Carbohydrate Processing by Bacterial Pathogens: Structural and Functional Analyses of Glycoside Hydrolases

by

Katie Jean Gregg
BSc, University of Victoria, 2005

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

Carbohydrate Processing by Bacterial Pathogens: Structural and Functional Analyses of Glycoside Hydrolases

by

Katie Jean Gregg
BSc, University of Victoria, 2005

Supervisory Committee

Dr. Alisdair B. Boraston (Department of Biochemistry and Microbiology)
Supervisor

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

Dr. Martin Boulanger (Department of Biochemistry and Microbiology)
Departmental Member

Dr. John S.Taylor (Department of Biology)
Outside Member

Abstract

Supervisory Committee
Dr. Alisdair B. Boraston (Department of Biochemistry and Microbiology)
Supervisor
Dr. Paul J. Romaniuk (Department of Biochemistry and Microbiology)
Departmental Member
Dr. Martin Boulanger (Department of Biochemistry and Microbiology)
Departmental Member
Dr. John S. Taylor (Department of Biology)
Outside Member

Carbohydrates are important in a large number of cellular, physiological, and pathological processes. Carbohydrates often function as the human host's first line of defence against pathogen invasion by coating surfaces of epithelial cells and as glycan-rich mucins which line the entrances to the body. Various pathogenic bacteria exploit their hosts by modifying their glycans through the production of carbohydrate-active enzymes. Two kinds of pathogenic bacteria that are notable for their production of carbohydrate-active enzymes are *Streptococcus pneumoniae* and *Clostridium perfringens*. Both *S. pneumoniae* and *C. perfringens* inhabit glycan-rich niches in the human body, the respiratory and gastrointestinal tracts, respectively. To properly colonize their human hosts both bacteria have developed an extensive repertoire of glycoside hydrolases (GHs) which are enzymes responsible for the breakdown of carbohydrates. These GHs have known or predicted specificities for human glycans, specifically those found in mucins. We chose *C. perfringens* and *S. pneumoniae* as model systems to study these enzymes due to their large complements of GHs, many of which are known virulence factors. The objectives are to probe the key features of the GHs from these two different kinds of bacteria that inhabit similar human niches and to study catalysis, modularity and overall enzyme structure. This work uses a multidisciplinary approach and provides molecular level insight into the *S. pneumoniae* and *C. perfringens* host-pathogen interaction.

Table of Contents

List of Tables

List of Figures

Dedication

This is for you mom and dad

Love, Katie

Chapter 1: General Introduction

1.1 Carbohydrates.

Carbohydrates are important in a large number of cellular, physiological, and pathological processes. The term carbohydrate was coined over one hundred years ago and literally refers to "hydrates of carbon". This term describes naturally occurring substances that have the formula $C_x (H_2O)_n$,where $_x$ does not necessarily equal $_n$, and which possess a carbonyl group. Carbohydrates are composed of the simplest of polyhydroxylated carbonyl compounds, monosaccharides, which have either an aldehyde group at the end of the hydroxylated carbon chain or an inner chain ketone. Monosaccharides cannot be hydrolyzed into simpler forms and can exist in open-chain or ring forms. Monosaccharides can join together through a glycosidic linkage to form oligosaccharides, with usually 2-20 monosaccharides, or into longer polymer chains termed polysaccharides. The monosaccharide building blocks have immense combinatorial diversity generated by the many different possible linkages, branch points and modifications that can form complex sugar structures. All cells in nature are covered in mono-, oligo- and polysaccharides which are generically referred to as glycans. The glycans that collectively cover a cell are referred to as the glycocalyx (Pries et al., 2000). The potential diversity of glycans that constitute the glycocalyx of a single cell only represents a very small portion of the overall glycan diversity available in nature.

Glycans have diverse functions ranging from non-essential roles to roles critical for proper development and function of an organism and for the organisms survival. Polysaccharides serve for not only the storage of energy, such as starch and glycogen, but also as structural components, such as cellulose in plants and chitin in arthropods. The monosaccharide ribose is very important in coenzymes, such as ATP, FAD and NAD, and in genetic molecules. Not only are glycans crucial for the function of humans and other multicellular organisms but are also important in interactions between hosts and symbionts, such as bacteria (Varki, 1993).

It is also very common for carbohydrates to be attached to non-carbohydrate macromolecules. In these cases, glycans consisting of one or more monosaccharide are attached covalently to a non-carbohydrate moiety such as proteins or lipids. In order to classify glycocongugates into more manageable groupings, they are described based on how and what type of moiety they are attached to, such as glycoproteins being proteins with glycans attached or glycolipids being lipids with attached glycans. Glycocongugates can have varying degrees of glycosylation whereby the glycan can contribute very little to the overall size of the molecule or can constitute a major portion of the overall mass. In fact, it is very common for the glycan portion of a glycocongugate to constitute the dominant portion (Varki et al., 2009).

1.1.1 Major Classes of Eukaryotic Glycocongugates and Glycans.

The following is an overview of the major classes of eukaryotic glycocongugates. The classes discussed here represent the most common classes but there are many other less common types not discussed here such as endoplasmic reticulum/golgi and nucleocytoplasmic glycosylations.

Proteoglycans are a class of glycocongugates that have glycosaminoglycan chains covalently attached via a xylose residue to a core protein through the hydroxyl group of a serine residue. Glycosaminoglycans are linear polysaccharides that consist of a repeating disaccharide unit of an *N*-acetylgalactosamine or *N*-acetylglucosamine and an uronic acid or galactose. Membrane proteoglycans can span the plasma membrane or be linked by a glycosylphospatidylinositol anchor (see below) or be secreted. Proteoglycans have diverse functions including being a major component of the extracellular matrix, contributing to the formation of basement membranes, acting as receptors, as well as having many more functions (Varki et al., 2009).

Glycosphingolipids are a class of glycolipids that consist of glycans attached to the terminal hydroxyl group of a lipid through a glucose or galactose residue. The lipid moiety is called a ceramide and consists of a long chain amino alcohol, called sphingosine, linked to a fatty acid. The ceramide can vary structurally in level of hydroxylation, saturation as well as length. The ceramide is typically linked to either

glucose or galactose through a β-linkage and is further decorated with other glucose and galactose residues as well as *N*-acetylgalactosamine. Glycosylation is often capped with one or more sialic acid residues. Some glycosphingolipids that do not have charged sugars are neutral and others are classified as sialylated, or anionic. Sialylated glycosphingolipids are typically referred to as gangliosides. Glycosphingolipids function in many ways, primarily in mediating cell-cell interactions or regulation of signal transduction (Hakomori, 2003).

Glycosylphospatidylinositol (GPI) anchors are glycolipids that are linked to the carboxy terminus of proteins. They consist of a phosphatidylinositol that is bridged through a glycan, consisting primarily of mannose and *N*-acetylglucosamine residues, to an ethanolamine that is linked to a protein. The fatty acids of the phosphatidylinositol anchor everything to the cell membrane (Ferguson and Williams, 1988). The GPI anchor functions to stably anchor the attached protein to the membrane with an anchoring device that is resistant to most proteases and lipases (Paulick and Bertozzi, 2008).

Glycoproteins are a class of glycocongugates that consist of a protein moiety with one or more glycans attached. The glycans can attach to the protein through a variety of linkages and are classified based on these linkages. The most typical linkages are N- and O-linkages, which will be discussed here; however, there are several other classes of glycosylation such as phospho-serine glycosylation and C-mannosylation that will not be covered.

N-glycans are proteins that have glycans attached covalently to an asparagine residue through an N-glycosidic bond (Pratt and Bertozzi, 2005). Not all asparagine residues can be N-glycosylated and the minimal amino acid core sequence Asn-X-Ser/Thr is required for glycosylation. The linkage commonly involves an *N*-acetylglucosamine which forms the base of the pentasaccharide core which consists of another *N*-acetylglucosamine followed by three mannose residues. The pentasaccharide core can be extended to generate high-mannose, hybrid or complex N-linked glycans (Figure 1). The high-mannose N-glycans consists of only mannose residues attached to the core structure in a

variety of linkages. Complex N-glycans consist of the core structure decorated with a multitude of other monosaccharide building blocks, not just mannose residues, with several different linkages possible. Hybrid N-glycans are, as their name implies, a hybrid of both high-mannose and complex N-glycans with regions of high-mannose content and regions with a variety of different monosaccharide residues. N-glycans are present in Archea, Bacteria and Eukaryotes and can have various functions ranging from protein folding, to structural elements, or, as carbohydrate receptors. N-glycans are abundant on the surfaces of epithelial cells in airways, on mucin layers, secreted cells and on bacterial cell surfaces (2).

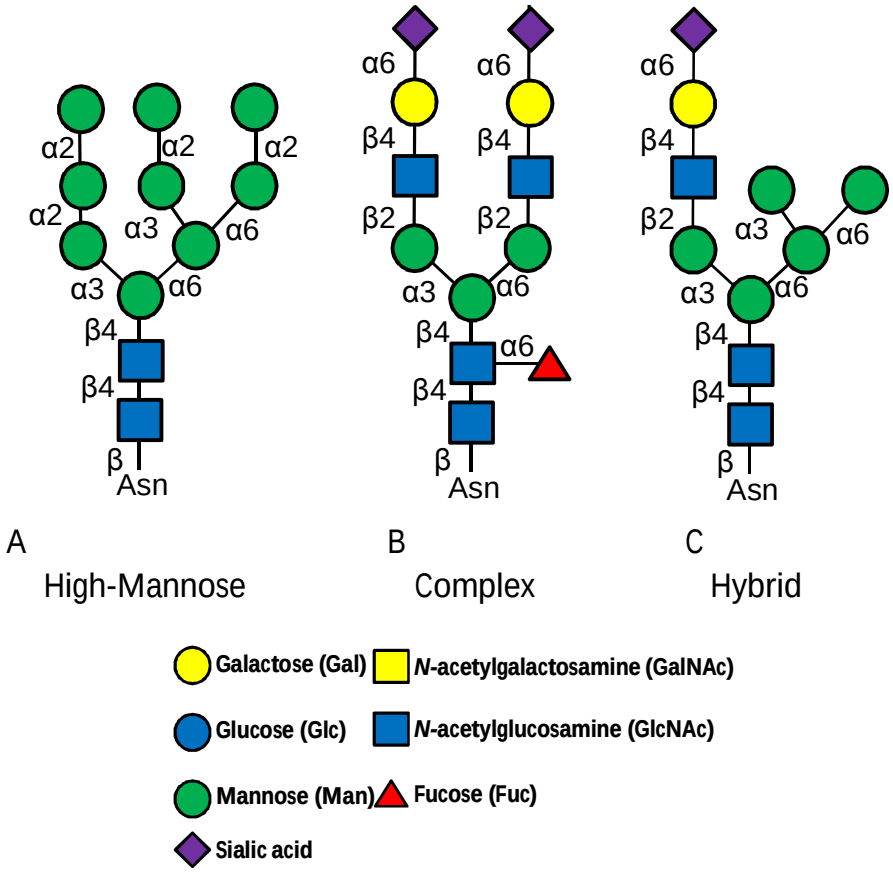


Figure 1: Schematic of the three main types of N-glycans.

A) high mannose B) complex and C) hybrid. The stereochemistry of the linkages indicated and sugars represented by symbols as follows based on accepted nomenclature from the Consortium for Functional Glycomics; shown in the legend.

O-glycans are a common form of glycoprotein that comprise a covalent α-linkage from the hydroxyl side chain of a serine or threonine residues to an *N*-acetylgalactosamine through an O-glycosidic bond (Pratt and Bertozzi, 2005). There are other types of O-glycans possible, including α-linked O-fucose and O-mannose, β-linked O-xylose and *N*-acetylglucosamine as well as both α- and β-linked O-galactose and O-glucose. In mucins, however, the predominant form is an α-linkage via an *N*-acetylgalactosamine and the hydroxyl group of a serine or threonine. Following the *N*-acetylgalactosamine, O-glycans are extended with monosaccharides including *N*-acetylglucosamine, galactose, fucose or sialic acid, but neither mannose nor glucose are found in mucin O-glycans (Figure 2). There are eight core structures found in mucin O-glycans with cores 1 through 4 being the most common. They can also be branched and the sugars can be modified leading to a huge variety of heterogeneous O-glycosylations possible. Similar to N-glycans, O-glycans are abundant on the surfaces of epithelial cells in airways, on mucin layers, secreted cells and on bacterial cell surfaces(Varki et al., 2009).

Mucins are the major glycoprotein components of mucous and they are heavily O-glycosylated (Yu et al., 2008). They consist of a core protein which is decorated with glycans to look like a “bottle brush” with the sugar content of mucin accounting for up to 90% of its weight (Perez-Vilar and Hill, 1999). They can be secreted or membrane anchored to cell surfaces by a hydrophobic transmembrane domain. They can also be connected to other mucins by cysteine-rich regions that lead to mucin polymerization creating huge complexes of mucin causing the characteristic viscosity of mucous. The huge variety of glycans decorating the core protein is very heterogeneous and closely spaced with hundreds of these glycan chains assembled on the mucins. Mucins coat the surfaces of epithelial cells lining the respiratory, gastrointestinal, and urogenital tracts, and in some amphibia, the skin. They protect epithelial cells from dehydration, physical and chemical damage, and provide protection against pathogens. Mucins can vary greatly structurally between organisms, organs and locations within the tract of the organ (Strous and Dekker, 1992).

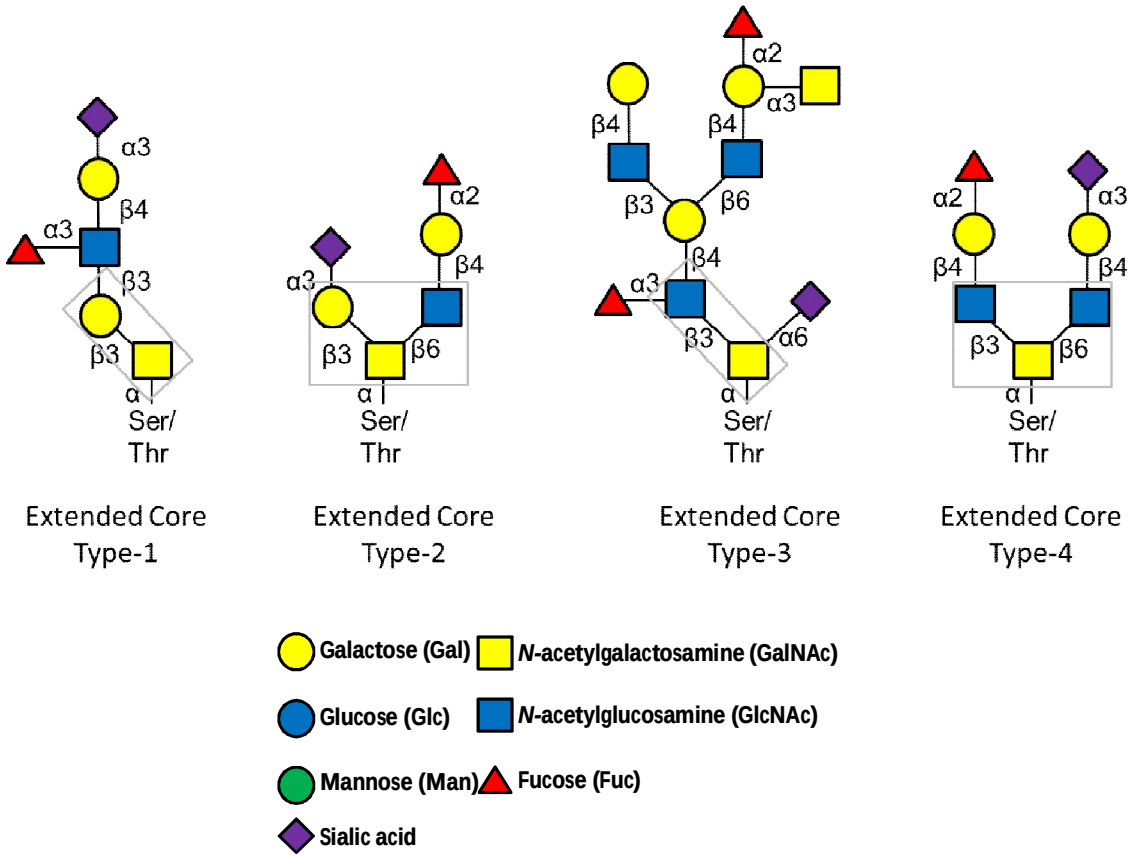


Figure 2: Schematic of complex O-glycans with different core structures.

Examples of O-glycans with extended core types 1, 2, 3, and 4 with core structures shown in boxes. The stereochemistry of the linkages indicated and sugars represented by symbols as follows based on accepted nomenclature from the Consortium for Functional Glycomics; shown in legend.

1.2 Carbohydrate-Active Enzymes.

The glycosidic linkage is extremely stable with an estimated half-life of 5 million year for the β-1,4-glucosidic bond of cellulose (Wolfenden et al., 1998). Despite this stability, carbohydrate polymers are dynamic molecules that are constantly being synthesized and broken down, which is achieved through the actions of enzymes. A huge variety of enzymes are necessary for the formation and breakdown of glycosidic linkages. These enzymes are organized and classified in a database called Carbohydrate-Active enZymes

(CAZy) which is dedicated to the display and analysis of genomic, structural and biochemical information (Cantarel et al., 2009). Carbohydrate-active enzymes account for 1-3% of the proteins encoded by the genomes of most organisms (Davies et al., 2005). The CAZy sequence-based classification system is an effective device for the annotation of function, structure and mechanism of open reading frames (ORFs) found from genome sequencing. From the annotation of these putative carbohydrate-active enzyme encoding ORFs, a framework is provided for the advancement of structural and mechanistic efforts to further our understanding of these enzymes. In the CAZy database, the carbohydrate-active enzymes, glycosyltransferases, polysaccharide lyases, carbohydrate esterases and glycoside hydrolases are classified into sequence based families.

Glycosyltransferases (GTs) are the enzymes responsible for the synthesis of glycosidic bonds. They transfer monosaccharide moieties via an activated donor sugar to a glycosyl acceptor. There is a range of donor species that GTs can utilize including nucleoside diphosphate sugars, nucleoside monophosphate sugars, lipid phosphates, and unsubstituted phosphate. The acceptor substrates of GTs are most commonly other carbohydrates but can also be lipids, proteins, nucleic acids, antibiotics, or other small molecules.

Transfer of glycosyls occurs not only to the nucleophilic oxygen of a hydroxyl substituent of the acceptor, it can also occur to nitrogen, such as in the formation of N-glycans, or to sulphur or carbon nucleophiles (Lairson et al., 2008). GTs form glycosidic bonds with the stereochemistry of the anomeric carbon being either inverted or retained. This results in the biosynthesis of disaccharides, oligosaccharides and polysaccharides as well as the addition of sugar moieties onto glycocongugates (Campbell et al., 1997)(Coutinho et al., 2003). There are over 12,000 GT sequences in the CAZy database that have been classified into 92 amino acid sequence-based families, which is constantly updated. Enzymes with different specificities, i.e. different donors or acceptors, are often found in the same family, complicating functional predictions and often making them unreliable. Despite structural similarities between GTs, they possess different

specificities for the activated sugar donor and acceptor substrate because of differences within loop regions surrounding the active site (Campbell et al., 1997).

Carbohydrate esterases are enzymes that catalyze the removal of ester-based O- and N-acylations from substituted sugars present in mono-, oligo- and polysaccharides. This sometimes facilitates further hydrolysis of complex glycans (Correia et al., 2008)(Cantarel et al., 2009). These enzymes use a Ser-Asp-His catalytic triad and a mechanism similar to protein and lipid esterases or a zinc catalyzed deacetylation mechanism. There are 16 different sequence based families of carbohydrate-esterases classified in the CAZy database (Lombard et al., 2010).

Both polysaccharide lyases and glycoside hydrolases cleave glycosidic bonds. Polysaccharide lyases proceed via a β-elimination mechanism cleaving uronic acid containing sugars. The majority of polysaccharide lyases are produced by bacteria that degrade plant cell walls and are active on the glucuronates, galacturonates and alginates from algaes and plant pectins (Lombard et al., 2010)(Garron and Cygler, 2010). These enzymes are also found as virulence factors produced by human pathogens with lyase activity on hyaluronan and heparin (Abbott and Boraston, 2008)(Li et al., 2000). There are 22 different sequence-based families of polysaccharide lyases and similarly to the GTs, these families are frequently polyspecific with enzymes having different substrates or generating different products (i.e. contain enzymes acting on different substrates or that generate different products).

1.2.1 Glycoside Hydrolases.

Glycoside hydrolases (GHs) are the largest group of carbohydrate-active enzymes. They hydrolyze the glycosidic bond between carbohydrates or between a carbohydrate and a non-carbohydrate. There are multiple methods of classifications of GHs including sequence-based, catalytic mechanism and endo vs exo action. Glycoside hydrolases are classified into families that are related by sequence and, by consequence, fold as well. There are currently 125 different GH families on the Carbohydrate-Active enzyme (CAZy) database with over 30,000 entries; this is constantly updated (Henrissat and

Davies, 1997). Sequence-based classification is very useful and powerful in that it can be predictive of not only enzyme fold but also mechanism and catalytic machinery, and is suggestive of function in a large number of families. An additional method of classification groups GH families into 14 structure-based clans in the absence of amino acid identity. A clan consists of families that have similar tertiary structure, catalytic residues and mechanism but have low or no amino acid sequence identity. This classification is also useful because the fold of a protein is better conserved than the sequence in many cases. It is useful in that it helps classify new GH enzymes whose amino-acid sequence may relate to more than one family.

There are two common reaction mechanisms found in GHs, which are distinguished by the stereochemistry of the anomeric carbon after hydrolysis, as is either retained or inverted (Zechel and Withers, 2000). Retention of the stereochemistry of the anomeric carbon is accomplished using a two-step double displacement mechanism. Hydrolysis involves the side chain of a catalytic residue which acts as a nucleophile and another that acts as an acid/base (Figure 3). These residues are typically glutamate or aspartate and are situated ~ 6 Å apart (Zechel and Withers, 2000). In the first glycosylation step, the nucleophile attacks the anomeric centre while the acid/base residue acts as an acid to protonate the glycosidic oxygen, passing through an oxocarbenium ion-like transition state to form a glycosyl enzyme intermediate. In the second deglycosylation step, the glycosyl enzyme intermediate is hydrolyzed as the acid/base residue, now acting as a base, deprotonates a water molecule which attacks the anomeric centre. In some retaining GHs, such as those active on sialic acids, the catalytic nucleophile is a tyrosine residue. Another form of retaining mechanism, called substrate-assisted catalysis, involves the hydrolysis of substrates that are *N*-acetylated (Terwisscha van Scheltinga et al., 1995; Macauley et al., 2005). These enzymes do not have a catalytic nucleophile but instead use the substrate acetamido group as an intramolecular nucleophile.

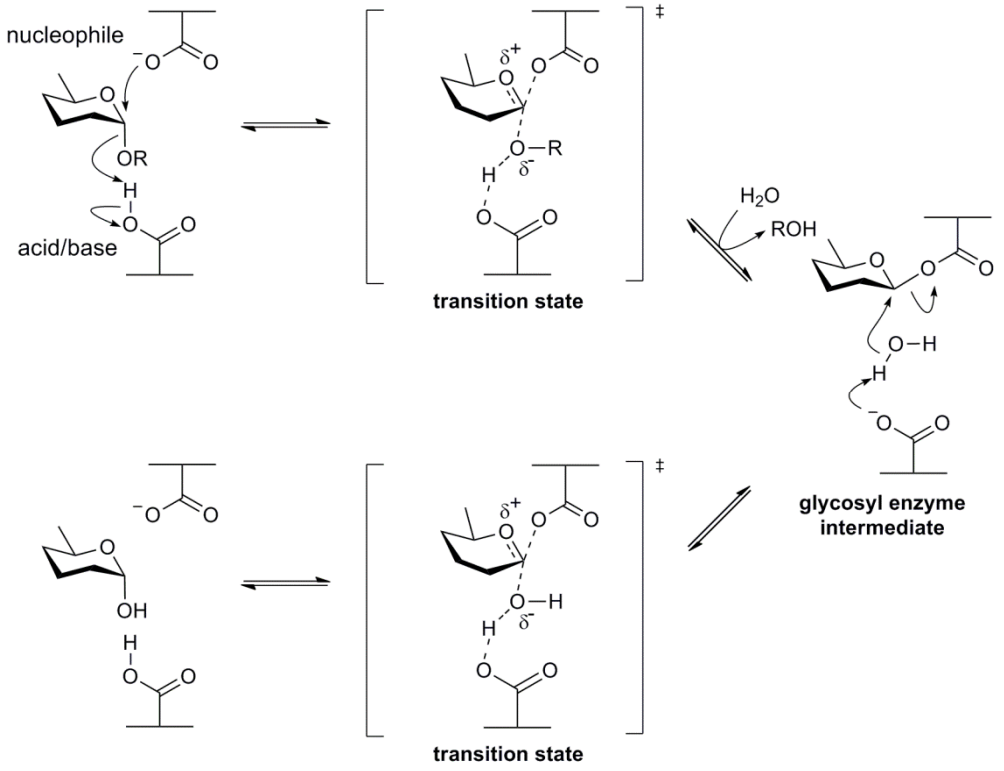


Figure 3: Retaining reaction mechanism of glycoside hydrolases.

Figure reproduced with permission from Withers, S. and Williams, S. "Glycoside hydrolases" in CAZypedia, available at URL http://www.cazypedia.org

GH mediated hydrolysis of sugars with inversion of anomeric configuration occurs by a single-displacement mechanism through an oxocarbenium ion-like transition state (Figure 4). In the inverting mechanism, two amino acid side chains, which act as a general acid and as a general base, typically glutamate or aspartate, are situated ~10 Å apart. A catalytic water is deprotonated by the general base, which then attacks the anomeric centre and the leaving group is protonated by the general acid (Zechel and Withers, 2000).

GHs are also classified based on their endo or exo acting mechanism. Exo-acting indicates hydrolysis at the end of a glycan, usually at the non-reducing end whilst endo-acting is hydrolysis in the middle of a polysaccharide chain.

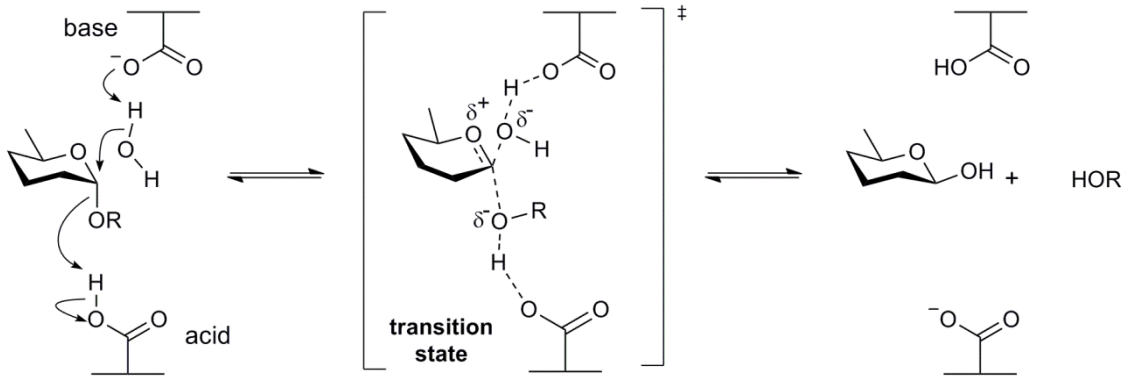


Figure 4: Inverting mechanism of glycoside hydrolases.

Figure reproduced with permission from Withers, S. and Williams, S. "Glycoside hydrolases" in CAZypedia, available at URL http://www.cazypedia.org/

1.2.2 Glycoside Hydrolase Multimodularity.

Glycoside hydrolases often have multiple modules in addition to the requisite catalytic module. A module is an amino acid sequence with a discreet, independent fold that is part of a larger sequence. There is a huge variety of ancillary modules found in GHs including many with sequences that encode proteins with unknown functions. The widespread presence of these ancillary modules in GHs suggests their importance.

The most prominent type of ancillary module found in GHs are carbohydrate-binding modules (CBMs). These CBMs are non-catalytic and, as their name implies, bind carbohydrate ligands and promote the association of the GH catalytic module with the substrate. They assist in the efficient degradation of carbohydrate substrates by binding to them and directing the catalytic module, ultimately increasing the specific activity of the enzyme. CBMs are also classified into sequence-based families in the CAZy database (Cantarel et al., 2009). It is very common for a GH to contain multiple CBMs from the same or even different families, which can be found in tandem or separated by other modules. Within one GH it is possible to have different CBMs with different substrate specificities (Boraston et al., 2004).

Also present in GHs are other modules such as fibronectin type-III (FN3) which are common constituents of proteins occurring in roughly 2% of all animal proteins including extracellular matrix molecules, cell-surface receptors, and intracellular proteins. FN3 modules have been identified in various carbohydrate-active enzymes from a diverse distribution of Gram-positive and Gram–negative bacteria. It has been postulated that these modules function to mediate protein-protein interactions although their functions are largely unknown (Bencharit et al., 2007; Varki et al., 2009).

A prominent example of modularity common in cellulolytic bacteria is the cellulosome (Figure 5). The cellulosome is a large, mega-dalton, multienzyme complex that is responsible for the efficient and synergistic breakdown of cellulose and hemicellulose found in plant cell walls. The classical cellulosome consists of a large non-catalytic scaffoldin protein that contains multiple copies of cohesin modules and one cellulose-binding CBM that targets the cellulosome to the substrate. The cohesin modules are involved in protein-protein interactions with their cognate binding partners, dockerin modules. The catalytic modules of the cellulases and hemicellulases GHs are tethered to the scaffoldin through their dockerin modules by the protein-protein contact of the interacting cohesin and dockerin modules and is termed a type-I interaction. This results in the coordination of the GHs onto a cohesin bearing scaffoldin through the action of the cohesin-dockerin interaction. A type-II cohesin·dockerin interaction links the cellulosome to the proteoglycan layer of the bacterial cell surface usually via an association with a cell-anchoring protein (Peer et al., 2009). The cellulosome assembly potentiates catalysis by enabling the enzyme synergy provided from spatial proximity and efficient substrate targeting (Shoham et al., 1999; Fontes and Gilbert, 2010).

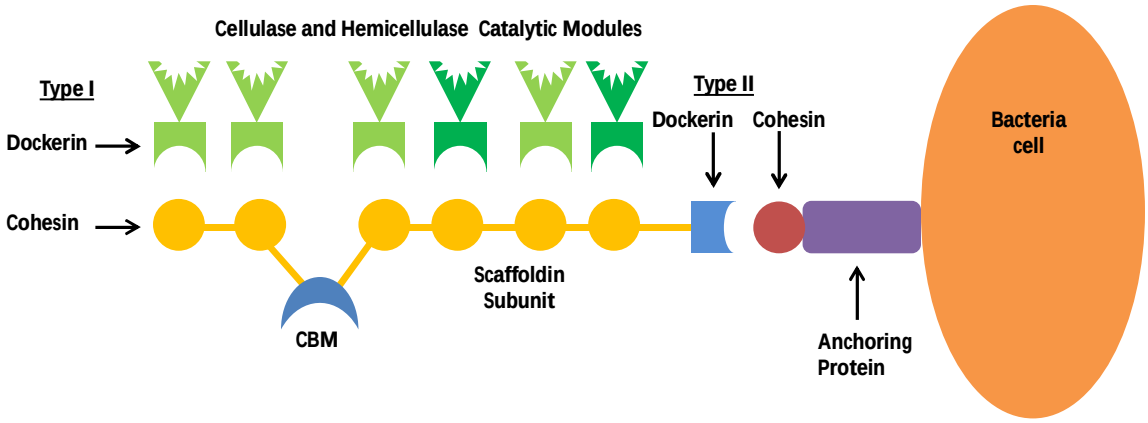


Figure 5: Simplified schematic cartoon of the *Clostridium thermocellum* classical cellulosome.

1.3 Bacterial Pathogens and Human Glycans.

The association between bacteria and humans can span a range of potential outcomes from being mutually beneficial, beneficial to the bacteria without compromising the health of the host, and lastly, benefitting the bacteria at the expense of the host. The latter refers to pathogenic bacteria and these types of interactions are often predicated on the ability of the pathogenic bacteria to exploit its hosts by modifying the host's glycans (Varki et al., 2009). Secreted carbohydrates and carbohydrates on the outer surfaces of cells control many biological processes including cell-cell, cell-extracellular matrix, cell-molecules, and cell-cell interactions from different organisms. Carbohydrates often function as the human host's first line of defence against pathogen invasion by coating surfaces of epithelial cells or in glycan-rich mucins lining the gastrointestinal and respiratory tracts. Two kinds of pathogenic bacteria notable for their production of carbohydrate-active enzymes are *Clostridium perfringens* and *Streptococcus pneumoniae.* The human niche of *C. perfringens* is in the gastro-intestinal tract and the human niche of *S. pneumoniae* is in the respiratory tract both of which are heavily coated with glycan-rich mucous and have glycan decorated cells.

1.3.1 *Clostridium perfringens.*

Clostridium perfringens is a pervasive Gram-positive bacterium commonly found in the gastrointestinal tract of animals, and ubiquitously throughout the environment, especially in soil. As a pathogen, *C. perfringens* infection can result in gas gangrene, necrotic

enteritis and gastroenteritis (Rood and Cole, 1991) (Songer, 1996). Gas gangrene is a significant health threat to people with diabetes and has other common risk factors such as smoking and alcoholism. Necrotic enteritis is a major agricultural problem in poultry farming causing increased mortality rates and reduced weight gain (Van Immerseel et al., 2004). *C. perfringens* is the third most frequent cause of food borne illness in the USA (McClane, 2001) and contributes significantly to the burden on healthcare and to agricultural health.

The five biotypes of *C. perfringens* are classified on the basis of the main toxins they produce (Petit et al., 1999). These toxins contribute to the pathogenic prowess of this organism and help it expeditiously destroy tissues during disease progression. The main toxin is the α-toxin which is a phospholipase that is active on glycolipids found attached to cell membranes (Titball et al., 1999). The β- and ε-toxins are pore-forming toxins which disrupt the integrity of cell membranes and the ι toxin is ADP-ribosylating (Shatursky et al., 2000; Petit et al., 1997). The μ-toxin, a glycoside hydrolase with hyaluronidase activity, is a putative virulence factor of *C. perfringens* that when injected intradermally with the α-toxin, potentiates the cytolytic effect of the α-toxin by expediting its spread (Smith, 1979). Other notable GHs produced by *C. perfringens* include two sialidases (NanI and NanJ) which have also been shown to potentiate the activity of the α-toxin by removing terminal sialic acid residues from cellular gangliosides significantly increasing the sensitivity of target cells to the cytotoxic effects of the α-toxin (Flores-Díaz et al., 2005); however, the *C. perfringens* sialidases are not essential for full virulence of the bacterium in a mouse myonecrosis model (Chiarezza et al., 2009).

In addition to the μ-toxin and the sialidases, sequencing of the *C. perfringens* (ATCC 13124) genome allowed for identification of 55 putative glycoside hydrolases (Shimizu et al., 2002). Of these putative enzymes, approximately half of them have classical Gram-positive signal peptides for secretion into the extracellular milieu and several are predicted to be attached to the cell wall (Figure 6). These putative carbohydrate-active enzymes have predicted substrate specificities for sugars that are components of complex

eukaryotic glycans in addition to several with undefined specificities. These enzymes likely play a role in the host-pathogen interaction due to their extracellular location, predicted specificities and their implication in virulence. The sialidases and the μ-toxin likely have combined efforts with the other glycan degrading putative GHs to destroy the glycans that form the body's first line of defense in mucins and cell-surface glycans (Chiarezza et al., 2009; Petit et al., 1999; Flores-Díaz et al., 2005; Canard et al., 1994). This likely contributes to the tissue destruction characteristic of *C. perfringens* infection and also provides the bacteria with a carbohydrate-based source of nutrition. The notable production of carbohydrate-active enzymes by *C. perfringens* helps it deal with the glycan rich mucins that coat the intestinal tract. There are several GH homologues that are present in other bacterial systems including *S. pneumoniae.* This bacterium is also distinguished in its production of human tissue degrading GHs that are both similar and different from those produced by *C. perfringens.*

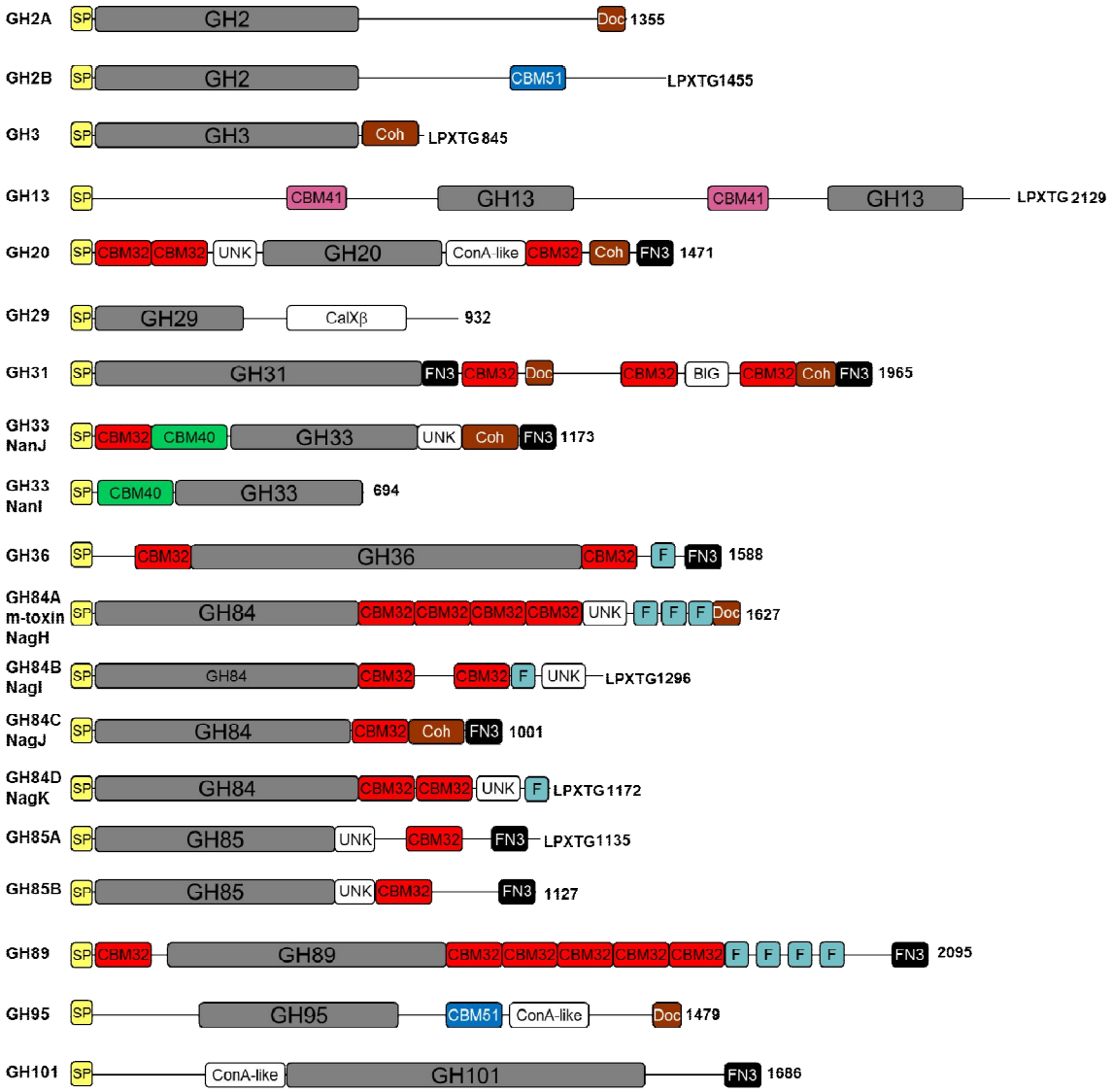


Figure 6: Schematic of the modular arrangement of examples of *C. perfringens* glycoside hydrolases.

SP, Signal peptide (yellow); CBM32 (red), CBM40 (green), CBM41 (pink), CBM51 (blue), denotes carbohydrate-binding modules from families 32, 40, 41 and 51 respectively; GH*XX*, glycoside hydrolase catalytic domains (grey) where *XX* represents the family number; UNK, modules of unknown function (white); BIG, bacterial Ig-Like fold (white); ConA-like, modules that are concanavalin-A like (white); CalXβ, calnexin-like (white); F denotes FIVAR, found-in-various–architectures (light blue), FN3, fibronectin type III module (black); LPXTG, LPXTG motif, sortase mediated cell-anchoring. Areas with no indicated

module have no sequence similarity to known other known sequences. The total number of amino acids are indicated.

1.3.2 *Streptococcus pneumoniae.*

Streptococcus pneumoniae is a significant human pathogen and is the major causative agent of pneumonia, an acute respiratory disease that is the most common cause of death from infection in developed countries. *S. pneumoniae* infection occurs most frequently in the elderly and the very young, responsible yearly for 1-2 million infant deaths worldwide (Bogaert et al., 2004; van der Poll and Opal, 2009). This Gram-positive bacterium is a cause of several other diseases and infections including meningitis, ear infection or otitis media, acute sinusitis, and septicaemia (Bogaert et al., 2004). *S. pneumoniae's* niche is the upper respiratory tract where it exists as a commensal organism in approximately 40% of humans (Kadioglu and Andrew, 2004; Tettelin et al., 2001). These bacteria have a polysaccharide capsule and the differences in this capsule permit serological differentiation between the >100 different serotypes. The serotypes vary in level of virulence, prevalence and the extent of antibiotic resistance. Despite its transient commensalism, *S. pneumoniae* can, through an unknown method, slip out of its passive role and into a pathogenic, disease causing state.

S. pneumoniae infection requires several steps including the penetration of the extracellular matrix, adherence to lung epithelium, infiltration of host cells, and subsequent dissemination of the bacteria throughout the tissue (Bergmann and Hammerschmidt, 2006; Hammerschmidt, 2006). This pathogen produces a number of virulence factors that contribute to its aggressiveness as a pathogen, the most important of which is the capsular polysaccharide. Several large-scale signature-tagged mutagenesis (STM) studies identified new virulence factors in *S. pneumoniae* and greatly expanded the repertoire of genes encoding virulence factors (Hava and Camilli, 2002; Polissi et al., 1998; Obert et al., 2006). Represented among these virulence factors are a noteworthy number of carbohydrate-active proteins. In the *S. pneumoniae* (TIGR4) genome there are 41 genes that encode known or putative glycoside hydrolases and of these 18 are required for full virulence. Many of these putative GHs are from families that are implicated in

the destruction of complex eukaryotic glycans, such as those found in mucins that line the respiratory tract of humans as well as epithelial cell-surface associated glycans (Figure 7).

There are many glycoside hydrolase virulence factors produced by *S. pneumoniae.* The enzymes that sequentially remove the terminal sugars from the distal arms of complex N-linked glycans are NanA, StrH, and BgaA, which are a sialidase, an *exo*-β-D-*N*-acetylglucosaminidase, and an *exo*-β-D-galactosidase, respectively. EndoD is an *endo*-β-D-*N*-acetylglucosaminidase that cleaves the chitobiose core of N-linked glycans smaller than Man5GlcNAc2 to remove the glycan from the protein scaffold. Despite increasing knowledge in this area it remains unclear how bacteria process the core mannose component of N-linked glycans (King et al., 2006; Dalia et al., 2010). In addition to these N-glycan degrading GHs are enzymes that have been implicated in O-glycan processing, such as an endo-*N*-acetylgalactosaminidase that cleaves a disaccharide from the polypeptide portion of the glycoprotein (Marion et al., 2009). However, a complete picture of O-glycan processing is still required. A pullulanase, SpuA has been implicated in *S. pneumoniae* virulence through its multivalent association and breakdown of host glycogen (Lammerts van Bueren et al., 2011). BgaC is a β-galactosidase that hydrolyzes host galactose moieties affecting adherence to the host cells so that binding of the bacteria is decreased. Notably, BgaC is expressed as a surface protein despite its lack of typical extracellular signal sequence or membrane anchoring motif (Jeong et al., 2009). The fucose utilization operon is critical to the virulence of the *S. pneumoniae* TIGR4 strain. One of the components of this operon is a GH from family 98 which is the only predicted extracellular component of this operon and thus very likely initiates this catabolic pathway by action on a host glycan (Higgins et al., 2009). Clearly, *S. pneumoniae* produces a broad range of GHs that are virulence factors from a variety of different families with diverse human glycan specificities. The activity of many of these GHs appears to be critical for full virulence of *S. pneumoniae.*

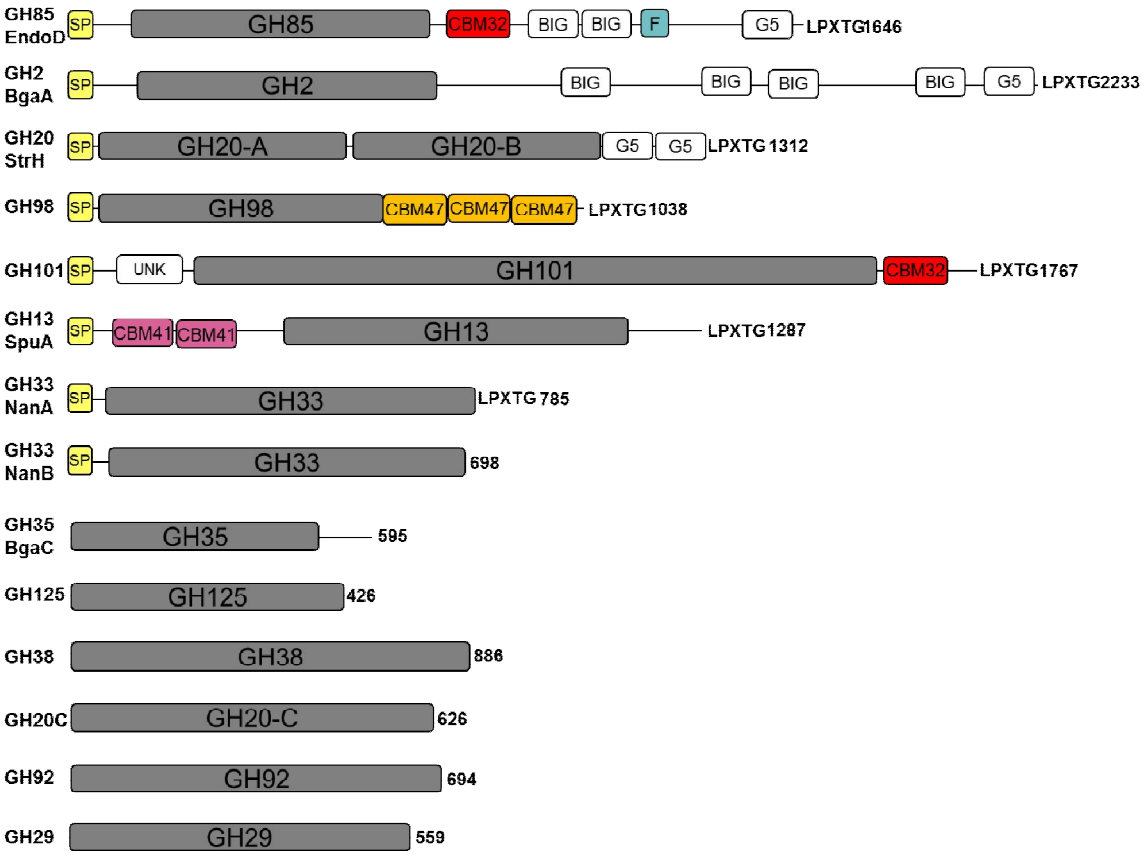


Figure 7: Schematic of the modular arrangement of examples of *S. pneumoniae* glycoside hydrolases.

SP, Signal peptide (yellow); CBM32 (red), CBM47 (dark yellow), CBM41 (pink), denotes carbohydrate-binding modules from families 32, 47, and 41 respectively; GH*XX*, glycoside hydrolase catalytic domains (grey) where *XX* represents the family number; UNK, modules of unknown function (white); BIG, bacterial Ig-Like fold (white); G5, unknown module (white); F denotes FIVAR, found-in-various–architectures (light blue); LPXTG, LPXTG motif, sortase mediated cell-anchoring. Areas with no indicated module have no sequence similarity to known other known sequences. The total number of amino acids is indicated.

1.4 Research Objectives.

Carbohydrate-metabolism is critical for both *S. pneumoniae* and *C. perfringens* which encode a large number of GHs involved in eukaryotic glycan processing. Many of these GHs are not only involved in carbohydrate-metabolism but also are implicated in virulence. Notably, the majority of the *C. perfringens* GHs have a significant amount of modularity with a catalytic module and up to eight ancillary modules. Unlike the majority of the *C. perfringens* GHs, the *S. pneumoniae* GHs have significantly less modularity with often the catalytic module being the only module. However, in other cases, some have CBM modules in addition to the catalytic modules as well as other modules with unknown function.

Interestingly, both *C. perfringens* and *S. pneumoniae* inhabit mucin-rich areas of the body, the gastrointestinal and respiratory tracts, respectively. Also, the genomes of both of these organisms harbor GHs from families that are implicated in the destruction of human glycans. The glycan-rich niches of both of these bacteria in combination with the predicted specificities of the GHs they produce, allows us to hypothesize that they are active on human glycans. Genome sequencing is providing a wealth of information that is leading to the discovery of many new putative carbohydrate-active enzymes and not surprisingly there are substantial gaps in knowledge with respect to the specificities, catalytic mechanism, overall structures and active site architectures of these enzymes; a situation that prevents a full understanding of their role in bacterial fitness and, in some cases, virulence. Also, in addition to GHs possessing modules with carbohydrate-hydrolysis activity, there are also modules that display carbohydrate-adherence and have modules of unknown function. Due to similarity with plant cell wall degrading enzymes, we hypothesize that some of these unknown modules could coordinate the organization of higher order complexes through interactions with other unknown modules from other GHs. We also hypothesize that the concerted actions of the different modules with diverse functions such as carbohydrate binding and hydrolysis, protein-protein interactions can be spatially coordinated to occur simultaneously. *C. perfringens* and *S. pneumoniae* will be used as model systems to study glycoside hydrolases due to their large complements of glycoside hydrolases, some of which are known virulence factors.

It is the global objective of this thesis to provide insights into human glycan processing by carbohydrate-active enzymes, specifically glycoside hydrolases, from the human pathogens Streptococcus pneumoniae and Clostridium perfringens.

This study will investigate the key features of glycoside hydrolases from *S. pneumoniae* and *C. perfringens* and their carbohydrate-hydrolysis, modularity and overall glycoside hydrolase structure. We will investigate *S. pneumoniae* glycoside hydrolases from different families and characterize the key elements that differentiate these enzymes and elucidate the molecular details that govern the processing of human glycans. As well, we seek to determine functions for some of the unknown ancillary modules present in the *C. perfringens* GHs and determine how these huge, multi-modular enzymes are arranged spatially to optimally perform the various functions of the different modules.

Specific research questions and objectives are addressed in chapters 2, 3, 4 and 5

Chapter 2: Structural analyses of substrate recognition of a family 101 glycoside hydrolase from *Streptococcus pneumoniae* revealing insights into O-glycan degradation

Adapted and expanded from: Acta Crystallography Section F Structural Biology Crystallization Communication. 2009; 65(Pt 2): 133-5.

Contributions to Research: Cloning, protein production and purification, crystallization and solution, manuscript and figure preparation.

2.1 Introduction.

The epithelial cells that line the gastrointestinal, respiratory and genitourinary tracts are typically coated in a secretion called mucous that protects the underlying cells from physical and chemical damage and pathogenic infiltration. Mucous is composed of mucin glycoproteins which are heavily O-glycosylated. The simplest mucin O-glycan consists of an *N*-acetylgalactosamine (GalNAc) that is α-linked to a serine or threonine residue and is often called the Tn antigen. The Tn antigen is antigenic and can be further glycosylated with a host of monosaccharides including galactose, *N*-acetylglucosamine, fucose and sialic acid to yield extended and diversely decorated, mucin O-glycans. There are a variety of common O-glycan core structures (Figure 2) with the most common being the core type-1 O-glycan which consists of the Tn antigen with a galactose (Gal) attached β1,3 to the GalNAc residue, called the T antigen (Varki et al., 2009).

Mucins not only hydrate and protect underlying epithelial cells but they are also responsible for providing one of the host's first lines of defense against invading pathogens. The pathogen mediated destruction of the mucin barrier can help the bacteria persist with the added benefit of providing a carbohydrate source of nutrition. *Streptococcus pneumoniae* is a formidable human pathogen that inhabits the respiratory tract of humans and causes serious diseases including pneumonia, meningitis, otitis media, and septicaemia. Genome sequencing, signature-tagged mutagenesis and other biochemical and genetic studies have revealed the reliance of *S. pneumoniae* on carbohydrate processing and metabolism for full virulence of the bacterium (Hava and Camilli, 2002; Tettelin et al., 2001; Boraston et al., 2006; Shelburne et al., 2008). *S. pneumoniae* produces a wide range of glycoside hydrolases that act in concert to degrade

host glycans, and a large portion of these have putative activity on sugars found in mucin-glycoproteins.

One component of the extracellular, cell wall attached armoury of enzymes in *S. pneumoniae* is an *endo*-α-*N*-acetylgalactosaminidase (*endo*-α-GalNAcase) that catalyzes the liberation of Galβ1,3GalNAc, core type-1 disaccharide, from serine or threonine residues of mucin glycoproteins. Based on sequence similarity this enzyme is classified as a family 101 glycoside hydrolase (GH101) together with homologues in the CAZy database (Henrissat, 1991). All 29 pneumococcal genome sequences that are currently available contain a GH101 homologue. To date, 69 GH101 *endo*-α-GalNAcase genes have been identified in various bacterial species including *Bifidobacterium longum, Clostridium perfringens, Propionibacterium acnes,* and *Enterococcus faecalis.* The *B. longum* GH101 is highly specific for the core type-1 O-glycan, releasing the disaccharide Galβ1,3GalNAc and therefore has comparable specificity to the *S. pneumoniae* GH101s (Fujita et al., 2005). The *C. perfringens, E. faecalis,* and *P. acnes* GH101s have been demonstrated to have broader substrate specificity in that they are able to catalyze the liberation of core type-1 disaccharide in addition to other O-glycan core types. The *C. perfringens* GH101 has broad substrate specificity and is able to liberate the core type-2 trisaccharide Galβ1,3(GlcNAcβ1,6)GalNAC, and the disaccharide Galα1,3GalNAc and the monosaccharide GalNAc (Ashida et al., 2008). The *P. acnes* GH101 can cleave the core type-3 disaccharide GlcNAcβ1,3GalNAc (Koutsioulis et al., 2008). *E. faecalis* GH101 is able to release trisaccharides from core type-2 (Galβ1,3[GlcNAcβ1,6]GalNAc), tetrasaccharide Gal-core type-2 and the core type-3 disaccharide GlcNAcβ1,3GalNAc (Goda et al., 2008). A common feature of GH101 family members is the ability to liberate the core type-1 disaccharide from serine/threonine.

The first reported three-dimensional X-ray structure of a GH101 was from *S. pneumoniae* R6 which revealed a multi-domain architecture of a ~170 kDa fragment of the enzyme (Caines et al., 2008). This structure was found to be analogous to the GH13 α-amylases in that the GH101 catalytic module has a distorted $(\beta/\alpha)_8$ barrel flanked by domains

composed of β-sheets (MacGregor et al., 2001). Functional characterization of GH101 from *B. longum* revealed that hydrolysis proceeds with retention of configuration which is indicative of a double displacement mechanism (Fujita et al., 2005). Subsequent structural characterization revealed the structure of the GH101 from *B. longum* and automated docking analyses and mutational analyses were performed in an attempt to investigate substrate interactions (Suzuki et al., 2009). Extensive kinetic and mechanistic analyses assigned putative catalytic residues of the GH101 from *S. pneumoniae* R6 and found further evidence for a double-displacement retaining mechanism (Willis et al., 2009).

Despite these structural and mechanistic studies, only structures for GH101 enzymes that lack bound substrate have been determined. There is very little information available as to the molecular basis of substrate recognition and why there are differences in specificities between the bacterial GH101s. We hypothesize that SpGH101 can accommodate the core type-1 O-glycan, Galβ1,3GalNAc in its active site in subsite positions -2 and -1. Also, we predict that the nature of the aglycon, i.e. the protein or polypeptide of the O-glycan, is not specific beyond the α-linked serine or threonine residue allowing the enzyme to liberate the disaccharide from a variety of core proteins. We also hypothesize that insight into how GH101s from other bacteria have differing specificities is due to the active site architecture.

The objectives of this study are to characterize the molecular basis of substrate recognition of a GH101 from S. pneumoniae.

This will be approached by attempting to obtaining complexes with substrates and substrate analogues to provide the first experimental report of a GH101 substrate-complex, thereby providing a structural basis for substrate recognition by GH101 and contributing to the elucidation of the molecular details that govern *endo*-α-*N*-acetylgalactosaminidase activity.

2.2 Experimental Procedures.

Cloning, production and purification of SpGH101. The gene fragment encoding the GH101 catalytic module (SpGH101) consisting of amino acids 317-1425 was PCR-amplified from *S. pneumoniae* TIGR4 genomic DNA (ATCC BAA-334D). Two sets of primers were used for cloning into pET-28a(+) and pET-22b(+) vector (Novagen) (Table 1). The PCR-amplified gene fragments were obtained using standard PCR methods using Phusion High-Fidelity DNA Polymerase (New England Biolabs). The products were digested with NdeI and XhoI restriction endonucleases and ligated to correspondingly digested pET-28a(+) or pET-22b(+), respectively, using standard cloning procedures. The resultant plasmids from pET-28a(+) and pET-22b(+), here called SpGH101N and SpGH101C, encode identical polypeptides consisting of residues 317–1425 of the protein. The SpGH101N clone contained an N-terminal six-histidine tag followed by a thrombin protease cleavage site and the SpGH101C clone contained a C-terminal non-cleavable six-histidine tag. The SpGH101C was cloned and produced in an effort to produce more soluble protein. PCR site-directed mutagenesis procedures (Hutchison et al., 1978) were used to introduce a D764N substitution into clone SpGH101C and standard cloning procedures were used resulting in plasmid, SpGH101Mut. The DNA sequence fidelity of all constructs was verified using bidirectional sequencing with nested primers (Table 1).

Name	Nucleotide sequence
GH101pET28For	GGCAGCCATATGGAAAAAGAAACAGGTCCTG
GH101pET28Rev	GGATCCCTCGAGTTACAACATCTTACCTG
GH101pET22For	TATACATATGGAAAAAGAAACAGGTCCTG
GH101pET22Rev	CGGCGTCTCGAGCAACATCTTACCTGTTAGGG
GH101D764NFor	CTTTATCTATGTGAACGTTTGGGGTAATGG
GH101D764NRev	CCATTACCCCAAACGTTCACATAGATAAAG
GH101NestedFor	GCGTATCGGTGGTGTCGAAGACTTCAAGACCC
GH101NestedRev	GGTGGTTACGGATAAAGCGGGTGATGGC

Table 1: Primers used for cloning of recombinant SpGH101 and nucleophile mutant.

SpGH101N, SpGH101C and SpGH101Mut plasmids were transformed into chemically competent *E. coli* BL21 STAR (DE3) cells (Novagen) and the proteins, SpGH101N,

SpGH101C and SpGH101Mut, were produced in Luria-Bertani media supplemented with 50 μg ml^{-1} kanamycin (Sigma). The cells were grown at 37°C to an optical density of 0.5 at A_{595} and induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside at 37°C for 4 hours. Cells were harvested by centrifugation at 6000 x g for 10 minutes, chemically lysed (Charlwood et al., 1998) and harvested by centrifugation at 27 000 x g for 45 minutes. The polypeptides were purified from cell-free extract using immobilized metal affinity chromatography following previously described methods (Boraston et al., 2001). The purity of fractions was assessed using SDS-PAGE and those deemed to be greater than 95% pure were pooled, concentrated and buffer exchanged into 20 mM Tris-HCl, pH 8.0, in a stirred ultra-filtration unit (Amicon) using a 10 kDa molecular weight cut-off (MWCO) membrane (Filtron). The SpGH101N protein was further purified by size-exclusion chromatography using Sephacryl S-200 (GE Biosciences) in 20 mM Tris–HCl pH 8.0 and the SpGH101C and SpGH101Mut proteins were purified by ion exchange chromatography using Resource Q column (GE Biosciences). The concentrations of purified proteins were determined from the UV absorbance at 280 nm using a calculated molar extinction coefficient of 240 420 M^{-1} cm^{-1} (Mach et al., 1992).

Selenomethionine-labeled (SelenoMet) SpGH101N was produced using the *E. coli* B834 (DE3) methionine auxotroph. *E. coli* colonies taken from an LB-agar plate were used to inoculate 1 liter of SelenoMet Medium Base (Molecular Dimensions Ltd.) supplemented with SelenoMet Nutrient Mix (Molecular Dimensions Ltd.) and l-selenomethionine (40 mg/liter). These cultures were grown, induced, and harvested, and the polypeptide was purified as described for the unlabeled protein.

Crystallization, Data Collection and Refinement. Prior to crystallization, the native and SeMet SpGH101N proteins were concentrated to 15 mg ml^{-1} in 20 mM Tris–HCl pH 8.0. SpGH101 crystals which had plate morphology grew within one week by adding 1 μl 25% polyethylene glycol (PEG) 1500 (Hampton Research) to 1 μl protein solution using the hanging-drop vapour-diffusion method at 292 K. Removal of the six-histidine tag was unnecessary for crystallization. Crystals were cryoprotected in 1 μl 33% PEG 1500

supplemented with 6% MPD (Hampton Research), and flash-cooled directly in a nitrogen-gas stream at 113 K.

SpGH101C and SpGH101Mut proteins were concentrated to 15 mg ml^{-1} in 20 mM Tris–HCl pH 8.0. Initial crystals were grown within one week by adding 1 µl of 18% polyethylene glycol (PEG) 3350 and 0.2 M ammonium citrate tribasic pH 7.0 (Hampton Research) to 1 µl protein solution using the hanging-drop vapour-diffusion method at 292 K. Subsequent micro-seeding with native SpGH101N crystals improved crystal size and diffraction quality. A complex of SpGH101C with a chemically synthesized substrate analogue, O-[3-O-(1-β-D-galactopyrano)-2-*N*-Acetyl-2-deoxy-*D*-galactopyranosylidene]amino-*N*-Phenylcarbamate (PUGT) (provided by Dr. Vocadlo; SFU), was produced by soaking native crystals in crystallization solution containing excess of PUGT. Crystals were cryoprotected in mother-liquor supplemented with 30% ethylene glycol and flash-cooled directly in a nitrogen-gas stream at 113 K. A complex of SpGH101Mut with serinyl-T antigen (serinyl-Tag; Galβ1,3GalNAc-α-serine), was produced by soaking the crystals in crystallization solution containing excess of serinyl-TAg. Crystals were cryoprotected in mother-liquor supplemented with 30% ethylene glycol and flash-cooled directly in a nitrogen-gas stream at 113 K.

A native data set was collected at Beam Line 9-2 at the Stanford Synchrotron Radiation Light Source, SSRL, and an optimized selenium single anomalous dispersion diffraction data set for selenomethionine-derivative SpGH101 was collected at CMCF1 at the Canadian Light Source, structure and refinement statistics are shown in Table 2. SHELXC/D was used to determine the substructure of twenty two selenium atoms, followed by refinement and phasing, and density modification and solvent flattening with the Phenix software suite, resulting in easily interpretable electron density maps (McCoy et al., 2007). Automatic model building of the selenium-substituted model was done with SOLVE/RESOLVE and yielded a partial model that was used as a starting point for the higher resolution native structure used for molecular replacement using MOLREP to find one molecule in the asymmetric unit (Vagin and Teplyakov, 2010). Model building was

done using COOT and refinement was done with REFMAC (Terwilliger, 2003; Murshudov et al., 1997).

Diffraction data for SpGH101 in complex with PUGT and SpGH101Mut in complex with serinyl-TAg were collected at SSRL Beam Line 9-2. Both complex structures were determined by molecular replacement using MOLREP (Vagin and Teplyakov, 2010) to find one molecule in the asymmetric unit and the native structure of SpGH101 as a search model. The initial models were corrected and completed manually by multiple rounds of building using COOT (99) and refinement using REFMAC (Murshudov et al., 1997). Water molecules were added using COOT:FINDWATERS and manually inspected after refinement.

In all data sets, 5% of the observations were flagged as ‘free’ and used to monitor refinement procedures (Brünger, 1992). Model validation was performed with SFCHECK (Vaguine et al., 1999), PROCHECK (Laskowski et al., 1993) and MOLPROBITY (Chen et al., 2010) and data collection, structure and refinement statistics are shown in Table 2.

	GH101 Native	GH101 SeMet	GH101 + PugT	GH101 + T antigen
Data Collection				
Beamline	SSRL 9.2	CMCF1	SSRL 9.2	SSRL 9.2
Wavelength	0.97946	0.97905	0.97901	0.86700
Space group	$P2_1$	$P2_1$	$P22_12_1$	$P22_12_1$
Cell dimensions: *a, b, c* (Å)	76.26, 89.13, 88.57	78.01, 89.62, 87.18	87.06, 122.12, 139.84	87.07, 121.96, 139.60
Resolution (Å)	30-1.85 (1.95-1.85)	20.00-2.45 (2.58-2.45)	50.0-1.46 (1.53-1.46)	50-1.80 (1.86-1.80)
R_{merge}	0.100 (0.376)	0.128 (0.442)	0.077 (0.375)	0.059 (0.445)
$I/\sigma I$	16.9 (4.6)	17.0 (4.4)	16.6 (4.9)	24.8 (4.3)
Completeness (%)	99.0 (98.2)	99.8 (100.0)	99.6 (98.1)	99.8 (99.5)
Redundancy	6.8 (6.5)	7.3 (7.3)	6.8 (6.0)	8.1 (7.6)
Refinement				
Resolution (Å)	1.85		1.46	1.80
No. of reflections	88793		242255	129750
R_{work}/R_{free}	0.15/0.19		0.17/0.20	0.16/0.18
No. of atoms				
Protein	8765		8957	8900
Ion	4		4	4
Ligand*	N/A		36	32
Water	1517		1803	1618
B-factors				
Protein	12.3		11.2	14.5
Ion	13.2		13.5	16.5
Ligand	N/A		14.7	15.6
Water	23.8		25.5	29.0
Root mean square deviations				
Bond lengths (Å)	0.012		0.010	0.009
Bond angles (degrees)	1.320		1.322	1.165
Ramachandran				
Preferred (%)	97.6		97.6	97.5
Allowed (%)	2.2		2.2	2.2
Disallowed (%)	0.2		0.2	0.4

Table 2: X-ray crystallographic data collection and structure refinement statistics for GH101 and complexes.
Values in parentheses are for the highest resolution bin. *Refers to carbohydrates and carbohydrate derivatives.

2.3 Results and Discussion.

The family 101 glycoside hydrolase from *S. pneumoniae* is a large multi-modular enzyme, as is common for glycoside hydrolases, comprising 1767 amino acids in three definable domains or modules sandwiched by an N-terminal secretion signal peptide and a C-terminal LPXTG cell wall attachment motif (Figure 7). The first module of the *S. pneumoniae* TIGR4 GH101 following the signal peptide comprises 278 amino acids and extends to residue 316 and has no putative conserved domain architecture. The following amino acids, 317–1425, comprise the catalytic domain of the enzyme, here called SpGH101, and neighbouring the catalytic module is a putative carbohydrate-binding module. In an effort to characterize the structure of this *S. pneumoniae* TIGR4 protein, we cloned the gene fragment that we predicted to contain the catalytic module (SpGH101), recombinantly produced the ~124 kDa polypeptide in *E. coli* and purified it

in high yields. The resulting polypeptide qualitatively displayed good activity towards the synthetic substrate *p*-nitrophenyl-2-acetamido-2-deoxy-3-O-(β-d-galactopyranosyl)-α-d-galactopyranoside (Toronto Research Chemical Inc.) (data not shown). This was consistent with the classification of SpGH101 TIGR4 as an *endo*-α-*N*-acetylgalactosaminidase similar to the previously characterized homologue, GH101 from *S. pneumoniae* R6 (Caines et al., 2008). To provide greater insight into the molecular features that govern the activity of SpGH101 from *S. pneumoniae* TIGR4 the structure of this protein was determined using X-ray crystallography in apo-form and in complex with a chemically synthesized substrate analogue, O-[3-O-(1-β-D-galactopyrano)-2-*N*-Acetyl-2-deoxy-*D*-galactopyranosylidene]amino-*N*-Phenylcarbamate (PUGT), and serinyl-T antigen (serinyl-TAg).

2.3.1 Apo-structure of SpGH101.

The very large size and multi-modularity of SpGH101 makes it recalcitrant to crystallization necessitating the dissection into smaller fragments that retain catalytic activity and crystallize readily. The active fragment of SpGH101 that consists of the putative catalytic module, here called SpGH101, was crystallized in native form and seleno-methionine derivative crystals were obtained of sufficient quality to enable the determination of a high-resolution crystal structure of this protein.

The structure of SpGH101 was solved by single wavelength anomalous dispersion in its apo-form to a resolution of 1.85 Å in space group $P2_1$ with one molecule in the asymmetric unit (Figure 8). Despite a structure of the R6 GH101 being present in the Protein Database, we chose to use the seleno-methionine derivative dataset that we had collected to solve the structure of SpGH101. This structure revealed the distorted $(\beta/\alpha)_8$ barrel flanked by domains consisting of β-sheet character similar to the structure that was found for the SpGH101 structure from the R6 strain (Caines et al., 2008). These two structures aligned with a root mean square deviation (RMSD) of 0.587 Å over 1091 Cα residues, after N-terminus truncation, consistent with the 99% sequence identity between these two homologues. Four manganese ions were found to be coordinated in this structure but their roles do not appear to be involved in hydrolysis due to their location being not near the catalytic site.

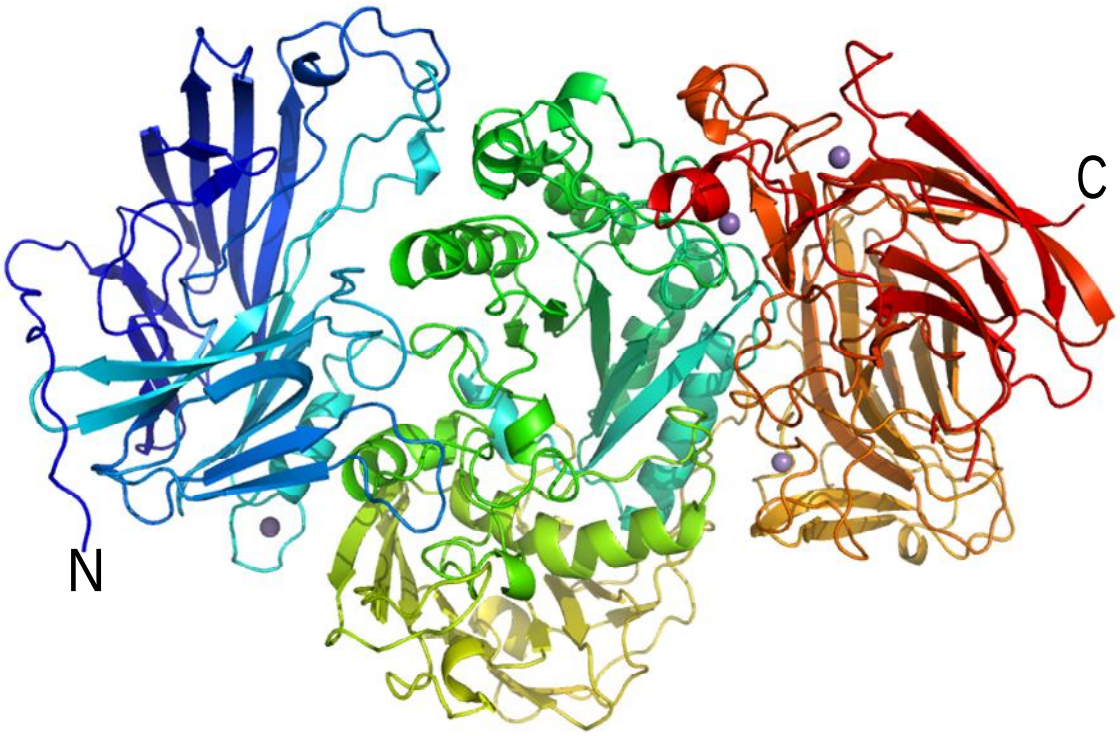


Figure 8: Apo-structure of *S. pneumoniae* TIGR4 GH101.

Ribbon representation of the SpGH101 shown in rainbow representation from the N- to C-termini coloured blue to red, respectively. The N- and C-termini are labeled accordingly. Four manganese ions are indicated by purple spheres.

2.3.2 Substrate analogue complex of SpGH101.

To this point, the structural basis of SpGH101 substrate interaction had only been defined based on automated docking analyses and substrate modelling. This motivated the pursuit of a substrate or substrate analogue structural complex for mechanistic expansion. A substrate analogue, O-[3-O-(1-β-D-galactopyrano)-2-*N*-Acetyl-2-deoxy-*D*-galactopyranosylidene]amino-*N*-phenylcarbamate (PUGT) (Figure 9) was chemically synthesized by Dr. Deng and Dr. Vocadlo at Simon Fraser University and was used as the candidate for complex formation with SpGH101. PUGT is very similar to a potent competitive inhibitor of β-*N*-acetylglucosaminidases called PUGNAc. One justification for the effectiveness of PUGNAC in inhibiting these other GHs is the sp^2 hydridization of the anomeric carbon which is thought to mimic the geometry of the dissociative transition state of GHs and is the reason for its design being very similar to many other inhibitors (Stubbs et al., 2006). In an attempt to ascertain if PUGT could inhibit the activity of

SpGH101 on the synthetic substrate *p*-nitrophenyl-2-acetamido-2-deoxy-3-O-(β-d-galactopyranosyl)-α-d-galactopyrano-side, inhibition was measured (data not shown). Unfortunately, the K_i was found to be too high to be detected indicating that PUGT was not an efficient or effective inhibitor of SpGH101. This is likely due to the planar nature of the anomeric carbon of PUGT which causes the phenylcarbamate to interfere with the enzyme active site residues.

Despite the lack of inhibition of SpGH101 by PUGT, a complex structure of SpGH101 and PUGT was obtained and provided substantial insight into how SpGH101 recognizes and hydrolyzes oligosaccharides. This high-resolution structure of SpGH101 in complex with PUGT was determined to 1.46 Å resolution, belonged to space group $P22_12_1$ and revealed clear electron density for PUGT (Figure 10).

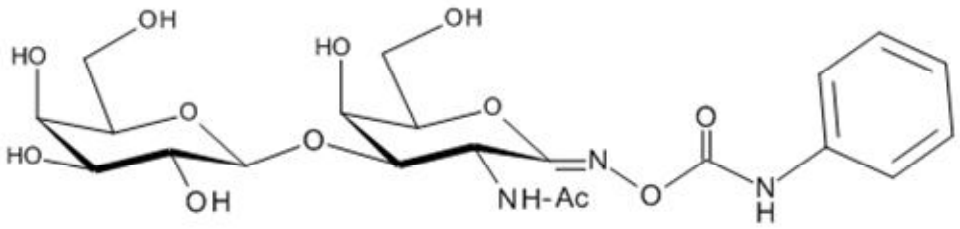


Figure 9: Schematic of O-[3-O-(1-β-D-galactopyrano)-2-*N*-Acetyl-2-deoxy-*D*-galactopyranosylidene]amino-*N*-Phenylcarbamate (PUGT).

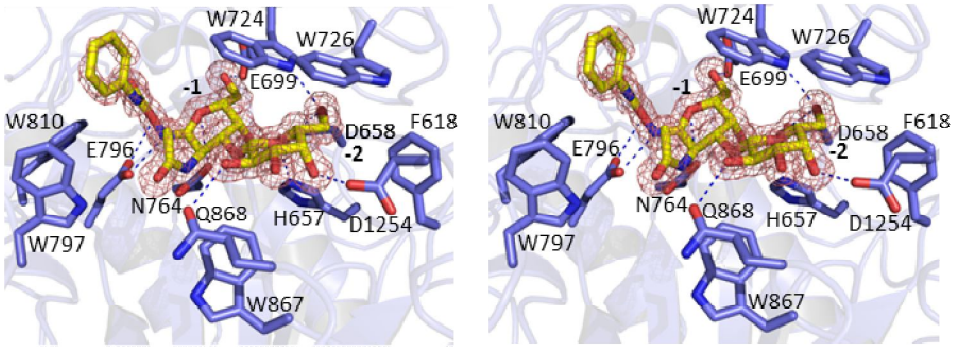


Figure 10: Active site representation of substrate analogue, PUGT, binding by SpGH101. A) Cartoon representation of SpGH101, blue, with bound PUGT, yellow. The red mesh shows the maximum likelihood /σ_a -weighted electron density maps contoured at 0.43 e/Å^{-3}. Key active site residues are shown in stick representation and coloured purple and PUGT coloured yellow. Active site subsite positions indicated as -2 and -1. Hydrogen bonds between the protein and substrate identified by using the criteria of proper geometry and a distance cutoff of 3.2 Å are shown as dotted blue lines.

Several of PUGT's hydroxyl groups are engaged in hydrogen bonds with residues in the SpGH101 active site. Notably, the PUGT hydroxyl groups that would be the equivalent of the O2, O4, O5 and O6 hydroxyls of the Gal residue of the native substrate interact with Gln868, Asp1254, His657, and Asp658, respectively and are buried in the active site (Figure 10). There are also many possibilities for van der Waals interactions between the Gal residue and the enzyme, consistent with the occupation of the -2 subsite by the Gal residue. There are also hydrogen bonds between the -1 subsite of the enzyme and hydroxyl groups from the PUGT GalNAc residue. These hydrogen bonds are between the amino acid side chains Asp658 and Glu699 with O5 and O6 of the GalNAc, respectively, and between the oxime N and O and the side chain of Glu796. Interestingly, only hydrogen bonds to waters were identified between the C2 acetyl of the PUGT GalNAc. Despite the lack of stabilizing hydrogen bonds between the GalNAc acetyl group and the enzyme active site, the acetyl appears to be positioned to form van der Waals interactions and is also surrounded by aromatic side chains that could have stabilizing effects.

The architecture of the SpGH101 active site is a shallow pocket with the -2 and -1 subsites embedded within the active site and lined by a large number of amino acid side

chains including several aromatic side chains. Aromatic residues that line the active site are Phe618, His657, Trp724, Trp726, Trp797, Trp810, and Trp867. Several of these aromatic side chains are positioned to provide hydrophobic platforms for interactions with the sugar rings. Trp726 and Trp724 are the most optimally positioned to stabilize the Gal and GalNAc from PUGT, respectively, and interestingly, both Trp724 and Trp726 appear to clamp down on these sugar rings compared to their relative positions in the native, apo-structure (Figure 11). In the apo-structure the electron density surrounding Trp726 is well defined but the side chain density for Trp724 appears to be disordered as judged by the sparse density surrounding the aromatic side chain; likely indicative of movement of this side chain. Comparing the native structure with the complex structure, it is apparent that the Trp724 and Trp726 Cα's move 1.8Å and 0.95Å, respectively, towards the PUGT in the complex structure concurrent with a side chain rotation of ~50° in relation to the Cα positions.

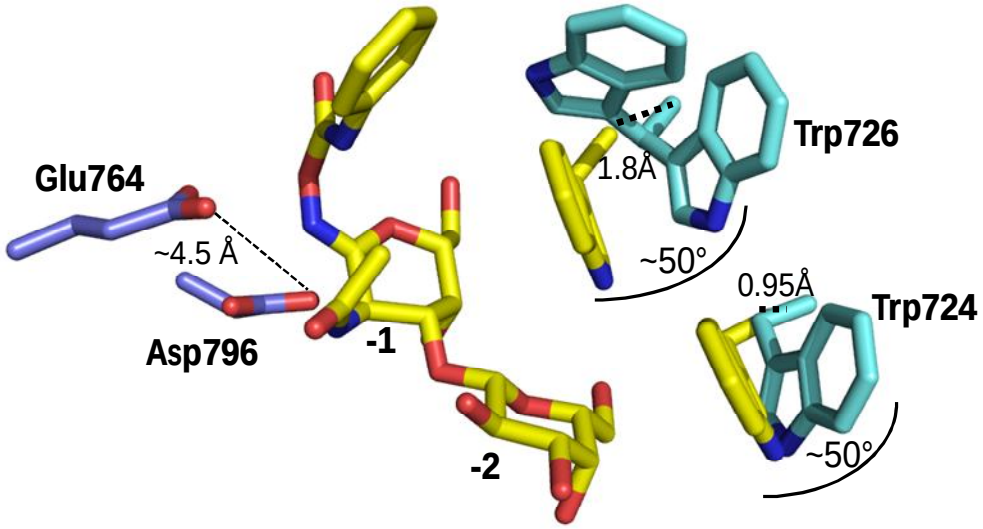


Figure 11: Induced fit movement of tryptophans 724 and 726 in SpGH101.
PUGT and Trp724 and 726 shown in stick representation coloured yellow. Position of apo-SpGH101 Trp724 and Trp726 shown in aqua blue. Positions of Glu764 and Asp796, the nucleophile and acid/base residues, respectively, shown in blue stick representation. Angles and distances between Cα positions indicated and subsites -1 and -2 indicated

Typically the -2 subsite of SpGH101 is occupied by a galactose residue, however if this galactose residue was substituted for a glucose residue it would not be structurally impeded by the active site architecture. Both the axial or equatorial position of the O4 hydroxyl could be accommodated but the hydrogen bond between the axial galactose O4 and Asp1254 would be abolished by this substitution. However, in the mucin O-glycan core structures (Figure 2), the galactose residue is not commonly found to be substituted for a glucose residue but a GlcNAc residue is commonly found in O-glycan core types 3 and 4. This lends significance to the ability of the enzyme to recognize acetylated sugars such as GlcNAc, and not unmodified glucose. There was a very small level of activity demonstrated by Kousioulis *et al.* for the GlcNAcβ1,3GalNAc, core type-3 structure that could easily be attributed to contamination of the commercial substrate or non-specific, residual cleavage (Koutsioulis et al., 2008). If the -2 subsite position of SpGH101 were occupied by a C2 acetylated residue such as GlcNAc, there would be spatial interference due to the side chain of Gln868. This would also have an effect on the hydrogen bond between Gln868 and the O2 of Gal, minimizing the ability of the enzyme to recognize other core structures, and would explain the favoured interaction of Galβ1,3GalNAc core type-1 structure over other core type structures.

A logical question is whether or not SpGH101 can efficiently hydrolyze the core type-1 monosaccharide, GalNAc, from serine or threonine eliminating the necessity of the -2 subsite to be occupied by Gal. Interestingly, there are considerably more hydrogen bonds between the enzyme and the Gal residue in the -2 subsite than the GalNAc residue in the -1 subsite of the enzyme. This indicates some significance to the -2 subsite being occupied for efficient hydrolysis. This importance was substantiated in a study by Willis *et al.*, where it was found that the activity of SpGH101 from strain R6 on Galβ1,3GalNAc had a 26,500-fold greater k_{cat}/K_m for the disaccharide substrate than just the monosaccharide, GalNAc (Willis et al., 2009). This large difference in selectivity was found to be due to an increase in affinity, as seen by the 70-fold decrease in K_m, and also due to a 400-fold increase in k_{cat}.

SpGH101 is incapable of accommodating a substrate longer than a disaccharide in its active site and an extended core type-1 O-glycan, such as a trisaccharide, would be sterically prevented from fitting in the active site. This specificity is probably due to the buried nature of the -2 subsite and the apparent lack of a -3 subsite. The structural inhibition of recognition of a longer glycon supports the finding that SpGH101 does not catalyze the liberation of longer, extended core type-1 O-glycans. This lack of activity on longer glycons would necessitate the O-glycan to be trimmed and processed by other enzymes to yield the disaccharide core type-1 structure. Marion *et al.* demonstrated that SpGH101 in conjunction with a sialidase, NanA, could deglycosylate the model glycocongugate fetuin (Marion et al., 2009). Fetuin consists of, on average, three N-glycan and three O-glycan chains per molecule with the O-glycans composed of sialylated core type-1 O-glycans, i.e. Siaα2,3Galβ1,3GalNac. That study illustrated that genetic knockouts of the genes encoding SpGH101 and NanA were unable to modify fetuin, but it did not clarify that the activity of SpGH101 was reliant upon the removal of sialic acid residues by NanA. While fetuin was a useful glycocongugate to test O-glycan deglycosylation, it is not an accurate representation of the diversity of O-glycans that are present in the mucin-rich niche of *S. pneumoniae.* However, it is apparent, that due to the specificity of SpGH101 to accommodate the serinyl-TAg and not longer glycons, pre-processing by other enzymes would be required. It is likely that the activity of other *S. pneumoniae* GHs such as GH2, GH20, and GH98, respectively a β-galactosidase, exo-β-*N*-acetylglucosaminidase, and endo-β-galactosidase in addition to sialidases and other GHs is required.

The residues that were previously assigned as the putative catalytic nucleophile and acid/base residues in *S. pneumoniae* R6 GH101 were Asp764 and Glu796, respectively, supported by the complete abrogation of enzyme activity on natural substrate when the catalytic residues were conservatively or non-conservatively mutated (Caines et al., 2008). These two catalytic residues are conserved in the TIGR4 SpGH101 and are situated ~4.5 Å apart and are positioned optimally for the double displacement retaining mechanism that was observed for the R6 GH101 and the *B. longum* GH101 (Willis et al., 2009; Fujita et al., 2005).

2.3.4 Serinyl-T antigen complex reveals aglycon specificity.

While the structure of the SpGH101 PUGT complex contributed insight into the occupation of the -2 and -1 subsites, there still remained questions as to the nature of the aglycon specificity. This was provided by a complex of SpGH101 with the putative nucleophile residue mutated, Asp764 to Asn764, with unhydrolyzed serinyl-Galβ1,3GalNAc (serinyl-TAg) (Figure 12) determined to 1.86 Å resolution (Figure 13).

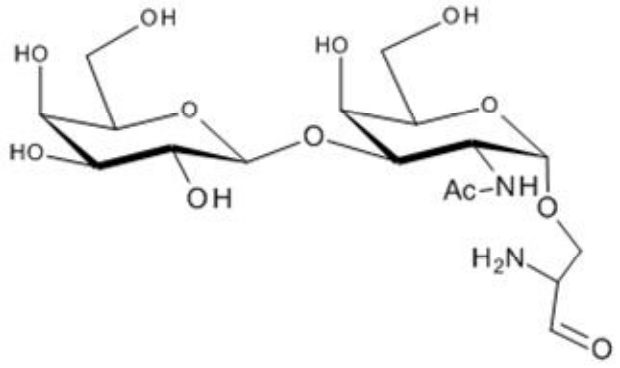


Figure 12: Schematic of serinyl-T antigen.

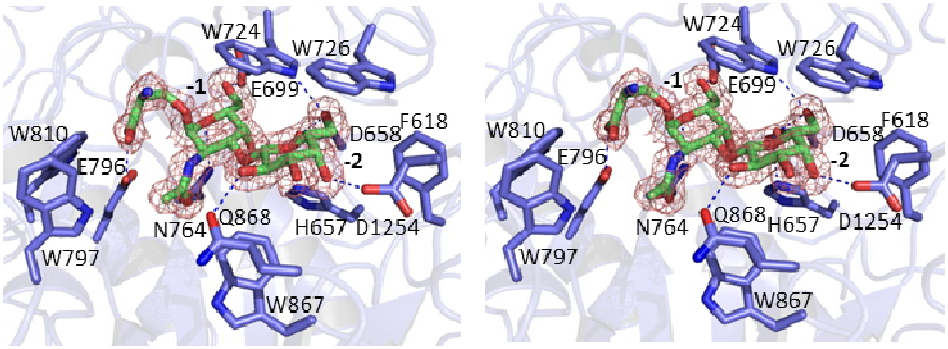


Figure 13: Active site representation of serinyl-T antigen binding by SpGH101.

Cartoon representation of SpGH101, blue, with bound serinyl-TAg, green. The red mesh shows the maximum likelihood /σ_a -weighted electron density maps contoured at 0.43 e/Å^{-3}. Key active site residues are shown in stick representation and coloured purple and serinyl-TAg coloured green. Active site subsite positions indicated as -2 and -1. Hydrogen bonds between the protein and substrate identified by using the criteria of proper geometry and a distance cutoff of 3.2 Å are shown as dotted blue lines.

In this structure the serinyl-TAg was well-ordered and positioned very similarly to the PUGT with the potential hydrogen bonds and van der Waals interactions between the active site and substrate conserved. Though well-ordered, the serine residue of the serinyl-TAg made very little direct interactions with the SpGH101 catalytic site and the serine residue was positioned protruding from the active site, suggesting that little specificity is contributed by this feature of the substrate and that the enzyme lacks a defined +1 subsite. Likewise, the position of the phenylcarbamate in the PUGT complex protrudes from the active site and is not accommodated in the catalytic pocket (Figure 13). There was one possible hydrogen bond identified between the serine carboxylic acid and Glu796, the putative acid/base residue. Based on this information it is plausible that the serine could be replaced by a threonine which is relevant considering that O-glycans are often found attached to protein α-linked to either serine or threonine. Beyond the requirement for the disaccharide to be linked to the protein by a serine or threonine residue, there does not appear to be any restriction as to the specific amino acid sequence of the protein/polypeptide. There does not appear to be any structural impediment for accommodating the protein portion of the O-glycan due to the broad, open surface of SpGH101 and the substantially open environment would provide room for mucousal interactions (Figure 14). It seems unlikely that SpGH101 would be selective based on different core protein sequences, more studies are required however to substantiate this claim. Nevertheless, it has been demonstrated that an endo-β-*N*-acetylglucosaminidase isolated from *S. pneumoniae* cultures was capable of hydrolyzing the O-glycosidic linkage between GalNAc and serine/threonine releasing the disaccharide Galβ1,3GalNAc from mucins (Umemoto et al., 1977). This indicates that the enzyme has the capability of liberating the disaccharide from O-glycan found in mucins and not simply a serine or threonine residue.

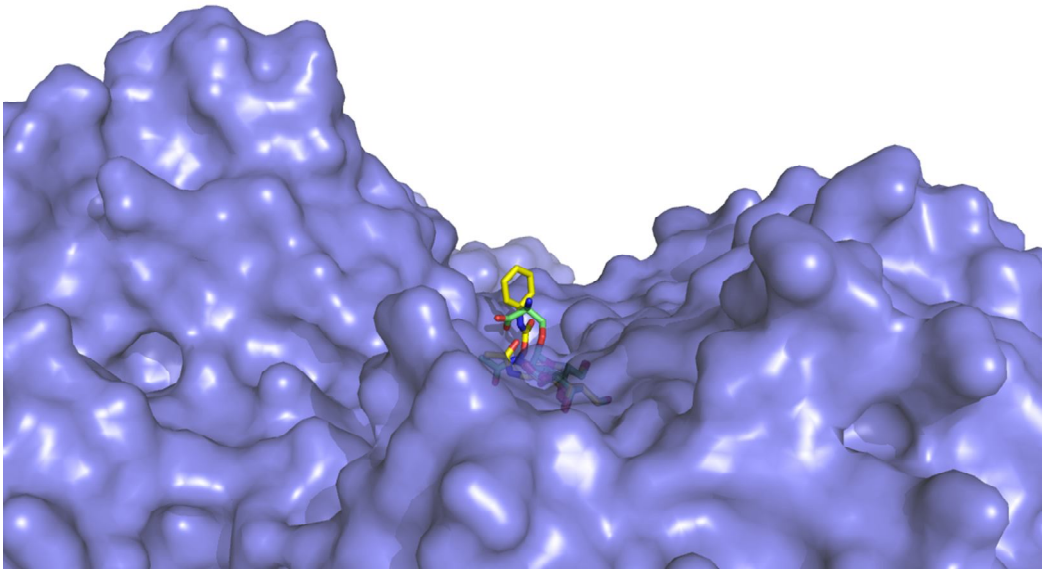


Figure 14: Surface representation of SpGH101 with PUGT and serinyl-T antigen bound. SpGH101 surface representation shown in blue. PUGT (yellow) and serinyl-TAg (green) shown in stick representation.

The only other structure of a glycoside hydrolase from family 101, aside from the GH101s from *S. pneumoniae,* is from *B. longum* (pdb accession code 2ZXQ). This *B. longum* GH101 (BfGH101) structure is very similar to that of the SpGH101s with 44% identity and an RMSD of 1.28 Å over 1024 Cα atoms (Suzuki et al., 2009). The active site residues are entirely conserved between the SpGH101 and the BfGH101 and the active site architecture is nearly identical (Figure 15) and the following active site residues are structurally conserved between both enzymes, listed SpGH101/BfGH101, respectively; Phe618/642, His657/681, Asp658/682, His661/685, Asn696/720, Glu699/723, Trp724/748, Trp726/750, Asp764/789, Glu796/822, Trp797/823, Trp810/836, Trp867/893, Gln868/894, and Asp1254/1295. As previously mentioned, the SpGH101 and BfGH101 have similar specificity for the core type-1 O-glycan disaccharide. This is further supported by the absolute conservation of active site residues.

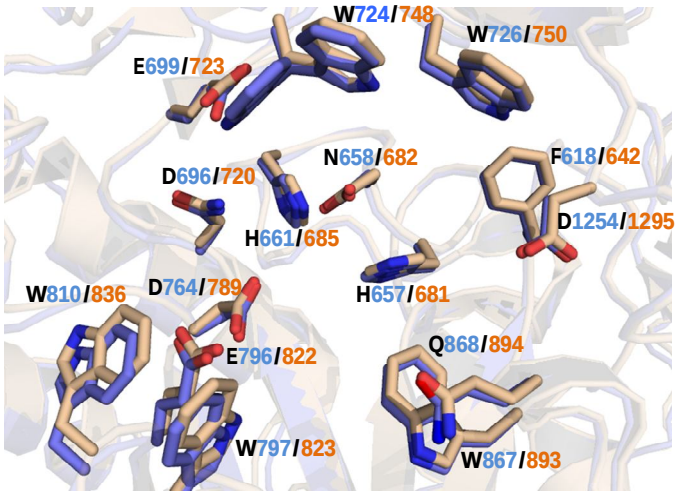


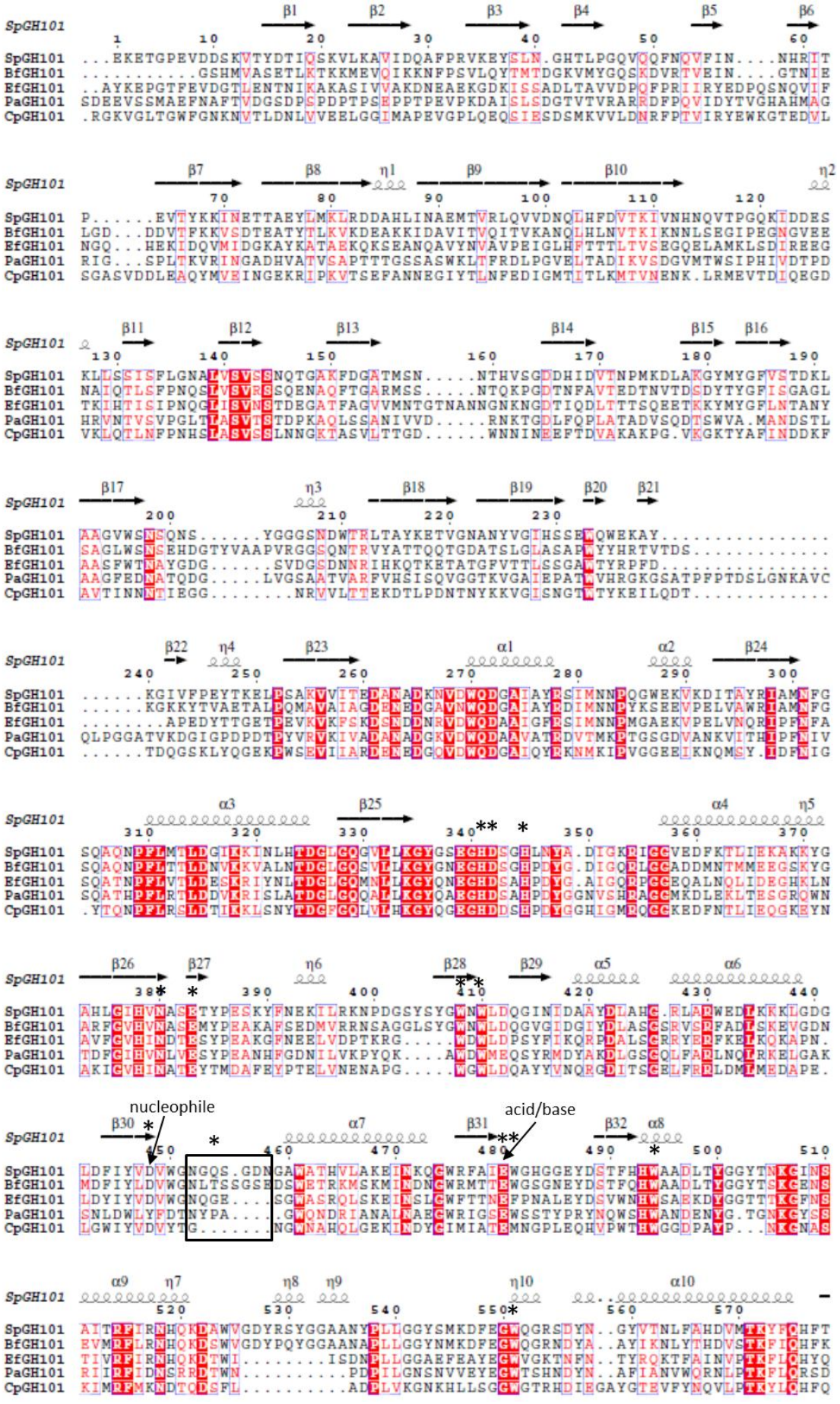
Figure 15: Structural overlay of SpGH101 and BfGH101 active sites.
Active sites of SpGH101 (blue) and BfGH101 (taupe) shown in cartoon representation with active site residues shown in stick representation and labeled and numbered SpGH101/BfGH101.

2.3.5 GH101 comparisons.

An alignment between the SpGH101, BfGH101, and GH101s from *C. perfringens* (CpGH101), *Propionibacterium acnes* (PaGH101), and *Enterococcus faecalis* (EfGH101) reveals sequence conservation of the key active site residues (Figure 16). One notable difference in the conservation of the active site residues is that in the PaGH101 the residue that aligns with the putative nucleophile is a tyrosine and not an aspartate. There are however, two other possible Asp residues that could be nucleophile candidates, Asp650 and Asp655.

Both the CpGH101 and EfGH10 are able to hydrolyze the core type-2 trisaccharide Galβ1,3(GlcNAcβ1,6)GalNAc. Upon analyzing the side chains involved in blocking the hydrolysis of this substrate by SpGH101 and BfGH101, amino acid Gln770 appeared to be responsible for blocking entry of the β1,6 branched GlcNAc. This Gln770 residue appears on a loop structure that connects a β-strand to an α-helix and, interestingly, the amino acids that corresponded to this loop are either absent or different in the CpGH101 and EfGH101 (Figure 16-indicated by a box) (Figure 17). It appears that this loop is responsible for preventing the recognition of core type-2 O-glycans and their β-1,6

branched GalNAc by physically blocking the recognition of the β-1,6 linked GalNAc. Also interesting is that the EfGH101 is able to hydrolyze the Gal extended core type-2 tetrasaccharide (Galβ1,4GlcNAcβ1,6[Galβ1,3]GalNAc). The absence of this specificity conferring loop likely is responsible for the accommodation of the Gal. This putative specificity loop is also absent in the PaGH101 but this enzyme cannot liberate the core type-2 structures. As previously mentioned the SpGH101 active site revealed that Gln868 would structurally inhibit the accommodation of an acetylated sugar at the -2 subsite, inhibiting recognition and hydrolysis of the core type-3 O-glycan. Based on the sequence alignment, the amino acids that likely correspond with Gln868 are a valine, threonine and a glycine in EfGH101, PaGH101, and CpGH101, respectively. These shorter side chains would likely allow for recognition of an acetylated residue in the -2 position. This was found to be the case for PaGH101 and EfGH101 which can liberate the core type-3 trisaccharide. The reason for CpGH101 not accommodating this core type is not understood but further structural studies should be useful in clarifying this.


SpGH101
β1 β2 β3 β4 β5 β6
SpGH101 ...EKETGPEVDDSKVTYDTIQSKVLKAVIDQAFPRVKEYSLN.GHTLPGQVQQFNQVFIN....NHRIT
BfGH101 ..........GSHMVASETLKTKKMEVQIKKNFPSVLQYTMTDGKVMYGQSKDVRTVEIN....GTNIE
EfGH101 ..AYKEPGTFEVDGTLENTNIKAKASIVVAKDNEAEKGDKISSADLTAVVDPQFPRIIRYEDPQSNQVIF
PaGH101 SDEEVSSMAEFNAFTVDGSDPSPDPTPSEPPTPEVPKDAISLSDGTVTVRARRDFPQVIDYTVGHAHMAG
CpGH101 .RGKVGLTGWFGNKNVTLDNLVVEELGGIMAPEVGPLQEQSIESDSMKVVLDNRFPTVIRYEWKGTEDVL
β7 β8 η1 β9 β10 η2
SpGH101 P.....EVTYKKINETTAEYLMKLRDDAHLINAEMTVRLQVVDNQLHFDVTKIVNHNQVTPGQKIDDES
BfGH101 LGD...DDVTFKKVSDTEATYTLKVKDEAKKIDAVITVQITVKANQLHLNVTKIKNNLSEGIPEGNGVEE
EfGH101 NGQ...HEKIDQVMIDGKAYKATAEKQKSEANQAVYNVAVPEIGLHFTTTLTVSEGQELAMKLSDIREEG
PaGH101 RIG...SPLTKVRINGADHVATVSAPTTTGSSASWKLTFRDLPGVELTADIKVSDGVMTWSIPHIVDTPD
CpGH101 SGASVDDLEAQYMVEINGEKRIPKVTSEFANNEGIYTLNFEDIGMTITLKMTVNENK.LRMEVTDIQEGD
β11 β12 β13 β14 β15 β16
SpGH101 KLLSSISFLGNALVSVSSNQTGAKFDGATMSN.....NTHVSGDDHIDVTNPMKDLAKGYMYGFVSTDKL
BfGH101 NAIQTLSFPNQSLVSVRSSQENAQFTGARMSS.....NTQKPGDTNFAVTEDTNVTDSDYTYGFISGAGL
EfGH101 TKIHTISIPNQGLISVNSTDEGATFAGVVMNTGTNANNGNKNGDTIQDLTTTSQEETKKYMYGFLNTANY
PaGH101 HRVNTVSVPGLTLASVTSTDPKAQLSSANIVVD.....RNKTGDLFQPLATADVSQDTSWVA.MANDSTL
CpGH101 VKLQTLNFPNHSLASVSSLNNGKTASVLTTGD.....WNNINEEFTDVAKAKPG.VKGKTYAFINDDKF
β17 η3 β18 β19 β20 β21
SpGH101 AAGVWSNSQNS.....YGGGSNDWTRLTAYKETVGNANYVGIHSSELQWEKAY.............
BfGH101 SAGLWSNSEHDGTYVAAPVRGGSQNTRVYATTQQTGDATSLGLASAPLYYHRTVTDS.............
EfGH101 AASFWTNAYGDG.....SVDGSDNNRIHKQTKETATGFVTTLSSGALTYRPFD...............
PaGH101 AAGFEDNATQDG.....LVGSAATVARFVHSISQVGGTKVGAIEPATWVHRGKGSATPFPTDSLGNKAVC
CpGH101 AVTINNNTIEGG.....NRVVLTTEKDTLPDNTNYKKVGISNGTTYTKEILQDT.............
β22 η4 β23 α1 α2 β24
SpGH101 ......KGIVFPEYTKELPSAKVVITEDANADKNVDWQDGAIAYRSIMNNPQGWEKVKDITAYRIAMNFG
BfGH101 ......KGKKYTVAETALPQMAVAIAGDENEDGAVNWQDGAIAYRDIMNNPYKSEEVPELVAWRIAMNFG
EfGH101 ........APEDYTTGETPEVKVKFSKDSNDDNRVDWQDAAIGFRSIMNNPMGAEKVPELVNQRIPFNFA
PaGH101 QLPGGATVKDGIGPDPDTPYVRVKIVADANADGKVDWQDAAVATRDVTMKPTGSGDVANKVITHIPFNIV
CpGH101 ......TDQGSKLYQGEKPWSEVIIARDENEDGQVDWQDGAIQYRKNMKIPVGGEEIKNQMSY.IDFNIG
α3 β25 α4 η5
SpGH101 SQAQNPFLMTLDGIKKINLHTDGLGQGVLLKGYGSEGHDSGHLNYA.DIGKPIGGVEDFKTLIEKAKKYG
BfGH101 SQAQNPFLTTLDNVKKVALNTDGLGQGSVLLKGYGNEGHDSGHPDYG.DIGQPLGGADDMNTMMEEGSKYG
EfGH101 SQATNPFLVTLDESKRIYNLTDGLGQGMNLLKGYQNEGHDSAHPDYG.AIGQPPGGEQALNQLIDEGHKLN
PaGH101 SQATHPFLRTLDDVKRISLATDGLGQGAALLKGYQAEGHDSAHPDYGGNVSHRAGGMKDLEKLTESGRQWN
CpGH101 .YTQNPFLRSLDTIKKLSNYTDGFGQGLVLHKGYQGEGHDDSHPDYGGHIGMPQGGKEDFNTLIEQGKEYN
β26 β27 η6 β28 β29 α5 α6
SpGH101 AHLGIHVNASETYPESKYFNEKILRKNPDGSYSYGWNWLDQGINIDAAYDLAHGRLAPWEDLKKKLGDG
BfGH101 ARFGVHVNASEMYPEAKAFSEDMVRRNSAGGLSYGWNWLDQGVGIDGIYDLASGSRVSPFADLSKEVGDN
EfGH101 AVFGVHINDTESYPEAKGFNEELVDPTKRG.....WDWLDPSYFIKQRPDALSGRRYEPFKELKQKAPN.
PaGH101 TDFGIHVNLVESYPEANHFGDNILVKPYQK....AWDWMEQSYRMDYAKDLGSGQLFARLNQLRKELGAK
CpGH101 AKIGVHINATETYTMDAFEYPTELVNENAPG....WGWLDQAYYVNQRGDITSGELFRRLDMLMEDAPE.
nucleophile
β30 α7 β31 acid/base β32 α8
SpGH101 LDFIYVDVWGNGQS.GDNGAWATHVLAKEINKQGWRFAIEWGHGGEYDSTFHHWAADLTYGGYTNKGINS
BfGH101 MDFIYLDVWGNLTSSGSEDSWETRKMSKMINDNGWRMTTEWGSGNEYDSTFQHWAADLTYGGYTSKGENS
EfGH101 LDYIYVDVWGNQGE....SGWASRQLSKEINSLGWFTTNEFPNALEYDSVWNHWSAEKDYGGTTTKGFNS
PaGH101 SNLDWLYFDTNYPA....GWQNDRIANALNAEGWRIGSEWSSTYPRYNQWSHWANDENYG.TGNKGYSS
CpGH101 LGWIYVDVYTG.....NGWNAHQLGEKINDYGIMIATEMNGPLEQHVPWTHWGGDPAYP...NKGNAS
α9 η7 η8 η9 η10 α10
SpGH101 AITRFIRNHQKDAWVGDYRSYGGAANYPLLGGYSMKDFEGWQGRSDYN..GYVTNLFAHDVMTKYFQHFT
BfGH101 EVMRFLRNHQKDSWVGDYPQYGGAANAPLLGGYNMKDFEGWQGRNDYA..AYIKNLYTHDVSTKFIQHFK
EfGH101 TIVRFIRNHQKDTWI.........ISDNPLLGGAEFEAYEGWVGKTNFN..TYRQKTFAINVPTKFLQHYQ
PaGH101 RIIRFIDNSRRDTWN.........PDPILGNSNVVEYEGWTSHNDYN..AFIANVWQRNLPTKFLQRSD
CpGH101 KIMRFMKNDTQDSFL.........ADPLVKGNKHLLSGGWGTRHDIEGAYGTEVFYNQVLPTKYLQHFQ

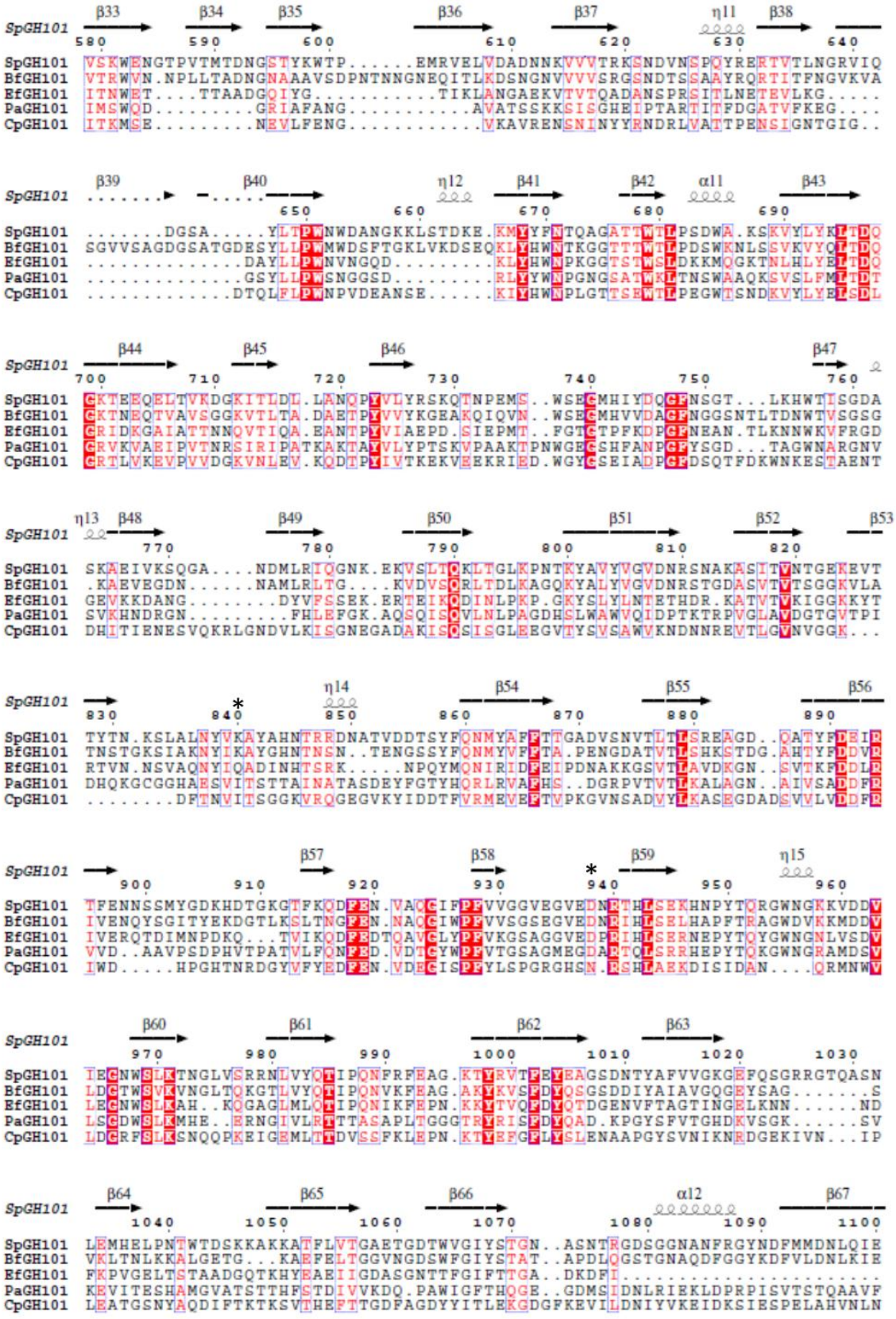

SpGH101
BfGH101
EfGH101
PaGH101
CpGH101

Figure 16: GH101 sequence alignment.

Alignment of amino acid sequences from SpGH101 *(S. pneumoniae* TIGR4), BfGH101 *(B. longum* NCC2705*)*, EfGH101 *(E. faecalis)*, PaGH101 (*P. acnes* KPA171202) and CpGH101 (*C. perfringens* ATCC13124) using ClustalW2. This figure was generated using ESPript with secondary structure elements indicated above sequence from structure file from SpGH101 TIGR4. Conserved active site residues denoted by asterisks (*). Sequence corresponding to loop region demarcated by box.

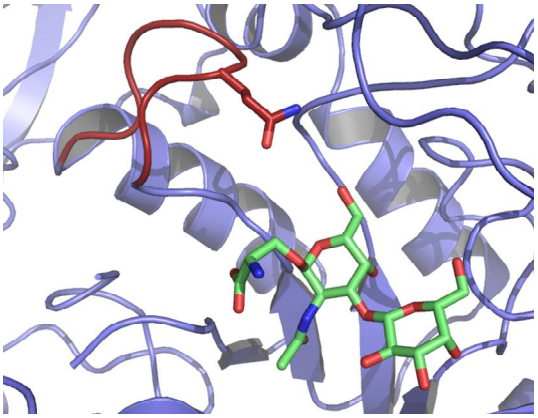


Figure 17: SpGH101 specificity conferring loop region.

Cartoon representation of SpGH101 (blue) with bound serinyl-TAg (green-stick representation). Loop region corresponding to amino acids 768-775 shown in red, with side chain of Gln770 shown in red stick representation.

acetylgalactosaminidase from glycoside hydrolase family 101 from *S. pneumoniae*. Our hypotheses were supported in that the SpGH101 active site has defined -2 and -1 subsites that accommodated the core type-1 disaccharide, Galβ1,3GalNAc, and that the nature of the aglycon was not specific beyond the α-linked serine or threonine residue. We speculated that this would allow the enzyme to liberate the disaccharide from a variety of core proteins found in mucin, but further structural and specificity studies are required. Amino acid sequence alignments with other bacterial GH101 homologues provided information as to why the other GH101 homologues have varied specificities but further structural studies are required.

The occurrence of glycoside hydrolases from family 101 in *S. pneumoniae* among other notable pathogens, such as *C. perfringens, E. faecalis* and other streptococci, imply that the breakdown of the core structures of O-glycans is of importance to pathogens. Surprisingly, there have only been 69 GH101s identified and all of them are bacterial lending further significance to the potential involvement of GH01s in bacterial fitness and pathogenesis. Also, all of the currently sequenced *S. pneumoniae* genomes have a gene with high identity to the SP_0368 that encodes SpGH101, suggesting that this enzyme is common to all strains. The importance of SpGH101 was demonstrated by Marion *et al.* to be involved in the adherence to human airway epithelial cells and establishment of colonization of the upper respiratory tract. This suggests that the cleavage of core type-1 O-glycans in mucin and on epithelial cells enhances the ability of *S. pneumoniae* to colonize its human niche and that degradation of O-glycans may contribute to the pathogenesis of *S. pneumoniae.*

Chapter 3: Analysis of a new family of metal-independent α-mannosidases provides unique insight into the processing of N-linked glycans

Katie Gregg [‡] , Wesley F. Zandberg [*] , Jan-Hendrik Hehemann [‡], Garrett E. Whitworth [*], Lehua Deng [*], David J. Vocadlo [*] and Alisdair B. Boraston [‡]

[‡]Department of Biochemistry and Microbiology, University of Victoria, Victoria, BC, Canada V8W 3P6;
*Department of Chemistry, Simon Fraser University, Burnaby, BC, Canada V5A 1S6

Adapted from: Journal of Biological Chemistry. 2011; 286(17): 15586-96

Contributions to Research: Cloning, protein production and purification, activity assays, crystallization and solution, manuscript and figure preparation

3.1 Introduction.

A feature of emerging importance to bacteria that colonize or infect humans is their capacity to process host glycans. *Streptococcus pneumoniae* is one notable human pathogen that relies on this ability for its full virulence (King, 2010). Among its known carbohydrate-active virulence factors are NanA, StrH, BgaA, and EndoD. NanA, StrH, and BgaA are a sialidase, an *exo*-β-D-*N*-acetylglucosaminidase, and an *exo*-β-D-galactosidase, respectively, which sequentially remove the terminal sugars from the distal arms of complex N-linked glycans. EndoD is an *endo*-β-D-*N*-acetylglucosaminidase that cleaves the chitobiose core of N-linked glycans smaller than Man5GlcNAc2 to remove the glycan from the protein scaffold. Despite increasing knowledge in this area it remains unclear how bacteria process the core mannose component of N-linked glycans.

α-mannosidases known to process N-glycans are found in families 38, 47, 76, 92 and 99. Very recent studies have shown the bacterial family 38 α-mannosidase from *Streptococcus pyogenes* (SpyGH38) to be a specific *exo*-α-1,3-mannosidase that is tolerant of the α-1,6-branches in N-glycans (Suits et al., 2010). Analysis of family 92 glycoside hydrolases from the human gut symbiont *Bacteroides thetaiotaomicron* revealed an expanded repertoire of α-mannosidases (Zhu et al., 2010). These enzymes

displayed activity primarily toward α-1,2- and α-1,3-mannosidic linkages with some having low α-1,6-mannosidase activity.

The genome of *S. pneumoniae* TIGR4 contains an ORF encoding a putative protein with high amino acid sequence identity to SpyGH38 (locus tag SP_2143) and an ORF encoding a putative protein with sequence similarity to family 92 glycoside hydrolases (locus tag SP_2145). These two genes are separated by an ORF encoding a putative protein of unknown function (SP_2144). Also present at this locus are a putative α-fucosidase (GH29, SP_2146), a putative ROK protein (SP_2142), and a putative α-hexosaminidase (SP_2141). This locus, which appears to be dedicated to glycan processing, is considered a virulence "hot-spot" as five of the genes (locus tags SP_2142 to SP_2146) have been identified as virulence factors in several screens (Polissi et al., 1998; Obert et al., 2006; Hava and Camilli, 2002). The presence of two genes encoding putative GH38 and GH92 α-mannosidases provides evidence that *S. pneumoniae* can process α-mannosides, likely those that are α-1,3- and α-1,2-linked. Of considerable interest is the unknown protein, SP_2144, whose association with a glycan processing locus suggests it is a carbohydrate-active protein belonging to an uncharacterized family. SP_2144 shows ~40% amino acid sequence identity to four proteins whose structures were deposited in the Protein Data Bank (2nvp from *C. perfringens* strain 13; 2p0v from *Bacteroides thetaiotaomicron;* 3p2c and 3on6, both from *Bacteroides ovatus*). These structures display an $(\alpha/\alpha)_6$ fold and have substantial structural identity with the family 15 glycoside hydrolases, which are primarily α-glucanases.

In addition to the established ability of *S. pneumoniae* to *exo*-hydrolytically process the distal arms of complex glycans, which comprise sialic acid, galactose, and *N*-acetylglucosamine, consideration of additional putative carbohydrate-active enzyme found in this organism suggests it can partly degrade the mannose component of N-glycans using enzymes similar to those found in *S. pyogenes* and *B. thetaiotaomicron*. Through these observations it has become clear that some bacteria, possibly including *S. pneumoniae*, have the capacity to process the mannose component of N-glycans. A noteworthy gap, however, in the known bacterial N-glycan degradation pathway is that

efficient α-1,6-mannosidases that would be required for complete N-glycan depolymerization have not yet been found. Thus, there is a clear possibility that further study of the bacterial glycan processing machineries might reveal new catabolic enzymes and additional insight into bacterial N-glycan degradation. The genomic context of SP_2144 and its similarity to proteins having structural identity to carbohydrate-active enzymes led to the hypothesis that SP_2144, and by extension its homologs, would encode a protein with activity on carbohydrates.

The objectives of this study are to structurally and functionally characterize the gene product of SP_2144 to determine if it has carbohydrate-activity and to determine its specificity and mechanism of hydrolysis.

This will be approached by analyzing SP_2144 from *S. pneumoniae* TIGR4 and a homolog from *C. perfringens* ATCC 13124. In addition to structural and functional studies we will seek to determine a substrate for these gene products and determine their specificity and catalytic mechanism.

3.2 Experimental Procedures.

Cloning, production and purification of GH125. The full-length *spgh125* gene (locus tag SP_2144) was PCR amplified from *S. pneumoniae* TIGR4 genomic DNA (ATCC BAA-334D) using the following forward and reverse oligonucleotide primers 5'- GGC AGC CAT ATG ATG GTT TAT TCG AAA G-3' and 5'- GTG GTG CTC GAG TTA GCG AAT ATC CAA GTA ATC C-3', respectively. The full length *cpgh125* gene (locus tag CPE0426) was PCR amplified from *C. perfringens* ATCC 13124 genomic DNA using the following forward and reverse oligonucleotide primers 5'- TAT ACC ATG GCC AGT TTA TCA ACT AAC GAA TT-3' and 5' GTG GTG CTC GAG TTT TTT ATT TAT AAC TTT CTC-3', respectively.

The PCR amplified gene fragments were obtained using standard PCR methods with Phusion High-Fidelity DNA polymerase (New England Biolabs). The amplified *spgh125* gene was cloned into pET-28a (+) (Novagen) via engineered 5' and 3' *NdeI* and *XhoI*

restriction sites, respectively. The amplified *cpgh125* gene was similarly cloned with 5' and 3' *NcoI* and *XhoI* restriction sites, respectively. Standard cloning procedures were used. The plasmid containing the *spgh125* gene encoded the polypeptide preceded by an N-terminal, thrombin cleavable six-histidine tag. The plasmid containing the *cpgh125* gene encoded the polypeptide followed by a C-terminal six-histidine tag. The DNA sequences of the constructs were verified by bidirectional sequencing.

The appropriate plasmids were transformed into chemically competent *E. coli* BL21 STAR (DE3) cells (Novagen). SpGH125 and CpGH125 proteins were produced in Luria-Bertani media containing 50 μg ml^{-1} kanamycin (Sigma). The cells were grown at 37°C to an optical density of 0.5 at A_{595} and induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside at 18°C for 14 hours. Cells were harvested by centrifugation at 27 000 x g and chemically lysed (Charlwood et al., 1998). Cell debris was pelleted using centrifugation at 27 000 x g for 45 minutes. The polypeptides were purified from cell-free extract using immobilized metal affinity chromatography using previously described methods (Boraston et al., 2001). The purity of fractions was assessed using SDS-PAGE and those deemed to be greater than 95% pure were pooled, concentrated and buffer exchanged into 20 mM Tris-HCl, pH 8.0, in a stirred ultra-filtration unit (Amicon) using a 10K molecular weight cut-off membrane (Filtron). Protein was further purified by size exclusion chromatography using Sephacryl S-200 (GE biosciences) in 20 mM Tris-HCl, pH 8.0. The concentration of purified protein was determined by UV absorbance at 280 nm using calculated molar extinction coefficients of 103 625 M^{-1} cm^{-1} and 91 010 M^{-1} cm^{-1} for SpGH125 and CpGH125, respectively (Gasteiger et al., 2003).

GH125 activity assay. The activities of SpGH125 and CpGH125 were initially screened on a variety of chromogenic aryl-glycosides (Table 3). These assays were performed with 500 nM enzyme and 0.5 mM substrate, all in phosphate buffered saline (PBS), pH 7.4, at 37°C. The reaction was monitored spectrophotometrically at 400 nM. Kinetic studies using the 2,4-dinitrophenyl α-D-mannopyranoside (DNP-Man) were carried out at 25°C using a Spectramax $Plus^{384}$ microplate reader (Molecular Devices, California). Standard-reaction mixtures comprised 200 μL reactions in PBS pH 7.5, containing 0.1% BSA

(Sigma), 2 µM enzyme, and 100 µM to 2.5 mM DNP-Man. The production of 2,4-dinitrophenolate was monitored at 400 nm over 30 minutes. The reaction velocity was determined by linear regression of the data within the early time period of the assay. Absorbance values were converted to concentration terms using an extinction coefficient (at λ=400 nm) of 10.91 mM^{-1} cm^{-1} for the dinitrophenolate at pH 7.4. To determine if the presence of EDTA would eliminate or reduce the catalytic activity of SpGH125 and CpGH125 toward DNP-Man, kinetic studies were carried out as described above in the presence of 10 mM EDTA with 1 mM DNP-Man as the substrate. Non-chromogenic substrates (Table 3) were tested for activity using high performance anion exchange chromatography using pulsed amperometric detection (HPAEC-PAD) to monitor glycoside cleavage. Assays were performed in 20 µL volumes at 37°C in PBS at pH 7.5. Reactions were initiated by the addition of 500 nM enzyme to 0.5 mM of each carbohydrate. Reactions were stopped by the addition of 140 µL of 100 mM sodium hydroxide. After centrifugation at 5 000 rpm for 5 minutes, 20 µL of the samples were analyzed by HPAEC-PAD using a Dionex ICS 3000 HPLC equipped with an ASI 100 Automated sample injector and an ED50 electrochemical detector (Dionex) with a gold working electrode and an Ag/AgCl reference electrode. Products were analysed using a PA-20 column set (analytical plus guard column) using an isocratic gradient of 100 mM NaOH.

Carbohydrates Tested	Activity
α(1,6)-mannobiose	+
α(1,2)-mannobiose	-
α(1,3)-mannobiose	-
α(1,3-1,6)-mannotriose	-
Galactose β(1,3) N-acetylgalactose	-
N-acetylgalctosamine α(1,3) Galactose	-
Galactobiose	-
Melibiose	-
Maltose	-
Isomaltose	-
Trehalose	-
Kojibiose	-
Nigerose	-
4-Nitrophenyl N-acetyl-β-D-glucosaminide	-
o-Nitrophenyl-N-acetyl-β-D-Galactosaminide	-
o-Nitrophenyl β-D-xylo-pyranoside	-
p-Nitrophenyl β-D-Gluco-pyranoside	-
p-Nitrophenyl α-D-Galactopyranoside	-
p-Nitrophenyl α-L-fucopyranoside	-
4-Nitrophenyl N-acetyl-α-D-glucosaminide	-
4-Nitrophenyl α-D-mannopyranoside	-
2,4-Dinitrophenyl α-D-mannopyranoside	+
p-Nitrophenyl β-D-galactopyranoside	-
4-Nitrophenyl N-acetyl-α-D-galactosaminide	-

Table 3: Carbohydrates tested for GH125 activity.

Crystallization and X-ray Data Collection. All crystals were obtained using hanging or sitting-drop vapour diffusion at 18°C. Prior to crystallization, SpGH125 and CpGH125 were concentrated to 15 mg ml^{-1} in 20 mM Tris pH 8.0. SpGH125 native crystals were obtained in 0.1 M Hepes, pH 8.0, and 1.2 M $LiSO_4$. CpGH125 native crystals were obtained in 0.1 M MES, pH 6.5, and 12% polyethylene glycol 20 000 (Hampton Research).

A complex of SpGH125 with the inhibitor 1-deoxymannojirimycin (dMNJ; Toronto Research Chemicals) was produced by soaking native crystals in crystallization solution

containing excess of dMNJ. A similar procedure was used for the non-productive substrate complex of SpGH125 with α-1,6-mannobiose; however, in this case artificial crystallization solution comprising 0.1 M CAPSO, pH 9.0 (i.e. raised pH), 1.2 M $LiSO_4$, and a large molar excess of α-1,6-mannobiose (Dextra Laboratories) was used. Native CpGH125 crystals were soaked in crystallization solution containing methyl-*S*-(α-D-mannopyranosyl)-(1-6)-α-D-mannopyranose (thiomannobiose) (a kind gift from Professor Geert-Jan Boons, CCRC, University of Georgia, USA) to obtain a complex representing bound substrate.

All crystals were cryoprotected with crystallization solution supplemented with 25% ethylene glycol (Hampton Research) and flash cooled directly in a nitrogen gas stream at 113 K. Data sets were collected either on a "home-beam" comprising a Rigaku R-AXIS 4++ area detector coupled to a MM-002 X-ray generator with Osmic "blue" optics and an Oxford Cryostream 700, BL9-2 at the Stanford synchrotron Radiation Laboratories, or CMCF1 at the Canadian Light Source as indicated in Table 4. All diffraction data were processed using MOSFLM/SCALA in the CCP4 suite of programs (Powell, 1999), (Collaborative Computational Project, Number 4, 1994). Data collection and processing statistics are shown in Table 4.

Structure Solution and Refinement. The structure of native SpGH125 was determined by molecular replacement using MOLREP (Vagin and Teplyakov, 2010) to find two molecules of SpGH125 in the asymmetric unit using the coordinates of CPE0426 from *C. perfringens* strain 13 (PDB ID 2nvp) as a search model. The initial model was corrected and completed manually by multiple rounds of building using COOT (Emsley and Cowtan, 2004) and refinement using REFMAC (Murshudov et al., 1997). The same approach was used to solve the structure of CpGH125 in complex with methyl-*S*-(α-D-mannopyranosyl)-(1-6)-α-D-mannopyranose. The completed model of native SpGH125 was used to solve the structures of SpGH125 in complex with dMNJ and mannobiose. These models were manually corrected and refined as above. Water molecules were added using COOT:FINDWATERS and manually inspected after refinement. In all data sets, 5% of the observations were flagged as 'free' and used to monitor refinement

procedures (Brünger, 1992). Model validation was performed with SFCHECK (Vaguine et al., 1999), PROCHECK (Laskowski et al., 1993) and MOLPROBITY (Chen et al., 2010). Structure and refinement statistics are shown in Table 4.

	*Sp*GH125 native	*Sp*GH125 + dMNJ	*Sp*GH125 + mannobiose	*Cp*GH125 native	*Cp*GH125 + thiomannobiose
Data Collection					
Beamline	SSRL 9.2	SSRL 9.2	Home-beam	Home-beam	CLS
Wavelength	0.97915	0.97946	1.54180	1.54180	0.9794
Space group	$P2_12_12_1$	$P2_1$	$P2_1$	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions: *a, b, c* (Å)	60.04, 101.61, 158.89	53.45, 158.78, 60.03	53.51, 159.04, 60.13	49.98, 96.11, 109.62	49.87, 96.59, 109.15
Resolution (Å)	48.39-2.15 (2.27-2.15)	20.00-1.75 (1.84-1.75)	20.0-2.10 (2.10-2.15)	20.00-2.35 (2.48-2.35)	40.00-2.05 (2.16-2.05)
R_{merge}	0.133 (0.356)	0.074 (0.176)	0.108 (0.328)	0.122(0.402)	0.091(0.324)
$I/\sigma I$	13.8 (6.1)	13.4 (5.6)	8.8 (3.4)	10.3 (3.2)	14.6 (5.7)
Completeness (%)	99.98 (100.00)	97.4 (83.4)	99.5 (100)	99.1 (99.6)	88.8 (78.9)
Redundancy	9.1 (9.1)	4.1 (3.3)	3.1 (3.1)	3.7 (3.8)	7.1 (6.7)
Refinement					
Resolution (Å)	2.15	1.75	2.10	2.35	2.05
No. of reflections	50972	88818	52816	21220	28513
R_{work}/R_{free}	0.16/0.22	0.17/0.22	0.16/0.22	0.16/0.22	0.16/0.20
No. of atoms					
Protein	3466(A); 3461(B)	3478(A); 3461(B)	3476(A); 3471(B)	3498	3492
Ligand*	N/A	11(ligA); 11(ligB)	23(ligA); 23(ligB)	N/A	23
Water	936	1388	993	401	403
B-factors					
Protein	11.5(A); 10.5(B)	9.8(A); 11.2(B)	12.4(A); 13.6(B)	17.4	20.6
Ligand	N/A	9.2(ligA); 12.9(ligB)	15.0(ligA); 19.9(ligB	N/A	21.6
Water	24.2	26.1	24.9	25.4	31.6
Root mean square deviations					
Bond lengths (Å)	0.015	0.015	0.015	0.013	0.016
Bond angles (degrees)	1.396	1.401	1.415	1.350	1.391
Ramachandran					
Preferred (%)	97.9	97.6	97.6	95.7	96.2
Allowed (%)	1.9	2.1	2.1	3.8	3.6
Disallowed (%)	0.2	0.2	0.2	0.5	0.2

Table 4: X-ray crystallographic data collection and structure refinement statistics for GH125s.

Values in parentheses are for the highest resolution bin. *Refers to carbohydrates and carbohydrate derivatives associated with the A and B protein chains in the asymmetric unit

Phylogeny. Searches for protein sequence similarity to *C. perfringens* ATCC 13124 CpGH125 were performed with BLASTp (Altschul et al., 1990) and the sequences with significant similarity (E-value cutoff of 10^{-5}) were extracted from the National Center for Biotechnology Information (NCBI) via blast explorer (www.phylogeny.fr) (Dereeper et al., 2010).

3.3 Results and Discussion.

3.3.1 GH125 from *S. pneumoniae* and *C. perfringens* are α-1,6-mannosidases.

Guided by both the genomic context of SpGH125 and its distant relationship to glucose processing enzymes to choose potential substrates, recombinant SpGH125 was screened against a variety of synthetic aryl-glycosides and unmodified disaccharides (Table 3). Of the aryl-glycosides tested SpGH125 displayed activity only on DNP-Man. Assessment of the kinetic parameters describing DNP-Man hydrolysis was complicated by an inability to reach substrate saturation making it possible to determine only the second order rate constant, k_{cat}/K_m (Higgins et al., 2009). The k_{cat}/K_m value was found to be 0.13 (± 0.01) $min^{-1}mM^{-1}$ for SpGH125 (Figure 18). To provide more general insight into the activity of this family of enzymes, the related protein CpGH125 was also assayed on DNP-man with the k_{cat}/K_m determined to be 0.58 (± 0.01) $min^{-1}mM^{-1}$ (Figure 18). Thus, both SpGH125 and CpGH125 exhibit relatively low activity on this substrate. The pattern of these enzymes having no observable activity on *para*-nitrophenyl α-D-mannopyranoside (pNP-Man) but measurable activity on DNP-Man is similar to the class I α-mannosidases, though this latter class of enzyme is specific for α-1,2-mannose linkages (Homer et al., 2001; Scaman et al., 1996). Unlike all of the currently characterized mannosidases, the activity of SpGH125 and CpGH125 displayed no sensitivity to the addition of EDTA (not shown), indicating a catalytic mechanism that does not rely on common metal ions used by other currently characterized mannosidases. Thus, SpGH125 and CpGH125 represent the founding members of a new family of glycoside hydrolases, family 125 that appears to possess properties distinct from other known α-mannosidases.

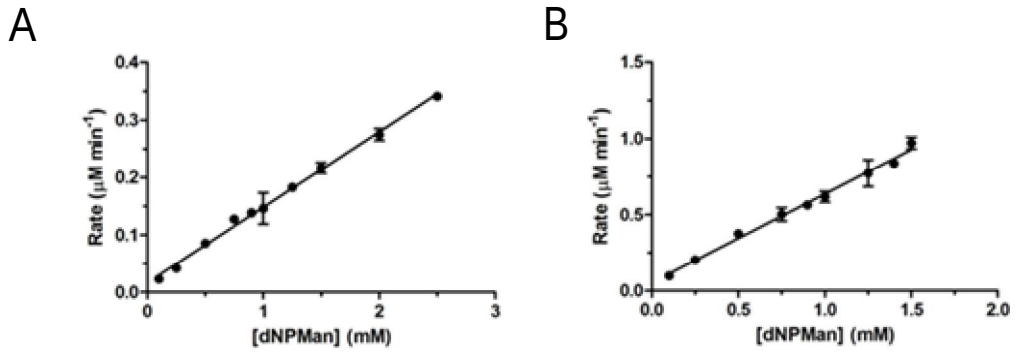


Figure 18: Kinetic plots of hydrolysis of 2,4-dinitrophenylate-α-1-mannoside.
A) SpGH125 and B) CpGH125

HPAEC-PAD analysis of carbohydrate hydrolysis revealed that SpGH125 could process α-1,6-mannobiose into the product mannose (Figure 19), CpGH125 was identical (data not shown). A thorough analysis using this method to detect activity on other α-mannosides showed that the enzyme had no detectable activity on α-1,2-mannobiose, α-1,3-mannobiose or α-(1,3)/(1,6)-mannotriose. To extend this analysis to additional family members, CpGH125 was also examined for activity on α-1,2-mannobiose, α-1,3-mannobiose, α-1,6-mannobiose and α-(1,3)/(1,6)-mannotriose with activity only being found on α-1,6-mannobiose (not shown). This analysis indicates the ability of the GH125 enzymes to act upon isolated fragments of N-linked glycans but leaves it unclear as to whether these enzymes are active on more complete N-glycans.

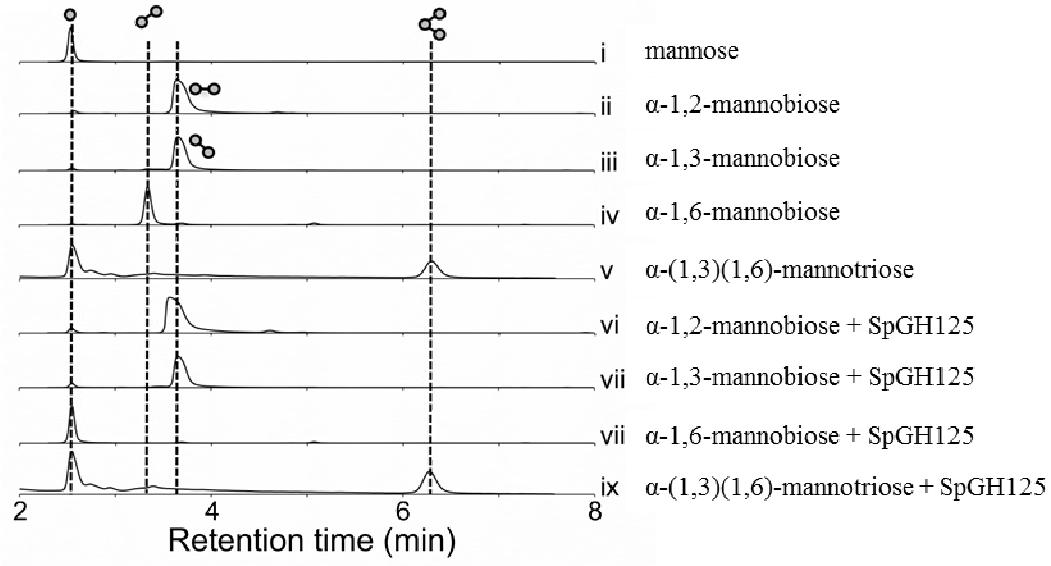


Figure 19: Analysis of GH125 specificity by HPAEC-PAD

HPAEC-PADS traces i-v show the elution profiles of mannose, a-1,2-mannobiose, a-1,3-mannobiose, a-1,6-mannobiose, and a-(1,3)(1,6)-mannotriose standards, respectively. Traces vi-ix show the elution profiles of α-1,2-mannobiose, a-1,3-mannobiose, a-1,6-mannobiose, and a-(1,3)(1,6)-mannotriose, respectively, treated with SpGH125.

To confirm the ability of SpGH125 and CpGH125 to act on carbohydrates representing more complete N-glycans and more firmly establish the specificity of the GH125 enzymes Capillary Electrophoresis was employed as a sensitive method to examine activity on high-mannose glycans (See Appendix A: Figure 41). Treatment of Man9, Man5 and Man3a with SpGH125 or CpGH125 revealed the resistance of these glycans to the activity of these enzymes. The use of a commercial α-(1,2)/(1,3)-mannosidase to first remove the terminal α-1,3-linked mannose residues from Man3a resulted in species having a CE mobility between that of Man1 and Man3a and was thus inferred to be Man2a; surprisingly this enzyme appeared to have a low level of α-1,6-mannosidase activity giving some of the Man1 species. After co-treatment of Man3a with the α-(1,2)/(1,3)-mannosidase, SpGH125 and CpGH125 were able to completely convert the glycan into Man1 indicating that the α-1,6-linkage can only be efficiently hydrolysed by these enzymes after removal of the α-1,3-linked mannose by the commercial enzyme.

Taken together these results clearly reveal that the GH125 enzymes are specific for α-1,6-mannosides; however, these enzymes are only able to process this linkage in the mannose core of N-glycans after the α-1,3-linked mannose is removed. The activity of these enzymes on DNP-Man and the tested glycans is suggestive of an *exo*-mode of action but this is not unequivocally demonstrated by these experiments. To provide greater insight into the molecular features that govern the activity of SpGH125 and CpGH125 the structures of these proteins were determined using X-ray crystallography.

3.3.2 The structural basis of α-1,6-mannoside recognition.

The structures of SpGH125 and CpGH125 were initially solved by molecular replacement in their uncomplexed forms to 2.15 Å and 2.35 Å, respectively. These structures reveal the expected $(\alpha/\alpha)_6$ fold previously observed for the homologous

hypothetical proteins from *C. perfringens* (PDB ID 2nvp), *B. thetaiotaomicron* (PDB ID 2p0v), and *B. ovatus* (PDB IDS 3p2c and 3on6) (Figure 20). In keeping with their high amino acid sequence identity of ~40%, SpGH125 and CpGH125 displayed a root mean square deviation (RMSD) of 1.2 Å.

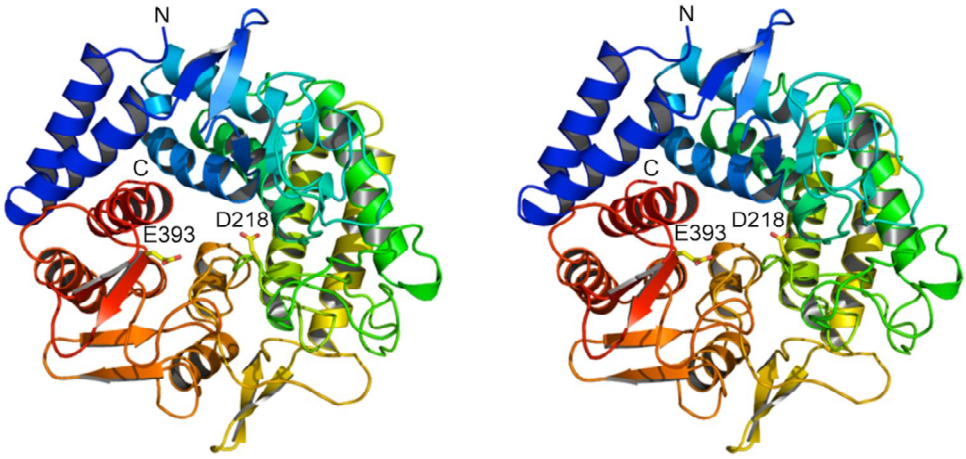


Figure 20: Structure of GH125.

Cartoon representation of SpGH125 showing its overall fold with N- and C-termini labelled and putative catalytic residues labelled and shown in yellow stick representation.

Insight into how SpGH125 and CpGH125 both recognize and hydrolyze mannooligosaccharides was gained through analysis of three structures in complex with different ligands: SpGH125 in complex with the inhibitor dMNJ (Legler and Jülich, 1984), a non-productive complex of SpGH125 with α-1,6-mannobiose, and a complex of CpGH125 with a non-hydrolyzable methyl-*S*-(α-D-mannopyranosyl)-(1-6)-α-D-mannopyranose (Zhong et al., 2008).

The structure of SpGH125 in complex with dMNJ was determined to 1.75 Å resolution and revealed clear electron density for the inhibitor (Figure 21A). All of the inhibitor's hydroxyl groups are engaged in hydrogen bonds with residues in the SpGH125 active site. Notably, the inhibitor hydroxyl groups that would be the equivalent of the O2, O3, and O4 hydroxyls on mannose interact with Arg62, Asp63, Asn302, and with the Pro216

main-chain carbonyl and are buried deep in the active site. This arrangement is consistent with occupation of the -1 subsite by a terminal mannose

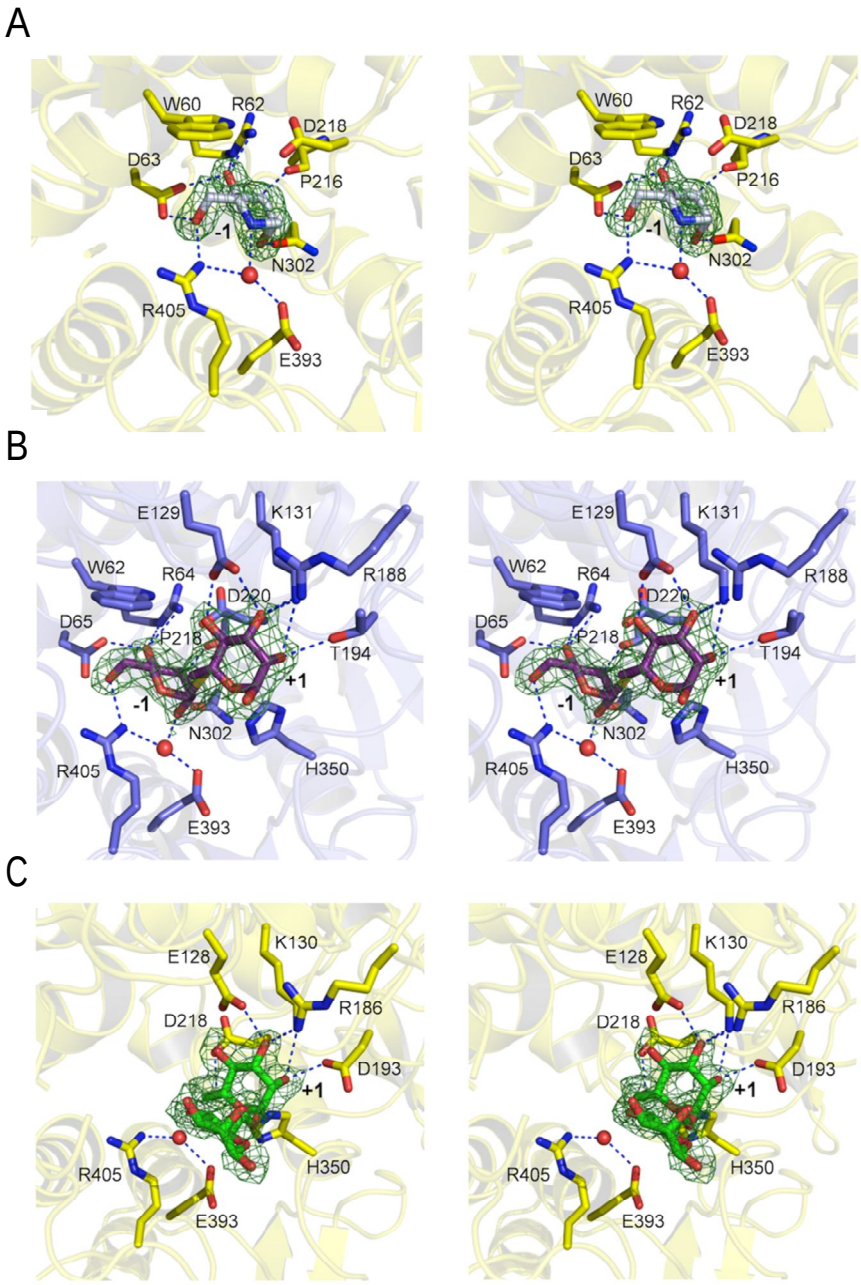


Figure 21: Carbohydrate recognition by GH125.

All panels shown in divergent stereo. A) Active site representations of inhibitor binding by SpGH125. The green mesh shows the maximum-likelihood /σ_a -weighted electron density maps contoured at 1σ (0.47 e/Å^3). Key active site residues are shown in stick representation

and coloured yellow and DMJ coloured grey. B) Active site representations of methyl-*S*-(a-D-mannopyranosyl)-(1-6)-a-D-mannopyranose binding by CpGH125. The electron density map is contoured at 1σ (0.32 e/Å^3)(40). Key active site residues are coloured blue and methyl-*S*-(a-D-mannopyranosyl)-(1-6)-a-D-mannopyranose coloured purple. C) Active site representations of α-(1,6)-mannobiose binding by SpGH125 at pH9. The electron density map is contoured at 1σ (0.40 e/Å^3). Key active site residues are coloured yellow and α-(1,6)-mannobiose coloured green. In all panels active site subsites are labelled according to the convention established by Davies *et al.* (Davies et al., 1997)**. Hydrogen bonds between the protein and compound are shown as dashed blue lines identified using the criteria of proper geometry and a distance cutoff of 3.2 Å.**

residue; it also precludes the recognition of mannose residues that are further modified on their non-reducing end and thus supports the proposal that the GH125s are *exo*-mannosidases.

Aglycon recognition appears to be a critical feature of the GH125 enzymes as they lack activity on the pNP-Man and have very poor activity on DNP-Man, despite these substrates possessing chemically activated leaving groups. Furthermore, these enzymes appear to show selectivity for terminal, unbranched α-1,6-dimannose motifs, again indicating a strict aglycon requirement. To better understand the aglycon features required by the GH125 enzymes, a complex was obtained of CpGH125 bound to methyl *S*-(α-D-mannopyranosyl)-(1-6)-α-D-mannopyranose. This 2.05 Å resolution structure revealed unambiguous electron density for the sugar residues occupying the -1 and +1 subsites of the CpGH125 active site but the methyl group (on O1) was too disordered to be modeled (Figure 21B). The non-reducing end of this disaccharide fits into the pocket of the -1 subsite and makes a range of interactions with the protein that are identical to those observed in the SpGH125-dMNJ complex. The reducing terminal mannose residue sits in the +1 subsite and engages the sidechains of Glu129, Lys131, Arg188, and Thr194 in an extensive hydrogen bond network; this subsite is also absolutely conserved with SpGH125. The -1 subsite is separated from the +1 subsite by ~6.5 Å and, like the -1 subsite, the +1 subsite almost fully encloses the mannose residue, leaving only room for additional substituents linked through reducing terminal O1 group. Thus the architecture

of the +1 subsite clearly excludes modification of this portion of the aglycon with an α-1,3-linked mannose branch. Furthermore, the separation of the subsites is roughly 1.5 Å larger than that is expected to separate subsites accommodating α-1,3-linked or α-1,2-linked mannobiose and thus confers specificity toward the more extended α-1,6-linkage that positions the two mannose residues further apart.

The structure of the GH125 active site is uniquely tailored to accept a terminal α-1,6-linked dimannose motif in its -1 and +1 subsites. However, the demonstrated activity of these enzymes on a larger glycan, i.e. Man2a, reveals the ability of the enzymes to tolerate additional reducing end sugars and the possibility of additional subsites on the aglycon side of the cleaved bond. Consideration of the CpGH125-thio-sugar complex shows the O1 of the mannose bound to the +1 subsite (Figure 21B) can likely accommodate a sugar residue such as the *N*-acetylglucosamine that is α-1,4-linked to the first mannose in the N-linked glycan core. A catalytically non-productive complex of SpGH125 with α-1,6-mannobiose, fortuitously trapped by raising the pH of the α-1,6-mannobiose containing solution used to soak the crystal, provides some additional insight into recognition of additional saccharide residues. In this 2.1 Å resolution structure the α-1,6-mannobiose was well-ordered with the non-reducing end mannose occupying the +1 subsite where the interactions were identical to those seen in the CpGH125 thio-sugar complex (Figure 21C). Though well-ordered and in the same conformation for both molecules in the asymmetric unit, the reducing end mannose residue made no direct interactions with the protein (Figure 21C). This mode of accommodating this disaccharide indicates that various structures terminating with Manα1,6Manβ1-OR (where R could be various structures) such as Manα1,6Manα1,6Manβ1,4GlcNAc2-Asn or Manα1,6Manα1,6(Manα1,3)Manβ1,4GlcNAc2-Asn. Thus, the *S. pneumoniae* and *C. perfringens* GH125 enzymes appear to lack a defined +2 subsite, which should provide some flexibility in accommodating various structures.

3.3.3 Comparison of GH125 structures.

Comparison of SpGH125 and CpGH125 structures with the homologs from *Bacteroides ovatus* and *Bacteroides thetaiotaomicron* gives RMSDs of approximately 1.5 Å for all of the comparisons, consistent with the ~40% amino acid sequence identity they all share (the example from *C. perfringens* strain 13, PDB ID 2nvp, was omitted due to its 100% sequence identity with CpGH125). All of these proteins have absolutely conserved -1 and +1 subsites indicating that the Bacteroides proteins are almost certainly also α-1,6-mannosidases (Figure 22).

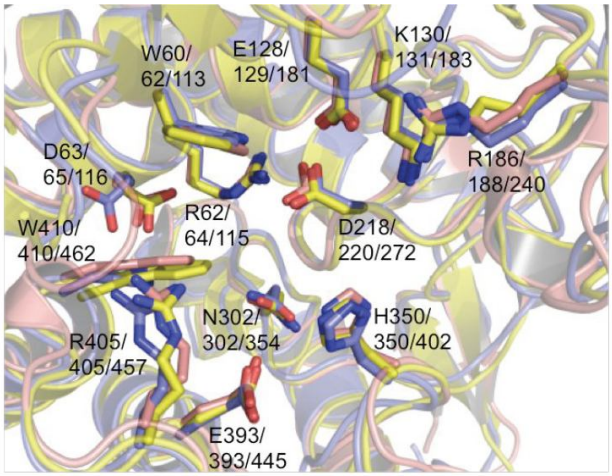


Figure 22: Comparison of GH125s.

A structural overlay of SpGH125 (yellow), CpGH125 (blue), and *Bacteroides ovatus* PDB Id 3p2c (salmon) active sites. Relevant side chains in the -1 and +1 subsites of the active site are shown in stick representation and labeled in order of SpGH125, CpGH125 and 3P2C, respectively.

3.3.4 α-glycoside hydrolysis on a conserved platform.

All of the GH125 enzymes have a high degree of structural similarity (RMSDs of ~2.7-3.0 Å) but low amino acid sequence identity (<14%) with members of glycoside hydrolase family 15. A comparison of the active site of CpGH125 with the GH family 15 *Arthrobacter globiformis* glucodextranase (G1d; PDB Id 1ulv) (Mizuno et al., 2004) shows not only the fold similarity but the conserved location of the active site (Figure 23 A). G1d and CpGH125 recognize α-configured glucose and mannose, respectively, in their -1 subsites; nevertheless, the -1 subsites of the two enzymes show remarkable conservation revealing the shared molecular features involved in the recognition of an α-configured pyranose ring having equatorial groups at C3, C4, and C5 (Figure 23B). The primary difference between the two different families of enzymes in this subsite is in regards to their recognition of O2, the orientation of which distinguishes glucose from mannose. In CpGH125 the axial O2 of mannose makes a hydrogen bond with N302. This interaction is not present in G1d but instead R567 of this protein makes a hydrogen bond with the equatorial O2 of glucose. In contrast, there is virtually no conservation between CpGH125 and G1d in the +1 subsites, which reflects the quite different aglycon requirements of these two families of enzymes (Figure 23C).

In addition to conservation in the -1 subsite, the known catalytic residues in GH family 15, as represented by G1d, are conserved in the GH125 enzymes, as is the position of a putative nucleophilic water molecule. In G1d, Glu430 acts as the catalytic acid while Glu628 acts as the catalytic base that activates a strategically positioned water molecule. In the CpGH125 structure, the α-1,6 S-linkage spans the -1 and +1 subsites where the carboxylate groups of Asp220 and Glu393 are structurally conserved with the GH15 catalytic residues, suggesting roles as catalytic acid and base, respectively, in the GH125 mechanism. The carboxylate group of Asp220 in CpGH125 is 3.0 Å from the sulfur of the S-linkage and appropriately oriented to donate a proton, consistent with a predicted role as the general acid (Figure 23D). In the -1 subsite Glu393 coordinates a water molecule that sits 3.2 Å directly beneath C1 of the mannose residue. This water molecule is perfectly positioned to attack the anomeric centre and displace the leaving group; identically positioned water is present in the SpGH125-dMNJ complex (Figure 23D).

Thus, elements of the GH125 active sites, including the catalytic machinery, are highly conserved within the active sites of GH15 enzymes, strongly suggesting that GH15 and GH125 enzymes have in common an inverting catalytic mechanism. Indeed, the shared catalytic features of the GH125 enzymes and GH15 indicate that family GH125 is a part of GH-L in the glycoside hydrolase clan classification (Cantarel et al., 2009).

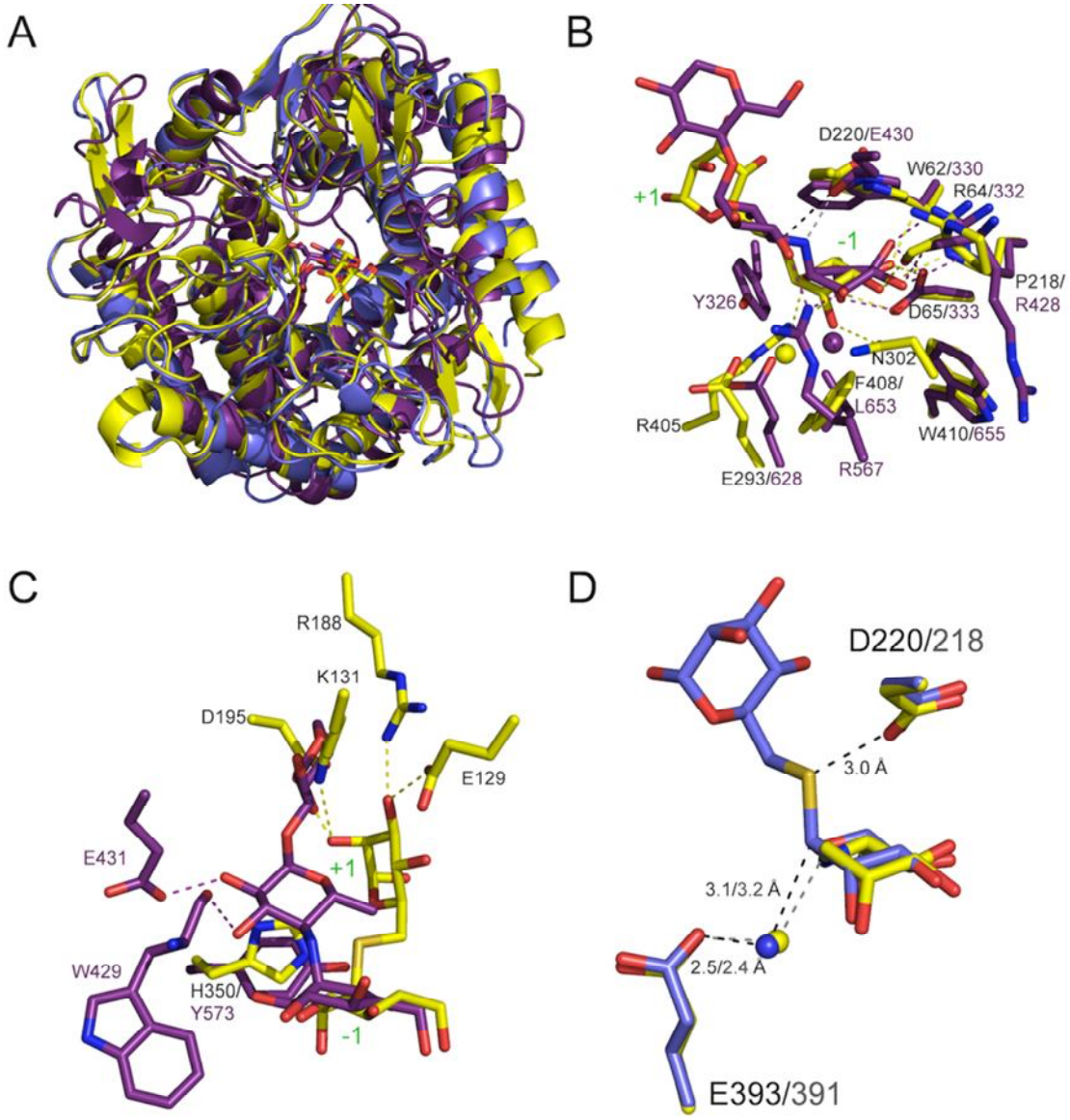


Figure 23: Similarities between GH family X and family 15.

A) A structural overlay of CpGH125 (yellow), SpGH125 (blue) and the GH15 *Arthrobacter globiformis* glucodextranase G1d (PDB ID 1ulv) (purple). Bound ligands are shown in stick representation. B) Overlay of the active sites of CpGH125 and G1d showing the -1 subsite and the +1 subsite (C). CpGH125 is coloured yellow (residue labels in black) and G1d

purple (residue labels in purple). Hydrogen bonds are shown as dashed lines. Putative catalytic waters are shown as spheres. Subsites are labelled in green. D) The relative orientations of the catalytic residues in CpGH125 (blue) and SpGH125 (yellow). Hydrogen bonds and distances are indicated. Labels and lines are color coded black and grey for CpGH125 and SpGH125, respectively. The putative catalytic waters, shown as spheres, do not hydrogen bond with C1 of the substrate – the dashed line is intended to indicate the putative trajectory of nucleophilic attack.

3.3.5 The GH125 enzymes use an inverting catalytic mechanism.

To provide clear support for the structure-based analysis of the GH125 catalytic mechanism, ^{1}H NMR spectroscopy was used to monitor SpGH125 and CpGH125 (See Appendix A: Figure 42 and 43) catalyzed cleavage of methyl 6-*O*-(α-D-mannopyranosyl)-*β*-D-mannopyranoside as a function of time. This analysis revealed that formation of the β-hemiacetal of mannose preceded the appearance of the α-hemiacetal. This confirmed a catalytic mechanism wherein the α-glycosidic linkage is hydrolysed via an inverting mechanism. In light of the observed architecture of the CpGH125 and SpGH125 active sites, and the similarity of this to the GH15 active sites, the general acid catalytic residue can be assigned as Asp218 in SpGH125 and D220 in CpGH125. Likewise, the general base residue can be assigned as Glu393 in both SpGH125 and CpGH125.

In none of the five SpGH125 and CpGH125 structures was a metal ion observed in the active site, nor was there any effect of EDTA on the activity of either enzyme. Thus, a notable feature of the GH125 catalytic mechanism is its lack of a participating metal ion, which distinguishes it from currently well-characterized α-mannosidases. Mannosidases in families GH38, GH47 and GH92, which are structurally and mechanistically distinct and hydrolyse a variety of α-mannosidic linkages, all require the coordination of a divalent metal ion in the catalytic site for activity. The divalent metal ion in the active site is thought to aid in distorting the substrate and stabilize the transition state by bridging the O2 and O3 hydroxyl groups of the mannoside bound in the -1 subsite (Suits et al., 2010; Zhu et al., 2010; Karaveg et al., 2005). The absence of a participating metal

ion in the GH125 mechanism reveals that divalent cation involvement is not a general requirement for efficient hydrolysis of α-mannosides.

3.3.6 Microbial N-glycan deconstruction.

The distribution of SpGH125 and CpGH125 homologues was tested by BLAST extraction of related sequences from the non-redundant protein database at NCBI. Three-hundred and sixty related proteins were identified with significant amino acid sequence similarity to SpGH125 and CpGH125 (E-value $<10^{-5}$). These sequences fall into a wide variety of prokaryotic and eukaryotic phyla.

In N-glycans, α-1,6-mannose branches always co-occur with α-1,3-mannose branches. GH125 enzymes, however, do not appear to tolerate the α-1,3-mannoside branches generating the requirement for pre-processing of these substrates with an α-1,3-mannosidase prior to GH125 action. This suggests the need for the evolutionary co-retention of an α-1,3-mannosidase with a GH125. Indeed, in the Firmicutes, except for *C. perfringens*, all of the bacteria possessing a gene encoding a GH125 also have a gene encoding a GH38 mannosidase immediately adjacent to the GH125-encoding gene (*C. perfringens* does have two GH38 encoding genes that are elsewhere in the genome). SpyGH38 from the firmicute *S. pyogenes* was demonstrated to be a specific exo-α-1,3-mannosidase that is tolerant of substrates having the α-1,6-branches found in N-linked glycans (Suits et al., 2010). The GH125-linked GH38s in the Firmicutes share no less than 42% overall amino acid sequence identity and have nearly identical active site residues with SpyGH38 suggesting that these enzymes have similar, if not identical, specificities. Therefore, in the N-glycan degrading Firmicutes, this implies a model whereby the α-1,3-branches in N-glycans are necessarily first processed by a GH38 followed by the activity of a GH125 to remove exposed and terminal α-1,6-mannose residues. This model is also supported by our studies of N-glycan processing using GH125 in combination with another mannosidase. This particular duo of enzymes is ideally suited to process the Man3GlcNAc2 core of N-glycans. The biological relevance of this remains largely uninvestigated; however, it is known that SpGH125 is a pneumococcal virulence factor, as is its accompanying putative GH38 (SP_2143), indicating that N-glycan degradation is important to the pneumococcal-host interaction (Hava and Camilli, 2002; Obert et al., 2006). The occurrence of these enzymes among

other notable pathogens, such as *Listeria monocytogenes, C. perfringens, S. pyogenes* and other streptococci, suggest that N-glycan depolymerisation is of relatively widespread importance to pathogens that infect a broad variety of human tissues.

The *Bacteroides* sp., with *Bacteroides thetaiotaomicron* being the primary model at present, are notable in that many are symbiotic inhabitants of the human gut. *B. thetaiotaomicron* is known to metabolise human glycans lining the gut under conditions of limited dietary polysaccharide intake by the host (Sonnenburg et al., 2005). The large expansion of α-mannosidases in this bacterium, as represented by the 23 family 92 α-mannosidases, is thought to contribute to this ability by enabling the metabolism of N-glycans (Zhu et al., 2010). These enzymes, however, lack substantial α-1,6-mannosidase activity, which would be required for complete N-glycan deconstruction. GH125s are relatively common among the *Bacteroides* sp., including three GH125s in *B. thetaiotaomicron,* suggesting these enzymes play a role in completing the machinery required to completely depolymerize N-glycans.

The function of GH125 enzymes in the fungi that contain them is less clear. These organisms can extensively post-translationally glycosylate their secreted proteins; both the N- and O-glycans can contain high-mannose content. Thus, the GH125 enzymes in these organisms may play a role in the maturation or recycling of fungal glycans. Alternatively, they may aid in foraging for mannose from the environment.

3.4 Conclusion.

A surprisingly small number of enzymes with specific α-1,6-mannosidase activity have been identified. A human core-specific lysosomal α-1,6-mannosidase that removes the α-1,6-mannose residue from Man3GlcNAc but not Man3GlcNAc2 indicating that it cannot be active on N-glycans still linked to proteins (Park et al., 2005). A mannosidase from *B. thetaiotamicron* (Bt3994) hydrolyzes the distal α-1,6-mannosidic linkage in high-mannose N-glycans irrespective of the presence of the α-1,3-mannose arm; however, this enzyme also shows some activity toward α-1,3-mannobiose (60). These examples highlight the unique specificity of GH125, which is made more remarkable as this specificity appears to be shared among the large number of family members.

Furthermore, the catalytic platform used by GH125 closely resembles that used by glucoamylases, possibly indicating a common ancestry; however, future mechanistic studies will be required to clarify the role of active site residues in the GH125 family. Nevertheless, the active site architecture clearly indicates that the GH125 enzymes utilize a catalytic mechanism that does not rely on the participation of a metal ion, making GH125 the only presently characterized metal-independent α-mannosidase family. Significantly, this reveals that metal-assisted catalysis is not a required feature of α-mannoside hydrolysis.

Chapter 4: *Clostridium perfringens* toxin complex formation through protein-protein interaction

Katie Gregg [†‡] , Jarrett J. Adams[*†] , Edward A. Bayer [§] , Alisdair B. Boraston [‡] and Steven P. Smith[¶‖]

[‡]Department of Biochemistry and Microbiology, University of Victoria, Victoria, BC, Canada V8W 3P6;
*Department of Molecular and Cellular Physiology, Stanford University, Stanford, CA 94305;
[§]Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel; and
[¶]Department of Biochemistry, Queen's University, Kingston, ON, Canada, K7L 3N6
[†]J.J.A. and K.G. contributed equally to the work.

Adapted with permission from: Proceedings of the National Academy of Sciences USA. 2008; 105(34): 12194–12199.

Contributions to Research: Protein production and purification, isothermal titration calorimetry, crystallization and solution, manuscript and figure preparation

4.1 Introduction.

Clostridium perfringens is a notable human and livestock pathogen that is a prolific producer of toxins. In addition to the production of its major toxins, *C. perfringens* devotes a huge portion of its genome to encoding carbohydrate-active enzymes, with a significant portion of these being glycoside hydrolases (Myers et al., 2006; Shimizu et al., 2002; Henrissat and Davies, 1997). The catalytic activities of many of these glycoside hydrolases are consistent with the breakdown of the mucosal layer of the human gut and other glycans that decorate the surface of epithelial cells. *C. perfringens* infection involves the combined action of this complex arsenal of glycoside hydrolases on glycans and glycocongugates, contributing to its virulence through tissue/glycan destruction to facilitate spread, potentiating the activity of the major toxins, and providing a source of nutrition for the bacterium.

The glycoside hydrolases are among the largest and most modular enzymes produced by *C. perfringens* (Figure 6). The majority of these enzymes are larger than 1000 amino

acids with several near or over 2000 amino acids. A dominating feature of these proteins is their extensive modularity in addition to the catalytic modules and the reoccurring appearance of various ancillary modules. The widespread presence of these ancillary modules, both in *C. perfringens* and in other pathogenic bacteria, suggests their importance. However, unlike the catalytic modules the ancillary module functions are generally not predictable with many having unknown functions. The recurring nature of some of these ancillary modules, including carbohydrate-binding modules (CBMs), suggests that these non-catalytic modules play roles that improve the efficacy of the catalytic module.

In this study, the μ-toxin, also called CpGH84A, was found to contain a dockerin-like sequence at its C-terminus, which comprises two conserved calcium-binding EF-hand loop sequences (Chitayat et al., 2007b)(Gilbert, 2007). Dockerin modules have been extensively characterized in cellulolytic bacteria, where they mediate the enzyme assembly and cell-surface attachment of the cellulose-degrading complex, termed the cellulosome, through interactions with its cognate binding partner, a cohesin module (Figure 5) (Gilbert, 2007; Adams et al., 2006; Bayer et al., 2004; Carvalho et al., 2007; Doi and Kosugi, 2004). A module of unknown structure or function called X82 from CpGH84C, a *C. perfringens* homolog of the μ-toxin with exo-β-D-*N*-acetylglucosaminidase activity (Ficko-Blean and Boraston, 2005), was extensively investigated in this study by amino acid sequence similarity comparisons, secondary structure predictions, and 3-D structure threading predictions indicating distant sequence and structural similarity to cohesin modules. The observation of both cohesin- and dockerin-like modules in *C. perfringens* glycoside hydrolases led to the hypothesis that these enzymes could form complexes through the cohesin·dockerin interaction.

The objective of this study is to determine if the cohesin- and dockerin-like modules identified in C. perfringens GHs can interact and to ascertain if this interaction could form complexes of GHs.

To address these objectives, a combination of biophysical and structural techniques will be used to assess the putative interaction between these modules and the relationship to the cellulolytic cohesin·dockerin interactions and the implications of cohesin·dockerin mediated toxin complexes in *C. perfringens* pathogenesis will be discussed.

4.2 Experimental Procedures.

Cloning, protein production and purification. The cloning of the C-terminal FIVAR (found-in-various-architectures)-dockerin modular pair fragment from the µ-toxin, CpGH84A, and the cohesin module of CpGH84C were performed as previously described (Chitayat et al., 2007a; Chitayat et al., 2007b), respectively. FIVAR-dockerin and CpGH84C cohesin methionine mutant constructs were generated, in which Leu-1533 and Val-1535 of the former and Leu-784 and Leu-856 of the latter were each mutated to methionine residues for the generation of a seleno-L-methionine (SeMet)-containing complex. The methionine mutations were engineered by using the QuikChange Site-Directed Mutagenesis Kit (Stratagene) and a modified PCR protocol as previously reported (Wang and Malcolm, 1999). Subcloning of the FIVAR-dockerin mutant construct made use of a single set of forward and reverse primers, whereas the generation of the cohesin methionine-encoding mutant required a unique set of primers for each mutation (Table 5). The respective methionine mutant-encoding plasmids were transformed into *Escherichia coli* XL1-Blue, isolated by using a QIAprep Spin Miniprep kit (Qiagen), and sequenced to assess mutagenesis.

Primer	Sequence
Wt µ-toxin FIVARdockerin-F	GGGAATTCCATATGGATAAGACAAATTTAGGTGAATTAATA
Wt µ-toxin FIVARdockerin-F	CCGCTCGAGATTTAGTATTCTATGATTTATAAAACT
L1533M/V1535M µ-toxin FIVARdockerin-F	GTGAATACCATAAAGGCGCTAAGGATGGAATGACAATGGAAATTAATAAGGCTGAAGAAG
L1533M/V1535M µ-toxin FIVARdockerin-R	CTTCTTCAGCCTTATTAATTTCCATTGTCATTCCATCCTTAGCGCCTTTATGGTATTCAC
L784M CpGH84C Cohesin-F	GAAGTGTATCTCTTGAAGCTATGGAGGAAGTTCAAGTTGGAG
L784M CpGH84C Cohesin-R	CTCCAACTTGAACTTCCTCCATAGCTTCAAGAGATACACTTC
L856M CpGH84C Cohesin-F	CTTTAACTGGAGAACCAATGCCAGCTAAAGAAGTTTTAG
L856M CpGH84C Cohesin-R	CTAAAACTTCTTTAGCTGGCATTGGTTCTCCAGTTAAAG
-F, forward; -R, reverse.	

Table 5: Primers used for cloning of recombinant modular protein constructs and mutagenesis for cohesin and dockerin constructs

Protein Production and Purification. Recombinant protein derivatives of the native FIVAR-dockerin from the μ-toxin and the cohesin module from CpGH84C were expressed and purified as described previously (Chitayat et al., 2007a; Chitayat et al., 2007b). Expression and purification of SeMet-labeled FIVAR-dockerin and cohesin was also performed in a manner similar to that previously described with the exception that the expression plasmids encoding the FIVAR-dockerin and cohesin methionine mutants were transformed into the auxotrophic *E. coli* strain DL41 (DE3) and the medium was supplemented with 50 mg SeMet (Cambridge Isotope Laboratories). Recombinant expression, refolding, and purification of the SeMet-FIVAR-dockerin and SeMet-CpGH84C cohesin mutant derivatives were performed as described for the native protein fragments. Formation of the FIVAR- dockerin·cohesin complex involved the combination of purified cohesin and FIVAR-dockerin at a molar ratio of 1.3:1 in 5 mM Hepes, pH 7.5; 50 mM NaCl; and 5 mM $CaCl_2$. The 1:1 FIVAR-dockerin·cohesin complex was purified from excess cohesin by application on a Hi-Load16/60 Superdex 75 size exclusion column (Amersham Pharmacia Biosciences) equilibrated with 5 mM Hepes, pH 7.5; 50 mM NaCl; and 5 mM
$CaCl_2$ and eluted in 2 ml fractions using the same buffer. Fractions containing the complex were identified by SDS PAGE, pooled, and concentrated by using a Millipore Amicon 10-kDa centrifugal device to a final concentration of 50 mg/ml and stored at 4°C.

Calorimetry. The FIVAR-dockerin and cohesin protein samples for calorimetry, prepared in 25 mM Tris HCl (pH 7.5), 50 mM NaCl, and 5 mM$CaCl_2$, were filtered and degassed at 21°C. Titration of FIVAR-dockerin (75–150 μM) into cohesin (10 μM) was performed by using a VP-ITC titration calorimeter from Microcal. Twenty-five to fifty injections of 5–10 μl were made, with 240 s of equilibration between injections. Titrations were done in triplicate at the following temperatures: 30°C, 35°C, and 37°C. The changes in enthalpy (ΔH) at these temperatures were determined manually from the total integrated heat generated in the experiment (corrected for heats of dilution), normalized to the amount of complex formed. Fitting a bimolecular binding model to the data gave ΔH

values that were in excellent agreement with the determined values (not shown). A change in heat capacity (ΔC_p) was determined to be -143.5 ± 11.10 cal·mol^{-1}·K^{-1} as calculated from the temperature dependence of ΔH. Using ΔC_p and assuming its temperature independence, we extrapolated ΔH to 88.9°C, the temperature at which the association constant was determined. Using the ΔC_p, ΔH (at 88.9°C), and K_a (at 88.9°C) (Appendix B) as references, we determined the association constant at 37°C with the integrated form of the van't Hoff equation (Liu and Sturtevant, 1995).

Crystallization, Data Collection and Refinement. Using the vapour diffusion hanging drop method at 21°C, we obtained crystals of the native FIVAR-dockerin·cohesin complex at 25 mg/ml in 20% (w/v) PEG 2000; 0.1 M sodium acetate, pH 4.5; and 0.2 M ammonium sulfate. Crystals of the SeMet-containing complex were obtained at 25 mg/ml in 19% (w/v) PEG 1500 and 0.1 M sodium acetate (pH 4.5).

Data for the native (wild-type, unlabeled) and SeMet-containing (mutant) FIVAR-dockerin·cohesin complex crystals were collected at the National Synchrotron Light Source (NSLS). The crystals were flash-cooled in a cryostream of N_2 gas at 100 K by using mother-liquor supplemented with 20% (vol/vol) glycerol (native) and 20% (vol/vol) PEG 400 (SeMet) as cryoprotectant. Processing of the FIVAR-dockerin·cohesin diffraction data was performed with HKL2000 (Otwinowski and Minor, 1997). Diffraction data for the selenium substituted FIVAR-dockerin·cohesin mutant were collected on beamline X6A at a wavelength optimized for f'' at the selenium edge. This structure was solved by SAD using SOLVE (Terwilliger, 2003) to determine selenium substructure (five atoms were found in the asymmetric unit) and RESOLVE for density modification and initial structure tracing. This initial model was used as input into ARP/wARP (Perrakis et al., 1999) with the native FIVAR-dockerin·cohesin diffraction data collected on beamline X8C. ARP/wARP was able to build a virtually complete model that was completed by manual building. All computing was done with CCP4 (Collaborative Computational Project, Number 4, 1994) unless otherwise stated. Manual graphical model building was performed with COOT (Emsley and Cowtan, 2004), and refinements were done with REFMAC (Murshudov et al., 1997). Water molecules were

added to all models by using the REFMAC implementation of ARP/wARP and were inspected visually before deposition. In all data sets, 5% of the observations were flagged as ‘‘free’’ and used to monitor refinement procedures (Brünger, 1992). All final model statistics are given in Table 6. All structural representations were generated in PyMOL (DeLano, 2002).

Statistic	Native Coh–Fivar-Doc	SeMet Coh–Fivar-Doc
Data collection		
Space group	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions		
a, *b*, *c*, Å	35.49, 74.59, 94.79	35.13, 74.63, 93.75
Wavelength, Å	1.1	0.9789
Resolution, Å	50.0–1.6 (1.66–1.60)	50.0–1.8 (1.86–1.80)
R_{sym}, %	6.2 (25.1)	8.4 (43.0)
$I/\sigma I$	38.3 (3.4)	47.3 (2.2)
Completeness, %	98.9 (93.2)	97.6 (83.0)
Redundancy	5.4 (3.0)	11.7 (3.6)
Refinement		
Resolution, Å	13.95–1.60	
No. reflections	31,904	
R_{work}/R_{free}, %	20.5/24.6	
No. of atoms		
Protein	1,997	
Ions	3	
Water	315	
B-factors, Å^2		
Protein	20.33	
Ions	25.85	
Water	33.90	
Root mean square deviations		
Bond lengths, Å	0.010	
Bond angles, °	1.254	

Values in parentheses indicate the statistics for the highest resolution shell.

Table 6: X-ray crystallographic data collection and structure refinement statistics for cohesin·dockerin-FIVAR

4.3 Results and Discussion.

The presence of ancillary modules in *C. perfringens* glycoside hydrolases with likeness to dockerin and cohesin modules suggests that they contribute important functionalities to the overall efficacy of these enzymes so long as they have similar binding properties to cellulolytic cohesins and dockerins. Dockerin modules are best known for their role in the cellulosome, a multienzyme complex produced by anaerobic bacteria to efficiently and synergistically degrade cellulose-based biomass. These calcium-binding modules

interact with their cognate cohesin modules, allowing the integration of their covalently attached catalytic domains onto the cellulosome scaffold and cell-surface attachment (Gilbert, 2007).

4.3.1 Interaction and identification of *C. perfringens* cohesin and dockerin modules.

Cohesin- and dockerin-like modules were identified in *C. perfringens* glycoside hydrolases, CpGH84C and the µ-toxin, respectively. To ascertain if other putative cohesin and dockerin modules are present in other *C. perfringens* glycoside hydrolases, bioinformatics analyses were conducted. Searches of the annotated genomic sequences of the myonecrotic *C. perfringens* strains resulted in identification of four other putative cohesin modules and three other putative dockerin modules (Figure 6 and 24). Single copies of cohesin-like sequences were detected in these enzymes with the exception of CpGH31 which contains both a cohesin and a dockerin module, similar to that seen in several other human gut microorganisms. These observations suggest that in most cases this bacterium only has the capacity to form enzyme pairs, but with CpGH31 is able to form more elaborate carbohydrate-active enzyme complexes.

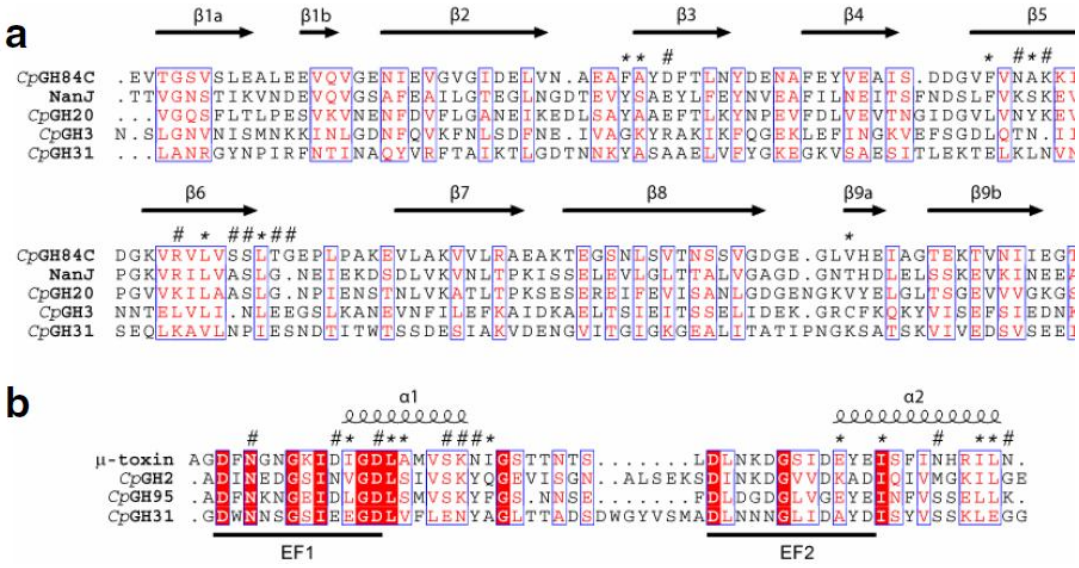


Figure 24: Structure-based sequence alignments *C. perfringens* cohesin and dockerin modules.

(*a*) Cohesin sequences from family 3 (CpGH3), 20 (CpGH20), 31 (CpGH31), 84 (CpGH84), and family 33 (NanJ) glycoside hydrolases. (b) Dockerin sequences from the family 84 (µ-

toxin), family 2 (CpGH2), 31 (CpGH31), and family 95 (CpGH95) glycoside hydrolases. Strictly conserved residues are shown as white letters on red background. Well conserved residues within a group, are displayed in red lettering, and residues displaying similarities across groups are boxed (blue). Secondary structures of the CpGH84C cohesin and μ-toxin dockerin modules, and residue positions involved in the cohesin·dockerin interface hydrogen bonding (#) and van der Waals contacts (*) are identified above the sequences.

A qualitative, ELISA-based assay was used to assess the binding ability and specificity of the five *C. perfringens* putative cohesin modules and four putative dockerin modules (See Appendix B) (Adams et al., 2008). Interactions involving three of the four putative dockerin modules, from CpGH84A (μ-toxin), CpGH2, and CpGH95 and three of the five putative cohesin modules, from NanJ, CpGH31, CpGH84C, and CpGH20 were identified. There was no binding ability identified for the cohesin modules from the family 3 and family 31 glycoside hydrolases (CpGH3 and CpGH31, respectively) and the dockerin module from a family 31 glycoside hydrolase (CpGH31). These results confirm the presence of cohesin and dockerin modules in *C. perfringens* and show that several of them have binding capabilities. It is possible that the modules for which no binding was observed were falsely hypothesized to be cohesin or dockerin modules.

The observed interaction of the *C. perfringens* cohesin and dockerin modules, and by extension their associated enzymes, could complex to form non-covalent enzyme complexes, in a manner similar to that seen in cellulolytic anaerobic bacteria (Gilbert, 2007; Bayer et al., 2004; Doi and Kosugi, 2004). These hypothetical enzyme complexes would allow the individual enzymes to function in concert to efficiently and rapidly degrade glycan substrates, which include the mucosal layer of the human gut, carbohydrates that comprise the extracellular matrix, and glycans of the surface of host cells.

To ascertain if the binding ability of the cohesin and dockerin modules was biologically relevant we sought to quantitatively analyze the interaction of the representative μ-toxin dockerin and CpGH84C cohesin modules. The isolated CpGH84C cohesin and a tandem derivative comprising the third FIVAR module and the dockerin module (FIVAR-

dockerin) from CpGH84A were produced and tested using isothermal titration calorimetry (ITC) and differential scanning calorimetry (DSC) studies (Figure 25 and Appendix B: Figure 44, respectively). A 1:1 binding stoichiometry was found, however the interaction was too strong to accurately determine an association constant (K_a) by ITC ($>10^9$ M^{-1}). The changes in enthalpy (ΔH) for the interaction determined by ITC at three different temperatures could be accurately established allowing the determination of the change in heat capacity (ΔC_p) to be assessed at 143.5 ± 11.1 cal·mol^{-1}·K^{-1}. Using the experimentally determined ΔC_p with the integrated form of the van't Hoff equation and extrapolating to physiological temperature, the K_a value for this interaction was determined by DSC to be 2.05 X 10^9 M^{-1} at 88.9°C or 1.44 X 10^{11} M^{-1} at 37°C.

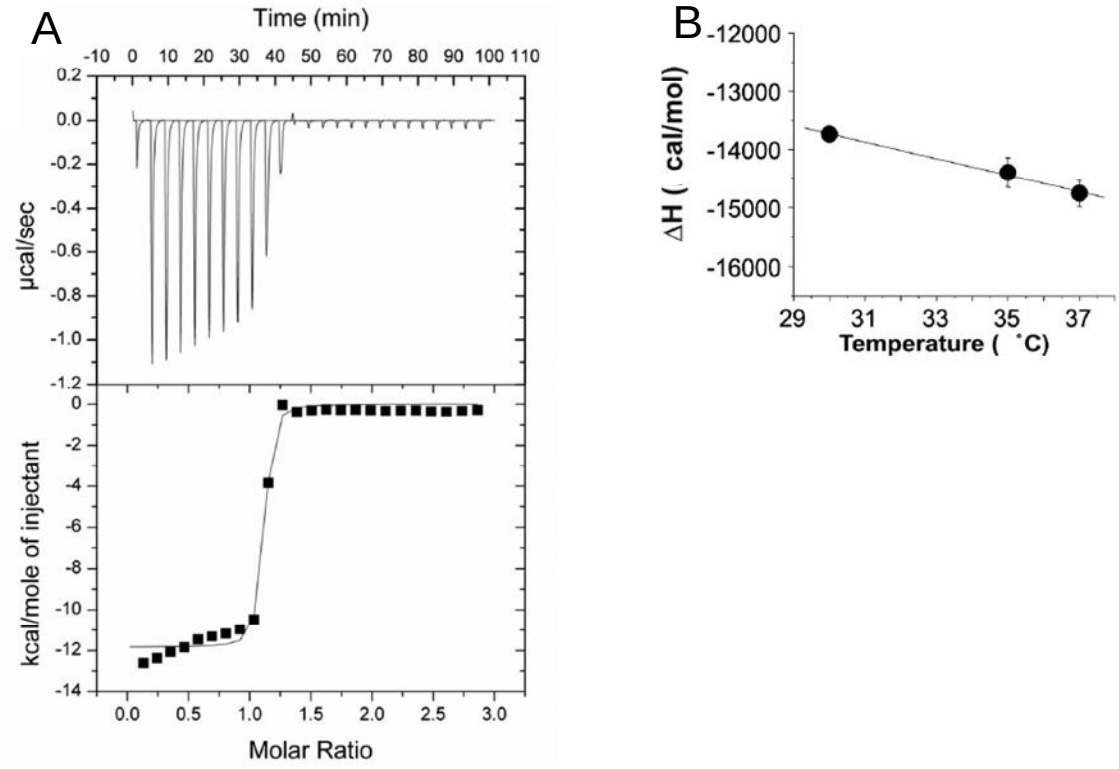


Figure 25: The ultrahigh affinity of the *C. perfringens Cp*GH84C cohesin and µ-toxin FIVAR-dockerin interaction.

A) Isothermal titration calorimetric analysis of the cohesin and FIVAR-dockerin interaction at 30°C. (Upper) Raw heat measurements. (Lower) Integrated heats after correction for heats of dilution as determined from the heats of ligand additions at the excess of saturation. The curves represent the best fit to a single-site model. B) Temperature

dependence of ΔH for the cohesin and FIVAR-dockerin interaction. ΔH values were determined at 30°C, 35°C, and 37°C.

4.3.2 Structural insights into *C. perfringens* cohesin·dockerin interaction.

The structural basis of the ultra-high affinity cohesin·dockerin interaction was examined using X-ray crystallography to determine the molecular basis of the interaction. A 1.6 Å resolution structure of the CpGH84C cohesin non-covalently complexed 1:1 with the μ-toxin, CpGH84A, FIVAR-dockerin double module was obtained. The cohesin module, in complex with the FIVAR-dockerin, has an expected β-sandwich structure comprised of nine strands in a jellyroll topology (Figure 26A, green). The two faces of the sandwich are comprised of β-strands 9a, 8, 3, 6, 5 (9a-8-3-6-5 face) and 9b, 1, 2, 7, 4 (9b-1-2-7-4 face) (See Figure 24 for numbering of β-strands) (Chitayat et al., 2008). An isolated CpGH84C cohesin structure (Chitayat et al., 2008) was compared with the cohesin from the complex structure revealing a backbone root mean square distance (RMSD) of 1.3 Å, illustrating that, similar to the cellulolytic cohesin modules (Adams et al., 2006; Carvalho et al., 2003), the *C. perfringens* cohesin module underwent only a very slight conformational rearrangement upon interacting with its cognate dockerin binding partner. The third FIVAR module (residues 1498–1577; Figure 26A, orange) from the μ-toxin of the FIVAR-dockerin double module protein fragment consists of three helices arranged in a left-handed three-helix bundle but no function can be inferred from the structure. The C-terminal dockerin from the μ-toxin has a fold typical for cellulolytic dockerins (residues 1578–1628; Figure 26A, light blue) with two antiparallel F-hand motifs separated by a nine amino acid linker. The two F-hand motifs each have an EF-hand calcium-binding loop that coordinates calcium in a pentagonal bipyramid configuration (Figure 26B). The two helices are aligned in such a way that the cohesin binding surface is an exposed and flush face.

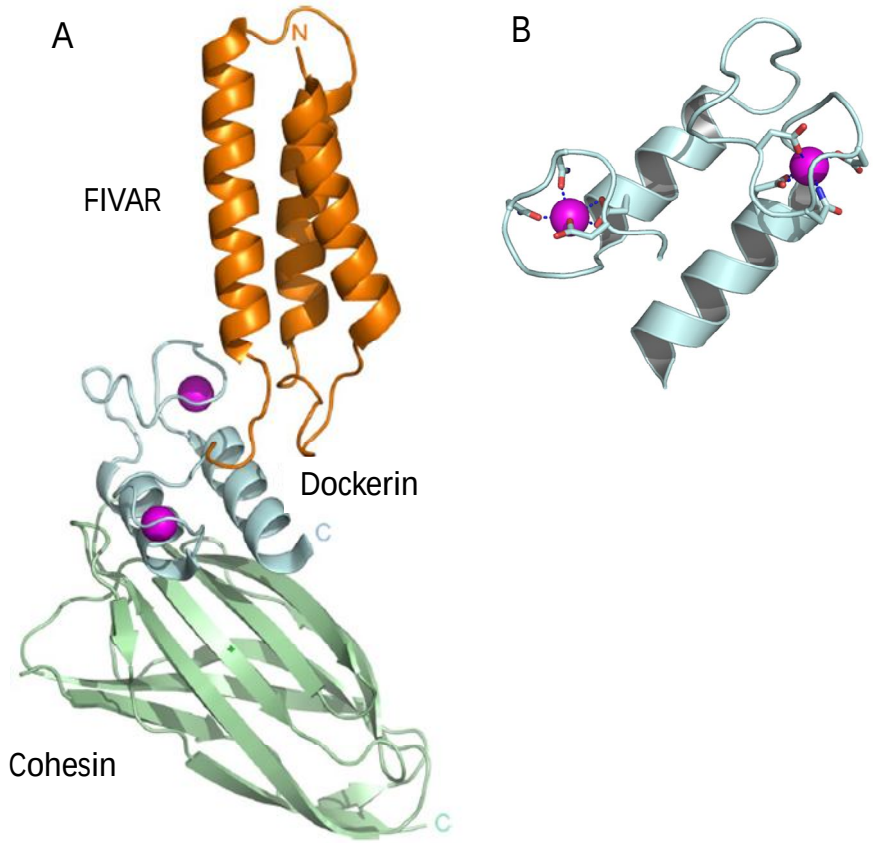


Figure 26: Structure of *C. perfringens* cohesin and dockerin.

A) Ribbon representation of the *Cp*GH84C cohesin·µ-toxin FIVAR-dockerin complex with the cohesin depicted in green, the FIVAR shown in orange, and the dockerin module in light blue. The N- and C-termini are coloured labeled accordingly. The calcium ions are depicted as purple spheres. B) Dockerin calcium coordination with the µ-toxin dockerin module shown in light blue. The calcium ions are depicted as purple spheres and the coordinating residues are shown in stick representation and the coordination indicated by blue dashes.

4.3.3 *C. perfringens* cohesin·dockerin complex binding interface.

The cohesin·dockerin binding interface is formed by both helices of the dockerin module and the β-strands 9a-8-3-6-5 of the cohesin module. Extensive hydrogen bonds and non-polar interactions between the non-covalently associated cohesin and dockerin residues accounts for the ultrahigh affinity of these modules for one another (Figure 27 and Table 7). The binding interface residues identified in this cohesin·dockerin interacting pair are

conserved among the three *C. perfringens* cohesin modules that were demonstrated to be functional, CpGH84C, NanJ, CpGH20 but poorly conserved in the CpGH3 andCpGH31 cohesin modules (Figure 24 and Appendix B: Figure 45). This is a likely explanation for the lack of observed interaction with the dockerin modules tested. Likewise, the dockerin residues that demonstrated cohesin binding capabilities are highly conserved among the functional dockerin modules, but not in the inactive CpGH31 dockerin module (Figure 24).

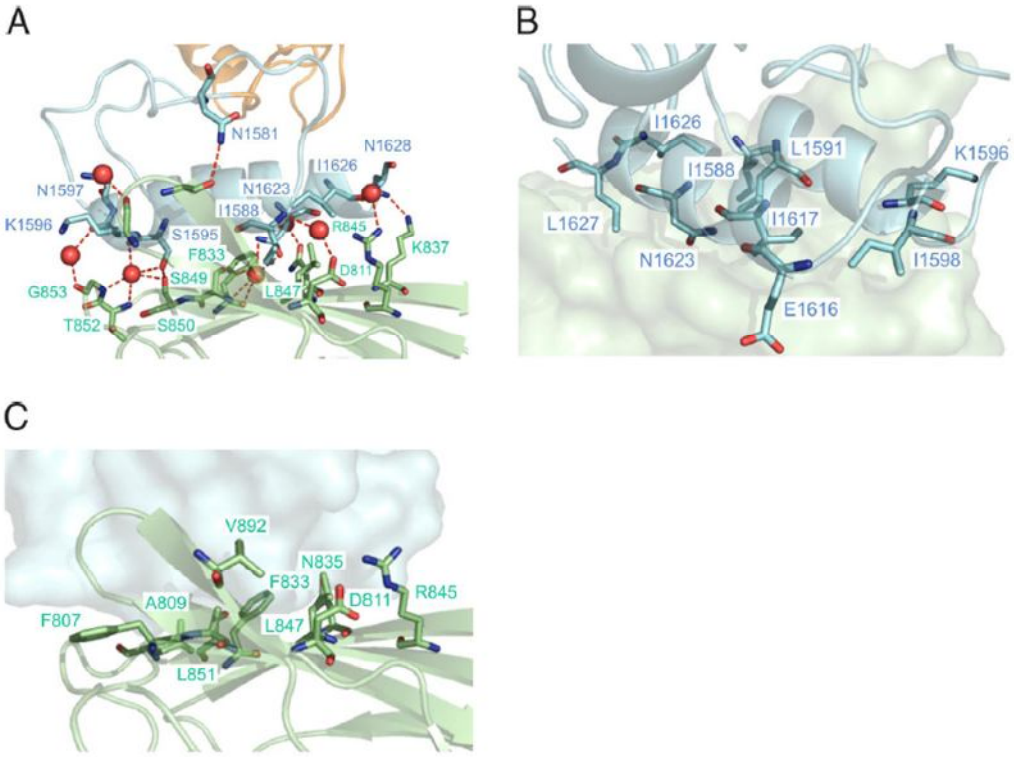


Figure 27: *C. perfringens* cohesin·dockerin intermolecular contacts.

A) CpGH84C cohesin and μ-toxin dockerin interface hydrogen-bonding network, with hydrogen-bond contacts shown as red dashed lines and water molecules as red spheres. B) Ribbon representation of dockerin, displaying residues involved in intermolecular van der Waals contacts as stick modules on the cohesin surface. C) Ribbon representation of cohesin, displaying residues involved in intermolecular Van der Waals contacts as stick modules on the dockerin surface. Cohesin and dockerin are coloured green and light blue, respectively. Residues are labeled in one-letter code, and numbered and coloured accordingly.

Coh residue	Residue no.	Atom	Water molecule	Doc residue	Residue no.	Atom
Ser	849	Oγ	4	Asn	1623	Nδ2
Phe	833	O	4	Asn	1623	Oδ2
Leu	847	O	4			
Ser	849	N	4			
Asp	811	Oδ2	15	Ile	1588	N
Arg	845	NH1	34	Ile	1626	O
Asn	835	Oδ1	177	Ser	1620	O
Gly	888	O	154	Asn	1597	Oδ1
Gly	853	O	179	Lys	1596	O
Ser	850	O	70	Ser	1595	Oγ
Thr	852	N	70	Ser	1595	O
Gly	853	N	70			

Table 7: FIVAR-dockerin·cohesin interface water coordination

Cellulolytic dockerin modules have two distinct roles in the classic cellulosome that both involve interaction with cohesin modules. The classical bacterial cellulosome contains a scaffoldin which is a large non-catalytic protein that functions as an anchor for the various enzymes and other cellulosomal components to tether to form a single functional entity (Figure 5). The scaffoldin typically consists of a cellulose-binding module that targets the cellulosome to the substrate and various cohesin modules that are usually arranged in tandem repeats. The individual cellulolytic enzymes (GHs) have dockerin modules in addition to their catalytic modules and are tethered to the scaffoldin by their dockerin modules interacting with the scaffoldin cohesin modules and is termed a type-I interaction. A type-II cohesin·dockerin interaction tethers the cellulosome to the proteoglycan layer of the bacterial cell surface usually via an association with a cell-anchoring protein (Peer et al., 2009). Considering that no multiple cohesin-bearing scaffoldins were identified in the *C. perfringens* glycoside hydrolases, it was not immediately evident which type of interaction the *C. perfringens* cohesins and dockerins would likely be most similar to. The CpGH84C cohesin and the μ-toxin dockerin interacting pair are similar to the *Clostridium thermocellum* type-II interaction, especially in terms of its non-polar character and both interactions involving direct contacts with the 9a-8-3-6-5 face of the cohesin (Adams et al., 2006). However, the *C. perfringens* dockerin module is oriented on the cohesin module at an angle of 180° when compared to the dockerin position in the *C. thermocellum* native type-I (Carvalho et al., 2003) and type-II complexes(Figure 28) (Adams et al., 2006). This orientation is similar to that of a *C. thermocellum* type-I dockerin mutant, which was generated to illustrate how the

internal sequential symmetry of the type-I dockerin would allow for plasticity in cohesin binding (Carvalho et al., 2007). No such internal symmetry exists for the µ-toxin dockerin, or the other *C. perfringens* dockerin modules, so it is not expected that they have a dual mode of cohesin recognition.

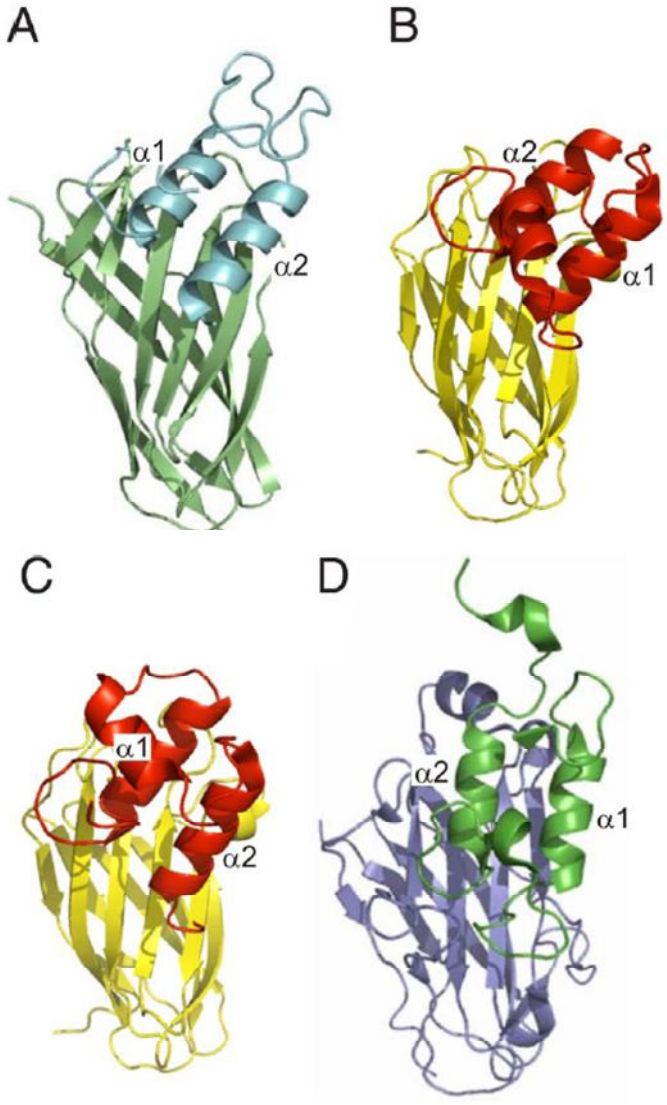


Figure 28: Variation of dockerin orientations in clostridial complexes.

Ribbon representations of (A) *C. perfringens* µ-toxin dockerin (light blue) on CpGH84C cohesin (light green); (B) native *C. thermocellum* type-I dockerin (red) on type-I cohesin (yellow) (16); (C) *C. thermocellum* Ser45Ala/Thr46Ala type-I dockerin mutant (red) on type-I cohesin (yellow) ; (D) *C. thermocellum* type-II dockerin(emerald green) on type-II

cohesin (slate blue). The helices within the two F-hand motifs of the dockerin modules are identified as 1 and 2, respectively.

4.3.4 Widespread distribution of noncellulosomal cohesin and dockerin modules.

The first identification of cohesin and dockerin modules in a noncellulolytic microorganism was in *Archaeoglobus fulgidus* in 1999 by Bayer *et al*. (Bayer et al., 1999). At the time it was unclear what the purpose of these modules was in this archaeon given that they were thought to be cellulosome signature sequences and this archaeon does not produce cellulases. The two cohesin and one dockerin modules were found to have binding activity in 2008 confirming their identification as cohesin and dockerin modules (Haimovitz et al., 2008). The next identification of noncellulolytic cohesin and dockerin modules was in *C. perfringens,* presented in this study, and was the first identification of cohesin and dockerin modules in pathogenic bacteria. Since the publication of this information (Adams et al., 2008), there have been extensive bioinformatic analyses that have revealed putative noncellulosomal cohesin and dockerin modules in a broad range of prokaryotes and to a lesser extent in eukaryotes. The wide distribution of these modules in archaea, bacteria and primitive eukaryotic species is thoroughly described by Peer *et al.* (Peer et al., 2009). Of the known archaeal and bacterial genomes, approximately 40% and 14%, respectively, contain either a putative cohesin- or/and a dockerin-like module and the majority of these genomes only encode for one or a few of these modules (Peer et al., 2009). The co-occurrence of both a cohesin and a dockerin module in many of these species appears to be uncommon, however, it is not known if this is due to difficulty in identifying the modules due to divergence of the sequences. There are many possible explanations for the lack of co-occurrence of putative cohesin and dockerin modules in noncellulolytic species. As mentioned, difficulty in identifying the sequences of the cognate binding partner or other types of binding partners that hereto have not been identified could account for this discrepancy. Another explanation is that these modules produced by organisms that occupy the same niche could participate in a synergistic relationship and possess interspecies cooperation or association. Also possible is that some of these putative cohesin or dockerin modules do not participate in any binding interaction at all, as was

demonstrated by some of the putative cohesin and dockerin modules identified in *C. perfringens* in this study.

4.4 Conclusion.

The current model of toxin delivery by *C. perfringens* suggests a process involving the dissemination of individual soluble enzymes. Given the highly heterogeneous nature of densely packed glycan moieties on the host cell surface glycocalyx and in gastrointestinal mucins, *C. perfringens* must require not only substantial production and secretion of carbohydrate-active enzymes, but also their coordinated effort to degrade these complex carbohydrates. In this context, the identification of cohesin and dockerin modules in *C. perfringens* glycoside hydrolases is very fitting with the necessity for efficient, concerted efforts to degrade the host glycans. The highly modular nature of the *C. perfringens* glycoside hydrolases and the recurring nature of various protein modules suggest important functional roles for these ancillary modules. These enzymes could act individually or simultaneously and in-concert, via the ultra-high-affinity cohesin·dockerin interaction, to impart their pathogenic effects. This synergistic action would allow for more efficient breakdown of complex heterogeneous glycans which would also be facilitated by the carbohydrate-binding modules (CBMs) found very commonly in glycoside hydrolases. These CBMs with varying specificities could tether the enzymes or cohesin·dockerin-mediated enzyme complexes to the glycan surface, a scenario that could also enhance substrate binding through an avidity effect. Our biochemical and structural characterization of *C. perfringens* cohesin·dockerin interactions represents an example of noncellulosomal cohesin·dockerin-mediated enzyme complexes, and provides a rationale for the expeditious rate of tissue damage associated with *C. perfringens* infections.

Chapter 5: Complete structural analysis of a *Clostridium perfringens* sialidase, NanJ

Contributions to Research: All experiments and data processing, writing and figure preparation.

5.1 Introduction.

Sialic acids are a group of N- or O-substituted derivatives of neuraminic acid, which are heavily modified, nine-carbon monosaccharides. The term sialic acid generally refers to the most common member of this group, *N*-acetylneuraminic acid. Sialic acids are often found capping branches of N- and O-glycans in animal tissues, plants, fungi, yeasts and bacteria. They are most commonly found in glycoproteins and gangliosides, and are sometimes found to cap side chains of GPI anchors (Vimr et al., 2004).

The enzymes responsible for removing these terminal sialic acid residues from a variety of glycocongugates are termed sialidases (or neuraminidases) and operate through a hydrolysis mechanism whereby the configuration of the anomeric carbon is typically retained. In bacteria, sialidases are important for nutrition, cellular interactions and have been implicated in a wide range of diseases (Varki, 1993). An example of this is the *Vibrio cholera* sialidase which enhances the activity of the cholera enterotoxin by acting synergistically with the toxin and increasing the binding and penetration of it to enterocytes (Galen et al., 1992). Other examples include the *Pseudomonas aeruginosa* sialidase which is involved in infection by contributing to the formation of biofilms while the *Tannerella forsythia* sialidase is involved in epithelial cell attachment (Soong et al., 2006). The action of two sialidases from *S. pneumoniae* have been shown to be virulence factors in animal models (Orihuela et al., 2004; Manco et al., 2006; Tong et al., 2000). NanA, the major pneumococcal sialidase, is very important for resistance to opsonophagocytic killing in assays using human neutrophils. NanA and two other surface-associated exoglycosidases are responsible for reducing deposition of complement component C3 on the pneumococcal surface, providing a mechanism for this resistance (Dalia et al., 2010). NanA is also involved in biofilm formation (Parker et al.,

2009). From these cases it is apparent that sialidases and their action of removing terminal sialic acid residues are often involved in bacterial pathogenesis.

The sialidases produced by the gangrene and necrotic-enteritis causing *Clostridium perfringens* have also been implicated in virulence. *C. perfringens* sialidases have been shown to potentiate the activity of the major toxin produced by this bacterium, the α-toxin. The α-toxin is a membrane-damaging sphingomyelinase that hydrolyzes phospholipids and is essential for the growth and spread of *C. perfringens* infection in the host (Petit et al., 1999). The action of the α-toxin is reliant upon its ability to interact with host membranes. The *C. perfringens* sialidases remove terminal sialic acid residues from cellular gangliosides, significantly increasing the sensitivity of target cells to the cytotoxic effects of the α-toxin (Flores-Díaz et al., 2005). While the sialidases potentiate the activity of the α-toxin, they are not essential for full virulence of the bacterium in a mouse myonecrosis model (Chiarezza et al., 2009).

C. perfringens strains encode combinations of up to three sialidase genes, *nanH*, *nanI*, and *nanJ*. The *nanH* gene product is a cytoplasmic, single module 43 kDa sialidase that is thought to be involved in bacterial nutrition (Kruse et al., 1996) and the *nanI* gene product is a 77 kDa secreted enzyme (Traving et al., 1994). NanI consists of a family 40 carbohydrate-binding module (CBM) and a catalytic module that belongs to family 33 glycoside hydrolases (Figure 6). The structure of the NanI catalytic domain was recently described and revealed a classic six bladed β-propeller in addition to a small β-barrel domain (Newstead et al., 2008). Several bacterial sialidases are found to be multi-modular containing carbohydrate-binding modules in addition to their catalytic modules, which have been suggested to increase the efficiency of catalysis by targeting the enzyme to glycan environments that contain sialic acids and other saccharides (Thobhani et al., 2003). NanJ is the largest of the sialidases with a molecular weight of 129 kDa and is the most modular amongst the *C. perfringens* sialidases, containing a total of six modules; from the N-terminus, a CBM from family 32, a family 40 CBM, a catalytic module from family 33 glycoside hydrolase, a module of unknown function, a cohesin module and a module with high amino acid identity to a fibronectin type III domain (FN3) (Figure 6).

The two N-terminal CBMs, CBM32 and CBM40, have been thoroughly studied and their individual structures determined (Boraston et al., 2007). They are different from one another in that they are classified into two distinct families, CBM32 and CBM40 and have differing binding specificities. The CBM32 has a binding preference for non-reducing terminal galacto-configured sugars and the CBM40 has a preference for terminal sialic acid. These specificities, which are consistent with the specificities characteristic of their families, are not the same as one another; the reason for the CBM32 having specificity for galactose is not entirely understood (Boraston et al., 2007). There is no structural information available for the remaining NanJ modules - the catalytic module, the unknown module, the cohesin module and the FN3 module.

The involvement of sialidases in bacterial pathogenesis and the ability of the *C. perfringens'* sialidases to potentiate the activity of the main toxin from this bacterium make a structural study of NanJ important. We hypothesize that the NanJ modules are spatially organized with respect to one another in orientations that would allow simultaneous coordination of catalysis, carbohydrate-binding and formation of cohesin-dockerin interactions.

The objective of this study is to structurally characterize the individual modules and modular combinations of the C. perfringens' NanJ to understand the spatial coordination of the modules relative to one another.

These objectives will be approached by applying a "dissect-and-build" approach using X-ray crystallography, structure prediction software and SAXS as methods for analysis of the individual modules and modular combinations of NanJ. The modular complexity and large size of NanJ necessitates the analyses to be performed by dissecting the enzyme into individual modules or more manageable modular combinations.

5.2 Experimental Procedures.

Cloning, protein production and purification. The DNA fragments that encode the modules and module combinations of NanJ (locus tag CPF_0532) were PCR amplified from *C. perfringens* ATCC 13124 genomic DNA using the oligonucleotide primers listed

in Table 8. Modular combinations produced are listed in Table 9. The PCR amplified gene fragments were obtained using standard PCR methods with Phusion High-Fidelity DNA polymerase (New England Biolabs). The amplified gene fragments were cloned into pET-28a (+) (Novagen) via engineered 5' and 3' *NdeI* and *XhoI* restriction sites, respectively, and the cohesin module was cloned using 5' and 3' *NheI* and *XhoI* restriction sites, respectively. Standard cloning procedures were used. The resultant plasmids encoded the polypeptides preceded by an N-terminal, thrombin cleavable, six-histidine tag. The DNA sequences of the constructs were verified by bidirectional sequencing.

Name	Nucleotide sequence
CBM32F	CAT ATG GCT AGC GCT ATT ATT GAA ACT GC
CBM40F	CAT ATG GCT AGC ATC AAA GGC GAA GTA GAT
CBM40R	GGA TCC CTC GAG TTA CTT AGT TTC CCC TGT TTT
GH33F	CAT ATG GCT AGC GCG CCA TCA GAG GAT AGT TTA TTG
GH33R	GGA TCC CTC GAG TTA TTC AGA AGC ATT AAA CTT AAG
UnkF	CAT ATG GCT AGC GAT TCA CCA TCA GCT TCA GTT C
UnkR	GGA TCC CTC GAG TTA TAC ATT AAT TCC TGT ATT ATC
CohF	CAT ATG CAT ATG CAG ATT GGT GAA TTA
CohR	CAC CAC CTC GAG TTA TGA AGC TTC TTC ATT
FN3R	GGA TCC CTC GAG TTA CCT AGC AGT TCT TAT AG

Table 8: Primers used for cloning of recombinant NanJ

Construct	Amino acid boundaries	Primers used to amplify gene fragment	Molar extinction coefficients ($M^{-1} cm^{-1}$)
CBM32-CBM40	42-367	CBM32F and CBM40R	31400
CBM40-GH33	181-826	CBM40F and GH33R	83785
CBM40-GH33-Unknown	181-938	CBM40F and UnkR	89745
GH33-Unknown	368-938	GH33F and UnkR	77825
Cohesin	939-1081	CohF and CohR	4470
Unknown-Cohesin-FN3	827-1173	UnkF and FN3R	21890

Table 9: NanJ modular combinations used in this study

The appropriate plasmids were transformed into chemically competent *E. coli* BL21 STAR (DE3) cells (Novagen) and the proteins were produced in Luria-Bertani media containing 50 μg ml^{-1} kanamycin (Sigma). The cells were grown at 37°C to an optical density of 0.5 at A_{595} and induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside at 18°C for 14 hours. Cells were harvested by centrifugation at 27 000 x g for 45 minutes

and chemically lysed (Charlwood et al., 1998). The polypeptides were purified from cell-free extract using immobilized metal affinity chromatography and previously described methods (Boraston et al., 2001). The purity of fractions was assessed using SDS-PAGE and those deemed to be greater than 95% pure were pooled, concentrated and buffer exchanged into 20 mM Tris-HCl, pH 8.0, in a stirred ultra-filtration unit (Amicon) using a 5 K molecular weight cut-off (MWCO) membrane (Filtron). The cohesin module protein was treated overnight at room temperature with thrombin. All proteins, including the cohesin module protein, were further purified by size exclusion chromatography using Sephacryl S-200 (GE biosciences) in 20 mM Tris-HCl, pH 8.0. The concentration of purified protein was determined by UV absorbance at 280 nm using calculated molar extinction coefficients listed in Table 9 (Gasteiger et al., 2003).

Crystallization and X-ray Data Collection. The isolated cohesin module and CBM40-GH33 double module crystals were obtained using sitting-drop vapour diffusion at 18^{o}C. Prior to crystallization both proteins were concentrated to 15 mg ml^{-1} in 20 mM Tris pH 8.0. Cohesin crystals were obtained in 20% polyethylene glycol (PEG) 4000, 0.2 M ammonium acetate and 0.1 M sodium acetate, pH 4.6. CBM40-GH33 crystals were obtained in 0.1 M Tris, pH 8.5, and 20% PEG 3350 and 0.1 M $CaCl_2$ (Hampton Research). Crystals were cryoprotected with crystallization solution supplemented with 20% and 25% ethylene glycol (Hampton Research) respectively, then flash cooled directly in a nitrogen gas stream at 113 K. Diffraction data for cohesin was collected on the "home-beam" comprising a Rigaku R-AXIS 4++ area detector coupled to a MM-002 X-ray generator with Osmic "blue" optics and an Oxford Cryostream 700. Diffraction data for CBM40-GH33 catalytic were collected at CMCF1 at the Canadian Light Source. All diffraction data were processed using MOSFLM/SCALA in the CCP4 suite of programs (Powell, 1999)(Collaborative Computational Project, Number 4, 1994). Data collection and processing statistics are shown in Table 10.

The cohesin structure was solved by molecular replacement using the coordinates from PDB accession code 2O4E from CpGH84C cohesin module and CBM40-GH33 catalytic was solved using molecular replacement using PDB accession code 2VK5 from NanI

from *C. perfringens* (Newstead et al., 2008). Molecular replacement was performed using MOLREP (Vagin and Teplyakov, 2010) and found one molecule of cohesin in the asymmetric unit and one molecule of CBM40-GH33 in the asymmetric unit as well. The initial models were corrected and completed manually by multiple rounds of building using COOT (Emsley and Cowtan, 2004) and refined using REFMAC (Murshudov et al., 1997). These models were manually corrected and refined as above. Water molecules were added to the cohesin structure using the refmac implementation of ARP/wARP and waters molecules were added to the CBM40-GH33 catalytic structure using COOT:FINDWATERS. In both data sets, 5% of the observations were flagged as 'free' and used to monitor refinement procedures (Brünger, 1992). Model validation was performed with SFCHECK (Vaguine et al., 1999), PROCHECK (Laskowski et al., 1993). Structure and refinement statistics are shown in Table 10.

	Cohesin module	CBM40-GH33 catalytic double module
Data Collection		
Wavelength	1.5418	0.97949
Space group	$P4_322$	$P2_12_12_1$
Cell dimensions: *a*, *b*, *c* (Å)	38.09, 38.09, 174.85	49.67, 73.65, 202.57
Resolution (Å)	20.0-1.7	59.57-2.0
R_{merge}	0.048 (0.379)	0.039 (0.419)
$I/\sigma I$	12.9 (3.2)	24.7 (3.2)
Completeness (%)	93.9 (88.9)	99.5 (98.9)
Redundancy	3.8 (3.7)	5.9 (3.9)
Refinement		
Resolution (Å)	1.7	2.0
No. of reflections	13479	48540
R_{work}/R_{free}	21.6/25.7	20.2/25.2
No. of atoms		
Protein	1109	4977
Ions	1	2
Water	132	453
B-factors		
Protein	11.8	32.7
Ions	7.5	32.5
Water	23.2	38.3
Bond lengths (Å)	0.02	0.011
Bond angles (degrees)	1.881	1.298
Ramachandran		
Preferred (%)	95.3	95.1
Allowed (%)	4.3	4.5
Disallowed (%)	0.4	0.5

Table 10: X-ray crystallographic data collection and structural refinement statistics for NanJ constructs.

Structure prediction; Unknown and FN3 modules. Protein Homology/analogY Recognition Engine (PHYRE) was used to predict the protein structure of the unknown module (Kelley and Sternberg, 2009). A structure of the unknown module was modelled against PDB accession code 1ULV residues 492-574 yielding an immunoglobulin-like β-sandwich fold. This structure was produced with an estimated precision of 85% and an E-

value of 2.66 e^{-01}. A structure of the FN3 module was generated using Swiss Model, a fully automated protein structure homology-modeling server (Arnold et al., 2006). The generated prediction was based on template PDB accession code 2W1N, the FN3 module from *C. perfringens* GH84C which has a sequence identity of 68% and an E-value of 2.10 e^{-27} was found.

*SAXS.*SAXS data were recorded on Beamline 4-2 (BL4-2) at the Stanford Synchrotron Radiation Laboratory using a MarCCD165 detector. The scattering patterns were measured with a data series of 10 frames with 3 minute exposure time at 20°C with wavelength of 1.127 Å with a sample-to-detector distance of 1.7 m, leading to scattering vectors q (defined as $q=4\pi/\lambda \sin\theta$, where 2θ is the scattering angle) ranging from 0.02 to 0.42 $Å^{-1}$. Bovine serum albumin at a concentration of 9.5 mg ml^{-1} was measured as a reference and for calibration. The concentrations of the protein samples ranged from 1-8 mg ml^{-1} (listed in Table 11) and each protein was measured at three or four different concentrations. The program SasTool was used for integration of the 2-dimensional scattering patterns and reduction to a 1-dimensional scattering profile, normalizing for incident beam intensity, correcting for detector response and averaging over the entire data series. The averaged buffer data was subtracted from the averaged sample data using the program Primus from the ATSAS suite of software, resulting in a buffer subtracted scattering curve (Konarev et al., 2003). The Guinier approximation [3], $\ln[I(q)]=\ln[I(0)]-q^2R_g^2/3$, where $I(q)$ is the scattered intensity and $I(0)$ is the forward scattered intensity extrapolated at zero angle, was used to determine the radius of gyration, R_g, and $I(0)$ by plotting data in the low q range as $\ln[I(q)]$ vs q^2 (Fournet, F). R_g is estimated from the slope and $I(0)$ from the intercept of the linear fit of this plot in the q-range $q \cdot R_g<1.3$. At low angles the scattered intensities were well approximated by the Guinier approximation. The pair-density distance distribution function, P(r), which provides real space information about the distances between electrons in the scattering sample, was calculated by indirect Fourier transform methods of the scattering intensity, $I(q)$, using the program GNOM (Svergun, 1992).

Low resolution envelopes of the proteins were generated *ab initio* using the program Dammif by simulated annealing using a single phase dummy atom model (Franke and Svergun, 2009). Twenty independent Dammif reconstruction runs were done with no shape constraints introduced. The *ab initio* models were aligned and averaged using DAMAVER (Volkov and Svergun, 2003) and the shapes were found to be highly similar in all cases, with normalized spatial discrepancy (NSD) values ranging between 0.62 and 0.81 (Putnam et al., 2007). The X-ray structures and the PHYRE and Swiss Model generated structures were positioned appropriately in the solution envelopes using PyMOL. Crysol was used to evaluate the solution scattering for the atomic molecules and fit the SAXS experimental curve, and calculate the goodness of fit by minimizing the discrepancy (Chi-square value) which is defined by Konarev *et al.* (Konarev et al., 2003; Svergun et al., 1995). Data are summarized in Table 11 and 12, the Crysol fits are shown in Figure 29.

Constructs	**[Protein] mg ml^{-1}**	**R_g (Guinier) Å**	**R_g (Gnom) Å**	**D_{max} Å**
CBM32-CBM40	2	25.7 ± 0.3	25.2 ± 0.1	75
	4	26.5 ± 0.4	25.5 ± 0.1	80
	6	25.4 ± 0.1	25.4 ± 0.1	75
CBM40-GH33	1	38.8 ± 1.0	34.8 ± 0.1	110
	2	38.5 ± 0.3	36.4± 0.1	110
	4	37.7 ± 0.2	35.9± 0.1	110
	6	36.1 ± 0.2	35.4± 0.1	110
CBM40-GH33-Unk	2	34.1 ±0.2	35.5± 0.1	110
	3	33.4 ± 0.1	35.1± 0.1	110
	4	32.9 ± 0.1	34.8± 0.1	110
	5	32.6 ± 0.1	34.6± 0.1	110
GH33-Unk	2	28.3 ± 0.9	29.7± 0.1	95
	4	32.5 ± 0.4	31.7± 0.1	100
	6	35.7 ± 0.3	32.7± 0.1	100
	8	31.7 ± 0.2	32.1± 0.1	105
Unk-Coh-FN3	2	46.3 ± 1.9	40.3 ± 0.1	115
	4	45 ± 0.7	39.8 ± 0.1	115
	6	50.3 ± 0.7	41.5 ± 0.1	120
	8	51.4 ± 0.5	40.1 ± 0.1	115

Table 11: SAXS parameters of NanJ constructs at different concentrations.

						Ab initio modelling		
Constructs	**Molecular Weight kDa**	**[Protein] used mg mL^{-1}**	**R_g (Guinier) Å**	**R_g (Gnom) Å**	**D_{max} Å**	**χ (Dammif)**	**NSD**	**χ (Crysol)**
CBM32-CBM40	35.5	4	26.5 ± 0.4	25.5 ± 0.1	80	1.3 ± 0.1	0.67 ± 0.04	2.4
CBM40-GH33	71.9	1	38.8 ± 1.0	34.8 ± 0.1	110	1.4 ± 0.1	0.62 ± 0.02	1.6
CBM40-GH33-Unk	84.1	2	34.1 ± 0.2	35.5± 0.	110	1.3 ± 0.1	0.71 ± 0.08	2.2
GH33-Unk	63.3	2	28.3 ± 0.9	29.7± 0.1	95	1.2 ± 0.1	0.81 ± 0.01	1.5
Unk-Coh-FN3	37.8	2	46.3 ± 1.9	40.3 ± 0.1	115	1.3 ± 0.1	0.73 ± 0.02	5.2

Table 12: Structural SAXS parameters and *Ab initio* modelling data

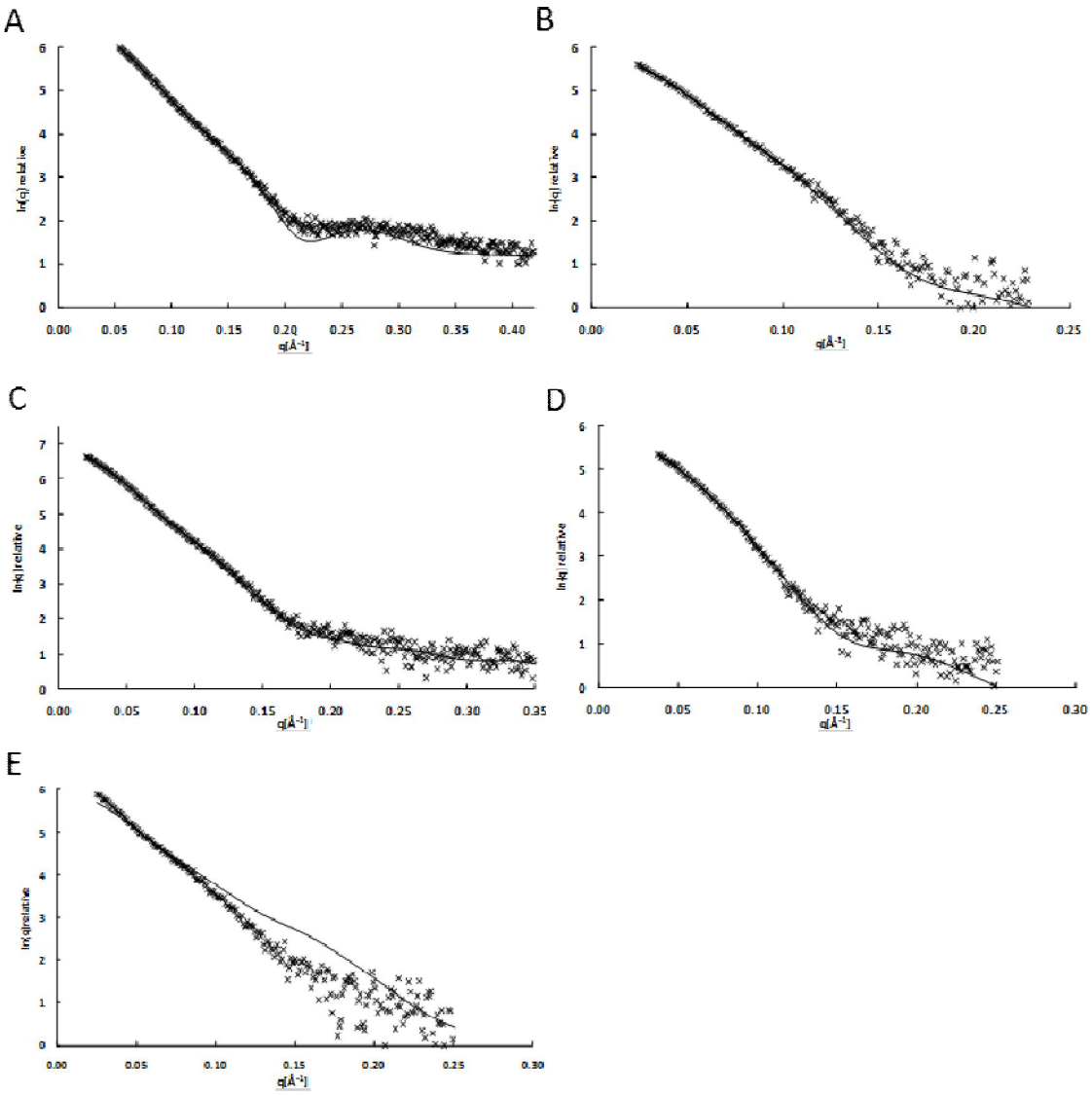


Figure 29: Crysol generated theoretical SAXS scattering curve fit to the experimentally generated SAXS scattering curve.

A) CBM32-CBM40 B) CBM40-GH33 C) CBM40-GH33-unknown D) GH33-unknown E) Unknown-cohesin-FN3

Calorimetry. The third FIVAR module from CpGH84A and the dockerin module, FIVAR-dockerin, produced in Section 4.3. and the NanJ cohesin protein samples were prepared for calorimetry and dialyzed into 25 mM Tris HCl (pH 7.5), 50 mM NaCl, and 5 mM$CaCl_2$, and were filtered and degassed at 21°C. Titration of FIVAR-dockerin (75 µM) into cohesin (5 µM) was performed by using a VP-ITC titration calorimeter from

Microcal. Twenty-five injections of 10 µl were made, with 240 s of equilibration between injections. The titrations were done in triplicate at 30°C.

5.3 Results and Discussion.

To study the individual modules and the overall architecture of *C. perfringens* NanJ, we dissected the enzyme into various manageable modular combinations. This study is similar to the complete structural analysis of a multi-modular GH84C from *C. perfringens* (Ficko-Blean et al., 2009). This model study revealed insight into how catalysis, carbohydrate-binding and the formation of cohesin-dockerin interactions resulting in enzyme complexes could be spatially coordinated in an enzyme that is involved in host-pathogen interaction. Using this strategy allowed us to successfully clone, recombinantly produce and purify soluble proteins. Crystallization of all constructs was attempted but was only successful for the CBM32 and CBM40 individual modules which were solved previously by Boraston *et al.* (Boraston et al., 2007), CBM40-GH33 catalytic module double construct, and the individual cohesin module. The remaining NanJ modules, the unknown module and the FN3 module, were analyzed by structure prediction software to provide an estimation of their 3D structure. Next, low resolution SAXS solution data was collected on the various modular combinations enabling the generation of SAXS *ab initio* envelopes. The X-ray crystallography or predicted structures were docked into the envelopes revealing the overall spatial coordination of the modules.

5.3.1 Positioning the CBM32 and CBM40 modules.

The solution conformation of the CBM32-CBM40 modular pair was analyzed by SAXS. The SAXS envelope was generated by multiple dammif envelopes which were averaged yielding an NSD value of 0.67 that represent highly similar forms. The *ab initio* generated molecular envelope reveals an elongated kidney shape (Figure 30A, B). The R_g values were also in very good agreement with similar guinier and gnom generated R_g values, 26.5 ± 0.4 and 25.5 ± 0.1 Å, respectively. The maximum linear dimension (D_{max}) was found to be ~80 Å which corresponded with the measured length of the longest dimension of 76.7 Å. The atomic structures of both CBMs were previously determined by X-ray crystallography by Boraston *et al.* (Boraston et al., 2007) and these structures

were placed in the SAXS envelope (Figure 30B). The individual structures were manually rotated and translated within the envelopes and the relative positions of the N- and C-termini of the domains were taken into account and used as a restraint for the positioning of the termini of the contiguous modules. The nine amino acid linker distance between the modules was kept in consideration and several viable positions for these modules were obtained within the SAXS envelope. This manipulation revealed a model that was in compliance with the geometrical limits imposed on this modular pair and gave a satisfactory χ_{crysol} value of 2.4. The orientation of the ligand binding sites of CBM32 and CBM40 were found to be ~56 Å apart and at an angle of ~ 106° from one another from the centre of the SAXS envelope (Figure 30C).

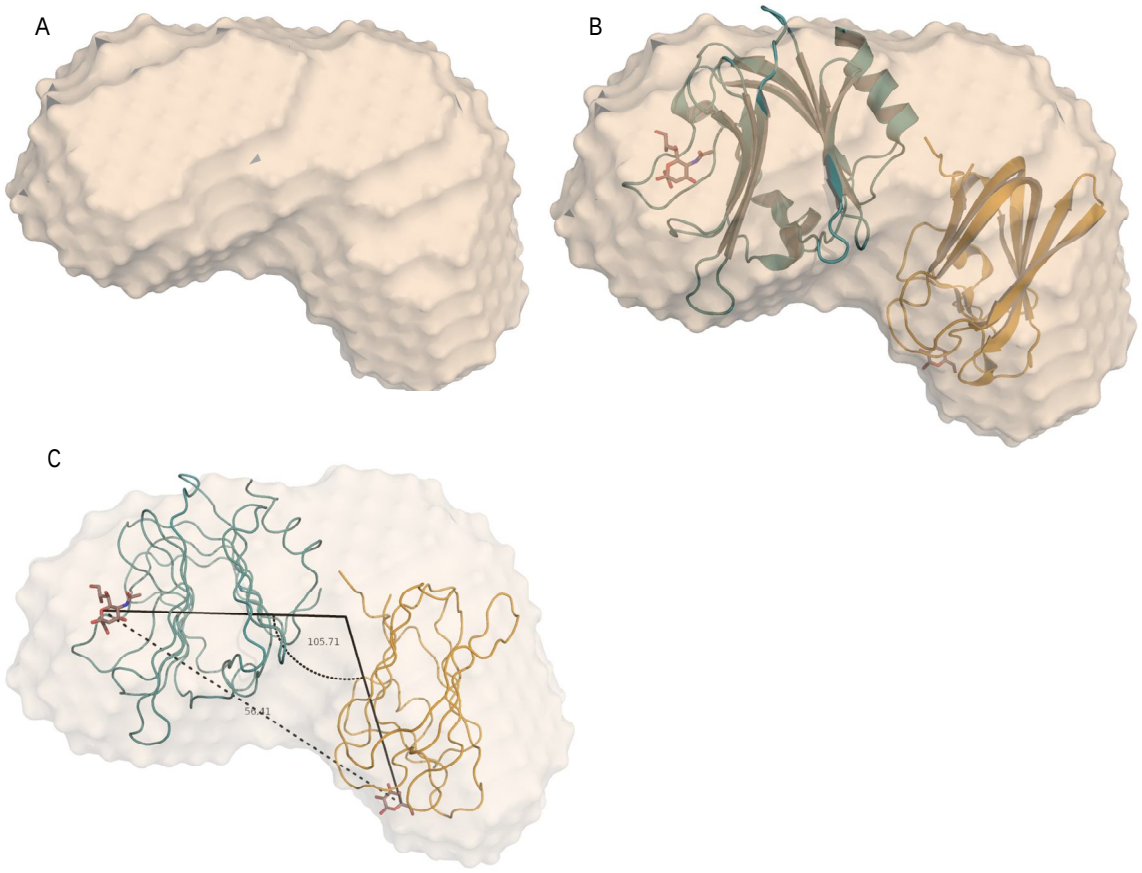


Figure 30: SAXS envelope, CBM40-CBM32.

A) *Ab initio* SAXS-generated envelope for double module CBM40-CBM32 B) The CBM40 and CBM32 modules were fit into the SAXS envelope. The CBM40 is shown in green bound to sialic acid in cartoon representation. The CBM32 is shown in orange bound galactose in stick representation C) Distance and angles between CBM32 and CBM40 bound ligands. CBM modules shown in ribbon representation.

5.3.2 Structure of the CBM40 - GH33 catalytic module double construct.

The CBM40-GH33 catalytic double module construct, here called CBM40-GH33, was crystallized and its structure determined by X-ray crystallography to a resolution of 2.0 Å (Figure 31A and Table 10). The structure of the CBM40 is, as expected, very similar to the CBM40 structure already reported and discussed (Boraston et al., 2007). The models align with a root mean square deviation (RMSD) of 0.464 Å over 187 Cα positions with the only visible difference between the two being the presence of two non-structural Ca^{2+} from the original crystallization conditions in the CBM40 structure reported by Boraston *et al*. The NanJ catalytic GH33 module structure is connected to the CBM40 by a linker of five amino acids for which the electron density was too sparse to accurately model. It was clear that the placement of these two modules, in the asymmetric unit, was the correct positioning because the short length of the linker prohibited any alternate positions in the symmetry molecule. Also, the short length of the linker would inhibit much movement between the two modules but could allow for marginal flexibility. The NanJ catalytic module structure has two distinct domains, a six-bladed β-propeller catalytic sialidase domain that is formed by residues 368-491 and 563-819 and a small β-barrel domain that is formed by residues 492-562. The backbone RMSD between the C. *perfringens* NanI and NanJ catalytic modules is 1.138Å for 432 Cα positions that are equivalent and the amino acid identity is 59% (Figure 31B, C and Figure 32). NanI coordinates two Ca^{2+} residues in the β-propeller by oxygen atoms very similar to the coordination found in NanI, however, these Ca^{2+} do not appear to play a role in substrate recognition or catalysis due to their positions which are distal from the active site. The structure of NanI, being so similar to NanJ, allows for some inference of the location of the active site and the putative catalytic residues as well as other active site residues of NanJ (Figure 31C). A key feature of sialidases is the clustering of three arginine residues and it is conserved in NanI and NanJ and consists of Arg266/399, Arg555/687, and Arg615/797, respectively, and in NanI interacts with the carboxylate group of the sialic acid. Another conserved feature of sialidase active sites is the hydrophobic pocket that accommodates the *N*-acetyl group of the substrate. In NanI and NanJ these residues are conserved and consist of Phe347/485, Phe353/491, Phe460/599, Thr345/483, Ile327/459,

respectively, and lastly Trp354 forms a cap on the hydrophobic pocket in NanI, which is replaced by Pro492 in NanJ. The catalytic nucleophile in NanI is Tyr655 and the equivalent residue in NanJ is Tyr787. The other catalytic residue is Asp291 in NanI and its equivalent in NanJ is Asp424. The proposed mechanism of catalysis of NanI was discussed in depth by Newstead *et al.* (Newstead et al., 2008) and occurs with a similar mechanism to other retaining glycoside hydrolases. The mechanism of NanJ will not be discussed here but is assumed to be very similar to that of NanI based on the conservation of catalytic and other active site residues.

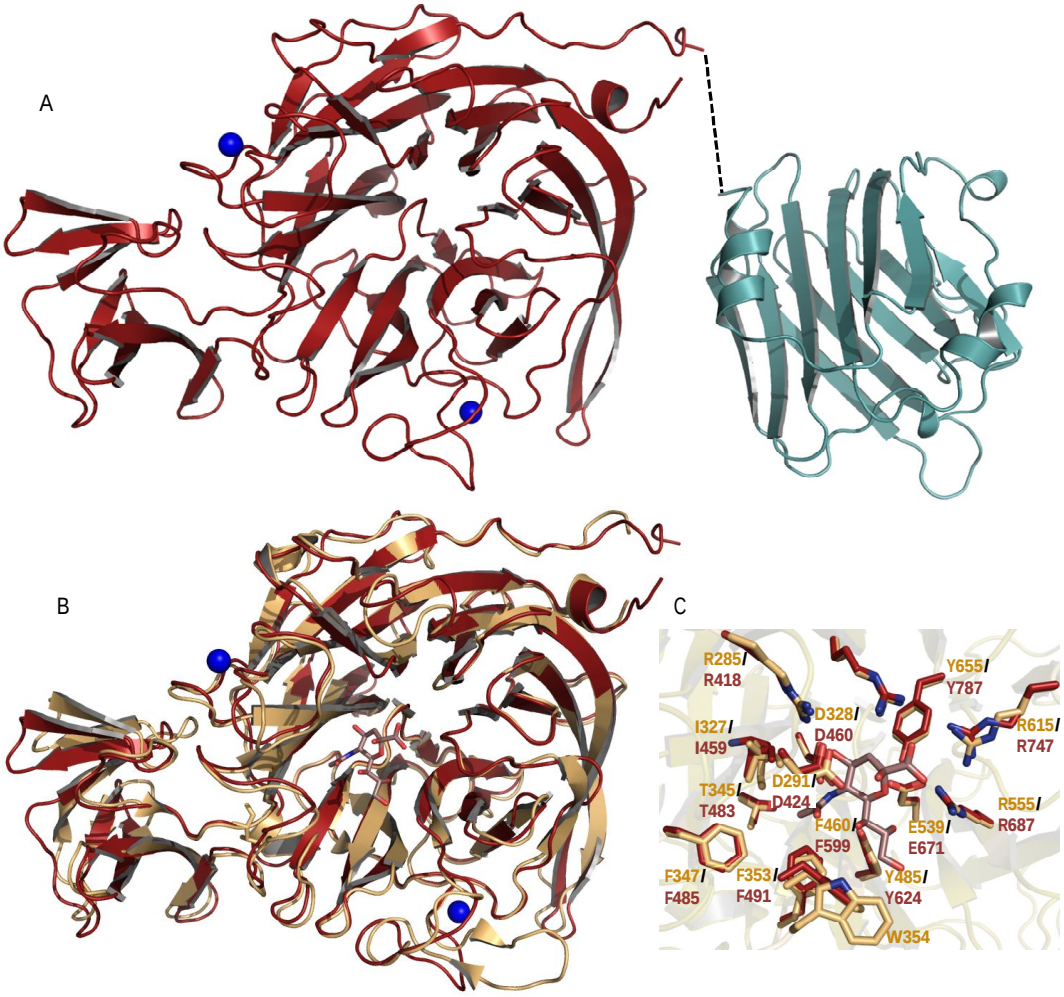


Figure 31: CBM40-GH33 catalytic double module construct X-ray structure and overlay with NanI.

A) Cartoon representation of the CBM40-GH33 catalytic double module construct X-ray structure. Catalytic module in red and CBM40 in blue. Coordinated calcium ions in blue. Unmodelled linker indicated by dashed line B) NanI and NanJ cartoon overlay. NanI shown

in gold with sialic acid shown in stick representation. NanJ shown in red. C) NanI and NanJ active site overlay depicting the catalytic residues of NanI (gold) and analogous residues in NanJ (red) in stick representation. Residues are identified by one-letter code, and coloured and numbered accordingly.

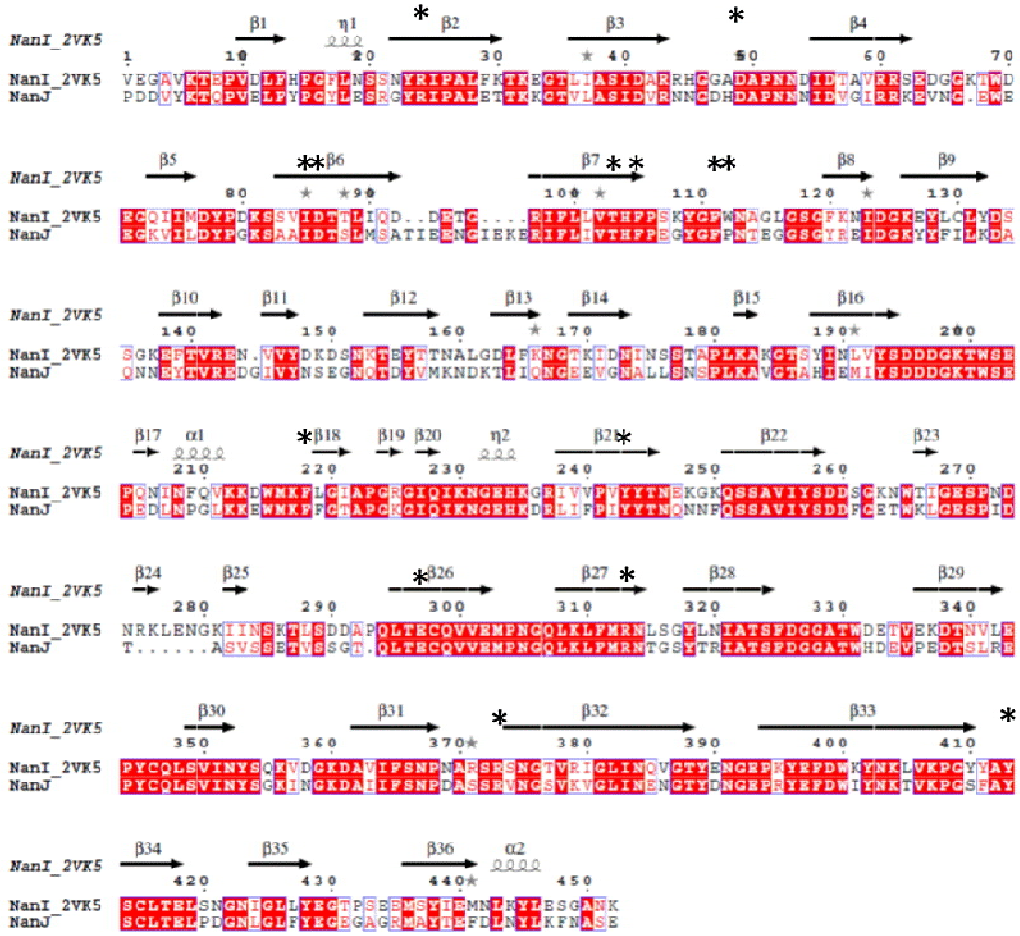


Figure 32: Amino acid sequence alignment of NanI and NanJ.

Alignment of amino acid sequences of NanI (pdb accession code 2VK5) and NanJ from *C. perfringens* using ClustalW2. This figure was generated using ESPript with secondary structure elements indicated above sequence from structure 2VK5. Conserved active site residues denoted by asterisks.

5.3.3 Positioning of CBM40-GH33 modular pair.

The solution conformation of the CBM40-GH33 catalytic double module construct, referred to as CBM40-GH33, was also analyzed via SAXS. The NSD value of 0.62 revealed that the dammif generated envelopes were highly similar and the R_g values

generated from guinier (38.8 ± 1.0 Å) and gnom (34.8 ± 0.1 Å) were also in agreement. The *ab initio* calculation of the SAXS molecular envelope yielded an extended form with a D_{max} of 110 Å (Figure 33). One end of the envelope was slightly narrower than the other, clearly able to restrict the catalytic module from fitting there but sufficient for CBM40 accommodation. To minimize bias, the coordinates of the individual modules were placed separately in the SAXS envelope. As with the CBM32-CBM40 double module construct, the modules were translated and rotated within the envelope and the N- and C-termini positions were taken into consideration. Based on the different positions tested, the best model was found to be that of the CBM40-GH33 catalytic double module X-ray crystal structure giving an excellent χ_{crysol} value of 1.6. The position of the CBM40 binding site and the GH33 catalytic site are found on opposite faces of the bimodular envelope with a distance of ~70 Å between them.

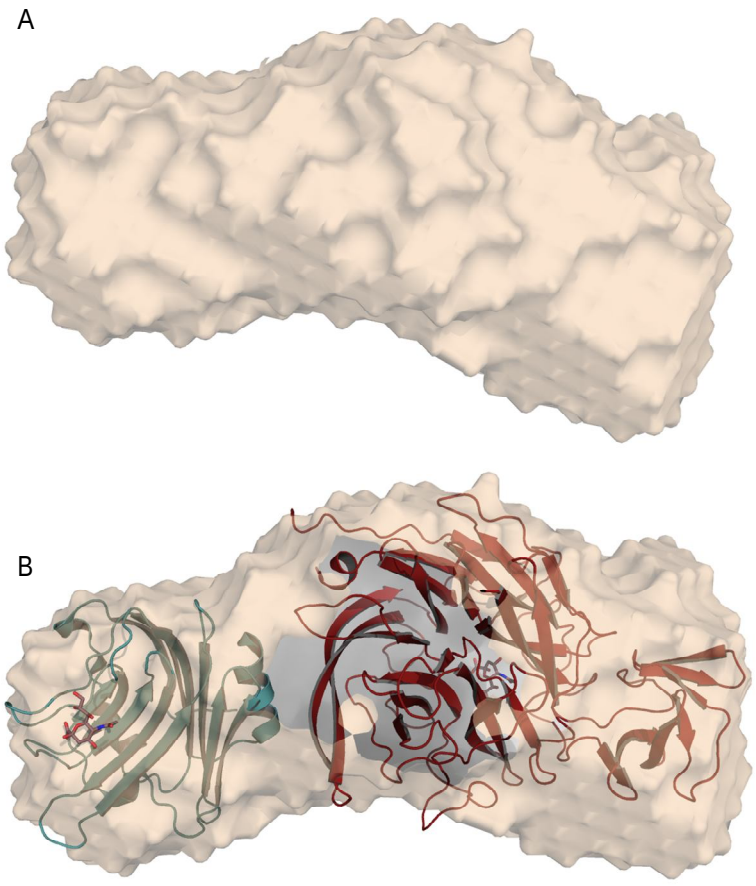


Figure 33: SAXS envelope, CBM40 and GH33 catalytic modules.
A) *Ab initio* SAXS-generated envelope for double module CBM40-GH33 B) The CBM40 and GH33 catalytic modules were fit into the SAXS envelope. The GH33 catalytic module is shown in red cartoon representation with the NanI sialic acid shown to indicate the active site. The CBM40 is shown in green bound to sialic acid in stick representation.

5.3.4 Positioning of CBM40-GH33-Unk triple module construct.

Further NanJ constructs consisting of multiple modules, aside from the double module CBM40-GH33 construct, proved recalcitrant to crystallization and so a strictly SAXS based approach was used to study the triple module construct, CBM40-GH33-unknown. The CBM40-GH33-unknown triple module envelope was very similar to that for the CBM40-GH33 construct but was ~18 Å wider at one end which is to be expected to allow for the addition of the unknown module but was very similar at the catalytic module end with the same, small protrusion for the β-barrel portion of the catalytic module (Figure 34). The length of the envelope was ~113 Å which is not only near-identical to the length found for the double module construct but also is fitting with the calculated D_{max} of 110 Å. The dammif generated envelopes were very similar to one another with an NSD of 0.71. Due to the fact that crystallization of the unknown module was not successful, either individually or in modular combinations, we used the structure prediction software, PHYRE, to estimate the structures of this module for the purpose of modeling into a low resolution SAXS envelope. The unknown module has no putative conserved domains and no known or hypothetical function and very little information was gained regarding a possible function for the unknown module. The coordinates of the individual modules- CBM40 and GH33 catalytic, and the structure prediction model of the unknown module - were manually placed in the envelope with similar restraints as previously mentioned. The best solution was found to yield an acceptable χ_{crysol} value of 2.2 and was generated by positioning the X-ray structure model from the CBM40-GH33 double module construct, oriented in the same way as previously described in Section 5.3.3, and then manually moving around the unknown module with the proper N-and C-termini orientation restricted based on the number of linking amino acids between the modules. The model generated revealed that the unknown module favours positioning in

alignment with the CBM40 module making the three modules appear in a V shape at an angle of ~33° with the catalytic module being at the vertex, instead of in a linear arrangement. There appears to be some leniency with the positioning on the unknown module allowing it to translate to be closer or farther from the catalytic module within the SAXS envelope, however, the model shown produced the best χ_{crysol} value, leading us to believe that it favours being positioned closer to the CBM40. Another factor that could contribute to the variability in the positioning of the unknown module is the fact that the unknown module structure being used is not an X-ray structure, but a structure prediction generated with relatively low confidence. The approximation of this structure could be a poor representation of the 3D shape of this module and further attempts at obtaining an atomic structure of this module are required.

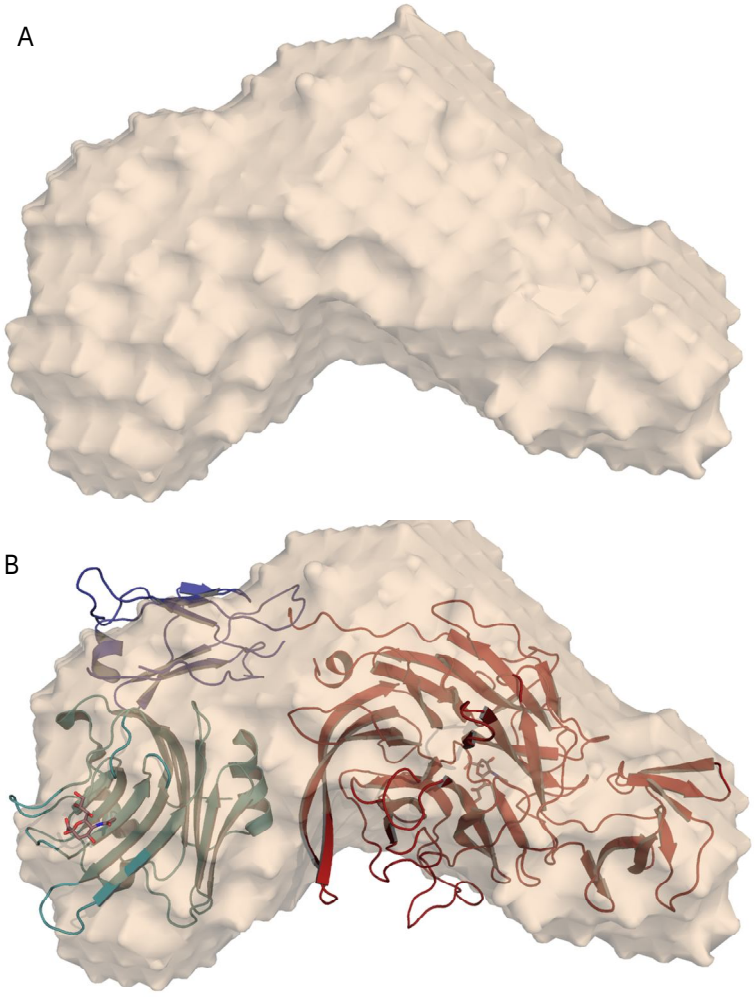


Figure 34: SAXS envelope, CBM40-GH33-unknown.

A) *Ab initio* SAXS-generated envelope for triple module construct, CBM40-GH33-unknown B)The CBM40 ,GH33 catalytic, and unknown modules were fit into the SAXS envelope. The GH33 catalytic module is shown in red cartoon representation with the NanI sialic acid shown to indicate the active site. The CBM40 is shown in green bound to sialic acid in stick representation. The unknown module is shown in dark blue.

5.3.5 Positioning of GH33-unknown modular pair.

Although the GH33-unknown modular pair was previously investigated in the triple module construct, the double module construct was also analyzed by SAXS to check for consistencies in the data. This double module pair SAXS generated *ab initio* envelope was ~95 Å long and had the same characteristic shape at one end which clearly fits the catalytic module with its β-propeller region fitting into the small protrusion, similarly to the triple module CBM40-GH33-unknown construct (Figure 35). Also, the NSD of 0.81 indicated agreement between the dammif generated shapes and the R_g values were very similar. The catalytic and unknown modules were manipulated within the envelope and the catalytic module was restricted to the one end of the envelope as described previously. Various positions of the unknown module were analyzed with CRYSOL and the best χ_{crysol} value generated was an excellent 1.5 and was found using the model used for the CBM40-GH33-unknown triple module construct with the CBM40 coordinates removed. There appears to be a large area where the unknown module could be positioned in the envelope allowing for movement of this module. This suggests some flexibility of the unknown module when not sterically restricted by the presence of the CBM40 module. To see if other positions of the unknown module would yield a better χ_{crysol} value, the unknown module was positioned in various conformations and the best model found gave a χ_{crysol} value of 2.1. This is a reasonable solution but the model above remains the best described.

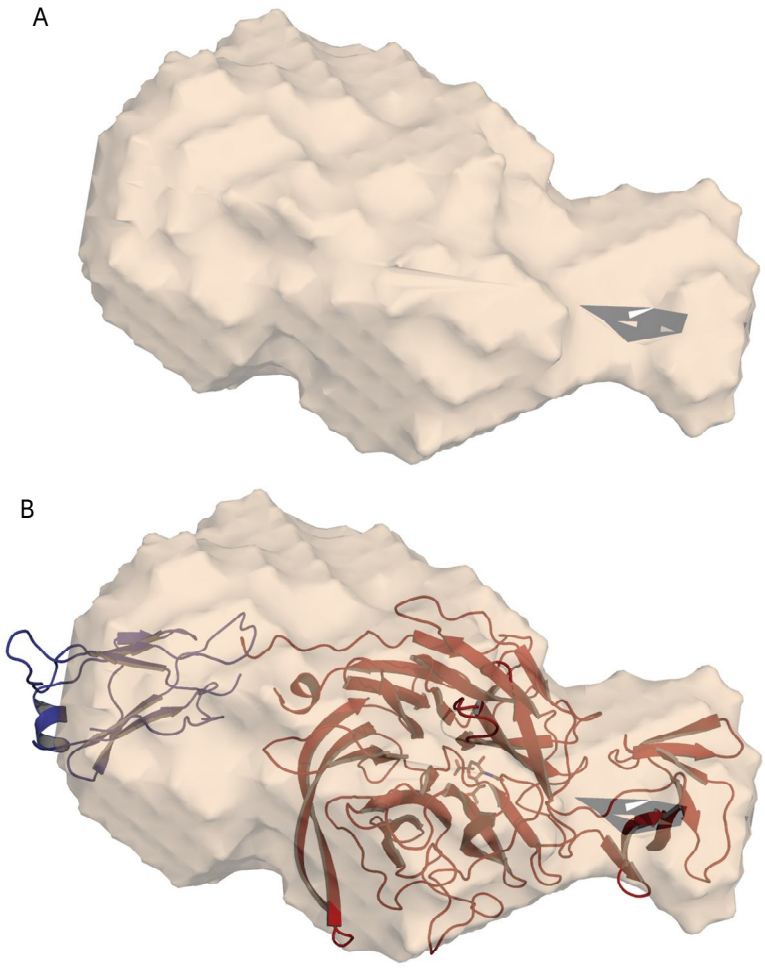


Figure 35: SAXS envelope, GH33 catalytic-unknown modules.

A) *Ab initio* SAXS-generated envelope of double module construct, GH33 catalytic-unknown. B) The GH33 catalytic and unknown modules were fit into the SAXS envelope. The GH33 catalytic module is shown in red cartoon representation with the NanI sialic acid shown to indicate the active site. The unknown module is shown in dark blue.

5.3.6 Structure of the cohesin module.

The cohesin module from NanJ was cloned as an isolated module, produced recombinantly, and crystallized. The structure of the cohesin module was solved by molecular replacement using CpGH84C cohesin module (pdb accession code: 2O4E) to a resolution of 1.7 Å (Table 10). The structure revealed an expected β-sandwich fold similar to the cohesin module described in Section 4.3.2 from CpGH84C. These two cohesin modules have a 33% amino acid identity and a backbone root mean square deviation (RMSD) of 1.2 Å over 143 Cα positions (Figure 36A). This cohesin module

also has a very well conserved dockerin binding interface (Figure 36B) indicating that this cohesin module could form a very similar ultra-tight cohesin-dockerin interaction to the one formed by the CpGH84C cohesin module with the CpGH84A μ-toxin dockerin. The dockerin recognition residues from CpGH84C cohesin module are conserved in the NanJ cohesin module and are listed here and numbered as CpGH84C cohesin residues/NanJ cohesin module; F808/Y974, A809/S975, D811/E977, F833/F1000, N835/K1002, K837/K1004, R845/R1012, L847/L1014, S849/A1016, S850/S1017, T851/L1018, T852/G1019, V892/T1058. The binding ability of this cohesin module was tested in an ELISA-based binding experiment which revealed an interaction with all of the putative *C. perfringens* dockerins that also bound the CpGH84C cohesin module (Appendix B: Figure 45) (Adams et al., 2008). In addition to this, isothermal titration calorimetry of the NanJ cohesin and the μ-toxin FIVAR-dockerin double module construct was performed and revealed an ultra-high interaction similar to the one found for CpGH84C cohesin and the μ-toxin FIVAR-dockerin described in Section 4.3.1 (Figure 37).

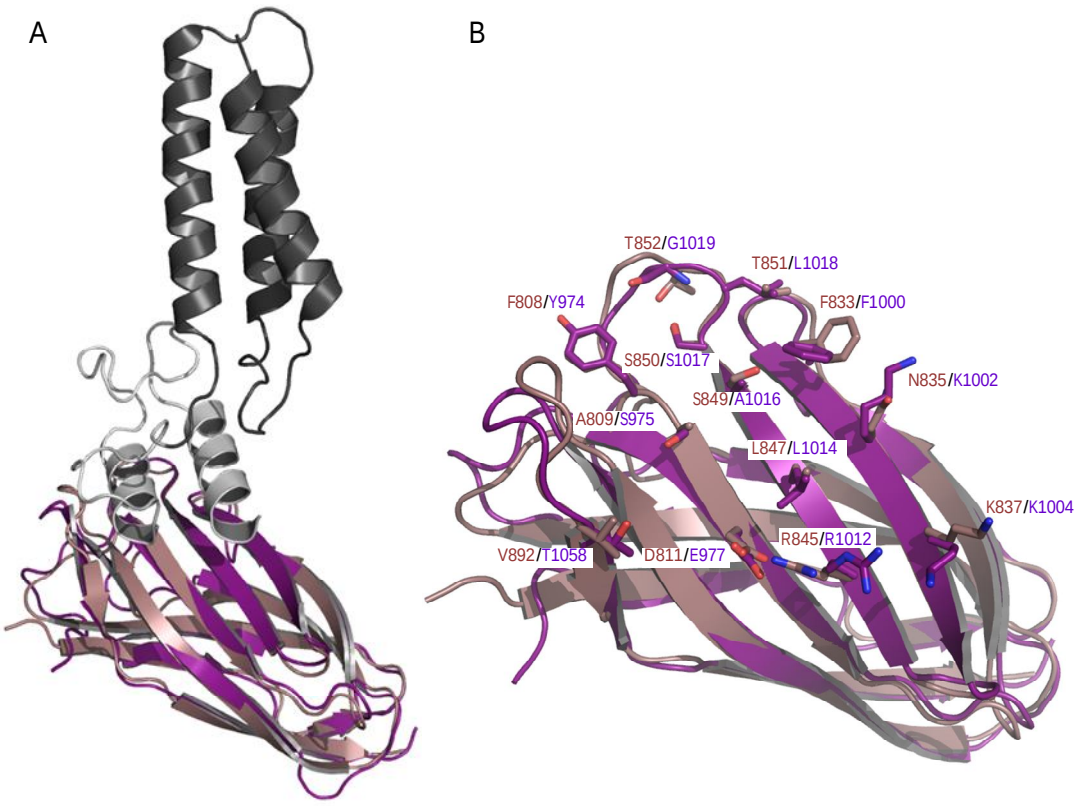


Figure 36: Structural homology and interface residue conservation displayed by the *C. perfringens* cohesin modules.

A) Backbone overlay of NanJ cohesin module (dark purple) with the cohesin module from *Cp*GH84C (salmon) in complex with the dockerin-FIVAR (grey and black, respectively) from *Cp*GH84A. B) Overlay of the NanJ and *Cp*GH84C cohesin modules depicting the dockerin recognition residues of *Cp*GH84C cohesin module and the analogous residues in NanJ cohesin module in stick representation. Residues are identified by one-letter code, and coloured and numbered accordingly.

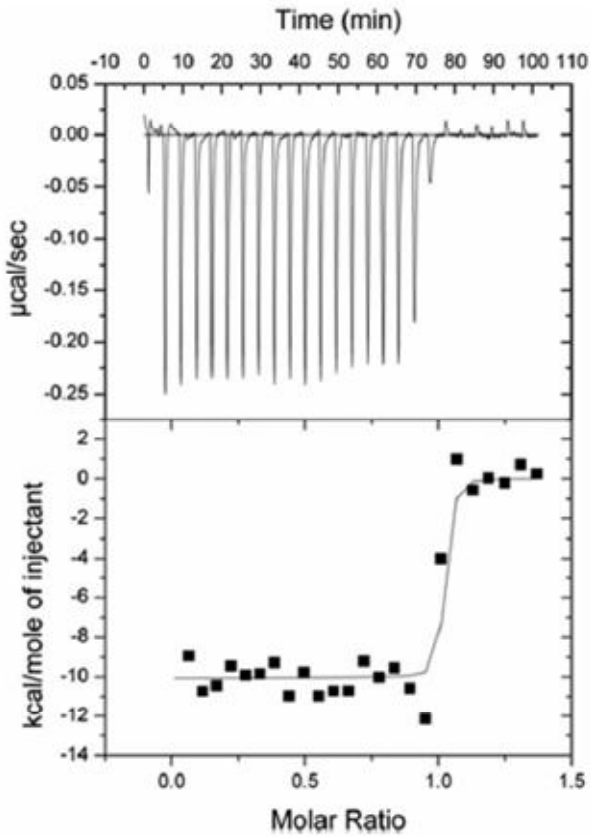


Figure 37: Isothermal titration calorimetric analysis of the NanJ cohesin·µ-toxin FIVAR-dockerin interaction at 30°C.
(*Upper*) Raw heat measurement. (*Lower*) Integrated heats after correction for heats of dilution as determined from the heat of ligand additions at the excess of saturation. The curve represents the best fit to a single-site model.

5.3.7 Positioning of Unknown-Cohesin-FN3 triple module construct.

Aside from the cohesin module, the C-terminal end of NanJ was recalcitrant to crystallization, whether in single or multiple modular constructs. The triple module construct, unknown-cohesin-FN3, was the only construct that incorporated the C-terminal end for which the protein produced was sufficiently stable for SAXS analysis.

Crystallization of the FN3 module was not successful, either individually or in modular combinations, necessitating the use of structure prediction software, Swiss Model, to obtain a model for this module. The FN3 module from CpGH84C has a sequence identity

of 68% with the NanJ FN3 producing a very confident protein structure prediction with the typical β-sandwich fold with seven β-sheets. The cohesin module from NanJ was overlapped with a cohesin-FN3 double module structure from CpGH84C, PDB accession code 2W1N, and used as a model for the relative placement of the FN3 module with some movement occurring between the two modules to better fit the SAXS molecular envelope (Ficko-Blean et al., 2009). Also, the NanJ cohesin module was overlapped with the high resolution crystal structure of the CpGH84C cohesin·CpGH84A dockerin-FIVAR complex (Adams et al., 2008), PDB accession 2OZN, to reveal the face of the cohesin module that would non-covalently interact with the dockerin module (Figure 36B). This association was deemed to be representative of the association between the NanJ cohesin module and the CpGH84A dockerin module due to the high overall identity and conservation of the binding face of the cohesin modules.

The SAXS envelope for this triple module construct was generated by multiple dammif envelopes which were averaged yielding an NSD value of 0.7 that represent highly similar forms and was long and extended with a maximum linear dimension of 115 Å (Figure 38). The quality of the models was deemed to be of good quality due to the reasonable agreement of the guinier and gnom Rg values of 46.3 ± 1.9 and 40.3 ± 0.1 Å. In determining the placements of the individual structures – structure prediction unknown, cohesin X-ray structure, and FN3 Swiss Model- the orientations of the N- and C-termini and the length of linking amino acids were taken into consideration and the modules were placed contiguously in a near linear alignment with minimal rotation between the modules. The relative positions of the unknown, cohesin, and FN3 modules reveals a long extended conformation with the relative angles between the modules being approximated, but their relative rotational orientations around these axes could not. This rotational symmetry could provide variability in the overall model, especially with the positioning of the cohesin-dockerin interface, and creates uncertainty as to the precise positions of the modules. A χ_{crysol} value of 5.2 was the best estimation found for the placements of these modules. The reason for the poor χ_{crysol} value indicates that the pseudo-atomic model generated from the structure predictions of the unknown and FN3 module and the X-ray structure of the cohesin module might be unreliable. This could be

partly due to the fact that the structures of the unknown module and the FN3 module were determined using prediction software. The structure of the FN3 can be assumed to be quite reliable given its high identity, but the unknown module structure was approximated with very little confidence. Another reason for this could be the innate flexibility between the modules which is a very common reason for poor χ_{crysol} values. The extreme flexibility between the cohesin and FN3 modules from CpGH84C made structural placement of these modules in the SAXS envelope difficult and it can be assumed due to the high identity of both of these modules with the equivalent ones from NanJ that there is a high degree of flexibility between these two modules as well (Ficko-Blean et al., 2009). Further studies to obtain an X-ray structure of the unknown module are required.

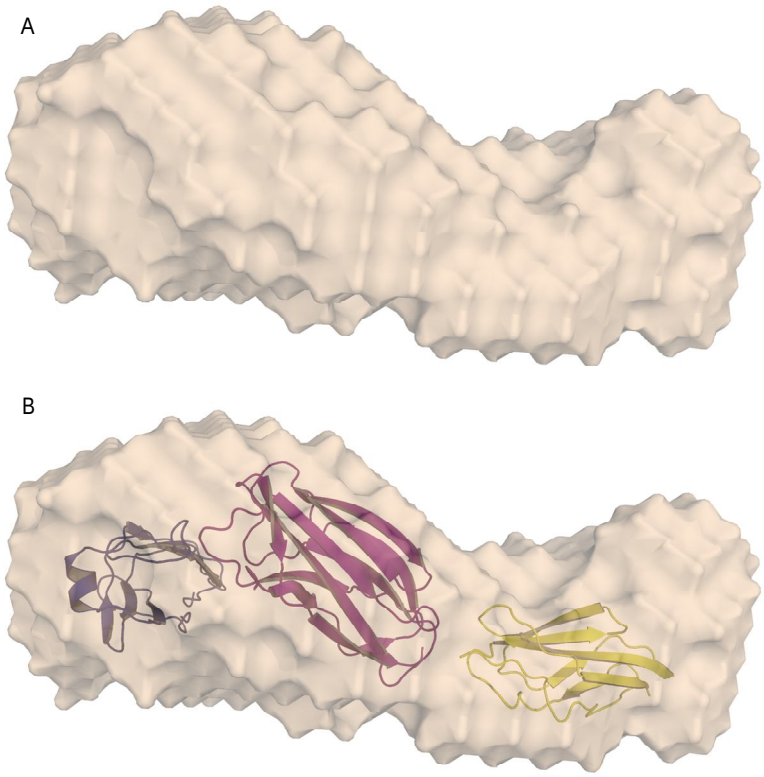


Figure 38: SAXS envelope, unknown-cohesin-FN3.

A) *Ab initio* SAXS-generated envelope for triple module construct, unknown-cohesin-FN3 B)The unknown, cohesin and FN3 modules were fit into the SAXS envelope. The unknown module is shown in dark blue cartoon representation, the cohesin in maroon, and the FN3 in yellow.

5.3.8 A composite model of NanJ.

The generation of a complete composite model of NanJ was done using an amalgamation of the X-ray crystallography structures, software prediction structures and SAXS generated *ab initio* envelopes. The overall structure of NanJ was generated by overlaying the individual modules in the multi-modular constructs to generate a whole enzyme representation (Figure 39). The CBM40 module of the CBM32-CBM40 double module construct was overlayed with the CBM40 module from the CBM40-GH33 and CBM40-GH33-unknown constructs. This generated a model that consisted of the CBM32, CBM40, GH33 catalytic and the unknown module, indicating their relative positions. The GH33 catalytic-unknown double module construct was not included in this overlay because it was redundant with the CBM40-GH33-unknown triple construct. Next, the unknown-cohesin-FN3 model was overlayed with the unknown module from the model just described. This indicated the placements of these C-terminal modules, cohesin and FN3. The cohesin module was then overlayed with the cohesin module from CpGH84C. This CpGH84C cohesin was in complex with the non-covalently associated dockerin module from the μ-toxin which was a double module construct with its neighbouring module the FIVAR. This overlay allowed for an indication of how NanJ could spatially associate with another enzyme through a cohesin·dockerin interaction. The overall composite structure reveals a model for the three-dimensional organization and coordination of the individual modules in NanJ as well as the proposed location of the cohesin·dockerin interface.

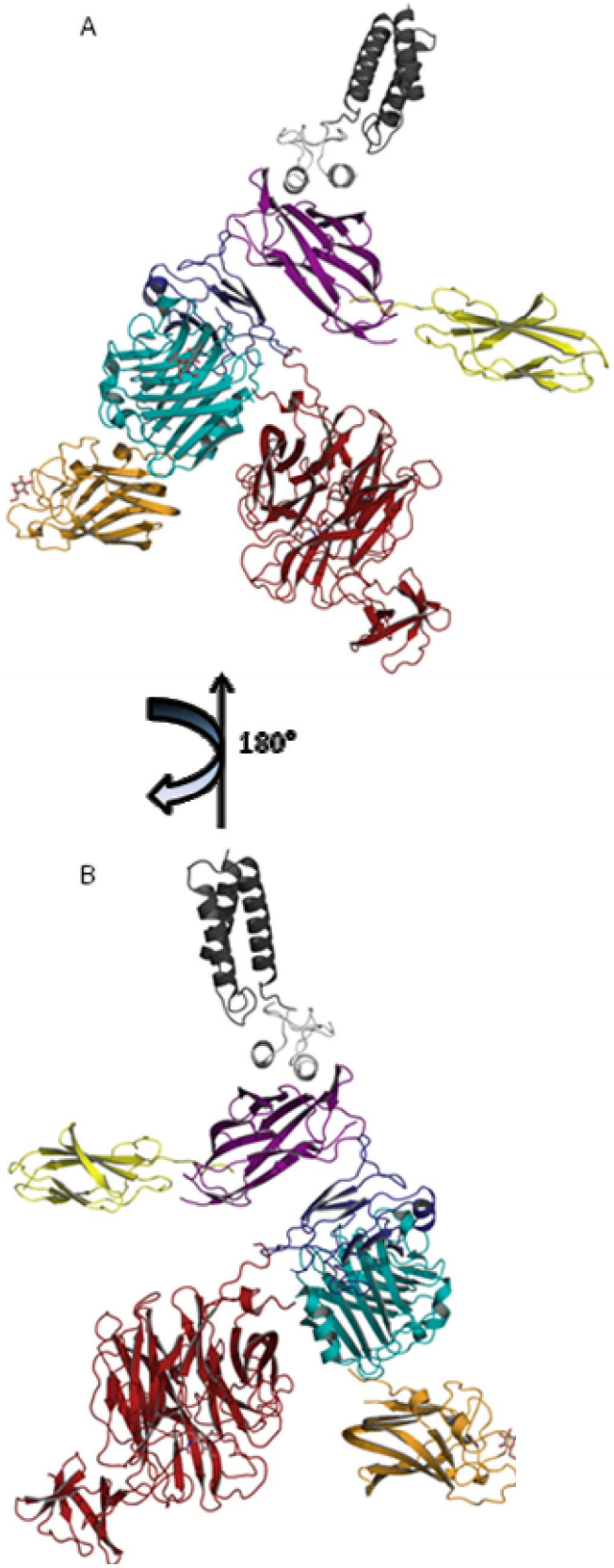


Figure 39: Composite structure of NanJ.

A) The overall structure of NanJ in complex with the FIVAR-Doc module from CpGH84A determined by the amalgamation of the SAXS data and crystallographic data. Modules coloured as in previous figures and stick representation of carbohydrate ligand/substrates. B) The overall structure shown rotated 180° around the vertical axis of the page.

One of the most striking features noted in the composite model is the clustering of the CBM32, CBM40 and GH33 catalytic modules to one 'end' of the whole enzyme. While the carbohydrate-adherence and hydrolysis portion of NanJ are localized together, their binding sites are localized on different faces and therefore do not appear to be able to function on a level homogeneous glycan surface. The heterogeneous glycans present on the outer surface of a host cell or in mucosal surfaces are the proposed substrates of *C. perfringens* and could be manipulated by adherence and hydrolysis by these three modules. The CBM32 has a binding preference for terminal galacto-configured sugars and CBM40 has a preference for terminal sialic acids. This is especially interesting considering that sialic acids are often found capping terminal galactose residues found in gangliosides and other host glycans. A possible scenario could involve the CBM40 tethering the enzyme to the glycan surface via its interaction with a terminal non-reducing sialic acid residue while the catalytic module hydrolyzes adjacent terminal sialic acid residues. After the uncapping of sialic acid residues, terminal non-reducing galacto-sugars would be exposed allowing CBM32 to further tether the enzyme to the glycan surface allowing processivity of other associated enzymes to adhere and hydrolyze the glycan surface without having to completely disassociate from the glycan surface. The removal of the terminal sialic acid residues is what has been shown to potentiate the activity of the *C. perfringens* α-toxin and contribute to the virulence and pathogenesis of this bacterium. The *Vibrio cholerae* sialidase has a family 40 CBM in addition to its catalytic module and the active sites of both of these modules are found on the same side of the bimodular pair but are not near one another with the CBM40 active site being off to the side (Connaris et al., 2009; Moustafa et al., 2004). It was proposed that the *V. cholerae* CBM would likely target the sialidase catalytic site to sialic acid-rich environments and enhance the efficiency of the enzyme despite the CBM binding site and the active site not being oriented in the exact same direction, likely a similar scenario would exist for NanJ.

There is very little information available for the structure or function of the NanJ unknown module. It does not have any conserved putative domains or hypothetical function and the immunoglobulin-like fold generated using PHYRE provided very little insight into what its biological function might be. Based on its position in the overall NanJ model, the unknown module could function as a kind of spatial divider to orient the other modules optimally for their proposed functions. Or, it could be involved in other protein-protein interactions or could function as a CBM by adhering to carbohydrate surfaces.

The cohesin module has the capacity to interact with its cognate binding partner, dockerin modules. Considering that the *C. perfringens* dockerin modules are not separate entities, but are found in GHs that contain multiple modules, it is important that this interacting pair be oriented in a fashion that allows the remaining modules of both enzymes to be accommodated. In NanJ, the position of the cohesin domain is such that the putative dockerin interacting face would orient the dockerin module to be spatially removed from the other modules of NanJ, thus allowing simultaneous carbohydrate-adherence and hydrolysis and recruitment of another enzyme through this interaction. The remaining modules of the dockerin recruited enzyme (in this hypothetical case, CpGH84A) are not spatially restricted by the other modules of NanJ and would be able to act on the glycan surface simultaneously with NanJ. Given that there are three known functional dockerin modules and three known functional cohesin modules and therefore several interacting pairs possible, there exists the possibility of multiple enzyme complexes. Given that many of these cohesin and dockerin containing enzymes also contain CBMs, simultaneous glycan adherence and hydrolysis of various different glycans along with enzyme complex formation could allow for a concerted attack on diverse host tissue glycans.

FN3 modules are found in 2% of all animal proteins (Bork and Doolittle, 1992) and have been found in many bacterial carbohydrate-active enzymes. They share a common fold, a β-sandwich consisting of seven β-strands that form two antiparallel β-sheets (Leahy,

1997). Little is known about their function but in eukaryotes they are present in proteins that mediate adhesion processes and migration, as well as surface hormone and cytokine receptors (Bencharit et al., 2007). The NanJ FN3 module is oriented away from the other modules and does not appear to interfere with adherence or hydrolysis. The very structurally similar FN3 module found in CpGH84C is also oriented away from the catalytic module and CBMs (Ficko-Blean et al., 2009). It was suggested by Ficko-Blean *et al.* that the FN3 module might act in recruiting other *C. perfringens* enzymes, similarly to the cohesin module through an uncharacterized interaction. Another suggested function was an interaction between the conserved basic residues at the distal tip of the FN3 modules with the acidic charge of the Gram-positive bacterial cell wall. This electrostatic interaction could bind the enzyme or cohesin·dockerin mediated enzyme complexes to the *C. perfringens* cell (Ficko-Blean et al., 2009).

5.4 Conclusion.

The concerted actions of all of these discrete functional units - carbohydrate binding and hydrolysis, protein-protein interaction complex formation, and potential bacterial cell wall interaction – form a highly sophisticated carbohydrate-degrading machine. The modular complexity of these enzymes is what enables the bacteria to effectively breakdown the carbohydrate rich areas of the host including the glycans on the surface of host cells and in mucosal surfaces of the gastrointestinal tract, weakening the host's first line of defense. The uncapping of sialic acids from glycoproteins and gangliosides significantly potentiates the activity of the cytotoxic *C. perfringens* main α-toxin, the action of which is reliant upon its ability to interact with host membranes. This enzyme, which contributes to *C. perfringens* virulence, provides information into CBM adherence to carbohydrates, carbohydrate hydrolysis via a catalytic module, enzyme-complex formation through a cohesin·dockerin interaction and allows for speculation into a possible adherence to bacterial cell walls. Other bacterial, modular glycoside hydrolases have a diverse arrangement of modules and thusly will perform a variety of different functions, not identical to the enzyme presented here, but they likely contain catalytic modules that are often associated with CBMs and other auxiliary modules similar to

NanJ, making this a model for the interpretation of other multi-modular bacterial carbohydrate-active enzymes.

Chapter 6: Discussion

The mutual efforts of genome sequencing projects, large-scale virulence factor screening, classical microbiology and biochemistry are providing evidence that a substantial amount of carbohydrate-active enzymes, such as glycoside hydrolases, are bacterial virulence factors. These enzymes are not well understood with a large portion of them being uncharacterized beyond the point of DNA sequencing. These glycoside hydrolases are important and in many cases there is a lack of knowledge as to the function, cellular location, mechanism of hydrolysis, functions of ancillary modules and overall structures of these large, often multi-modular enzymes. We chose *C. perfringens* and *S. pneumoniae* as model systems to study these enzymes due to their large complements of glycoside hydrolases, many of which are known virulence factors. The objectives of this study were formulated to probe the key features of glycoside hydrolases from these two kinds of bacteria and their carbohydrate-hydrolysis, modularity and overall glycoside hydrolase structure.

C. perfringens is a prevalent human and animal pathogen that produces a collection of glycoside hydrolases that are involved in the breakdown of the complex eukaryotic glycans that it encounters in its gastrointestinal tract niche. There are 88 predicted carbohydrate-processing enzymes produced by *C. perfringens* ATCC 13124. This includes 55 glycoside hydrolases, 20 glycosyltransferases, 3 polysaccharide lyases and 10 carbohydrate esterases. The 55 putative and known GHs produced by *C. perfringens* are classified into 27 different families with a broad range of predicted specificities. Many of these specificities are sugars that are found in eukaryotic glycans. These GHs are known or putative sialidases, α-N-acetylglucosaminidase, α-*N*-acetylgalacotosaminidases, β-N-acetylglucosaminidase, hyaluronidase, β-hexosaminidases, α-L-fucosidases, β-glucosidases, β-galactosidases, and α-galactosidases, to name some examples. There are also likely many more GHs that have not yet been described which will likely increase the total number of *C. perfringens* GHs and perhaps even broaden the range of sugar specificities just mentioned. All of these enzymes encompass an efficient bacterial

system that is capable of breaking down complex eukaryotic glycans that are found on the surface of cells and in gastrointestinal mucins.

Similar to *C. perfringens* that inhabits the mucin-rich gastrointestinal tract, *S. pneumoniae* occupies the mucin-rich respiratory tract and encodes many carbohydrate-active enzymes that allow it to degrade the complex eukaryotic glycans. *S. pneumoniae* TIGR4 produces 78 carbohydrate-active enzymes including 41 glycoside hydrolases, 29 glycosyltransferases, 3 polysaccharide lyases, and 5 carbohydrate esterases. Among the 41 putative glycoside hydrolases produced by *S. pneumoniae,* there are 18 that are predicted to be virulence factors (Hava and Camilli, 2002; Polissi et al., 1998; Obert et al., 2006). These GHs have a broad range of different specificities and many of these specificities are sugars that are found in eukaryotic glycans such as sialidases, β-galactosidases, α-*l*-fucosidases, mannosidases, *N*-acetylglucosaminidase, β-*N*-acetylglucosaminidase, and pullulanases to name some examples.

Eukaryotic glycans have immense structural diversity which can be attributed to the huge variety of monosaccharides that are linked to one another and to other molecules through many different types of linkages. To efficiently degrade these complex carbohydrates and glycoconjugates, an arsenal of enzymes with varying specificities is required. The *C. perfringens* and *S. pneumoniae* GH-mediated breakdown of complex sugars into simpler sugars likely has multiple purposes. One purpose is for carbohydrate metabolism by the bacteria. The simple sugars are likely transported into the cell and used as carbohydrate energy increasing the bacteria's fitness. Other possible purposes for hydrolysis are to expose receptors for other enzymes and to remove glycan barriers allowing other toxins access to underlying tissues. In *S. pneumoniae* there are several host cell receptors that have been associated with pathogenesis including cell surface proteins. Adherence to and subsequent penetration through the extracellular matrix, the mucin coating, and the glycocalyx are essential for uncovering these host receptors and efficient adherence of *S. pneumoniae* on the cell surface (Hammerschmidt, 2006; Bergmann and Hammerschmidt, 2006). In both *S. pneumoniae* and *C. perfringens* the removal of terminal sialic acids from host glycans has been shown to potentiate the activity of other toxins. It has been

suggested in several studies that cleavage of terminal sialic acid by NanA, an *S. pneumoniae* sialidase, may reveal an unknown receptor on the surface of epithelial cells. Decreased binding of *S. pneumoniae* to human epithelial cells was demonstrated after addition of sialylated glycocongugates and binding to chinchilla tracheas was increased after sialidase treatment (Tong et al., 2002; Barthelson et al., 1998). Binding to epithelial cells was significantly reduced in a *nanA* mutant but binding was found to be restored by the activity of purified sialidase (King et al., 2006). Similarly, in *C. perfringens,* the hydrolysis of sialic acids from glycans increases the sensitivity of host cells to the activity of the α-toxin which is the major toxin of the bacteria and has phospholipase C and sphingomyelinase activity (Flores-Díaz et al., 2005). The breakdown of the glycans found in the extracellular matrix, mucins or the glycocalyx allow the bacteria and its toxins easier access to the tissues, potentiating their effects and helping the bacteria spread throughout the host.

The *S. pneumoniae* GHs have substantially less modularity than the *C. perfringens* GHs, the reason for this is not entirely understood (Figure 6 and Figure 7). It is common for *S. pneumoniae* GHs to consist of only a catalytic module. The decreased level of modularity is exemplified by the 15 and 45 putative carbohydrate-binding modules found in *S. pneumoniae* and *C. perfringens*, respectively. There are also considerably less ancillary modules represented in the *S. pneumoniae* GHs. There does not appear to be any putative cohesin- or dockerin-like modules found in the *S. pneumoniae* GHs, unlike the *C. perfringens* GHs. Also FIVAR modules, which are modules of unknown functions, are very common and appear frequently in *C. perfringens* but are very infrequently identified in *S. pneumoniae* GHs. The situation is similar for FN3 modules which are very common in *C. perfringens* GHS.

A ubiquitous mechanism for Gram-positive anchoring is the LPXTG motif of surface-proteins which is recognized by sortase, a transpeptidase, and covalently links them to the cell wall peptidoglycan. Other cell-anchoring methods are through lipoproteins and through the phosphorylcholine of the pneumococcal cell wall which is bound by surface proteins that possess a choline-binding domain. In addition to the predicted surface

proteins produced by *S. pneumoniae,* there are non-classical surface proteins that have been identified which lack Gram-positive signal peptides for secretion into the extracellular milieu and known cell-anchoring motifs. The localization of these secreted, membrane anchored proteins is confounding and they have received considerable attention not only for their inexplicable secretion and anchoring but also for their contribution to virulence. An example of these surface-proteins is a β-galactosidase *S. pneumoniae* virulence factor, called BgaC, which was found to be surface localized despite a lack of signal peptide and known anchoring motifs or domains. As of yet, the mechanism of secretion and anchoring of these proteins is unknown for pneumococci and other Gram-positive pathogens (Jeong et al., 2009).

The degree of modular diversity and the presence of signal peptides and cell-anchoring motifs in *S. pneumoniae* GHs make them intriguing. In this study we investigated the activity of a *S. pneumoniae* GH that has a signal peptide, an LPXTG motif for cell-anchoring and considerable modularity. Another GH from *S. pneumoniae* was also characterized in this study that does not have a signal peptide, known-anchoring motif or any ancillary modules. Both of these enzymes had activity on eukaryotic glycans, O- and N-glycans, respectively. While both of these studies provided insight into the structure and function of these very distinct enzymes, there still remain unanswered questions. An explanation for the discrepancies in modularity between GHs in *S. pneumoniae* is lacking. From a functional standpoint it seems logical for GHs to have ancillary modules that can tether the enzyme to carbohydrate surfaces, to other GHs, or to bacterial or host cells/surfaces. Also, a complete representation of the cellular localization of the GHs that do not possess the necessary units to be secreted or surface-bound is required. This is of primary interest because many of these GHs have been implicated in virulence and it seems unlikely that these enzymes are solely involved in microbial fitness. Furthermore, as these enzymes have homologs that are virulence factors in other bacterial pathogens this research is afforded considerable applicability beyond just *S. pneumoniae*.

The *C. perfringens* GHs, as mentioned, are considerably more modular than the *S. pneumoniae* enzymes and they contain many CBMs, especially CBMs from family 32.

The family 32 CBMs have specificities for galacto-configured sugars which are prevalent throughout the gastrointestinal tract glycans, especially in mucins. These CBMs, along with CBMs from different families, help keep the catalytic modules of the GHs in proximity to their glycan substrate. In addition to these CBM modules are many fibronectin type III (FN3) modules. The function of these FN3 modules remains unknown but we speculate that they might be involved in protein-protein interactions to recruit other GHs. Another possibility that is even more likely is that they bind the negatively charge techoic acids in the Gram-positive cell wall through an electrostatic interaction with the basic charged distal tip of the module. The protruding position of the FN3 module at the C-terminus of *C. perfringens* GHs suggests this exposure is important to its function, which could allow for contact with the bacterial cell wall. The effect of this putative molecular interaction would be to concentrate the enzyme activity to the surface of the *C. perfringens* cell. Interestingly, the GHs that have a C-terminal FN3 module do not have an LPXTG motif for cell wall anchoring, and many of the GHs that do not have an FN3 module have an LPXTG motif. It would appear that the enzymes do not have a redundancy in regards to cell-anchoring.

There are many other modules that have undescribed structures and/or functions. In this study, we sought to characterize the function and structure of some of these unknown modules. Modules with distant similarities to protein-protein interacting modules typical of cellulosomal cohesin and dockerin modules were identified in several *C. perfringens* GHs. The identification of cohesin and dockerin modules in non-cellulolytic bacteria was especially intriguing and provided evidence for formation of GH complex formation. The implications of this would mean an added level of organization of these GHs, some of which are believed to be virulence factors of this bacterium. In addition to ascribing functions to unknown modules, a complete structural model of the organization of a GH implicated in *C. perfringens* virulence was studied. The organization of the modules in this GH provided a glimpse into how it could coordinate simultaneous carbohydrate-adherence by CBMs, carbohydrate hydrolysis by the catalytic module, GH complex formation through cohesin·dockerin interaction and lastly, putative association of GH complexes to bacterial cell surfaces through FN3 modules. Despite the fact that not all of

the *C. perfringens* GHs have the ability to form complexes, these studies provided evidence that it is possible to have higher order structural organization of these enzymes.

The information provided by this and other works allows for the design of a model for GH organization in *C. perfringens* and *S. pneumoniae* (Figure 40). In *C. perfringens*, the GHs could be attached to the cell wall *via* their FN3 modules or through sortase-mediated attachment of the LPXTG motif, or some enzymes might not be anchored at all. The catalytic modules of the GHs would hydrolyze glycan surfaces, such as gastrointestinal mucin, and the CBM modules would attach the enzymes to the carbohydrate surface. Additional association would be caused by the ultra-tight protein-protein interactions of the cohesin and dockerin domains. The GHs that contain cohesin modules also contain FN3 modules ensuring attachment to the cell wall and recruitment of other dockerin containing enzymes. In *S. pneumoniae,* the GHs could be attached to the cell wall through their LPXTG motifs or by other, unknown methods. Other enzymes would not be anchored to the cell wall and would be able to move freely. Like in the *C. perfringens* GHs the CBMs could attach the enzymes to the carbohydrate surface and the catalytic modules could hydrolyze the glycan surface, such as the mucin lining the respiratory tract.

Both *S. pneumoniae* and *C. perfringens* are formidable human pathogens that inhabit glycan-rich niches. To properly colonize their human hosts they have both developed extensive carbohydrate-active enzymes to process the rich variety of glycans that they will encounter. While they both encode GHs that are highly similar, they also have distinct sets of enzymes that are unique from one another. This work provided completely new molecular level insight into the *S. pneumoniae* and *C. perfringens* host interactions using a multidisciplinary approach to studying carbohydrate-active enzymes and their components. The results of this research will aid in driving future studies of GHs from these bacteria and other organisms. Though our research is principally at a fundamental level, the long term implications at an applied level may contribute to the foundation of knowledge that could lead to the development of therapeutic treatments for the diseases associated with carbohydrate hydrolysis.

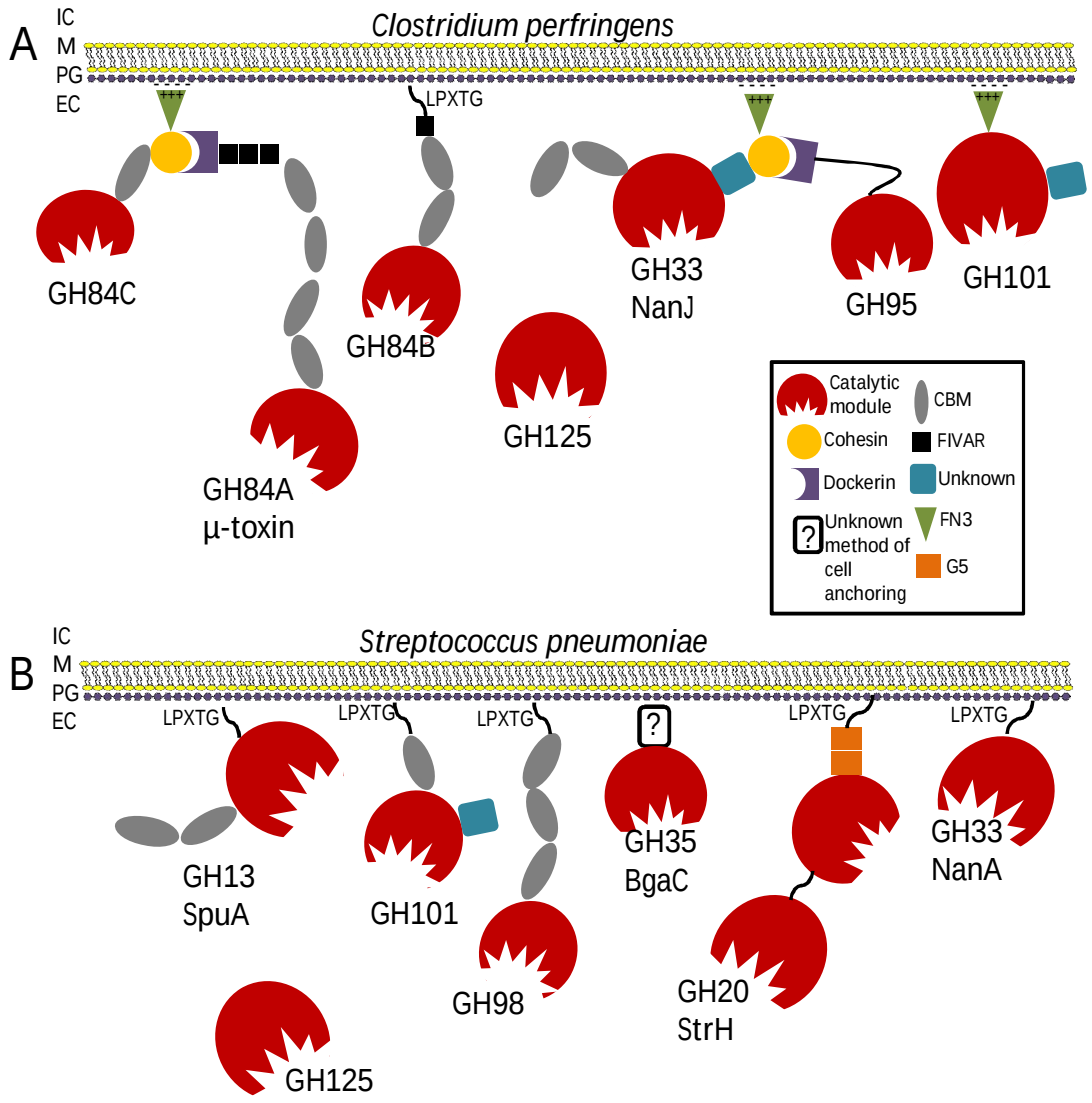


Figure 40: Model for GH organization in *C. perfringens* and *S. pneumoniae*.

A) *C. perfringens* glycoside hydrolase model B) *S. pneumoniae* glycoside hydrolase model. Glycoside hydrolase modules designated as shapes, as described in legend. IC, M, PG, and EC; intracellular, membrane, peptidoglycan, and extracellular, respectively. Adherence to the bacterial cell wall via LPXTG anchoring, putative FN3 anchoring, unknown method of anchoring or not cell anchored. Adherence to the cell wall ensures the enzymes are near the surface of the bacterial cell for optimal scavenging efficiency. Complex formation of GHs through cohesin·dockerin interaction. Catalytic modules and CBMs positioned for hydrolysis of heterogeneous glycan surface. CBMs maintain association to the carbohydrate surface.

Bibliography

Abbott, D. W., and Boraston, A. B. (2008). Structural Biology of Pectin Degradation by Enterobacteriaceae. Microbiol. Mol. Biol. Rev *72*, 301-316.

Adams, J. J., Gregg, K., Bayer, E. A., Boraston, A. B., and Smith, S. P. (2008). Structural basis of *Clostridium perfringens* toxin complex formation. Proc. Natl. Acad. Sci. U.S.A *105*, 12194-12199.

Adams, J. J., Pal, G., Jia, Z., and Smith, S. P. (2006). Mechanism of bacterial cell-surface attachment revealed by the structure of cellulosomal type II cohesin-dockerin complex. Proc. Natl. Acad. Sci. U.S.A *103*, 305-310.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol *215*, 403-410.

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

Ashida, H., Maki, R., Ozawa, H., Tani, Y., Kiyohara, M., Fujita, M., Imamura, A., Ishida, H., Kiso, M., and Yamamoto, K. (2008). Characterization of two different endo-alpha-N-acetylgalactosaminidases from probiotic and pathogenic enterobacteria, *Bifidobacterium longum* and *Clostridium perfringens*. Glycobiology *18*, 727-734.

Barak, Y., Handelsman, T., Nakar, D., Mechaly, A., Lamed, R., Shoham, Y., and Bayer, E. A. (2005). Matching fusion protein systems for affinity analysis of two interacting families of proteins: the cohesin-dockerin interaction. J. Mol. Recognit *18*, 491-501.

Barthelson, R., Mobasseri, A., Zopf, D., and Simon, P. (1998). Adherence of *Streptococcus pneumoniae* to respiratory epithelial cells is inhibited by sialylated oligosaccharides. Infect. Immun *66*, 1439-1444.

Bayer, E. A., Belaich, J.-P., Shoham, Y., and Lamed, R. (2004). The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. Annu. Rev. Microbiol *58*, 521-554.

Bayer, E. A., Coutinho, P. M., and Henrissat, B. (1999). Cellulosome-like sequences in *Archaeoglobus fulgidus*: an enigmatic vestige of cohesin and dockerin domains. FEBS Lett *463*, 277-280.

Bencharit, S., Cui, C. B., Siddiqui, A., Howard-Williams, E. L., Sondek, J., Zuobi-Hasona, K., and Aukhil, I. (2007). Structural insights into fibronectin type III domain-mediated signaling. J. Mol. Biol *367*, 303-309.

Bergmann, S., and Hammerschmidt, S. (2006). Versatility of pneumococcal surface proteins. Microbiology (Reading, Engl.) *152*, 295-303.

Bogaert, D., De Groot, R., and Hermans, P. W. M. (2004). *Streptococcus pneumoniae* colonisation: the key to pneumococcal disease. Lancet Infect Dis *4*, 144-154.

Boraston, A. B., Bolam, D. N., Gilbert, H. J., and Davies, G. J. (2004). Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem. J *382*, 769-781.

Boraston, A. B., Creagh, A. L., Alam, M. M., Kormos, J. M., Tomme, P., Haynes, C. A., Warren, R. A., and Kilburn, D. G. (2001). Binding specificity and thermodynamics of a family 9 carbohydrate-binding module from *Thermotoga maritima* xylanase 10A. Biochemistry *40*, 6240-6247.

Boraston, A. B., Ficko-Blean, E., and Healey, M. (2007). Carbohydrate recognition by a large sialidase toxin from *Clostridium perfringens*. Biochemistry *46*, 11352-11360.

Boraston, A. B., Wang, D., and Burke, R. D. (2006). Blood group antigen recognition by a *Streptococcus pneumoniae* virulence factor. J. Biol. Chem *281*, 35263-35271.

Bork, P., and Doolittle, R. F. (1992). Proposed acquisition of an animal protein domain by bacteria. Proc. Natl. Acad. Sci. U.S.A *89*, 8990-8994.

Brandts, J. F., and Lin, L. N. (1990). Study of strong to ultratight protein interactions using differential scanning calorimetry. Biochemistry *29*, 6927-6940.

Brünger, A. T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature *355*, 472-475.

Caines, M. E. C., Zhu, H., Vuckovic, M., Willis, L. M., Withers, S. G., Wakarchuk, W. W., and Strynadka, N. C. J. (2008). The structural basis for T-antigen hydrolysis by *Streptococcus pneumoniae*: a target for structure-based vaccine design. J. Biol. Chem *283*, 31279-31283.

Campbell, J. A., Davies, G. J., Bulone, V., and Henrissat, B. (1997). A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. Biochem. J *326 (Pt 3)*, 929-939.

Canard, B., Garnier, T., Saint-Joanis, B., and Cole, S. T. (1994). Molecular genetic analysis of the nagH gene encoding a hyaluronidase of *Clostridium perfringens*. Mol. Gen. Genet *243*, 215-224.

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res *37*, D233-238.

Carvalho, A. L., Dias, F. M. V., Nagy, T., Prates, J. A. M., Proctor, M. R., Smith, N., Bayer, E. A., Davies, G. J., Ferreira, L. M. A., Romão, M. J., et al. (2007). Evidence for a dual binding mode of dockerin modules to cohesins. Proc. Natl. Acad. Sci. U.S.A *104*, 3089-3094.

Carvalho, A. L., Dias, F. M. V., Prates, J. A. M., Nagy, T., Gilbert, H. J., Davies, G. J., Ferreira, L. M. A., Romão, M. J., and Fontes, C. M. G. A. (2003). Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. Proc. Natl. Acad. Sci. U.S.A *100*, 13809-13814.

Charlwood, J., Birrell, H., and Camilleri, P. (1998). Efficient carbohydrate release, purification, and derivatization. Anal. Biochem *262*, 197-200.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr *66*, 12-21.

Chiarezza, M., Lyras, D., Pidot, S. J., Flores-Díaz, M., Awad, M. M., Kennedy, C. L., Cordner, L. M., Phumoonna, T., Poon, R., Hughes, M. L., et al. (2009). The NanI and NanJ sialidases of *Clostridium perfringens* are not essential for virulence. Infect. Immun *77*, 4421-4428.

Chitayat, S., Adams, J. J., and Smith, S. P. (2007a). NMR assignment of backbone and side chain resonances for a dockerin-containing C-terminal fragment of the putative mu-toxin from *Clostridium perfringens*. Biomol NMR Assign *1*, 13-15.

Chitayat, S., Ficko-Blean, E., Adams, J. J., Gregg, K., Boraston, A. B., and Smith, S. P. (2007b). NMR assignment of backbone and side chain resonances for a putative protein-protein interaction module from a family 84 glycoside hydrolase of *Clostridium perfringens*. Biomol NMR Assign *1*, 7-9.

Chitayat, S., Gregg, K., Adams, J. J., Ficko-Blean, E., Bayer, E. A., Boraston, A. B., and Smith, S. P. (2008). Three-dimensional structure of a putative non-cellulosomal cohesin module from a *Clostridium perfringens* family 84 glycoside hydrolase. J. Mol. Biol *375*, 20-28.

Collaborative Computational Project, Number 4 (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr *50*, 760-763.

Connaris, H., Crocker, P. R., and Taylor, G. L. (2009). Enhancing the receptor affinity of the sialic acid-binding domain of *Vibrio cholerae* sialidase through multivalency. J. Biol. Chem *284*, 7339-7351.

Correia, M. A. S., Prates, J. A. M., Brás, J., Fontes, C. M. G. A., Newman, J. A., Lewis, R. J., Gilbert, H. J., and Flint, J. E. (2008). Crystal Structure of a Cellulosomal Family 3 Carbohydrate Esterase from *Clostridium thermocellum* Provides Insights

into the Mechanism of Substrate Recognition. Journal of Molecular Biology *379*, 64-72.

Coutinho, P. M., Deleury, E., Davies, G. J., and Henrissat, B. (2003). An evolving hierarchical family classification for glycosyltransferases. J. Mol. Biol *328*, 307-317.

Dalia, A. B., Standish, A. J., and Weiser, J. N. (2010). Three surface exoglycosidases from Streptococcus pneumoniae, NanA, BgaA, and StrH, promote resistance to opsonophagocytic killing by human neutrophils. Infect. Immun *78*, 2108-2116.

Davies, G. J., Gloster, T. M., and Henrissat, B. (2005). Recent structural insights into the expanding world of carbohydrate-active enzymes. Curr. Opin. Struct. Biol *15*, 637-645.

Davies, G. J., Wilson, K. S., and Henrissat, B. (1997). Nomenclature for sugar-binding subsites in glycosyl hydrolases. Biochem. J *321 (Pt 2)*, 557-559.

DeLano, W. (2002). The Pymol Molecular Graphics System (Palo Alto, CA: DeLano Scientific).

Dereeper, A., Audic, S., Claverie, J.-M., and Blanc, G. (2010). BLAST-EXPLORER helps you building datasets for phylogenetic analysis. BMC Evol. Biol *10*, 8.

Doi, R. H., and Kosugi, A. (2004). Cellulosomes: plant-cell-wall-degrading enzyme complexes. Nat. Rev. Microbiol *2*, 541-551.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr *60*, 2126-2132.

Evangelista, R. A., Guttman, A., and Chen, F. T. (1996). Acid-catalyzed reductive amination of aldoses with 8-aminopyrene-1,3,6-trisulfonate. Electrophoresis *17*, 347-351.

Ferguson, M. A. J., and Williams, A. F. (1988). Cell-Surface Anchoring of Proteins Via Glycosyl-Phosphatidylinositol Structures. Annu. Rev. Biochem. *57*, 285-320.

Ficko-Blean, E., and Boraston, A. B. (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*, 834-836.

Ficko-Blean, E., Gregg, K. J., Adams, J. J., Hehemann, J.-H., Czjzek, M., Smith, S. P., and Boraston, A. B. (2009). Portrait of an enzyme, a complete structural analysis of a multimodular β-N-acetylglucosaminidase from *Clostridium perfringens*. J. Biol. Chem *284*, 9876-9884.

Flores-Díaz, M., Alape-Girón, A., Clark, G., Catimel, B., Hirabayashi, Y., Nice, E., Gutiérrez, J.-M., Titball, R., and Thelestam, M. (2005). A cellular deficiency of gangliosides causes hypersensitivity to *Clostridium perfringens* phospholipase C. J. Biol. Chem *280*, 26680-26689.

Fontes, C. M. G. A., and Gilbert, H. J. (2010). Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. Annu. Rev. Biochem. *79*, 655-681.

Fournet, F, G., A.F. Small Angle Scattering of X-rays (New York: Wiley Interscience).

Franke, D., and Svergun, D. I. (2009). DAMMIF, a program for rapid ab-initi shape determination in small-angle scattering. J Appl Crystallogr *42*, 342-346.

Fujita, K., Oura, F., Nagamine, N., Katayama, T., Hiratake, J., Sakata, K., Kumagai, H., and Yamamoto, K. (2005). Identification and molecular cloning of a novel glycoside hydrolase family of core 1 type O-glycan-specific endo-alpha-N-acetylgalactosaminidase from *Bifidobacterium longum*. J. Biol. Chem *280*, 37415-37422.

Galen, J. E., Ketley, J. M., Fasano, A., Richardson, S. H., Wasserman, S. S., and Kaper, J. B. (1992). Role of *Vibrio cholerae* neuraminidase in the function of cholera toxin. Infect. Immun *60*, 406-415.

Garron, M.-L., and Cygler, M. (2010). Structural and mechanistic classification of uronic acid-containing polysaccharide lyases. Glycobiology *20*, 1547-1573.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., and Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res *31*, 3784-3788.

Gilbert, H. J. (2007). Cellulosomes: microbial nanomachines that display plasticity in quaternary structure. Mol. Microbiol *63*, 1568-1576.

Goda, H. M., Ushigusa, K., Ito, H., Okino, N., Narimatsu, H., and Ito, M. (2008). Molecular cloning, expression, and characterization of a novel endo-alpha-N-acetylgalactosaminidase from *Enterococcus faecalis*. Biochem. Biophys. Res. Commun *375*, 441-446.

Gregg, K. J., Zandberg, W. F., Hehemann, J.-H., Whitworth, G. E., Deng, L., Vocadlo, D. J., and Boraston, A. B. (2011). Analysis of a New Family of Widely Distributed Metal-independent α-Mannosidases Provides Unique Insight into the Processing of N-Linked Glycans. J. Biol. Chem *286*, 15586-15596.

Guttman, A., Chen, F. T., Evangelista, R. A., and Cooke, N. (1996). High-resolution capillary gel electrophoresis of reducing oligosaccharides labeled with 1-aminopyrene-3,6,8-trisulfonate. Anal. Biochem *233*, 234-242.

Haimovitz, R., Barak, Y., Morag, E., Voronov-Goldman, M., Shoham, Y., Lamed, R., and Bayer, E. A. (2008). Cohesin-dockerin microarray: Diverse specificities between two complementary families of interacting protein modules. Proteomics *8*, 968-979.

Hakomori, S. (2003). Structure, organization, and function of glycosphingolipids in membrane. Curr. Opin. Hematol *10*, 16-24.

Hammerschmidt, S. (2006). Adherence molecules of pathogenic pneumococci. Curr. Opin. Microbiol *9*, 12-20.

Hava, D. L., and Camilli, A. (2002). Large-scale identification of serotype 4 *Streptococcus pneumoniae* virulence factors. Mol. Microbiol *45*, 1389-1406.

Henrissat, B. (1991). A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J *280 (Pt 2)*, 309-316.

Henrissat, B., and Davies, G. (1997). Structural and sequence-based classification of glycoside hydrolases. Curr. Opin. Struct. Biol *7*, 637-644.

Higgins, M. A., Whitworth, G. E., El Warry, N., Randriantsoa, M., Samain, E., Burke, R. D., Vocadlo, D. J., and Boraston, A. B. (2009). Differential recognition and hydrolysis of host carbohydrate antigens by *Streptococcus pneumoniae* family 98 glycoside hydrolases. J. Biol. Chem *284*, 26161-26173.

Homer, K. A., Roberts, G., Byers, H. L., Tarelli, E., Whiley, R. A., Philpott-Howard, J., and Beighton, D. (2001). Mannosidase production by viridans group streptococci. J. Clin. Microbiol *39*, 995-1001.

Hutchison, C. A., Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P., and Smith, M. (1978). Mutagenesis at a specific position in a DNA sequence. Journal of Biological Chemistry *253*, 6551 -6560.

Van Immerseel, F., De Buck, J., Pasmans, F., Huyghebaert, G., Haesebrouck, F., and Ducatelle, R. (2004). *Clostridium perfringens* in poultry: an emerging threat for animal and public health. Avian Pathol *33*, 537-549.

Jeong, J. K., Kwon, O., Lee, Y. M., Oh, D.-B., Lee, J. M., Kim, S., Kim, E.-H., Le, T. N., Rhee, D.-K., and Kang, H. A. (2009). Characterization of the *Streptococcus pneumoniae* BgaC protein as a novel surface beta-galactosidase with specific hydrolysis activity for the Galbeta1-3GlcNAc moiety of oligosaccharides. J. Bacteriol *191*, 3011-3023.

Kadioglu, A., and Andrew, P. W. (2004). The innate immune response to pneumococcal lung infection: the untold story. Trends Immunol *25*, 143-149.

Karaveg, K., Siriwardena, A., Tempel, W., Liu, Z.-J., Glushka, J., Wang, B.-C., and Moremen, K. W. (2005). Mechanism of class 1 (glycosylhydrolase family 47) α-

mannosidases involved in N-glycan processing and endoplasmic reticulum quality control. J. Biol. Chem *280*, 16197-16207.

Kelley, L. A., and Sternberg, M. J. E. (2009). Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protocols *4*, 363-371.

King, S. J. (2010). Pneumococcal modification of host sugars: a major contributor to colonization of the human airway? Mol Oral Microbiol *25*, 15-24.

King, S. J., Hippe, K. R., and Weiser, J. N. (2006). Deglycosylation of human glycoconjugates by the sequential activities of exoglycosidases expressed by *Streptococcus pneumoniae.* Mol. Microbiol *59*, 961-974.

Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J., and Svergun, D. I. (2003). PRIMUS : a Windows PC-based system for small-angle scattering data analysis. J Appl Crystallogr *36*, 1277-1282.

Koutsioulis, D., Landry, D., and Guthrie, E. P. (2008). Novel endo-alpha-N-acetylgalactosaminidases with broader substrate specificity. Glycobiology *18*, 799-805.

Kruse, S., Kleineidam, R. G., Roggentin, P., and Schauer, R. (1996). Expression and purification of a recombinant "small" sialidase from *Clostridium perfringens* A99. Protein Expr. Purif *7*, 415-422.

Lairson, L. L., Henrissat, B., Davies, G. J., and Withers, S. G. (2008). Glycosyltransferases: Structures, Functions, and Mechanisms. Annu. Rev. Biochem. *77*, 521-555.

Lammerts van Bueren, A., Ficko-Blean, E., Pluvinage, B., Hehemann, J.-H., Higgins, M. A., Deng, L., Ogunniyi, A. D., Stroeher, U. H., El Warry, N., Burke, R. D., et al. (2011). The conformation and function of a multimodular glycogen-degrading pneumococcal virulence factor. Structure *19*, 640-651.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr *26*, 283-291.

Leahy, D. J. (1997). Implications of atomic-resolution structures for cell adhesion. Annu. Rev. Cell Dev. Biol *13*, 363-393.

Legler, G., and Jülich, E. (1984). Synthesis of 5-amino-5-deoxy-D-mannopyranose and 1,5-dideoxy-1,5-imino-D-mannitol, and inhibition of alpha- and beta-D-mannosidases. Carbohydr. Res *128*, 61-72.

Li, S., Kelly, S. J., Lamani, E., Ferraroni, M., and Jedrzejas, M. J. (2000). Structural basis of hyaluronan degradation by *Streptococcus pneumoniae* hyaluronate lyase. EMBO J *19*, 1228-1240.

Liu, Y., and Sturtevant, J. M. (1995). Significant discrepancies between van't Hoff and calorimetric enthalpies. II. Protein Sci *4*, 2559-2561.

Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P. M., and Henrissat, B. (2010). A hierarchical classification of polysaccharide lyases for glycogenomics. Biochem. J *432*, 437-444.

Macauley, M. S., Whitworth, G. E., Debowski, A. W., Chin, D., and Vocadlo, D. J. (2005). O-GlcNAcase uses substrate-assisted catalysis: kinetic analysis and development of highly selective mechanism-inspired inhibitors. J. Biol. Chem *280*, 25313-25322.

MacGregor, E. A., Janecek, S., and Svensson, B. (2001). Relationship of sequence and structure to specificity in the alpha-amylase family of enzymes. Biochim. Biophys. Acta *1546*, 1-20.

Mach, H., Middaugh, C. R., and Lewis, R. V. (1992). Statistical determination of the average values of the extinction coefficients of tryptophan and tyrosine in native proteins. Anal. Biochem *200*, 74-80.

Manco, S., Hernon, F., Yesilkaya, H., Paton, J. C., Andrew, P. W., and Kadioglu, A. (2006). Pneumococcal neuraminidases A and B both have essential roles during infection of the respiratory tract and sepsis. Infect. Immun *74*, 4014-4020.

Marion, C., Limoli, D. H., Bobulsky, G. S., Abraham, J. L., Burnaugh, A. M., and King, S. J. (2009). Identification of a pneumococcal glycosidase that modifies O-linked glycans. Infect. Immun *77*, 1389-1396.

McClane, B. (2001). *Clostridium perfringens*. In In Food Microbiology: Fundamentals and Frontiers (Washington DC: ASM Press), pp. 351-372.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007). Phaser crystallographic software. J Appl Crystallogr *40*, 658-674.

Mizuno, M., Tonozuka, T., Suzuki, S., Uotsu-Tomita, R., Kamitori, S., Nishikawa, A., and Sakano, Y. (2004). Structural insights into substrate specificity and function of glucodextranase. J. Biol. Chem *279*, 10575-10583.

Moustafa, I., Connaris, H., Taylor, M., Zaitsev, V., Wilson, J. C., Kiefel, M. J., von Itzstein, M., and Taylor, G. (2004). Sialic acid recognition by *Vibrio cholerae* neuraminidase. J. Biol. Chem *279*, 40819-40826.

Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997). Refinement of Macromolecular Structures by the Maximum-Likelihood Method. Acta Crystallogr D Biol Crystallogr *53*, 240-255.

Myers, G. S. A., Rasko, D. A., Cheung, J. K., Ravel, J., Seshadri, R., DeBoy, R. T., Ren, Q., Varga, J., Awad, M. M., Brinkac, L. M., et al. (2006). Skewed genomic variability in strains of the toxigenic bacterial pathogen, *Clostridium perfringens*. Genome Res *16*, 1031-1040.

Newstead, S. L., Potter, J. A., Wilson, J. C., Xu, G., Chien, C.-H., Watts, A. G., Withers, S. G., and Taylor, G. L. (2008). The structure of *Clostridium perfringens* NanI sialidase and its catalytic intermediates. J. Biol. Chem *283*, 9080-9088.

Obert, C., Sublett, J., Kaushal, D., Hinojosa, E., Barton, T., Tuomanen, E. I., and Orihuela, C. J. (2006). Identification of a Candidate *Streptococcus pneumoniae* core genome and regions of diversity correlated with invasive pneumococcal disease. Infect. Immun *74*, 4766-4777.

Orihuela, C. J., Gao, G., Francis, K. P., Yu, J., and Tuomanen, E. I. (2004). Tissue-specific contributions of pneumococcal virulence factors to pathogenesis. J. Infect. Dis *190*, 1661-1669.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collection in oscillation mode. In Methods in Enzymology, Volume 276: Macromolecular Crystallography, Part A (Academic Press), pp. 307-326.

Park, C., Meng, L., Stanton, L. H., Collins, R. E., Mast, S. W., Yi, X., Strachan, H., and Moremen, K. W. (2005). Characterization of a human core-specific lysosomal α-1,6-mannosidase involved in N-glycan catabolism. J. Biol. Chem *280*, 37204-37216.

Parker, D., Soong, G., Planet, P., Brower, J., Ratner, A. J., and Prince, A. (2009). The NanA neuraminidase of *Streptococcus pneumoniae* is involved in biofilm formation. Infect. Immun *77*, 3722-3730.

Paulick, M. G., and Bertozzi, C. R. (2008). The Glycosylphosphatidylinositol Anchor: A Complex Membrane-Anchoring Structure for Proteins. Biochemistry *47*, 6991-7000.

Peer, A., Smith, S. P., Bayer, E. A., Lamed, R., and Borovok, I. (2009). Noncellulosomal cohesin- and dockerin-like modules in the three domains of life. FEMS Microbiol. Lett *291*, 1-16.

Perez-Vilar, J., and Hill, R. L. (1999). The structure and assembly of secreted mucins. J. Biol. Chem *274*, 31751-31754.

Perrakis, A., Morris, R., and Lamzin, V. S. (1999). Automated protein model building combined with iterative structure refinement. Nat. Struct. Biol *6*, 458-463.

Petit, L., Gibert, M., and Popoff, M. R. (1999). *Clostridium perfringens*: toxinotype and genotype. Trends Microbiol *7*, 104-110.

Petit, L., Gibert, M., Gillet, D., Laurent-Winter, C., Boquet, P., and Popoff, M. R. (1997). *Clostridium perfringens* epsilon-toxin acts on MDCK cells by forming a large membrane complex. J. Bacteriol *179*, 6480-6487.

Polissi, A., Pontiggia, A., Feger, G., Altieri, M., Mottl, H., Ferrari, L., and Simon, D. (1998). Large-scale identification of virulence genes from *Streptococcus pneumoniae*. Infect. Immun *66*, 5620-5629.

van der Poll, T., and Opal, S. M. (2009). Pathogenesis, treatment, and prevention of pneumococcal pneumonia. Lancet *374*, 1543-1556.

Powell, H. R. (1999). The Rossmann Fourier autoindexing algorithm in MOSFLM. Acta Crystallogr D Biol Crystallogr *55*, 1690-1695.

Pratt, M. R., and Bertozzi, C. R. (2005). Synthetic glycopeptides and glycoproteins as tools for biology. Chem Soc Rev *34*, 58-68.

Pries, A. R., Secomb, T. W., and Gaehtgens, P. (2000). The endothelial surface layer. Pflugers Archiv European Journal of Physiology *440*, 653-666.

Putnam, C. D., Hammel, M., Hura, G. L., and Tainer, J. A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q. Rev. Biophys *40*, 191-285.

Rood, J. I., and Cole, S. T. (1991). Molecular genetics and pathogenesis of *Clostridium perfringens*. Microbiol. Rev *55*, 621-648.

Scaman, C. H., Lipari, F., and Herscovics, A. (1996). A spectrophotometric assay for alpha-mannosidase activity. Glycobiology *6*, 265-270.

Shatursky, O., Bayles, R., Rogers, M., Jost, B. H., Songer, J. G., and Tweten, R. K. (2000). *Clostridium perfringens* beta-toxin forms potential-dependent, cation-selective channels in lipid bilayers. Infect. Immun *68*, 5546-5551.

Shelburne, S. A., Davenport, M. T., Keith, D. B., and Musser, J. M. (2008). The role of complex carbohydrate catabolism in the pathogenesis of invasive streptococci. Trends Microbiol *16*, 318-325.

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

Shoham, Y., Lamed, R., and Bayer, E. A. (1999). The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides. Trends Microbiol *7*, 275-281.

Smith, L. D. (1979). Virulence factors of *Clostridium perfringens*. Rev. Infect. Dis *1*, 254-262.

Songer, J. G. (1996). Clostridial enteric diseases of domestic animals. Clin Microbiol Rev *9*, 216-234.

Sonnenburg, J. L., Xu, J., Leip, D. D., Chen, C.-H., Westover, B. P., Weatherford, J., Buhler, J. D., and Gordon, J. I. (2005). Glycan foraging in vivo by an intestine-adapted bacterial symbiont. Science *307*, 1955-1959.

Soong, G., Muir, A., Gomez, M. I., Waks, J., Reddy, B., Planet, P., Singh, P. K., Kaneko, Y., Kanetko, Y., Wolfgang, M. C., et al. (2006). Bacterial neuraminidase facilitates mucosal infection by participating in biofilm production. J. Clin. Invest *116*, 2297-2305.

Strous, G. J., and Dekker, J. (1992). Mucin-Type Glycoproteins. Critical Reviews in Biochemistry and Molecular Biology *27*, 57-92.

Stubbs, K. A., Zhang, N., and Vocadlo, D. J. (2006). A divergent synthesis of 2-acyl derivatives of PUGNAc yields selective inhibitors of O-GlcNAcase. Org. Biomol. Chem. *4*, 839.

Suits, M. D. L., Zhu, Y., Taylor, E. J., Walton, J., Zechel, D. L., Gilbert, H. J., and Davies, G. J. (2010). Structure and kinetic investigation of *Streptococcus pyogenes* family GH38 alpha-mannosidase. PLoS ONE *5*, e9006.

Suzuki, R., Katayama, T., Kitaoka, M., Kumagai, H., Wakagi, T., Shoun, H., Ashida, H., Yamamoto, K., and Fushinobu, S. (2009). Crystallographic and Mutational Analyses of Substrate Recognition of Endo-α-N-acetylgalactosaminidase from *Bifidobacterium longum*. Journal of Biochemistry *146*, 389 -398.

Svergun, D. I. (1992). Determination of the regularization parameter in indirect-transform methods using perceptual criteria. J Appl Crystallogr *25*, 495-503.

Svergun, D., Barberato, C., and Koch, M. H. J. (1995). CRYSOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. J Appl Crystallogr *28*, 768-773.

Terwilliger, T. C. (2003). SOLVE and RESOLVE: automated structure solution and density modification. Meth. Enzymol *374*, 22-37.

Terwisscha van Scheltinga, A. C., Armand, S., Kalk, K. H., Isogai, A., Henrissat, B., and Dijkstra, B. W. (1995). Stereochemistry of chitin hydrolysis by a plant chitinase/lysozyme and X-ray structure of a complex with allosamidin: evidence for substrate assisted catalysis. Biochemistry *34*, 15619-15623.

Tettelin, H., Nelson, K. E., Paulsen, I. T., Eisen, J. A., Read, T. D., Peterson, S., Heidelberg, J., DeBoy, R. T., Haft, D. H., Dodson, R. J., et al. (2001). Complete

genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science *293*, 498-506.

Thobhani, S., Ember, B., Siriwardena, A., and Boons, G.-J. (2003). Multivalency and the mode of action of bacterial sialidases. J. Am. Chem. Soc *125*, 7154-7155.

Titball, R. W., Naylor, C. E., and Basak, A. K. (1999). The *Clostridium perfringens* alpha-toxin. Anaerobe *5*, 51-64.

Tong, H. H., Blue, L. E., James, M. A., and DeMaria, T. F. (2000). Evaluation of the virulence of a *Streptococcus pneumoniae* neuraminidase-deficient mutant in nasopharyngeal colonization and development of otitis media in the chinchilla model. Infect. Immun *68*, 921-924.

Tong, H. H., Liu, X., Chen, Y., James, M., and Demaria, T. (2002). Effect of neuraminidase on receptor-mediated adherence of *Streptococcus pneumoniae* to chinchilla tracheal epithelium. Acta Otolaryngol *122*, 413-419.

Traving, C., Schauer, R., and Roggentin, P. (1994). Gene structure of the "large" sialidase isoenzyme from *Clostridium perfringens* A99 and its relationship with other clostridial nanH proteins. Glycoconj. J *11*, 141-151.

Umemoto, J., Bhavanandan, V. P., and Davidson, E. A. (1977). Purification and properties of an endo-alpha-N-acetyl-D-galactosaminidase from *Diplococcus pneumoniae*. Journal of Biological Chemistry *252*, 8609 -8614.

Vagin, A., and Teplyakov, A. (2010). Molecular replacement with MOLREP. Acta Crystallogr. D Biol. Crystallogr *66*, 22-25.

Vaguine, A. A., Richelle, J., and Wodak, S. J. (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*, 191-205.

Varki, A. (1993). Biological roles of oligosaccharides: all of the theories are correct. Glycobiology *3*, 97 -130.

Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., and Etzler, M. E. (2009). Essentials of Glycobiology Second. (Cold Spring Harbour, NY: Cold Spring Harbour Laboratory Press).

Vimr, E. R., Kalivoda, K. A., Deszo, E. L., and Steenbergen, S. M. (2004). Diversity of microbial sialic acid metabolism. Microbiol. Mol. Biol. Rev *68*, 132-153.

Volkov, V. V., and Svergun, D. I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. J Appl Crystallogr *36*, 860-864.

Wang, W., and Malcolm, B. A. (1999). Two-stage PCR protocol allowing introduction of multiple mutations, deletions and insertions using QuikChange Site-Directed Mutagenesis. BioTechniques *26*, 680-682.

Whitworth, G. E., Zandberg, W. F., Clark, T., and Vocadlo, D. J. (2010). Mammalian Notch is modified by D-Xyl-alpha1-3-D-Xyl-alpha1-3-D-Glc-beta1-O-Ser: implementation of a method to study O-glucosylation. Glycobiology *20*, 287-299.

Willis, L. M., Zhang, R., Reid, A., Withers, S. G., and Wakarchuk, W. W. (2009). Mechanistic investigation of the endo-alpha-N-acetylgalactosaminidase from *Streptococcus pneumoniae* R6. Biochemistry *48*, 10334-10341.

Wolfenden, R., Lu, X., and Young, G. (1998). Spontaneous Hydrolysis of Glycosides. Journal of the American Chemical Society *120*, 6814-6815.

Yu, S.-Y., Khoo, K.-H., Yang, Z., Herp, A., and Wu, A. M. (2008). Glycomic mapping of O- and N-linked glycans from major rat sublingual mucin. Glycoconj. J *25*, 199-212.

Zechel, D. L., and Withers, S. G. (2000). Glycosidase mechanisms: anatomy of a finely tuned catalyst. Acc. Chem. Res *33*, 11-18.

Zhong, W., Kuntz, D. A., Ember, B., Singh, H., Moremen, K. W., Rose, D. R., and Boons, G.-J. (2008). Probing the substrate specificity of Golgi alpha-mannosidase II by use of synthetic oligosaccharides and a catalytic nucleophile mutant. J. Am. Chem. Soc *130*, 8975-8983.

Zhu, Y., Suits, M. D. L., Thompson, A. J., Chavan, S., Dinev, Z., Dumon, C., Smith, N., Moremen, K. W., Xiang, Y., Siriwardena, A., et al. (2010). Mechanistic insights into a Ca2+-dependent family of alpha-mannosidases in a human gut symbiont. Nat. Chem. Biol. *6*, 125-132.

Appendix A

[1]Capillary Electrophoresis experiments were performed by Garrett E. Whitworth and Dr. David Vocadlo and NMR experiments were performed by Wesley F. Zandberg and Dr. David Vocadlo

[1]Department of Chemistry, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada.

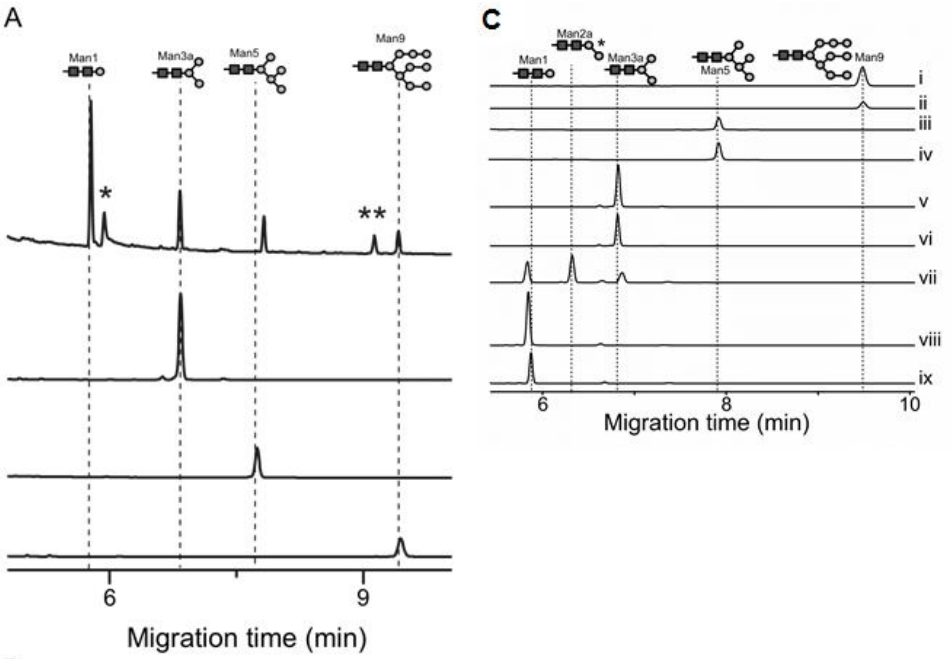
Capillary Electrophoresis (Gregg et al., 2011). The oligosaccharide standards Man9, Man3a and Man1 were obtained from V-Labs Inc., while Man5 was purchased from Glyko/Prozyme. All standards were dissolved in H_2O and stored at -20°C until use. α-(1,2)/(1,3)-mannosidase was acquired from New England BioLabs. 8-aminopyrene-1,3,6-trisulfonic acid (APTS) was purchased as its sodium salt from Beckman-Coulter while $NaBH_3CN$, of the highest available purity, was purchased from Fluka. 200 mg Hypercarb solid-phase extraction cartridges were from Thermo and conditioned with 1 M NaOH (3 mL), H_2O (3 mL), 1 M formic acid (3 mL), 50% (v/v) CH_3CN + 0.1% TFA (3mL) and 5% (v/v) CH_3CN + 0.1% TFA (6 mL) before use. 10 µL Hypercarb-containing SpinTips (Thermo) were conditioned with 50% (v/v) CH_3CN + 25 mM TFA (50 µL) and H_2O (3 x 50 µL) prior to use. All extraction procedures used Milli-Q-purified H_2O (18 MΩ/cm) and HPLC-grade CH_3CN (Sigma).

Lyophilized oligosaccharide standards, Man-9, Man-5, Man-3a, and Man-1 were labelled with APTS by reductive amination exactly as previously described (Evangelista et al., 1996). Excess labelling reagents and impurities were removed from APTS-labelled standards by fluorophore-assisted carbohydrate electrophoresis as described previously (Whitworth et al., 2010) with minor modifications. Briefly, labelled material was resolved by gel electrophoresis at a constant voltage (200 V) on 0.75 mm thick, 20% polyacrylamide gels, at 4°C in the dark. Bands corresponding to N-glycan standards, or their mannosidase-digested products, were excised, transferred to separate 15 mL centrifuge tubes, mixed with 0.4 mL H_2O, and sonicated in a water bath sonicator for 2

hours. The H_2O was removed, and the gel slices were extracted a second time. All extracts were lyophilized, dissolved in H_2O and analyzed by capillary electrophoresis (CE) before they were desalted on Hypercarb SpinTips as described by Whitworth *et al.* (Whitworth et al., 2010). APTS-labelled material eluting in 50 % (v/v) CH_3CN+25 mM TFA was immediately flash-frozen in liquid nitrogen and lyophilized.

Digestion of APTS-labelled glycans with commercial α-(1,2)/(1,3)-mannosidase was performed following the manufacturer`s protocols. SpGHX and CpGHX, both in PBS, were tested at 5 mg/mL. All digests contained 0.1% BSA and were incubated for ~24 hours. Double digested samples were pre-treated with α-(1,2)/(1,3)-mannosidase for 3 hours prior to adding SpGHX or CpGHX for an additional 18 hours. All reactions were quenched by the addition of 3 volumes of -20°C ethanol. After further incubation at -20°C for one hour, samples were centrifuged (4°C, 10 000 rpm, 20 min), to pellet the precipitated protein, and then analyzed directly by CE.

All CE separations were carried out using a ProteomeLab PA800 (Beckman-Coulter) equipped with a laser-induced fluorescence detector and a 488 nm argon ion laser. Separations were performed on PVA-coated capillaries (Beckman Coulter) filled with NCHO separation buffer (Beckman Coulter) under reversed-polarity conditions as described previously (Figure 41) (Guttman et al., 1996).



B

Glycan	Average migration times (min)	Relative standard deviation (%)
Man-1	5.92	0.60
Man-3a	6.85	2.10
Man-5	7.90	1.20
Man-9	9.47	0.61

Figure 41: Analysis of GH125 specificity by capillary electrophoresis.

A) Migration times of APTS fluorescently-labelled and FACE purified N-glycan standards; Man1, Man-3a, 5, and 9. Contamination denoted by asterisk. B) Table representation of average migration times (minutes) for commercial standards. Also reported are the relative standard deviations (%). n=6 for each standard. C) CE traces of Man9 treated with SpGH125 (i), CpGH125 (ii), Man5 treated with SpGH125 (iii), CpGH125 (iv), and Man3a treated with SpGH125 (v), CpGH125 (vi), α-(1,2)(1,3) mannosidase (vii), SpGH125 + α-(1,2)(1,3) mannosidase (viii) and CpGH125 + α-(1,2)(1,3) mannosidase (ix). The structures of the N-glycans are shown above the traces. The Man2a product indicated with an asterisk is inferred. The identities of the peaks were determined from the mobilities of standards.

NMR Experiments (Gregg et al., 2011). ^{1}H NMR spectroscopy (600 MHz Bruker AMX spectrometer) was used to follow the progress and identify the products of the SpGHX and CpGHX catalyzed reactions. The reactions were carried out in ~ 0.2 ml [21ºC, PBS buffer (pH 6.5)] containing 7.0 mM methyl 6-O-(α-D-mannopyranosyl)-β-D-mannopyranoside. Initiation of the reaction was done by the addition of 15 µl of a 10 mg/mL stock of recombinant SpGHX or CpGHX. The hydrolysis of the disaccharide was monitored until the reaction reached equilibrium. An initial spectrum (referred to as time 0) containing substrate and buffer was acquired before the addition of enzyme.

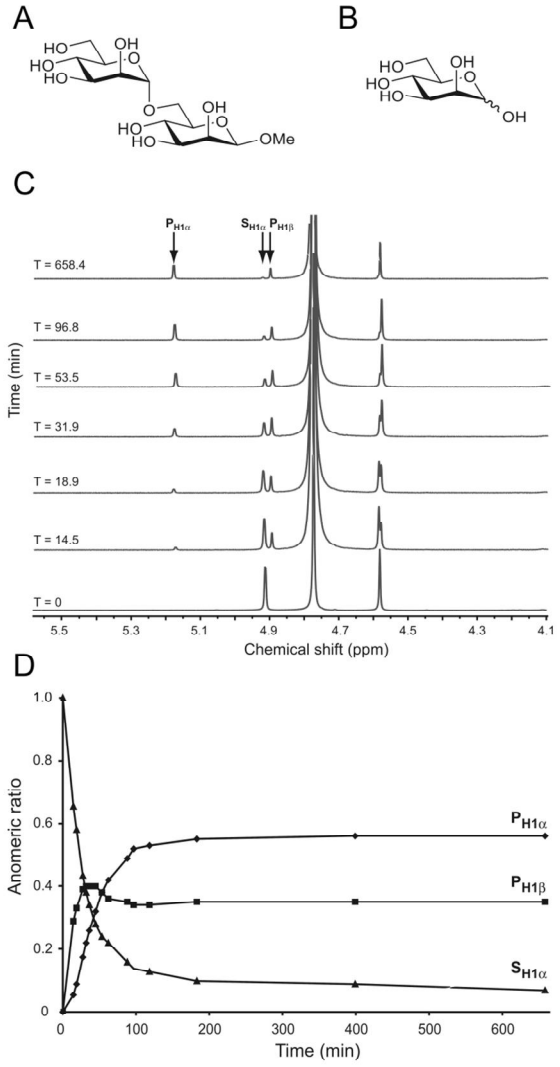


Figure 42: NMR analysis of SpGH125.

SpGH125 catalyzed cleavage of methyl 6-*O*-(*α*-D-mannopyranosyl)-*β*-D-mannopyranoside (Man*α*1-6Man*β*1-OMe) proceeds by an inverting catalytic mechanism. A) Methyl 6-*O*-(*α*-D-mannopyranosyl)-*β*-D-mannopyranoside (Man*α*1-6Man*β*1-OMe), the substrate monitored by ^{1}H NMR spectroscopy. B) Mannose, the product monitored by ^{1}H NMR spectroscopy. C) The SpGH125 catalyzed cleavage of Man*α*1-6Man*β*1-OMe was monitored as a function of time by ^{1}H NMR spectroscopy. A stacked plot shows hydrolysis of the substrate, with $S_{H1\alpha}$ representing the resonance of the anomeric proton of the non-reducing, terminal mannose unit of Man*α*1-6Man*β*1-OMe, to first form the mannose hemiacetal having the *β*-configuration at the anomeric center, indicated by $P_{H1\beta}$. $P_{H1\alpha}$ represents the anomeric proton of the *α*-anomer of the mannose product and arises from spontaneous mutarotation of the first formed *β*-anomer. D) Graphical representation of the anomeric ratio of $S_{H1\alpha}$ (▲), $P_{H1\beta}$ (■), and $P_{H1\alpha}$ (♦) with respect to time illustrate that the *β*-anomer is formed first (Gregg et al., 2011)**.**

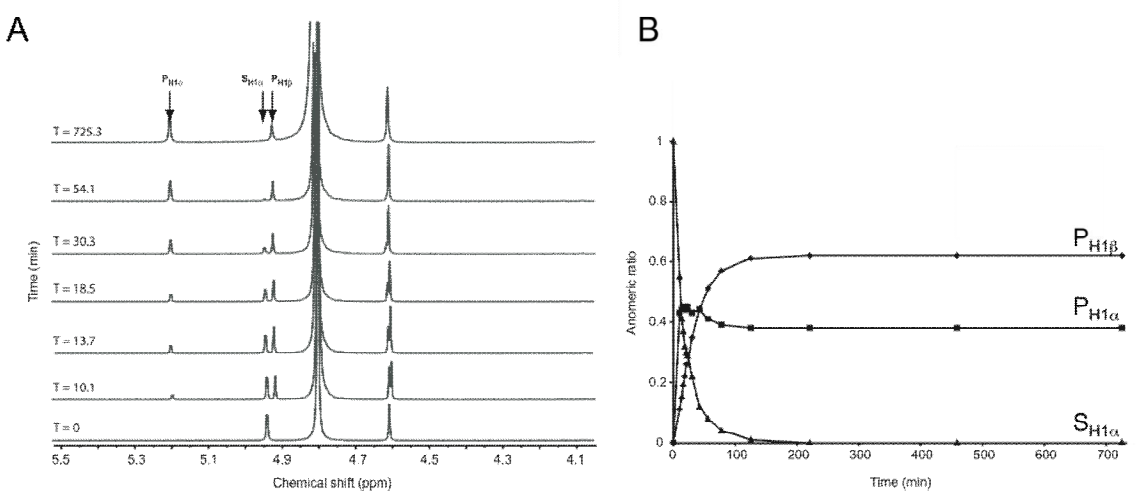


Figure 43: NMR analysis of CpGH125.

CpGH125 catalyzed cleavage of methyl 6-*O*-(*α*-D-mannopyranosyl)-*β*-D-mannopyranoside (Man*α*1-6Man*β*1-OMe) proceeds by an inverting catalytic mechanism. A) The CpGH125 catalyzed cleavage of Man*α*1-6Man*β*1-OMe was monitored as a function of time by ^{1}H NMR spectroscopy. A stacked plot shows hydrolysis of the substrate, with $S_{H1\alpha}$ representing the resonance of the anomeric proton of the non-reducing, terminal mannose unit of Man*α*1-6Man*β*1-OMe, to first form the mannose hemiacetal having the *β*-configuration at the anomeric center, indicated by $P_{H1\beta}$. $P_{H1\alpha}$ represents the anomeric proton of the *α*-anomer of the mannose product and arises from spontaneous mutarotation of the first formed *β*-anomer. B) Graphical representation of the anomeric ratio of $S_{H1\alpha}$ (▲), $P_{H1\beta}$ (■), and $P_{H1\alpha}$ (♦) with respect to time illustrate that the *β*-anomer is formed first (Gregg et al., 2011)**.**

Appendix B

[1]ELISA assays and differential scanning calorimetry experiments were performed by Jarett J. Adams and Dr. Steven P. Smith

[1] Department of Biochemistry, Queen's University, Kingston, ON, Canada, K7L 3N6

Differential Scanning Calorimetry (DSC) (Adams et al., 2008). The FIVAR-dockerin and cohesin protein samples for calorimetry, prepared in 25 mM Tris HCl (pH 7.5), 50 mM NaCl, and 5 mMCaCl2, were filtered and degassed at 21°C. The heat capacity measurements of 171 µM cohesin, 198 µM FIVAR-dockerin, and 18.2 µM FIVAR-dockerin•cohesin complex were performed from 20°C to 110°C with a scan rate of 45°C per hour by using a VP-DSC calorimeter from MicroCal. The thermodynamic parameters of the single-unfolding transitions(Tm, ΔC_p, and ΔH) were calculated with Origin 5.0 software (MicroCal). The binding association constant (K_a) for the interaction was calculated as previously described (i.e., the melting temperature of the complex is higher than the melting temperatures of the individual components, taking into account the temperature shift of both transitions) (Brandts and Lin, 1990).

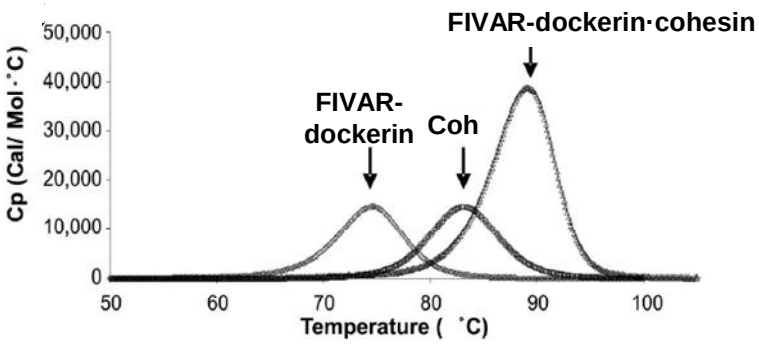


Figure 44: Differential scanning calorimetric denaturation profiles of cohesin and FIVAR-dockerin.

171 µM cohesin (Coh), 198 µM FIVAR-dockerin, and 18.2 µM FIVAR-dockerin·cohesin complex. All samples contained 5 mM $CaCl_2$.

ELISA Assays (Adams et al., 2008). The regions encoding the cohesin modules from the genes Cpe1234 (NagJ), Cpe1364, Cpe0553 (NanJ), and Cpe0266 and the regions encoding the putative dockerin modules from Cpe0191 (µ-toxin), Cpe1266, Cpe1875, and Cpe1046 were amplified from the *C. perfringens strain* 13 genomic DNA with engineered XhoI and BamHI, and BamHI and KpnI restriction sites, respectively. The cohesin- encoding fragments were cloned into a CBM3a-fusion construct, and the dockerin encoding fragments were cloned into a His-tagged heat stable xylanase (Xyn)-fusion derivative as previously described (Barak et al., 2005). The resultant fusion proteins were recombinantly expressed in E. coli and purified on amorphous cellulose (CBM containing protein derivatives) or by metal affinity chromatography (His-tagged protein derivatives), and the ELISA-based experiments were carried out similar to the method previously described (Barak et al., 2005),with incubation steps of 1 hour at 37°C for each step. The ELISA plate was coated with 100 ng of CBM-X82 and assayed with 100 µL of each of the Xyn-dockerin derivatives (0–50 ng/mL). The final colorimetric assay was developed for 1 min. All assays were done in duplicate.

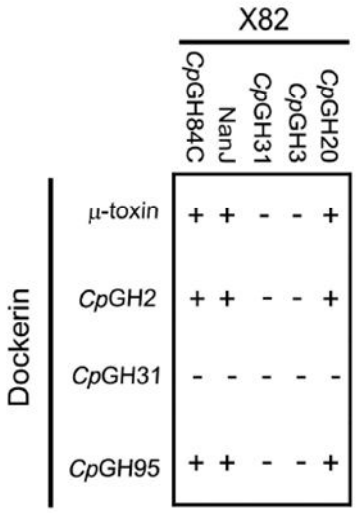


Figure 45: ELISA-based binding specificities.

Cohesin modules from family 3 (CpGH3), 20(CpGH20), 31 (CpGH31), 84 (CpGH84C) and 33 (NanJ) glycoside hydrolases fused to a CBM were probed with dockerin modules from the µ-toxin (CpGH84A), and family 2 (CpGH2), 31 (CpGH31), and 95 (CpGH95) glycoside hydrolases, which were fused to a *G. stearothermophilus* xylanase T6. (+) and (-) indicate detectable and no detectable binding, respectively (Adams et al., 2008)**.**