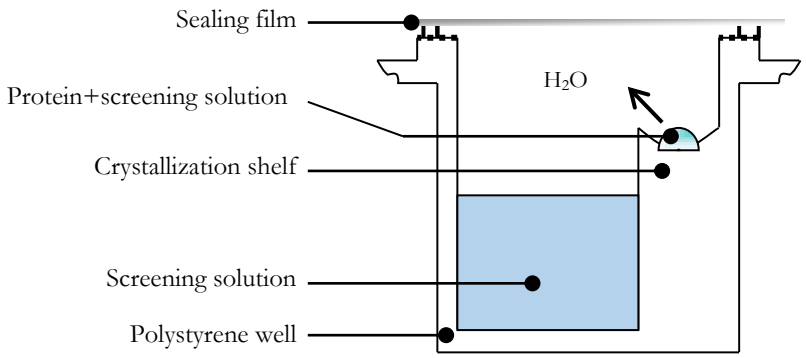


Figure II.3 - Set-up of a sitting drop vapour diffusion system.

II.1.32 Isothermal titration calorimetry (ITC)

Isothermal Titration Calorimetry (ITC) was used to assess the binding of recombinantly expressed putative carbohydrate binding modules to sugars. This was routinely performed using a MicroCal™ VP-Isothermal Titration Calorimeter (Microcal, USA). Recombinant proteins were extensively dialyzed overnight against a buffer of choice and then filtered alongside the dialyses buffer using a sterile 1.2 μm filter (acrodisc). Sugars to be tested were dissolved in the filtered dialysis buffer to desired concentration. Dissolving in the same buffer helps to minimize heats of dilution during titration into the recombinant protein. Filtered proteins and sugars were then degassed and applied to the micro-calorimeter according to the manufacturers' instructions. Typically 27 injections (10 μl per injection) of the degassed ligand into the protein solution in the reaction cell were made with rapid stirring (307 rpm), at 300 s intervals. Following each injection, the heat evolved (in exothermic reactions) or absorbed (as in endothermic reactions) due to the interaction of the protein and the ligand is calculated from the electrical power required to maintain the temperature of the reaction cell against that of the reference cell. Data were fitted using the MicroCal Origin software (version 7.0) by non-linear regression and applying a simple one-site binding model yielding the association constant (K_a), stoichiometry of binding (n), the enthalpy of binding (ΔH) and the entropy of binding (ΔS). These were then used to calculate other thermodynamic parameters such as ΔG and TΔS using the standard thermodynamic equation shown below.

$$-RT\ln K_a = \mathbf{\Delta}G = \mathbf{\Delta}H - T\mathbf{\Delta}S$$

Where R = gas constant (1.99 $cal.K^{-1}.mol^{-1}$), T = temperature in Kelvin (298.15 K), **Δ**G = change in Gibbs free enthalpy, **Δ**H = enthalpy change, **Δ**S = entropy of binding.

II.1.33 Thin Layer Chromatography (TLC)

Enzyme catalysed hydrolysis reactions were also analysed by TLC which enables the chromatographic separation of low molecular weight sugars. Samples analysed by TLC were initially boiled for at least 5 min at 98 °C, allowed to cool to room temperature and later centrifuged by pulsing. 3 μl of each sample was applied separately (1 cm apart) to appropriately cut Silicagel 60 TLC plates (Merck) at a distance of 1 cm from the edge of the plate. Spots where dried using a BaByliss® hair dryer (BaByliss®) and another 3 μl of same samples re-applied to dried spots. After drying for the second time, TLC plates were placed in a glass chromatography tank (23 × 23 × 7.5) containing a <1cm high solvent mixture of 1-butanol/acetic acid/water (2:1:1, v/v). Freshly prepared solvent was often initially allowed for at least 2 h before use to

allow vapours to equilibrate in the tank. Plates were positioned perpendicularly to the bottom of the tank so that only the 1 cm region between the edge of the plate and the point where samples were applied on the plate was immersed in the solvent. The solvent migrated up the plates carrying spotted samples and the experiment was stopped after the solvent reached at least 1cm close to the top edge of the plate. Plates were carefully dried using a hair dryer and placed back in the solvent tank for another run. After the second run, plates were dried and this time fully immersed for a few seconds in either orcinol sulphuric acid reagent (sulphuric acid/ethanol/water 3:70:20 v/v, orcinol 1 ‰), or DPA (1.7% w/v Diphenylamine, 1.7% v/v Aniline, 85% v/v acetone, and 11% Phosphoric acid) solution (Anderson *et al.*, 2000) if dealing with highly charged sugars. The plates were dried again and finally taken to a 120 °C oven for at least 10 min or >30 min if using DPA. Revealed sugar spots on plates were photographed using Canon PowerShoot A75 camera (Canon). Sugar standards were included to help with the identification of unknown spots.

II.1.34 High performance liquid chromatography (HPLC)

Enzyme catalysed hydrolysis reactions were also analysed by a more sensitive approach involving HPLC. The HPLC column used was the Dionex CARBOPAC™ PA-100 column within an automated Dionex DX500 and ICS3000 system (Dionex). Sugars were detected by pulsed amperometric detection (PAD) with settings E1= +0.05, E2= +0.6, E3= -0.6. Samples analysed were initially boiled and centrifuged at 13 000 rpm for 5 min (using a HERAEUS, PICO17 benchtop centrifuge) leaving behind supernatants that were applied to the HPLC machine. Standards were also included to help with the identification of unknown peaks. HPLC data was analysed using the Chromeleon TM chromatography software (Version 6.8) (Dionex) and GraphPad Prism (Version 7.0) (Prism).

II.1.35 Concentrating purified sugars by freeze drying

Purified oligosaccharide and polysaccharide sugars were frozen to -80 °C and then lyophilised in a Christ Alpha 1-2 Freeze Drier (Martin Christ Gefriertrocknungsanlagen GmbH) at - 60 °C.

II.1.36 Mucinase assays

Enzymatic degradation of mucin substrates was evaluated using a combination of SDS-agarose gel electrophoresis (SAGE), Western blotting and lectin-based detection techniques. Mucin substrates used in this study included bovine submaxillary mucin Type I-S (BSM) and porcine

stomach or gastric mucin Type III [bound sialic acid 0.5-1.5 % (PGMIII)] and Type II [bound sialic acid, ~1% (PGMII)] (Sigma, UK). Except otherwise stated, BSM and PGM stock solutions were prepared in Talon buffer (Section II.1.24.2) and 50 µl of each was incubated in a 37 °C oven overnight with various concentrations of the enzyme to be tested in a final volume of 200 µl, made up with Talon buffer. 50 mM EDTA was also added to similarly prepared samples to test the metal dependency of recombinant proteins with mucinase activity. The next day, samples were collected and pulsed and 5 µl of SDS sample buffer added to 10 µl of each sample. Samples were boiled at 98 °C for 3 min after which they were allowed to cool to room temperature and gently centrifuged by pulsing. 9 µl of each sample was then separated on SAGE gels (Section II.1.28) at 100 V for about an hour in TAE buffer. Samples were then blotted onto appropriately sized cut PVDF membranes (Amersham) as described in Section II.1.29. Blots were blocked in excess PBS containing 0.5% Tween 20 (PBS-Tween 20) for 1 h before probing with 1: 1000 dilution of 1 µg/µl biotinylated wheat germ agglutinin (WGA, Sigma) for another hour. Washing was performed again with excess PBS-Tween 20 for 1 h, this time replacing the washing solution with new solution every 10 min (6 x 10 min changes). Washed blots were then treated with a 1:2000 dilution of ExtrAvidin®−Peroxidase conjugate (Sigma cat.; E2886) in washing solution for 1 h. This was followed by another wash step with PBS-Tween 20 (6 x 10 min changes) before chemiluminescence detection with luminol/enhancer from the Biorad Immun-Star™ WesternC™ Chemiluminescence kit (Bio-rad cat.; Kit #170-5070). Luminol and enhancer solutions were mixed in equal proportions (1 ml each), spread on washed membranes and chemiluminescent signals recorded using a ChemiDoc XRS system (Bio-rad).

II.1.37 Periodic Acid-Schiff staining

Periodic Acid-Schiff (PAS) is general glycoprotein stain that allows for the detection of peripheral sugars in glycoproteins. Using this method, gels containing electrophoresed mucins could be stained instantly without the need for Western blotting. Staining was done using the Sigma glycoprotein detection kit (Sigma cat.; GLYCOPRO-1KT) according to the manufacturers' instructions. A summary of the staining protocol is provided in Table II.10. Samples destined for analyses by PAS detection were often electrophoresed using 4-15% gradient gels (Ready Gel Tris-HCl Gel, 4-15% linear gradient, 10-well, 30 #l, 8.6 x 6.8 cm (W x L) purchased from Biorad.

Steps	Time for gel thickness 0.5-0.75 mm or for membrane	Time for gel thickness 1.0-1.5 mm
1. Fixing	30 min.	60 min.
2. Washing	2 x 10 min.	2 x 20 min.
3. Oxidation	30 min.	60 min.
4. Washing	2 x 10 min.	2 x 20 min.
5. Staining	1-2 hours or until bands turn magenta	1-2 hours or until bands turn magenta
6. Reduction	60 min.	120 min.
7. Washing	Band color will intensify with changes of fresh water	Band color will intensify with changes of fresh water
8. Storage	overnight	overnight

Table II.10 – Summary of steps followed during the PAS staining procedure

II.1.38 IgA protease assays

These were performed to determine if expressed recombinant proteins had IgA protease activity. The protocol was a combination of SDS-PAGE, Western blotting and immunochemical detection techniques. Human myeloma IgA1 (cat.; 400109-500UG) and IgA2 isoforms (400110-500UG) (calbiochem), were separately incubated with the desired concentration of the recombinant putative IgA protease enzyme (Table II.11) and the final volume of the mixture made up with Talon buffer (Table I.II). Mixtures were incubated at 37 ^{o}C in a water bath for different time periods depending on the aim of the experiment. Samples, post incubation were centrifuged by pulsing and 10 µl of each treated with 5 µl of SDS-PAGE sample buffer and boiled at 98 ^{o}C for 3 min. After SDS-PAGE and Western blotting of samples (II.1.27 and II.1.29 respectively, PVDF membranes containing blotted proteins were washed in excess PBS-Tween 20 (Section II.1.36) for 1 h at room temperature before application of a 1:2000 dilution of primary mouse Anti-Human IgA1 antibodies conjugated to biotin (SouthernBiotech cat.; 9140-08). This was followed by another washing step this time for 30 min and replacing the washing solution with new solution every 10 min (3 x 10 min changes). Washed blots were then treated with a 1: 2000 dilution of ExtrAvidin®−Peroxidase conjugate (Sigma cat.; E2886) in washing solution for 1 h. A final washing step was performed as above before chemiluminescence detection with luminol/enhancer solutions from the Biorad Immun-Star™ WesternC™ Chemiluminescence kit (Bio-rad cat.; Kit #170-5070). Luminol and enhancer solutions were mixed in equal proportions (1 ml each), spread on washed membranes and chemiluminescent signals recorded using a ChemiDoc XRS system (Bio-rad). To determine the effect of deglycosylation on IgA1 protease activity, IgA1 was initially treated with various enzymes including a commercial sialidase/neuraminidase enzyme from *Clostridium perfringens* (*C. welchii*) (Sigma cat.; N2876), a β-galactosidase and α-N-acetylgalactosaminidase enzymes (expressed during the course of this study) in different combinations overnight before incubation with the putative IgA protease enzyme. IgA deglycosylation was monitored using a biotin - conjugated

Helix aspersa (garden snail) agglutinin (HAA) (Sigma cat.; L8764) that binds N-acetylgalactosamine (GalNAc). This was performed on a separate blot containing replicate samples and using a 1: 1000 dilution of the lectin.

Set-up	A	B	C	D	E	F	G	H
IgA1 (0.5 mg/ml)	10µl	10µl	10µl	10µl	10µl	10µl	10µl	10µl
Sialidase/neuraminidase (5 U/ml)			2.5µl	2.5µl	2.5µl	2.5µl	2.5µl	2.5µl
GH2 (0.5 mg/ml)					2.5µl	2.5µl	2.5µl	2.5µl
GH109 (0.5 mg/ml)							2.5µl	2.5µl
Talon (pH 8.0)	2.5µl		2.5µl		2.5µl		2.5µl	
Putative IgA protease (0.5 mg/ml)		2.5µl		2.5µl		2.5µl		2.5µl
Talon (pH 8.0)	7.5µl	7.5µl	5µl	5µl	2.5µl	2.5µl		

Table II.11 - Example of an experimental set-up to test protease activity against IgA1 and the effect of sequential IgA1 deglycosylation on protease activity. GH2 and GH109 display β-galactosidase and α-N-acetylgalactosaminidase activities respectively.

II.1.39 N-terminal sequencing of proteins

N-terminal Edman sequencing of digested IgA1 fragments was performed using ABI high-throughput 'Procise' 494 HT sequencers (AltaBioscience, UK). 10 µg of the IgA1 sample that had been digested for over 48 hr at 37 ºC was electrophoresed by SDS PAGE. After electrophoresis, gels were stained with Coomassie blue stain for about 2 h and later on washed in 50% methanol solution. Target bands containing digested IgA1 fragments were excised using a scalpel blade, placed into 1.5 ml eppendorf tubes and posted to AltaBioscience for protein sequencing. Samples were also blotted on PVDF membranes and sent for N-terminal Edman sequencing at Alphalyse (Alphalyse, Denmark).

II.1.40 Universal Protease Activity Assay

To determine if recombinant proteins had general protease activity or where specific to particular substrates, universal protease activity assays were performed following the protocol described for the Sigma Universal Protease Activity Assay (Sigma). According to the principle, the treatment of casein with a general protease leads to the release of tyrosine residues which react with Folin & Ciocalteu's Phenol Reagent (Sigma cat.; F9252) to yield a colour change that can be quantified spectrophotometrically (A600nm). This can be compared to a standard curve (Figure II.4)

prepared by measuring the absorbance of known μmole quantities of tyrosine. Some important modifications to the protocol included an increase in the incubation period of the test protein-casein mixture to 24 h rather than 3 h as recommended in the manufacturer's protocol. This was to allow enough time for degradation of casein. Secondly absorbance readings were all measured at 600 nm instead of 660 nm due to technical limitations of the machine [Eppendorf Biophotometer (Eppendorf)].

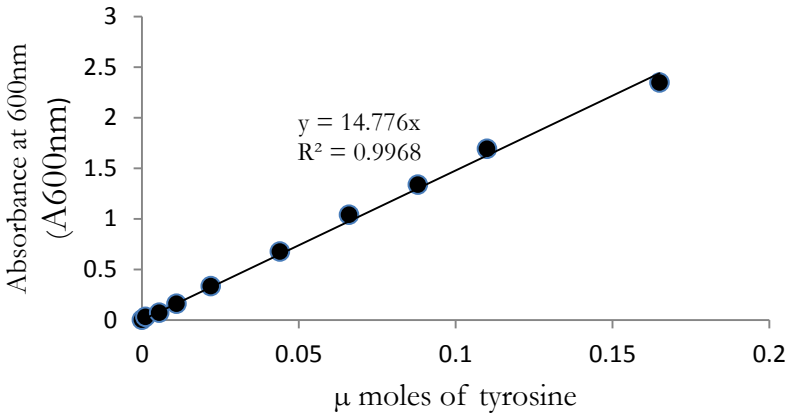


Figure II.4 - Standard curve used in Universal Protease Activity Assays.

II.1.41 Sucrose density gradient centrifugation (SDGC)

Inner and outer membrane lipid bilayers of *B. thetaiotaomicron* were separated from each other by sucrose density gradient centrifugation (SDGC) (Kotarski and Salyers, 1984). 100 ml of *B. thetaiotaomicron* cells grown to stationary phase i.e. $A600_{nm}$ > 0.9 in minimal medium containing 1%PGMIII (1%MM - PGMIII) were harvested by centrifugation (12, 000 x g for 15 min) at 4°C. Cells were re-suspended in 20 ml of 10 mM HEPES buffer pH 7.4 at 4 °C and harvested as before. The washed cell pellet was drained and resuspended to 1.33 ml of 10 mM HEPES buffer pH 7.4 containing 10% sucrose. 0.25 mg of RNase A and Pancreatic DNase I were added to the cell suspension and the cells sonicated for 2 min (0.5 second cycling) on ice using a B. Braun Labsonic U sonicator (B. Braun, Germany) set at low intensity (~45 watts). The cell lysate was centrifuged at 17,000 x g for 5 min followed by a further centrifugation of the resulting supernatant in new centrifuge tube to remove excess cell debris. The resulting CFE was diluted up to 1.7 ml with 10 mM HEPES buffer and layered onto a two-step sucrose gradient containing 440 μl of 70% sucrose at bottom of a 2.0 ml solution of 37 % sucrose (Figure II.5). High speed centrifugation using a Beckman SW60 rotor at 140,000 x g for 1 h was then performed on the

sample and membranes in the resulting solution above the 37% sucrose solution and at the 10%/37% sucrose interface ("yellow material" in Figure II.5) where collected and diluted to 4 ml with HEPES buffer. This sample was further pelleted through a 108 µl volume of a 37% sucrose cushion by centrifugation in a Beckman SW60 rotor at 140,000 x g for 3 h. The pellet material from the 37% sucrose pad was diluted to 1 ml with HEPES buffer and this constituted the mixed membrane fraction (MM). The resulting supernatant was further centrifuged at 300,000 x g for 3 h using a Beckman 70 Ti rotor and the pellet or inner membrane fraction resuspended in 1 ml HEPES buffer. Membranes trapped in the 37% and 70% sucrose gradient of the original two-step gradient ("white material" in Figure II.5) were diluted to 4 ml with HEPES buffer and centrifuged (140, 00 x g for 3 h at 4ºC) through 108 µl of a 57% sucrose cushion in using a Beckman SW60 rotor. The outer membrane enrichment which constituted the pellet of the centrifuged sample was collected and resuspended/diluted to a final volume of 1 ml with HEPES buffer. Please see below in Figure II.5 for a summary of the procedure. All samples were stored at -20 ºC and used later for SDS-PAGE and Western blotting analyses.

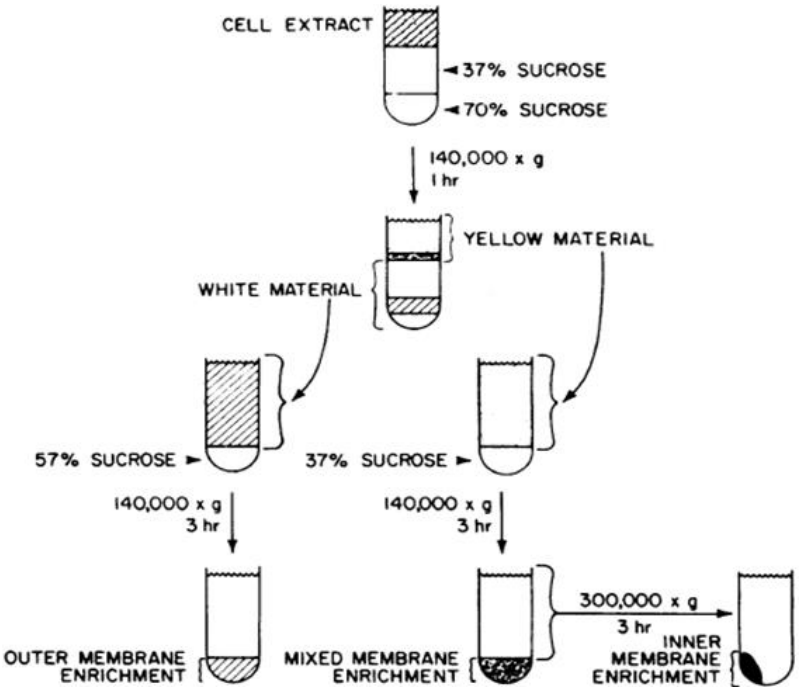


Figure II.5- Summary of the protocol for the purification of inner and outer membrane lipid bilayers of *B. thetaiotaomicron.* Details of the method are described above in the preceding text. Please see further details of the protocol in Kotarski and Salyers, (1984) from where the above figure was adapted.

II.1.42 Immunofluorescence assays (*B. thetaiotaomicron*)

Cells were grown on defined substrate in this case 5 ml of minimal medium containing 1% PGMII to an $A600_{nm}$ of 0.4. PGMII instead of PGMIII was used due to unexplained high background signals observed during fluorescence imaging of cells when using the latter. 5 ml of a 9% formalin solution prepared in PBS was added to the culture and cells were fixed by rocking overnight at 4 ^{o}C. Cells were then harvested by centrifuging at 5000 rpm for 5 min [using a fixed angle bench centrifuge (Hettich Lab technology)] and washed three times in 5 ml of PBS. After washing they were resuspended in 5 ml of blocking solution containing 2% normal goat serum (Invitrogen cat.; PCN5000) and 0.02% NaN_3 in PBS overnight at 4 ^{o}C while rocking slowly. The next day, cells were centrifuged at 13,000 rpm for 1 min (using a HERAEUS, PICO17 benchtop centrifuge).To cells harvested from 1 ml of blocking solution, 1 ml of a 1:500 dilution of primary antibody solution was added in a 1.5 ml eppendorf. Cells were allowed for 2 h in primary antibody solution at room temperature while rocking. The primary antibody solution was removed by centrifugation at 13,000 rpm for 1 min (using a HERAEUS, PICO17 benchtop centrifuge) and cells resuspended in fresh PBS. Resuspended cells were washed by rocking 10 min followed by centrifugation as before. The process was repeated three times every 10 min for 30 min. After washing a 500 µl of a 1:200 dilution of secondary goat anti-rabbit Alexa Fluor® 594 antibodies (Invitrogen cat.; A-11037) was added to harvested cells and allowed for 1h at room temperature while rocking. After another washing step cells were finally resuspended in 100 µl of PBS and a single drop of ProLong® Gold antifade reagent (Invitrogen) added before cells were taken for microscopy. All phase contrast and fluorescence images were captured using an Andor iXonEM+ 885 EMCCD camera coupled to a Nikon Ti-E microscope (Nikon) using a 100x/NA 1.4 oil immersion objective. Images were acquired with NIS-ELEMENTS software (Nikon) and processed using ImageJ tool (http://rsbweb.nih.gov/ij/)

II.1.43 Colorimetric assays

Colorimetric assays involving chromogenic para-nitrophenyl (*p*NP) linked substrates were used to qualitatively and quantitatively measure the activity of glycoside hydrolases analysed in this study.

II.1.43.1 *p*NP-substrate screens for β-galactosidase enzyme

*p*NP-substrate screens were carried out following a modification of the protocol described for the OZ BIOSCIENCES ONPG β -Galactosidase Assay Kit (OZ BIOSCIENCES, France). In brief, 4 mg/ml of each *p*NP substrate was prepared in Talon buffer (20 mM Tris/HCl pH 8.0 containing

100 mM NaCl). 10 µl of 0.5 mg/ml enzyme was added to the solution and incubated at 37 °C for 5 min. The reaction was stopped by addition of 250 µl of 1M Na_2CO_3. Absorbance readings were recorded from 400 µl volumes of each reaction mixture. These were compared with similar reactions containing 0.5 mg/ml of BSA in place of the enzyme as control.

II.1.43.2 Measurement of enzyme kinetic parameters using pNP-substrates

To measure the rate of *p*NP –substrate hydrolysis, different concentrations of the *p*NP substrate in a total volume of about 400 µl were incubated with 100 µl of the enzyme at a desired concentration. Reactions were allowed to proceed at 37 °C in quartz cuvettes and *p*NP release was measured using a Pharmacia Ultrospec 4000 spectrophotometer at 420 nm (A420nm). Graphs of reaction velocity versus substrate concentration were generated and fitted to the Michaelis–Menten kinetics model using GraphPad Prism (version 6.0), allowing for the calculation of enzyme kinetic parameters.

II.1.44 β-galactosidase activity assays

β-galactosidase activity was measured using the Megazyme galactose detection kit (Megazyme). A summary of the principle of the method is shown in Figure II.6

Principle

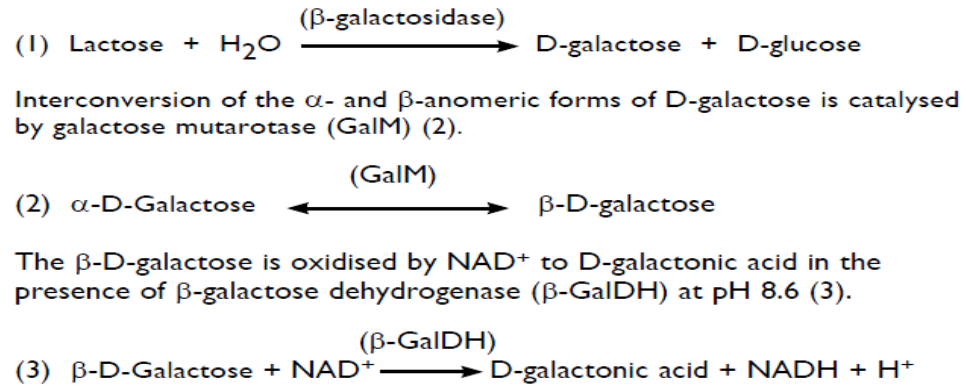


Figure II.6 – Principle of β-galactosidase assays. GalM: galactose mutarotase, GalDH; β-Galactose dehydrogenase, (Source: Megazyme Galactose detection kit)

The modified reaction set-up is shown in Table II.12. Reactions were allowed to proceed at 37 °C in quartz cuvettes and galactose release was monitored by measuring the rate of NADH production (molar extinction coefficient; 6200 M^{-1} cm^{-1}) over time using a Pharmacia Ultrospec 4000 spectrophotometer at $A340_{nm.}$ All experiments were performed in triplicates and kinetic

parameters were estimated in the same manner as described for colorimetric assays using GraphPad prism 6.0 (Prism).

Components	Volume	Volume /2 (used)
Water	2 ml	1 ml
Buffer (20 mM Tris pH 8.6)	200 µl	100 µl
Glycan (10 mM)	200 µl	100 µl
GalDH/GalM	20 µl	10 µl
NAD+	100 µl	50 µl
Enzyme (0.5 mg/ml)	200 µl	100 µl
Total	2.72 ml	1.36 ml (1360 µl)

Table II.12 - Set-up for β-galactosidase activity assays. GalM: galactose mutarotase, GalDH; β-Galactose dehydrogenase, NAD+: nicotinamide adenine dinucleotide.

II.1.45 Sugar kinase assays

The ability of recombinant proteins to phosphorylate various sugars including amino sugars present in mucins was evaluated by performing sugar kinase assays as described in Reith *et al.*, 2011. The general principle and the reaction set-up are summarized in Figure II.7 and Table II.13 respectively.

Principle

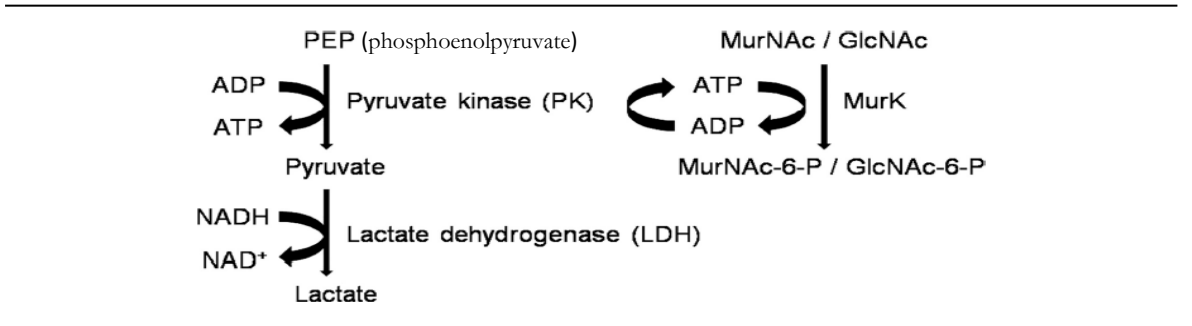


Figure II.7 – General principle of sugar kinase assays. The target sugars for phosphorylation in this case are MurNAc or GlcNAc for N-acetylmuramic Acid and N-acetylglucosamine respectively, and the kinase enzyme is MurK, a MurNAc/GlcNAc kinase from *Clostridium acetobutylicum*. A similar scheme was followed in this study to test the ability of a putative kinase enzyme BT4240 from *Bacteroides thetaiotaomicron* to phosphorylate various mucin/non-mucin sugars. The above image was adapted from (Reith *et al.*, 2011)

Components	Original concentration	Volume per 1ml reaction
1. PEP	10 mM	100 µl
2. NADH	2 mM	100 µl
3. ATP	50 mM	100 µl
4. MgCl2	100 mM	100 µl
5. Tris –HCL (pH 8.0)	1000 mM	100 µl
6. PK	100 U/ml	100 µl
7. LDH	80 U/ml	100 µl
8. Substrate	Varies	100 µl
9. Enzyme	Varies	200 µl
Total	**-**	**1000 µl**

Table II.13 - Set-up for Amino sugar kinase assays. During the assay, samples were mixed chronologically up till LDH followed by addition of the desired substrate. 400 µl of the mixture containing the substrate was placed in a Pharmacia Ultrospec 4000 spectrophotometer for 1 min at 37 ^{o}C before addition of 100 µl of enzyme solution. Change in NADH was recorded by absorbance at 340 nm (A340nm). PEP: phosphoenolpyruvate, NADH: Nicotinamide adenine dinucleotide, ATP: Adenosine Tri-Phosphate, PK: pyruvate kinase, LDH: lactate dehydrogenase.

II.1.46 *B.thetaiotaomicron* counter-selectable gene deletion and competition experiments

B. thetaiotaomicron deletion mutants were created by an adaptation of the standard counter-selectable allelic exchange procedure (Modified by EC Martens of the University of Michigan Medical School, USA). It is based on the sensitivity of wild type *B. thetaiotaomicron* species to the deoxyuridine nucleotide analogue analog, 5-fluoro-2-deoxy-uridine (FUdR) (Yagil *et al.*, 1971). Phosphorylation of FUdR by a thymidine kinase enzyme (Tdk or BT2275) produced by *B. thetaiotaomicron* forms 5-fluoro-2-deoxyuridine monophosphate (FdUMP) which is capable of poisoning the de- novo thymidine biosynthetic pathway through its irreversible binding to ThyA (BT2047), another enzyme in the pathway. The *B. thetaiotaomicron* strain used in this protocol contained a deletion of *tdk* gene (gene that encodes the Tdk enzyme) which still has a functional *de novo* pyrimidine pathway. For convenience, this strain will be referred to in this study as 'wild type' (WT) or *B. thetaiotaomicron tdk*. The counter-selectable suicide vector used was the pExchange-*tdk* vector that contains amongst other components a copy of the the *tdk* gene (Appendix C, Figure C.2). Chromosomal re-integration of the *tdk* gene during the experiment restores FUdR sensitivity allowing for counter-selection against pExchange-*tdk* cells.

II.1.46.1 Building the knockout construct

The chromosomal region to be deleted referred to here as X (Figure II.8) was identified and two sets of primers (A and B for flank 1 and C and D for flank 2) designed to amplify about 1 kbp sized flanking regions of X. Primers A and D contained engineered enzyme restriction sites to allow for subsequent cloning into the pExchange-*tdk* counter-selectable suicide vector. Primers B and C contained a complement of each other towards their 5' ends to enable stitching of flanks during sewing PCR. Equal sized PCR - amplified flanks were analyzed by agarose gel electrophoresis followed by gel purification. 1 µl of each purified flank (~90 ng/µl) eluted with ~ 10 µl of EB were used in sewing PCR reactions. The standard sewing PCR was as in Section II.1.15 without genomic DNA and primers but flanks 1 and 2 (i.e. final volume was ~41 µl). After running ~35 PCR cycles, 5 µl of each primer A and D was added for another 35 PCR cycles under the same conditions. At this stage flanks 1 and 2 would have been sewn together (~2 kbp) and ready for cloning into the pExchange-*tdk* vector. 5 µl of sewing reaction was also analyzed by agarose gel electrophoresis for further confirmation. The cloning protocol was as described from sections II.1.14 to II.1.24 except that the *E.coli* strain CC118 λ *-pir* was used in place of One Shot™ TOP10 for the propagation of pExchange plasmids. After cloning, the recombinant vector containing the sewn flanks was sequenced and transformed into *E. coli* S17-1λ *pir* strains (Table II. 1) to yield a strain denoted here as the 'donor'

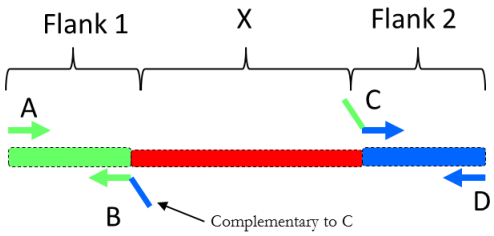


Figure II.8 – Primer design for gene deletion experiments. Outer primers A and D contain restriction sites for cloning into the pExchange-*tdk* suicide vector while inner primers B and C are partially complementary to each other. The latter feature is important for sewing PCR reactions. X represents the chromosomal region or gene targeted for deletion.

II.1.46.2 Conjugation into *B. thetaiotaomicron* and recombinant mutant isolation

The donor strain from Section II.1.46.1 (S17-1 λ *pir*/pExchange-*tdk* + flank1-2) was grown overnight in LB medium containing 50 µg/ml ampicillin (while shaking, 180 rpm, 37 °C) at the same time with a *B. thetaiotaomicron tdk* strain (recipient) in TYG medium (static culture, at 37 °C). Three dilutions, 1:50, 1:100 and 1:200 of each strain in their respective media (5 ml new media)

were prepared, and each s*et a*llowed to grow for 5 h at 37 °C. Cells were collected and centrifuged at 5000 rpm for 5 min [using a fixed angle bench centrifuge (Hettich Lab technology)]. This step was repeated again with cells resuspended in 5 ml TYG. Equal-sized pellets from the donor and recipient were resuspended in a final volume of 1 ml TYG and spread evenly on the surface of BHI/His-Hem agar [BHI medium plus 1:1000 dilution of His-Hem (Section II.1.3)]. Plates containing the conjugation mix were then taken to 37 °C incubator and cells allowed to grow anaerobically for 16-24 h. After growth, the lawn of bacterial biomass from the conjugation plates were scraped into 5 ml of TYG and 100 µl of suspension (alongside a 1:10 dilution of same mixture) plated on BHI/His-Hem agar containing gentamicin (200 µg/ml) and erythromycin (25 µg/ml) antibiotics. Cells were grown anaerobically at 37°C for 2 days. Anaerobic conditions were created using GasPak EZ anaerobe container system sachets (BD) placed in an air-tight anaerocult container containing plated cells. 5-10 individual colonies after growth were re-streaked on fresh BHI-blood agar containing the same antibiotics as before and grown anaerobically for another 2 days. The resulting colonies represented single-recombinant strains in which the pExchange-*tdk* knockout plasmid had recombined with the *B. thetaiotaomicron* genome via one of the two flanks (flank 1 or 2 above). 10 colonies of the single recombinants were then grown separately in 5 ml TYG overnight and equal amounts of each (1 ml) pooled into one volume. 100 µl of the pooled stock (alongside a 10 fold dilution of same stock) were then plated on BHI/His-Hem containing FUdR (200 µg/ml). The cells were grown anaerobically at 37 °C for 2-3 days and 10 FUdR resistant colonies re-streaked on BHI/His-Hem with FUdR (200 µg/ml) for another 2 days. 10 randomly picked colonies after growth were cultured in 5 ml TYG for genomic DNA extraction the next day using the GenElute™ Bacterial Genomic DNA Kit (Sigma) according to the manufacturer's instructions. Genomic DNAs from various colonies were screened for the desired genotype by PCR (using primers A and D) and agarose gel electrophoresis. Positive deletions yielded ~2 kbp fragment following agarose gel electrophoresis. DNAs from positive clones were also sequenced to confirm deletions before use in subsequent experiments.

II.1.46.3 *B. thetaiotaomicron* strain signature tagging with *pNBU2-tetQb*

All *B. thetaiotaomicron* strains used for competition experiments were initially tagged with short oligonucleotide sequences (~24 bp) to enable the differentiation and quantification of strains after co-culture by quantitative real-time PCR (qPCR) (Section II.1.46.5). An overnight culture of the recipient strain (i.e. wild type or deletion mutant) was prepared alongside the S17-1 λ *pir E. coli* strain harboring *p*NBU2-*bla-tetQb* with the desired tag in TYG and LB (150 µg/mL

ampicillin) respectively. Three dilutions, 1:50, 1:100 and 1:200 of each strain in their respective media (5 ml new media and no cysteine in the new TYG) were prepared, and each s*et a*llowed to grow for 5 h at 37 °C. Cells were collected and centrifuged at 5000 rpm for 5 min [using a Hettich Zentrifugen fixed angle bench centrifuge (Hettich Lab technology)]. Equal-sized pellets from the recipient and donor were resuspended in a final volume of 1 ml TYG and spread evenly on the surface of BHI/His-Hem agar (BHI medium plus 1:1000 dilution of His-Hem). Cells were grown anaerobically at 37 °C for 16-24 h. The lawn of bacterial biomass from the conjugation plates were then scraped into 5 ml of TYG and 100 µl of suspension (alongside a 1:10 dilution of same mixture) plated on BHI/His-Hem agar containing gentamicin (200 µg/ml) and tetracycline (25 µg/ml) antibiotics. Cells were grown anaerobically at 37 °C for 2 days and 10 randomly picked colonies after growth were cultured in 5 ml TYG for genomic DNA extraction the next day using the GenElute™ Bacterial Genomic DNA Kit (Sigma) according to the manufacturer's instructions. Genomic DNAs from various colonies were screened for tag insertions by PCR using primers listed in Appendix Table A.5. A tag insertion can occur at one of two serine tRNA sites (NBU2-*att1* and NBU2-*att2*) in the genome of *B. thetaiotaomicron*. PCR screens using NBU2-*att* primers (Appendix Table A.5) were also performed to confirm the insertion site of each tag. Strains used for competition experiments had their tags inserted at similar sites (in this case at NBU2-*att1*) to avoid any off-target effects. Were tag insertion occurred at the NBU2-*att1* site, PCR amplification using NBU2-*att1* primers yielded no products by AGE due to destruction of the primer site while wild type or untagged strains yielded 900 kbp DNA fragments.

II.1.46.4 *In-vitro* competition of wild type and knockout strains on mucins

Competition experiments were performed to evaluate the contribution of specific genetic factors to *B. thetaiotaomicron* fitness on porcine gastric mucin type III (PGMIII). Approximately equal amounts of the signature tagged wild type and deletion mutant initially grown in TYG were mixed and 100 µl used to inoculate minimal medium containing 1% glucose (MM-Glc). 100 µl of cells were subcultured every day for 5 days before switching to new medium containing 1% PGM (MM-PGMIII). Every day before subculturing, 2 ml of MM-PGMIII and 1 ml of MM-Glc cells were collected and frozen at -80 °C for later enumeration by qPCR.

II.1.46.5 qPCR Enumeration of competing strains *in-vitro*

Genomic DNA was extracted from the 2 ml samples of competing cells (Section II.1.43) every day and 10 ng of each days DNA sample assayed in duplicates using a Roche Light Cycler 480

real-time PCR system. The run parameters are shown in Table II.14. A standard 10 µl qPCR mix 5µl of SYBR green I master mix (Roche), 1 µl each of 5 µM forward and reverse primers 2 µl of DNA template and 1 µl of distilled water. To quantify the amount of each tagged strain present, standards prepared from purified genomic DNA of each signature tagged strain were also included. These ranged from 0 – 100ng of purified DNA and were used to create a standard curve to help calculate the percentage representation of each strain in various samples analysed. All data analyses were carried out using the LightCycler ® 480 version 1.5.0.39 software (Roche)

Experiment

Creation Date	02/05/2013 17:57:10	Last Modified Date	02/05/2013 19:04:02
Operator	Didier Ndeh	Owner	Didier Ndeh
Start Time	02/05/2013 17:57:39	End Time	02/05/2013 18:58:20
Run State	Completed	Software Version	LCS480 1.5.0.39
Macro		Macro Owner	
Macro Status			
Templates	Test 1 210113	Plate ID	
Test ID		Lot ID	
Color Comp ID			
Run Notes			

Programs

Program Name	Denature		
Cycles	1	Analysis Mode	None

Target (°C)	Acquisition Mode	Hold (hh:mm:ss)	Ramp Rate (°C/s)	Acquisitions (per °C)	Sec Target (°C)	Step size (°C)	Step Delay (cycles)
95	None	00:05:00	4.40		0	0	0

Program Name	Cycling		
Cycles	45	Analysis Mode	Quantification

Target (°C)	Acquisition Mode	Hold (hh:mm:ss)	Ramp Rate (°C/s)	Acquisitions (per °C)	Sec Target (°C)	Step size (°C)	Step Delay (cycles)
95	None	00:00:05	4.40		0	0	0
50	None	00:00:05	2.20		0	0	0
72	Single	00:00:10	4.40		0	0	0

Program Name	Melt curve		
Cycles	1	Analysis Mode	Melting Curves

Target (°C)	Acquisition Mode	Hold (hh:mm:ss)	Ramp Rate (°C/s)	Acquisitions (per °C)	Sec Target (°C)	Step size (°C)	Step Delay (cycles)
95	None	00:00:05	4.40		0	0	0
65	None	00:01:00	2.20		0	0	0
95	Continuous		0.11	5	0	0	0

Program Name	Cool		
Cycles	1	Analysis Mode	None

Target (°C)	Acquisition Mode	Hold (hh:mm:ss)	Ramp Rate (°C/s)	Acquisitions (per °C)	Sec Target (°C)	Step size (°C)	Step Delay (cycles)

Day1 (50) 02/05/2013 Page 1 of 9

Table II.14 – Typical qPCR run parameters used for competition experiments

II.2 Bioinformatic tools

II.2 .1 LipoP 1.0 Server: Lipoprotein signal peptide prediction in gram negative bacteria (http://www.cbs.dtu.dk/services/LipoP/)

II.2.2 SignalP 4.1 Server: Signal peptide prection in gram positive/negative and eukaryotic sequences (http://www.cbs.dtu.dk/services/SignalP/)

II.2.3 Phobius: Lipoprotein and transmembrane sequence prediction (http://phobius.sbc.su.se/)

II.2.4 PyMol: Protein structure analyses http://www.pymol.org/

II.2.5 Multalin: Multiple sequence alignment tools (http://multalin.toulouse.inra.fr/multalin/)

II.2.6 Xtream: Tandem repeat identification and architecture modeling (http://jimcooperlab.mcdb.ucsb.edu/xstream/)

II.2.7 ProtParam tool Estimation of protein molecular weight and extinction coifficients **(**http://web.expasy.org/protparam/)

II.2.8 NetOGlyc 4.0 Server: prediction of mucin type O-glycosylation sites in mammalian proteins (http://www.cbs.dtu.dk/services/NetOGlyc/)

II.2.9 SWISS-MODEL server: automated protein structure homology-modeling (http://swissmodel.expasy.org/)

CHAPTER III

Biochemical Characterisation of Putative *B. thetaiotaomicron* and *T. vaginalis* M60-like Proteases

III.1 Introduction

B. thetaiotaomicron and *T. vaginalis* represent important human mucosal microbes capable of colonising the human gut and urogential tracts respectively. While *T. vaginalis* is generally pathogenic in nature, the activities of *B. thetaiotaomicron* and a host range of other mutualistic human mucosal microbes can either positively or negatively influence our health, thus making both organisms important public health concerns (Chapter I). As discussed in Chapter I, both organisms are also known to share some gene families including those encoding proteins containing the novel family of domains termed "M60-like/PF13402" domains (Nakjang *et al.*, 2012). The taxonomic distribution, sequence features, domain content and supportive evidence from literature on some members of the M60-like/PF13402 family led us to hypothesize that they could indeed represent important surface zinc metalloprotease enzymes in these organisms processing extracellular glycoprotein targets. This is further supported by evidence suggesting that M60-like domain-containing proteins represent distant relatives of viral-enhancin proteases known to degrade insect mucins (Nakjang *et al.*, 2012, Wang and Granados 1997).

The metabolism of host derived glycoproteins such as mucins has been shown to be crucial for enhanced microbial fitness and colonization at mucosal surfaces (Martens *et al.*, 2008), where a significant number of M60-like positive microbes thrive. Human mucins alongside other prominent mucosal surface glycoproteins such as secretory IgA (SIgA) and underlying epithelial cell glycocalyx components (Ouwerkerk *et al.*, 2013), thus represent potential targets for candidate M60-like proteases. Indeed there is evidence that, in addition to belonging to polysaccharide utilisation loci (PULs) and encoding proteins that are potentially surface exposed (Sections I.8.3), some *B. thetaiotaomicron* M60-like genes including BT_4244 and BT_3015 (Section II.8.3) are also upregulated following exposure of the organism to porcine mucosal glycans (Martens *et al.*, 2008). Upregulation of BT_0277, another M60-like-encoding gene has been observed during *B. thetaiotaomicron* growth on plant (larch) - derived arabinogalactans which are often linked to proteins [arabinogalactan proteins (AGP)] (Martens *et al.*, 2011). *T. vaginalis* M60-like domain-containing proteins on the other hand, including those encoded by the

TVAG_189150 and TVAG_339720 genes of strain G3, are also reportedly surface exposed (de Miguel *et al*., 2010).

Furthermore, family 32 carbohydrate binding modules (CBM32) and PA14 domains detected in several M60-like proteins (Section I.8.2), have been reported in other mucosal microbes such as *Clostridium perfringens* and *Candida albicans,* respectively, to target galacto-configured sugars (Ficko-Blean and Boraston, 2006, Ficko-blean *et al*., 2012, Maestre-Reyna *et al*., 2012, Zupancic *et al*., 2008). Galacto-configured sugars are not only prominent components of several human glycoproteins but also constitute sections of the peripheral and core units of mucins (Section I.2.1.1.1.1, and Figure I.5), further presenting mucin/mucin-related glycoproteins as strong potential targets for candidate M60-like proteases. The association of CBMs [typically appended to carbohydrate acting enzymes (Boraston *et al*., 2004)] with proteases also represents a novel functional context worthy of study.

As earlier indicated, *B. thetaiotaomicron* is a prokaryotic gut mutatualist as opposed to *T. vaginalis* which is a eukaryotic protozoan parasite, implying that studies on M60-like proteins from both organisms will not only enhance our understanding of their role in these separate organisms but also across phyla and in the context of the different symbiotic relationships they are involved in.

III.2 Objectives

This chapter thus aims to analyse the findings of experiments that were designed to test the hypothesis that M60-like domain-containing proteins from *B. thetaiotaomicron* and *T. vaginalis* are glycoprotein targeted extracellular zinc metalloproteases. It includes data from experiments studying the proteolytic, carbohydrate binding activity and cellular localisation of *B thetaiotaomicron and T. vaginalis* M60-like domain-containing proteins.

III.3 Results

III.3.1 Bioinformatics and selection of M60-like entries for biochemical characterisation

A collection of all *B. thetaiotaomicron* and *T. vaginalis* M60-like protein entries from the PFAM database is given in Figure III.1. Given the numbers, only M60-like sequences that were representative of many of the group's features were selected for further biochemical studies. The selection criteria thus included the presence of a zincin motif (HEXXH), a putative carbohydrate binding domain and evidence suggesting possible extracellular localisation [e.g. signal peptides (SP), transmembrane domains (TMD)]. Proteins with supporting information from literature were also given priority. For the purpose of clarity throughout this study, proteins were differentiated from their encoding genes by removing the underscore symbol from the original gene name or locus tag e.g BT4244 to represent the protein encoded by the BT_4244 gene of *B. thetaiotaomicron* and TVAG339720 to represent the protein encoded by the TVAG_339720 gene of *T. vaginalis*

As observed in Figure III.1, each *B. thetaiotaomicron* entry (4 in total) contains one or more putative carbohydrate binding domains including CBM32 and/or BACON domains, an M60-like domain and a type II signal peptide (except for BT0227 with an N-terminal transmembrane domain). All sequences except BT0277 also contain a gluzincin motif (Section I.8.1.1.2) within their M60-like domain and hence the entry was not included for biochemical characterisation in this study.

On the other hand, there were over 25 putative M60-like protein encoding entries from *T. vaginalis* in the PFAM database (Figure III.1). Eleven of these contain a gluzincin motif two of which had been detected in *T. vaginalis* cell membrane extracts (TVAG339720 and TVAG189150). A third protein; TVAG199300 was included based on its similarity in domain content and size to the above two.

In summary, the proteins shortlisted for biochemical characterisation in this study (starred below in Figure III.1) included BT4244, BT3015, BT4272 (from *B. thetaiotaomicron*), and TVAG339720, TVAG199300 and TVAG189150 (from *T. vaginalis*).

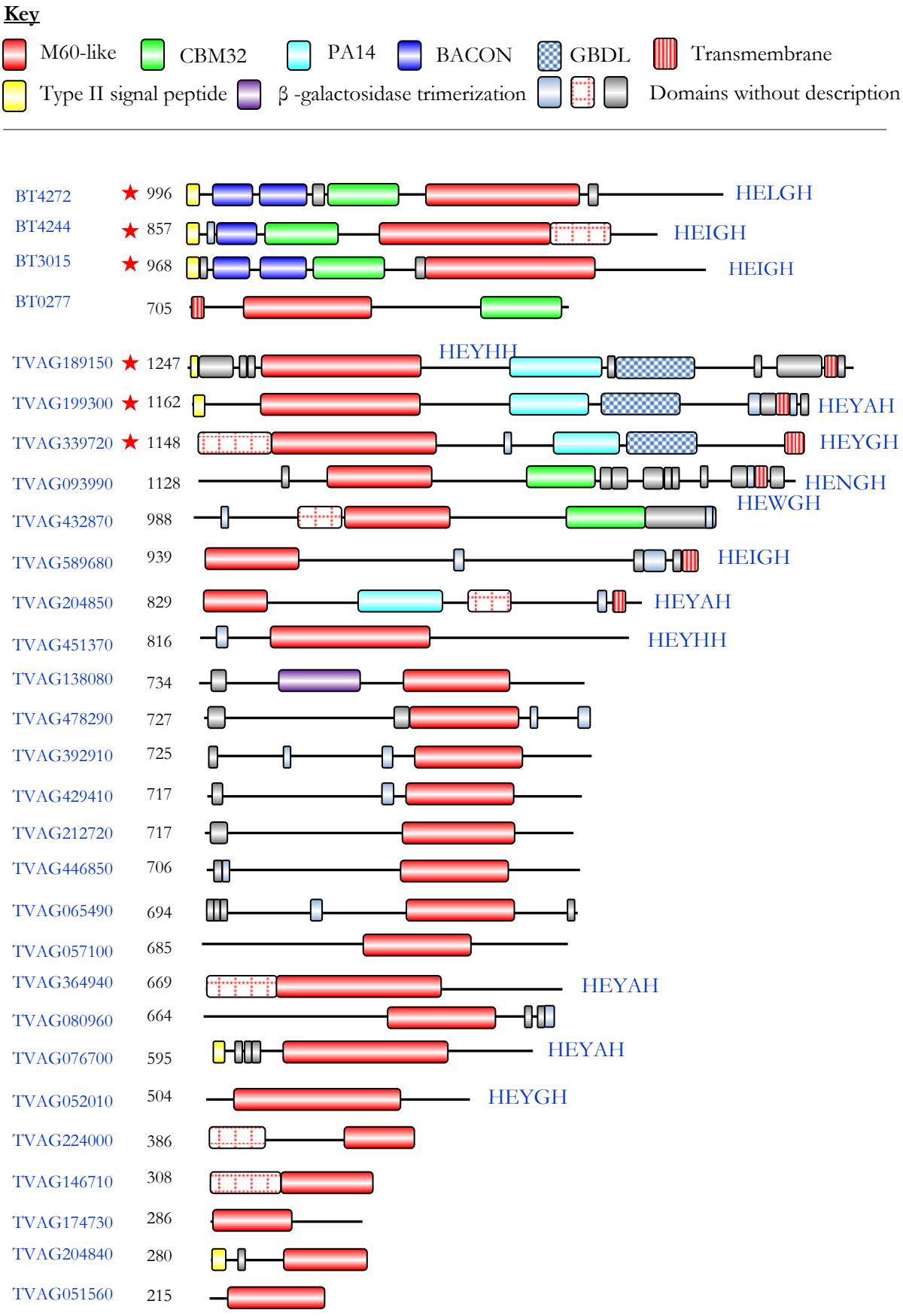


Figure III.1 - Domain architecture of *B. thetaiotaomicron* and *T. vaginalis* M60-like entries from the PFAM database. Starred sequences are those that were targeted for biochemical characterisation in the current study. Sequences commencing with BT are from *B. thetaiotaomicron* while those with TVAG are from *T. vaginalis*.

III.3.2 Gene cloning and expression

Gene fragments encoding various domains from shortlisted *B. thetaiotaomicron* and *T. vaginalis* M60-like entries namely BT4244, BT3015, BT4272 (from *B. thetaiotaomicron*), TVAG339720, TVAG199300 and TVAG189150 (from *T. vaginalis*) (Figure III.1) were PCR amplified from the genomic DNAs of the respective organisms using oligonucleotide primers listed in Appendix Table A.1. Amplified DNA containing engineered restriction sites were then cloned into expression plasmids (Table III.1) for later expression in an *E. coli* BL21 (DE3) expression host.

Gene	**Target domain**	**Region cloned from DNA**	**Cloning Vector**	**Cloning sites in vector**	**Code name for Recombinant Protein**	**Theoretical Weight (kDa)**
BT_4272	M60-like	1120-2991	pET-28a(+)	*BamHI-XhoI*	BT4272-M60L	74.6
BT_4244	M60-like	820-2574	*mini*PRSET A	*BamHI-EcoRI*	BT4244-M60L	69.2
BT_3015	M60-like	1105-2901	pET-28a(+)	*BamHI-EcoRI*	BT3015-M60L	71.1
BT_4244	Full length	70-2574	*mini*PRSET A	*BamHI-EcoRI*	BT4244-FL	97.0
BT_4272	CBM32	694-1110	pET-28a(+)	*BamHI-XhoI*	BT4272-CBM32	19.1
BT_4244	BACON	70-420	*mini*PRSET A	*BamHI-EcoRI*	BT4244-BACON	14.8
BT_**4244**	CBM32	394-819	pET-28a(+)	*NcoI-XhoI*	BT4244-CBM32	17.1
BT_4244	BACON-CBM32	70-819	*mini*PRSET A	*BamHI-EcoRI*	BT4244-BC	29.7
BT_3015	CBM32	682-1260	*mini*PRSET A	*BamHI-EcoRI*	BT3015-CBM32	23.6
TVAG_339720	M60-like	61 – 1842	pET-43.1a(+)	*BamHI/XhoI*	TVAG339720-M60L	129.1
TVAG_189150	M60-like	61 -1761	pET-43.1a(+)	*BamHI/SalI*	TVAG189150-M60L	126.4
TVAG_199300	M60-like	67-1605	pET-43.1a(+)	*SacI/SalI*	TVAG199300-M60L	120.1
TVAG_339720	PA14	1711-2442	pET-28a(+)	*BamHI/XhoI*	TVAG339720-PA14	32.1
TVAG_339720	GBDL	2407-3327	pET-43.1a(+)	*BamHI/XhoI*	TVAG339720-GBDL	95.3
TVAG_199300	PA14-GBDL	1633-3327	pET-43.1a(+)	*SacI/XhoI*	TVAG199300-CBD	124.9

Table III.1 - Cloning strategy and details of various gene/gene fragments analysed in this study.

Theoretical molecular weights and extinction coifficients (Appendix B) of proteins were estimated using the Expasy Protparam tool at http://web.expasy.org/protparam/

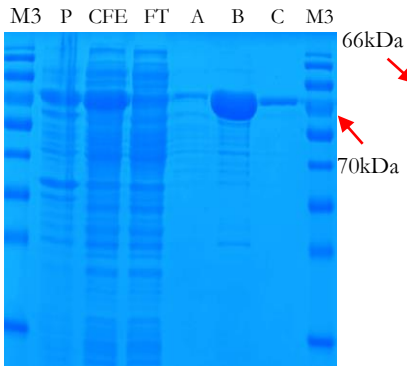
The full length BT_4244 gene excluding the region encoding the type II signal peptide sequence was also cloned in a similar expression plasmid (*mini*PRSET A) as the M60-like domain encoding region of same gene for comparative studies. The cloning strategy was designed to allow the introduction of an N- or C-terminal poly-histidine tag (His_6) in each recombinant protein following expression. A single point mutation was also introduced in the zincin motif of BT4244-FL yielding the mutant protein BT4244-FL–E575D containing an aspartic acid residue (D) in place of the putative active site glutamic acid (E) residue in the gluzincin motif of the protein. The glutamic acid residue is the catalytically active amino acid in the gluzincin motif typical of zinc metalloproteases (Section I.8.1.1.2) and hence this mutation is likely to inhibit the activity of the enzyme if indeed it belongs to gluzincin family of proteases. Primers used for this mutation are also included in Table A.1, Appendix A.

An *E. coli* BL21 (DE3) expression host was routinely used for the expression of cloned genes. Protein expression in all cases was induced with 1 mM IPTG overnight at 16 ^{o}C, except for BT4244-FL whose expression was often induced at 37 ^{o}C for 4 h after reaching the mid log growth phase. This strategy helped to reduce excessive production of truncated products or background often observed following purification of the protein. Purification of recombinant N-/C-terminal His-tagged proteins was carried out using immobilized metal affinity chromatography (IMAC) (Section II.1.24.2), Ion-exchange and gel filtration chromatography techniques (II.1.24.3). The concentrations of purified proteins were measured by absorbance (A280nm) using the estimated molecular weight and extinction coefficient of each purified recombinant protein (Appendix B). Please see Figure III.2 for all protein expression data

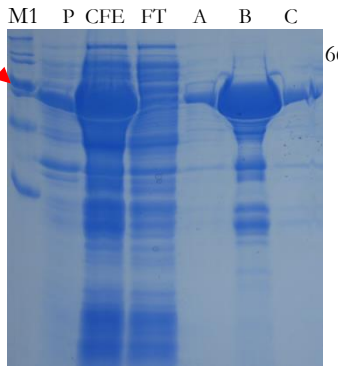
Recombinant soluble proteins were successfully expressed for all *B. thetaiotaomcrion* M60-like entries as opposed to *T. vaginalis* were significant amounts of the expressed proteins formed inclusion bodies. The pET-43.1a (+) vector (Figure C.2, Appendix C) was used during the cloning of *T. vaginalis* proteins due to this problem but this only partially improved the solubility of the expressed proteins. This strategy introduces a NUS tag (with an additional molecular weight of ~ 60 kDa at the N-terminal of the protein (Figure C.2, Appendix C). While increasing the culture volume to 1 or 2 L significantly increased the amount of the solubly expressed TVAG199300-M60L, this didn't make any difference in the case of TVAG189150-M60L and TVAG339720-M60L proteins and hence only TVAG199300-M60L could be reliably used for enzymatic assays.

A

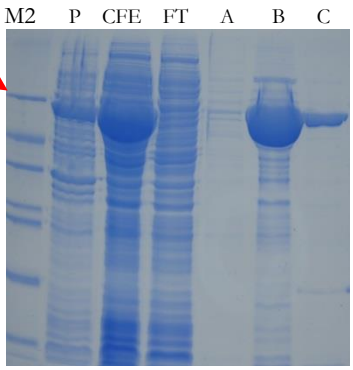
------- **BT4272-M60L** ----------



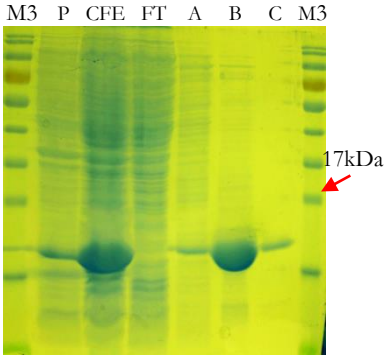
----- **BT4244-M60L** --------



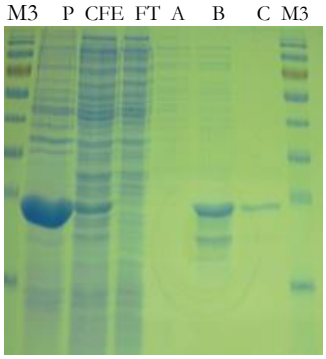
------- **BT3015-M60L** --------



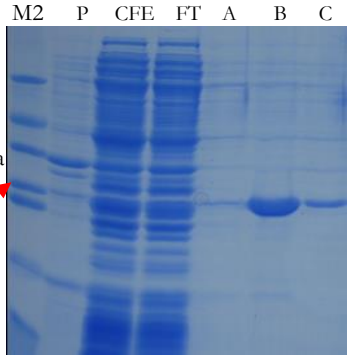
------ **BT4244-CBM32**-----



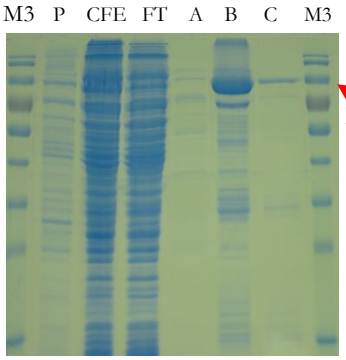
----- **BT3015-CBM32 - -----**



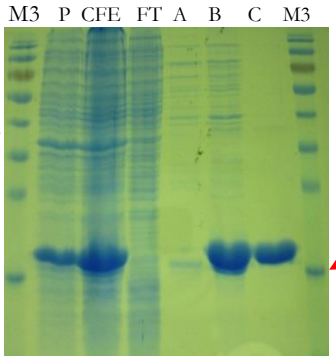
--- **BT4244-BACON-CBM32--**



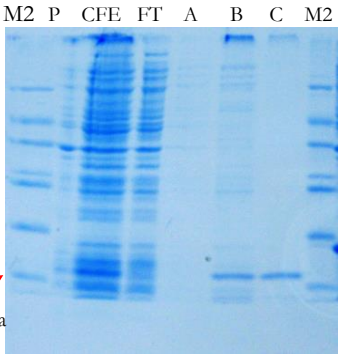
------- **BT4244-FL** ----------



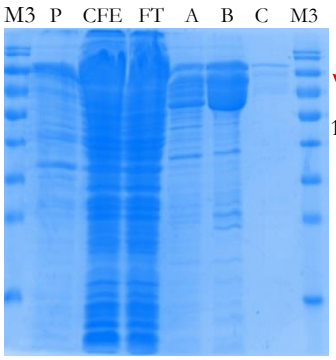
----- **BT4272-CBM32**--------



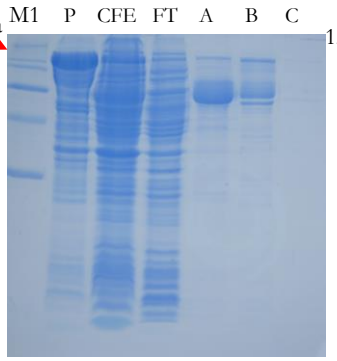
------- **BT4244-BACON** ---



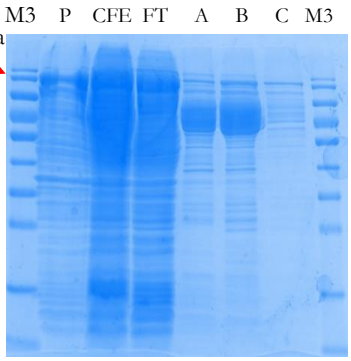
---- **TVAG199300-PROT**-----



----- **TVAG339720-PROT** --------



------- **TVAG189150-PROT** ------



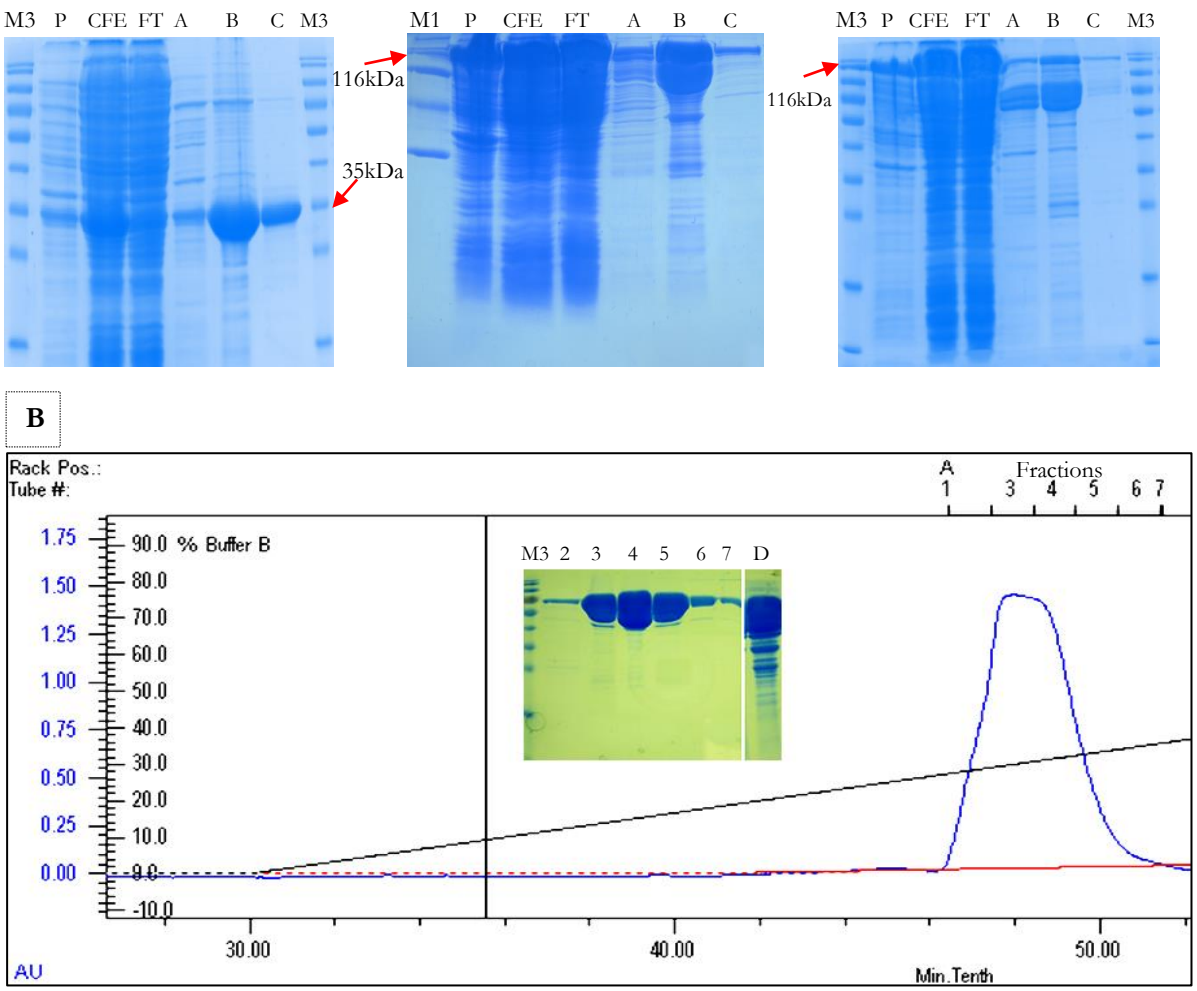


Figure III.2 - Expression and purification of protein domains from selected *B. thetaiotaomicron* and *T. vaginalis* M60-like/PF13402 proteins. A: Purification of recombinant proteins by immobilized metal affinity chromatography (IMAC). Protein expression and purification was carried out as described in Section II.1.26. All *B. thetaiotaomicron* proteins alongside TVAG339720-PA14 were produced from 100 ml of *E.coli* BL21 (DE3) culture while the rest of *T. vaginalis* proteins were produced from at least a litre of culture. 5 µl of the insoluble pellet fraction (P) re-suspended in 10ml of Talon buffer (20 mM Tris/HCl pH 8.0 containing 100 mM NaCl), 10 µl of cell free extract (CFE), 10 µl of flow through (FT), 10 µl of the fraction eluted with 10 mM imidazole (A), 10 µl of fractions sequentially eluted with 100 mM imidazole (B and C) were analysed by SDS PAGE (12.5%) in all cases. The results generally showed high expression of *B. thetaiotaomicron* M60-like proteins compared to *T. vaginalis*. Red arrows point to the positions of bands with the indicated molecular weights. Also see Section II.1.27 for the molecular weights of various markers in M1, M2 and M3. **B**: Example of an anion exchange chromatogram for the purification of BT4244-M60L. Following purification of proteins from about 6 L of culture by IMAC, fractions corresponding to B and C were pooled and concentrated by centrifugation into a 4 ml volume (~25 mg/ml protein). An initial anion exchange chromatography (AEC) run using 500 mM NaCl in 10 mM Tris-HCl, pH 8.0 was performed on the sample and impure fractions pooled again (Lane D on embedded SDS PAGE image). Sample D

was then used for a second AEC run yielding samples of lanes 2, 3, 4, 5, 6, 7 corresponding to fractions indicated on chromatogram.

III.3.3 Mucinase assays

Mucinase assays were performed to investigate the potential mucinase activity of expressed recombinant M60-like proteins. Substrates used included mucins from bovine submaxillary glands Type I-S (BSM) (Sigma, UK), and porcine stomach [Type III or PGMIII (bound sialic acid 0.5-1.5 %) and Type II or PGMII (bound sialic acid, ~1%)]. Following the treatment of various mucins with recombinantly expressed M60–like proteins, samples collected at different times post incubation were analysed by a combination of SDS - agarose gel electrophoresis (SAGE), Western blotting and Wheat germ agglutinin (WGA) or Periodic Acid-Schiff (PAS) detection techniques (Sections II.1.36 and II.1.37). WGA is a lectin from the wheat; *Triticum vulgaris* capable of binding N- acetylglucosamine (GlcNAc) present in mucins. PAS detection on the other hand allows for the staining of vicinal diol groups on peripheral sugar moieties in glycoproteins.

The clearance of WGA and PAS reactive bands relative to controls without enzyme was interpreted as evidence of mucinase activity. This is assuming that mucin cleavage results in low molecular weight fragments that migrate further down each lane compared to the controls as observed in a similar study involving the TagA secreted mucin protease of *Vibrio cholera* (Szabady *et al.*, 2011). By analyses, there was evidence of mucinase activity from all M60-like domains recombinantly expressed from *B. thetaiotaomicron* including the recombinant full length BT4244 protein (BT4244-FL) which degraded both BSM and PGMIII in a time dependent manner (Figure III.3). PGMIII degradation in general however seemed to be less dramatic compared to BSM following treatment with the same proteins over the same periods of time.

The mucinase activity of the recombinant BT4244-FL protein was also shown to be inhibited in the presence of a metal chelator such as Ethylenediaminetetraacetic acid (EDTA) and in the mutant version of the full length BT4244 protein (BT4244-FL-E575). The TVAG199300-M60L from *T. vaginalis* on the other hand, which was the only reliably expressed soluble M60-like protein of the three selected M60-like sequences from *T. vaginalis* failed to degrade BSM and PGMIII.

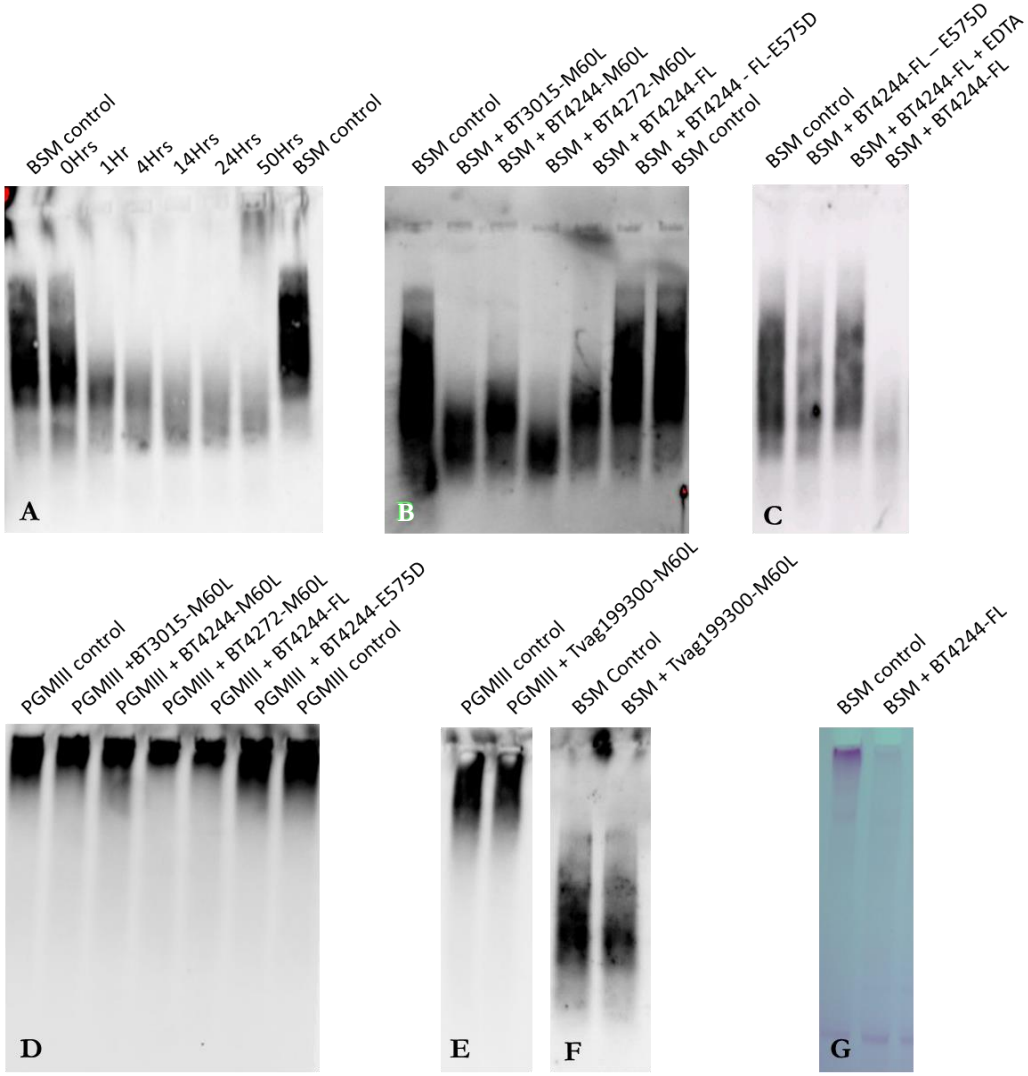


Figure III.3 – Mucinase activity of recombinantly expressed *B. thetaiotaomicron and T. vaginalis* M60-like/PF13402 proteins. **A:** Time course degradation of bovine submaxillary mucin by recombinant BT4244-FL. 0.05% of BSM was incubated with 1 µM of the "full length" recombinant BT4244-FL protein in a final volume of 500 µl in Talon buffer (20 mM Tris/HCl pH 8.0 containing 100 mM NaCl) over different time periods. Samples were electrophoresed post incubation by SDS-agarose gel electrophoresis, transferred to PVDF membranes and probed with wheat germ agglutinin (WGA) (Section II.1.36). Clearance of WGA reactive bands/smears was seen as evidence of mucin degradation by BT4244-FL**.** **B:** Mucinase activity of recombinant M60-like domains from *B. thetaiotaomicron* proteins. 0.05% of BSM was treated with 0.01 mg/ml (~0.1 µM of BT4244-FL) of various expressed recombinant proteins including the BT4244-FL-E575 mutant for 24 h before analyses as above. There was evidence of mucinase activity for all recombinant proteins tested except for the BT4244-FL-E575 mutant. **C:** Evidence for inhibition of BT4244 –FL mucinase activity by EDTA (a metal chelator). 0.05% of BSM samples had been treated with 1 µM of enzyme for over 48 h in this case. **D:** Same as B except that substrate used in this case was 0.08% PGMIII. **E**: Same as D except that the protein evaluated for mucinase activity in this case was the recombinantly expressed M60-like domain of the *T. vaginalis* TVAG199300 protein (TVAG199300-M60L) **F**: Same as in B except that the protein evaluated for mucinase activity in this case was the recombinant TVAG199300-M60L

protein. **G**: Evaluation of BT4244-FL mucinase activity by PAS staining. In this case, 0.05% of BSM samples were treated with 1 µM of enzyme for about 60 h before analyses on 4-20% gradient gels purchased from Biorad.

III.3.4 IgA1 protease activity of BT4244

Secretory IgA, like mucins are prominent mucosal surface glycoproteins also encountered by mucosal microbes. There exists two main isotypes of the IgA molecule, namely IgA1 and IgA2 (Section I.2.1.2). IgA1 is more similar to mucins compared to IgA2, as it contains a mucin like hinge insertion sequence (Figure III.4) which also happens to be the site of mucin-type O-glycosylations (Section I.2.1.2, Figure I.9). As glycoproteins, both IgA1 and 2 represent potential substrates for *B. thetaiotaomicron and T. vaginalis* m60-like proteins.

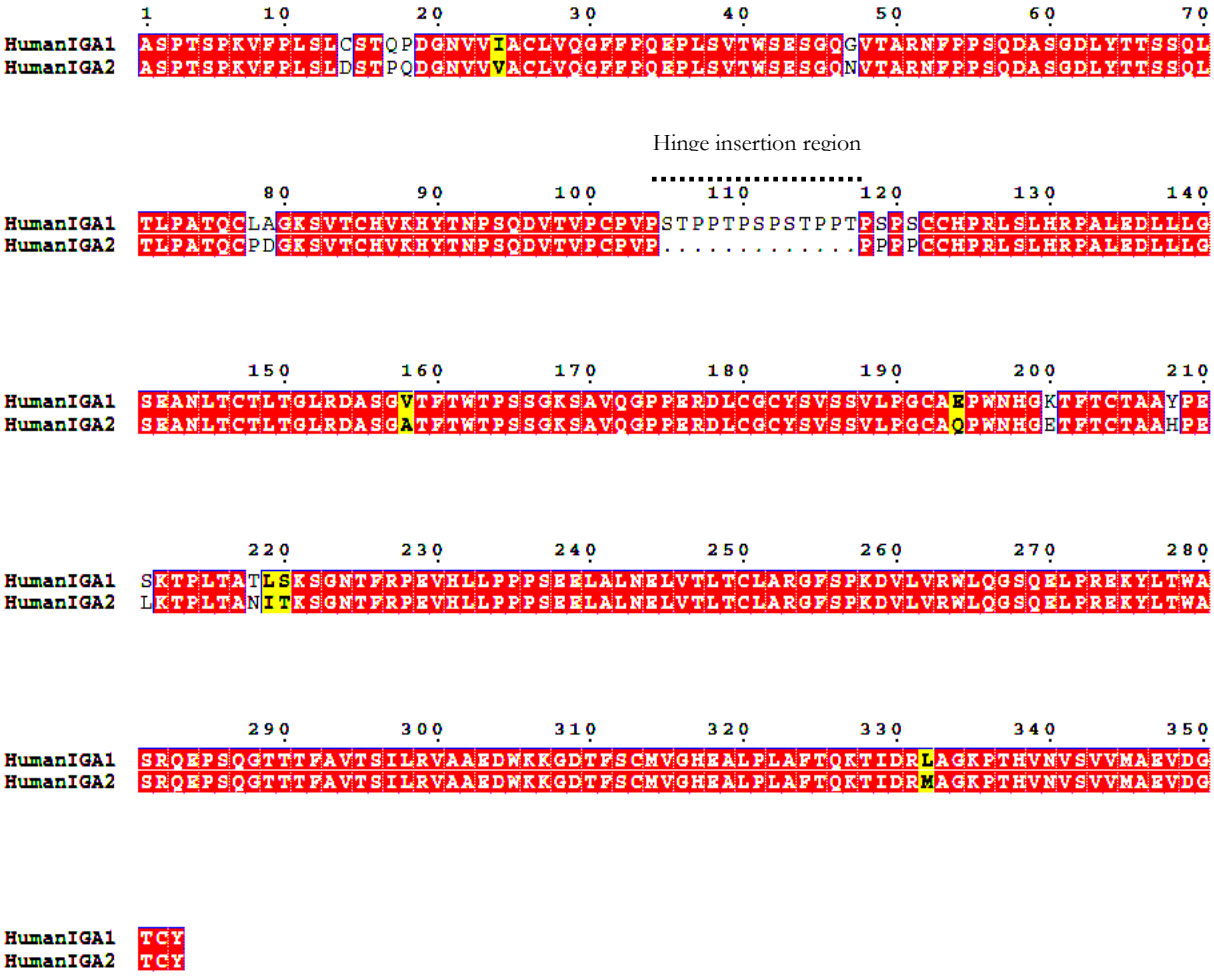


Figure III.4 - Alignment of human IgA1 and IgA2 alpha (α) chain C regions to show the mucin-like hinge insertion sequence of IgA1. The hinge region is indicated by dotted lines and contains mucin-like PTS repeats representing potential O-glycosylation sites (Section I.2.1.1). [Human IGA1 Uniprot ID = P01876 (IGHA1_HUMAN), Human IGA2 Uniprot ID = P01877 (IGHA2_HUMAN]. Sequence alignments were viewed using the ESPript 2.2 utility at http://espript.ibcp.fr/ESPript/ESPript/ with a global similarity score threshold of

0.7. Red highlights are for amino acid residues showing 100% conservation while yellow highlights are for residues showing less than 100% conservation but above the global score threshold (Gouet *et al.*, 1999).

Following incubation of both IgA isoforms (human myeloma IgA1 and IgA2) with the recombinant BT4244-FL protein, the degradation of IgA1 but not IgA2 was observed. Cleavage of IgA1 by the enzyme yielded two fairly distinct low molecular weight bands, both of which were very close to the 35 kDa protein marker as shown in the SDS PAGE image of Figure III.5A. Interestingly, even after incubation with excess amounts of the enzyme (up to 2.8 μM of enzyme) for extensive periods of time (up to 48 h), complete degradation of IgA1 was not observed. This was later found to be the consequence of IgA1 sialylation as prior treatment of IgA1 with a neuraminidase enzyme from *Clostridium perfringens (C. welchii)* (Sigma cat. N2876) led to an increase in the amount of released degradation fragments (Figure III.5C). The neuraminidase enzyme is capable of cleaving terminal sialyl-α(2→3), α(2→6), and α(2→8)-linkages some of which are present in mucin and IgA glycans shown in Figure III.5B.

III.3.5 Cleavage site of BT4244 on IgA1

To determine the specific site of cleavage on IgA1 by the recombinant BT4244-FL enzyme, cleavage fragments following treatment with BT4244-FL were separated by on 12.5% SDS PAGE gels under reducing conditions and stained with Coomassie blue. The two bands indicated by the red and black arrows in Figure III.5A, were excised from the gel and subjected to N-terminal Edman sequencing. The sequencing data revealed that cleavage occurred at possibly two identical sites between the CH1 and CH2 domains of the IgA1 heavy chain between residues Pro_{-223} and Ser_{-224} or Pro_{-231} and Ser_{-232} in the glycosylated hinge region (Figures III.5 A and B). The raw N-terminal sequencing data from Alphalyse A/S (Denmark) is available in Appendix F. Based on these findings, the upper fragment indicated by the red arrow in Figure III.5 corresponds to the Fc-α fragment of IgA1 while the lower fragment indicated by the black arrow corresponds to Fab-α fragment of IgA1 (Section I.2.1.2). N-terminal sequencing data was only obtained for the Fc-α fragment (red arrow in Figure III.5A) and not the Fab-α fragment implying the N-terminal of the latter maybe modified.

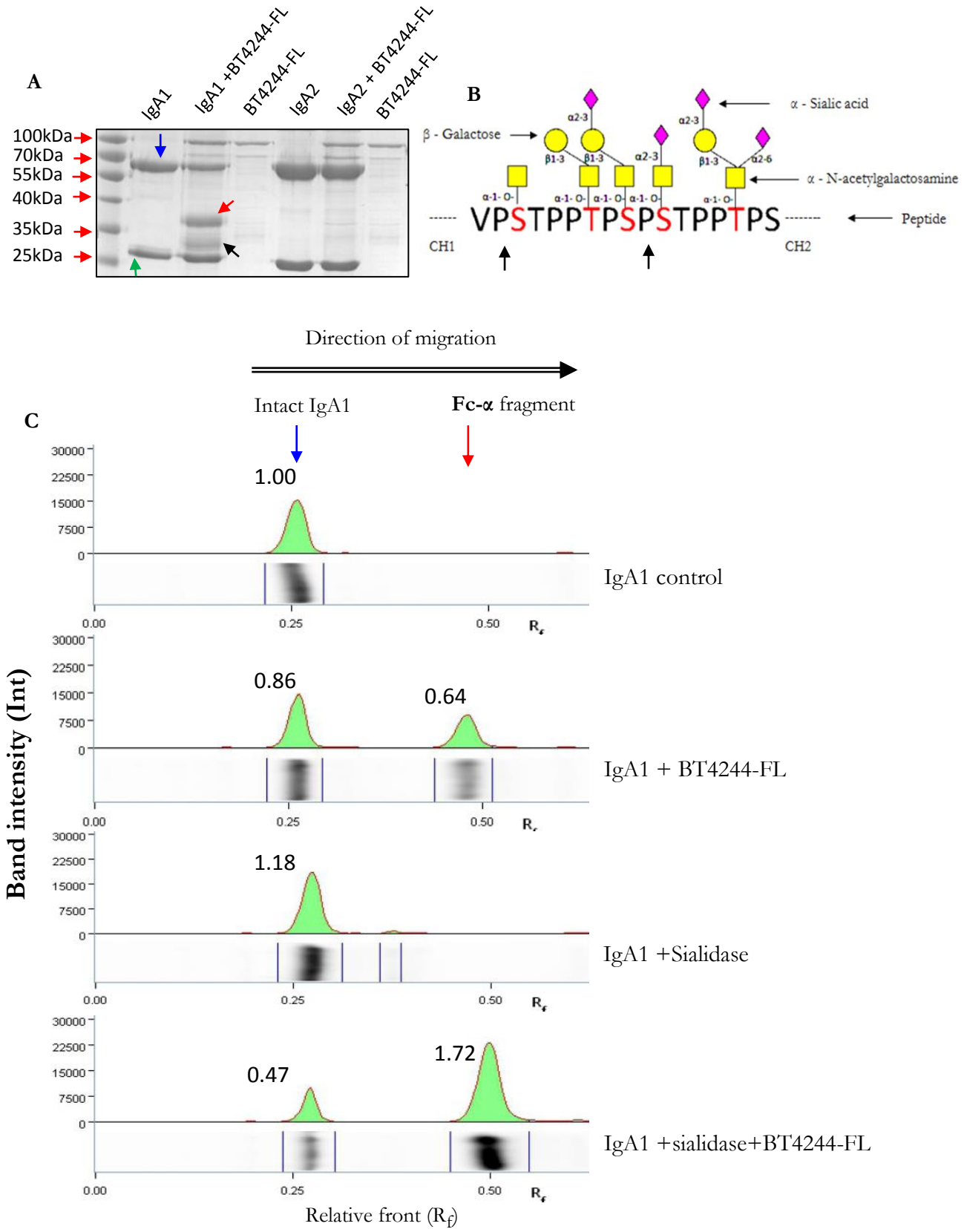


Figure III.5 - Cleavage of human myeloma IgA1 by recombinant BT4244-FL. **A**: BT4244-FL degrades IgA1 but not IgA2. 0.33 mg/ml of human myeloma IgA1 and IgA2 samples were treated with 0.8 µM of the recombinant BT4244-FL protein in Talon buffer (20 mM Tris/HCl pH 8.0 containing 100 mM NaCl) at 37 °C for 48 h. Samples were electrophoresed on 12.5% SDS PAGE gels before staining with Coomassie blue reagent.

Digested products corresponding to Fc-α and Fab-α fragments used for N-terminal Edman sequencing are indicated by red and black arrows respectively. The blue and green arrow point to the intact IgA1 heavy and light chains respectively (Figure I.9) before protease treatment. **B**: Hinge insertion sequence of IgA1 showing glycan structures and BT4244-FL cleavage sites (black arrows) determined by N-terminal Edman sequencing (Section II.1.39). **C:** Effect of IgA1 desialylation on BT4244-FL IgA protease activity. 0.25 mg/ml of IgA1 was initially treated for approximately 15 h with a 0.63 U/ml of a *C. perfringens* neuraminidase enzyme at 37 °C. 0.625 μM of recombinant BT4244-FL was later incubated with the sample at 37 °C for another ~15 h. Samples were then electrophoresed on 12.5% SDS PAGE gels followed by transfer to PVDF membranes. Chemiluminescence detection of degradation products was carried out using a 1:2000 dilution of primary mouse Anti-Human IgA1 heavy chain antibodies conjugated to biotin followed by a 1: 2000 dilution of ExtrAvidin®−Peroxidase (Section II.1.38). Band intensities were quantified using the Image Lab™ tool (Biorad) after 5 seconds of exposure to chemiluminescent substrate. Red arrows point to the Fc-α product while the blue arrow points to the position of the intact human myeloma IgA1 glycoprotein.

III.3.6 Universal Protease Activity Assay with BT4244-FL

To determine whether BT4244-FL exhibits general or non-specific proteolytic activity, universal protease assays were performed using the Sigma Universal Protease Activity Assay kit (Sigma) according to the manufacturers' instructions. The general protease substrate used was casein and the principle is based on the fact that casein degradation by a non-specific protease leads to tyrosine release which can be quantified spectrophotometrically ($A600_{nm}$) after reaction with Folin & Ciocalteu's Phenol Reagent (Section II.1.40). The amount of tyrosine released following treatment of casein with the recombinant BT4244–FL was not only very low compared to general protease enzyme controls but also similar to the amount released by the inactive BT4244-FL-E575 mutant and BSA controls (Figure III.6). The data thus suggest BT4244-FL does not exhibit general protease activity.

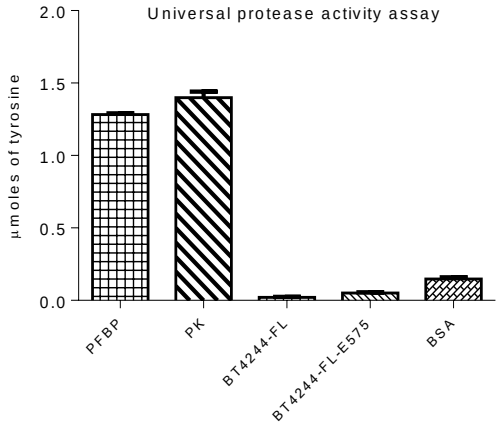


Figure III.6 - Universal protease activity assay to test the specificity of recombinant BT4244-FL. 0.54% of casein substrate was incubated with ~0.2 μM of BT4244-FL and its mutant BT4244-FL-E575D (~0.02 mg/ml for each) alongside positive controls [~0.02 mg/ml of proteinase K (PK, 0.05U/ml) and protease from

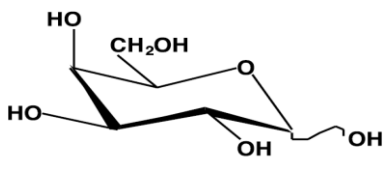
bovine pancreas (PFBP, 0.5mU/ml)] and negative controls (0.02 mg/ml BSA) at 37 °C for 24 h. Protease activity was estimated by measuring the amount of released tyrosine using a tyrosine standard curve (Section II.1.40). 0.2 μM of BT4244-FL is an amount double that which had earlier been shown to be capable of significantly degrading BSM over same 24 h (Figure III.3B). Tyrosine release by BT4244-FL was very low and comparable to the amount released by the inactive mutant (BT4244-FL-E575) and the negative control (BSA).

III.3.7 Carbohydrate binding modules of *B. thetaiotaomicron* and *T. vaginalis* M60-like proteins

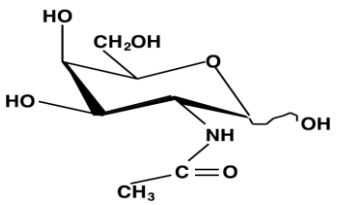
Isothermal titration calorimetry (ITC, Section II.1.32) was used to study the carbohydrate binding properties of various recombinantly expressed putative carbohydrate binding modules identified in *B. thetaiotaomicron* and *T. vaginalis* M60-like proteins. As mucins and IgA had been hypothesized, and at least demonstrated to be targets for recombinant M60-like proteins from *B. thetaiotaomicron*, their glycans were given priority when screening for potential ligands recognized by various carbohydrate modules. Mucin sugars amongst other sugars tested ranged from monosaccharide through disaccharide to complex glycan structures within the core, backbone and peripheral units of mucin type - O glycans (Section I.2.1.1.1).

III.3.8 BACON and CBM32 domains of *B. thetaiotaomicron* M60-like proteins

ITC studies on recombinant CBM32 domains from *B. thetaiotaomicron* M60-like proteins revealed affinity for galacto-configured sugars with higher preference for the monosaccharide sugar; N-acetylgalactosamine (GalNAc) over more complex glycans containing the same sugar (Figure III.8 and Table III.2 and III.3). Binding in all cases was however in the low affinity range (Ka $<10^3$ M^{-1}) with significant saturation of the recombinant proteins only observed at ligand concentrations as high as 20 – 40 mM (Figure III.8, Table III.3). Recombinant BACON domains on the other hand failed to bind any of the mucin sugars tested (Table III.2). The binding affinity towards GalNAc and other sugars tested was not affected when the recombinant BT4244-BC protein containing both the BACON and CBM32 domains of BT4244 was used (Table III.2)



Galactose (Gal)



N-acetylgalactosamine (GalNAc)

Figure III.7 – Structure of Galactose (Gal) and N-acetylgalactosamine (GalNAc). GalNAc binds weakly to recombinant BT4244-CBM32, BT3015-CBM32 and BT4272-CBM32 proteins (Figure III.8)

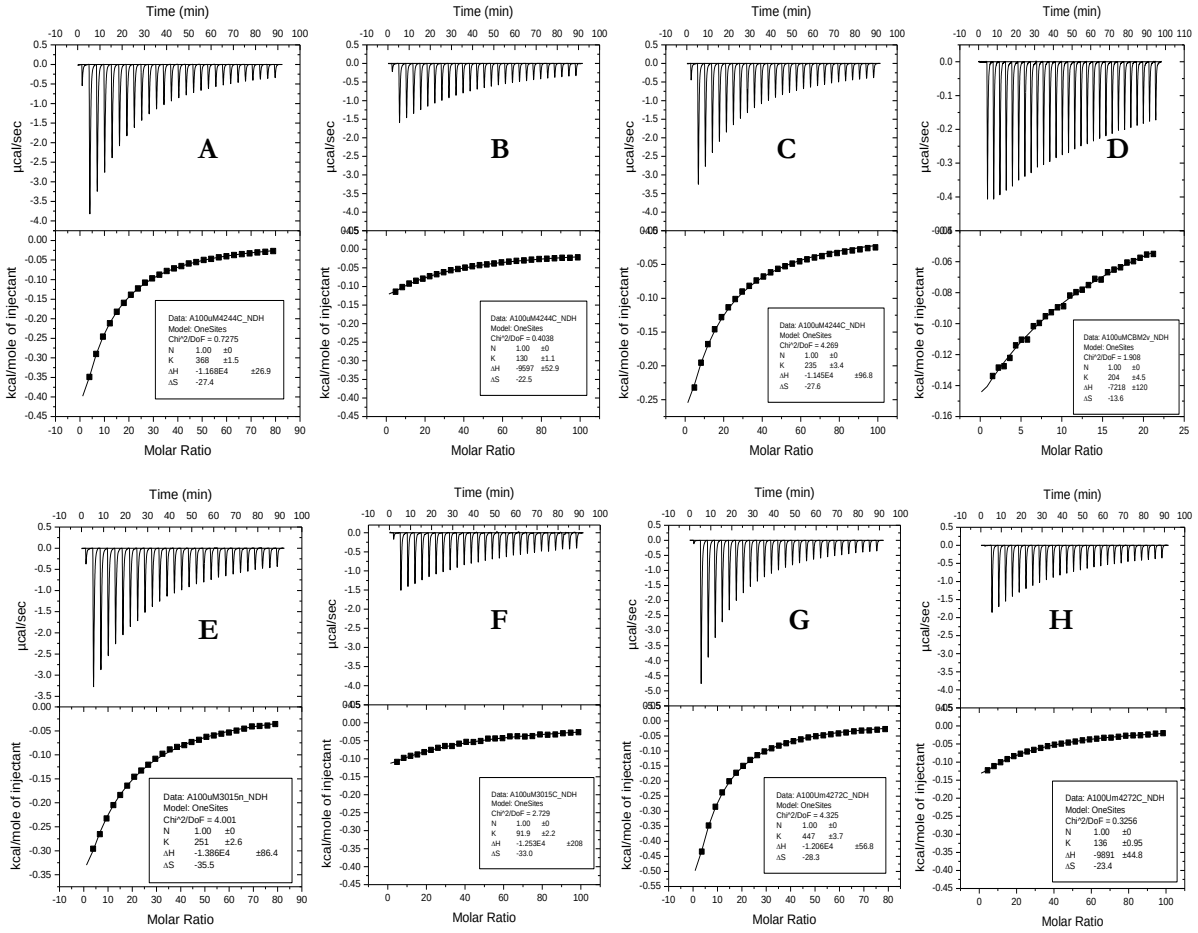


Figure III.8 - Analyses of carbohydrate binding by recombinant CBM32 domains of *B. thetaiotaomicron* M60-like proteins. Graphs at the top are representative ITC binding data of various CBM32 domains to mucin/non-mucin sugars. For these experiments, 100 µM of the purified recombinant protein was dialysed overnight into 20 mM Tris pH 8.0. Samples were filtered the next day alongside ligands that had been prepared in filtered dialyses buffer. The sugar in the syringe was titrated (27 injections) into the cell containing the protein sample and thermodynamic data analysed using Origin version 7.0 tool. The top half of each panel shows the raw power data while the bottom half are integrated peak areas fitted to a single-site binding model and stoichiometry fixed at 1 (n≈1). **A**: 100 µM BT4244-CBM32 vs 40mM GalNAc, **B:** 100 µM BT4244-CBM32 vs 50mM Gal, **C:** 100 µM BT4244-CBM32 vs 50mM Lac. **D**: BT4244-CBM32 vs 10mM Galβ1-3GalNAc (core-1) **E:** 100 µM BT3015 vs 40mM GalNAc **F:** 100 µM BT3015 vs 50mM Gal, **G:** 100 µM BT4272 vs 40mM GalNAc, **H:** 100 µM BT4272 vs 50mM Gal.

Ligand	BT4244-BC	BT4244-CBM32	BT4244-BACON	BT4272-CBM32	BT3015-CBM32
Fuc (10mM)	-				
Gal (50mM)	±	±	-	±	±
GalNAc (20mM, 40mM)	+	+	-	+	+
GlcNAc (20mM)	-				
NeuAc (10mM)	-				
Man (20mM)	-				
Xyl (40mM)	-				
Gal-6S (2.5mM)	-				
Fucα1-2Gal (2mM)	-				
GalNAcα1-Ser (2.5mM)	±				
Galβ1-3GalNAc (10mM)	±				
Galα1-4Glc (20mM, 50mM)	+	+	-	+	±
Galβ1-4GlcNAc (10mM)	±				
Galβ1-3GlcNAc (10mM)	±				
Fucα1-2Galβ1-4GlcNAc (2.5mM)	-				
BSM (0.1%)	-				
PGMIII (0.1%)	-				

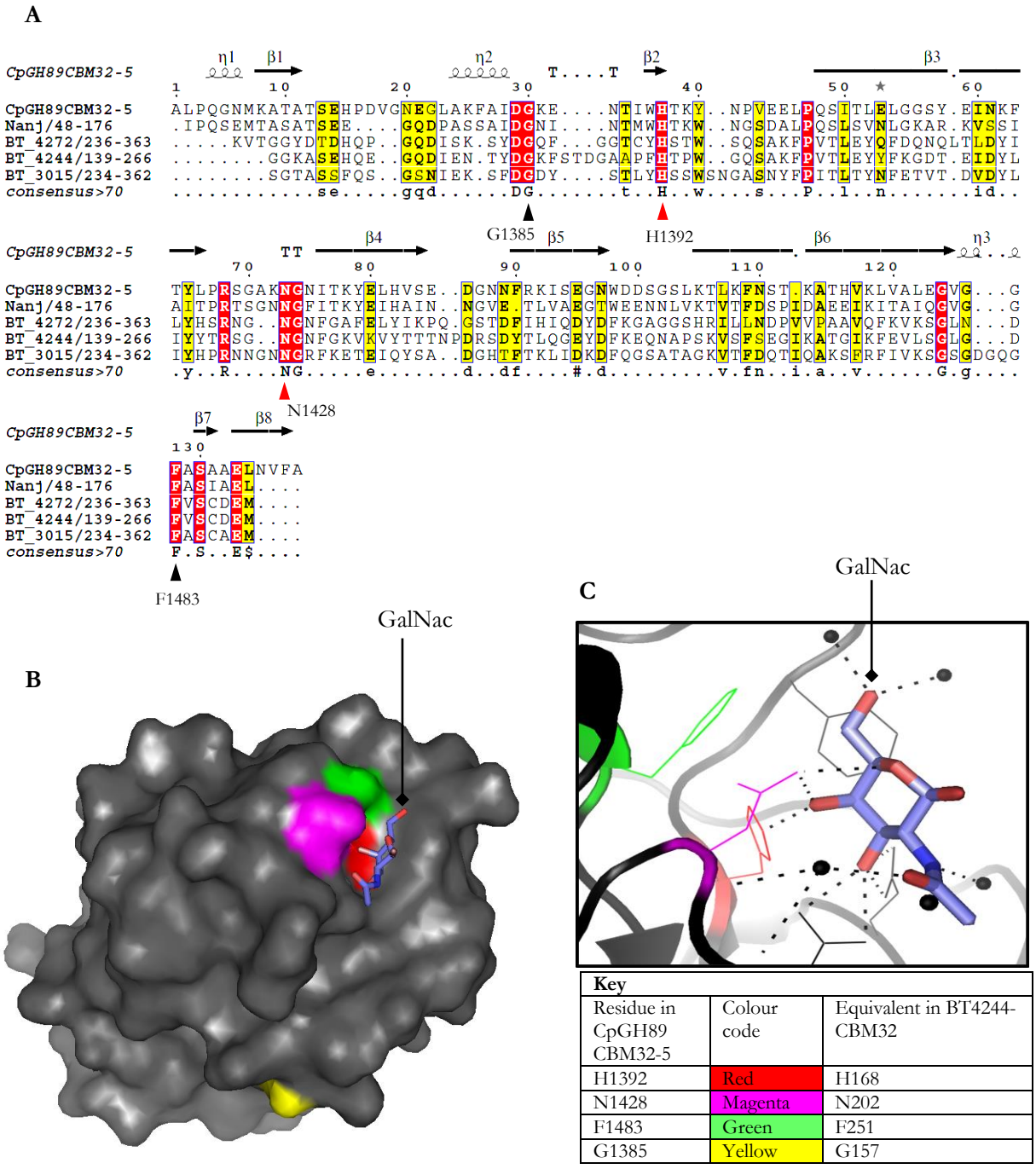
Table III.2 - Summary of qualitative ITC binding data for various CBMs from *B. thetaiotaomicron* M60-like proteins against simple and complex mucin sugars. The plus sign (+) implies binding was observed and curve fitting could be performed while the ± indicates saturation of the CBM by the ligand was observed but too weak to be reliably fitted at the concentration of the substrate indicated. The minus (-) sign implies no binding was observed.

Recombinant protein	$K_a \times 10^3$ (M^{-1})	ΔG (kcal mol^{-1})	ΔH (kcal mol^{-1})	$T\Delta S$ (kcal mol^{-1})	n
BT4244-CBM32	0.36 ± 1.5	-3.47	-11.7 ±26.9	-8.23	1
BT3015-CBM32	0.25 ± 2.6	-3.26	-13.9 ±86.4	-10.64	1
BT4272-CBM32	0.45 ± 3.7	-3.60	-12.1 ±56.8	-8.50	1

Table III.3 - Affinity and thermodynamic parameters of N-acetylgalactosamine binding to recombinant CBM32 domains form various *B. thetaiotaomicron* M60-like proteins. Thermodynamic parameters were calculated as described in Section II.1.32 using the standard thermodynamic equation $-RT\ln K_a = \Delta G = \Delta H - T\Delta S$, where R = gas constant (1.99 $cal.K^{-1}.mol^{-1}$), T = temperature in Kelvin (298.15 K), ΔG = change in free enthalpy, ΔS = entropy of binding. n = stoichiometry of binding.

III.3.9 GalNAc recognition by CBM32 domains of *B. thetaiotaomicron* M60-like proteins

All GalNAc binding CBM32 proteins in Table III.3 show sequence similarity (at least 50%) to each other by alignment (Figure III.9), suggesting they might utilise a conserved GalNAc recognition mechanism. An attempt to crystallize one of them (the recombinant BT4244-CBM32) for structural insights into the mechanism through X-ray crystallography, however failed. In any case, amino acid sequence alignments revealed that this group of proteins share important sequence and seemingly structural features with the well- characterised GalNAc–binding CBM32 domains of CpGH89, exo-α-D-N-acetylglucosaminidase (CpGH89CBM32-5, PDB: 4AAX) and NanJ sialidase enzymes (NanJCBM32, PDB: 2v72) from the mucosal microbe; *Clostridium perfringens* (Boraston *et al*., 2007, Ficko-Blean *et al*., 2012). Strikingly, aromatic and non-aromatic amino acid residues including the CpGH89CBM32-5 Histidine-1392 (H1392), Arginine-1423 (R1423), Asparagine–1428 (N1428), Phenylalanine-1483 (F1483), some of whose side chains directly interact through intermolecular hydrogen bonds with GalNAc, were strictly conserved in all the CBM32 domains analysed (Figure III.9A).



Key		
Residue in CpGH89 CBM32-5	Colour code	Equivalent in BT4244-CBM32
H1392	Red	H168
N1428	Magenta	N202
F1483	Green	F251
G1385	Yellow	G157

Figure III.9 - Comparing BT4244, BT3015 and BT4272 CBM32 sequences with the GalNAc binding CpGH89CBM32-5 (PDB: 4AAX) and NanJCBM32 (NanJCBM32: PDB 2v72) CBM32 sequences of *Clostridium perfringens*. A: Structure based-sequence alignment. The secondary structure of CpGH89CBM32-5 (PDB: 4AAX, Ficko-blean *et al.*, 2012) is indicated above the alignments. The position of four highly conserved residues (G1385, H1392, N1428 and F1483) mutated in the current study are indicated by black and red triangles. Three of these residues (H1392, N1428 and F1483) are located in the binding site of the molecule and two of them (H1392, N1428) indicated by the red triangles make direct intermolecular hydrogen bond interactions with GalNAc in the binding site. **B**: Surface representation of the structure of CpGH89CBM32-5 showing the positions of the four highly conserved residues discussed in A. Please see key below panel C for guide with colour coding. The equivalents of these conserved residues in BT4244-CBM32 are also provided in the key. **C:** Stick representation of the binding site of CpGH89CBM32-5 showing

interactions of various binding site residues including H1392, N1428 with GalNAc. Sequence alignments were viewed using the ESPript 2.2 utility at http://espript.ibcp.fr/ESPript/ESPript/ (Gouet *et al.*, 1999) with a global similarity score threshold of 0.7. Red highlights are for amino acid residues showing 100% conservation while yellow highlights are for residues showing less than 100% conservation but above the global score threshold.

To determine if these similarities were an indication that GalNAc recognition by *B. thetaiotaomicron* CBM32 domains is similar to what was observed for the *C. perfringens* CpGH89CBM32-5 and NanJCBM32 CBM32 domains, mutational studies involving the four conserved residues identified in BT4244-CBM32 (key in Figure III.9) were carried out. Each of the residues in the wild type BT4244-CBM32 protein was substituted with alanine by site-directed mutagenesis giving four separate mutants namely G157A, H168A, N202A and F251A (see Quickchange™ mutagenesis primers in Appendix Table A.2). The results of ITC experiments involving the four mutants and GalNAc are shown below in Figure III.10.

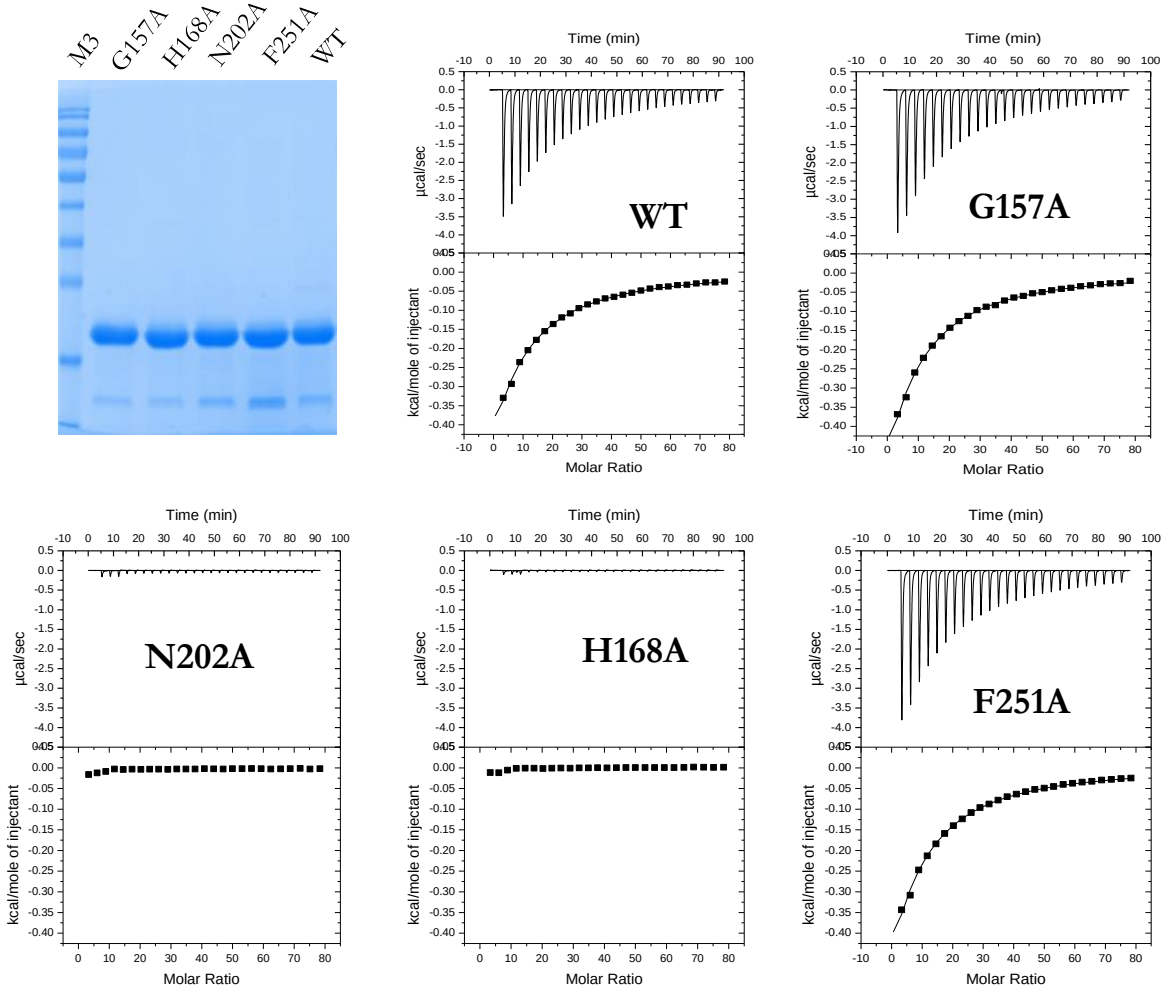


Figure III.10 – Impact of selected mutations on BT4244-CBM32 binding to GalNAc.

The set-up of the experiment was as described in Figure III.8, i.e. titrating ~40mM GalNAc sugar (~27 injections) into 100 µM of the recombinant protein at 25 °C. Wild type and mutant proteins where all expressed under the same conditions and buffer

exchanged into Talon buffer (20 mM Tris/HCl pH 8.0 containing 100 mM NaCl) . G157A and F251A mutants had similar binding affinities for GalNAc compared to the wild type (WT) (Ka ~ 0.4 x 103 M^{-1}) while N202A and H168A mutants completely failed to bind GalNAc.

ITC binding data for the various mutants indicated that N202 and H168 residues likely interact with GalNAc in a manner similar to their equivalents in CpGH89CBM32-5, while G157A and F251A like their equivalents in the same protein may not directly interact with GalNAc. Predicted three dimensional structures for the various CBM32 domains based on their alignment with the *C. perfringens* CpGH89CBM32-5 CBM32 domain are shown in Figure III.11. All three structures display typical CBM Type C features (Boraston *et al.*, 2004) i.e. containing a β-sandwich fold and a small sugar binding pocket (GalNAc binding pocket) in the loops at the edge of the sandwich.

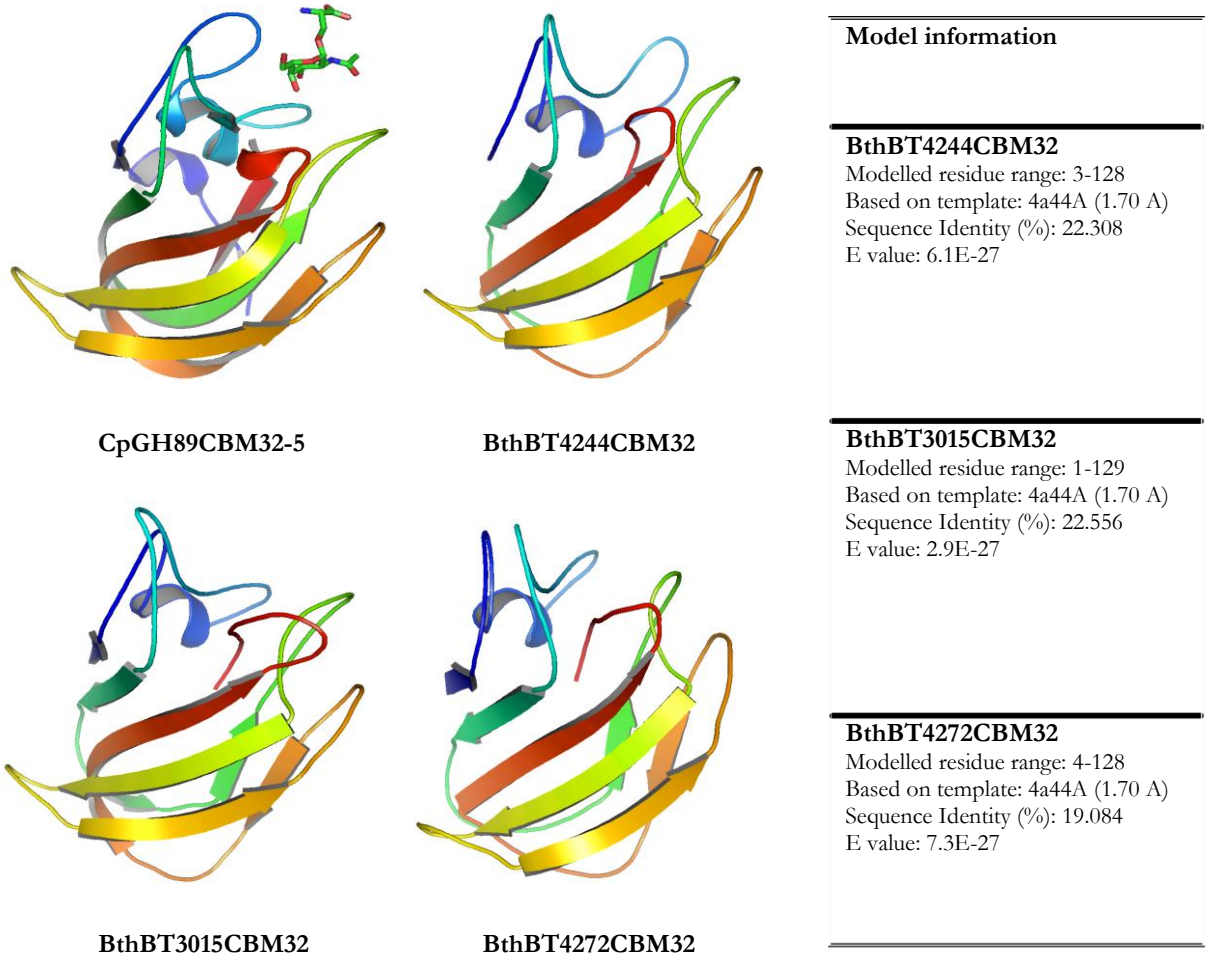
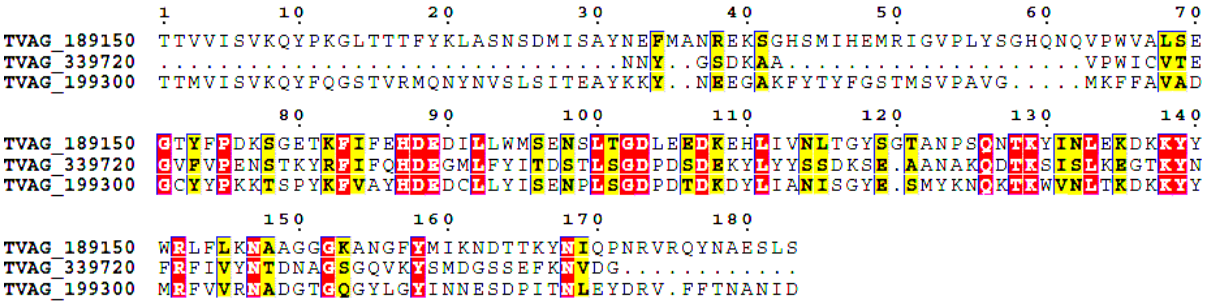


Figure III.11 – SWISS model prediction of the three-dimesional structures of GalNAc binding CBM32 domains from various *B. thetaiotaomicron* M60-like proteins. The template used for the production three-dimensional structures was the *C. perfringens* CpGH89CBM32-5 CBM32 domain in complex with the Tn antigen (GalNAc-Ser) (PDB 4A44A) (top left). CBM32 sequences used were retrieved from the PFAM database and models where produced using the SWISS MODEL comparative modelling server (Arnold *et al.*, 2006) . *B. thetaiotaomicron* CBM32 structures are indicated by the abbreviation **Bth** followed by the protein and domain names.

III.3.10 Putative carbohydrate binding domains of *T. vaginalis* M60-like proteins

GBDL and PA14 sequences of the three selected *T. vaginalis* G3 proteins in this study (TVAG189150, TVAG339720 and TVAG199300) are aligned below in Figure III.12 and contain several highly conserved residues. Despite the generally poor expression of *T. vaginalis* M60-like proteins in the BL21(DE3) *E. coli* host, sufficient amounts of the recombinant TVAG338720-PA14 and TVAG199300-CBD proteins for ITC experiments could be obtained by scaling up the culture volume to ~2 L. Both recombinant proteins however failed to bind any of the mucin sugars they were tested against including Gal, GalNAc, NeuAc, GlcNAc, Fuc, Galβ1-3GlcNAc (LNB), and Galβ1-3GalNAc (T- antigen). Instead the TVAG339720-PA14 was found to bind the glycosaminoglycan sugar - heparin (Section I.2.2., Figure I.11), with a preference for the more sulphated versions of the sugar (Figure III.13, Table III.4).

PA14 sequences



GBDL sequences

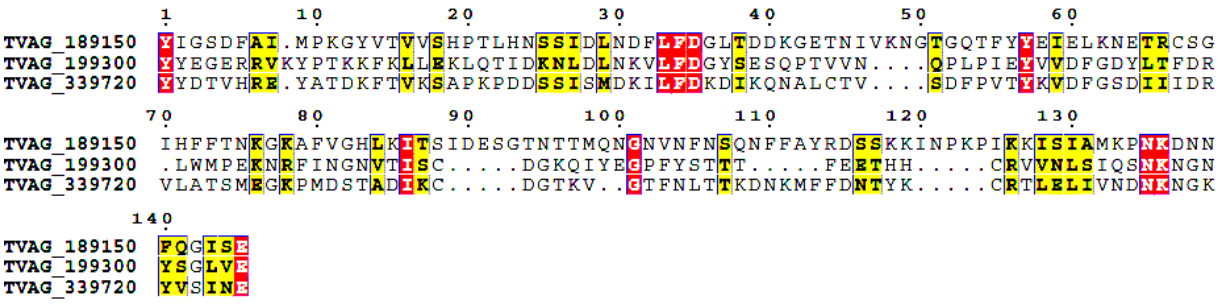
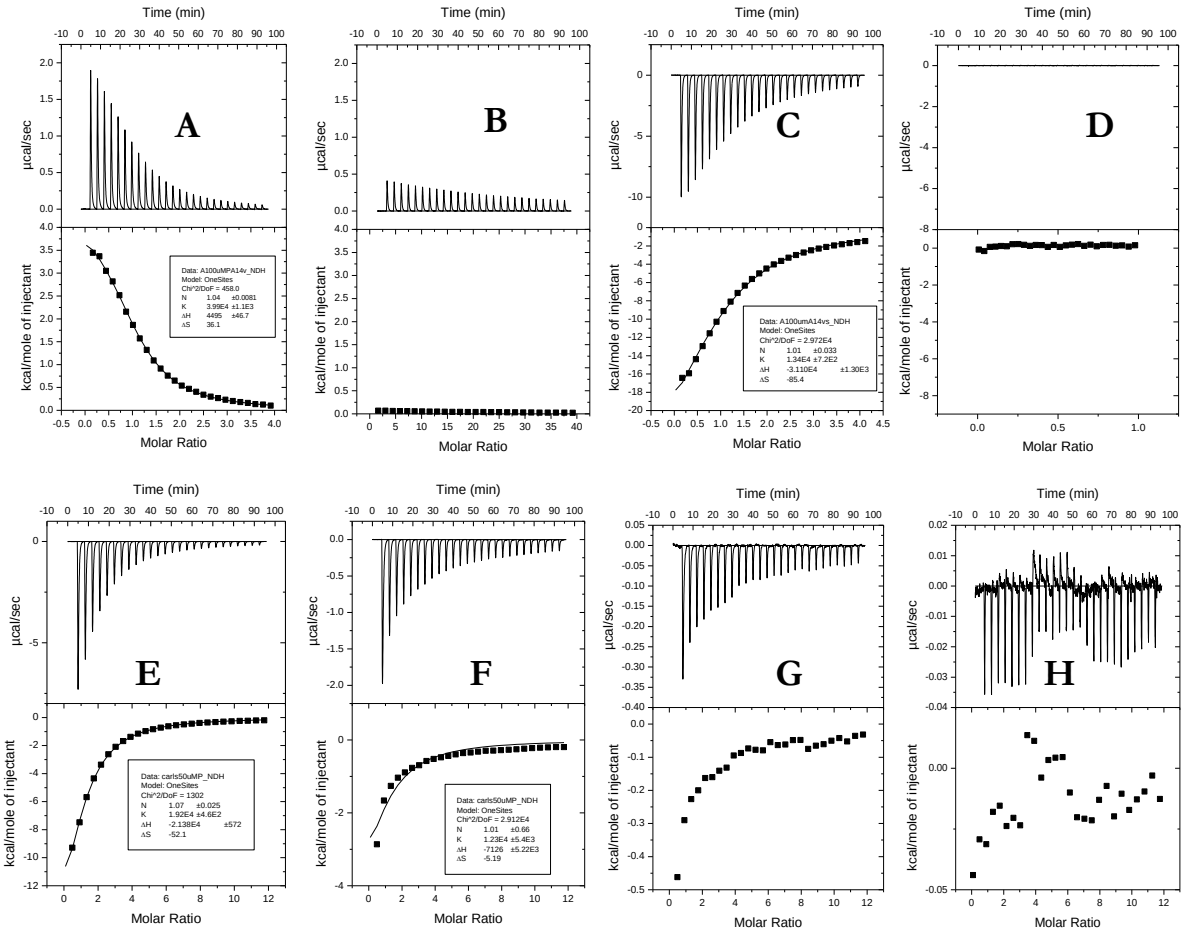
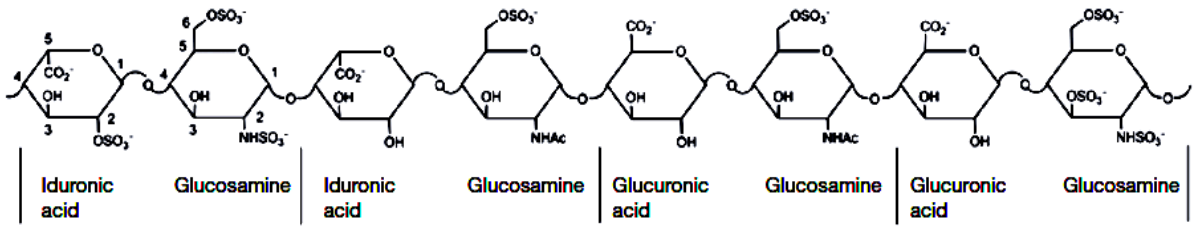


Figure III.12 - Alignment of PA14 sequences from *T. vaginalis* M60-like proteins characterised in this study. Red highlights are amino acid residues showing 100% conservation while yellow highlights are for partially conserved residues. Sequence alignments were viewed using the ESPript 2.2 utility at http://espript.ibcp.fr/ESPript/ESPript/ (Gouet *et al.*, 1999) with a global similarity score threshold of 0.7. Red highlights are for amino acid residues showing 100% conservation while yellow highlights are for residues showing less than 100% conservation but above the global score threshold.





Heparin building blocks

Figure III.13 - Representative ITC data of TVAG339720–PA14 binding to heparin and heparin derivatives. **A, B, C and D:** Heparin interaction with TVAG339720 – PA14 in different buffer conditions. 0.5% of heparin was prepared in 20 mM Tris buffer (pH 7.0) that had earlier been used to dialyse 100 µM of recombinant TVAG339720 – PA14 protein. Using MicroCal™ VP-Isothermal Titration Calorimeter, 27 equal injections of the ligand were made into the protein sample (A) or buffer without protein sample [B (control reaction)] at a cell temperature of 25 ^{o}C. The thermodynamic data obtained was analysed using the Origin version 7 tool. The reactions were exothermic in nature and the steep saturation trend observed for the protein sample compared to the control without the protein was suggestive of a binding interaction. Similar reactions were performed using samples prepared in 50 mM Sodium phosphate buffer (pH 7.0) as shown in C (with protein sample) and D [without protein

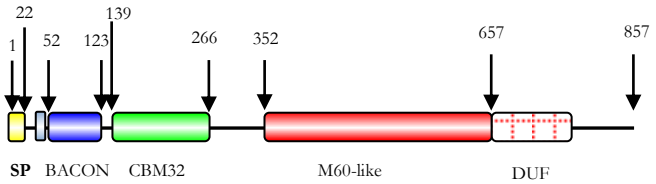
sample (control)]. Reactions were same as in A and B but endothermic in this case. **E, F, G, and H:** Comparing ITC data obtained using 50 μM of TVAG339720-PA14 versus 0.5% of heparin derivatives in 50 mM Sodium phosphate buffer pH 7.0. **E**: Heparin, **F**: N-Acetyl-heparin (Heparin I-A) sodium salt **G**: De-N-sulfated heparin (Heparin I-H) sodium salt **H**: N-Acetyl-de-O-sulfated heparin (Heparin IV-A) sodium salt. The structure below ITC binding curves is a representation of the four disaccharide building blocks of the heparin polysaccharide. It provides information on various sulphation and acetylation events within the molecule and heparin complexity in general (Lee *et al.*, 2013).

Ligand	Binding	$K_a \times 10^4$ (M^{-1})	ΔG (kcal mol^{-1})	ΔH (kcal mol^{-1})	$T\Delta S$ (kcal mol^{-1})	n
Heparin	+	1.92 ±4.6e^2	-5.82	-21.38 ± 572	-15.56	~1.07
N-Acetyl-heparin	+	1.23 ±5.4e^3	-5.56	-7.13 ± 5.22e^3	-1.56	~1.01
De-N-sulfated heparin	±	n/a	n/a	n/a	n/a	n/a
N-Acetyl-de-O-sulfated heparin	x	n/a	n/a	n/a	n/a	n/a

Table III.4 - Affinity and thermodynamic parameters of TVAG339720-PA14 binding to heparin and heparin derivatives. Thermodynamic parameters were calculated as described in Section II.1.32 using the standard thermodynamic equation -RTlnKa = ΔG = ΔH-TΔS, where R = gas constant (1.99 cal.K-1.mol-1), T = temperature in Kelvin (298.15 K), ΔG = change in free enthalpy, ΔS = entropy of binding. n = stoichiometry of binding. In the binding column, the + sign indicates binding was observed for the sugar tested. The ± sign indicates the raw power data showed saturation of the protein following addition of sugar which was suggestive of binding but the affinity was too low for curve fitting and thermodynamic calculations to be carried out confidently. The x sign indicates no saturation trend and hence no binding was observed for the sugar tested.

III.3.11 Substrate induced expression and cellular localisation of BT4244

BT4244 protein belongs to a PUL (PULBT_4240-50) and contains a type II signal peptide in its structure towards the N-terminus between amino acids 22 and 23 based on LipoP and SignalP analyses (Section II.2 and Figure III.14). Close relatives of BT4244 such as BT3015 and BT4272 also contain N-terminal lipoprotein signal peptides (Figure III.1) suggesting that they may all be anchored to either of the two lipid bilayer membranes (Dalbey *et al.,* 2012) of the gram negative *B. thetaiotaomicron* cell envelope. This was investigated for the native BT4244 protein in this study using polyclonal antibodies generated in a rabbit host against the soluble recombinant M60-like domain of the protein (BT4244-M60L, Section III.3.2).




```
# Sequence SpII score=7.10465 margin=0.90515 cleavage=22-23 Pos+2=K
# Cut-off=-3
Sequence        LipoP1.0:Best   SpII    1       1       7.10465
Sequence        LipoP1.0:Margin SpII    1       1       0.90515
Sequence        LipoP1.0:Class  SpI     1       1       6.1995
Sequence        LipoP1.0:Class  CYT     1       1       -0.200913
Sequence        LipoP1.0:Signal CleavII 22      23      7.10465 # TVSLA|CKDTE Pos+2=K
Sequence        LipoP1.0:Signal CleavI  24      25      5.47016 # SLACK|DTEKS
Sequence        LipoP1.0:Signal CleavI  22      23      4.668   # TVSLA|CKDTE
Sequence        LipoP1.0:Signal CleavI  25      26      1.02277 # LACKD|TEKSI
Sequence        LipoP1.0:Signal CleavI  23      24      0.50236 # VSLAC|KDTEK
Sequence        LipoP1.0:Signal CleavI  20      21      -2.84572        # LFTVS|LACKD
```

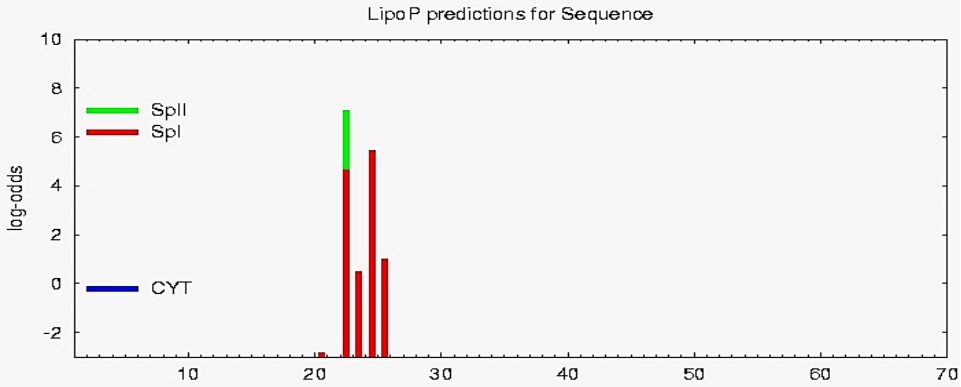


Figure III.14 - Modular representation (top) and LipoP analyses of the native protein encoded by the BT_4244 gene of *B. thetaiotaomicron* (bottom). The encoded protein displays features of a membrane anchored lipoprotein (i.e. high probability for an N-terminal type II signal peptide (spII) over a type I signal peptide (spI) sequence behind a cysteine residue in a lipobox (Section I.4. and Dalbey *et al.,* 2012). The LipoP version 1.0 tool (Section II.2) was used for this analyses. Also see link in Section II.2 for more information on the interpretation of other score values.

III.3.12 Native BT4244 expression

Polyclonal rabbit antibodies generated against BT4244-M60L were very specific as they could reliably distinguish the protein from closely related members such as BT3015-M60L and BT4272-M60L following Western blotting and immunochemical detection (Figure III.15A). Using these antibodies, the native BT4244 protein was only detected in lysates from *B. thetaiotaomicron* cells that had earlier been cultured in minimal medium containing PGMII or III as opposed to glucose (Figure III.15B). These findings were in accordance with data from earlier gene transcription studies suggesting higher mRNA levels for the gene during the late phase of *B. thetaiotaomicron* growth on purified porcine mucosal glycans (Martens *et al.*, 2008). As an additional control *B. thetaiotaomicron* cells with in frame deletions to the BT_4244 gene (ΔBT_4244) were also included in the experiment and native BT4244 was not detected in these strains (Figure III.15C). See Chapter V for more on the creation of ΔBT_4244 deletion mutants

III.3.13 Native BT4244 cellular localisation

A combination of sucrose density gradient centrifugation (SDGC) and immunofluorescence assay (IFA) techniques were used to investigate the cellular localisation of the native BT4244 protein in *B. thetaiotaomicon.* The native protein was found to be enriched in cell membrane fractions of *B. thetaitaiomcron* and not in the soluble fraction suggesting it is likely to be membrane anchored, probably as a lipoprotein as predicted in Figure III.14. To determine what membrane the protein is anchored to, inner and outer membrane fractions from lysed *B. thetaiotaomicon* cells cultured in minimal medium containing 1% PGMIII were obtained by the method of SDGC and analysed by Western blotting and immunochemical detection using the rabbit polyclonal anti-BT4244-M60L antibodies. The native ~80 kDa BT4244 protein was found to be more enriched in the outer- compared to the inner membrane fractions. This was further supported by data from proteinase K digestion (data not included) (Shipman *et al.,* 1999) and immunofluorescence assays (Figure III.16) which both revealed that native BT4244 proteins are indeed surface-exposed.

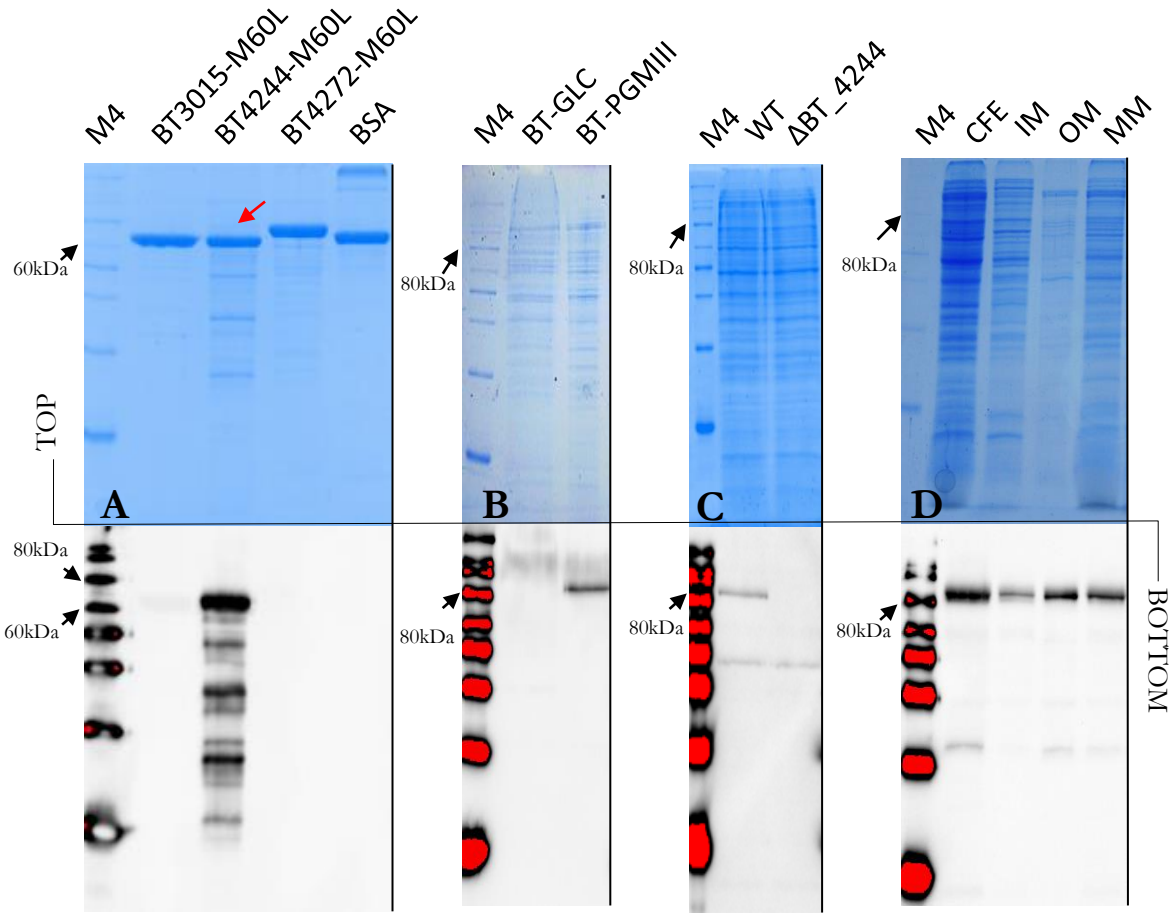


Figure III.15 - Expression and cellular localisation of native BT4244 protein

A: Specificity of polyclonal rabbit anti-BT4244-M60L antibodies. 2 µg of each recombinant protein was subjected to SDS PAGE using 12.5% gels and stained with Coomassie blue reagent (top section) or electro-blotted onto PVDF membranes (bottom section). Bound proteins were probed with a 1:10 000 dilution of primary polyclonal rabbit anti-BT4244-M60L serum and a 1:5000 dilution of secondary HRP conjugated donkey anti-rabbit antibodies. This was followed by chemiluminescence detection using luminol reagent (Biorad) (bottom section). Anti-BT4244-M60L serum was very specific for the ~60 kDa recombinant BT4244-M60L protein (red arrow). See full details of M4 markers in Section II.1.27, Figure II.2. BSA stands for bovine serum albumin **B**: Induction of native BT4244 protein expression by PGMIII. *B. thetaiotaomicron* cells were grown separately in minimal medium containing 1% glucose or PGMIII to an OD600nm of ~ 0.9. Cells were harvested by centrifugation at 13 000 rpm for 1 min into Talon buffer (20 mM Tris/HCl pH 8.0 containing 100 mM NaCl). Fairly equal amounts of cells were lysed in sample buffer by boiling at 98 °C for 3 min and the samples subjected to SDS PAGE and Western blotting as in A. The native ~80 kDa BT4244 protein was only detected in samples from cells cultured in PGMIII (BT-PGMIII) as opposed to glucose (BT-Glc). **C:** Control experiment for the detection of native BT4244 protein using a *B. thetaiotaomicron* wild type (WT) and a knockout strain containing an in-frame deletion of the BT_4244 ORF (ΔBT_4244). Cells were grown in 5 ml of minimal medium with 1% PGMIII to an OD of ~0.6. Cells were harvested as in panel B and 10 µl analysed by SDS PAGE and Western blotting using anti-BT4244-M60L serum. **D**: Detection of native BT4244

protein in *B. thetaiotaomicron* cellular fractions separated by SDGC (Section II.1.41). Cells were grown to OD600nm ~ 0.9 in minimal medium containing 1% PGM II. Cells were harvested from 100 ml of culture and sonicated in sucrose before separation by SDGC. The ~80 kDa band corresponding to the native protein was more enriched in the outer membrane fractions (OM) compared to the inner membrane fraction (IM). CFE: Cell free extract, MM: mixed membrane fraction (Section I.1.41)

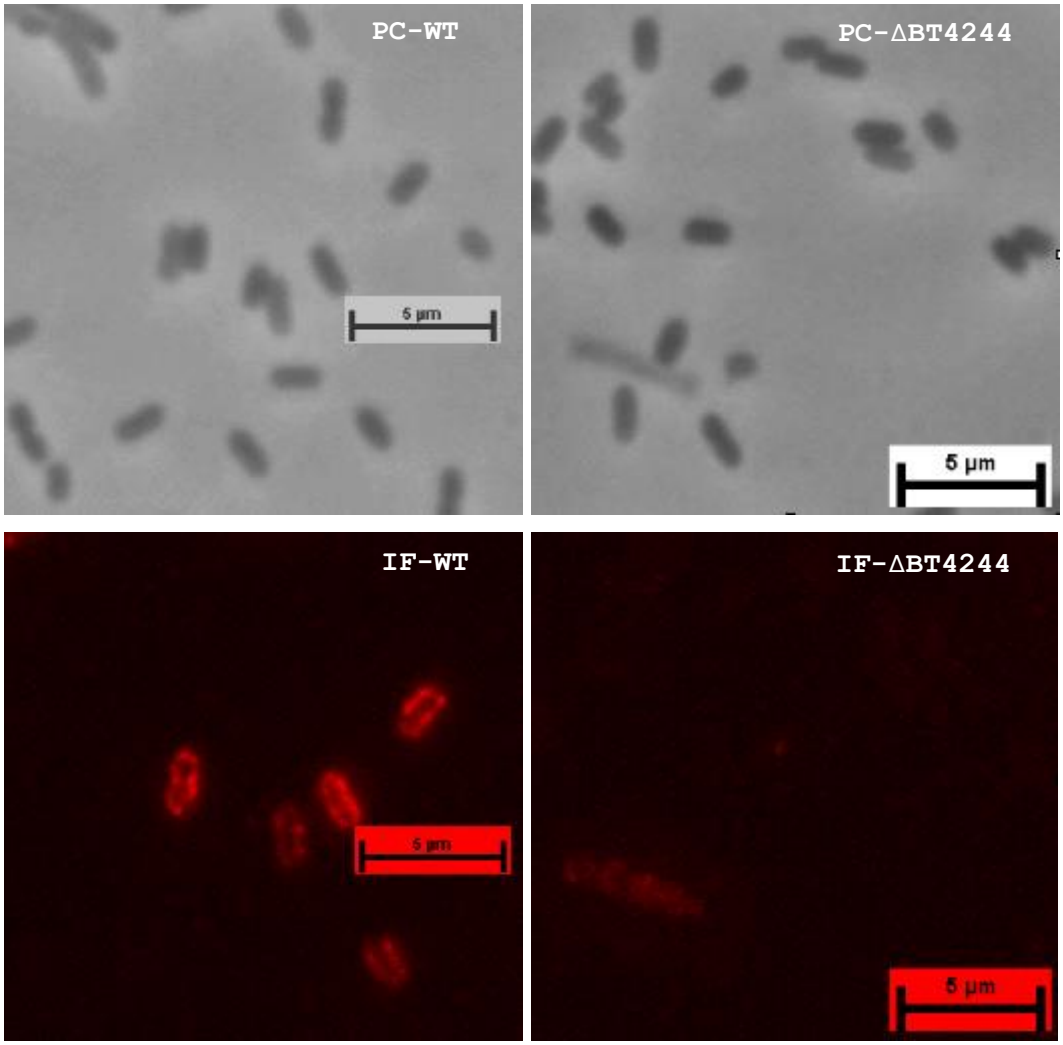


Figure III.16 - Detection of native BT4244 by immunofluorescence microscopy. Non permeabilized cells earlier cultured in minimal medium containing 1% PGMII (to an OD600nm of 0.4) were probed for native BT4244 using a 1:500 dilution of polyclonal rabbit anti-BT4244-M60L serum and a 1: 200 dilution of secondary anti-rabbit Alexa flour 594 antibodies (Section II.1.42). Flourescent images were captured using an Andor iXonEM+ 885 EMCCD camera coupled to a Nikon Ti-E microscope (Nikon) using a 100x/NA 1.4 oil immersion objective. **PC-WT:** phase contrast image for wild type cells, **PC-ΔBT4244:** phase contrast image for mutant cells containing a deletion to the gene encoding the BT4244 protein, **IF-WT** immunofluorescent image of wild type cells, **IF-ΔBT4244:** immunofluorescent image for mutant cells containing a deletion to the gene encoding the BT4244 protein.

III.4 Discussion

III.4.1 M60-like domains of *B. thetaiotaomicron* M60-like/PF13402 proteins

The recombinant M60-like domains of all three *B. thetaiotaomicron* proteins (BT4244, BT3015 and BT4272) were shown in this study to exhibit mucinase activity. BSM and PGMs are very similar to human mucins and both groups are typically O-glycosylated (Section I.2.1.3). BSM glycosylation is relatively less complex (with short, less branched polysaccharide chains) compared to PGMs and most human mucins which possess more complex glycan structures (Karlsson *et al.,* 1997). Glycosylation primarily confers resistance against proteolytic enzymes by limiting access to the core peptide in the glycoprotein structure and hence a likely explanation for the somewhat reduced rate of PGM degradation by recombinant M60-like proteins. Unfortunately due to very high inter-assay variability and the difficulty in accurately quantifying smeared WGA lectin reactive mucin bands after Western blotting, the impact of mucin glycosylation on mucinase activity could not be reliably evaluated using the SDS-agarose gel electrophoresis/WGA lectin detection approach. Nevertheless, the ability of BT4244-FL to cleave human myeloma IgA1 at the O-glycosylated, mucin-like hinge insertion region of the glycoprotein provided an important opportunity to test the effect of deglycosylation on the activity of the enzyme. Firstly, O-glycosylation at the hinge insertion sequence of IgA1 is much less complex compared to typical mucins and secondly, IgA1 unlike mucins runs as distinct bands on SDS PAGE gels making them easier to quantify. As observed in Figure III.5, IgA1 cleavage by recombinant BT4244-FL was clearly potentiated by desialylation implying that glycosylation indeed hinders BT4244-FL proteolytic activity. N-glycan linked sialic acid residues are also present elsewhere in the IgA1 structure (Royle *et al*., 2003) and their removal may also mean reduced steric hindrance against the protease enzymes. This observation was clearly in contrast to the behaviour of leucocyte glycoprotein targeted serine proteases from *Shigella flexneri* whose activities are markedly reduced by desialyation of the glycoprotein substrate (Gutierrez-Jimenez *et al*., 2008, Ruiz-perez *et al*., 2011), implying that attached glycans could also be exploited during substrate recognition.

Substrate cleavage by the recombinant BT4244-FL protein was prevented by a metal chelator such as EDTA or by a single glutamic acid to aspartic acid substitution at position 575 (E575) of the recombinant BT4244-FL protein. E575 is part of what is referred to as the gluzincin motif (HEIGH) of the BT4244 protein, a characteristic feature of most zinc-dependent metalloproteases (Section 1.8.1.1.2) in which the motif glutamic acid acts as the active site amino

acid. These findings thus suggest that *B. thetaiotaomicron* M60-like domain-containing proteins are metal dependent metalloproteases.

Although the BT4244-FL protein was capable of degrading both mucins and IgA1 substrates, gut mucins theoretically represent a more important source of energy based on their glycan content compared to IgA1. The latter is based on the assumption that the role of BT4244-FL which also happens to be a member of a PUL is that of nutrient acquisition and utilisation. It is also worth noting that substrate promiscuity is a very common phenomenon with this group of proteases, especially involving other O-glycosylated substrates that are very similar to the mucins themselves (Table III.5). This however does not rule out the possibility of a dual functional role for this group of proteins in *B. thetaiotaomicron* i.e. in mucin utilisation and immune evasion through IgA proteolysis.

Mucin protease	**Source**	**Variety of target substrates**	**Reference**
Pic serin protease	*Shigella flexneri 2a,* Uropathogenic and Enteroaggregative *Escherichia coli*	CD43, CD44, CD45, CD93, CD162 , ovomucin BSM, Intestinal mucins	Gutie´rrez-Jime´ nez *et al*., 2008, Ruiz-Perez *et al*., 2011
Tsh_serine protease autotransporter	*Escherichia coli*	BSM, Chicken tracheal mucin, PGMIII, Coagulation factor V, Casein	Kobayashi*et al*., 2007, Kostakioti and Stathopoulos, 2004.
Hap Zn(2+)-dependent metalloprotease	*Vibrio cholerae*	Mucin, fibronectin, lactoferrin	Silva *et al*., 2003, Finkelstein *et al*., 1983
ToxR-activated gene A (TagA)	*Vibrio cholerae*	CD43, MUC7, BSM and PGMII/III, LS174T goblet cell surface mucin	Szabady *et al*., 2011
StcE zinc metalloprotease	Enterohemorrhagic *Escherichia coli*	C1-INH, gp340 and MUC7	Grys *et al*., 2005
CP39, Cysteine proteinase	*Trichomonas vaginalis*	Collagens I, III, IV, and V fibronectin, hemoglobin, and IgA and IgG	Hernández-Gutiérrez *et al*., 2004,
Cysteine proteinase	*Trichomonas vaginalis*	BSM and Porcine stomach mucin (PSM), IgG, IgM, IgA	Lehker and Sweeney, 1999, Provenzano and Alderete ., 1995
Elastase B	*Pseudomonas aeruginosa*	Hog/porcine gastric mucin MUC5AC and MUC5B	Aristoteli and Willcox, 2003, Henke *et al.,* 2011
Sap2p (Secretory Aspartyl Proteinase)	*Candida albicans*	PGMIII	Colina *et al*., 1996
Cysteine proteases	*Entamoeba histolytica*	MUC2	Lidell *et al*., 2006 , Moncada *et al*., 2003
Enhancin	*Trichoplusia ni* granulosis virus (TnGV)	Invertebrate intestinal mucin (IIM)	Wang and Granados, 1997
unidentified mucin protease	*Campylobacter pyloridis*	Gastric mucins, albumin	Slomiany *et al*., 1987

Table III.5 - Substrate promiscuity of microbial mucin proteases

IgA1 also seems to contain more than one cleavage site for the BT4244-FL enzyme. Although this was not surprising, given that this is a highly repetitive region in the molecule, it was clearly in contrast to many other well-known microbial IgA1 degrading enzymes which cleave at unique sites within the same hinge region including an IgA1 protease from *Bacteroides melaninogenica* which happens to cleave at same site like BT4244-FL (Figure III.17).

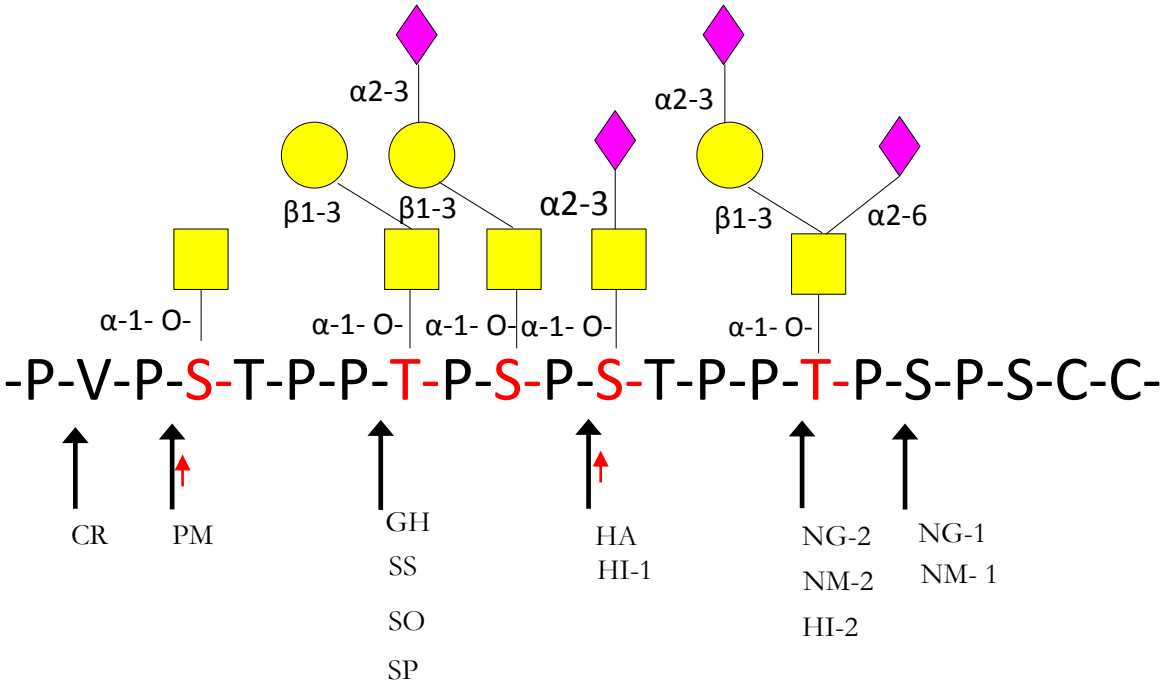


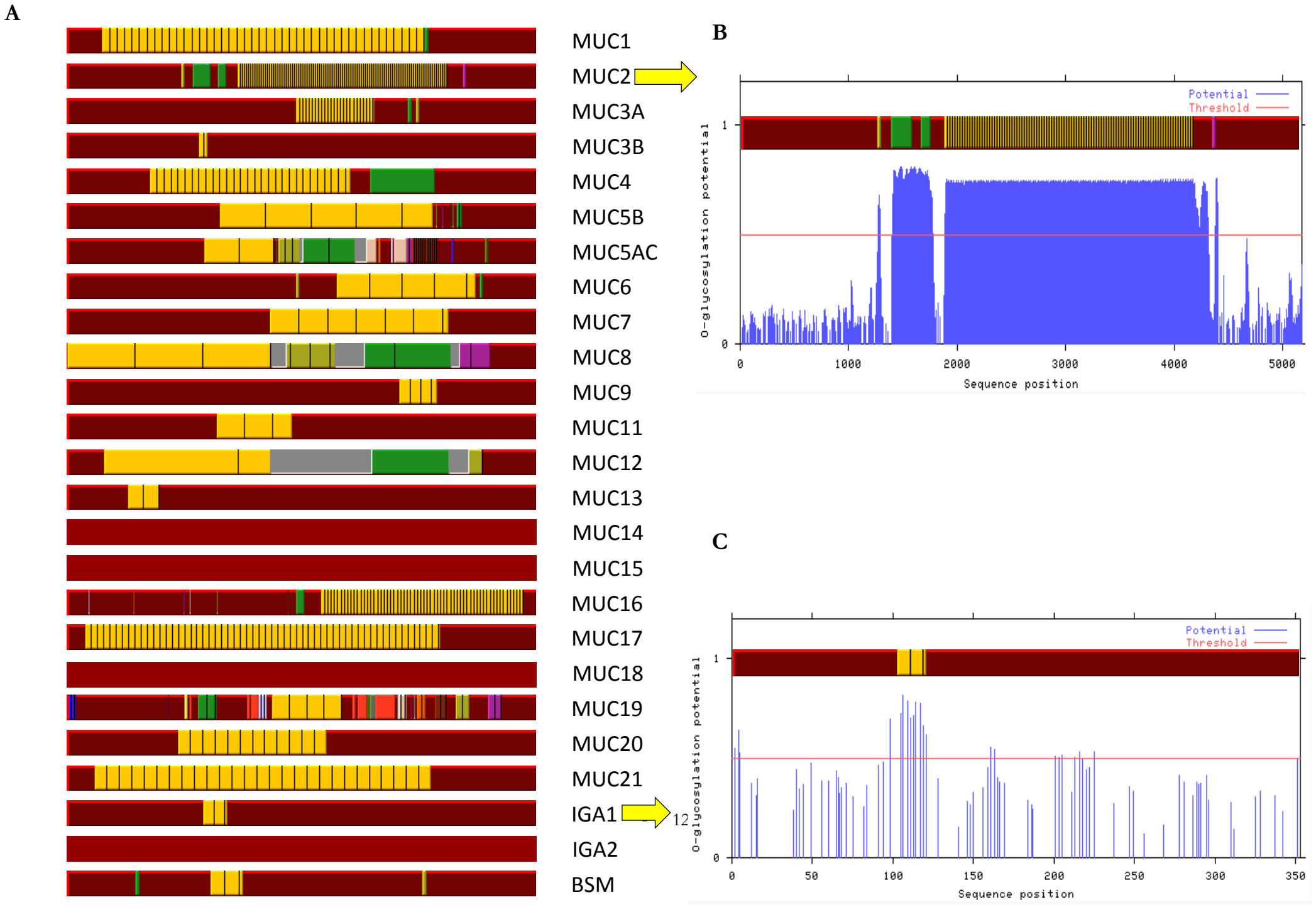
Figure III.17 – Cleavage sites of some microbial proteases at the IgA1 hinge insertion sequence
Red arrows point to sites of *B. thetaiotaomicron* BT4244-FL cleavage (Modified from Mortensen and Kilian, 1984, Gilbert *et al.*, 1991, Qiu *et al.*, 1996). CR: *Clostridium ramosum*, PM: *Prevotella melaninogenica,* GH: *Gemella haemolysans* SS: *Streptococcus sanguinis* SO: *Streptococcus oralis* SP*: Streptococcus pneumonia,* HA*: Haemophilus aegyptius,* HI-1: *Haemophilus influenzae I,* HI-2: *Haemophilus influenzae 2,* NG-1: *Neisseria gonorrhoeae 1,* NG-2: *Neisseria gonorrhoeae 2,* NM-1: *Neisseria meningitides-1, NM-2: Neisseria meningitides-2*

Several mucin glycoproteins including human MUC2 which are highly expressed in the human colon where *B. thetaiotaomicron* thrives, contain high amounts of PTS repeat sequences similar to those of the IgA1 hinge region (Appendix D). In-silico analyses also revealed that just like the IgA1 hinge region, these regions are heavily O - glycosylated (Figure III.18) hence making MUC2 and other colonic mucins potential targets for BT4244-FL. Unfortunately due to the paucity of

commercial human colonic mucins including the human MUC2 glycoprotein, *in-vitro* experiments to test BT4244-FL protease activity on these substrates could not be performed.

III.4.2 Carbohydrate binding modules of *B. thetaiotaomicron* M60-like proteins

The association of carbohydrate binding modules with proteases is a novel functional context for CBMs which are often rather appended to carbohydrate acting enzymes (Section I.8.2). The CBM32 domains of *B. thetaiotaomicron* M60-like proteases exhibited very low affinity binding to mucin sugar N-acetylgalactosamine (GalNAc) ($Ka < 10^3\ M^{-1}$). The low binding affinity was however not surprising as such behaviour is typical of this family of small sugar binding CBMs monosaccharide binding CBMs proteins [Type C CBMs (Boraston *et al.*, 2004, Ficko-blean and Boraston, 2006, 2009, Ficko-blean *et al.*, 2012)]. In any case, binding to GalNAc, a major sugar anchor for a majority of human mucin glycans further strengthened the argument that mucins are targets for *B. thetaiotaomicron* M60-like proteases. GalNAc is capable of forming an α-1-O glycosidic bond with the apomucin peptide and this proximity to the core mucin peptide implies that CBM32 binding to this sugar may enhance contact between the catalytic M60-like domain of the protease and the target core peptide. The specificity for GalNAc could also mean that cleavage is restricted around O-glycosylation sites in the mucin structure where GalNAc is present. The CBM32 domain may thus influence proteolytic activity through proximity and targeting effects (Boraston *et al.*, 2004). Although the BT4244-CBM32 recombinant protein was recalcitrant to crystallisation for X-ray crystallographic studies, clues to its GalNAc recognition mechanism could still be obtained by exploiting its similarity to the already characterised *C. perfringens* CBM32 domains (PDB: 4AAX). The failure of the H168A and N202A alanine substituted mutants of BT4244-CBM32 to bind to GalNAc suggested that the targeted residues (H168 and N202), like their conserved counterparts in CpGH89CBM32-5 may form direct hydrogen bonds with the GalNAc at their binding site. Alanine lacks the imidazole ring of histidine containing the epsilon nitrogen residue that makes hydrogen bonds with the O4 hydroxyl group of GalNAc.



Please see title and legend over page

Figure III.18 - Human mucins contain several uniform/non-uniform PTS rich tandem repeat sequences. **A**: Comparing the distribution of tandem repeat sequences in human mucins, BSM and IgA. Repeat regions are indicated as coloured vertical bars. Largely identical repeats are represented by similarly coloured bars while non-identical repeats are represented with different colours. **B1 and B2**: O – glycosylation analyses using NetOGlyc 4.0 Server (Section II.2, Steentoft *et al.,* 2013) to show that tandem repeat sequences correlates with O-glycosylation. Shown (to the right) are the results for MUC2 and IgA1 using the tool. Repeat region identification was performed using the Xtream repeat region identifier tool (Section II.2)

N202 on the other hand is the equivalent of N1428 in *C. perfringens* that makes two direct hydrogen interactions with GalNAc, one with the O4 hydroxyl group of GalNAc and the other with the ring oxygen. Although equally conserved, the F251 and G157 mutations were not expected to prevent GalNAc binding if indeed both proteins utilise a similar GalNAc recognition mechanism. The equivalents of F251 and G157 in CpGH89CBM32-5 are F1483 and G1385 respectively and neither of these form hydrogen bonds with the GalNAc ligand in CpGH89CBM32-5. G157 is located far off from the binding site of GalNAc while F1483 is solvent exposed and simply forms part of the shallow cleft in the loops at the edges of the β-sandwich that accommodates the carbohydrate.

BT4244-BACON domains on the other hand failed to bind any of the mucin sugars tested in this study (Section III.3.8). BACON domains (pfam:PF13004) are a new family of putative carbohydrate binding domains (Mello *et al.*, 2010) and although they are predicted to be mucin binding, this property has not yet been proven for any member of the group. BACON domains are present in good number of predicted surface proteins from members of the Bacteroidetes group including carbohydrate acting enzymes and proteases (Mello *et al.*, 2010). Assuming these domains actually bind to mucins, then the overall binding affinity of the parent enzyme which will be the combined affinity of the individual CBMs in this case the BACON and CBM32 domains would be higher (avidity effect) (Bolam *et al.*, 2001, Abbott *et al.*, 2008), thus compensating for the low CBM32 affinity.

The current study also provided biochemical evidence that native BT4244 proteins in *B. thetaiotaomicron* are surface-exposed. Its close relatives (BT3015 and BT4272) are also likely to be surface –exposed based on the marked sequence and functional similarities that exist between all three of them. Their localisation to the surface of the cell is consistent with the fact that these proteins have to target substrates in this case mucins or IgA glycoproteins which are extracellular.

In summary, human mucins and IgA represent very important elements of the mucosal immune system and resident or invading microbes must develop strategies to evade or take advantage of these systems. For a mutualistic mucosal microbe like *B. thetaiotaomicron*, mucin proteases could play an important role in the effective colonisation and survival at mucosal surfaces by promoting the utilisation of host derived glycans as well as easing the penetration of the colonic mucus. They have indeed been reported in several mucosal pathogens including *Vibrio cholerae* (Szabady *et al*., 2011), *Entamoeba histolytica* (Lidell *et al*., 2006), *Naegleria fowleri* (Cervantes-Sandoval *et al*. 2008) and *Trichomonas vaginalis* (Lehker and Sweeney, 1999) where they are thought to perform these functions. Interestingly, unlike the above, *B. thetaiotaomicron* is a mutualist and being the first mucin proteases to be characterised from this organism, this study makes an important contribution towards our understanding of host-microbial interactions involving mutualists.

III.4.3 M60-like proteins of *T. vaginalis*

Most of the *T. vaginalis* recombinant proteins expressed in the BL21 (DE3) *E.coli* host strain in this study were insoluble. An attempt to express the M60-like domain of TVAG339720 in a yeast system (*Pichia pastoris*) equally proved futile hence preventing functional studies with the protein. The only solubly expressed recombinant M60-like domain protein (TVAG199300-M60L) also failed to degrade any of the mucin substrates (BSM and PGM) tested against the protein. The PA14 domain of TVAG339720 (TVAG339720-PA14) was also solubly expressed in significant amounts allowing for binding assays to be performed. PA14 domains are so named after the presence of this domain in the protective antigen (PA) region of the complex anthrax toxin (Rigden *et al*., 2004). It is has been identified in a variety of proteins including bacterial adhesins, toxins, proteases, amidases glycosyltransferases, glycosidases (Rigden *et al*., 2004). PA14 domains in adhesins produced by the opportunistic pathogen *Candida glabrata* have been shown to target galacto-configured residues including mucin glycans like the T antigen (Galβ1-3GalNAc) (Zupancic *et al*., 2008, Maestre-Reyna *et al*., 2012). A pectin binding PA14 domain has also been reported in *Clostridium thermocellum* (Kahel-Raifer *et al*., 2010).The inability of TVAG199300-M60L to degrade mucins coupled with TVAG339720-PA14 binding to heparin instead of mucin glycans as earlier hypothesized was an indication that putative M60-like metalloproteases from *T. vaginalis* unlike their counterparts in *B. thetaiotaomicron* may be used for a purpose other than mucin degradation. TVAG339720-PA14 binding to heparin was also relatively higher (~10 times) compared CBM32 binding to GalNAc. The preference for highly sulphated heparin

derivatives was an indication that sulphate groups may play a role in heparin recognition by the PA14 domain.

Heparin is produced by connective tissue-type mast cells, often existing as a proteoglycan called serglycin (Kolset and Tveit, 2008, Rabenstein *et al.*, 2002). Serglycin is known to play important roles in inflammation including the storage and retention of mast cell inflammatory mediators (Kolset and Tveit, 2008, Humphries *et al.*, 1999). Mucosal mast cells which provide defence against mucosal pathogens (Urb and Sheppard, 2012) however synthesize the heparin related chondroitin sulphate glycosaminoglycan instead of heparin (Enerback *et al.*, 1985, Kusche *et al.*, 1988, Kolset and Tveit, 2008) and it is currently not very clear the circumstances under which *T. vaginalis* comes in contact with heparin. An important hypothesis is that the actual target for the M60-like PA14 domains are heparan sulphate glycosaminoglycans which are not only very similar in many ways to heparin glycosaminoglycans but also form part of the epithelial cell glycocalyx at mucosal surfaces e.g in syndecans and glypicans which are well-known epithelial cell surface proteoglycans (Rodgers *et al.*, 2008, Sarrazin *et al.*, 2011, Carlsson and Kjellen 2012). Perlecans, Agrins and Collagen XVIII heparin proteoglycans of the extracellular matrix also represent potential targets for the putative *T. vaginalis* M60-like proteases (Sarrazin *et al.*, 2011). Any of these could serve as important cell surface receptors for *T. vaginalis* during infection alongside other known receptors such as human cervical galectin 1 (Okumura *et al.*, 2008).

Native TVAG339720 and TVAG189150 proteins had been detected in *T. vaginalis* cell membrane extracts and the extracellular localisation of the TVAG339720 protein has been confirmed through immunofluorescence assays by our collaborators in the U.S.A (personal communication, Prof. Robert Hirt). Our collaborators have also shown that overexpressing the latter protein in *T. vaginalis* itself increases vaginal epithelial cell cytolysis *in-vitro*. In summary, although it is not quite clear yet what the M60-like domain-containing proteins of *T. vaginalis* target, data accumulated so far suggests that TVAG339720 and its close relatives may represent important virulence factors for the organism.

III.5 Future work

Some suggestions for future work to further our understanding of the activities and role of *B. thetaiotaomicron* and *T. vaginalis* M60-like proteins include; a) Pursue further studies if possible structural studies on the carbohydrate binding properties of identified M60-like CBMs and b) Screen other potential human mucin/non-mucin glycoprotein targets for the *B. thetaiotaomicron* and *T. vaginalis* M60-like proteins including colonic mucins in the case of the M60-like proteins from *B. thetaiotaomicron* and proteoglycans in addition to urogenital tract mucins in the case of *T. vaginalis* M60-like proteins.

CHAPTER IV

Functional Mechanism of the Sus-like System (BT4240-50) Containing the BT4244 M60-like Protease

IV.1 Introduction

Evidence from the functional characterisation of the fructan and starch utilisation systems (Sus) in *B. thetaiotaomicron* suggests that polysaccharide utilisation loci (PULs) or Sus-like systems represent an important adaptation for the efficient acquisition and utilisation of complex and diverse glycan substrates in the organism (Section I.5.1). Human mucins are complex and diverse glycoproteins and despite their importance in host – microbial interactions (Section I.3, Johansson *et al.*, 2011, McGuckin *et al.*, 2011, Skoog *et al.*, 2012.) vital details of the molecular mechanisms by which individual mucinolytic bacteria such as *B. thetaiotaomicron* or the gut microbial community as a whole interact and carry out mucin utilisation are lacking. On account of their complexity and ability to upregulate the expression of several PUL-associated genes in *B. thetaiotaomicron* (Figure IV.1), mucins and other complex glycan substrates encountered by the organism in the human gut such as glycosaminoglycans, arabinans, arabinogalactans and pectins are also thought to be metabolised by Sus-like systems (Koropatkin *et al.*, 2012, Table IV. 1).
Interestingly, except for the fructan and starch utilisation systems, the functional mechanisms of many of the 88 predicted PULs in *B. thetaiotaomicron* (Martens *et al.*, 2008) are poorly understood, unknown or better still only hypothesised based on gene annotation and transcriptional data. Specifically, this has not been described for any of the 38 host glycan inducible PULs in *B. thetaiotaomicron*, albeit it is thought to be a prolific user of host derived glycans including mucin O- glycans (Martens *et al.*, 2008). Part of the difficulty in the case of mucins is their structural complexity, but this is further complicated by a similarly complex and redundant PUL response observed when the organism is made to rely on host mucin glycans as its sole source of carbon in vitro (Martens *et al.*, 2008, 2011). Mucin-type-O glycans alone for example contain about 12 unique sugar linkages to which the organism upregulates approximately 15 different PULs for their metabolism while starch on the other hand contains just two unique linkages and a single PUL dedicated to its metabolism (Table IV.1). The transcriptional profile highlighting the complexity of *B. thetaiotaomicron*'s response following growth on purified porcine mucosal glycans is shown in Figure IV.1.

In the previous chapter, mucins were identified as targets for various gluzincin M60-like proteases in *B. thetaiotaomicron.* All the proteases (BT4244, BT3015 and BT4272) also happen to be associated with Sus-like systems namely; BT4240-50, BT3010-17 and BT4266-72 respectively encoded by PULs within the genome of the organism (Section I.8.3). For the purpose of clarity, the corresponding PULs encoding the various M60-like systems will be represented here as PULBT_4240-50, PULBT_3010-17 and PULBT_4266-72 respectively. By analogy to the prototypic starch and fructan utilisation systems, the mucinase activities exhibited by various M60-like proteases could be regarded as part of a concerted action by members of the respective Sus-like systems to which they belong to metabolise their target mucin substrate. Supposing this is true, then knowledge of the activities of their functional partners or in general the Sus-like systems to which they belong will not only improve our understanding of the role of M60-like proteases but also PUL mediated mucin utilisation in *B. thetaiotaomicron.*

Glycan	**Unique linkages**	**Degrading species**	**Number of PULs in the species**	**Number of enzymes in the system**
Pectic galactan (β1,4-galactan)	1	*B.thetaiotaomicron*	1	2
Levan (β2,6-fructan)	1	*B. thetaiotaomicron*	1	3
Inulin (β2,1-fructan)	1	*B. ovatus* / *B. caccae*	1	4
Starch	2	*B. thetaiotaomicron*	1	3
Barley β-glucan	2	*B. ovatus*	1	3
Galactomannan and glucomannan	3	*B. ovatus*	1	4
Homogalacturonan	4	*B. thetaiotaomicron*	1	7
Arabinan	4	*B. thetaiotaomicron*	2	6
Xyloglucan	4	*B. ovatus*	1	8
Arabinogalactan	4	*B. thetaiotaomicron*	2	8
Yeast α-mannan	4	*B. thetaiotaomicron*	3	12
Heparin	5	*B. thetaiotaomicron*	1	5
Hyaluronan, dermatan and chondroitin sulphates	7	*B. thetaiotaomicron*	1	5
Xylan	11	*B. ovatus*	2	21
Mucin *O*-linked glycans	12	*B. thetaiotaomicron*	15	17
Rhamnogalacturonan I	13	*B. thetaiotaomicron*	1	20
Rhamnogalacturonan II	22	*B. thetaiotaomicron*	1	32

Table IV. 1 – List of some complex glycans targeted by PULs in *B. thetaiotaomicron* and its close relatives *B. ovatus and B. caccae.* **A**dapted from Koropatkin *et al.,* (2012)

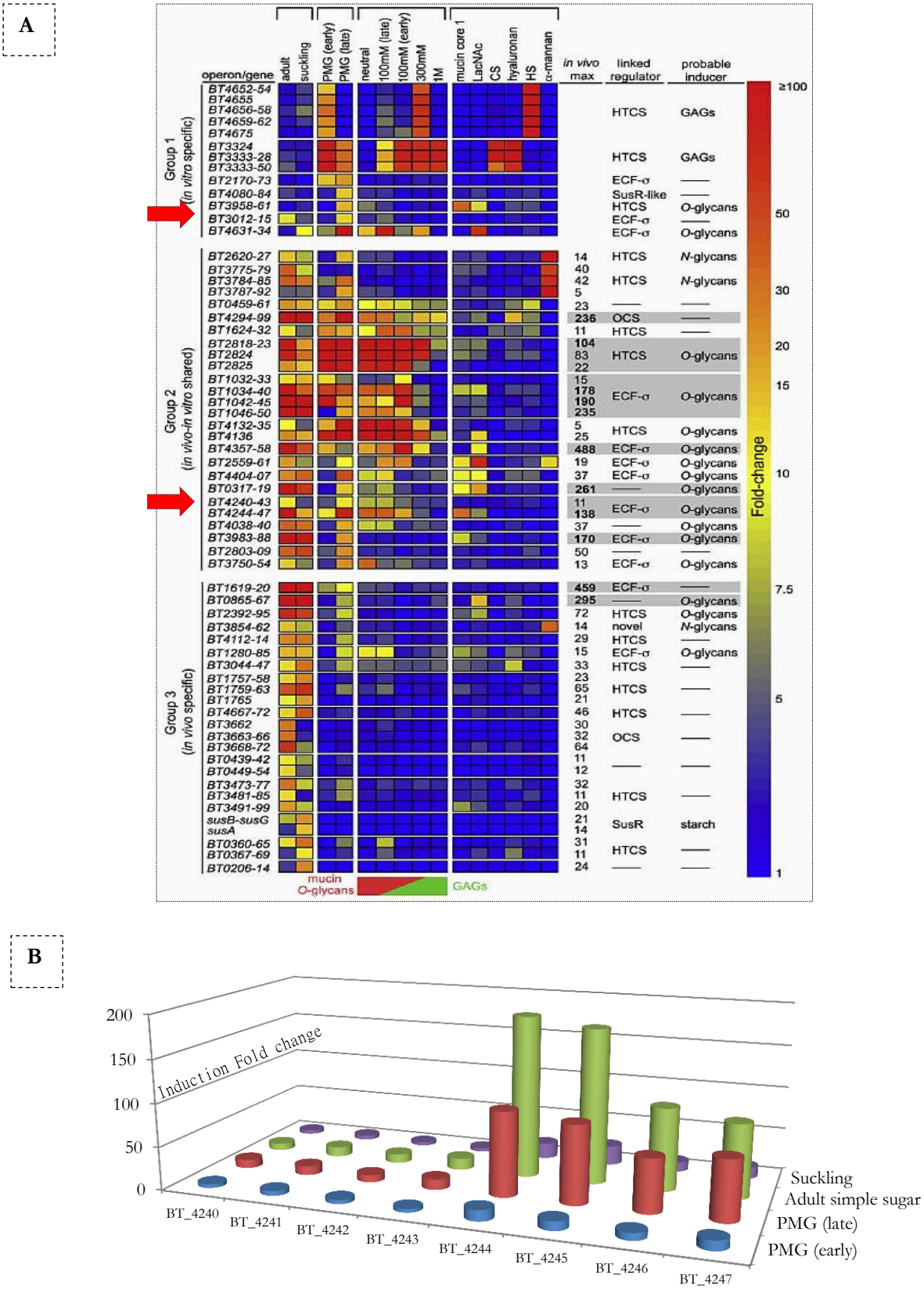


Figure IV.1 - Induction of *B. thetaiotaomicron* host glycan sensitive PULs in different nutrient conditions. A: Heat map showing induction of all host glycan sensitive PULs. Red arrows to the left of the heat map point to

various PULs containing M60-like proteases. Transcriptional profiles are shown for *B. thetaiotaotaomicron* growth on fractionated porcine mucosal glycans (PMG) including early and late phase growth on the unfractionated substrate (PMG early/late). The composition of various fractions were as follows; neutral - mucin O-glycans, 100 mM - mucin O-glycans, N-glycans, GAGs, 300 mM GAGs > O-glycans, 1M – GAGs. CS and HS stand for chondroitin and heparin sulphates respectively. Adult stands for mice fed simple sugar diet while suckling stands for suckling mice. **B:** Fold induction of genes within the putative PUL containing genes from locus BT_4240 chronologically to BT4250 (PULBT_4240-50) relative to levels observed when *B. thetaiotaomicron* is grown in minimal medium containing glucose as sole carbon source. The seemingly low levels of BT_4240-43 induction are due to already high levels of expression in glucose. The transcriptional data were were reproduced from Martens e al., (2008).

The transcriptional data in Figure IV.1 shows upregulation of the PULs encoding BT4244 and BT3015 M60-like proteases (PULBT_4240-50 and PULBT_3010-17 respectively) in one or more conditions of *B. thetaiotaomicron* growth on substrates tested including porcine mucosal glycans (PMG). Several components of PULBT_4240-50 were ~ 100 times up-regulated *in-vitro* in the late phase of growth on PMG while PULBT_3010-17 was ~10 times up regulated under similar conditions. The PUL containing the gene for the BT4272 M60-like protease (PULBT_4266-72) was not shown to be up regulated under any of the conditions tested. PULBT_4240-50 was also upregulated *in-vivo* as opposed to PULBT_3010-17 which was only induced *in-vitro*. An important difference between the two PULs is that unlike the latter PULBT_4240-50 is upregulated to the mucin core 1 glycan or T antigen (Galβ1-3GalNAc) present in many mucin O-type glycans including human colonic mucin -2 (MUC2), IgA1, BSM and PMGs themselves (Section I.2.1.1.1.1). PULBT_4240-50 also contains the highest number of putative gene functional partners (11 in total) including genes encoding putative family 2 and 109 glycoside hydrolases, compared to PULBT_3010-17 and PULBT_4266-72 (Table I.5 and Figure IV.1). As glycoproteins, mucin degradation is likely to involve amongst others the collective action of glycoside hydrolases and mucin proteases and hence taken together the PULBT_4240-50 locus offers a unique opportunity to study the interactions of *B. thetaiotaomicron* with host mucins.

IV.2 Objectives

This chapter is an attempt to elucidate the functional mechanism the BT4240-50 Sus-like system (encoded by PULBT_4240-50) containing the BT4244 M60-like protease, with the aim of putting into context the mucinase activity earlier observed for the enzyme (BT4244-FL) in Chapter III, as well as furthering our understanding of mucin utilisation in *B. thetaiotaomicron*.

IV.3 Results

IV.3.1 Comparing the PULBT_4240-50 and Sus gene loci (PULBT_3698-05)

The prototypic Sus system involved in starch utilisation contains a total of eight protein coding genes occupying a 15.821kbp genomic region (GC content; 45%) in the genome of *B. thetaiotaomicron*. Except for the *sus*R gene, all *sus* genes are unidirectionally transcribed. These include three outer membrane starch binding proteins (SusD, E, and F), two GH13 proteins (SusG and SusA) with alpha amylase activity, a GH97 SusA protein with alpha (α)-glucosidase activity, a TonB dependent SusC protein involved in the transport of starch oligosaccharides into the cell and a SusR regulator protein (Section I.6.1.1)

The PULBT_4240-50 locus (Figure IV.2) on the other hand consists of ~ 11 protein coding genes occupying an 18.320kbp region of the *B. thetaiotaomicron* genome (GC ~ 42%). PULBT_4240-50 genes are divided between two operons, the first from BT_4240 to BT_4243 and the second from BT_4244 to BT_4250 (Martens *et al.*, 2008). Members of the BT4240-50 Sus-like system (encoded by PULBT_4240-50) include, a putative aminoglycoside phosphotransferase family protein (BT4240), two putative glycoside hydrolase family proteins (BT4241 and BT4243), a putative transporter (BT4242), a hypothetical protein (BT4244) earlier shown to possess mucin protease activity (Chapter III), a putative outer membrane protein (BT4245), SusC and D homologues (BT4247 and BT4246, respectively) and regulatory elements (BT4248, BT4249 and BT4250) (Martens *et al.*, 2008, 2009a).

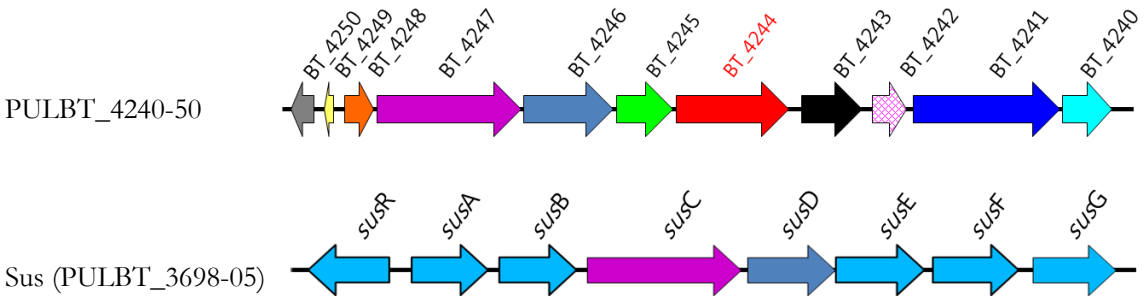


Figure IV.2 – Organisation of the gene loci encoding components of the Sus (PULBT_3698-05) and BT4240-50 (PULBT_4240-50) systems. Homologues of Sus components (mainly SusC and D) present in PULBT_4240-50 are coded by colour. The gene encoding the M60-like protease (BT_4244) is shown in red. Note that only genes within the same PUL are drawn to scale relative to each other and entire PULs are not. See Sus gene IDs or locus tags in Section I.6.1.1. The Sus locus structure was modified from Koropatkin *et al.*, (2012).

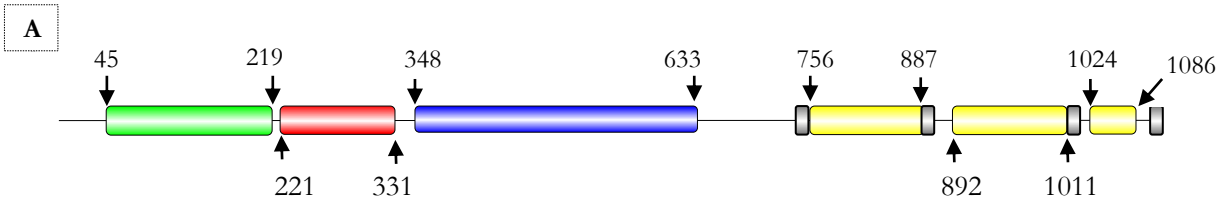
IV.3.1.1 The protein encoded by the BT_4241 gene (BT4241)

IV.3.1.1.2 Features and recombinant protein expression

BT4241 is a large multimodular protein (1103 a.a.) annotated as a beta (β)-galactosidase [EC:3.2.1.23] (PFAM) belonging to the GH2 family of glycoside hydrolases (Chapter 1). Other family 2 glycoside hydrolases include β-glucuronidases (EC 3.2.1.31); β-mannosidases (EC 3.2.1.25); mannosylglycoprotein endo-β-mannosidases (EC 3.2.1.152); or exo-β-glucosaminidases (EC 3.2.1.165) (CAZy). β-galactosidases are enzymes capable of hydrolysing β-glycosidic bonds between galactose and other sugars. Many prokaryotic β-galactosidases have been described (Zahner and Hakenbeck, 2000, Ashida *et al.*, 2001, Jeong *et al.*, 2009, Terra *et al.*, 2010) most of which target the Galβ1-3/4GlcNAc linkage present in human glycoconjugates such as human milk oligosaccharides (HMOs) and mucin glycoproteins.

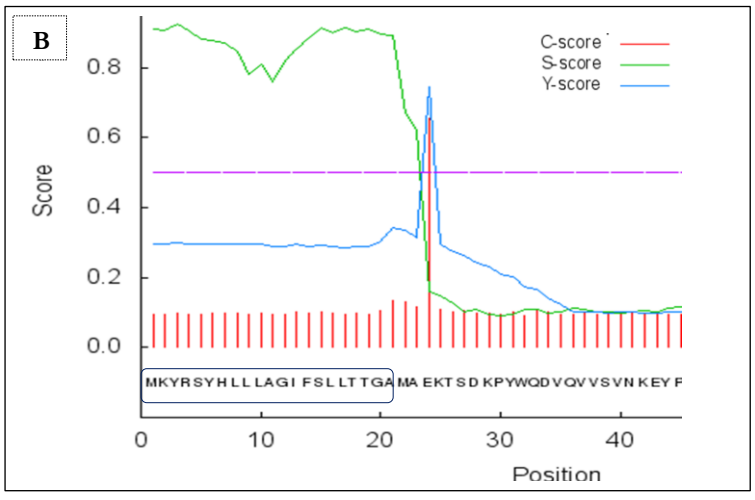
The predicted open reading frame for BT4241 on the KEGG database [http://www.genome.jp/kegg/ (Kanehisa *et al.*, 2012)] encodes no signal peptide. However, upstream of the currently annotated "start" methionine in the putative encoded protein is a potential type I signal peptide (SPI) (Figure IV.3), suggesting there might have been an error in the prediction of the actual start codon for the BT4241 gene (personal communication, Dr David Bolam). BT4241 may thus be periplasmic given that SPI is a typical feature of periplasmic proteins (Cameron *et al.,* 2010, Dalbey *et al.,* 2012). Homologues of BT4241 (>70% sequence identity) are present in related gut *Bacteroides* species such as *B. xylanisolvens, B. fragilis*, and *B. vulgatus*.

To functionally characterise the BT4241 protein, it was initially over-expressed in *E. coli* BL21 (DE3) cells. In the process, the full length BT_4241 gene was amplified from *B. thetaiotaomicron* VPI-5482 genomic DNA by PCR (Section II.1.15) using the primers containing engineered *Bam*HI and *Xho*I restriction sites (Appendix Table A.3). The amplified gene was then cloned into the same sites in the pET-28a (+) vector (Novagen) and expression carried out in *E. coli* BL21 (DE3) host cells (Section II.1.26). Protein expression was induced with 1 mM IPTG at 16 °C overnight. The N-terminal His_6-tagged ~129 kDa recombinant protein (BT4241-FL) was purified by immobilised metal affinity chromatography (IMAC) as described in Section II.1.26 (Figure IV.3). Protein concentration was routinely measured by absorbance at 280 nm (A280nm) using the estimated molar extinction coefficient (240685 M^{-1} cm^{-1}) of the protein.



Key

Source	Domain	Start	End	E-value	
				Sequence	Domain
Pfam A	Glyco_hydro_2_N (sugar binding domain)	45	219	3.1e-48	8.6e-48
disorder	n/a	109	113	n/a	n/a
disorder	n/a	115	119	n/a	n/a
Pfam A	Glyco_hydro_2	221	331	2.1e-15	4.9e-15
Pfam A	Glyco_hydro_2_C (TIM barrel domain)	348	633	2.3e-81	4.1e-81
disorder	n/a	363	364	n/a	n/a
disorder	n/a	538	540	n/a	n/a
disorder	n/a	739	752	n/a	n/a
Pfam A	Bgal_small_N (betagalactosidase small chain)	756	877	2e-54	2.2e-10
disorder	n/a	806	809	n/a	n/a
disorder	n/a	873	874	n/a	n/a
disorder	n/a	878	887	n/a	n/a
Pfam A	Bgal_small_N(betagalactosidase small chain)	892	1011	2e-54	5.7e-29
disorder	n/a	934	935	n/a	n/a
disorder	n/a	954	956	n/a	n/a
disorder	n/a	963	969	n/a	n/a
disorder	n/a	1007	1008	n/a	n/a
disorder	n/a	1010	1033	n/a	n/a
Pfam A	Bgal_small_N(betagalactosidase small chain)	1024	1086	2e-54	0.00063
disorder	n/a	1063	1065	n/a	n/a
disorder	n/a	1099	1101	n/a	n/a



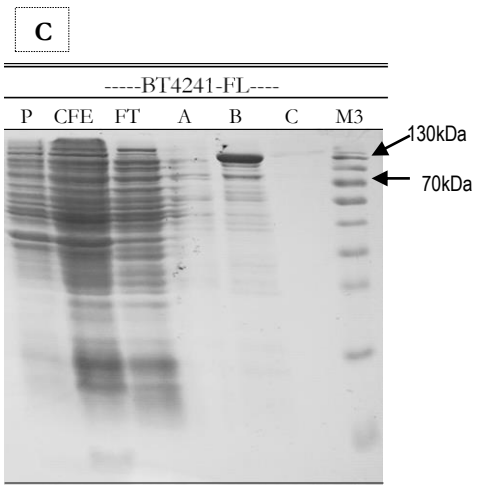


Figure IV.3 - The protein encoded by the BT_4241 gene of *B. thetaiotaomicron*. **A:** Protein features. BT4241 is a large multimodular protein with several GH2 domains including a putative sugar binding and TIM barrel domain. TIM barrel domains represent the catalytic domains of several GH family enzymes (Rigden *et al.*, 2003). See colour coded key below figure for other details relating to various domains in the protein. E-value is the

probability that each sequence or domain is detected by chance. **B:** SignalP analyses of the product containing the predicted BT4241 sequence and 21 amino acids upstream of the sequence (boxed area) suggests the presence of a putative type I signal peptide. Please follow link in Section II.2 for more information on the interpretation of score values. **C:** SDS PAGE analyses of recombinant BT4241 (BT4241-FL) purified by immobilized metal affinity chromatography (IMAC)**.** Protein expression and purification was carried out as described in Section II.1.26. BT4241-FL was purified from 100 ml of *E. coli* BL21 (DE3) culture volume. 5 µl of the insoluble pellet fraction (P) re-suspended in 10 ml of Talon buffer (20 mM Tris/HCl pH 8.0 containing 100 mM NaCl), 10 µl of cell free extract (CFE), 10 µl of flow through (FT), 10 µl of the fraction eluted with 10 mM imidazole (A), 10 µl of fractions sequentially eluted with 100 mM imidazole (B and C) were all analysed by SDS PAGE, using a 12.5% polyacrylamide gel. Please see Section II.1.27 for the molecular weights of various markers in lane M3. The theoretical molecular weight of the recombinant N-terminal His_6-tagged protein is ~129kDa

IV.3.1.1 .3 Screening for candidate sugar targets of BT4241

IV.3.1.1.3.1 Colorimetric assays

The β-galactosidase activity of BT4241 was initially tested using colorimetric assays (Section II.1.43) involving chromogenic synthetic para-nitrophenyl linked β-galactose substrates such as 4/*p*-Nitrophenyl β-D-galactopyranoside (Galβ1-4*p*NP) and 2/*o*-Nitrophenyl β-D-galactopyranoside (Galβ1-2*o*NP). 4-Nitrophenyl β-D-glucopyranoside , 4-Nitrophenyl β-D-mannopyranoside, 4-Nitrophenyl- β-L-fucopyranoside, 4-Nitrophenyl β-D-mannopyranoside, 4-Nitrophenyl α-D-galactopyranoside, 4-Nitrophenyl α-D-glucopyranoside and 4-Nitrophenyl α-L-fucopyranoside were also included in the screens. β-galactosidase activity was detected by measuring the release of *p*/*o*-nitrophenol at 420 nm (A420nm) following treatment of the substrates with the recombinant BT4241-FL protein. A colour change was observed for the Galβ1-4*p*NP and Galβ1-2*o*NP substrates with absorbance readings at least 3 times higher (A420nm >3.0) than the control [BSA, (Section II.1.43)]. In contrast, the other *p*NP substrates showed only very little colour changes (A420nm <0.2) (data not shown).

IV.3.1.1.3.2 Thin layer chromatography

After β-galactosidase activity was confirmed by the above method, several synthetic galactose containing disaccharide sugars, most from mucin type O – glycans were tested against the enzyme. These included galactose containing sugars from the core, backbone and peripheral units of mucin O-glycans such as Galβ1-3GalNAc, Galβ1-3GlcNAc, Galβ1-4GlcNAc and Galα1-3Gal. Digestion products were analysed by thin layer chromatography (TLC) (Section II.1.33). Results from the TLC analyses following a one hour incubation of the substrates (~8 mM each) with the enzyme (0.8 µM) showed complete degradation of the Galβ1-3GalNAc and Galβ1-3GlcNAc substrates. Galactose release was also observed from non mucin substrates such

as α/β-lactose under the same conditions highlighting a broad substrate spectrum for the enzyme (Figure IV.4).

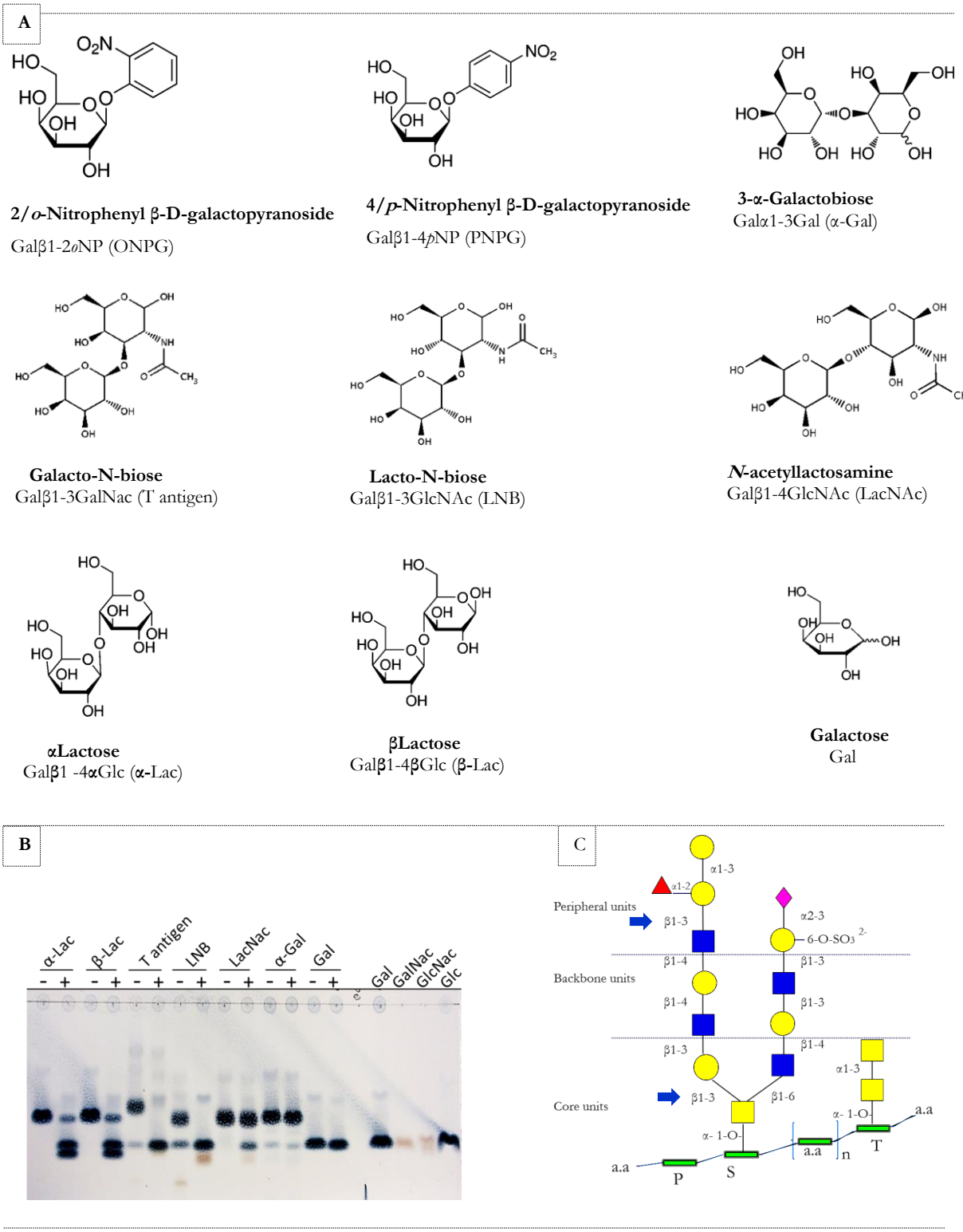


Figure IV.4 – Screening for potential targets of BT4241 in mucin O-glycans. **A**: Structures of various synthetic sugar substrates screened. **B**: TLC analyses of BT4241-FL substrate specificity. 8 mM of each sugar

substrate was prepared and incubated with 0.8 µM of the recombinant BT4241-FL enzyme in 20 mM Tris/HCl pH 8.0 containing 100 mM NaCl at 37°C for 1h. 25 µl of each reaction was boiled at 98 °C for 10 min and 6 µl of each reaction applied to silica gel TLC plates. 6 µl of 2 mg/ml standards were also applied alongside. Plates were stained with DPA reagent (Section II.1.33) after chromatography of the reaction products. Standards were; Gal for galactose, Glc for glucose, GlcNAc for N-acetylglucosamine, GalNAc for N-acetylgalactosamine **C**: Model mucin structure showing the location of bonds (blue arrows) targeted by BT4241-FL based on the TLC data from B**.** See more details of the structure in Section I.2.1.1.1.1.

IV.3.1.1.4 Enzyme kinetics and pH dependence

Equal amounts of the BT4241-FL enzyme, within an hour completely degraded both the Galβ1-3GalNAc (T-antigen) and Galβ1-3GlcNAc (LNB) sugar substrates (Figure IV.4). To determine which substrate was the most preferred by the enzyme, rate experiments were performed using the protocol described in the Megazyme galactose detection kit (Megazyme) (Section II.1.44). The highest activity of the enzyme was observed against the T antigen substrate with Km, Kcat and Vmax values of 0.36 mM, 14 s^{-1} and 243 µMmin^{-1} respectively (Figure IV.5). The pH dependency of the enzyme activity was also analysed using colorimetric assays (Section II.1.43). BT4241-FL had maximal activity between pH 7.0 and 8.0 retaining about half of this activity at about pH 5.5 and 9.0.

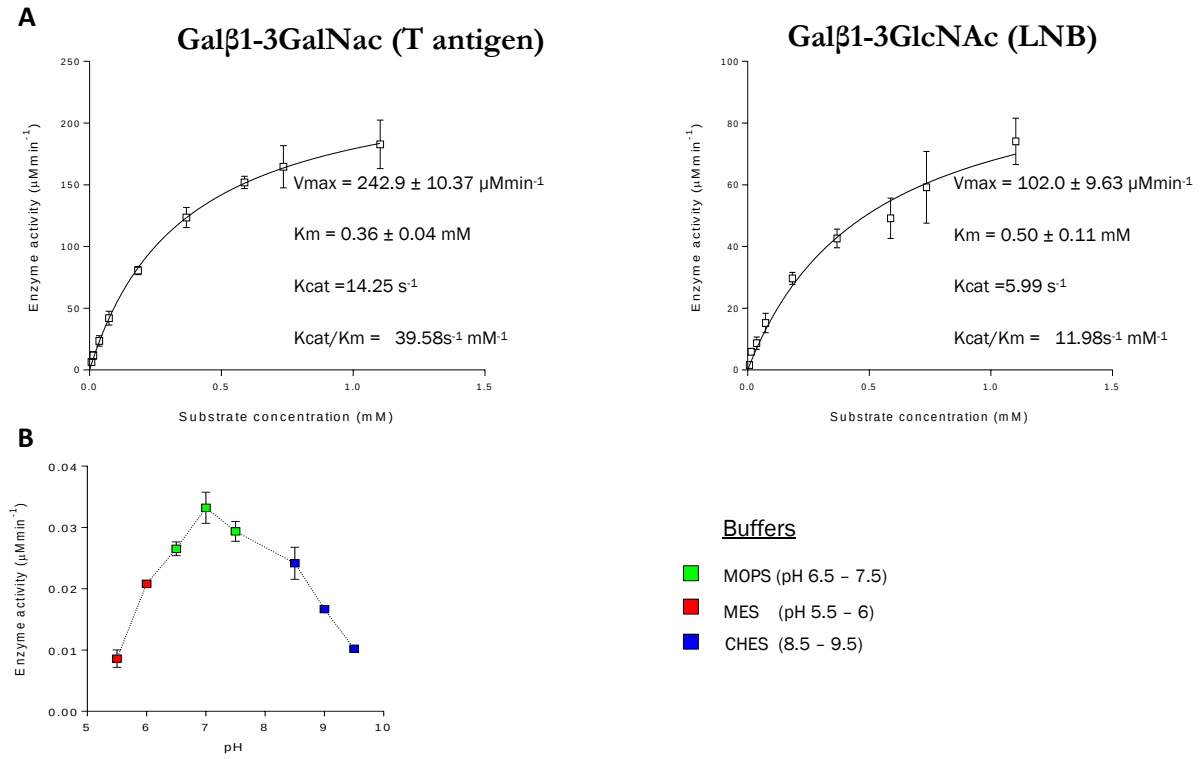


Figure IV.5 – Kinetics and pH dependency of BT4241-FL β-galactosidase activity. A: Kinetics of BT4241-FL activity against T antigen and LNB. Rates were measured using a modification of the assay protocol described for Megazyme galactose detection kit (Section II.1.44). **B**: pH dependency of BT4241-FL β-galactosidase

activity at 37 °C. This was assessed through colorimetric assays (Section II.1.43) using Galβ1-2*o*NP (ONPG) as substrate in various buffers.

IV.3.1.1.5 BT4241-FL β-galactosidase activity against natural glycan substrates

BT4241 was capable of releasing galactose from natural glycan containing substrates such as porcine gastric mucins (PGMII and III) and bovine sialofetuin (BF). BF just like PGM contains both O and N-linked oligosaccharides but in relatively lower amounts (Karlsson *et al.*, 2002). It was found to be entirely exo-acting as galactose release was only observed following defucosylation and desialylation of PGM and BF substrates respectively (Figure IV.6). Commercial PGMIII and II contain just about ~ 0.5-1.5% and 1% bound sialic acid respectively (Sigma) with several fucosylated glycans as opposed to bovine sialofetuin where a significant amount of its N-glycans are sialylated (Karlsson *et al.*, 2002, Royle *et al.*, 2002, Yamada *et al.*, 2007).

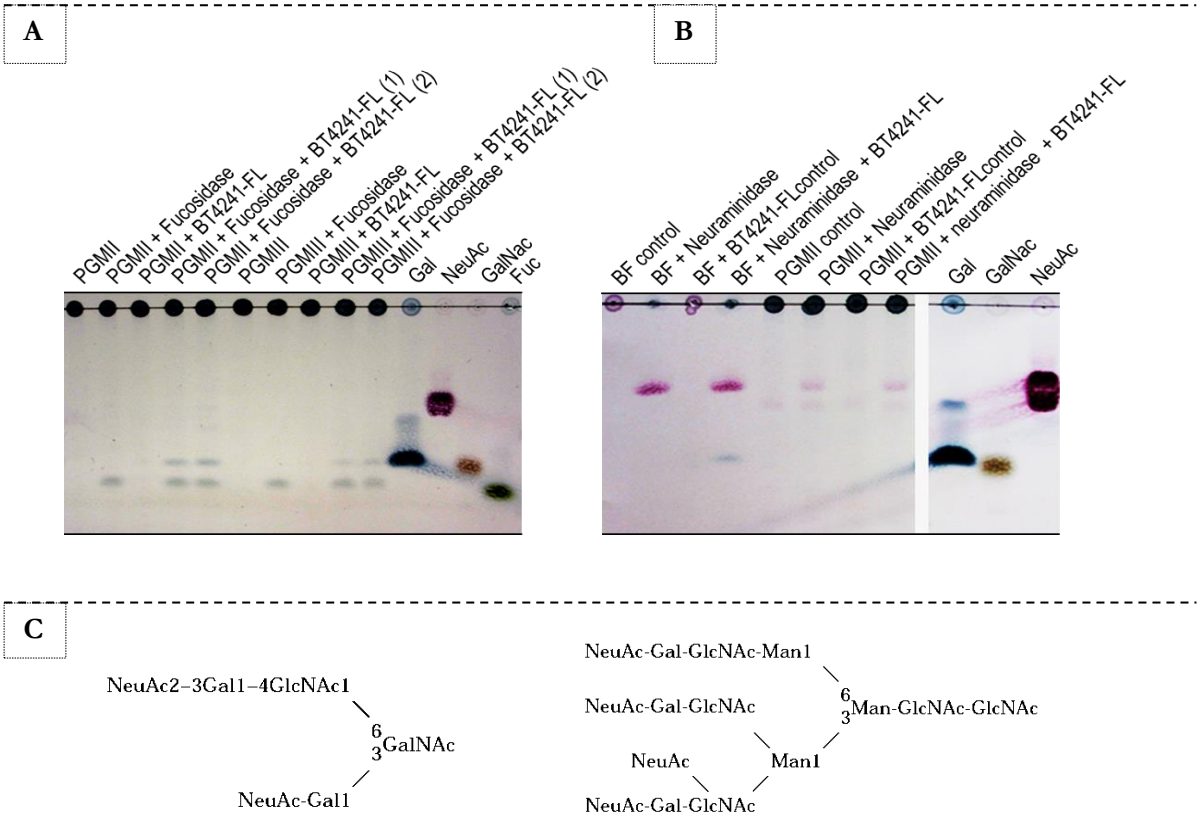


Figure IV.6 – Release of galactose from natural glycoprotein substrates. A: TLC analyses of PGMII and III treated with BT4241-FL and fucosidase enzymes. Incubation reactions contained 4 mg/ml PGMII and III, 0.9 μM α-fucosidase enzyme (GenBank: AAQ72464.1) from Dr Arthur rogowski and 0.32 μM of recombinant

BT4241-FL in 20 mM Tris-HCl pH 7.5. Reactions were carried out at 37 °C overnight. Were reactions contained both fucosidase and BT4241-FL enzymes, the reaction was either initially incubated with the fucosidase enzyme for about 5 h at at 37 °C, then boiled at 98 °C for 3 min, before addition of BT4241-FL (2) or both enzymes were included at the same time and incubated overnight . Samples were collected the next day and boiled at 37 °C for 3 min before applying to TLC plates alongside standards. 6 µl of each standard [Gal (galactose)– 2 mg/ml, NeuAc (N-acetylneuraminic acid)– 2.5 mg/ml, GalNAc (N-acetylgalactosamine) -4 mg/ml, Fuc (Fucose) - 5 mg/ml) was applied to TLC plates. **B:** TLC analyses of bovine sialofetuin (BF) treated with BT4241-FL and neuraminidase enzymes. Incubation reactions contained 6.25 mg/ml BF, 0.31 U/ml of α2-3,6,8,9-Neuraminidase from *Arthrobacter ureafaciens* (Calbiochem) and 1.2 µM of BT4241-FL in 20 mM Tris-HCl pH 7.5. Reactions were carried out at 37 °C overnight. PGMIII samples contained 8 mg/ml PGMIII and same amounts of neuraminidase and BT4241-FL enzyme. Samples after incubation were treated as in **A** before TLC analyses. The NeuAc standard concentration in this case was 5 mg/ml. **C:** Typical sialylated N and O-glycan structures found in bovine sialofetuin. Most of the β1-3/4 linked galactose residues are protected by sialic acids (adapted from Karlsson *et al* 2002). See structures of fucosylated PGM glycans in Section I.2.1.3 (Chapter I)

IV.3.1.2 The protein encoded by the BT_4243 gene (BT4243)

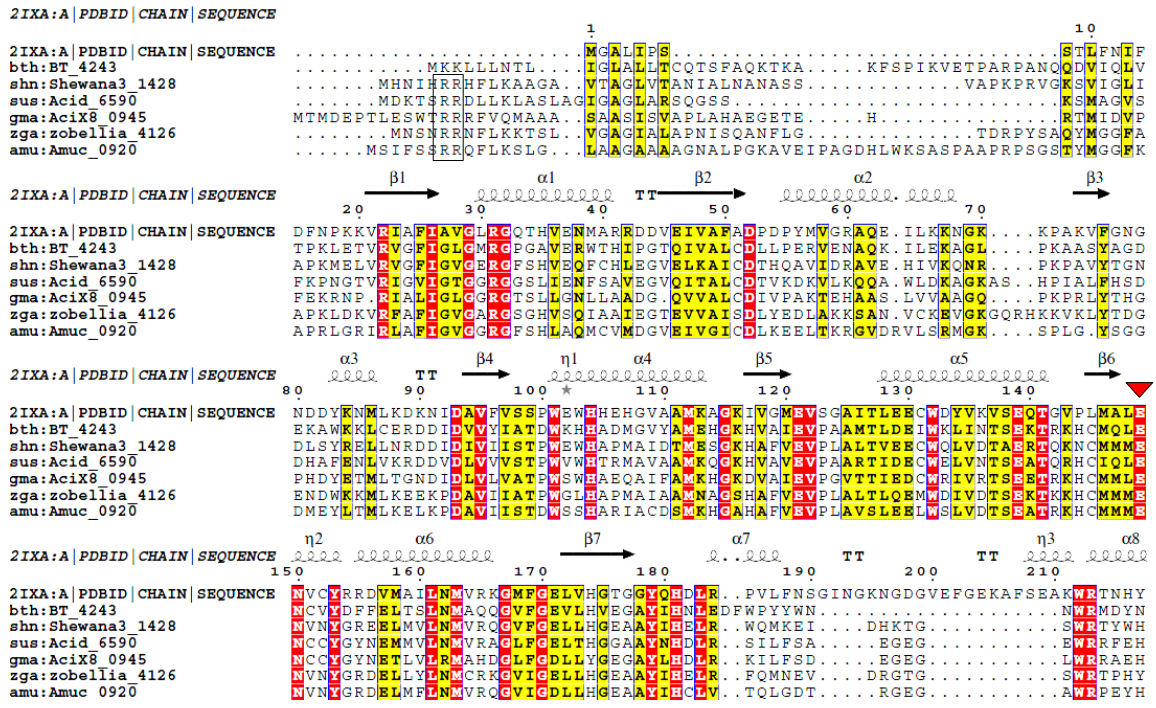
IV.3.1.2.1 Features and recombinant protein expression

BT4243 is a 467 amino acid long protein belonging to the GH109 family of glycoside hydrolases (CAZy). Members of this family are thought to exhibit α-N-acetylgalactosaminidase activity and by virtue are potential candidates for use in the production of universal red blood cells (Liu *et al.*, 2007, Olsson *et al.*, 2004). The encoded protein contains a type I signal peptide sequence towards its N-terminal [hence maybe periplasmic (Cameron *et al.,* 2010, Dalbey *et al.,* 2012)]) and a large domain termed GFO_IDH_MocA, defined as oxidoreductase family with a NAD+ binding Rossmann fold (Figure IV.8). A similar N-terminal dinucleotide binding Rossman fold has been reported in the structure of the α-N-acetylgalactosaminidase enzyme (NagA) encoded by the nagA gene of *Elizabethkingia meningoseptica* (PDB: 2IXA, Liu *et al.*, 2007). NAD+ is thought to be important in the catalytic mechanism of this group of enzymes (CAZy, Liu *et al.*, 2007). As a putative periplasmic NAD+ binding protein, BT4243 ideally will be secreted in a completely folded state requiring a secretion system such as the TAT system (twin-arginine translocase secretion system) as opposed to the Sec secretion system of bacteria for its export into the periplasm (Posey *et al.*, 2006, Palmer *et al.*, 2012). The TAT secretion system is so named after the twin-arginine secretion signal present in the N-terminal of TAT secreted proteins. Intriguingly, BT4243 lacks this motif and rather contains two lysine residues aligning with the putative twin-arginine motif of its close homologues (Figure IV.7). Either these lysine residues

are equally recognisable TAT secretion signals or the protein uses a completely different, yet unknown secretion system for its export into the periplasm.

Just like BT4241, homologues of BT4241 (>70% sequence identity) are present in other gut *Bacteroides* species such as *B. fragilis* and *B. vulgatus*. An alignment of BT4243 with the *E. meningosepticum* NagA protein and selected homologues from other organisms showing less than 50% sequence identity revealed conservation of the several residues including Tyr-307, Tyr-225, His-228 and Glu-149 thought to be important for interactions with N-acetylgalactosamine (GalNAc) in the active site of NagA (Figure IV.7).

To functionally characterise the BT4243 protein, it was initially over-expressed in an *E. coli* BL21 (DE3) host. In the process, the full length BT_4243 gene excluding the region encoding the predicted type I signal peptide sequence was amplified from *B. thetaiotaomicron* genomic DNA by PCR (Section II.1.15) using the following primers using the primers containing engineered *Bam*HI and *XhoI* restriction sites (Appendix Table A.3). The amplified gene was then cloned into the same sites of the pET-28a(+) vector (Novagen) and expression carried out in *E. coli* BL21 (DE3) host cells (Section II.1.26). Protein expression was induced with 1 mM IPTG at 16 °C overnight. The N-terminal His_6-tagged ~54kDa recombinant protein (BT4243-FL) was purified by immobilised metal affinity chromatography (IMAC) as described in Section II.1.26 (Figure IV.8). Protein concentration was routinely measured by absorbance at 280 nm (A280nm) using the estimated molar extinction coefficient (81290 M^{-1} cm^{-1}) of the protein.



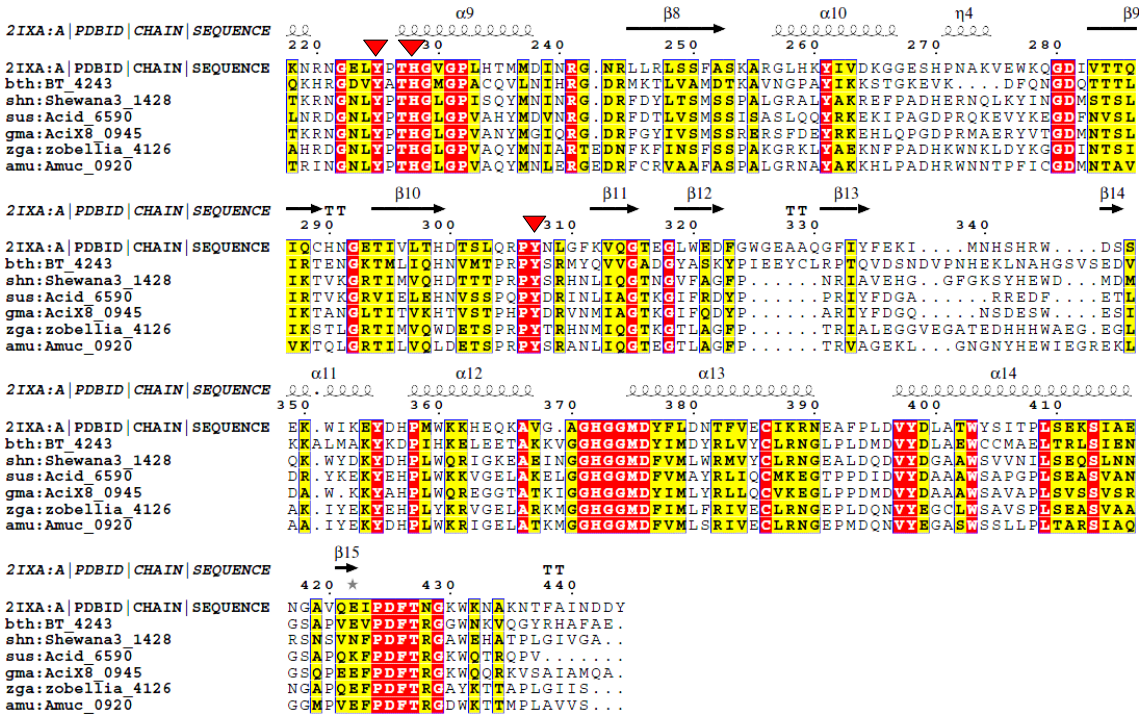
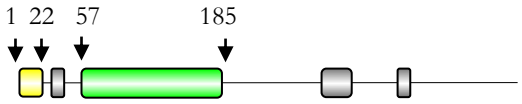


Figure IV.7 - Alignment of *E. meningoseptica* NagA against BT4243 and its homologues from different organisms. Apart from BT4243 protein (denoted in the figure as bth: BT_4243), Only homologues with less than 50% identity to BT4243 were selected so as to demonstrate the strong conservation of structural and catalytic site residues within this group. Red inverted triangles point to conserved residues Tyr-307, Tyr-225, His-228 and Glu-149, thought to be important for interactions with GalNAc in the active site of the *E. meningoseptica* NagA enzyme (PDB: 2IXA, Liu *et al.*, 2007). Also included (above the alignment) is the secondary structure of NagA showing regions forming alternating α-helices and β-sheets of the N-terminal β-α-β-α-β motif typical of NAD+ binding enzymes (Lesk *et al.*, 1995). Sequences used in the alignment came from the following sources; bth: BT_4243: *Bacteroides thetaiotaomicron,* shn: Shewana3_1428: *Shewanella sp.* ANA-3, sus: Acid_6590 : *Candidatus Solibacter usitatus,* gma: AciX8_0945: *Granulicella mallensis,* zga: zobellia_4126: *Zobellia galactanivorans* and amu: Amuc_0920: *Akkermansia muciniphila.* The putative twin-arginine secretion motif (RR) in various aligned sequences excluding BT4243 is shown in the boxed area towards the N-terminal of the proteins. Notice that this motif is replaced by a twin lysine motif in BT4243. Sequence alignments were viewed using the ESPript 2.2 utility at http://espript.ibcp.fr/ESPript/ESPript/ (Gouet *et al.*, 1999) with a global similarity score threshold of 0.7. Red highlights are for amino acid residues showing 100% conservation while yellow highlights are for residues showing less than 100% conservation but above the global score threshold.

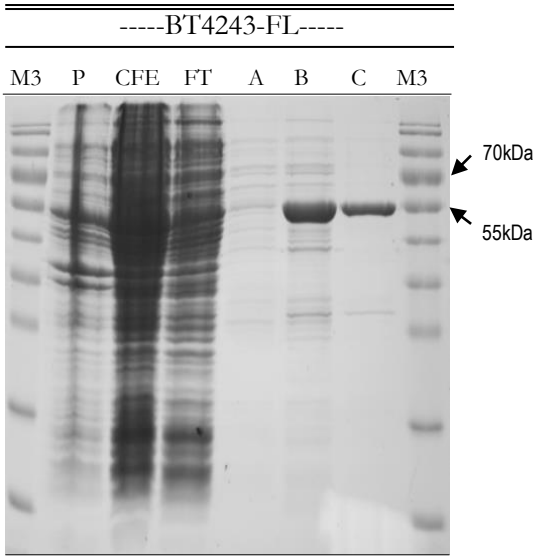
A



Key

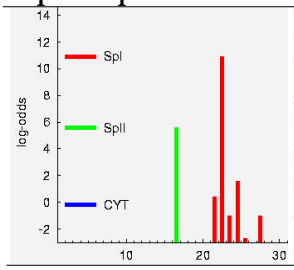
Source	Domain	Start	End	E-value	
				Sequence	Domain
sig_p	n/a	1	22	n/a	n/a
Low complexity	n/a	4	15	n/a	n/a
disorder	n/a	32	34	n/a	n/a
Pfam A	GFO_IDH_MocA (oxidoreductase family, NAD-binding Rossmann fold)	57	185	5.7e-13	1e-12
disorder	n/a	285	288	n/a	n/a
disorder	n/a	291	299	n/a	n/a
disorder	n/a	302	303	n/a	n/a
disorder	n/a	353	363	n/a	n/a

C



B

LipoP report




```
# Sequence SpI score=10.9023 margin=5.32938 cleavage=22-23
# Cut-off=-3
Sequence        LipoP1.0:Best   SpI     1       1       10.9023
Sequence        LipoP1.0:Margin SpI     1       1       5.32938
Sequence        LipoP1.0:Class  SpII    1       1       5.57292
Sequence        LipoP1.0:Class  CYT     1       1       -0.200913
Sequence        LipoP1.0:Signal CleavI  22      23      10.8981 # QTSFA|QKTKA
Sequence        LipoP1.0:Signal CleavI  24      25      1.58844 # SFAQK|TKAKF
Sequence        LipoP1.0:Signal CleavI  21      22      0.408115        # CQTSF|AQKTK
Sequence        LipoP1.0:Signal CleavI  23      24      -1.00742        # TSFAQ|KTKAK
Sequence        LipoP1.0:Signal CleavI  27      28      -1.02921        # QKTKA|KFSPI
Sequence        LipoP1.0:Signal CleavI  25      26      -2.72002        # FAQKT|KAKFS
Sequence        LipoP1.0:Signal CleavII 16      17      5.57292 # LALLT|CQTSF Pos+2=Q
```

Figure IV.8 - The protein encoded by the BT_4243 gene of *B. thetaiotaomicron*. **A**: Protein features. BT4243 is a 476 amino acid long protein containing a putative N-terminal type I signal peptide sequence and an NAD+ binding Rossmann fold also present in the NagA α-N-acetylgalactosaminidase of *E. meningoseptica* (Liu *et al.*, 2007). See colour coded key below figure for other details relating to various domains in the protein. E-value is the probability that each sequence or domain is detected by chance **B**: LipoP analyses show a high probability for a type I signal peptide (spI) over a type II signal peptide (spII) sequence. Please follow link in Section II.2 for more information on the interpretation of other score values. **C:** SDS PAGE analyses of recombinant BT4243 (BT4243-FL) purified by IMAC (section II). Protein expression and purification was carried out as described in Section II.1.26. BT4243-FL was purified from 100 ml of *E. coli* BL21 (DE3) culture volume. 5ul of the insoluble pellet fraction (P) re-suspended in 10ml of Talon buffer (20 mM Tris/HCl pH 8.0 containing 100 mM NaCl), 10 µl of cell free extract (CFE), 10 µl of flow through (FT), 10 µl of the fraction eluted with 10 mM imidazole (A), 10 µl of fractions sequentially eluted with 100 mM imidazole (B and C) were analysed by SDS PAGE,

using a 12.5% polyacrylamide gel. Also see Section II.1.27 for the molecular weights of various markers in lane M3. The theoretical molecular weight of the recombinant N-terminal His_6-tagged protein is ~54 kDa

IV.3.1.2.2 Enzymatic activity

IV.3.1.2.2.1 Hydrolysis of GalNAc –PNP substrates

BT4243-FL was evaluated for α-N-acetylgalactosaminidase activity by initially testing against the synthetic chromogenic substrate; 4-Nitrophenyl N-acetyl-α-D-galactosaminide or *p*-Nitrophenyl 2-acetamido-2-deoxy-α-D-galactopyranoside (GalNAcα1-4*p*NP) in colorimetric assays (Section II.1.43). Hydrolysis of GalNAc α1-4*p*NP by recombinant BT4243-FL produced a yellow colour that was quantified at 420nm (A420nm). Just like NagA, BT4243 was also capable of hydrolysing the β - configuration of the sugar (GalNAc β1-4*p*NP) although the Km was ~3 times higher compared to the α- configuration (Figure IV.9).

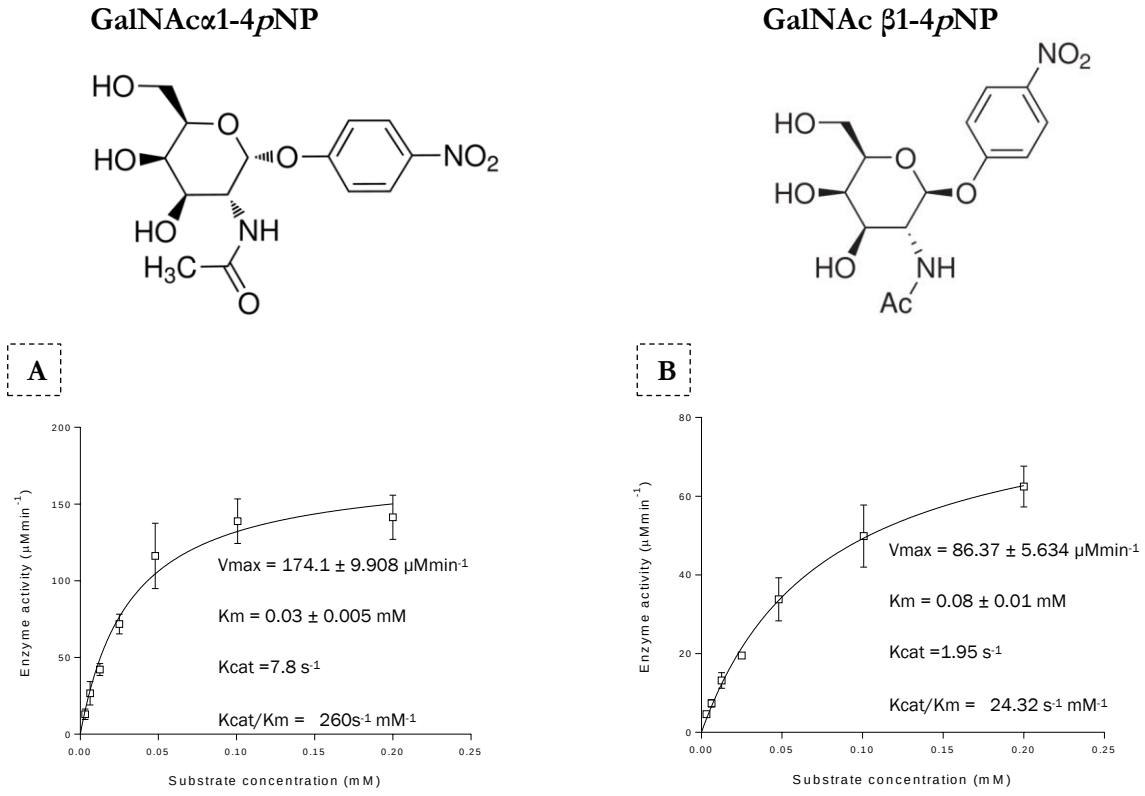
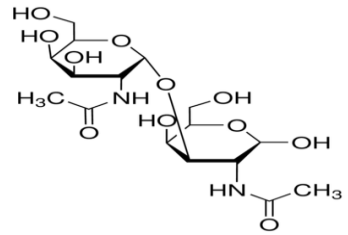


Figure IV.9 - Kinetics of GalNAcα1-4*p*NP and GalNAcβ1-4*p*NP hydrolysis by recombinant BT4243-FL enzyme. **A**: Graph of BT4243-FL enzyme activity versus GalNAcα1-4pNP substrate concentration **B**: Graph of BT4243-FL enzyme activity versus GalNAcβ1-4*p*NP substrate concentration. Kinetic parameters were calculated using data obtained from colorimetric assays (Section II.1.43) involving various *pNP* substrates shown above respective graphs in panels **A** and **B**. Various substrate concentrations were prepared and mixed with the enzyme (0.37 μM for GalNAcα1-4*p*NP and 0.74 μM for GalNAc β1-4*p*NP) in 20 mM Tris/HCl pH 8.0 containing 100 mM NaCl at 37 °C. *pNP* release over time was quantified by absorbance at 420 nm (A420nm) and the data used to calculate enzymatic activity and other kinetic parameters.

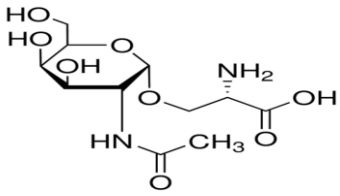
IV.3.1.2.2.2 Synthetic mucin sugar targets of BT4243-FL

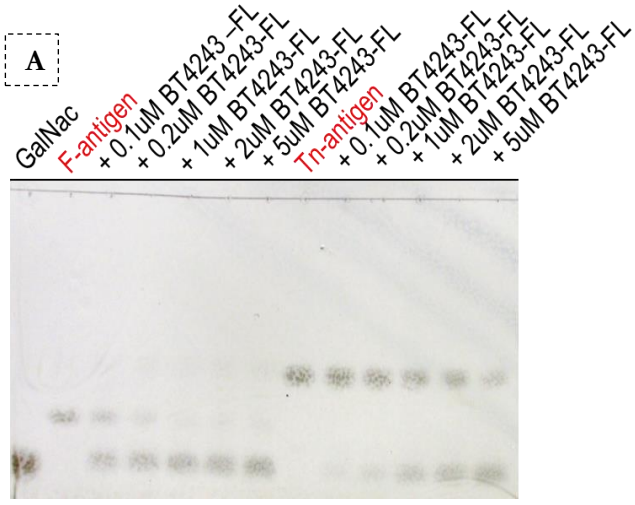
GH109 enzymes have potential applications in medicine as enzymes for the production of universal donor RBCs due to their ability to cleave the GalNAcα1-3Gal linkage present in blood group A antigens (Liu *et al.*, 2007, Olsson *et al.*, 2004). Mass spectrometric analyses of human colonic MUC2 has revealed structures present in the molecule with similar linkages (Larsson *et al.*, 2009). These include the Forsmann disaccharide or F antigen (GalNAcα1-3GalNAc) and Tumor or Tn antigens (GalNAcα1-O-Lserine). The latter is also a component of several mucin O-glycoproteins including BSM and PGM (Section I.2.1.3). As evidenced in Figure IV.10, BT4243-FL was found to be capable of degrading both substrates. Unfortunately, due to the paucity and cost of these substrates, detailed rate experiments could not be performed.

GalNacα1-3GalNac (F antigen)



GalNAcα1-O-L serine (Tn antigen)





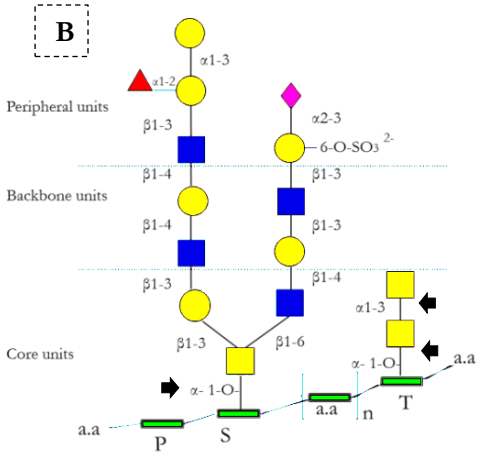


Figure IV.10 – Degradation of F and Tn antigens by recombinant BT4243 (BT4243-FL). **A**: TLC analyses of products released by BT4243-FL following incubation with F and Tn antigens (structures above panels A and B). 1 mg/ml of each sugar substrate was prepared and incubated with various concentrations of the enzyme (as indicated on the TLC image) in 20 mM Tris/HCl pH 8.0 containing 100 mM NaCl at 37 °C for 1 h. 40 µl of each reaction was boiled at 98 °C for 5 min and 6 µl of each reaction applied to silica gel TLC plates alongside a GalNAc standard (4 mg/ml). **B:** Model mucin structure showing the location of bonds (black arrows) targeted by BT4243-FL based on the TLC data in A. See more details of the structure in Section I.2.1.1.1.1.

IV.3.1.2.2.3 Activity of BT4243-FL against native BSM and interactions with other PUL functional partners

Both TLC and HPLC were used to evaluate the activity of BT4243-FL against native mucin substrates and interactions with recombinant versions of its PUL functional partners BT4241-FL and BT4244-FL. The data revealed that BT4243-FL is capable of degrading BSM releasing two products; products 1 and 2, (as shown in the HPLC data of Figure IV.11) one of which was identified as N-acetylgalactosamine (GalNAc) (product 1). The enzyme was also found to be exo-acting as an increase in the release of GalNAc from BSM which contains ~50% sialyl Tn antigen (Neuα2-6GalNAcα1-O-LSer)(Tsuji and Osawa,1986) was observed following pre-treatment of the mucin with a neuraminidase enzyme (Figure IV.11). Access to the Tn antigen of PGM on the other hand is often shielded by fucosyl-galactose (Karlsson *et al.*, 2002) implying substantial GalNAc release from PGM would require additional enzymes such as α-fucosidases and β-galactosidases. Interestingly the latter activity is encoded by its previously characterised PUL partner BT4241, suggesting that both enzymes may work concertedly to degrade their target in the mucin structure (Figure IV.12). This also provided an opportunity to test the effect of their activities (deglycosylation) on the activity of the BT4244 – M60-like protease (Chapter III) using IgA1 as substrate. The data revealed that further deglycosylation of IgA1 after desialylation does not increase BT4244-FL activity (Figure IV.13), an indication that BT4244 may be located upstream of its putative functional partners BT4243 and BT4241 in the PULs metabolic pathway of mucin degradation.

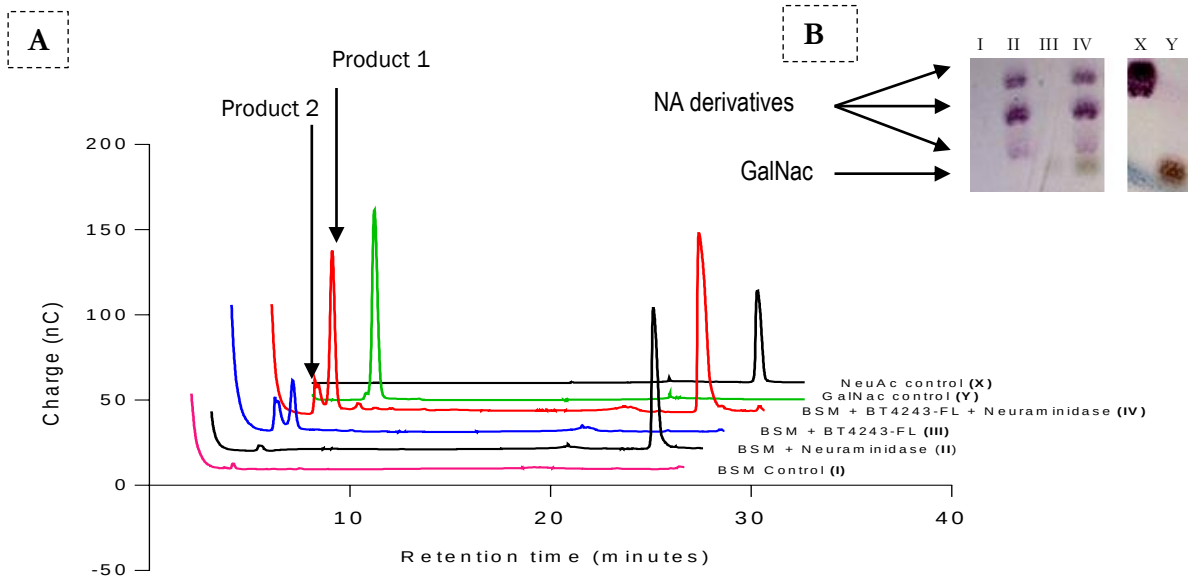


Figure IV.11 – Degradation of BSM by recombinant BT4243-FL and effect of BSM desialylation on BT4243 activity. A: HPLC analyses of products released by BT4243-FL from BSM. Incubation reactions contained 4

mg/ml of BSM, 0.9 µM of BT4243-FL, and 0.1 U/ml of neuraminidase enzyme in 20 mM Tris-HCl pH 7.5 containing 100 mM NaCl. The final volume of each reaction was 500 µl and reactions were carried out at 37°C for ~15 h after which samples were boiled at 98 °C for 3 min and later frozen at -20 °C. Samples were later freeze-dried and concentrated into half the original volume of the reaction before analysing 100 µl of each by HPLC. The concentration of standards included were 1 mM for GalNAc and 0.8 mM for NeuAc. **B.** Example of a TLC profile with samples prepared in a similar manner as in A. The lane numbers/codes shown in B correspond to the those shown in the HPLC data in A. The neuraminidase enzyme used in B however was the α2-3,6,8,9-Neuraminidase from *Arthrobacter ureafaciens* (Calbiochem) while that in A was the Sigma neuraminidase enzyme from Clostridium perfringens (*C. welchii*) (Sigma cat. N2876). In both cases, GalNAc release was increased following prior treatment of the BSM substrate with the neuraminidase enzymes.

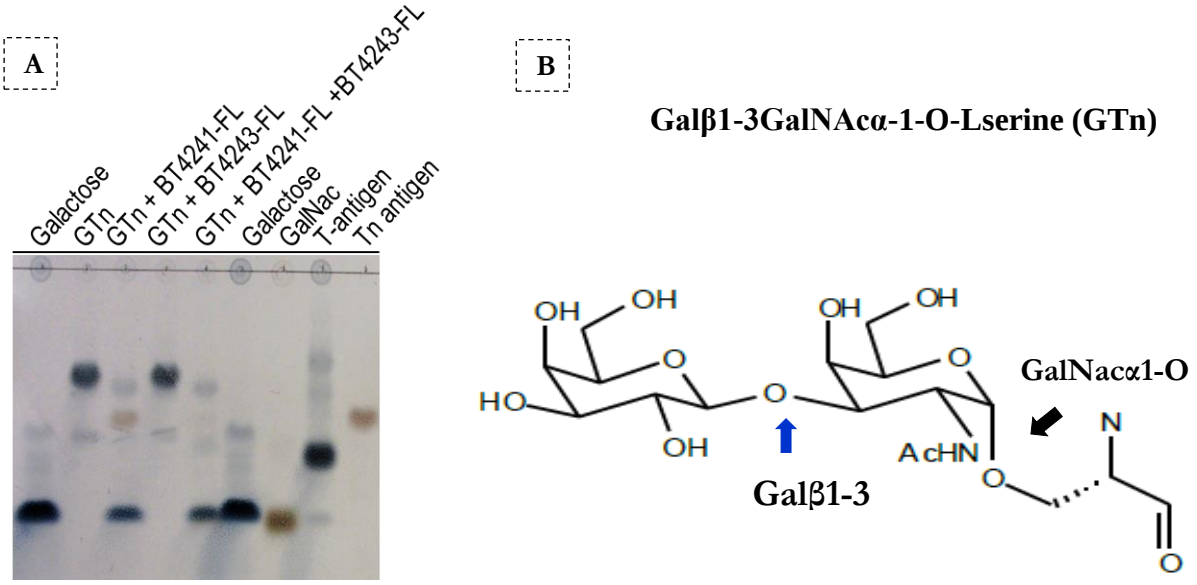
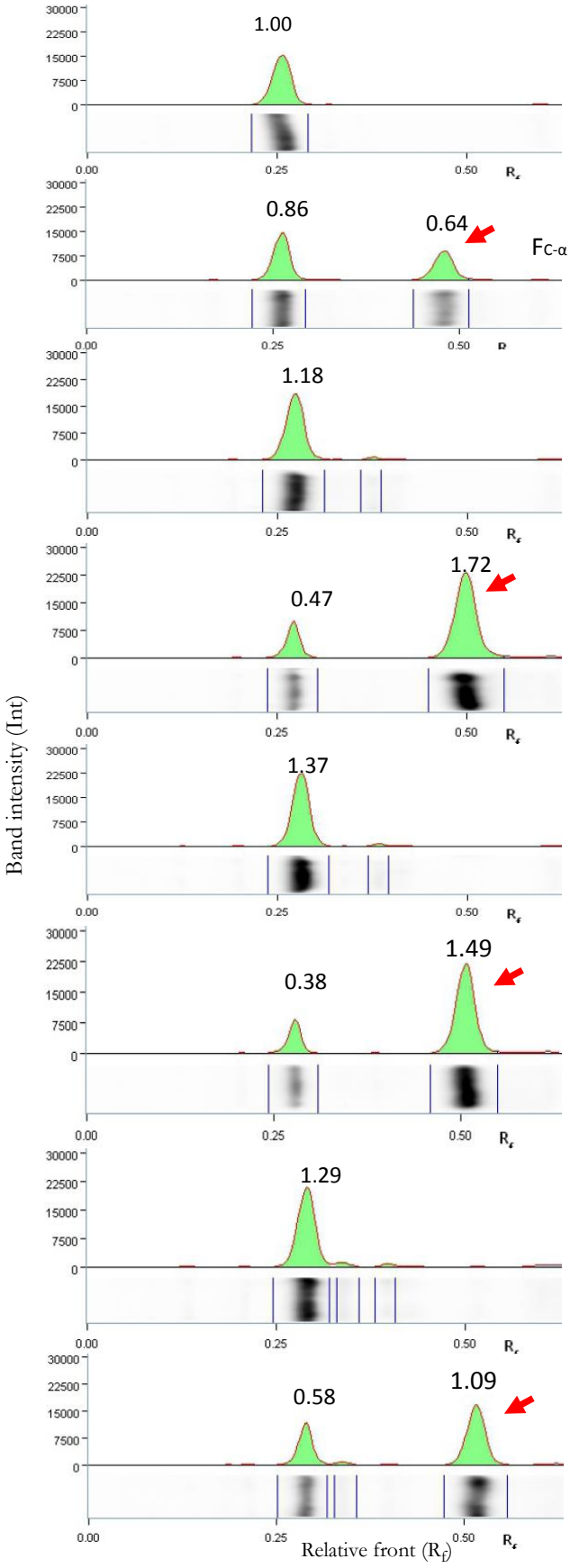
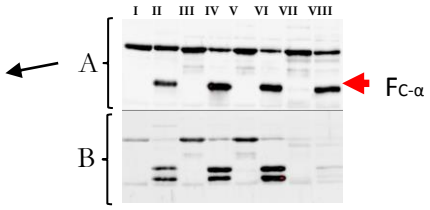


Figure IV.12 - Evidence for cooperation between BT4241-FL and BT4243-FL enzymes in the cleavage of Galactosyl-Tn antigen (GTn). A: Cleavage of the GalNAc-α-1-O-glycosididc bond by BT4243-FL requires prior removal of the capping sugar, galactose by the recombinant BT4241 β-galactosidase. Reactions contained3 mg/ml of GTn, 0.8 µM of BT4241-FL and 1.9 µM BT4243-FL in 20 mM Tris-HCl pH 7.5 containing 100 mM NaCl at 37 °C for 25 h. Samples post incubation were collected and boiled at 98 °C for 2 min before application (6 µl each including standards) to TLC plates. The concentration of the Gal and GalNAc standards were 2 mg/ml and 4 mg/ml respectively. **B:** Structure of Galactosyl-Tn antigen showing linkages targeted by BT4241-FL (blue) and BT4243-FL enzymes (black). The structure of the GTn sugar was obtained from Rougé *et al.*, (2011).





See legend overpage

Figure IV.13 – Effect of BT4241-FL and BT4243-FL mediated deglycosylation of IgA1 on the activity of BT4244-FL. The protocol for this experiment was same as detailed in Section II.1.38. Released Fc fragment was detected by Western blotting and immunochemical detection with Anti-human IgA1 myeloma antibodies (A - top right) and band intensities were quantified using the Image LabTM tool (Biorad) after 5 seconds of exposure to chemiluminescent substrate. Various lanes of the Western blot (I-VIII) are indicated in the different graphs on the left hand side. The deglycosylation of IgA1 at every stage of the reaction was monitored using a biotin - conjugated lectin from *Helix aspersa* (HAA) that binds specifically to GalNAc present in the hinge insertion region of IgA1 (**B** - top right). Complete loss of reactivity with the lectin as seen in lanes VII and VIII was an indication that the hinge insertion region of the IgA1 molecule (also the target site of BT4244-FL) had at that stage been completely deglycosylated by the sialidase, BT4241-FL and BT4243-FL enzyme combination. Red arrows point to the Fc-α product released by the BT4244-FL M60-like protease.

IV.3.1.3 The protein encoded by the BT_4240 gene (BT4240)

IV.3.1.3.1 Features and recombinant protein expression

BT4240 is 362 a.a. long protein containing a single motif annotated as APH standing for aminoglycoside phosphotransferase family (PF01636). Sequence Similarity DataBase (SSDB) and Basic Local Alignment Search Tool (BLAST) searches revealed significant sequence similarities between BT4240 and protein entries from different organisms, most of which were annotated as phosphotransferases, desulfatases and hypothetical proteins. The protein also showed about 39% sequence identity to the *Bifidobacterium longum* Blo_BL1642 or NahK gene product (*lnpB*) which although is annotated as a desulfatase enzyme (Table IV.2), had been shown to rather exhibit sugar kinase activity (Nishimoto and Kitaoka 2007).

Entry	len	Identity
bxy:BXY_33320 Phosphotransferase enzyme family.	362	0.928
bfg:BF638R_0909 putative desulfatase	361	0.928
bfr:BF0928 hypothetical protein	361	0.928
bfs:BF0850 desulfatase	361	0.928
bhl:Bache_0572 aminoglycoside phosphotransferase	360	0.858
pgn:PGN_0702 hypothetical protein	362	0.772
pgt:PGTDC60_1792 hypothetical protein	362	0.764
tva:TVAG_054020 hypothetical protein	359	0.708
bvu:BVU_1819 hypothetical protein	360	0.678
pdi:BDI_1946 hypothetical protein	365	0.660
tfo:BFO_3327 mucin-desulfating sulfatase	362	0.660
pmz:HMPREF0659_A6445 mucin-desulfating sulfatase (EC:3.	363	0.607
pdn:HMPREF9137_2290 mucin-desulfating sulfatase (EC:3.1	363	0.601
bsa:Bacsa_2592 aminoglycoside phosphotransferase	370	0.613
osp:Odosp_0077 aminoglycoside phosphotransferase	410	0.574
pdt:Prede_1778 putative homoserine kinase type II (prot	362	0.582
afd:Alfi_2564 homoserine kinase type II (protein kinase	367	0.476
scn:Solca_2716 putative homoserine kinase type II (prot	381	0.443
shg:Sph21_3292 aminoglycoside phosphotransferase	373	0.452
amu:Amuc_0030 hypothetical protein	364	0.405
cpe:CPE1200 hypothetical protein	368	0.424
cpf:CPF_1410 MdsC protein	368	0.421
csh:Closa_3988 aminoglycoside phosphotransferase	370	0.417
rum:CK1_01160 Phosphotransferase enzyme family.	371	0.396
era:ERE_03150 dTDP-glucose pyrophosphorylase	668	0.408
cpy:Cphy_0570 hypothetical protein	370	0.400
tbe:Trebr_2348 aminoglycoside phosphotransferase	375	0.386
ssm:Spirs_1754 aminoglycoside phosphotransferase	373	0.380
ert:EUR_30720 dTDP-glucose pyrophosphorylase	668	0.406
ere:EUBREC_0139 hypothetical protein	668	0.403
frt:F7308_0778 hypothetical protein	369	0.373
zpr:ZPR_2899 phosphotransferase	371	0.382
blm:BLLJ_1622 phosphotransferase	359	0.392
blb:BBMN68_1674 homoserine kinase type ii	359	0.391
blf:BLIF_1690 phosphotransferase	359	0.391
blg:BIL_03760 Phosphotransferase enzyme family.	359	0.391
blk:BLNIAS_00371 phosphotransferase	359	0.391
bll:BLJ_1682 aminoglycoside phosphotransferase	359	0.391
☑ blo:BL1642 desulfatase	359	0.391
tpi:TREPR_3846 MdsC protein	376	0.375

Table IV.2 - Sequence Similarity DataBase (SSDB) search results for the *B. thetaiotaomicron* BT4240 protein. All identified homologues of BT4240 are listed on the left hand side of the table. Details of their lengths and the levels of similarity (expressed as a fraction of 1) are provided to the right of each entry. The sugar kinase homologue of BT4240 from *B. longum* (lnpB) is besides the checked box towards the bottom of the list i.e. blo : BL1642 (second to the last entry)

The full length BT_4240 gene was amplified from *B. thetaiotaomicron* genomic DNA by PCR (Section II.1.15) using the primers containing engineered *Bam*HI and *Xho*I restriction sites (Appendix Table A.3). The amplified gene was then cloned into the same sites a pET-28a (+) vector (Novagen) and expression carried out in *E. coli* BL21 (DE3) host cells (Section II.1.26). Protein expression was induced with 1 mM IPTG at 16 °C overnight. The N-terminal His_6-

tagged ~45kDa recombinant protein (BT4240-FL) was purified by IMAC as described in Section II.1.26 (Figure IV.14). Protein concentration was routinely measured by absorbance at 280 nm using the estimated molar extinction coefficient (36705 M^{-1} cm^{-1}) of the protein.

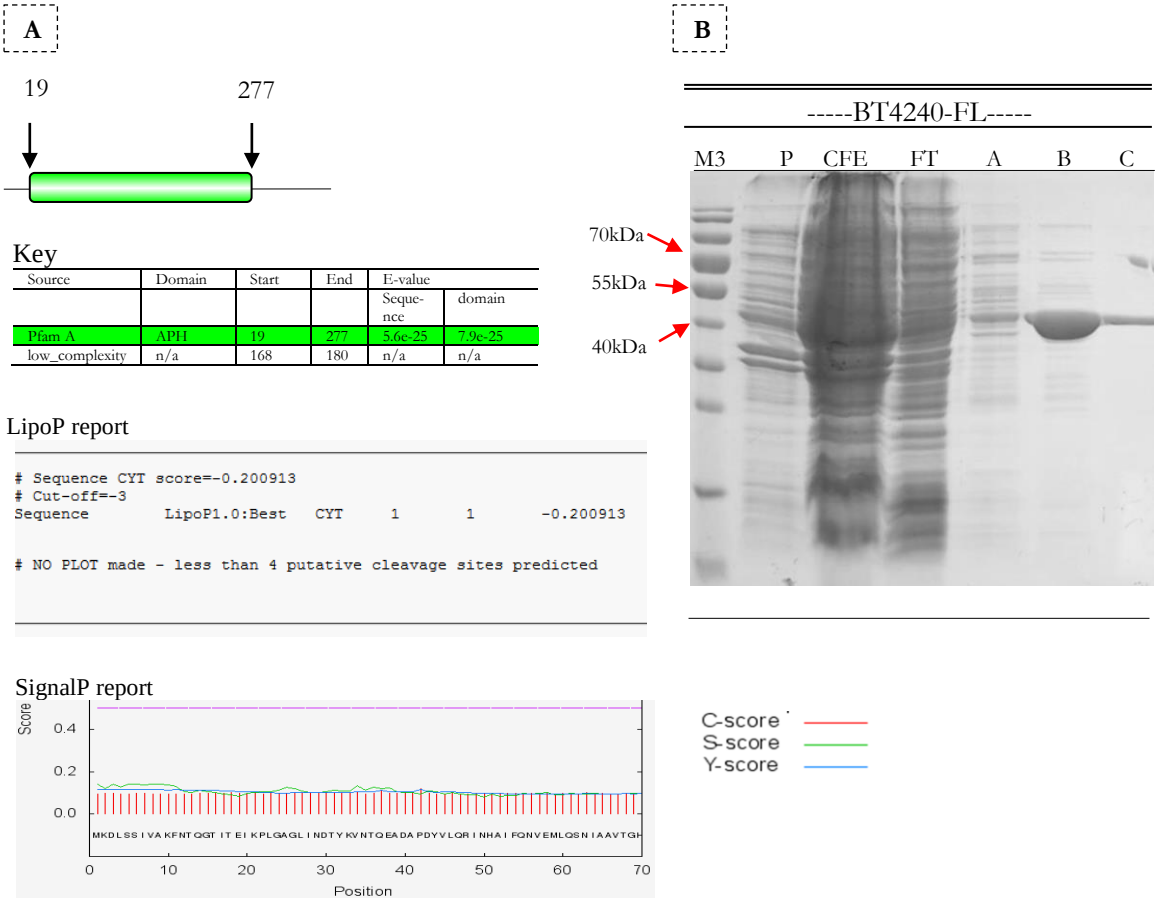


Figure IV.14 - The protein encoded by the BT_4240 gene of *B. thetaiotaomicron*. **A**: Protein features. BT4240 contains a single APH domain. See colour coded key below figure for other details of the APH domain in the protein. E-value is the probability that each sequence or domain is detected by chance. **B:** LipoP and SignalP reports show no evidence of a signal peptide sequence. Please follow link in Section II.2 for more information on the interpretation of other score values. **C:** SDS PAGE analyses of recombinant BT4240 (BT4240-FL) purified by IMAC**.** Protein expression and purification was carried out as described in Section II.1.26. BT4240-FL was purified from 100ml of *E. coli* BL21 (DE3) culture volume. 5 µl of the insoluble pellet fraction (P) re-suspended in 10ml of Talon buffer (20 mM Tris/HCl pH 8.0 containing 100 mM NaCl), 10ul of cell free extract (CFE), 10 µl of flow through (FT), 10 µl of the fraction eluted with 10 mM imidazole (A), 10 µl of fractions sequentially eluted with 100 mM imidazole (B and C) were analysed by SDS PAGE, using a 12.5% polyacrylamide gel. Also see Section II.1.27 for the molecular weights of various markers in lane M3. The theoretical molecular weight of the recombinant N-terminal His_6-tagged protein is ~45 kDa

IV.3.1.3.2 Phosphorylation substrates of BT4240

The coupled enzyme ATPase assay procedure (described in Section II.1.45 and in Reith *et al.*, 2011) was used to screen for potential phosphorylation substrates of BT4240. ATP was used as a phosphoryl donor and substrate phosphorylation was recorded as a continuous fall in the absorbance of the reaction mixture at 340nm (A340nm) due to conversion of NADH to NAD+ after addition of the recombinant BT4240-FL enzyme (Figure IV.15B). BT4240 was found to be capable of phosphorylating the mucin amino sugars N-acetylgalactosamine (GalNAc) and N-acetylglucosamine (GlcNAc) (Figure IV.15B).

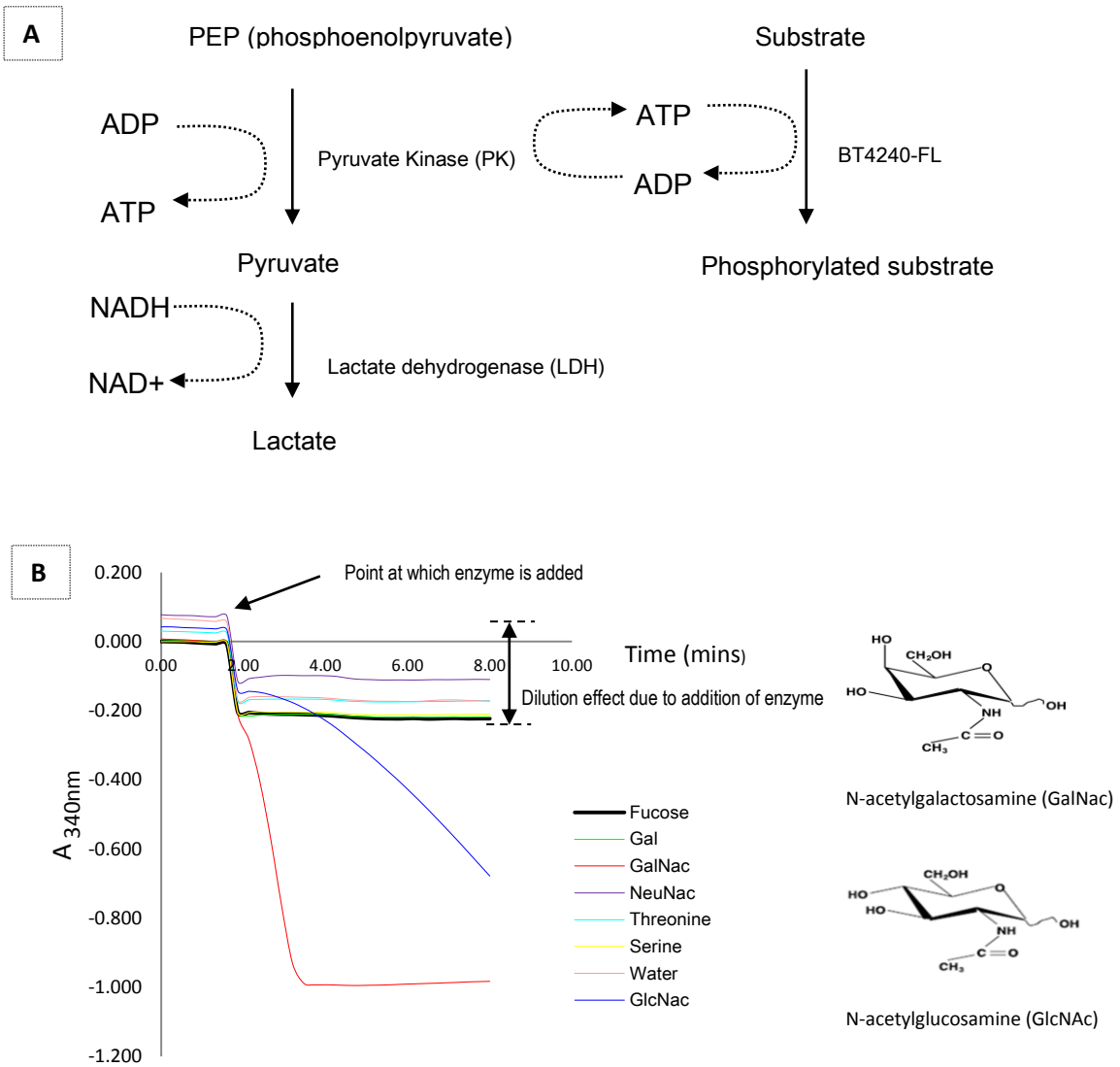


Figure IV.15 – Screening for potential phosphorylation substrates of BT4240. A: Scheme of coupled enzyme ATPase assay used in the screening process (Reith *et al.*, 2011). **B:** Change in A340nm over time following addition of BT4240 – FL to reactions containing various substrates tested. Each substrate (sugar/amino acid) was added to a final concentration of 4 mM in a standard reaction mixture containing 1 mM PEP, 0.2 mM NADH, 5

mM ATP, 10 mM $MgCl_2$, 10 U/ml proteinase K, 8 U/ml LDH, 0.54 µM of BT4240-FL and 100 mM Tris-HCl pH 8.0. Reactions were all carried out at 37 °C.

IV.3.1.3.3 Thin layer chromatography and enzyme kinetics

Phosphorylation of GalNAc by BT4240 was also confirmed by thin layer chromatography, allowing for the visualisation of the phosphorylated product (GalNAc-x-phosphate). The kinetics of the phosphorylation reaction was also studied using the coupled enzyme ATPase assay procedure discussed in Section IV.3.1.3.2 above. The Km of GlcNAc phosphorylation by BT4240-FL was ~30 times higher than for GalNAc

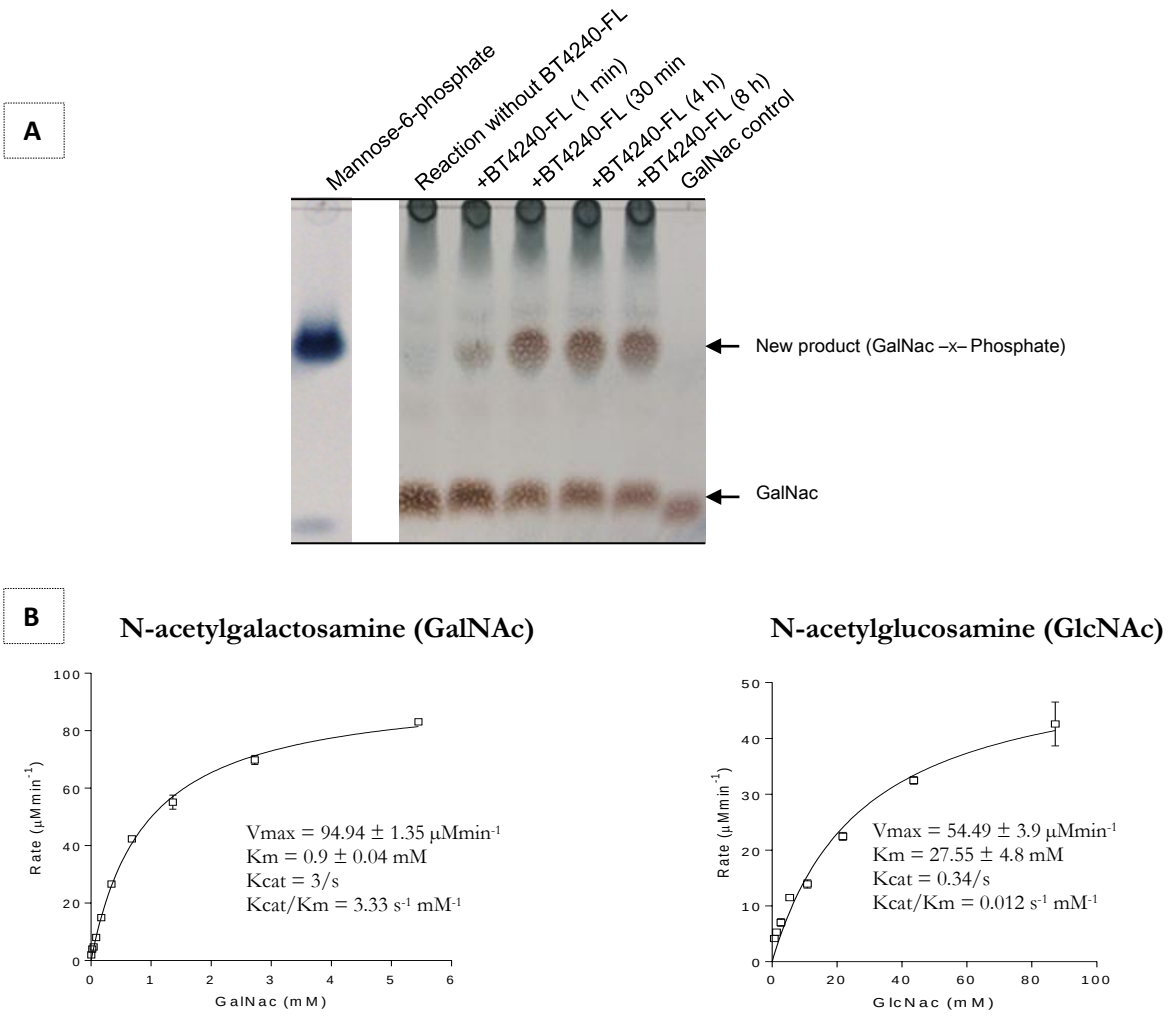


Figure IV.16 - Amino sugar kinase activity of recombinant BT4240. A: Time course phosphorylation of GalNAc by BT4240-FL analysed by TLC. 50 mM of GalNAc was treated with 0.2 mg/ml (~4.4 µM) of recombinant BT4240-FL plus 10 mM $MgCl_2$ and 0.5 M ATP/$NADP^+$ in 0.1 M Tris-HCl pH 8.0, at 37 °C for the various time periods indicated. Samples post-incubation were boiled at 98 °C for 2 min before application (6

µl/sample) to TLC plates alongside the GalNAc standard (4 mg/ml). A mannose phosphate standard was rather used for comparison due to scarcity of GalNAc phosphate. **B:** Kinetics of GalNAc and GlcNAc phosphorylation by BT4240-FL. Reactions were set-up as in Figure IV.15 and rates were measured at different substrate concentrations from the A340nm data

IV.3.1.3.4 Position of BT4240 in the mucin degradation pathway

The aim of this section was to determine the position of the BT4240 enzyme in the mucin degradation pathway of the BT4240-50 Sus-like system in relation to its putative functional partner BT4243 by investigating when GalNAc phosphorylation occurs. The set-up of the experiment, which is a modification of the set-up used to investigate the kinetic properties of BT4240 (Section IV.3.1.3.3) is shown in Figure IV.17, using the Tn antigen as substrate. Phosphorylation of GalNAc, recorded as a fall in A340nm mainly occurred following release of the sugar from the Tn antigen, implying BT4240 is likely upstream of BT4243 in the mucin degradation pathway.

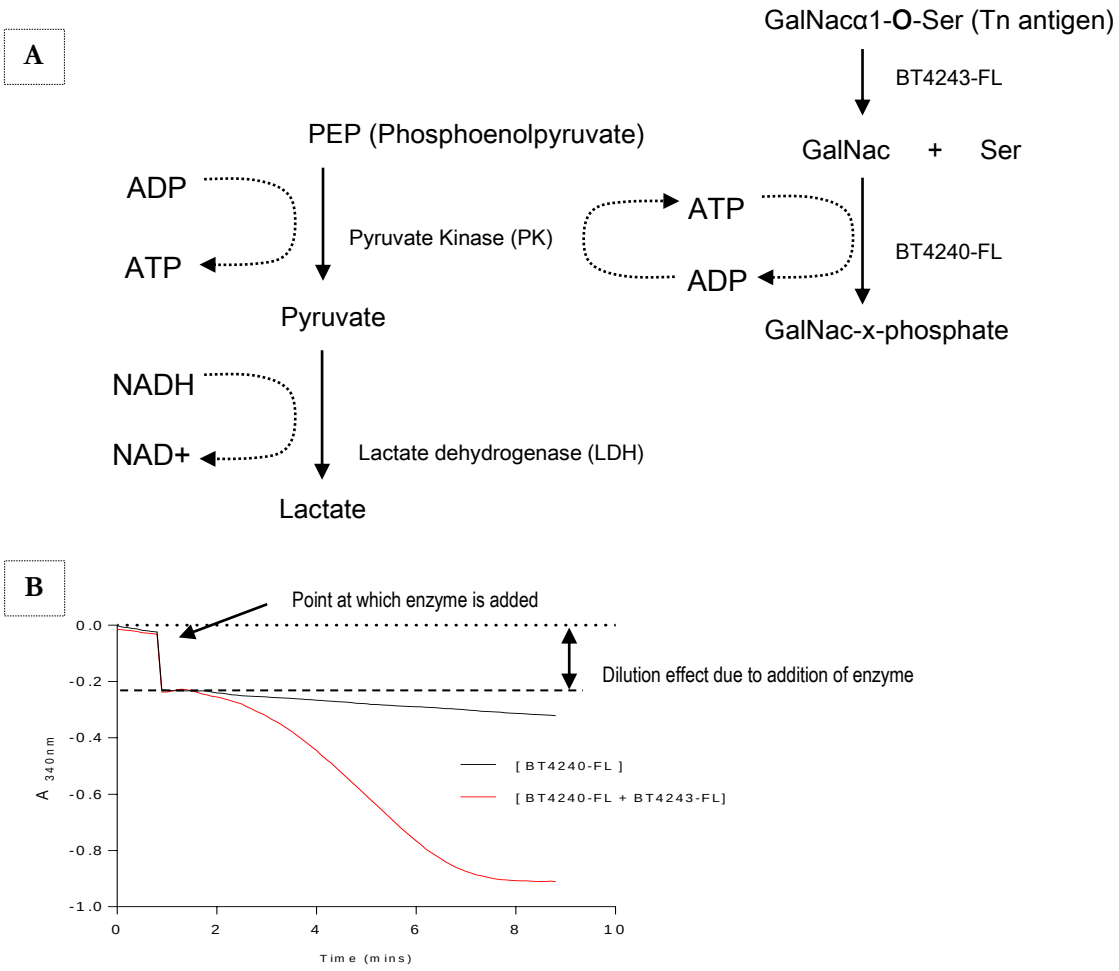


Figure IV.17 – Phosphorylation of GalNAc by BT4240-FL mainly occurs after release of the sugar from the mucin glycoprotein. A: Scheme to test the interaction of BT4240-FL and BT4243-FL during the

utilisation of the Tn antigen**. B:** Change in A340nm over time following phosphorylation of Tn antigen by BT4240-FL in the presence and absence of BT4243-FL. The observed steep fall in the A340nm over time in the presence of both enzymes as opposed to BT4240 alone was an indication that BT4240-FL preferentially phosphorylates the free GalNAc sugar released by BT4243-FL. The slight fall in A280nm is likely to be very negligible as up to 8 µM of the BT4240 enzyme was used in the reaction. The concentrations of the Tn antigen and BT4243 enzymes were 50 mM and 0.9 µM respectively.

IV.3.1.4 The protein encoded by the BT_4245 gene (BT4245)

IV.3.1.4.1 Features and recombinant protein expression

Unlike BT4243, BT4245 contains a type II signal peptide and two other domains including a C-terminal F5/8 type C or CBM32 domain [also found in BT_4244 (Chapter III)] and an N-terminal domain of unknown function (DUF1735) (Figure IV.18 and IV.19). BT4245 is also unique in that it shows only about 30% sequence identity to its closest homologues, compared to BT4240, BT4243, and BT4241 which show at least 70% identity to their closest homologues. The CBM32 sequence of the protein shows some similarities to the C. *perfringens* CpGH89CBM32-5 and *B. thetaiotaomicron* M60-like CBM32 sequences previously characterised in Chapter III. This includes conservation of the GalNAc interacting residues; Histidine-1392 (H1392), Arginine-1423 (R1423), Asparagine–1428 (N1428), Phenylalanine-1483 (F1483) of CpGH89CBM32-5.

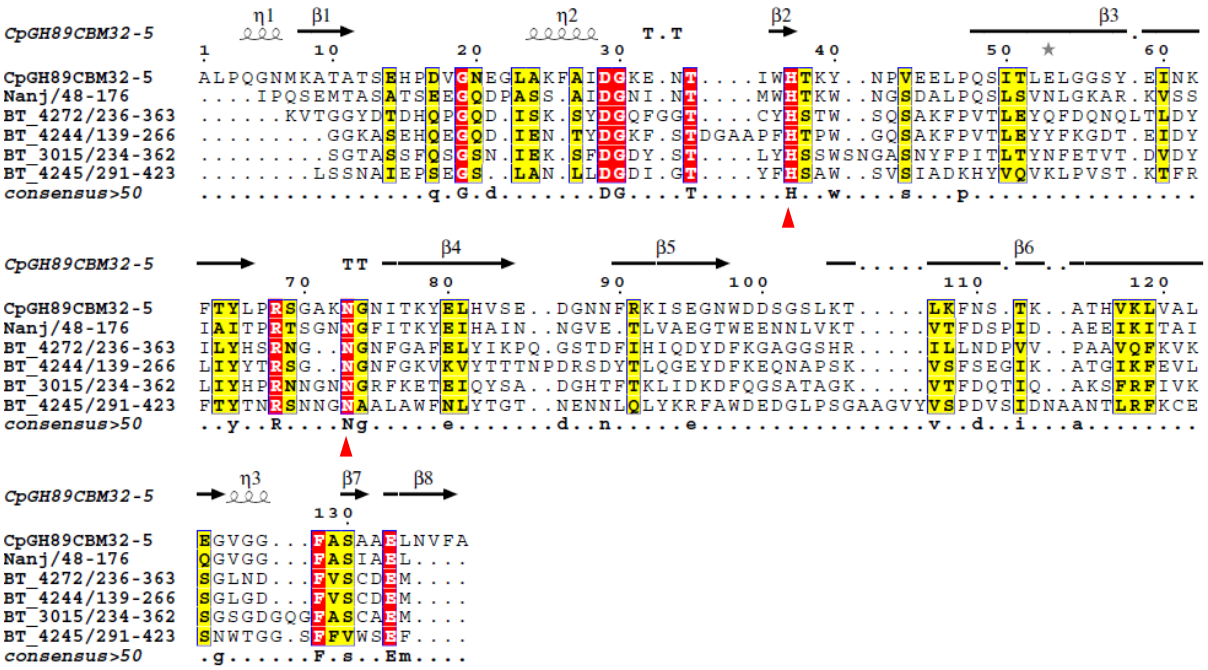


Figure IV.18 –Comparing the BT4245 CBM32 sequence with the GalNAc binding BT4244, BT3015 BT4272, CpGH89CBM32-5 (PDB: 4AAX) and NanJCBM32 (NanJCBM32: PDB 2v72)

CBM32 sequences (described in Section III.3.9). The CBM32 sequence of BT4245 possesses highly conserved residues (indicated with red inverted triangles) earlier shown to be important for GalNAc recognition in the BT4244 –FL M60-like protease (Section III.3.9). The secondary structure of CpGH89CBM32-5 (PDB: 4AAX, Ficko-blean *et al.*, 2012) is shown above the alignments. Sequence alignments were viewed using the ESPript 2.2 utility at http://espript.ibcp.fr/ESPript/ESPript/ (Gouet *et al.*, 1999) with a global similarity score threshold of 0.7. Red highlights are for amino acid residues showing 100% conservation while yellow highlights are for residues showing less than 100% conservation but above the global score threshold.

The full length BT_4245 gene excluding the region encoding the predicted type II signal peptide sequence was amplified from *B. thetaiotaomicron* genomic DNA by PCR (Section II.1.15) using the primers containing engineered *Nco*I and *Xho*I restriction sites (Appendix Table A.3). The amplified gene was then cloned into the same sites of the pET-28a (+) vector (Novagen) and expression carried in *E. coli* BL21 (DE3) host cells. Protein expression was induced with 1 mM IPTG at 16 °C overnight. The C-terminal His_6-tagged ~46 kDa recombinant protein (BT4245-FL) was purified by IMAC as described in Section II.1.26 (Figure IV.19). BT4245 expression from 100 ml of *E. coli* BL21 (DE3) host cells was very poor (Figure IV.19C) compared to BT4240 for example (Figure IV.14), with very high amounts of contaminating background proteins. Attempts to express the CBM32 domain alone failed. The protein was thus routinely expressed in very large volumes ~5 L *E. coli* BL21 (DE3) cells for experimental work. Very high levels of purity could also be achieved following a combination of IMAC and anion exchange chromatography (AEC). Protein concentration was measured by absorbance at 280 nm (A280nm) using the estimated molar extinction coefficient (60070 M^{-1} cm^{-1}) of the protein.

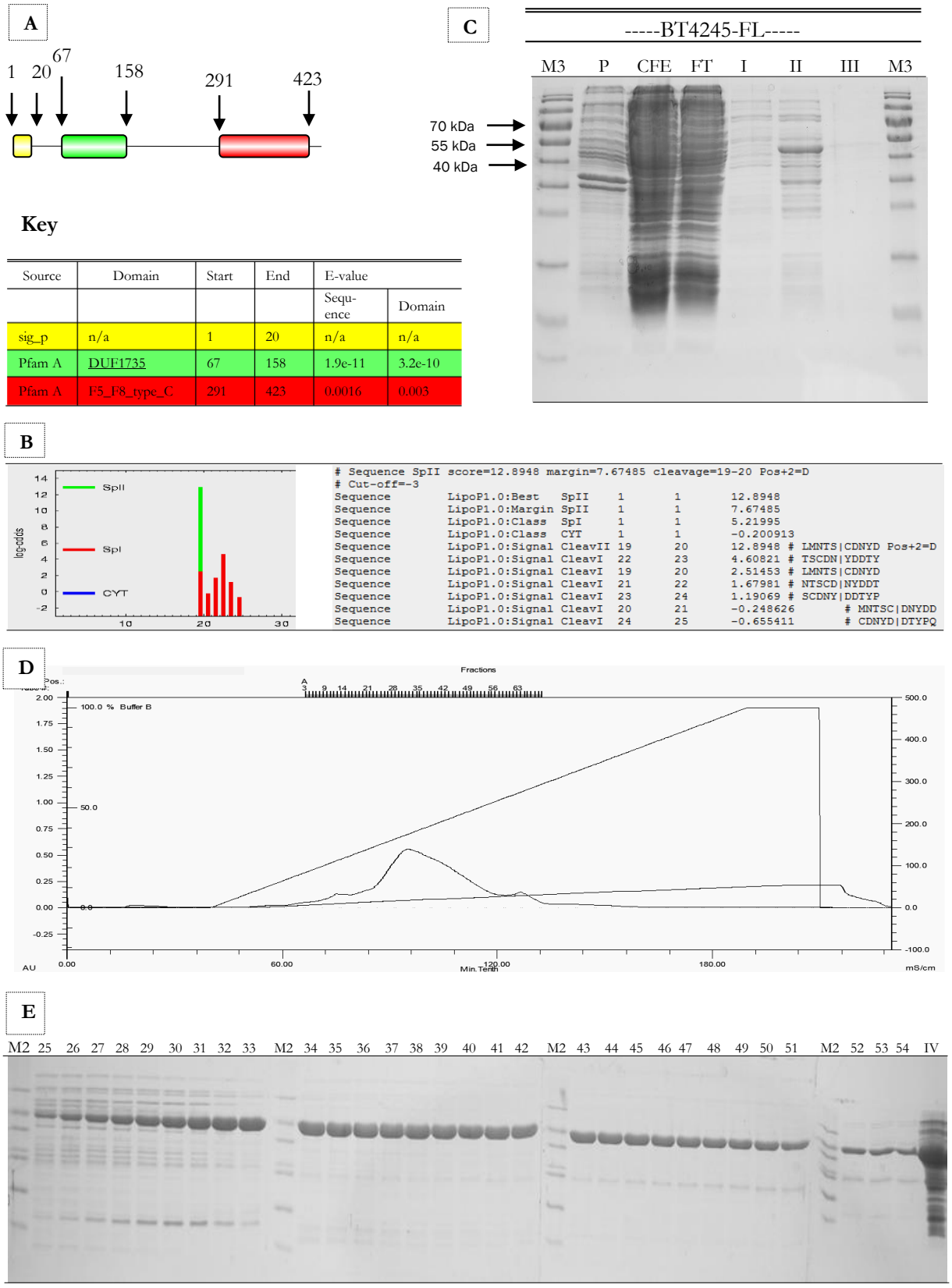


Figure IV.19 - The protein encoded by the BT_4245 gene of *B. thetaiotaomicron*. **A:** Protein features. The BT4245 protein contains three distinct features including an N-terminal type II signal peptide

sequence, an F5_F8_type_C domain (CBM32) and a domain of unknown function (DUF). See colour coded key below figure for other details of various domains in the protein. E-value is the probability that each sequence or domain is detected by chance **B:** LipoP analyses showing high probability of a type II signal peptide sequence in BT4245. Please follow link in Section II.2 for more information on the interpretation of other score values. **C:** Purification of recombinant BT4245 (BT4245-FL) by IMAC. Protein expression and purification was carried out as described in Section II.1.26. BT4245-FL was purified from 100 ml of *E. coli* BL21 (DE3) culture volume. 5 µl of the insoluble pellet fraction (P) re-suspended in 10 ml of Talon buffer (20 mM Tris/HCl pH 8.0 containing 100 mM NaCl), 10 µl of cell free extract (CFE), 10 µl of flow through (FT), 10 µl of the fraction eluted with 10 mM imidazole (A), 10 µl of fractions sequentially eluted with 100 mM imidazole (II and III) were analysed by SDS PAGE, using a 12.5% polyacrylamide gel. **D:** Example of an anion exchange chromatogram for the purification of BT4245-FL. Following purification of proteins from ~ 10 L of culture by IMAC, fractions corresponding to II and III (Figure IV.19C) were pooled and concentrated by centrifugation into a 4ml volume (~ 100 µM total proteins) before analyses anion exchange chromatography (AEC). Elution of protein during AEC was carried out using 500 mM NaCl in 10 mM Tris, pH 8.0.**E:** SDS PAGE analyses of the proteins purified by anion exchange chromatography (AEC). 10 µl each from 1ml fractions collected after AEC was boiled in 5 µl of SB buffer before application to 12% SDS PAGE gels. Sample IV is the pooled II and III fractions from IMAC before AEC purification while samples 25-55 came from fractions after AEC purification. Also see Section II.1.27 for the molecular weights of various markers in lanes M2 and M3 in panels C and E respectively. The theoretical molecular weight of the recombinant N-terminal His_6-tagged protein is ~46 kDa

IV.3.1.4.2 Carbohydrate binding properties of BT4245

Isothermal titration calorimetry (ITC, section II) was used to study the carbohydrate binding properties of recombinant BT4245-FL. GalNAc containing sugars were initially screened as potential ligands for the protein owing to similarities between the CBM32 sequence of BT4245 and other GalNAc binding proteins (Figure IV.18). BT4245 is also induced during *B. thetaiotaomicron* growth on the T antigen (Martens *et al.*, 2008) which was earlier identified as a target for its PUL member BT4241. It was found to be capable of saturating Gal and GalNAc monosacharides as well as their disaccharide sugars T (Galβ1-3GalNAc) and F (GalNAcα1-3GalNAc) antigens in ITC experiments (Figure IV.20). BT4245 just like the CBM32 domains of the *B. thetaiotaomicron* M60-like proteins (Chapter III) also showed greater preference for the GalNAc monosaccharide (Table IV.3) although this was found to be ~5 times higher than observed for the latter group of proteins. BT4245 did not bind other mucin sugars such as GlcNAc, NeuAc, and Fuc or the commercial mucins PGMII / III and BSM. Although BT4245 was not recalcitrant to crystallisation, the few crystals obtained failed to diffract to a useful resolution for structural studies.

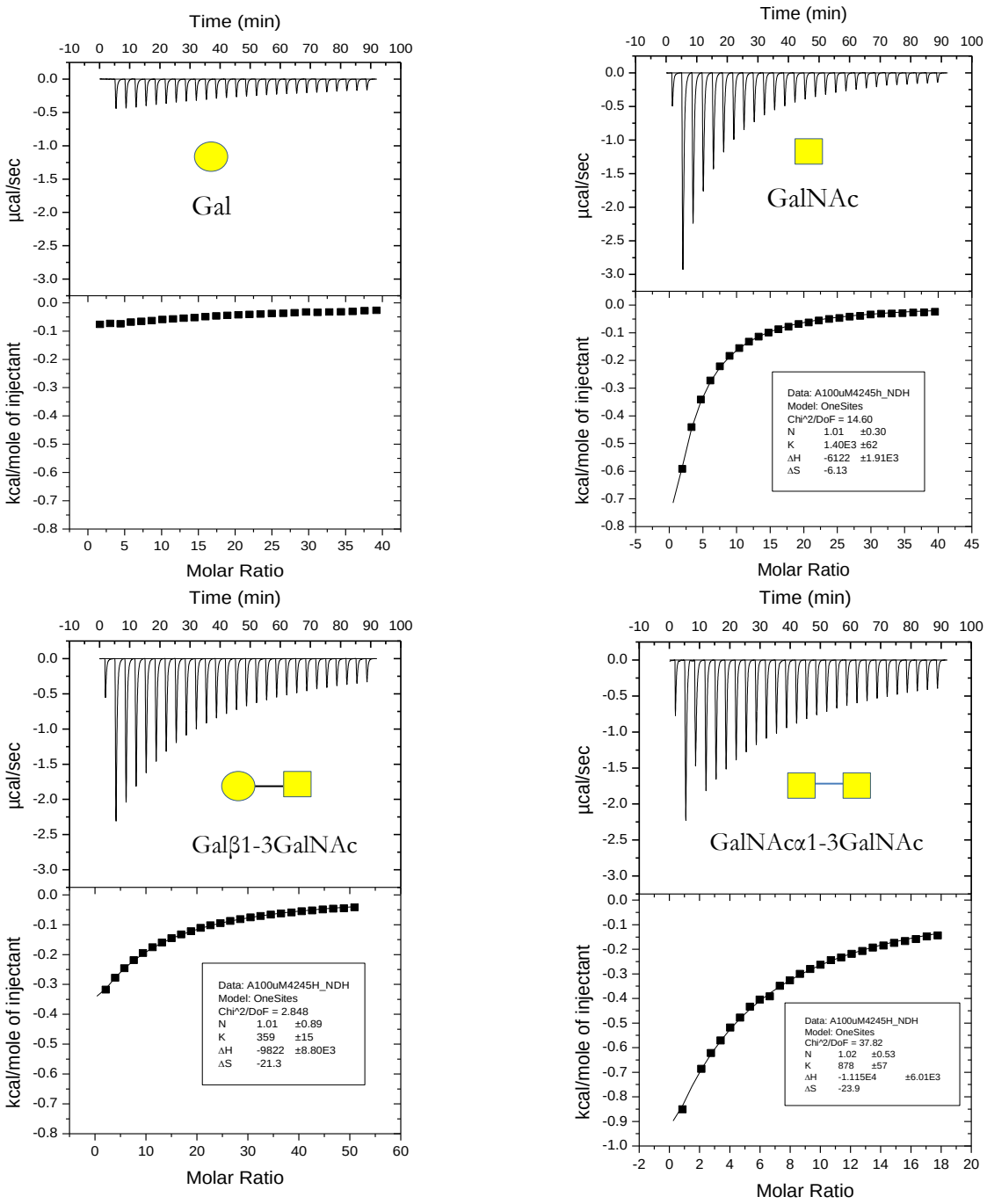


Figure IV.20 - Representative ITC data for BT4545-CBM32 binding to selected mucin sugars. Each ligand [20 mM Galactose (Gal), 20 mM GalNAc (N-acetylgalactosamine), 26 mM Galβ1-3GalNAc (T antigen) and 9 mM GalNAcα1-3GalNAc (F antigen)] was titrated (27 injections) into 100 µM of the recombinant protein BT4244-CBM32 in Tris (20 mM Tris – HCl pH 8.0) or Talon buffer (20 mM Tris/HCl pH 8.0 plus 100 mM NaCl) at 25 ºC using a MicroCal™ VP-Isothermal Titration Calorimeter as described in Section II.1.32. ITC data obtained was analysed using Origin, version 7.0. The top half of each panel shows the raw power data while the bottom half

are integrated peak areas fitted to a single-site binding model and stoichiometry fixed at 1 (n≈1). Note: Only a single ITC run was performed with GalNAcα1-3GalNAc due to the paucity of the substrate.

Ligand	$K_a \times 10^3$ (M^{-1})	ΔG (kcal mol^{-1})	ΔH (kcal mol^{-1})	$T\Delta S$ (kcal mol^{-1})	n
GalNAc	1.40 ± 62	-4.27	-6.12 ± 1.91e^3	-1.85	~1.01
Galβ1-3GalNAc	0.36 ±15	-3.47	-9.82 ± 8.8e^3	-6.35	~1.01
GalNAcα1-3GalNAc	0.88	-4.0	-11.15 ± 6.01e^3	-7.2	~1.02

Table IV.3 - Affinity and thermodynamic parameters of BT4545-CBM32 binding various substrates analysed in Figure IV.20. Thermodynamic parameters were calculated as described in Section II.1.32 using the standard thermodynamic equation -RTlnK_a = **Δ**G = **Δ**H-T**Δ**S, where R = gas constant (1.99 cal.K^{-1}.mol^{-1}), T = temperature in Kelvin (298.15 K), **Δ***G* = change in free enthalpy, **Δ**S = entropy of binding. n = stoichiometry of binding. Thermodynamic paramters could not be estimated for the rection invovling galactose due to poor saturation of the protein by the sugar.

IV.4 Discussion

IV.4.1. Functional units within the BT4240-50 Sus-like system

The *in-silico* and biochemical data accumulated from the characterisation of the functional partners of BT4244 revealed vital information that was used to piece together a possible functional mechanism of the BT4240-50 Sus-like system, as well as hypothesize the role of the protease as a member of the system. Based on the data, BT4240-50 consists of three important machineries, the carbohydrate binding / transport, enzymatic, and sensory/regulatory machineries.

IV.4.1.1 The carbohydrate binding / transport machinery

Biochemical evidence from Chapters III and IV suggest that the carbohydrate binding and transport machinery of BT4240-50 is partly contributed by the GalNAc – binding family 32 carbohydrate binding modules (CBM32) of the BT4244 and BT4245 proteins. Although the binding affinity of the BT4245 protein was 5 times higher than its BT4244 counterpart, the binding affinities of both proteins were in the low affinity range (Ka ~10^3 M^{-1}) which is typical of this family of carbohydrate binding modules [Type C CBMs (Boraston *et al.*, 2004, Ficko blean 2006, 2009, 2012). BT4245 likely employs a similar GalNAc recognition mechanism like the CBM32 domains of the *B. thetaiotaomicron* M60-like proteases and the *C. perfringens* CpGH89CBM32-5 (PDB: 4AAX) and NanJCBM32 (NanJCBM32: PDB 2v72) proteins based on the conservation of GalNAc interacting residues (Chapter III.3.9). Theoretically, the carbohydrate binding / transport machinery also includes the BT4246, putative SusD protein. Although BT4246 was successfully expressed in this study, it failed to bind with any measurable affinity to any of the mucin sugars it was tested against. These included GalNAc, F antigen and even the T antigen which had been shown to induce BT4246 expression (Martens *et al.*, 2008). Typically a SusD protein interacts with a SusC protein to bind and channel captured glycans into the cell (Cho and Salyers 2001). The putative SusC protein in this case is BT4247. The BT4242 protein (296.a.a) also exhibits features of a typical multi-pass transmembrane transporter protein (Section I.4, Figure I.12C) as evidenced in the Phobius prediction below (Figure IV.21). In summary the proteins BT4242, BT4244, BT4245, BT4246 and BT4247 all constitute the sugar binding and transport machinery of the BT4240-50 Sus-like system (Figure IV.22). There is likelihood however, of other possibly unidentified or uncharacterised sugar binding domains that

exist within other PUL proteins e.g the putative sugar binding domain of the BT4241 β-galactosidase (Section IV.3.1.1).

A

SEQENCE ID	TM	SP	PREDICTION
BT4240	0	0	i
BT4241	0	Y	n8-18c23/24o
BT4242	6	0	i20-39o54-79i86-106o118-139i160-180o186-204i
BT4243	0	Y	n4-17c22/23o
BT4244	0	Y	n6-17c22/23o
BT4245	0	Y	n4-15c20/21o
BT4246	0	Y	n10-19c24/25o
BT4247	0	Y	n17-28c32/33o
BT4248	0	0	i
BT4249	0	0	i
BT4250	0	0	i

B

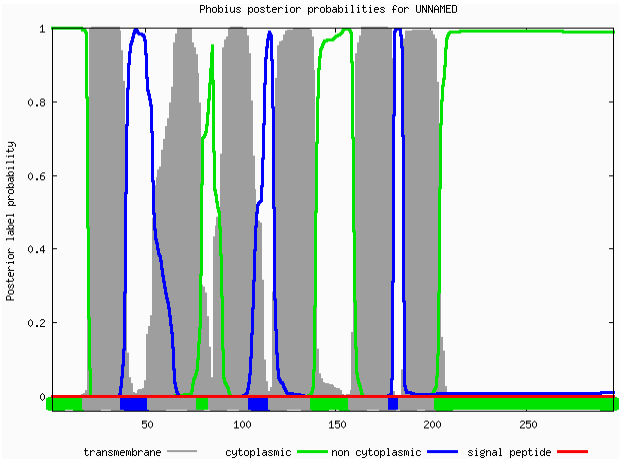


Figure IV.21 – Predicting the cellular localisation of BT4240-50 entries. **A:** Summarised /short prediction for all PUL entries. TM: Transmembrane domain, SP: signal peptide Y: signal peptide detected 0: no signal peptide detected. **B:** Graphical representation of BT4242 protein topology and interaction with the cell. Analyses were performed using the combined transmembrane topology and signal peptide predictor tool, Phobius (Kall *et al.*, 2004).

IV.4.1.2 The enzymatic machinery

Recombinant forms of the following proteins, BT4240, BT4241, BT4243 and BT4244 were all shown during the course of this study to exhibit enzymatic activity. The BT4244 M60-like protein was identified in Chapter III as a mucin protease. Data from the current chapter suggest that BT4240 is an amino sugar (GalNAc and GlcNAc) kinase while BT4241 and BT4243 are T antigen/LNB targeted β-galactosidase and Tn/F antigen targeted glycoside hydrolases respectively.

To our knowledge, BT4240 is the first GalNAc kinase to be biochemically characterised from the gut microbe *B. thetaiotaomicron*. Its functionally characterised homologue, the NahK N-acetylhexosamine 1-kinase encoded by the BL1642/*lnpB* gene of Bifidobacterium *longum* JCM1217 is also capable of catalysing the phosphorylation of GalNAc and GlcNAc phosphates (Nishimoto and Kitaoka 2007). Comparing both enzymes, the *B. thetaiotaomicron* kinase phosphorylates GalNAc at a much higher rate (~30 times lower Km) than GlcNAc while NahK phosphorylates both substrates at similar rates. Although the specific site of phosphorylation of the GalNAc sugar by BT4240 was not confirmed in the current study, it is very likely that this

event occurs at the reducing end carbon -1 (C-1) as observed with NahK given that GalNAc phosphorylation only occurred following release from the Tn antigen where this position is blocked. This still however needs to be investigated.

BT4241 happens to exhibit a rare specificity as only a few enzymes including a *Bifidobacterium bifidum* DSM 20082/*longum* Lacto-*N*-biose phosphorylase (LNBP encoded by BL1641 in *B. longum*) and a *Clostridium perfringens* endo-β-galactosidase (Endo-β-Gal_{GnGa}) are known to target the Galβ1-3GalNAc linkage (Derensy-Dron *et al.*, 1999, Kitaoka *et al.*, 2005, Nishimoto and Kitaoka, 2007, Ashida *et al.*, 2001). Endo-β-Gal_{GnGa} unlike BT4241 is endo-acting and is capable of releasing GlcNAcα1-4Gal from the PGM sugar GlcNAcα1-4Galβ1-3GalNAcα1-Ser/Thr and the oligosaccharide GlcNAcα1-4Galβ1-4GlcNAcβ1-6(GlcNAcα1-4Galβ1-3) GalNAc while LNBP shares the same linkage specificities with the BT4241 enzyme.

The BT4243 protein exhibited properties of an exo-acting α-N-acetylgalactosaminidase capable of releasing GalNAc from synthetic and natural mucin substrates (Section III.1.2). Its ability to target both the F and Tn antigens implies that BT4240-50 maybe involved in the metabolism of both sugars. While core-1(Galβ1-3GalNAc) has been shown to upregulate several components of the PUL (Martens *et al.*, 2008), it is not yet known whether the F antigen (GalNAcα1-3GalNAc) is equally capable of doing so and hence it is important to verify the physiological significance of the observed activity against the F antigen. α-N-acetylgalactosaminidases targeting the O-glycosidic bond of the Tn antigen enzymes have also been reported in the enteric species *Bifidobacterium bifidum/longum* and *Clostridium perfringens* (Kiyohara *et al.*, 2011, Ashida *et al.*, 2008). Just like the *E. meningosepticum* NagA protein, BT4243 may represent a potential candidate for the production of universal red blood cells (Liu *et al.*, 2007, Olsson *et al.*, 2004).

IV.4.1.3 Sensory/regulatory machinery

PULBT_4240-50 like many other mucin O-glycan targeted PULS are regulated by ECF-σ/anti-σ systems (extra-cytoplasmic function sigma/anti sigma factor systems) (Martens *et al.*, 2008, 2009a). These systems generally operate by a mechanism termed trans-envelope signalling analogous to the ferric citrate uptake (Fec) system in *E.coli* (Mahren *et al.*, 2002, Koebnik *et al.*, 2005). In this sensory /regulatory system, an incoming sugar or signal, in this case probably core-1 or the Fn antigen is detected by an outer membrane TonB-dependent transporter protein (in this case the BT4247 putative SusC protein) which triggers the activation of a cytoplasmic ECF-σ transcription factor (in this case the BT4250 protein) through its interactions with a

periplasmic/cytoplasmic anti-σ factor protein (in this case contributed by the product of the intact BT_4248/49 gene) leading to induction of gene transcription.

IV.4.2 Mechanism of the BT4240-50 Sus-like system based on the Sus prototype

IgA1, BSM, PGM and other mucin derived sugars including the T, F antigens and GalNAc were all identified as targets for one or more BT4240-50 proteins. Based on the evidence from the biochemical characterisation of various BT4240-50 components in this Chapter and the previous (Chapter III), is can be conclude that this Sus-like system is targeted at mucin O-glycoprotein structures containing the Forssmann disaccharide or F antigen (GalNAcα1-3GalNAc) and Core-1 or T antigen (Galβ1-3GalNAc) common in intestinal mucins (Robbe *et al.,* 2004, Larsson *et al.,* 2009). As with the Sus system, the binding or capture of the substrate in this case the T or F antigen containing glycoprotein to the surface of the *B. thetaiotaomicron* cell represents first step in the metabolism of the sugar. Based on the findings of this study and in comparison to the Sus system, BT4240-50 will achieve this initially through its surface localised GalNAc binding proteins BT4245 and BT4244.

The BT4245 gene is SusE-positioned (downstream of the SusD homologue; BT4246) implying it may operate in a similar manner in BT4240-50 as the SusE of the Sus system. As to whether it performs a specialised binding role like the SusE of the Sus system (Cameron *et al*., 2012) will depend on whether the uncharacterised DUF domain towards the N-terminal of the protein exhibits enzymatic activity (Section III.1.4.1). Enzymatic activity has been reported for SusG and the products of *sus*G/F - positioned genes (downstream of SusE) which like SusE proteins are also surface exposed and in the case of SusG contains a CBM that recognizes glycan components of the target substrate (Shipman *et al*., 1999, Koropatkin *et al*., 2010, Sonnenburg *et al*., 2010, Cameron *et al*., 2012) . So far in the BT4240-50 Sus-like system, the BT4244 M60-like protease is the only protein that fits the description of a SusG-like protein. Its probable function in comparison to SusG and the BT1760 endo-levanase enzyme of the fructan utilisation locus (Sonnenburg *et al*., 2010) is that it creates internal cuts in the captured glycoprotein, releasing short glycopeptide structures that can be easily imported by the SusCD complex into the cell.

Based on the findings of this study, the imported glycopeptides contain the T and F antigens which will then be acted upon by the periplasmic exo-acting BT4241 and BT4243 enzymes (Figure IV.22). To metabolise the T antigen of the imported glycopeptide, BT4241 will be

required initially to cleave the β1-3 linked galactose capping sugar from GalNAc exposing the GalNAcα1-O glycosidic bond of the glycopeptide to the BT4243 exo-acting α-N-acetylgalactosaminidase (Section IV.3.1.2.2.3, Figure IV.22). In the case of the F antigen, BT4241 will not be required as BT4243 is capable of cleaving both the GalNAcα1-3 and GalNAcα1-O linkages in the glycopeptide containing this sugar (Figure IV.22). In comparison to the Sus system BT4241 and BT4243 enzymes thus represent SusA and B proteins respectively (Section I.6.1.1, D'Elia and Salyers, 1996).

Finally, free or released GalNAc was shown to be a target for the BT4240 kinase. BT4240 is likely cytoplasmic (Figure IV.22) given that the ATP required for its activity in the cell will be cytoplasmic (Wülfing and Plückthun, 1994). GalNAc sugar must therefore be transported from the periplsam into the cytoplasm. A likely candidate to perform this function is the putative multipass transmembrane BT4242 protein which in effect should be an inner membrane protein to achieve this function. Also, logically there will be no need for an additional outer membrane transporter protein in the presence of outermembrane SusCD binding/transport system. The fate of the phosphorylated GalNAc from here on is not quite clear as the pathway for GalNAc utilisation and amino sugar metabolism in general in *B. thetaiotaomicron* is poorly understood. There is however strong evidence at least in other organisms such as the proteobacterial species *E.coli* and *Shewanella* spp. that phosphorylated GalNAc can be further metabolised to glycolytic pathway intermediates such as the triose phosphates; dihydroxyacetone phosphate and glyceraldehyde-3-phosphate (Reizer *et al.*, 1996, Leyn *et al.*, 2012). GalNAc has also been reported in cell wall polymers such as techoic acids linked to the peptidoglycan network of gram positive bacteria (Freymond *et al.*,2006, Hermoso *et al.*, 2007). Please see Table IV.4 and Figure IV.22 for a summary of the proposed model for the mechanism of the BT4240-50 Sus-like system.

Locus tag	Pfam domains	SP, TMD	Metabolic role
BT4240	PF01636 - APH	–	**Phosphorylation**. BT4240 is a kinase capable of phosphorylating amino sugars GalNAc and GlcNAc present in mucin glycoproteins
BT4241	PF02837 - Glyco_hydro_2_N	SPI	**Glycoprotein deglycosylation**. BT4241 is a β-galactosidase that cleaves Galβ1-3GalNac (T-antigen) and Galβ1-3GlcNac (LN) present in mucin glycoproteins
	PF00703 - Glyco_hydro_2		
	PF02929 - Bgal_small_N		
	PF02836 - Glyco_hydro_2_C		
BT4242	PF02588 - DUF161	TMD	**Transport –** BT4242 is annotated as a transporter and maybe involved in the transport of mucin sugars into the cell cytoplasm
	PF10035 - DUF2179		
BT4243	PF01408 - GFO_IDH_MocA	SPI	**Glycoprotein deglycosylation**. BT4243 is an N-acetylgalactosaminidase that cleaves GalNAcα1-3GalNac and GalNAcα1-O-Ser present in mucin glycoproteins
BT4244	PF13004 - BACON	SPII	**Glycoprotein binding and proteolysis**. BT4244 exhibits mucinase, IgA protease, Gal and GalNAc binding activity
	PF00754 - F5.F8.type.C, CBM32		
	PF13402 - M60-like		
BT4245	PF08522 - DUF1735	SPII	**Glycoprotein binding**. BT4245 binds GalNAc, and F-antigen all present in mucin glycoproteins
	PF00754 - F5.F8.type.C, CBM32		
BT4246	PF14322 - SusD-like_3	SPII	**Glycoprotein binding and transport**. BT4247 and BT4246 are homologues of SusC and D respectively and together likely perform a sugar binding and transport role like the SusCD complex of *B. thetaiotaomicron* (Figure I.18). BT4246 very weakly saturates Galβ1-3GalNac (T-antigen) in ITC experiments (Data not shown).
	PF07980 - SusD		
BT4247	PF07660 - STN	SPI	
	PF13715 - Cna_B_2		
	PF07715 - Plug		
	PF00593 - TonB_dep_Rec		
BT4248	PF04773	-	**Regulation.** BT4248, BT4249 and BT4250 constitute the ECF-σ/anti- σ regulatory system for PULBT_4240-50 (Martens *et al*., 2008, 2009b)
	PF07660		
BT4249	PF05306	-	
BT4250	PF07638	-	

Table IV.4 – Summary of the annotation and function of various components of the BT4240-50 Sus-like system. SP and TMD stand for signal peptide and transmembrane domains respectively. These were predicted using a combination of LipoP, Phobius and SignalP tools (Section II.2). Protein family (PFAM) annotations were obtained from the KEGG database resource at http://www.genome.jp/kegg/ or http://www.kegg.jp/ (Kanehisa *et al*., 2012)

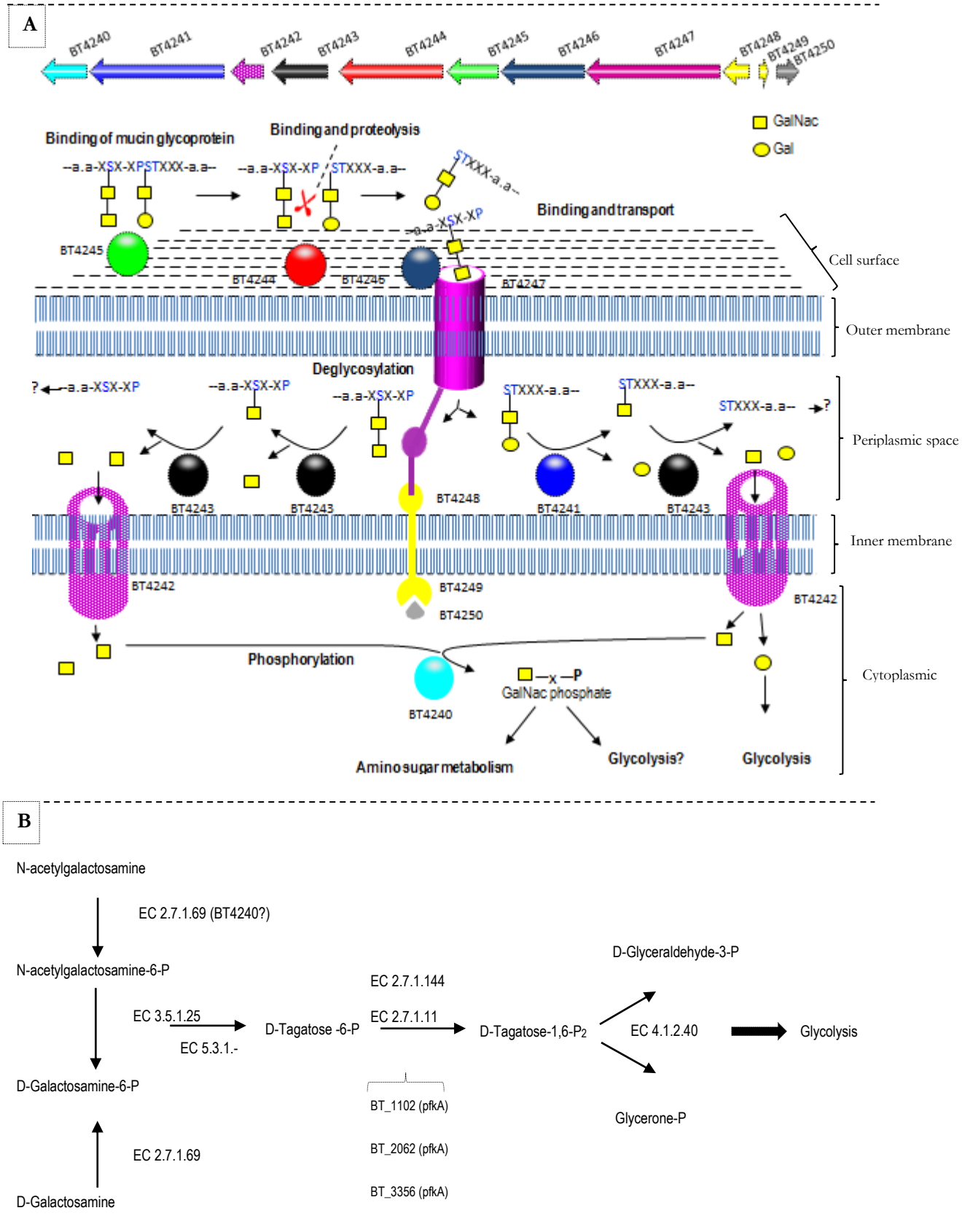


Figure IV.22 – Proposed model for the acquisition and metabolism of T and F antigen containing mucin glycoproteins by the BT4240-50 Sus-like system. Glycoprotein capture, partial digestion

and import into the periplasmic space is achieved through the concerted action of members of an outer membrane Sus-like apparatus consisting of a putative SusE GalNAc binding protein BT4245, a putative SusG GalNAc binding and mucin protease enzyme BT4244 (Chapter III) and the putative SusC and D proteins BT4246 and BT4247. Sequential deglycosylation of the imported glycopeptides is achieved by periplasmic putative SusA and B glycoside hydrolases BT4241 and BT4243 respectively. Released GalNAc is imported into the cell cytoplasm from the periplasmic space by the putative multi-pass transmembrane protein BT4242. Cytoplasmic GalNAc is then phosphorylated and probably used as a substrate for amino sugar metabolism or glycolysis. The latter is likely true for imported galactose (Gal). **B:** KEGG pathway for the conversion of GalNAc to glycolysis intermediates in *B. thetaiotaomicron (*KEGG pathway ID: bth00052*).* The KEGG database resource is available at http://www.genome.jp/kegg/ or http://www.kegg.jp/ (Kanehisa *et al.*, 2012) The above pathway is very similar to what has been described for GalNAc utilisation in *E.coli* and *Shewanella spp*, (Reizer *et al.*, 1996, and Leyn *et al.*, 2012), however so far, only putative 6-phosphofructokinase 1 enzymes (EC:2.7.1.11) alongside the BT4240 kinase enzyme (characterised in this study), have been identified in *B. thetaitoamicron.*

In summary, the BT4240-50 Sus-like system presents a comprehensive machinery for the efficient acquisition and utilisation of F and T antigen containing host-derived glycoproteins. The functional significance of the mechanism is evident from the conservation of the PUL structure and the presence of several homologues of PULBT_4240-50 genes in closely related mucin degrading species (Figure IV.23). In one species, the PUL structure is ~90% conserved (Figure IV.23B). This is the case in *Bacteroides caccae* (PULBACCAC_01835-51) which was also found to contain several other variably conserved copies of the PUL. PULBT_4240-50 may thus represent an important adaptation for host glycan foraging in *Bacteroides* species. *Bacteroides caccae* also contains over sixteen copies of M60-like domain containing proteins, two of which are present within PULBACCAC_01835-51 (Figure IV.23B). These attributes together may confer additional metabolic advantages to the organism during growth on mucins.

In conclusion, the findings of the current study have not only helped to explain the context of the mucinase activity that was observed for the BT4244-M60-like protease in Chapter III but also have provided some novel insights into the mechanism of host glycan foraging involving PULs in *B. thetaiotaomicron* and possibly its close relatives. Broadly speaking, it makes an important contribution to our general understanding of host microbial interactions in the human gut.

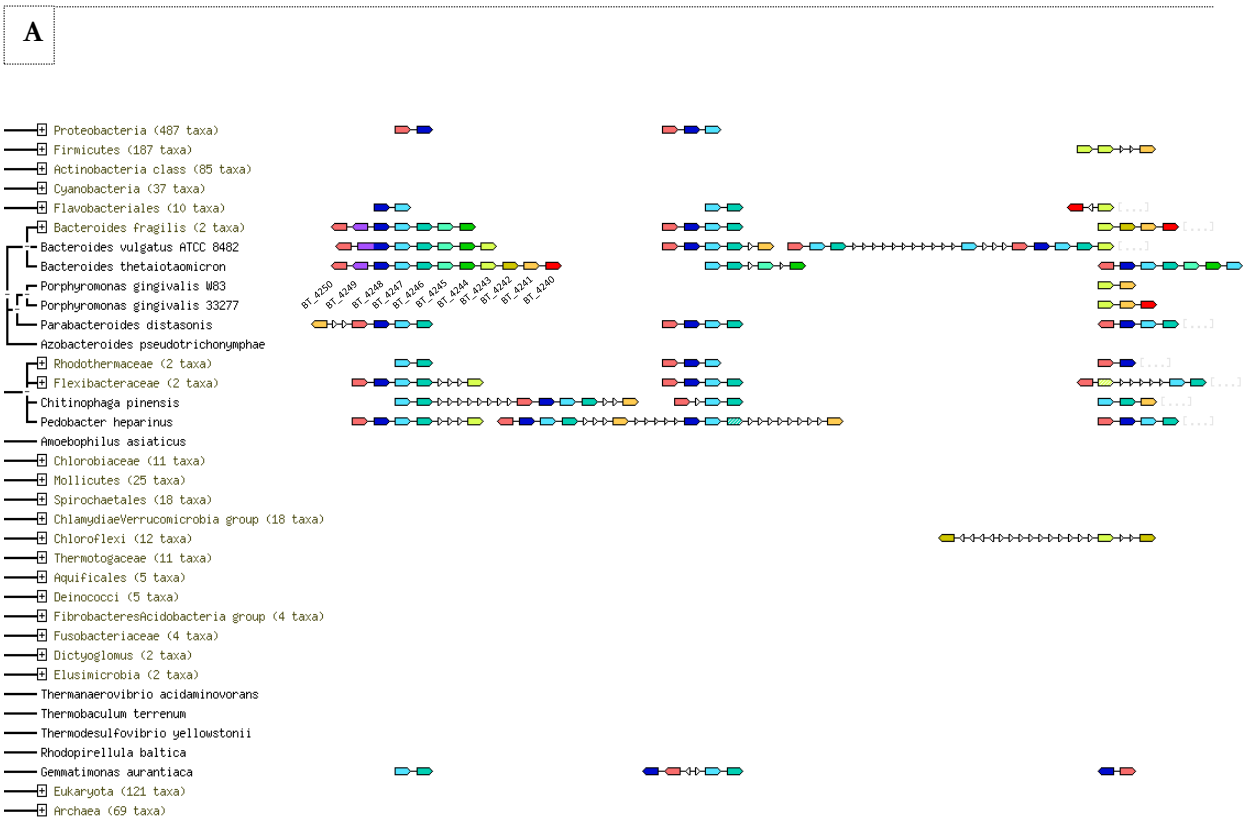


Figure IV.23 - Evidence for homologues of PULBT_4240-50 components in close and distant relatives of *B. thetaiotaomicron*. A: Results of comparative genomic analyses using the search tool for the retrieval of interacting genes/proteins (STRING). PULBT_4240-50 components are present and variably conserved in many *Bacteroides* species and members of other taxa. Note that only organisms with genomic DNA data in the STRING database (STRING 9.05) (von Mering *et al.*, 2005, Szklarczyk *et al.*, 2011) could be used in these analyses. Genomic data for *B. caccae* for example is not currently available in this database. The different genes and their

respective homologues, are colour coded throughout the figure and defined for *B. thetaiotaomicron* **B**: Structure of the most conserved PULBT_4240-50 – like gene cluster (PULBACCAC_01835-51) in *B. caccae* retrieved from sequenced DNA scaffolds from the PathoSystems Resource Integration Center (PATRICK) database (Gillespie *et al.*, 2011).

IV.5 Summary of future work

Our understanding of BT4240-50 function and host microbial interactions in general is still limited and it may be of interest to;

1- Establish the function of the DUF domain of the BT4245 protein and identify ligands for the putative SusD protein; BT4246.
2- Determine the fate of the mucin peptide that is released after deglycosylation by BT4240-50 enzymes and elucidate the pathway of GalNAc utilisation in *B. thetaiotaomicron* after phosphorylation by BT4240.
3- Investigate the biological relevance of the BT4243-FL α-N-acetylgalactosaminidase activity against the F antigen. Despite being a plausible target for the BT4240-50 Sus-like system, it is currently it is not known whether this substrate is capable of up-regulating the PULBT_4240-50.

CHAPTER V

In-vitro contribution of the BT4240-50Sus-like system to *B. thetaiotaomicron* fitness and survival on mucins

V.1 Introduction

Mucin utilization is important for microbial virulence and the competitive colonisation of mucosal surfaces. This has been shown for several mucosal microbes including recently for the pathogenic gastrointestinal microbes *Campylobacter jejuni* and *E. coli* whose ability to utilize mucin monosaccharides confers upon them a competitive advantage as well as enhances their virulence (Stahl *et al.*, 2011, Bertin *et al.*, 2013). In mutualistic mucosal microbes, mucin utilisation is mainly for the purposes of nutrition and competitive colonisation, although it was recently shown for *B. thetaiotaomicron* that this activity can also modulate virulence gene expression in pathogenic *E. coli* (Pacheco *et al.*, 2012, McGuckin *et al.*, 2011). Mucins thus represent an important fitness factor whose utilisation has significant implications on gut microbial ecology as well as human health (Koropatkin *et al.*, 2012). This also implies that microbial genetic factors driving mucin adaptation hold great potential as future tools for the manipulation of the complex gut microbial community for various health reasons.

In-vitro experimental data from the previous chapter suggested that the BT4240-50 Sus-like system (encoded by PULBT_4240-50) of *B. thetaiotaomicron* is capable of metabolising mucin glycoproteins containing the T and F antigens (Galβ1-3GalNAc and GalNAcα1-3GalNAc respectively). However, the requirement of BT4240-50 and its components in mucin utilisation has not been confirmed. The importance of doing so in this case in particular lies in the fact that several other PULs in *B. thetaiotaomicron* (about 10 in total) are also upregulated to the T antigen (Martens *et al.*, 2009). A similar degenerate response had been reported by Marcobal *et al.*, (2011) during *B. thetaiotaomicron* growth on human milk oligosaccharides (HMO) and indeed attempts to prove the requirement of several HMO sensitive PULs proved futile. This included a deletion to the entire PUL locus containing the putative fucosidase enzyme BT4136 (PULBT_4232-36) which failed to cause any growth defect in the organism on HMOs, despite being one of the most upregulated PULs to the substrate *in-vitro* (Marcobal *et al.*, 2011).

Of the ~ 10 PULs upregulated in response to the T antigen, 7 of them including PULBT_4240-50 are about 10 times upregulated with some containing several uncharacterized glycoside hydrolases (Table V.1). Interestingly one of the PULs, PULBT_3957-65 which appears to be as up regulated as PULBT_4240-50 also encodes a putative protease enzyme [BT3960, annotated as Peptidase_C2 (PF00648)] in addition to several putative GH families (Martens *et al.*, 2008). Although most of these activities have not been characterised biochemically, such observations immediately raise questions over the actual functional significance of PULBT_4240-50.

PULBT_3983-94

BT3983	susC-like
BT3984	susD-like
BT3985	hypothetical protein
BT3986	putative patatin-like protein
BT3987	Glycoside Hydrolase Family 18
BT3988	putative peptidoglycan bound protein
BT3989	hypothetical protein
BT3990	Glycoside Hydrolase Family 92
BT3991	Glycoside Hydrolase Family 92
BT3992	anti-sigma factor
BT3993	ECF-type sigma factor
BT3994	Glycoside Hydrolase Family 92

PULBT_3957-65

BT3957	Hybrid two-component system regulator
BT3958	susC-like
BT3959	susD-like
BT3960	hypothetical protein (peptidase_C2 protease)
BT3961	hypothetical protein
BT3962	Glycoside Hydrolase Family 92
BT3963	Glycoside Hydrolase Family 92
BT3964	putative secretory protein
BT3965	Glycoside Hydrolase Family 92

PULBT_4402-07

BT4402	ECF-type sigma factor
BT4403	anti-sigma factor
BT4404	susC-like
BT4405	susD-like
BT4406	hypothetical protein
BT4407	hypothetical protein

PULBT_1032-51

BT1032	Glycoside Hydrolase Family 92
BT1033	hypothetical protein
BT1034	putative signal transducer
BT1035	hypothetical protein
BT1036	hypothetical protein
BT1037	hypothetical protein
BT1038	hypothetical protein
BT1039	susD-like
BT1040	susC-like
BT1041	tyrosine DNA recombinase
BT1042	susC-like
BT1043	susD-like
BT1044	Glycoside Hydrolase Family 18
BT1045	hypothetical protein
BT1046	susC-like
BT1047	susD-like
BT1048	Glycoside Hydrolase Family 18
BT1049	putative patatin-like protein
BT1050	hypothetical protein
BT1051	Glycoside Hydrolase Family 20

PULBT_0317-19

BT0317	susC-like
BT0318	susD-like
BT0319	susD-like

PULBT_2259-62

BT2559	susD-like
BT2560	susC-like
BT2561	anti-sigma factor
BT2562	ECF-type sigma factor

Table V.1 - List of T antigen (Core 1) sensitive PULs in *B. thetaiotaomicron* showing at least 10 times upregulation to the substrate *in-vitro*. Please see list of PULBT_4240-50 components (not included here) in the previous Chapter. Highest upregulation values in response to the T antigen have been reported for PULBT_3957-61 and PULBT_4240-50 (Martens *et al.*, 2008).

Martens *et al.*, (2008) showed that a combined deletion of five ECF-σ/anti-σ factor regulators controlling mucin O-targeted PULs including the regulator of PULBT_4240-50 negatively impacts on the organism's fitness in vivo. However, it was not clear specifically which of the affected PULs or set of genes were driving this phenotype. More importantly, it was also discovered that PULBT_4240-50 and PULBT_4356-58 were not under the sole control of their respective ECF regulators as several genes within these PULs still showed upregulation after the ECF deletions.

V.2 Objective

The purpose of this chapter is to evaluate the contribution of BT4240-50 Sus-like system (encoded by PULBT_4240-50) and its components to *B. thetaiotaomicron* fitness and survival on mucins by analysing data from various *in-vitro* growth experiments involving the organism and various deletion mutants.

V.3 Results:

V.3.1 Requirement of the BT4240-50 Sus-like system and BT4244 in mucin utilisation

V.3.1.1 *In-vitro* non-competition growth experiments with ΔBT_4244 and ΔPULBT_4240-50 deletion strains

B. thetaiotaomicron has been shown to be capable of utilising porcine gastric mucins (PGMIII) *in-vitro* as a sole source of carbon and nitrogen (Martens *et al.*, 2009a, Marcobal *et al.*, 2011). In this study, we initially tested the requirement of BT4240-50 and its M60-like protease component, (BT4244) in mucin utilisation *in-vitro* in a non-competitive environment. To perform this, B. *thetaiotaomicron* mutant strains containing in-frame deletions of either the gene encoding the BT4244 M60-like protease (ΔBT_4244) or the entire PULBT_4240-50 locus encoding the BT4240-50 Sus-like system (ΔPULBT_4240-50) were created.

V.3.1.1.1 Generating ΔBT_4244 and ΔPULBT_4240-50 deletion strains

ΔBT_4244 and ΔPULBT_4240-50 deletion mutants were created using the method of allelic exchange as described in in Section II.1.46. The genomic region flanking PULBT_4240-50 is shown below in Figure V.1. Deletion primers used during the process are listed in Appendix Table A.4. Outer primers A and D contained engineered *Bam* HI and *Xba*I sites allowing for cloning of the sown flanks into similar sites of the pExchange-*tdk* suicide vector (Section II.1.46).

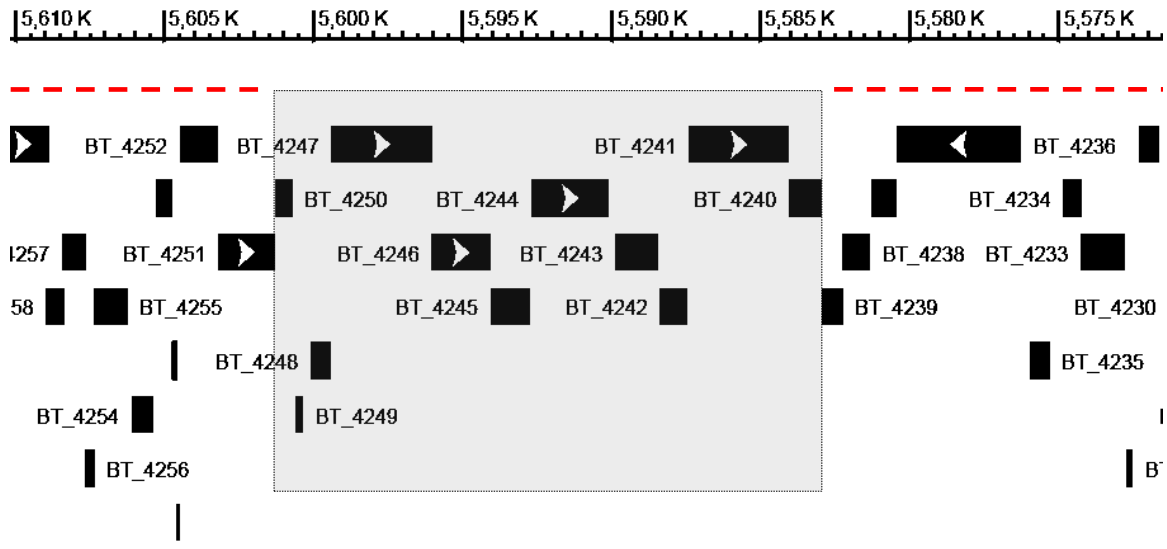


Figure V.1 – Genomic context of PULBT_4240-50 and BT_4244. The PULBT_4240-50 gene locus is highlighted in gray and flanking regions are indicated by red dotted lines. The same deletion strategy for PULBT_4240-50 was used for the production of BT_4244 deletion mutants as described in Section II.1.46.

During the isolation of recombinant mutants (Section II.1.46), up to 10 *B. thetaiotaomicron* colonies were screened for the desired deletions. Clones positive for the desired deletion in each case produced a ~ 2000 bp DNA fragment by agarose gel electrophoresis (AGE) after PCR with outer primers A and D (Figure V.2). Because PCR amplification conditions were often aimed at the 2000 bp fragment, the large fragments of the wild type or 'failed mutants' were not usually completely amplified. PCR was also often performed on wild type DNA (WT) and included as control.

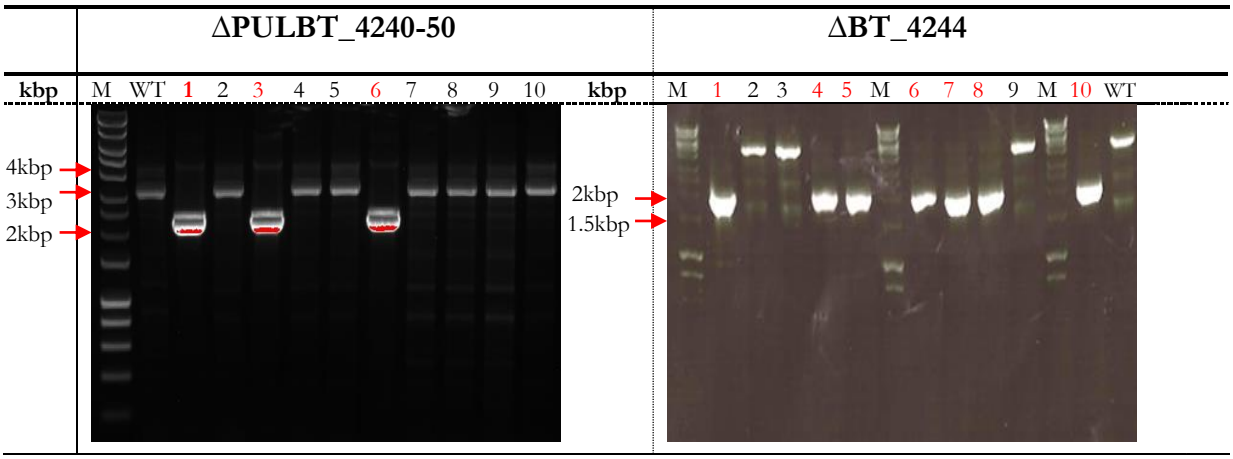


Figure V.2 - Screening for ΔPULBT_4240-50 and ΔBT_4244 deletion mutants. Clones positive for each deletion are indicated with red numbers and yielded a ~ 2000 bp fragment by AGE [after PCR on extracted genomic DNA with primers A and D (Appendix Table A.4)] equivalent to the size of fused flanking regions earlier cloned into the pExchange-*tdk* suicide vector (See Section II.1.46)

V.3.1.1.2 Growth of ΔPULBT_4240-50 and ΔBT_4244 deletion mutants on glucose and porcine gastric mucins *in-vitro*

To investigate the effect of various deletions on the ability of *B. thetaiotaomicron* to utilise mucins, wild type and deletion strains were cultured separately in minimal media supplemented with either 1% glucose (MM-Glc) or porcine gastric mucins (PGMIII). Growths were monitored by measuring the optical density of the cultures at 600nm every hour for a maximum of 24 hours. Unlike ΔBT_4244 the growth of ΔPULBT_4240-50 was partially retarded on PGMIII in comparison to the wild type (Figure V.3). A similar experiment was carried out using bovine submaxillary mucins as substrate but both wild type and knockout strains failed to grow on the substrate (data not shown).

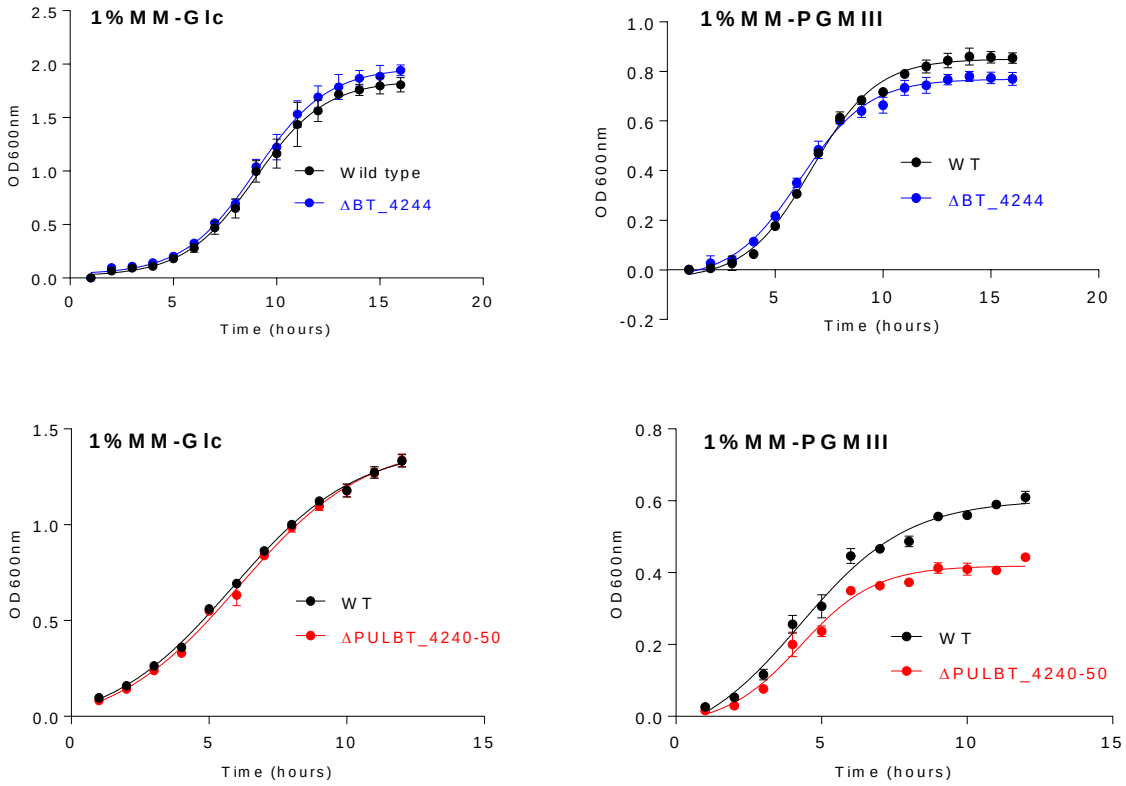


Figure V.3 - Comparing the growth of *B. thetaiotaomicron* wild type (WT), ΔPULBT_4240-50 and ΔBT_4244 deletion mutants on glucose and porcine gastric mucins. WT and deletion strains were all initially cultured in minimal medium containing 0.5% glucose (MM-Glc) to approximately equal optical densities (OD600nm ~1.5-2) before transfer to 1% MM-Glc or minimal medium containing 1% porcine gastric mucins (MM-PGMIII) for growth monitoring. Growth data for each strain was obtained from three individual replicate cultures. Error bars indicate the standard deviation of the OD600nm for each triplicate data set.

V.3.1.2 *In-vitro* competition experiments with ΔBT_4244 and ΔPULBT_4240-50 deletion strains

To investigate the impact of various deletions on the *B. thetaiotaomicron* fitness *in-vitro*, competition experiments were performed with various deletion strains against the wild type. Strains were initially labelled with a unique 24 bp signature/tag sequence to enable their differentiation and quantification from co-cultures.

V.3.1.2.1 Tagging of *B. thetaiotaomicron* WT and deletion strains

The procedure for tagging strains is as described in Section II.1.46. Tagging sequences and primers used for their detection are listed in Appendix Table A.5. The naming of tags (Tag1 and 11) was decided by Eric C. Martens of the University of Michigan Medical School (USA). WT strains were tagged with the tag1 sequence while both ΔBT_4244 and ΔPULBT_4240-50 deletion strains were tagged with tag11. Positive clones yield 200 bp fragments by AGE after PCR with tag primers. Strains selected for competition experiments all had their tags integrated in the NBU2 *att1*site (tRNASer between BT_4680 and BT_4681), confirmed using NBU2-*att1* primers (Appendix Table A.5). Theoretically these should yield no product on agarose gels after PCR with NBU2-*att1* primers due to loss of the primer site after integration while WT and strains with tags integrated in the NBU2 *att2* site site (between BT4738 and BT4739) should yield a 900 bp fragment (Figure V.4).

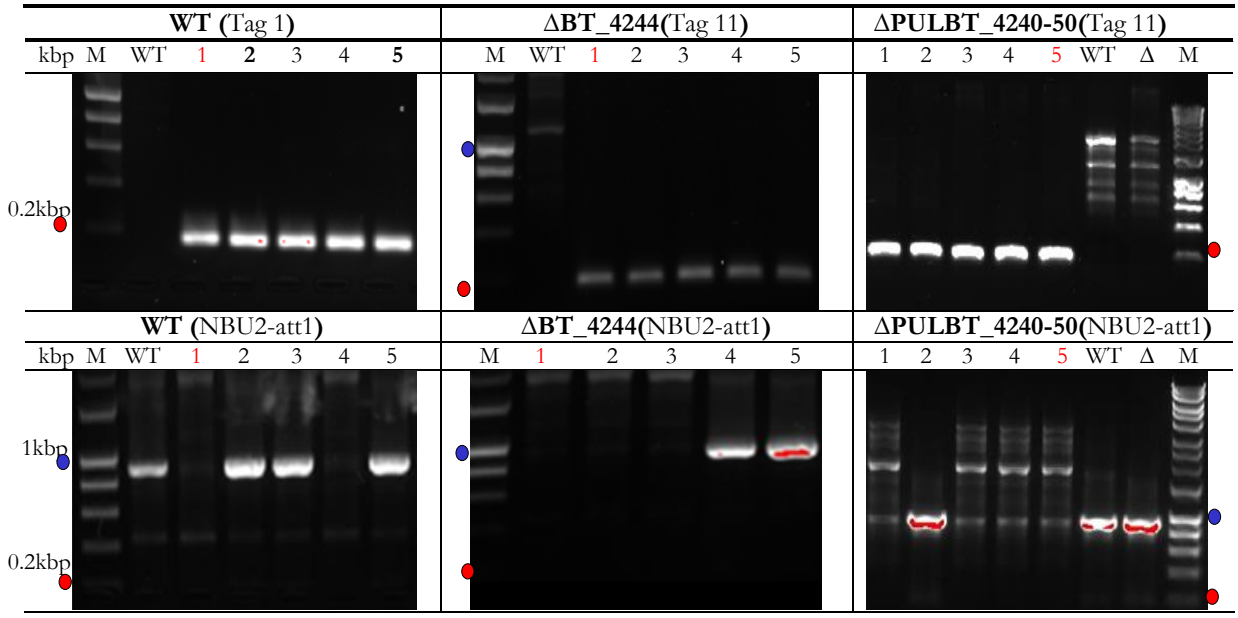


Figure V.4 – Tagging of *B. thetaiotaomicron* WT and deletion strains and determination of tag integration sites. Top agarose gels are PCR results for various clones using corresponding tag detection primers listed in (Appendix Table A.5). Positive clones with integrated tags yield a 200 bp product. Bottom gels are PCR results for the same clones using NBU-2 *att1* primers to confirm the site of integration of various tags detected in the top gels. Positive clones with tags integrated in the NBU-2 *att2* site yield a 900 bp product while those with tags integrated in the NBU-2 *att1* site yield no product using NBU-2 *att1* primers (Appendix Table A.5). PCR conditions are detailed in the materials and methods section. Clone numbers indicated in red were those selected for competition experiments. The details are as follows WT: clone 1, tag1, NBU2-*att1*, ΔBT_4244: clone 1, tag 11, NBU2-*att1,* ΔPULBT_4240-50: clone 5, tag11, NBU2-*att1*. Red circles indicate position of the 200 bp (0.2 kbp) marker while blue circles indicate the position of the 1 kbp marker.

V.3.1.2.2 Co-culture growth of tagged *B. thetaiotaomicron* wild type and deletion strains

Selected wild type and deletion strains [WT: clone 1, tag1, NBU2-*att1*, ΔBT_4244: clone 1, tag 11, NBU2-*att1,* ΔPULBT_4240-50: clone 5, tag11, NBU2-*att1* (Figure V.4)] were initially cultured in TYG rich medium overnight and mixed in approximately equal proportions before inoculating into MM-Glc. Cells were cultured over a period of 5 days, subculturing every day. 2 ml of culture were also collected every day and frozen at -80 °C for DNA extraction. After 5 days, cells were switched to MM-PGMIII and subcultured every day for about 10 days. 2 ml of culture were also saved every day as with the glucose samples. qPCR enumeration of competing strains was performed using genomic DNA extracted from the frozen 2 ml cultures as described in Section II.1.46. The results showed that wild type and deletion strains in both cases grew normally with similar percentage representations for the first five days in glucose. However, there was a consistent increase in the percentage representation of the wild type against the ΔPULBT_4240-50 strain after switching to growth on MM-PGMIII (Figure V.5). This was not the case for the WT and ΔBT_4244 combination, were the percentage representation of each stayed fairly constant after switching to MM-PGMIII. PULBT_4240-50 thus confers a competitive advantage to B. *thetaiotaomicron* during growth on PGMIII *in-vitro*.

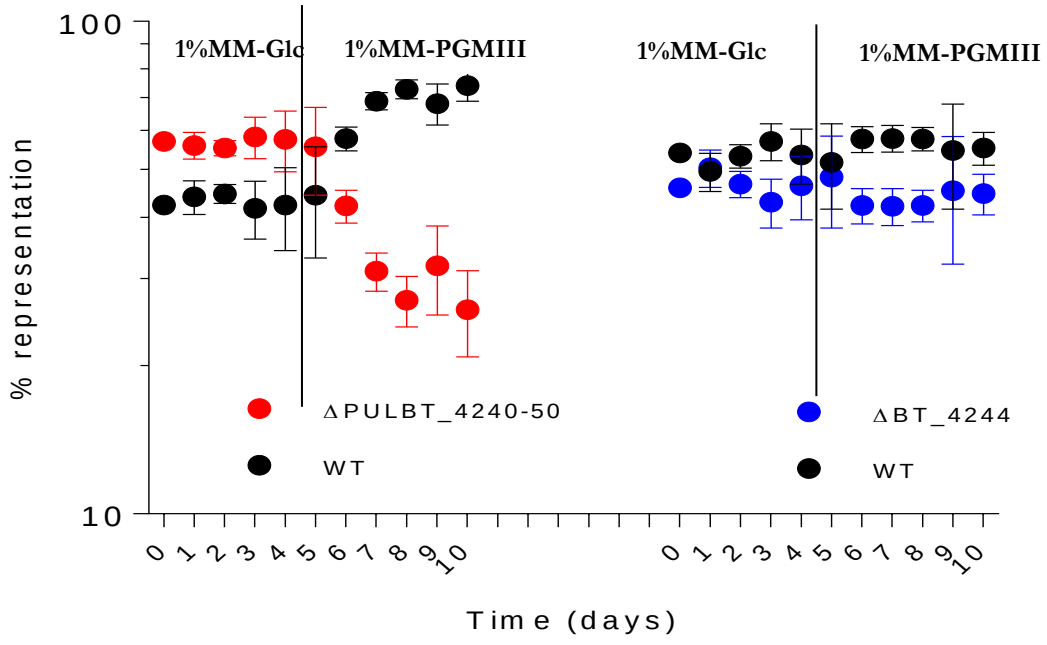


Figure V.5 –Impact of PULBT_4240-50 and BT_4244 deletions on the ability *B. thetaiotaomicron* to compete with the wild type *in-vitro*. WT strains contained tag1 inserted into their NBU2-*att1* sites while both ΔBT_4244 and ΔPULBT_4240-50 deletion mutants contained tag11 in the same sites. Equal

proportions of each strain were co-cultured initially in 1%MM-Glc over a period of 5 days, subculturing every day before switching to 1%MM-PGMIII. DNA was extracted from a fraction of each culture every day and the percentage representation of each strain from co-cultures enumerated by qPCR using respective tag primers (listed in Appendix Table A.5). qPCR conditions are given in Section II.1.46.5

V.3.2 Requirement of BT4244 homologues BT3015 andBT4272 in mucin utilisation

V.3.2.1 *In-vitro* non-competition experiments

Recombinant M60-like domains of BT4272 and BT3015 just like the recombinant BT4244 protein were all shown to exhibit mucinase activity in chapter I. The PUL containing BT_3015 (PULBT_3010-17) is also about 10 times upregulated during *B.thetaiotaomicron* growth in the late phase of growth on PGM. The absence of a growth defect on mucins by the ΔBT_4244 mutant and its persistence in co-culture with the wild type (Figures V.3 and V.5) led us to question whether these other mucin proteases (BT3015 and BT4272) could have been providing 'back-up' for the deleted BT4244 enzyme. We sought to investigate this by testing the growth of mutants containing deletions to the genes encoding these proteins on mucins. A mutant containing deletions to all M60-like protease genes (BT_3105, BT_4244 and BT_4272) was also tested.

V.3.2.1.1 Generating ΔBT_3015, ΔBT_4272 and ΔBT_3015ΔBT_4244ΔBT_4272 mutants

The genomic contexts of the BT_3015 and BT_4272 genes are shown in Figure V.6. ΔBT_3015, ΔBT_4272 were created in the same manner as ΔBT_4244 and the ΔBT_4240-50 using the method of allelic exchange. To create the ΔBT_3015ΔBT_4244ΔBT_4272 deletion mutant, an initial deletion of the BT_3015 gene was performed on the ΔBT_4244 strain to yield a double deletion mutant ΔBT_3015ΔBT_4244 on which a third deletion, this time to the ΔBT_4272 was made. Deletion primers used during the process are listed in Appendix Table A.4.

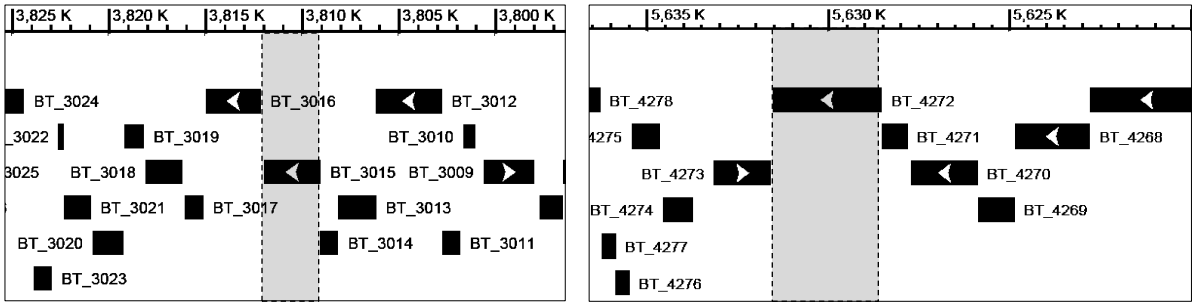


Figure V.6 - Genomic context of BT_3015 and BT_4272. Genes targeted for deletion are shown in gray rectangles. The same deletion strategy for the PULBT_4240-50 locus deletion was used to produce ΔBT_3015, ΔBT_4272 and ΔBT_3015ΔBT_4244ΔBT_4272 single and triple deletion mutants respectively.

V.3.2.1.2 Growth ΔBT_3015, ΔBT_4272 and ΔBT_3015ΔBT_4244ΔBT_4272 mutants on glucose and porcine gastric mucins in-vitro

The same procedure as described in Section V.3.1.1.2, this time involving the above mutants (ΔBT_3015, ΔBT_4272 and ΔBT_3015ΔBT_4244ΔBT_4272) was followed. The growth data is as shown in Figure V.7. None of the deletion mutants showed any growth defect on PGMIII.

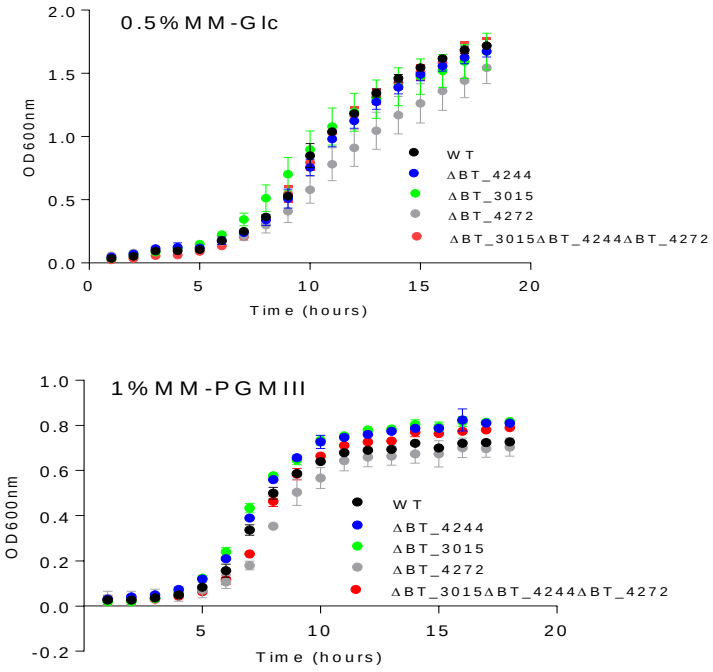


Figure V.7 - Comparing the growth of *B. thetaiotaomicron* wild type (WT), ΔBT_3015, ΔBT_4272 and ΔBT_3015ΔBT_4244ΔBT_4272 deletion mutants on glucose and porcine gastric mucins. WT and deletion strains were all initially cultured in minimal medium containing 0.5% glucose (MM-Glc) to approximately equal optical densities (OD600nm ~1.5-2) before transfer to 0.5% MM-Glc or minimal medium

containing 1% porcine gastric mucins (MM-PGMIII) for growth monitoring. Growth data for each strain was obtained from three individual replicate cultures. Error bars indicate the standard deviation of the OD600nm for each triplicate data set.

V.3.3 Role of BT4240-50 in N-acetylgalactosamine (GalNAc) utilisation

GalNAc utilisation in *B. thetaiotaomicron* is currently poorly understood. Galβ1-3GalNAc (T antigen) and GalNAcα1-3GalNAc (F antigen) were identified as the main targets for BT4240-50 (Chapter I), and hence were interesting targets to test the requirement of the PUL. This was our intention in the later part of this study, but due to the paucity of these substrates; such experiments could not be performed. We however wondered what effect the PUL deletion would have on the organisms' ability to utilise their monosaccharide components (Gal and GalNAc). A rather unanticipated observation was made as the mutant completely lost its ability to utilise GalNAc but not Gal (Figure V.8), leading us to question the role of the PUL in GalNAc utilisation. The importance of addressing this question also lies in the fact that this observation was contrary to the general notion that PULs are primarily targeted at complex sugars not monosaccharides.

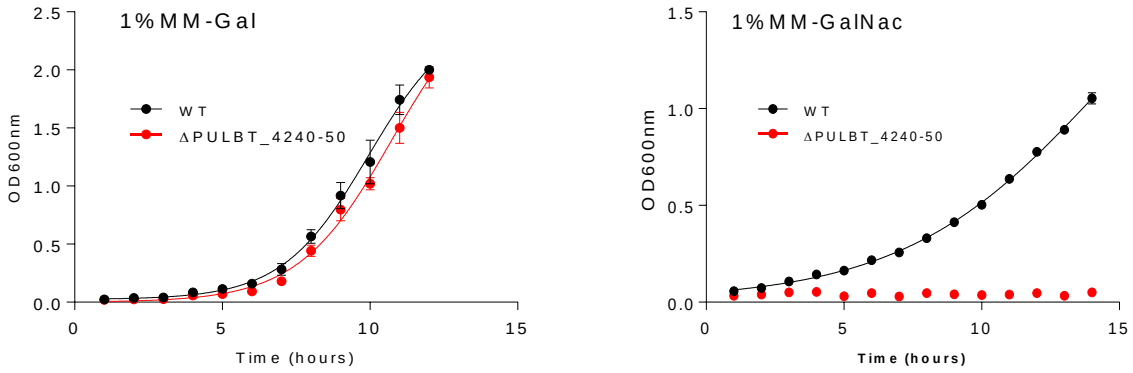


Figure V.8 – Growth of WT and ΔPULBT_4240-50 on T and F antigen monosaccharides (Gal and GalNAc). WT and deletion strains were all initially cultured in nutrient rich medium (TYG) to approximately equal optical densities (OD600nm ~1.5-2) before transfer to minimal media containing 1% galactose (MM-Gal) or N-acetylgalactosamine (MM-GalNAc). Growth data for each strain was obtained from three individual replicate cultures. Error bars indicate the standard deviation of the OD600nm for each triplicate data set.

V.3.3.1 Growth of ΔBT_4240 and ΔBT_4242 deletion mutants on GalNAc and GlcNAc

The inability of ΔPULBT_4240-50 to grow on GalNAc was an indication that there are important steps in GalNAc utilisation in *B. thetatiotaomicron* that are exclusive to BT4240-50. Drawing on from the discussion in Chapter IV, it could be due to its inability to import GalNAc through the putative SusCD complex (i.e. BT4246 and BT4247) which is lost in ΔPULBT_4240-50, implying that the sugar is not available for the rest of the PUL or other possible systems that may be involved with its metabolism (e.g. pathway 2 of Figure V.9). However, this was deemed to be unlikely as the outer membrane SusCD complex is only partially upregulated to GalNAc (Martens *et al.*, 2008), although it could also be argued as the reason for organism's slow growth on GalNAc compared to other substrates like Gal and glucose. A second reason could be due to loss of the putative inner membrane transporter BT4242 (Chapter IV), i.e. assuming that transport of GalNAc from the extracellular space into the periplasm can occur through other outer membrane GalNAc transporters (e.g. pathway 3). This option was worth investigating because BT4242 unlike the putative SusCD complex shows high basal levels of expression i.e. during growth on glucose (Martens *et al.*, 2008). Finally it could be due to loss of the BT4240 GalNAc kinase which also shows high basal levels of expression, implying that GalNAc may get imported into the cell through BT4242 or another unknown transport system (e.g. pathway 4). To investigate the latter two possibilities we tested the ability of strains containing deletions of gene BT_4242 and BT_4242 to grow on GalNAc. Appendix Table A.4 contains a list of primers used during the creation of various mutants. Evidence from the growth data suggested that only the strain containing a deletion to the BT4240 kinase was unable to utilise GalNAc (Figure V.9), implying that GalNAc is likely imported through an unknown transport system (e.g. pathway 4).

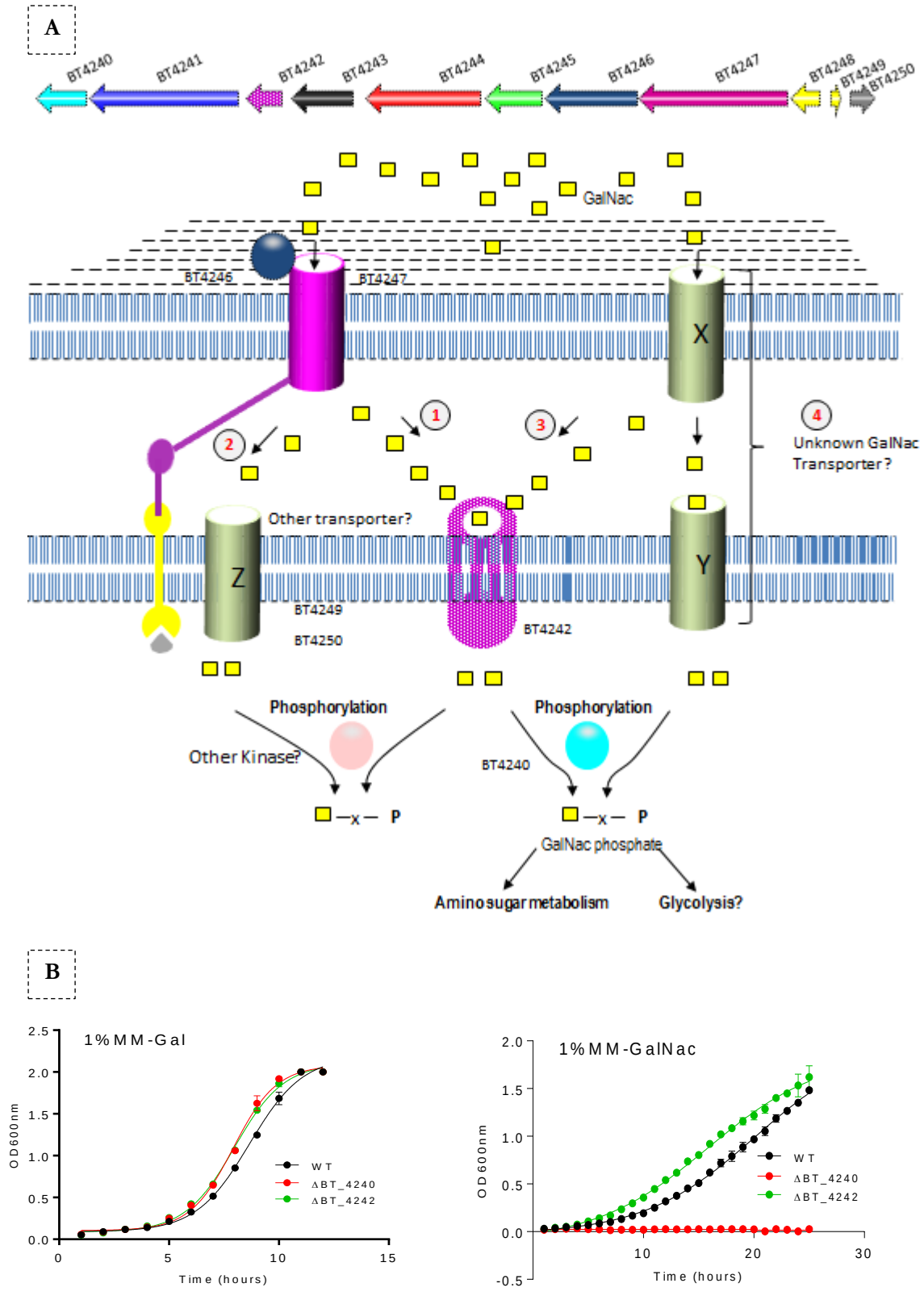


Figure V.9 – Growth of ΔBT_4240 and ΔBT_4242 deletion mutants on Gal and GalNAc. A: Analysing the possible cause of *B. thetaiotaomicron* strain ΔPULBT_4240-50's growth defect on GalNAc to identify potential key GalNAc utilisation genes and targets for deletion. The putative SusCD complex (BT4246 and BT4247) is only partially upregulated to GalNAc *in-vitro* implying that the inability of ΔPULBT_4240-50 to utilise GalNAc

could be simply as a result of the loss of the putative inner membrane transporter BT4242 that may use GalNAc imported through an unknown outer membrane transporter (X) (e.g. pathway 3) or due to loss of the cytoplasmic GalNAc kinase BT4240 which could phosphorylate GalNAc imported through BT4242 or an unknown outer and inner membrane X-Y transport system (pathway 4). **B:** Growth of ΔBT_4240 and ΔBT_4242 on T and F – antigen monosaccharides (Gal and GalNAc). WT and deletion strains were all initially cultured in nutrient rich medium (TYG) to approximately equal optical densities (OD600nm ~1.5-2) before transfer to minimal media containing 1% galactose (MM-Gal) or N-acetylgalactosamine (MM-GalNAc). Growth data for each strain was obtained from three individual replicate cultures and error bars indicate the standard deviation of the OD600nm for each triplicate data set.

V.3.3.2 Growth of ΔBT_4240 and ΔBT_4242 deletion mutants on chondroitin sulphate and porcine gastric mucins

ΔBT_4240's inability to grow on GalNAc was not only suggestive that phosphorylation is central to GalNAc utilisation but also that this function is exclusive to the BT4240 GalNAc kinase. It is however also possible that other GalNAc kinases are produced by the cell as components of different PULs but are tightly controlled. We sought to investigate this by evaluating the growth of ΔBT_4240 on a GalNAc containing substrate chondroitin sulphate (CS) that upregulates a completely different PUL in *B. thetaiotaomicron* (PULBT_3324-50). CS contains about 50% GalNAc and hence we anticipated that if the GalNAc released by the Sus-like system encoded by this PUL (BT3324-50) can only be phosphorylated by BT4240 from BT4240-50, then the deletion mutant will always grow at 50% the rate of the wild type at all times during the experiment as confirmed below (Figure V.10). Finally, ΔBT_4240 and ΔBT_4242 growth on PGMIII was also assessed. ΔBT_4240 showed a slight growth defect on the substrate while ΔBT_4242's growth was clearly unaffected.

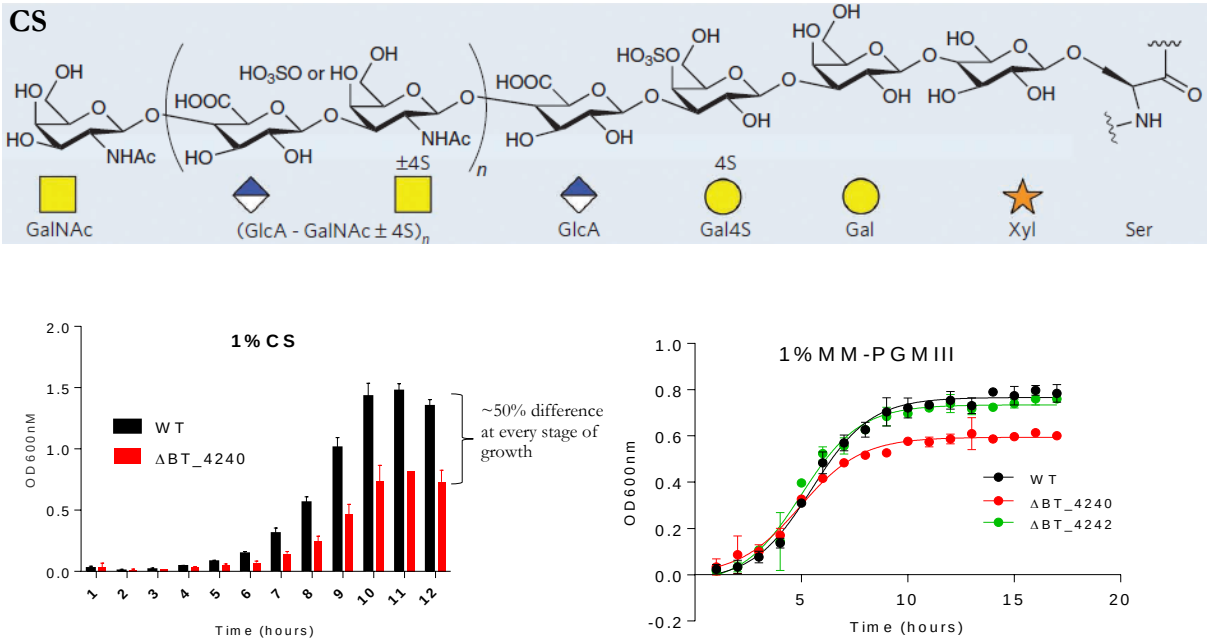


Figure V.10 - Effect of BT_4240 and BT_4242 deletions on *B. thetaiotaotomcin* growth on GalNAc containing substrates Chondroitin sulphate (CS) and porcine gastric mucins (PGMIII). WT and deletion strains were all initially cultured in nutrient rich medium (TYG) to approximately equal optical densities (OD600nm ~1.5-2) before transfer to minimal media containing 1% CS or PGMIII. Growth data for each strain was obtained from three individual replicate cultures and error bars indicate the standard deviation of the OD600nm for each triplicate data set. A typical structure of CS adapted from Ly *et al.*, 2011, is shown above growth curves

V.4 Discussion

The current study revealed an important involvement of the BT4240-50 Sus-like system in mucin utilisation in *B. thetaiotaomicron*. Although deletion of encoding PUL (PULBT_4240-50) did not result in a profound growth phenotype in vitro, evidence from the current study suggests that it is an important requirement for growth on mucins in a competitive environment. A deletion of the BT4244 M60-like protease failed to produce any significant growth defect even in a competitive environment. BT4244 is functionally similar to the prototypic SusG (BT3698) and the BT1760 endo-levanase enzyme of the fructan utilisation PUL which have all been shown genetically to be important for growth on their respective substrates (Reeves *et al.*, 1997, Sonnenburg *et al.*, 2010). However it is worth mentioning that unlike BT_4244's PUL (which is not the only mucin sensitive PUL in *B. thetaiotaomicron*) the PULs to which the BT_3698 and BT_1760 genes belong seem to be only responsive PULs to their respective substrates, implying there is a stricter requirement for them. A typical SusG protein functions to create internal cuts in the complex substrate facilitating its uptake through the SusCD complex (Koropatkin *et al.*, 2012). Given the importance of this function it is very likely that there are other SusG-type proteins within other T antigen sensitive PULS excluding the BT4272 and BT3015 M60-like proteases which were shown to be dispensable for growth on mucins. One candidate is the BT3960 hypothetical protein annotated as peptidase_C2 protein of PULBT_3957-65 (Table V.1). This protease is not only present within apparently the most upregulated PUL to the T antigen (Martens *et al.*, 2008) but also shows some remarkable similarities with the BT4242 M60-like protease (Figure V.11) and the SusG protein.

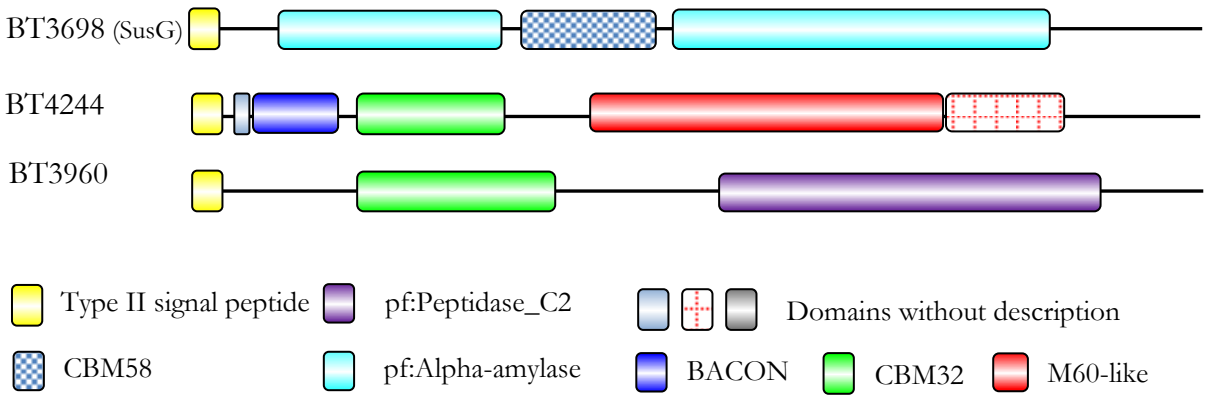


Figure V.11 - Comparing the domain content and organisation of BT4244, BT3960 and the SusG protein (BT3968). All proteins contain a type II signal peptide sequence, one or more putative or confirmed carbohydrate binding module(s) and one or more putative or confirmed enzymatic domain(s). Note: The

domains are roughly to scale relative to each other within same protein but individual proteins are not drawn to scale relative to one another. The actual length of BT3698 is 692 amino acids , that of BT4244 is 857 amino acids, while that of BT3960 is 508 amino acids.

The discovery that BT4240-50 through the BT4240 kinase plays a crucial role in GalNAc utilisation makes an important contribution towards our understanding of GalNAc utilisation in *B. thetaiotaomicron* which is currently poorly understood. Evidence form the data suggests that GalNAc phosphorylation is crucial to GalNAc utilisation in *B. thetaiotaomicron*. The data also suggest that free monosaccharide GalNAc may not be imported into the cell through BT4240-50 but through an alternative transporter for sugar. This is not only due to the very low expression of PULBT_4240-50 components during growth on the GalNAc monosaccharide (Martens *et al.*, 2008) but also due to the absence of a growth defect following the deletion of the gene encoding the putative BT4242 inner membrane transporter of the BT4240-50 Sus-like system. BT4240-50 as expected is likely specialised for complex GalNAc containing substrates rather than monosaccharides. It is however important to further confirm this in growth experiments with strains containing a deletion to the SusCD complex. In GalNAc utilising proteobacterial species such as *E.coli* and *Shewanella* spp., GalNAc import is thought to be mediated by specialised membrane transporters. These include PTS (phosphotransferase system) transporters in the case of *E.coli* and a combination of outer membrane TonB-dependent and inner membrane permease transporters in the case of *Shewanella* spp (Reizer *et al.*, 1996, Leyn *et al.*, 2012)

GalNAc is an important component of dietary chondroitin sulphate as well as host derived mucin O-glycans and its use likely confers a significant competitive advantage to *B. thetaiotaomicron* during growth on these substrates. Both T and F antigens are known to be important cancer associated antigens including colonic cancer as their production increases during the disease (Brockhausen, 2006, Lescar *et al.*, 2007). Theoretically this should alter the gut microbiota in these patients in favour of species most adapted to utilise these substrates. Interestingly, the percentage of *Bacteroides* and *Prevotella* species which happen to contain homologues of the BT4240 kinase (Appendix E) was found to be significantly higher than normal in these patients (Sobhani *et al.*, 2011) and could be the reason behind their ability to persist in the GalNAc rich environment.

Two homologues of the gene are also present in the eukaryotic protozoa *T. vaginalis* (Appendix E) and may have been shared through horizontal gene transfer (Alsmark *et al.*, 2013). These

homologues may be performing a similar phosphorylation role in GalNAc utilisation (as observed for BT4240 *B. thetaiotomicron*) in this mucin degrading mucosal microbe (Wiggins *et al.*, 2001).

Finally the absence of a growth defect in the ΔBT_4242 deletion mutant during growth on complex PGMIII *in-vitro* left even more important questions to be answered. Ideally if BT4240-50 is targeted at complex GalNAc containing sugars, one would expect that the deletion of BT_4242 which possibly encodes the sole inner membrane transporter within the PUL should cause the accumulation of released Gal and GalNAc within the periplasmic space leading to defective growth. However this was not the case as observed in Figure V.9. One possibility is that the released periplasmic sugars are channelled into the cytoplasm by other Gal or GalNAc inner membrane transporters yet to be defined. It may also include transporters from other T or F antigen targeted PULs. All these however need to be investigated experimentally.

V.5 Future work

Some suggestions for future studies to further our understanding of the requirement of BT4240-50 in mucin utilisation include

1) Identify and characterise other F and T antigen targeted PULs in *B. thetaiotaomicron*. An interesting PUL to start with is PULBT_3957-65 which happens to be just as regulated as PULBT_4240-50 during growth on PGMIII
2) Perform more deletion experiments on PULBT_4240-50 components including the SusCD complex (BT4246 and BT4247) to identify other crucial contributions made by the PUL in mucin utilisation.
3) If possible perform growth experiments to confirm the PULs role in metabolising its biochemically proven targets, the T and F antigens (Chapter IV)

CHAPTER VI

Final discussion

Many of the mechanisms underlying host microbial interactions still remain unclear and the difficulty with understanding the interactions of just a few clinically important members from the trillions-strong microbial communities inhabiting the human system as was the case in this study reflects just how much work there is still to be done in this area.

A large number of host-microbial interactions occurring at human mucosal surfaces involve mucosal surface components such as the extracellular secreted mucus and the epithelial cell glycocalyx which together constitute an important protective barrier against resident and foreign microbes (Ouwerkerk *et al.*, 2013). Mucosal microbes are known to encode biological molecules including binding or adhesion proteins and enzymes such as sulphatases, glycosidases and proteases that target and metabolize components of this protective barrier (McGuckin *et al.*, 2011, Chapter I). In the case of pathogenic mucosal microbes, these are regarded as colonization and virulence factors because of their role in aiding the organisms circumvent the barrier during infection (McGuckin *et al.*, 2011). Proteases and glycosidases that degrade integral mucus components such as mucins for this purpose for example have been reported in several pathogenic mucosal microbes (Wiggins *et al.*, 2001, McGuckin *et al.*, 2011). In non-invasive bacteria which constitute a significant proportion of the human gut microbiota microflora (Eckburg *et al.*, 2005, Dethlefsen *et al.*, 2007), the same elements are regarded as mere colonisation factors (McGuckin *et al.*, 2011)..

The novel family of M60-like domain-containing proteins were only recently discovered in 2012 (Nakjang *et al.*, 2012) and much of the focus in this study has been about investigating their role in the biology of mucosal microbes where they are more prominent. Our hypothesis that these proteins are surfaced-exposed glycoprotein targeted proteases in both pathogenic and non-pathogenic mucosal microbes was backed by evidence from data in Chapter III in the case of the gut commensal *B. thetaiotaomicron.* Their ability to degrade mucins and human myeloma IgA1 which are prominent components of secreted colonic mucus suggests an important role in mucosal colonisation. Investigating this for the pathogenic urogenital tract microbe *T. vaginalis* was however problematic, partly due to the difficulty in expressing soluble forms of these proteins in *E. coli.* The only solubly expressed proteins from this group where the M60-like

domain of TVAG199300 and the PA14 domain of TVAG339720. The recombinant M60-like domain however failed to degrade BSM and PGM glycoproteins contrary to several lines of evidence pointing to mucins as the most probable glycoprotein target for the proteins. Rather the PA14 domain of TVAG339720 in Chapter III was shown to bind heparin and its sulphated derivatives. This observation was only made after several failed ITC trials with mucin sugars. Although this was unanticipated, it will not be surprising that an invasive pathogen like *T. vaginalis* could rather be using these putative proteases against proteoglycan structures that constitute the glycocalyx of mucosal epithelial cells. Although there is need to further explore these findings, data from our collaborators in the USA are already suggesting that these proteins indeed play an important role in *T. vaginalis* pathogenesis.

In line with the overall aim of this study to contribute to our understanding of HMIs at mucosal surfaces, we also investigated putative functional partners of the BT4244 – M60-like protease from *B. thetaiotaomicron* in Chapter IV. This was not only important to further our understanding of the role of M60-like proteases in *B. thetaiotaomicron* which are all members of Sus-like systems in the organism, but also to improve our knowledge of the functioning of mucin targeted PULs, none of whose functional mechanism had been elucidated before this study.
Glycosylation is generally known to protect proteins against proteolysis (Russell *et al.*, 2009) and most mucin proteases produced by mucosal microbes are thought to target the less glycosylated regions of the glycoprotein (Wiggins *et al.*, 2001, Moncada *et al.*, 2003, Lidell *et al.*, 2006, Hasnain *et al.*, 2012). Hence we initially reasoned that the association of the BT4244 M60-like protease with the BT4240-50 Sus-like system in *B. thetaiotaomicron* is to exploit the deglycosylating activity of its associated glycoside hydrolases which eventually should ease access for the protease to the protein core. Rightly so, this is supported by evidence in Chapter III showing that desialylation of human myeloma IgA1 enhances the activity of the enzyme. However, *in-silico* cellular localisation data suggested that all BT4240-50 glycoside hydrolases are periplasmic while the BT4244 protein is surface localised implying that the protein rather initially comes in contact with the substrate before the glycoside hydrolases. Logically this did not appear very convincing, hence giving another important reason to biochemically characterise the entire PUL.

Based on the data presented in Chapter III, the PUL to which BT4244 belongs operates in a manner similar to the prototypic Sus system of *B. thetaiotaomicron*. Comparing the activities of various enzymes within the PUL to the Sus system, there was compelling evidence that BT4244

acts as a SusG-like protein. As a SusG-like protein, BT4244's role is likely to create internal cuts in complex extracellular mucin target to facilitate import through the SusCD complex (Section IV.4.2). Evidence that BT4240-50 glycoside hydrolases (BT4241 and BT4243) are downstream of BT4244 is also provided in Chapter IV, where it is shown that further deglycosylation of the protein by its functional partners BT4243 and BT4241 following desialylation of the IgA1 glycoprotein substrate does not enhance BT4244's IgA protease activity. BT4244 is indeed well adapted to perform its SusG-like function containing a GalNAc binding CBM32 domain to enable it trap the mucin glycoprotein and the M60-like protease domain that targets the glycosylated region of the glycoprotein. This is very similar to the prototypic SusG protein containing a CBM35 domain in addition to catalytic G site - containing domain (Cameron *et al.*, 2012 and Koropatkin *et al.,* 2009). Currently, efforts are still being made to identify target ligands for the BACON domains. There is also need to further characterise the BT4246 SusD protein of PULBT_4240-50. This is because it is not clear yet exactly what substrate is imported by the PUL into the periplasmic space of the cell envelop. Based on the data provided in Chapter IV, they are likely glycopeptide structures containing the F and T antigens. However, the recombinant BT4246 SusD positioned protein was only poorly saturated by the T antigen sugar in ITC binding experiments. It is worth noting that this protein, just like the GalNAc binding BT4244 and BT4245 proteins have all been shown to be highly upregulated during *B. thetaiotaomicron* growth on the T antigen (Martens *et al.*, 2008). Other considerations for future work on the PUL are listed in the discussion section of Chapter IV.

Finally in Chapter IV we sought to evaluate the importance of the PUL to which the BT4244 protease belongs through a series of gene deletion and *in-vitro* growth experiments with native mucins and their constituent sugars. Data from experiments with a strain containing a deletion to the complete BT4240-50 gene locus (PULBT_4240-50) under non-competitive conditions *in-vitro* on porcine gastric mucins revealed only a slight defect in the organism's ability to utilise the complex substrate. However, there was a significant effect on the organisms' ability to compete against the wild type *in-vitro* under the same conditions suggesting that the PUL is an important fitness factor. BT4244 deletion mutants on the other hand showed very limited growth defect in either condition. This was not very surprising given the massive expansion of mucin targeted genes in the organism (Martens *et al.*, 2008, Koropatkin *et al.*, 2010). The unanticipated loss of ΔPULBT_4240-50's ability to utilise the GalNAc monosaccharide came as a surprise and provided an opportunity to further explore GalNAc utilisation in *B. thetaiotaomicron*. GalNAc

utilisation in *B. thetaiotaomicron* and many other organisms in general is poorly understood and the proposed KEGG pathway for the metabolism of GalNAc (KEGG pathway ID: bth00052) still contains a lot of unanswered questions (Section V.4). Our data demonstrated that GalNAc phosphorylation is central to its utilisation in the organism and that the PUL contains a unique GalNAc kinase (BT4240) that performs this function, capable of phosphorylating GalNAc from several sources including from mucin and non mucin sources such as chondroitin sulphate.

References

Abbott, D. W., Eirin-Lopez, J. M., & Boraston, A. B. (2008). Insight into ligand diversity and novel biological roles for family 32 carbohydrate-binding modules. *Molecular Biology and Evolution, 25*(1), 155-167. doi:10.1093/molbev/msm243

Abdolrasouli, A., Amin, A., Baharsefat, M., Roushan, A., & Mofidi, S. (2007). Persistent urethritis and prostatitis due to *Trichomonas vaginalis*: A case report. *The Canadian Journal of Infectious Diseases & Medical Microbiology = Journal Canadien Des Maladies Infectieuses Et De La Microbiologie Medicale / AMMI Canada, 18*(5), 308-310.

Addis, M. F., Rappelli, P., & Fiori, P. L. (2000). Host and tissue specificity of *Trichomonas vaginalis* is not mediated by its known adhesion proteins. *Infection and Immunity, 68*(7), 4358-4360.

Aksoy, N., Corfield, A. P., Paraskeva, C., Thornton, D. J., & Sheehan, J. K. (2000). Alterations of MUC2 mucin in colorectal adenocarcinoma. *Turkish Journal of Medical Sciences, 30*(4), 359-366.

Allen, A., Hutton, D. A., & Pearson, J. P. (1998). The MUC2 gene product: A human intestinal mucin. *The International Journal of Biochemistry & Cell Biology, 30*(7), 797-801.

Allen, A., & Snary, D. (1972). The structure and function of gastric mucus. *Gut, 13*(8), 666-672.

Alsmark, C., Foster, P. G., Sicheritz-Ponten, T., Nakjang, S., Martin Embley, T., & Hirt, R. P. (2013). Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *Genome Biology, 14*(2), R19. doi:10.1186/gb-2013-14-2-r19

Alvarez-Sanchez, M. E., Avila-Gonzalez, L., Becerril-Garcia, C., Fattel-Facenda, L. V., Ortega-Lopez, J., & Arroyo, R. (2000). A novel cysteine proteinase (CP65) of *Trichomonas vaginalis* involved in cytotoxicity. *Microbial Pathogenesis, 28*(4), 193-202. doi:10.1006/mpat.1999.0336

Ambort, D., van der Post, S., Johansson, M. E., Mackenzie, J., Thomsson, E., Krengel, U., & Hansson, G. C. (2011). Function of the CysD domain of the gel-forming MUC2 mucin. *The Biochemical Journal, 436*(1), 61-70. doi:10.1042/BJ20102066; 10.1042/BJ20102066

Anderson, K., Li, S. C., & Li, Y. T. (2000). Diphenylamine-aniline-phosphoric acid reagent, a versatile spray reagent for revealing glycoconjugates on thin-layer chromatography plates. *Analytical Biochemistry, 287*(2), 337-339. doi:10.1006/abio.2000.4829

Andrianifahanana, M., Moniaux, N., & Batra, S. K. (2006). Regulation of mucin expression: Mechanistic aspects and implications for cancer and inflammatory diseases. *Biochimica Et Biophysica Acta, 1765*(2), 189-222. doi:10.1016/j.bbcan.2006.01.002

Ao, Z., Quezada-Calvillo, R., Sim, L., Nichols, B. L., Rose, D. R., Sterchi, E. E., & Hamaker, B. R. (2007). Evidence of native starch degradation with human small intestinal maltase-glucoamylase (recombinant). *FEBS Letters, 581*(13), 2381-2388. doi:10.1016/j.febslet.2007.04.035

Arai, T., Fujita, K., Fujime, M., & Irimura, T. (2005). Expression of sialylated MUC1 in prostate cancer: Relationship to clinical stage and prognosis. *International Journal of Urology : Official Journal of the Japanese Urological Association, 12*(7), 654-661. doi:10.1111/j.1442-2042.2005.01112.x

Aristoteli, L. P., & Willcox, M. D. (2003). Mucin degradation mechanisms by distinct pseudomonas aeruginosa isolates in vitro. *Infection and Immunity, 71*(10), 5565-5575.

Arnold K., Bordoli L., Kopp J., and Schwede T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics, 22,195-201.

Ashida, H., Anderson, K., Nakayama, J., Maskos, K., Chou, C. W., Cole, R. B., . . . Li, Y. T. (2001). A novel endo-beta-galactosidase from clostridium perfringens that liberates the disaccharide GlcNAcalpha 1-->Gal from glycans specifically expressed in the gastric gland mucous cell-type mucin. *The Journal of Biological Chemistry, 276*(30), 28226-28232. doi:10.1074/jbc.M103589200

Ashida, H., Maki, R., Ozawa, H., Tani, Y., Kiyohara, M., Fujita, M., . . . Yamamoto, K. (2008). Characterization of two different endo-alpha-N-acetylgalactosaminidases from probiotic and pathogenic enterobacteria, bifidobacterium longum and clostridium perfringens. *Glycobiology, 18*(9), 727-734. doi:10.1093/glycob/cwn053; 10.1093/glycob/cwn053

Asker, N., Axelsson, M. A., Olofsson, S. O., & Hansson, G. C. (1998). Dimerization of the human MUC2 mucin in the endoplasmic reticulum is followed by a N-glycosylation-dependent transfer of the mono- and dimers to the golgi apparatus. *The Journal of Biological Chemistry, 273*(30), 18857-18863.

Atuma, C., Strugala, V., Allen, A., & Holm, L. (2001). The adherent gastrointestinal mucus gel layer: Thickness and physical state in vivo. *American Journal of Physiology.Gastrointestinal and Liver Physiology, 280*(5), G922-9.

Babu, M. M., Priya, M. L., Selvan, A. T., Madera, M., Gough, J., Aravind, L., & Sankaran, K. (2006). A database of bacterial lipoproteins (DOLOP) with functional assignments to predicted lipoproteins. *Journal of Bacteriology, 188*(8), 2761-2773. doi:10.1128/JB.188.8.2761-2773.2006

Backstrom, M., Thomsson, K. A., Karlsson, H., & Hansson, G. C. (2009). Sensitive liquid chromatography-electrospray mass spectrometry allows for the analysis of the O-glycosylation of immunoprecipitated proteins from cells or tissues: Application to MUC1 glycosylation in cancer. *Journal of Proteome Research, 8*(2), 538-545. doi:10.1021/pr800713h; 10.1021/pr800713h

Bansil, R., & Turner, B. S. (2006). Mucin structure, aggregation, physiological functions and biomedical applications. *Current Opinion in Colloid & Interface Science, 11*(2–3), 164-170. doi:http://dx.doi.org/10.1016/j.cocis.2005.11.001

Barnich, N., Carvalho, F. A., Glasser, A. L., Darcha, C., Jantscheff, P., Allez, M., . . . Darfeuille-Michaud, A. (2007). CEACAM6 acts as a receptor for adherent-invasive E. coli, supporting

ileal mucosa colonization in crohn disease. *The Journal of Clinical Investigation, 117*(6), 1566-1574. doi:10.1172/JCI30504

Baruch, A., Hartmann, M., Yoeli, M., Adereth, Y., Greenstein, S., Stadler, Y., . . . Wreschner, D. H. (1999). The breast cancer-associated MUC1 gene generates both a receptor and its cognate binding protein. *Cancer Research, 59*(7), 1552-1561.

Bastida-Corcuera, F. D., Okumura, C. Y., Colocoussi, A., & Johnson, P. J. (2005). *Trichomonas vaginalis* lipophosphoglycan mutants have reduced adherence and cytotoxicity to human ectocervical cells. *Eukaryotic Cell, 4*(11), 1951-1958. doi:10.1128/EC.4.11.1951-1958.2005

Bell, S. L., Xu, G., & Forstner, J. F. (2001). Role of the cystine-knot motif at the C-terminus of rat mucin protein Muc2 in dimer formation and secretion. *The Biochemical Journal, 357*(Pt 1), 203-209.

Bell, S. L., Xu, G., Khatri, I. A., Wang, R., Rahman, S., & Forstner, J. F. (2003). N-linked oligosaccharides play a role in disulphide-dependent dimerization of intestinal mucin Muc2. *The Biochemical Journal, 373*(Pt 3), 893-900. doi:10.1042/BJ20030096

Benjdia, A., Martens, E. C., Gordon, J. I., & Berteau, O. (2011). Sulfatases and a radical S-adenosyl-L-methionine (AdoMet) enzyme are key for mucosal foraging and fitness of the prominent human gut symbiont, *bacteroides thetaiotaomicron*. *The Journal of Biological Chemistry, 286*(29), 25973-25982. doi:10.1074/jbc.M111.228841; 10.1074/jbc.M111.228841

Bergman, E. N. (1990). Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. *Physiological Reviews, 70*(2), 567-590.

Bertin, Y., Chaucheyras-Durand, F., Robbe-Masselot, C., Durand, A., de la Foye, A., Harel, J., . . . Martin, C. (2013). Carbohydrate utilization by enterohaemorrhagic escherichia coli O157:H7 in bovine intestinal content. *Environmental Microbiology, 15*(2), 610-622. doi:10.1111/1462-2920.12019; 10.1111/1462-2920.12019

Bewsey, K. E., Johnson, M. E., & Huff, J. P. (1991). Rapid isolation and purification of DNA from agarose gels: The phenol-freeze-fracture method. *Biotechniques, 10*(6), 724-725.

Bolam, D. N., Xie, H., White, P., Simpson, P. J., Hancock, S. M., Williamson, M. P., & Gilbert, H. J. (2001). Evidence for synergy between family 2b carbohydrate binding modules in cellulomonas fimi xylanase 11A. *Biochemistry, 40*(8), 2468-2477.

Boraston, A. B., Bolam, D. N., Gilbert, H. J., & Davies, G. J. (2004). Carbohydrate-binding modules: Fine-tuning polysaccharide recognition. *The Biochemical Journal, 382*(Pt 3), 769-781. doi:10.1042/BJ20040892

Boraston, A. B., Ficko-Blean, E., & Healey, M. (2007). Carbohydrate recognition by a large sialidase toxin from clostridium perfringens. *Biochemistry, 46*(40), 11352-11360. doi:10.1021/bi701317g

Boris, S., Suarez, J. E., Vazquez, F., & Barbes, C. (1998). Adherence of human vaginal lactobacilli to vaginal epithelial cells and interaction with uropathogens. *Infection and Immunity, 66*(5), 1985-1989.

Boshell, M., Lalani, E., Pemberton, L., Burchell, J., Gendler, S., & Taylor-Papadimitriou, J. (1992). The product of the human MUC1 gene when secreted by mouse cells transfected with the full-length cDNA lacks the cytoplasmic tail. *Biochemical and Biophysical Research Communications, 185*(1), 1-8. doi:http://dx.doi.org/10.1016/S0006-291X(05)80946-5

Braun-Fahrlander, C., Riedler, J., Herz, U., Eder, W., Waser, M., Grize, L., . . . Allergy and Endotoxin Study Team. (2002). Environmental exposure to endotoxin and its relation to asthma in school-age children. *The New England Journal of Medicine, 347*(12), 869-877. doi:10.1056/NEJMoa020057

Brayman, M., Thathiah, A., & Carson, D. D. (2004). MUC1: A multifunctional cell surface component of reproductive tissue epithelia. *Reproductive Biology and Endocrinology : RB&E, 2*, 4. doi:10.1186/1477-7827-2-4

Brockhausen, I. (2006). Mucin-type O-glycans in human colon and breast cancer: Glycodynamics and functions. *EMBO Reports, 7*(6), 599-604. doi:10.1038/sj.embor.7400705

Brockhausen, I., Yang, J. M., Burchell, J., Whitehouse, C., & Taylor-Papadimitriou, J. (1995). Mechanisms underlying aberrant glycosylation of MUC1 mucin in breast cancer cells. *European Journal of Biochemistry / FEBS, 233*(2), 607-617.

Brown, D., & Waneck, G. L. (1992). Glycosyl-phosphatidylinositol-anchored membrane proteins. *Journal of the American Society of Nephrology : JASN, 3*(4), 895-906.

Buleon, A., Colonna, P., Planchot, V., & Ball, S. (1998). Starch granules: Structure and biosynthesis. *International Journal of Biological Macromolecules, 23*(2), 85-112.

Cameron, E. A., Maynard, M. A., Smith, C. J., Smith, T. J., Koropatkin, N. M., & Martens, E. C. (2012). Multidomain carbohydrate-binding proteins involved in *bacteroides thetaiotaomicron* starch metabolism. *The Journal of Biological Chemistry, 287*(41), 34614-34625. doi:10.1074/jbc.M112.397380; 10.1074/jbc.M112.397380

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. 2009 Jan;37(Database issue):D233-8. http://www.ncbi.nlm.nih.gov/pubmed/18838391

Carciofi, M., Blennow, A., Jensen, S. L., Shaik, S. S., Henriksen, A., Buleon, A., . . . Hebelstrup, K. H. (2012). Concerted suppression of all starch branching enzyme genes in barley produces amylose-only starch granules. *BMC Plant Biology, 12*, 223-2229-12-223. doi:10.1186/1471-2229-12-223; 10.1186/1471-2229-12-223

Carlsson, P., & Kjellen, L. (2012). Heparin biosynthesis. *Handbook of Experimental Pharmacology, (207):23-41. doi*(207), 23-41. doi:10.1007/978-3-642-23056-1_2; 10.1007/978-3-642-23056-1_2

Carlton, J. M., Hirt, R. P., Silva, J. C., Delcher, A. L., Schatz, M., Zhao, Q., . . . Johnson, P. J. (2007). Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science (New York, N.Y.), 315*(5809), 207-212. doi:10.1126/science.1132894

Carson, C. A., Christiansen, J. M., Yampara-Iquise, H., Benson, V. W., Baffaut, C., Davis, J. V., . . . Fales, W. H. (2005). Specificity of a *bacteroides thetaiotaomicron* marker for human feces. *Applied and Environmental Microbiology, 71*(8), 4945-4949. doi:10.1128/AEM.71.8.4945-4949.2005

Carson, D. D., DeSouza, M. M., Kardon, R., Zhou, X., Lagow, E., & Julian, J. (1998). Mucin expression and function in the female reproductive tract. *Human Reproduction Update, 4*(5), 459-464.

Cebra, J. J. (1999). Influences of microbiota on intestinal immune system development. *The American Journal of Clinical Nutrition, 69*(5), 1046S-1051S.

Celli, J., Gregor, B., Turner, B., Afdhal, N. H., Bansil, R., & Erramilli, S. (2005). Viscoelastic properties and dynamics of porcine gastric mucin. *Biomacromolecules, 6*(3), 1329-1333. doi:10.1021/bm0493990

Cervantes-Sandoval, I., Serrano-Luna Jde, J., Garcia-Latorre, E., Tsutsumi, V., & Shibayama, M. (2008). Mucins in the host defence against naegleria fowleri and mucinolytic activity as a possible means of evasion. *Microbiology (Reading, England), 154*(Pt 12), 3895-3904. doi:10.1099/mic.0.2008/019380-0; 10.1099/mic.0.2008/019380-0

Chen, Y., Zhao, Y. H., Kalaslavadi, T. B., Hamati, E., Nehrke, K., Le, A. D., . . . Wu, R. (2004). Genome-wide search and identification of a novel gel-forming mucin MUC19/Muc19 in glandular tissues. *American Journal of Respiratory Cell and Molecular Biology, 30*(2), 155-165. doi:10.1165/rcmb.2003-0103OC

Cho, K. H., Cho, D., Wang, G. R., & Salyers, A. A. (2001). New regulatory gene that contributes to control of *bacteroides thetaiotaomicron* starch utilization genes. *Journal of Bacteriology, 183*(24), 7198-7205. doi:10.1128/JB.183.24.7198-7205.2001

Cho, K. H., & Salyers, A. A. (2001). Biochemical analysis of interactions between outer membrane proteins that contribute to starch utilization by *bacteroides thetaiotaomicron*. *Journal of Bacteriology, 183*(24), 7224-7230.

Chou, K. C., & Elrod, D. W. (1999). Prediction of membrane protein types and subcellular locations. *Proteins, 34*(1), 137-153.

Claus, S. P., Ellero, S. L., Berger, B., Krause, L., Bruttin, A., Molina, J., . . . Nicholson, J. K. (2011). Colonization-induced host-gut microbial metabolic interaction. *Mbio*, 2(2), e00271-10. doi:10.1128/mBio.00271-10; 10.1128/mBio.00271-10

Clemente, J. C., Ursell, L. K., Parfrey, L. W., & Knight, R. (2012). The impact of the gut microbiota on human health: An integrative view. *Cell, 148*(6), 1258-1270. doi:10.1016/j.cell.2012.01.035; 10.1016/j.cell.2012.01.035

Cohen, S. N., Chang, A. C., & Hsu, L. (1972). Nonchromosomal antibiotic resistance in bacteria: Genetic transformation of escherichia coli by R-factor DNA. *Proceedings of the National Academy of Sciences of the United States of America, 69*(8), 2110-2114.

Coleman, J. E. (1998). Zinc enzymes. *Current Opinion in Chemical Biology, 2*(2), 222-234.

Colina, A. R., Aumont, F., Deslauriers, N., Belhumeur, P., & de Repentigny, L. (1996). Evidence for degradation of gastrointestinal mucin by candida albicans secretory aspartyl proteinase. *Infection and Immunity, 64*(11), 4514-4519.

Conze, T., Carvalho, A. S., Landegren, U., Almeida, R., Reis, C. A., David, L., & Soderberg, O. (2010). MUC2 mucin is a major carrier of the cancer-associated sialyl-tn antigen in intestinal metaplasia and gastric carcinomas. *Glycobiology, 20*(2), 199-206. doi:10.1093/glycob/cwp161; 10.1093/glycob/cwp161

Crago, S. S., Kutteh, W. H., Moro, I., Allansmith, M. R., Radl, J., Haaijman, J. J., & Mestecky, J. (1984). Distribution of IgA1-, IgA2-, and J chain-containing cells in human tissues. *Journal of Immunology (Baltimore, Md.: 1950), 132*(1), 16-18.

Cummings, J. H. (1981). Short chain fatty acids in the human colon. *Gut, 22*(9), 763-779.

Dalbey RE, Wang P, van Dijl JM (2012). Membrane proteases in the bacterial protein secretion and quality control pathway. *Microbiology and molecular biology reviews*,*76*(2),311-30.doi: 10.1128/MMBR.05019-11

Davies, G. J., & Sinnott, M. L. (2008). Sorting the diverse: The sequence-based classifications of carbohydrate-active enzymes. *Biochemical Journal.* doi:10.1042/BJ20080382

de Miguel, N., Lustig, G., Twu, O., Chattopadhyay, A., Wohlschlegel, J. A., & Johnson, P. J. (2010). Proteome analysis of the surface of *Trichomonas vaginalis* reveals novel proteins and strain-dependent differential expression. *Molecular & Cellular Proteomics : MCP, 9*(7), 1554-1566. doi:10.1074/mcp.M000022-MCP201; 10.1074/mcp.M000022-MCP201

Delacroix, D. L., Dive, C., Rambaud, J. C., & Vaerman, J. P. (1982). IgA subclasses in various secretions and in serum. *Immunology, 47*(2), 383-385.

D'Elia, J. N., & Salyers, A. A. (1996). Contribution of a neopullulanase, a pullulanase, and an alpha-glucosidase to growth of *bacteroides thetaiotaomicron* on starch. *Journal of Bacteriology, 178*(24), 7173-7179.

Derensy-Dron, D., Krzewinski, F., Brassart, C., & Bouquelet, S. (1999). Beta-1,3-galactosyl-N-acetylhexosamine phosphorylase from bifidobacterium bifidum DSM 20082: Characterization, partial purification and relation to mucin degradation. *Biotechnology and Applied Biochemistry, 29 (Pt 1)*(Pt 1), 3-10.

DeSouza, M. M., Surveyor, G. A., Price, R. E., Julian, J., Kardon, R., Zhou, X., . . . Carson, D. D. (1999). MUC1/episialin: A critical barrier in the female reproductive tract. *Journal of Reproductive Immunology, 45*(2), 127-158.

Dethlefsen, L., McFall-Ngai, M., & Relman, D. A. (2007). An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature, 449*(7164), 811-818. doi:10.1038/nature06245

Dharmani, P., Srivastava, V., Kissoon-Singh, V., & Chadee, K. (2009). Role of intestinal mucins in innate host defense mechanisms against pathogens. *Journal of Innate Immunity, 1*(2), 123-135. doi:10.1159/000163037

Dick, G., Akslen-Hoel, L. K., Grondahl, F., Kjos, I., & Prydz, K. (2012). Proteoglycan synthesis and golgi organization in polarized epithelial cells. *The Journal of Histochemistry and Cytochemistry : Official Journal of the Histochemistry Society, 60*(12), 926-935. doi:10.1369/0022155412461256; 10.1369/0022155412461256

Din, N., Gilkes, N. R., Tekant, B., Miller, R. C., Warren, R. A. J., & Kilburn, D. G. (1991). Non–Hydrolytic disruption of cellulose fibres by the binding domain of a bacterial cellulase. *Nature Biotechnology, 9*(11), 1096–9.

Dohrman, A., Miyata, S., Gallup, M., Li, J. D., Chapelin, C., Coste, A., . . . Basbaum, C. (1998). Mucin gene (MUC 2 and MUC 5AC) upregulation by gram-positive and gram-negative bacteria. *Biochimica Et Biophysica Acta, 1406*(3), 251-259.

Dos Santos, A. L. (2011). Protease expression by microorganisms and its relevance to crucial physiological/pathological events. *World Journal of Biological Chemistry, 2*(3), 48-58. doi:10.4331/wjbc.v2.i3.48; 10.4331/wjbc.v2.i3.48

DPDx. (2009). Life Cycle, *Trichomonas vaginalis*. Available: http://dpd.cdc.gov/dpdx/HTML/Trichomoniasis.htm. Last accessed 26/08/2013.

Du, Y., He, Y. X., Zhang, Z. Y., Yang, Y. H., Shi, W. W., Frolet, C., . . . Chen, Y. (2011). Crystal structure of the mucin-binding domain of Spr1345 from streptococcus pneumoniae. *Journal of Structural Biology, 174*(1), 252-257. doi:10.1016/j.jsb.2010.10.016

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., . . . Relman, D. A. (2005). Diversity of the human intestinal microbial flora. *Science (New York, N.Y.), 308*(5728), 1635-1638. doi:10.1126/science.1110591

Edwards, R. (1997). Resistance to beta-lactam antibiotics in bacteroides spp. *Journal of Medical Microbiology, 46*(12), 979-986.

Egberts, H. J., Koninkx, J. F., van Dijk, J. E., & Mouwen, J. M. (1984). Biological and pathobiological aspects of the glycocalyx of the small intestinal epithelium. A review. *The Veterinary Quarterly, 6*(4), 186-199. doi:10.1080/01652176.1984.9693936

Eisenhaber, B., Bork, P., & Eisenhaber, F. (2001). Post-translational GPI lipid anchor modification of proteins in kingdoms of life: Analysis of protein sequence data from complete genomes. *Protein Engineering, 14*(1), 17-25.

Elofsson, A., & von Heijne, G. (2007). Membrane protein structure: Prediction versus reality. *Annual Review of Biochemistry, 76*, 125-140. doi:10.1146/annurev.biochem.76.052705.163539

Embley, T. M., van der Giezen, M., Horner, D. S., Dyal, P. L., Bell, S., & Foster, P. G. (2003). Hydrogenosomes, mitochondria and early eukaryotic evolution. *IUBMB Life, 55*(7), 387-395. doi:10.1080/15216540310001592834

Enerback, L., Kolset, S. O., Kusche, M., Hjerpe, A., & Lindahl, U. (1985). Glycosaminoglycans in rat mucosal mast cells. *The Biochemical Journal, 227*(2), 661-668.

Engbring, J. A., & Alderete, J. F. (1998). Characterization of *Trichomonas vaginalis* AP33 adhesin and cell surface interactive domains. *Microbiology (Reading, England), 144 (Pt 11)*(Pt 11), 3011-3018.

Engelmann, K., Kinlough, C. L., Muller, S., Razawi, H., Baldus, S. E., Hughey, R. P., & Hanisch, F. G. (2005). Transmembrane and secreted MUC1 probes show trafficking-dependent changes in O-glycan core profiles. *Glycobiology, 15*(11), 1111-1124. doi:10.1093/glycob/cwi099

Englyst, H. N., & Macfarlane, G. T. (1986). Breakdown of resistant and readily digestible starch by human gut bacteria. *Journal of the Science of Food and Agriculture, 37*(7), 699-706. doi:10.1002/jsfa.2740370717

Esko, J. D., Kimata, K., & Lindahl, U. (2009). Proteoglycans and sulfated glycosaminoglycans. In A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, . . . M. E. Etzler (Eds.), *Essentials of glycobiology* (2nd ed.,). Cold Spring Harbor (NY): The Consortium of Glycobiology Editors, La Jolla, California.

Ferguson, A. D., & Deisenhofer, J. (2002). TonB-dependent receptors-structural perspectives. *Biochimica Et Biophysica Acta, 1565*(2), 318-332.

Fernandez, M., Liu, X., Wouters, M. A., Heyberger, S., & Husain, A. (2001). Angiotensin I-converting enzyme transition state stabilization by HIS1089: Evidence for a catalytic mechanism distinct from other gluzincin metalloproteinases. *The Journal of Biological Chemistry, 276*(7), 4998-5004. doi:10.1074/jbc.M009009200

Ficko-Blean, E., & Boraston, A. B. (2006). The interaction of a carbohydrate-binding module from a clostridium perfringens N-acetyl-beta-hexosaminidase with its carbohydrate receptor. *The Journal of Biological Chemistry, 281*(49), 37748-37757. doi:10.1074/jbc.M606126200

Ficko-Blean, E., & Boraston, A. B. (2009). N-acetylglucosamine recognition by a family 32 carbohydrate-binding module from clostridium perfringens NagH. *Journal of Molecular Biology, 390*(2), 208-220. doi:10.1016/j.jmb.2009.04.066

Ficko-Blean, E., Stuart, C. P., Suits, M. D., Cid, M., Tessier, M., Woods, R. J., & Boraston, A. B. (2012). Carbohydrate recognition by an architecturally complex alpha-N-acetylglucosaminidase from clostridium perfringens. *PloS One, 7*(3), e33524. doi:10.1371/journal.pone.0033524; 10.1371/journal.pone.0033524

Fiebrig, I., Harding, S. E., Rowe, A. J., Hyman, S. C., & Davis, S. S. (1995). Transmission electron microscopy studies on pig gastric mucin and its interactions with chitosan. *Carbohydrate Polymers, 28*(3), 239-244. doi:http://dx.doi.org/10.1016/0144-8617(95)00105-0

Finkelstein, R. A., Boesman-Finkelstein, M., & Holt, P. (1983). Vibrio cholerae hemagglutinin/lectin/protease hydrolyzes fibronectin and ovomucin: F.M. burnet revisited. *Proceedings of the National Academy of Sciences of the United States of America, 80*(4), 1092-1095.

Franco, A. A., Buckwold, S. L., Shin, J. W., Ascon, M., & Sears, C. L. (2005). Mutation of the zinc-binding metalloprotease motif affects bacteroides fragilis toxin activity but does not affect propeptide processing. *Infection and Immunity, 73*(8), 5273-5277. doi:10.1128/IAI.73.8.5273-5277.2005

Freymond, P. P., Lazarevic, V., Soldo, B., & Karamata, D. (2006). Poly(glucosyl-N-acetylgalactosamine 1-phosphate), a wall teichoic acid of bacillus subtilis 168: Its biosynthetic pathway and mode of attachment to peptidoglycan. *Microbiology (Reading, England), 152*(Pt 6), 1709-1718. doi:10.1099/mic.0.28814-0

Fukasawa, K. M., Hata, T., Ono, Y., & Hirose, J. (2011). Metal preferences of zinc-binding motif on metalloproteases. *Journal of Amino Acids, 2011*, 574816. doi:10.4061/2011/574816; 10.4061/2011/574816

Gerbase, A. C., Rowley, J. T., Heymann, D. H., Berkley, S. F., & Piot, P. (1998). Global prevalence and incidence estimates of selected curable STDs. *Sexually Transmitted Infections, 74 Suppl 1*, S12-6.

Gerber, L. D., Kodukula, K., & Udenfriend, S. (1992). Phosphatidylinositol glycan (PI-G) anchored membrane proteins. amino acid requirements adjacent to the site of cleavage and PI-G attachment in the COOH-terminal signal peptide. *The Journal of Biological Chemistry, 267*(17), 12168-12173.

Gerken, T. A. (1993). Biophysical approaches to salivary mucin structure, conformation and dynamics. *Critical Reviews in Oral Biology and Medicine : An Official Publication of the American Association of Oral Biologists, 4*(3-4), 261-270.

Gilbert, J. V., Plaut, A. G., & Wright, A. (1991). Analysis of the immunoglobulin A protease gene of streptococcus sanguis. *Infection and Immunity, 59*(1), 7-17.

Gilkes, N. R., Warren, R. A., Miller, R. C.,Jr, & Kilburn, D. G. (1988). Precise excision of the cellulose binding domains from two cellulomonas fimi cellulases by a homologous protease and the effect on catalysis. *The Journal of Biological Chemistry, 263*(21), 10401-10407.

Gillespie, J. J., Wattam, A. R., Cammer, S. A., Gabbard, J. L., Shukla, M. P., Dalay, O., . . . Sobral, B. W. (2011). PATRIC: The comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infection and Immunity, 79*(11), 4286-4298. doi:10.1128/IAI.00207-11; 10.1128/IAI.00207-11

Godl, K., Johansson, M. E., Lidell, M. E., Morgelin, M., Karlsson, H., Olson, F. J., . . . Hansson, G. C. (2002). The N terminus of the MUC2 mucin forms trimers that are held together within a trypsin-resistant core fragment. *The Journal of Biological Chemistry, 277*(49), 47248-47256. doi:10.1074/jbc.M208483200

Goldstein, E. J. (1996). Anaerobic bacteremia. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America, 23 Suppl 1*, S97-101.

Gomis-Ruth, F. X. (2003). Structural aspects of the metzincin clan of metalloendopeptidases. *Molecular Biotechnology, 24*(2), 157-202. doi:10.1385/MB:24:2:157

Gouet, P., Courcelle, E., Stuart, D.I. and Metoz, F. (1999). ESPript: multiple sequence alignments in PostScript. *Bioinformatics*. 15 305-8

Grys, T. E., Siegel, M. B., Lathem, W. W., & Welch, R. A. (2005). The StcE protease contributes to intimate adherence of enterohemorrhagic escherichia coli O157:H7 to host cells. *Infection and Immunity, 73*(3), 1295-1303. doi:10.1128/IAI.73.3.1295-1303.2005

Guillen, D., Sanchez, S., & Rodriguez-Sanoja, R. (2010). Carbohydrate-binding domains: Multiplicity of biological roles. *Applied Microbiology and Biotechnology, 85*(5), 1241-1249. doi:10.1007/s00253-009-2331-y; 10.1007/s00253-009-2331-y

Gutierrez-Jimenez, J., Arciniega, I., & Navarro-Garcia, F. (2008). The serine protease motif of pic mediates a dose-dependent mucolytic activity after binding to sugar constituents of the mucin substrate. *Microbial Pathogenesis, 45*(2), 115-123. doi:10.1016/j.micpath.2008.04.006; 10.1016/j.micpath.2008.04.006

Hanisch, F. G., & Muller, S. (2000). MUC1: The polymorphic appearance of a human mucin. *Glycobiology, 10*(5), 439-449.

Harp, D. F., & Chowdhury, I. (2011). Trichomoniasis: Evaluation to execution. *European Journal of Obstetrics, Gynecology, and Reproductive Biology, 157*(1), 3-9. doi:10.1016/j.ejogrb.2011.02.024; 10.1016/j.ejogrb.2011.02.024

Harrison R J (2013). Encyclopædia Britannica Online. Encyclopædia Britannica Inc. accessed August 26, 2013,

Hasnain, S. Z., McGuckin, M. A., Grencis, R. K., & Thornton, D. J. (2012). Serine protease(s) secreted by the nematode trichuris muris degrade the mucus barrier. *PLoS Neglected Tropical Diseases, 6*(10), e1856. doi:10.1371/journal.pntd.0001856; 10.1371/journal.pntd.0001856

Hattrup, C. L., & Gendler, S. J. (2008). Structure and function of the cell surface (tethered) mucins. *Annual Review of Physiology, 70*, 431-457. doi:10.1146/annurev.physiol.70.113006.100659

Henke, M. O., John, G., Rheineck, C., Chillappagari, S., Naehrlich, L., & Rubin, B. K. (2011). Serine proteases degrade airway mucins in cystic fibrosis. *Infection and Immunity, 79*(8), 3438-3444. doi:10.1128/IAI.01252-10; 10.1128/IAI.01252-10

Hermoso, J. A., Garcia, J. L., & Garcia, P. (2007). Taking aim on bacterial pathogens: From phage therapy to enzybiotics. *Current Opinion in Microbiology, 10*(5), 461-472. doi:10.1016/j.mib.2007.08.002

Hernandez-Gutierrez, R., Avila-Gonzalez, L., Ortega-Lopez, J., Cruz-Talonia, F., Gomez-Gutierrez, G., & Arroyo, R. (2004). *Trichomonas vaginalis*: Characterization of a 39-kDa cysteine proteinase found in patient vaginal secretions. *Experimental Parasitology, 107*(3-4), 125-135. doi:10.1016/j.exppara.2004.05.004

Herrero, M., de Lorenzo, V., & Timmis, K. N. (1990). Transposon vectors containing non-antibiotic resistance selection markers for cloning and stable chromosomal insertion of foreign genes in gram-negative bacteria. *Journal of Bacteriology, 172*(11), 6557-6567.

Herve, C., Rogowski, A., Blake, A. W., Marcus, S. E., Gilbert, H. J., & Knox, J. P. (2010). Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. *Proceedings of the National Academy of Sciences of the United States of America, 107*(34), 15293-15298. doi:10.1073/pnas.1005732107; 10.1073/pnas.1005732107

Hirt, R. P., Noel, C. J., Sicheritz-Ponten, T., Tachezy, J., & Fiori, P. L. (2007). *Trichomonas vaginalis* surface proteins: A view from the genome. *Trends in Parasitology, 23*(11), 540-547. doi:10.1016/j.pt.2007.08.020

HOLLANDER, F. (1963). The electrolyte pattern of gastric mucinous secretions: Its implication for cystic fibrosis. *Annals of the New York Academy of Sciences, 106*, 757-766.

Hooper, L. V. (2009). Do symbiotic bacteria subvert host immunity? *Nature Reviews.Microbiology, 7*(5), 367-374. doi:10.1038/nrmicro2114

Hooper, L. V., & Gordon, J. I. (2001). Glycans as legislators of host-microbial interactions: Spanning the spectrum from symbiosis to pathogenicity. *Glycobiology, 11*(2), 1R-10R.

Hooper, L. V., Midtvedt, T., & Gordon, J. I. (2002). How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annual Review of Nutrition, 22*, 283-307. doi:10.1146/annurev.nutr.22.011602.092259

Hooper, L. V., Stappenbeck, T. S., Hong, C. V., & Gordon, J. I. (2003). Angiogenins: A new class of microbicidal proteins involved in innate immunity. *Nature Immunology, 4*(3), 269-273. doi:10.1038/ni888

Humphries, D. E., Wong, G. W., Friend, D. S., Gurish, M. F., Qiu, W. T., Huang, C., . . . Stevens, R. L. (1999). Heparin is essential for the storage of specific granule proteases in mast cells. *Nature, 400*(6746), 769-772. doi:10.1038/23481

Hurlimann, J., & Darling, H. (1971). In vitro synthesis of immunoglobulin-A by salivary glands from animals of different species. *Immunology, 21*(1), 101-111.

Hytonen, J., Haataja, S., Gerlach, D., Podbielski, A., & Finne, J. (2001). The SpeB virulence factor of streptococcus pyogenes, a multifunctional secreted and cell surface molecule with strepadhesin, laminin-binding and cysteine protease activity. *Molecular Microbiology, 39*(2), 512-519.

Inouye, S., Wang, S., Sekizawa, J., Halegoua, S., & Inouye, M. (1977). Amino acid sequence for the peptide extension on the prolipoprotein of the escherichia coli outer membrane. *Proceedings of the National Academy of Sciences of the United States of America, 74*(3), 1004-1008.

Ishiguro, K., Baba, E., Torii, R., Tamada, H., Kawate, N., Hatoya, S., . . . Inaba, T. (2007). Reduction of mucin-1 gene expression associated with increased escherichia coli adherence in the canine uterus in the early stage of dioestrus. *Veterinary Journal (London, England : 1997), 173*(2), 325-332. doi:10.1016/j.tvjl.2005.11.009

Jedrzejas, M. J. (2001). Pneumococcal virulence factors: Structure and function. *Microbiology and Molecular Biology Reviews : MMBR, 65*(2), 187-207 ; first page, table of contents. doi:10.1128/MMBR.65.2.187-207.2001

Jensen, P. H., Kolarich, D., & Packer, N. H. (2010). Mucin-type O-glycosylation--putting the pieces together. *The FEBS Journal, 277*(1), 81-94. doi:10.1111/j.1742-4658.2009.07429.x

Jeong, J. K., Kwon, O., Lee, Y. M., Oh, D. B., Lee, J. M., Kim, S., . . . Kang, H. A. (2009). Characterization of the streptococcus pneumoniae BgaC protein as a novel surface beta-galactosidase with specific hydrolysis activity for the Galbeta1-3GlcNAc moiety of oligosaccharides. *Journal of Bacteriology, 191*(9), 3011-3023. doi:10.1128/JB.01601-08; 10.1128/JB.01601-08

Jiang, W., Gupta, D., Gallagher, D., Davis, S., & Bhavanandan, V. P. (2000). The central domain of bovine submaxillary mucin consists of over 50 tandem repeats of 329 amino acids. chromosomal localization of the BSM1 gene and relations to ovine and porcine counterparts. *European Journal of Biochemistry / FEBS, 267*(8), 2208-2217.

Johansson, M. E., Larsson, J. M., & Hansson, G. C. (2011). The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. *Proceedings of the National Academy of Sciences of the United States of America, 108 Suppl 1*, 4659-4665. doi:10.1073/pnas.1006451107

Johansson, M. E., Phillipson, M., Petersson, J., Velcich, A., Holm, L., & Hansson, G. C. (2008). The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Proceedings of the National Academy of Sciences of the United States of America, 105*(39), 15064-15069. doi:10.1073/pnas.0803124105

Juarez Tomas, M. S., Ocana, V. S., Wiese, B., & Nader-Macias, M. E. (2003). Growth and lactic acid production by vaginal lactobacillus acidophilus CRL 1259, and inhibition of uropathogenic escherichia coli. *Journal of Medical Microbiology, 52*(Pt 12), 1117-1124.

Juliano, C., Cappuccinelli, P., & Mattana, A. (1991). In vitro phagocytic interaction between *Trichomonas vaginalis* isolates and bacteria. *European Journal of Clinical Microbiology & Infectious Diseases : Official Publication of the European Society of Clinical Microbiology, 10*(6), 497-502.

Kahel-Raifer, H., Jindou, S., Bahari, L., Nataf, Y., Shoham, Y., Bayer, E. A., . . . Lamed, R. (2010). The unique set of putative membrane-associated anti-sigma factors in clostridium thermocellum suggests a novel extracellular carbohydrate-sensing mechanism involved in

gene regulation. *FEMS Microbiology Letters, 308*(1), 84-93. doi:10.1111/j.1574-6968.2010.01997.x; 10.1111/j.1574-6968.2010.01997.x

Kaiserlian, D., Cerf-Bensussan, N., & Hosmalin, A. (2005). The mucosal immune system: From control of inflammation to protection against infections. *Journal of Leukocyte Biology, 78*(2), 311-318. doi:10.1189/jlb.0105053

Kall L, Anders K, and Erik L. Sonnhammer L.A Combined Transmembrane Topology and Signal Peptide Prediction Method.*Journal of Molecular Biology*, 338(5):1027-1036, May 2004.

Kamada, N., Seo, S. U., Chen, G. Y., & Nunez, G. (2013). Role of the gut microbiota in immunity and inflammatory disease. *Nature Reviews.Immunology, 13*(5), 321-335. doi:10.1038/nri3430; 10.1038/nri3430

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research,* 40 (Database issue), D109-14. doi: 10.1093/nar/gkr988

Karlsson, N. G., Nordman, H., Karlsson, H., Carlstedt, I., & Hansson, G. C. (1997). Glycosylation differences between pig gastric mucin populations: A comparative study of the neutral oligosaccharides using mass spectrometry. *The Biochemical Journal, 326 (Pt 3)*(Pt 3), 911-917.

Karlsson, N. G., & Packer, N. H. (2002). Analysis of O-linked reducing oligosaccharides released by an in-line flow system. *Analytical Biochemistry, 305*(2), 173-185. doi:10.1006/abio.2002.5657

Kazeeva, T. N., & Shevelev, A. B. (2007). Unknown functions of immunoglobulins A. *Biochemistry.Biokhimiia, 72*(5), 485-494.

Kitaoka, M., Tian, J., & Nishimoto, M. (2005). Novel putative galactose operon involving lacto-N-biose phosphorylase in bifidobacterium longum. *Applied and Environmental Microbiology, 71*(6), 3158-3162. doi:10.1128/AEM.71.6.3158-3162.2005

Kiyohara, M., Tanigawa, K., Chaiwangsri, T., Katayama, T., Ashida, H., & Yamamoto, K. (2011). An exo-alpha-sialidase from bifidobacteria involved in the degradation of sialyloligosaccharides in human milk and intestinal glycoconjugates. *Glycobiology, 21*(4), 437-447. doi:10.1093/glycob/cwq175; 10.1093/glycob/cwq175

Kobayashi, R. K., Gaziri, L. C., Venancio, E. J., & Vidotto, M. C. (2007). Detection of tsh protein mucinolytic activity by SDS-PAGE. *Journal of Microbiological Methods, 68*(3), 654-655. doi:10.1016/j.mimet.2006.10.002

Koebnik, R. (2005). TonB-dependent trans-envelope signalling: The exception or the rule? *Trends in Microbiology, 13*(8), 343-347. doi:10.1016/j.tim.2005.06.005

Kolset, S. O., & Tveit, H. (2008). Serglycin--structure and biology. *Cellular and Molecular Life Sciences : CMLS, 65*(7-8), 1073-1085. doi:10.1007/s00018-007-7455-6

Konno, M., Baba, S., Mikawa, H., Hara, K., Matsumoto, F., Kaga, K., . . . Ubukata, K. (2006). Study of upper respiratory tract bacterial flora: First report. variations in upper respiratory tract bacterial flora in patients with acute upper respiratory tract infection and healthy subjects and variations by subject age. *Journal of Infection and Chemotherapy : Official Journal of the Japan Society of Chemotherapy, 12*(2), 83-96. doi:10.1007/s10156-006-0433-3

Koropatkin, N. M., Cameron, E. A., & Martens, E. C. (2012). How glycan metabolism shapes the human gut microbiota. *Nature Reviews.Microbiology, 10*(5), 323-335. doi:10.1038/nrmicro2746; 10.1038/nrmicro2746

Koropatkin, N. M., Martens, E. C., Gordon, J. I., & Smith, T. J. (2008). Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure (London, England : 1993), 16*(7), 1105-1115. doi:10.1016/j.str.2008.03.017; 10.1016/j.str.2008.03.017

Koropatkin, N. M., & Smith, T. J. (2010). SusG: A unique cell-membrane-associated alpha-amylase from a prominent human gut symbiont targets complex starch molecules. *Structure (London, England : 1993), 18*(2), 200-215. doi:10.1016/j.str.2009.12.010; 10.1016/j.str.2009.12.010

Kostakioti, M., & Stathopoulos, C. (2004). Functional analysis of the tsh autotransporter from an avian pathogenic escherichia coli strain. *Infection and Immunity, 72*(10), 5548-5554. doi:10.1128/IAI.72.10.5548-5554.2004

Kotarski, S. F., & Salyers, A. A. (1984). Isolation and characterization of outer membranes of *bacteroides thetaiotaomicron* grown on different carbohydrates. *Journal of Bacteriology, 158*(1), 102-109.

Kovacs-Simon, A., Titball, R. W., & Michell, S. L. (2011). Lipoproteins of bacterial pathogens. *Infection and Immunity, 79*(2), 548-561. doi:10.1128/IAI.00682-10

Kuberski, T. T. (1981). Ankylosing spondylitis associated with *Trichomonas vaginalis* infection. *Journal of Clinical Microbiology, 13*(5), 880-881.

Kusche, M., Lindahl, U., Enerback, L., & Roden, L. (1988). Identification of oversulphated galactosaminoglycans in intestinal-mucosal mast cells of rats infected with the nematode worm nippostrongylus brasiliensis. *The Biochemical Journal, 253*(3), 885-893.

Laemmli UK (1970). "Cleavage of structural proteins during the assembly of the head of bacteriophage T4". *Nature* **227** (5259): 680–685. doi:10.1038/227680a0. PMID 5432063.

Lagow, E., DeSouza, M. M., & Carson, D. D. (1999). Mammalian reproductive tract mucins. *Human Reproduction Update, 5*(4), 280-292.

Lancaster, C. A., Peat, N., Duhig, T., Wilson, D., Taylor-Papadimitriou, J., & Gendler, S. J. (1990). Structure and expression of the human polymorphic epithelial mucin gene: An expressed VNTR unit. *Biochemical and Biophysical Research Communications, 173*(3), 1019-1029.

Larsen, B., & Monif, G. R. (2001). Understanding the bacterial flora of the female genital tract. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America, 32*(4), e69-77. doi:10.1086/318710

Larsson, J. M., Karlsson, H., Crespo, J. G., Johansson, M. E., Eklund, L., Sjovall, H., & Hansson, G. C. (2011). Altered O-glycosylation profile of MUC2 mucin occurs in active ulcerative colitis and is associated with increased inflammation. *Inflammatory Bowel Diseases, 17*(11), 2299-2307. doi:10.1002/ibd.21625; 10.1002/ibd.21625

Larsson, J. M., Karlsson, H., Sjovall, H., & Hansson, G. C. (2009). A complex, but uniform O-glycosylation of the human MUC2 mucin from colonic biopsies analyzed by nanoLC/MSn. *Glycobiology, 19*(7), 756-766. doi:10.1093/glycob/cwp048; 10.1093/glycob/cwp048

Lebeer, S., Vanderleyden, J., & De Keersmaecker, S. C. (2010). Host interactions of probiotic bacterial surface molecules: Comparison with commensals and pathogens. *Nature Reviews.Microbiology, 8*(3), 171-184. doi:10.1038/nrmicro2297

Lee, S., Raw, A., Yu, L., Lionberger, R., Ya, N., Verthelyi, D., . . . Woodcock, J. (2013). Scientific considerations in the review and approval of generic enoxaparin in the united states. *Nature Biotechnology, 31*(3), 220-226. doi:10.1038/nbt.2528; 10.1038/nbt.2528

Lehker, M. W., & Sweeney, D. (1999). Trichomonad invasion of the mucous layer requires adhesins, mucinases, and motility. *Sexually Transmitted Infections, 75*(4), 231-238.

Lescar, J., Sanchez, J. F., Audfray, A., Coll, J. L., Breton, C., Mitchell, E. P., & Imberty, A. (2007). Structural basis for recognition of breast and colon cancer epitopes tn antigen and forssman disaccharide by helix pomatia lectin. *Glycobiology, 17*(10), 1077-1083. doi:10.1093/glycob/cwm077

Lesk, A. M. (1995). NAD-binding domains of dehydrogenases. *Current Opinion in Structural Biology, 5*(6), 775-783.

Levine M. (2011). The Zincins: Collagen Fiber Processing and Degradation. In: Topics in Dental Biochemistry. Berlin Heidelberg: Springer Berlin Heidelberg. pp113-128.

Levitin, F., Stern, O., Weiss, M., Gil-Henn, C., Ziv, R., Prokocimer, Z., . . . Wreschner, D. H. (2005). The MUC1 SEA module is a self-cleaving domain. *The Journal of Biological Chemistry, 280*(39), 33374-33386. doi:10.1074/jbc.M506047200

Ley, R. E., Backhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., & Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America, 102*(31), 11070-11075. doi:10.1073/pnas.0504978102

Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell, 124*(4), 837-848. doi:10.1016/j.cell.2006.02.017

Leyn, S. A., Gao, F., Yang, C., & Rodionov, D. A. (2012). N-acetylgalactosamine utilization pathway and regulon in proteobacteria: Genomic reconstruction and experimental

characterization in shewanella. *The Journal of Biological Chemistry, 287*(33), 28047-28056. doi:10.1074/jbc.M112.382333; 10.1074/jbc.M112.382333

Lidell, M. E., Moncada, D. M., Chadee, K., & Hansson, G. C. (2006). Entamoeba histolytica cysteine proteases cleave the MUC2 mucin in its C-terminal domain and dissolve the protective colonic mucus gel. *Proceedings of the National Academy of Sciences of the United States of America, 103*(24), 9298-9303. doi:10.1073/pnas.0600623103

Ligtenberg, M. J., Kruijshaar, L., Buijs, F., van Meijer, M., Litvinov, S. V., & Hilkens, J. (1992). Cell-associated episialin is a complex containing two proteins derived from a common precursor. *The Journal of Biological Chemistry, 267*(9), 6171-6177.

Lin, M. Y., Yen, C. L., & Chen, S. H. (1998). Management of lactose maldigestion by consuming milk containing lactobacilli. *Digestive Diseases and Sciences, 43*(1), 133-137.

Ling, Z., Liu, X., Luo, Y., Yuan, L., Nelson, K. E., Wang, Y., . . . Li, L. (2013). Pyrosequencing analysis of the human microbiota of healthy chinese undergraduates. *BMC Genomics, 14*, 390-2164-14-390. doi:10.1186/1471-2164-14-390; 10.1186/1471-2164-14-390

Liu, Q. P., Sulzenbacher, G., Yuan, H., Bennett, E. P., Pietz, G., Saunders, K., . . . Clausen, H. (2007). Bacterial glycosidases for the production of universal red blood cells. *Nature Biotechnology, 25*(4), 454-464. doi:10.1038/nbt1298

Ly, M., Leach, F. E.,3rd, Laremore, T. N., Toida, T., Amster, I. J., & Linhardt, R. J. (2011). The proteoglycan bikunin has a defined sequence. *Nature Chemical Biology, 7*(11), 827-833. doi:10.1038/nchembio.673; 10.1038/nchembio.673

Ma, L., Meng, Q., Cheng, W., Sung, Y., Tang, P., Hu, S., & Yu, J. (2011). Involvement of the GP63 protease in infection of *Trichomonas vaginalis*. *Parasitology Research, 109*(1), 71-79. doi:10.1007/s00436-010-2222-2; 10.1007/s00436-010-2222-2

MacDonald, T. T. (2003). The mucosal immune system. *Parasite Immunology, 25*(5), 235-246.

Macfarlane, G. T., Blackett, K. L., Nakayama, T., Steed, H., & Macfarlane, S. (2009). The gut microbiota in inflammatory bowel disease. *Current Pharmaceutical Design, 15*(13), 1528-1536.

Macfarlane, S., Woodmansey, E. J., & Macfarlane, G. T. (2005). Colonization of mucin by human intestinal bacteria and establishment of biofilm communities in a two-stage continuous culture system. *Applied and Environmental Microbiology, 71*(11), 7483-7492. doi:10.1128/AEM.71.11.7483-7492.2005

Mackenzie, A. K., Pope, P. B., Pedersen, H. L., Gupta, R., Morrison, M., Willats, W. G., & Eijsink, V. G. (2012). Two SusD-like proteins encoded within a polysaccharide utilization locus of an uncultured ruminant bacteroidetes phylotype bind strongly to cellulose. *Applied and Environmental Microbiology, 78*(16), 5935-5937. doi:10.1128/AEM.01164-12; 10.1128/AEM.01164-12

Macpherson, A. J., Hunziker, L., McCoy, K., & Lamarre, A. (2001). IgA responses in the intestinal mucosa against pathogenic and non-pathogenic microorganisms. *Microbes and Infection / Institut Pasteur, 3*(12), 1021-1035.

Maestre-Reyna, M., Diderrich, R., Veelders, M. S., Eulenburg, G., Kalugin, V., Bruckner, S., . . . Essen, L. O. (2012). Structural basis for promiscuity and specificity during candida glabrata invasion of host epithelia. *Proceedings of the National Academy of Sciences of the United States of America, 109*(42), 16864-16869. doi:10.1073/pnas.1207653109; 10.1073/pnas.1207653109

Mahren, S., Enz, S., & Braun, V. (2002). Functional interaction of region 4 of the extracytoplasmic function sigma factor FecI with the cytoplasmic portion of the FecR transmembrane protein of the escherichia coli ferric citrate transport system. *Journal of Bacteriology, 184*(13), 3704-3711.

Malik, S. B., Brochu, C. D., Bilic, I., Yuan, J., Hess, M., Logsdon, J. M.,Jr, & Carlton, J. M. (2011). Phylogeny of parasitic parabasalia and free-living relatives inferred from conventional markers vs. Rpb1, a single-copy gene. *PloS One, 6*(6), e20774. doi:10.1371/journal.pone.0020774; 10.1371/journal.pone.0020774

Mansfeld, Johanna. (2007). Metalloproteases. In: Polaina, Julio.MacCabe, Andrew P. Industrial Enzymes. Netherlands: Springer Netherlands. p221-242.

Marcobal, A., Barboza, M., Sonnenburg, E. D., Pudlo, N., Martens, E. C., Desai, P., . . . Sonnenburg, J. L. (2011). Bacteroides in the infant gut consume milk oligosaccharides via mucus-utilization pathways. *Cell Host & Microbe, 10*(5), 507-514. doi:10.1016/j.chom.2011.10.007; 10.1016/j.chom.2011.10.007

Martens, E. C., Chiang, H. C., & Gordon, J. I. (2008). Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host & Microbe, 4*(5), 447-457. doi:10.1016/j.chom.2008.09.007

Martens, E. C., Koropatkin, N. M., Smith, T. J., & Gordon, J. I. (2009a). Complex glycan catabolism by the human gut microbiota: The bacteroidetes sus-like paradigm. *The Journal of Biological Chemistry, 284*(37), 24673-24677. doi:10.1074/jbc.R109.022848

Martens, E. C., Lowe, E. C., Chiang, H., Pudlo, N. A., & Wu, M. (2011). Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biology, 9*(12), e1001221.

Martens, E. C., Roth, R., Heuser, J. E., & Gordon, J. I. (2009b). Coordinate regulation of glycan degradation and polysaccharide capsule biosynthesis by a prominent human gut symbiont. *The Journal of Biological Chemistry, 284*(27), 18445-18457. doi:10.1074/jbc.M109.008094; 10.1074/jbc.M109.008094

Martinez, I., Muller, C. E., & Walter, J. (2013). Long-term temporal analysis of the human fecal microbiota revealed a stable core of dominant bacterial species. *PloS One, 8*(7), e69621. doi:10.1371/journal.pone.0069621; 10.1371/journal.pone.0069621

Maslowski, K. M., & Mackay, C. R. (2011). Diet, gut microbiota and immune responses. *Nature Immunology, 12*(1), 5-9. doi:10.1038/ni0111-5; 10.1038/ni0111-5

Matthews, B. W. (1988). Structural basis of the action of thermolysin and related zinc peptidases. *Accounts of Chemical Research, 21*(9), 333-340.

McConville, M. J., & Ferguson, M. A. (1993). The structure, biosynthesis and function of glycosylated phosphatidylinositols in the parasitic protozoa and higher eukaryotes. *Biochemical Journal, 294*(Pt 2), 305-324.

McGuckin, M. A., Linden, S. K., Sutton, P., & Florin, T. H. (2011). Mucin dynamics and enteric pathogens. *Nature Reviews.Microbiology, 9*(4), 265-278. doi:10.1038/nrmicro2538; 10.1038/nrmicro2538

McNeil, N. I. (1984). The contribution of the large intestine to energy supplies in man. *The American Journal of Clinical Nutrition, 39*(2), 338-342.

Meddens, M. J., Thompson, J., Eulderink, F., Bauer, W. C., Mattie, H., & van Furth, R. (1982). Role of granulocytes in experimental streptococcus sanguis endocarditis. *Infection and Immunity, 36*(1), 325-332.

Mello, L. V., Chen, X., & Rigden, D. J. (2010). Mining metagenomic data for novel domains: BACON, a new carbohydrate-binding module. *FEBS Letters, 584*(11), 2421-2426. doi:10.1016/j.febslet.2010.04.045

Mendoza-Lopez, M. R., Becerril-Garcia, C., Fattel-Facenda, L. V., Avila-Gonzalez, L., Ruiz-Tachiquin, M. E., Ortega-Lopez, J., & Arroyo, R. (2000). CP30, a cysteine proteinase involved in *Trichomonas vaginalis* cytoadherence. *Infection and Immunity, 68*(9), 4907-4912.

Meyers, J. A., Sanchez, D., Elwell, L. P., & Falkow, S. (1976). Simple agarose gel electrophoretic method for the identification and characterization of plasmid deoxyribonucleic acid. *Journal of Bacteriology, 127*(3), 1529-1537.

Meza-Cervantez, P., Gonzalez-Robles, A., Cardenas-Guerra, R. E., Ortega-Lopez, J., Saavedra, E., Pineda, E., & Arroyo, R. (2011). Pyruvate:Ferredoxin oxidoreductase (PFO) is a surface-associated cell-binding protein in *Trichomonas vaginalis* and is involved in trichomonal adherence to host cells. *Microbiology (Reading, England), 157*(Pt 12), 3469-3482. doi:10.1099/mic.0.053033-0; 10.1099/mic.0.053033-0

Miquel, S., Martin, R., Rossi, O., Bermudez-Humaran, L. G., Chatel, J. M., Sokol, H., . . . Langella, P. (2013). Faecalibacterium prausnitzii and human intestinal health. *Current Opinion in Microbiology, 16*(3), 255-261. doi:10.1016/j.mib.2013.06.003; 10.1016/j.mib.2013.06.003

Miyoshi, S., & Shinoda, S. (2000). Microbial metalloproteases and pathogenesis. *Microbes and Infection / Institut Pasteur, 2*(1), 91-98.

Moncada, D., Keller, K., & Chadee, K. (2003). Entamoeba histolytica cysteine proteinases disrupt the polymeric structure of colonic mucin and alter its protective function. *Infection and Immunity, 71*(2), 838-844.

Moran, A. P., Gupta, A., & Joshi, L. (2011). Sweet-talk: Role of host glycosylation in bacterial pathogenesis of the gastrointestinal tract. *Gut,* doi:10.1136/gut.2010.212704

Mortensen, S. B., & Kilian, M. (1984). Purification and characterization of an immunoglobulin A1 protease from bacteroides melaninogenicus. *Infection and Immunity, 45*(3), 550-557.

Morton, H. C., Atkin, J. D., Owens, R. J., & Woof, J. M. (1993). Purification and characterization of chimeric human IgA1 and IgA2 expressed in COS and chinese hamster ovary cells. *Journal of Immunology (Baltimore, Md.: 1950), 151*(9), 4743-4752.

Muller, M. (1993). The hydrogenosome. *Journal of General Microbiology, 139*(12), 2879-2889.

Muller, S., Alving, K., Peter-Katalinic, J., Zachara, N., Gooley, A. A., & Hanisch, F. G. (1999). High density O-glycosylation on tandem repeat peptide from secretory MUC1 of T47D breast cancer cells. *The Journal of Biological Chemistry, 274*(26), 18165-18172.

Nakayama Y, Nakamura N, Tsuji D, Itoh K and Kurosaka A (2013). Genetic Diseases Associated with Protein Glycosylation Disorders in Mammals, Genetic Disorders, Prof. Maria Puiu (Ed.), ISBN: 978-953-51-0886-3, InTech, DOI: 10.5772/54097. Available from: http://www.intechopen.com/books/genetic-disorders/genetic-diseases-associated-with-protein-glycosylation-disorders-in-mammals

Nakjang, S., Ndeh, D. A., Wipat, A., Bolam, D. N., & Hirt, R. P. (2012). A novel extracellular metallopeptidase domain shared by animal host-associated mutualistic and pathogenic microbes. *PloS One, 7*(1), e30287. doi:10.1371/journal.pone.0030287; 10.1371/journal.pone.0030287

Nakjang, S., Ndeh, D. A., Wipat, A., Bolam, D. N., & Hirt, R. P. (2012). A novel extracellular metallopeptidase domain shared by animal host-associated mutualistic and pathogenic microbes. *PloS One, 7*(1), e30287. doi:10.1371/journal.pone.0030287; 10.1371/journal.pone.0030287

Naumoff, D. G. (2011). Hierarchical classification of glycoside hydrolases. *Biochemistry.Biokhimiia, 76*(6), 622-635. doi:10.1134/S0006297911060022; 10.1134/S0006297911060022

Nettleship J E (2012). Structural Biology of Glycoproteins, Glycosylation, Dr. Stefana Petrescu (Ed.), ISBN: 978-953-51-0771-2, InTech, DOI: 10.5772/48154. Available from: http://www.intechopen.com/books/glycosylation/structural-biology-of-glycoproteins

Nishimoto, M., & Kitaoka, M. (2007). Identification of N-acetylhexosamine 1-kinase in the complete lacto-N-biose I/galacto-N-biose metabolic pathway in bifidobacterium longum. *Applied and Environmental Microbiology, 73*(20), 6444-6449. doi:10.1128/AEM.01425-07

Novak, J., Julian, B. A., Tomana, M., & Mestecky, J. (2008). IgA glycosylation and IgA immune complexes in the pathogenesis of IgA nephropathy. *Seminars in Nephrology, 28*(1), 78-87. doi:10.1016/j.semnephrol.2007.10.009

Noverr, M. C., & Huffnagle, G. B. (2005). The 'microflora hypothesis' of allergic diseases. *Clinical and Experimental Allergy : Journal of the British Society for Allergy and Clinical Immunology, 35*(12), 1511-1520. doi:10.1111/j.1365-2222.2005.02379.x

Okumura, C. Y., Baum, L. G., & Johnson, P. J. (2008). Galectin-1 on cervical epithelial cells is a receptor for the sexually transmitted human parasite *Trichomonas vaginalis*. *Cellular Microbiology, 10*(10), 2078-2090. doi:10.1111/j.1462-5822.2008.01190.x; 10.1111/j.1462-5822.2008.01190.x

Olsson, M. L., Hill, C. A., de la Vega, H., Liu, Q. P., Stroud, M. R., Valdinocci, J., . . . Kruskall, M. S. (2004). Universal red blood cells--enzymatic conversion of blood group A and B antigens. *Transfusion Clinique Et Biologique : Journal De La Societe Francaise De Transfusion Sanguine, 11*(1), 33-39. doi:10.1016/j.tracli.2003.12.002

Ouwerkerk, J. P., de Vos, W. M., & Belzer, C. (2013). Glycobiome: Bacteria and mucus at the epithelial interface. *Best Practice & Research.Clinical Gastroenterology, 27*(1), 25-38. doi:10.1016/j.bpg.2013.03.001; 10.1016/j.bpg.2013.03.001

Pacheco, A. R., Curtis, M. M., Ritchie, J. M., Munera, D., Waldor, M. K., Moreira, C. G., & Sperandio, V. (2012). Fucose sensing regulates bacterial intestinal colonization. *Nature, 492*(7427), 113-117. doi:10.1038/nature11623; 10.1038/nature11623

Parry, S., Hanisch, F. G., Leir, S. H., Sutton-Smith, M., Morris, H. R., Dell, A., & Harris, A. (2006). N-glycosylation of the MUC1 mucin in epithelial cells and secretions. *Glycobiology, 16*(7), 623-634. doi:10.1093/glycob/cwj110

Perez-Vilar, J., & Hill, R. L. (1999). The structure and assembly of secreted mucins. *The Journal of Biological Chemistry, 274*(45), 31751-31754.

Petrin, D., Delgaty, K., Bhatt, R., & Garber, G. (1998). Clinical and microbiological aspects of *Trichomonas vaginalis*. *Clinical Microbiology Reviews, 11*(2), 300-317.

Pope, P. B., Mackenzie, A. K., Gregor, I., Smith, W., Sundset, M. A., McHardy, A. C., . . . Eijsink, V. G. (2012). Metagenomics of the svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci. *PloS One, 7*(6), e38571. doi:10.1371/journal.pone.0038571; 10.1371/journal.pone.0038571

Posey JE,Shinnick TM,Quinn FD, (2006).Characterization of the twin-arginine translocase secretion system of Mycobacterium smegmatis. *Journal of bacteriology*, *188* (4), doi: 1332-40. 10.1128/JB.188.4.1332-1340.2006

Potapenko, I. O., Haakensen, V. D., Luders, T., Helland, A., Bukholm, I., Sorlie, T., . . . Borresen-Dale, A. L. (2010). Glycan gene expression signatures in normal and malignant breast tissue; possible role in diagnosis and progression. *Molecular Oncology, 4*(2), 98-118. doi:10.1016/j.molonc.2009.12.001; 10.1016/j.molonc.2009.12.001

Premaratne, P., Welen, K., Damber, J. E., Hansson, G. C., & Backstrom, M. (2011). O-glycosylation of MUC1 mucin in prostate cancer and the effects of its expression on tumor growth in a prostate cancer xenograft model. *Tumour Biology : The Journal of the International*

Society for Oncodevelopmental Biology and Medicine, 32(1), 203-213. doi:10.1007/s13277-010-0114-9; 10.1007/s13277-010-0114-9

Provenzano, D., & Alderete, J. F. (1995). Analysis of human immunoglobulin-degrading cysteine proteinases of *Trichomonas vaginalis*. *Infection and Immunity, 63*(9), 3388-3395.

Punta M., P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn, The Pfam protein families database Nucleic Acids Research (2012) Database Issue 40:D290-D301

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., . . . Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature, 464*(7285), 59-65. doi:10.1038/nature08821; 10.1038/nature08821

Qiu, J., Brackee, G. P., & Plaut, A. G. (1996). Analysis of the specificity of bacterial immunoglobulin A (IgA) proteases by a comparative study of ape serum IgAs as substrates. *Infection and Immunity, 64*(3), 933-937.

Rabenstein, D. L. (2002). Heparin and heparan sulfate: Structure and function. *Natural Product Reports, 19*(3), 312-331. doi:10.1039/B100916H

Ramasarma, T., & Joshi, N. (2001). Transmembrane domains. *Els* () John Wiley & Sons, Ltd. doi:10.1038/npg.els.0005051

Ramos, ,O.H.P., & Selistre-de-araujo, H.S. (2001). Identification of metalloprotease gene families in sugarcane. *Genet.Mol.Biol, 24*(1-4), 285-290.

Rawlings, N.D., Barrett, A.J. & Bateman, A. (2012) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res 40*, D343-D350. Last accessed 29/08/13

Redondo, M. C., Arbo, M. D., Grindlinger, J., & Snydman, D. R. (1995). Attributable mortality of bacteremia associated with the bacteroides fragilis group. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America, 20*(6), 1492-1496.

Reeves, A. R., D'Elia, J. N., Frias, J., & Salyers, A. A. (1996). A *bacteroides thetaiotaomicron* outer membrane protein that is essential for utilization of maltooligosaccharides and starch. *Journal of Bacteriology, 178*(3), 823-830.

Reeves, A. R., Wang, G. R., & Salyers, A. A. (1997). Characterization of four outer membrane proteins that play a role in utilization of starch by *bacteroides thetaiotaomicron*. *Journal of Bacteriology, 179*(3), 643-649.

Reith, J., Berking, A., & Mayer, C. (2011). Characterization of an N-acetylmuramic acid/N-acetylglucosamine kinase of clostridium acetobutylicum. *Journal of Bacteriology, 193*(19), 5386-5392. doi:10.1128/JB.05514-11; 10.1128/JB.05514-11

Reithmeier, R. A. (2001). Membrane proteins. *Els* () John Wiley & Sons, Ltd. doi:10.1038/npg.els.0000624

Reizer, J., Ramseier, T. M., Reizer, A., Charbit, A., & Saier, M. H.,Jr. (1996). Novel phosphotransferase genes revealed by bacterial genome sequencing: A gene cluster encoding a putative N-acetylgalactosamine metabolic pathway in escherichia coli. *Microbiology (Reading, England), 142 (Pt 2)*(Pt 2), 231-250.

Rendon-Maldonado, J. G., Espinosa-Cantellano, M., Gonzalez-Robles, A., & Martinez-Palomo, A. (1998). *Trichomonas vaginalis*: In vitro phagocytosis of lactobacilli, vaginal epithelial cells, leukocytes, and erythrocytes. *Experimental Parasitology, 89*(2), 241-250. doi:10.1006/expr.1998.4297

Rigden, D. J., Jedrzejas, M. J., & de Mello, L. V. (2003). Identification and analysis of catalytic TIM barrel domains in seven further glycoside hydrolase families. *FEBS Letters, 544*(1-3), 103-111.

Rigden, D. J., Mello, L. V., & Galperin, M. Y. (2004). The PA14 domain, a conserved all-beta domain in bacterial toxins, enzymes, adhesins and signaling molecules. *Trends in Biochemical Sciences, 29*(7), 335-339. doi:10.1016/j.tibs.2004.05.002

Robbe, C., Capon, C., Coddeville, B., & Michalski, J. C. (2004). Structural diversity and specific distribution of O-glycans in normal human mucins along the intestinal tract. *The Biochemical Journal, 384*(Pt 2), 307-316. doi:10.1042/BJ20040605

Rodgers, K. D., San Antonio, J. D., & Jacenko, O. (2008). Heparan sulfate proteoglycans: A GAGgle of skeletal-hematopoietic regulators. *Developmental Dynamics : An Official Publication of the American Association of Anatomists, 237*(10), 2622-2642. doi:10.1002/dvdy.21593; 10.1002/dvdy.21593

Rogemond, V., & Guinet, R. M. (1986). Lectinlike adhesins in the bacteroides fragilis group. *Infection and Immunity, 53*(1), 99-102.

Rojas, M., Ascencio, F., & Conway, P. L. (2002). Purification and characterization of a surface protein from lactobacillus fermentum 104R that binds to porcine small intestinal mucus and gastric mucin. *Applied and Environmental Microbiology, 68*(5), 2330-2336.

Ronnqvist, D., Forsgren-Brusk, U., Husmark, U., & Grahn-Hakansson, E. (2007). Lactobacillus fermentum ess-1 with unique growth inhibition of vulvo-vaginal candidiasis pathogens. *Journal of Medical Microbiology, 56*(Pt 11), 1500-1504. doi:10.1099/jmm.0.47226-0

Roos, S., & Jonsson, H. (2002). A high-molecular-mass cell-surface protein from lactobacillus reuteri 1063 adheres to mucus components. *Microbiology (Reading, England), 148*(Pt 2), 433-442.

Rose, M. C., & Voynow, J. A. (2006). Respiratory tract mucin genes and mucin glycoproteins in health and disease. *Physiological Reviews, 86*(1), 245-278. doi:10.1152/physrev.00010.2005

Rougé P. (2011). Chapter 8. In: Cohen I R, Lajtha A, Lambris J D, Paoletti R Advances in Experimental Medicine and Biology. Online: Springer Science+Business Media, LLC. 144.

Rousseau, K., Byrne, C., Kim, Y. S., Gum, J. R., Swallow, D. M., & Toribara, N. W. (2004). The complete genomic organization of the human MUC6 and MUC2 mucin genes. *Genomics, 83*(5), 936-939. doi:10.1016/j.ygeno.2003.11.003

Rousset, E., Harel, J., & Dubreuil, J. D. (1998). Sulfatide from the pig jejunum brush border epithelial cell surface is involved in binding of escherichia coli enterotoxin b. *Infection and Immunity, 66*(12), 5650-5658.

Roy, R., & Baek, M. G. (2002). Glycodendrimers: Novel glycotope isosteres unmasking sugar coding. case study with T-antigen markers from breast cancer MUC1 glycoprotein. *Journal of Biotechnology, 90*(3-4), 291-309.

Royle, L., Mattu, T. S., Hart, E., Langridge, J. I., Merry, A. H., Murphy, N., . . . Rudd, P. M. (2002). An analytical and structural database provides a strategy for sequencing O-glycans from microgram quantities of glycoproteins. *Analytical Biochemistry, 304*(1), 70-90. doi:10.1006/abio.2002.5619

Royle, L., Roos, A., Harvey, D. J., Wormald, M. R., van Gijlswijk-Janssen, D., Redwan, e. M., . . . Rudd, P. M. (2003). Secretory IgA N- and O-glycans provide a link between the innate and adaptive immune systems. *The Journal of Biological Chemistry, 278*(22), 20140-20153. doi:10.1074/jbc.M301436200

Ruggeri, Z. M., & Ware, J. (1993). Von willebrand factor. *The FASEB Journal, 7*(2), 308-316.

Ruiz-Perez, F., Wahid, R., Faherty, C. S., Kolappaswamy, K., Rodriguez, L., Santiago, A., . . . Nataro, J. P. (2011). Serine protease autotransporters from shigella flexneri and pathogenic escherichia coli target a broad range of leukocyte glycoproteins. *Proceedings of the National Academy of Sciences of the United States of America, 108*(31), 12881-12886. doi:10.1073/pnas.1101006108; 10.1073/pnas.1101006108

Russell, D., Oldham, N. J., & Davis, B. G. (2009). Site-selective chemical protein glycosylation protects from autolysis and proteolytic degradation. *Carbohydrate Research, 344*(12), 1508-1514. doi:10.1016/j.carres.2009.06.033; 10.1016/j.carres.2009.06.033

Rye, C. S., & Withers, S. G. (2000). Glycosidase mechanisms. *Current Opinion in Chemical Biology, 4*(5), 573-580. doi:http://dx.doi.org/10.1016/S1367-5931(00)00135-6

Sadler, J. E. (1998). Biochemistry and genetics of von willebrand factor. *Annual Review of Biochemistry, 67*, 395-424. doi:10.1146/annurev.biochem.67.1.395

Sansonetti, P. J. (2004). War and peace at mucosal surfaces. *Nature Reviews.Immunology, 4*(12), 953-964. doi:10.1038/nri1499

Sarrazin, S., Lamanna, W. C., & Esko, J. D. (2011). Heparan sulfate proteoglycans. *Cold Spring Harbor Perspectives in Biology, 3*(7), 10.1101/cshperspect.a004952. doi:10.1101/cshperspect.a004952; 10.1101/cshperspect.a004952

Savage, D. C. (1977). Microbial ecology of the gastrointestinal tract. *Annual Review of Microbiology, 31*, 107-133. doi:10.1146/annurev.mi.31.100177.000543

Scheppach, W. (1994). Effects of short chain fatty acids on gut morphology and function. *Gut, 35*(1 Suppl), S35-8.

Schmid, G., Narcisi, E., Mosure, D., Secor, W. E., Higgins, J., & Moreno, H. (2001). Prevalence of metronidazole-resistant *Trichomonas vaginalis* in a gynecology clinic. *The Journal of Reproductive Medicine, 46*(6), 545-549.

Schneider, R. E., Brown, M. T., Shiflett, A. M., Dyall, S. D., Hayes, R. D., Xie, Y., . . . Johnson, P. J. (2011). The *Trichomonas vaginalis* hydrogenosome proteome is highly reduced relative to mitochondria, yet complex compared with mitosomes. *International Journal for Parasitology, 41*(13-14), 1421-1434. doi:10.1016/j.ijpara.2011.10.001; 10.1016/j.ijpara.2011.10.001

Schroeder, J. A., Masri, A. A., Adriance, M. C., Tessier, J. C., Kotlarczyk, K. L., Thompson, M. C., & Gendler, S. J. (2004). MUC1 overexpression results in mammary gland tumorigenesis and prolonged alveolar differentiation. *Oncogene, 23*(34), 5739-5747. doi:10.1038/sj.onc.1207713

Schwebke, J. R., & Burgess, D. (2004). Trichomoniasis. *Clinical Microbiology Reviews, 17*(4), 794-803, table of contents. doi:10.1128/CMR.17.4.794-803.2004

Seema, S., & Arti, K. (2008). An update on *Trichomonas vaginalis*. *Indian Journal of Sexually Transmitted Diseases and AIDS, 29*(1), 7-14.

Sekirov, I., Russell, S. L., Antunes, L. C., & Finlay, B. B. (2010). Gut microbiota in health and disease. *Physiological Reviews, 90*(3), 859-904. doi:10.1152/physrev.00045.2009; 10.1152/physrev.00045.2009

Sheehan, J. K., Kirkham, S., Howard, M., Woodman, P., Kutay, S., Brazeau, C., . . . Thornton, D. J. (2004). Identification of molecular intermediates in the assembly pathway of the MUC5AC mucin. *The Journal of Biological Chemistry, 279*(15), 15698-15705. doi:10.1074/jbc.M313241200

Shipman, J. A., Cho, K. H., Siegel, H. A., & Salyers, A. A. (1999). Physiological characterization of SusG, an outer membrane protein essential for starch utilization by *bacteroides thetaiotaomicron*. *Journal of Bacteriology, 181*(23), 7206-7211.

Shoseyov, O., Shani, Z., & Levy, I. (2006). Carbohydrate binding modules: Biochemical properties and novel applications. *Microbiology and Molecular Biology Reviews : MMBR, 70*(2), 283-295. doi:10.1128/MMBR.00028-05

Silva, A. J., Pham, K., & Benitez, J. A. (2003). Haemagglutinin/protease expression and mucin gel penetration in el tor biotype vibrio cholerae. *Microbiology (Reading, England), 149*(Pt 7), 1883-1891.

Singer, S. J. (1974). The molecular organization of membranes. *Annual Review of Biochemistry, 43*(0), 805-833. doi:10.1146/annurev.bi.43.070174.004105

Sipaul, F., Birchall, M., & Corfield, A. (2011). What role do mucins have in the development of laryngeal squamous cell carcinoma? A systematic review. *European Archives of Oto-Rhino-Laryngology : Official Journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : Affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery, 268*(8), 1109-1117. doi:10.1007/s00405-011-1617-8; 10.1007/s00405-011-1617-8

Skoog, E. C., Sjoling, A., Navabi, N., Holgersson, J., Lundin, S. B., & Linden, S. K. (2012). Human gastric mucins differently regulate helicobacter pylori proliferation, gene expression and interactions with host cells. *PloS One, 7*(5), e36378. doi:10.1371/journal.pone.0036378; 10.1371/journal.pone.0036378

Skorupski, K., & Taylor, R. K. (1996). Positive selection vectors for allelic exchange. *Gene, 169*(1), 47-52.

Slomiany, B. L., Bilski, J., Sarosiek, J., Murty, V. L., Dworkin, B., VanHorn, K., . . . Slomiany, A. (1987). Campylobacter pyloridis degrades mucin and undermines gastric mucosal integrity. *Biochemical and Biophysical Research Communications, 144*(1), 307-314.

Sobhani, I., Tap, J., Roudot-Thoraval, F., Roperch, J. P., Letulle, S., Langella, P., . . . Furet, J. P. (2011). Microbial dysbiosis in colorectal cancer (CRC) patients. *PloS One, 6*(1), e16393. doi:10.1371/journal.pone.0016393; 10.1371/journal.pone.0016393

Sonnenburg,J.L., Xu, J., Leip, D.D., Chen, C.H., Westover,B.P., Weatherford,J., Buhler,J.D. and Gordon,J.I. (2005) Glycan foragingin vivo by an intestine-adapted bacterial symbiont. *Science*, 307, 1955-1959.

Sonnenburg, E. D., Zheng, H., Joglekar, P., Higginbottom, S. K., Firbank, S. J., Bolam, D. N., & Sonnenburg, J. L. (2010). Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell, 141*(7), 1241-1252. doi:10.1016/j.cell.2010.05.005; 10.1016/j.cell.2010.05.005

Sorvillo, F., & Kerndt, P. (1998). *Trichomonas vaginalis* and amplification of HIV-1 transmission. *Lancet, 351*(9097), 213-214. doi:10.1016/S0140-6736(05)78181-2

Stahl, M., Friis, L. M., Nothaft, H., Liu, X., Li, J., Szymanski, C. M., & Stintzi, A. (2011). L-fucose utilization provides campylobacter jejuni with a competitive advantage. *Proceedings of the National Academy of Sciences of the United States of America, 108*(17), 7194-7199. doi:10.1073/pnas.1014125108; 10.1073/pnas.1014125108

Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester-Christensen MB, Schjoldager KT, Lavrsen K, Dabelsteen S, Pedersen NB, Marcos-Silva L, Gupta R, Bennett EP, Mandel U, Brunak S, Wandall HH, Levery SB, Clausen H.EMBO J, 32(10):1478-88, May 15, 2013. (doi: 10.1038/emboj.2013.79. Epub 2013 Apr 12)

Studier, F. W., & Moffatt, B. A. (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Biology, 189*(1), 113-130.

Sutcliffe, S. (2010). Sexually transmitted infections and risk of prostate cancer: Review of historical and emerging hypotheses. *Future Oncology (London, England), 6*(8), 1289-1311. doi:10.2217/fon.10.95

Swallow, D. M. (2003). Genetic influences on carbohydrate digestion. *Nutrition Research Reviews, 16*(1), 37-43. doi:10.1079/NRR200253; 10.1079/NRR200253

Szabady, R. L., Yanta, J. H., Halladin, D. K., Schofield, M. J., & Welch, R. A. (2011). TagA is a secreted protease of vibrio cholerae that specifically cleaves mucin glycoproteins. *Microbiology (Reading, England), 157*(Pt 2), 516-525. doi:10.1099/mic.0.044529-0

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., . . . von Mering, C. (2011). The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research, 39*(Database issue), D561-8. doi:10.1093/nar/gkq973; 10.1093/nar/gkq973

Tai, S. S. (2006). Streptococcus pneumoniae protein vaccine candidates: Properties, activities and animal studies. *Critical Reviews in Microbiology, 32*(3), 139-153. doi:10.1080/10408410600822942

Taylor-Papadimitriou, J., Burchell, J., Miles, D. W., & Dalziel, M. (1999). MUC1 and cancer. *Biochimica Et Biophysica Acta (BBA) - Molecular Basis of Disease, 1455*(2–3), 301-313. doi:http://dx.doi.org/10.1016/S0925-4439(99)00055-1

Teng, L. J., Hsueh, P. R., Tsai, J. C., Liaw, S. J., Ho, S. W., & Luh, K. T. (2002). High incidence of cefoxitin and clindamycin resistance among anaerobes in taiwan. *Antimicrobial Agents and Chemotherapy, 46*(9), 2908-2913.

Terra, V. S., Homer, K. A., Rao, S. G., Andrew, P. W., & Yesilkaya, H. (2010). Characterization of novel beta-galactosidase activity that contributes to glycoprotein degradation and virulence in streptococcus pneumoniae. *Infection and Immunity, 78*(1), 348-357. doi:10.1128/IAI.00721-09; 10.1128/IAI.00721-09

Tester, R. F., Karkalas, J., & Qi, X. (2004). Starch—composition, fine structure and architecture. *Journal of Cereal Science, 39*(2), 151-165. doi:http://dx.doi.org/10.1016/j.jcs.2003.12.001

Thomsson KA, Prakobphol A, Leffler H, Reddy MS, Levine MJ, Fisher SJ, Hansson GC (2002). The salivary mucin MG1 (MUC5B) carries a repertoire of unique oligosaccharides that is large and diverse. *Glycobiology,* 12 (2), 1-14.

Thornton, D. J., Rousseau, K., & McGuckin, M. A. (2008). Structure and function of the polymeric mucins in airways mucus. *Annual Review of Physiology, 70*, 459-486. doi:10.1146/annurev.physiol.70.113006.100702

Thornton, D. J., & Sheehan, J. K. (2004). From mucins to mucus: Toward a more coherent understanding of this essential barrier. *Proceedings of the American Thoracic Society, 1*(1), 54-61. doi:10.1513/pats.2306016

Tomas, M. S., Claudia Otero, M., Ocana, V., & Elena Nader-Macias, M. (2004). Production of antimicrobial substances by lactic acid bacteria I: Determination of hydrogen peroxide. *Methods in Molecular Biology (Clifton, N.J.), 268*, 337-346. doi:10.1385/1-59259-766-1:337

Torano, A., & Putnam, F. W. (1978). Complete amino acid sequence of the alpha 2 heavy chain of a human IgA2 immunoglobulin of the A2m (2) allotype. *Proceedings of the National Academy of Sciences of the United States of America, 75*(2), 966-969.

Tran, D. T., & Ten Hagen, K. G. (2013). Mucin-type O-glycosylation during development. *The Journal of Biological Chemistry, 288*(10), 6921-6929. doi:10.1074/jbc.R112.418558; 10.1074/jbc.R112.418558

Tremaroli, V., & Backhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. *Nature, 489*(7415), 242-249. doi:10.1038/nature11552; 10.1038/nature11552

GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens Giardia lamblia and *Trichomonas vaginalis*. 2009 Jan;37(Database issue):D526-30. Aurrecoechea C, *et al.*

Tsuji, T., & Osawa, T. (1986). Carbohydrate structures of bovine submaxillary mucin. *Carbohydrate Research, 151*, 391-402.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., & Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature, 444*(7122), 1027-1031. doi:10.1038/nature05414

Turner, B. S., Bhaskar, K. R., Hadzopoulou-Cladaras, M., & LaMont, J. T. (1999). Cysteine-rich regions of pig gastric mucin contain von willebrand factor and cystine knot domains at the carboxyl terminal(1). *Biochimica Et Biophysica Acta, 1447*(1), 77-92.

Underdown, B. J., & Schiff, J. M. (1986). Immunoglobulin A: Strategic defense initiative at the mucosal surface. *Annual Review of Immunology, 4*, 389-417. doi:10.1146/annurev.iy.04.040186.002133

Van der Sluis, M., De Koning, B. A., De Bruijn, A. C., Velcich, A., Meijerink, J. P., Van Goudoever, J. B., . . . Einerhand, A. W. (2006). Muc2-deficient mice spontaneously develop colitis, indicating that MUC2 is critical for colonic protection. *Gastroenterology, 131*(1), 117-129. doi:10.1053/j.gastro.2006.04.020

Varki A, Sharon N. Historical Background and Overview. In: Varki A, Cummings RD, Esko JD, *et al.*, editors. Essentials of Glycobiology. 2nd edition. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2009. Chapter 1. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1931/

Velcich, A., Yang, W., Heyer, J., Fragale, A., Nicholas, C., Viani, S., . . . Augenlicht, L. (2002). Colorectal cancer in mice genetically deficient in the mucin Muc2. *Science (New York, N.Y.), 295*(5560), 1726-1729. doi:10.1126/science.1069094

Vinall, L. E., Hill, A. S., Pigny, P., Pratt, W. S., Toribara, N., Gum, J. R., . . . Swallow, D. M. (1998). Variable number tandem repeat polymorphism of the mucin genes located in the complex on 11p15.5. *Human Genetics, 102*(3), 357-366.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., . . . Bork, P. (2005). STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research, 33*(Database issue), D433-7. doi:10.1093/nar/gki005

Vonk, R. J., Hagedoorn, R. E., de Graaff, R., Elzinga, H., Tabak, S., Yang, Y. X., & Stellaard, F. (2000). Digestion of so-called resistant starch sources in the human small intestine. *The American Journal of Clinical Nutrition, 72*(2), 432-438.

Vuong, T. V., & Wilson, D. B. (2010). Glycoside hydrolases: Catalytic base/nucleophile diversity. *Biotechnology and Bioengineering, 107*(2), 195-205. doi:10.1002/bit.22838; 10.1002/bit.22838

Wang AC, Fudenberg HH, (1972). Fc and Fab fragments from IgG 2 human immunoglobulins characterized. *Nature: New biology*, 240 (96), 24-6

Wang, P., & Granados, R. R. (1997). An intestinal mucin is the target substrate for a baculovirus enhancin. *Proceedings of the National Academy of Sciences of the United States of America, 94*(13), 6977-6982.

Wesseling, J., van der Valk, S. W., & Hilkens, J. (1996). A mechanism for inhibition of E-cadherin-mediated cell-cell adhesion by the membrane-associated mucin episialin/MUC1. *Molecular Biology of the Cell, 7*(4), 565-577.

Wexler, H. M. (2007). Bacteroides: The good, the bad, and the nitty-gritty. *Clinical Microbiology Reviews, 20*(4), 593-621. doi:10.1128/CMR.00008-07

Wiggins, R., Hicks, S. J., Soothill, P. W., Millar, M. R., & Corfield, A. P. (2001). Mucinases and sialidases: Their role in the pathogenesis of sexually transmitted infections in the female genital tract. *Sexually Transmitted Infections, 77*(6), 402-408.

Wrzosek, L., Miquel, S., Noordine, M. L., Bouet, S., Chevalier-Curt, M. J., Robert, V., . . . Thomas, M. (2013). *Bacteroides thetaiotaomicron* and faecalibacterium prausnitzii influence the production of mucus glycans and the development of goblet cells in the colonic epithelium of a gnotobiotic model rodent. *BMC Biology, 11*, 61-7007-11-61. doi:10.1186/1741-7007-11-61; 10.1186/1741-7007-11-61

Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., . . . Gordon, J. I. (2003). A genomic view of the human-*bacteroides thetaiotaomicron* symbiosis. *Science (New York, N.Y.), 299*(5615), 2074-2076. doi:10.1126/science.1080029

Yagil, E., & Rosner, A. (1971). Phosphorolysis of 5-fluoro-2'-deoxyuridine in escherichia coli and its inhibition by nucleosides. *Journal of Bacteriology, 108*(2), 760-764.

Yamada, K., Hyodo, S., Matsuno, Y. K., Kinoshita, M., Maruyama, S. Z., Osaka, Y. S., . . . Kakehi, K. (2007). Rapid and sensitive analysis of mucin-type glycans using an in-line flow glycan-releasing apparatus. *Analytical Biochemistry, 371*(1), 52-61. doi:10.1016/j.ab.2007.06.013

Yanagibashi, T., Hosono, A., Oyama, A., Tsuda, M., Hachimura, S., Takahashi, Y., . . . Kaminogawa, S. (2009). Bacteroides induce higher IgA production than lactobacillus by increasing activation-induced cytidine deaminase expression in B cells in murine peyer's patches. *Bioscience, Biotechnology, and Biochemistry, 73*(2), 372-377.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., . . . Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature, 486*(7402), 222-227. doi:10.1038/nature11053; 10.1038/nature11053

Yoo, E. M., & Morrison, S. L. (2005). IgA: An immune glycoprotein. *Clinical Immunology (Orlando, Fla.), 116*(1), 3-10. doi:10.1016/j.clim.2005.03.010

Zahner, D., & Hakenbeck, R. (2000). The streptococcus pneumoniae beta-galactosidase is a surface protein. *Journal of Bacteriology, 182*(20), 5919-5921.

Zhang, H., DiBaise, J. K., Zuccolo, A., Kudrna, D., Braidotti, M., Yu, Y., . . . Krajmalnik-Brown, R. (2009). Human gut microbiota in obesity and after gastric bypass. *Proceedings of the National Academy of Sciences of the United States of America, 106*(7), 2365-2370. doi:10.1073/pnas.0812600106

Zhou, X., Bent, S. J., Schneider, M. G., Davis, C. C., Islam, M. R., & Forney, L. J. (2004). Characterization of vaginal microbial communities in adult healthy women using cultivation-independent methods. *Microbiology (Reading, England), 150*(Pt 8), 2565-2573. doi:10.1099/mic.0.26905-0

Zrihan-Licht, S., Baruch, A., Elroy-Stein, O., Keydar, I., & Wreschner, D. H. (1994). Tyrosine phosphorylation of the MUC1 breast cancer membrane proteins. cytokine receptor-like molecules. *FEBS Letters, 356*(1), 130-136.

Zupancic, M. L., Frieman, M., Smith, D., Alvarez, R. A., Cummings, R. D., & Cormack, B. P. (2008). Glycan microarray analysis of candida glabrata adhesin ligand specificity. *Molecular Microbiology, 68*(3), 547-559. doi:10.1111/j.1365-2958.2008.06184.x

Appendices

Appendix A – Cloning, mutation, deletion and tag detection primers

Recombinant protein code	Primers		
	Type	**RE**	**Sequence**
BT_4272-M60L	Forward	*BamHI*	CGCG GGATCC GACGAACAACTGATAACAGTATTC
	Reverse	*XhoI*	CCGG CTCGAG CTATTTATTTGCTTTAATACGTTGCC
BT_4272-CBM32	Forward	*BamHI*	CGCG GGATCC GATATAAAGTTGAAAGTTACCGGTG
	Reverse	*XhoI*	CCGG CTCGAG TTATTCCGCAGCATGGAAAAATTCCAT
BT_4244-M60L	Forward	*BamHI*	CGCG GGATCC GAT AAA ACA CTG GAT AAA CAA CTT C
	Reverse	*EcoRI*	CCGG GAATTC TTA TAA CAG AAT ACG TTT TCC GTC
BT_4244-CBM32	Forward	*NcoI*	CGCG CCATGG ACATCAAGGTTACACCAACC
	Reverse	*XhoI*	CCGG CTCGAG CGTATTTGTTTTGTAAAATTCCATTTC
BT_3015-M60L	Forward	*BamHI*	CGCG GGATCC CCCGTGAACTTTGATTATTCG
	Reverse	*EcoRI*	CCGG GAATTC TTACATTGGAATGTCGATACGTTC
BT_3015-CBM32	Forward	*BamHI*	CGCG GGATCC ATCAAGATCAAAATAGTAAGCGG
	Reverse	*EcoRI*	CCGG GAATTC TTAACGGAACTCTGCGGGATACTT
BT4244-FL	Forward	*BamHI*	CGCG GGATCC AAG GAT ACC GAA AAA TCG ATT ATA
	Reverse	*EcoRI*	CCGG GAATTC TTA TAA CAG AAT ACG TTT TCC GTC
BT4244-BACON	Forward	*BamHI*	CGCG GGATCC AAG GAT ACC GAA AAA TCG ATT ATA
	Reverse	*EcoRI*	CCGG GAATTC TTAGCCTCCGGTTGGTGTAAC
BT4244-BACON CBM32	Forward	*BamHI*	CGCG GGATCC AAG GAT ACC GAA AAA TCG ATT ATA
	Reverse	*EcoRI*	CCGG GAATTC TTA CGT ATT TGT TTT GTA AAA TTC CAT TTC
BT4244-FL-E575D	Forward	none	GGGACCAGCTCATGATATTGGCCATGTTCATCAGGCAGC
	Reverse	one	GCTGCCTGATGAACATGGCCAATATCATGAGCTGGTCCC
TVAG199300-M60L	Forward	*SacI*	CGCG GAGCTC GT GAT ACT GTT CAA GCA CAA GAG
	Reverse	*SalI*	CCGG GTCGAC TTATGGAATTCTGAATGGTTTGGC
TVAG199300CBD	Forward	*SacI*	CGCG GAGCTC GTGAAAGATATAAGAGATCAAGAGAAG
	Reverse	*XhoI*	CCGG CTCGAG TTACAAGTTATCTCCGTTATCATTTTC
TVAG189150-M60L	Forward	*BamHI*	CGCG GGATCC ACAGATTCAGTAATGATAAGAGAC
	Reverse	*SalI*	CCGG GTCGAC TTATAACTCTTTTCTATCGCCAGG
TVAG339720-M60L	Forward	*BamHI*	CGCG GGATCC GGC ATC AAT ACA GTT CAA GTA C
	Reverse	*XhoI*	CCGG CTCGAG TTATTTCTCTCCATTTACTTTATCTTTAAG
TVAG339720-PA14	Forward	*BamHI*	CGCG GGA TCC ATGCATGCATTTGAGTTCGATG

	Reverse	*XhoI*	CCGG CTCGAG TTATGGATCGTTCTCAAGCTTTG
TVAG339720-GBDL	Forward	*BamHI*	CGCG GGATCC GACCACATCTTCAAGCCAAAG
	Reverse	*XhoI*	CCGG CTCGAG TTATCTATAATAAGCGTTTCCATCATC

Table A.1 - Cloning primers for various *B. thetaiotaomicron* and *T. vaginalis* M60-like encoding genes. RE stands for restriction enzyme used**.** Underlined sequences are corresponding restriction enzyme cleavage sites

Recombinant protein code	**Primers**	
	Type	**Sequence**
G157A	Forward	GAAAATACATATGATGCTAAATTTTCAACTGATGGAGC
	Reverse	GCTCCATCAGTTGAAAATTTAGCATCATATGTATTTTC
H168A	Forward	GCTGCCCCTTTCGCTACTCCGTGGGGACAATCCGCCAA
	Reverse	TTGGCGGATTGTCCCCACGGAGTAGCGAAAGGGGCAGC
N202A	Forward	CACGATCCGGTGCTGGAAACTTCGGAAAGGTCAAAG
	Reverse	CTTTGACCTTTCCGAAGTTTCCAGCACCGGATCGTG
F251A	Forward	GACCGGAATCAAAGCTGAAGTATTAAGCGGTCTGGGTG
	Reverse	CACCCAGACCGCTTAATACTTCAGCTTTGATTCCGGTC

Table A.2– List of BT4244-CBM32 mutation primers

Recombinant protein code	**Primers**		
	Type	**RE**	**Sequence**
BT4240-FL	Forward	*Bam*HI	CGCGGGATCC ATGAAAGATTTATCAAGTATTGTAGC
	Reverse	*Xho*I	CCGGCTCGAG TTATCCATTAACCAAGCACTCATTG
BT4241-FL	Forward	*Bam*HI	CGCGGGATCC ATGGCCGAAAAGACATCCGACAA
	Reverse	*Xho*I	CCGGCTCGAG TTATCGATAATCATATTTGGCGGC
BT4243-FL	Forward	*Bam*HI	CGCGGGATCC CAAAAGACAAAAGCAAAGTTCTCT
	Reverse	*Xho*I	CCGGCTCGAG TTATTCGGCAAAAGCATGTCTGTA
BT4245-FL	Forward	*Nco*I	CGCGCCATGGACAATTATGACGATACCTATCC
	Reverse	*Xho*I	CCGGCTCGAGTTCGGACAGTATGAACAGACT

Table A.3 - Cloning primers for selected members of the BT4240-50 Sus-like system. RE stands for restriction enzyme used **.** Underlined sequences are corresponding restriction enzyme cleavage sites

Primer name	Primer/code	RE	Primer sequence
ΔBT_4244Flank 1	Forward/A	*BamH*I	CGCGGGATCCCATGAGTTGCATATACATCACCAC
	Reverse/B	None	TCTTAAAAACTAATACAGCAAAAAAGAATAATTTAACTTAACACACATTAC
ΔBT_4244 Flank 2	Forward/C	None	GTAATGTGTGTTAAGTTAAATTATTCTTTTTTGCTGTATTAGTTTTTAAGA
	Reverse/D	*Xba*I	CGCGTCTAGACTCTTGAAGAGGGACAAACGTG
ΔPULBT_4240-50 Flank 1	Forward/A	*BamH*I	CGCGGGATCCGGATCGATTTCAAGCATAACAAAT
	Reverse/B	None	ATGGGTATGAATACGTTTGAGCCTTGGAGAATGGAAAATGGATAATTAGT
ΔPULBT_4240-50 Flank 2	Forward/C	None	ACTAATTATCCATTTTCCATTCTCCAAGGCTCAAACGTATTCATACCCAT
	Reverse/D	*Xba*I	CGCGTCTAGAGTATCTTCATTTACCACAGTATGGAT
ΔBT_3015 Flank 1	Forward/A	*BamH*I	CGCGGGATCCGGCTACTAATCAATGATGAATAACAAT
	Reverse/B	None	GTTATCTTGAATGTATCTGAGATACGATCTTTTTTTTTATTCAAGTACGACCATTG
ΔBT_3015 Flank 2	Forward/C	None	CAATGGTCGTACTTGAATAAAAAAAAAGATCGTATCTCAGATACATTCAAGATAAC
	Reverse/D	*Xba*I	CGCGTCTAGATGGCGCTGAACATCCGTTTGTAA
ΔBT_4272 Flank 1	Forward/A	*BamH*I	CGCGGGATCCATACATTCGCTTTATTCCGGAAGC
	Reverse/B	None	GATTGGGGATATTCAGTAAGTTTCCAGTATTTATAATTTTTGTTGTTTTTCCAT
ΔBT_4272 Flank 2	Forward/C	None	ATGGAAAAACAACAAAAATTATAAATACTGGAAACTTACTGAATATCCCCAATC
	Reverse/D	*Xba*I	CGCGTCTAGACAAACAATAGCGAGCTTATGGG
ΔBT_4240 Flank 1	Forward/A	*BamH*I	CGCGGGATCCCTCATACAAAGTGGATGGAACC
	Reverse/B	None	AATTATCCATTTTCCATTCTCCAGGGTCTTTTCTTTAAAATTTATA
ΔBT_4240 Flank 1	Forward/C	None	TATAAATTTTAAAGAAAAGACCCTGGAGAATGGAAAATGGATAATT
	Reverse/D	*Xba*I	CGCGTCTAGAACCAGTTCTTTATCATGGGAACC
ΔBT_4242 Flank 1	Forward/A	*BamH*I	CGCGGGATCCGATGAAATCTGGAAGCTAATCAAC
	Reverse/B	None	GTTTTTCAATCTTGTTTCTTTTTATAGGTGATTATAATAATCAGTA
ΔBT_4242 Flank 2	Forward/C	None	TACTGATTATTATAATCACCTATAAAAAGAAACAAGATTGAAAAAC
	Reverse/D	*Xba*I	CGCGTCTAGATTACCAGCTTGTAAAGGTTCGG

Table A.4 – Deletion primers for PULBT_4240-50 (PUL encoding the BT4240-50 Sus-like system) and its components from the genome of *B. thetaiotaomicron*. RE stands for restriction enzyme used**.** Underlined sequences are corresponding restriction enzyme cleavage sites

Tag	Primers	
	Type	Sequence
Tag1	Forward (Tag1_F)	ATGTCGCCAATTGTCACTTTCTCA (also Tag1 signature sequence)
	Reverse(Universal reverse)	CACAATATGAGCAACAAGGAATCC
Tag11	Forward(Tag11_F)	ATGCCGCGGATTTATTGGAAGAAG (also tag11 signature sequence)
	Reverse(Universal reverse)	CACAATATGAGCAACAAGGAATCC
Site		
NBU2-att1	Forward (NBU2att1_F)	CCTTTGCACCGCTTTCAACG
	Reverse (NBU2att1_R)	TCAACTAAACATGAGATACTAGC

Table A.5 - Primers used to confirm the integration of tagging sequences and their sites of integration in the genome of WT, ΔBT_4240-50 and ΔBT_4244 deletion mutants.

Appendix B – Extinction coefficients

Code name for recombinant protein	Extinction coefficient (M^{-1} cm^{-1})
BT4272-M60L	119555
BT4244-M60L	125390
BT3015-M60L	102610
BT4272-FL	162875
BT4272-CBM32	17545
BT_4244-BACON	16960
BT4244-CBM32	20400
BT_4244-BC	37485
BT3015-CBM32	23505
TVAG339720-M60L	159810
TVAG189150-M60L	139870
TVAG199300 -M60L	181685
TVAG339720-PA14	40590
TVAG339720-GBDL	54780
TVAG199300-CBD	110020
BT4240-FL	36705
BT4241-FL	240685
BT4243-FL	81290
BT4245-FL	60070

Table B.1 - Recombinant protein extinction coefficients

Appendix C – Vector maps

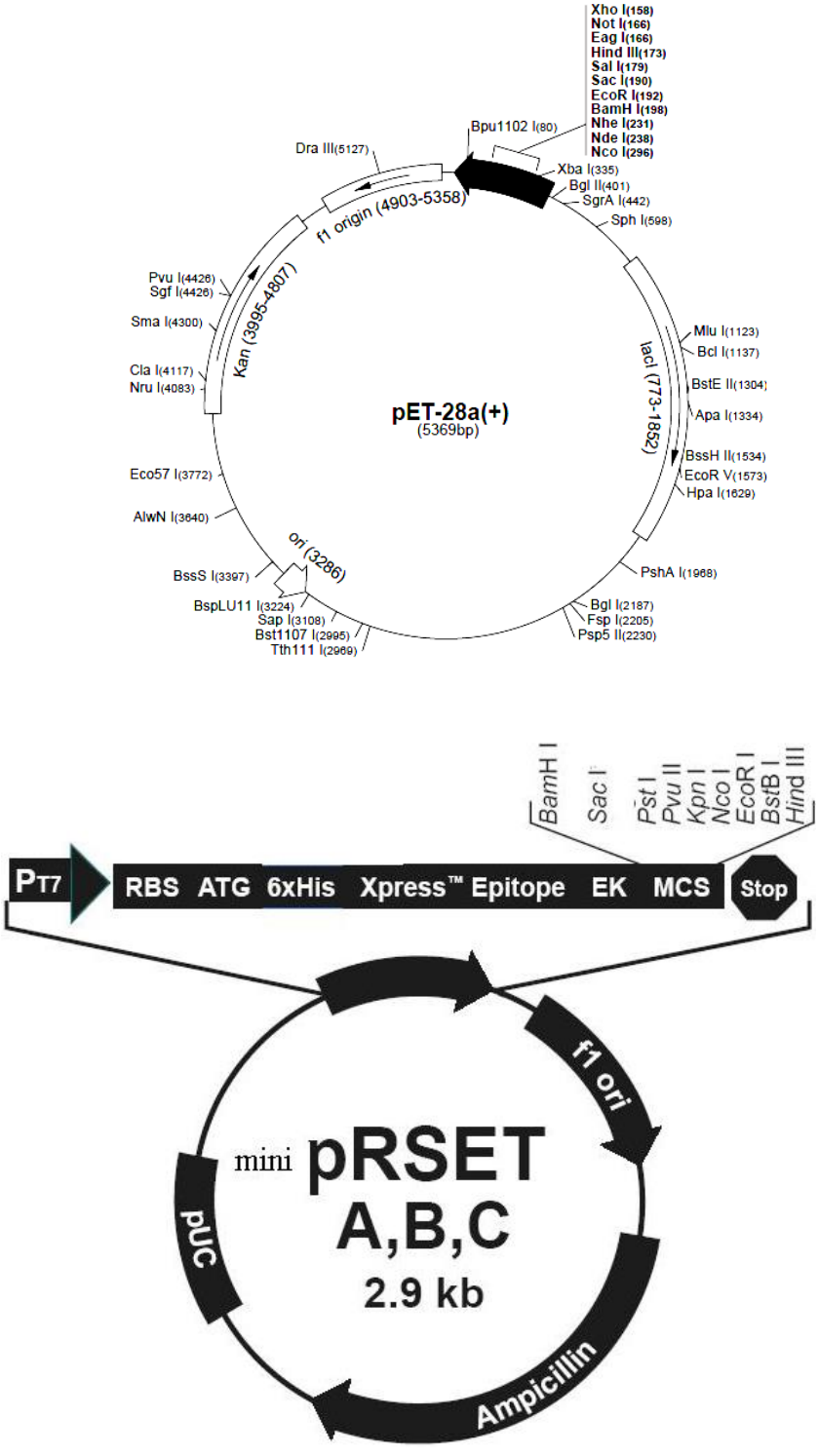


Figure C.1 – pET-28a (+) and *mini*PRSET A, B, C

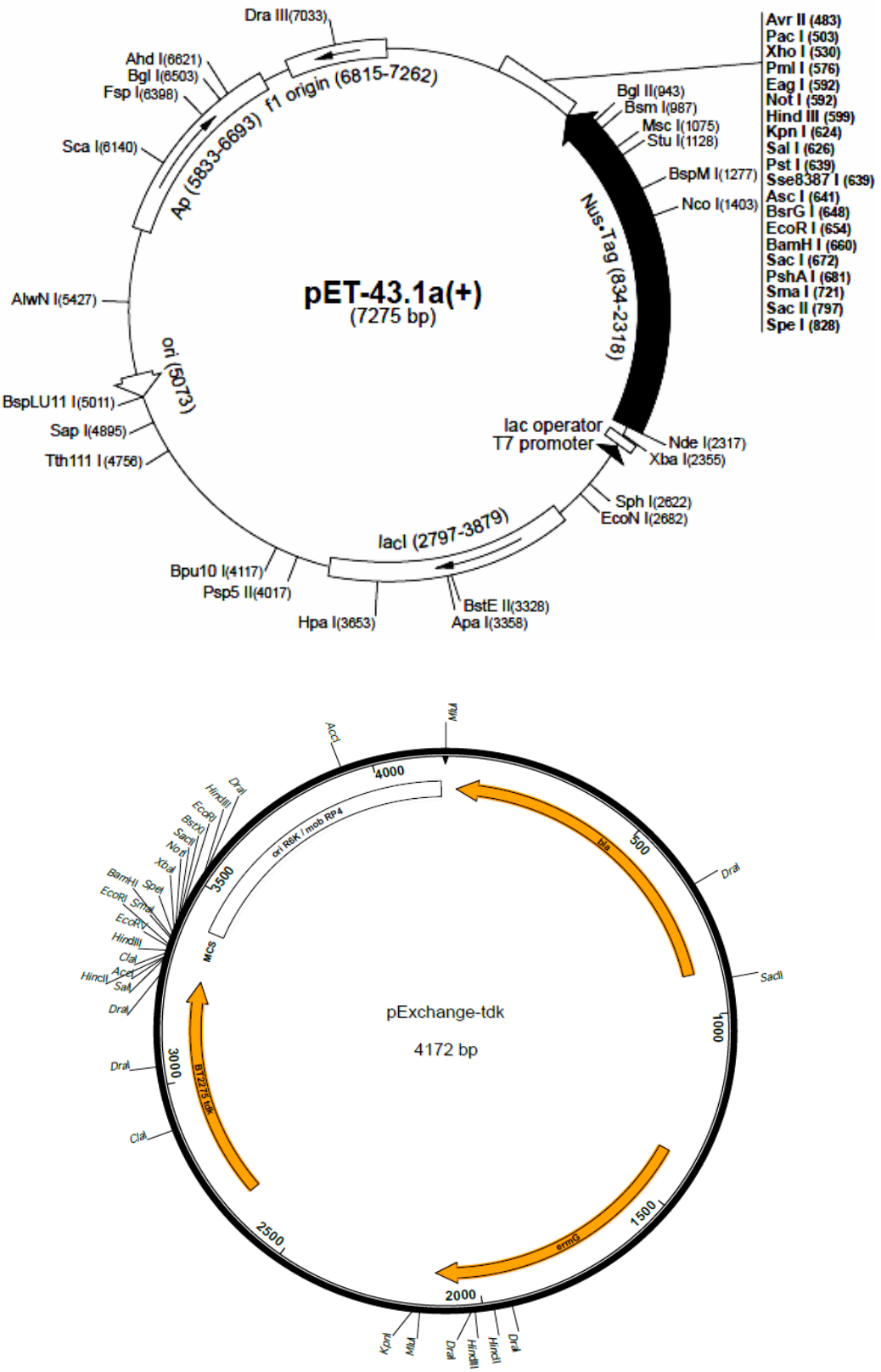


Figure C.2 - pET-43.1a (+) and pExchange-*tdk*

Appendix D – MUC 2 information

Human MUC2 sequence

>sp|Q02817|MUC2_HUMAN Mucin-2 OS=Homo sapiens GN=MUC2 PE=1 SV=2

MGLPLARLAAVCLALSLAGGSELQTEGRTRYHGRNVCSTWGNFHYKTFDGDVFRFPGLCD
YNFASDCRGSYKEFAVHLKRGPGQAEAPAGVESILLTIKDDTIYLTRHLAVLNGAVVSTP
HYSPGLLIEKSDAYTKVYSRAGLTLMWNREDALMLELDTKFRNHTCGLCGDYNGLQSYSE
FLSDGVLFSPLEFGNMQKINQPDVVCEDPEEEVAPASCSEHRAECERLLTAEAFADCQDL
VPLEPYLRACQQDRCRCPGGDTCVCSTVAEFSRQCSHAGGRPGNWRTATLCPKTCPGNLV
YLESGSPCMDTCSHLEVSSLCEEHRMDGCFCPEGTVYDDIGDSGCVPVSQCHCRLHGHLY
TPGQEITNDCEQCVCNAGRWVCKDLPCPGTCALEGGSHITTFDGKTYTFHGDCYYVLAKG
DHNDSYALLGELAPCGSTDKQTCLKTVVLLADKKKNAVVFKSDGSVLLNQLQVNLPHVTA
SFSVFRPSSYHIMVSMAIGVRLQVQLAPVMQLFVTLDQASQGQVQGLCGNFNGLEGDDFK
TASGLVEATGAGFANTWKAQSTCHDKLDWLDDPCSLNIESANYAEHWCSLLKKTETPFGR
CHSAVDPAEYYKRCKYDTCNCQNNEDCLCAALSSYARACTAKGVMLWGWREHVCNKDVGS
CPNSQVFLYNLTTCQQTCRSLSEADSHCLEGFAPVDGCGCPDHTFLDEKGRCVPLAKCSC
YHRGLYLEAGDVVVRQEERCVCRDGRLHCRQIRLIGQSCTAPKIHMDCSNLTALATSKPR
ALSCQTLAAGYYHTECVSGCVCPDGLMDDGRGGCVVEKECPCVHNNDLYSSGAKIKVDCN
TCTCKRGRWVCTQAVCHGTCSIYGSGHYITFDGKYYDFDGHCSYVAVQDYCGQNSSLGSF
SIITENVPCGTTGVTCSKAIKIFMGRTELKLEDKHRVVIQRDEGHHVAYTTREVGQYLVV
ESSTGIIVIWDKRTTVFIKLAPSYKGTVCGLCGNFDHRSNNDFTTRDHMVVSSELDFGNS
WKEAPTCPDVSTNPEPCSLNPHRRSWAEKQCSILKSSVFSICHSKVDPKPFYEACVHDSC
SCDTGGDCECFCSAVASYAQECTKEGACVFWRTPDLCPIFCDYYNPPHECEWHYEPCGNR
SFETCRTINGIHSNISVSYLEGCYPRCPKDRPIYEEDLKKCVTADKCGCYVEDTHYPPGA
SVPTEETCKSCVCTNSSQVVCRPEEGKILNQTQDGAFCYWEICGPNGTVEKHFNICSITT
RPSTLTTFTTITLPTTPTSFTTTTTTTTPTSSTVLSTTPKLCCLWSDWINEDHPSSGSDD
GDREPFDGVCGAPEDIECRSVKDPHLSLEQHGQKVQCDVSVGFICKNEDQFGNGPFGLCY
DYKIRVNCCWPMDKCITTPSPPTTTPSPPPTTTTTLPPTTTPSPPTTTTTTPPPTTTPSP
PITTTTTPLPTTTPSPPISTTTTPPPTTTPSPPTTTPSPPTTTPSPPTTTTTTPPPTTTP
SPPMTTPITPPASTTTLPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTT
TPSPPPTTTTTPPPTTTPSPPTTTTPSPPTITTTTPPPTTTPSPPTTTTTTPPPTTTPSP
PTTTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTTPSPPITTTTTPPPTTT
PSSPITTTPSPPTTTMTTPSPTTTPSSPITTTTTPSSTTTPSPPPTTMTTPSPTTTPSPP
TTTMTTLPPTTTSSPLTTTPLPPSITPPTFSPFSTTTPTTPCVPLCNWTGWLDSGKPNFH
KPGGDTELIGDVCGPGWAANISCRATMYPDVPIGQLGQTVVCDVSVGLICKNEDQKPGGV
IPMAFCLNYEINVQCCECVTQPTTMTTTTTENPTPPTTTPITTTTTVTPTPTPTGTQTPT
TTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTP
TPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPIT
TTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGT
QTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTV
TPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTT
TPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPT
PTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITT
TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQ
TPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVT
PTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTT
PITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTP
TGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTT
TTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQT
PTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTP
TPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTP
ITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPT
GTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTT
TVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTP
TTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPT
PTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPI
TTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTG
TQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTT
VTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPT

TTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTP
TPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPIT
TTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGT
QTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTV
TPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTT
TPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPT
PTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITT
TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQ
TPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVT
PTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTT
PITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTP
TGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTT
TTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQT
PTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTP
TPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTGPPTH
TSTAPIAELTTSNPPPESSTPQTSRSTSSPLTESTTLLSTLPPAIEMTSTAPPSTPTAPT
TTSGGHTLSPPPSTTTSPPGTPTRGTTTGSSSAPTPSTVQTTTTSAWTPTPTPLSTPSII
RTTGLRPYPSSVLICCVLNDTYYAPGEEVYNGTYGDTCYFVNCSLSCTLEFYNWSCPSTP
SPTPTPSKSTPTPSKPSSTPSKPTPGTKPPECPDFDPPRQENETWWLCDCFMATCKYNNT
VEIVKVECEPPPMPTCSNGLQPVRVEDPDGCCWHWECDCYCTGWGDPHYVTFDGLYYSYQ
GNCTYVLVEEISPSVDNFGVYIDNYHCDPNDKVSCPRTLIVRHETQEVLIKTVHMMPMQV
QVQVNRQAVALPYKKYGLEVYQSGINYVVDIPELGVLVSYNGLSFSVRLPYHRFGNNTKG
QCGTCTNTTSDDCILPSGEIVSNCEAAADQWLVNDPSKPHCPHSSSTTKRPAVTVPGGGK
TTPHKDCTPSPLCQLIKDSLFAQCHALVPPQHYYDACVFDSCFMPGSSLECASLQAYAAL
CAQQNICLDWRNHTHGACLVECPSHREYQACGPAEEPTCKSSSSQQNNTVLVEGCFCPEG
TMNYAPGFDVCVKTCGCVGPDNVPREFGEHFEFDCKNCVCLEGGSGIICQPKRCSQKPVT
HCVEDGTYLATEVNPADTCCNITVCKCNTSLCKEKPSVCPLGFEVKSKMVPGRCCPFYWC
ESKGVCVHGNAEYQPGSPVYSSKCQDCVCTDKVDNNTLLNVIACTHVPCNTSCSPGFELM
EAPGECCKKCEQTHCIIKRPDNQHVILKPGDFKSDPKNNCTFFSCVKIHNQLISSVSNIT
CPNFDASICIPGSITFMPNGCCKTCTPRNETRVPCSTVPVTTEVSYAGCTKTVLMNHCSG
SCGTFVMYSAKAQALDHSCSCCKEEKTSQREVVLSCPNGGSLTHTYTHIESCQCQDTVCG
LPTGTSRRARRSPRHLGSG

Neutral
Gal-3GalNAcol
GlcNAc-3GalNAcol
GalNAc-3GalNAcol
Fuc-Gal-3GalNAcol
Gal-GlcNAc-3GalNAcol[4]
Gal-GlcNAc-3GalNAcol[4]
Gal-(Fuc)GlcNAc-3GalNAcol
HexNAc-Gal-GlcNAc-3GalNAcol
Fuc-Gal-(Fuc)GlcNAc-3GalNAcol
Fuc-Gal-3(Gal-GlcNAc-6)GalNAcol

Monosialylated
NeuAc-6GalNAcol

Gal-3(NeuAc-6)GalNAcol
GlcNAc-3(NeuAc-6)GalNAcol
GalNAc-3(NeuAc-6)GalNAcol
Fuc-Gal-3(NeuAc-6)GalNAcol
Fuc-GlcNAc-3(NeuAc-6)GalNAcol
NeuAc-Gal-3(GlcNAc-6)GalNAcol
Gal-GlcNAc-3(NeuAc-6)GalNAcol
NeuAc-Gal-(Fuc)GlcNAc-3GalNAcol
Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol
Fuc-Gal-GlcNAc-3(NeuAc-6)GalNAcol
$[NeuAc]_1$HexNAc-Gal-GlcNAc-3GalNAcol[5]
$[NeuAc]_1$HexNAc-Gal-GlcNAc-3GalNAcol[5]
Fuc-Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol
HexNAc-(Fuc)Gal-(NeuAc)GlcNAc-GalNAcol
GalNAc-(NeuAc)Gal-(Fuc)GlcNAc-GalNAcol[3]
Gal-GlcNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol
HexNAc-(Fuc)Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol
Gal-GlcNAc-Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol
Monosulfated
SO_3^--Gal-GlcNAc-3GalNAcol
SO_3^--Gal-(Fuc)GlcNAc-3GalNAcol[4]
SO_3^--Gal-(Fuc)GlcNAc-3GalNAcol[4]
Gal-3(SO_3^--Gal-GlcNAc-6)GalNAcol
$[SO_3^-]_1[Gal]_2[GlcNAc]_1$GalNAcol
Gal-3(SO_3^--Gal-(Fuc)GlcNAc-6)GalNAcol[6]
SO_3^--Gal-(Fuc)GlcNAc-Gal-3GalNAcol[6]
HexNAc-(SO_3^-)Gal-(Fuc)GlcNAc-3GalNAcol
Fuc-Gal-3(SO_3^--Gal-(Fuc)GlcNAc-6)GalNAcol
Gal-GlcNAc-(SO_3^-)Gal-(Fuc)GlcNAc-3GalNAcol
SO_3^--Gal-(Fuc)GlcNAc-Gal-(Fuc)GlcNAc-3GalNAcol
Multiple acidic residues
SO_3^--Gal-GlcNAc-3(NeuAc-6)GalNAcol
SO_3^--Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol

SO_3^--Gal-GlcNAc-Gal-3(NeuAc-6)GalNAcol^6
Gal-(SO_3^-)GlcNAc-Gal-3(NeuAc-6)GalNAcol^6
SO_3^--GlcNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol
NeuAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^4
NeuAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^4
SO_3^--Gal-(Fuc)GlcNAc-Gal-3(SO_3^--GlcNAc-6)GalNAcol
NeuAc-Gal-3(Gal-(Fuc)(SO_3^-)GlcNAc-6)GalNAcol
NeuAc-Gal-(Fuc)(SO_3^-)GlcNAc-3(GlcNAc-6)GalNAcol
Fuc-GlcNAc-(SO_3^-)Gal-GlcNAc-3(NeuAc-6)GalNAcol
NeuAc-Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol
$[NeuAc]_1SO_3^-$-Gal-GlcNAc-Gal-GlcNAc-3GalNAcol^5
SO_3^--Gal-GlcNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^4
SO_3^--Gal-GlcNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^4
$[NeuAc]_1$GalNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^{4,5}
$[NeuAc]_1$GalNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^{4,5}
$[NeuAc]_1$GalNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^{4,5}
SO_3^--Gal-(Fuc)GlcNAc-3(SO_3^--Gal-(Fuc)GlcNAc-6)GalNAcol
$[NeuAc]_1SO_3^-$-Gal-GlcNAc-Gal-3(NeuAc-6)GalNAcol^5
$[NeuAc]_1$NeuAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol
SO_3^--Gal-GlcNAc-Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol^4
SO_3^--Gal-GlcNAc-Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol^4
SO_3^--Gal-(Fuc)GlcNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol
$[NeuAc]_1$Gal-GlcNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^{4,5}
$[NeuAc]_1$Gal-GlcNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^{4,5}
$[NeuAc]_1SO_3^-$-Gal-GlcNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^{4,5}
$[NeuAc]_1SO_3^-$-Gal-GlcNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^{4,5}
NeuAc-Gal-GlcNAc-Gal-(SO_3^-)GlcNAc-3(NeuAc-6)GalNAcol
Gal-(Fuc)GlcNAc-(SO_3^-)Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol
SO_3^--Gal-(Fuc)GlcNAc-Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol
GalNAc-(NeuAc)Gal-GlcNAc-(SO_3^-)Gal-(Fuc)GlcNAc-GalNAcol^3
NeuAc-Gal-GlcNAc-Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol^4
NeuAc-Gal-(Fuc)GlcNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol
NeuAc-Gal-GlcNAc-Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol^4
$[NeuAc]_1$HexNAc-Gal-GlcNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^{4,5}
$[NeuAc]_1$HexNAc-Gal-GlcNAc-Gal-GlcNAc-3(NeuAc-6)GalNAcol^{4,5}

$[NeuAc]_1HexNAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{4,5}$
$[NeuAc]_1SO_3^-\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}(Fuc)GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{5}$
$[NeuAc]_1SO_3^-\text{-}Gal\text{-}(Fuc)GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{5}$
$NeuAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}(Fuc)(SO_3^-)GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{4}$
$NeuAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}(Fuc)(SO_3^-)GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{4}$
$[NeuAc]_1NeuAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{4,5}$
$[NeuAc]_1NeuAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{4,5}$
$[NeuAc]_1[Fuc]_1Gal\text{-}(Fuc)GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{5}$
NeuAc-Gal-(Fuc)GlcNAc-Gal-(Fuc)GlcNAc-3(NeuAc-6)GalNAcol
$[NeuAc]_1HexNAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}(Fuc)GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{4,5}$
$[NeuAc]_1HexNAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}(Fuc)GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{4,5}$
$[NeuAc]_1NeuAc\text{-}Gal\text{-}(Fuc)GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(GlcNAc\text{-}6)GalNAcol^{5}$
$[NeuAc]_1NeuAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}(Fuc)GlcNAc\text{-}3(GlcNAc\text{-}6)GalNAcol^{4,5}$
$[NeuAc]_1NeuAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}(Fuc)GlcNAc\text{-}3(GlcNAc\text{-}6)GalNAcol^{4,5}$
$[NeuAc]_1Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{5}$
$[NeuAc]_1SO_3^-\text{-}Gal\text{-}(Fuc)GlcNAc\text{-}Gal\text{-}(Fuc)GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{5}$
$GalNAc\text{-}(NeuAc)Gal\text{-}GlcNAc\text{-}(SO_3^-)Gal\text{-}(Fuc)GlcNAc\text{-}3(NeuAc\text{-}6)\text{-}GalNAcol^{3}$
$[NeuAc]_1NeuAc\text{-}Gal\text{-}(Fuc)GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{5}$
$[NeuAc]_1SO_3^-\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{4,5}$
$[NeuAc]_1SO_3^-\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{4,5}$
$[NeuAc]_1GalNAc\text{-}(NeuAc)Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{3,5}$
$[NeuAc]_1HexNAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{5}$
$[SO_3^-]_1[NeuAc]_1[Fuc]_1[Gal]_3[GlcNAc]_3(NeuAc\text{-}6)GalNAcol$
$[NeuAc]_1[Fuc]_1GalNAc\text{-}(NeuAc)Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{3,5}$
$[NeuAc]_2SO_3^-\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}3(NeuAc\text{-}6)GalNAcol^{5}$
$[NeuAc]_2HexNAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc\text{-}Gal\text{-}GlcNAc(NeuAc\text{-}6)GalNAcol^{5}$
$[NeuAc]_2[HexNAc]_2[Gal]_3[GlcNAc]_3(NeuAc\text{-}6)GalNAcol$

Table D.1 - **MUC2 glycans from mass spectrometry.** Please see Larsson *et al.,* 2009 for more details

Appendix E - Taxonomic distribution of BT4240 homologues

cellular organisms ..... 227 hits 188 orgs [root]
. Bacteroidales ..... 225 hits 187 orgs [Bacteria; Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia]
. . Bacteroides ..... 165 hits 135 orgs [Bacteroidaceae]
. . . Bacteroides thetaiotaomicron ..... 3 hits 2 orgs
. . . . Bacteroides thetaiotaomicron VPI-5482 ..... 2 hits 1 orgs
. . . . Bacteroides thetaiotaomicron dnLKV9 ..... 1 hits 1 orgs
. . . Bacteroides sp. 1_1_6 ..... 2 hits 1 orgs
. . . Bacteroides sp. 1_1_14 ..... 2 hits 1 orgs
. . . environmental samples ..... 28 hits 28 orgs
. . . . Bacteroides thetaiotaomicron CAG:40 ..... 1 hits 1 orgs
. . . . Bacteroides faecis CAG:32 ..... 1 hits 1 orgs
. . . . Bacteroides caccae CAG:21 ..... 1 hits 1 orgs
. . . . Bacteroides sp. CAG:754 ..... 1 hits 1 orgs
. . . . Bacteroides finegoldii CAG:203 ..... 1 hits 1 orgs
. . . . Bacteroides ovatus CAG:22 ..... 1 hits 1 orgs
. . . . Bacteroides fragilis CAG:47 ..... 1 hits 1 orgs
. . . . Bacteroides fragilis CAG:558 ..... 1 hits 1 orgs
. . . . Bacteroides sp. CAG:189 ..... 1 hits 1 orgs
. . . . Bacteroides uniformis CAG:3 ..... 1 hits 1 orgs
. . . . Bacteroides sp. CAG:633 ..... 1 hits 1 orgs
. . . . Bacteroides cellulosilyticus CAG:158 ..... 1 hits 1 orgs
. . . . Bacteroides stercoris CAG:120 ..... 1 hits 1 orgs
. . . . Bacteroides eggerthii CAG:109 ..... 1 hits 1 orgs
. . . . Bacteroides intestinalis CAG:315 ..... 1 hits 1 orgs
. . . . Bacteroides sp. CAG:598 ..... 1 hits 1 orgs
. . . . Bacteroides sp. CAG:661 ..... 1 hits 1 orgs
. . . . Bacteroides sp. CAG:20 ..... 1 hits 1 orgs
. . . . Bacteroides sp. CAG:98 ..... 1 hits 1 orgs
. . . . Bacteroides vulgatus CAG:6 ..... 1 hits 1 orgs
. . . . Bacteroides sp. CAG:714 ..... 1 hits 1 orgs
. . . . Bacteroides plebeius CAG:211 ..... 1 hits 1 orgs
. . . . Bacteroides coprophilus CAG:333 ..... 1 hits 1 orgs
. . . . Bacteroides sp. CAG:875 ..... 1 hits 1 orgs
. . . . Bacteroides sp. CAG:443 ..... 1 hits 1 orgs
. . . . Bacteroides sp. CAG:1076 ..... 1 hits 1 orgs
. . . . Bacteroides sp. CAG:530 ..... 1 hits 1 orgs
. . . . Bacteroides coprocola CAG:162 ..... 1 hits 1 orgs
. . . Bacteroides faecis ..... 1 hits 1 orgs
. . . Bacteroides caccae ..... 3 hits 3 orgs
. . . . Bacteroides caccae ATCC 43185 ..... 1 hits 1 orgs
. . . . Bacteroides caccae CL03T12C61 ..... 1 hits 1 orgs
. . . Bacteroides finegoldii ..... 4 hits 3 orgs
. . . . Bacteroides finegoldii CL09T03C10 ..... 1 hits 1 orgs
. . . . Bacteroides finegoldii DSM 17565 ..... 1 hits 1 orgs
. . . Bacteroides ovatus ..... 9 hits 7 orgs
. . . . Bacteroides ovatus SD CMC 3f ..... 1 hits 1 orgs
. . . . Bacteroides ovatus ATCC 8483 ..... 1 hits 1 orgs
. . . . Bacteroides ovatus CL02T12C04 ..... 1 hits 1 orgs
. . . . Bacteroides ovatus CL03T12C18 ..... 1 hits 1 orgs
. . . . Bacteroides ovatus 3_8_47FAA ..... 1 hits 1 orgs
. . . . Bacteroides ovatus SD CC 2a ..... 1 hits 1 orgs
. . . Bacteroides sp. 3_1_23 ..... 1 hits 1 orgs
. . . Bacteroides sp. D2 ..... 1 hits 1 orgs
. . . Bacteroides sp. 2_2_4 ..... 2 hits 1 orgs
. . . Bacteroides sp. 1_1_30 ..... 2 hits 1 orgs
. . . Bacteroides xylanisolvens ..... 6 hits 4 orgs
. . . . Bacteroides xylanisolvens CL03T12C04 ..... 1 hits 1 orgs
. . . . Bacteroides xylanisolvens SD CC 1b ..... 1 hits 1 orgs
. . . . Bacteroides xylanisolvens XB1A ..... 2 hits 1 orgs
. . . Bacteroides sp. D1 ..... 2 hits 1 orgs
. . . Bacteroides sp. 2_1_22 ..... 1 hits 1 orgs
. . . Bacteroides sp. D22 ..... 2 hits 1 orgs

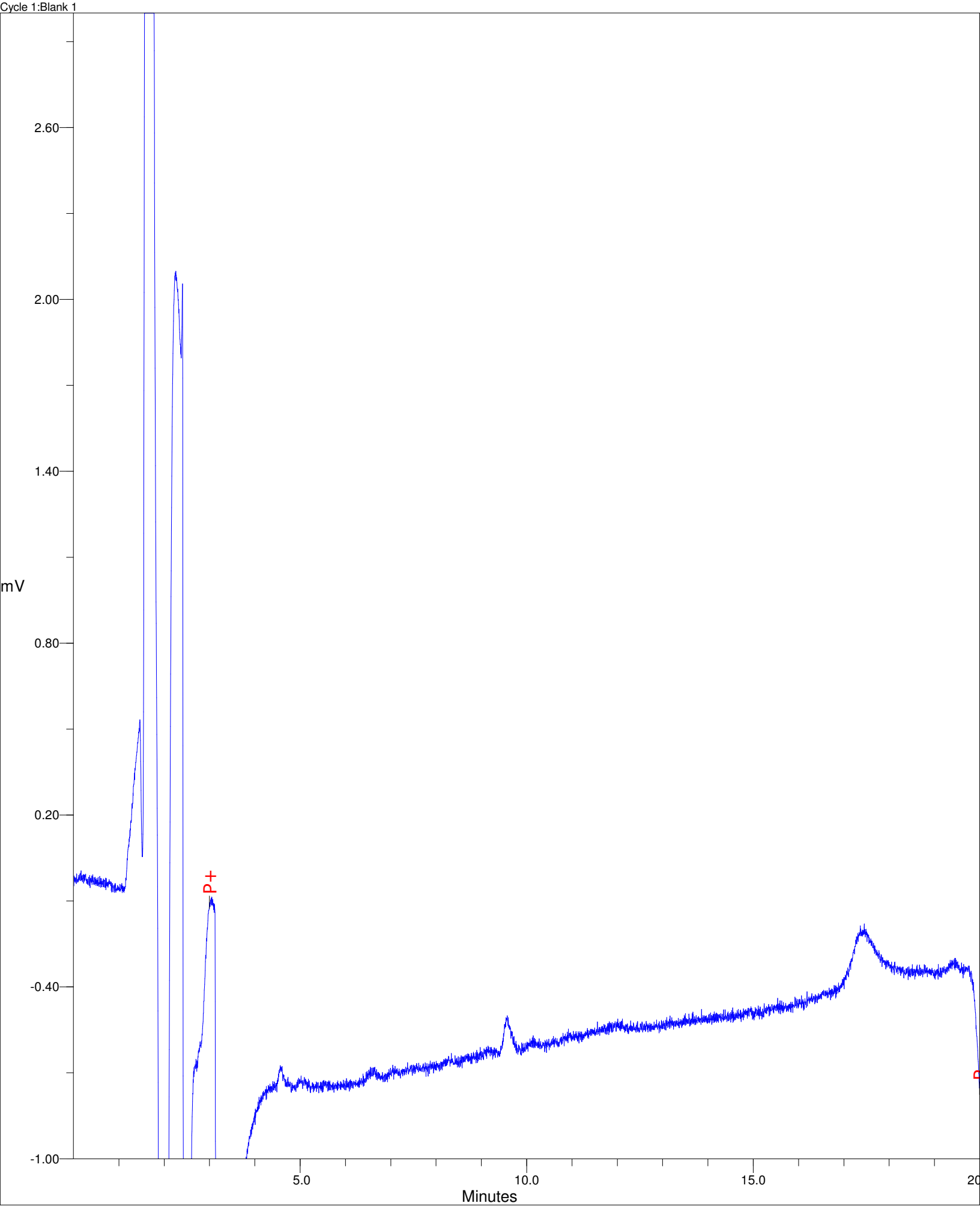
. . . Bacteroides fragilis ..... 19 hits 14 orgs
. . . . Bacteroides fragilis YCH46 ..... 2 hits 1 orgs
. . . . Bacteroides fragilis NCTC 9343 ..... 2 hits 1 orgs
. . . . Bacteroides fragilis 638R ..... 2 hits 1 orgs
. . . . Bacteroides fragilis CL07T00C01 ..... 1 hits 1 orgs
. . . . Bacteroides fragilis CL03T12C07 ..... 1 hits 1 orgs
. . . . Bacteroides fragilis CL03T00C08 ..... 1 hits 1 orgs
. . . . Bacteroides fragilis CL05T00C42 ..... 1 hits 1 orgs
. . . . Bacteroides fragilis CL07T12C05 ..... 1 hits 1 orgs
. . . . Bacteroides fragilis CL05T12C13 ..... 1 hits 1 orgs
. . . . Bacteroides fragilis HMW 615 ..... 1 hits 1 orgs
. . . . Bacteroides fragilis 3_1_12 ..... 1 hits 1 orgs
. . . . Bacteroides fragilis HMW 616 ..... 1 hits 1 orgs
. . . . Bacteroides fragilis HMW 610 ..... 1 hits 1 orgs
. . . Bacteroides sp. 3_2_5 ..... 1 hits 1 orgs
. . . Bacteroides sp. 2_1_16 ..... 1 hits 1 orgs
. . . Bacteroides sp. 2_1_56FAA ..... 1 hits 1 orgs
. . . Bacteroides nordii ..... 2 hits 2 orgs
. . . . Bacteroides nordii CL02T12C05 ..... 1 hits 1 orgs
. . . Bacteroides sp. HPS0048 ..... 2 hits 1 orgs
. . . Bacteroides salyersiae ..... 3 hits 3 orgs
. . . . Bacteroides salyersiae CL02T12C01 ..... 1 hits 1 orgs
. . . . Bacteroides salyersiae WAL 10018 = DSM 18765 = JCM 12988 ..... 1 hits 1 orgs
. . . Bacteroides sp. D20 ..... 2 hits 1 orgs
. . . Bacteroides uniformis ..... 6 hits 5 orgs
. . . . Bacteroides uniformis CL03T12C37 ..... 1 hits 1 orgs
. . . . Bacteroides uniformis CL03T00C23 ..... 1 hits 1 orgs
. . . . Bacteroides uniformis dnLKV2 ..... 1 hits 1 orgs
. . . . Bacteroides uniformis ATCC 8492 ..... 1 hits 1 orgs
. . . Bacteroides sp. 4_1_36 ..... 2 hits 1 orgs
. . . Bacteroides helcogenes ..... 3 hits 2 orgs
. . . . Bacteroides helcogenes P 36-108 ..... 2 hits 1 orgs
. . . Bacteroides fluxus ..... 2 hits 2 orgs
. . . . Bacteroides fluxus YIT 12057 ..... 1 hits 1 orgs
. . . Bacteroides cellulosilyticus ..... 4 hits 3 orgs
. . . . Bacteroides cellulosilyticus CL02T12C19 ..... 1 hits 1 orgs
. . . . Bacteroides cellulosilyticus DSM 14838 ..... 1 hits 1 orgs
. . . Bacteroides stercoris ..... 3 hits 3 orgs
. . . . Bacteroides stercoris ATCC 43183 ..... 1 hits 1 orgs
. . . . Bacteroides stercoris CC31F ..... 1 hits 1 orgs
. . . Bacteroides clarus ..... 2 hits 2 orgs
. . . . Bacteroides clarus YIT 12056 ..... 1 hits 1 orgs
. . . Bacteroides eggerthii ..... 4 hits 3 orgs
. . . . Bacteroides eggerthii DSM 20697 ..... 1 hits 1 orgs
. . . . Bacteroides eggerthii 1_2_48FAA ..... 1 hits 1 orgs
. . . Bacteroides intestinalis ..... 2 hits 2 orgs
. . . . Bacteroides intestinalis DSM 17393 ..... 1 hits 1 orgs
. . . Bacteroides oleiciplenus ..... 2 hits 2 orgs
. . . . Bacteroides oleiciplenus YIT 12058 ..... 1 hits 1 orgs
. . . Bacteroides gallinarum ..... 1 hits 1 orgs
. . . Bacteroides coprosuis ..... 2 hits 2 orgs
. . . . Bacteroides coprosuis DSM 18011 ..... 1 hits 1 orgs
. . . Bacteroides vulgatus ..... 7 hits 5 orgs
. . . . Bacteroides vulgatus ATCC 8482 ..... 2 hits 1 orgs
. . . . Bacteroides vulgatus PC510 ..... 1 hits 1 orgs
. . . . Bacteroides vulgatus CL09T03C04 ..... 1 hits 1 orgs
. . . . Bacteroides vulgatus dnLKV7 ..... 1 hits 1 orgs
. . . Bacteroides massiliensis ..... 4 hits 3 orgs
. . . . Bacteroides massiliensis B84634 = Timone 84634 = DSM 17679 = JCM 13223 . 1 hits 1 orgs
. . . . Bacteroides massiliensis dnLKV3 ..... 1 hits 1 orgs
. . . Bacteroides sp. 3_1_40A ..... 2 hits 1 orgs
. . . Bacteroides sp. 4_3_47FAA ..... 1 hits 1 orgs
. . . Bacteroides dorei ..... 7 hits 6 orgs
. . . . Bacteroides dorei 5_1_36/D4 ..... 1 hits 1 orgs
. . . . Bacteroides dorei CL03T12C01 ..... 1 hits 1 orgs
. . . . Bacteroides dorei DSM 17855 ..... 1 hits 1 orgs

. . . . Bacteroides dorei CL02T00C15 ..... 1 hits 1 orgs
. . . . Bacteroides dorei CL02T12C06 ..... 1 hits 1 orgs
. . . Bacteroides sp. 3_1_33FAA ..... 1 hits 1 orgs
. . . Bacteroides sp. 9_1_42FAA ..... 2 hits 1 orgs
. . . Bacteroides barnesiae ..... 1 hits 1 orgs
. . . Bacteroides plebeius ..... 2 hits 2 orgs
. . . . Bacteroides plebeius DSM 17135 ..... 1 hits 1 orgs
. . . Bacteroides coprophilus ..... 2 hits 2 orgs
. . . . Bacteroides coprophilus DSM 18228 = JCM 13818 ..... 1 hits 1 orgs
. . . Bacteroides sp. 2_1_33B ..... 1 hits 1 orgs
. . . Bacteroides sp. 3_1_19 ..... 1 hits 1 orgs
. . . Bacteroides sp. 20_3 ..... 1 hits 1 orgs
. . Porphyromonadaceae ..... 34 hits 26 orgs
. . . Porphyromonas ..... 11 hits 6 orgs
. . . . Porphyromonas gulae ..... 1 hits 1 orgs
. . . . Porphyromonas gingivalis ..... 10 hits 5 orgs
. . . . . Porphyromonas gingivalis JCVI SC001 ..... 1 hits 1 orgs
. . . . . Porphyromonas gingivalis ATCC 33277 ..... 2 hits 1 orgs
. . . . . Porphyromonas gingivalis W50 ..... 1 hits 1 orgs
. . . . . Porphyromonas gingivalis TDC60 ..... 2 hits 1 orgs
. . . Barnesiella intestinihominis ..... 2 hits 2 orgs [Barnesiella]
. . . . Barnesiella intestinihominis YIT 11860 ..... 1 hits 1 orgs
. . . Tannerella ..... 7 hits 5 orgs
. . . . Tannerella sp. 6_1_58FAA_CT1 ..... 2 hits 1 orgs
. . . . environmental samples ..... 2 hits 2 orgs
. . . . . Tannerella sp. CAG:51 ..... 1 hits 1 orgs
. . . . . Tannerella sp. CAG:118 ..... 1 hits 1 orgs
. . . . Tannerella forsythia ..... 3 hits 2 orgs
. . . . . Tannerella forsythia ATCC 43037 ..... 2 hits 1 orgs
. . . Parabacteroides ..... 14 hits 13 orgs
. . . . Parabacteroides distasonis ..... 5 hits 4 orgs
. . . . . Parabacteroides distasonis ATCC 8503 ..... 2 hits 1 orgs
. . . . . Parabacteroides distasonis CL03T12C09 ..... 1 hits 1 orgs
. . . . . Parabacteroides distasonis CL09T03C24 ..... 1 hits 1 orgs
. . . . Parabacteroides sp. D13 ..... 1 hits 1 orgs
. . . . Parabacteroides sp. D25 ..... 1 hits 1 orgs
. . . . environmental samples ..... 2 hits 2 orgs
. . . . . Parabacteroides sp. CAG:2 ..... 1 hits 1 orgs
. . . . . Parabacteroides johnsonii CAG:246 ..... 1 hits 1 orgs
. . . . Parabacteroides johnsonii ..... 3 hits 3 orgs
. . . . . Parabacteroides johnsonii DSM 18315 ..... 1 hits 1 orgs
. . . . . Parabacteroides johnsonii CL02T12C29 ..... 1 hits 1 orgs
. . . . Parabacteroides merdae ..... 2 hits 2 orgs
. . . . . Parabacteroides merdae CL03T12C32 ..... 1 hits 1 orgs
. . Prevotella ..... 24 hits 24 orgs [Prevotellaceae]
. . . environmental samples ..... 6 hits 6 orgs
. . . . Prevotella sp. CAG:755 ..... 1 hits 1 orgs
. . . . Prevotella sp. CAG:617 ..... 1 hits 1 orgs
. . . . Prevotella sp. CAG:474 ..... 1 hits 1 orgs
. . . . Prevotella sp. CAG:487 ..... 1 hits 1 orgs
. . . . Prevotella sp. CAG:891 ..... 1 hits 1 orgs
. . . . Prevotella sp. CAG:5226 ..... 1 hits 1 orgs
. . . Prevotella buccalis ..... 2 hits 2 orgs
. . . . Prevotella buccalis ATCC 35310 ..... 1 hits 1 orgs
. . . Prevotella timonensis ..... 2 hits 2 orgs
. . . . Prevotella timonensis CRIS 5C-B1 ..... 1 hits 1 orgs
. . . Prevotella oralis ..... 3 hits 3 orgs
. . . . Prevotella oralis ATCC 33269 ..... 1 hits 1 orgs
. . . . Prevotella oralis HGA0225 ..... 1 hits 1 orgs
. . . Prevotella saccharolytica ..... 2 hits 2 orgs
. . . . Prevotella saccharolytica F0055 ..... 1 hits 1 orgs
. . . Prevotella sp. oral taxon 317 ..... 2 hits 2 orgs
. . . . Prevotella sp. oral taxon 317 str. F0108 ..... 1 hits 1 orgs
. . . Prevotella bergensis ..... 2 hits 2 orgs
. . . . Prevotella bergensis DSM 17361 ..... 1 hits 1 orgs
. . . Prevotella loescheii ..... 1 hits 1 orgs

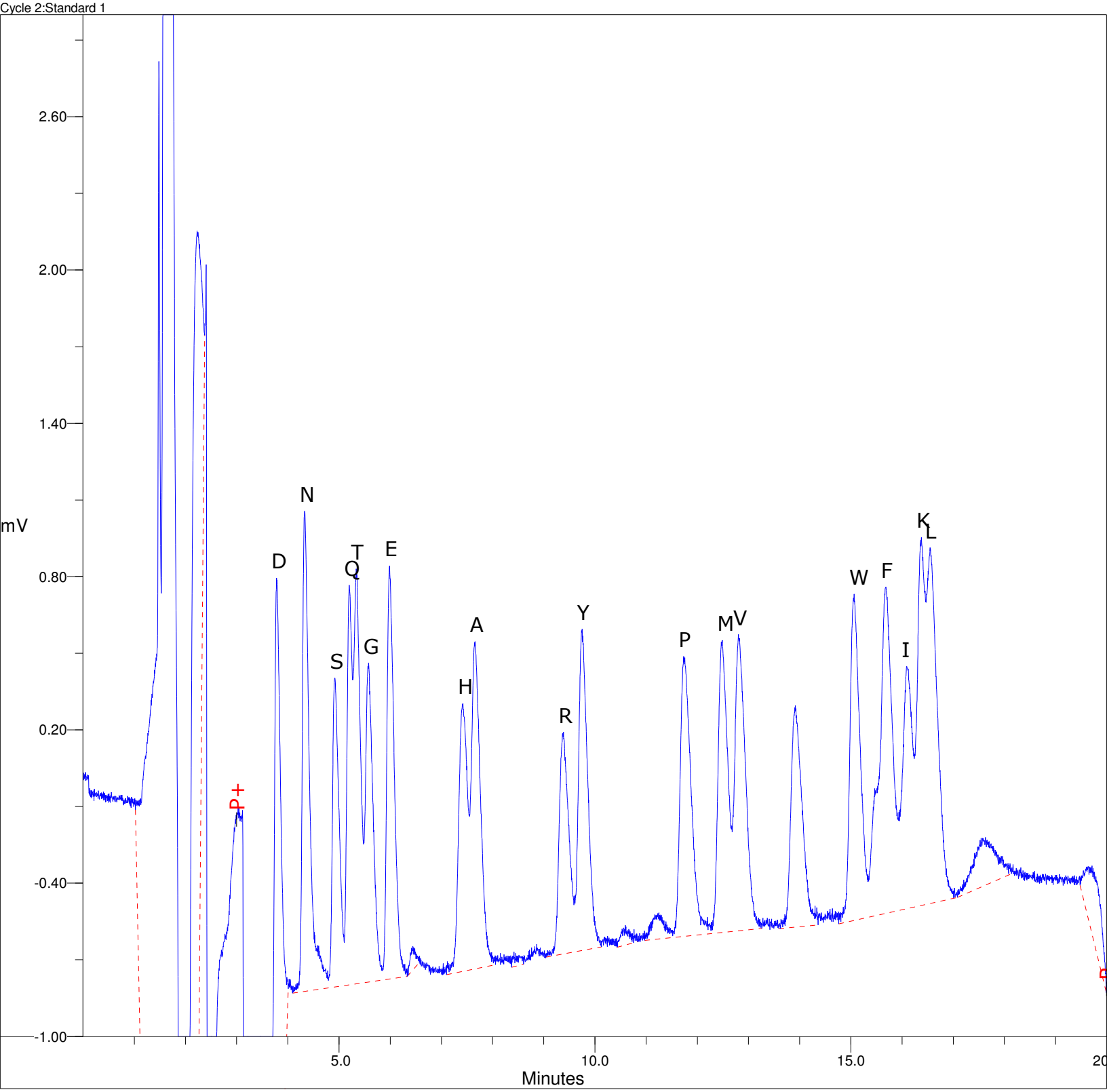
. . . Prevotella sp. oral taxon 472 ..... 2 hits 2 orgs
. . . . Prevotella sp. oral taxon 472 str. F0295 ..... 1 hits 1 orgs
. . . Prevotella multisaccharivorax ..... 2 hits 2 orgs
. . . . Prevotella multisaccharivorax DSM 17128 ..... 1 hits 1 orgs
. . environmental samples ..... 2 hits 2 orgs [Rikenellaceae; Alistipes]
. . . Alistipes sp. CAG:435 ..... 1 hits 1 orgs
. . . Alistipes sp. CAG:514 ..... 1 hits 1 orgs
. *Trichomonas vaginalis* G3 ..... 2 hits 1 orgs [Eukaryota; Parabasalia; Trichomonadida; Trichomonadidae; Trichomonas; *Trichomonas vaginalis*]

Appendix F – N-terminal Edman sequencing data for the Fc-α fragment of human myeloma IgA1 after digestion with the BT4244-FL protease

Cycle 1:Blank 1

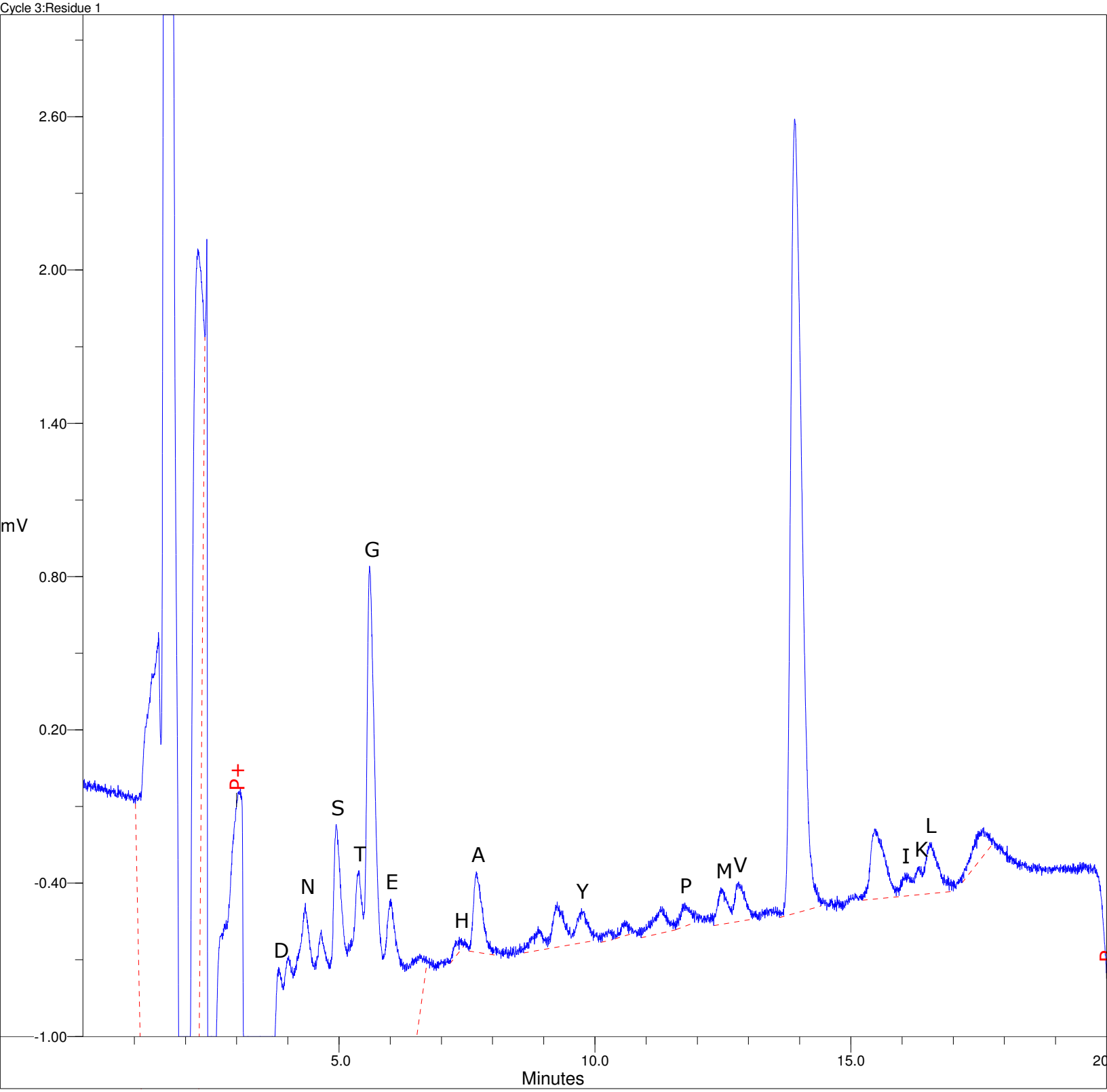


Cycle 2:Standard 1



PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT	PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	1.48		7.561		R	9.38	9.38	0.869	10.000
	1.65		14.911		Y	9.75	9.75	1.259	10.000
	2.23		4.086			10.59		0.079	
	3.03		6.054			11.25		0.106	
D	3.78	3.78	3.011	10.000	P	11.73	11.73	1.099	10.000
N	4.33	4.33	1.881	10.000	M	12.49	12.49	1.146	10.000
S	4.92	4.92	1.210	10.000	V	12.80	12.80	1.161	10.000
Q	5.20	5.20	1.565	10.000		13.91		0.868	
T	5.33	5.33	1.626	10.000	W	15.06	15.06	1.278	10.000
G	5.58	5.58	1.250	10.000	F	15.69	15.69	1.278	10.000
E	5.99	5.99	1.621	10.000	I	16.09	16.09	0.949	10.000
	6.44		0.093		K	16.38	16.38	1.441	10.000
H	7.42	7.42	1.051	10.000	L	16.55	16.55	1.393	10.000
A	7.65	7.65	1.283	10.000		17.63		0.187	
	8.42		0.058			19.71		0.267	

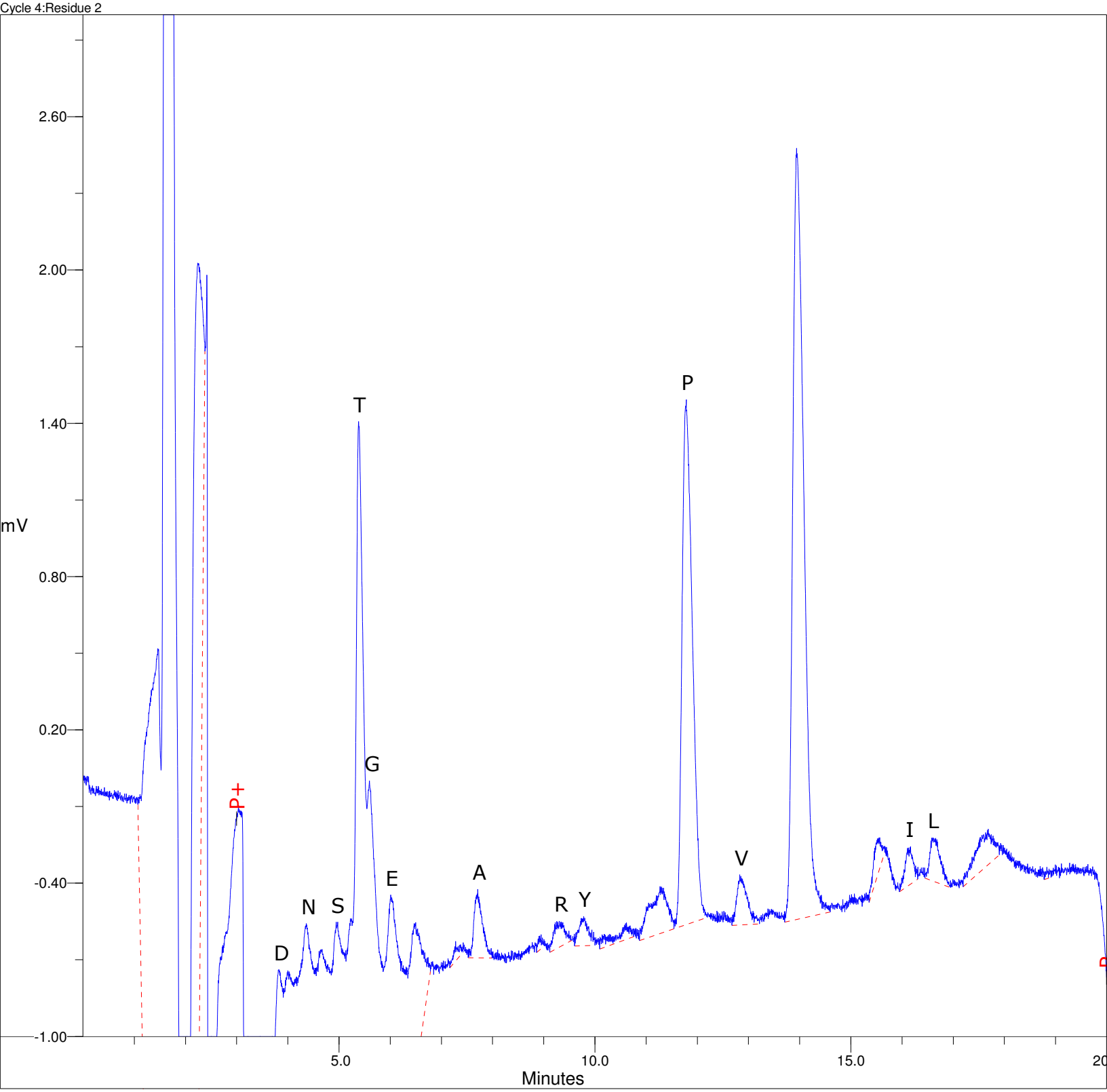
Cycle 3:Residue 1



PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	1.47		5.171	
	1.65		14.794	
	2.24		3.834	
	3.08		5.992	
D	3.82	3.78	4.200	13.948
	4.02		3.957	
N	4.34	4.33	3.691	19.625
	4.65		3.131	
S	4.95	4.92	3.121	25.788
T	5.39	5.33	2.288	14.077
G	5.60	5.58	3.183	25.460
E	6.01	5.99	1.284	7.918
	6.58		0.236	
H	7.35	7.42	0.059	0.558
A	7.68	7.65	0.317	2.467

PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	8.89		0.095	
	9.26		0.178	
Y	9.74	9.75	0.136	1.079
	10.27		0.046	
	10.58		0.062	
	11.28		0.102	
P	11.75	11.73	0.090	0.819
M	12.46	12.49	0.146	1.277
V	12.81	12.80	0.157	1.349
	13.90		3.110	
	15.47		0.275	
I	16.10	16.09	0.094	0.989
K	16.34	16.38	0.111	0.772
L	16.55	16.55	0.200	1.439
	17.60		0.110	

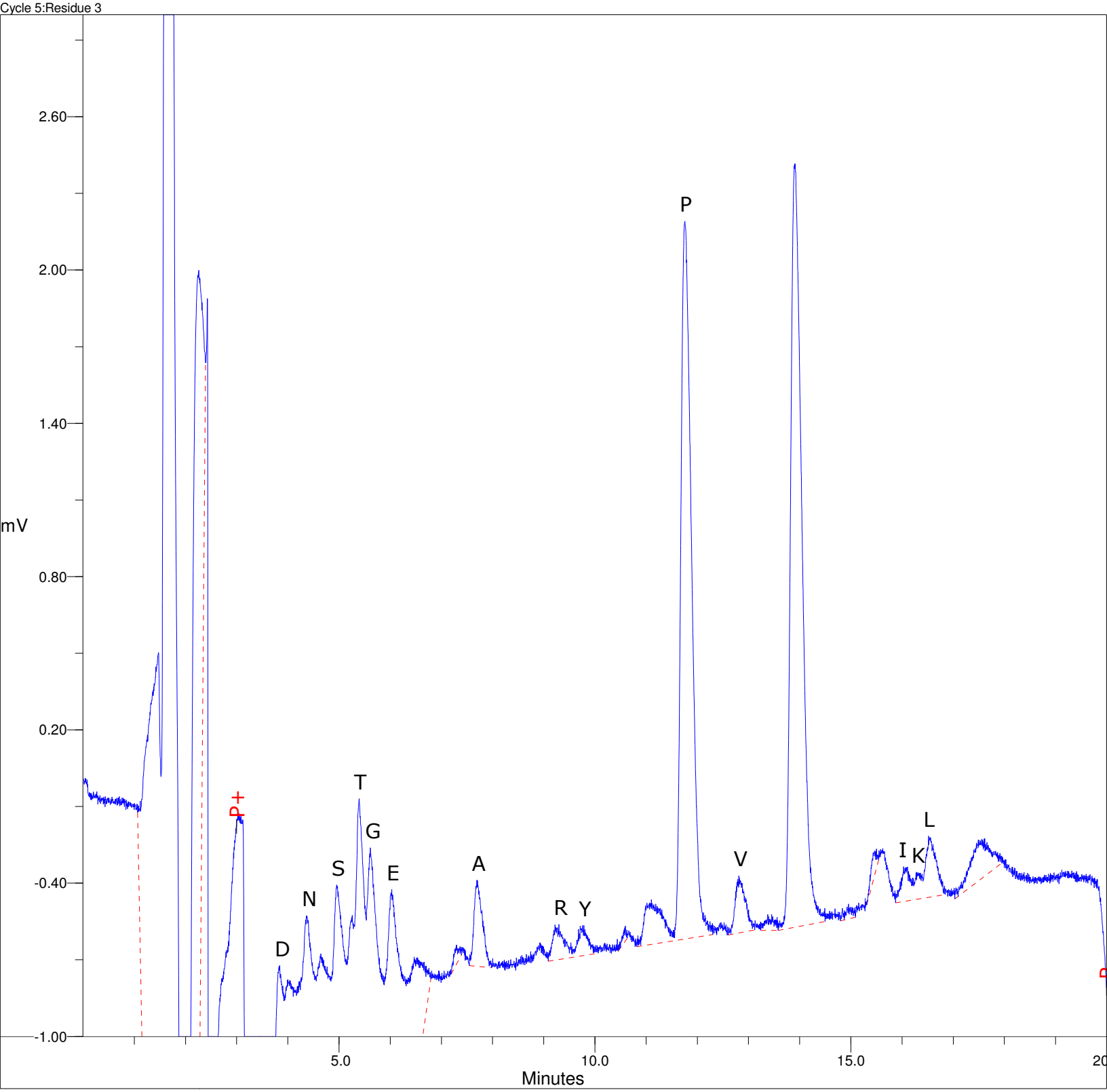
Cycle 4:Residue 2



PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	1.46		4.752	
	1.65		14.576	
	2.24		3.809	
	3.04		5.924	
D	3.83	3.78	4.172	13.854
	4.00		3.920	
N	4.36	4.33	3.600	19.144
	4.67		3.060	
S	4.95	4.92	2.779	22.970
T	5.38	5.33	4.123	25.358
G	5.59	5.58	2.423	19.380
E	6.01	5.99	1.390	8.571
	6.50		0.597	
	7.28		0.077	
A	7.71	7.65	0.270	2.100

PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	8.95		0.057	
R	9.30	9.38	0.103	1.180
Y	9.78	9.75	0.116	0.923
	10.18		0.052	
	10.62		0.067	
	11.27		0.186	
P	11.78	11.73	2.056	18.717
V	12.83	12.80	0.197	1.695
	13.94		3.020	
	15.56		0.122	
I	16.17	16.09	0.147	1.554
L	16.60	16.55	0.171	1.228
	17.68		0.140	
	18.83		0.046	

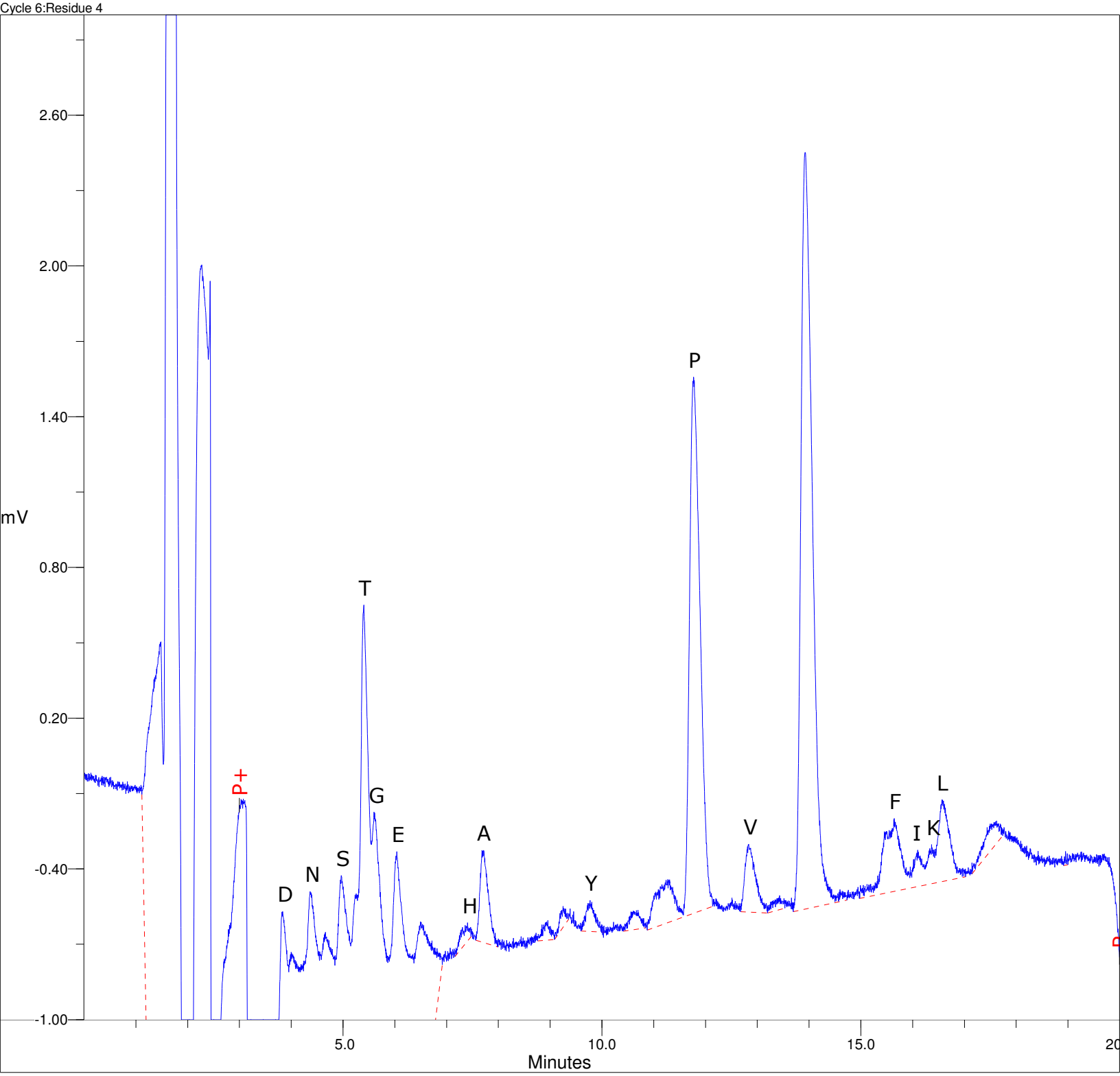
Cycle 5:Residue 3



PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	1.47		4.927	
	1.65		14.525	
	2.26		3.519	
	3.04		5.963	
D	3.84	3.78	4.239	14.078
	4.00		3.955	
N	4.37	4.33	3.680	19.564
	4.65		3.141	
S	4.95	4.92	2.974	24.580
T	5.40	5.33	2.691	16.549
G	5.61	5.58	2.188	17.506
E	6.03	5.99	1.438	8.873
	6.47		0.553	
	7.29		0.078	
A	7.69	7.65	0.338	2.634

PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
R	9.30	9.38	0.138	1.590
Y	9.78	9.75	0.118	0.939
	10.60		0.059	
	11.10		0.175	
P	11.75	11.73	2.810	25.575
V	12.81	12.80	0.224	1.934
	13.45		0.063	
	13.90		2.989	
	14.83		0.047	
	14.95		0.058	
	15.50		0.076	
I	16.03	16.09	0.136	1.439
K	16.28	16.38	0.107	0.743
L	16.52	16.55	0.240	1.724
	17.62		0.147	

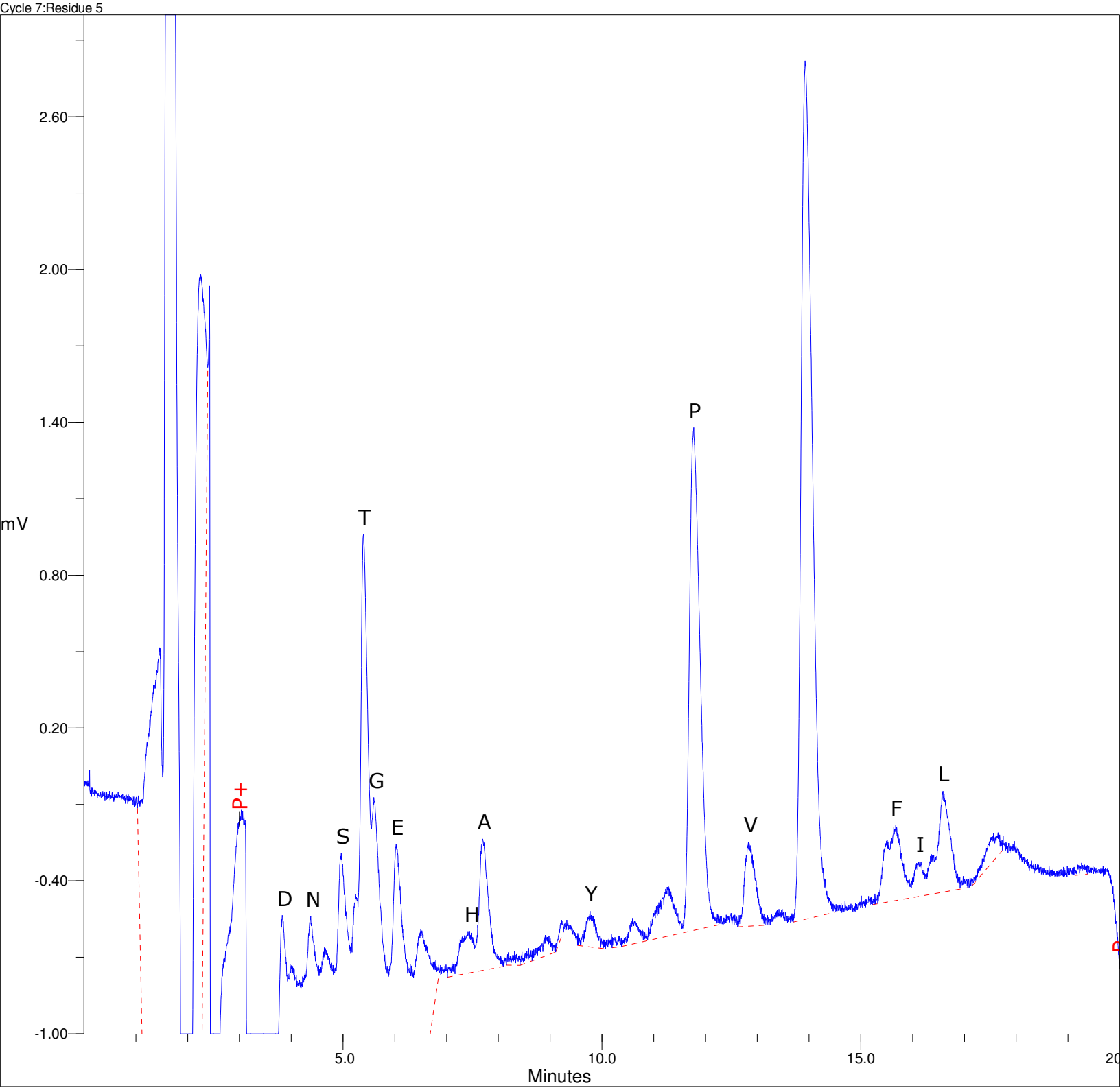
Cycle 6:Residue 4



PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	1.48		4.667	
	1.66		14.287	
	2.27		10.576	
	3.08		7.105	
D	3.83	3.78	5.383	17.877
	4.00		4.942	
N	4.36	4.33	4.574	24.321
	4.66		3.913	
S	4.97	4.92	3.625	29.961
T	5.40	5.33	3.975	24.448
G	5.60	5.58	2.816	22.523
E	6.04	5.99	1.925	11.877
	6.50		0.879	
H	7.40	7.42	0.073	0.696
A	7.68	7.65	0.366	2.850
	8.96		0.080	
	9.25		0.084	

PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
Y	9.77	9.75	0.125	0.995
	10.70		0.077	
	11.27		0.173	
P	11.77	11.73	2.138	19.457
V	12.83	12.80	0.271	2.332
	13.43		0.056	
	13.93		3.013	
	14.60		0.048	
	15.19		0.050	
	15.50		0.247	
F	15.65	15.69	0.291	2.277
I	16.10	16.09	0.146	1.542
K	16.37	16.38	0.157	1.089
L	16.57	16.55	0.326	2.344
	17.62		0.091	
	18.95		0.056	

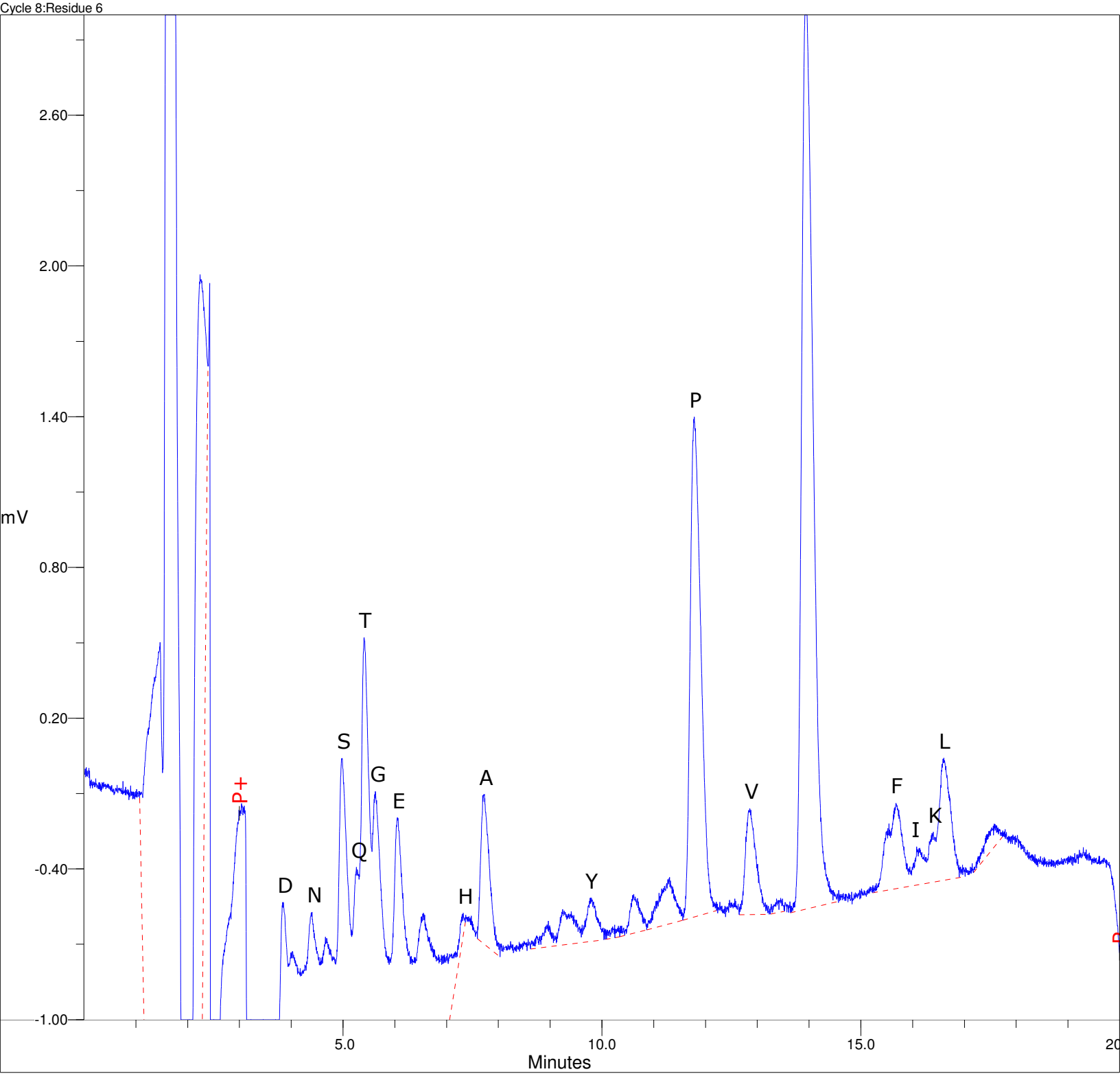
Cycle 7:Residue 5



PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	1.46		4.995	
	1.65		14.505	
	2.25		3.637	
	3.04		5.950	
D	3.83	3.78	4.441	14.748
	3.98		4.034	
N	4.38	4.33	3.678	19.558
	4.64		3.187	
S	4.97	4.92	3.100	25.617
T	5.40	5.33	3.752	23.080
G	5.60	5.58	2.442	19.530
E	6.03	5.99	1.664	10.264
	6.50		0.666	
H	7.44	7.42	0.164	1.562
A	7.70	7.65	0.517	4.031
	8.31		0.055	

PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	8.58		0.056	
	8.90		0.084	
	9.22		0.065	
Y	9.77	9.75	0.140	1.116
	10.32		0.043	
	10.59		0.102	
	11.28		0.197	
P	11.77	11.73	1.975	17.977
V	12.84	12.80	0.333	2.872
	13.92		3.370	
	15.51		0.244	
F	15.68	15.69	0.295	2.310
I	16.16	16.09	0.135	1.424
L	16.59	16.55	0.397	2.852
	17.65		0.091	
	19.26		0.046	

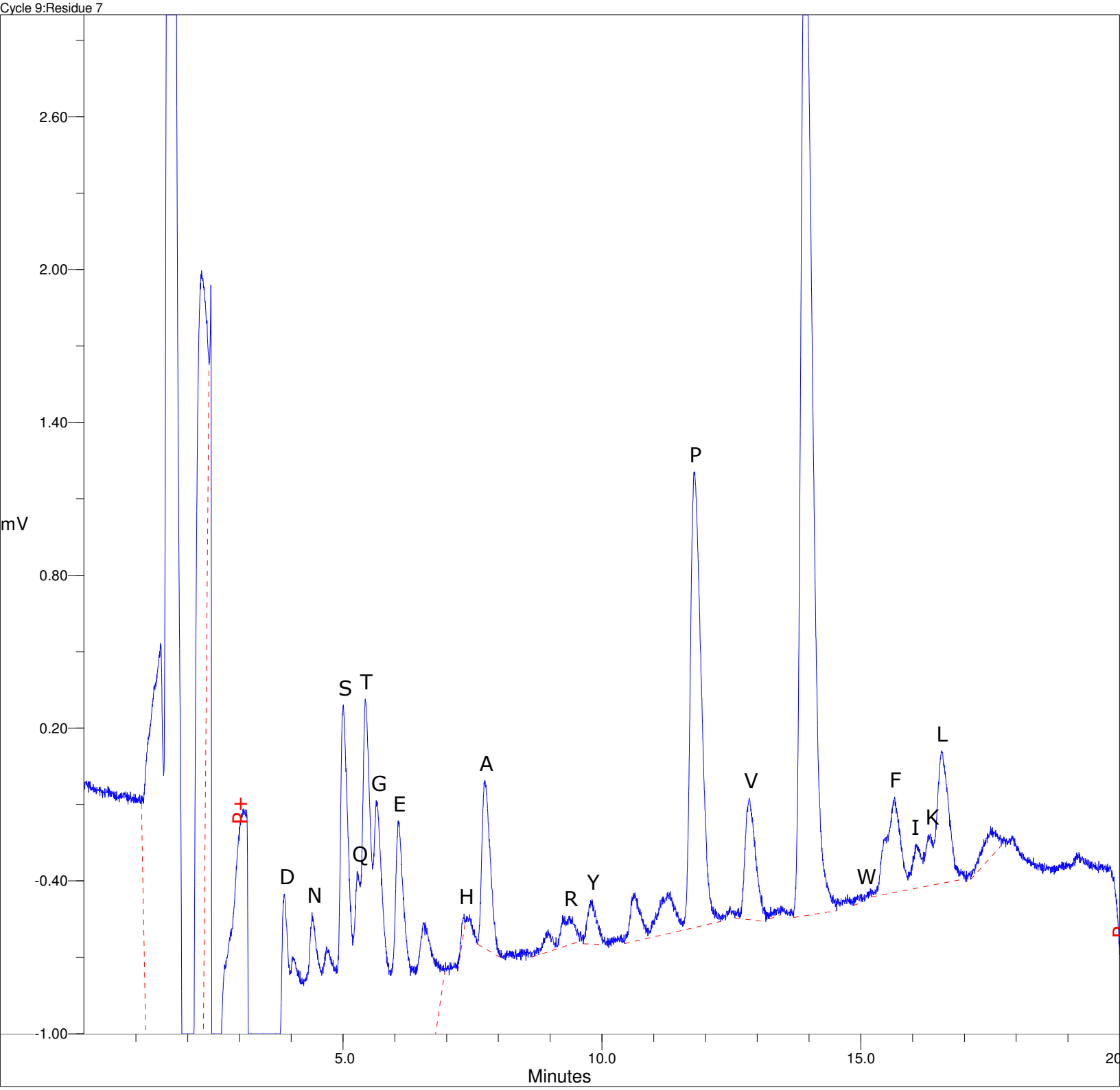
Cycle 8:Residue 6



PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	1.47		4.817	
	1.65		14.375	
	2.24		4.036	
	3.04		5.981	
D	3.83	3.78	4.578	15.204
	4.03		4.140	
N	4.40	4.33	3.817	20.297
	4.66		3.384	
S	4.98	4.92	3.685	30.453
Q	5.26	5.20	2.902	18.540
T	5.41	5.33	3.626	22.300
G	5.63	5.58	2.735	21.880
E	6.05	5.99	2.094	12.914
	6.55		1.070	
H	7.32	7.42	0.090	0.858
A	7.73	7.65	0.599	4.671
	8.75		0.049	

PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	8.97		0.103	
	9.24		0.142	
Y	9.78	9.75	0.174	1.385
	10.25		0.052	
	10.32		0.044	
	10.62		0.155	
	11.30		0.186	
P	11.78	11.73	1.993	18.137
V	12.85	12.80	0.423	3.647
	13.45		0.055	
	13.93		3.687	
F	15.68	15.69	0.342	2.674
I	16.08	16.09	0.154	1.621
K	16.40	16.38	0.202	1.402
L	16.60	16.55	0.487	3.499
	17.57		0.092	

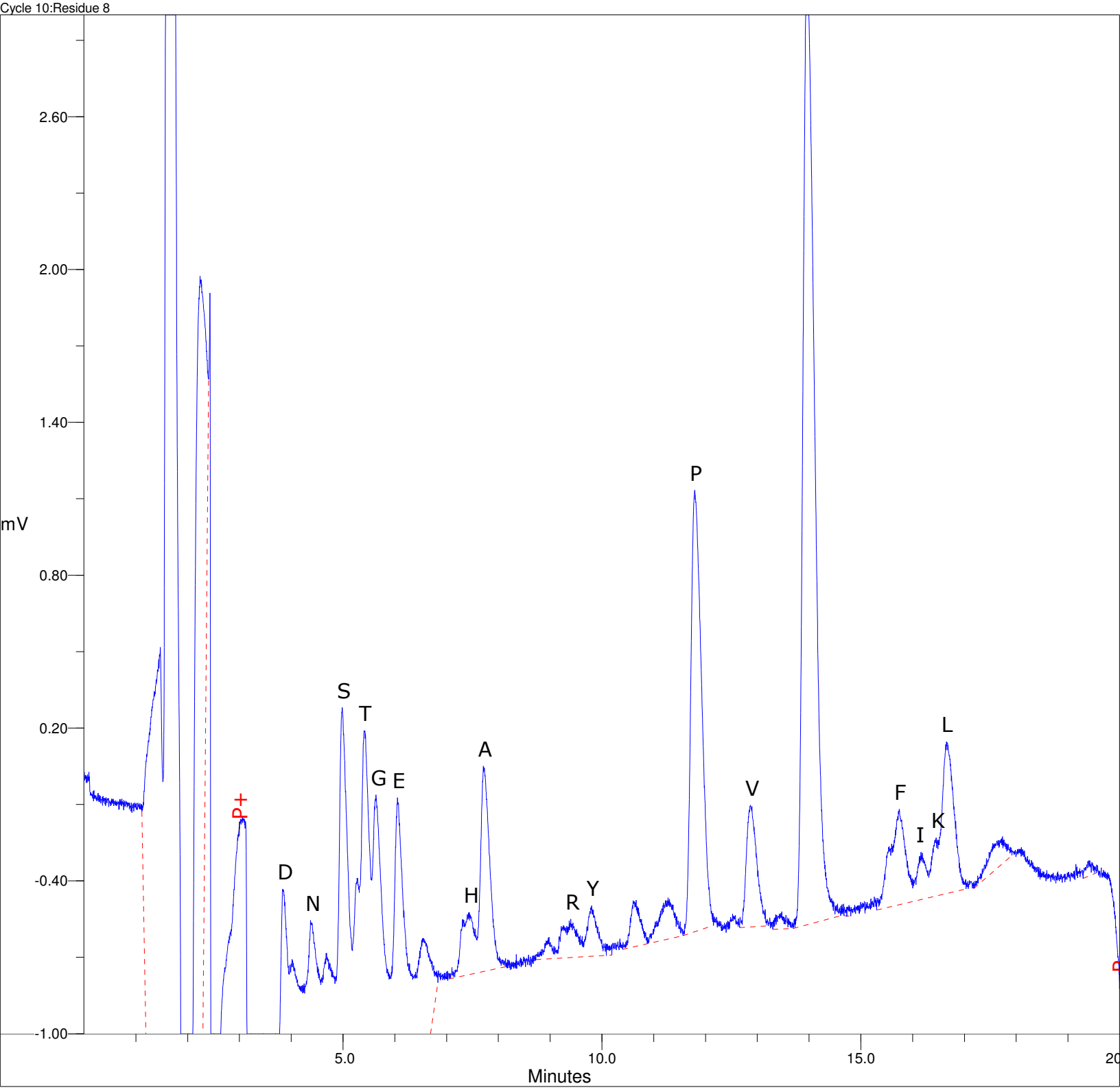
Cycle 9:Residue 7



PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	1.47		4.596	
	1.67		14.348	
	2.27		3.740	
	3.08		5.996	
D	3.87	3.78	4.573	15.187
	4.03		4.107	
N	4.40	4.33	3.765	20.017
	4.68		3.240	
S	5.01	4.92	3.747	30.963
Q	5.28	5.20	2.725	17.409
T	5.43	5.33	3.185	19.590
G	5.64	5.58	2.495	19.955
E	6.06	5.99	1.835	11.320
	6.59		0.713	
H	7.34	7.42	0.064	0.614
A	7.73	7.65	0.659	5.138

PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	8.96		0.089	
	9.24		0.124	
R	9.36	9.38	0.115	1.327
Y	9.81	9.75	0.176	1.399
	10.64		0.191	
	11.32		0.164	
P	11.78	11.73	1.789	16.285
V	12.85	12.80	0.478	4.119
	13.26		0.047	
	13.92		4.095	
W	15.02	15.06	0.034	0.269
F	15.65	15.69	0.374	2.930
I	16.07	16.09	0.174	1.832
K	16.35	16.38	0.203	1.412
L	16.56	16.55	0.521	3.738
	17.51		0.121	

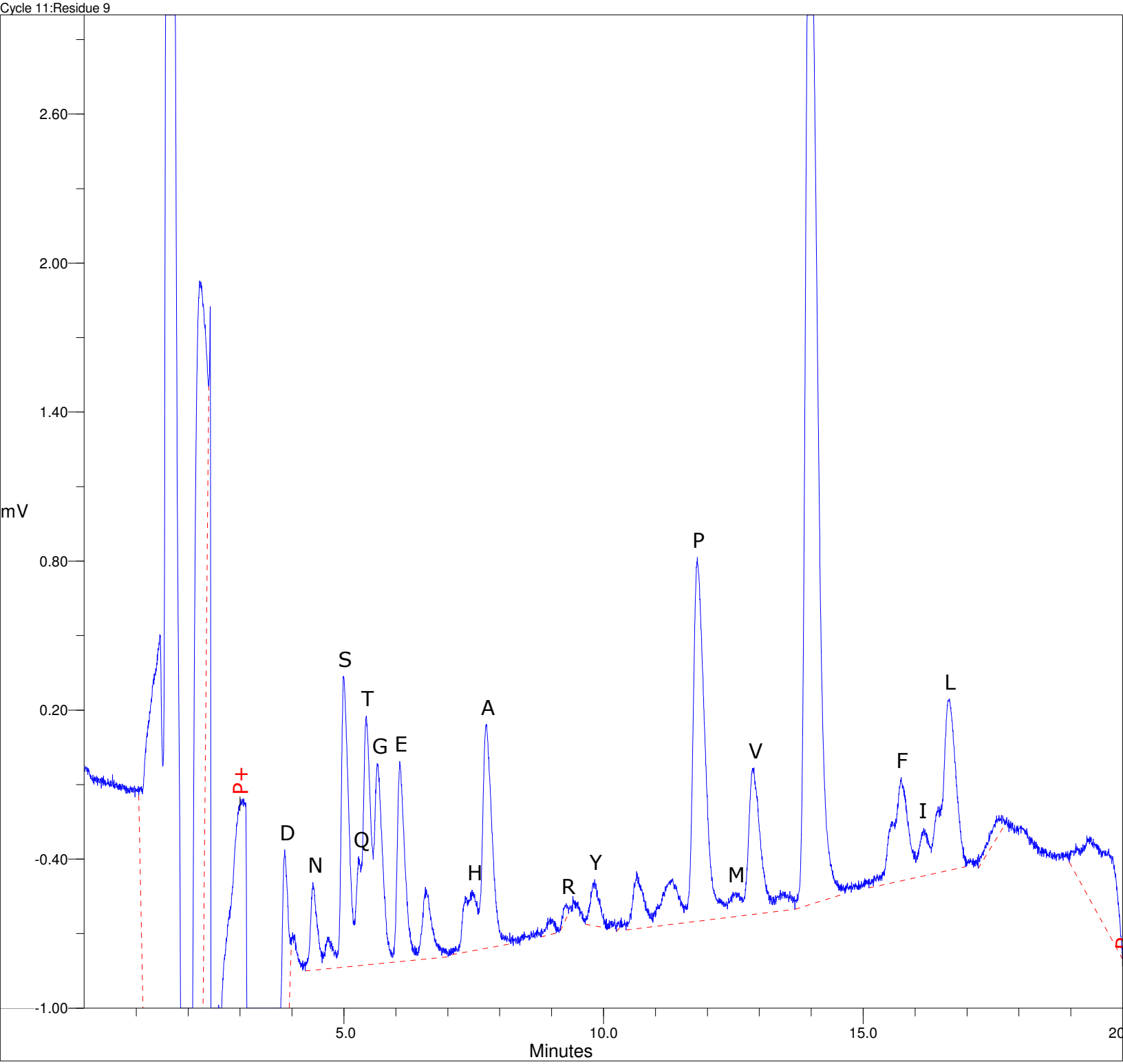
Cycle 10:Residue 8



PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	1.47		4.626	
	1.65		14.143	
	2.24		4.282	
	3.06		5.974	
D	3.83	3.78	4.600	15.275
	4.02		4.065	
N	4.37	4.33	3.718	19.770
	4.68		3.149	
S	4.98	4.92	3.690	30.494
T	5.41	5.33	3.000	18.451
G	5.64	5.58	2.424	19.392
E	6.05	5.99	1.831	11.296
	6.55		0.581	
H	7.43	7.42	0.244	2.324
A	7.71	7.65	0.807	6.291
	8.98		0.085	

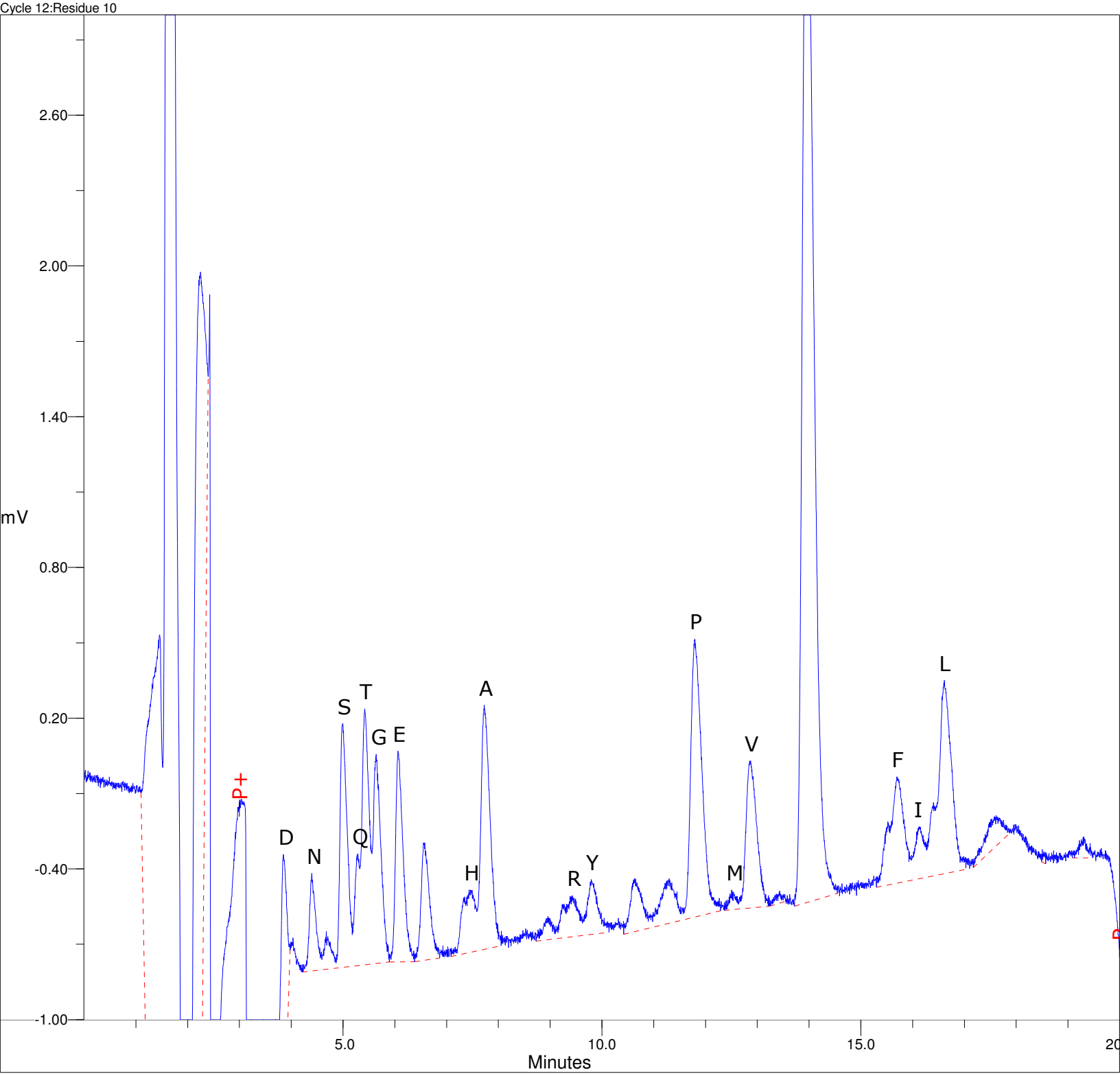
PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	9.26		0.129	
R	9.40	9.38	0.151	1.738
Y	9.80	9.75	0.200	1.586
	10.66		0.184	
	11.28		0.163	
P	11.80	11.73	1.734	15.787
V	12.87	12.80	0.478	4.120
	13.48		0.068	
	13.96		3.767	
	14.77		0.042	
F	15.74	15.69	0.376	2.943
I	16.16	16.09	0.189	1.990
K	16.45	16.38	0.230	1.598
L	16.65	16.55	0.597	4.284
	17.73		0.107	
	19.40		0.061	

Cycle 11:Residue 9



PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT	PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	1.45		4.925		Y	9.83	9.75	0.188	1.496
	1.64		14.188			10.30		0.030	
	2.22		4.484			10.64		0.232	
	3.07		5.826			11.33		0.184	
D	3.86	3.78	1.189	3.948	P	11.81	11.73	1.466	13.345
N	4.40	4.33	0.355	1.887	M	12.49	12.49	0.097	0.846
	4.71		0.128		V	12.90	12.80	0.591	5.088
S	4.99	4.92	1.176	9.716		13.40		0.081	
Q	5.28	5.20	0.441	2.817		13.98		4.404	
T	5.43	5.33	1.005	6.183		15.54		0.250	
G	5.64	5.58	0.809	6.473	F	15.73	15.69	0.417	3.267
E	6.08	5.99	0.807	4.980	I	16.17	16.09	0.193	2.031
	6.57		0.292			16.45		0.269	
	7.34		0.226		L	16.66	16.55	0.690	4.958
H	7.47	7.42	0.247	2.347		17.66		0.062	
A	7.74	7.65	0.902	7.033		19.10		0.124	
	8.97		0.072			19.31		0.235	
R	9.28	9.38	0.061	0.699					

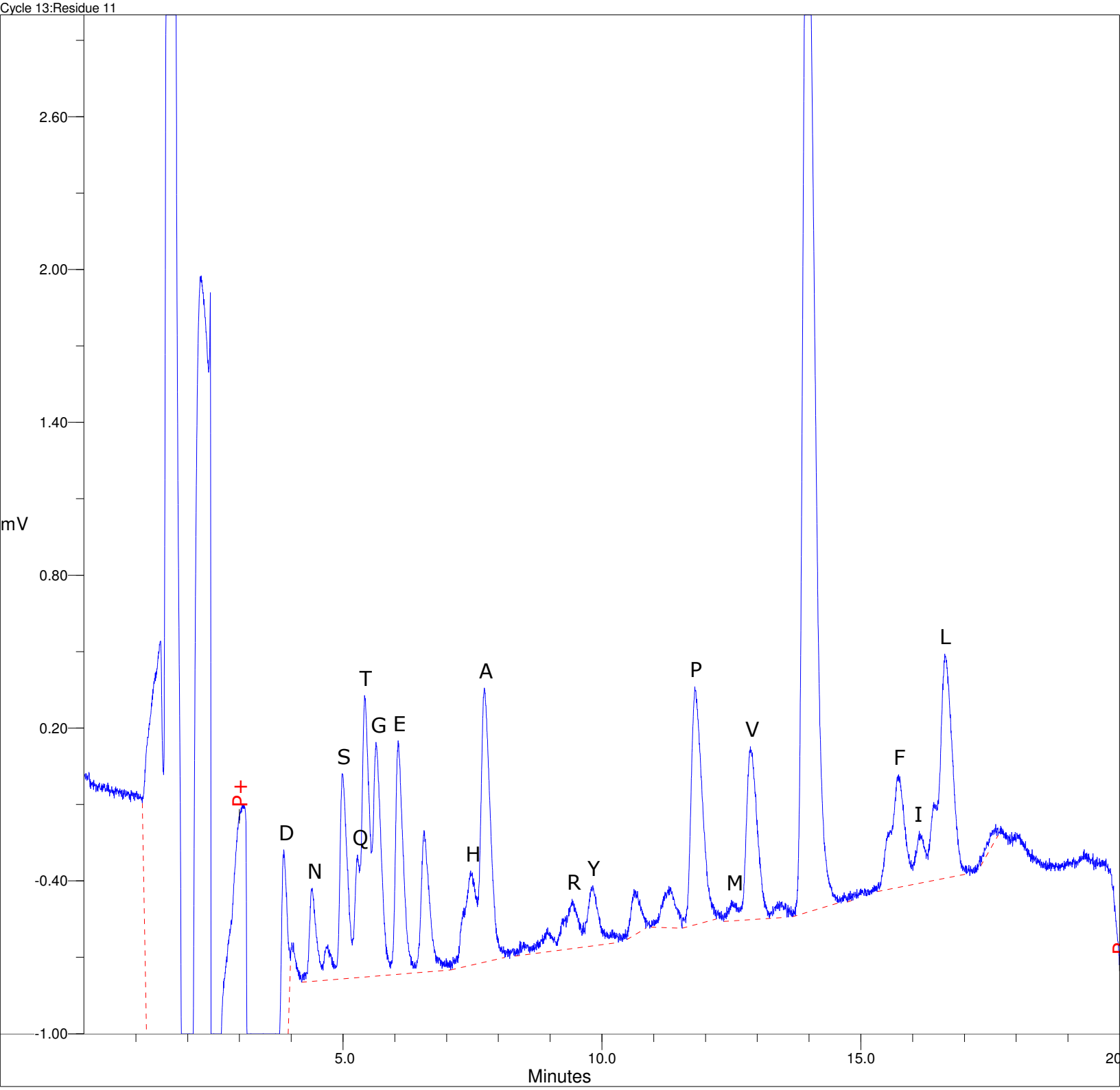
Cycle 12:Residue 10



PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	1.46		4.608	
	1.65		14.154	
	2.25		3.887	
	3.04		5.855	
D	3.85	3.78	1.201	3.987
N	4.40	4.33	0.390	2.074
	4.68		0.152	
S	4.99	4.92	0.971	8.025
Q	5.28	5.20	0.448	2.862
T	5.42	5.33	1.019	6.270
G	5.64	5.58	0.834	6.669
E	6.06	5.99	0.841	5.186
	6.58		0.468	
	7.35		0.226	
H	7.44	7.42	0.252	2.396
A	7.72	7.65	0.973	7.580
	8.96		0.093	

PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	9.25		0.136	
R	9.43	9.38	0.165	1.897
Y	9.79	9.75	0.220	1.750
	10.63		0.212	
	11.30		0.179	
P	11.80	11.73	1.107	10.073
M	12.50	12.49	0.083	0.721
V	12.85	12.80	0.587	5.053
	13.43		0.049	
	13.95		4.409	
F	15.70	15.69	0.424	3.319
I	16.13	16.09	0.208	2.193
L	16.61	16.55	0.771	5.536
	17.61		0.123	
	18.50		0.051	
	19.30		0.084	

Cycle 13:Residue 11



PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	1.48		4.569	
	1.66		14.127	
	2.26		10.202	
	3.08		5.899	
D	3.85	3.78	1.262	4.192
N	4.41	4.33	0.364	1.936
	4.71		0.132	
S	4.98	4.92	0.795	6.572
Q	5.28	5.20	0.466	2.980
T	5.42	5.33	1.089	6.699
G	5.64	5.58	0.899	7.189
E	6.06	5.99	0.896	5.525
	6.57		0.546	
H	7.46	7.42	0.373	3.544
A	7.72	7.65	1.079	8.407
	8.42		0.057	

PEAK ID	R.TIME (mins)	C.TIME (mins)	HEIGHT (mV)	PMOL HT
	8.95		0.095	
R	9.43	9.38	0.194	2.235
Y	9.82	9.75	0.238	1.888
	10.21		0.055	
	10.62		0.181	
	11.29		0.165	
P	11.80	11.73	0.935	8.511
M	12.50	12.49	0.082	0.718
V	12.87	12.80	0.680	5.861
	13.43		0.067	
	13.97		4.414	
	14.80		0.040	
F	15.73	15.69	0.444	3.473
I	16.13	16.09	0.206	2.168
L	16.62	16.55	0.882	6.334
	17.60		0.061	