Structural and Functional Studies on Secreted Glycoside Hydrolases Produced by *Clostridium perfringens*

by

Elizabeth Ficko-Blean
BSc Hon, University of British Columbia, 2003

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Biochemistry and Microbiology

Supervisory Committee

Structural and Functional Studies on Secreted Glycoside Hydrolases Produced by *Clostridium perfringens*

by

Elizabeth Ficko-Blean
BSc Hon, University of British Columbia, 2003

Supervisory Committee

Dr. Alisdair B. Boraston, Department of Biochemistry and Microbiology
Supervisor

Dr. Stephen V. Evans (Department of Biochemistry and Microbiology)
Departmental Member

Dr. Paul J. Romaniuk (Department of Biochemistry and Microbiology)
Departmental Member

Dr. Cornelia Bohne (Department of Chemistry)
Outside Member

Abstract

Supervisory Committee
Dr. Alisdair B. Boraston, Department of Biochemistry and Microbiology
Supervisor
Dr. Stephen V. Evans (Department of Biochemistry and Microbiology)
Departmental Member
Dr. Paul J. Romaniuk (Department of Biochemistry and Microbiology)
Departmental Member
Dr. Cornelia Bohne (Department of Chemistry)
Outside Member

Clostridium perfringens is a gram positive spore forming anaerobe and a causative agent of gas gangrene, necrotic enteritis (pig-bel) and food poisoning in humans and other animals. This organism secretes a battery of exotoxins during the course of infection as well as a variety of virulence factors which may help to potentiate the activities of the toxins. Among these virulence factors is the μ-toxin, a family 84 glycoside hydrolase which acts to degrade hyaluronan, a component of human connective tissue. *C. perfringens* has 53 open reading frames encoding glycoside hydrolases. About half of these glycoside hydrolases are predicted to be secreted. Among these are CpGH84C, a paralogue of the μ-toxin, and CpGH89. CpGH89 shares sequence similarity to the human α-*N*-acetylglucosaminidase, NAGLU, in which mutations can cause a devastating genetic disease called mucopolysaccharidosis IIIB.

One striking feature of the secreted glycoside hydrolase enzymes of *C. perfringens* is their modularity, with modules predicted to be dedicated to catalysis, carbohydrate-binding, protein-protein interactions and cell wall attachment. The extent of the modularity is remarkable, with some enzymes containing up to eight ancillary modules. In order to help understand the role of carbohydrate-active enzymes produced by bacterial pathogens, this thesis will focus on the structure and function of the modular extracellular glycoside hydrolase enzymes secreted by the disease causing bacterium, *C. perfringens*. These structure function studies examine two family 32 CBMs (carbohydrate-binding modules), one from the μ-toxin and the other from CpGH84C. As well we examine the complete structure of CpGH84C in order to help further our understanding of the structure of carbohydrate-active enzymes as a whole. Finally, the

catalytic module of CpGH89 is characterized and its relationship to the human NAGLU enzyme is discussed.

Table of Contents

List of Tables

List of Figures

Acknowledgments

First and foremost, I would like to thank my mentor and supervisor Dr. Alisdair Boraston, without whom none of this research would have been possible. Thank you!

Thank you also to my committee, Dr. Stephen Evans, who has given me much guidance in regards to X-ray Crystallography, Dr. Paul Romaniuk, who has given me much guidance in life and Dr. Cornelia Bohne and Dr. David Zechel.

Big thanks to my Mom and Dad, Devon Blean and Don Ficko. My mother gave me my love of learning from an early age. My father showed me the true meaning of work ethic, even when the chips are down. And of course thank you Charlotte Ficko-Blean, my favorite/only sister and best friend. We have shared things that only we can understand. And you have provided us little Izzy, he's a light in my life.

Thank you to my Grandparents for their support: Grandpa Hal, Grandma Jean, Grandpa Gordon, Grandma Helen and Grandpa Gus.

Thank you also to my great friends throughout my life. Dawn Slater, an unexpected bonus when I moved to Victoria. Dr. Alicia Lammerts van Bueren, you rock! Carmeen de Wit, with whom I share many memories. Tracy Kendrick, we like to create our own drama. Peter Duong, Hasan Keskic and Jessica and Stephanie Irwin, your friendship has meant a great deal to me.

Many thanks to the 10 am coffee crew: Katie Gregg, Melanie Higgins, Ron Finn, Dr. Marty Boulanger, Dr. Wade Abbott and all the others who have passed through.

Finally, I would like to acknowledge my funding sources: the Michaeal Smith Foundation for Health Research, the Canadian Institutes of Health Research and the National Sciences and Engineering Council of Canada.

Dedication

To my Grandma Edith Wieland, who at the time of writing is quite sick. Your never-ending love and support have been invaluable.

Chapter 1: General Introduction

1.1 Glycoside Hydrolases

1.1.1 Glycoside Hydrolases are Modular

Study of starch, chitin and plant cell wall degrading glycoside hydrolases produced by microorganisms has contributed much to our current knowledge on the structure, function and mechanism of these enzymes. Investigation has revealed that they are often modular, containing the prerequisite catalytic module as well as ancillary modules, like the common carbohydrate binding modules or CBMs. Other modules, such as those involved in protein-protein interactions, have been characterized in the multi-enzyme cellulose degrading complexes of cellulolytic Clostridia.

1.1.2 Glycoside Hydrolase Activity

The glycosidic bond, the bond linking two carbohydrates together or a carbohydrate to another compound, is extraordinarily stable. The β-1,4-glucosidic bond in cellulose has an estimated half life of almost 5 million years (Wolfenden *et al*. 1998). Glycoside hydrolases act to hydrolyze glycosidic bonds, increasing bond cleavage by up to ~ 10^{17} fold (Wolfenden *et al*. 1998). Other carbohydrate degrading enzymes are the polysaccharide lyases, which cleave the glycosidic bond *via* a β-elimination mechanism, and the carbohydrate esterases, which de-O-acetylate or de-N-acetylate substituted saccharides. Glycosyl transferases provide the converse reaction and catalyze the transfer of sugar moieties from activated donor molecules to specific acceptor molecules to form a glycosidic bond.

Potential target substrates for glycoside hydrolases abound. These substrates include but are not limited to, plant cell wall polysaccharides, such as cellulose, hemicellulose, β-1,3-glucans and lignin. Other target substrates may include chitin and starch and glycogen. Glycoside hydrolases can also act on the diverse glycan components found in the human body such as those described in section 1.2.2.

1.1.3 Glycoside Hydrolase Classification

Glycoside hydrolases are classified in 114 families based on amino acid sequence identity (Cantarel *et al*. 2008), though this number is constantly increasing as more families are identified. Glycoside hydrolases are also broadly grouped into two categories based on the stereochemistry of the anomeric carbon after hydrolysis (discussed in Section 1.1.4), either retaining or inverting. Because 3D folds can be conserved between different families, though the amino acid sequences diverge, glycoside hydrolases can be further divided into 14 structure based clans. Structural folds found in the glycoside hydrolases clans include: $(\beta/\alpha)_8$, β-jelly roll, 6-fold β-propeller, 5-fold β-propeller, $(\alpha/\alpha)_6$, α + β and the β-helix (Figure 1).

1.1.4 Glycoside Hydrolase Mechanisms

Glycoside hydrolases act by adding water across the glycosidic bond. Asp or Glu are most commonly found as the catalytically active residues, though there are exceptions (Davies *et al*. 2005; Lodge *et al*. 2003; Watts *et al*. 2003; Yip *et al*. 2004). There are 3 fundamental mechanisms by which these enzymes proceed (Figure 2) and these are described below.

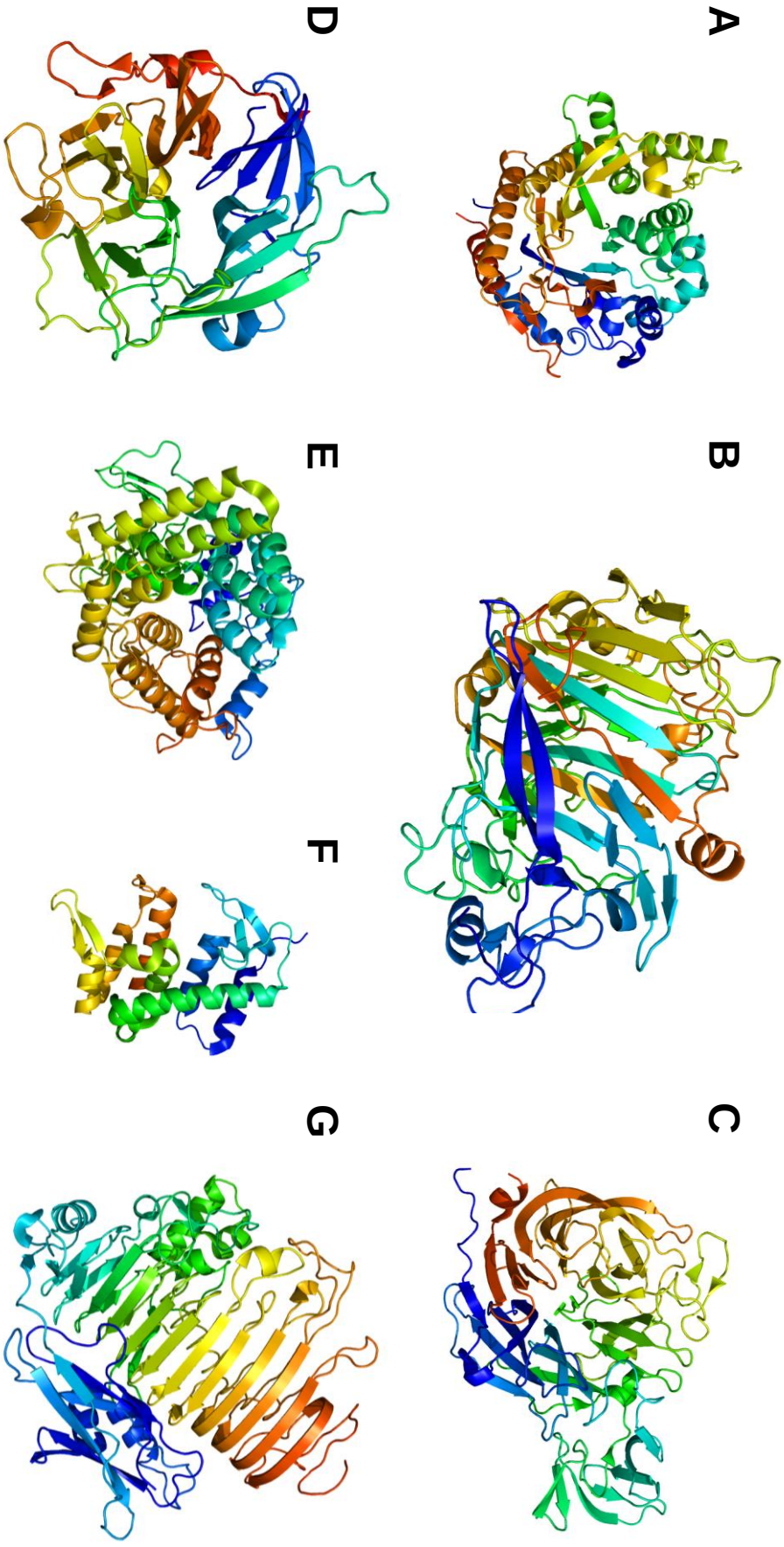

A
B
C
D
E
F
G

Figure 1 (Previous page). Clan fold families of the glycoside hydrolases. (A) $(\beta/\alpha)_8$. Family 1 β-glycosylceraminidase/β-glucosidase from *Homo sapiens* (PDB code 2ZOX) (B) β-jelly roll. Family 7 cellobiohydrolase I/chitosanase from *Hypocrea jecorina* L27 (PDB code 2CEL) (C) 6-fold β-propeller. Family 33 sialidase from *C. perfringens* strain 13 (PDB code 2VK6) (D) 5-fold β-propeller. Family 43 arabinanase from *Cellvibrio japonicus* (PDB code 1GYE) (E) $(\alpha/\alpha)_6$. Family 37 trehalase from *E. coli* K12 (PDB code 2JGO) (F) α + β. Family 24 lysozyme from Enterobacteria phage P1 (PDB code 1XJT) G. β-helix. Family 28 *exo*-polygalacturonosidase from *Yersinia enterocolitica* (PDB code 2UVE). Images were created using PyMol.

Retaining Mechanisms

Glycoside hydrolases which use a retaining mechanism generally have the nucleophile and catalytic acid/base residues ~6 Å apart (Zechel and Withers 2000). This reaction proceeds *via* a double displacement mechanism (Figure 2A). The first step requires nucleophilic attack at the anomeric centre with simultaneous protonation of the glycosidic oxygen by the general acid. This occurs through a transition state that has oxocarbenium ion character and results in the formation of a glycosyl-enzyme intermediate. Hydrolysis of the intermediate occurs as the general base deprotonates an incoming water molecule, which attacks the anomeric centre causing hydrolysis of the glycosyl-enzyme intermediate. This step also proceeds through a transition state with oxocarbenium ion character and results in retention of stereochemistry at the anomeric carbon (Zechel and Withers 2000).

When the substrate plays a role in catalysis this is referred to as anchimeric assistance or substrate assisted catalysis, which is also a retaining mechanism (Macauley *et al*. 2005). The enzyme positions the C2 acetamido group of the substrate, an N-acetylated sugar, for nucleophilic attacks on the anomeric carbon. The glycosidic oxygen of the leaving group is protonated by the catalytic acid/base. Unlike above, there is not the formation of an enzyme-glycosyl intermediate, rather, a bicyclic oxazoline intermediate is formed. This intermediate is hydrolyzed by an incoming water molecule resulting in overall retention of stereochemistry (Mark *et al*. 2001; Vocadlo and Withers 2005; Zechel and Withers 2000).

Inverting Mechanism

Inverting glycoside hydrolases require a larger distance in order to accommodate the direct attack of a water molecule and the substrate, thus their general base and general acid residues are separated by ~10 Å (Zechel and Withers 2000). This reaction proceeds *via* a single displacement mechanism (Figure 2B). Water is deprotonated by the general base and attacks the anomeric center while the general acid concomitantly protonates the leaving group. The

inverting mechanism proceeds *via* an oxacarbenium ion-like transition state and results in overall inversion of stereochemistry at the anomeric carbon (Zechel and Withers 2000).

1.1.5 Carbohydrate-Binding Modules

Definition

CBMs are defined as independently folding modules, found in carbohydrate-active enzymes, which function to bind carbohydrate, though there are some exceptions to this. CBMs were originally identified in plant cell wall degrading bacteria and described as cellulose-binding domains (CBDs) as examples bound solely to cellulose. As more diverse carbohydrate ligands were identified for new CBMs, the term CBDs was changed to CBMs. The first identification of a cellulose binding domain occurred in 1986 with the proteolytic degradation of a cellulase produced by the fungus *Trichoderma reesei* and resulted from the observation that one domain retained the cellulase activity and the other had cellulose-binding

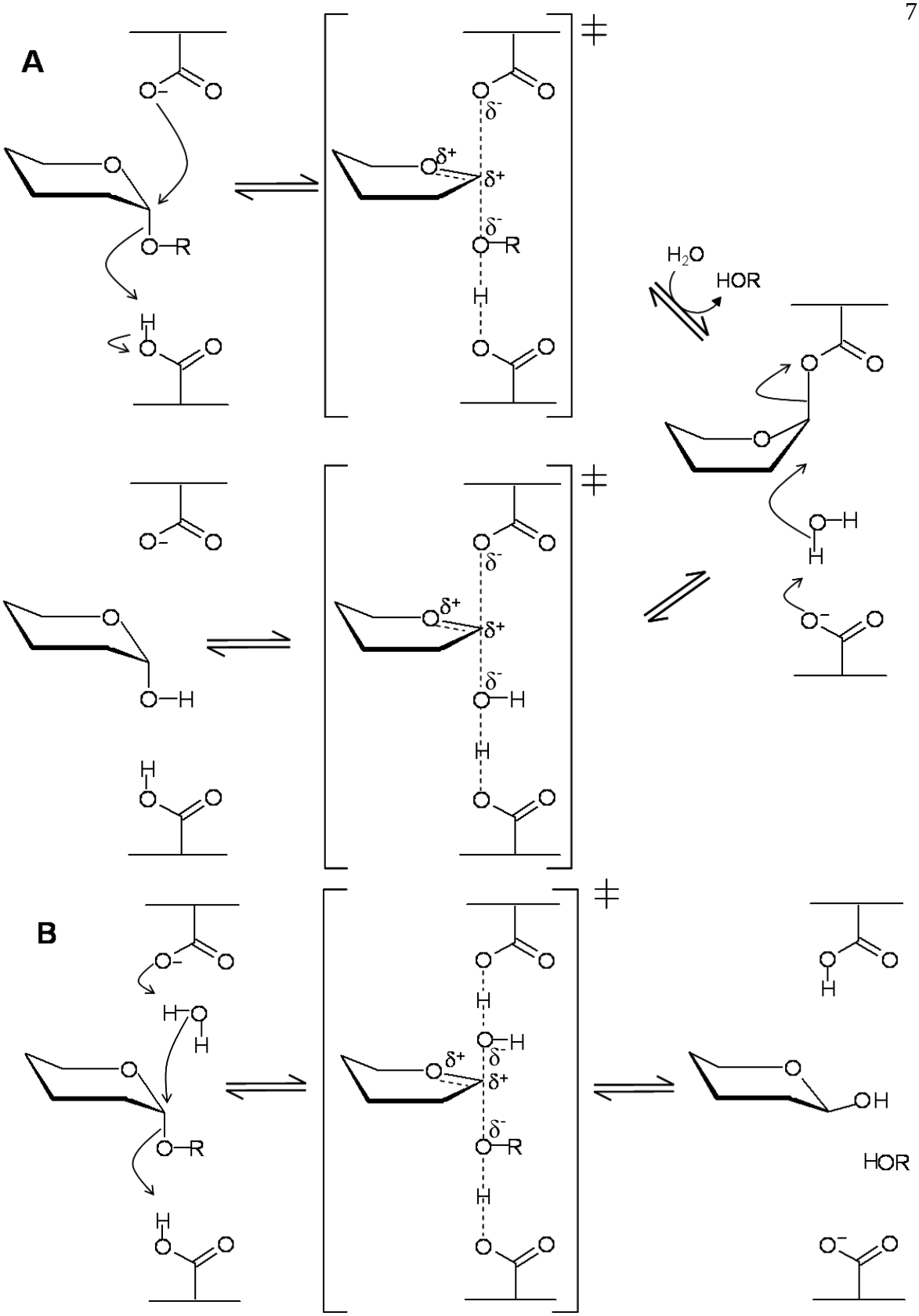


Figure 2. Mechanisms of glycoside hydrolase hydrolysis. (A) Retaining mechanism. (B) Inverting Mechanism.

function (Tomme *et al*. 1988; Vantilbeurgh *et al*. 1986). In general, CBMs act by binding carbohydrate ligands thereby targeting the enzyme to substrate (Araki *et al*. 2004; Hong *et al*. 2002), though there are some exceptions, for example, in the non-catalytic scaffoldin structure found in cellulolytic Clostridia (Bayer *et al*. 1998). This definition helps separate these modules from other carbohydrate-binding proteins, such as lectins, antibodies, and proteins involved in transport.

Sequence Classification

Currently, there are 53 sequence-based CBM families (Cantarel *et al*. 2008) with specificities ranging from the relatively simple, such as for crystalline cellulose, chitin, the β-1,3-glucans, xylan, or mannan to more complex, such as for the human blood group antigens (Gregg *et al*. 2008). Characterization of these families continues with the identification of ligands and structure-function studies. Much work has been done on plant cell wall binding CBMs as these were the first identified and therefore characterized. CBM families expected to have more "exotic" binding specificities are now being revealed in the sequences of carbohydrate-active enzymes in human pathogens such as *Streptococcus pneumoniae* and *Clostridium perfringens*.

Binding Site Architecture

The sequence-based families have been further subdivided into three types based on mechanism of ligand binding and binding site architecture (Figure 3). Type A CBMs bind to insoluble crystalline cellulose or chitin and have flat platform-like binding sites. The

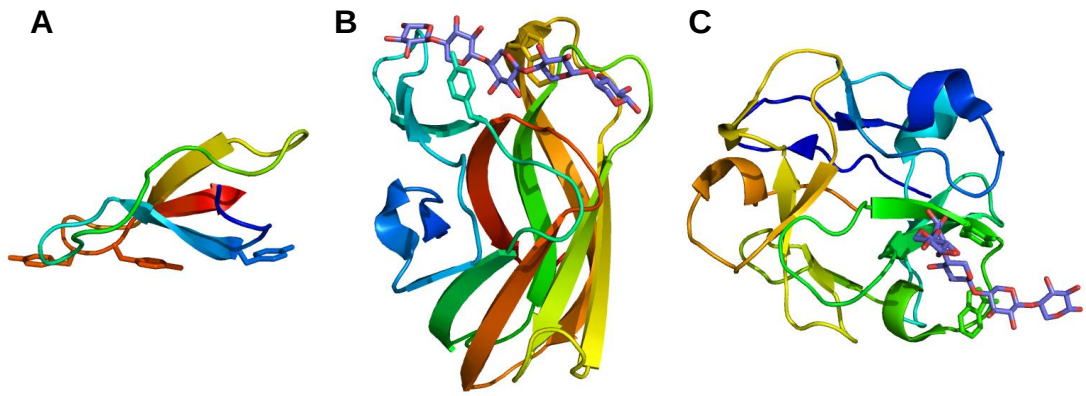


Figure 3. Three types of ligand binding and binding site architecture found in CBMs. Aromatic amino acids involved in binding are shown in stick A. Type A, cellulose binding HjCBM1 from *Hypocrea jecorina* (PDB code 1CBH) B. Type B, CtCBM6 from *Clostridium thermocellum* in complex with xylopentaose (PDB code 1UXX) C. Type C, SlCBM13 from *Streptococcus lividans* in complex with xylopentaose (PDB code 1MC9).

planar arrangement of aromatic residues within the binding site allow the CBM to rest on the flat hydrophobic surface of a crystalline ligand. Type B CBMs have a binding site that exists as an extended groove, within the groove there are multiple sugar subsites. These CBMs bind chains of soluble polysaccharides and both polar and apolar interactions drive specificity of binding. Type C CBMs, considered lectin-like, have a shallow binding pocket ideal for binding mono-, di- or tri-saccharides. As with the Type B CBMs both polar and apolar interactions drive specificity; however, in general the hydrogen bonding interactions are more extensive for the type C CBMs (Boraston *et al*. 2004). Members of the Type C CBMs include families: 9, 13, 14, 18, 32, 47 and 51. Family 32 carbohydrate-binding modules are a focus of this study and will be discussed further in Chapter 2.

Fold Families

Currently there are seven fold families described for CBMs (Cantarel *et al*. 2008). The seven different folds include the β-sandwich fold, the β-trefoil fold, the oligonucleotide-carbohydrate binding fold (OB), the knottin fold, the hevein fold, the hevein-like fold and a unique fold. The most frequent is the β-sandwich fold which is common to plant legume lectins, and animal galectins. The second most common fold is the β-trefoil fold.

CBM Function

CBMs increase the ability of glycoside hydrolases to efficiently degrade their polysaccharide substrates. This occurs through three different roles that CBMs can play in polysaccharide breakdown.

Proximity Effect

The proximity effect describes the binding of the CBM to bring the catalytic module into close association (or proximity) with substrate for a prolonged period. This effect is seen primarily on insoluble substrates such as cellulose and xylan (Bolam *et al*. 1998; Boraston *et al*. 2004; Boraston *et al*. 2003a; Charnock *et al*. 2000; Tomme *et al*. 1988).

Targeting Effect

The second effect is the targeting effect. This is described as a finer specificity for carbohydrate substructures. For example, there are many cellulose specific CBMs; however, type A cellulose binding CBMs bind crystalline cellulose whereas, type B cellulose binding CBMs bind non-crystalline components of cellulose. Thus, two different binding mechanisms are driving the recognition different components of cellulose substructure. This targeting effect would drive hydrolysis in specific regions of cellulose, rather than just bringing the catalytic modules into proximity as described above (Boraston *et al*. 2004; Carrard *et al*. 2000).

Disruptive Effect

Finally, there has been a disruptive effect described in the literature. This may arise due to disruption of polysaccharide fibers due to CBM permeation within the fibers. It is suggested that the CBMs bind and disrupt crystalline cellulose (Din *et al*. 1994; Gao *et al*. 2001) or chitin (Vaaje-Kolstad *et al*. 2005) allowing release of any non-covalently attached fibers thereby exposing further sites for hydrolysis.

CBMs in Glycosyl Transferases

CBM13 modules have been identified in family 27 glycosyl transferases suggesting a role for CBMs in anabolism (Cantarel *et al*. 2008). The first step in the O-glycosylation of mucin is the transfer of UDP activated GalNAc to a serine or threonine residue on the Muc protein to form the base for any further glycosyl decorations. The transfer reaction is catalyzed by GalNAc transferases (UDP-GalNAc: polypeptide α-*N*-acetylgalactosaminyltransferases). Recognition of partially glycosylated substrates is mediated by the CBM (Gerken *et al*. 2008; Raman *et al*. 2008). Different GalNAc transferases have different modes of recognition, likely driven by CBM specificity.

CBMs and Multivalency

Often, more than one CBM may be found within a glycoside hydrolase. CBMs can occur in tandem with one another within the enzyme though this is not always the case as the CBMs may also be separated by other modules. More than one CBM from the same family may occur in an enzyme; however, CBMs from different families may also occur within the same enzyme. CBMs in tandem may show increased affinity for ligand over that of the individual CBMs though this is not always true (Bolam *et al*. 2001; Boraston *et al*. 2002; Connaris *et al*. 2009; Ponyi *et al*. 2000; Tomme *et al*. 1996). Multivalent binding can help maintain the CBMs proximity to the carbohydrate surface or fine tune targeting.

1.1.6 Cohesin-Dockerin Interactions

Plant cell wall degrading Clostridial species employ a cohesin-dockerin system of protein-protein interactions to build polysaccharolytic superstructures in excess of 1 megadalton molecular weight and comprising multiple glycoside hydrolases (Bayer *et al*. 1998; Hammel *et al*. 2004). Cohesin and dockerin modules have been well characterized in these cellulolytic bacteria where their high affinity interactions help mediate enzyme assembly and attachement to the cell surface contributing to the huge cellulose degrading complex called the cellulosome (Bayer *et al*. 1998; Hammel *et al*. 2004). This organization acts to condense enzyme activities near the bacterial cell surface and increases the efficiency of cellulose degradation.

Recently, modules sharing sequence identity to the cohesins and dockerins from cellulolytic Clostridia were identified in pathogenic *C. perfringens* glycoside hydrolases, suggesting the formation of higher order structures between these enzymes (Adams *et al*. 2008; Chitayat *et al*. 2008b).

1.2 Carbohydrates

1.2.1 Carbohydrate Function

Carbohydrates are ubiquitous throughout our environment and the most abundant biomolecule on Earth (Peters 2006). They can be found in their simplest form, the monosaccharide, in a wide variety of stereochemical forms, to more complicated combinations such as the complex mannose type glycans that are N-linked to glycoproteins. The most abundant polysaccharide is cellulose (a simple polymer of β-1,4-linked glucose) which is found in the cell wall of plants where it has a critical structural role (Mutwil *et al*. 2008). Another structural polysaccharide is chitin, the principle component in the exoskeletons of arthropods (Vogan *et al*. 2008), and it is also found in fungal cell walls (Duo-Chuan 2006). Bacterial cell walls contain peptidoglycan, a polymer of repeating disaccharide residues, cross linked by short peptides, which are important for rigidity and strength (Deva *et al*. 2006). Other polysaccharides such as starch (amylose) and glycogen are important for energy storage in plants and animals (Flatt 1995; Geigenberger *et al*. 2005).

Carbohydrates play a crucial role in numerous biological processes. Carbohydrates and their conjugates can play roles in signal transduction (Li *et al*. 2006), adhesion (Jabbar *et al*. 2006), cell death (Raymond and Le Stunff 2006) (Thon *et al*. 2005), viral entry (Shukla and Spear 2001) and bacterial pathogenesis (host-pathogen interactions) (Landry *et al*. 2006) as well as for structural purposes (Mutwil *et al*. 2008).

1.2.2 Mammalian Complex Glycans

Carbohydrates decorate our cell surfaces and can also be found in the extracellular milieu. They may form variety of glycosidic linkages through different monosaccharides to generate a multitude of structures. Carbohydrates can be found within the context of: glycoproteins, glycolipids, glycosaminoglycans, and proteoglycans. O-linked and N-linked glycosylation, which help form glycoconjugates, is discussed below.

O-linked glycosylation

The most abundant type of O-glycosylation is that found in mucin (Strous and Dekker 1992). Here, GalNAc is linked in the α-configuration to a serine or threonine residue; this is referred to as the Tn antigen and forms the base for the eight core mucin structures shown in Figure 4 (Pratt and Bertozzi 2005). These core structures can themselves be quite decorated with sialic acid, fucose, sulphation and other decorations. The highly glycosylated and hydrated mucin glycoproteins, which are a constituent of mucus, provide a sticky protective lining to the surface of airways, the urogenital tract and the gastrointestinal tract providing a physical barrier against the entry of pathogens (Hounsell and Feizi 1982; Thornton and Sheehan 2004). Mucins are quite extended and have many sites of O-linked glycosylation as well as sites of N-linked glycosylation (Ho *et al*. 2003; Strous and Dekker 1992).

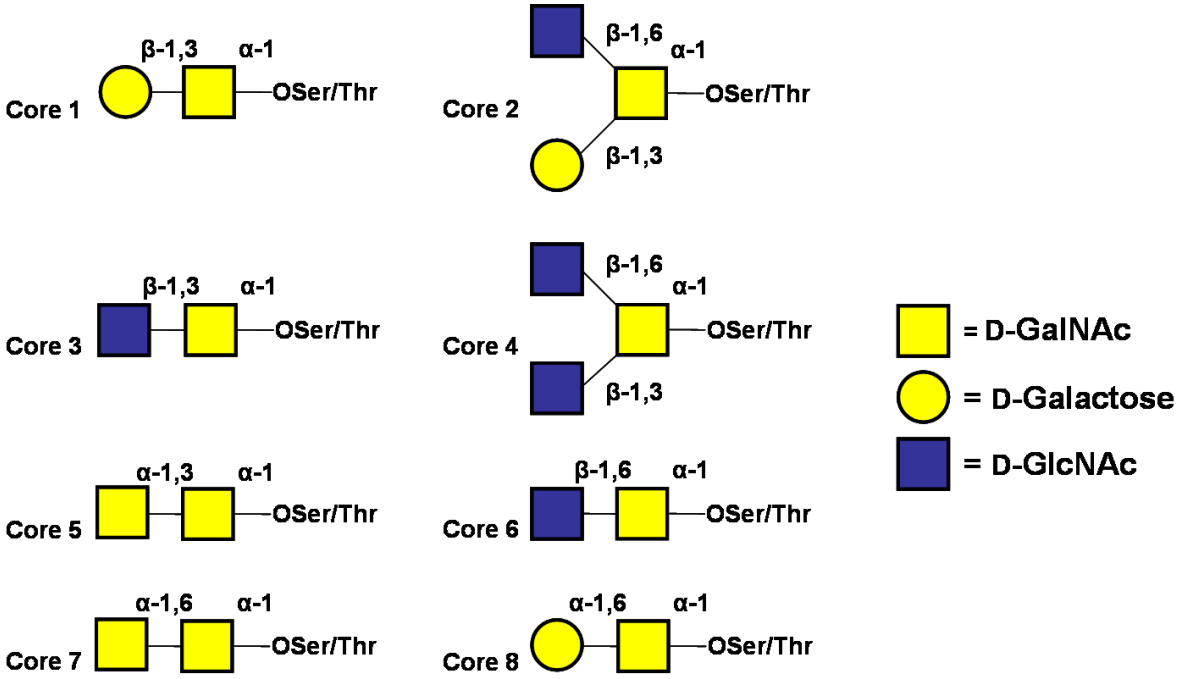


Figure 4. The eight core structures of mucin.

Another form of O-linked glycosylation is the dynamic β-O-GlcNAc modifications on serine and threonine residues as seen on cytosolic and nuclear proteins. The transfer of GlcNAc to a polypeptide is mediated by a family 41 glycosyl transferase, uridine diphosphate-*N*-acetyl-D-glucosamine: polypeptidyltransferase or OGT (Martinez-Fleites *et al*. 2008). A family 84 glycoside hydrolase, O-GlcNAcase, catalyzes the removal of the GlcNAc residues (Slawson *et al*. 2006). This modification can also have dynamic interplay with phosphorylation (Comer and Hart 2000). O-GlcNAc modifications play a role in cell signalling events (Slawson *et al*. 2006) and O-GlcNAc abnormalities have been linked to Alzheimer's (Liu *et al*. 2004), diabetes (Akimoto *et al*. 2007) and cancer (Chou and Hart 2001).

O-linked glycosylation also occurs with glycosaminoglycans which are found appended to proteoglycans *via* a β-linked xylose residue (Pratt and Bertozzi 2005). Serine and threonine residues have also been shown to carry β-O-linked glucose, α-O-linked mannose and α-O-linked fucose (Pratt and Bertozzi 2005).

N-linked glycosylation

N-linked glycosylation occurs on secreted or membrane bound proteins at the amino acid sequence NXS/T where X is any amino acid but proline. An oligosaccharide transferase in the ER transfers an oligosaccharide to the terminal amide of Asn during translation. Further modification of the glycan, such as trimming and the addition of decorations, occur in the ER and Golgi. The pentasaccharide core $Man_3GlcNAc_2$βN-Asn forms the base for all N-linked glycans (Figure 5A) (Pratt and Bertozzi 2005). This core can be

extended to generate high mannose, hybrid or complex N-linked glycans (Figure 5ABC). Further modifications can occur with galactose, GalNAc, sialic acid, GlcNAc and sulphation. Intracellularly, N-linked glycans can function in protein folding and trafficking. Extracellularly, N-linked glycans may be structural or they may act as carbohydrate receptors (Pratt and Bertozzi 2005).

Connective tissue

Connective tissue is made up of cells and the extracellular products they secrete. Complex glycans are an important component of connective tissue and can be found as glycosaminoglycans (acid mucopolysaccharides), proteoglycans (glycosaminoglycans bound to protein) and glycoproteins. Glycosaminoglycans are long unbranched polysaccharides that are composed of a repeating disaccharide. The following glycosaminoglycans are some of the most common that can be found in connective tissue (Figure 6): chondroitin sulphate, keratan sulphate, heparan sulphate, and hyaluronan. These glycosaminoglycans have a structural role within our extracellular matrix though they may play a part in other processes as well. For example, hyaluronan has a signalling role impacting cell proliferation (Kosaki *et al*. 1999) and cell motility (Ichikawa *et al*. 1999; Lee and Spicer 2000).

Blood Antigens

Probably the most well known complex glycans are the ABO histo-blood group antigenic determinants. The type O blood group, or H antigen, is the base upon which type A and B are formed (Figure 7). There are five types of minimal disaccharide precursors:

A
α-1,6
α-1,3
α-1,6
α-1,3
α-1,3
α-1,2
β-1,4
β-1,4
β-1
NAsn
B
β-1,4
β-1,2
α-1,6
α-1,3
β-1,4
β-1,4
β-1
NAsn
α-2,6
β-1,4
β-1,2
C
α-1,6
α-1,3
α-1,6
α-1,3
β-1,4
β-1,4
β-1
NAsn
β-1,4
β-1,2
= D-GalNAc
= D-Galactose
= D-GlcNAc
= D-Mannose
= Neuraminic acid

Figure 5 (Previous page). Schematic of potential N-linked glycan linkages. The core pentasaccharide is surrounded by an orange box. The three classes of N-linked glycans are (A) high mannose (B) complex and (C) hybrid.

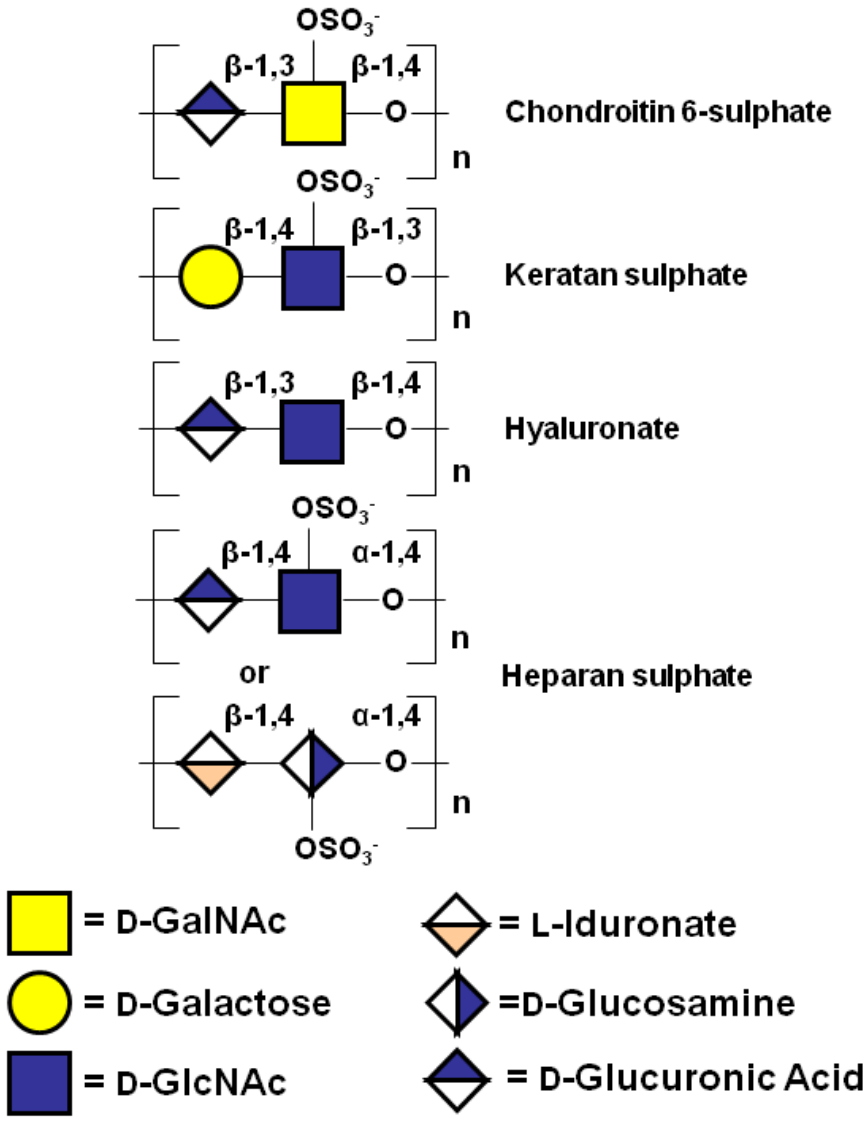


Figure 6. Components of mammalian connective tissue.

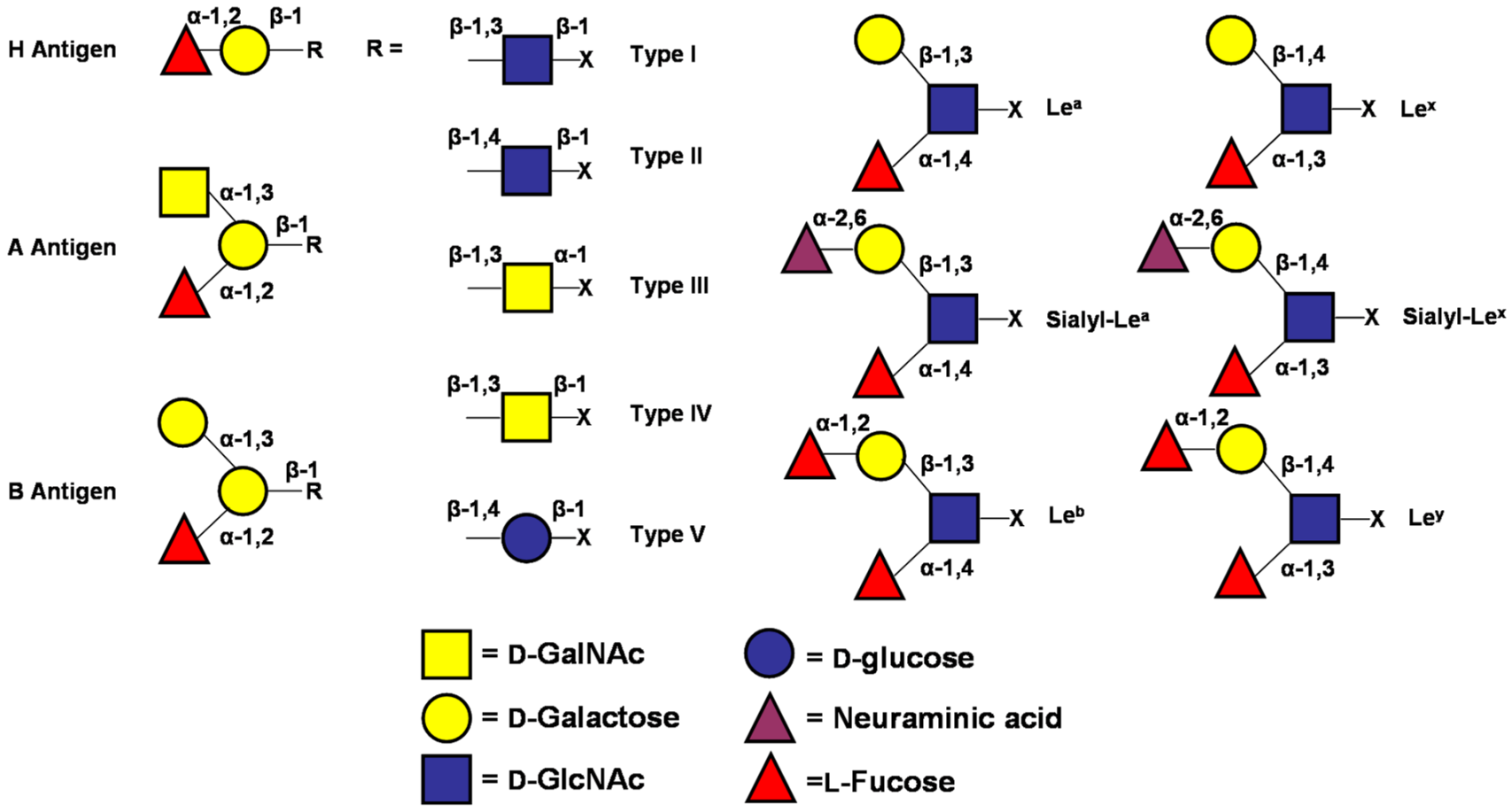

H Antigen
α-1,2
β-1
R
A Antigen
α-1,3
β-1
R
α-1,2
B Antigen
α-1,3
β-1
R
α-1,2
R =
β-1,3
β-1
X
Type I
β-1,4
β-1
X
Type II
β-1,3
α-1
X
Type III
β-1,3
β-1
X
Type IV
β-1,4
β-1
X
Type V
β-1,3
X
Le^a
α-1,4
β-1,4
X
Le^x
α-1,3
α-2,6
β-1,3
X
Sialyl-Le^a
α-1,4
α-2,6
β-1,4
X
Sialyl-Le^x
α-1,3
α-1,2
β-1,3
X
Le^b
α-1,4
α-1,2
β-1,4
X
Le^y
α-1,3
= D-GalNAc
= D-Galactose
= D-GlcNAc
= D-glucose
= Neuraminic acid
=L-Fucose

Figure 7 (Previous page). ABO histo-blood group structures and the Lewis antigens. X represents another glycan component.

Type 1, Gal-β-1,3-GlcNAc-β-1-R (D-galactose-β-1,3-*N*-acetyl-D-glucosamine-β-1-R); Type 2, Gal-β-1,4-GlcNAc-β-1-R; Type 3, Gal-β-1,3-GalNAc-α-1-R (D-galactose-β-1,3-*N*-acetyl-D-galactosamine-α-1-R); Type 4, Gal-β-1,3-GalNAc-β-1-R; and type 5, Gal-β-1,4-Glc-β-1-R (D-galactose-β-1,4-D-glucose-β-1-R) (Marionneau *et al*. 2001). The addition of an α-1,2 linked fucose is catalyzed by an α-1,2-fucosyltranferase to form the H antigen (Marionneau *et al*. 2001). To form the A antigen a *N*-acetylgalactosamine residue is linked α-1,3 to the galactose moiety of the H antigen by glycosyltransferase A (GTA) (Letts *et al*. 2006). To form the B antigen a galactose residue is linked α-1,3 to the galactose by glycosyltransferase B (GTB) (Letts *et al*. 2006). These antigens are synthesized by glycosyl transferases in a stepwise manner.

The histo-blood group antigens can be found on the surface of erythrocytes as well as in most epithelial tissues and secretions (Ravn and Dabelsteen 2000). The expression of these antigens varies between cells and organs. The histo-blood group antigens can be linked to many different core carbohydrate structures; however, the ABO antigen is maintained regardless of the core (Ravn and Dabelsteen 2000). An understanding of blood type is important in transfusion medicine; however, the biological function of ABO antigens remains unknown.

The Lewis antigens are related to the human blood group antigens as they also use the Type I and Type II core structures as bases for fucosylation. The first type, Type I, is formed from Gal-β-1,3-GlcNAc which is the backbone for Lewis^{a} (Le^{a}), sialyl-Le^{a}, and Lewis^{b} (Le^{b}). The second type, Type II, is formed from Gal-β-1,4-GlcNAc (*N*-

acetyllactosamine or LacNAc) which is the backbone for Lewis^x (Le^x), sialyl-Le^x and Le^y (Figure 3) (Moran 2008). Similar to the human blood group antigens, the expression varies between tissues (Ravn and Dabelsteen 2000). Abnormal expression of the Lewis antigens has been found to be correlated to cancer (Lloyd 2000) and these antigens are important in some host-pathogen interactions (See section 1.3).

1.3 Bacterial Pathogens and Binding Interactions to Host Cell Glycans

1.3.1 Adhesins

Adhesins are proteins at the cell surface of a microorganism that promote attachment through specific binding interactions to receptors. Adhesins help mediate host-pathogen interaction in a variety of systems and adherence is often seen as a crucial step for successful colonization and subsequent infection of a host. The following is a brief discussion on adhesin mediated protein-carbohydrate interactions in bacterial pathogens.

The Lewis antigens have been implicated in bacterial pathogenesis, particularly with the bacterium *Helicobacter pylori*. *H. pylori* colonizes the gastroduodenal tract in humans and can cause gastritis, gastric and duodenal ulcers and an increased risk of gastric adenocarcinoma and gastric lymphoma (Ernst and Gold 2000) (Moran 2008). Le^a, Le^b, Le^x and Le^y are expressed in the human stomach, the ecological niche of *H. pylori* (Moran 2008). The best characterized adhesin-carbohydrate interaction of *H. pylori* is that of outer membrane adhesin BabA which binds host Le^b and related fucosylated antigens. Another *H. pylori* adhesin, SabA recognizes sialyl-Le^x and sialyl-Le^a antigens. These adhesins have been implicated in gastroduodenal pathogenesis by *H. pylori* (Yamaoka 2008). *H. pylori*, in turn, carries the Le^x and terminal Le^x and Le^y units at the surface of its lipopolysaccharide. Some strains additionally carry Le^a, Le^b, sialyl-Le^x and ABO histo-blood group units (Moran 2008). Initially, mimicry was thought to play a role in immune evasion; however, this expression of human type glycosylation patterns has now been shown to be important to gastric colonization and bacterial adhesion. Galectin-3, a lectin bound to cell surfaces and within the extracellular matrix of the host (Sato and

Hughes 1994), is a gastric receptor for polymeric Le^x on the bacterium (Fowler *et al*. 2006). Their interaction helps drive adhesion of the bacterium to gastric mucosa.

Bacteria, such as *Escherichia coli*, have surface associated fimbrial lectins which help mediate adhesion to surface oligosaccharides of the host cell. Infection by *E. coli* has been implicated in urinary tract infections, sepsis, diarrhea and newborn meningitis (Hacker 1990; Lindhorst *et al*. 1998). Type 1 fimbriae recognize mannosyl ligands and are composed of two subunits, the FimA subunit, which is the major repeating subunit, and FimH, the mannose specific adhesin found along the shaft of the fimbriae and at the lateral end. There are between 100-400 type 1 fimbriae found on the surface of *E. coli* cells (Lindhorst *et al*. 1998) which allows these adhesins to adhere in a multivalent manner, *via* the clustered lectin domains, to the multivalent mannosyl cell surface glycoconjugates (Kiessling and Pohl 1996; Lindhorst *et al*. 1998).

Another opportunistic pathogen, *Pseudomonas aeruginosa*, is a major colonizer of the human lungs. This organism is most frequently found in the lungs of cystic fibrosis patients and in patients with chronic lung diseases. *P. aeruginosa* binds directly to the mucus layer covering the lungs through multiple adhesion events. The Le^x, Le^y and Le^a antigens as well as other complex carbohydrate epitopes have been identified as carbohydrate receptors (Ramphal and Arora 2001; Scharfman *et al*. 2001).

Interactions with the complex glycans found in the human body allow adherence of these bacterial pathogens, to name just a few, to the host cell tissue which helps to provide an

anchoring point for infection to set it. As seen in *H. pylori*, bacterial mimicry of host cell glycans in order to interact with the host's galectins can also drive adherence.

1.3.2 Carbohydrate Binding Modules

Most of what we know about CBMs and their interactions with carbohydrate comes from plant cell wall degrading glycoside hydrolases; however, more CBM families with predicted specificity for complex carbohydrates, such as eukaryotic glycans, are now being identified. CBM families that bind more complex glycans such as human blood group antigens, terminal sialic acid residues and fucose, include 13, 32, 40, 47, and 51.

Vibrio cholerae produces a sialidase with a family 40 CBM that binds sialic acid. The effect of sialidase activity is to hydrolyze terminal sialic acid residues from gangliosides thereby exposing GM1, the receptor for the cholera toxin (Moustafa *et al*. 2004a). Thobhani *et al* demonstrated that catalytic efficiency of the *Vibrio cholerae* sialidase with its sialic acid binding CBM is increased on a polyvalent substrate relative to a monovalent substrate. These researchers were able to design a polyvalent inhibitor that targeted carbohydrate binding and catalysis (Thobhani *et al*. 2003).

Binding to blood group A/B antigens has been described by a CBM51 module, GH98CBM51 (Gregg *et al*. 2008) from *C. perfringens*. Binding to galactose has been described by another member of this family GH95CBM51. GH98 acts as an *endo*-β-galactosidase on human blood group A and B antigens (Anderson *et al*. 2005), while GH95 is an α-L-fucosidase (Katayama *et al*. 2004). The binding to complex human

glycans by these CBMs likely targets enzymes with specificities for such glycans to help in the efficient breakdown of tissue.

CBM47 modules from *Streptococcus pneumonia* bind the fucosylated blood group ABO antigens and to the Le^y antigen (Boraston *et al*. 2006). These CBMs are found appended as a triplet to a family 98 glycoside hydrolase, a putative virulence factor predicted to be active on host lung tissue (Boraston *et al*. 2006). The studies on the triplet of CBMs revealed a multivalent avidity effect on binding fucosylated resin *in vitro* suggesting a purpose for the replication of this motif within the enzyme.

1.4 *Clostridium perfringens*

The gram positive, spore forming, anaerobic bacterium, *Clostridium perfringens*, is found in the gastro-intestinal tract of animals, and ubiquitously throughout the environment (Hatheway 1990; McDonel 1980). *C. perfringens* is a causative agent of gas gangrene, necrotic enteritis, and food poisoning (Hatheway 1990; McDonel 1986; Rood and Cole 1991). Requirements for *C. perfringens* infection are necrotic tissue and an anaerobic environment (Smith 1975). There are 5 biotypes of *C. perfringens* (A-E) which are characterized on the basis of the toxins they produce (Table 1).

Biotype	Toxins	Diseases
A	α, μ, cpe	gas gangrene, food poisoning
B	α, β, ε, μ, cpe	enteritis in older lambs, hemorrhagic enteritis in neonatal calves and foals, hemorrhagic enterotoximia in adult sheep
C	α, β, μ, cpe	Necrotic enteritis in humans, fowl, pigs, lambs, goats calves, foals
D	α, ε, μ, cpe	enterotoximia in lambs and calves
E	α, ι, μ, cpe	enteritis in dogs and pigs

Table 1. The five disease causing biotypes of *C. perfringens* and some of the associated toxins.

C. perfringens secretes a battery of exotoxins which contribute to pathogenicity of the organism. Their secretion plays a role in the swift spread of the bacterium throughout the host. These virulence factors include, but are not limited to: the α toxin, which has phospholipase C activity (Titball *et al*. 1999); the pore forming β toxin (Shatursky *et al*. 2000); the pore forming ε toxin (Petit *et al*. 1997; Petit *et al*. 2003; Petit *et al*. 2001); the ADP-ribosylating ι toxin (Nagahama *et al*. 2000; Tsuge *et al*. 1999)); the μ-toxin, a glycoside hydrolase active on hyaluronan (Canard *et al*. 1994); and other putative glycoside hydrolases.

During World War I and World War II gas gangrene was a common killer. Improper wound debridement followed by immediate suturing provided an ideal growth environment for this anaerobe. Infections can occur spontaneously, post-operatively, or post-traumatically (Stevens 1997). Due to the rapid advancement of the disease sometimes the only option to prevent death is the removal of a limb.

Some bacteria that inhabit the gut are harmless such as the Bifidobacteria (Katayama *et al*. 2005). The appropriate balance of bacteria within the gut is thought to be important for proper digestion (Simon and Gorbach 1986). When this ecosystem is disturbed, by colonization with harmful bacteria, the balance can be destroyed and illness may result. Bacteria, such as Salmonella, *H. pylori* and *C. perfringens*, when they colonize, may cause discomfort, serious illness or even death.

Food poisoning or necrotic enteritis (pig-bel) may be incurred when eating improperly cooked or cooled down food that is contaminated with *C. perfringens*. Type A *C. perfringens* food poisoning events average over 100 people/outbreak as often the outbreaks occur in hospital or seniors centre type settings (McClane 2000; McClane 2001). Necrotic enteritis is endemic to Papua New Guinea where their diet is low in protein and high in sweet potato, which contains high levels of trypsin inhibitors. Usually infection occurs after the consumption of food (often pork) contaminated with type C *C. perfringens*. Due to tryptic inhibition combined with low levels of intestinal trypsin because of a protein poor diet, the pore forming β-toxin is not degraded and *C. perfringens* is able to obtain a foothold, which allows the spread of the infection, often causing severe tissue necrosis and death.

Much of what we know about polysaccharide degradation comes from plant cell wall degrading bacteria. These microorganisms produce enzyme systems, that is, enzymes with multiple specificities, in order to more efficiently degrade a polysaccharide. The plant cell wall degrading bacteria usually produce highly complex systems of polysaccharolytic enzymes as plant cell walls contain a more diverse mixture of polysaccharides than starch or chitin (Warren 1996). Curiously, recent sequencing of the genome of *C. perfringens*, a bacterial pathogen not active on plant polysaccharides, has allowed identification of 53 putative glycoside hydrolases. Of these, approximately half are predicted to be secreted and to have specificities for components found in complex glycans including hexosaminidases, galactosidases, hyaluronidases, sialidases, a fucosidase and putative glycoside hydrolases of several undefined specificities (Figure

8)(Cantarel *et al*. 2008) suggesting these pathogens contain saccharide degrading enzyme systems specific for host glycans. A striking feature of these glycoside hydrolases in *C. perfringens* is their modularity. They frequently comprise a catalytic module and up to 8 ancillary modules (Figure 8). Another observation is the repetition of the CBM32 motif throughout these enzymes, suggesting an important role to enzyme function (Section 1.1.5). *C. perfringens* is a known colonizer of the human gastrointestinal tract and invasive strains can cause massive tissue necrosis. This leads us to postulate that since these enzymes are secreted, the CBMs from this human pathogen might reveal complex binding specificities consistent with the ligand arrays found in the human body. Often, more than one CBM32 can be found in these enzymes making them potentially multivalent binders. The enzymes also contain modules that share sequence similarity with the interacting cohesin and dockerin modules from cellulolytic bacteria (Section 1.1.6). This suggests the formation of higher order structures between these enzymes, a phenomenon not described before in other glycoside hydrolases produced by pathogens. Investigation into the function of the catalytic and accessory modules of the glycoside hydrolases of *C. perfringens* will give a better understanding of the modular nature of these enzymes and possibly of other glycoside hydrolases in other organisms. ***The objective of this research is to characterize the key elements of the carbohydrate-active glycoside hydrolases of C. perfringens, particularly, CpGH84C and CpGH89.*** Understanding the role the modules play in the enzymes will help in understanding the role of the enzymes and protein-carbohydrate recognition in bacterial virulence. This research provides a glimpse into how a notorious human pathogen, known for its massive tissue destruction, coordinates such an event.

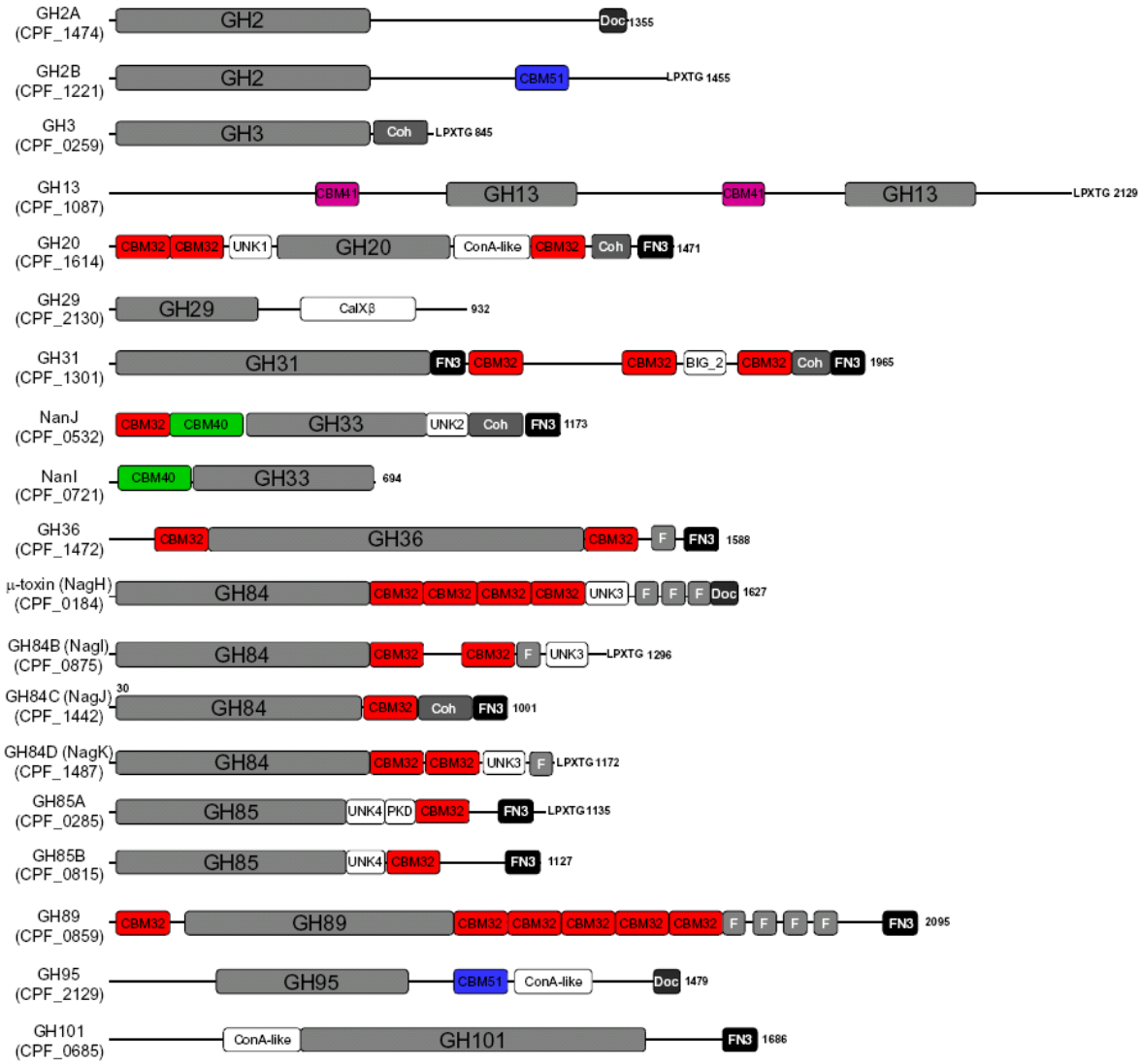


Figure 8. Modular schematics of the extracellular *C. perfringens* strain ATCC 13124 glycoside hydrolases. Protein designations are shown on the left with their locus tags and the full number of amino acids in the protein on the right. Catalytic modules are shown in grey and numbered by their CAZy classification. Red, green, pink, and blue indicate carbohydrate-binding modules which are labeled by their CAZy classification. Other module designations are F, FIVAR modules; Doc, dockerin modules; FN3, fibronectin-type III modules; and Coh, cohesin modules. Modules labeled UNK are of unknown function as are the PKD, ConA-like, BIG, and CalXb modules, which were defined on the basis of InterproScan analyses. All of these proteins possess an N terminal secretion signal sequence (not shown). Those proteins which have a C-terminal LPXTG motif to direct sortase-mediated cell-wall attachment are indicated.

Chapter 2: Structure-Function Studies on the Family 84 Glycoside Hydrolase Enzymes of *C. perfringens*

2.1 Introduction

The recognition of carbohydrates by proteins is a biological phenomenon of fundamental importance. It is critical to numerous events including the movement and interactions of cells and proteins in many organisms, recycling of plant carbohydrates, and interactions between hosts and disease causing organisms. Microbial and viral invaders of the human body often exploit host glycans to aid in adherence and then must contend with the protective and structural sugar layers to enable invasion and further spread of the infection. The exploitation of host carbohydrates as receptors for the adherence of pathogenic bacteria *via* non-catalytic lectin-like adhesins is relatively well understood for a number of different bacterial species, including uropathogenic *Escherichia coli* (Bouckaert *et al*. 2005; Zhou *et al*. 2001) and *Helicobacter pylori* (Lelwala-Guruge *et al*. 1992; Robinson *et al*. 1990). Some of the more spectacular bacterial infections, such as the severe myonecrotic infections caused by *Streptococcus pyogenes* and *Clostridium perfringens*, involve extensive tissue destruction (Bryant *et al*. 2005; Hynes 2004; Smedley *et al*. 2004; Stevens and Bryant 2002). The tissue destruction and bacterial spread appears to be aided by a variety of carbohydrate-active enzymes, which break down the polysaccharides of the extracellular matrix or potentiate the activity of other cytolytic toxins (Canard *et al*. 1994; Flores-Diaz *et al*. 2005; Gerding 1997; Sheldon *et al*. 2006). A large number of other bacterial pathogens also feature carbohydrate-active enzymes as important virulence factors that figure in a variety of roles related to degrading and modifying host glycans (Figura 1997; Galen *et al*. 1992; Shelburne *et al*. 2008). For example, enzymes active on sialic acid, a sugar typically implicated in

adherence events, and hyaluronic acid, a structural component of animal extracellular matrix, are commonly associated with the virulence of *Clostridium* sp. and *Streptococcus* sp. (Canard *et al*. 1994; King *et al*. 2006).

Mucus endows the surfaces of airways, the urogenital tract and the gastrointestinal tract with a protective layer and physical barrier to the entry of pathogens (Thornton and Sheehan 2004; Brayman *et al*. 2004; Einerhand *et al*. 2002). Key constituents of mucus are the mucins, which are highly hydrated glycoproteins comprising up to 80% carbohydrate. The glycan structures on gastric mucins vary from simple to very complex and differ in structure/composition depending on their location in the gastrointestinal track (Robbe *et al*. 2004). Recent studies of *Bacteroides thetaiotaomicron*, a symbiotic bacterial inhabitant of the human intestine, have revealed that its genome contains on the order of 250 genes that encode carbohydrate degrading enzymes (Comstock and Coyne 2003). Elegant gene expression studies showed that this bacterium preferentially expressed genes encoding enzymes specific to the degradation of dietary polysaccharides (Salyers *et al*. 1977; Salyers and Kotarski 1980; Sonnenburg *et al*. 2005). However, in the absence of dietary polysaccharides the expression of genes encoding enzymes with more "exotic" specificities (e.g. sialidases, β-hexosaminidases, mannosidases, and fucosidases) was up-regulated indicating a switch to the oligosaccharide side-chains of mucin as an alternative carbohydrate source (Sonnenburg *et al*. 2005). Though *B. thetaiotaomicron* has forged a relationship of symbiosis with its host, rather than one of pathogenesis, this organism does provide some useful parallels with the pathogenic bacterium *Clostridium perfringens*.

Determination of the genome sequence of *C. perfringens* (strain 13) (Shimizu *et al*. 2001) has revealed 53 open reading frames encoding putative glycoside hydrolases falling into 24 known glycoside hydrolase families (Cantarel *et al*. 2008; Coutinho 1999). Many encode intracellular proteins likely involved in the latter stages of sugar metabolism or proteins involved in peptidoglycan remodelling. However, roughly one-half are predicted to be secreted with many of these having specificities analogous to those suggested to be involved in muco-oligosaccharide metabolism in *B. thetaiotaomicron*. In contrast to *B. thetaiotaomicron*, *C. perfringens* appears to lack the enzymes capable of dietary polysaccharide degradation. Thus, though *C. perfringens* is most frequently thought of as a "flesh-eater", its most common niche in humans is the gastrointestinal tract and it appears that it is well-equipped to attack the diverse sugar structures of the mucins in this environment. Indeed, mucosal necrosis is associated with severe *C. perfringens* caused enteritis (Gerding 1997). This may be in part due to the arsenal of *C. perfringens* glycoside hydrolases. In turn, breaking down the mucosal barrier could improve access of other toxins, such as the pore forming cpe (*Clostridium perfringens* enterotoxin) and α-toxins, to the epithelial layer.

The family 84 glycoside hydrolases are a group of enzymes found in pathogens, such as *Clostridium perfringens* and *Streptococcus pyogenes*, the common commensal gut bacterium *Bacteroides thetaiotaomicron* and even in humans. The most notorious of these is the μ-toxin or CpGH84A (EC 3.2.1.35), which has been implicated as a putative virulence factor and preliminarily characterized as a hyaluronidase due to its activity on hyaluronan (Canard *et al*. 1994), a glycosaminoglycan found in human connective tissue

comprising the repeating disaccharide unit β-1,4-glucuronic acid-β-1,3-GlcNAc (Figure 6). There are four other family 84 glycoside hydrolases, initially presumed to be hyaluronidases, in *Clostridium perfringens* (CpGH84B, CpGH84C, CpGH84D and CpGH84E all with EC 3.2.1.52) identified through sequence similarity to CpGH84A. The recurring presence of these enzymes suggests their importance to *C. perfringens* survival and virulence. Currently, evidence is emerging that these enzymes are in fact not hyaluronidases but β-*N*-acetylglucosaminidases (Ficko-Blean and Boraston 2005; Macauley *et al*. 2005; Sheldon *et al*. 2006).

Macauley *et al* first determined, in a sophisticated study, that the human family 84 glycoside hydrolase operates using substrate assisted catalysis (Macauley *et al*. 2005). Subsequently the structure of the bacterial CpGH84C catalytic module was published with PUGNAc [O-(2-acetamido-2-deoxy-D-glucopyranosylidene)amino-N-phenylcarbamate], a putative transition state mimic (Rao *et al*. 2006). The bacterial homologue, which has two adjacent catalytically important Asp residues, also proceeds *via* anchimeric assistance. The structure of a family 84 glycoside hydrolase from *Bacteroides thetaiotaomicron* was also published providing more evidence that hydrolysis also proceeds through anchimeric assistance within the family 84 glycoside hydrolases (Dennis *et al*. 2006).

Glycoside hydrolases are often modular, containing one or more ancillary modules. Carbohydrate binding modules (CBMs) are the most common ancillary module found in glycoside hydrolases. The lectin-like family 32 CBMs have been identified in pathogenic

and non-pathogenic bacteria and in eukaryotic species. The structure of a family 32 CBM has previously been solved in complex with galactose from a sialidase produced by the non-pathogenic bacterium *Micromonospora viridifaciens* (Gaskell *et al*. 1995; Newstead *et al*. 2005). Sequence similarity permitted the identification of numerous family 32 CBMs in *Clostridium perfringens*. CBM32s are prominent in the five family 84 glycoside hydrolases identified within *C. perfringens* as well as in other secreted glycoside hydrolases. CBMs can occur N or C terminally and often more than one occur in tandem. It is likely that the family 32 CBMs play an important role in enzyme function due to their prominence among these enzymes. Their presence throughout the genome of a bacterial pathogen leads us to hypothesize the following: ***the function of complex carbohydrate binding can be ascribed to the CBM32 modules found within the C. perfringens glycoside hydrolases.***

The focus of this project is the family 84 enzymes, particularly the μ-toxin and CpGH84C (Figure 8). The μ-toxin was chosen for study as it has been previously characterized as a toxin and therefore holds more impact than a predicted toxin. CpGH84C was chosen for characterization due to its relatively small size (1001 amino acids) and simple modular nature; containing only one appended family 32 CBM. Finally it is the only enzyme where all the ancillary modules share sequence identity to known polypeptide sequences, thus there are no unknown modules. Despite the mounting evidence that a large number of carbohydrate degrading enzymes are virulence factors, relatively little is known about the structure-function relationship of the ancillary modules and how they might affect enzyme function. We can hypothesize that: ***the complete structure of the***

modules in relationship to one another is related to the function of the individual modules and to the function of the enzyme as a whole.

The complexity of the glycoside hydrolase enzymes of *C. perfringens* is revealed by the structure-function studies described within this chapter, contributing to our understanding on carbohydrate-active enzymes.

2.2 The Interaction of a Carbohydrate-Binding Module from a Bacterial *N*-Acetyl-β-Hexosaminidase with its Carbohydrate Receptor

Elizabeth Ficko-Blean and Alisdair B. Boraston

Adapted from: The interaction of a carbohydrate-binding module from a Clostridium perfringens N-acetyl-beta-hexosaminidase with its carbohydrate receptor. J Biol Chem. 2006 Dec 8;281(49):37748-57

Biochemistry & Microbiology, University of Victoria, Victoria, British Columbia V8W 3P6, Canada

2.2.1 Abstract

Clostridium perfringens is a notable colonizer of the human gastrointestinal tract. This bacterium is quite remarkable for a human pathogen by the number of glycoside hydrolases found in its genome. The modularity of these enzymes is striking as is the frequent occurrence of modules having amino acid sequence identity with family 32 carbohydrate-binding modules (CBMs), often referred to as F5/8 domains. Here we report the properties of a family 32 CBM from a *C. perfringens N*-acetyl-β-D-hexosaminidase, CpGH84C. Macroarray, UV difference and ITC binding studies indicate a preference for the disaccharide *N*-acetyllactosamine. The molecular details of the interaction of CpCBM32C with galactose, *N*-acetyllactosamine, and the type II blood group H-trisaccharide are revealed by X-ray crystallographic studies at 1.49 Å, 2.4 Å, and 2.3 Å resolution respectively.

2.2.2 Introduction

Thirteen of the predicted *C. perfringens* (strain 13) glycoside hydrolases [and notably 13 glycoside hydrolases for each of the sequenced *Bacteroides* sp. genomes (*thetaiotaomicron*, *fragilis* YCH46, and *fragilis* 25285)] have, in addition to catalytic domains, modules with amino acid sequence identity to family 32 carbohydrate-binding modules (Figure 8) (Cantarel *et al*. 2008). This hints at the possible importance of the putative non-catalytic carbohydrate-binding modules in the functions of enzymes from these gastrointestinal inhabitants/pathogens. In order to better understand these glycoside hydrolases and their CBMs we initiated studies of CpGH84C from *C. perfringens* (strain ATCC 13124). The predicted modular structure of CpGH84C is as follows: there is a secretion signal at the N-terminus of the protein followed by the family 84 catalytic module, C-terminal to the catalytic module is an internal family 32 CBM (CpCBM32C). Next, a module which shares sequence identity with the cohesin domains from other Clostridial species and at the C-terminus of the protein is a module that shares sequence similarity with fibronectin type III repeats (Figure 8).

To facilitate structure-function studies, we dissected this protein at the genetic level to recombinantly produce isolated CpCBM32C. The experimental results reveal the ability of CpCBM32C to bind to the terminal LacNAc (β-D-galactosyl-1,4-β-D-*N*-acetylglucosamine) glycotopes commonly found in elaborated O- and complex N-linked glycans (Gupta *et al*. 1996; Morelle *et al*. 2000; Robbe *et al*. 2004). The X-ray crystal structures of CpCBM32C in complex with sugar help uncover the molecular details that confer this binding ability.

2.2.3 Materials and Methods

Materials

Unless otherwise stated, chemicals, carbohydrates, glycoproteins, and polysaccharides were purchased from Sigma.

Cloning

The DNA fragment encoding the family 32 CBM (Figure 8) of *Cp*GH84C was amplified by PCR from *C. perfringens* genomic DNA (Sigma; ATCC 13124) using previously described methods. Nucleotides 1873-2301 of the *cpgh84c* gene, which corresponds to the CBM (amino acid residues 625-767), was amplified with the oligonucleotide primers 5'-CACCAATCCAAGAACAGTAAAG-3' (CBMF) and 5'-CTTTTATCCATGAACATTAACCTC-3' (CBMR). The amplified gene fragment was ligated directly into the pET-150 TOPO Directional Cloning kit (Invitrogen, San Diego, CA) to generate pCBM32. The polypeptide (called CpCBM32C) encoded by pCBM32 comprises a His_6 tag fused to the CBM32 module by an enterokinase protease cleavage site.

Protein production and purification

BL21star (DE3) *Escherichia coli* expression strain (Invitrogen, San Diego) was transformed with pCBM. A 1.5 L culture was grown in Luria-Bertani (LB) media, supplemented with ampicillin (100 μg/ml), was grown to an OD ~ 1 and induced with 1 mM IPTG then grown overnight at 37 ˚C. The cells were harvested at 4000 x g and resuspended in 20 ml of binding buffer containing 20 mM Tris-HCl, pH 8.0, and 0.5 M NaCl. Cells were lysed using a French pressure cell. Cell debris was removed by

centrifugation for an hour at 26915 x g. The supernatant was applied to His-Select resin followed by step elution with binding buffer containing imidazole concentrations between 5 and 500 mM. Samples were run on a 15% SDS gel and fractions containing the polypeptide of interest were pooled. Proteins were concentrated and buffer exchanged in a stirred ultra-filtration unit (Amicon, Beverly, MA) using a 5K molecular weight cut-off (MWCO) membrane (Filtron, Northborough, MA). Purity, assessed by SDS-PAGE, was greater than 95%.

Determination of protein concentration

The concentrations of purified proteins were determined by UV absorbance (280 nm) using calculated molar extinction coefficients (18450 $M^{-1}cm^{-1}$) (Gasteiger 2005).

UV Difference

Automated UV difference titrations of CpCBM32C were performed as described previously (Boraston *et al*. 2001b). Difference spectra were examined for peak and trough wavelengths and values at the appropriate wavelengths extracted for further analysis. The wavelengths for the maximum peak to trough differences were determined individually for each sugar solution. The peak-to-trough heights at three wavelength pairs were calculated by subtraction of the trough values from the peak values and the dilution corrected data plotted against total carbohydrate concentration. Data for the three wavelength pairs was analyzed simultaneously with MicroCal Origin (v.7.0) using a one site binding model accounting for ligand depletion. Experiments were performed at 20

^{o}C in 50 mM Tris-HCl, pH 7.5. The data reported are the averages and standard deviations of three independent titrations.

Isothermal titration calorimetry

Isothermal titration calorimetry (ITC) was performed as described previously (Boraston *et al*. 2001a; Lammerts van Bueren and Boraston 2004) using a VP-ITC (MicroCal, Northampton, MA). Protein samples were extensively dialyzed against buffer (50 mM Tris-HCl, pH 7.5). Sugar solutions were prepared by mass in buffer saved from the final protein dialysis step. Both protein and sugar solutions were filtered and degassed immediately prior to use. Protein concentrations were determined by UV absorbance as described above. Aliquots of 10 mM LacNAc were titrated into CpCBM32C at 350 μM, which gave C-values > 5 (Wiseman *et al*. 1989). Aliquots of 5.0 mM human type II H-trisaccharide (α-L-fucosyl-1,2-β-D-galactosyl-1,4-β-D-*N*-acetylglucosamine) were titrated into CpCBM32C at 245 μM. In this case, the C-value was less than 1 due to the low binding affinity. Based on the 1:1 binding observed in the crystal structure, the stoichiometry was fixed at 1 in the analysis of this data.

Crystallization and data collection

All crystallization experiments were performed using the hanging-drop vapour-diffusion method. Prior to crystallization, the H6-tag was removed from CpCBM32C by treatment with enterokinase over a 4 day period. The digested sample was run through a Novagen His-bind Quick 900 cartridge to remove the His tag and any undigested protein from the solution. Samples were concentrated and exchanged as above into 20 mM Tris-HCl, pH

8.0. Co-crystals of CpCBM32C (10.5 mg/ml) with galactose (~10 mM) were obtained with 0.2M $MgCl_2$, 25 %, polyethylene glycol monomethylether 2000 (PEG 2000 MME), and 0.1 M Tris pH 7.5. These crystals were cryoprotected with 15% glycerol in mother liquor. 1.5 M sodium/potassium phosphate was used to co-crystallize CpCBM32C (20 mg/ml) LacNAc and the type II blood group H-trisaccharide. Optimization of this condition determined the ideal NaH_2PO_4/K_2HPO_4 concentration to be 1.5 M (at a 1:100 ratio) and the optimal protein concentration to be 5 mg/ml. The cryoprotectant used was 1.45 M Na/K_2HPO_4 with 27 % ethylene glycol.

Diffraction data were collected with a Rigaku R-AXIS IV++ area detector coupled to a MM-002 X-ray generator with Osmic "blue" optics and an Oxford Cryostream 700. Data were processed with Crystal Clear/d*trek (Pflugrath 1999). All data collection statistics are given in Table 2.

Structure determination

CpCBM32C was solved by molecular replacement using the family 32 galactose binding module from the *Micromonospora viridifaciens* sialidase (pdb code 1EUT) (Gaskell *et al*. 1995) as a search model. The program molrep (Vagin and Teplyakov 1997) was able to find one clear rotation/translation solution corresponding to the single molecule in the asymmetric unit. This initial model was corrected by successive rounds of building and ligand was added using COOT (Emsley and Cowtan 2004). Refinement was done using REFMAC (Murshudov *et al*. 1997). Water molecules were added using the REFMAC

Dataset	Galactose	LacNAc	H-trisaccharide
Space group	$P3_221$	$P4_32_12$	$P4_32_12$
Unit cell	a = 43.84, b = 43.84, c = 139.47; α = 90.00, β = 90.00, γ = 120.00	a = 77.38, b = 77.38, c = 73.42; α = 90.00, β = 90.00, γ = 90.00	a = 77.32, b = 77.32, c = 74.16; α = 90.00, β= 90.00, γ = 90.00,
Asymmetric unit contents	Monomer	Monomer	Monomer
Resolution range	20-1.49 (1.53-1.49)	20-2.40 (2.49-2.40)	20-2.30 (2.36-2.30)
Number of measured reflections	158,269	39,833	50,836
Number of unique reflections	23,619	9,130	10,441
R_{merge} (%)	3.5 (33.7)	9.9 (26.0)	11.9 (30.7)
Completeness (%)	89.6 (50.1)	99.6 (99.9)	99.9 (99.5)
$I/\sigma I$	21.6 (3.0)	8.6 (4.0)	7.8 (3.5)
Redundancy	6.70 (3.11)	4.36 (4.42)	4.87 (4.93)
R value (%)	20.6	17.5	18.1
R_{free} value (%)	23.3	25.2	24.0
r.m.s. bond lengths (Å)	0.018	0.019	0.015
r.m.s. bond angles (deg.)	1.544	1.696	1.881
r.m.s. chiral-centre restraints (Å^3)	0.108	0.107	0.088
Overall *B*-factors (Å)			
All atoms	22.40	27.79	30.67
Protein atoms	20.88	26.57	29.08
Water molecules	32.24	34.79	36.66
Sugar	16.38	29.13	45.64
No. residues	143	142	143
No. waters	172	183	218
No. sugar atoms	12	26	36
PDB code	2j1a	2j1e	2j1f

Table 2 (Previous page). X-ray data collection and refinement statistics for CpCBM32C. Values in parentheses are for the highest resolution shells.

implementation of ARP/wARP and inspected visually prior to deposition. The final model lacking waters was used as a starting model to solve the structures of the other CpCBM32C sugar complexes. Initial models were corrected, ligand(s) added and waters added as above. Residue numbering conforms to the numbering in the complete CpGH84C enzymes. Final model statistics are given in Table 2.

2.2.4 Results and Discussion

Carbohydrate-binding properties of CpCBM32C from CpGH84C

Based on its amino acid sequence identity (~30%) with the family 32 galactose-binding module from the *M. viridifaciens* sialidase we postulated that the CBM32 module in *Cp*GH84C is indeed a carbohydrate binding protein. We assessed the binding of various monosaccharides, disaccharides and trisaccharides to CBM32 by qualitative and quantitative UV difference experiments. The addition of D-galactose and GalNAc (*N*-acetyl-D-galactosamine) to CpCBM32C resulted in large perturbations of the UV difference spectra indicative of the involvement of tryptophan in sugar binding (Boraston *et al*. 2001b) (Figure 9A). L-fucose, D-glucose, D-mannose, and GlcNAc (*N*-acetyl-D-glucosamine) were also tested but did not influence the UV absorption of CpCBM32C and thus are unlikely primary ligands of CpCBM32C. Quantitative studies by UV difference titrations showed an affinity of roughly 1 x 10^3 M^{-1} for galactose-based monosaccharides and a preference for β rather than α configured O-methyl-galactose (Figure 9B and Table 3). The presence of the acetamido group of GalNAc did not appear to confer any advantage to binding. CpCBM32C preferred lactose (β-D-galactose-1,4-β-

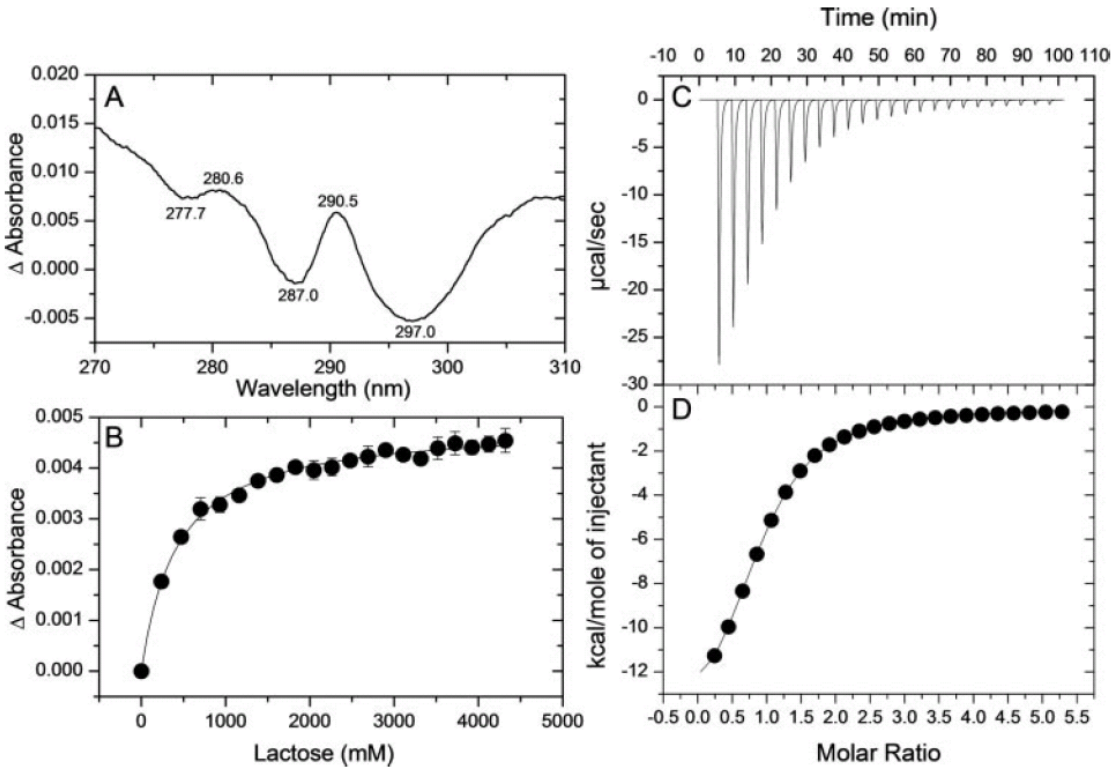


Figure 9. Assessment of CpCBM32C binding. (A) UV difference spectrum produced by the addition of excess galactose to CpCBM32C (60 µM). Wavelengths at which peaks and troughs occur are labeled. (B) An example of UV difference binding isotherm produced by titrating lactose into CpCBM32 (45 µM) in 50 mM Tris-HCl, pH 7.5. The solid line shows the fit to a one-binding site model. (C) and (D), representative ITC experiment of LacNAc titrated into CpCBM32. (C) shows the raw calorimetric data while (D) shows the integrated heats. The solid line shows the fit to a one-binding site model.

Sugar	K_a
	$M^{-1} \times 10^{-3}$
D-Galactose	0.98 (±0.17)
D-GalNAc	0.86 (±0.12)
Methyl-α-D-galactose	0.57 (±0.25)
Methyl-β-D-galactose	1.31 (±0.08)
Lactose	2.49 (±0.06)
LacNAc	9.09 (±2.98)

Table 3. Characterization of CpCBM32C binding by UV difference at 20 °C in 50 mM Tris-HCl, pH 7.5

Sugar	Temperature	n	K_a	ΔH	ΔS	ΔG
	°C		$10^{-4} \times M^{-1}$	*kcal/mol*	*cal/mol/K*	*kcal/mol*
LacNAc	25	0.89 ± 0.0	1.11 (±0.02)	−15.6 (±0.0)	−33.6 (±0.5)	−5.5 (±0.0)
	20[a]	0.89 ± 0.0	1.77 (±0.02)	−15.1 (±0.1)	−32.1 (±0.5)	−5.7 (±0.0)
	17.5[a]	0.89 ± 0.0	2.26 (±0.02)	−14.8 (±0.1)	−30.8 (±0.5)	−5.8 (±0.0)
H-trisaccharide	25[a]	1.0[b]	0.12 (±0.00)	−6.6 (±0.1)	−8.3 (±0.1)	−4.2 (±0.0)

[a] Only single titrations were performed, and errors are those from the data-fitting process. Otherwise, errors are the standard deviations determined from three independent experiments.

[b] This value was fixed as a constant for the data-fitting process.

Table 4 (Previous page). Binding constants and thermodynamic parameters determined by ITC Calorimetry at 25 ºC in 50 mM Tris-HCl, pH 7.5.

D-glucose) and LacNAc over galactose by factors of approximately 2.5 and 10-fold, respectively. The increased affinity of LacNAc (*cf.* lactose) suggested the specific involvement of the 2'-acetamido group of the GlcNAc moiety in binding. Along similar lines, the α-1,2-linked L-fucose of fucosyllactose (α-L-fucosyl-1,2-β-D-galactosyl-1,4-β-D-glucose) substantially reduced the binding affinity to levels below those we could quantify due to limiting quantities of sugar, but did not entirely legislate against binding.

The binding to LacNAc and the type II H-trisaccharide was further investigated by isothermal titration calorimetry (Figure 9C and D; Table 4). The values for LacNAc revealed the enthalpically driven binding process common to protein-carbohydrate interactions. The ΔH of binding was temperature dependent, the analysis of which allowed the approximation of the change in heat capacity (ΔC_p) to be -105 (±5) cal/mol/K. Again, this small negative ΔC_p is consistent with the majority of protein-carbohydrate interactions. The affinity of CpCBM32C for the type II H-trisaccharide was too low to accurately deconvolute the stoichiometry of binding so, on the basis of the LacNAc binding data and X-ray crystallography data (see below), this value was fixed at 1 for the analysis (Turnbull and Daranas 2003). Like with fucosyllactose, the fucose residue of the type II H-trisaccharide is detrimental to binding relative to LacNAc or lactose, though it does not destroy binding. The roughly 1.3 kcal/mol increase in free energy due to the fucose moiety on the H-trisaccharide appears to come by virtue of a substantial enthalpic penalty (~ +9.0 kcal/mol) which is partially offset by a favourable contribution to entropy (~ +25 cal/mol/K) (note: assumption of up to a 25% error in the estimate of stoichiometry due to errors in either the sugar or protein concentration does

change the magnitudes of the calculated thermodynamic penalties but does not change their sign and, thus, qualitative interpretation of the data is unaffected).

The fold of CpCBM32C

CpCBM32C was crystallized in the presence of galactose and its X-ray crystal structure solved by molecular replacement at high resolution (1.49 Å). The fold is that of a β-sandwich comprising a β-sheet of three anti-parallel β-strands opposing a β-sheet of five anti-parallel β-strands (Figure 10). The closest structural neighbour is MvCBM32, which has an RMSD of 1.2 $Å^2$ (over 141 matched Cα) with CpCBM32C. More distantly related are the family 6 and 36 CBMs as well as the *Anguilla* sp. fucolectin. CBM32 coordinates one metal ion, likely calcium, at a site that appears to be conserved in this fold (Figure 10). This ion does not appear to play a direct role in binding carbohydrate as it is quite removed from the carbohydrate binding site; however, it may be structurally important.

CBM32 in complex with carbohydrates

The electron density map of the galactose complex clearly revealed a single bound molecule of galactose. Trp 661 and Phe757 create a relatively hydrophobic pocket which cradles the C6-hydroxylmethyl group. Trp 661 "stacks" against the flat, apolar surface created by carbons 3-6 on the B-face of D-galactose with the O4 pointing away from the aromatic residue (Figure 11**)**. Specificity for the non-reducing end of D-galactose (and presumably GalNAc) is conferred by 3 potential hydrogen bonds to the axial O4 of the sugar from the terminal δO of Asn 695, a terminal guanido nitrogen of Arg 690, and the εN from the imidazole ring of His 658 (Figure 11 and 12). Additional hydrogen bonds

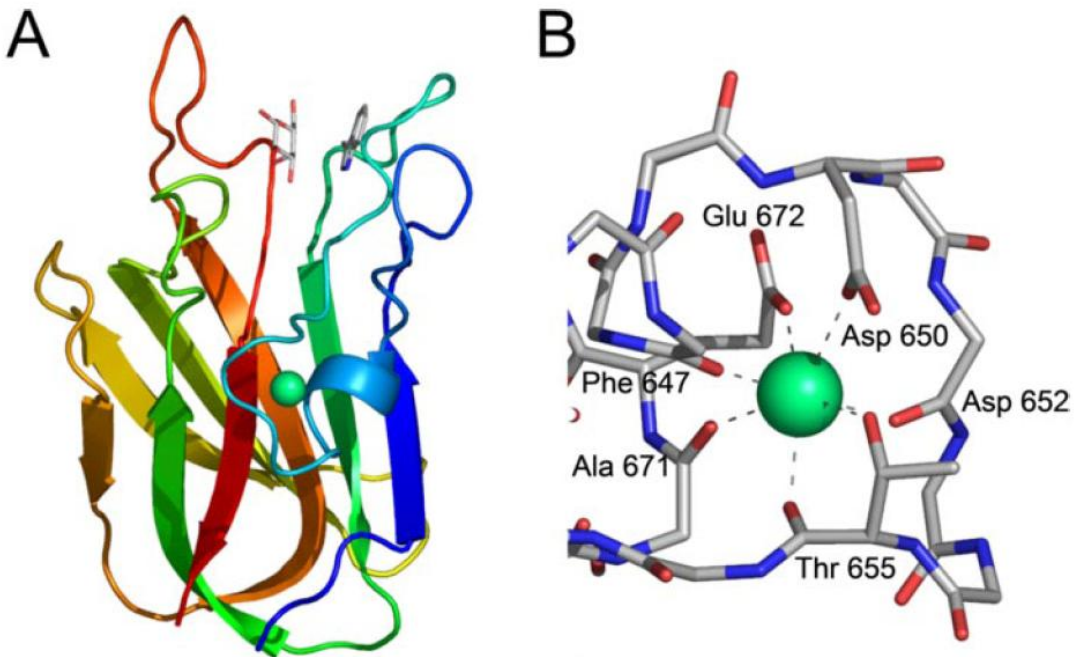


Figure 10. Structural features of CpCBM32C. (A) Secondary structure stereo-view representation of CpCBM32C. The tryptophan in the carbohydrate binding site and the bound galactose are shown in gray stick representation. (B) Coordination of the bound metal ion. Only the side chains involved in binding the metal are shown for clarity. In both panels the metal ion is shown in green

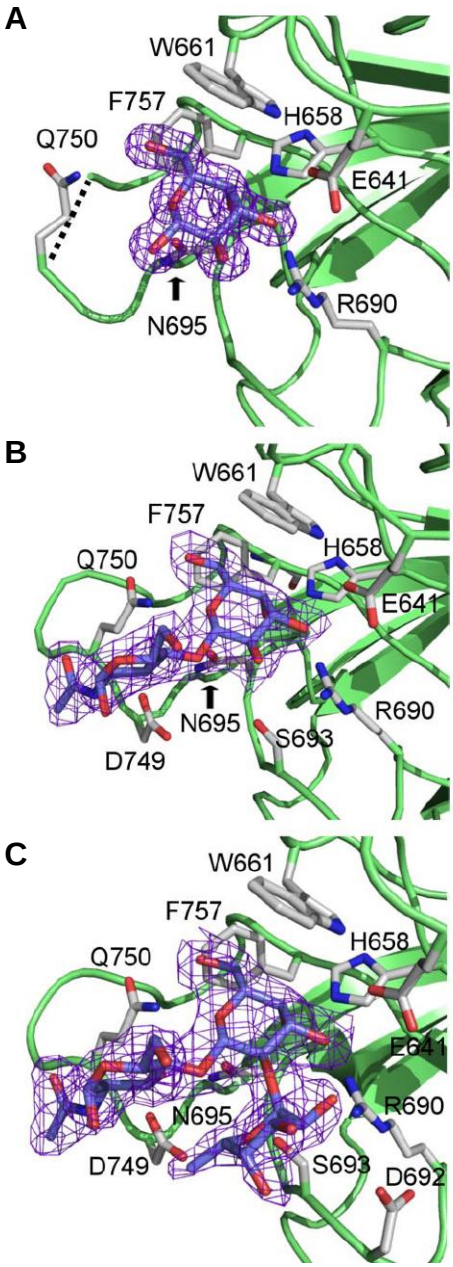


Figure 11. Representative electron density and interaction of CpCBM32C with galactose (A) LacNAc (B) and the type II blood group H-trisaccharide (C). All maps are maximum-likelihood/σ_A -weighted 2Fobs -Fcalc electron density maps contoured at 1σ (0.35, 0.30, and 0.29 electrons/Å^3 for galactose, LacNAc, and the H-trisaccharide, respectively). The disordered loop in the galactose complex is shown with a dashed line. Amino acid side chains involved in binding the sugars are shown in gray stick representation and labeled.

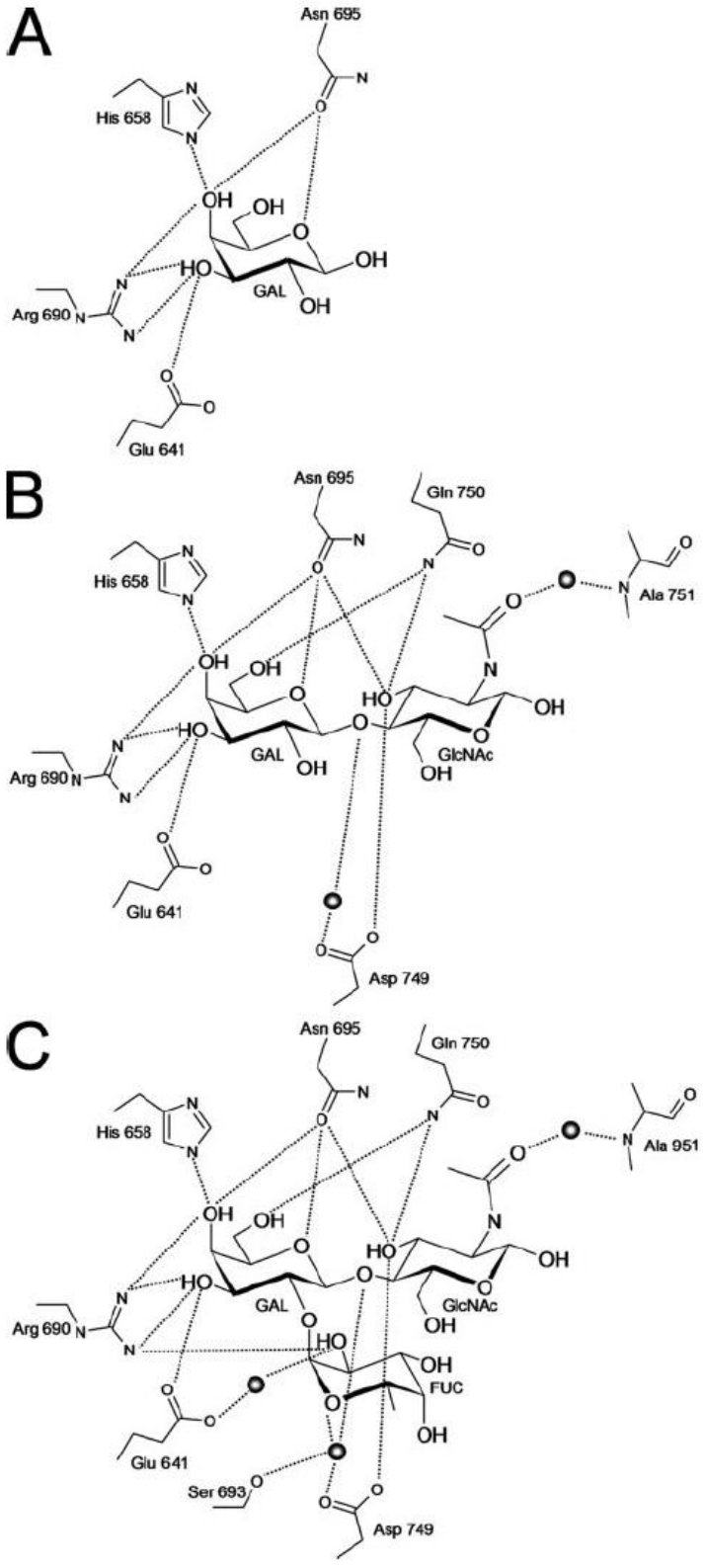


Figure 12. Schematics showing the interactions of CpCBM32C with galactose (A), LacNAc (B) and the type II blood group H-trisaccharide (C). A distance of 3.2 Å was used as the cut-off for determination of significant hydrogen bonds. Water molecules are shown as shaded spheres. For simplicity, not all hydrogen atoms are shown.

are made between the O3 of D-galactose and Arg 690 and Glu 641; the endocyclic oxygen of galactose makes a hydrogen bond with the amide nitrogen of Asn 695. The burial of the galactose O3 legislates against binding to the AB blood group antigens. Examination of the CpCBM32C-galactose complex revealed that the O1, O2, and O6 groups of the bound galactose were solvent exposed, thus hinting at how this protein might accommodate the additional sugar groups of LacNAc, melibiose and the type II H-trisaccharide. We probed the presence of additional subsites by co-crystallizing CpCBM32C with LacNAc and the type II H-trisaccharide – possible biological ligands of CpCBM32C (Figure 11). The binding of LacNAc appears to induce some changes relative to the galactose structure. In the galactose structure the loop comprising residues 750-753 were disordered and, in fact, residues 751 and 752 could not be modelled. In the LacNAc structure, this loop becomes ordered and the side chains of Asp 749 and Gln 750 close in on the sugar to make additional hydrogen bonds with both the galactose and GlcNAc residues (Figure 11 and 12). One terminal δO of Asp 749 is situated to hydrogen bond with the O3 of GlcNAc. The other terminal δO of Asp 749 along with the δO of Ser 693 positions an ordered water for a water mediated hydrogen bond to the oxygen of the β1,4-linked glycosidic bond. The terminal amide nitrogen of Gln 750 is positioned to hydrogen bond with the O6 of galactose and the O3 of GlcNAc. The 2' acetamido group of the GlcNAc does not appear to be positioned to make any additional hydrogen bonds. However, this chemical group does sit above a relatively apolar platform made by the planar conformations of the Asp 749 and Gln 750 side chains and makes a number of van der Waals contacts. A water mediated hydrogen bond occurs between the carbonyl group oxygen of GlcNAc and the main chain N of Ala 751. These

additional interactions are the likely source of the increased affinity of CpCBM32C for LacNAc *vs.* lactose.

The crystal structure of CpCBM32C in complex with the type II blood group H-trisaccharide revealed the interactions between the LacNAc core of this sugar and the protein to be identical to the LacNAc-CpCBM32C interactions (Figure 11). The fucose residue of the type II H-trisaccharide occupied a subsite located above Arg 690 and hydrogen bonding interactions with the protein were limited to two water mediated hydrogen bonds: the first provided by the δO of Asp 749 and the δO of Ser 693 (Figure 12). The same water is also involved in a water mediated hydrogen bond to the glycosidic bond between galactose and GlcNAc. The second water is positioned for potential hydrogen bonding with the O2 of fucose by the εO of Glu 641.

Molecular determinants of CBM32 specificity

The CBM32 module from *Cp*GH84C appears to have three, possibly four, subsites that accommodate the monosaccharide units of oligosaccharide ligands. The primary subsite binds galactose, but is also able to accommodate GalNAc, and likely provides the bulk of the binding free energy and thus provides the "anchor" for the interaction. β-linked substituents on the O1 of this galactose result in improved binding affinity. In the case of LacNAc, where the GlcNAc is β-1,4-linked to the galactose, the affinity improves by an order of magnitude relative to galactose and is influenced positively by the presence of the acetamido group of GlcNAc. Modelling of lacto-N-biose (β-D-galactosyl-1,3-β-D-N-acetylglucosamine; a component of the type I blood group H-trisaccharide antigen) into

the CpCBM32C binding site using LacNAc as a guide suggested that CpCBM32C may also accommodate this sugar with the β-1,3-linked GlcNAc bound in the same secondary subsite as for the GlcNAc of LacNAc (not shown). In the case of lacto-*N*-biose, the C6-hydroxymethyl group of the GlcNAc residue in this disaccharide would be positioned similarly to the acetamido group of LacNAc. Though structurally there is apparently nothing to hinder the binding of lacto-*N*-biose, it is unknown if C6-hydroxymethyl group interaction of this sugar would be energetically equivalent to those of the LacNAc acetamido group. Along an identical line of argument, the core 1 O-glycan (D-galactosyl-1,3-β-*N*-acetyl-D-galactosamine) would be accommodated in a similar manner to lacto-*N*-biose.

The structure of CpCBM32C with the type II H-trisaccharide revealed an additional secondary subsite that accommodates the β-1,2-linked fucosyl residue. However, this appeared to incur a substantial energetic penalty. The thermodynamic values indicated an enthapically unfavourable but partially offsetting entropically favourable contribution from the occupation of this subsite. This signature is consistent with the removal of ordered waters from the protein surface. Indeed, when comparing the binding site water networks in the galactose, LacNAc and type II H-trisaccharide structures, the latter sugar results in the displacement of at least one, possibly two, waters that would otherwise be ordered on the protein surface in the absence of the fucosyl residue. Thus, though the fucosyl residue of the blood group H-antigen is accommodated by the CBM, it is not a preferred feature.

CpCBM32C may have a fourth subsite that accommodates substitutions linked to the O6 of the anchoring galactose. The O6 of galactose is solvent exposed which may permit the linkage of another sugar to that moiety. Sialic acid is found linked α-2,6 to LacNAc, typically at the terminus of a glycan, and to GalNAc, often as an elaboration of O-glycan cores (Figure 5 and 7) (Robbe *et al*. 2004). It is possible that CpCBM32C may accommodate a sialic acid residue within a fourth binding site.

Comparison with other family 32 CBMs

The architecture of the CBM32 carbohydrate-binding site and the constellation of protein-carbohydrate interactions are highly conserved with the *Micromonospora viridifaciens* family 32 CBM, MvCBM32 (Figure 13). Consistent with this is the affinity of CpCBM32C for galactose-based monosaccharides as determined by UV difference titrations (Table 3), which was very similar to the affinity of MvCBM32 for D-galactose and lactose (~$1x10^3$ M^{-1}) (Boraston *et al*. 2003b). This galactose recognition machinery is well conserved among several other Clostridial CBM32s despite having amino acid sequence identity as low as 25% (Figure 13A). The highest degree of sequence variability occurs in the regions of the polypeptides that contribute to the proposed additional binding subsites (Figure 13A). This suggests that among these reasonably closely related CBM32s that galactose is the primary ligand but that different substitutions on the galactose may be optimally accommodated by the different CBMs. This added diversity could complement the varied specificities of the enzymes to which these CBM32 modules are appended. However, precisely how CBM32 ligand preference might complement the enzyme specificity is unclear. For example, the receptor specificity of

the CBM32 from *Cp*GH84C for terminal LacNAc motifs does not match the specificity of *Cp*GH84C, which evidence indicates is an *exo*-β-D-*N*-acetylglucosaminidase (Dennis *et al*. 2006; Ficko-Blean and Boraston 2005; Rao *et al*. 2006)**.**

Using amino acid sequence identity criteria less stringent than a 25% cut-off allows the identification of numerous additional family 32 CBMs in *C. perfringens* and other bacteria. However, at these lower levels of sequence identity the conservation of functional residues, including those involved in galactose binding, is lost and only amino acids contributing to structure remain conserved (not shown). This suggests that the specificity of family 32 CBMs is not necessarily centred on galactose. In section 2.3 we investigate the specificity of a CBM32 module lacking conservation of binding site residues with CpCBM32C.

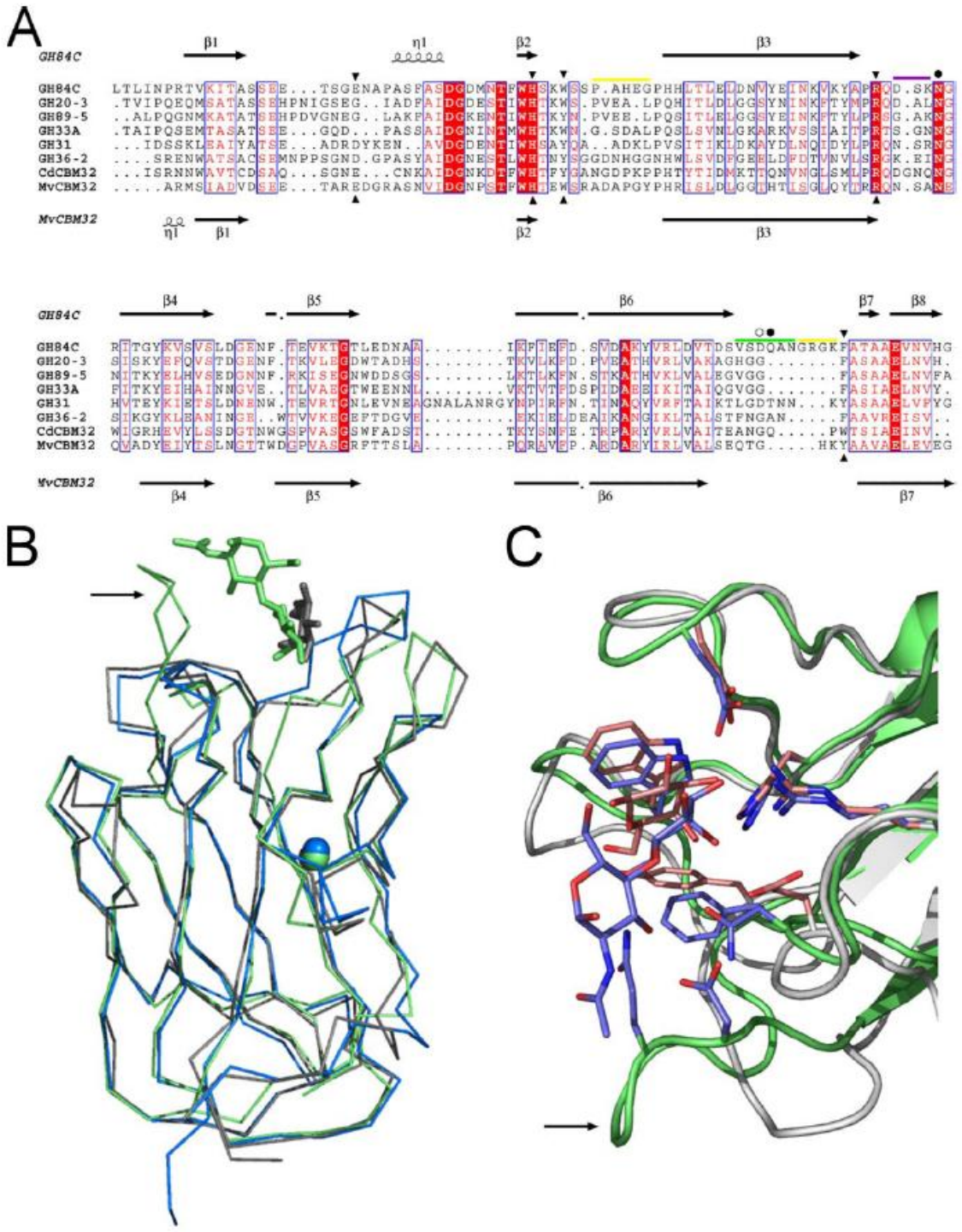

A
GH84C
β1
η1
β2
β3
GH84C LTLINPRTV KITASSEE...TSGENAPASF ASDGDMNTF WHSKWSSP.AHEGPHHLTL ELDNVYEINK VKYAEPQD.SKNVS
GH20-3 .TVIPQEQM SATASSEHPNIGSEG..IADF AIDGDESTI WHTKWS.PVEA.LPQHITL DLGGSYNINK FTYMPRQD.ALNVS
GH89-5 ..ALPQGNM KATATSEHPDVGNEG..LAKF AIDGKENTI WHTKYM.PVEE.LPQSITL ELGGSYEINK FTYLPRSG.AKNVS
GH33A .TAIPQSEM TASATSEE...GQD...PASS AIDGNINTM WHTKWN.G.SDALPQSLSV NLGKARKVSS IAITPRTS.GNNVS
GH31 ...IDSSKL EAIYATSE...ADRDYKEN..AVDGDENTI WHSAYQA..ADKLPVSITI KLDKAYDLNQ IDYLPRQN.SRNVS
GH36-2 ....SRENW ATSACSEMNPPSGND.GPASY AIDGNESTL WHTNYSGGDNHGGNHW LSVDFGEELD FDTVNVLSRG K.EINT
CdCBM32 ...ISRNNW AVTCDSAQ...SGNE...CNKA IDGNKDTFW HTFYGANGDPKPPHT YTIDMKTTQN VNGLSMLPRQ DGNQNT
MvCBM32 .....ARMS IADVDSEE...TAREDGRASN VIDGNPSTF WHTEWSRADAPGYPHR ISLDLGGIHT ISGLQYLPRQ N.SANE
MvCBM32
η1
β1
β2
β3
GH84C
β4
β5
β6
β7
β8
GH84C RTTGYKVSV SLDGENF.TEVKTGTLEDNAA.....IRFIEFD.SVDAKYVRLDVTDSVSDQANGRGKFATAAEVNVHG
GH20-3 SISKYEFQV STDGENF.TKVLEGDWTADHS.....TKVLKFD.PVEATHVRLVAKAGHGG.....FASAAELNVH.
GH89-5 NITKYELHV SEDGNNF.RKISEGNWDDSGS.....LKTLKFN.STKATHVKLVALEGVGG.....FASAAELNVFA
GH33A FITKYEIHA INNGVE..TLVAEGTWEENNL.....VKTVTFDSPIDAEEIKITAIQGVGG.....FASIAELNVY.
GH31 HVTEYKIET STLDNENW.TEVRTGNLEVNEAGNALANRGYNPIRFN.TINAQYVRFTAIKTLGDTNN...KYASAAELVFYG
GH36-2 SIKGYKLEA NINGE..WTVVKEGEFTDGVE.....EKIELDEAIKANGIKLTALSTFNGAN.....FAAVREISV..
CdCBM32 WIGRHEVYL SSDGTNWGSPVASGSWFADST.....TKYSNFE.TRPARYVRLVAITEANGQ.....PWTSIAEINV..
MvCBM32 QVADYEIYT SLNGTTWDGPVASGRFTTSLA.....PQRAVFP.ARDARYIRLVALSEQTG.....HKYAAVAELEVEG
MvCBM32
β4
β5
β6
β7
B
C

Figure 13 (Previous page). Comparison of CpCBM32C with other family 32 CBMs. Panel (A) shows an amino acid sequence alignment of the CpGH84C CBM32 (labeled GH84C) with CBM32 modules found in other *C. perfringens* glycoside hydrolases using a cut-off of 25% amino acid sequence identity. Labels refer to the enzyme family from which the CBMs originate (e.g. GH20 refers to the CBM32 from a *C. perfringens* family 20 glycoside hydrolase). A number appended to the label indicates that more than one CBM32 is found in that enzyme, and the number indicates the position of the CBM from the N terminus relative to the other CBMs. CdCBM32 refers to the CBM from the *C. dendroides* galactose oxidase, and MvCBM32 refers to the CBM from the *M. viridifaciens* sialidase. Above the alignment is the

secondary structure of CpCBM32C. Residues involved in galactose binding are indicated by a triangle; a closed circle indicates residues involved in binding both galactose and GlcNAc (of LacNAc); the open circle indicates the residue involved only in GlcNAc (of LacNAc) binding. Yellow bars above the sequence show regions potentially involved in the formation of a subsite for accommodated sugars linked to the O6 of galactose. Likewise, the green bar shows the O4 subsite, and the purple bar the O2 subsite. The secondary structure of MvCBM32 is shown beneath the alignment. (B) Structural overlap of CpCBM32C (green), MvCBM32 (grey; PDB code 1BZD (Newstead et al. 2005)), and CdCBM32 (blue; 1GOF (Ito et al. 1991)). Bound metal ions are shown as spheres and bound ligands as stick models. (C) Structural overlap of CpCBM32C with MvCBM32. CpCBM32 is depicted in lime with its structural calcium in light orange. Galactose and amino acids involved in binding are in blue stick representation. MvCBM32 is depicted in white, and its structural sodium is colored pale yellow. Its amino acids and bound galactose are shown in pink stick representation. Arrows in B and C indicate the loop discussed in the text.

2.3 Unique Binding Mechanism and Ligand Specificity of a Family 32 CBM from the μ-toxin Secreted by *Clostridium perfringens*

Elizabeth Ficko-Blean and Alisdair Boraston

Biochemistry & Microbiology, University of Victoria, Victoria, British Columbia V8W 3P6, Canada

2.3.1 Abstract

The μ-toxin, secreted by *Clostridium perfringens*, acts on the glycosaminoglycan hyaluronan, a component of our connective tissue. CpCBM32-2 is one of four putative tandem CBM32 modules found in the μ-toxin. This CBM does not share the galatose binding machinery so far described for CBM32 modules and was therefore chosen for study as a potential unique member of this family. CpCBM32-2 demonstrates a similar β-sandwich scaffold to other members of the family. Unusually, the ligand binding loops at the terminus of the β-sandwich are structurally and sequentially divergent from the other characterized members of this family. X-ray crystallographic analysis of the module at resolutions from 1.45-2.0 Å and in complex with two disaccharides reveals that its mode of sugar recognition is quite different from that observed for galactose specific CBM32s. Here we report a unique binding specificity and mode of ligand binding for a family 32 CBM.

2.3.2 Introduction

The lectin-like family 32 CBMs, from various bacteria, have been studied in some detail. These CBM32s bind complex carbohydrate components such as, LacNAc (β-D-galactosyl-1,4-β-D-*N*-acetylglucosamine), the type II H-trisaccharide, D-galactose, GalNAc (*N*-acetyl-D-galactosamine) and D-galacturonic acid (Abbott *et al*. 2008; Abbott *et al*. 2007; Boraston *et al*. 2007; Ficko-Blean and Boraston 2006; Gaskell *et al*. 1995) (See Section 2.2). One consistent feature of these CBMs has been the requirement for a galacto-configured monosaccharide (axial O4) as the primary anchoring sugar, without which binding to the other components of the complex di- or tri- saccharides could not be detected.

Centeno *et al* have studied, using affinity gel electrophoresis, a CBM32 from *Cellvibrio mixtus* that binds weakly to laminarin and pustulan (Centeno *et al*. 2006). This would imply that galactose is not a prerequisite for binding for all members of this family of CBMs. CBM32 modules have a wide range of sequence variability, yet they share a similar 3 dimensional fold (Abbott *et al*. 2008). Given what we know of the variety of ligands already determined for members of this CBM family, this suggests further ligand diversity might yet be determined.

The μ-toxin is a family 84 glycoside hydrolase (GH84) produced by *C. perfringens*, and has previously been identified as a hyaluronidase by Canard *et al* (Canard *et al*. 1994). At the N-terminus is a secretion signal, followed by the family 84 catalytic module, next, in tandem, are four putative family 32 carbohydrate-binding modules (CBM32s). These

CBMs are described here as CpCBM32-1, CpCBM32-2, CpCBM32-3, and CpCBM32-4. C-terminal to the CBMs are three FIVAR repeats and finally a dockerin module involved in forming ultra-tight protein-protein interactions with other *C. perfringens* glycoside hydrolases (Figure 8) (Adams *et al*. 2008). Based on sequence analysis, CpCBM32-2 has been identified as lacking the key residues involved in galactose binding that are common among other characterized members of the CBM32 family, such as those described in Section 2.2, implying a unique mechanism of ligand binding for this member of the family 32 CBMs.

Until now, only the molecular determinants for the galactose binding specificity of family 32 CBMs have been described. Here we describe the characterization of CpCBM32-2 from the *C. perfringens* μ-toxin. This CBM has a unique binding specificity as it binds preferentially to the non-reducing end of gluco-configured oligosaccharides. CpCBM32-2 has a second, more promiscuous, subsite which appears fashioned to accommodate more than one sugar conformer. Our study has revealed a CBM32 with a unique binding specificity and mechanism of carbohydrate recognition which sets it apart from other members of its family and furthers our knowledge of the molecular basis of protein-complex carbohydrate recognition.

2.3.3 Materials and Methods

Materials

Unless stated otherwise, all chemicals were purchased from Sigma.

Cloning CpCBM32-2

C. perfringens ATCC 13124 (Sigma) genomic DNA was used as a template to PCR amplify the gene fragment of the *nagH* gene encoding CpCBM32-2. Primers were designed with engineered restriction sites for recombinant cloning purposes. The forward primer: CAT ATG GCT AGC AAT CCA AGT TTA ATA AGA AGT GAA TCT TGG CAA GTT contains a NheI restriction endonuclease site. The reverse primer: GAA TTC GGA TCC TTA CTC TTT ATT TCC TGC ATT TTC TAA TTC ATC ACT contains a *Bam*HI restriction endonuclease site. The gene fragment (nucleotides 2419-2865 comprising amino acid residues 806-955) was cloned into pET28a(+) (Invitrogen) *via* the *Nhe*I and *Bam*HI restriction sites generating pCBM32-2. The recombinant polypeptide has an N-terminal six-histidine tag followed by a thrombin protease cleavage site and finally the CBM.

CpCBM32-2 Protein Production and Purification

For protein production purposes, pCBM32-2 was transformed into the expression strain *Escherichia coli* BL21 Star (DE3) (Novagen). Luria-Burtani broth (3L) containing 50 μg/mL kanamycin was inoculated with the transformed cells and grown at 37ºC to an OD~1 at 595 nm whereupon protein production was induced using 0.5 mM isopropyl 1-thio-β-D-galactopyranoside. Cultures were incubated overnight at 37ºC. Cells were harvested at 5180 x g for 10 minutes and lysed in Binding Buffer (20 mM Tris-HCl, pH 8.0 and 0.5 mM NaCl) using a using a French Pressure Cell. Lysate was clarified by centrifugation at 48384 x g for 45 minutes and the supernatant loaded onto a Ni-NTA IMAC column (Amersham). Protein was eluted with binding buffer containing

increasing concentration of imidazole (0-500mM). Samples from fractions were run using SDS PAGE and those fractions judged to have purity > 95% were pooled, concentrated and buffer exchanged into 20 mM Tris-HCl, pH 8.0, using a stirred cell concentrator (Amicon).

Protein to be used for crystallography was digested with thrombin (Novagen) overnight to remove the N-terminal six-histidine tag. The sample was further purified by size exclusion chromatography using a Sephacryl S-200 (GE). Fractions containing CpCBM32-2 were pooled and concentrated in a stirred cell concentrator.

In order to produce seleno-methionine labeled CpCBM32-2 the methionine auxotrophic strain *Escherichia coli* 834 (DE3) (Novagen) was transformed with pCBM32-2. Seleno-methionine minimal media (AthenaES) containing 40 mg of L-seleno-methionine and kanamycin to 50 μg/mL, was inoculated with the transformed *E. coli* 834 (DE3) cells. Protein production and purification continued as described above for unlabeled CpCBM32-2.

Determination of Protein Concentration

The concentrations of purified protein was determined by UV absorbance (280 nm) using the calculated molar extinction coefficient of 36130 $M^{-1}cm^{-1}$ (Gasteiger 2005).

Crystallization and Data Collection

The hanging drop vapour diffusion method was used for all crystallization experiments. The optimized condition that gave diffraction quality crystals for the native and seleno-methionine preparation was 0.1M Bis-Tris, pH 6.0, 0.2 M $MgCl_2$, 25% PEG 2000 MME. The crystallization solution supplemented with 10% ethylene glycol was used as a cryoprotectant. CpCBM32-2 was co-crystallized in complex with 5 mM GlcNAc-β-1,3-GalNAc (Toronto Research Chemicals) in 0.1 M Tris pH 8.5, 0.2M NaAcetate, and 30 % PEG 4000. CpCBM32-2 was co-crystallized in the same condition with 5 mM GlcNAc-β-1,2-mannose (*N*-acetyl-D-glucosamine-β-1,2-D-mannose). In both cases the crystallization solution containing 15% ethylene glycol was used as a cryoprotectant. In all cases, crystals were frozen in the cryo-stream.

Data was collected with a Rigaku R-AXIS IV++ area detector coupled to an MM-002 X-ray generator with Osmic "blue" optics and an Oxford Cryostream 700. Data was processed with Crystal Clear/d*trek (Pflugrath 1999). Data collection statistics are given in Table 5.

Structure Determination

ShelxC/D (Schneider and Sheldrick 2002) was used to determine the selenium atom substructure using isomorphous differences between the native and seleno-methionine derivative datasets. Two selenium sites were identified, which were then used for single isomorphous replacement phasing with SHARP (Bricogne *et al*. 2003). There were four

	CpCBM32-2	Selenium derivative	GlcNAc-β-1,2-mannose	GlcNAc-β-1,3-GalNAc
Data collection statistics				
Space Group	$P2_12_12_1$	$P2_12_12_1$	$P3_12_1$	$P3_12_1$
Unit Cell (a,b,c) (Å)	56.1, 61.5, 83.0	56.3, 61.9, 83.0	91.3, 91.3, 132.7	91.2, 91.2, 132.6
Asymmetric Unit Contents	dimer	dimer	tetramer	tetramer
Resolution Range	49.39-1.60 (1.64-1.60)	20.00-1.45 (1.49-1.45)	39.52-2.03 (2.10-2.03)	20.00-2.00 (2.05-2.00)
R-merge (%)	3.9 (25.6)	5.7 (37.5)	9.4 (33.9)	9.9 (30.3)
Completeness (%)	93.8 (86.9)	99.1	99.7 (100.0)	98.6 (99.9)
I/σI	19.2 (3.9)	15.7 (3.2)	11.6 (5.1)	10.4 (4.6)
Redundancy	3.5 (2.7)	7.1 (4.2)	7.0 (7.2)	7.0 (6.7)
Total Reflections	125786	773883	291472	299940
Unique Reflections	36193	109314	41797	43009
Refinement statistics				
R-Value (%)	17.5	20.1	19.4	16.9
R-free Value (%)	21.1	24.4	26.0	21.3
r.m.s. bond lengths (Å)	0.015	0.010	0.010	0.015
r.m.s. bond angles (deg.)	1.469	1.243	1.299	1.508
r.m.s. chiral -centre restraints ($Å^3$)	0.107	0.094	0.079	0.111
Overall B factors ($Å^2$)	18.19	17.79	27.93	19.44
Protein	14.71 (A); 16.92 (B);	14.28 (A); 16.61 (B)	32.68 (A); 24.39 (B); 24.48 (C); 27.80 (D)	21.19 (A); 17.18 (B); 17.07 (C); 17.41 (D)
Water molecules	29.55	28.50	32.44	26.73
Ligands	13.48 (Ca); 14.79(Na); 12.55 (BTB)	13.53 (Ca); 16.51 (Na); 11.53 (BTB)	45.53 (A); 34.36 (B); 26.00 (C); 34.73 (D); 24.40 (Ca)	30.74 (A); 23.39 (B); 16.41 (C); 24.65 (D); 16.5 (Ca); 32.2 (ACT)
No. atoms total	2720	2739	5092	5331
No. protein atoms	1121 (A); 1106 (B)	1127 (A); 1095 (B);	1100 (A); 1161 (B); 1117 (C); 1128 (D);	1100 (A); 1153 (B); 1109 (C); 1136 (D)
No. Waters	475	499	478	704
No. ligand atoms	2 (Ca); 2 (Na); 14 (BTB)	2 (Ca); 2 (Na) 14 (BTB)	26 (A); 26 (B); 26 (C); 26 (D); 4 (Ca)	30 (A); 34 (B); 34 (C); 34 (D); 4 (Ca); 3 (ACT)
Ramachandran statistics				
Most favoured	88.0	88.8	88.2	89.3
Additional allowed	12.0	11.2	11.8	10.7
Disallowed	0	0	0	0

Table 5. X-ray data collection and refinement statistics for CpCBM32-2. Values in parentheses are for the highest resolution shells.

monomers in the asymmetric unit with two Se sites. Refinement of the two sites resulted in a phasing powers of 0.59 and 0.79 for centric and acentric reflections, respectively.

Solvent flattening using DM (Cowtan and Main 1998) with a solvent content of 50 % resulted in a figure-of-merit of 0.84. ARP/wARP (Morris *et al*. 2002; Perrakis *et al*. 1999) was able to build nearly complete models of the two molecules in the asymmetric unit. These were then completed manually using successive rounds of refinement and model building using the program COOT and REFMAC (Emsley and Cowtan 2004; Murshudov *et al*. 1997). Water molecules were added using the ARP/wARP option within REFMAC (Murshudov *et al*. 1997). . Model validation was performed with SFCHECK (Vaguine *et al*. 1999) and PROCHECK(Laskowski 1993). Final model statistics are given in Table 5.

CpCBM32-2 in complex with GlcNAc-β1,3-GalNAc was solved by molecular replacement using Molrep (Vagin and Teplyakov 1997) with the native structure of CpCBM32-2 as a search model. CpCBM32-2 complex underwent successive rounds of model building using Coot and refinement using Refmac. Final data processing statistics are given in Table 5.

UV difference experiments

UV difference scans were performed by taking a baseline scan on protein in solution (~30 μM, 20 mM Tris-HCl, pH 8.0) between 270-300 nm. A second scan was performed after the addition of excess monosaccharide such as: glucose, GlcNAc, L-fucose, galactose,

GalNAc, and mannose. The baseline scan was subtracted from the second scan resulting in a difference scan. The difference scan was examined for the characteristic peaks and troughs associated with binding.

UV difference titrations were done as described previously (Boraston *et al*. 2001b; Ficko-Blean and Boraston 2006). GlcNAc (25 mM) was titrated into CpCBM32-2 (33.5 μM in 20 mM Tris-HCl, pH 8.0). GlcNAc-β-1,3-mannose (25 mM) and GlcNAc-β-1,2-mannose (20 mM) were titrated into CpCBM32-2 (28.5 μM in 20 mM Tris-HCl, pH 8.0) The absorbance was measured between 270 nm and 300 nm. Peak to trough differences were determined by subtracting trough from peak within the spectra for three wavelength pairs (292.1 nm, 278.4 nm; 284.4 nm, 278.4 nm; 292.1 nm, 289.1 nm). This value was plotted as a function of ligand concentration. Data was analyzed using MicroCal Origin software (version 7.0) using a one site binding model. Data is reported as an average of independent experiments and error reported is the standard deviation.

Isothermal Titration Calorimetry

ITC Calorimetry was performed as described previously (Boraston *et al*. 2001a; Ficko-Blean and Boraston 2006) using a VP-ITC (MicroCal, Northampton, MA). Protein was dialyzed in 20 mM Tris-HCl, pH 8.0. Ligand was prepared by weight in buffer saved from dialysis. 25 injections of GlcNAc-β-1,3-GalNAc (2.75 mM) were titrated in 10 μL aliquots into CpCBM32-2 (200.28 μM). Similarly 25 injections of GlcNAc (17.5 mM) were titrated in 10 μL aliquots into CpCBM32-2 (200.28 μM). Heats of dilution, determined by titration of ligand into buffer, were subtracted from the appropriate

experimental run. Due to the low affinities, and thus low C-values (Wiseman *et al*. 1989), the stoichiometries (n) were fixed at 1, justified by the crystal data which shows one binding site in the CBM. Data was fit with a one site binding model using MicroCal Origin software (version 7.0).

2.3.4 Results and Discussion

There has been some suggestion in the literature (Centeno *et al*. 2006) that other members of CBM family 32 might bind to carbohydrates other than galactose, however, this has not been conclusively demonstrated. Using amino acid sequence alignments, CpCBM32-2 from the μ-toxin produced by *C. perfringens* was identified as sharing structural residues associated with the CBM32 family; however, the galactose binding machinery was lacking. Therefore this CBM was chosen for study.

Carbohydrate-Binding Properties of CpCBM32-2

In order to examine the specificity of CpCBM32, qualitative UV difference scans were performed with a variety of sugars. These scans identified CpCBM32-2 as binding D-GlcNAc (Figure 14A) and a very weak signal was seen for D-glucose. Scans were attempted with other monosaccharides, such as fucose, galactose, GalNAc, and mannose, but no change in absorbance upon ligand addition was detected. UV difference titrations on GlcNAc titrated into CpCBM32-2 gave a low association constant (K_a) of 2.22 (± 0.25) x 10^3 M^{-1} (Figure 14B). Glucose, again, gave a weak signal but the affinity was too low to determine accurately using this method. Mucin lining the intestinal walls is one predicted target for the secreted glycoside hydrolases of the enteric bacterium, *C.*

perfringens. The most abundant core structure found in O-linked intestinal mucin is core 3, GlcNAc-β-1,3-GalNAc (Figure 4) (Robbe *et al*. 2004), therefore, we also determined the affinity for this more complex ligand. ITC using GlcNAc-β-1,3-GalNAc resulted in an affinity of 5.91 (± 0.33) x 10^3 M^{-1} (Figure 14D), an approximately 3 fold higher affinity than the ITC determined affinity for GlcNAc (1.88 (± 0.02) x 10^3 M^{-1}) (Figure 2C). Additionally, N-linked glycans are also found in much smaller amount in intestinal mucin (Figure 5) (Ho *et al*. 2003; Klomp *et al*. 1994; McCool *et al*. 1999). The K_a of CpCBM32-2 for chitobiose (GlcNAc-β−1,4-GlcNAc), a motif found at the core of glycans N-linked to proteins, was determined by UV difference titrations to be 1.25 (± 0.13) x 10^3 M^{-1} and thus not substantially different from the CBM's affinity for the monosaccharide GlcNAc. A motif common to N-linked glycans is the disaccharide GlcNAc-β-1,2-mannose. The affinity for this disaccharide was determined by UV difference titrations to be 7.86 (± 0.37) x 10^3 M^{-1} 3.5 fold the affinity of GlcNAc. We were also interested in whether the CBM could accommodate the β-1,3 linkage of this disaccharide and therefore determined the affinity for GlcNAc-β-1,3-mannose to be 3.98 (±1.08) x 10^3 M^{-1} almost 2 fold the affinity for GlcNAc. CpCBM32-2 is able to accommodate both a galactose (axial C4 OH) through a β-1,3 linkage and a mannose (axial C2 OH) through either a β-1,3 or β-1,2 linkage with GlcNAc or even another β-1,4 linked GlcNAc albeit with the same affinity.

Crystal Structure of CpCBM32-2

In order to explore the molecular basis of CpCBM32-2's carbohydrate-binding specificity we determined its structure by x-ray crystallography. A seleno-methionine derivative of

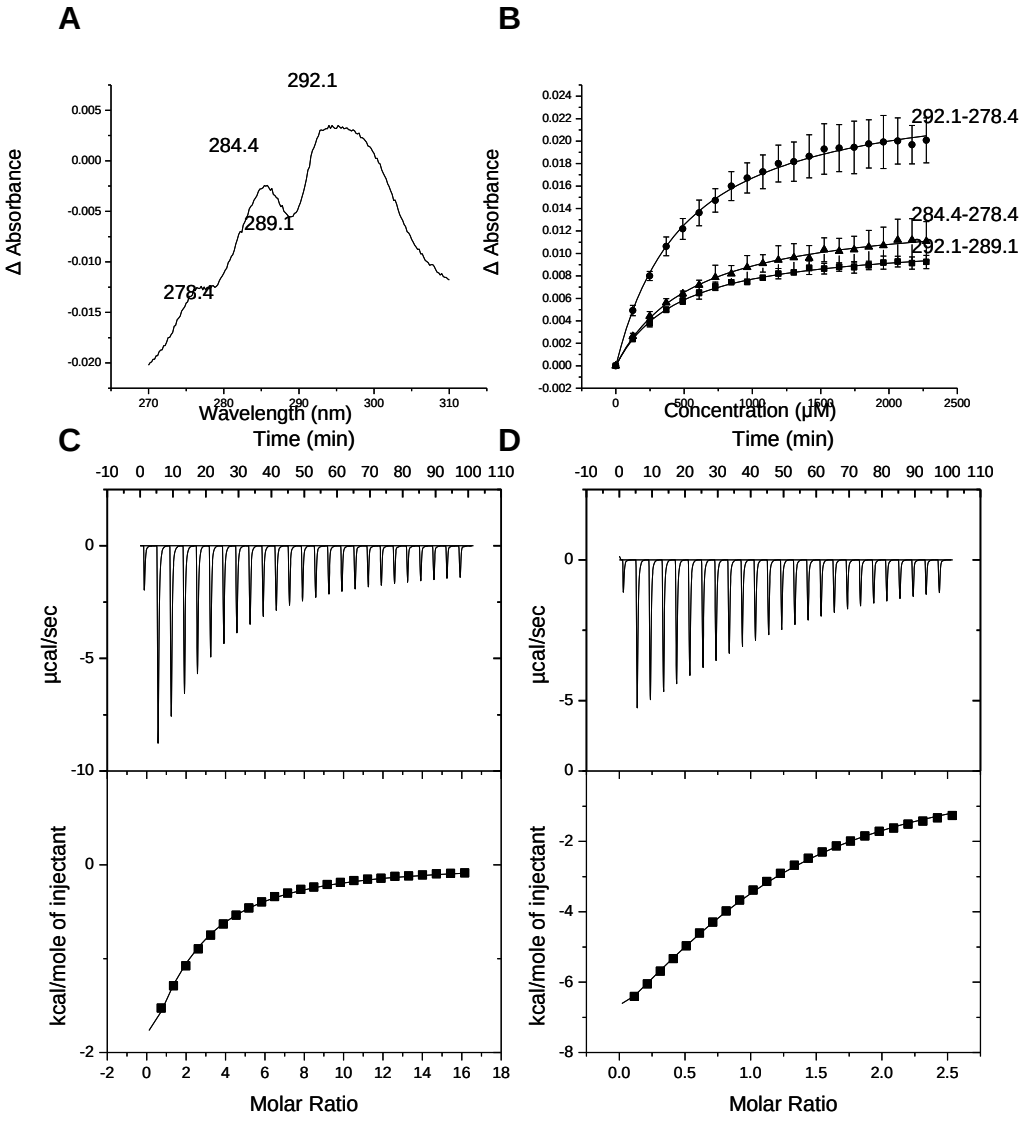


Figure 14. CpCBM32-2 binding studies. (A) UV difference spectrum upon CpCBM32-2 binding to excess GlcNAc. Wavelengths of peaks and troughs are labeled. (B) UV difference isotherms generated by the titration of GlcNAc (25mM) into CpCMB32-2 (33.5 μM). (C) Top- Raw calorimetry data from ITC Calorimetry titrations of GlcNAc (17.51 mM) titrated into CpCBM32-2 (200.28 μM). Bottom- Integrated heats for the calorimetry run after correction for heats of dilution. (D) Top- Raw calorimetry data from ITC Calorimetry titrations of GlcNAc-β-1,3-GalNAc (2.75 mM) titrated into CpCBM32-2 (200.28 μM). Bottom- Integrated heats for the calorimetry run. Non-linear curve fitting was used to fit a one site binding model and the fit is represented by a solid line.

CpCBM32-2 was crystallized to exploit the presence of two methionine residues within the protein for successful structure solution using single isomorphous replacement (SIR) with the native crystals. The seleno-methionine derivative structure of CpCBM32-2 provided the highest resolution structure obtained during these experiments at 1.45 Å.

CpCBM32-2 adopts a β-sandwich and is composed of 5 anti-parallel β-strands overtop 5 anti-parallel β strands (Figure 15). This CBM is clearly structurally related to previously characterized family 32 CBMs such as CpCBM32C from the family 84 β-*N*-acetylglucosaminidase CpGH84C produced by *Clostridium perfringens* (See Section 2.2), CpGH33ACBM32 from the family 33 sialidase GH33A, produced by *Clostridium perfringens*, and MvCBM32 from family 33 sialidase MvNedA produced by *Micromonospora viridifaciens* (Figure 15). (Abbott *et al*. 2008; Abbott *et al*. 2007; Ficko-Blean and Boraston 2006; Gaskell *et al*. 1995). CpCBM32-2 shares rmsd with CpCBM32C, CpGH33ACBM32, and MvCBM32 of 2.53 Å, 2.54 Å, and 2.58 Å respectively. These CBMs have a common structural Ca^{2+} (or in the case of MvCBM32 a structural Na^{+} has been modeled). The ion is distal to the binding site and is unlikely to play a role in binding carbohydrate. For CpCBM32-2 the metal ion is coordinated by 7 oxygen atoms, consistent with Ca^{2+} being the metal ion (Figure 15). Also when Ca^{2+} is modelled in the B factors of Ca^{2+} in are similar to the B factors of the surrounding atoms.

The structure of CpCBM32-2 was solved in complex with GlcNAc-β-1,3-GalNAc and GlcNAc-β-1,2-mannose. Both reveal the binding site of CpCBM32-2 to be specific for the non-reducing end of GlcNAc and clear electron density is apparent for the

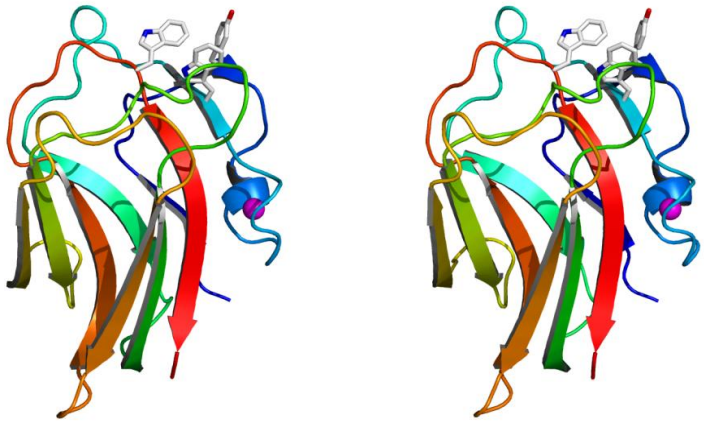


Figure 15. Divergent stereo color ramped cartoon representation of the high-resolution X-ray crystal structure of CpCBM32-2. Aromatic residues in the carbohydrate-binding site are shown in stick representation. The bound metal ion is shown as a magenta sphere.

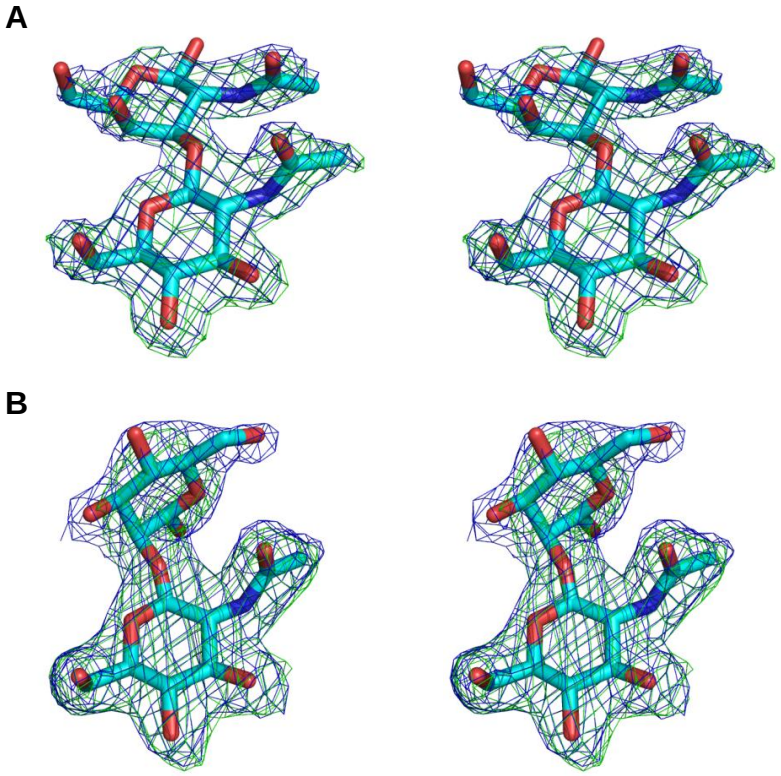


Figure 16. Electron density of (A) GlcNAc-β-1,3-GalNAc and (B) GlcNAc-β-1,2-mannose bound to CpCBM32-2. Sugars are shown in cyan stick representation. Blue mesh represents the maximum-likelihood/σ_A-weighted $2F_{obs}$-F_{calc} electron density maps contoured at 1σ (0.41 electrons/Å^3 for GlcNAc-β-1,3-GalNAc and 0.31 electrons/Å^3 for GlcNAc-β-1,2-mannose). Green mesh represents the maximum-likelihood/σ_A-weighted F_{obs}-F_{calc} electron density maps obtained from refinements with the sugar atoms omitted and contoured at 3σ (0.25 electrons/Å^3 for GlcNAc-β-1,3-GalNAc and 0.18 electrons/Å^3 for GlcNAc-β-1,2-mannose). For simplicity, only the sugars bound to monomer A (of four monomers) in the asymmetric unit are shown. Images are shown in divergent stereo.

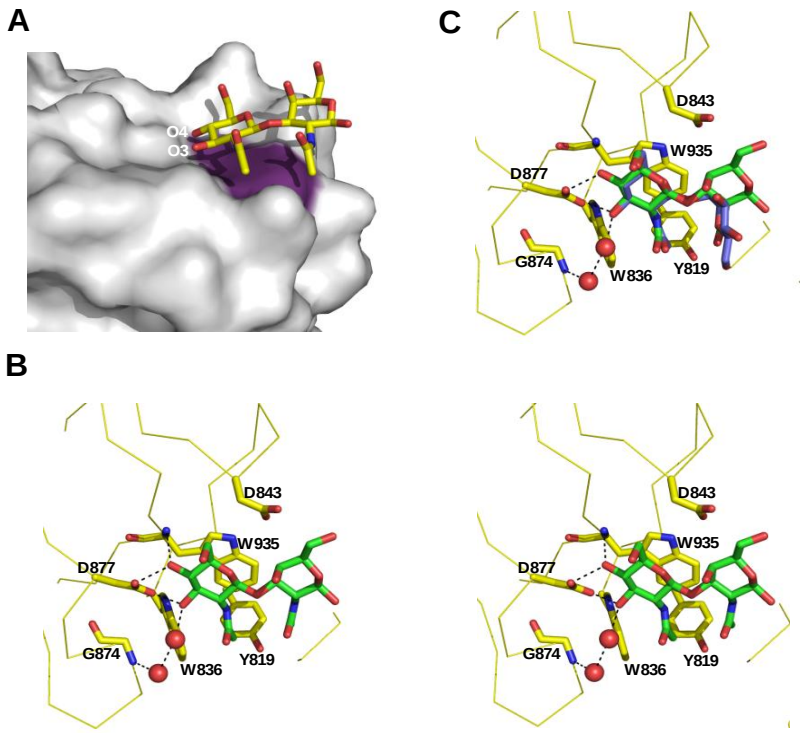


Figure 17. Structural properties of ligand binding by CpCBM32-2. (A) Surface properties of the CpCBM32-2 binding site. The solvent accessible surface of CpCBM32-2 is shown in grey with the surface contributed by a Trp935 colored in purple. The sugar, GlcNAc-β-1,3-GalNAc is shown in yellow stick representation. O3 and O4 of the non-reducing end are labeled for reference. (B) Divergent stereo representation of the interactions made in the active site. Residues in the binding site involved in binding GlcNAc-β–1,3-GalNAc are shown in yellow stick representation and labeled with the backbone of the protein shown in ribbon representation. The sugar is shown in green stick representation. Possible hydrogen bonds are shown as black dotted lines. Relevant waters are shown as red spheres. These interactions are identical for the four molecules in the binding site. (C) Overlap of GlcNAc-β–1,3-GalNAc (green) with GlcNAc-β–1,2-mannose (blue; only from monomer A). Only the side chains from the GlcNAc-β–1,3-GalNAc complex are shown as the conformations of the side chains in both complexes were nearly identical.

disaccharides within the binding site (Figure 16). The requirement of the CBM for chain ends having terminal GlcNAc would also place the associated enzymatic machinery of the μ-toxin near the non-reducing termini of the complex carbohydrate components of the human body.

Interactions with GlcNAc are essentially the same between GlcNAc-β-1,3-GalNAc and GlcNAc-β-1,2-mannose. Trp935 acts as a hydrophobic platform that interacts with the B face of the GlcNAc sugar ring. The specificity for the equatorial O4 of GlcNAc rather than the axial O4 of GalNAc at the first subsite is provided by two hydrogen bonds with the main chain amide N of Trp935 and the carboxylate group of Asp877 (Figure 17). This is two of only four hydrogen bonds that form an interaction with the sugar moiety. The axial O4 of galactose would not be accessible to form a hydrogen bond with Trp935 or Asp977 as the OH group would be pointing out towards the bulk water. Asp877 also provides a hydrogen bond to the O3 of GlcNAc. Finally, one indirect water mediated interaction occurs between Gly874, two water molecules and the O3 of β-D-GlcNAc.

The acetamido group of GlcNAc is oriented so that it tucks into a shallow hydrophobic pocket provided by Tyr819, Trp836 and Trp935. The preferential binding of GlcNAc over glucose can likely be attributed to glucose lacking the acetamido arm of GlcNAc and so is unable to exploit the hydrophobic interaction; the polar nature of the O2 group of glucose may also electrostatically hinder binding at this site.

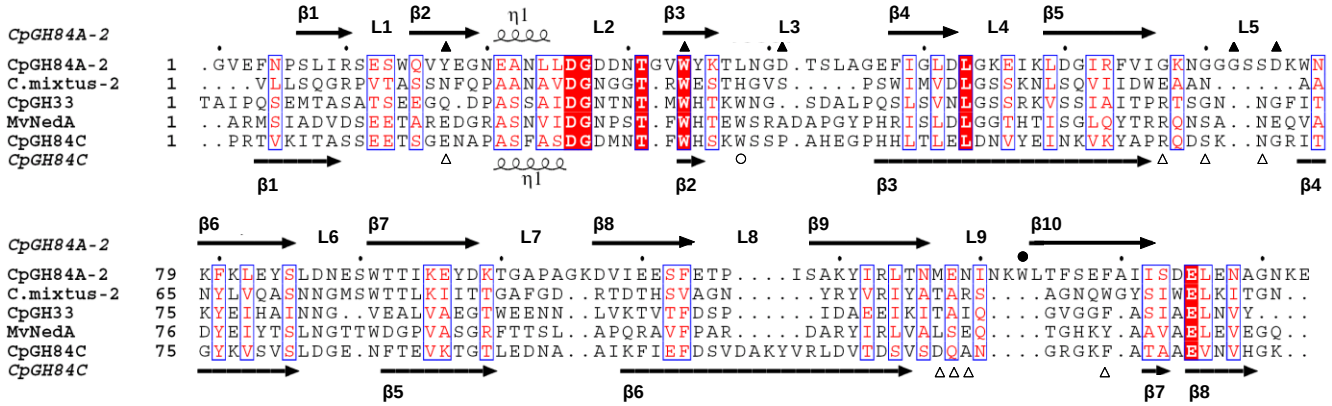
Though GlcNAc-β-1,3-GalNAc binds with 3 fold higher affinity over GlcNAc; the second subsite where GalNAc lies does not appear to contribute greatly to the specificity of the CBM. There are no direct or indirect hydrogen bonds that occur with GalNAc in the second subsite, the interactions that do occur are hydrophobic. The hydrophobic platform provided by Trp935 partially extends over the glycosidic bond between the GlcNAc-β-1,3-GalNAc disaccharide. Asp843 acts as a hydrophobic arm against which the C4-C5-C6 surface edge of the GalNAc ring lies forming van der Waals interactions. The acetamido group of GalNAc nestles against the hydrophobic surface provided by the aromatic residues Tyr819 and Trp935 (Figure 17). Since there appears to be no selection at this site for a specific sugar this suggests the second binding site of CpCBM32-2 could tolerate other monosaccharides.

The structure of CpCBM32-2 in complex with GlcNAc-β-1,2-mannose reveals four protein monomers in the asymmetric unit, each binding slightly differently to the mannose residue. This perhaps reflects the lack of specificity of binding at the second subsite as well as the transient nature of binding. Crystal packing constraints may also play a role in the slight differences seen within the 2nd binding site. In all four protein monomers of the asymmetric unit the planar platform of the carboxylate group of Asp 843 and the arm of its R group lie along the along the glycosidic bond linking mannose to GlcNAc (Figure 17C). Trp 935 also provides a hydrophobic interaction along the β-1,2 glycosidic bond linkage. Monomer A forms no direct or indirect water mediated hydrogen bonds with mannose. However, in monomer B and C a water mediated hydrogen bond occurs between Asp 843, two water molecules, and O1 of mannose.

Finally, Tyr 819 provides a water mediated hydrogen bond to the O6 of mannose in monomer D only. The mannose moiety of GlcNAc-β-1,2-mannose lies perpendicular when compared to the GalNAc moiety of GlcNAc-β-1,3-GalNAc, though GlcNAc binds in the 1st subsite similarly for both disaccharides (Figure 17C). The 2nd subsite of CpCBM32-2 is non-specific and due to its shallowness, can accommodate more than one sugar and more than one disaccharide conformation.

CBM32s have been characterized as type C CBMs and have lectin-like binding pockets designed to bind mono, di- or tri-saccharides (Boraston *et al*. 2004). The β-sandwich provides the structural core of the protein and β-strands 2 and 3 (Figure 18 A) contribute Tyr819 and Trp836 to the binding site of CpCBM32-2. The binding site is found at the terminus where loops L3, L5 and L9 supply the remaining amino acids involved in binding (Figure 18A). This loop region provides the ligand binding site for all of the known CBM32 complexes determined thus far; however, it is here where the structural differences due to amino acid variability within this family of CBMs becomes strongly apparent. In some cases, such as CpCBM32C, the loop region is more extended, forming a deeper cleft as the binding site. CpCBM32-2 is atypical of the CBM32s as the binding pocket is only a shallow indentation at the terminus of the CBM (Figure 17A). The non-reducing end of GlcNAc binds deeper into the pocket of the first subsite while the second subsite resides essentially at the protein surface.

A


CpGH84A-2
β1 L1 β2 ηl L2 β3 L3 β4 L4 β5 L5
CpGH84A-2 1 .GVEFNPSLIRSESWQVYEGNEANLLDGDDNTGVWYKTLNGD.TSLAGEFIGLDLGKEIKLDGIRFVIGKNGGGSSDKWN
C.mixtus-2 1 ....VLLSQGRPVTASSNFQPAANAVDGNGGT.RWESTHGVS.....PSWIMVDLGSSKNLSQVIIDWEAAN.....AA
CpGH33 1 TAIPQSEMTASATSEEGQ.DPASSAIDGNTNTT.MWHTKWNG..SDALPQSLSVNLGSSRKVSSIAITPRTSGN..NGFIT
MvNedA 1 ..ARMSIADVDSEETAREDGRASNVIDGNPST.FWHTEWSRADAPGYPHRISLDLGGTHTISGLQYTRRQNSA..NEQVA
CpGH84C 1 ..PRTVKITASSEETSGENAPASFASDGDMNT.FWHSKWSSP.AHEGPHHLTLELDNVYEINKVKYAPRQDSK..NGRIT
CpGH84C
β1 ηl β2 β3 β4
β6 L6 β7 L7 β8 L8 β9 L9 β10
CpGH84A-2 79 KFKLEYSLDNESWTTIKEYDKTGAPAGKDVIEESFETP....ISAKYIRLTNMENINKWLTFSEFAIISDELENAGNKE
C.mixtus-2 65 NYLVQASNNGMSWTTLKIITTGAFGD..RTDTHSVAGN.....YRYVRIYATARS....AGNQWGYSIWELKITGN..
CpGH33 75 KYEIHAINNG..VEALVAEGTWEENN..LVKTVTFDSP.....IDAEEIKITAIQ....GVGGF.ASIAELNVY....
MvNedA 76 DYEIYTSLNGTTWDGPVASGRFTTSL..APQRAVFPAR.....DARYIRLVALSEQ....TGHKY.AAVAELEVEGQ..
CpGH84C 75 GYKVSVSLDGE.NFTEVKTGTLEDNA..AIKFIEFDSVDAKYVRLDVTDSVSDQAN....GRGKF.ATAAEVNVHGK..
CpGH84C
β5 β6 β7 β8

B

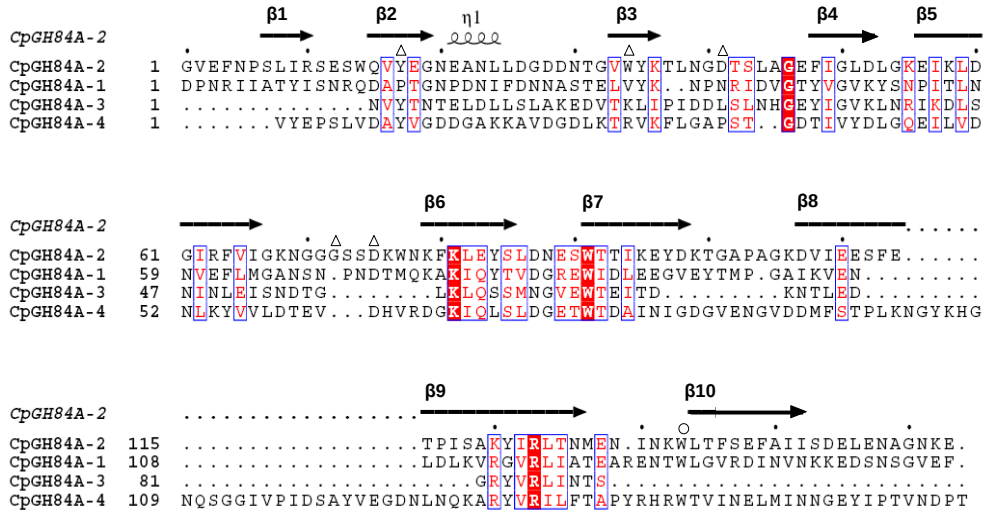

CpGH84A-2
β1 β2 ηl β3 β4 β5
CpGH84A-2 1 GVEFNPSLIRSESWQVYEGNEANLLDGDDNTGVWYKTLNGDTSLAGEFIGLDLGKEIKLD
CpGH84A-1 1 DPNRIIATYISNRQDAPTGNPDNIFDNNASTELVYK..NPNRIDVGTYVGVKYSNPITLN
CpGH84A-3 1 .............NVYTNTELDLLSLAKEDVTKLIPIDDLSLNHGEYIGVKLNRIKDLS
CpGH84A-4 1 .......VYEPSLVDAYVGDDGAKKAVDGDLKTRVKFLGAPST..GDTIVYDLGQEILVD
CpGH84A-2
β6 β7 β8
CpGH84A-2 61 GIRFVIGKNGGGSSDKWNKFKLEYSLDNESWTTIKEYDKTGAPAGKDVIEESFE.....
CpGH84A-1 59 NVEFLMGANSN.PNDTMQKAKIQYTVDGREWIDLEEGVEYTMP.GAIKVEN.....
CpGH84A-3 47 NINLEISNDTG.....LKLQSSMNGVEWTEITD.....KNTLED.....
CpGH84A-4 52 NLKYVVLDTEV...DHVRDGKIQLSLDGETWTDAINIGDGVENGVDDMFSTPLKNGYKHG
CpGH84A-2
β9 β10
CpGH84A-2 .....
CpGH84A-2 115 .....TPISAKYIRLTNMEN.INKWLTFSEFAIISDELENAGNKE.
CpGH84A-1 108 .....LDLKVRGVRLIATEARENTWLGVRDINVNKKEDSNSGVEF.
CpGH84A-3 81 .....GRYVRLINTS.....
CpGH84A-4 109 NQSGGIVPIDSAYVEGDNLNQKARYVRILFTAPYRHRWTVINELMINNGEYIPTVNDPT

Figure 18. Sequence comparison of CpCBM32-2 with (A) the other CBM32 modules that have been structurally characterized in complex with ligand and CmCBM32-2 from *Cellvibiro mixtus*. Labels refer to the glycoside hydrolase family from which the CBM32 originates. CpCBM32-2 is from CpGH84A (μ-toxin) produced by *C. perfringens*. CmCBM32-2 is from cellulase C produced by *Cellvibrio mixtus*. CpGH33 is a sialidase produced by *C. perfringens*. CpGH84C is a β-*N*-acetylglucosaminidase produced by *C. perfringens*. MvNedA is a sialidase produced by *Micromonospora viridifaciens.* The secondary structure is represented above (CpCBM32-2) and below (CpCBM32C) Closed circles represent the tryptophan residue involved in providing a hydrophobic platform for GlcNAc in CpCBM32-2. Closed triangles represent the other amino acid residues involved in coordinating GlcNAc-β-1,3-GalNAc. Open circles represent the tryptophan residue involved in providing a hydrophobic platform for galactose in CpCBM32C. Open triangles represent the other amino acid residues involved in coordinating LacNAc by CpCBM32C. (B) The predicted CpCBM32 modules from the μ-toxin: CpCBM32-1, CpCBM32-2, CpCBM32-3, CpCBM32-4. Closed circles represent the tryptophan residue involved in providing a hydrophobic platform for GlcNAc in CpCBM32-2. Closed triangles represent the other amino acid residues involved in coordinating GlcNAc-β-1,3-GalNAc.

This promiscuity may play an important role for these CBMs as their role in the enzymes as a whole is likely to target the catalytic module to the diverse glycans of mucin, on the surface of tissues, or to glycan components of the extracellular matrix. In the case of CpCBM32-2, specificity is driven by the first subsite with its requirement for an anchoring GlcNAc residue, without which there is no appreciable binding to sugars in the accompanying subsite. Promiscuity of binding occurs in the second subsite. The ability of CpCBM32-2 to accommodate both GalNAc and mannose in its second subsite suggests that the μ-toxin, characterized as a hyaluronidase, may also act on substrates other than hyaluronan. The second subsite of CpCBM32-2 does not provide any consistent hydrogen bonds that might help drive specificity for a particular sugar conformer or linkage. Modeling of glucose and GlcNAc into the second subsite through the β-1,2 and β-1,3 linkages indicate it might also be accommodated. The second subsite seems to have few restrictions on possible sugar residues which could be accommodated.

Structural relationship to other CBM32s.

The binding site for all CBM32 modules thus far has been found at the terminal loop region of the β-sandwich. A major determinant in the location of the first subsite is the hydrophobic platform provided by a solvent exposed tryptophan residue which lies against the B face the monosaccharide. At this first subsite of CpCBM32-2 there is a loop that contributes to the specificity for gluco-configured sugars providing a hydrogen bond between the amide N of Trp935 and the O4 of GlcNAc. Notably, in the three structurally characterized galactose binding CBMs, CpCBM32C from CpGH84C, CpCBM32 from CpGH33 and MvCBM32 from MvNedA, this subsite is sterically

blocked by either a phenylalanine or tyrosine residue which prevents sugar binding. Therefore, within the CBM32 family the positioning of the first subsite is influenced by loop orientation and the composition of amino acids within the loops (Figure 19). Loop orientation and amino acid composition within these loops vary significantly and is the major source of structural diversity in this family. The binding site for GlcNAc-β-1,3-GalNAc, though located similarly at the terminal variable loop within the CBM32s, does not share structural or sequential overlap with the binding sites of the other structurally characterized galactose binding CBM32s described within this paper (Figure 18). In fact none of the amino acids that are involved in carbohydrate binding between CpCBM32-2 and CpCBM32C are conserved between the two family 32 CBMs. CpCBM32-2 demonstrates a new ligand binding site to accommodate gluco-configured ligands (Figure 19). Interestingly, none of the binding site residues involved in binding GlcNAc are conserved in CmCBM32-2, which was described by Centeno *et al* as binding weakly to laminarin and pustulan.

Sequence alignments of CpCBM32-2 with the other three CBM32 modules (CpCBM32-1, CpCBM32-3, CpCBM32-4) within the μ-toxin indicate that though related, the ligand binding machinery is not shared with CpCBM32-2 (Figure 18B). This is particularly exemplified with CpCBM32-4 which has a large insert not found in the other CBM modules, though its effect on binding remains to be elucidated.

The role of CpCBM32-2 (and the three other putative CBM32 modules) on the activity of the μ-toxin remains to be determined. This would provide further understanding of the role of protein-complex carbohydrate interactions on carbohydrate hydrolysis.

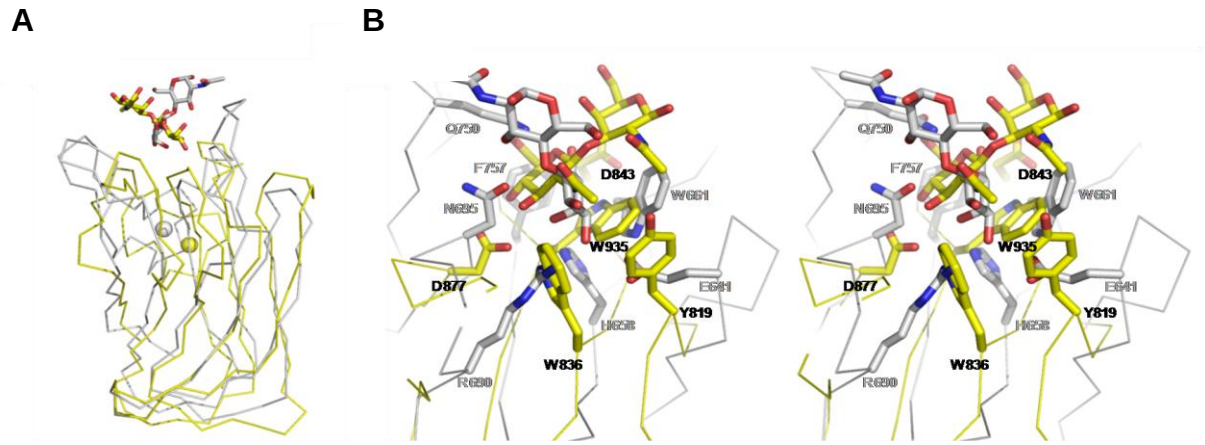


Figure 19. Structural overlay of CpCBM32-2 (yellow) and CpCBM32C (grey). Bound sugars are shown in stick representation and metal ions as spheres. (B) The overlapped active sites of CpCBM32-2 (yellow) and CpCBM32C (grey) shown in divergent stereo. Bound ligands and residues that interact with these are shown in stick representation. Residues in CpCBM32-2 are labeled in black and those in CpCBM32C labeled in grey.

2.4 Portrait of an Enzyme: a Complete Structural Analysis of a Multi-Modular α-*N*-Acetylglucosaminidase from *Clostridium perfringens*

Elizabeth Ficko-Blean[1], Katie J. Gregg[1], Jarrett J. Adams[2], Jan-Hendrik Hehemann[3], Mirjam Czjzek[3], Steven J. Smith[2] and Alisdair B. Boraston[1]

Adapted from: Portrait of an enzyme: A complete structural analysis of a multi-modular beta-*N*-acetylglucosaminidase from *Clostridium perfringens*. J Biol Chem. 2009 Feb 4.

[1]Biochemistry & Microbiology, University of Victoria, Victoria, British Columbia V8W 3P6, Canada
[2]Department of Biochemistry, Queen's University, Kingston, Ontario, K7L 3N6, Canada
[3]Station Biologique de Roscoff, Marine Plants and Biomolecules, CNRS-UPMC Paris6, BP74, 29682 Roscoff, France

2.4.1 Abstract

Common features of the extracellular carbohydrate-active virulence factors involved in host-pathogen interactions are their large sizes and modular complexities. This has made them recalcitrant to structural analysis and therefore our understanding of the importance of modularity in these vital proteins is lagging. *C. perfringens* is a prevalent human pathogen that harbours a wide array of large, extracellular carbohydrate-active enzymes and is an excellent and relevant model system to approach this problem. Here, we describe the complete structure of *C. perfringens* CpGH84C (NagJ), a 1001 amino acid multi-modular homolog of the *C. perfringens* μ-toxin, which was determined using a combination of small angle X-ray scattering and X-ray crystallography. The resulting structure reveals unprecedented insight into how catalysis, carbohydrate-specific adherence, and the formation of molecular complexes with other enzymes *via* an ultra-tight protein-protein interaction are spatially coordinated in an enzyme involved in a host-pathogen interaction.

2.4.2 Introduction

A common feature of microbial carbohydrate-active enzymes involved in processing eukaryotic glycans is their structural complexity; such proteins are frequently very large (over 1000 amino acids) and can comprise numerous modules and domains (Abbott *et al*. 2008; Adams *et al*. 2008; Chitayat *et al*. 2008b; Ficko-Blean and Boraston 2006; Thobhani *et al*. 2003). *C. perfringens*, a myonecrotic ("flesh-eating") bacterial pathogen and a leading cause of food-borne gastrointestinal illness, deploys numerous protein exotoxins that work together to increase the pathogen's virulence (Rood and Cole 1991; Smedley *et al*. 2004). Among its exotoxin proteins are a considerable battery of large extracellular carbohydrate-active enzymes (Figure 8), including the hyaluronidase μ-toxin, which destroys the polysaccharide hyaluronan, and the large sialidases (NanJ and NanI), which remove terminal sialic acid sugars and enhance the lethal cytolytic phospholipase activity of the α-toxin (Boraston *et al*. 2007; Canard *et al*. 1994; Flores-Diaz *et al*. 2005) (See section 2.2). The individual modules of such carbohydrate-active enzymes perform a variety of functions, most commonly catalysis, protein-carbohydrate interactions or protein-protein interactions, which contribute to the overall function and efficiency of the protein. However, because of their large size and flexibility, properties that make them recalcitrant to structural analyses, and their multifunctional natures, such large proteins have required a reductionist approach to their study whereby individual modules are heterologously produced for thorough structural and functional analysis. This has provided considerable insight into the individual functions of the modules. However, how the functions of the modules are spatially coordinated to cooperatively influence the overall activity of the full-length enzymes - knowledge that would have

considerable biological relevance - is largely unknown as the complete structures of large carbohydrate-active bacterial virulence factors, for the most part, resist determination.

The considerable number of large and multimodular carbohydrate-active enzymes produced by *C. perfringens* makes this organism an excellent and relevant model system for the study of complex carbohydrate-active enzymes involved in bacterial pathogenesis. CpGH84C (NagJ) from *C. perfringens* is a homolog of the μ-toxin that comprises 1001 amino acids and four distinct modules: an N-terminal catalytic module followed by a family 32 carbohydrate-binding module (CBM32), a cohesin module (Coh), and a C-terminal fibronectin type III module (FN3) (Figure 8). The structural and biochemical properties of the isolated catalytic module, CBM32, and Coh have been described separately (Adams *et al*. 2008; Chitayat *et al*. 2008b; Ficko-Blean and Boraston 2006; Rao *et al*. 2006). These informative studies have revealed that the catalytic module is similar to human O-GlcNAcase and functions as an *exo*-β-D-*N*-acetylglucosaminidase while the CBM32 preferentially recognizes the non-reducing terminus of *N*-acetyllactosamine bearing carbohydrate receptors (Ficko-Blean and Boraston 2006; Rao *et al*. 2006). The Coh module is perhaps the most unique component as it functions to recognize and bind ultra-tightly to dockerin modules (Doc), such as that present at the C-terminus of the μ-toxin, and plays a role in forming higher-order complexes with other large *C. perfringens* exotoxins, all of which are thought to contribute to the virulence of this bacterium (Adams *et al*. 2008). However, the overall three-dimensional structure and therefore spatial distribution of CpGH84C's cognate modules remains unknown, as does the structure and function of the FN3 module. This information would provide

considerable insight into the structural coordination of modular functions in this protein and provide a powerful model for understanding modularity in *C. perfringens'* carbohydrate-active enzymes as well as carbohydrate-active virulence factors in other bacterial pathogens, which also frequently display modular complexity.

Intact CpGH84C proved intractable to heterologous production and *in vitro* analysis; however, a less conventional approach that combined X-ray crystallography and small-angle X-ray scattering (SAXS) was applied in a "dissect-and-build" approach to generate a solution structure model of the complete enzyme. The reconstructed architecture of CpGH84C in complex with a Doc-containing fragment of the μ-toxin reveals unprecedented insight into how the recognition of substrate by the catalytic module, interaction of a CBM with a carbohydrate receptor, and recruitment of other carbohydrate-active enzymes through a high affinity protein-protein interaction are spatially coordinated in an enzyme involved in a host-pathogen interaction. Furthermore, the methodology employed here should be broadly applicable to a range of large and flexible multi-modular proteins.

2.4.3 Materials and Methods

Materials

Unless otherwise stated, chemicals, carbohydrates, glycoproteins, and polysaccharides were purchased from Sigma.

Cloning, protein production and purification

The DNA fragments encoding the modules and combinations thereof from CpGH84C were amplified by PCR from *C. perfringens* genomic DNA (ATCC 13124) and, unless previously cloned, cloned into pET 28a(+) using described methods and the primers listed in Tables 6 and 7 (Ficko-Blean and Boraston 2005; Ficko-Blean and Boraston 2006; Ficko-Blean *et al*. 2008). With the exception of Coh-FN3 and FIVAR-Doc, all constructs encoded polypeptides comprising a hexa-histidine tag fused to the desired module(s) by an N-terminal thrombin protease cleavage site. Coh-FN3 and FIVAR-Doc had C-terminal, non-cleavable hexa-histidine tags.

The CpGH84C catalytic module and FIVAR-Doc were produced, purified by immobilized metal affinity chromatography, buffer exchanged and concentrated using previously described methods (Ficko-Blean and Boraston 2005) (Adams *et al*. 2008; Chitayat *et al*. 2008a). This process was followed by further purification by size exclusion chromatography using a Sephacryl S-200 size exclusion column (Amersham). The CBM32-Coh:FIVAR-Doc complex was prepared by mixing the two polypeptides with the FIVAR-Doc in ~1.5 fold molar excess. The complex was purified using a Sephacryl S-200 size exclusion column. Protein concentrations were determined using UV absorbance at 280 nm and calculated extinction coefficients (Gasteiger *et al*. 2003).

Name	Nucleotide sequence
GH84F	CAT ATG GCT AGC GTA GGA CCT AAA ACT GGG
GH84R	GAA TTC GGA TCC TTA TAT TAA TGT TAA ATC AAA ACT TAA AGC
CBMF	CAT ATG GCT AGC AAT CCA AGA ACA GTA AAG
CBMR	GAA TTC GGA TCC TTA TCC ATG AAC ATT AAC CTC TGC
CohF	GGG AAT TCC ATA GGT TCA TGG AAA GTT AAA AGA AGC T
CohR	GAA TTC GGA TCC TTA ACT AGT TCC TTC TAT TAT ATT AAC
FN3R	CCG CTC GAG TCT TGC AGT TCT TAA AGT TAG GC

Table 6. Primers used in this study.

Construct	Amino acid boundaries	Primers used to amplify gene fragment
GH84C Catalytic	31-624	GH84F and GH84R
GH84C-CBM32	31-767	GH84F and CBMR
GH84C-CBM32-Coh	31-909	GH84F and CohR
CBM32-Coh	625-909	CBMF and CohR
Coh-FN3	765-1001	CohF and FN3R

Table 7. CpGH84C modular constructs made for this study.

X-ray crystallography

Crystallization and data collection procedures have been described previously for the CpGH84C catalytic module (Ficko-Blean and Boraston 2005). The hexa-histidine tag was removed from the CpGH84C-CBM32 construct by incubation with thrombin protease for 16 h. The mixture was separated on a Sephacryl S-200 size exclusion column in 20 mM Tris-HCl, pH 8.0. All protein constructs were concentrated to 10-25 mg/ml for crystallization trials. Crystallization was performed using the hanging drop vapour diffusion method. CpGH84C-CBM32 was crystallized in 0.1 M Tris-HCl, pH 7.5, 0.1 M $MgCl_2$, 15% (w/v) polyethylene glycol (PEG) 4000. These crystals were cryoprotected in the crystallization solution supplemented with 20 % ethylene glycol (v/v). The Coh-FN3 modular pair was crystallized in 0.1 M sodium acetate, pH 3.5, 4% (w/v) PEG 8000. These crystals were cryoprotected in 20% (w/v) PEG 400 with mother liquor. Diffraction data for CpGH84C-CBM32 was collected with a Rigaku R-AXIS IV++ area detector coupled to an MM-002 X-ray generator with Osmic "blue" optics and an Oxford Cryostream 700 at 113 K. Diffraction data for Coh-FN3 were collected at 100 K on beamline X6A at the NSLS (Brookhaven National Laboratories). All data were processed with Crystal Clear/d*trek (Pflugrath 1999). Data collection statistics are given in Table 8.

The structure of the CpGH84C catalytic module was solved by molecular replacement using PHASER (Zwart *et al*. 2008) and the preliminary coordinates of the family 84 glycoside hydrolase from *Bacteroides thetaiotaomicron* as a search model, kindly provided by Dr. Gideon Davies prior to deposition (PDB accession 2J47), (Dennis *et al*. 2006). Three iterations of manual model correction using COOT

	CpGH84C-CBM32	CpGH84C Catalytic	Coh-FN3
Data collection			
Space group	$P2_1$	$I2_12_12_1$	C2
Cell dimensions			
a, *b*, *c* (Å)	46.85, 69.54, 136.78	130.39, 150.05, 155.43	148.25, 48.64, 36.90
α, β, γ (°)	90.00, 95.78, 90.00	90.00, 90.00, 90.00	90.00, 96.19, 90.00
Resolution (Å)	20.00-3.30 (3.39-3.30)	20.00-2.10 (2.15-2.10)	73.72-1.80 (1.85-1.80)
R_{merge}	0.114 (0.287)	0.103 (0.397)	0.051 (0.495)
$I / \sigma I$	4.4 (2.2)	12.4 (4.8)	30.2 (1.9)
Completeness (%)	92.8 (94.7)	99.9 (100.0)	94.0 (94.5)
Redundancy	2.5 (2.5)	16.1 (15.7)	3.7 (3.5)
Refinement			
No. reflections	31497	1384042	27684
R_{work} / R_{free}	0.327/0.393	0.195/0.255	0.224/0.294
No. atoms			
Protein	5677	4644 (A); 4632 (B)	1750
Ligands	3 (Ca)	4 (Ca); 3 (Na); 25 (CAC)	4 (ACT)
Water	71	1195	318
B-factors			
Protein	51.8	61.2	28.3
Ligands	73.6 (Ca)	44.3 (Ca); 22.0 (Na); 71.7 (CAC)	30.2 (ACT)
Water	21.2	34.3	38.8
R.m.s. deviations			
Bond lengths (Å)	0.006	0.010	0.018
Bond angles (°)	0.934	1.389	1.732

Table 8. X-ray crystal diffraction data collection and refinement statistics for CpGH84C-CBM, CpGH84C catalytic and Cohesin-FN3. Values in parentheses are for the highest resolution shells.

(Emsley and Cowtan 2004), refinement with REFMAC (Murshudov *et al*. 1997), and solvent flattening with two-fold NCS averaging using DM (Cowtan 1994) were required to obtain a model of roughly 60% completeness with partially built side chains. Phases from this process were then sufficient for ARP/wARP (Morris *et al*. 2002) to build a virtually complete model with docked side chains, which was then manually completed using COOT and refinement with REFMAC. Waters were added using the ARP/wARP option in REFMAC.

The CpGH84C-CBM32 modular pair structure was solved by molecular replacement by first running MOLREP (Vagin 1997) using the CpGH84C catalytic module as a search model. MOLREP was run a second time to place the CBM32 model (PDB accession 2J7M) in the asymmetric unit (See Section 2.2). Successive rounds of model correction were done using COOT and refinement with REFMAC. Waters were added manually.

The structure of the Coh-FN3 modular pair was determined by molecular replacement using PHASER and the coordinates of the isolated Coh module (PDB accession 2O4E) as a search model (Chitayat *et al*. 2008b). Though the molecular replacement solution contained only ~50% of the asymmetric unit contents, the initial phases from restrained refinement with REFMAC were of sufficient quality for ARP/wARP to build a complete model, including the FN3 module, with docked side chains. Model correction was done manually using COOT and refinement with REFMAC. Waters were added using the ARP/wARP option in REFMAC.

In all cases, 5% of the reflections were flagged as "free" to monitor refinement procedures and judge model quality (Brunger 1993). Model validation was performed with SFCHECK (Vaguine *et al*. 1999) and PROCHECK (Roman A Laskowski 1993). All model statistics are shown in Table 9. The coordinates and structure factors for the CpGH84C catalytic module, CpGH84C-CBM32, and Coh-FN3 have been deposited with the pdb codes of 2v5c, 2v5d, and 2w1n, respectively.

Small Angle X-Ray Scattering

Small angle X-ray scattering data were collected at the X33 beamline of the EMBL (DESY, Hamburg) using a MAR345 image plate detector or a Pilatus 500K detector. A 4.2 mg/ml solution of BSA was measured as a reference and for calibration. The scattering patterns were measured with an exposure time of 2 min at 288 K. The wavelength was 1.5 Å. The sample-to-detector distance was set at 2.4 m, leading to scattering vectors q (defined as $q=4\pi/\lambda \sin\theta$, where 2θ is the scattering angle) ranging from 0.06 $Å^{-1}$ to 0.5 $Å^{-1}$. The concentration of the protein samples ranged from 0.92 mg/ml to 13.48 mg/ml, depending on the protein, and each protein was measured at three to five concentrations. Background scattering was measured after each protein sample using the buffer solution and subsequently subtracted from the protein scattering patterns after proper normalisation and correction from detector response.

The radii of gyration (R_g) were derived from the Guinier approximation: $I(q) = I(0) \exp(-q^2R_g^2/3)$, where $I(q)$ is the scattered intensity and $I(0)$ is the forward scattered intensity (Guinier 1955). The radius of gyration and $I(0)$ are inferred from the slope and the

intercept, respectively, of the linear fit of *Ln[I(q)] vs* q^2 in the *q*-range *q*.Rg < 1.12. At low angles, the scattered intensities were very well approximated by the Guinier law, and revealed some repulsive interparticle interactions at high concentrations. All scattering curves were indicative of monomeric states of the molecules in solution. The distance distribution function *P(r)* was calculated on the merged curve by the Fourier inversion of the scattering intensity *I(q)* using GNOM (Svergun 1992) and GIFT (Bergmann 2000).

The low-resolution shapes of the protein constructs were determined *ab initio* from the scattering curve using the program GASBOR (Svergun 2001). Several independent fits were run with no symmetry restriction and the stability of the solution was checked. These solutions were subsequently compared with the program DAMAVER (Volkov 2003), which computes the normalized spatial discrepancy (NSD) value for the various obtained shapes (Putnam *et al*. 2007). In all cases, calculations led to highly similar forms with NSD values ranging between 0.8 and 1.3. The atomic crystallographic structures of the individual modules were positioned in the envelopes using PyMOL. For each structural model obtained the theoretical SAXS profile, the R_G, and the corresponding fit to the experimental data were calculated using the program CRYSOL (Petoukhov 2002). SAXS data are summarized in Table 9 and experimental SAXS curves and their fits are shown in Figures 20 and 21.

Construct (module boundaries[1])	Mw[2] (kDa)	Number of aa[2]	R_G (Å)	D_{max} (Å)	*ab initio* modeling	
					d_χ (GASB)	NSD
CpGH84C-CBM32-Coh (31-909)	100	902	37±1	116±3	1.0	1.3
CpGH84C-CBM32 (31-767)	82.3	760	35±6	103±6	2.3	1.3
CBM32-Coh (625-909)	32.4	307	26±3	85±3	1.7	0.9
CBM32-Coh/FIVARDoc (N/A)	47.9	447	31±3	95±3	1.4	1.2
Coh-FN3 (765-1001)	27.0	246	37±6	118±6	nd[3]	nd
BSA	66.3	582	31±1	90±2	nd	nd

[1] construct boundaries are shown relative to the amino acid numbering in the intact CpGH84C protein.
[2] values are calculated including the N- or C-terminal six-histidine tags.
[3] nd, not determined

Table 9. Structural parameters of CpGH84C constructs obtained by SAXS.

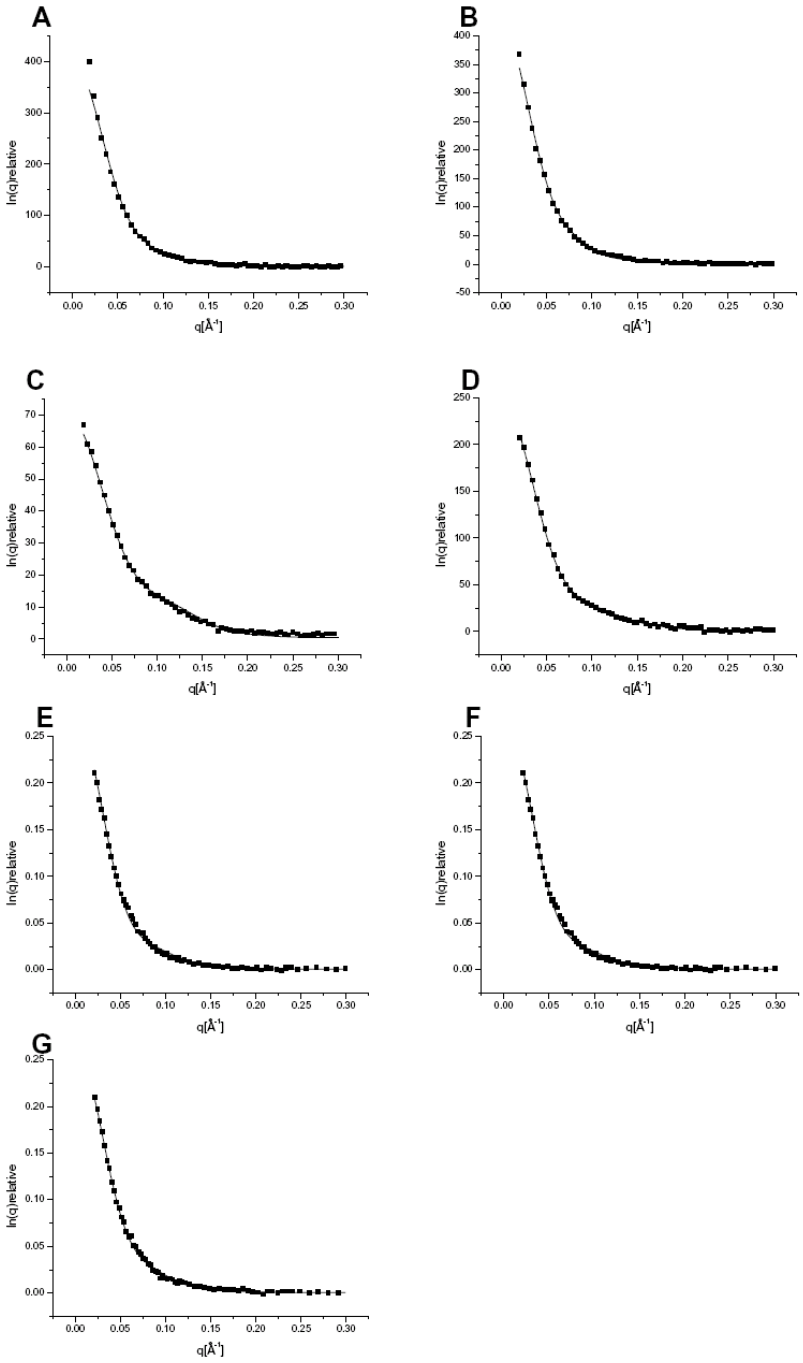


Figure 20. Crysol generated theoretical SAXS scattering curve fit to the experimentally generated SAXS scattering curve for (A) CpGH84C-CBM crystal structure (B) CpGH84C-CBM with some manual movement of structural components. (C) CBM-Coh (D) CBM-Coh:FIVARDoc (E) CpGH84C-CBM-Coh created by overlapping the SAXS structure of CBM-Coh. (F) CpGH84C-CBM-Coh created by overlapping the SAXS structure of CBM-Coh:FIVARDoc, though FIVARDoc is omitted from the calculation. (G) CpGH84C-CBM-Coh created by fitting to the *ab initio* GASBOR form.

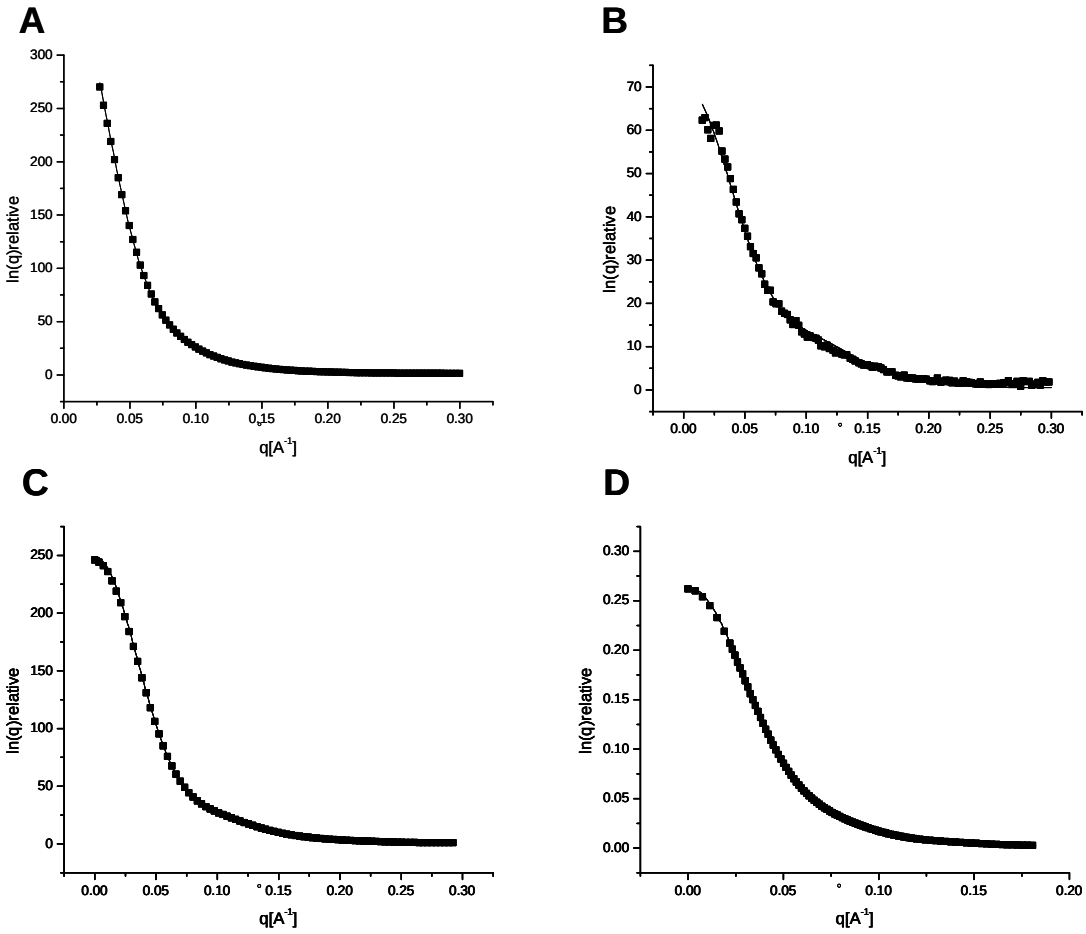


Figure 21. Experimental SAXS curves and the scattering profiles computed from the GASBOR models of the different constructs. The respective scattering curves were fitted with an *ab initio* model using the program GASBOR for (A) CpGH84C-CBM (B) CBM-Coh (C) CBM-Coh-FIVARDoc (D) CpGH84C-CBM-Coh

2.4.4 Results and Discussion

Structure of the CpGH84C catalytic module

To determine the overall architecture of the enzyme we dissected CpGH84C into a series of modular combinations. These constructs were successfully overproduced in *Escherichia coli* and purified in sufficient quantities for structural analysis by X-ray crystallography and SAXS (see Experimental Procedures for details).

The catalytic module construct of CpGH84C comprising the residues 41-624 of the full protein sequence was crystallized and its structure determined by X-ray crystal diffraction to 2.1 Å (Table 8, Figure 22A). This fragment of CpGH84C comprises three domains that are not distinguishable through amino acid-based sequence comparisons. The N-terminal domain (amino acids 41-177) is a mixed β-sheet structure consisting of six β-strands sandwiched between three α-helices; two helices on one face of the β-sheet and one helix on the other. A central domain (amino acids 178-470) adopts a $(\beta/\alpha)_8$ barrel lacking the 7^{th} helix while the C-terminal domain (amino acids 471-624) is an elongated five-helix bundle (Figure 2A). This is identical to the structure of a CpGH84C fragment from *C. perfringens* strain 13 reported by Rao *et al*. to 2.25 Å and thus the additional structural and functional properties of this catalytic region have been discussed in detail previously (Rao *et al*. 2006).

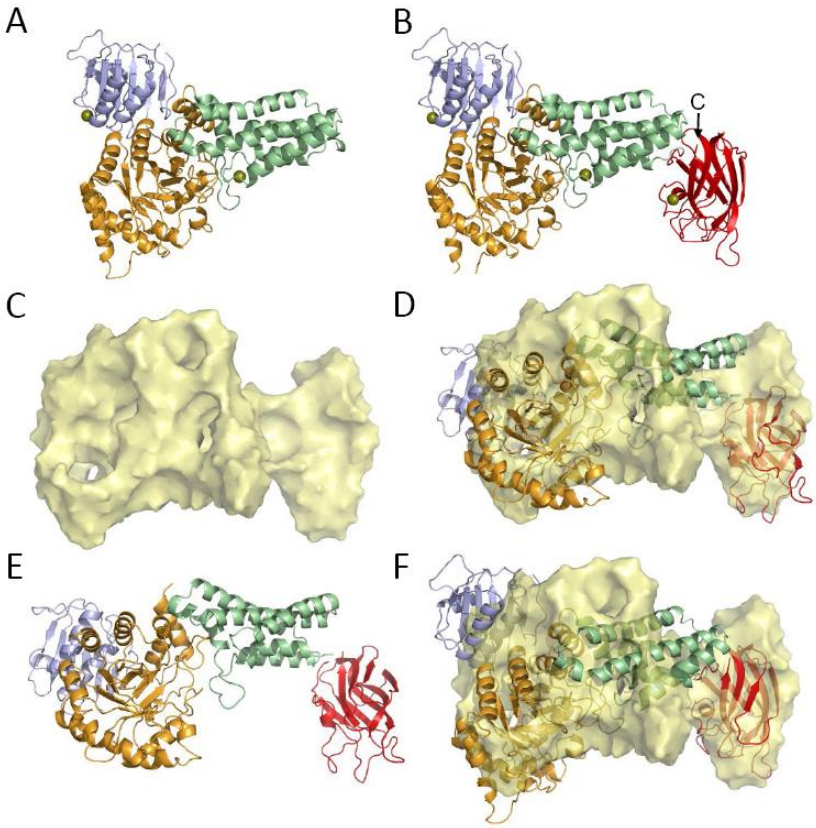


Figure 22. Structures of CpGH84C catalytic module and CpGH84C-CBM as determined using X-ray crystallography and SAXS. Panels (A) and (B) show the crystal structures of CpGH84C catalytic module and CpGH84C-CBM32, respectively, in a ribbon representation. The arrow in panel (B) shows the C-terminus of the CBM. Panel (C) shows the GASBOR generated SAXS envelope of CpGH84C-CBM while panel (D) shows the modules of CpGH84C-CBM manually fit into the SAXS envelope. Panel (E) shows the model in panel (D) without the SAXS form. Panel (F) shows the unmodified X-ray crystal structure, shown in panel (B) fit into the SAXS generated envelope. All of the structures are shown from identical orientations. The N-terminal domain is pictured in light blue, the catalytic TIM barrel in orange, the helical bundle in pale green, and the CBM in red.

Positioning the carbohydrate-binding module

A CpGH84C fragment comprising the catalytic module and the adjacent CBM32, here called CpGH84C-CBM32, was also crystallized and its structure determined by X-ray crystallography to 3.3 Å resolution (Table 8; Figure 22B). Though this data was measured to comparatively low resolution the availability of the high-resolution structures of the isolated catalytic module and CBM32, the latter previously determined at resolutions as high as 1.4 Å (Ficko-Blean and Boraston 2006) (See Section 2.2), made the placement of the modules by molecular replacement relatively facile and accurate. The CpGH84C-CBM32 structure reveals the positioning of the CBM32 at the extremity of the rigid α-helical linker domain with the long axis of the CBM32 at roughly right angles to the axis of the α-helical bundle (Figure 22B). The α-helical bundle and the CBM32 are separated by a short linker, which was clearly visible in electron density maps thus allowing unambiguous assignment of the CBM32 to the correct catalytic module in the crystal lattice. The C-terminus of the CBM32 is positioned approximately at the tip of the α-helical bundle where, in the intact enzyme, the Coh module would immediately begin (Figure 22B).

The solution conformation of the CpGH84C-CBM32 modular pair was also analyzed by SAXS (Table 9). The *ab initio* calculation of SAXS molecular envelopes for CpGH84C-CBM32 consistently yielded extended forms comprising a large globular region and a smaller region (Figure 22C). An atomic model of CpGH84C-CBM32 based on the SAXS data was derived by two methods. In the first method, the coordinates of the composite modules were individually placed in the *ab initio* SAXS envelope with an attempt at

minimizing bias towards the positioning of the modules in the crystal structure (Figure 22D and 22E). The crystal structure of the CpGH84C catalytic module was first placed in the SAXS envelope followed by rotation and translation of its domains to better fit the envelope. The coordinates of the CBM32 was then placed in the envelope in relation to the catalytic module. The relative positions of the N- and C-termini of contiguous domains were used as a constraint for placement. This produced a model with acceptable geometry in terms of the relative placement of the modules and gave a satisfactory χ_{crysol} value of 2.0. Given the structural similarity of this model to the CpGH84C-CBM32 crystal structure, we also modeled the entire CpGH84C-CBM32 crystal structure into the *ab initio* SAXS envelope as a rigid body (Figure 22F). The χ_{crysol} value of the model generated in this manner was 2.1 and thus not substantially different from the approach where the modules were positioned individually. The similarity of the unbiased SAXS-derived model to the crystal structure and the goodness of the fit of the unmodified crystal structure to the SAXS data indicate that the CpGH84C-CBM32 crystal structure is indeed representative of the solution conformation of this polypeptide.

Orientation of the cohesin module

Modular constructs of CpGH84C larger than CpGH84C-CBM32 proved either recalcitrant to crystallization (CpGH84C-CBM32-Coh) or to heterologous production (full length CpGH84C). Thus, we utilized a strictly SAXS-based approach to determine the structure of CpGH84C-CBM32-Coh and its complex with a fragment of the μ-toxin comprising its C-terminal Doc module and the single FIVAR module (found in various architectural regions) that precedes it, hereafter called FIVAR-Doc (Figure 8). These

studies were initiated with the CBM32-Coh construct to provide a means of determining the relative orientations of the CBM32 and Coh modules. The maximum SAXS-derived dimension of this construct, as determined by the D_{max}, was ~85 Å (Table 9), which was somewhat smaller than the sum of the known maximum lengths of the isolated CBM32 (49 Å) and Coh (56 Å) modules determined from their crystal structures. The structure of this modular pair, determined by modeling the two modules into the *ab initio* SAXS envelope using the positions of the N- and C-termini as a constraint, supports an end-to-end arrangement with an ~115^{o} angle between the two modules (Figure 23A). The angling of the two modules creates a more compact structure, which explains the D_{max} that is less than the sum of the end-to-end sizes of the modules. The relatively high χ_{crysol} value of 2.9 for this model suggests some flexibility between these two modules that has not been accounted for in this single representative conformation. Consistent with this observation was the CBM32-Coh polypeptide's complete recalcitrance to crystallization, implying structural heterogeneity and/or flexibility that prevented packing in a crystalline lattice. Similar flexibility of bimodular carbohydrate-active enzymes has previously been observed for cellulosomal components (Hammel *et al*. 2004). The D_{max} for the CBM32-Coh:FIVAR-Doc complex was ~95 Å and thus only slightly more extended than the CBM32-Coh structure (Table 9). An atomic model for this complex was generated by fitting the known high resolution X-ray crystal structure of the Coh:FIVAR-Doc complex (Adams *et al*. 2008) as a rigid body into the *ab initio* SAXS envelope followed by placement of the CBM32 module, again using the positions of the N- and C-termini as a constraint (Figure 23B). The modeled complex fit well into the envelope as judged by a

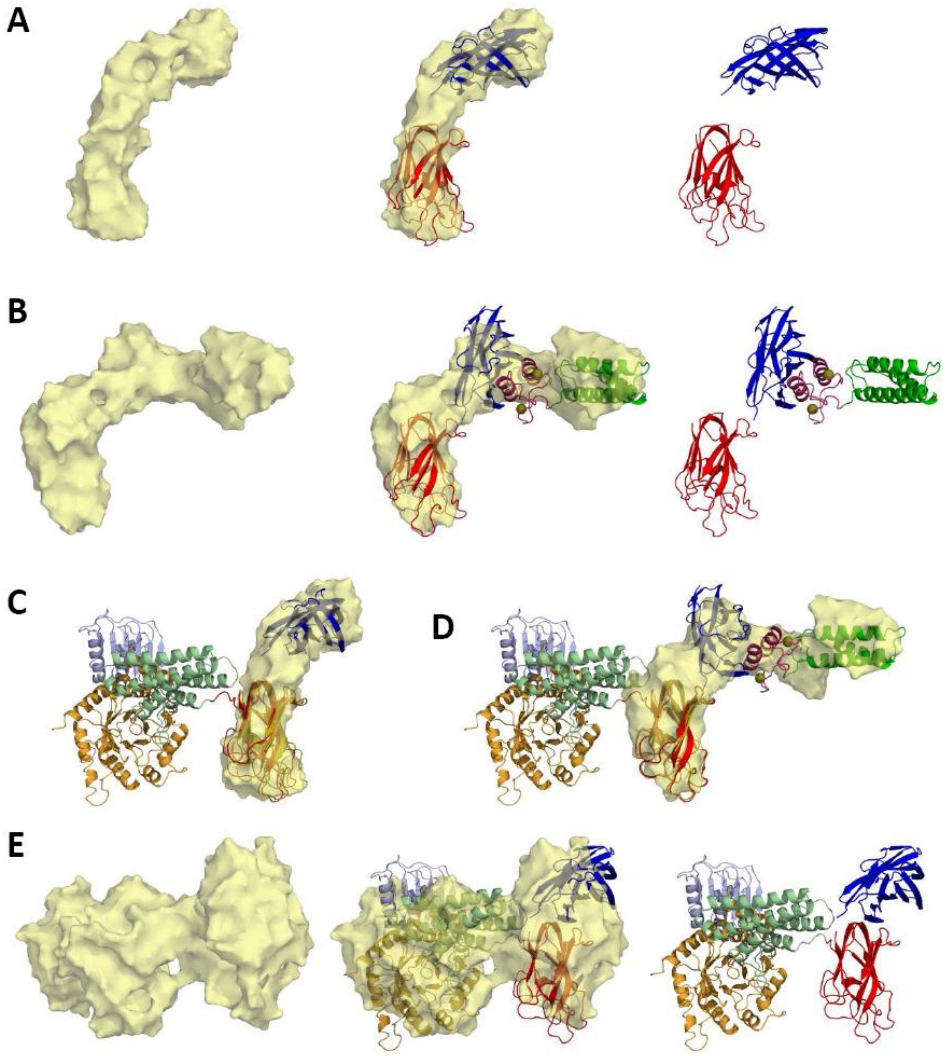


Figure 23. Structure of CBM-Coh and CBM32-Coh;FIVARDoc and CpGH84C-CBM-Coh as determined using SAXS. Panel (A) shows, from left to right, the GASBOR generated SAXS envelope for CBM-Coh, the SAXS form with the structures of the CBM (red) and Coh (blue) modules fit into it, and the ribbon representation of the SAXS derived model without the SAXS envelope. Panel (B) shows the same for the CBM-Coh module in complex with FIVAR-Doc from the μ-toxin (green denotes the FIVAR, pink the Doc and yellow spheres the calcium atoms). Panel (C) shows the crystal structure of CpGH84C-CBM that was overlapped with the GASBOR generated SAXS envelope and model of CBM-Coh and panel (D) shows the crystal structure of CpGH84C-CBM that was overlapped with the GASBOR generated SAXS envelope and model of the CBM-Coh:FIVARDoc complex. Panel (E) shows, from left to right, the GASBOR generated SAXS envelope for CpGH84CCBM-Coh, the SAXS form with the structures of the catalytic module (purple, orange, and green), the CBM (red) and Coh (blue) modules fit into it, and the ribbon representation of the SAXS derived model without the SAXS envelope.

χ_{crysol} value of 1.2 and perhaps reflected a more structured relative arrangement of CBM32 and Coh modules induced by the binding of the FIVAR-Doc component.

An initial putative model of CpGH84C-CBM32-Coh was generated by overlaying the CBM32 of the CBM32-Coh modular pair and CBM32-Coh:FIVAR-Doc SAXS models with the CBM32 of the CpGH84C-CBM32 crystal structure; the orientations of Coh and Coh:FIVAR-Doc portions of the model were maintained relative to the CBM32 but placed to minimize clashes with helical bundle of the CpGH84C catalytic module (Figure 23C and 23D). The two resulting models, with the FIVAR-Doc portion removed from the second model, agreed well with SAXS data collected on CpGH84C-32CBM-Coh with χ_{crysol} values of 1.5 and 1.6, respectively. Indeed, the measured D_{max} of ~116 Å was consistent with the maximum dimensions of 124 Å for both of the two models (not considering the modeled FIVAR-Doc portions) (Table 9). To provide additional support for this modular arrangement the *ab initio* SAXS envelope for this polypeptide was also calculated using GASBOR (Svergun 2001). The CpGH84C catalytic module was modeled as a rigid body while the CBM32 and Coh modules were rotated and translated relative to this component to best fit the envelope. The CpGH84C-CBM32 crystal structure was used as a guide for the placement of the CBM32 and, as above, the positions of the N- and C-termini of the component modules were used to constrain the potential orientations and spatial separation of the modules. The resulting model agreed very well with the scattering data and yielded a χ_{crysol} of 1.2.

Generating a complete model of CpGH84C

The structural studies of the various CpGH84C modules were completed by analysis of the Coh-FN3 modular pair, whose crystal structure was determined at a resolution of 1.8 Å (Table 8; Figure 24A). The Coh module adopts a β-sandwich fold comprising a β-sheet of 4 anti-parallel β-strands packing against a β-sheet of 5 anti-parallel β-strands, identical to that previously described for the module in isolation (Adams *et al*. 2008; Chitayat *et al*. 2008b). The FN3 module also adopts a β-sandwich fold of 3 anti-parallel β-strands atop 4 anti-parallel β-strands, which is characteristic of this family of modules (Bisig *et al*. 1999; Leahy *et al*. 1996; Leahy *et al*. 1994). Overall, the crystal structure reveals an elongated structure (109 Å long) with an extended eight-residue linker separating the two modules (Figure 24A). SAXS analysis of this modular pair yielded a D_{max} of ~115 Å, consistent with the crystal structure and indicative of this polypeptide's tendency to adopt an extended structure (Table 9). However, standard *ab initio* determination of SAXS envelopes yielded inconsistent results, suggesting there may be considerable conformational flexibility between the modules, as one might expect from the inter-modular linker.

Incorporation of the Coh-FN3 crystal structure into the crystal/SAXS-based structure of CpGH84C-CBM32-Coh enabled the generation of a complete composite model of CpGH84C (Figure 25). The Coh module of the Coh-FN3 structure was overlapped with the Coh module of the CpGH84C-CBM32-Coh model with the lowest χ_{crysol} value to

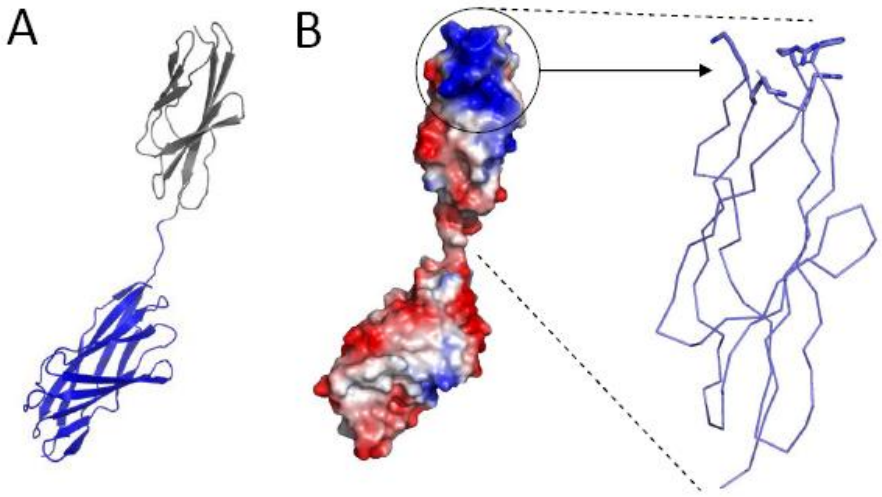


Figure 24. Structural features of the Coh-FN3 modular pair. Panel (A) shows a ribbon representation of the 1.8 Å crystal structure of Coh-FN3. The Coh module is depicted in blue and FN3 is shown in black. Panel (B) shows the surface representation of Coh-FN3 colored according to electrostatic potential (red is negative, blue is positive). The basic patch of FN3 is circled and expanded to show a patch of basic residues.

provide an indication of the placement of the FN3 module. The Coh module of this model was then overlapped with the high-resolution crystal structure of the Coh:FIVAR-Doc complex to incorporate the non-covalently associated FIVAR-Doc modules in the complete model. The resulting overall model clearly shows the splayed architecture of this complex (Figure 25). However, this model represents only one potential conformation of the full-length enzyme. The low resolution of the SAXS analyses made it impossible to determine the absolute relative orientations of the CBM32 and Coh modules. While the relative angles between the individual long axes of the modules could be approximated their relative rotational orientations around these axes could not. Therefore, assuming the CBM32 is appropriately positioned based on the X-ray crystal structure of CpGH84C-CBM32 this leaves some uncertainty as to the exact relative positioning of the Doc-binding site on the Coh module. However, given the steric constraints from the closely neighbouring CBM32, other potential orientations of the Coh module would likely have the Doc-binding site rotated away from the CBM32 (Figure 25). Indeed, considering the complex structural architecture of this enzyme and the somewhat limited potential for inter-modular contacts there is likely some capacity for movement between the modules, such as that suggested by the SAXS analysis of the CBM32-Coh structure. Inter-modular flexibility is likely to be particularly pronounced within the linker between the Coh and FN3. Thus, the FN3 in this enzyme likely samples a large number of spatial conformations relative to the remainder of the protein (Figure 25).

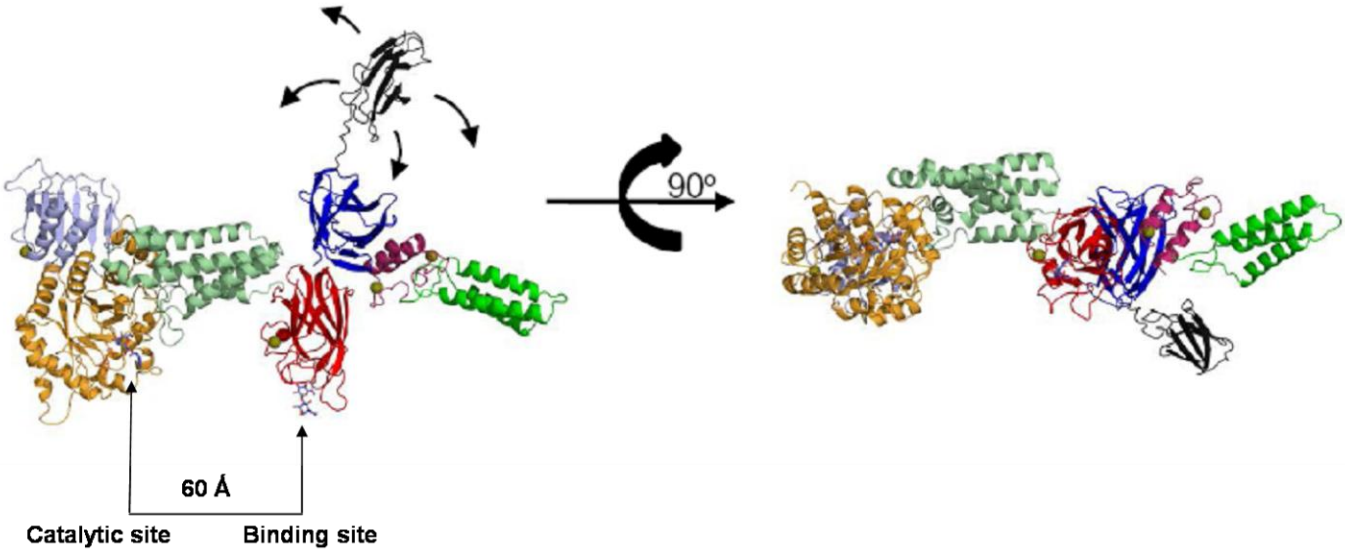


Figure 25. Composite structure of CpGH84C. The overall structure of CpGH84C in complex with the FIVAR-Doc module from the μ-toxin determined by the amalgamation of the SAXS data and crystallographic data is shown. Two Asp residues, 297 proposed to be involved in stabilizing the conformation of the acetamido group of GlcNAc and 298 proposed to be the catalytic acid, of CpGH84C are shown as sticks interacting with the competitive inhibitor PUGNAc, which was modeled in on the basis of the *C. perfringens* CpGH84C complex determined by Rao *et al* (Rao et al. 2006). The CBM32 module is shown interacting with its carbohydrate ligand, *N*-acetyllactosamine, which was modeled in on the basis of the previously determined complex of the CBM with this sugar (Ficko-Blean and Boraston 2006). Arrows near the FN3 module represents possible motion of this module due to the flexible linker region. The structure is also shown rotated by 90º around the horizontal axis running parallel to the page (right). The N-terminal domain is pictured in light blue, the catalytic TIM barrel is coloured orange, the helical bundle in pale green, the CBM32 in red, the cohesin module in blue, and the FN3 module in black. The FIVAR and Doc modules from the μ-toxin are shown in green and pink, respectively. Calcium atoms are shown in yellow.

Spatial coordination of modular function

The overall composite structure of CpGH84C is exceptionally revealing of how the known functions of the individual modules of CpGH84C are coordinated in three-dimensional space. The most immediate and striking observation is the relative positioning of the catalytic site and the binding site of the CBM32. CpGH84C is an *exo*-β-D-*N*-acetylglucosaminidase whose activity is likely directed towards terminal *N*-acetylglucosamine residues presented by host tissues (Rao *et al*. 2006). CBM32 binds carbohydrates bearing non-reducing terminal *N*-acetylgalactosamine disaccharide motifs but also binds other terminally galactosylated sugars (Ficko-Blean and Boraston 2006). It is widely maintained that the CBMs present in carbohydrate-active enzymes target and hold it in proximity to substrate (Bolam *et al*. 1998; Boraston *et al*. 2004); however, there are surprisingly few intact structures of carbohydrate-active enzymes, and no solution structures, that provide insight into how this adherence and catalysis might be coordinated. Our model of CpGH84C, which is based on both the crystal and solution conformations of the protein, reveals that the active site of the catalytic module and the carbohydrate-binding site of the CBM32, separated by ~60 Å, are located on the same side of the protein and are oriented in the same direction (Figure 25). This suggests an optimal arrangement to simultaneously interact with and hydrolyze terminal glycans presented on a surface, such as those clustered on the outer surface of a host cell or mucosal surface. While such a coordinated attack of binding and catalysis has been implicated in CBM-containing carbohydrate-active enzymes, this suggestion has been mainly limited to enzymes that hydrolyze insoluble plant cell-wall polysaccharides and has not been directly supported by solution structure data. The available structures of

multi-modular carbohydrate-active enzymes involved in bacterial pathogenesis have been primarily restricted to the sialidases from *Vibrio cholerae* (VCNA) and, more recently, *Streptococcus pneumoniae* (NanB) (Gut *et al*. 2008; Moustafa *et al*. 2004b). These proteins both contain a β-propeller catalytic module, which is fused to two CBM40 modules in VCNA and one CBM40 in NanB. In both of these proteins, the sialic acid-binding site of the CBM is oriented in the same direction as the catalytic site, as in CpGH84C; however, the distance between catalytic sites and carbohydrate binding sites is only ~25-30 Å. Though these multi-modular enzymes lack the complexity typical of extracellular carbohydrate-active enzymes from bacterial pathogens, the arrangement of their catalytic sites and CBM carbohydrate-binding sites do support a model of coordinated catalysis and adherence.

One of the most fascinating and newly discovered features of *C. perfringens* enzymes is their capacity to interact with each other *via* an ultra-tight Coh:Doc interaction (Adams *et al*. 2008). The model of CpGH84C shows that the Doc-binding site on the Coh module is spatially removed from the carbohydrate recognizing components of the enzyme thus allowing simultaneous substrate recognition and recruitment of other carbohydrate-active enzymes through the Coh:Doc interaction. Nevertheless, the structural placement of the Coh would permit another associated enzyme to lie in the same plane as the substrate-presenting surface thus allowing the second component to act on the carbohydrate bearing surface as well. Indeed, three *C. perfringens* glycoside hydrolases, including the μ-toxin, are known to contain functional dockerin modules and two of these also contain CBMs (Adams *et al*. 2008) (Figure 8). This latter observation raises the possibility of

substantial avidity effects resulting from simultaneous binding to host glycans by two or more CBMs present in these enzyme complexes. The harmonized formation of such complexes would also harness the diversity of specificities displayed by *C. perfringens* enzymes to synergistically degrade host tissue complex glycans.

FN3 modules are found in numerous bacterial carbohydrate-active enzymes, yet little is known about their function(s). Indeed, the function of the FN3 module of CpGH84C remains enigmatic. In eukaryotic systems, FN3s are involved in a variety of molecular recognition processes such as cell adhesion, cell surface hormone receptors, and cytokine receptors by facilitating protein–protein interactions (Bencharit *et al*. 2007; Hansen *et al*. 2008; Yang *et al*. 2008) The structural opposition of the FN3 from the active site of CpGH84C suggests that the FN3 module does not play a role in substrate recognition (Figure 25). However, given the historical role of FN3s in protein-protein interactions and its spatial placement in the enzyme it is conceivable that the CpGH84C FN3 module, like the Coh module, may mediate recruitment of other *C. perfringens* enzymes through an as yet unidentified interaction. Another plausible function is suggested by the unique surface charge of the FN3 module. The tip of the FN3 module that is most distal from the body of the enzyme has a distinct patch of basic residues leading to a cluster of positive charge at physiological pH (Figure 4B). This charge is complementary to the negative charge of Gram-positive bacterial cell walls, which is imparted by the phosphodiester bonds of techoic acid moieties on the exterior of organisms, such as *C. perfringens* (Myers *et al*. 2006; Weidenmaier and Peschel 2008) Considering the spatial disposition of the FN3 and its associated basic patch, FN3s may function to associate the enzyme

with the bacterial cell wall *via* an electrostatic interaction. In this scenario, the Doc-containing enzymes, which invariably lack FN3 modules or other obvious cell-wall attachment motifs, could also be tethered to the cell wall through their interaction with Coh-containing proteins, which always contain FN3 modules (Figure 8).

2.5 Discussion on Molecular Interactions and the Potential for Avid Binding among the Glycoside Hydrolase Enzymes Secreted by *C. perfringens*

Both CpGH84C and the μ-toxin are highly modular enzymes secreted among an assortment of other glycoside hydrolases with varying specificities by pathogenic *C. perfringens*. One thing that makes these glycoside hydrolase enzymes unique is the extent of their modularity. CpGH84C is one of the simplest with four modules which function in catalysis, CBM-carbohydrate interaction, protein-protein interaction, and predicted cell wall association. In turn, the μ-toxin has one of the more complex modular architectures with a catalytic module, four modules predicted to be involved in CBM-carbohydrate interaction; and a dockerin module involved in protein-protein interactions. In fact, threaded throughout the secreted glycoside hydrolases are ancillary modules that function in forming intermolecular interactions (Figure 8). We can question then, what influence might the ancillary modules of *C. perfringens* have on the enzymes? Avidity is a term used to describe the effect of multiple synergistic binding events and represents the combined strengths of the interactions. Avid binding may increase affinity relative to a single component binding event. With respect to protein-carbohydrate interactions avidity can be found driving antibody-carbohydrate interactions, lectin-carbohydrate interactions and also other CBM-carbohydrate interactions. The following discussion focuses on the molecular interactions contributed by the ancillary modules from the secreted glycoside hydrolase enzymes of *C. perfringens* and their potential effect on avid binding.

CBM32s have the general property of binding relatively weakly to their ligands. This weak affinity may be mitigated by the presence of multiple CBMs within the enzyme

which could exploit avidity when binding to a carbohydrate covered surface. Examples of CBMs overcoming weak affinity through avidity can be found among some of the plant cell wall degrading glycoside hydrolases (Bolam *et al*. 2001; Boraston *et al*. 2002). Interchain and intrachain co-operative binding to soluble xylan and insoluble cellulose was demonstrated with a triplet of CBM6 modules from a xylanase produced by *C. stercorarium*. The avid binding by these CBMs increased the affinity ~20 to ~40 fold relative to the individual modules (Boraston *et al*. 2002). Similarly, two CBM2 modules were shown to act in synergy when binding acid swollen cellulose and xylan. Synergistic binding increased the affinity ~18-20 fold (Bolam *et al*. 2001). Multivalent binding effects have been described from a *S. pneumonia* glycoside hydrolase virulence factor's CBMs with a triplet of CBM47 modules. The triplet bound to immobilized fucose resin more tightly than the doublet which, in turn, bound more tightly than a single module (Boraston *et al*. 2006).

The repetition of CBMs within an enzyme can take on two forms. The first is homo-ligand specificity (Figure 26A), in this case the tandem or linked CBMs are specific for the same ligand. For avid binding of the multivalent CBM construct to occur it would require a multivalent ligand displaying the same binding epitope. The second is hetero-ligand specificity (Figure 26B), here, the tandem or linked CBMs have two or more different ligands. For avid binding to occur it would require a multivalent ligand displaying multiple epitopes. Since the four CBM32s within the μ-toxin lack the same binding machinery, it is unlikely they all share identical specificities (Figure 18B). This raises the interesting possibility of multivalent binding to a hetero-ligand display. This

also raises the possibility of there being a higher order of specificity. For if there is repetition of a multivalent carbohydrate epitope that the tetramer of CBMs is able to recognize, then no longer is it only the individual CBM driving specific binding but the tetramer in concert. These CBMs might display binding to a multivalent epitope. Four potential CBMs with differing specificities might also maximize the potential of binding a carbohydrate bearing surface and therefore promote the enzyme coming into contact with substrate (Figure 27A). Given the unique and diverse carbohydrate environment in the body, there may be increased likelihood of binding if an enzyme has four CBMs of different binding specificities rather that only one. Further exploration of the binding specificities of other members within this family should reveal further ligand diversity and modes of binding.

Not only do the enzymes interact with carbohydrate through their CBMs but they also interact with one another through the highly specific ultra-tight cohesin-dockerin protein-protein inter-molecular interactions (Figure 27B). The cohesin-dockerin binding provides the framework to form higher order structures, related to those of the plant cell wall degrading Clostridial sp. which form multi-component macromolecular cellulose degrading complexes called cellulosomes. *C. perfringens* lacks the other necessary components to form a cellulosome yet the Coh:FDoc interaction is retained. The strength of the cohesin-dockerin interaction is in the order of ~10^{11} M^{-1}, this intermolecular interaction drives the association of CpGH84C (containing one CBM32 module) with the μ-toxin (containing one known and three predicted) CBM32 modules. The potential for avid binding to carbohydrate is now shared between CpGH84C and the μ-toxin with at

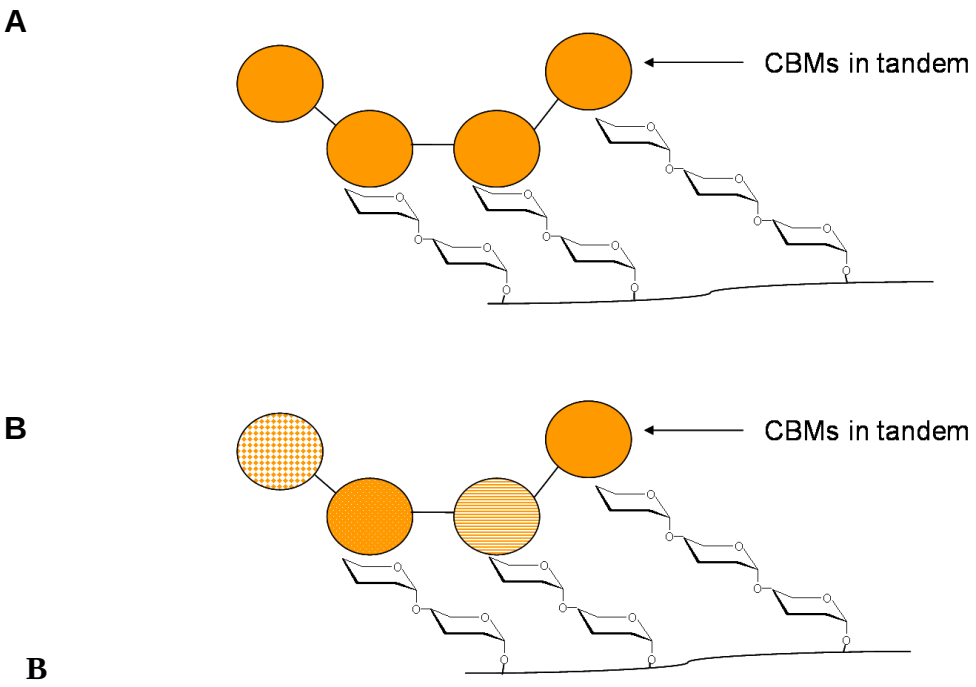


Figure 26. Multivalent binding by CBMs. (A) Homo-ligand specificity. The CBMs bind the same multivalent ligand epitopes. (B) Hetero-ligand specificity. The CBMs bind different multivalent ligand epitopes. Ligands are attached to a carbohydrate bearing surface.

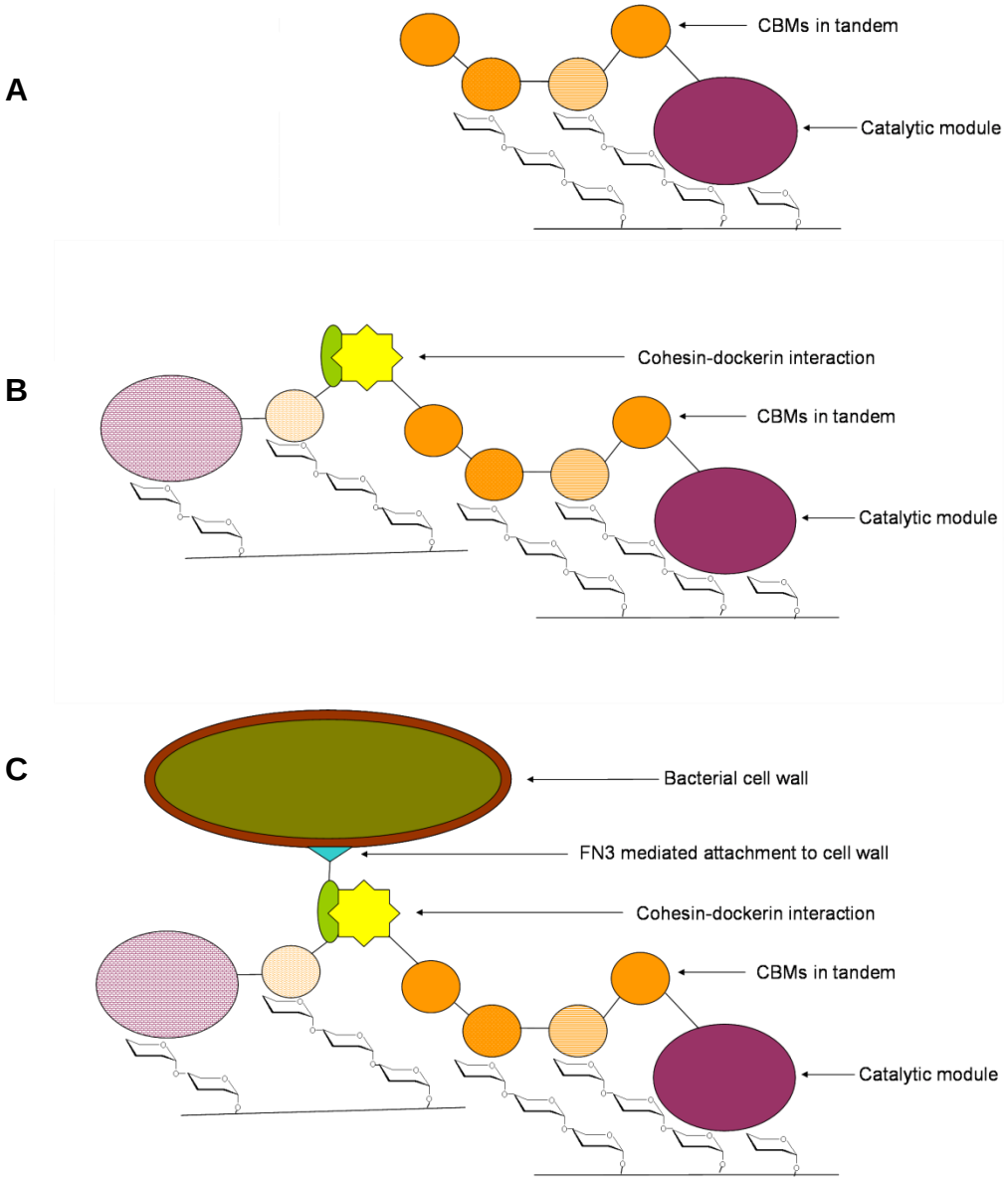


Figure 27. Higher order structure formation. (A) Multivalent Binding Effects. If the CBMs have the same specificity this implies homo-ligand specificity and therefore the CBMs bind the same multivalent carbohydrate epitope. If the CBMs have different specificities this implies hetero-ligand specificity and therefore the CBMs bind different multivalent carbohydrate epitopes. (B) Association of the glycoside hydrolases through a cohesin-dockerin interaction. CBMs from both enzymes now maintain association to the carbohydrate surface. (C) FN3 adherence to the bacterial cell wall. This ensures the enzymes are near the surface of the bacterial cell for optimal scavenging efficiency. Multivalent binding of the CBMs maintains enzyme and bacterial association near the carbohydrate surface.

the very least two different carbohydrate binding specificities (i.e. galacto- and gluco-configured ligands respectively). The cohesin/dockerin motif is found throughout the extracellular glycoside hydrolases of *C. perfringens* (Figure 8). In some cases the glycoside hydrolases contain both a cohesin and a dockerin suggesting the ability to form multiple higher order structures through high affinity interactions.

The prominent exposure of the FN3 module at the C-terminus of multiple enzymes within the secreted glycoside hydrolase enzymes of *C. perfringens* suggests this exposure is important to its function, which, if our hypothesis is correct, is to maintain intimate contact with the bacterial cell wall (Figure 27C). The overall effect of this molecular interaction would be to concentrate the enzyme activity at the cell surface of *C. perfringens*. Interestingly, many of the secreted glycoside hydrolases that don't have terminal FN3 modules do carry the LPXTG amino acid sequence which is known to direct sortase-mediated cell wall attachment (Figure 8). Since these enzymes already have the LPXTG motif they likely do not require an FN3 at their terminus to bind the cell wall. The sum of this information provides a new paradigm for bacterial virulence. The bacterial cell wall might well be coated with these enzymes attached *via* their FN3 modules or through sortase-mediated attachment. The enzymes protrude from the cell wall as their CBMs latch onto a carbohydrate surface. Further association is driven by the ultra-tight protein-protein interactions of the cohesin and dockerin domains. All the glycoside hydrolase enzymes that contain cohesin modules also contain FN3 modules (Figure 8) ensuring attachment to the cell wall and recruitment of other dockerin containing enzymes. Now neighbouring enzymes are also promoting association to the

multivalent carbohydrate surface through their CBMs. The secreted modular glycoside hydrolases of *C. perfringens* give more involved meaning to the concept of avidity. Multi-modularity drives the ability of these massive enzymes to form molecular interactions, which, in turn, may act to promote the association of enzyme and of neighbouring enzyme to a multivalent carbohydrate bearing surface. Essentially, these enzymes are made to interact with one another, with carbohydrate and to the cell wall of the invading bacteria. This is a potentially synergistic situation, where combined components act more efficiently than the individual components. For example, four enzymes, a sialidase, a β-galactosidase, a mannosidase and a β-hexosaminidase acting together (in synergy) on an N-terminal glycan displaying different monosaccharide units, would show increased efficiency over that of the individual enzymes making random contacts with the same glycan. This might, particularly in the harsh colonized environment of the intestines, provide increased scavenging efficiency which in turn would promote bacterial growth and cell division and thus, successful colonization. Further study is required to verify the possibilities discussed here within.

The human body provides accommodation to a greater number of bacterial cells than it has of its own cells (Hooper and Gordon 2001; Sears 2005). The majority of niches within the body that are occupied by these bacteria, such as the mucosal surfaces of gastro-intestinal tract, the genito-urinary tract, and the upper respiratory tract, display enormous carbohydrate content and diversity. Not surprisingly, commensal bacteria, which are often also opportunistic pathogens, as well as other dedicated pathogens, have developed extensive carbohydrate-degrading machineries to enable their persistence in

the host and to promote infection. The enzymes comprising these carbohydrate-degrading systems are often large and complex and thus poorly understood. Here, we have provided a detailed look at the complete structure of such a complex enzyme from *C. perfringens*. While the numerous bacterial carbohydrate-active virulence factors do not display an identical arrangement of modules they do invariably contain catalytic modules, which are often associated with CBMs, and frequently numerous other modules of unknown functions. Investigation into the functions of the glycoside hydrolase modules has allowed us to develop a model for adherence not described before in a bacterial pathogen. This research provides a paradigm to help guide the analysis and interpretation of other important multi-modular microbial enzymes.

Chapter 3: Structural and Mechanistic insights into the Basis of Mucopolysaccharidosis IIIB

Elizabeth Ficko-Blean[1], Keith A. Stubbs[2], Oksana Berg[1], David J. Vocadlo[2] and Alisdair B. Boraston[1]

Adapted from: Structural and mechanistic insight into the basis of mucopolysaccharidosis IIIB. Proc Natl Acad Sci U S A. 2008 May 6;105(18):6560-5.

[1]Biochemistry & Microbiology, University of Victoria, PO Box 3055 STN CSC, Victoria, BC, V8W 3P6, Canada. [2] Department of Chemistry, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada

3.1 Abstract

Mucopolysaccharidosis III (MPS III), or Sanfilippo syndrome, has four forms (A-D) that result from build up of an improperly degraded glycosaminoglycan, heparan, in lysosomes. MPS IIIB is attributable to the decreased activity of a lysosomal α-N-acetylglucosaminidase (HsGH89 or NAGLU). Here we describe the structure, catalytic mechanism, and inhibition of a close bacterial homolog of HsGH89 from *Clostridium perfringens*, CpGH89. The structure of CpGH89 enables the generation of a homology model for human HsGH89, an enzyme which has resisted structural studies despite having been studied for over 20 years. This model reveals which mutations that give rise to MPS IIIB map to the active site and which map to regions distant from the active site. The identification of potent inhibitors of CpGH89 and the structures of these inhibitors in complex with the enzyme suggest small molecule candidates for use as chemical chaperones. These studies therefore illuminate the genetic basis of MPS IIIB, provide a clear biochemical rationale for the necessary sequential action of heparan degrading enzymes, and open the door to the design and optimization of chemical chaperones for treating MPS IIIB arising from mutations distal from the active site.

3.2 Introduction

Lysosomes are acidified vesicles, within human and animal cells, containing enzymes functioning in the digestion of proteins, polysaccharides, nucleic acids and lipids. Lysosomal storage disorders are an heritable group of heterogeneous disorders characterized by a genetic defect in a lysosomal enzyme, receptor target, activator protein, membrane protein or transporter (Wilcox 2004). The inability to catabolically process a substrate results in the build up of the substrate in the lysosomes. As the lysosomes become progressively more distended they occupy more cytosolic space, consequently interfering with normal cellular function.

Mucopolysaccharidosis (MPS) is one group of autosomal recessive lysosomal storage disorders caused by the storage of undegraded glycosaminoglycans in lysosomes. MPS-III or Sanfilippo syndrome is characterized by the reduced function of one of four lysosomal enzymes involved in the sequential degradation of heparan sulfate. MPS-IIIA is caused by the reduced activity of heparan sulfate sulfamidase, MPS-IIIB by the reduced activity of an α-*N*-acetyl-D-glucosaminidase (NAGLU, here referred to as HsGH89), MPS-IIIC by the reduced activity of acetylCoA:α-glucosaminidase-*N*-acetyltransferase, and MPS-IIID by the reduced activity of *N*-acetyl-D-glucosamine 6-sulfatase (Yogalingam and Hopwood 2001). The four Sanfilippo subtypes occur about every 1 in 24000 births (van de Kamp *et al*. 1981). The disease ultimately leads to CNS degeneration whose early symptoms include aggressiveness and hyperactivity and later progressive mental retardation. Patients generally survive into early adulthood, and although attenuated forms of the disease have been described with longer survival rates, there is currently no cure or effective treatment for this disease.

HsGH89 is a glycoside hydrolase that is placed into family 89 of the glycoside hydrolase classification (URL: http://www.cazy.org). As an α-*N*-acetylglucosamine specific glycoside hydrolase, it is responsible for cleaving the glycosidic bond in the backbone of heparan and thus plays a key role in heparan recycling. Upwards of 100 different mutations, including missense, nonsense, and deletions, in the *hsgh89* (*naglu*) gene have been associated with the MPS-IIIB phenotype and biochemical studies have confirmed the deleterious effects of these mutations (Yogalingam and Hopwood 2001; Beesley *et al*. 2005; Beesley *et al*. 1998; Emre *et al*. 2002; Tanaka *et al*. 2002; Tessitore *et al*. 2000; Schmidtchen *et al*. 1998; Beesley *et al*. 2005; Beesley *et al*. 1998; Bunge *et al*. 1999; Weber *et al*. 1999). The catalytic mechanism of HsGH89 and the structure of this enzyme, however, remain unknown. Despite having been cloned over 10 years ago (Zhao *et al*. 1996), no structural or mechanistic data for HsGH89 has been obtained, hindering the development of potential therapeutic strategies to treat patients suffering from MPSIIIB. To address this information gap we cloned CpGH89, a bacterial family 89 glycoside hydrolase from *C. perfringens* to use as a model for HsGH89 since they share a relatively high level of amino acid sequence identity (~30%). ***We hypothesize that CpGH89, based on its sequence similarity to the human homologue which has known function, is active as an α-N-acetylglucosaminidase.*** Here we detail the structure of this enzyme and a homology model of the human enzyme. These structures enable us to map the known mutations causing MPS IIIB onto the structure. Further, identification of potent inhibitors of this enzyme, the structures of the enzyme in complex with these

inhibitors, and detailing of the catalytic mechanism will facilitate the rational development of chemical chaperones that could be used to treat this tragic genetic illness.

3.3 Materials and Methods

All reagents are from Sigma-Aldrich unless otherwise stated.

Cloning, Protein Production and Purification

The gene fragment comprising nucleotides 76-2748 of the *cpgh89* gene (locus tag CPF_0859), which encodes a polypeptide with a putative N-terminal family 32 carbohydrate-binding module followed by the catalytic module, was PCR amplified from *C. perfringens* ATCC 13124 genomic DNA (Sigma) and cloned into pET-28a(+) (Novagen) *via* 5' and 3' *Nhe*I and *Xho*I restriction endonuclease sites to generate pCBM1GH89 using standard PCR cloning procedures. The polypeptide encoded by this recombinant gene comprises an N-terminal H6 tag, followed by a thrombin protease cleavage site fused to the N-terminus of the CBM32/GH89 modules and is referred to as CpGH89. "Mega-primer" (Barik 1996) PCR site directed mutagenesis procedures were used to introduce the E483A and E601A substitutions. The DNA sequence fidelity of all constructs was verified by bi-directional sequencing. *Cp*GH89 was produced in BL21 star (DE3) *Escherichia coli* cells (Invitrogen) harboring the pCBM1GH89 and purified by immobilized metal affinity chromatography using methods described previously (See Section 2.2).

Enzyme kinetics

All steady state kinetic studies were carried out at 25 ^{o}C in a Cary/Varian 300 Bio UV-Visible Spectrophotometer. The pH dependence of CpGH89 was determined using 133 nM enzyme and a pNP-α-GlcNAc concentration of 1 mM. 0.1 M sodium acetate was used to buffer within the pH range 4.6-6.0; 0.1 M sodium phosphate was used to buffer between pH range 6.0-8.0; 0.1 M glycine-glycine was used to buffer between the pH range 8.0-9.8. At the transition between buffers one or two pHs were overlapped to control for effects of the buffer on the enzyme activity. Standard reaction mixtures for the determination of kinetic constants were done at the optimal pH (0.1 M sodium phosphate, pH 7.3) in 750 μL volumes containing 15.2 nM CpGH89 or 390 nM Glu483Ala, 1 mg/ml BSA, and 0-2.5 mM pNP-α-GlcNAc. 4-nitrophenolate production was measured at 400 nm over a 5 minute time period. Rate of release was determined by linear regression over the linear period of release on the curve. All experiments were performed in triplicate. Michaelis-Menten parameters were determined by non-linear curve fitting using Origin7. Similarly determination of the kinetic constants for dNP-α-GlcNAc, synthesized as described previously (Chen and Withers 2007), was done in 0.1 M sodium phosphate, pH 7.3 in 750 μL volumes containing 240 nM *Cp*GH89, 10 % BSA, and 0-3.7 mM dNP-α-GlcNAc

NMR Experiments

See Appendix A

Inhibitor Binding Studies

The K_i value for CpGH89 binding to PUGNAc was determined by linear regression of Dixon plots for inhibitor concentrations ranging from 1/3 to 3 times K_i.

Isothermal Titration Calorimetry (ITC) was performed with a VP-ITC (MicroCal, Northampton, MA) as described previously (Ficko-Blean and Boraston 2006; Lammerts van Bueren and Boraston 2004). CpGH89 was dialyzed extensively against 0.1 M sodium phosphate pH 7.3. Solid inhibitor was resuspended in buffer saved from the dialysis. Twenty-five 10 μL injections of 100 μM 2-acetamido 1,2 dideoxynojirimycin were titrated into 6 μM of protein. Similarly, twenty-five 5 μL injections of 1700 μM PUGNAc were titrated into 50 μM CpGH89. C-values (Wiseman *et al*. 1989) were maintained between 10 and 1000. Binding stoichiometry, enthalpy, and the equilibrium association constant were determined by fitting the heat of dilution corrected data to a one site binding model. All values reported are averages and standard deviations of experiments performed in triplicate.

Crystallization, Data Collection, Structure Solution and Refinement

CpGH89 in 20 mM Tris HCl, pH 8.0, was digested with thrombin overnight at room temperature. Digested protein was applied to an S-200 Sephacryl High Resolution Size Exclusion column and fractions containing pure CpGH89 (as assessed by SDS-PAGE) were pooled and concentrated as above. Crystal trials were set up using the hanging drop vapour diffusion method with both thrombin treated and untreated CpGH89. Diffraction

quality crystals were obtained in 2.0 M $(NH_4)_2SO_4$ with 3% glycerol (no buffer). Both thrombin treated and untreated CpGH89 yielded crystals.

The crystals were cryoprotected by a 10 second immersion in a solution containing 2 M ammonium sulfate and 30 % glycerol followed by freezing in a nitrogen gas cryostream. Diffraction data were collected with a Rigaku R-AXIS IV++ area detector coupled to an MM-002 x-ray generator with Osmic "blue" optics and an Oxford Cryostream 700. Data were processed with Crystal Clear/d*trek (Pflugrath 1999).

A holmium (III) derivative was obtained by soaking native CpGH89 crystals in mother liquor supplemented with 1 mM $HoCl_3$ for a period of 3 days. This crystal was cryoprotected as above. The heavy atom substructure of two holmium sites was determined with the programs ShelxC/D(Schneider and Sheldrick 2002) using the isomorphous differences between the derivative and a native dataset using data to a resolution of 4.5 Å. Refinement of heavy atom parameters and initial phasing to a resolution of 3.0 Å was performed with SHARP (Bricogne *et al*. 2003). Density modification with DM (Cowtan and Main 1998) was used to improve and extend the phases to 2.4 Å and ultimately yielded interpretable electron density maps. ARP/wARP (Perrakis *et al*. 1999) was able to build the backbone of ~40% of the molecule. Manual model building in COOT (Emsley and Cowtan 2004) allowed the initial model to be corrected and extended to ~65% completeness with some side chains. This was used as input into ARP/wARP which was then able to build a model of ~90% completeness with

~80% of the side chains correctly docked. The model was manually completed in COOT followed by refinement with REFMAC (Murshudov *et al*. 1997).

Inhibitor complexes (2-acetamido-1,2-dideoxynojirimycin and PUGNAc) and product complexes were obtained by adding excess inhibitor or substrate (pNP-α-GlcNAc) to mother liquor and soaking for a period of 1 hr to up to one week. Crystals were cryo-protected as per the native crystal. These complexes were solved using the native CpGH89 model as a starting point. Ligands were built into the models using COOT. When necessary, appropriate REFMAC library files containing the restraints for the ligands were obtained using the PRODRG Server (van Aalten *et al*. 1996).

In all data sets, the same five percent of the observations were flagged as "free" (Brunger 1993) and used to monitor refinement procedures. Water molecules were added using the REFMAC implementation of ARP/wARP and inspected visually prior to deposition. See Table 10 for all data collection, refinement and model validation statistics.

Homology Modeling

See Appendix B

Accession codes

Coordinates and structure factors have been deposited with the PDB codes of 2VCC for the native structure, 2VCA for the β-D-GlcNAc complex, 2VC9 for the 2AcDNJ complex, and 2VCB for the PUGNAc complex.

	Native	Holmium derivative	GlcNAc	2AcDNJ	PUGNAc
Data collection					
Space group	$P6_1$	$P6_1$	$P6_1$	$P6_1$	$P6_1$
Cell dimensions					
a, b, c, Å	90.71, 90.71, 252.66	90.88, 90.88, 252.65	90.77, 90.77, 252.78	90.96, 90.96, 252.40	90.70, 90.70, 252.47
a, b, g, °	90.00, 90.00, 120.00	90.00, 90.00, 120.00	90.00, 90.00, 120.00	90.00, 90.00, 120.00	90.00, 90.00, 120.00
Resolution, Å	20.00–2.00 (2.05–2.00)	49.26–3.00 (3.11–3.00)	29.71–2.05 (2.12–2.05)	19.83–2.36 (2.44–2.36)	29.69–2.20 (2.28–2.20)
R_{sym} or R_{merge}	0.126 (0.435)	0.118 (0.288)	0.080 (0.417)	0.134 (0.398)	0.071 (0.321)
$I/\sigma I$	7.8 (3.2)	13.0 (6.6)	8.1 (2.5)	4.9 (2.3)	8.7 (2.9)
Completeness, %	96.4 (92.1)	100.0 (99.8)	99.1 (98.7)	90.7 (97.1)	90.3 (91.9)
Redundancy	4.5 (4.4)	11.0 (10.8)	4.0 (3.9)	3.3 (3.5)	2.6 (2.4)
Refinement					
Resolution, Å	2.00	3.00	2.05	2.36	2.20
No. reflections	345557	258579	291398	143110	139907
R_{work}/R_{free}	0.21/0.26	—	0.19/0.23	0.21/0.25	0.21/0.25
No. atoms					
Protein	7,228	—	7,197	7,158	7,155
Ligand	—	—	15	14	25
Water	578	—	896	391	338
B factors					
Protein	31.9	—	29.5	34.9	39.4
Ligand	—	—	25.4	26.0	44.2
Water	34.2	—	39.5	34.9	37.1
rmsd					
Bond lengths, Å	0.007	—	0.008	0.008	0.008
Bond angles, °	1.058	—	1.321	1.157	1.125

Values in parentheses are for the highest resolution shell.

Table 10. Data collection and refinement statistics for CpGH89. Values in parentheses are for the highest resolution shells.

3.4 Results and Discussion

Activity and structure of CpGH89

The genome of *Clostridium perfringens* ATCC 13124 contains an ORF (CPF_0859) that encodes a putative 2095 amino acid protein. This protein is classified into glycoside hydrolase family 89 on the basis of amino acid sequence similarity between it and the known family 89 glycoside hydrolase from *Homo sapiens*. After molecular dissection of the modular structure of the *C. perfringens* GH89 we were able to recombinantly overproduce a truncated construct comprising residues 25-916 (an N-terminal family 32 carbohydrate binding module followed by the family 89 catalytic module) in *E. coli* and purify the resulting polypeptide, which for simplicity will be referred to as CpGH89 in this work.

CpGH89 was screened for catalytic activity against a panel of synthetic substrates and activity was detected only on *para*-nitrophenyl-α-*N*-acetyl-D-glucosamine (pNP-α-GlcNAc), consistent with the known activity of HsGH89. This substrate produced a bell-shaped activity-pH profile suggesting that two residues with kinetic pKa values of approximately 6.3 and 8.0 must be maintained in appropriate ionization states (Figure 28). Further kinetic characterization at the optimum pH of 7.3 was carried out to determine the values of K_m (1.1 ± 0.1 mM), k_{cat} [(2.6 ± 0.1) x 10^{-1} s^{-1}], as well as k_{cat}/K_m [(2.5 ± 0.3) x 10^{-1} $mM^{-1}s^{-1}$] of the enzyme toward pNP-α-GlcNAc. The substrate dNP-α-GlcNAc (dinitrophenyl-α-*N*-acetyl-D-glucosamine) (Chen and Withers 2007) is turned over significantly more quickly with the corresponding values for K_m, k_{cat}, and k_{cat}/K_m of (7.4 ± 1.1) x 10^{-1} mM, 8.5 ± 0.4 s^{-1} and 11.6 ± 2.3 $mM^{-1}s^{-1}$, respectively.

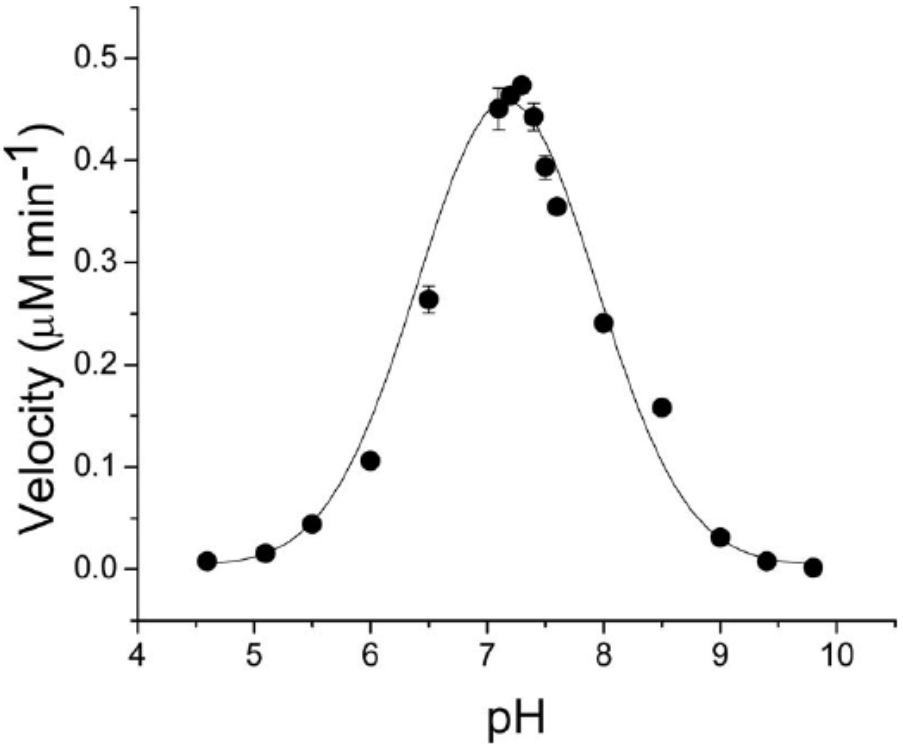


Figure 28. The pH/activity profile of CpGH89 using pNP-α-GlcNAc as a substrate.

The structure of CpGH89 was solved by SIR using a $HoCl_3$ derivative (see Materials and Methods). A single molecule of CpGH89, which is composed of 4 distinct domains, was present in the asymmetric unit. The N-terminal domain (residues 26-155) is a putative family 32 CBM having the typical β-sandwich fold adopted by many CBMs. The catalytic region comprises a small α/β domain (residues 170 – 280), an elaborated $(\alpha/\beta)_8$ barrel (residues 280-620), and an all α-helical domain that packs against the first three domains (residues 621-916)(Figure 1a). The elaborations on the $(\alpha/\beta)_8$ barrel comprise extensions of defined secondary structure that pile onto one face of the $(\alpha/\beta)_8$ core (see green in Figure 29A).

Family 89 glycoside hydrolases use a retaining catalytic mechanism

Soaking crystals of CpGH89 with pNP-α-GlcNAc produced a product complex with a single molecule of β-D-GlcNAc in the active site (Figure 29B). The active site of CpGH89 is a "sock" shaped cavity lined primarily with aromatic amino acid side chains (Figure 29B and C). The bottom of this pocket accommodates the terminal non-reducing sugar unit of the substrate glycoside. The plane formed by the pyranose ring of the sugar sits roughly perpendicular to entrance path into the active site. As well, the active site architecture limits the enzyme to act strictly as an *exo*-glycosidase. The positioning of the Glu483 and Glu601 side chains and their separation of ~6.5 Å in this complex suggested them to be the catalytic residues (Figure 29AB and C). Glu601 resides below the A-face of the sugar ring with Oε ~2.8-3.1 Å from the C1 while Glu483 is positioned directly above the B-face of the pyranose ring at what would be a distance of ~2.5-3.0 Å from the oxygen in a α-glycosidic bond. Glu601 should therefore be suitably poised for

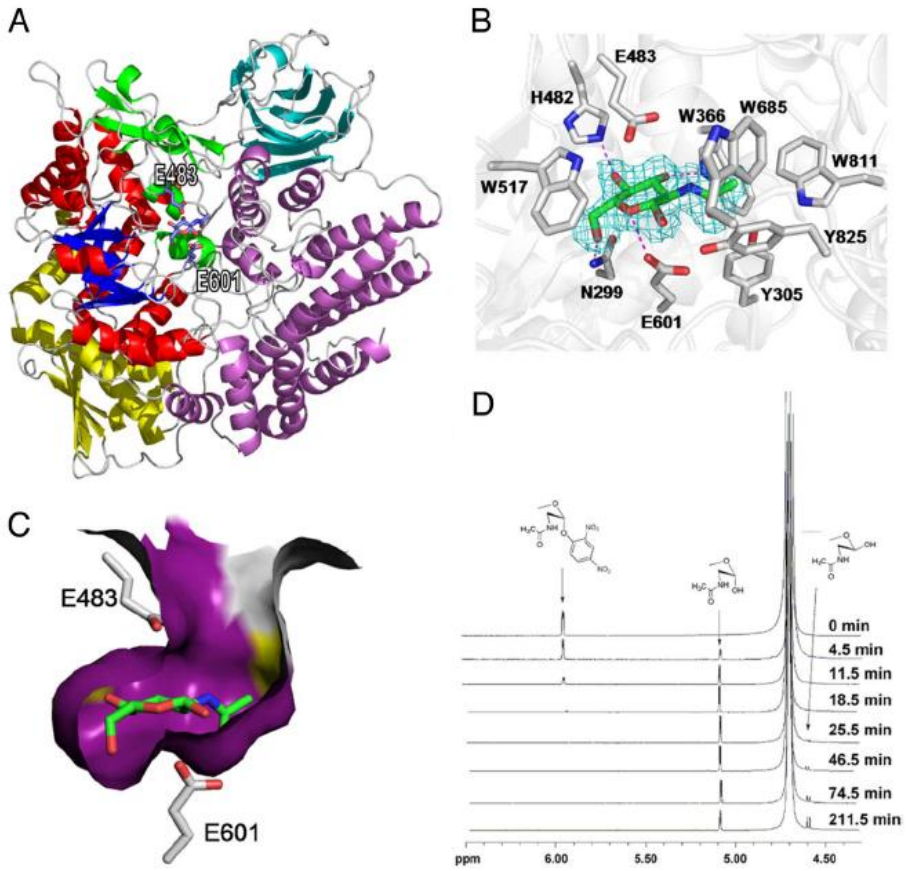


Figure 29. The structure and mechanism of *Cp*GH89. (A) A cartoon representation of CpGH89showing its overall fold. Its constituent domains are colored as follows: light blue, N-terminal domain with amino acid sequence similarity to family 32 carbohydrate-binding domains; yellow, linking α/β domain; red and blue (with small elaborations shown in green), core $(\alpha/\beta)_8$ barrel; purple, C-terminal all α-domain. A bound *N*-acetyl-D-glucosamine in the product complex is shown in blue stick representation, and the putative catalytic residues are shown in gray stick representation (labeled). (B) Active-site representation for the product complex *N*-acetyl-D-glucosamine. The blue mesh shows the maximum likelihood (Murshudov et al. 1997)/σA (Read 1986)-weighted electron density maps contoured at 0.26 e-/Å^3. The product (green) and key active site residues (gray), including the putative catalytic residues Glu-483 and Glu-601, are shown in stick representation. Putative hydrogen bonds between the protein and ligand, identified by using the criteria of proper geometry and a distance cutoff of 3.2 Å, are shown as dotted magenta lines. (C) The *N*-acetyl-D-glucosamine of the product complex in active site whose solvent accessible surface has been rendered and partially cut away to show the deep pocket of the catalytic site. The catalytic residues are shown in gray stick representation and the *N*-acetyl-D-glucosamine in green stick representation. (D) The hydrolysis of dNP-α-GlcNAc measured as a function of time by ^{1}H NMR spectroscopy.

nucleophilic attack of the anomeric carbon of a bound substrate while Glu483 acts as the acid/base. This architecture suggested a two-step catalytic mechanism in which the stereochemistry at C1 of the glycon is retained. We tested this hypothesis by NMR analysis of the hydrolysis of dNP-α-GlcNAc (Figure 29D). The results clearly revealed the rapid appearance of the α-anomer of GlcNAc (α-D-GlcNAc) thus supporting a retaining catalytic mechanism. The gradual appearance of the β-anomer of GlcNAc was due to the slow mutarotation of α-D-GlcNAc, which likely explains the observed β-anomer in the product complex structure.

Mutation of Glu601 to alanine abolished activity as far as could be evaluated by our enzyme assays while the analogous substitution at Glu483 reduced the value of k_{cat} 10-fold ($3.2 \pm 0.4 \times 10^{-2}$ s^{-1}), and increased the value of K_m two-fold (2.3 ± 0.4 mM), thus decreasing k_{cat}/K_m by roughly 20 fold [$(1.4 \pm 0.3) \times 10^{-2}$ $mM^{-1}s^{-1}$] confirming the participation of these residues in catalysis. This apparent lack of activity of the Glu601Ala CpGH89 mutant also supports assignment of this residue as the catalytic nucleophile. This assignment is consistent with previous studies of other retaining glycoside hydrolases that reveal deletion of the catalytic nucleophile results in massive impairments in catalysis (typically $>10^3$-fold) (Tarling *et al*. 2003; Vallmitjana *et al*. 2001; Wan *et al*. 2007; Zechel *et al*. 2003). Mutagenesis of Glu483, on the other hand, results in a much smaller impairment in catalysis than the wild-type enzyme. Modest changes in the Michaelis-Menten parameters resulting from mutation of acid/base residues is consistent with cleavage of the glycosidic linkage of these activated aryl glycosides substrates not greatly benefiting from general acid catalysis (Vallmitjana *et al*.

2001; Wan *et al*. 2007; Zechel *et al*. 2003). Furthermore, Glu483 is further from C1 (approximately 3.6 Å) but is appropriately positioned to hydrogen bond with the axially oriented glycosidic oxygen of a bound substrate, suggesting that this group is the general acid/base catalytic residue. Given that the enzyme is a retaining α-glycosidase there is no precedent, and thus does it does not appear likely, that the 2-acetamido group could be productively involved in catalysis in a direct manner as a catalytic group. Indeed, there is no structural evidence to imply participation of the 2-acetamido group. Therefore, it is most likely that CpGH89, and by extension all members of GH89 (Henrissat and Davies 1997), use a double displacement retaining mechanism (Figure 30) (Davies and Henrissat 1995; Vasella *et al*. 2002; Zechel and Withers 2000). In CpGH89 we propose Glu483 as the general acid/base catalytic residue and Glu601 as the catalytic nucleophile, residues which are conserved across all known members of the GH89 family.

Insights into mucopolysaccharidosis IIIB

There is substantial genetic information describing mutations in HsGH89 gene that give rise to MPS-IIIB and some biochemical characterization of these mutant HsGH89 enzymes (Schmidtchen *et al*. 1998; Yogalingam *et al*. 2000). There is, however, no structure of HsGH89 or any homologue that can link this genetic information to the enzyme structure. CpGH89 has approximately 30% overall amino acid sequence identity to HsGH89 (Figure 31) allowing the construction of a complete model of HsGH89. This model shows the overall structural conservation (Figure 32A) while an examination of the active site residues of CpGH89 overlaid with the predicted HsGH89

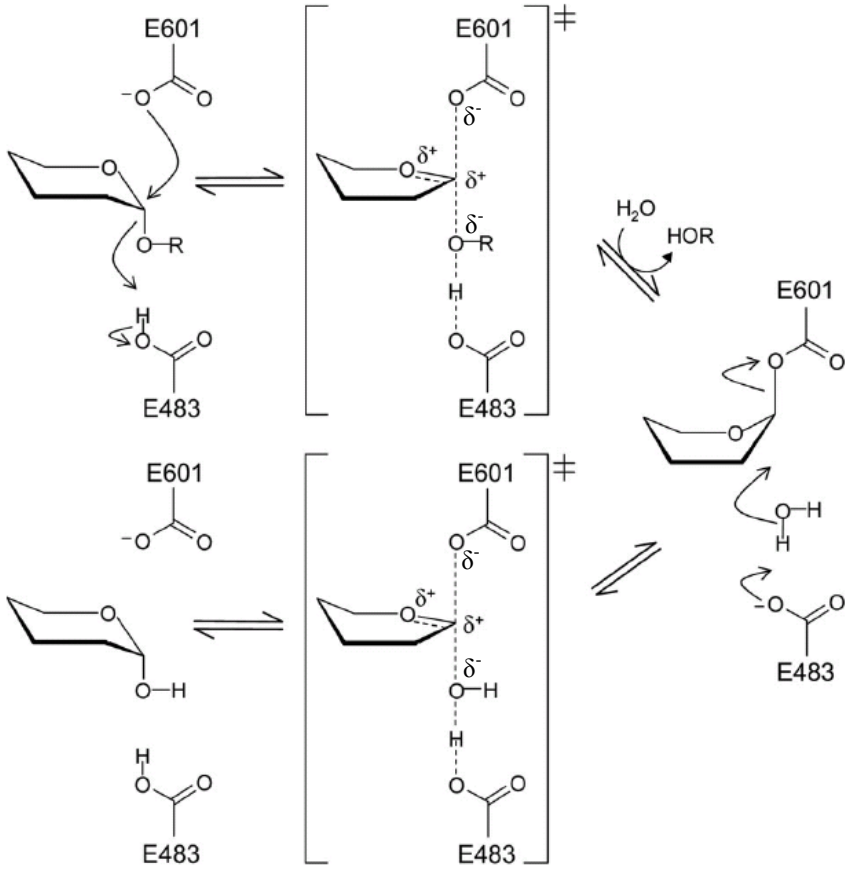


Figure 30. Schematic of the retaining catalytic mechanism used by CpGH89. Glu-483 acts as an acid in the first step, protonating the oxygen of the glycosidic bond as Glu-601 concomitantly targets the anomeric carbon in a nucleophilic attack. In the predicted oxocarbenium ion-like transition state, the anomeric carbon has sp2 hybridization, and a resonance stabilized partial positive charge develops at the endocyclic oxygen and the anomeric carbon. This property results in the formation of a covalent glycosyl-enzyme intermediate and the departure of the leaving group (HOR). Water is activated by Glu-483 and attacks at the anomeric center, forming a second oxocarbenium ion-like transition state. Bond cleavage thereby occurs via a retaining double displacement mechanism.

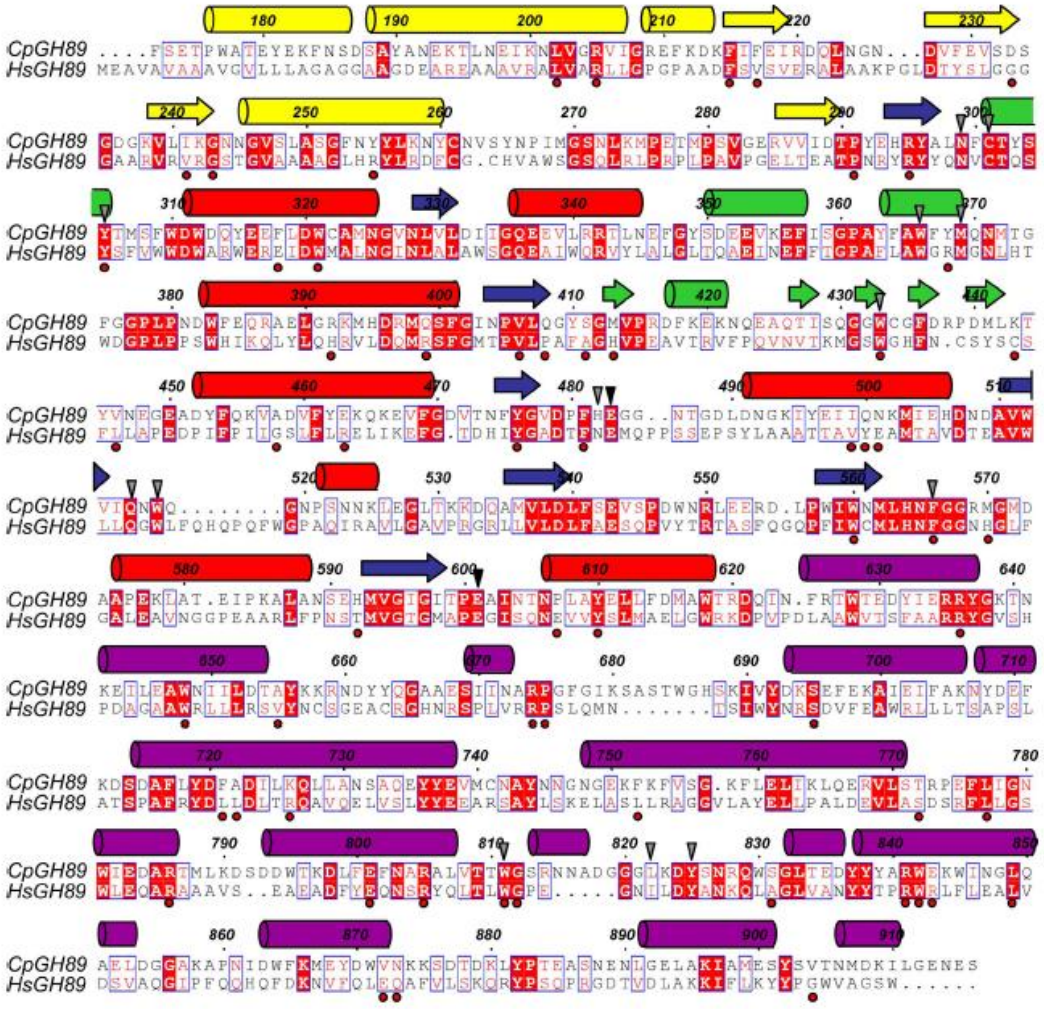


Figure 31. The amino acid alignment of CpGH89 with its human homolog, NAGLU (labeled here as HsGH89). The CBM32-like portion from *Cp*GH89 has been omitted because this module is not present in HsGH89. The secondary structure representation is given above the sequences, and coloring corresponds to the coloring in Figure 29A. Active site residues are indicated above the CpGH89 sequence with inverted triangles. The black triangles specifically represent the catalytic residues. Red circles underneath the NAGLU (HsGH89) sequence denote residues in the human sequence, where substitutions resulting in MPS IIIB have been noted in the literature.

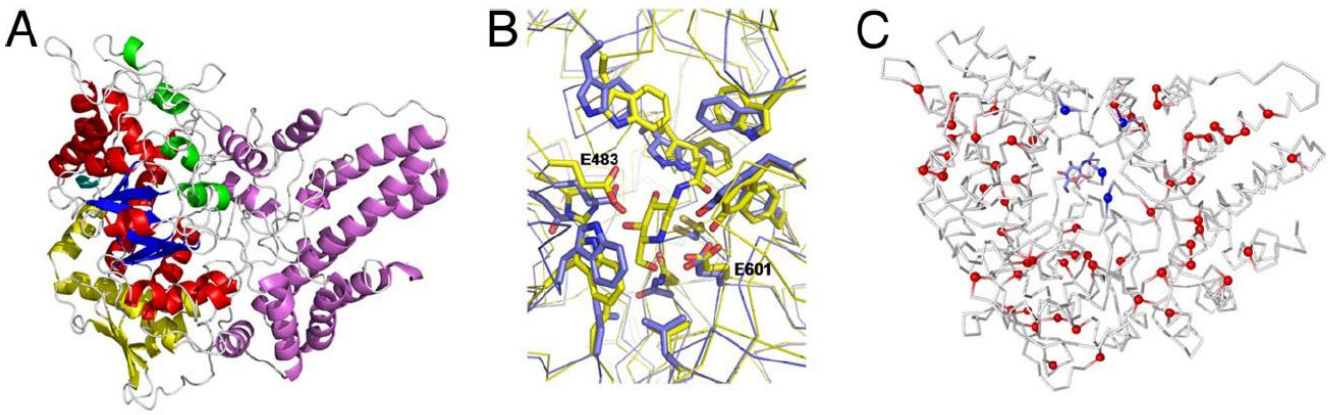


Figure 32. Structural location of naturally occurring mutations in NAGLU. (A) A cartoon representation of the homology model of NAGLU showing its overall fold. The coloring of the domains is as for CpGH89 in Figure 29A. (B) A structural overlay of the active site of CpGH89 and NAGLU. The catalytic residues are labeled. (C) A view of the ribbon trace of the NAGLU model with the structural location of the known mutations that lead to MPS IIIB. Sites of mutations are shown as spheres. Blue coloring indicates active site residues, whereas red coloring indicates non-active site residues.

active site shows that architecture of the active sites are almost entirely conserved (Figures 31 and 32B).The only residue that is not conserved is His482 in CpGH89, which is conservatively substituted by an asparagine in HsGH89 (Figure 32B). Accordingly, we believe that the active sites of the two enzymes are extremely similar and that CpGH89 is an excellent model of HsGH89 in terms of substrate recognition and catalysis, as well as overall fold.

CpGH89, and by extension HsGH89, are obligate *exo*-acting enzymes with an active site architectures that appear to have no ability to accommodate additional modifications, such as sulfations, to the terminal *N*-acetyl-D-glucosamine residue, a property that has been suggested in the literature but not experimentally demonstrated (Hopwood and Morris 1990). This strict specificity necessitates a route to generate non-reducing terminal α-*N*-acetylglucosamine residues on heparan. Thus, the degradation and recycling of heparan in the human body is initiated by the sequential action of three additional enzymes: heparan sulfate sulfamidase, acetylCoA:α-glucosaminidase-*N*-acetyltransferase, and *N*-acetyl-D-glucosamine 6-sulfatase (Hopwood and Morris 1990). These enzymes are responsible for generating non-reducing terminal *N*-acetyl-D-glucosamine residues and deficiency in the activity of one of these three enzymes results in mucopolysaccharidosis IIIA, C or D, respectively. Given the apparent strict requirement of the α-*N*-acetyl-D-glucosaminidase for unmodified non-reducing α-linked *N*-acetyl-D-glucosamine residues it becomes clear how a deficiency in any of these four enzymes could stall heparan sulfate depolymerization.

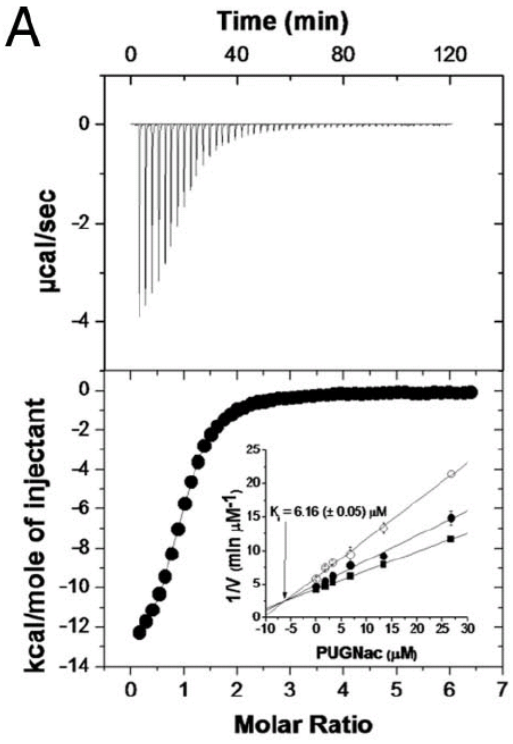
The mucopolysaccharidosis IIIB phenotype is associated with numerous missense, nonsense and deletion mutations, with the missense mutations far outnumbering the rest. Mapping the positions of these known missense mutations onto the model of HsGH89 is very revealing. These missense mutations are randomly scattered throughout the protein with only four mutations coinciding with residues in the active site (Figures 32C). However, there are no mutations of the catalytic residues, perhaps because these mutations are so deleterious as to be prenatally lethal. By analogy to other lysosomal storage disorders, such as Tay Sachs disease (Mahuran 1999), these mutations likely influence the protein by reducing its stability and resulting in less functional enzyme reaching the lysosome. Indeed, Yogalingam *et al* (Yogalingam *et al*. 2000) recently demonstrated that mutations in HsGH89 do indeed affect the stability and transiting of the enzyme through the secretory pathway. It is a well described phenomenon that many point mutations in proteins cause changes in folding that abolish proper trafficking (Bychkova and Ptitsyn 1995; Ellgaard *et al*. 1999). Ironically, in many cases, these physiologically disastrous mutations do not destroy the catalytic activity and may only cause slight folding anomalies (Welch and Brown 1996).

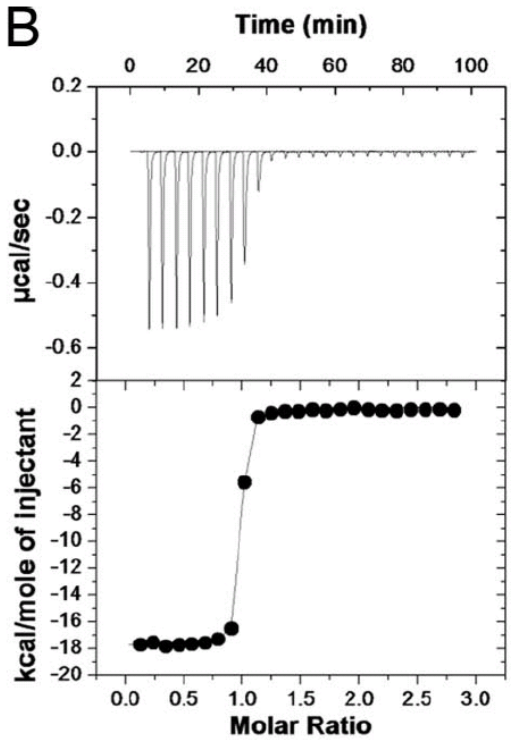
One approach for treating diseases, that stem from decreased protein stability, involves the use of small molecules (Amaral 2006). Compounds called chemical chaperones have been described that bind to the newly biosynthesized protein and thereby increase its stability. These ‘chaperoned’ mutant proteins are more stable and can therefore be transported out of the ER thereby avoiding intracellular degradation (Amaral 2006; Asano *et al*. 2001; Burrows *et al*. 2000; Fan *et al*. 1999; Gregersen 2006; Lieberman *et*

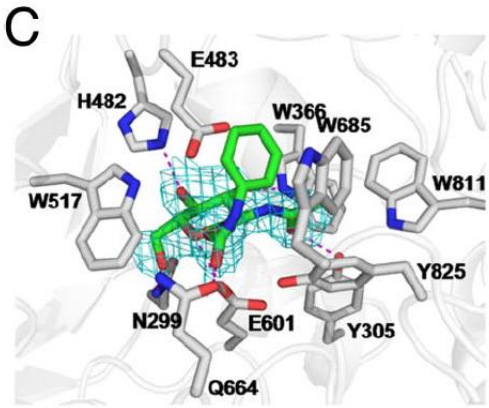
al. 2007; Sawkar *et al*. 2002; Steet *et al*. 2006; Tropak *et al*. 2004). With this in mind, we sought out potential CpGH89 inhibitors that may aid in identifying, designing and developing efficient chemical chaperones to treat mucopolysaccharidosis IIIB.

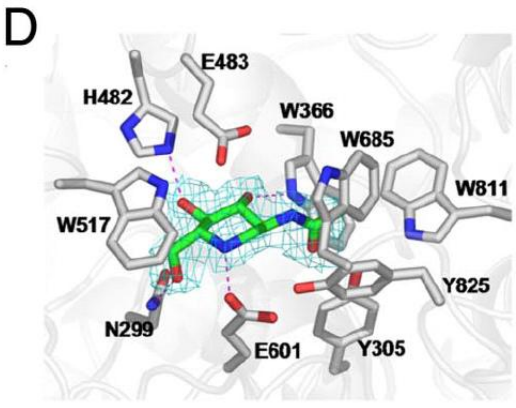
GH89 inhibitors – potential chemical chaperones?

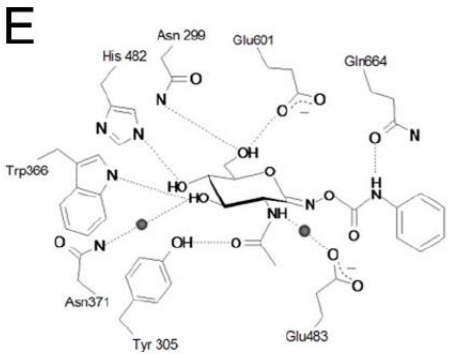
We assessed the binding of two known *N*-acetyl-D-glucosaminidase inhibitors, 2-acetamido-1,2-dideoxynojirimycin (2AcDNJ) (Horsch *et al*. 1991) and O-(2-acetamido-2-deoxy-D-glucopyranosylidene)amino-*N*-phenylcarbamate (PUGNAc) (Beer *et al*. 1990), to CpGH89 by isothermal titration calorimetry (ITC) and kinetic methods (Figure 33A and B). The K_i for PUGNAc was determined by Dixon plot analysis to be 6.2 ± 0.1 μM (Figure 33A inset), which agreed with the dissociation constant (K_d) of 5.5 ± 0.3 μM determined by ITC (Figure 33A). Since the artificial substrates available for CpGH89 are turned over quite slowly high concentrations of enzyme (67 nM) were required for kinetic assays therefore defining the lower limit at which K_i values could be measured. However, the dissociation constant for the tight binding inhibitor 2AcDNJ was determined by ITC to be 3.6 ± 0.3 nM, indicating approximately 1000-fold tighter binding of this inhibitor with CpGH89 relative to PUGNAc. Further, a pattern of competitive inhibition is observed for PUGNAc (Figure 33A inset) and since these two inhibitors bind to the same location (see structures below) it seems most likely that 2AcDNJ also binds in a competitive manner.


A
Time (min)
0
40
80
120
µcal/sec
0
-2
-4
kcal/mole of injectant
0
-2
-4
-6
-8
-10
-12
-14
0
1
2
3
4
5
6
7
Molar Ratio
1/V (min µM-1)
Ki = 6.16 (± 0.05) µM
PUGNac (µM)


B
Time (min)
0
20
40
60
80
100
µcal/sec
0.2
0.0
-0.2
-0.4
-0.6
kcal/mole of injectant
2
0
-2
-4
-6
-8
-10
-12
-14
-16
-18
-20
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Molar Ratio


C
E483
H482
W366
W685
W811
W517
Y825
N299
E601
Y305
Q664


D
E483
H482
W366
W685
W811
W517
Y825
N299
E601
Y305


E
Asn 299
Glu601
His 482
Gln664
Trp366
Asn371
Tyr 305
Glu483

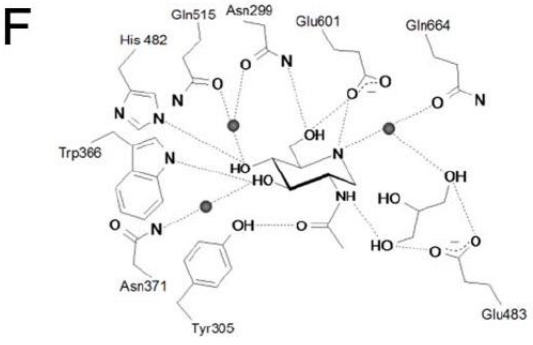

F
Gln515
Asn299
Glu601
Gln664
His 482
Trp366
Asn371
Tyr305
Glu483

Figure 33. Inhibitor binding by CpGH89. (A and B) Isotherms of CpGH89 binding to PUGNAc (A) and 2AcDNJ (B) produced by ITC. (*Upper*) Raw heat measurements. (*Lower*) Integrated heats. The solid lines show the fit of a one site binding model to the data. (A *Inset*) Dixon plot analysis of CpGH89 inhibition by PUGNAc. The intersection point on the graph corresponding to the *K*i (absolute value of the X value at the intersection) is indicated by an arrow. (C and D) Active site representations are shown for PUGNAc (C) and 2AcDNJ (D). The blue mesh shows the maximum likelihood(Murshudov et al. 1997)/σ_A (Read 1986)-weighted electron density maps contoured at 0.23 e-/Å^3 and 0.22 e-/Å^3 for 2AcDNJ and PUGNAc, respectively. Key active site residues, including the putative catalytic residues Glu-483 and Glu-601, are shown in stick representation and colored gray. Ligands are shown in green stick representation. Putative hydrogen bonds between the protein and ligand identified by using the criteria of proper geometry and a distance cutoff of 3.2 Å are shown as dotted magenta lines. (E and F) Schematics showing the interactions within the active site of CpGH89 with PUGNAc (E) and 2AcDNJ (F). A distance of 3.2 Å was used as the cutoff for significant hydrogen bonds. Water molecules are shown as shaded spheres. Protons on the amino acids are omitted for clarity.

Diffraction data from crystals soaked with 2AcDNJ or PUGNAc yielded excellent electron density for the sugar rings in the active site pocket (Figure 33 C and D). However, the phenyl ring substituent of PUGNAc was so disordered as to render its electron density undetectable (Figure 33C). Numerous polar interactions with the equatorial substituents at C3–C5 likely confer high specificity for the gluco-configured sugars (Figure 33 E and F). The basis for recognition of the 2-acetamido group also is clearly revealed in these complexes. A tight non-polar pocket formed by the side chains of Trp-366, Trp-685, Tyr-825, and Tyr-305 accommodates the *N*-acetyl group (Figure 33 C and D). Selectivity for the *N*-acetyl group of substrate and these inhibitors is likely imparted by the close van der Waals interactions between these aromatic residues and the substituent, as well as a favorably oriented 2.6-Å hydrogen bond between the carbonyl and Tyr-305 (Figure 33 C–F). The structural basis for the comparatively weak binding of PUGNAc is unclear, although it may result from difficulties of the active site in accommodating the phenylcarbamate group.

The structure of CpGH89 and the model of NAGLU clearly reveal that mutations leading to MPS IIIB are scattered throughout the tertiary structure of the enzyme, in keeping with the notion that the instability of mutated NAGLU can result in the MPS IIIB phenotype (Yogalingam *et al*. 2000). Mutant lysosomal enzymes that are unstable, including point mutants of NAGLU, can be retained in the endoplasmic reticulum by quality control mechanisms and thereby fail to traffic to lysosomes (Ellgaard *et al*. 1999; Fan *et al*. 1999; Yogalingam *et al*. 2000). When used at low concentrations, inhibitors of lysosomal glycosidases can be used as ‘‘chemical chaperones’’ to bind to and stabilize these mutant

enzymes, thereby facilitating their maturation and passage to lysosomes (Asano *et al*. 2001; Fan *et al*. 1999; Lieberman *et al*. 2007; Sawkar *et al*. 2002; Steet *et al*. 2006; Tropak *et al*. 2004). Once in the lysosome, these mutant enzymes have catalytic activities that support normal cellular functioning (Asano *et al*. 2001; Fan *et al*. 1999; Lieberman *et al*. 2007; Sawkar *et al*. 2002; Steet *et al*. 2006; Tropak *et al*. 2004). Accordingly, a chemical chaperone approach to treating MPS IIIB is likely viable, and our results indicate that 2AcDNJ is an effective inhibitor that may provide an initial lead for developing such chemical chaperones. Indeed, this inhibitor has also been shown to inhibit NAGLU (Zhao and Neufeld 2000) *in vitro* with nanomolar potencies (*K*i for 2AcDNJ=450 nM), in keeping with the results we observe here for CpGH89. These data, in combination with the near-complete conservation of active site residues, strongly support the validity of using the structure of CpGH89 as a guide for inhibitor design. However, the use of this particular compound may be limited because it is known to be a potent inhibitor of several other human enzymes. Nevertheless, there remains the potential to generate more selective inhibitors (Asano *et al*. 2001; Macauley *et al*. 2005; Ogawa *et al*. 2007).

In summary, the data presented here reveal the structural basis underlying the specificity of NAGLU for terminal non-reducing *N*-acetyl-D-glucosamine residues as well as its place within the glycosaminoglycan catabolic pathway. The extensive stereochemically specific interactions with the sugar hydroxyl groups, coupled with the sock-shaped architecture of the active site, strictly defines selectivity for the α-configured glycosidic

linkage of unornamented terminal *N*-acetyl-D-glucosamine residues. Further, the collective structural, mechanistic, and inhibition data presented here provide a clear template that will facilitate the design of effective chemical chaperones that specifically target NAGLU.

Chapter 4: Bacterial Pathogens and their Arsenals of Carbohydrate-Active Enzymes

Bacteria employ extensive systems of enzymes to help in the degradation and synthesis of saccharide structures. Plant cell wall degrading bacteria and bacteria that target the diverse glycans found in the animal body often have more complex systems of carbohydrate-active enzymes than their counterparts who use starch or chitin as their carbon source (Warren 1996). Plant cell wall degrading bacteria may have enzymes that act as pectinases, cellulases, xylanases, galacturonases, and more, as their target polysaccharides are cellulose, hemicellulose, pectin, lignin, and other components of the plant cell wall. The plant cell wall degrading glycoside hydrolases have been quite well studied and have given us insight into the modular nature of these enzymes. These enzymes have been shown to use CBMs to bring the catalytic modules into proximity of their ligand and to target the catalytic module to specific areas within a polysaccharide. Within the plant cell wall degrading Clostridia sp the clustering of glycoside hydrolases through protein-protein interactions demonstrates how grouping enzymes increases their synergistic effect on a substrate.

The genomes of most organisms on average contain between 1-3% carbohydrate-active enzymes (Davies *et al*. 2005). For example, the plant cell wall degrading *Xanthomonas campestris* has 153 enzymes involved in processing carbohydrates. These enzymes compose 1.16 % of the bacterial genome. Among the carbohydrate-active enzymes of *X. campestris* there are at least 9 cellulases, 5 xylanases, 5 pectate lyases and 4 pectin esterases (Agrios 2005; Cantarel *et al*. 2008). This is an example of a bacterial system which functions in the efficient degradation of plant cell walls.

Bacteria that are able to occupy the human gastro-intestinal tract tend to have a higher percentage of their genome devoted to carbohydrate-active enzymes (Davies *et al*. 2005). *B. thetaiotaomicron*, an animal commensal, uses an astonishing 6.6 % of its genome for carbohydrate-active enzymes. This organism is capable of digesting plant cell wall polysaccharides, and depending on growth conditions and polysaccharide availability, genes involved in the breakdown of host-glycans are induced (Salyers *et al*. 1977),(Salyers and Kotarski 1980; Sonnenburg *et al*. 2005). *B. thetaiotaomicron* has many large groups of paralogous genes involved in efficient and effective carbohydrate degradation for carbon and energy sources (Cantarel *et al*. 2008; Comstock and Coyne 2003) and differential induction of these genes involved in carbohydrate processing occurs depending on growth conditions and polysaccharide availability, thus, nutrient sensing allows this organism to utilize both dietary polysaccharide and/or to exploit host glycans (Comstock and Coyne 2003; Salyers and Kotarski 1980; Salyers *et al*. 1977). Analysis of the genome sequence of *B. thetaiotaomicron* reveals that its glycoside hydrolases are often modular and often contain the CBM32 motif common to the glycoside hydrolase enzymes of *C. perfringens*. Characterization of a family 84 glycoside hydrolase from *B. thetaiotaomicron* has revealed it to be similar in structure and function to CpGH84C (Dennis *et al*. 2006). Therefore these two gut microbes share at least some similarities between their carbohydrate-active enzymes, not surprising considering the ecological niche they both inhabit.

C. perfringens, thought by some to be the most widely distributed pathogen in nature (McDonel 1980; Shimizu *et al*. 2002), lacks the capacity to degrade dietary polysaccharides as the plant cell wall carbohydrate degrading machinery is absent in this organism; however, *C. perfringens* does employ a massive carbohydrate-active enzyme system for efficient and effective utilization of host carbohydrate sources and has demonstrated capability to degrade pig gastric and colonic mucins (Stanley *et al*. 1986). The *C. perfringens* histotoxic strain 13 devotes 3.7 % of its genome to carbohydrate-active enzymes. This reflects a larger proportion of carbohydrate active enzymes when compared to typical organisms which use 1-3%. The *C. perfringens* food poisoning strain SM101 devotes 2.7 % of its genome to carbohydrate-active enzymes and the *C. perfringens* gas gangrene isolate ATCC 13124 (Myers *et al*. 2006) devotes 2.2 % of its genome to carbohydrate-active enzymes; a relative paucity when compared to strain 13. Thus, even between biotypes of *C. perfringens* there is differences in the percentage of the genome devoted to carbohydrate-active enzymes. Figure 34 shows the breakdown in annotated carbohydrate-active enzymes between the three sequenced strains of *C. perfringens.* The differences in carbohydrate-active enzymes produced may reflect the ambition of the organism, whether it is to inhabited the gastrointestinal tract or for more invasive purposes. It is notable that the invasive strains (ATCC 13124 and strain 13) produce significantly more glycoside hydrolases than the food poisoning strain (SM101). We could postulate that the invasive strains require a more extensive system of

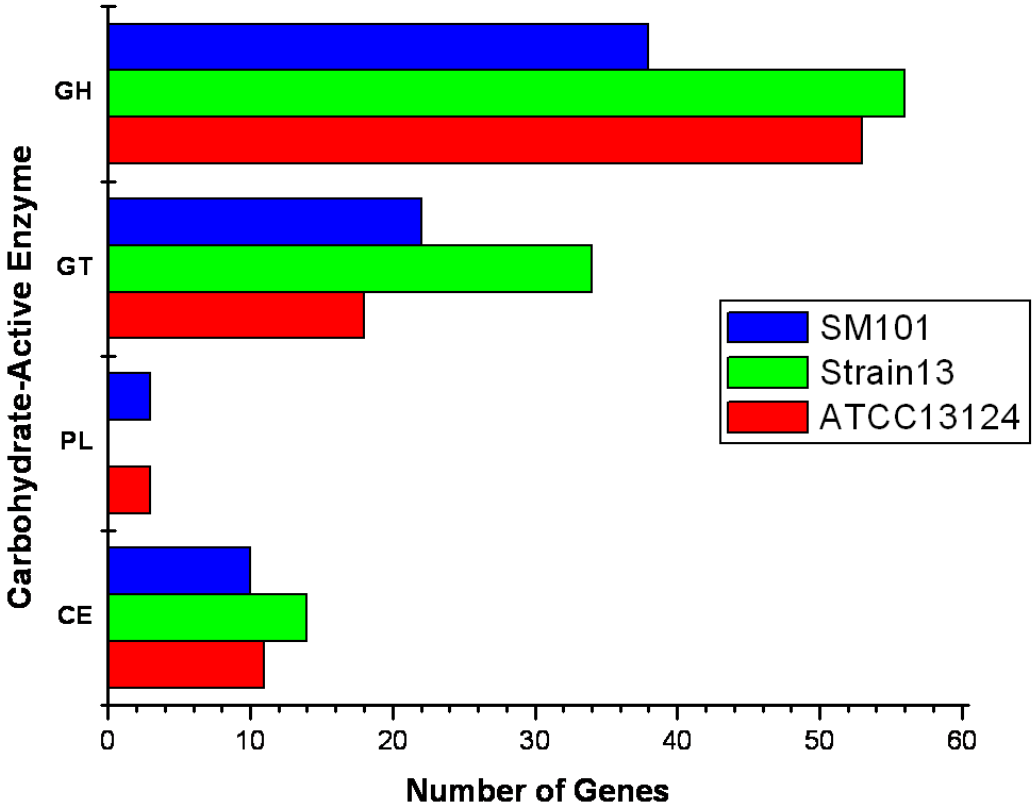


Figure 34. Carbohydrate-active enzymes and the number of annotated sequences found within the genome of *C. perfringens* strain SM101 (2723 total genes within the genome), strain 13 (2850 total genes within the genome), and strain ATCC 13124 (3018 total genes within the genome). GH= glycoside hydrolases, GT= glycosyl transferases, PL= polysaccharide lyases, CE= carbohydrate esterases.

glycoside hydrolases to tend to the diversity of glycans found throughout the animal body.

C. perfringens strain ATCC 13124 has 85 predicted enzymes involved in processing carbohydrate. 53 of these are glycoside hydrolases, 18 are glycosyltransferases, 3 are polysaccharide lyases and 11 are carbohydrate esterases. The genes for these enzymes compose 2.2 % of the bacterial genome. Among the carbohydrate-active enzyme of *C. perfringens* are known sialidases, an α-N-acetylglucosaminidase, a β-N-acetylglucosaminidase, and a hyaluronidase. Predicted enzyme specificities included, β-hexosaminidases, α-L-fucosidases, β-glucosidases, β-galactosidases, α-galactosidases, α-*N*-acetylgalacotosaminidases, heparan sulphate lyases and a hyaluronate lyase (Cantarel *et al*. 2008). This is an example of a bacterial system which is designed to function in the efficient degradation of complex animal glycans.

Animal glycans are structurally diverse containing a variety of monosaccharides as well as an assortment of glycosidic linkages between them. Glycosylation is abundant and for efficient degradation of complex carbohydrates and glycoconjugates an arsenal of enzymes with varying specificities is required, specificities which reflect their targeted carbon and energy source. Of the 53 known sequences for glycoside hydrolases in *C. perfringens* ATCC 13124 approximately half of these are predicted to be secreted and active on the complex glycans found in the animal host. Among these, to name a few, 2 are from family 2 with predicted β-galactosidase activity, 2 are from family 33 and act as sialidases, 4 are from family 84 (including CpGH84C) which are β-D-*N*-

acetylglucosaminidases, and only one member resides in family 89 (CpGH89) which has activity as an α-D-*N*-acetylglucosaminidase (Figure 8) (Chapter 3). Again we see repetition amongst the members of the same families composing these huge carbohydrate degrading enzyme systems. The remaining glycoside hydrolases are not predicted to be secreted and are likely involved in the latter stages of sugar processing for energy and carbon usage.

C. perfringens saccharolytic enzymes act to degrade larger more complex sugars into simpler sugars. The purpose for this is likely three fold. The first and perhaps most obvious is for consumption and utilization. These sugars are likely imported into the cell where they may be used for energy or other purposes. *C. perfringens* is an anaerobic fermenter which drives energy production. As well, this generates abundant gas production in the form of CO_2 and H_2 gas. This gas production in turn enhances the anaerobic environment which is the preferred growth environment for the organism. Secondly, hydrolysis may expose receptors for other enzymes. This is exemplified by the desialylation of the normally sialylated Thomsen-Friedenreich antigen on patient's erythrocytes leading to life threatening hemolysis (Batge *et al*. 1992; Klein *et al*. 1986; Placzek and Gorst 1987). The hydrolysis of sialic acids from glycans has been shown to increase a cells sensitivity to the effect of the α-toxin, which acts as a phospholipase C and sphingomyelinase, and is a major cytolytic toxin produced by *C. perfringens* (Boraston *et al*. 2007; Canard *et al*. 1994; Flores-Diaz *et al*. 2005). Finally, these enzymes may act as spreading factors, efficient tissue destruction may improve access for other toxins, thereby potentiating their effect and aiding in the spread of the organism.

The purpose of this research has been to structurally and functionally characterize two of the secreted glycoside hydrolases produced by *C. perfringens*. One predicted glycoside hydrolase virulence factor is CpGH89 (EC 3.2.1.50), a family 89 α-*N*-acetylglucosaminidase found not only in the myonecrotic *C. perfringens* strain ATCC 13124 and histo-toxic strain 13, but also in the enterotoxin-producing food poisoning strain SM101. CpGH89 is active on terminal α-GlcNAc moieties, which may be found decorating the mucins lining the intestinal tract, specifically those of class III mucins (Nakamura *et al*. 1998; Zhang *et al*. 2001), thus this enzyme may be involved in colonization of the intestine during infection (Nakayama *et al*. 1999). CpGH89 may also hydrolyze the α1,4 linked GlcNAc of heparan sulfate, a sulfated glycosaminoglycan and component of human connective tissue, in concert with other glycoside hydrolases that catalyze the hydrolysis of the β-1,4 linkage, such as a D-glucuronidase or L-iduronidase. Another glycoside hydrolase produced by these disease causing Clostridia is CpGH84C, which acts as an *exo*-β-*N*-acetylglucosaminidase. β-linked GlcNAc residues can be found modifying the surface of cytosolic and nuclear proteins in O-linked glycosylation (Comer and Hart 2000), in the pentasaccharide core of all N-linked glycans, $Man_3GlcNAc_2$(β-N)Asn, and in hybrid and complex N-linked glycans (Pratt and Bertozzi 2005) (Section 1.2.2). β-linked GlcNAc residues are found throughout the gastrointestinal mucins, as well as in the core structures of mucin, particularly core 2, 3, 4 and 6 (Pratt and Bertozzi 2005; Robbe *et al*. 2004) (Section 1.2.2). β-linked GlcNAc is also found in the human blood group antigens (Jenkins and O'Donnell 2006) and in the extracellular matrix as components of hyaluronan, heparan sulfate, and keratan sulfate

(See Section 1.2.2). These enzymes, with their "exotic" specificities, have potential substrates found throughout the human body. And in the event of infection, with the arsenal of glycoside hydrolases that *C. perfringens* produces, and its battery of toxins, this may contribute to the rapid tissue necrosis associated with both gas gangrene and necrotic enteritis.

The modularity seen in the Clostridial glycoside hydrolases is remarkable. CpGH84C is one of the simplest glycoside hydrolases produced by *C. perfringens* with modules shown to be involved in catalysis, carbohydrate-binding, and protein-protein interactions (Figure 8). CpGH89 has a more complex modular architecture, with six predicted family 32 carbohydrate binding modules as well as a cohesin module predicted to be involved in protein-protein interactions. The repetition of the family 32 CBM motif throughout these enzymes is significant, pointing to their importance in carbohydrate binding and enzyme activity. Roughly 1/3 of the bacterial enzymes containing family 32 CBMs belong to a varied range of glycoside hydrolases produced by the bacterial gut colonizers from species of *Clostridium* and *Bacteroides*. The role of these enzymes is likely in scavenging carbohydrates from dietary sugars or muco-oligosaccharides. By analogy to the plant cell wall glycoside hydrolases, the role of the CBM32s in these enzymes is probably to attach the enzyme to a carbohydrate-bearing surface, such as the mucin layer of the intestines, targeting the secreted glycoside hydrolases to their substrates. For "flesh-eating" strains of *C. perfringens* this function might be extended to binding and targeting enzyme to carbohydrate on tissue surfaces and within connective tissue. This evokes an image of the enzyme being attached to a carbohydrate-bearing surface *via* the

CBM while allowing the catalytic module to "graze" on that substrate. Given the diversity of enzymes which contain family 32 carbohydrate binding modules, one might expect, like with plant cell wall hydrolases and their CBMs, that the diversity of catalytic specificities is mirrored by the diversity of CBM32 specificities. The repetition of CBM motifs is also common to plant cell wall degrading glycoside hydrolases and multivalency within these enzymes has demonstrated, in some cases, an increase in affinity for polysaccharide. The CBM32 motif is replicated multiple times within many of the glycoside hydrolases of *C. perfringens* suggesting a mechanism of avid binding for these enzymes. Again, similar to the plant cell wall degrading glycoside hydrolases, we see further potential within this organism for synergistic attack on a carbohydrate substrate, mediated by association through protein-protein interactions (See section 2.4). The association of enzymes *via* cohesin-dockerin interactions and the predicted attachment of the enzymes to the cell wall through FN3 modules would maintain the enzymes in intimate contact and in association with the bacterium. The coordinated functions of carbohydrate-binding and potentially avid binding to ligand, protein-protein interaction and bacterial cell wall attachment may function to increase the effectiveness of the enzymes in a synchronized attack upon substrate thereby allowing more successful uptake of nutrients by the organism. In conclusion, the research described throughout this thesis has provided detailed characterization of three of the secreted glycoside hydrolase enzymes produced by *C. perfringens* and has contributed to our knowledge on protein-complex carbohydrate interactions in a bacterial pathogen.

Bibliography

Abbott DW, Eirin-Lopez JM, Boraston AB. 2008. Insight into ligand diversity and novel biological roles for family 32 carbohydrate-binding modules. Mol Biol Evol 25(1):155-67.

Abbott DW, Hrynuik S, Boraston AB. 2007. Identification and characterization of a novel periplasmic polygalacturonic acid binding protein from *Yersinia enterolitica*. J Mol Biol 367(4):1023-33.

Adams JJ, Gregg K, Bayer EA, Boraston AB, Smith SP. 2008. Structural basis of *Clostridium perfringens* toxin complex formation. Proc Natl Acad Sci U S A 105(34):12194-9.

Agrios GN. 2005. Genetics of Plant Disease. Plant Pathology: Academic Press. p 125-174.

Akimoto Y, Hart GW, Wells L, Vosseller K, Yamamoto K, Munetomo E, Ohara-Imaizumi M, Nishiwaki C, Nagamatsu S, Hirano H and others. 2007. Elevation of the post-translational modification of proteins by O-linked N-acetylglucosamine leads to deterioration of the glucose-stimulated insulin secretion in the pancreas of diabetic Goto-Kakizaki rats. Glycobiology 17(2):127-40.

Amaral MD. 2006. Therapy through chaperones: Sense or antisense? - Cystic fibrosis as a model disease. Journal of Inherited Metabolic Disease 29(2-3):477-487.

Anderson KM, Ashida H, Maskos K, Dell A, Li SC, Li YT. 2005. A clostridial endo-beta-galactosidase that cleaves both blood group A and B glycotopes: the first member of a new glycoside hydrolase family, GH98. J Biol Chem 280(9):7720-8.

Araki R, Ali MK, Sakka M, Kimura T, Sakka K, Ohmiya K. 2004. Essential role of the family-22 carbohydrate-binding modules for beta-1,3-1,4-glucanase activity of *Clostridium stercorarium* Xyn10B. FEBS Lett 561(1-3):155-8.

Asano N, Kato A, Watson AA. 2001. Therapeutic applications of sugar-mimicking glycosidase inhibitors. Mini Rev Med Chem 1(2):145-54.

Barik S. 1996. Site-directed mutagenesis in vitro by megaprimer PCR. Methods Mol Biol 57:203-15.

Batge B, Filejski W, Kurowski V, Kluter H, Djonlagic H. 1992. Clostridial sepsis with massive intravascular hemolysis: rapid diagnosis and successful treatment. Intensive Care Med 18(8):488-90.

Bayer EA, Shimon LJ, Shoham Y, Lamed R. 1998. Cellulosomes-structure and ultrastructure. J Struct Biol 124(2-3):221-34.

Beer D, Maloisel JL, Rast DM, Vasella A. 1990. Synthesis of 2-Acetamido-2-Deoxy-D-Gluconhydroximolactone-Derived and Chitobionhydroximolactone-Derived N-Phenylcarbamates, Potential Inhibitors of Beta-N-Acetylglucosaminidase. Helvetica Chimica Acta 73(7):1918-1922.

Beesley CE, Jackson M, Young EP, Vellodi A, Winchester BG. 2005. Molecular defects in Sanfilippo syndrome type B (mucopolysaccharidosis IIIB). J Inherit Metab Dis 28(5):759-67.

Beesley CE, Young EP, Vellodi A, Winchester BG. 1998. Identification of 12 novel mutations in the alpha-N-acetylglucosaminidase gene in 14 patients with Sanfilippo syndrome type B (mucopolysaccharidosis type IIIB). J Med Genet 35(11):910-4.

Bencharit S, Cui CB, Siddiqui A, Howard-Williams EL, Sondek J, Zuobi-Hasona K, Aukhil I. 2007. Structural insights into fibronectin type III domain-mediated signaling. J Mol Biol 367(2):303-9.

Bergmann A, Fritz, G. & Glatter, O. 2000. Solving the generalized indirect Fourier transformation (GIFT) by Boltzmann simplex simulated annealing (BSSA). J. Appl. Cryst.(33):1212-1216.

Bisig D, Weber P, Vaughan L, Winterhalter KH, Piontek K. 1999. Purification, crystallization and preliminary crystallographic studies of a two fibronectin type-III domain segment from chicken tenascin encompassing the heparin- and contactin-binding regions. Acta Crystallogr D Biol Crystallogr 55(Pt 5):1069-73.

Bolam DN, Ciruela A, McQueen-Mason S, Simpson P, Williamson MP, Rixon JE, Boraston A, Hazlewood GP, Gilbert HJ. 1998. Pseudomonas cellulose-binding domains mediate their effects by increasing enzyme substrate proximity. Biochem J 331 (Pt 3):775-81.

Bolam DN, Xie H, White P, Simpson PJ, Hancock SM, Williamson MP, Gilbert HJ. 2001. Evidence for synergy between family 2b carbohydrate binding modules in *Cellulomonas fimi* xylanase 11A. Biochemistry 40(8):2468-77.

Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. 2004. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem J 382(Pt 3):769-81.

Boraston AB, Creagh AL, Alam MM, Kormos JM, Tomme P, Haynes CA, Warren RA, Kilburn DG. 2001a. Binding specificity and thermodynamics of a family 9 carbohydrate-binding module from *Thermotoga maritima* xylanase 10A. Biochemistry 40(21):6240-7.

Boraston AB, Ficko-Blean E, Healey M. 2007. Carbohydrate recognition by a large sialidase toxin from *Clostridium perfringens*. Biochemistry 46(40):11352-60.

Boraston AB, Kwan E, Chiu P, Warren RA, Kilburn DG. 2003a. Recognition and hydrolysis of noncrystalline cellulose. J Biol Chem 278(8):6120-7.

Boraston AB, McLean BW, Chen G, Li A, Warren RA, Kilburn DG. 2002. Co-operative binding of triplicate carbohydrate-binding modules from a thermophilic xylanase. Mol Microbiol 43(1):187-94.

Boraston AB, Notenboom V, Warren RA, Kilburn DG, Rose DR, Davies G. 2003b. Structure and ligand binding of carbohydrate-binding module CsCBM6-3 reveals similarities with fucose-specific lectins and "galactose-binding" domains. J Mol Biol 327(3):659-69.

Boraston AB, Wang D, Burke RD. 2006. Blood group antigen recognition by a Streptococcus pneumoniae virulence factor. J Biol Chem 281(46):35263-71.

Boraston AB, Warren RA, Kilburn DG. 2001b. beta-1,3-Glucan binding by a thermostable carbohydrate-binding module from *Thermotoga maritima*. Biochemistry 40(48):14679-85.

Bouckaert J, Berglund J, Schembri M, De Genst E, Cools L, Wuhrer M, Hung CS, Pinkner J, Slattegard R, Zavialov A and others. 2005. Receptor binding studies disclose a novel class of high-affinity inhibitors of the *Escherichia coli* FimH adhesin. Mol Microbiol 55(2):441-55.

Brayman M, Thathiah A, Carson DD. 2004. MUC1: a multifunctional cell surface component of reproductive tissue epithelia. Reprod Biol Endocrinol 2:4.

Bricogne G, Vonrhein C, Flensburg C, Schiltz M, Paciorek W. 2003. Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. Acta Crystallogr D Biol Crystallogr 59(Pt 11):2023-30.

Brunger AT. 1993. Assessment of phase accuracy by cross validation: the free R value. Methods and applications. Acta Crystallogr D Biol Crystallogr 49(Pt 1):24-36.

Bryant AE, Bayer CR, Chen RY, Guth PH, Wallace RJ, Stevens DL. 2005. Vascular dysfunction and ischemic destruction of tissue in *Streptococcus pyogenes* infection: the role of streptolysin O-induced platelet/neutrophil complexes. J Infect Dis 192(6):1014-22.

Bunge S, Knigge A, Steglich C, Kleijer WJ, van Diggelen OP, Beck M, Gal A. 1999. Mucopolysaccharidosis type IIIB (Sanfilippo B): identification of 18 novel alpha-N-acetylglucosaminidase gene mutations. J Med Genet 36(1):28-31.

Burrows JA, Willis LK, Perlmutter DH. 2000. Chemical chaperones mediate increased secretion of mutant alpha 1-antitrypsin (alpha 1-AT) Z: A potential pharmacological

strategy for prevention of liver injury and emphysema in alpha 1-AT deficiency. Proc Natl Acad Sci U S A 97(4):1796-801.

Bychkova VE, Ptitsyn OB. 1995. Folding intermediates are involved in genetic diseases? FEBS Lett 359(1):6-8.

Canard B, Garnier T, Saint-Joanis B, Cole ST. 1994. Molecular genetic analysis of the nagH gene encoding a hyaluronidase of *Clostridium perfringens*. Mol Gen Genet 243(2):215-24.

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2008. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res.

Carrard G, Koivula A, Soderlund H, Beguin P. 2000. Cellulose-binding domains promote hydrolysis of different sites on crystalline cellulose. Proc Natl Acad Sci U S A 97(19):10342-7.

Centeno MS, Goyal A, Prates JA, Ferreira LM, Gilbert HJ, Fontes CM. 2006. Novel modular enzymes encoded by a cellulase gene cluster in *Cellvibrio mixtus*. FEMS Microbiol Lett 265(1):26-34.

Charnock SJ, Bolam DN, Turkenburg JP, Gilbert HJ, Ferreira LM, Davies GJ, Fontes CM. 2000. The X6 "thermostabilizing" domains of xylanases are carbohydrate-binding modules: structure and biochemistry of the *Clostridium thermocellum* X6b domain. Biochemistry 39(17):5013-21.

Chen HM, Withers SG. 2007. Facile synthesis of 2,4-dinitrophenyl alpha-D-glycopyranosides as chromogenic substrates for alpha-glycosidases. Chembiochem 8(7):719-22.

Chitayat S, Adams JJ, Furness HS, Bayer EA, Smith SP. 2008a. The solution structure of the C-terminal modular pair from *Clostridium perfringens* mu-toxin reveals a noncellulosomal dockerin module. J Mol Biol 381(5):1202-12.

Chitayat S, Gregg K, Adams JJ, Ficko-Blean E, Bayer EA, Boraston AB, Smith SP. 2008b. Three-dimensional structure of a putative non-cellulosomal cohesin module from a *Clostridium perfringens* family 84 glycoside hydrolase. J Mol Biol 375(1):20-8.

Chou TY, Hart GW. 2001. O-linked N-acetylglucosamine and cancer: messages from the glycosylation of c-Myc. Adv Exp Med Biol 491:413-8.

Comer FI, Hart GW. 2000. O-Glycosylation of nuclear and cytosolic proteins. Dynamic interplay between O-GlcNAc and O-phosphate. J Biol Chem 275(38):29179-82.

Comstock LE, Coyne MJ. 2003. Bacteroides thetaiotaomicron: a dynamic, niche-adapted human symbiont. Bioessays 25(10):926-9.

Connaris H, Crocker PR, Taylor GL. 2009. Enhancing the receptor affinity of the sialic acid-binding domain of *Vibrio cholerae* sialidase through multivalency. J Biol Chem.

Coutinho PMH, B. 1999. Carbohydrate-active enzymes: an integrated database approach. In: H.J. Gilbert GD, B. Henrissat and B. Svensson, editor. Recent Advances in Carbohydrate Bioengineering". Cambridge: The Royal Society of Chemistry. p 3-12.

Cowtan K. 1994 dm: An automated procedure for phase improvement by density modification. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography:34-38.

Cowtan K, Main P. 1998. Miscellaneous algorithms for density modification. Acta Crystallogr D Biol Crystallogr 54(Pt 4):487-93.

Davies G, Henrissat B. 1995. Structures and mechanisms of glycosyl hydrolases. Structure 3(9):853-9.

Davies GJ, Gloster TM, Henrissat B. 2005. Recent structural insights into the expanding world of carbohydrate-active enzymes. Curr Opin Struct Biol 15(6):637-45.

Dennis RJ, Taylor EJ, Macauley MS, Stubbs KA, Turkenburg JP, Hart SJ, Black GN, Vocadlo DJ, Davies GJ. 2006. Structure and mechanism of a bacterial beta-glucosaminidase having O-GlcNAcase activity. Nat Struct Mol Biol 13(4):365-71.

Deva T, Baker EN, Squire CJ, Smith CA. 2006. Structure of *Escherichia coli* UDP-N-acetylmuramoyl:L-alanine ligase (MurC). Acta Crystallogr D Biol Crystallogr 62(Pt 12):1466-74.

Din N, Damude HG, Gilkes NR, Miller RC, Jr., Warren RA, Kilburn DG. 1994. C1-Cx revisited: intramolecular synergism in a cellulase. Proc Natl Acad Sci U S A 91(24):11383-7.

Duo-Chuan L. 2006. Review of fungal chitinases. Mycopathologia 161(6):345-60.

Einerhand AW, Renes IB, Makkink MK, van der Sluis M, Buller HA, Dekker J. 2002. Role of mucins in inflammatory bowel disease: important lessons from experimental models. Eur J Gastroenterol Hepatol 14(7):757-65.

Ellgaard L, Molinari M, Helenius A. 1999. Setting the standards: quality control in the secretory pathway. Science 286(5446):1882-8.

Emre S, Terzioglu M, Tokatli A, Coskun T, Ozalp I, Weber B, Hopwood JJ. 2002. Sanfilippo syndrome in Turkey: Identification of novel mutations in subtypes A and B. Hum Mutat 19(2):184-5.

Emsley P, Cowtan K. 2004. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 60(Pt 12 Pt 1):2126-32.

Ernst PB, Gold BD. 2000. The disease spectrum of *Helicobacter pylori*: the immunopathogenesis of gastroduodenal ulcer and gastric cancer. Annu Rev Microbiol 54:615-40.

Fan JQ, Ishii S, Asano N, Suzuki Y. 1999. Accelerated transport and maturation of lysosomal alpha-galactosidase A in Fabry lymphoblasts by an enzyme inhibitor. Nat Med 5(1):112-5.

Ficko-Blean E, Boraston AB. 2005. Cloning, recombinant production, crystallization and preliminary X-ray diffraction studies of a family 84 glycoside hydrolase from *Clostridium perfringens*. Acta Crystallogr Sect F Struct Biol Cryst Commun 61(Pt 9):834-6.

Ficko-Blean E, Boraston AB. 2006. The interaction of a carbohydrate-binding module from a *Clostridium perfringens* N-acetyl-beta-hexosaminidase with its carbohydrate receptor. J Biol Chem 281(49):37748-57.

Ficko-Blean E, Stubbs KA, Nemirovsky O, Vocadlo DJ, Boraston AB. 2008. Structural and mechanistic insight into the basis of mucopolysaccharidosis IIIB. Proc Natl Acad Sci U S A 105(18):6560-5.

Figura N. 1997. *Helicobacter pylori* factors involved in the development of gastroduodenal mucosal damage and ulceration. J Clin Gastroenterol 25 Suppl 1:S149-63.

Flatt JP. 1995. Use and Storage of Carbohydrate and Fat. American Journal of Clinical Nutrition 61(4):S952-S959.

Flores-Diaz M, Alape-Giron A, Clark G, Catimel B, Hirabayashi Y, Nice E, Gutierrez JM, Titball R, Thelestam M. 2005. A cellular deficiency of gangliosides causes hypersensitivity to *Clostridium perfringens* phospholipase C. J Biol Chem 280(29):26680-9.

Fowler M, Thomas RJ, Atherton J, Roberts IS, High NJ. 2006. Galectin-3 binds to *Helicobacter pylori* O-antigen: it is upregulated and rapidly secreted by gastric epithelial cells in response to H. pylori adhesion. Cell Microbiol 8(1):44-54.

Galen JE, Ketley JM, Fasano A, Richardson SH, Wasserman SS, Kaper JB. 1992. Role of *Vibrio cholerae* neuraminidase in the function of cholera toxin. Infect Immun 60(2):406-15.

Gao PJ, Chen GJ, Wang TH, Zhang YS, Liu J. 2001. Non-hydrolytic Disruption of Crystalline Structure of Cellulose by Cellulose Binding Domain and Linker Sequence of Cellobiohydrolase I from *Penicillium janthinellum*. Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai) 33(1):13-18.

Gaskell A, Crennell S, Taylor G. 1995. The three domains of a bacterial sialidase: a beta-propeller, an immunoglobulin module and a galactose-binding jelly-roll. Structure 3(11):1197-205.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. 2003. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31(13):3784-8.

Gasteiger E. HC, Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. 2005. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker JM, editor. The Proteomics Protocols Handbook: Humana Press. p 571-607.

Geigenberger P, Kolbe A, Tiessen A. 2005. Redox regulation of carbon storage and partitioning in response to light and sugars. Journal of Experimental Botany 56(416):1469-1479.

Gerding SJaDN. 1997. Enterotoxemic Infections. In: Julian Rood BAM, J. Glenn Songer, Richard W. Titball, editor. The Clostridia Molecular Biology and Pathogenesis. London: Harcourt Brace & Company. p 117-140.

Gerken TA, Hagen KG, Jamison O. 2008. Conservation of peptide acceptor preferences between Drosophila and mammalian polypeptide-GalNAc transferase ortholog pairs. Glycobiology 18(11):861-70.

Gregersen N. 2006. Protein misfolding disorders: pathogenesis and intervention. J Inherit Metab Dis 29(2-3):456-70.

Gregg KJ, Finn R, Abbott DW, Boraston AB. 2008. Divergent modes of glycan recognition by a new family of carbohydrate-binding modules. J Biol Chem 283(18):12604-13.

Guinier AF, F. 1955. Small angle scattering of X-rays. New York: Wiley Interscience.

Gupta D, Kaltner H, Dong X, Gabius HJ, Brewer CF. 1996. Comparative cross-linking activities of lactose-specific plant and animal lectins and a natural lactose-binding immunoglobulin G fraction from human serum with asialofetuin. Glycobiology 6(8):843-9.

Gut H, King SJ, Walsh MA. 2008. Structural and functional studies of *Streptococcus pneumoniae* neuraminidase B: An intramolecular trans-sialidase. FEBS Lett.

Hacker J. 1990. Genetic determinants coding for fimbriae and adhesins of extraintestinal *Escherichia coli*. Curr Top Microbiol Immunol 151:1-27.

Hammel M, Fierobe HP, Czjzek M, Finet S, Receveur-Brechot V. 2004. Structural insights into the mechanism of formation of cellulosomes probed by small angle X-ray scattering. J Biol Chem 279(53):55985-94.

Hansen SM, Kohler LB, Li S, Kiselyov V, Christensen C, Owczarek S, Bock E, Berezin V. 2008. NCAM-derived peptides function as agonists for the fibroblast growth factor receptor. J Neurochem 106(5):2030-41.

Hatheway CL. 1990. Toxigenic clostridia. Clin Microbiol Rev 3(1):66-98.

Henrissat B, Davies G. 1997. Structural and sequence-based classification of glycoside hydrolases. Curr Opin Struct Biol 7(5):637-44.

Ho JJ, Jaituni RS, Crawley SC, Yang SC, Gum JR, Kim YS. 2003. N-glycosylation is required for the surface localization of MUC17 mucin. Int J Oncol 23(3):585-92.

Hong TY, Cheng CW, Huang JW, Meng M. 2002. Isolation and biochemical characterization of an endo-1,3-beta-glucanase from *Streptomyces sioyaensis* containing a C-terminal family 6 carbohydrate-binding module that binds to 1,3-beta-glucan. Microbiology 148(Pt 4):1151-9.

Hooper LV, Gordon JI. 2001. Commensal host-bacterial relationships in the gut. Science 292(5519):1115-8.

Hopwood JJ, Morris CP. 1990. The mucopolysaccharidoses. Diagnosis, molecular genetics and treatment. Mol Biol Med 7(5):381-404.

Horsch M, Hoesch L, Vasella A, Rast DM. 1991. N-acetylglucosaminono-1,5-lactone oxime and the corresponding (phenylcarbamoyl)oxime. Novel and potent inhibitors of beta-N-acetylglucosaminidase. Eur J Biochem 197(3):815-8.

Hounsell EF, Feizi T. 1982. Gastrointestinal mucins. Structures and antigenicities of their carbohydrate chains in health and disease. Med Biol 60(5):227-36.

Hynes W. 2004. Virulence factors of the group A streptococci and genes that regulate their expression. Front Biosci 9:3399-433.

Ichikawa T, Itano N, Sawai T, Kimata K, Koganehira Y, Saida T, Taniguchi S. 1999. Increased synthesis of hyaluronate enhances motility of human melanoma cells. J Invest Dermatol 113(6):935-9.

Ito N, Phillips SE, Stevens C, Ogel ZB, McPherson MJ, Keen JN, Yadav KD, Knowles PF. 1991. Novel thioether bond revealed by a 1.7 A crystal structure of galactose oxidase. Nature 350(6313):87-90.

Jabbar AA, Kazarian T, Hakobyan N, Valentino LA. 2006. Gangliosides promote platelet adhesion and facilitate neuroblastoma cell adhesion under dynamic conditions simulating blood flow. Pediatr Blood Cancer 46(3):292-9.

Jenkins PV, O'Donnell JS. 2006. ABO blood group determines plasma von Willebrand factor levels: a biologic function after all? Transfusion 46(10):1836-44.

Katayama T, Fujita K, Yamamoto K. 2005. Novel bifidobacterial glycosidases acting on sugar chains of mucin glycoproteins. J Biosci Bioeng 99(5):457-65.

Katayama T, Sakuma A, Kimura T, Makimura Y, Hiratake J, Sakata K, Yamanoi T, Kumagai H, Yamamoto K. 2004. Molecular cloning and characterization of *Bifidobacterium bifidum* 1,2-alpha-L-fucosidase (AfcA), a novel inverting glycosidase (glycoside hydrolase family 95). J Bacteriol 186(15):4885-93.

Kiessling LL, Pohl NL. 1996. Strength in numbers: non-natural polyvalent carbohydrate derivatives. Chem Biol 3(2):71-7.

King SJ, Hippe KR, Weiser JN. 2006. Deglycosylation of human glycoconjugates by the sequential activities of exoglycosidases expressed by *Streptococcus pneumoniae.* Mol Microbiol 59(3):961-74.

Klein RL, Novak RW, Novak PE. 1986. T-cryptantigen exposure in neonatal necrotizing enterocolitis. J Pediatr Surg 21(12):1155-8.

Klomp LW, van Rens L, Strous GJ. 1994. Identification of a human gastric mucin precursor: N-linked glycosylation and oligomerization. Biochem J 304 (Pt 3):693-8.

Kosaki R, Watanabe K, Yamaguchi Y. 1999. Overproduction of hyaluronan by expression of the hyaluronan synthase Has2 enhances anchorage-independent growth and tumorigenicity. Cancer Res 59(5):1141-5.

Lammerts van Bueren A, Boraston AB. 2004. Binding sub-site dissection of a carbohydrate-binding module reveals the contribution of entropy to oligosaccharide recognition at "non-primary" binding subsites. J Mol Biol 340(4):869-79.

Landry RM, An D, Hupp JT, Singh PK, Parsek MR. 2006. Mucin-*Pseudomonas aeruginosa* interactions promote biofilm formation and antibiotic resistance. Mol Microbiol 59(1):142-51.

Laskowski RA, Macarthur, M. W., Moss, D. S. & Thornton, J. M. . 1993. Procheck - a Program to Check the Stereochemical Quality of Protein Structures. Journal of Applied Crystallography Journal of Applied Crystallography (26):283-291.

Leahy DJ, Aukhil I, Erickson HP. 1996. 2.0 A crystal structure of a four-domain segment of human fibronectin encompassing the RGD loop and synergy region. Cell 84(1):155-64.

Leahy DJ, Erickson HP, Aukhil I, Joshi P, Hendrickson WA. 1994. Crystallization of a fragment of human fibronectin: introduction of methionine by site-directed mutagenesis to allow phasing *via* selenomethionine. Proteins 19(1):48-54.

Lee JY, Spicer AP. 2000. Hyaluronan: a multifunctional, megaDalton, stealth molecule. Curr Opin Cell Biol 12(5):581-6.

Lelwala-Guruge J, Ljungh A, Wadstrom T. 1992. Haemagglutination patterns of *Helicobacter pylori*. Frequency of sialic acid-specific and non-sialic acid-specific haemagglutinins. Apmis 100(10):908-13.

Letts JA, Rose NL, Fang YR, Barry CH, Borisova SN, Seto NO, Palcic MM, Evans SV. 2006. Differential recognition of the type I and II H antigen acceptors by the human ABO(H) blood group A and B glycosyltransferases. J Biol Chem 281(6):3625-32.

Li DH, Tung JW, Tarner IH, Snow AL, Yukinari T, Ngernmaneepothong R, Martinez OM, Parnes JR. 2006. CD72 Down-Modulates BCR-Induced Signal Transduction and Diminishes Survival in Primary Mature B Lymphocytes. J Immunol 176(9):5321-8.

Lieberman RL, Wustman BA, Huertas P, Powe AC, Jr., Pine CW, Khanna R, Schlossmacher MG, Ringe D, Petsko GA. 2007. Structure of acid beta-glucosidase with pharmacological chaperone provides insight into Gaucher disease. Nat Chem Biol 3(2):101-7.

Lindhorst TK, Kieburg C, Krallmann-Wenzel U. 1998. Inhibition of the type 1 fimbriae-mediated adhesion of *Escherichia coli* to erythrocytes by multiantennary alpha-mannosyl clusters: the effect of multivalency. Glycoconj J 15(6):605-13.

Liu F, Iqbal K, Grundke-Iqbal I, Hart GW, Gong CX. 2004. O-GlcNAcylation regulates phosphorylation of tau: a mechanism involved in Alzheimer's disease. Proc Natl Acad Sci U S A 101(29):10804-9.

Lloyd KO. 2000. The chemistry and immunochemistry of blood group A, B, H, and Lewis antigens: past, present and future. Glycoconj J 17(7-9):531-41.

Lodge JA, Maier T, Liebl W, Hoffmann V, Strater N. 2003. Crystal structure of *Thermotoga maritima* alpha-glucosidase AglA defines a new clan of NAD+-dependent glycosidases. J Biol Chem 278(21):19151-8.

Macauley MS, Whitworth GE, Debowski AW, Chin D, Vocadlo DJ. 2005. O-GlcNAcase uses substrate-assisted catalysis: kinetic analysis and development of highly selective mechanism-inspired inhibitors. J Biol Chem 280(27):25313-22.

Mahuran DJ. 1999. Biochemical consequences of mutations causing the GM2 gangliosidoses. Biochim Biophys Acta 1455(2-3):105-38.

Marionneau S, Cailleau-Thomas A, Rocher J, Le Moullac-Vaidye B, Ruvoen N, Clement M, Le Pendu J. 2001. ABH and Lewis histo-blood group antigens, a model for the meaning of oligosaccharide diversity in the face of a changing world. Biochimie 83(7):565-73.

Mark BL, Vocadlo DJ, Knapp S, Triggs-Raine BL, Withers SG, James MN. 2001. Crystallographic evidence for substrate-assisted catalysis in a bacterial beta-hexosaminidase. J Biol Chem 276(13):10330-7.

Martinez-Fleites C, Macauley MS, He Y, Shen DL, Vocadlo DJ, Davies GJ. 2008. Structure of an O-GlcNAc transferase homolog provides insight into intracellular glycosylation. Nat Struct Mol Biol 15(7):764-5.

McClane B. 2000. The action, genetics and synthesis of *Clostridium perfringens* enterotoxin. In: J. Cary MS, D. Bhatnager, editor. Microbial Foodborne Diseases: Mechanisms of Pathogenesis and Toxin Synthesis. lancaster, Pa: Technomic Press. p 247-272.

McClane B. 2001. *Clostridium perfringens*. In: M. Doyle LB, and T. Montville, editor. Food Microbiology: Fundamentals and Frontiers Second Edition. Washington DC: ASM Press. p 351-372.

McCool DJ, Okada Y, Forstner JF, Forstner GG. 1999. Roles of calreticulin and calnexin during mucin synthesis in LS180 and HT29/A1 human colonic adenocarcinoma cells. Biochem J 341 (Pt 3):593-600.

McDonel JL. 1980. *Clostridium perfringens* toxins (type A, B, C, D, E). Pharmacol Ther 10(3):617-55.

McDonel JL. 1986. Toxins of Clostridium perfringens types A,B,C,D and E. In: Drews FDaJ, editor. Pharmacology of Bacterial Toxins. Oxford: Permagon Press. p 477–517.

Moran AP. 2008. Relevance of fucosylation and Lewis antigen expression in the bacterial gastroduodenal pathogen *Helicobacter pylori*. Carbohydr Res 343(12):1952-65.

Morelle W, Haslam SM, Ziak M, Roth J, Morris HR, Dell A. 2000. Characterization of the N-linked oligosaccharides of megalin (gp330) from rat kidney. Glycobiology 10(3):295-304.

Morris RJ, Perrakis A, Lamzin VS. 2002. ARP/wARP's model-building algorithms. I. The main chain. Acta Crystallogr D Biol Crystallogr 58(Pt 6 Pt 2):968-75.

Moustafa I, Connaris H, Taylor M, Zaitsev V, Wilson JC, Kiefel MJ, von Itzstein M, Taylor G. 2004a. Sialic acid recognition by *Vibrio cholerae* neuraminidase. Journal of Biological Chemistry 279(39):40819-40826.

Moustafa I, Connaris H, Taylor M, Zaitsev V, Wilson JC, Kiefel MJ, von Itzstein M, Taylor G. 2004b. Sialic acid recognition by *Vibrio cholerae* neuraminidase. J Biol Chem 279(39):40819-26.

Murshudov GN, Vagin AA, Dodson EJ. 1997. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53(Pt 3):240-55.

Mutwil M, Debolt S, Persson S. 2008. Cellulose synthesis: a complex complex. Curr Opin Plant Biol 11(3):252-7.

Myers GS, Rasko DA, Cheung JK, Ravel J, Seshadri R, DeBoy RT, Ren Q, Varga J, Awad MM, Brinkac LM and others. 2006. Skewed genomic variability in strains of the toxigenic bacterial pathogen, *Clostridium perfringens*. Genome Res 16(8):1031-40.

Nagahama M, Sakaguchi Y, Kobayashi K, Ochi S, Sakurai J. 2000. Characterization of the enzymatic component of *Clostridium perfringens* iota-toxin. J Bacteriol 182(8):2096-103.

Nakamura N, Ota H, Katsuyama T, Akamatsu T, Ishihara K, Kurihara M, Hotta K. 1998. Histochemical reactivity of normal, metaplastic, and neoplastic tissues to alpha-linked N-acetylglucosamine residue-specific monoclonal antibody HIK1083. J Histochem Cytochem 46(7):793-801.

Nakayama J, Yeh JC, Misra AK, Ito S, Katsuyama T, Fukuda M. 1999. Expression cloning of a human alpha1, 4-N-acetylglucosaminyltransferase that forms GlcNAcalpha1-->4Galbeta-->R, a glycan specifically expressed in the gastric gland mucous cell-type mucin. Proc Natl Acad Sci U S A 96(16):8991-6.

Newstead SL, Watson JN, Bennet AJ, Taylor G. 2005. Galactose recognition by the carbohydrate-binding module of a bacterial sialidase. Acta Crystallogr D Biol Crystallogr 61(Pt 11):1483-91.

Ogawa S, Kanto M, Suzuki Y. 2007. Development and medical application of unsaturated carbaglycosylamine glycosidase inhibitors. Mini Rev Med Chem 7(7):679-91.

Perrakis A, Morris R, Lamzin VS. 1999. Automated protein model building combined with iterative structure refinement. Nature Structural Biology 6(5):458-463.

Peters D. 2006. Carbohydrates for fermentation. Biotechnol J 1(7-8):806-14.

Petit L, Gibert M, Gillet D, Laurent-Winter C, Boquet P, Popoff MR. 1997. *Clostridium perfringens* epsilon-toxin acts on MDCK cells by forming a large membrane complex. J Bacteriol 179(20):6480-7.

Petit L, Gibert M, Gourch A, Bens M, Vandewalle A, Popoff MR. 2003. *Clostridium perfringens* epsilon toxin rapidly decreases membrane barrier permeability of polarized MDCK cells. Cell Microbiol 5(3):155-64.

Petit L, Maier E, Gibert M, Popoff MR, Benz R. 2001. *Clostridium perfringens* epsilon toxin induces a rapid change of cell membrane permeability to ions and forms channels in artificial lipid bilayers. J Biol Chem 276(19):15736-40.

Petoukhov MV, Eady, N. A., Brown, K. A. & Svergun, D. I. 2002. Biophys. J.(83):3113-3125.

Pflugrath JW. 1999. The finer things in X-ray diffraction data collection. Acta Crystallogr D Biol Crystallogr 55(Pt 10):1718-25.

Placzek MM, Gorst DW. 1987. T activation haemolysis and death after blood transfusion. Arch Dis Child 62(7):743-4.

Ponyi T, Szabo L, Nagy T, Orosz L, Simpson PJ, Williamson MP, Gilbert HJ. 2000. Trp22, Trp24, and Tyr8 play a pivotal role in the binding of the family 10 cellulose-binding module from *Pseudomonas xylanase* A to insoluble ligands. Biochemistry 39(5):985-91.

Pratt MR, Bertozzi CR. 2005. Synthetic glycopeptides and glycoproteins as tools for biology. Chemical Society Reviews 34(1):58-68.

Putnam CD, Hammel M, Hura GL, Tainer JA. 2007. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q Rev Biophys 40(3):191-285.

Raman J, Fritz TA, Gerken TA, Jamison O, Live D, Liu M, Tabak LA. 2008. The catalytic and lectin domains of UDP-GalNAc:polypeptide alpha-N-Acetylgalactosaminyltransferase function in concert to direct glycosylation site selection. J Biol Chem 283(34):22942-51.

Ramphal R, Arora SK. 2001. Recognition of mucin components by *Pseudomonas aeruginosa*. Glycoconj J 18(9):709-13.

Rao FV, Dorfmueller HC, Villa F, Allwood M, Eggleston IM, van Aalten DM. 2006. Structural insights into the mechanism and inhibition of eukaryotic O-GlcNAc hydrolysis. Embo J 25(7):1569-78.

Ravn V, Dabelsteen E. 2000. Tissue distribution of histo-blood group antigens. APMIS 108(1):1-28.

Raymond MN, Le Stunff H. 2006. Involvement of de novo ceramide biosynthesis in macrophage death induced by activation of ATP-sensitive P2X7 receptor. FEBS Lett 580(1):131-6.

Read RJ. 1986. Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Crystallographica Section A A42:140-149.

Rempel BP, Clarke LA, Withers SG. 2005. A homology model for human alpha-l-iduronidase: insights into human disease. Mol Genet Metab 85(1):28-37.

Robbe C, Capon C, Coddeville B, Michalski JC. 2004. Structural diversity and specific distribution of O-glycans in normal human mucins along the intestinal tract. Biochem J 384(Pt 2):307-16.

Robinson J, Goodwin CS, Cooper M, Burke V, Mee BJ. 1990. Soluble and cell-associated haemagglutinins of *Helicobacter (Campylobacter) pylori*. J Med Microbiol 33(4):277-84.

Roman A Laskowski MWM, David S Moss and Janet M Thornton. 1993. J. App. Cryst(26):283.

Rood JI, Cole ST. 1991. Molecular genetics and pathogenesis of *Clostridium perfringens*. Microbiol Rev 55(4):621-48.

Salyers AA, Kotarski SF. 1980. Induction of chondroitin sulfate lyase activity in *Bacteroides thetaiotaomicron*. J Bacteriol 143(2):781-8.

Salyers AA, Palmer JK, Wilkins TD. 1977. Laminarinase (beta-glucanase) activity in Bacteroides from the human colon. Appl Environ Microbiol 33(5):1118-24.

Sato S, Hughes RC. 1994. Regulation of secretion and surface expression of Mac-2, a galactoside-binding protein of macrophages. J Biol Chem 269(6):4424-30.

Sawkar AR, Cheng WC, Beutler E, Wong CH, Balch WE, Kelly JW. 2002. Chemical chaperones increase the cellular activity of N370S beta -glucosidase: a therapeutic strategy for Gaucher disease. Proc Natl Acad Sci U S A 99(24):15428-33.

Scharfman A, Arora SK, Delmotte P, Van Brussel E, Mazurier J, Ramphal R, Roussel P. 2001. Recognition of Lewis x derivatives present on mucins by flagellar components of *Pseudomonas aeruginosa*. Infect Immun 69(9):5243-8.

Schmidtchen A, Greenberg D, Zhao HG, Li HH, Huang Y, Tieu P, Zhao HZ, Cheng S, Zhao Z, Whitley CB and others. 1998. NAGLU mutations underlying Sanfilippo syndrome type B. Am J Hum Genet 62(1):64-9.

Schneider TR, Sheldrick GM. 2002. Substructure solution with SHELXD. Acta Crystallographica Section D-Biological Crystallography 58:1772-1779.

Schwede T, Kopp J, Guex N, Peitsch MC. 2003. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 31(13):3381-5.

Sears CL. 2005. A dynamic partnership: celebrating our gut flora. Anaerobe 11(5):247-51.

Shatursky O, Bayles R, Rogers M, Jost BH, Songer JG, Tweten RK. 2000. *Clostridium perfringens* beta-toxin forms potential-dependent, cation-selective channels in lipid bilayers. Infect Immun 68(10):5546-51.

Shelburne SA, Davenport MT, Keith DB, Musser JM. 2008. The role of complex carbohydrate catabolism in the pathogenesis of invasive streptococci. Trends Microbiol 16(7):318-25.

Sheldon WL, Macauley MS, Taylor EJ, Robinson CE, Charnock SJ, Davies GJ, Vocadlo DJ, Black GW. 2006. Functional analysis of a group A streptococcal glycoside hydrolase Spy1600 from family 84 reveals it is a beta-N-acetylglucosaminidase and not a hyaluronidase. Biochem J 399(2):241-7.

Shimizu T, Ohshima S, Ohtani K, Hayashi H. 2001. Genomic map of *Clostridium perfringens* strain 13. Microbiol Immunol 45(2):179-89.

Shimizu T, Ohtani K, Hirakawa H, Ohshima K, Yamashita A, Shiba T, Ogasawara N, Hattori M, Kuhara S, Hayashi H. 2002. Complete genome sequence of *Clostridium perfringens,* an anaerobic flesh-eater. Proc Natl Acad Sci U S A 99(2):996-1001.

Shukla D, Spear PG. 2001. Herpesviruses and heparan sulfate: an intimate relationship in aid of viral entry. J Clin Invest 108(4):503-10.

Simon GL, Gorbach SL. 1986. The human intestinal microflora. Dig Dis Sci 31(9 Suppl):147S-162S.

Slawson C, Housley MP, Hart GW. 2006. O-GlcNAc cycling: How a single sugar post-translational modification is changing the way we think about signaling networks. Journal of Cellular Biochemistry 97(1):71-83.

Smedley JG, 3rd, Fisher DJ, Sayeed S, Chakrabarti G, McClane BA. 2004. The enteric toxins of *Clostridium perfringens.* Rev Physiol Biochem Pharmacol 152:183-204.

Smith L. 1975. Clostridial wound infections. In: Smith LDS, editor. Pathogenic Anaerobic Bacteria p321-324.

Sonnenburg JL, Xu J, Leip DD, Chen CH, Westover BP, Weatherford J, Buhler JD, Gordon JI. 2005. Glycan foraging in vivo by an intestine-adapted bacterial symbiont. Science 307(5717):1955-9.

Stanley RA, Ram SP, Wilkinson RK, Roberton AM. 1986. Degradation of pig gastric and colonic mucins by bacteria isolated from the pig colon. Appl Environ Microbiol 51(5):1104-9.

Steet RA, Chung S, Wustman B, Powe A, Do H, Kornfeld SA. 2006. The iminosugar isofagomine increases the activity of N370S mutant acid beta-glucosidase in Gaucher fibroblasts by several mechanisms. Proc Natl Acad Sci U S A 103(37):13813-8.

Stevens D. 1997. Necrotizing soft tissue infections. In: Rood Jea, editor. The Clostridia Molecular Biology and Pathogenesis. p 141-151.

Stevens DL, Bryant AE. 2002. The role of clostridial toxins in the pathogenesis of gas gangrene. Clin Infect Dis 35(Suppl 1):S93-S100.

Strous GJ, Dekker J. 1992. Mucin-type glycoproteins. Crit Rev Biochem Mol Biol 27(1-2):57-92.

Svergun D. 1992. Determination of the regularization Parameter in Indirect-Transform Methods using perceptual criteria. J. Appl. Cryst.(25):495-503.

Svergun DI, Petoukhov, M. V. & Koch, M. H. 2001. Determination of Domain Structure of Proteins from X-ray Solution Scattering. Biophys. J.(80):2946-2953.

Tanaka A, Kimura M, Lan HT, Takaura N, Yamano T. 2002. Molecular analysis of the alpha-N-acetylglucosaminidase gene in seven Japanese patients from six unrelated families with mucopolysaccharidosis IIIB (Sanfilippo type B), including two novel mutations. J Hum Genet 47(9):484-7.

Tarling CA, He S, Sulzenbacher G, Bignon C, Bourne Y, Henrissat B, Withers SG. 2003. Identification of the catalytic nucleophile of the family 29 alpha-L-fucosidase from *Thermotoga maritima* through trapping of a covalent glycosyl-enzyme intermediate and mutagenesis. J Biol Chem 278(48):47394-9.

Tessitore A, Villani GR, Di Domenico C, Filocamo M, Gatti R, Di Natale P. 2000. Molecular defects in the alpha-N-acetylglucosaminidase gene in Italian Sanfilippo type B patients. Hum Genet 107(6):568-76.

Thobhani S, Ember B, Siriwardena A, Boons GJ. 2003. Multivalency and the mode of action of bacterial sialidases. J Am Chem Soc 125(24):7154-5.

Thon L, Mohlig H, Mathieu S, Lange A, Bulanova E, Winoto-Morbach S, Schutze S, Bulfone-Paus S, Adam D. 2005. Ceramide mediates caspase-independent programmed cell death. Faseb J 19(14):1945-56.

Thornton DJ, Sheehan JK. 2004. From mucins to mucus: toward a more coherent understanding of this essential barrier. Proc Am Thorac Soc 1(1):54-61.

Titball RW, Naylor CE, Basak AK. 1999. The *Clostridium perfringens* alpha-toxin. Anaerobe 5(2):51-64.

Tomme P, Creagh AL, Kilburn DG, Haynes CA. 1996. Interaction of polysaccharides with the N-terminal cellulose-binding domain of *Cellulomonas fimi* CenC. 1. Binding specificity and calorimetric analysis. Biochemistry 35(44):13885-94.

Tomme P, Van Tilbeurgh H, Pettersson G, Van Damme J, Vandekerckhove J, Knowles J, Teeri T, Claeyssens M. 1988. Studies of the cellulolytic system of *Trichoderma reesei* QM 9414. Analysis of domain function in two cellobiohydrolases by limited proteolysis. Eur J Biochem 170(3):575-81.

Tropak MB, Reid SP, Guiral M, Withers SG, Mahuran D. 2004. Pharmacological enhancement of beta-hexosaminidase activity in fibroblasts from adult Tay-Sachs and Sandhoff Patients. J Biol Chem 279(14):13478-87.

Tsuge H, Nagahama M, Nishimura T, Sakaguchi Y, Katunuma N, Sakurai J. 1999. Crystallization and preliminary X-ray studies of the Ia component of *Clostridium perfringens* iota toxin complexed with NADPH. J Struct Biol 126(2):175-7.

Turnbull WB, Daranas AH. 2003. On the value of c: can low affinity systems be studied by isothermal titration calorimetry? J Am Chem Soc 125(48):14859-66.

Vaaje-Kolstad G, Houston DR, Riemen AH, Eijsink VG, van Aalten DM. 2005. Crystal structure and binding properties of the *Serratia marcescens* chitin-binding protein CBP21. J Biol Chem 280(12):11313-9.

Vagin A, Teplyakov A. 1997. MOLREP: an automated program for molecular replacement. Journal of Applied Crystallography 30:1022-1025.

Vagin AT, A. 1997. J. Appl. Cryst.(30):1022±1025.

Vaguine AA, Richelle J, Wodak SJ. 1999. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallogr D Biol Crystallogr 55(Pt 1):191-205.

Vallmitjana M, Ferrer-Navarro M, Planell R, Abel M, Ausin C, Querol E, Planas A, Perez-Pons JA. 2001. Mechanism of the family 1 beta-glucosidase from Streptomyces sp: catalytic residues and kinetic studies. Biochemistry 40(20):5975-82.

van Aalten DM, Bywater R, Findlay JB, Hendlich M, Hooft RW, Vriend G. 1996. PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. J Comput Aided Mol Des 10(3):255-62.

van de Kamp JJ, Niermeijer MF, von Figura K, Giesberts MA. 1981. Genetic heterogeneity and clinical variability in the Sanfilippo syndrome (types A, B, and C). Clin Genet 20(2):152-60.

Vantilbeurgh H, Tomme P, Claeyssens M, Bhikhabhai R, Pettersson G. 1986. Limited Proteolysis of the Cellobiohydrolase I from *Trichoderma-Reesei* - Separation of Functional Domains. Febs Letters 204(2):223-227.

Vasella A, Davies GJ, Bohm M. 2002. Glycosidase mechanisms. Curr Opin Chem Biol 6(5):619-29.

Vocadlo DJ, Withers SG. 2005. Detailed comparative analysis of the catalytic mechanisms of beta-N-acetylglucosaminidases from families 3 and 20 of glycoside hydrolases. Biochemistry 44(38):12809-18.

Vogan CL, Powell A, Rowley AF. 2008. Shell disease in crustaceans - just chitin recycling gone wrong? Environ Microbiol 10(4):826-35.

Volkov VVS, D. I. 2003. Uniqueness of ab initio shape determination in small-angle scattering. Journal of Applied Crystallography (36):860-864.

Wan CF, Chen WH, Chen CT, Chang MD, Lo LC, Li YK. 2007. Mutagenesis and mechanistic study of a glycoside hydrolase family 54 alpha-L-arabinofuranosidase from *Trichoderma koningii*. Biochem J 401(2):551-8.

Warren RA. 1996. Microbial hydrolysis of polysaccharides. Annu Rev Microbiol 50:183-212.

Watts AG, Damager I, Amaya ML, Buschiazzo A, Alzari P, Frasch AC, Withers SG. 2003. *Trypanosoma cruzi* trans-sialidase operates through a covalent sialyl-enzyme intermediate: tyrosine is the catalytic nucleophile. J Am Chem Soc 125(25):7532-3.

Weber B, Guo XH, Kleijer WJ, van de Kamp JJ, Poorthuis BJ, Hopwood JJ. 1999. Sanfilippo type B syndrome (mucopolysaccharidosis III B): allelic heterogeneity corresponds to the wide spectrum of clinical phenotypes. Eur J Hum Genet 7(1):34-44.

Weidenmaier C, Peschel A. 2008. Teichoic acids and related cell-wall glycopolymers in Gram-positive physiology and host interactions. Nat Rev Microbiol 6(4):276-87.

Welch WJ, Brown CR. 1996. Influence of molecular and chemical chaperones on protein folding. Cell Stress Chaperones 1(2):109-15.

Wilcox WR. 2004. Lysosomal storage disorders: the need for better pediatric recognition and comprehensive care. J Pediatr 144(5 Suppl):S3-14.

Wiseman T, Williston S, Brandts JF, Lin LN. 1989. Rapid measurement of binding constants and heats of binding using a new titration calorimeter. Anal Biochem 179(1):131-7.

Wolfenden R, Lu XD, Young G. 1998. Spontaneous hydrolysis of glycosides. Journal of the American Chemical Society 120(27):6814-6815.

Yamaoka Y. 2008. Roles of *Helicobacter pylori* BabA in gastroduodenal pathogenesis. World J Gastroenterol 14(27):4265-72.

Yang F, West AP, Jr., Allendorph GP, Choe S, Bjorkman PJ. 2008. Neogenin interacts with hemojuvelin through its two membrane-proximal fibronectin type III domains. Biochemistry 47(14):4237-45.

Yip VL, Varrot A, Davies GJ, Rajan SS, Yang X, Thompson J, Anderson WF, Withers SG. 2004. An unusual mechanism of glycoside hydrolysis involving redox and elimination steps by a family 4 beta-glycosidase from *Thermotoga maritima.* J Am Chem Soc 126(27):8354-5.

Yogalingam G, Hopwood JJ. 2001. Molecular genetics of mucopolysaccharidosis type IIIA and IIIB: Diagnostic, clinical, and biological implications. Hum Mutat 18(4):264-81.

Yogalingam G, Weber B, Meehan J, Rogers J, Hopwood JJ. 2000. Mucopolysaccharidosis type IIIB: characterisation and expression of wild-type and mutant recombinant alpha-N-acetylglucosaminidase and relationship with sanfilippo phenotype in an attenuated patient. Biochim Biophys Acta 1502(3):415-25.

Zechel DL, Reid SP, Stoll D, Nashiru O, Warren RA, Withers SG. 2003. Mechanism, mutagenesis, and chemical rescue of a beta-mannosidase from *Cellulomonas fimi*. Biochemistry 42(23):7195-204.

Zechel DL, Withers SG. 2000. Glycosidase mechanisms: anatomy of a finely tuned catalyst. Acc Chem Res 33(1):11-8.

Zhang MX, Nakayama J, Hidaka E, Kubota S, Yan J, Ota H, Fukuda M. 2001. Immunohistochemical demonstration of alpha1,4-N-acetylglucosaminyltransferase that forms GlcNAcalpha1,4Galbeta residues in human gastrointestinal mucosa. J Histochem Cytochem 49(5):587-96.

Zhao HG, Li HH, Bach G, Schmidtchen A, Neufeld EF. 1996. The molecular basis of Sanfilippo syndrome type B. Proc Natl Acad Sci U S A 93(12):6101-5.

Zhao KW, Neufeld EF. 2000. Purification and characterization of recombinant human alpha-N-acetylglucosaminidase secreted by Chinese hamster ovary cells. Protein Expr Purif 19(1):202-11.

Zhou G, Mo WJ, Sebbel P, Min G, Neubert TA, Glockshuber R, Wu XR, Sun TT, Kong XP. 2001. Uroplakin Ia is the urothelial receptor for uropathogenic *Escherichia coli:* evidence from in vitro FimH binding. J Cell Sci 114(Pt 22):4095-103.

Zwart PH, Afonine PV, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, McKee E, Moriarty NW, Read RJ, Sacchettini JC and others. 2008. Automated structure solution with the PHENIX suite. Methods Mol Biol 426:419-35.

Appendix A: NMR

^{1}H NMR experiments were performed by Dr. Keith Stubbs and Dr. David Vocadlo[2].

[1]Department of Chemistry, The University of Western Australia 35 Stirling Highway Crawley WA 6009, Australia. [2] Department of Chemistry, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada

^{1}H NMR spectroscopy (600 MHz Bruker AMX spectrometer) was employed to identify the products of the enzyme-catalyzed reaction. The reaction was carried out in approximately 0.6 mL (25 °C, 0.1 M Gly-Gly pD 7.7) containing 3 mM dNP-α-GlcNAc (Chen and Withers 2007). Initiation of the reaction was done by addition of 30 μL of a 70 μM stock of the enzyme, CpGH89. The hydrolysis of the dNP-α-GlcNAc was monitored until the reaction reached equilibrium. Spectral data were collected with 64000 data points (64k) over a spectral width of 7211 Hz (12.01 ppm), with a relaxation delay of 1 s and 32 scans. An initial spectrum (referred to as time zero) containing substrate and buffer was acquired before the addition of enzyme.

Appendix B: Molecular Modeling

Homology model was built by Dr. Alisdair B. Boraston[1].

[1]Biochemistry & Microbiology, University of Victoria, PO Box 3055 STN CSC, Victoria, BC, V8W 3P6, Canada.

A homology model of HsGH89 was constructed from the CpGH89-product complex coordinates using methods essentially identical to those described previously (Rempel *et al*. 2005) using DeepView and the Swiss-Model Automated Comparative Protein Modeling Server (Schwede *et al*. 2003). The resulting model was subjected to ten rounds of model idealization with REFMAC and the output model manually examined residue-by-residue. The model was validated by tools within COOT and PROCHECK (Roman A Laskowski 1993).