# Molecular and Thermodynamic Determinants of Carbohydrate Recognition by Carbohydrate-Binding Modules and a Bacterial Pullulanase

by

Alicia Lammerts van Bueren
BSc, University of Victoria, 2003

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

DOCTOR OF PHILOSOPHY

in the Faculty of Science/Department of Biochemistry and Microbiology

# Supervisory Committee

Molecular and Thermodynamic Determinants of Carbohydrate Recognition by Carbohydrate-Binding Modules and a Bacterial Pullulanase

by

Alicia Lammerts van Bueren
BSc, University of Victoria, 2003

**Supervisory Committee**

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

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

Dr. Juan Ausio (Department of Biochemistry and Microbiology)
**Departmental Member**

Dr. Penelope W. Codding (Department of Chemistry)
**Outside Member**

Dr. Steven P. Smith (Department of Biochemistry, Queen's University, ON, Canada)
**Additional Member**

# Abstract

**Supervisory Committee**
Dr. Alisdair B. Boraston, Department of Biochemistry and Microbiology
**Supervisor**
Dr. Stephen V. Evans, Department of Biochemistry and Microbiology
**Departmental Member**
Dr. Juan Ausio, Department of Biochemistry and Microbiology
**Departmental Member**
Dr. Penelope W. Codding, Department of Chemistry
**Outside Member**
Dr. Steven P. Smith, Department of Biochemistry, Queen's University, ON, Canada)
**Additional Member**

Protein-carbohydrate interactions are pivotal to many biological processes, from plant cell wall degradation to host-pathogen interactions. Many of these processes require the deployment of carbohydrate-active enzymes in order to achieve their intended effects. One such class of enzymes, glycoside hydrolases, break down carbohydrate substrates by hydrolyzing the glycosidic bond within polysaccharides or between carbohydrates and non-carbohydrate moieties. The catalytic efficiency of glycoside hydrolases is often enhanced by carbohydrate-binding modules (CBMs) which are part of the modular structure of these enzymes. Understanding the carbohydrate binding function of these modules is often key to studying the catalytic properties of the enzyme. This thesis investigates the molecular determinants of carbohydrate recognition by CBMs that share similar amino acid sequences and overall three-dimensional structures and thus fall within the same CBM family. Specifically this research focused on two families; plant cell wall binding family 6 CBMs and the α-glucan binding family 41 CBMs. Through X-ray crystallography, isothermal titration calorimetry and other biochemical experiments, the structural and biophysical properties of CBMs were analyzed. Studying members of CBM family 6 allowed us to establish the overall picture of how similar CBMs interact

with a diverse range of polysaccharide ligands. This was found to be due to changes in the topology of the binding site brought about by changes in amino acid side chains in very distinct regions of the binding pocket such that it adopted a three-dimensional shape that is complementary to the shape of the carbohydrate ligand. Members of CBM family 41 were shown to have nearly identical modes of starch recognition as found in starch-binding CBMs from other families. However family 41 CBMs are distinct as they are found mainly in pullulanases (starch debranching enzymes) and have developed binding pockets which are able to accommodate α-1,6-linkages, unlike other starch-binding CBM families. These are the first studies comparing multiple CBMs from within a given CBM family at the molecular level whose results allow us to examine the distinct modes of carbohydrate recognition within a CBM family.

Analysis of the family 41 CBMs revealed that these CBMs are mainly found in pullulanases from pathogenic bacteria. Members from *Streptococcal* species were shown to specifically interact with glycogen stores within mouse lung tissue, leading us to investigate the role of α-glucan degradation by the pullulanase SpuA in the pathogenesis of *Streptococcus pneumoniae*. SpuA targets the α-1,6-branches in glycogen granules, forming α-1,4-glucan products of varying lengths. The overall three-dimensional structure of SpuA in complex with maltotetraose was determined by X-ray crystallography and showed that its active site architecture is optimal for interacting with branched substrates. Additionally, the N-terminal CBM41 module participates in binding substrate within the active site, a novel feature for CBMs. This is the first study of α-glucan degradation by a streptococcal virulence factor and aids in explaining why it is crucial for full virulence of the organism.

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

β-CD: β-cyclodextrin or β-cycloheptaamylose

ΔG: change in free energy

ΔH: change in enthalpy

ΔS: change in entropy

AGE: affinity gel electrophoresis

CaZY: carbohydrate-active enzymes

CBD: cellulose binding domain

CBH: cellobiohydrolase

CBM: carbohydrate-binding module

CBM6: carbohydrate-binding module family 6

CCD: charged coupled device

CE: carbohydrate esterase

CPS: capsular polysaccharide

Da: Daltons

DAPI: 4',6-diamidino-2-phenylindole

DNA: deoxyribonucleic acid

ECM: extracellular matrix

EMBL: European Molecular Biology Laboratory

FACE: fluorophore assisted carbohydrate electrophoresis

FITC: fluoresceine isothiocyanate

FN3: Fibronectin type III

FOM: figure of merit

GalNAc: N-acetylgalactosamine

GAS: Group A Streptococcus

GH: glycoside hydrolase

GH32: glycoside hydrolase family 32

GLC: glucose

GM3: $6^3$-α-glucosylmaltotriose

GM3M3: $6^3$-α-glucosylmaltotriosyl-maltotriose

GPM: glucopyranosyl moranoline

GT: glycosyltransferase

GT-A: glycosyltransferase fold A

GT-B: glycosyltransferase fold B

HPA: human pancreatic amylase

HTMD: hemi thiol maltodextrin

Ig: immunoglobulin

IMAC: immobilized metal affinity column

IPTG: isopropyl β-D-thiogalactopyranoside

IUPAC: International Union of Pure and Applied Chemistry

ITC: isothermal titration calorimetry

$K_a$: affinity constant

KDa: kiloDaltons

LacNAc: N-acetyl-lactosamine

LB: Luria Bertani

MWCO: molecular weight cut off

$N_o$: binding capacity

NMR: Nuclear magnetic resonance

NTA: nitrilotriacetic acid

OD: optical density

O-GlcNAc: O-linked N-acetylglucosamine

PBS: phosphate buffered saline

PCR: polymerase chain reaction

PDB: Protein Data Bank

PEG: polyethylene glycol

PL: polysaccharide lyase

ProSP-C: prosurfactant C

RMSD: root mean square deviation

SAD: single anomalous dispersion

SeMet: selenomethionine

SIRAS: single isomorphous replacement with anomalous signal

SAXS: small angle X-Ray scattering

SBD: starch-binding domain

s.d.: standard deviation

SDS: sodium dodecylsulfate

SDS-PAGE: sodium dodecylsulfate polyacrylamide gel electrophoresis

SP: signal peptide

STM: signature tagged mutagenesis

TLC: thin layer chromatography

US: United States

UV: ultraviolet

WGA: wheat germ agglutinin

# Acknowledgments

Firstly, I would like to thank my supervisor, Dr. Alisdair Boraston, who gave me the incredible opportunity to pursue graduate studies in his lab. His knowledge and guidance throughout my Ph.D. has been invaluable. Not only has he been an excellent supervisor, he is also a mentor and a friend. After a rough introduction to biochemistry, you showed me that research can actually be fun. I admire your work ethic which I hope to carry with me into my future laboratory endeavors.

My committee Steve Evans, Juan Ausio and Penny Codding for their guidance throughout my PhD studies. I would also like to acknowledge Steve Evans for all his help with X-ray crystallography and also for his personal guidance on science and family. His personal insight really helped put life into perspective.

Thanks to all of the collaborators in the carbohydrate active enzymes field: Professors Harry Gilbert, Gideon Davies and Mirijam Czjzek. Also thanks to Dr. Robert Burke and Diana Wang for their help with flourscence microscopy.

I would also like to acknowlednge the members of the lab who have also been like family over the past five years. First, I want to thank Elizabeth Ficko-Blean whose surprise appearance in Al's lab after losing contact from back in our college days has developed into a profound academic and personal relationship. Thanks to Dr. Wade Abbott , Melanie Higgins, Katie Gregg and Ami Bitschy for all their help and support.

I also would like to acknowledge coffee and beer Friday discussions with Dr.'s Paul Romaniuk and Marty Boulanger which led to many useful and insightful research talks.

# Dedication

First and foremost I want to thank my husband Jason who has supported me throughout my entire education. In fact, as long as we have known each other I have been a student, from dating, to marriage and finally the birth of our daughter, so it must be a great relief to him to know that his support has finally proven its worth. I am forever grateful for your love and support and even though it was incredibly rough at times, you were always there to help me focus on the important things.

I would also like to thank my family for their support; specifically I want to thank Sharon Lammerts van Bueren who has always been there to help during my educational pursuits. Your love has always been motivational and a geat support net.

Finally, I would like to dedicate this thesis to two very important people in my life. First, my Opa who passed away in 1998 who taught me that I am able to do anything. His struggle and eventual defeat from esophageal cancer gave me the motivation to do something greater for the world and hopefully I will be able to fulfill that in my future endeavors. Second, and most importantly, I dedicate this thesis to my daughter Saskia who blessed our lives a bit sooner than we intended. She has shown me that there is more to life than science which can be very consuming. Her smiles and curiosity have taught me to slow down a bit and enjoy the simpler things that life has to offer.

# Chapter 1: General Introduction

## 1.1 Carbohydrates and the Environment

Carbohydrates are the most abundant biomolecules on the planet. They are ubiquitous in nature as they are found in places such as plant biomass, insect exoskeletons, bacterial cell surfaces and biofilms, and mammalian cell surfaces. The functionality of carbohydrates are determined by their overall three-dimensional shape, which is also dependent on length of the carbohydrate, its sugar composition, the position of the anomeric carbon and the type of glycosidic linkages that can be formed between sugar monomers. These create a platform for millions of different possible combinations of carbohydrate structures, and each structure is suited for serving its function in nature.

### 1.1.1 Plant and Fungal Polysaccharides

Plant cell-wall material is the main component of terrestrial biomass [1]. The bulk of plant cell wall material is cellulose, a homopolymer of β-1,4-glucose which takes on an overall linear shape. Cellulose is hypothesized to exist in two forms, crystalline and amorphous. In the crystalline form cellulose chains self associate via intra- and intermolecular hydrogen bonds and van der Waals forces to form cellulose fibrils and microfibrils, which are extremely insoluble and provide the majority of tensile strength to the plant cell wall. Amorphous regions lack this higher order structure and are more susceptible to increased degradation. Plant cell walls also contain a number of other sugar polymers termed hemicellulose which includes xylan (β-1,4-linked xylose), laminarin (β-1,3-linked glucose), mannan (β-1,4-linked mannose) and lichenan (mixed β-1,3-1,4-

linked glucose). The other main structure found within the plant cell wall are pectins, substituted heteropolysaccharides composed of a α-1,4-D-galacturonic acid backbone with rhamnose, galactose and arabinose substituents. Cellulose contains regions of attachment for hemicellulose and pectins, forming a complex interwoven amalgam within the cell wall. Together they form a rigid structure which provides a barrier that is highly resistant to environmental forces and biological attack. Seaweeds, which include algae and kelp, have cellulose and β-1,3-glucans such as laminarin within their cell wall structures but also contain specific unique polysaccharides such as alginic acid, agarose and carageenan.

Fungal cell walls are mainly composed of chitin, a linear polymer of β-1,4-linked N-acetyl-glucosamine, which provides rigidity to the cells and helps stabilize long filamentous cells such as hyphae and mycelia. Chitin is also the main component of exoskeletons found in arthropods such as insects (beetles, spiders, etc.) and crustaceans (crab, shrimp, etc.). These exoskeletons serve as a solid barrier for protection from dessication and other environmental forces. Other fungal cell wall polysaccharides include β-1,3-glucans, chitosan (a polymer of β-1,4-linked glucosamine) in addition to a small percentage of cellulose.

### 1.1.2 Bacterial polysaccharides

Bacterial cells are surrounded with carbohydrate coatings which serve as a protective barrier for the cell. Gram positive and gram negative bacteria contain a thick wall of peptidoglycan, a repeating unit of N-acetyl-glucosamine and N-acetyl-muramic acid connected by a β-1,4-linkage. Peptidoglycan layers are connected via oligopeptide

chains and the overall three-dimensional structure aids in maintaining the bacterial cell structure. Gram negative bacteria have a thin wall of peptidoglycan followed by an outer membrane containing lipopolysaccharide (LPS), a unique bacterial species-specific sugar polymer attached to the cell by a lipid anchor. Many gram positive pathogenic bacteria have carbohydrate capsules attached to the peptidoglycan layer such as lipoarabino-mannan from *Mycobacterium tuberculosis* [2]and capsular polysaccharide from *Streptococcal* species [3; 4; 5]. The functions of these capsules serve to protect bacteria from the immune system and are implicated in attachment to host cells during infection [4]. For example, the hyaluronic acid capsule of *Streptococcus pyogenes* mimics that found in its human host and helps the organism hide from the immune system [6].

Bacterial biofilms consist of bacterial cells and associated extracellular polymeric substances (EPS) which include many carbohydrate structures[7]. This exopolysaccharide matrix is secreted by bacteria to create a platform for the attachment of many bacterial cells, creating a multicellular entity, which aids in bacterial resistance to environmental forces such as dessication and antibiotics. Most bacteria live in biofilms, which are found everywhere in the environment in conditions with a solid substrate that is exposed to aqueous solutions. Biofilm formation by *Pseudomonas aeruginosa* can lead to chronic infection in lung epithelia [8]. The most well known biofilm is that found in dental plaque of which a large portion is made up of a dextran matrix deposited onto the teeth by the bacterium *Streptococcus mutans* [9]. Without proper removal, the biofilm contributes to the loss of tooth enamel causing ailments such as cavities and gingivitis. Biofilms are also problematic in the colonization of hospital equipment, which can lead to hospital-acquired bacterial infections in patients.

### 1.1.3 Energy storage by α-glucans

Most organisms are capable of metabolizing glucose as an energy source for cellular processes. In plants and animals, glucose is stored as starch and glycogen, respectively. Starch is composed of amylose, a homogenous polysaccharide of α-1,4-linked glucose, and amylopectin, which is similar to amylose but with additional α-1,6-branch points occurring every approximately 20 glucose residues. Starch is stored within amyloplasts within the seeds, roots and stems. Glycogen is of similar composition to amylopectin but with α-1,6-branches occurring more frequently every 8 – 12 residues and is mainly found in the liver hepatocytes where it makes up ~8% of liver mass. Due to the α-1,4-linkages, portions of starch and glycogen take on a double helix shape forming compact granules for efficient storage of glucose [10]. Other common α-glucans include pullulan, dextran and mutan. Pullulan is a linear water soluble polymer of repeating α-1,6-linked maltotriose (three α-1,4-linked glucose monomers) occuring naturally in the plant fungus *Aureobasidium pullulans*. It is generated from starch for the production of blastospores and hyphae [11]. Dextran is a homogenous polymer of α-1,6-linked glucose and is produced from the lactic acid fermentation of sucrose by *S. mutans* for extracellular energy storage [12]. Mutan is a water insoluble polymer of α-1,3-linked glucose generated from starch in some tubers and can also be found in the cell wall of some fungal species[13].

### 1.1.4 Mammaliam cells and Complex Glycans

Complex glycans can be found attached to the surface of mammalian cells as glycolipids and glycoproteins or as soluble entities and serve many important functions in cell recognition, cell signaling, cell development, and cell-matrix interactions. The surfaces of mammalian cells are coated in varied complex glycans and are differentially expressed during stages of growth and maturation. Fully mature cell often have carbohydrate structures which are characteristic to their cell type, permitting cell recognition. Often changes in these surface carbohydrates are indicative of many malignant forms of cancer [14].

Surface glycoproteins are classified as either N-linked or O-linked. N-linked sugars are attached to proteins via the amine group of asparagines, forming an aspartylglycosylamine linkage with an N-acetylglucosamine. N-linked glycans are varied in their composition but all have characteristic high mannose content and terminate in fucose or sialic acid. They also contain an identical N-acetylchitobiose-trimannosyl core structure. O – linked glycans are attached to hydroxyl groups of serine and threonine side chains and are common in mucin and mucopolysacharides lining lung and GI tract epithelium and the epithelium of the reproductive tract. They also have a common core structure composed of GalNAc substituted with Gal and GlcNAc residues to form a backbone structure for the attachment of peripheral carbohydrate antigens such as $Lewis^A$, $Lewis^Y$ and Tn antigen as well as sialylated and branched forms of these sugars. Mucins provide a highly hydrated surface which act as a barrier between body fluids and epithelium. Many pathogenic bacteria use these glycans as receptors for invasion and as a nutritional source to promote growth and spread throughout the body (see section 1.6).

Heparan sulfate and chondroitin sulfate are components of proteoglycans and

occur in all animal tissues, making up a majority of the extracellular matrix. They are involved in regulating the movement of molecules through the ECM which aids in many cell regulatory processes.

Perhaps the most well known complex glycans are those of the ABO blood group system comprised of antigens in the form of carbohydrates found on the surface of red blood cells. Type O, also known as the H-antigen, is a chain of α-fucose-(1,2)-β-D-galactose linked to β-N-acetyl-glucosamine and β-D-galactose. The H antigen serves as the base for types A and B with type A having an α-1,3-N-acetyl-galactosamine and type B having an α-1,3-galactose attached to the galactose of the terminal fucosylgalactose moiety. Their functional role remains unknown; however, these antigens play an important immunological role in recognition of self and can cause severe reactions in an individual who receives blood of the wrong type.

Cell signaling events in response to environmental stimuli are often triggered by the modification of target proteins, including phosphorylation, acetylation, ubiquitination and methylation. More recently the importance of O-GlcNAc modification of proteins has become apparent in signaling events [15]. It was once thought that proteins within the nucleus and cytoplasm were not glycosylated, but now it is known that O-GlcNAc is a dynamic modification occurring on cellular proteins, often competing with phosphorylation sites at serine and threonine residues [16; 17]. O-GlcNAc modifications of cellular proteins have been identified in regulating events such as chromatin rearrangement, transcription, translation, regulation of glucose levels and maintaining cell shape. It also has many implications in diseases such as diabetes, neurodegeneration, and many forms of cancer [18; 19; 20; 21].

## 1.2 Carbohydrate-Active Enzymes

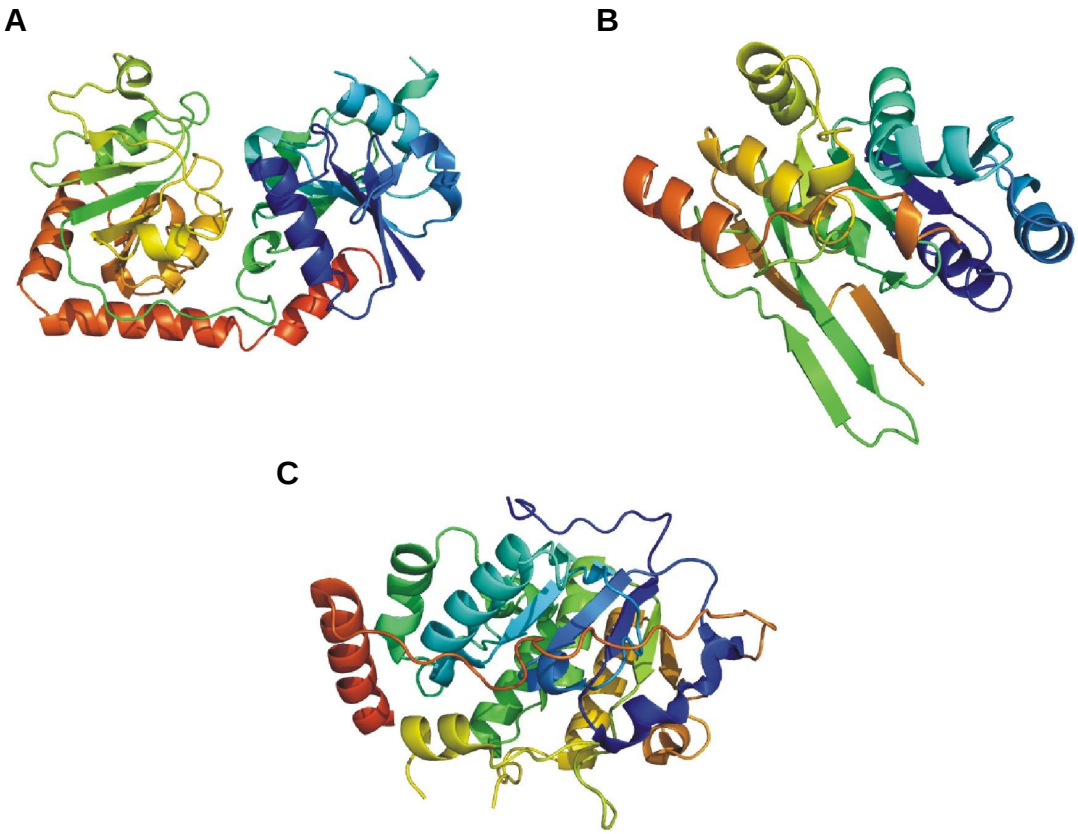
Carbohydrates are dynamic molecules that are constantly being synthesized and broken down. To achieve this, organisms contain genes encoding a variety of enzymes that are involved in glycosidic bond formation and cleavage. These include glycosyltransferases, which are mainly responsible for the formation of the glycosidic bond in the biosynthesis of carbohydrates, and polysaccharide lyases, carbohydrate esterases and glycoside hydrolases, which are involved in the breakdown of polysaccharides and carbohydrate moieties. These carbohydrate-active enzymes are grouped into over 250 families based on amino acid sequence similarity and are all listed in the continually updated Carbohydrate-Active Enzyme (CAZy) database (www.cazy.org) [22]. Closer analysis of the genomes listed within the database reveals the importance of carbohydrate metabolism to life on Earth as 1-3% of the genome of most organisms is devoted to encoding glycosyltransferases (GTs) and glycoside hydrolases (GHs) [23]. This provides a wealth of gene sequences in which to study the structure and function of carbohydrate-active enzymes to better understand how these enzymes function in nature.

### 1.2.1 Glycosidic Bond Formation

Glycosyltransferases (GTs) are responsible for the biosynthesis of carbohydrates from the formation of plant cell wall polysaccharides to detailed glycoconjugates found on cell surfaces (see Section 1.1.4). They catalyze the transfer of a sugar moiety via an activated donor sugar molecule onto an acceptor (which can be either a carbohydrate,

protein or lipid molecule) forming a glycosidic bond with either retention or inversion of the anomeric carbon. In the CAZy database there are currently over 12,000 GT sequences grouped into 91 amino acid sequence-based families with structures representing only 29 of these families [24; 25]. GTs are the least studied of the carbohydrate-active enzymes due to difficulty in expressing and purifying these enzymes for crystallization. So far all GTs share a high degree of structural similarity despite the low amino acid sequence identity between families and fall into either the GT-A or GT-B fold clan (Figure 1A&B). The first characterized GT-B fold was reported in 1994 of the bacteriophage T4-glucosyltransferase (GT family 63) which catalyzes the transfer of glucose to phage-modified DNA [26] (Figure 1A). The enzyme contained two domains with a characteristic Rossmann nucleotide-binding motif, which was shown to interact with the activated nucleotide sugar donor molecule. In 1999 the first GT-A fold was revealed by the X-Ray crystal structure of SpsA, a glycosyltransferase implicated in the synthesis of *B. subtilis* spore coat [27] (Figure 1B). The GT-A fold of this family 2 GT is also a two-domain enzyme with an N-terminal Rossmann motifs and a C-terminal DxD motif with mixed $\alpha/\beta/\alpha$-sandwich. Recently a third fold was identified in an $\alpha$-2,3- sialyltransferase from *Campylobacter jejuni* from GT family 42, which was found to be similar to GT-A with a seven-stranded $\beta$-sheet but no DxD motif [28] (Figure 1C). Despite the structural similarities in GTs, they show exquisite specificity for both the activated sugar donor and the acceptor substrate due to modifications within loop regions surrounding the active site.

Figure 1: The three GT folds observed in glycosyltransferases (A) GT-B fold from bacteriophage T4 glucosyltransferase (GT63) (PDB Code 2BGU) [26] (B) GT-A fold from *Bacillus subtilis* spore coat forming glycosyltransferase SpsA (GT2) (PDB code 1QG8) [27]. (C) A new fold recently revealed from an α-2,3-sialyltransferase from *Campylobacter jejuni* (GT42), a modified GT-A fold (PDB code 2P2V) [28].



### 1.2.2 Carbohydrate breakdown

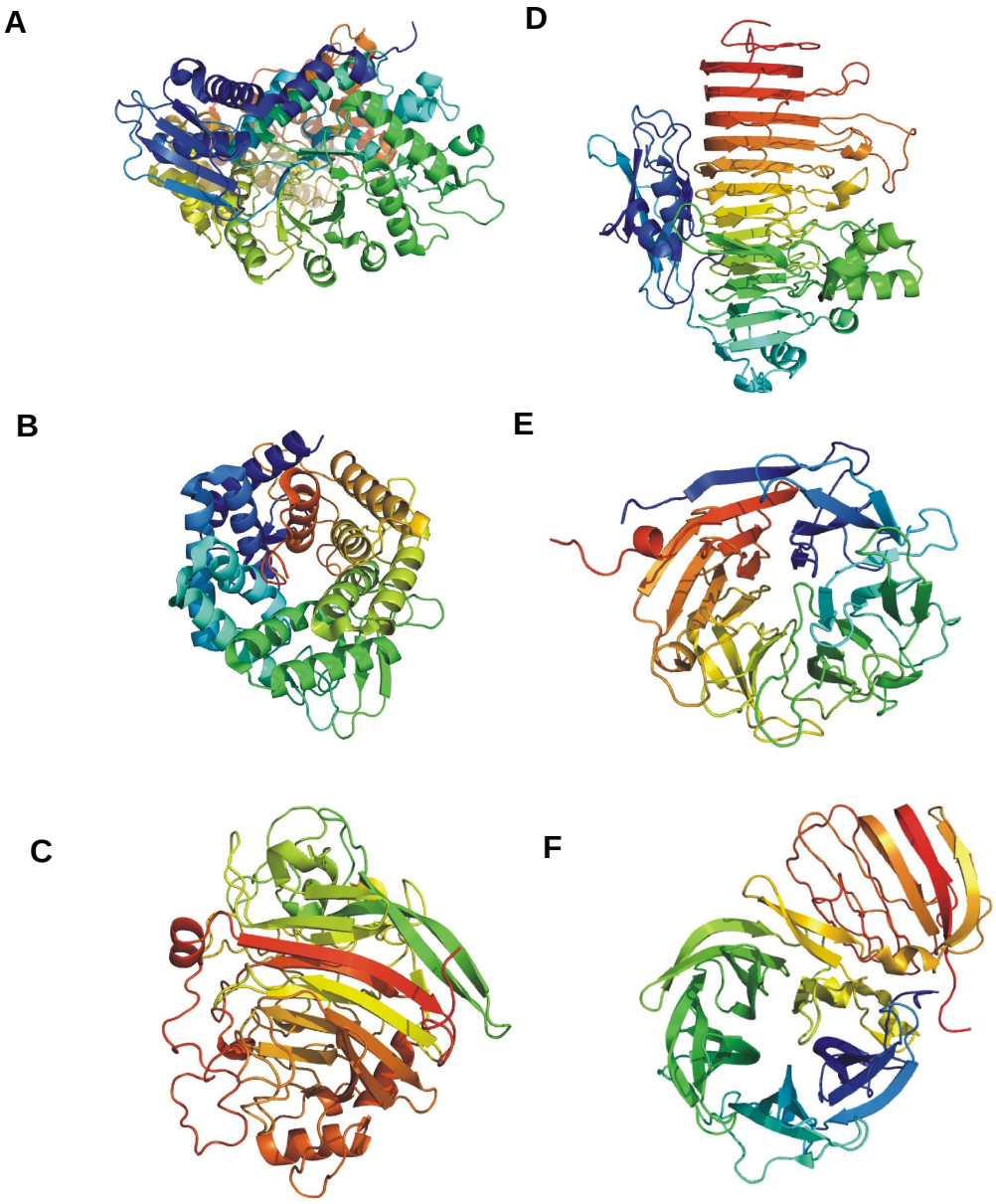
Carbohydrate esterases, polysaccharide lyases and glycoside hydrolases are all enzymes involved in the breakdown of polysaccharides. Carbohydrate esterases are a class of enzymes that catalyze the de-O or de-N-acetylation of substituted sugars using a catalytic mechanism similar to protein and lipid esterases that utilize a Ser-Asp-His catalytic triad (CE families 1, 3, 5, 7, 10, 12) [29]. Other families have been shown to use a $Zn^{2+}$ catalyzed deacetylation method (4, 9, 11, 14) [30]. There are 15 sequence-based families with structures for thirteen families, most often showing a classic serine-protease ($\beta\alpha\beta$) sandwich fold [22]. CE's are most commonly implicated in the deacetylation of chitin [31], peptidoglycan modification [32], and the deacetylation of acetylated plant xylans and glucans [33].

Polysaccharide lyases cleave glycosidic bonds via β-elimination which results in the formation of a double bond at the newly formed non-reducing end between C4 and C5. Most PLs are of bacterial origin and are active on uronic acid sugars such as glucuronates, galacturonates and alginates which are found in plant pectins and algaes. These enzymes participate in plant biomass degradation and as virulence factors in plant and human pathogens, as in pectin degradation in soft rot by Erwinia species [34]. Their presence in human pathogens often mimics the activity of hyaluronate lyases and heparin lyases [35], however, the presence of polygalacturonate pathways in human pathogens is ambiguous. Entire pectin utilization pathways are found in a variety of human pathogens from *Enterobacteraciae* [36]; specifically the foodborne pathogen *Yersinia enterocolitica.* *Y. enterocolitica* produces a pectate lyase (PL-2) and polygalacturonase and it is

rationalized that they allow the bacteria to scavenge pectin found within the human intestine to be used as a nutritional source [37]. There are 18 sequence based families with structures representing 14 families [22]. A selection of folds are observed within PLs, such as the β-helix, β-jelly roll folds and α-toroid, while the active sites remain structurally conserved [37].

Glycoside hydrolases are by far the most prevalent class of carbohydrate-active enzyme with over 30,000 entries in 112 amino acid sequence-based families [38]. Structures have been determined for 76 of these families. The mechanisms of glycosidic bond hydrolysis by glycoside hydrolases have been extensively studied. In general, glycosidic bond cleavage results in either inversion or retention of the anomeric carbon[39]. Inversion occurs in a single step while retention is a two-step mechanism involving an oxacarbenium ion-like transition state. A third mechanism which also results in retention of the anomeric configuration is substrate-assisted catalysis where the N-acetyl group of the sugar acceptor takes the place of the catalytic nucleophile, forming an oxazolinium intermediate [39]. Unlike GTs, structural data on GHs has revealed several different folds, such as the $(\alpha/\alpha)_6$, β-helix, β-propellor, β-jelly roll and the $(\alpha/\beta)_8$ TIM barrel motif, of which the latter is found in the majority of GHs to date (Figure 2A-F). Enzymes within a family have similar structures, mechanisms of hydrolysis, and conserved catalytic residues, therefore we can often predict the activity of a GH within a given family. Because of fold similarities between GH families, GHs have been grouped into 14 structure-based clans which helps classify new GH enzymes whose categorization based on amino-acid sequence may relate to more than one family [40].

Figure 2: Folds observed within Glycoside hydrolases. (A) TIM barrel $(\alpha/\beta)_8$ motif from *Clostridium perfringens* α-N-acetylglucosaminidase (GH89) (PDB Code 2VCA) [41]. (B) $(\alpha/\alpha)_6$ toroid motif from *Bacillus sp.* unsaturated glucuronyl hydrolase (GH88) (PDB Code 1VD5) [42]. (C) β-jelly roll motif from *Trichoderma reesei* cellobiohydrolase I (GH7) (PDB Code 1CEL) [43]. (D) β-helix fold from *Yersinia enterocolitica* exo-polygalacturonase (GH28) (PDB Code 2UVE) [44]. (E) 6-fold β-propeller motif from *Micromonospora viridifaciens* sialidase (GH33) (PDB Code 1EUR) [45] (F) 5-fold β-propeller motif from *Thermotoga maritima* β-fructosidase (GH32) (PDB Code 1UYP).



GH activity is, in general terms, opposite to GTs in that GHs hydrolyze the glycosidic bond between two sugar molecules or a carbohydrate and non-carbohydrate moiety. Therefore it is no surprise that GHs and GTs work together in many dynamic processes. Examples include the synthesis and breakdown of polysaccharides in plant growth and differentiation and in meeting energy requirements in the case of starch and glycogen. The dynamic O-GlcNAc modification of cellular proteins is regulated by OGlcNAc-transferase [46], a GT41 that transfers an O-GlcNac onto serine and threonine sidechains, and OGlcNAcase from family GH84 which catalyzes the removal of these sugars [47]. The importance of synergism between these two enzymes is apparent in many diseases, where an abundance of O-GlcNac modified cellular proteins can lead to diabetes, Alzheimers and cancers [19; 20; 21].

GHs are important in plant cell wall degradation which represents the largest reservoir of organic carbon in the biosphere, and thus cell wall degradation by microbial enzymes is pivotal to many biological and industrial processes [48]. However, the polysaccharide composite of plant cell walls is relatively recalcitrant to enzymatic degradation and as a result microbes have evolved complex enzymatic systems in order to tackle this problem. For example, *Clostridium thermocellum* and *Clostridium cellulolyticum* secrete a megadalton multimodular enzyme complex called the cellulosome [49]. It is an extracellular enzyme complex, which functions to degrade plant cell wall tissue. The multienzyme arrangement is mediated by a scaffoldin protein base containing cohesin domains, which interact with dockerin domains of the hydrolytic enzymes, forming an ensemble of various catalytic subunits [50].

More recent research into human pathogenic bacteria has identified GHs as virulence factors. *Clostridium perfringens* secretes a battery of hydrolytic enzymes as exotoxins for pathogenesis, of which many are glycoside hydrolases [51], and recent experiments suggest that these enzymes may form multimodular complexes like those found in the cellulosome[52]. Also many GHs appear as virulence factors in *Streptococcus pneumoniae* [53] where they may participate directly in hydrolysis of host glycans. This relatively new area of glycoside hydrolase virulence factor research will likely translate to many other pathogens and their importance in pathogenesis may lead to novel targets for drug therapies to treat infections and combat bacterial antibiotic resistance.

### 1.3 Glycoside Hydrolases and their modularity

To increase the efficiency of degradation, glycoside hydrolases often have complex modular architectures consisting of a catalytic module fused with one or more ancillary modules *via* linker peptides. A module is defined as a contiguous amino acid sequence within a larger sequence that folds independently (Figure 3) and has an individual function but together increase the overall efficiency of the enzyme. The first indication that these enzymes contained distinct independent functioning modules was from the limited papain digestion of cellobiohydrolase I and II from the fungus *Trichoderma reesei* [54; 55]. Proteolytic cleavage identified two functional N and C terminal domains where the hydrolytic activity of N-terminal domain remained active but the specific activity on cellulose decreased to 50% of initial activity. The C-terminal domain independently bound to crystalline cellulose but had no catalytic activity, suggesting that

Figure 3: Modularity of glycoside hydrolases as shown by the sialidase from *Micromonospora viridifaciens* (PDB Code 1EUT) [45]. Catalytic module (GH33) shown in green, linker (Ig fold) shown in yellow and carbohydrate-binding module (CBM32) shown in red.

the C-terminal carbohydrate-binding activity complemented the catalytic module. With the implementation of bioinformatics came the ability to find distinct regions within glycoside hydrolases that share sequence and secondary structure similarities to other protein motifs, including cohesins, dockerins and FN3 motifs which potentially mediate protein-protein interactions. However, the most frequently found modules are the carbohydrate-binding modules (CBMs) which interact with the target carbohydrate substrate. Because large multimodular enzymes can be difficult to work with in the laboratory, the popular molecular biological approach has been to dissect the modular structure of glycoside hydrolases and study the activity of the individual modules independently, allowing researchers to then fit the pieces together and determine how the enzyme functions as a whole.

### 1.4 Carbohydrate-Binding Modules

Carbohydrate binding modules are the most prominent accessory module found in glycoside hydrolases [56]. They are classified as non-catalytic modules that assist in the efficient degradation of the targeted carbohydrate substrate. This is accomplished by binding to the target substrate and directing the catalytic module to the cleavage site, which in turn increases the specific activity of the enzyme. The modular structure of a glycoside hydrolase may include a single or multiple CBMs, either from the same or different CBM families and may also include other functional modules, such as those that mediate protein-protein interactions. The importance of CBMs to carbohydrate degradation is significant, as demonstrated by the immense distribution of these modules in glycoside hydrolases (www.cazy.org) [22] and by biochemical analysis. The contribution

of CBM research to the field of protein-carbohydrate interactions has been invaluable and has led to CBM use in many biotechnological applications [57; 58; 59; 60].

CBMs have three main roles in carbohydrate degradation: targeting the enzyme to its substrate, localizing the catalytic module of the enzyme in close proximity of the substrate, and disruption of the carbohydrate surface [56]. The targeting effect of CBMs allows for the specific interaction with polysaccharide substructures within higher-order polymers, such as cellulose-specific CBMs that are suited to interact with either crystalline or amorphous regions of cellulose. The proximity effect of CBMs helps in directing the enzyme within the proximity of the target substrate rather than in solution. The disruptive effect by CBMs increases the accessibility of the enzyme to the target substrate by disrupting the surfaces of insoluble substrates such as cellulose and chitin [61]. We have recently revealed a potential fourth function for CBMs, an anchoring effect where an appended CBM anchors a secreted glycoside hydrolase onto the surface carbohydrates of bacterial cells (ALVB and ABB, unpublished data). A modification of the proximity effect, the anchoring of the glycoside hydrolase onto the bacterial surface keeps the enzyme in close proximity of the bacterial cell rather than the substrate, acting as a possible defense mechanism against antimicrobials present in the bacterium's environment.

Glycoside hydrolases often contain multiple CBMs, which may or may not be from the same family and may exist in tandem or be separated by other modules. Multiple CBMs show an increased affinity for their target substrate over the single modules, which is brought on by an avidity effect with the ligand [62; 63]. The first study to characterize the role of bifunctional CBMs in cellulose binding was Linder *et. al.* who showed that two

recombinantly fused CBDs had an increased affinity for cellulose over the single modules[64]. This is attributed to an additive effect of the free energies of binding for the individual CBDs plus the coupling free energy (the free energy caused by the increased probability of the second CBD binding with ligand after the first CBD is bound). The first investigation into the role of multiple CBMs within an enzyme was of *Cellulomonas fimi* xylanase, which contains one internal and one C-terminal CBMs from family 2b [65]. When both CBMs were incorporated into a single polypeptide chain, either within the enzyme or joined by a polypeptide linker, they had an 18-20 fold increase in affinity for soluble and insoluble xylan and insoluble cellulose over the individual modules. Multiple CBMs also may occur in tandem within the modular structure of glycoside hydrolases. The *Clostridium stercorarium* xylanase contains a tandem triplet of CBM6s where the individual modules interact with xylan with an affinity of ~$10^3$ – $10^5$ $M^{-1}$ but together have a cooperative effect and have an overall 20-40 fold increase in affinity for xylan [62].

CBMs are able to bind specifically with their target ligand but may also accommodate other sugars with a slightly decreased affinity. An example is CBM29-2 from *Piromyces equi* NCP-1, a non-catalytic protein which is part of the cellulase/hemicellulase complex, that is able to accommodate both gluco- and manno- configured sugars [66]. This flexibility in ligand recognition permitted the cellulase/hemicellulase complex to target a range of different components within the plant cell wall. Mutagenic studies pin pointed CBM29-2's specificity for both sugars to a glutamate residue within the binding pocket [67]. When glutamate is mutated to an arginine, CBM29-2 loses its affinity for manno-oligosaccharides, showing a possible evolutionary link in promiscuous ligand binding. Another interesting example of CBM promiscuity is *Tm*CBM9-2 from *Thermotoga*

*maritima* xylanase Xyn11A [68], which is able to bind tightly with reducing end glucose and xylose residues. In both sugars, the hydroxyl groups are positioned equatorially, making the same contacts with the protein, and furthermore, the C6 hydroxymethyl group in cellulose does not make any direct contacts with the protein [69]. Since xylan and cellulose are intimately associated within the plant cell wall, the promiscuity in *Tm*CBM9-2 ligand binding would increase the number of binding sites in the enzyme while maintaining its specificity for xylan. CBM family 4 contains many examples of modules with varied polysaccharide binding specificities, including *Tm*CBM4-2, from a *T. maritima* laminarinase with specificity for β-1,3-linked glucans, mixed β-1,3-1,4-linked glucans and cellooligosaccharides [70], *Rm*CBM4-1 and 4-2 from *Rhodothermus marinus* xylanase with specificities for xylan, cellulose and mixed β-1,3-1,4-linked glucans [71], and *Cf*CBM4-1 and 4-2, from *Cellulomonas fimi* cellulase specific for mixed β-1,3-1,4-linked glucans in addition to amorphous cellulose and cellooligosaccharides [72]. Recent work with family 32 CBMs from secreted glycoside hydrolase exotoxins from the pathogen *Clostridium perfringens* showed that the appended CBM32s are able to accommodate galactose and substituted galactose moieties [51] which might be beneficial to the enzymes as it would increase their chance of interacting with complex human glycans, their intended target substrate.

Although CBMs are mainly associated with glycoside hydrolases, there are examples of CBMs appended to glycosyltransferases. Many family 27 GTs have C-terminal family 13 CBMs. The CBM13 module from a polypeptide-N-acetylgalactosaminyltransferase (GalNAc transferase) involved in the O-glycosylation of mucin biosynthesis was shown

to interact with GalNAc and inactivation of this module prevented attachment of the enzyme to its acceptor [73].

There are a few examples of CBMs that exist as independent modules and are not found in the context of a catalytic module. CBM family 14 and 18 contains members from non-catalytic proteins that interact with chitin. The fungus *Cladosporium fulvum* utilizes a CBM14 as a means of protecting itself from plant chitinases [74]. YeCBM32 from *Yersinia enterocolitica,* interacts with pectin fragments within the periplasm of the organism as a means to retain these fragments in the periplasm for further degradation and transport into the cytoplasm [75]. Because independent CBMs have similar properties to lectins and lectins can be classified as CBMs based on amino acid sequence similarities (such as ricin B-chain from CBM family 13 and wheat germ agglutinin from CBM family 18), it is sometimes difficult to make the distinction between CBMs and lectins.

### 1.4.1 CBM Structure

Initially CBMs were classified as cellulose binding domains based on their discovery in enzymes that are active on cellulose. Since they are also found in enzymes not active on cellulose, the term carbohydrate-binding module has become a more widely accepted term to classify these modules. Similar to GHs and GTs, CBMs are grouped into 52 sequence-based families [56], which may be found in the continuously updated carbohydrate-active enzyme data base at www.cazy.org [22]. A new family is created once the carbohydrate-binding activity of a putative CBM has been demonstrated. Other putative members are then added based on amino acid sequence similarities. The CBM classification system enables CBMs to be grouped according to structure; however,

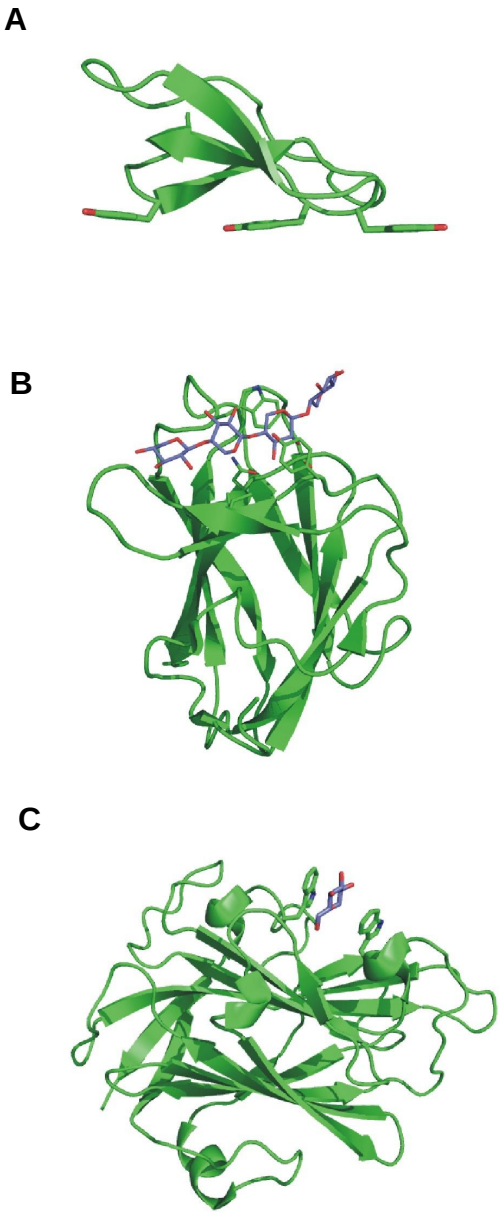
binding specificity is dependent on differences within loops and side chains. Therefore belonging to a specific family may (eg: CBM41) or may not (eg: CBM6) be predictive of binding function.

Like catalytic modules from glycoside hydrolases, CBMs also exhibit a variety of folds. Seven different folds have been identified which include the β-sandwich, β-trefoil, cysteine knot, unique, OB, hevein, and hevein-like folds [56]. By far the most dominant fold is the β-sandwich, a fold shared with some lectins which is comprised of two overlapping β-sheets, each containing 3-6 antiparallel β-strands. Most often the binding site is located on the concave face of one of the β-sheets but may also be found at the apex of the protein within the loops joining the β-strands (families 6, 32, 47, 51). All β-sandwich CBMs have an associated metal ion that helps in maintaining the overall structure of the protein, however there are also examples of CBMs with additional metal-assisted ligand binding site properties. The interactions of xylan with a family 36 CBM from *Paenibacillus polymyxa* xylanase [76] and with a family 35 CBM from *Cellvibrio japonicus* xylanase [77]are calcium dependent as well as the interaction of a CBM35 from *Amycolatopsis orientalis* exo-β-D-glucosaminidase with glucuronic acid (ALvB and ABB, unpublished data). The adaptability of this fold in carbohydrate-binding proteins makes it an ideal scaffold for protein-carbohydrate interactions with a diverse range of polysaccharides. The cysteine knot, unique and OB folds only appear in “type A” CBMs (see below) that interact with crystalline cellulose and chitin.

Along with the sequence and fold-based classification systems, CBM binding function can be classified into three different types based on the topology of the binding

sites which reflects the macromolecular structure of the target ligand [56] (Figure 4). Type A CBMs contain a planar hydrophobic ligand binding surface that interacts with crystalline polysaccharides such as cellulose and chitin and are found in families 1,2,3,5 and 10 (Figure 4A). Mainly aromatic side chains are responsible for forming a platform that mediates hydrophobic stacking interactions with cellulose chains by overlapping with the pyranose rings of glucose [78]. Hydrogen bond formation by type A CBMs does not appear to be important because mutating amino acids involved in hydrogen bond formation does not affect its affinity for ligand [79]. Type B CBMs, which make up the majority of CBM families due to their frequent presence in plant cell wall degrading enzymes, contain clefts that accommodate single polysaccharide chains (Figure 4B). The three-dimensional structure of all Type B CBMs determined to date have a β-sandwich fold with a single ligand binding site comprised of a shallow extended cleft on the concave surface of the protein or at the apex of the protein within loops joining the β-sheets (eg: family 6). Type C CBMs, comprising families 9, 13, 14, 18, 32, 47 and 51 interact with mono- or disaccharides in a lectin-like manner (Figure 4C). The most well studied type C CBM is TmCBM9-2 from *Thermotoga maritima* xylanase which interacts with the reducing end of glucose or xylose polymers [68; 69] (see promiscuity and CBMs above). Only recently has there been more information on type C CBMs since very few type Cs are involved in plant cell wall recognition and they appear to be more prevalent in bacterial exotoxins and enzymes active on complex glycans [51, 80,81]. Like the type B CBMs, Type C most commonly have a β-sandwich fold with a short binding site on the concave surface (such as families 9, 14, 18) or at the apex of the protein (such as families

Figure 4: CBM types based on binding site topology. (A) Type A CBMs have a planar binding surface for interacting with crystalline ligands: CBM1 from Trichoderma reesei cellobiohydrolase I (PDB Code 1CBH) [82]. (B) Type B CBMs have an extended binding pocket for interacting with extended sugars: CBM6 from *Clostridium stercorarium* xylanase in complex with xylotetraose (blue) (PDB Code 1UY4) [83]. (C) Type C CBMs have short binding pockets for interacting with mono-, di-, or trisaccharides: CBM9 from *Thermotoga maritima* xylanase in complex with cellobiose (blue) (PDB Code 1I82) [69].



32, 47, 51), while family 13 CBMs have a β-trefoil fold that resembles the ricin toxin fold [84]. This fold has three antiparallel β-sheet repeats with three potential binding sites for ligand interaction which is optimal for multivalent interaction with target ligands. In both types B and C CBMs, interactions with ligand are mediated by hydrophobic stacking interactions between aromatic side chains and the face of the sugar molecules. Unlike in type A CBMs, direct and water-mediated hydrogen bonds play a significant role in ligand binding. Classification of a CBM into type B or C includes the number of subsites within the binding site, where 1-3 subsites are classified as type C and >3 subsites are type B. The number of direct hydrogen bonds formed per Å of buried polar surface area is another criterion; type Cs follow a lectin-like pattern of hydrogen bonding with ~3.7 hydrogen bonds per 100 $Å^2$ of buried polar surface area, while type B CBMs have ~2 hydrogen bonds per 100 $Å^2$ of buried polar surface area. Reasons for this remain unknown, however, they may involve the role of the bulk solvent in protein-ligand interactions by the different CBM types as well as the need to accommodate highly decorated plant cell wall ligands [56]. Sometimes classification into types can be ambiguous, as seen with the starch-binding families 20, 25, 26, 34, 41. Modules from these families fall between type B and C as they have folds and extended binding pockets similar to type B CBMs, however, they have a hydrogen bonding pattern similar to type C with ~3.4 hydrogen bonds per 100 $Å^2$ of buried polar surface with only two subsites for direct interaction with glucose molecules [63; 85].

### 1.4.2 Plant specific CBMs: a historical perspective

CBM research originated in the late 80's with the discovery that limited proteolytic cleavage of cellobiohydrolase I (CBHI) from *Trichoderma reseei* yielded two

functional domains, one that acts as a binding site for insoluble cellulose at the carboxy terminus and another, termed the protein core, which contains the active site for hydrolytic activity on cellulose [54]. However, only the hydrolytic activity of the protein core on crystalline cellulose was affected whereas its activity on smaller molecular mass substrates remained the same, suggesting that the C-terminal domain aids in adsorption of the enzyme onto crystalline cellulose. Further studies on *T. reseei* cellobiohydrolase II (CBHII) revealed a similar binding domain at the N-terminus which was also involved in adsorption of the enzyme onto cellulose. Researchers were also able to identify the modular boundaries of these binding domains in CBHI and CBHII and suggested that these modules were important in synergism with the catalytic core in hydrolyzing cellulose. They first proposed that these "secondary substrate binding sites" are key to efficient cellulose hydrolysis [55]. Preliminary structural studies of both CBHI and CBHII using small angle x-ray scattering (SAXS) showed that the enzymes were tadpole shaped with an ellipsoid hydrolytic "core" and an elongated tail comprised of the binding domains which were in a position to anchor the hydrolytic core onto cellulose [86; 87]. Studies on several other cellulolytic enzymes from bacterial and fungal origin also identified similar binding domains with independent binding function. These experiments confirmed that cellulolytic enzymes contained discrete domains that fold independently of one another and work synergistically to effectively break down cellulose. These domains became known as cellulose binding domains (CBDs) and were grouped into families based on sequence similarities and binding properties (CBD I – XIII) [88]. Soon after CBDs were identified, similar secondary binding domains were found in enzymes that were active on other plant cell wall hemicellulose [89; 90]. A new classification system

for these domains was established to include domains with specificity for polysaccharides other than cellulose and they became known as carbohydrate binding modules (CBMs). The family base classification system of CBMs initially established in 1999 has since grown to include 52 amino acid sequence based families (www.cazy.org) [22]. Of the 52 sequence based families, at least 36 are involved in recognizing plant cell wall glycans.

### 1.4.3 CBMs and complex glycans: the wave of the future

It has long been known that CBMs aid in the efficient degradation of plant cell wall polysaccharides by glycoside hydrolases. More recently it has become apparent that CBMs are also potentially involved in the degradation of complex glycans by glycoside hydrolases from pathogenic bacteria in human hosts [91; 92; 93]. Recently, new CBM families have been discovered in secreted or surface-associated glycoside hydrolases from bacteria and these glycoside hydrolases are often key virulence factors in pathogenesis [80,81]. CBMs that bind to complex human glycans belong to the families 32, 40, 47, and 51. They have demonstrated binding function on complex sugars such as sialylated glycoproteins, blood group A/B antigens and Lewis$^{Y}$ antigen. Sialidases, or neuraminidases, are key virulence factors in bacteria and viruses and have been shown to remove terminal sialic acid residues from complex glycans, unmasking receptors for invasion into host cells [94]. Often these enzymes have appended CBMs that aid in the removal of sialic acid, such as the large sialidase toxin with CBMs from family 32 and 40 that interact with galactose and sialic acid respectively, allowing the enzyme to be targeted to glycan regions containing these sugars [95]. In fact, many exotoxins secreted by *C. perfringens* contain family 32 CBMs. A detailed study of these CBM32s showed that they are able to interact with terminal galactose commonly found in highly decorated N

and O-linked glycans, such as LacNAc and type II H-trisaccharide (a precursor to the blood group A/B antigens) [51]. Their role in pathogenesis appears to allow for colonization of mucosal surfaces and spread into surrounding tissues, utilizing the carbohydrates as a nutritional source by the bacteria.

Blood group antigens also are a target for bacterial virulence factors. A family 98 GH from the fucose utilization operon, a known virulence factor in *Streptococcus pneumoniae,* contains a triplet of CBM47s at the C-terminus which interact with fucosylated sugars found in the ABH blood group antigens and Lewis$^{Y}$ antigen as well as with the surface of mouse lung tissue [80]. It is speculated that virulence is conveyed through the catalytic activity of the enzyme on host lung tissue. Recently a new family, CBM51, was identified in a putative α-fucosidase and blood group specific endo-β-galactosidase exotoxins of *C. perfringens* [81]. Their specificity for host glycans also conveys the importance of these CBMs in pathogenesis of the organism.

The combined effect of CBMs from glycoside hydrolases in the recognition of host glycans by bacteria for pathogenesis, colonization, as a nutritional source, and evading the hosts immune system, defines a new avenue of CBM research apart from plant cell wall recognition.

## 1.5 Relevance of PhD Research

### 1.5.1 Evolution of CBM research

Initially CBMs were identified by proteolytic cleavage of glycoside hydrolases active on plant cell walls (see Section 1.5) where additional binding domains attached to the hydrolytic core enhanced the catalytic activity of the enzyme. Subsequent

experiments to find additional domains with similar function within glycoside hydrolases included proteolytic cleavage of enzymes and recombinantly producing truncated enzymes lacking the binding domain. These experiments demonstrated that enzymatic activity on substrate decreased for the truncated enzymes as compared to wild type enzyme and established the importance of CBMs in polysaccharide degradation.

In addition to determining the presence and function of these CBMs within the context of glycoside hydrolases, research in the 1990's began focusing on the structural properties of CBMs. The first structure of an independent CBM was the NMR structure of the C-terminal CBD (or CBM family 1) from *T. reseei* CBHI in 1989 (now known as a Type A CBM) [82]. Following were several NMR and X-ray crystal structures on several CBMs from different CBM families. The general goal for obtaining structural data at the time was to establish the fold for each CBM family by obtaining a structure for one or two members of a given family. Because all members of a given CBM family share a high degree of amino acid sequence homology, the fold would be representative of the overall fold of a family.

Once folds were established and the ability to obtain structural data became more conventional, research focused on the structural basis of ligand recognition by CBMs. By 2000, only two CBM crystal structures were solved in complex with ligand (family 13 and 18, ricin B chain and WGA respectively)[96; 97] and one NMR structure of a family 20 CBM in complex with β-cyclodextrin [98]. Only since 2001 have CBM structures in complex with ligand become a key aspect of CBM research. The structures allowed for the observation of the molecular determinants that drive a tight binding interaction and

established the importance of hydrogen bonding networks and hydrophobic stacking interactions between the sugar and amino acid side chains within the binding pocket.

Previously, all work on CBMs has involved looking at a single member within a given family. The intent of this PhD work beginning in 2004 was to look at diversity of ligand recognition within a CBM family by obtaining structural and biochemical data for multiple CBMs in a given family. This has allowed us to observe how different members within a CBM family impart specificity for their ligand while maintaining similar folds and amino acid sequences. Family 6 and family 41 were our representative families. CBM family 6 is exemplary because they share similar amino acid sequences and overall structural folds and binding sites but members bind to a structurally diverse range of plant cell wall polysaccharides, including cellulose, xylan, β-1,3-glucans and mixed β-glucans. *Our objective was to study the molecular basis of ligand recognition by CBM6s and to further elucidate how family members accommodate the variability in plant cell wall polysaccharide structure* (see Section 2).

Family 41 is a new CBM family (2004) that interacts with α-glucans but is distinguishable from other α-glucan binding CBM families found in glucanases and amylases. Members from family 41 are mainly found in pullulanases, also known as starch-debranching enzymes, and have evolved a binding site that is able to accommodate α-1,6-linkages found in pullulan. *Our objective was to study the molecular basis of α-glucan recognition by CBM41.* By studying multiple members within CBM family 41, we have been able to observe how they have evolve binding sites suited for interacting with pullulan compared to starch which is primarily α-1,4-linked glucose, and also

observed a novel bivalent architecture that is optimally suited for interacting with α-glucan chains (see Section 3).

### 1.5.2 Evolution of starch degradation

Bacterial starch recognition has long been established as a means of biomass conversion using plant-based starch granules as a carbohydrate source. Activity of bacterial, fungal and yeast amylases, glucanases and cyclodextrinases on starch granules break down starch into smaller glucose units that can be utilized by the organism as a nutritional source. Its activity is also exploited as a means of producing ethanol in the production of food and biofuels. Therefore research on bacterial α-glucan active enzymes has focused primarily on starch degradation from environmental sources. Our research on the α-glucan binding Family 41 CBMs from bacterial pullulanases identified additional family members mainly from bacteria that are human pathogens, such as *Streptococcus pneumoniae, Streptococcus pyogenes,* and *Klebsiella pneumoniae* [85]. Often they are essential for viability of the organism in their host. Since some of these pathogenic bacteria have no known environmental niche, *it was our objective to study the mechanism of α-glucan degradation by a pullulanase from* S. pneumoniae *and how it may contribute to virulence of the organism.*

The N-terminal CBM41 modules of the pullulanases SpuA and PulA from *S. pneumoniae* and *S. pyogenes* respectively have demonstrated glycogen binding activity and, like the fucose specific CBM47s [80], interact with mouse lung tissue, however, were shown to localize specifically to glycogen granules in type II alveolar cells (see Section 3.4). In addition to providing the bacteria with a nutritional source, glycogen degradation also appears to be a means of evading the host immune system during invasion (see

section 4). This may have pharmaceutical importance in developing new drug targets to combat *Streptococcal* infections. This research is the first to establish starch degradation activity as a means of bacterial virulence, expanding the field of α-glucan active enzymes to include activity on starch from animal sources by pathogenic bacteria.

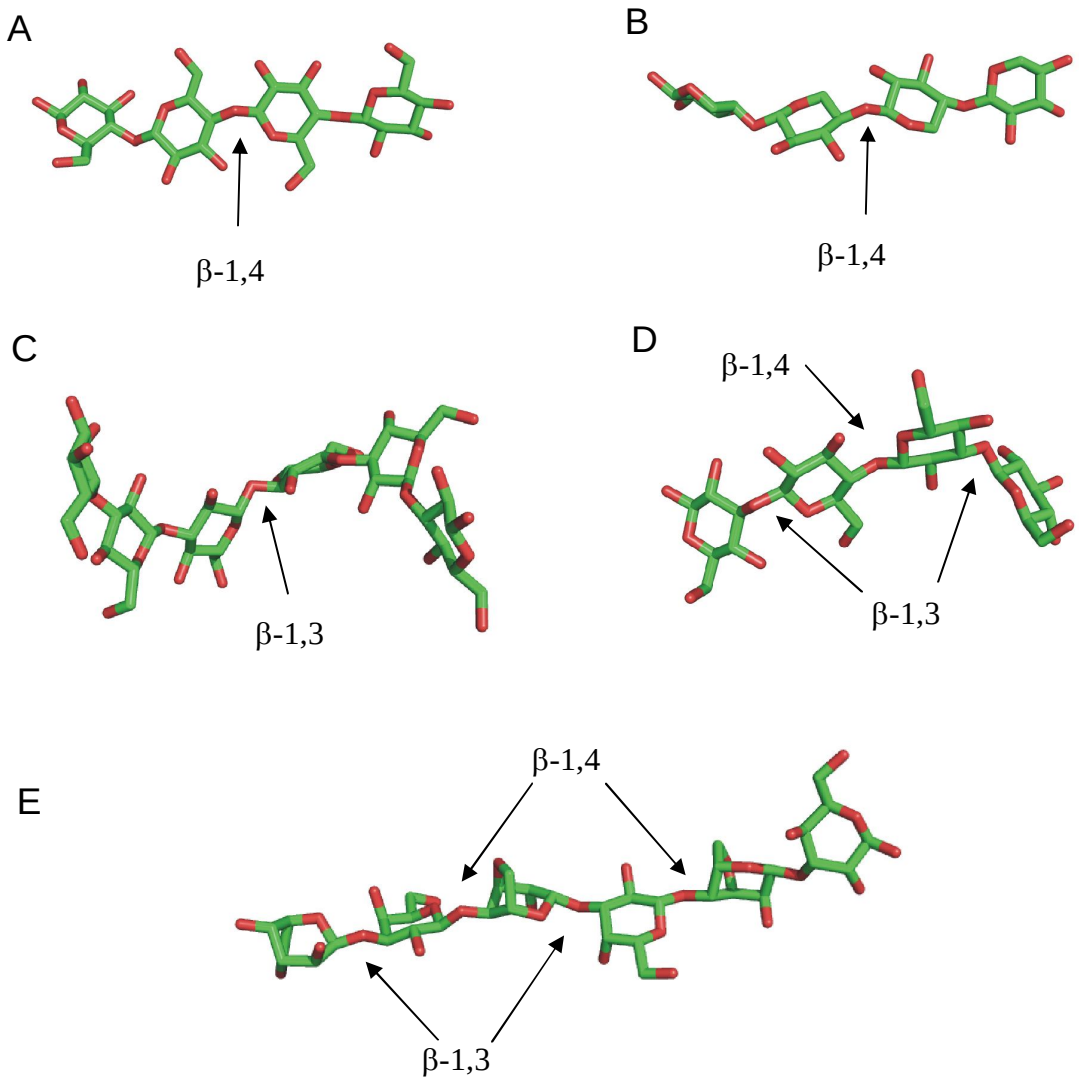
# Chapter 2: Molecular Determinants of Carbohydrate Recognition by the β-Glucan Binding Family 6 CBM's

## 2.1: Introduction

Every sugar has a three-dimensional structure whose shape is determined by the different linkages between each monomer. The three-dimensional structure of polysaccharides is important for plant cell wall structure. For example, the β-1,4-linkages between glucose monomers in cellulose form linear polymers that are suitable for self association, forming rigid fibrils that provide the majority of the tensile strength to the cell wall. The hemicellulose xylan is β-1,4-linked xylose, and the loss of the C6 hydroxymethyl group causes the polysaccharide to form a three-fold linear helix [99]. It is closely associated with cellulose fibrils to further increase strength of the plant cell wall. Other plant polysaccharides include β-1,3-glucans such as laminarin which form a large U-shaped coiled structure while mixed β-1,3-1,4-glucans such as lichenan have an extended two-fold helix (Figure 5). These contribute to an overall triple-helix structure within the plant cell wall. The three dimensional structure of a polysaccharide is key when discussing the specificity of a CBM-carbohydrate interaction as these polypeptides have evolved binding pockets that are contoured to the shape of the ligand, driving high specificity and affinity interactions.

CBM family 6 is a large family containing approximately 150 members from ~35 different types of enzymes, mainly from bacterial origin. They are associated with enzymes that have activity on a diverse range of β-linked plant polysaccharides

Figure 5: Three dimensional shapes of some plant polysaccharides. (A) cellulose (beta-1,4-glucose) (1J84) (B) xylan (beta-1,4-xylose) (1UY4) (C) laminarin (beta-1,3-glucose) (1W9W) (D) lichenan (mixed beta-1,3-1,4-glucose) (1UYO) (E) agarose (3,6-anhydro-α-L-galactose-(1,3)-β -D-galactopyranose) (2CDP).

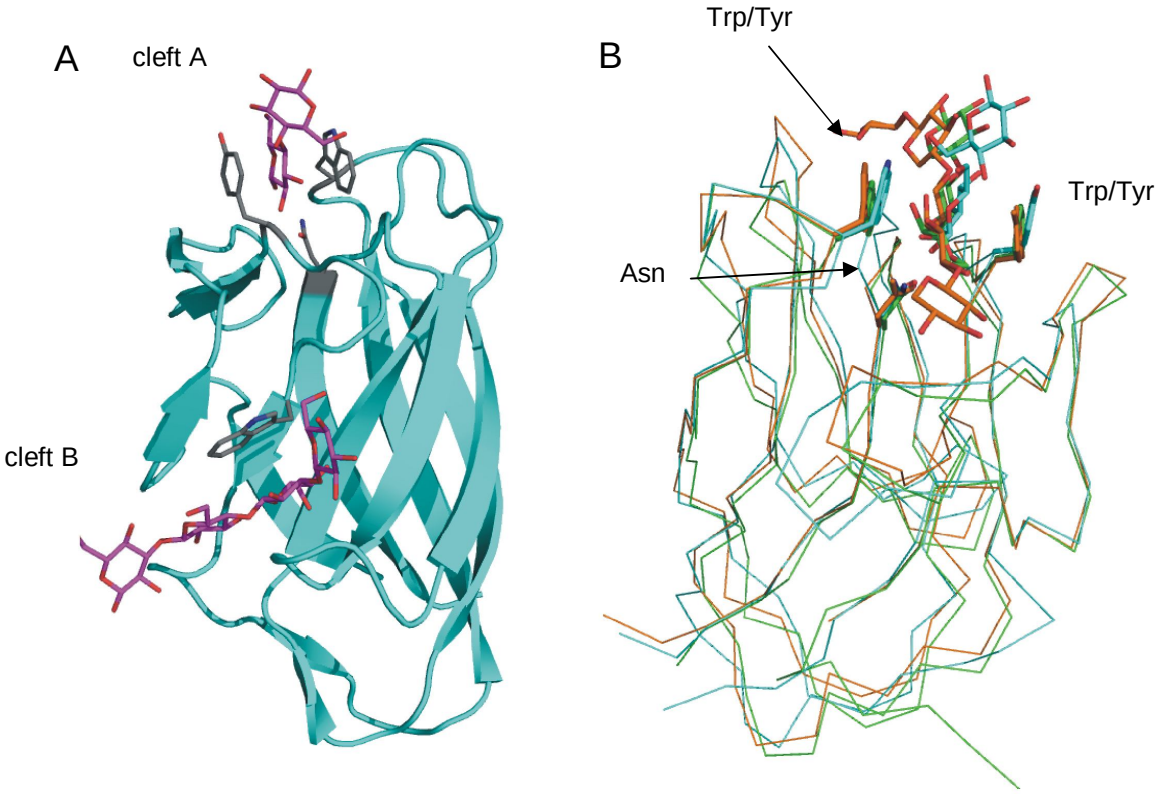


(www.cazy.org) [22] and ligand specificity often parallels the target substrate of the catalytic module. Before the implementation of the CBM classification system, this family was known as the cellulose binding domain (CBD) family VI with cellulose binding function demonstrated on amorphous cellulose and xylan [100]. CBM6s are approximately 120 amino acid residues in length and share similar amino acid sequences, with ~50-70 % sequence similarity and 20-60% sequence identity.

The first CBM6 structure of the C-terminal CBM6 of *Clostridium thermocellum* xylanase Xyn11A was solved in 2001 [101]. *Ct*CBM6 was shown to have a β-sandwich fold with a lectin-like β-jelly roll topology and two possible binding clefts as indicated by solvent exposed tryptophan residues; cleft A on the apex of the protein within the loops connecting the β-strands and cleft B on the concave face of the jelly roll (Figure 6A). Mutagenic studies determined that cleft A was the binding site, which distinguished this family of CBMs from other families with similar folds such as CBM4, 15, 17, and 22 where ligand binding occurs across the face of the protein in "cleft B" [56]. *Ct*CBM6 bound to xylooligosaccharides with high affinity and to cellohexaose with ~100x lower affinity. The affinity for xylooligosaccharides increased as the degree of polymerization increased up to xylopentaose indicating at least 5 possible subsites for interacting with xylan. In 2003 the structure of *Cs*CBM6-3, a module belonging to a triplet of CBM6s found in a *Clostridium stercorarium* xylanase Xyn11A, was solved in complex with xylotriose, cellobiose and laminaribiose, and exhibited similar affinities for cellulose and xylan [102]. The protein displayed the same β-jelly roll fold with ligand bound within the cleft A binding site for all structures. The complex with xylotriose showed three binding subsites with the possibility of up to six binding subsites available to interact with xylan. The

possibility of ligand binding in cleft B in both *Ct*CBM6 and *Cs*CBM6-3 was prevented due to a proline residue blocking accessibility to this site. Therefore it was hypothesized that cleft B was utilized by other family 6 members whose specificity was different from these two modules. In 2004 the structure of a C-terminal CBM6 from *Cellvibrio mixtus* lichenase was solved in complex with a variety of plant cell wall glucans to address this hypothesis [103]. *Cm*CBM6-2 could accommodate cello-, and xylooligosaccharides in both cleft A and B, where cleft A binds terminal sugars and cleft B binds internal sugars. Both clefts acted synergistically in binding to insoluble cellulose [104]. Interestingly, mixed β-1,3-1,4-linked glucans only bound within cleft B with four binding subsites where the β-1,3-linkage is specifically located at subsite 4 (Figure 6A). Thus cleft B is utilized by *Cm*CBM6-2 and it appears to be specific for mixed β-1,3-1,4-glucans however this research failed to examine the interaction of a pure β-1,3-glucan. In these studies cleft A interacts indiscriminantly with both cello- and xylo-oligosaccharides. In all structures so far there are three structurally conserved amino acids that interact with the sugars; two hydrophobic side chains (Trp/Tyr) and an Asn (Figure 6B). **This leads to the question: Does cleft A bind only to β-1,4-linked glucans and xylans?** Cleft B may also act as a binding site when it is accessible to solvent but so far has only been shown to be an active binding site in *Cm*CBM6-2 (Figure 6A). **The question is then posed: Is cleft B utilized by other members of this family?** The global question is how is specificity defined by members of family 6 CBMs for different plant β-glucans when they have similar amino acid sequences and similar protein structures and binding sites? The research in this chapter aims to address these questions. ***Our overall hypothesis is that cleft A is the***

Figure 6: Binding clefts of family 6 CBMs (A) Cleft A and B of *Cm*CBM6-2. Cleft A bound with xylotriose (1UYX) and cleft B bound with mixed β-glucan GLC-1,3-GLC-1,4-GLC-1,3-GLC (1UY0). Ligands shown in magenta and residues shown in grey stick representation. Cartoon representation of overall β-sandwich fold shown in cyan. (B) Structural overlaps of complexed *Ct*CBM6 (orange, PDB code 1UXX), *Cs*CBM6-3 (green, PBD code 1NAE) and *Cm*CBM6-2 (cyan, xylotriose PDB code 1UYX) with mainchain in ribbon to show overall structural similarities between these CBM6s. Conserved amino acids in cleft A and ligands shown in stick representation. Images generated using PyMol.



***primary binding site and the topology of the cleft A binding site is altered to accommodate the different three- dimensional shapes of ligands.*** Alterations to binding site topology are imparted through amino acid modifications within the binding site, thus altering the ligand specificity of the module. The objective of this research is to study other members of this family and acquire more structural and biochemical data for comparison with the known CBM6 data, from this a general model of how the binding site accommodates the many different plant polysaccharide ligands may be generated.

The impact of this research may have implications in industry since many CBMs are being utilized as biotechnological tools. Learning how this family interacts with ligand may be useful in engineering CBMs for use as bioprobes as well as creating designer CBMs with enhanced binding sites in generating hybrid enzymes for use in biomass conversion.

The following sections present studies on two separate CBM6s. The first study is on the first family 6 module found in the C-terminal CBM6 triplet from *Clostridium stercorarium* xylanase Xyn11A, named *Cs*CBM6-1. This study addressed the contribution of individual subsites within cleft A of *Cs*CBM6-1 by dissecting the thermodynamic driving forces of each subsite towards ligand binding to xylooligosaccharides of increasing length. The hypothesis is that binding subsites determine the position of the sugar within the binding pocket and as chain length of the ligand increases, more subsites are occupied and further contacts with the sugar are made, driving a higher affinity interaction. The next study was carried out on a C-terminal family 6 CBM from a *Bacillus halodurans* laminarinase (β-1,3-glucanase), where structural and biochemical data was obtained for comparison. Previous research on

CBM6s did not fully examine pure β-1,3-glucans and this was the first study to include structural and functional data on an extended β-1,3-glucan ligand. This study also demonstrated how cleft A binding site architecture is able to differentiate between β-1,4-linked xylan and β-1,3-linked glucan..

The results obtained from studying these two CBMs are placed in the context of data in the literature and more recently obtained, unpublished data, allowing us to make a more detailed comparison of CBM6 binding site architecture. Closer inspection of the binding sites identified 5 specific regions in cleft A where amino acid side chains contribute to forming a binding pocket whose overall topology accommodates the three-dimensional structure of its corresponding ligand.

## 2.2 Binding Sub-site Dissection of a Carbohydrate-binding Module Reveals the Contribution of Entropy to Oligosaccharide Recognition at "Non-primary" Binding Subsites

**Alicia Lammerts van Bueren and Alisdair B. Boraston**

Department of Biochemistry and Microbiology, University of Victoria, P.O. Box 3055 STN CSC, Victoria, BC, Canada V8W 3P6

Adapted from the Journal of Molecular Biology. Published 2004 Jul 16;340(4):869-79.

Contribution to work: ITC, X-Ray data collection, structure solving and structure refinement, preparation of figures and writing.

### 2.2.1 Abstract

The optimal ligands for many carbohydrate-binding proteins are often oligosaccharides comprising two, three, or more monosaccharide units. The binding affinity for these sugars is increased incrementally by contributions from binding subsites on the protein that accommodate the individual monosaccharide residues of the oligosaccharide. Here, we use *Cs*CBM6-1, a xylan-specific type B carbohydrate-binding module (CBM) from *Clostridium stercorarium* falling into amino acid sequence family CBM6, as a model system to investigate the structural and thermodynamic contributions of binding subsites in this protein to carbohydrate recognition. The three-dimensional structures of uncomplexed *Cs*CBM6-1 (at 1.8 Å resolution) and bound to the oligosaccharides xylobiose, xylotriose, and xylotetraose (at 1.70 Å, 1.89 Å, and 1.69 Å resolution, respectively) revealed the sequential occupation of four subsites within the binding site in the order of subsites 2, 3, 4 then 1. Overall, binding to all of the xylooligosaccharides tested was enthalpically favourable and entropically unfavourable, like most protein–carbohydrate interactions, with the primary subsites 2 and 3 providing the bulk of the free energy and enthalpy of binding. In contrast, the contributions to the changes in entropy of the non-primary subsites 1 and 4 to xylotriose and xylotetraose binding, respectively, were positive. This observation is remarkable, in that it shows that the 10–20-fold improvement in association constants for oligosaccharides longer than a disaccharide is facilitated by favourable entropic contributions from the non-primary binding subsites.

### 2.2.2 Introduction

Carbohydrate binding modules target polysaccharides with selectivity based on the architecture of their binding site where its shape is complementary to the shape of the ligand. For example type A CBMs have a planar binding surface to bind with crystalline polysaccharides while type B and C CBMs have binding grooves to accommodate glycan chains, with the former having elongated grooves for longer glycan chains and the latter having an abridged binding site for "lectin-like" binding of small sugar molecules. Much of the research into CBM function is driven by the desire to unravel their biological roles for their use as model systems to study fundamental aspects of sugar recognition [72; 105; 106].

It was shown previously how the binding-site topography of two evolutionarily related family 4 CBMs (both type B CBMs) seated the different ligand conformations of β-1,4-glucans and β-1,3-glucans uniquely [72]. Furthermore, the specific roles of apolar and hydrogen bonding residues in glycan recognition by type B CBMs have been investigated through site-directed mutagenesis [70; 105; 106]. These and similar studies have revealed that this class of CBM typically binds extended glycan chains through interactions between the protein and sugar at up to five binding subsites within the binding groove, each subsite accommodating an individual monosaccharide residue of the oligosaccharide. While CBMs make an ideal model system for studying this phenomenon it is not unique to CBMs: many carbohydrate-binding proteins have oligosaccharides comprising three or more monosaccharides as their optimal ligands [97; 107; 108]. Despite this, a somewhat poorly investigated question is what are the structural and energetic contributions of each subsite to ligand binding. Dissection of this may yield insight into how the free energy of carbohydrate binding is increased incrementally with increasing oligosaccharide length.

This will yield important information regarding fundamental aspects of protein–carbohydrate interactions, and it may provide a basis of information for engineering CBMs specific to particular biotechnological applications.

*Clostridium stercorarium* (NCIB 11745) harbours a putative xylanase comprising a catalytic module followed by triplicate family 6 CBMs [62]. The X-ray crystal structure of the last CBM in the triplet, *Cs*CMB6-3, has been solved in complex with xylotriose, but complexes of it with longer sugar species could not be obtained due to the influence of intermolecular crystal contacts [102]. Here, we describe the crystal structure of the first CBM in this triplet, *Cs*CBM6-1. The advantage of this CBM as a model system is that it could be co-crystallized with xylooligosaccharides of varying length without apparent interference of crystal contacts. Using a combination of X-ray crystallography and isothermal titration calorimetry (ITC) to study the interaction of *Cs*CBM6-1 with xylooligosaccharides of varying length we were able to dissect the structural and thermodynamic contributions of the individual subsites in its binding site.

### 2.2.3 Materials and Methods

Xylooligosaccharides were obtained from Megazyme International Ireland Ltd. (Bray, Co. Wicklow, Ireland). *Cs*CBM6-1 was produced and purified as described [62]. *Cs*CBM6-1 was dialyzed extensively into distilled water and then lyophilized.

*Determination of protein concentration*- The concentration of purified protein was determined by measuring UV absorbance (280 nm) using a calculated molar extinction coefficient [109]of 12,950 $M^{-1}cm^{-1}$.

*Isothermal titration calorimetry (ITC)*- ITC was performed as described, [68] using a VP-ITC (MicroCal, Northampton, MA). *Cs*CBM6-1 was resuspended in 50 mM potassium phosphate buffer (pH 7.0). Carbohydrate solutions were prepared by mass using the same stock of buffer used to resuspend the protein. Protein and carbohydrate solutions were filtered (0.2 µm pore size) and degassed prior to use. Titrations were performed at 25 °C by injecting 4–10 µl samples of oligosaccharide solutions at 3–15 mM into the ITC sample cell containing 200–600 µM *Cs*CBM6-1. The concentrations of protein were chosen such that they were in threefold or greater excess of the dissociation constants (i.e. *C*-values greater than 3) except in the case of xylobiose, where the concentration of protein was approximately equal to the dissociation constant. Because the solutions of the xylobiose and protein were prepared by mass and the concentrations of protein confirmed by UV absorbance, we have confidence in their concentrations. Furthermore, for xylobiose the regressed stoichiometry was reproducible and essentially the same as for longer xylooligosaccharides and, thus, the *n* value was judged to be reasonably accurate. A recent study suggests that these conditions are sufficient to provide acceptable confidence in the thermodynamic parameters regressed from ITC experiments with low *C*-values [110]. Heats of dilution upon titration of carbohydrate into buffer, buffer into buffer, or buffer into protein were negligible. Binding stoichiometries, enthalpies, and equilibrium association constants were determined by fitting the data to a one-site binding model with MicroCal Origin 7. All data show the average and standard deviation of three independent titrations. Binding to xylose was quantified by UV difference spectroscopy as described [111].

*Crystallization of CsCBM6-1*- *Cs*CBM6-1 was prepared for crystallization by overnight treatment with thrombin at room temperature in 25 mM Tris–HCl (pH 8.0) to remove the N-terminal $His_6$ tag. This reaction was concentrated and buffer-exchanged in a 10 ml stirred ultra-filtration device. Crystals of *Cs*CBM6-1 (25 mg $ml^{-1}$) were grown at 20 °C using the vapour-phase diffusion technique from hanging drops in 25% (w/v) polyethylene glycol 2000 monomethylether, 0.2 KSCN, 0.1 M sodium acetate (pH 4.5). Crystals of *Cs*CBM6-1 in complex with sugars were prepared by co-crystallization in the above conditions. Due to its low affinity for xylose (Table 2), *Cs*CBM6-1 would not crystallize in conditions containing a concentration of this sugar high enough to obtain a complex.

*Data collection, structure solution and refinement*- All crystals were frozen at 160 K after a short soak in artificial mother liquor supplemented with 20% (v/v) glycerol (final concentration). Crystals of *Cs*CBM6-1 belonged to the spacegroup $P4_12_12$ with general cell dimensions of *a*=83.4 Å, *b*=83.4 Å, and *c*=44.7 Å, with one protein molecule in the asymmetric unit. Data were collected with a Rigaku R-AXIS 4++ area detector coupled to a MM-002 X-ray generator with Osmic "blue" optics and an Oxford Cryostream 700. Data were processed using the Crystal Clear/d*trek software provided with the instrument. In each data set, 5% of the reflections were flagged as "free" to monitor refinement procedures [112]. The same reflections were flagged for all of the data sets. Statistics are given in Table 1 for the crystals and data sets used in the structure solution and refinement.

Table 1: Data collection and structure statistics for *Cs*CBM6-1

| | Uncomplexed | xylobiose | xylotriose | xylotetraose |
|---|---|---|---|---|
| **Data collection** | | | | |
| Resolution (Å) | 20-1.80 (1.86-1.80) * | 20-1.70 (1.74-1.70) | 20-1.89 (1.96 – 1.89) | 20-1.69 (1.75-1.69) |
| $R_{merge}$ | 0.068 (0.336) | 0.056 (0.378) | 0.070 (0.354) | 0.045 (0.345) |
| $I / \sigma I$ | 16.3 (5.6) | 17.8 (4.6) | 16.0 (5.2) | 22.3 (5.0) |
| Completeness (%) | 99.0 (100.0) | 99.6 (99.9) | 99.9 (100.0) | 99.1 (100.0) |
| Redundancy | 8.2 (7.8) | 7.7 (6.9) | 8.2 (8.0) | 8.6 (7.5) |
| | | | | |
| **Refinement** | | | | |
| $R_{work}$ / $R_{free}$ | 0.140/0.178 | 0.149/0.189 | 0.132/0.167 | 0.134/0.177 |
| No. residues | | | | |
| Protein | 132 | 131 | 132 | 132 |
| Ligand/ion atoms | N/A | 19 | 28 | 37 |
| Water molecules | 154 | 161 | 156 | 166 |
| *B*-factors | | | | |
| Protein | 26.7 | 25.2 | 27.0 | 22.8 |
| Ligand/ion | N/A | 35.5 | 43.7 | 40.6 |
| Water | 43.6 | 42.3 | 43.5 | 39.4 |
| R.m.s deviations | | | | |
| Bond lengths (Å) | 0.019 | 0.018 | 0.018 | 0.018 |
| Bond angles (°) | 1.622 | 1.687 | 1.631 | 1.679 |
| PDB Code | 1UY1 | 1UY2 | 1UY3 | 1UY4 |

*Highest resolution shell is shown in parenthesis.

The program molrep [113] was used to find a molecular replacement solution using the data for the xylotetraose complex and the crystal structure of the family 6 CBM from *Clostridium thermocellum* (PDB code 1gmm) as a search model [101]. This initial model was corrected by successive rounds of building using XtalView [114] and refinement with REFMAC [115]. The native and other *Cs*CBM6-1 complexes were built and refined using the refined xylotetraose complex as a starting point. As was done previously, the direction of the xylooligosaccharide from reducing end to non-reducing end was determined by potential hydrogen bonding of the O5 atoms and the difference between the B-factors of C5 and O5 (a relatively large discrepancy is found if the sugar is built and refined in the wrong orientation) [102], [116]. Water molecules were added using REFMAC/ARP-WARP and inspected visually prior to deposition. Unless stated otherwise, computing was done using the CCP4 suite [117]. All final model statistics are given in Table 1.

All surface area calculations were computed with GETAREA 1.1 [118].Figure 7, Figure 8, Figure 9 and Figure 11 were prepared with PyMOL (http://pymol.sourceforge.net/) and are shown in divergent stereo.

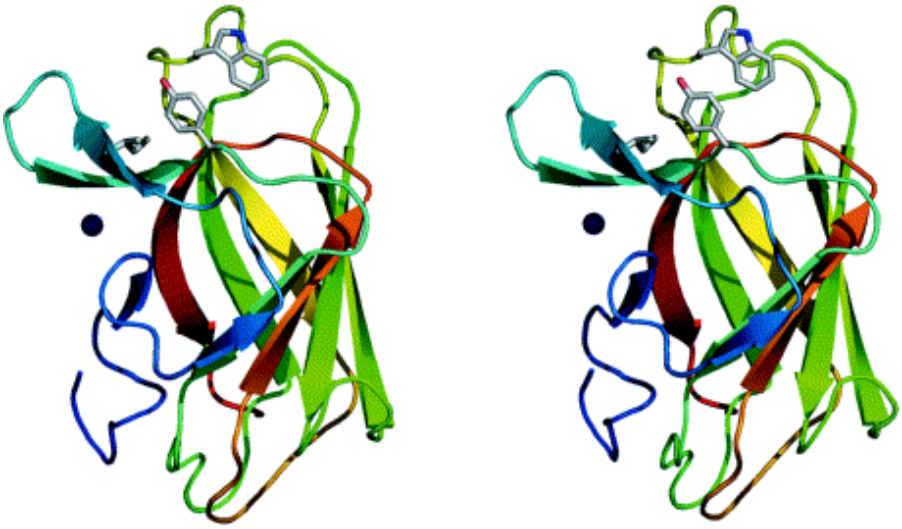
*Protein Data Bank accession codes*- Atomic coordinates and structure factors have been deposited with the Protein Data Bank and can be accessed through the PDB codes given in Table 1.

### 2.2.4 Results and Discussion

*Structure of CsCBM6-1 in the absence of ligand* - *Cs*CBM6-1 is the fourth family 6 CBM to have its three-dimensional structure determined. Like the family 6 CBM from *Clostridium thermocellum*, *Ct*CBM6, [101]*Cm*CBM6 from *Cellvibrio mixtus* endoglucanase 5A [104] and *Cs*CBM6-3 [102], *Cs*CBM6-1 is an 11-stranded β-sandwich with five β-strands forming one β-sheet of the sandwich and four β-strands forming the opposing sheet (Figure 7). Two additional β-strands form a finger-like structure that extends away from the β-sandwich core and creates a portion of the carbohydrate-binding site (Figure7).

Two regions of electron density corresponding to two metal ions were found in the electron density maps of *Cs*CBM6-1. The first, which was modelled as $Ca^{2+}$, was coordinated by the side-chains of Glu25, Glu27 (bidentate), and Asp137. Interactions with the backbone carbonyl oxygen atoms of Arg37, Asp137 and a single water molecule completed the coordination. The placement and coordinating residues of this metal ion were very well conserved with the calcium atoms in *Ct*CBM6 and *Cs*CBM6-3. An additional metal ion, modelled as $Na^{+}$ on the basis of its *B*-factor and coordination (though it may be $K^{+}$), was found at the interface of symmetry-related molecules and coordinated by groups present in both molecules.

Figure 7: Three-dimensional structure of uncomplexed *Cs*CBM6-1. The overall secondary structure of the protein is shown with the apolar amino acid side-chains in the binding site (Trp107, Tyr51, and Ile40) shown in “licorice”. The bound metal ion is shown as a blue sphere.



As expected, the structures of *Cs*CBM6-3, *Ct*CBM6, *Cm*CBM6 and *Cs*CBM6-1 were extremely similar. *Cs*CBM6-1 had a root-mean-square-deviation (rmsd) with *Cs*CBM6-3 of 0.52 $Å^2$ over 121 matched $C^{\alpha}$ atoms (as determined with PyMol). The rmsd between *Cs*CBM6-1 and *Ct*CBM6 was 0.54 $Å^2$ over 108 matched $C^{\alpha}$ atoms, while the same measurement between *Cs*CBM6-1 and *Cm*CBM6 was 2.1 $Å^2$ over 95 matched $C^{\alpha}$ atoms. The fold of family 6 CBMs is shared by CBMs in families 2, 3, 4, 15, 17, 22, 27, 28, 29, and 32 and, to a lesser degree, with lectins from a variety of sources [102].

*Structure of CsCBM6-1 in complex with xylooligosaccharides* - *Cs*CBM6-1 co-crystallized with xylooligosaccharides from xylobiose to xylohexaose. Clear electron density for each sugar residue in the ligands xylobiose to xylotetraose was evident (Figure 8). Examination of the surface of *Cs*CBM6-1 with xylotetraose occupying the binding site suggested space for an additional xylose residue at the non-reducing end of the oligosaccharide (Figure 9) But this space was not occupied by the non-reducing terminal sugar(s) of xylopentaose or xylohexaose, despite this region being surrounded by a large solvent channel leaving ample room for additional sugar residues (not shown). Clear density for only four sugar residues in xylopentaose and xylohexaose was visible, and those corresponded to the same residues as those observed for xylotetraose. Disordered density at low levels of contouring was visible extending from the reducing ends of these two oligosaccharides into the solvent channels of the crystal (not shown).

Figure 8: Observed electron density for (A) xylobiose, (B) xylotriose and (C) xylotetroase bound to *Cs*CBM6-1 (next page). All maps are maximum-likelihood-[115]/$\sigma_A$-[119]weighted $2F_{obs}-F_{calc}$ electron density maps contoured at 1$\sigma$ (0.27, 0.26, and 0.28 electrons/Å$^3$ for xylobiose, xylotriose, and xylotetraose, respectively). Trp107, Tyr51, Ile40 and Pro133 are shown in "licorice". The red sphere and its electron density indicate the water molecule at the base of the binding cleft that is conserved with *Cs*CBM6-3. The $C^{\alpha}$ trace of the protein backbone is shown in grey.

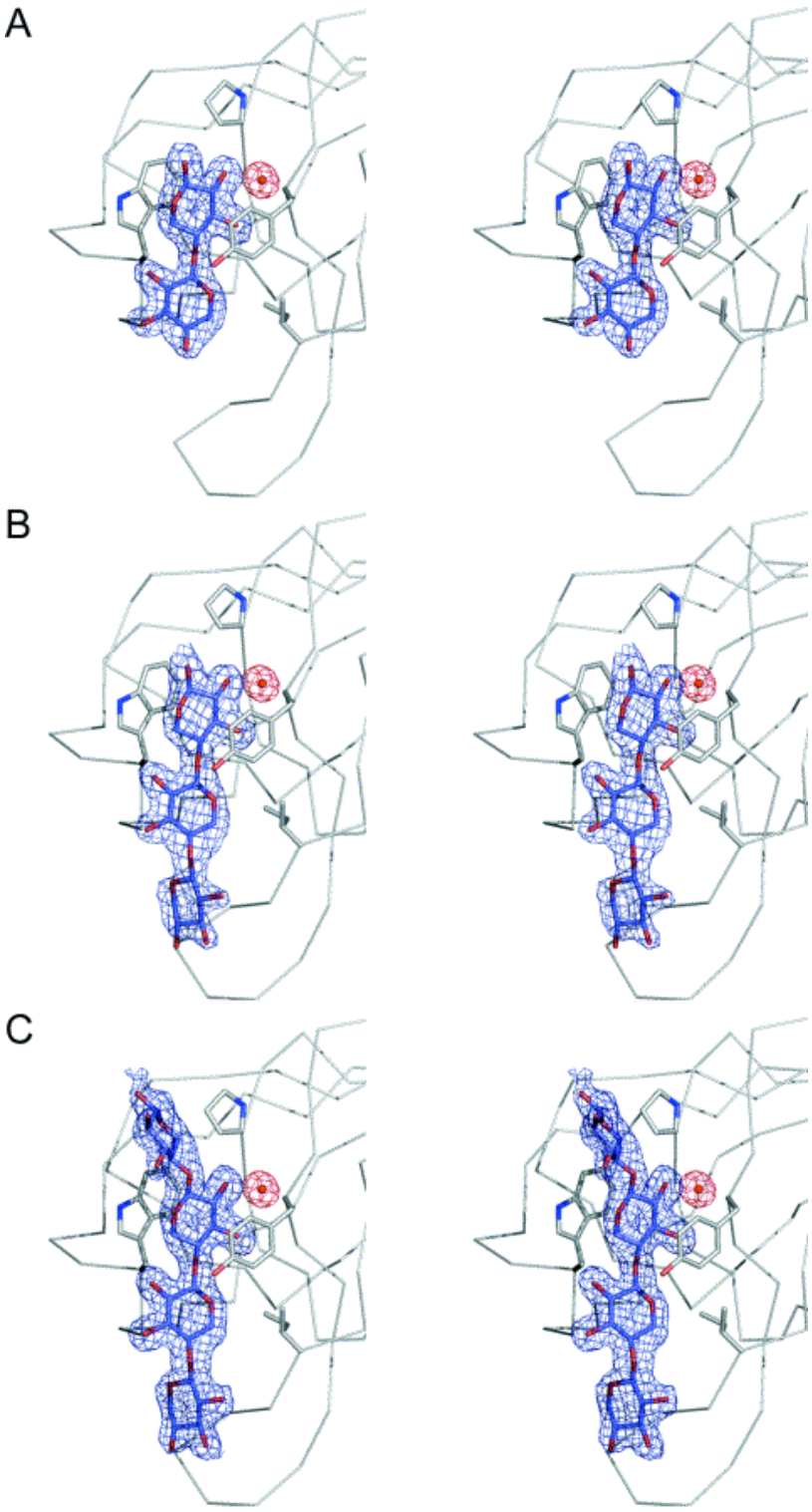
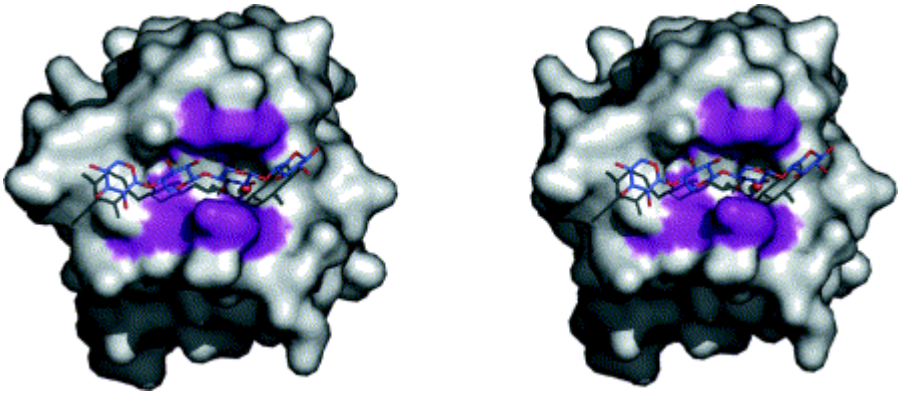

A
B
C

Figure 9: Solvent-accessible surface of *Cs*CBM6-1 complexed with xylotetraose. Purple regions indicate the surface contributed by the binding site apolar amino acid side-chains. The sugar molecule is shown in blue and red “licorice”. This surface reveals the pocket where a well-ordered water molecule (red sphere) that bridges multiple interactions between the ligand and protein is present.

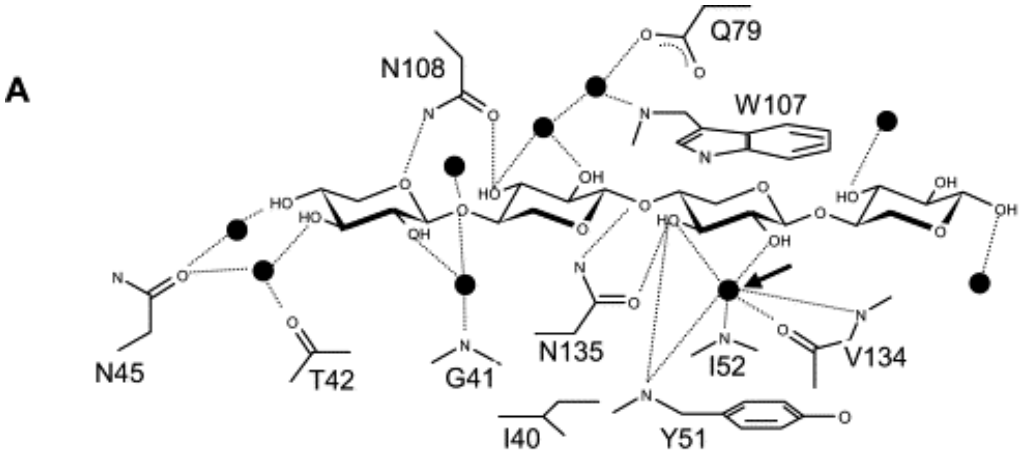


These additional residues could not be modelled confidently and, because the xylopentaose and xylohexaose complexes did not offer any additional information over the xylotetraose, these structures are not discussed any further nor were they deposited. No obvious structural differences in the polypeptide in its liganded and unliganded states could be found.

On the basis of these results, it appeared that the binding site of *Cs*CBM6-1 had four subsites that accommodated the individual sugar residues of xylotetraose (Figure 10). Increasing the sugar length from xylobiose to xylotetraose resulted in the sequential accommodation of subsites in the order 2+3, 4 then 1 (Figure 10). In the refined structures containing xylobiose, xylotriose and xylotetraose, the positions of sugar residues occupying subsites 2, 3 and 4 overlapped with only insubstantial differences (not shown). An apolar cradle was formed by Trp107 and Tyr51 in subsite 2 and Ile40 in subsite 3 (Figures 8, 9 and 10). Only five putative direct hydrogen bonds were made with xylotetraose: three in subsite 1 and one in each of subsites 3 and 4. Eight water-mediated hydrogen bonds were distributed approximately equally over subsites 2–4. Of note is the very ordered water moecule in subsite 2 that has the potential to bridge a number of interactions (Figures 8, 9 and 10). No obvious interactions between the xylose residue occupying subsite 1 and the protein were evident, except for potential van der Waals interactions with Pro133.

*Comparison with other xylan-specific CBM6s* - As mentioned, *Cs*CBM6-1 exists as part of a triplet of CBMs and we have described the structure of *Cs*CBM6-3 in complex with xylotriose. An overlap of the *Cs*CBM6-1–xylotetraose complex and the *Cs*CBM6-3–

Figure 10: A schematic showing the interactions of *Cs*CBM6-1 with xylooligosaccharides. A, The interactions with xylotetraose. Binding subsites referred to in the text are shown underneath the schematics with brackets and are numbered in accordance with IUPAC nomenclature. The water molecule conserved in the *Cs*CBM6-3 binding site is indicated with an arrow. B, The occupation of subsites by xylooligosaccharides of different lengths. The thermodynamic contributions of the subsites are given immediately below in the table. These values were obtained by subtraction of the thermodynamic values of the xylooligosaccharide with length *n* from the oligosaccharide with length *n*+1 (e.g. $\Delta\Delta H$(subsite1) = $\Delta H$(xylotetraose) − $\Delta H$(xylotriose) ). See Table 2 for a complete list of thermodynamic values determined by isothermal titration calorimetry.

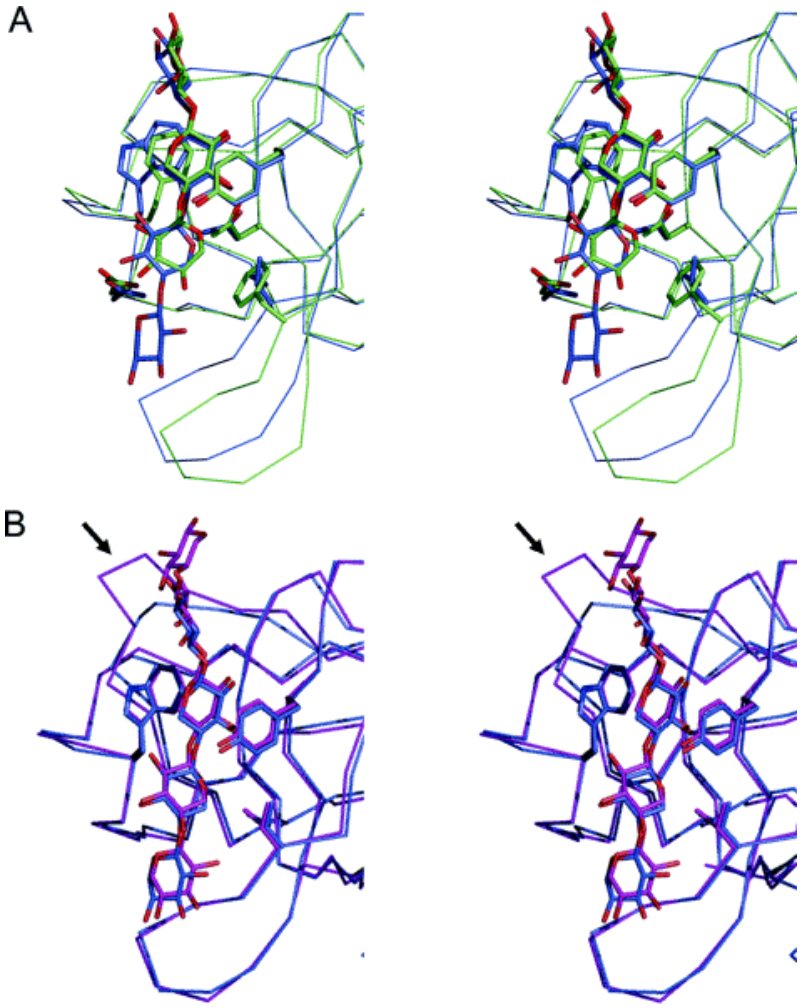



| | 4 | 3 | 2 | 1 |
|---|---|---|---|---|
| ΔΔG (kcal/mole) | -1.11 (±0.02) | -1.87 (±0.09) | -2.61 (±0.09) | -0.58 (±0.03) |
| ΔΔH (kcal/mole) | 0.40 (±0.17) | -9.46 (±0.11) | | -0.06 (±0.19) |
| TΔΔS (kcal/mole) | 1.51 (±0.16) | -4.98 (±0.01) | | 0.52 (±0.03) |

xylotriose complex shows that the xylose residues sandwiched between the aromatic amino acid side-chains in subsite 2 overlap extremely well (Figure 11). The constellation of hydrogen bonds formed in this subsite is very similar in the two proteins (not shown), including an ordered water molecule that mediates multiple interactions (this water molecule is indicated in (Figures 8, 9 and 10). The most notable difference in this subsite is the substitution of Trp107 for Phe112 in *Cs*CBM6-1 *versus Cs*CBM6-3 (Figure 11A). In subsite 3, Ile40 is substituted for Phe45 in *Cs*CBM6-1 *versus Cs*CBM6-3. This appears to cause the xylose residue in subsite 3 of *Cs*CBM6-1 to be rotated approximately 5° around an axis perpendicular to the plane defined by the pyranose ring and pushed up ~ 1–2 Å from the binding site relative to the equivalent sugar residue in *Cs*CBM6-3 (Figure 11A). The result is that ND2 of Asn135 in *Cs*CBM6-1 hydrogen bonds with the oxygen atom of the glycosidic bond between subsites 2 and 3, whereas the equivalent atom in *Cs*CBM6-3 hydrogen bonds with the endocyclic O5 of the xylose residue in subsite 3. In both proteins, subsite 1 contributes no obvious protein–carbohydrate interaction.

The structure of *Ct*CBM6 in complex with xylopentaose was solved recently and an overlap of the *Ct*CBM6 and *Cs*CBM6-1 binding sites hints at their extremely similar modes of xylooligosaccharide binding (Figure 11B). Indeed, the arrangements of protein–carbohydrate interactions are nearly identical (not shown). The primary difference is the presence of a slightly extended loop around amino acid residues 63–66 in *Ct*CBM6 that allows Asp64 and Thr65 to make additional interactions with the reducing-end

Figure 11: Overlap of the binding sites of A) *Cs*CBM6-1 (blue) and *Cs*CBM6-3 (green) with bound xylotetraose and xylotriose, respectively, and (B) *Cs*CBM6-1 (blue) and *Ct*CBM6 (magenta) with bound xylotetraose and xylopentaose, respectively. Trp107, Tyr51, Ile40, Asn135, and Asn108 of CsCBM6-1 are shown in "licorice" as are their direct counterparts in *Cs*CBM6-3: Phe112, Tyr56, Phe45, Asn140, and Asp112, respectively. Ile23, Tyr34, and Trp92 are shown in licorice for *Ct*CBM6. The loop in *Ct*CBM6 containing residues 63–66 that is discussed in the text is indicated by arrows.



monosaccharide residue of xylopentaose [104]. The creation of an additional binding subsite makes *Ct*CBM6 able to best accommodate a pentasaccharide. The presence of a fifth subsite in *Ct*CBM6 is consistent with it having a fourfold higher affinity for xylopentaose relative to xylotetraose, [101] whereas *Cs*CBM6-1 with only four subsites binds xylopentaose negligibly better than xylotetraose (~1.5-fold increase in the association constant; Table 2).

*Cs*CBM6-1 and *Cs*CBM6-3 both bind weakly to cellooligosaccharides but *Cs*CBM6-3 has the ability to bind non-crystalline cellulose, whereas *Cs*CBM6-1 does not. On the basis of the *Cs*CBM6-3 structure, we had proposed that this difference in binding specificity may be due to the Trp/Phe and Ile/Phe substitutions giving different binding-site topographies [102]. In light of the *Cs*CBM6-1 structure, which shows only very subtle differences between it and *Cs*CBM6-3, this now seems unlikely. On the basis of the recent discovery that *Cm*CBM6 has two separate binding sites, each with subtly different binding specificities,[103] we are investigating the possibility of a secondary binding site in *Cs*CBM6-3 that contributes to cellulose binding.

*Thermodynamic dissection of binding subsite properties* - The addition of xylose to *Cs*CBM6-1 induced perturbations in the UV difference spectrum with peaks at 292.2 nm, 285.2 nm, and 275.6 nm and troughs at 288.8 nm and 279.9 nm (not shown). This UV difference pattern is indicative of the movement of a tryptophan side-chain into a more apolar environment [120]. There is only one such residue in *Cs*CBM6-1, Trp107 in subsite 2, making this observation most consistent with the occupation of subsite 2 by xylose. Thus, we propose the sequential occupation of subsites in the order 2, 3, 4, then 1 with

xylose-based ligands of incremental lengths (see Figure 9 for a schematic). Using quantitative UV difference titrations, it was clear that the binding of xylose to subsite 2 was weak and had a correspondingly small free energy of binding ($\Delta G$) relative to longer sugar species (Table 2). By comparing the $\Delta G$ values of xylooligosaccharides of incremental lengths, we saw diminishing improvements of $\Delta G$ (i.e. $\Delta\Delta G$) of $-2.61$ kcal mol$^{-1}$ for the occupation of subsite 2, $-1.87$ kcal mol$^{-1}$ for subsite 3, $-1.11$ kcal mol$^{-1}$ for subsite 4, and $-0.58$ kcal mol$^{-1}$ for subsite 1 (Figure 9B). Clearly, occupation of subsites 2 and 3 provides the majority of the free energy of binding for the interaction of *Cs*CBM6-1 and xylooligosaccharides.

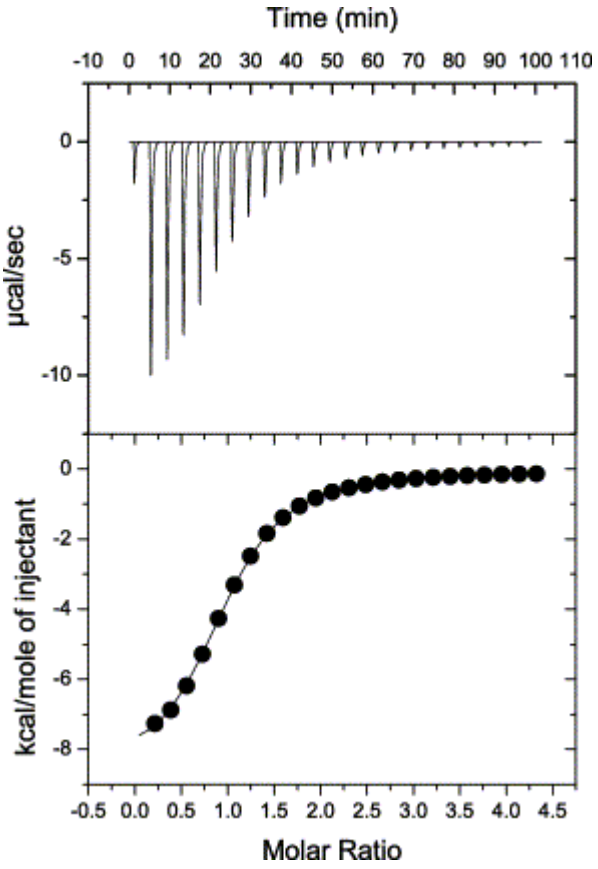
The titration of xylotetraose into *Cs*CBM6-1 resulted in the release of substantial heats as measured by isothermal titration calorimetry (ITC) at 25 °C (Fig 12). Analysis of the ITC binding isotherms for xylobiose, xylotriose, and xylotetraose indicated enthalpically favourable binding with partially offsetting unfavourable changes in entropy ($\Delta S$; Table 2), a thermodynamic signature that was broadly similar to the interactions of all soluble glycan binding CBMs studied to date [111; 121]. There was little difference in the changes in enthalpy ($\Delta H$) for these sugars, indicating very small unfavourable or negligible contributions to $\Delta H$ from subsites 4 and 1, respectively (Fig 10B). Therefore, subsites 2 and 3 provide the majority of the $\Delta G$ of binding and they provide the bulk of the driving $\Delta H$. This may not be surprising considering that these two subsites

Table 2: Thermodynamics of *Cs*CBM6-1 binding to xylooligosaccharides determined by isothermal titration calorimetry at 25 °C in 50 mM potassium phosphate (pH 7.0)

| Ligand | Ka ($x10^4$ $M^{-1}$) | ΔG (kcal $mol^{-1}$) | ΔH (kcal $mol^{-1}$) | ΔS (cal $mol^{-1}$ $K^{-1}$) | *n* |
|---|---|---|---|---|---|
| Xylose[a] | 0.01 (± 0.00) | -2.61 (±0.09) | ND | ND | 1[a] |
| Xylobiose | 0.19 (± 0.02) | -4.48 (±0.02) | -9.46 (±0.11) | -16.70 (±0.33) | 0.89 (±0.01) |
| Xylotriose | 1.27 (±0.01) | -5.59 (±0.00) | -9.06 (±0.12) | -11.63 (±0.43) | 0.95 (±0.01) |
| Xylotetraose | 3.48 (±0.09) | -6.17 (±0.03) | -9.11 (±0.15) | -9.88 (±0.56) | 0.93 (±0.01) |
| Xylopentaose | 5.29 (±0.13) | -6.44 (±0.01) | -7.92 (±0.08) | -4.96 (±0.28) | 0.93 (±0.01) |

[a] stoichiometry was set as 1 in the data analysis

Figure 12: An isotherm of *Cs*CBM6-1 binding to xylotetraose obtained by isothermal titration calorimetry at 25 °C in 50 mM potassium phosphate buffer (pH 7.0). The upper panel shows the raw calorimetric data. The lower panel shows the integrated data (spheres) and the fit of a bimolecular interaction model (continuous line).



contribute four of the five direct hydrogen bonds and five of the eight water-mediated hydrogen bonds. The occupation of subsites 2 and 3 by xylobiose results in the burial of 166 $Å^2$ and 284 $Å^2$ of polar and apolar surface area, respectively. Current structural energetic predictions derived from lectin–carbohydrate interactions [107; 122]and applied to the *Cs*CBM6-1–xylobiose interaction suggest that burial of such surface area would result in a binding $\Delta H$ of ~−6 kcal, only two-thirds of the observed $\Delta H$. This disagreement between structural energetic predictions based on lectin–carbohydrate interactions and our experimental measurements may arise from differences in hydrogen bonding between lectins and CBMs or, more enigmatically, differences in the participation of solvent in binding.

Under these conditions, the majority of the enhanced $\Delta G$ for xylotriose and xylotetraose relative to xylobiose comes by virtue of favourable contributions to $\Delta S$ (i.e. positive values of $\Delta\Delta S$; Fig 10B). Relative to xylobiose, the occupation of subsite 4 by xylotriose buries an added 37 $Å^2$ and 122 $Å^2$ of polar and apolar surface area, respectively, while occupation of subsite 1 adds another 25 $Å^2$ of polar and 31 $Å^2$ apolar buried surface areas. Though, overall, xylotriose binds with a thermodynamic signature similar to that of xylobiose, the buried surface area of subsite 4 is dominated by apolar groups. The thermodynamic contribution of this site to binding is entirely entropic and might be considered consistent with the release of ordered water from the apolar surface back into bulk solvent. It would appear that the direct and water-mediated hydrogen bonds in this subsite do little to enhance $\Delta H$, as the small enthalpic contribution from this subsite is unfavourable. The contribution of subsite 1 to binding is less clear. Overall, the total buried surface area in this subsite is low (~56 $Å^2$ cf. ~150 $Å^2$ buried per xylose residue in

the other three subsites). This is consistent with the observation that the sugar residue occupying this site makes few contacts with the protein. However, it raises questions as to the aetiology of the favourable entropic contribution that results from occupation of this subsite. This may be due to the primary interaction in this subsite being between the aliphatic C5 carbon atom of the xylose residue in subsite 1 and the aliphatic carbon atoms in the ring of Pro133. Thus, like in subsite 4, the desolvation of these apolar surfaces may result in the favourable contribution to the $\Delta S$ of binding.

On the basis of the X-ray crystal studies, the reducing sugar at the terminus of xylopentaose appears to extend into the solvent and makes no contact with the protein. Yet, this sugar molecule binds more tightly than xylotetraose, due to gains in entropy with partially offsetting losses in enthalpy (Table 2). A similar phenomenon was observed recently with the family 36 CBM from *Paenibacillus polymyxa* GH43 [76]. Improvements in the free energy of binding to longer oligosaccharides in the apparent absence of interactions with terminal sugar residues have been seen with other CBMs [72; 116 123]. There is currently no verifiable explanation for this; however, it may relate to the stabilization of optimal binding conformations for non-terminal sugar residues by the presence of flanking sugars.

An interesting feature of xylooligosaccharide recognition by *Cs*CBM6-1 was the presence of only very limited enthalpy–entropy compensation. A plot of $\Delta H$ *versus* $T\Delta S$ for the four xylooligosaccharides (xylobiose to xylopentaose) yielded a slope of 0.43 ($R^2$=0.84) (not shown). With many protein–carbohydrate interactions, such a plot gives a slope that approximates to unity, reflecting nearly fully compensating changes in $\Delta H$ and

$T\Delta S$ [124; 125]. In contrast, the limited enthalpy–entropy compensation for *Cs*CBM6-1 indicates that ligand length has a more profound effect on $\Delta S$ than it does on $\Delta H$, again suggesting a highly entropically influenced binding mechanism.

*Subsites and specificity* - The structure of *Cs*CBM6-1 demonstrates that it is a type B CBM with four binding subsites comprising its xylan-binding site. These subsites are occupied in the order 2, 3, 4, then 1 by xylooligosaccharides of increasing length. Xylose appears to occupy subsite 2 and provides the "glue" for binding. This is congruous with subsite 2 providing the bulk of the protein–carbohydrate interactions in the form of stacking interactions with aromatic amino acid side-chains, direct hydrogen bonds, and water-mediated hydrogen bonds. However, complete specificity is not conferred, as only the 2′, 3′ and 4′ hydroxyl groups make hydrogen bonds with the CBM, potentially allowing glucose to bind in this subsite. In addition to making a large contribution to $\Delta G$, the occupation of subsite 3 contributes binding specificity. Steric hindrance legislates against the presence of a C6-hydroxymethyl group (e.g. such as in glucose), as would any glycosidic linkage other than β(1−4) between subsites 2 and 3. Thus, this subsite ensures that the surfaces to be "glued" together are appropriately complementary. With respect to plant cell walls, this essentially limits the tight binding of *Cs*CBM6-1 to xylan, as it is the only polysaccharide with β-1,4-linked five-carbon pyranose sugar monomers. This polysaccharide frequently has arabino- or glucurono- substituents on its O2 and/or O3 groups. The structure of *Cs*CBM6-1 in complex with xylotetraose suggests that these groups may be accommodated when on the xylose residues that occupy subsites 3 and 4, as the 2′ and 3′ hydroxyl groups of these residues are solvent-exposed. However, it is

clear that subsite 4 contributes less free energy to binding, suggesting that the decorations on xylan would be better tolerated in this position than in subsite 3.

*Implications* - To our knowledge, this is the first study of a protein–carbohydrate interaction that marries X-ray crystallography and ITC to dissect out the structural and energetic contributions of binding subsites. The results show that the binding subsites were occupied in an orderly and sequential manner. In this case, where the ligand was a repeating polymer of β-1,4-linked xylose, subsites were occupied in the order 2, 3, 4 then 1. Only a fragment of the optimal ligand was required to provide the bulk of the free energy of binding. This gives the impression of a "zipper-like" mode of recognition where a small portion of the ligand provides an anchor for binding at the primary subsites (i.e. subsites 2 and 3) and the interaction is stabilized as the zipper closes and the non-primary subsites (i.e. subsites 1 and 4) are occupied by additional monosaccharide residues of the oligosaccharide ligand. Somewhat remarkably, the interactions at the non-primary subsites were found to be almost entirely entropic in their contributions to the overall free energy. The vast majority of protein–carbohydrate interactions, including *Cs*CBM6-1, are characterized by overall thermodynamic signatures of favourable changes in binding enthalpy, offset partly by unfavourable changes in binding entropy. In these cases, the unique contributions of subsites can be masked by the dominating energetics accompanying occupation of the primary subsites. Thus, this highlights the power of the approach used here to better understand the fine details of oligosaccharide recognition.

Overall, the Δ*G* of binding xylooligosaccharides by *Cs*CBM6-1 is built in increments of decreasing magnitude as additional subsites are occupied, a pattern that is broadly coherent with a decreasing number of protein–carbohydrate interactions in the subsites additional to the primary subsite, subsite 2. As family 6 CBMs may make a suitable platform for the directed evolution of small molecule capture agents, this study suggests that the choice of binding subsites to be evolved can be prioritized on the basis of their potential contributions to ligand binding.

## 2.3 Family 6 Carbohydrate Binding Modules Recognize the Non-reducing End of β-1,3-Linked Glucans by Presenting a Unique Ligand Binding Surface

**Alicia Lammerts van Bueren‡, Carl Morland§, Harry J. Gilbert§, and Alisdair B. Boraston‡**

From the ‡Department of Biochemistry and Microbiology, University of Victoria, P. O. Box 3055 STN CSC, Victoria, British Columbia V8W 3P6, Canada and the §School of Biomedical Sciences, University of Newcastle upon Tyne, Newcastle upon Tyne NE2 4HH, United Kingdom

Adapted from the Journal of Biological Chemistry. Published 2005 Jan 7;280(1):530-7.

Contribution to work: Crystallization of complexed protein, data collection and structure refinement, isothermal titration calorimetry, writing

### 2.3.1 Abstract

Enzymes that hydrolyze insoluble complex polysaccharide structures contain non-catalytic CBMs that play a pivotal role in the action of these enzymes against recalcitrant substrates. Family 6 CBMs (CBM6s) are distinct from other CBM families in that these protein modules contain multiple distinct ligand binding sites, a feature that makes CBM6s particularly appropriate receptors for the β-1,3-glucan laminarin, which displays an extended U-shaped conformation. To investigate the mechanism by which family 6 CBMs recognize laminarin, we report the biochemical and structural properties of a CBM6 (designated *Bh*CBM6) that is located in an enzyme, which is shown, in this work, to display β-1,3-glucanase activity. *Bh*CBM6 binds β-1,3-glucooligosaccharides with affinities of ~1 x $10^5$ $M^{-1}$. The x-ray crystal structure of this CBM in complex with laminarihexaose reveals similarity with the structures of other CBM6s but a unique binding mode. The binding cleft in this protein is sealed at one end, which prevents binding of linear polysaccharides such as cellulose, and the orientation of the sugar at this site prevents glycone extension of the ligand and thus conferring specificity for the non-reducing ends of glycans. The high affinity for extended β-1,3-glucooligosaccharides is conferred by interactions with the surface of the protein located between the two binding sites common to CBM6s and thus reveals a third ligand binding site in family 6 CBMs. This study therefore demonstrates how the multiple binding clefts and highly unusual protein surface of family 6 CBMs confers the extensive range of specificities displayed by this protein family. This is in sharp contrast to other families of CBMs where variation in specificity between different members reflects differences in the topology of a single binding site.

### 2.3.2 Introduction

It is well established that the orientation of the aromatic residues in the binding site of CBMs confers specificity for the planar and 3-fold helical conformations of cellulose and xylan, respectively [56]. The mechanism by which CBMs recognize other polysaccharides, which display more elaborate conformations, is unclear. An example of polysaccharides that exhibit complex conformations is provided by the β-1,3-linked glucose polymer laminarin, which adopts an extensive U-shaped conformation, and thus cannot be accommodated in CBMs that contain linear clefts [56]. The two potential ligand binding sites in CBM6s may present structural features that make these proteins ideally suited to accommodate polysaccharides with extended U-shaped conformations. To assess this hypothesis we have determined the structure-function relationship of a CBM6, designated *Bh*CBM6, located in an enzyme that displays laminarinase activity. *Bh*CBM6 does indeed bind to laminarin displaying maximum affinity for ligands with a d.p. >5. Uniquely, the protein displays absolute specificity for the non-reducing end of laminarin chains, and the crystal structure of *Bh*CBM6 in complex with laminarihexaose shows that the ligand extends out of cleft A interacting with the surface of the protein that does not encompass either cleft A or cleft B. These data demonstrate that the remarkable flexibility in ligand specificity displayed by CBM6s reflects variation in the location of the carbohydrate interacting sites on the surface of the protein, and is not exclusively the result of differences in the topology of a single binding site.

### 2.3.3 Materials and Methods

*Carbohydrates and Polysaccharides*—Xylooligosaccharides, laminarioligosaccharides, *Konjac* glucomannan, wheat arabino-xylan, *Tamarind* xyloglucan, and oat β-glucan were obtained from Megazyme International Ireland Ltd. (Bray Co., Wicklow, Ireland). All other carbohydrates and glycoproteins were purchased from Sigma.

*Cloning of Catalytic Domain and BhCBM6 of the Laminarinase*—The DNA fragment (nucleotides 76-2328) of the laminarinase gene (see GenBank™ AP001507; open reading frame BH0236) encoding the catalytic domain of the enzyme was amplified by PCR from *Bacillus halodurans* (C-125) genomic DNA (ATCC BAA-125) using the method of Boraston *et al.* [68] employing primers 5'-CACCTCCCCTCATGCGGTGAGC-3' CTT**TTA**GCCAATATTAAAGCTATG-3' (stop codon in bold). The amplified product was ligated into pET-151 TOPO (Invitrogen, San Diego, CA) to generate pAB1. The encoded polypeptide contained an N-terminal $His_6$/V5 epitope tag and a TEV protease cleavage site. The DNA fragment encoding *Bh*CBM6 (nucleotides 2367-2775 of the laminarinase gene) was amplified using the primers 5'-CATATGGCTAGCGATTTGAAAAATCCTTACGAG-3' (an NheI site is underlined), and 5' GCGGCCGCAAGCTT**TTA**GCCGTTTGCTCGGAAAAC 3' (a HindIII site is underlined; stop codon in bold) and cloned into NheI- and HindIII-digested pET28a to give pAB2. The encoded polypeptide contains an N-terminal $His_6$ tag and a thrombin cleavage site.

*Expression and Purification of the Catalytic Domain and BhCBM6 of the Laminarinase*—The catalytic domain of the laminarinase was produced in *Escherichia*

*coli* strain TUNER (Novagen) containing pAB1, and the protein was purified from cell-free extracts by immobilized metal ion affinity chromatography (IMAC) following the method of Freelove *et al.* [126] except the recombinant protein was eluted with 10 mM imidazole. *Bh*CBM6 was produced in 4-liter cultures of *E. coli* BL21(DE3) containing pAB2 as described previously [68], and the protein was purified by IMAC following the method of Boraston *et al.* [68]. Purified polypeptides were concentrated and exchanged into distilled water in a stirred ultrafiltration unit (Amicon, Beverly, MA) on a 5000 molecular weight cut-off (MWCO) membrane (Filtron, Northborough, MA). Purity, assessed by SDS-PAGE, was greater than 95%.

*Enzyme Activity Assay*—Enzyme reactions were carried out as described previously [127] using 0.2% substrate. The concentration of purified protein was determined by UV absorbance (280 nm) using calculated molar extinction coefficients [109] of 184,060 $M^{-1}$ $cm^{-1}$ and 36,130 $M^{-1}$ $M^{-1}$, for the catalytic domain of the laminarinase and *Bh*CBM6, respectively.

*UV Difference Titrations*—Automated UV difference titrations were performed as described previously [70]using a USB2000 CCD spectrometer (Ocean Optics, Dunedin, FL) with a diffraction grating providing measurements at 2048 approximately evenly spaced wavelengths between 234 and 395 nm. Difference spectra were examined for peak and trough wavelengths, and values at the appropriate wavelengths extracted for further analysis. The peak-to-trough heights at the wavelength pairs 289.3/301.0 nm, 289.3/294.5 nm, and 282.8/287.4 nm were calculated by subtraction of the trough values from the peak values, and the dilution-corrected data were plotted against total carbohydrate

concentration. Data for the three wavelength pairs were analyzed simultaneously with MicroCal Origin (v.7.0) using a one site binding model accounting for ligand depletion [70]. Experiments were performed at 20 °C in 50 mM Tris, pH 7.5. The data reported are the averages and standard errors of the means of three independent titrations.

*Isothermal Titration Calorimetry*—Isothermal titration calorimetry (ITC) was performed as described previously [68] using a VP-ITC (MicroCal, Northampton, MA) in 50 mM potassium phosphate buffer (pH 7.0) at 25 °C using 100-250 mM *Bh*CBM6 in the reaction cell and 1-5 mM oligosaccharide in the syringe, which gave C-values >10. Reverse titrations were performed by injecting 3-μl samples of 2.35 mM *Bh*CBM6 into 100 μM laminarin (based on an average degree of polymerization of 25; [128]). All data show the average and standard deviation of two or three independent titrations.

*Crystallization of BhCBM6—Bh*CBM6 was treated overnight with thrombin at room temperature, concentrated, and buffer exchanged into water in a 10-ml stirred ultrafiltration device using a 5000 MWCO membrane. Crystals of *Bh*CBM6 (25 mg/ml) were grown at 18 °C using the vapor-phase diffusion technique from hanging drops in 24% polyethylene glycol 2000 monomethylether, 0.2 M sodium citrate, 0.1 M MES, pH 6.5, and 3% glycerol. Crystals of *Bh*CBM6 in complex with xylobiose were prepared by adding xylobiose powder directly to hanging drops containing crystals and allowing these to equilibrate for ~72 h. Crystals of *Bh*CBM6 (20 mg/ml) in complex with excess laminarihexaose were grown using the same technique in 10-μl drops with 0.1 M MES, pH 6.5, containing 1.8 M ammonium sulfate as the mother liquor.

*Data Collection, Structure Solution, and Refinement*—All computing was done using the CCP4 suite [117] unless otherwise stated. Uncomplexed or xylobiose complexed crystals were frozen at 113 K after a short soak in artificial mother liquor supplemented with glycerol at 20% (v/v). Crystals of *Bh*CBM6 in complex with laminarihexaose were cryo-protected in the same manner with mother liquor containing ethylene glycol at 20% (v/v). Data were collected with a Rigaku R-AXIS 4++ area detector coupled to a MM-002 x-ray generator with Osmic "blue" optics and an Oxford Cryostream 700. Data were processed using the Crystal Clear/d*trek [129] software provided with the instrument. In all data sets, five percent of the observations were flagged as free [112] and used to monitor refinement procedures. In the case of the uncomplexed *Bh*CBM6 and xylobiose complex, the same reflections were flagged as free. Statistics are given in Table 3 for those crystals and data sets used in the structure solution and refinement.

Using the data for the triclinic uncomplexed crystals and the coordinates of the CBM6 from the *Clostridium thermocellum* xylanase Xyn10B (PDB ID 1GMM ; [101]) as a search model, the program molrep [113] was able to find two molecular replacement solutions corresponding to the two *Bh*CBM6 molecules in the asymmetric unit. One molecule was corrected by successive rounds of building using XtalView [114] and refinement with REFMAC [115]. This corrected model was used to replace the second molecule in the unit cell followed by additional rounds of building and refinement. This model was used directly as a starting point in the building and refinement of the xylobiose complex. Molecular replacement using the refined uncomplexed coordinates was used to solve the structure of the laminarihexaose complex. This model was corrected, and the laminarihexaose molecule was built manually in XtalView followed by refinement with

REFMAC. Water molecules were added using REFMAC/ARP-WARP and inspected visually prior to deposition. All final model statistics are given in Table 3. Figs. 16, 17, and 19 were prepared with PyMOL (see URL pymol.sourceforge.net/) and are shown in divergent stereo. The native BhCBM6, xylobiose complex and laminariohexaose complex have been deposited with the PDB codes of 1W9S, 1W9T, and 1W9W, respectively.

### 2.3.4 Results and Discussion

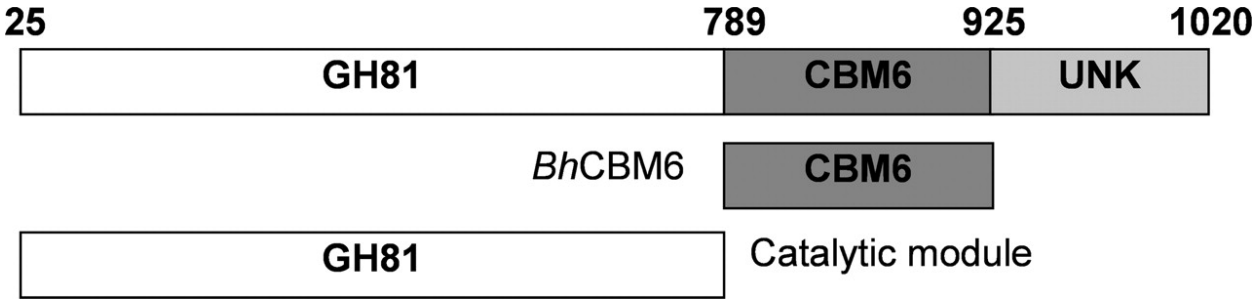
*The Modular Architecture and Catalytic Activity of the B. halodurans Laminarinase*— The alkalophilic bacterium *Bacillus halodurans* contains an open reading frame (BH0236) that encodes a 1020 amino acid of unknown function. Based on PSI-BLAST amino acid sequence alignments [130] the encoded protein appears to comprise three modules (Fig 13). The N-terminal module of this protein shows similarity with family 81 glycoside hydrolases, a family of proteins in which some members display β-1,3-glucanase activity, exemplified by Eng1p from *Saccharomyces cerevisiae*, which is involved in cell separation [131; 132], whereas others, an example of which is the Glycine max β-glucan elicitor receptor, are devoid of catalytic activity but do bind β-1,3-glucans [133]. The putative glycoside hydrolase catalytic module from the *B. halodurans* protein is the only bacterial example in this family. The recombinant catalytic module of this enzyme hydrolyzed laminarin with an activity of 14,231 $min^{-1}$ (14,231 mol of reducing sugar produced per mol of enzyme per min) but displays no detectable activity against oat-spelt xylan, wheat arabinoxylan, xyloglucan, lichenan, galactan, arabinan, carob galactomannan, konjac glucomannan, β-glucan, polygalacturoniuc acid (pectin), amylose,

Table 3: Data collection and structure statistics for *Bh*CBM6

| | Uncomplexed | Xylobiose | Laminarihexaose |
|---|---|---|---|
| **Data collection** | | | |
| Space group | P1 | P1 | $P4_32_12$ |
| Cell dimensions | | | |
| *a*, *b*, *c* (Å) | 30.8, 40.8, 51.6 | 30.7, 41.0, 56.0 | 61.6, 61.6, 121.7 |
| α, β, γ (°) | 108.9, 105.9, 90 | 108.9, 105.9, 90 | 90.00, 90.00, 90.00 |
| Resolution (Å) | 20-1.59 (1.65-1.59)* | 20-1.62 (1.68-1.62) | 20-2.10 (2.17-2.10) |
| $R_{merge}$ | 0.030 (0.168) | 0.048 (0.237) | 0.055 (0.372) |
| $I / \sigma I$ | 24.9 (5.7) | 14.9 (4.5) | 17.8 (4.4) |
| Completeness (%) | 90.1 (68.5) | 91.9 (84.4) | 95.6 (93.5) |
| Redundancy | 3.7 (3.2) | 3.6 (3.2) | 5.8 (6.0) |
| **Refinement** | | | |
| $R_{work}$ / $R_{free}$ | 0.119/0.166 | 0.121/0.179 | 0.209/0.269 |
| No. residues | | | |
| Protein | 266 | 266 | 134 |
| Ligand/ion atoms | N/A | 72 (in 4 xylobiose) | 67 ( in 1 laminarihexaose) |
| Water | 350 | 372 | 121 |
| *B*-factors | | | |
| Protein | 15.2 | 16.7 | 50.8 |
| Ligand/ion | 16.7 (Na and glycerol) | 27.7 (Na and sugar) | 49.0 (Na and sugar) |
| Water | 30.7 | 31.6 | 53.9 |
| R.m.s deviations | | | |
| Bond lengths (Å) | 0.018 | 0.019 | 0.018 |
| Bond angles (°) | 1.615 | 1.687 | 1.935 |

*Highest resolution shell is shown in parenthesis.

Figure 13: Modular organization of the *B. halodurans* laminarinase. Amino acid numbers corresponding to the module boundaries are shown *above* the schematic. The individual module constructs used in this study are also shown. *UNK* represents a module of unknown function.



amylopectin, hydroxyethyl cellulose, or carboxymethyl cellulose. Thus, the *B. halodurans* protein is clearly a β-1,3-glucanase (or laminarinase), similar to its eukaryotic homologues.

The C-terminal ~100 amino acid module has no identity to proteins of known function. Separating the C-terminal module and the N-terminal catalytic module is a module of ~ 140 amino acids module having ~36% identity to the xylan binding CBM6 from *C. thermocellum* xylanase 10A [101]. Based on its identity with CBM6s and its presence in a functional β-1,3-glucanase, we hypothesized that this module, defined as *Bh*CBM6, is indeed a CBM with β-1,3-glucan binding specificity.

*Analysis of BhCBM6 Binding Specificity*—Using a standard depletion binding analysis [79] no binding to the insoluble polysaccharides regenerated cellulose or pachyman (an insoluble β-1,3-glucan from *Poria cocos*) was evident (data not shown). Similarly, native affinity gel electrophoresis did not reveal significant binding to soluble preparations of wheat arabinoxylan, amylopectin, oat β-glucan, or lichenan, the latter two of which are commonly recognized by β-1,3-glucan binding CBMs [70; 103; 134]. Using a modificationof the macroarray method of McCartney *et al.* [57] *Bh*CBM6 showed weak binding to wheat arabino-xylan, birchwood glucurono-xylan, and pectic galactan, but did not interact with polysaccharides containing β-(1,3)(1,4) linked glucose, *i.e.* oat β-glucan (data not shown).

*UV Difference Studies of BhCBM6 Binding*—Xylose, xylooligosaccharides, and *O*-methyl-β-D-xylose induced relatively large changes in the UV absorbance difference

spectrum indicating binding to these sugars. The addition of glucose, sophorose (β-1,2-glucobiose), and β-1,3-glucooligosaccharides (laminarioligosaccharides) also gave UV difference signals (see Fig. 14A for a representative UV difference spectrum), apparently at odds with the inability to detect binding to glucose-based polymers by affinity electrophoresis and macroarray assays. No perturbation of the UV absorbance spectrum was observed when cellobiose, cellotriose, sucrose, mannose, galactose, fucose, *N*-acetylglucosamine, or *N*-acetylgalactosamine was added to *Bh*CBM6.

Three wavelength pairs in the UV difference spectra were used to monitor the dependence of the UV absorbance spectra on ligand concentration and quantify binding to carbohydrates (Fig. 14B). *Bh*CBM6 bound to glucose and xylose with association constants ($K_a$) of ~8 x $10^2$ $M^{-1}$ (Table 4). Xylobiose was bound ~2-fold more tightly indicating a small dependence of binding on ligand length. No additional gains in affinity were observed when the ligand length was increased to xylotriose (Table 4). Gentiobiose (β-1,6-glucobiose) and sophorose (β-1,2-glucobiose) were bound with affinities similar to that for xylobiose (Table 4). The β-1,3-glucooligosaccharides laminaribiose, laminaritetraose, and laminarihexaose were bound with association constants increasing from ~1 x $10^4$ $M^{-1}$ to ~1 x $10^5$ $M^{-1}$, indicating the strong preference of *Bh*CBM6 for β-1,3-glucooligosaccharides and the dependence of affinity on sugar length. Thus, despite the inability to detect binding to polysaccharides containing β-1,3-linked glucose residues by depletion binding, affinity electrophoresis, and macroarray experiments, *Bh*CBM6 does indeed appear to be primarily a β-1,3-glucan-specific CBM, consistent with its presence in a β-1,3-glucanase.

Figure 14: UV difference and ITC analysis of *Bh*CBM6 binding. *Panel A*, UV difference spectra collected with the indicated concentrations of added laminarihexaose. Peak and trough wavelengths are shown. *Panel B*, isotherm of laminarihexaose titrated into *Bh*CBM6. The curves show the data at the wavelength pairs of 289.3/301.0 (*open circles*), 289.3/294.5 nm (*closed squares*), and 282.8/287.4 nm (*closed circles*). *Solid lines* show the global fits to a one-site binding model. *Error bars* represent the standard errors of 3 measurements. *Panel C*, isotherm of *Bh*CBM6 binding to laminarin obtained by isothermal titration calorimetry (see "Materials and Methods" for experimental details). The *upper panel* shows the raw calorimetric data. The *lower panel* shows the integrated data (*closed circles*) and the results of a heat of dilution experiment (*open circles*).

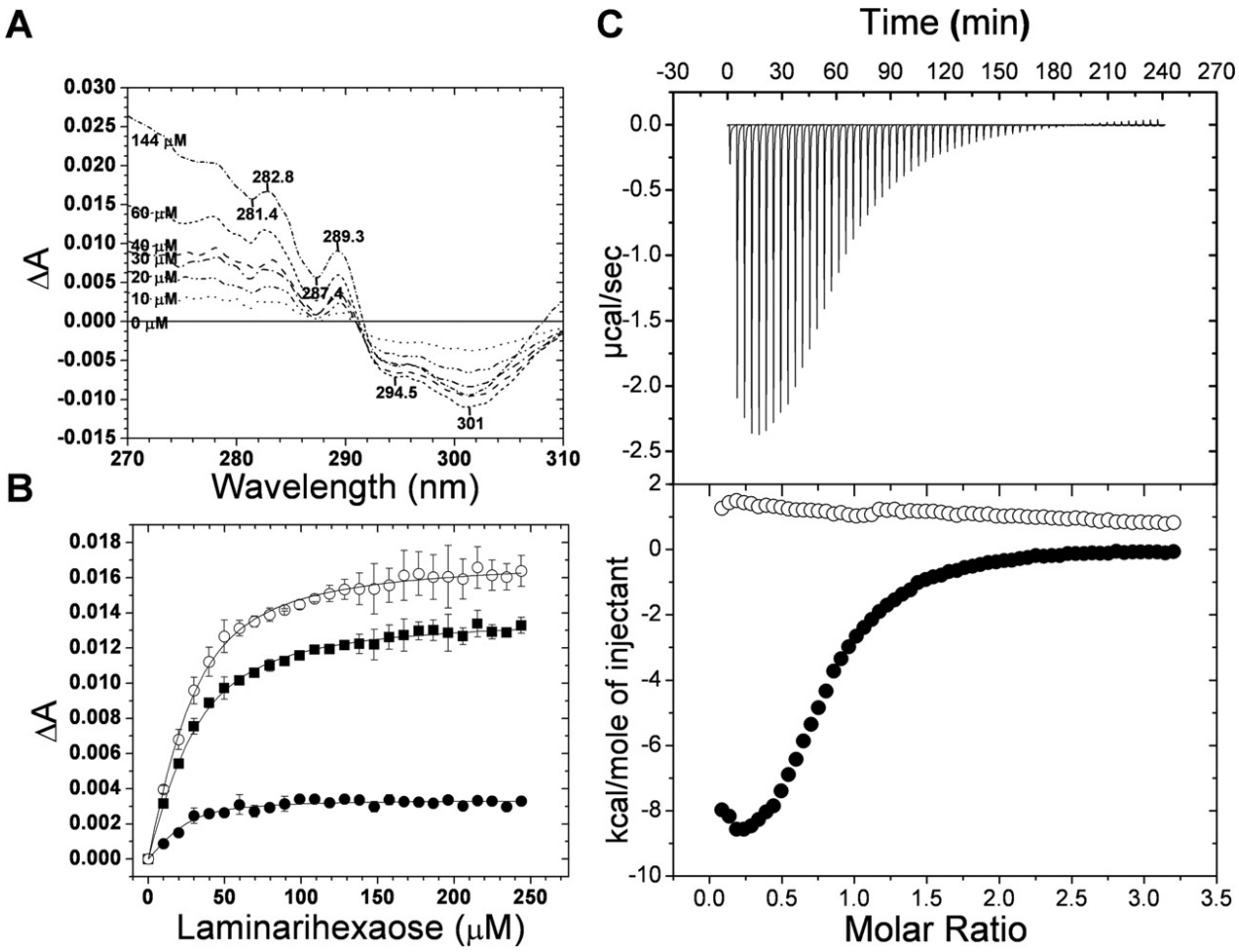


Table 4: Affinity of BhCBM6 for sugars determined by UV difference titrations at 20 °C in 50 mM Tris, pH 7.5

| Carbohydrate | $K_a$ ($x10^4$ $M^{-1}$) |
|---|---|
| β-D-Glucose | 0.1 (±0.0) |
| Laminaribiose | 0.9 (±0.1) |
| Laminaritetraose | 1.8 (±0.4) |
| Laminarihexaose | 10.2 (±2.9) |
| Gentiobiose | 0.3 (±0.1) |
| Sophorose | 0.6 (±0.2) |
| β-D-Xylose | 0.1 (±0.0) |
| O-Methyl-β-D-xylose | 0.1 (±0.0) |
| β-1,4-Xylobiose | 0.3 (±0.0) |
| β-1,4-Xylotriose | 0.3 (±0.1) |

*Thermodynamics of* β*-1,3-Glucan Recognition*—Isothermal titration calorimetry revealed the interaction of *Bh*CBM6 with β-1,3-glucooligosaccharides to be enthalpically favorable and entropically unfavorable at 25 °C (Table 5), like most protein-carbohydrate interactions studied to date. The near unitary stoichiometries indicated the formation of a simple 1:1 protein to carbohydrate macromolecular complex. Typically, CBMs that bind to extended glycan chains (Type B CBMs, [56]) show a consistent increase in binding affinity with increasing sugar length. *Bh*CBM6 showed a decrease in affinity when comparing laminaribiose with laminaritriose though the affinity consistently increased when moving from laminaritriose up to laminarihexaose (Table 5). The loss of affinity when binding laminaritriose *versus* laminaribiose appeared to be caused by an enthalpic penalty that more than offset a favorable gain in entropy. The reason for this is unclear, even in light of the structure of *Bh*CBM6 in complex with laminarihexaose (see below), which revealed no obvious peculiarity in how this third sugar interacts with the protein.

*Bh*CBM6s binding to laminarin was assessed by ITC in a mode where the protein was titrated into the polysaccharide. The resulting binding isotherms, which were highly reproducible, were not consistent with a single class of binding interaction (Fig. 14C). At least two binding phases were visually evident: one at low molar ratios (up to ~0.2) and another at higher molar ratios. A Scatchard analysis of these data, which were non-linear (not shown), confirmed the complexity of the isotherms. These results suggest either multiple classes of non-equivalent binding sites present in the laminarin population or cooperativity in the binding. We cannot conclusively comment on this and because of the apparently complex nature of laminarin recognition the kinetic and thermodynamic

Table 5: Affinity of BhCBM6 for sugars determined by isothermal titration calorimetry at 25 °C in 50 mM potassium phosphate, pH 7.0

| Carbohydrate | *n* | $K_a$ ($x10^4$ $M^{-1}$) | ΔH *kcal/mol* | ΔS *cal/mol/K* | ΔG *kcal/mol* |
|---|---|---|---|---|---|
| Laminaribiose | 0.98 (± 0.05) | 6.46 (±0.15) | -12.80 (± 0.14) | -20.80 (± 0.42) | -6.55 (± 0.01) |
| Laminaritriose | 1.08 (± 0.00) | 4.38 (±0.01) | -11.33 (± 0.02) | -16.80 (± 0.08) | -6.33 (± 0.00) |
| Laminaritetraose | 1.01 (± 0.01) | 6.59 (±0.03) | -10.89 (± 0.02) | -14.50 (± 0.07) | -6.58 (± 0.00) |
| Laminaripentaose | 1.09 (± 0.00) | 18.67 (±0.45) | -11.79 (± 0.04) | -15.40 (± 0.15) | -7.19 (± 0.01) |
| Laminarihexaose | 1.14 (± 0.00) | 29.03 (±0.11) | -11.95 (± 0.01) | -15.10 (± 0.03) | -7.45 (± 0.00) |

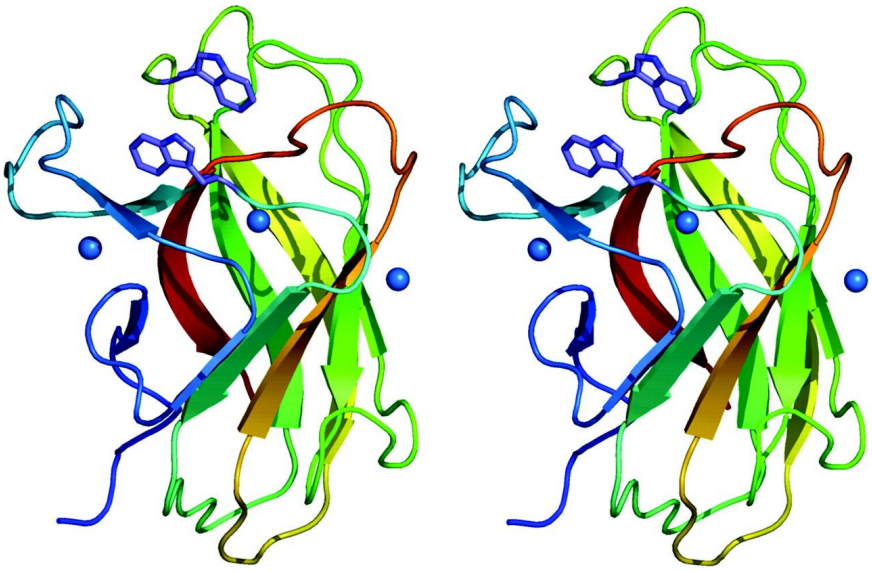
parameters of binding for the individual interactions could not be accurately deconvoluted. However, estimates from the analysis of the cumulative heats indicated association constants in the range of ~$10^5$ $M^{-1}$, consistent with the affinities for laminarioligosaccharides. The total number of binding sites was approximated by graphical determination of the stoichiometric limit of the interaction in the isotherms plotted in derivative form (standard presentation of ITC data) and cumulative form (not shown). This gave an *n* value of ~0.8 (*i.e.* ~1 CBM molecule per laminarin chain), which is consistent with *Bh*CBM6 binding only to the ends of laminarin (see below). The binding to xylan and pectic galactan was too weak to quantify by this method.

*Structure of BhCBM6* —In order to gain insight into the mechanism of carbohydrate recognition by *Bh*CBM6 we solved its three-dimensional structure by x-ray crystallography (see "Materials and Methods"). The final model of *Bh*CBM6 in the absence of ligand consisted of two *Bh*CBM6 molecules (133 amino acids each), six sodium atoms, two glycerol molecules, and 350 water molecules (refinement statistics are given in Table 3). Like other CBM6s whose structures have been determined, *Bh*CBM6 adopts a β-sandwich fold with a 5-stranded β-sheet opposing a 4-stranded β-sheet (Figure 15). This fold was highly similar to the xylan-binding CBM6 from a *C. thermocellum* xylanase (root mean-square-deviations (r.m.s.d.) of 0.81 $Å^2$ over 98 matched $C_\alpha$) [101]; the two xylan-binding CBM6s from a *Clostridium stercorarium* xylanase (r.m.s.d. of 0.71 $Å^2$ over 107 matched $C_\alpha$ and 0.83 $Å^2$ over 108 matched $C_\alpha$, respectively) [83; 102] and the glucan binding CBM6 from a *Cellvibrio mixtus*, endoglucanase (r.m.s.d. of 1.04 $Å^2$ over 95 matched $C_\alpha$) [104].

Each *Bh*CBM6 monomer coordinated 3 metal ions (Figure 15). The first is coordinated by the side chain oxygens of $Gln^{16}$, $Glu^{18}$, and $Asn^{134}$. The coordination is completed by the backbone carbonyl oxygens of $Asn^{134}$, $Asp^{38}$, and a single water molecule. The placement of this metal ion is conserved with the four other CBM6 structures, where these metal ions were modeled as calcium atoms. However, in the case of *Bh*CBM6, the B-factor refined to an unreasonably high value when this atom was modeled as calcium, suggesting a metal with fewer electrons. Based on the presence of a relatively high concentration of sodium atoms in the conditions used to crystallize *Bh*CBM6 the electron density corresponding to the metal ion was modeled as sodium. The second metal ion, also modeled as sodium, was coordinated by the side-chain oxygens of $Asn^{44}$ and $Asp^{47}$. The backbone carbonyl oxygens of $Trp^{42}$, $Gly^{25}$, and $Thr^{84}$ (of a separate, neighboring molecule) also participate in binding this atom. The third bound sodium ion is bound to the protein by one side chain oxygen of $Asp^{90}$ and four water molecules and, thus, is somewhat tenuously associated. Indeed, while the first bound sodium is likely structurally significant, as its position appears to be conserved in this protein family, the significance of the other two is unclear.

*BhCBM6 in Complex with Xylobiose*—A complex of *Bh*CBM6 was obtained by soaking unliganded P1 crystals in excess xylobiose. Electron density for two xylobiose molecules bound to each of the two monomers of *Bh*CBM6 in the unit cell was clearly evident. The secondary binding site contained a somewhat disordered sugar molecule (not shown). The xylobiose molecule in this site made few interactions with the protein, with one being a potential hydrogen bond with a second neighboring *Bh*CBM6 molecule. Thus, the biological significance of this binding site is uncertain and will not be discussed further.

Figure 15: Three-dimensional structure of uncomplexed *Bh*CBM6. The overall secondary structure of the protein is shown with the aromatic amino acid side chains in the binding site ($Trp^{42}$ and $Trp^{99}$) shown in a *licorice* representation. Bound metal ions are shown as *blue spheres*.



The other binding site, which accommodated a glycerol in the unliganded structure, is the conserved binding site among xylan binding CBM6s. The electron density for the complete xylobiose molecule bound to this site (in both of the *Bh*CBM6 monomers) allowed modeling of all of the atoms in this xylobiose molecule (Figure 16). However, because of the inability to discriminate between the positions of C-5 and O-5, the direction of the sugar (*i.e.* reducing end *versus* non-reducing end) was somewhat ambiguous. This was resolved to some extent by examining the B-factors of C-5 and O-5; one orientation resulted in large B-factor discrepancies whereas in the other they were approximately equal. Furthermore, two well ordered water molecules were observed to be properly positioned to hydrogen bond to what was assumed to be the O-5 atoms in each of the xylose residues when oriented based on the B-factors. Thus, the carbohydrate was modeled as having the non-reducing sugar sandwiched between tryptophans 42 and 99 (Figure 16), while the hydrogen bonding schematic reveals five potential direct hydrogen bonds (Figure 17). Three potential water-mediated hydrogen bonds are present with one making numerous potential interactions and is structurally conserved with water molecules in other CBM6s (Figures 16, 17, and 18A). The ability to bind in this orientation is supported by the capacity of *Bh*CBM6s to interact with *O*-methyl-β-D-xylose (Table 4). This sugar, with a blocked reducing end, bound better than its unmodified counterpart, indicating that *Bh*CBM6 must be able to accommodate the non-reducing end of the sugar. Despite this, we currently have no evidence that *B*hCBM6 cannot also bind the reducing end of xylobiose. It is apparent that the non-reducing terminus of the sugar is oriented in the binding pocket such that the oligosaccharide chain cannot extend past $Gln^{39}$ (*i.e.* over a wall of the binding pocket)

Figure 16: Observed electron density for xylobiose (*A*) and laminarihexaose (*B*) bound to *Bh*CBM6. All maps are maximum-likelihood (25)/$\sigma_A$ (38) weighted $2F_{obs}$ - $F_{calc}$ electron density maps contoured at 1$\sigma$ (0.40 and 0.13 electrons/Å$^3$ for xylobiose and laminarihexaose, respectively). Asn$^{132}$, Trp$^{42}$, and Trp$^{99}$ are shown in a *licorice* representation. The *red sphere* and its electron density indicate the water molecule at the base of the binding cleft that is conserved among CBM6s.

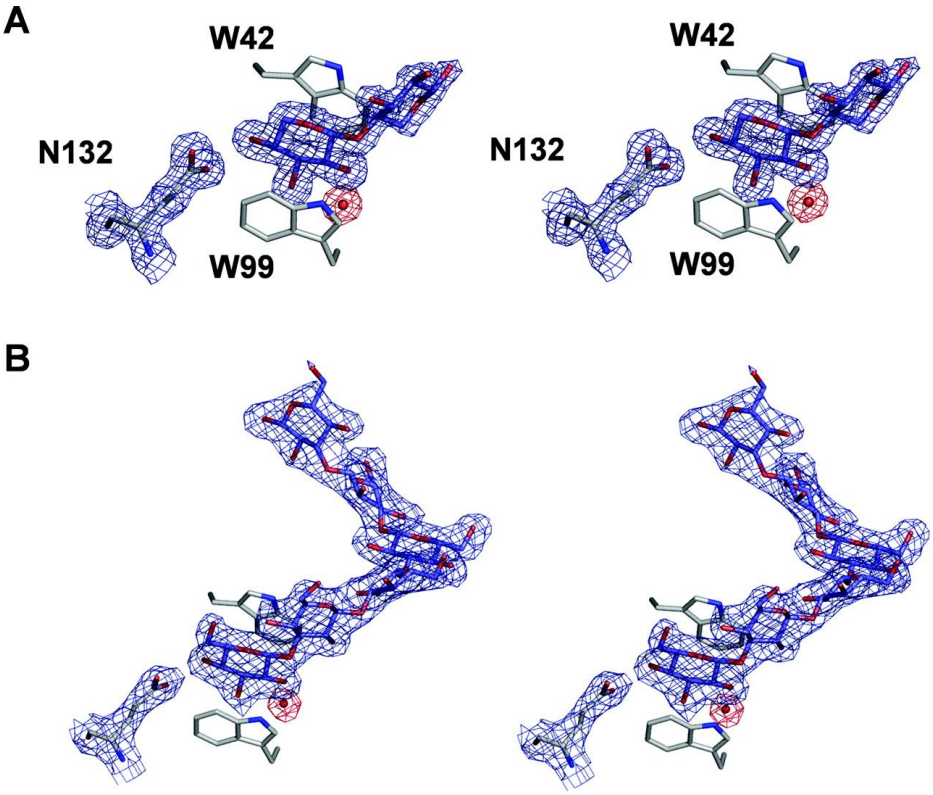


Figure 17: A schematic showing the interactions of *Bh*CBM6 with xylobiose (*A*) and laminarihexaose (*B*). Binding subsites referred to in the text are shown above the schematics with *brackets* and are numbered in accordance with IUPAC nomenclature. The water molecule conserved in the cleft A binding site of CBM6s is indicated with an *arrow*.

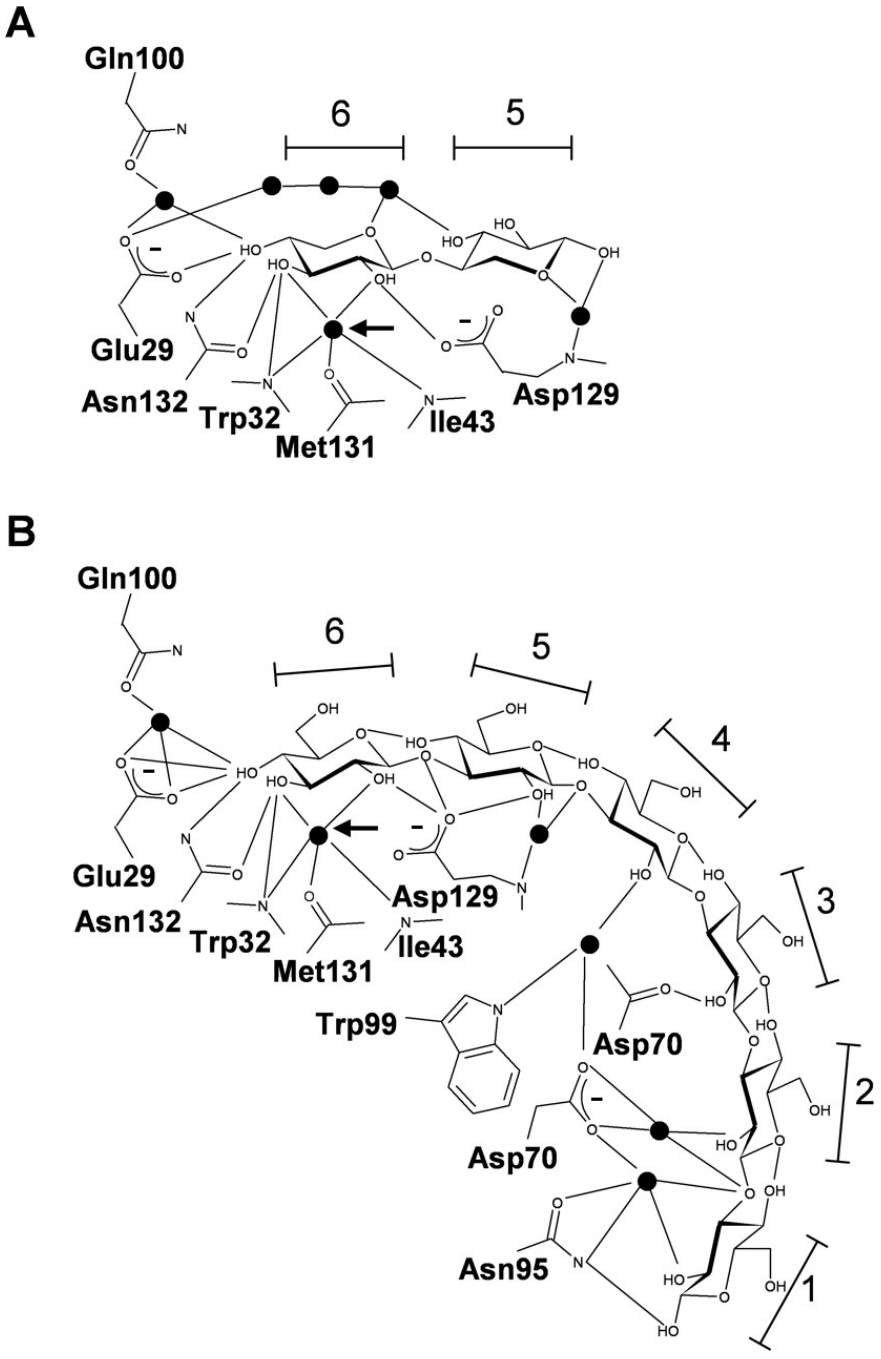
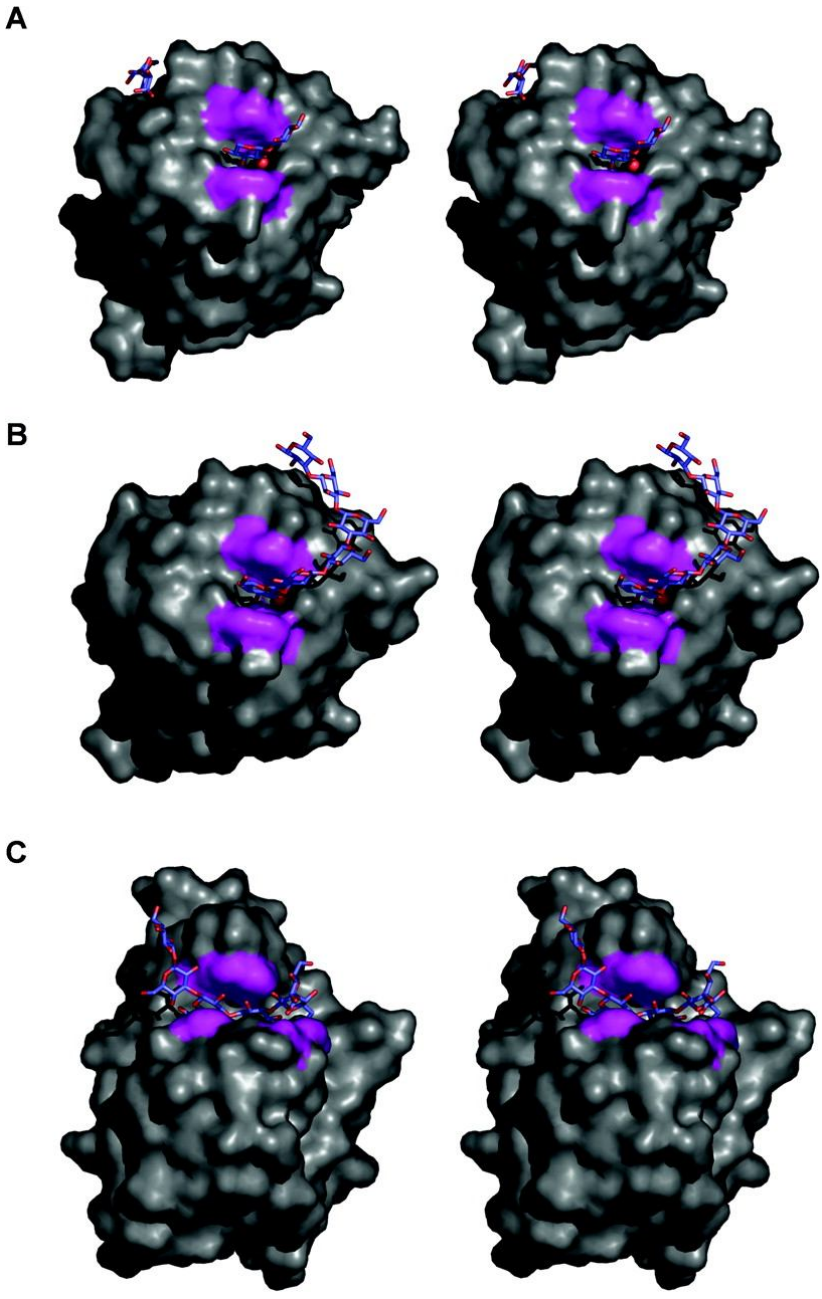


Figure 18: Solvent accessible surface of *Bh*CBM6 complexed with xylobiose (*A*) and laminarihexaose (*B*) and the family 4 CBM from *T. maritima*, *Tm*CBM4-2, in complex with laminarihexaose (*C*). *Purple regions* indicate the surface contributed by the binding site apolar amino acid side chains. The sugar molecules are shown in *blue* and *red licorice* representations. The surfaces in *panels A* and *B* reveal the pocket where a well ordered water molecule (*red sphere*) that is conserved in the cleft A binding site of CBM6s and bridges multiple interactions between the ligand and protein.



without substantial distortions to the sugar. In contrast, the reducing end appears free to extend out into solvent.

*BhCBM6 in Complex with Laminarihexaose—Bh*CBM6 co-crystallized with laminarihexaose in the space group $P4_32_12$ with a single protein and sugar molecule in the asymmetric unit. All six glucose residues of the laminarihexaose molecule could be unambiguously modeled (Figure 16). The terminal glucose residue at the non-reducing end of this oligosaccharide sandwiched between tryptophans 42 and 99, as was modeled for xylobiose (Figure 16).

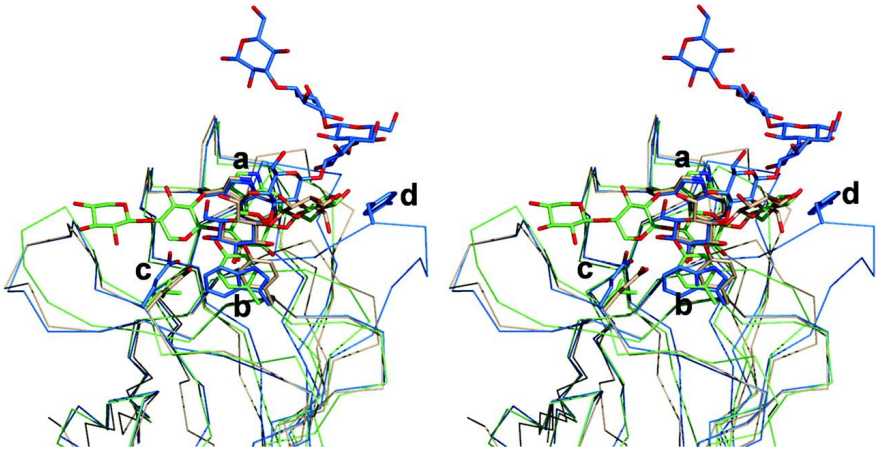
The complex of *Bh*CBM6 with laminarihexaose revealed six binding subsites (Figure 17). The extended nature of this binding site and *Bh*CBM6s preference for oligosaccharides with a degree of polymerization >4 classify this CBM as a Type B CBM [56]. Relatively few direct potential hydrogen bonds (~10) were distributed throughout these subsites, consistent with the relatively low density of direct hydrogen bonds observed with other Type B CBMs [56]. Numerous additional potential hydrogen bonds were mediated by five water molecules (Figure 17).

The bound laminarihexaose molecule adopts a U-shape very similar to that observed in the x-ray crystal structure of the laminarin-binding CBM from a *Thermotoga maritima* laminarinase (Figure 18 and Ref. [72]) and similar to that predicted for β-1,3-glucans in solution [135]. Similar to the laminarin-binding CBM4, the constellation of interactions between *Bh*CBM6 and laminarihexaose are unique to the conformation of this sugar and legislate against high affinity for other sugars. Unlike the family 4 CBM, which binds β-1,3-glucans in a deep binding groove (Figure 18*C*), the *Bh*CBM6 binding site begins with

a small "slot" only sufficient to accommodate the glucose residue at the reducing end of the oligosaccharide (Figure 18*B*). The remainder of the oligosaccharide curls around the CBM to form a crown (Figure 18*B*). The U shaped, or open helical, conformation of laminarin enables the ligand to curl over the walls that form the slot-like binding site and maintain interactions with the protein. In contrast, polysaccharides that have a 2-fold or 3-fold linear axis, such as cellulose and xylan, respectively, would simply extend out into solvent and not make any further direct interactions with the protein surface, explaining why xylose displays the same affinity for the protein as xylooligosaccharides and no increase in affinity is observed for xylose polymers. The sugar at the non-reducing end of laminarihexaose interacts with the terminal region of the CBM6 binding site such that O-3 is pointing directly at the protein surface sterically occluding extension of the sugar polymer and thus conferring specificity for the non-reducing end of the polysaccharide.

*Comparison of CBM6 Structures That Lead to the Differences in Specificity—* Comparison of the structures of CBM6 modules that recognize xylan and mixed linked β-(1,4)(1,3) glucans, respectively, with *Bh*CBM6, which binds to laminarin, provides novel insights into the structural basis for the extensive range of ligand specificities displayed by this family of proteins. In the xylan binding CBM6s cleft A is open at both ends explaining why these proteins are able to bind to the internal regions of the xylose polymers, with the central sugar in this site sandwiched between two aromatic residues (Figure 19). The surface of the cleft is likely to clash with the C-6 hydroxymethyl group of the pyranose ring at subsites 4 and 5, legislating against tight binding to glucose-containing ligands. By contrast, the primary ligand binding site in the *C. mixtus* CBM6

Figure 19: Overlap of cleft A region of *Bh*CBM6 in complex with laminarihexaose (*blue*), *Cm*CBM6 from *C. mixtus* LicA in complex with cellobiose (*beige*, Ref. 98), and *Cs*CBM6-1 from *C. stercorarium* in complex with xylotetraose (*green*, Ref.83). Relevant residues are labeled as follows: (*a*) $Trp^{99}$, $Trp^{92}$, and $Trp^{94}$; (*b*) $Trp^{42}$, $Tyr^{33}$, and $Tyr^{38}$; (*c*) $Gln^{29}$, $Gln^{20}$, and $Ile^{27}$ in *Bh*CBM6, *Cm*CBM6, and *Cs*CBM6-1, respectively. The label *d*, which indicates $Tyr^{128}$ in *Bh*CBM6, also shows the loop comprising residues 124-129 that is discussed in the text.



that recognizes β-(1,4)(1,3) mixed linked glucans is in cleft B, which lacks the extended loop that occludes this binding site in the other CBM6 modules. Cleft A in the *C. mixtus* CBM does display weak affinity for the terminal sugars of both xylose and glucose-containing oligosaccharides, but not the internal regions of the respective polysaccharides. The monosaccharide is again sandwiched between the parallel aromatic residues but is oriented 90° relative to the position of the sugar in the xylan binding modules and thus C-1 is pointing at the surface of the protein and sugars attached to O-4 extend into solvent and thus do not interact with the protein (Figure 19). The topology of cleft A in *Bh*CBM6 is most similar to the *C. mixtus* CBM (Figure 19). A glutamine residue, which at this position is typically an isoleucine or phenylalanine in xylan binding CBM6s, blocks off one end of the binding site in both the *C. mixtus* CBM6 and *Bh*CBM6. However, while in the *C. mixtus* CBM this residue hydrogen bonds with either terminus of the sugar, in *Bh*CBM6 it interacts specifically with the non-reducing sugar. The unique feature of the *Bh*CBM6 binding site is an extended loop comprising residues 124-129. This loop, and most notably the side chain of $Tyr^{128}$, creates a raised platform that follows the U-shaped curvature of the laminarioligosaccharide up and out of cleft A and along a surface that is distinct from the usual cleft A of CBM6s (Figure 19). Thus, remarkably, the primary specificity of *Bh*CBM6 for laminarin is conferred by a binding surface that is not only different from cleft A and cleft B, but has a convex shape, while the ligand binding site of all other Type B CBMs conform to concave clefts. These data therefore reveal a third ligand binding site in the CBM6 family of proteins that exhibits a unique topology. These data demonstrate how the multiple binding clefts and highly unusual protein surface of CBM6s confers the extensive range of specificities displayed by this

protein family. This is in sharp contrast to other families of CBMs where variation in specificity between different members is conferred by differences in the topology of a single binding site, while the range of ligand recognition observed in CBM6 is the result of variation in the location of the ligand binding site in different members of this family.

*Conclusions*—The biological rationale for the targeting of *Bh*CBM6 to the non-reducing ends of β-1,3-glucan chains is intriguing and rather counterintuitive as the molar concentration of available binding sites will be considerably less than for the majority of Type B CBMs, which bind to the internal regions of polysaccharides. Similar targeting, but to the reducing end termini of plant structural polysaccharides is mediated, however, by *Tm*CBM9-2 from the *T. maritima* xylanase10A [68; 69]. While it is possible that localization of the *B. halodurans* laminarinase to the ends of polysaccharide chains may reflect an exo-mode of action by the enzyme, the reaction products generated by its catalytic domain are consistent with a typical endo-mode of action; the enzyme releases oligosaccharides that display a range of different sizes (data not shown), whereas exo-acting glycoside hydrolases produce a single reaction product. The targeting of *B. halodurans* laminarinase to the ends of laminarin may reflect the complexity of the macromolecular structure that contains this polysaccharide, which, as a consequence, is recalcitrant to enzymatic attack. It is possible that disrupted regions of the plant cell wall, through either mechanical damage or the action of other enzymes, will contain a relatively large number of polysaccharide termini and be susceptible to laminarinase attack. *Bh*CBM6, by targeting the enzyme to these susceptible regions may potentiate its catalytic activity.

## 2.4: Discussion: Molecular determination of ligand specificity within the Family 6 CBMs.

The objective of this research was to study other CBMs within family 6 to acquire more structural and biochemical data for proposing a general model of how CBM6 binding sites accommodates the many different plant polysaccharide ligands. Including the CBMs from this research there are 8 known structures in complex with 18 different ligands: *Bh*CBM6 from *B. halodurans* laminarinase [136], *Cs*CBM6-1 and *Cs*CBM6-3 from *C. stercorarium* xylanase Xyn11A [83; 102], *Cm*CBM6-2 from *C. mixtus* lichenase [104], *Ct*CBM6 from *C. thermocellum* xylanase Xyn11A [101], *Sd*CBM6-2 from *Sarcophagus degradans* agarase AgaB [137], *Cc*CBM6 from a putative galactosidase from *Clostridium cellulolyticum* GH59 (Elizabeth Ficko-Blean; unpublished data) and finally a close relative of family 6 CBMs, a family 35 CBM, *Ao*CBM35 from *A. orientalis* exo-beta-D-glucosaminidase (Alicia Lammerts van Bueren; unpublished data). With the structural and biochemical information that has been accumulated on Family 6 CBMs we are able to observe how the structurally similar proteins within a CBM family are specific for a wide range of plant polysaccharides. Our hypothesis was that the topology of the cleft A binding site is altered to accommodate the different three-dimensional shapes of ligands and that alterations to binding site topology are imparted through amino acid modifications within the binding site, thus altering the ligand specificity of the module.

*Overall 3-dimensional structures*

Structural overlaps of all known CBM6s show that these proteins are all very similar with four antiparallel β-strands overlapping five antiparallel β-strands in a β-

sandwich fold with a β-jelly roll topology (see Figure 6A and Figure 20). They share anywhere from 19% - 60% amino acid sequence identity and RMSD's for all CBM6s fall between 0.72 and 1.89 Å over the entire protein (~120 amino acids) (Table 6). *Bh*CBM6, *Cs*CBM6-1 and *Sd*CBM6, *Cc*CBM6 and *Ao*CBM35 have only one binding site occupied by ligand at the apex of the protien within the loops connecting β-strands, making cleft A the only binding site in these modules which supports our hypothesis that cleft A is the main binding cleft in family 6 CBMs. Cleft B is thus far only an active binding site in *Cm*CBM6-2 [104] (Figure 6A). It is shown bound to a mixed β-1,3-1,4-glucan by an exposed Trp39 on the edge of one face of the β-sandwich [104]. *Bh*CBM6 Trp48 is structurally conserved with *Cm*CBM6-2 Trp39 however it is stacked against Pro83, preventing binding with ligand, as was similarly seen in *Ct*CBM6 (ref). The same was observed in *Sd*CBM6-2 where structurally conserved Tyr68 is blocked by the side chain of Arg140 which would also prevent ligand binding. *Cc*CBM6 Tyr59 is conserved with *Sd*CBM6-2 Y68 but is also blocked by an adjacent Ser94, preventing interaction with sugar. Both *Cs*CBM6-1 and *AoCBM35* do not have an apparent cleft B site.

The β-sandwich fold is characteristic of type B CBMs, which bind to extended sugar molecules, and in the majority of Type C CBMs that bind mono-, di- or tri saccharides, and indeed all CBM6s have binding within the binding cleft A (See section 1.4.1 in introduction). Upon closer observation of the cleft A binding site of CBM6s, we can identify 5 regions that contribute to ligand specificity: three sites of amino acid conservation and two molecular "hotspots" that alter the binding site to allow for the individual ligands to fit tightly within the binding pocket (Figure 21 and Table 7).

Figure 20: Structural overlaps of all known family 6 CBMs and AoCBM35. CsCBM6-1 (blue, PDB code 1UY4), BhCBM6 (burnt yellow, PDB code 1W9W), SdCBM6-2 (magenta, PDB code 2CDP), CsCBM6-3 (red, PBD code 1NAE), CmCBM6 (cyan, PDB code 1UYX), CtCBM6 (orange, PDB code 1UXX) and AoCBM35 (green). C-alpha backbone is shown in ribbon. Generated using PyMol.

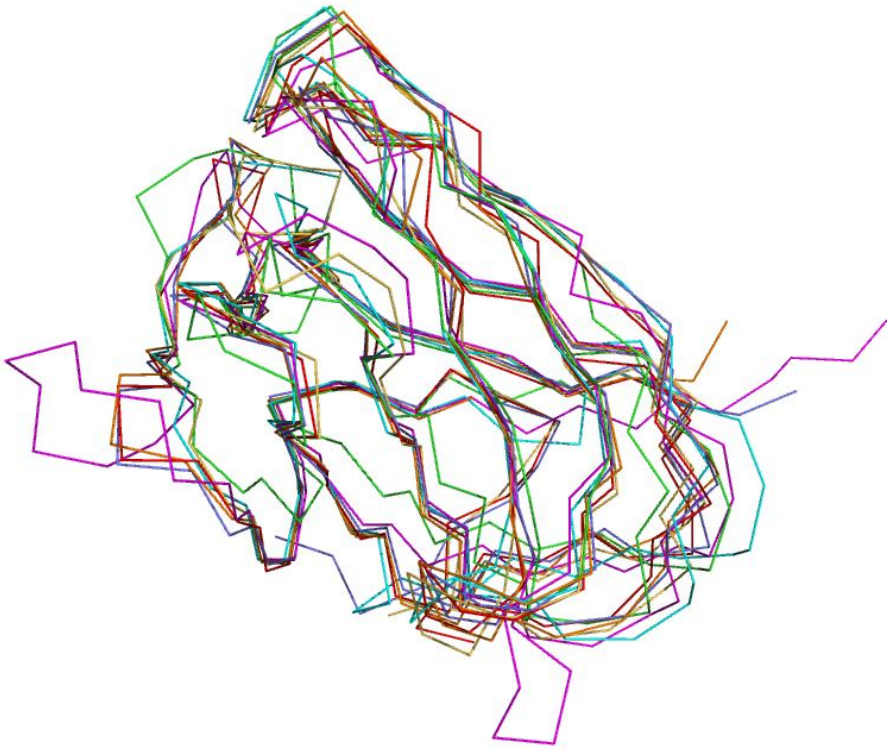


Table 6: (A) Percentage of amino acid sequence identity and (B) RMSD's for all structures of members of Family 6 and Family 35.

| A (%) | *Cs*CBM6-3 | *Cm*CBM 6-2 | *Cs*CBM6-1 | *Bh*CBM6 | *Sd*CBM6-2 | *Cc*CBM6 | *Ao*CBM35 |
|---|---|---|---|---|---|---|---|
| *Ct*CBM6 | 49 | 24 | 57 | 40 | 22 | 47 | 19 |
| *Cs*CBM6-3 | - | 26 | 60 | 38 | 17 | 44 | 24 |
| *Cm*CBM6-2 | - | - | 23 | 35 | 24 | 34 | 28 |
| *Cs*CBM6-1 | - | - | - | 38 | 20 | 44 | 22 |
| *Bh*CBM6 | - | - | - | - | 25 | 45 | 21 |
| *Sd*CBM6-2 | - | - | - | - | - | 24 | 22 |
| *Cc*CBM6 | - | - | - | - | - | - | 16 |

| B (Å) | *Cs*CBM6-3 | *Cm*CBM 6-2 | *Cs*CBM6-1 | *Bh*CBM6 | *Sd*CBM6-2 | *Cc*CBM6 | *Ao*CBM35 |
|---|---|---|---|---|---|---|---|
| *Ct*CBM6 | 0.91 | 1.46 | 0.72 | 1.12 | 1.52 | 1.17 | 2.11 |
| *Cs*CBM6-3 | - | 1.41 | 0.76 | 0.98 | 1.54 | 1.09 | 1.88 |
| *Cm*CBM6-2 | - | - | 1.49 | 1.28 | 1.61 | 1.39 | 1.97 |
| *Cs*CBM6-1 | - | - | - | 1.25 | 1.52 | 1.09 | 1.83 |
| *Bh*CBM6 | - | - | - | - | 1.49 | 0.78 | 1.92 |
| *Sd*CBM6-2 | - | - | - | - | - | 1.56 | 1.79 |
| CcCBM6 | - | - | - | - | - | - | 2.12 |

Figure 21: Structural overlaps of individual binding sites showing the regions of differentiation thought to be important in specific ligand interactions. *Cs*CBM6-1 (blue, PDB code 1UY4), *Bh*CBM6 (burnt yellow, PDB code 1W9W), *Sd*CBM6-2 (magenta, PDB code 2CDP), *Cs*CBM6-3 (red, PBD code 1NAE), *Cm*CBM6 (cyan, PDB code 1UYX), *Ct*CBM6 (orange, PDB code 1UXX), *Cc*CBM6 (grey, EFB unpublished) and *Ao*CBM35 (green, ALVB unpublished). Residues are shown in licorice representation. Regions and labeled A-E and the conserved water residue is shown as a blue sphere. Image generated using PyMol.

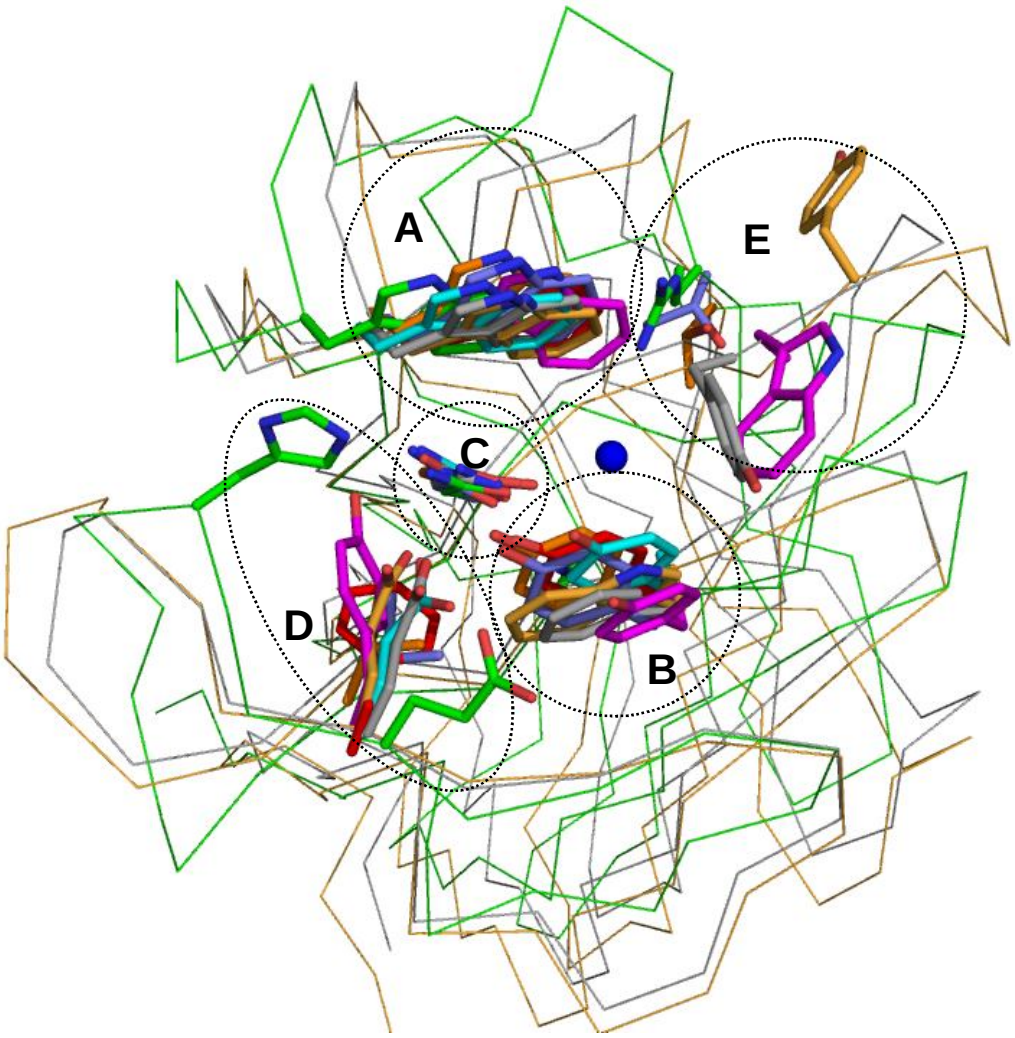


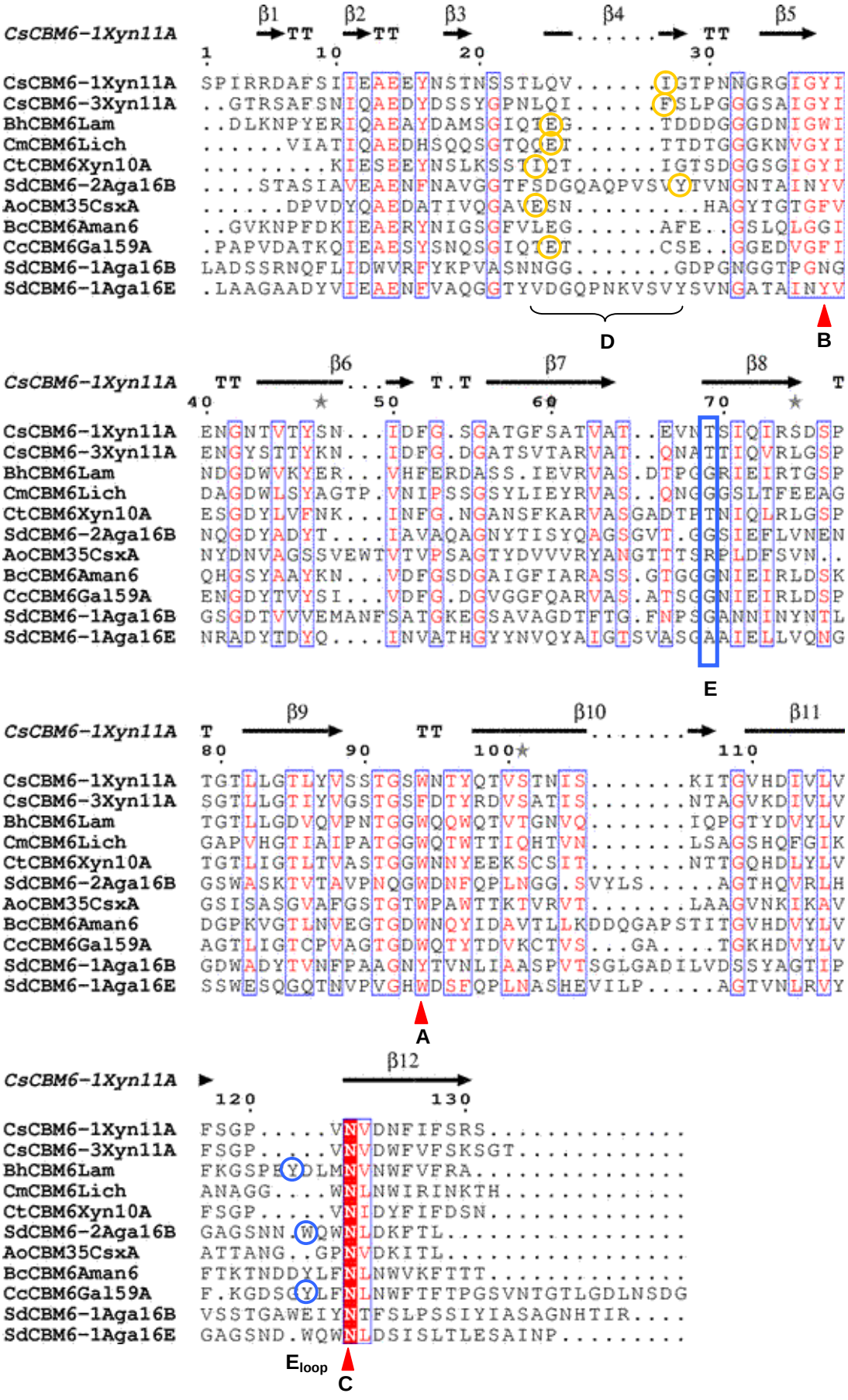
Table 7: Important Residues for Sugar binding by CBM6s

| Region of ligand binding | Protein | Region A | B | C | D | E | Ligand |
|---|---|---|---|---|---|---|---|
| Internal extended sugar | *Ct*CBM6 | W | Y | N | I | T | xylan |
| | *Cs*CBM6-1 | W | Y | N | I | T | xylan |
| | *Cs*CBM6-3 | F | Y | N | F | T | xylan |
| Non-reducing end of Extended sugars | *Bh*CBM6 | W | W | N | E | Y | β-1,3-glucan |
| | *Sd*CBM6-2 | W | Y | N | Y | W | agarose |
| | *Cm*CBM6 | W | Y | N | E | G | cellulose |
| Terminal sugars | *Cc*CBM6 | W | F | N | E | Y | xylose |
| | *Ao*CBM35 | W | N/Y | N | E/H | R | Glucuronic acid |

*Conserved amino acids within Cleft A*

Regions A and B consist of structurally conserved hydrophobic amino acid side chains that functionally allow for hydrophobic stacking interactions with a sugar monomer. Region A is usually a tryptophan, the exception being *Cs*CBM6-3 where it is a phenylalanine, and region B a tyrosine except for *Bh*CBM6 where it is a tryptophan and *Cc*CBM6 is a phenylalanine. *Ao*CBM35 is a unique case because Tyr33 is flipped away from the binding site and instead a calcium ion is present bound by the mainchain hydroxyl group of Try33 and Asn32. The calcium coordinates with the uronate group and the C4 hydroxyl group of GlcUA. Both the presence of calcium and Asn32 were shown to be essential for ligand binding. In all instances region C contains a structurally conserved asparagine whose side chain forms a direct hydrogen bond from the same sugar monomer bound by region A and B and the oxygen of the glycosidic linkage. These three regions are also conserved in the amino acid sequences of members of family 6 whose structures are unknown and are likely involved in all ligand interactions (Figure 22). Also within the binding site of all CBM6 structures is a conserved water molecule that hydrogen bonds with the conserved Asn carbonyl group in region C and to the equatorial C2 or C3 hydroxyl groups of the single sugar that is sandwiched between Regions A & B. The exception in this case is *Sd*CBM6-2 where due to the cyclization of galactose at C3 and C6 the oxygen is no longer in the correct position to form a water-mediated hydrogen bond. In *Ao*CBM35 the calcium is positioned in this region instead of a water molecule which has two additional waters coordinated with the calcium ion, which are displaced when bound to glucuronic acid. Although not directly in the binding pocket, *Sd*CBM6-2 also requires a calcium for binding agarose which stabilizes the

Figure 22: Amino acid sequence alignments of family 6 CBMs (next page). Sequences marked with (*) represent those with known structures. The sequences without structures were chosen at random from the CaZY database to demonstrate the relevance of binding site residues in CBM6s whose structures are unknown. Regions A, B and C (see Figure 21) are highlighted with a red arrow. Region D residues are highlighted in yellow. Region E residues are surrounded by a blue box. Region E residues in CBM6s with additional loop regions ($E_{loop}$) are circled in blue


CsCBM6-1Xyn11A
β1 β2 β3 β4 β5
CsCBM6-1Xyn11A SPIRRDAFSIIEAEEYNSTNSSTLQV.....IGTPNNGRGIGYI
CsCBM6-3Xyn11A ..GTRSAFSNIQAEDYDSSYGPNLQI.....FSLPGGGSAIGYI
BhCBM6Lam ..DLKNPYERIQAEAYDAMSGIQTEG.....TDDDGGGDNIGWI
CmCBM6Lich .....VIATIQAEDHSQQSGTQQET.....TTDTGGGKNVGYI
CtCBM6Xyn10A .....KIESEEYNSLKSSTIQT.....IGTSDGGGSGIGYI
SdCBM6-2Aga16B ....STASIAVEAENFNAVGGTFSDGQAQPVSVYTVNGNTAINYV
AoCBM35CsxA .....DPVDYQAEDATIVQGAVESN.....HAGYTGTGFV
BcCBM6Aman6 ..GVKNPFDKIEAERYNIGSGFVLEG.....AFE..GSLQLGGI
CcCBM6Gal59A .PAPVDATKQIEAESYSNQSGIQTET.....CSE..GGEDVGFI
SdCBM6-1Aga16B LADSSRNQFLIDWVRFYKPVASNNGG.....GDPGNGGTPGNG
SdCBM6-1Aga16E .LAAGAADYVIEAENFVAQGGTYVDGQPNKVSVYSVNGATAINYV
D
B
β6 β7 β8
CsCBM6-1Xyn11A ENGNTVTYSN...IDFG.SGATGFSATVAT..EVNTSIQIRSDSP
CsCBM6-3Xyn11A ENGYSTTYKN...IDFG.DGATSVTARVAT..QNATTIQVRLGSP
BhCBM6Lam NDGDWVKYER...VHFERDASS.IEVRVAS.DTPGGRIEIRTGSP
CmCBM6Lich DAGDWLSYAGTP.VNIPSSGSYLIEYRVAS..QNGGGSLTFEEAG
CtCBM6Xyn10A ESGDYLVFNK...INFG.NGANSFKARVASGADTPTNIQLRLGSP
SdCBM6-2Aga16B NQGDYADYT....IAVAQAGNYTISYQAGSGVT.GGSIEFLVNEN
AoCBM35CsxA NYDNVAGSSVEWTVTVPSAGTYDVVVRYANGTTTSRPLDFSVN..
BcCBM6Aman6 QHGSYAAYKN...VDFGSDGAIGFIARASS.GTGGGNIEIRLDSK
CcCBM6Gal59A ENGDYTVYSI...VDFG.DGVGGFQARVAS.ATSGGNIEIRLDSP
SdCBM6-1Aga16B GSGDTVVVEMANFSATGKEGSAVAGDTFTG.FNPSGANNINYNTL
SdCBM6-1Aga16E NRADYTDYQ....INVATHGYYNVQYAIGTSVASGAAIELLVQNG
E
β9 β10 β11
CsCBM6-1Xyn11A TGTLLGTLYVSSTGSWNTYQTVSTNIS.....KITGVHDIVLV
CsCBM6-3Xyn11A SGTLLGTIYVGSTGSFDTYRDVSATIS.....NTAGVKDIVLV
BhCBM6Lam TGTLLGDVQVPNTGGWQQWQTVTGNVQ.....IQPGTYDVYLV
CmCBM6Lich GAPVHGTIAIPATGGWQTWTTIQHTVN.....LSAGSHQFGIK
CtCBM6Xyn10A TGTLIGTLTVASTGGWNNYEEKSCSIT.....NTTGQHDLYLV
SdCBM6-2Aga16B GSWASKTVTAVPNQGWDNFQPLNGG.SVYLS.....AGTHQVRLH
AoCBM35CsxA GSISASGVAFGSTGTWPAWTTKTVRVT.....LAAGVNKIKAV
BcCBM6Aman6 DGPKVGTLNVEGTGDWNQYIDAVTLLKDDQGAPSTITGVHDVYLV
CcCBM6Gal59A AGTLIGTCPVAGTGDWQTYTDVKCTVS...GA....TGKHDVYLV
SdCBM6-1Aga16B GDWADYTVNFPAAGNYTVNLIAASPVTSGLGADILVDSSYAGTIP
SdCBM6-1Aga16E SSWESQGQTNVPVGHWDSFQPLNASHEVILP.....AGTVNLRVY
A
β12
CsCBM6-1Xyn11A FSGP.....VNVDNFIFSRS.....
CsCBM6-3Xyn11A FSGP.....VNVDWFVFSKSGT.....
BhCBM6Lam FKGSPEYDLMNVNWFVFRA.....
CmCBM6Lich ANAGG....WNLNWIRINKTH.....
CtCBM6Xyn10A FSGP.....VNIDYFIFDSN.....
SdCBM6-2Aga16B GAGSNN.WQWNLDKFTL.....
AoCBM35CsxA ATTANG..GPNVDKITL.....
BcCBM6Aman6 FTKTNDDYLFNLNWVKFTTT.....
CcCBM6Gal59A F.KGDSGYLFNLNWFTFTPGSVNTGTLGDLNSDG
SdCBM6-1Aga16B VSSTGAWEIYNTFSLPSSIYIASAGNHTIR....
SdCBM6-1Aga16E GAGSND.WQWNLDSISLTLESAINP.....
Eloop
C

amino acid side chains important for binding [137]; however, so far the involvement of calcium in ligand binding is restricted to these three CBMs. It appears that in all cases the regions bound by A, B and C provide most of the driving force for ligand binding since most of the interactions between a single sugar molecule and the protein are made within this site. This was confirmed in our study of *Cs*CBM6-1 binding subsite dissection where this region, termed subsite 2, was occupied by xylose, however at least two sugars were required to study the thermodynamic properties suggesting that the interaction of Asn in region C with the oxygen of the glycosidic linkage is an important driving force in the interaction [83].

*Molecular "hotspots" within cleft A*

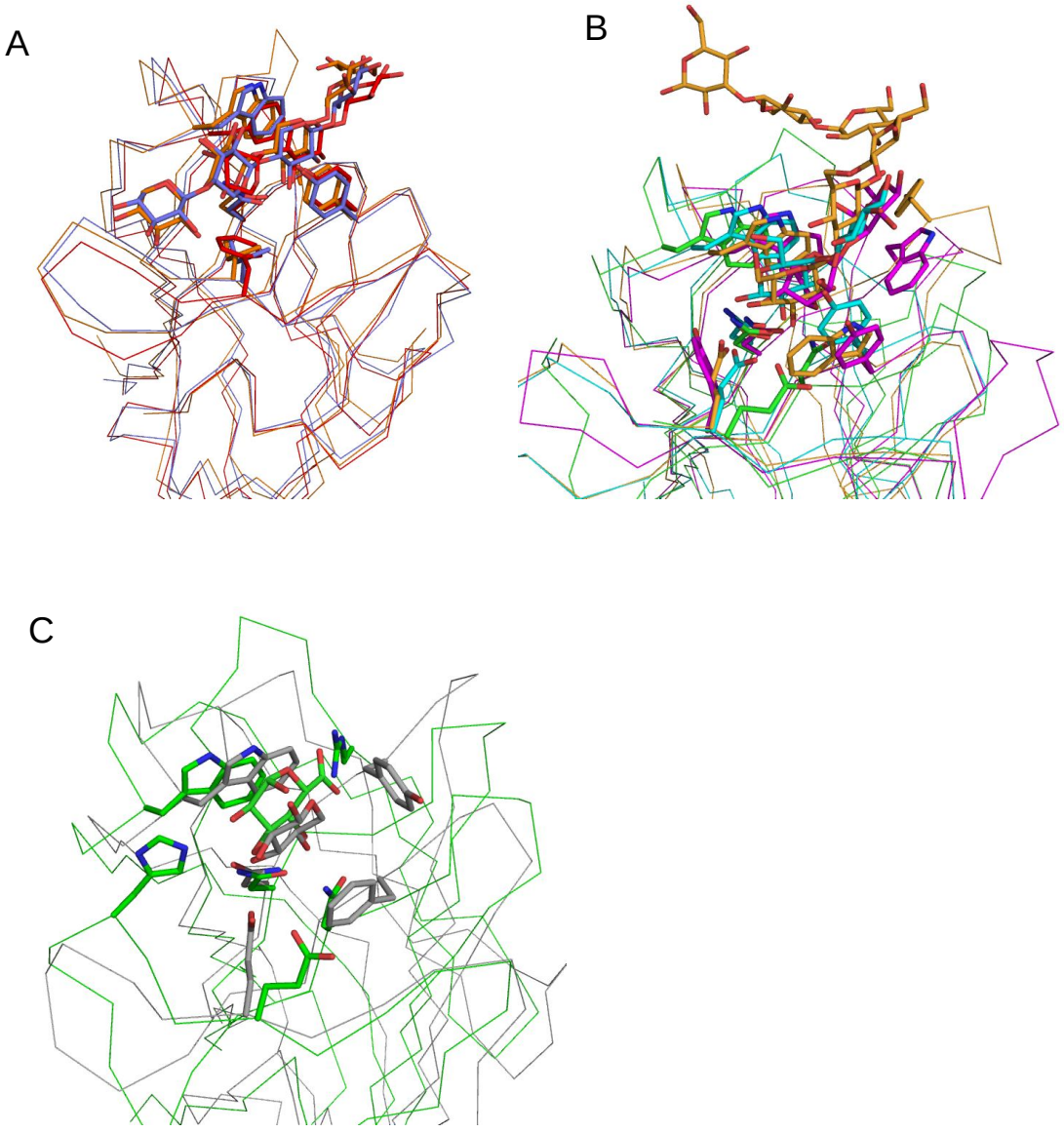
Regions D and E appear to act as "hot spots" where specific modifications act to confer specificity of this family to their respective ligands. The purpose of region D appears to be to strategically position amino acid side chains at the end of the binding site to either open up the binding site to surrounding bulk solvent or blocking the end of the binding site (Figure 21 and Table 7). This affects whether the CBM optimally binds to extended internal regions within a polysaccharide or to their terminal ends. *Cs*CBM6-1, *Cs*CBM6-3 and *Ct*CBM6 all bind to internal stretches of polysaccharide within xylan, having at least 4 binding subsites available for interacting with the ligand. The subsite within Region D is occupied by an isoleucine in *Cs*CBM6-1 and *Ct*CBM6 while *Cs*CBM6-3 has phenylalanine (Figure 23A). By lying parallel to the binding pocket, isoleucine does not occlude the binding site, allowing xylan to extend out into the surrounding bulk solvent. *Cs*CBM6-3 has a phenylalanine in this region which also lies

parallel with the ligand, also allowing the ligand to extend out into the surrounding bulk solvent. The terminal sugar binders *Bh*CBM6, *Sd*CBM6-2, *Cm*CBM6, *Cc*CBM6 and *Ao*CBM35 all have amino acid side chains that occlude the binding pocket within Region D (Fig. 23B and C): a glutamate in *Bh*CBM6, *Cm*CBM6-2 and *Cc*CBM6 and a tyrosine in *Sd*CBM6-2 are all directed into the binding site. *Ao*CBM35 shares this glutamate residue with *Bh*CBM6, *Cm*CBM6-2 and *Cc*CBM6; however, it forms a water-mediated hydrogen bond with C3 of glucuronic acid. Histidine 22 directly interacts with C4 and also contributes to binding terminal glucuronic acid residues.

The amino acid in Region D interacts directly with the ligand in the case of *Bh*CBM6, *Cm*CBM6, and *Cc*CBM6 where the glutamate hydrogen bonds with the C4 hydroxyl of the terminal sugar. In the case of *Cs*CBM6-3 and *Sd*CBM6-2 the phenyl and tyrosyl rings lie perpendicular to the sugar molecule and does not interact directly with the ligand.

Region E is important for accommodating the three dimensional shape of the ligand, whether it interacts within a glycan chain or to terminal sugars (Figure 21 and Table 7). The relatively small side chain provided by a threonine residue in *Ct*CBM6, *Cs*CBM6-3 and *Cs*CBM6-1 allows for xylan to extend out into bulk solvent on the opposite end of region D (Figure 21 and Fig 23A). *Cm*CBM6-2 has glycine residues in this region also allowing the ligand to direct out into the bulk solvent (Fig 23A) where van der Waals forces contribute to ligand stability in this region. In *Bh*CBM6, *Sd*CBM6-2 and *Cc*CBM6 the protein adjusts its three-dimensional structure in order to support the ligand with the addition of hydrophobic side chains. *Bh*CBM6 has an extended loop region and an additional Tyr residue forming a U shaped cleft to provide specificity

Figure 23: (A) Structural overlaps of CBM6s that bind internally to sugars CtCBM6 (orange, PDB code 1UXX), CsCBM6-1 (blue, PDB code 1UY4) and CsCBM6-3 (red, PBD code 1NAE). (B) Structural overlaps of CBM6s that bind to non-reducing end of ligands CmCBM6 (cyan, PDB code 1UYX), BhCBM6 (burnt yellow, PDB code 1W9W), SdCBM6-2 (magenta, PDB code 2CDP). (C) Structural overlap of CBM6s that bind to terminal sugars *Cc*CBM6 (gray-unpublished data) and AoCBM35 (green- unpublished data)



for the coiled structure of laminarin and not linear xylan (Fig 23B). *Sd*CBM6-2 contains an additional Trp residue in this region providing a second subsite for stacking with the galactose sugar in 3,6-anhydro-α-L-galactose-(1,3)-β -D-galactopyranose which was demonstrated to be necessary for ligand binding (Fig 23B).

The two CBM6s that bind to terminal sugars, *AoCBM35* and *Cc*CBM6, have bulky sidechains that block this end of the binding pocket (Figure 23C). *Ao*CBM35 contains an Arg67 which forms direct hydrogen bonds with the uronate group of a terminal GlcUA, allowing it to accommodate this decoration at the C6 position where otherwise would be a hydroxymethyl group. *Cc*CBM6 has a loop containing Y139 which is shifted ~6 Å up into the binding pocket and positions the Tyr so that it blocks this end of the binding pocket for accommodating terminal xylose residues.

Although amino acids within regions D and E structurally align, the amino acid sequence alignments using ClustalW did not align the amino acids within hotspot D (Figure 22). However, region E did align in the amino acid sequence. CBMs with additional loop regions (*Bh*CBM6, *Sd*CBM6-2 and *Cc*CBM6) have a conserved glycine in region E. Structurally, the glycine makes room for the additional loop to extend into this region. The aromatic amino acids present in the loop region also align in the sequences (Figure 22, $E_{loop}$). Therefore region E can be used to predict how other unknown CBM6s structurally interact with their ligand. For example, *Bc*CBM6 from a predictive mannanase (*Bc*CBM6A*man6*), and both *Sd*CBM6-1 modules from agarases Aga16B and 16E contains a glycine or alanine in region E and a tyrosine or tryptophan in region $E_{loop}$. Therefore it is likely that these CBM6s also contain the additional loop region. Interestingly, based on the amino acid sequence alignments, it appears as though

the additional loop region is common to most CBM6s except for those that interact with xylan (*Ct*CBM6, *Cs*CBM6-1, *Cs*CBM6-3) and in *Cm*CBM6, the only CBM6 thus far to have a functional cleft B binding site.

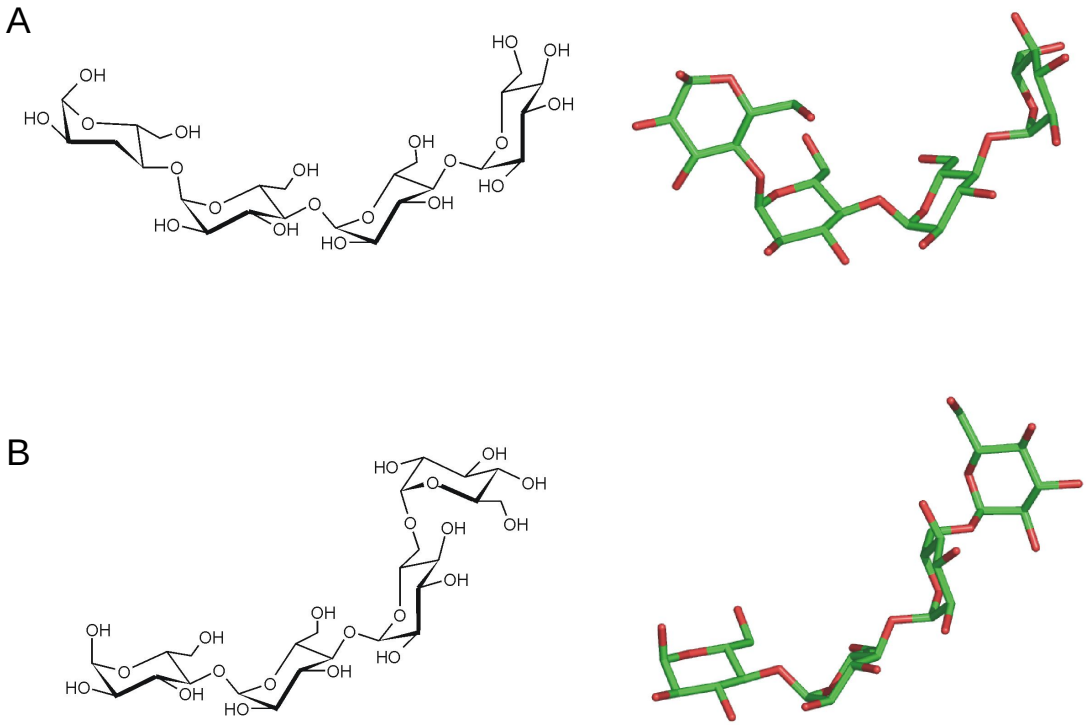
*Overall Conclusion* – This CBM family has specialized regions within the binding site that at the molecular level are able to discriminate between the many three dimensional structures of plant polysaccharides. The amino acids within these regions form an overall binding site topology that is complementary to the three-dimensional shape of the ligand. Looking at the fundamental molecular determinants in carbohydrate recognition within family 6 has implications in predicting important amino acids for carbohydrate binding in members whose structures are unknown and provides an understanding of how ligand specificity is determined by CBM binding site architecture. This information may also be useful as a biotechnological tool in modifying CBMs for optimal ligand binding.

# Chapter 3: Molecular Determinants of α-Glucan Recognition by Family 41 CBMs

## 3.1 Introduction

Although starch is a complex polysaccharide, it is limited in composition when compared to plant cell wall polysaccharides. A granule of starch is an amalgam of amylose (α-1,4-glucose) and amylopectin (α-1,4-glucose with α-1,6-branch points approximately every 16-20 glucose units). Glycogen is very similar in structure to amylopectin with α-1,6-branches occurring more frequently approximately every 8-12 glucose units. Unlike plant polysaccharides which have varied structures due to the presence of multiple linkages and varied sugar compositions, starch and glycogen are mainly α-1,4-linked glucose, giving rise to a double helix structure with occasional α-1,6 branches (Figure 24). As a result, there are fewer CBM families involved in the recognition of starch than those that interact with plant polysaccharides. Only 9 of the 51 families have demonstrated α-glucan binding activity, with 8 of those shown to interact with α-1,4-linked glucose (families 20, 21, 25, 26, 34, 41, 45 and 48) and some families are able to tolerate an α-1,6-linkage (41, 48). One family, family 24, has demonstrated α-1,3-glucan binding activity from a mutanase (α-1,3-glucanase). Often called starch binding domains (SBDs), they are associated with α-glucan degrading enzymes such as α and β amylases and glucoamylases (CBM families 20, 21, 25, 26 and 34) and pullulanases (CBM families 41 and 48). There are also examples of modules that

Figure 24: Three-dimensional structure of starch components. (A) Maltotetraose (M4) (alpha-1,4-linked glucose) and (B) Glucosyl-maltotriose (GM3) (maltotriose-alpha-1,6-glucose). Graphical representation on left and structural representation on right. (M4 PDB code 2J72) (GM3 PDB code 2J73).



are involved in glycogen synthesis and glycogen-binding AMP-activated protein kinases (CBM family 48) and in α-glucan water-dikinases (CBM family 45).

The first SBDs were initially identified in 1989 as distinct domains within larger starch-degrading enzymes when it was shown that amylases with little sequence homology to each other have similar C-terminal sequence motifs [138]. In parallel to the initial studies on plant CBMs and their importance in the hydrolysis of cellulose, starch-degrading enzymes lacking these C-terminal domains have unchanged activity on soluble glucans but have a reduced activity on granular starch [139]. Now known as CBM family 20, this is the most studied starch binding CBM family with structures for ~8 different modules from bacterial cyclodextrin glucanotransferases, α-amylases and a eukaryotic glucoamylase from *Aspergillus niger*.

All starch-binding CBM families studied, with the exception of family 45, have structural representatives, revealing similar three-dimensional protein structures and similar modes of starch recognition. The first solution structure of a starch-binding CBM in complex with ligand was the C-terminal CBM from *A. niger* glucoamylase in complex with β-cyclodextrin [98]. The structure revealed two binding sites: site one at the apex of the protein with a shortened binding pocket, and site 2 on the side of the β-sandwich with an elongated binding pocket for recognizing longer chains of starch. The N-terminal CBM34 from *Thermoactinomyces vulagris* α-amylase also has two binding sites allowing for an enhanced binding affinity to granular starch [140]. With the exception of CBM20 and CBM34, all other starch-binding CBMs studies to date have identified single binding sites located on the side of the β-sandwich. Some α-glucan degrading enzymes have multiple CBMs to efficiently interact with granular starch. An elegant study on the

starch binding activity of the tandem C-terminal family 25 and 26 modules from *Bacillus halodurans* α-amylase demonstrate an enhanced affinity in binding starch through an avidity effect [63]. The structures of *Bh*CBM25 and 26 revealed that they bound to maltooligosaccharides through a conserved mode of starch recognition to each other and other known starch-binding CBMs. To date all research on α-glucan recognition by CBMs from different families illustrate a very similar mode of starch recognition despite sharing very little amino acid sequence identity as shown in the CaZY database [22].

The family 41 CBMs were previously known as family X28 modules with unknown function and were first studied in this thesis research. They are appended to bacterial pullulanases from family 13 glycoside hydrolases. They are also known as starch debranching enzymes as they have demonstrated activity on α-1,6-linkages in starch and pullulan. All previous research shows that many starch binding domains have a decreased affinity for maltooligosccharides containing α-1,6-linkages [63; 140]. In pullulan, the frequency of α-1,6-linkages is much greater than that found in starch as they occur every three α-1,4-linked glucose units. Because CBM41s are associated with pullulanases, the chance of these modules encountering an α-1,6-linkage is much greater than CBMs within other α-glucan degrading enzymes, such as amylases and α-glucanases. **The question asked is, how do CBM41's from pullulanases accommodate the frequency of α-1,6-linkages found in pullulan?** This research involved studying the starch binding activity of the family 41 CBMs from *Thermotoga maritima, Streptococcus pneumoniae* and *Streptococcus pyogenes* pullulanases. ***Our hypothesis is that they share similar structure and modes of starch recognition with the other families of starch binding modules but have a modified binding site architecture which would allow them***

***to accommodate an α-1,6-linkage.*** To test this hypothesis, biochemical and structural analysis was carried out on these modules and compared with the known structures of other starch binding CBMs to see if the mode of starch recognition is conserved in this family.

The first study on the modular structure of pullulanase PulA from *T. maritima* identified a new family of CBMs with α-glucan binding function. It was shown from thermodynamic studies that the N-terminal CBM41 can interact with high affinity to maltooligosaccharides and can accommodate an α-1,6-linkage. It wasn't until the structure of *Tm*CBM41 was solved in complex with maltotetraose and glucosyl-α-(1,6)-maltotriose that we were able to identify how the binding site architecture accommodates the α-1,6-linkage, distinguishing this family from other starch binding domains and providing a rationale for why these modules are only found in pullulanases. In establishing CBM41 as a new family, we identified other potential family members using PSI-BLAST with amino acid sequence similarity to *Tm*CBM41. Surprisingly, our search yielded an abundance of CBM41 modules in pullulanases that were associated with pathogenic strains of bacteria, of which over half occur in duplicate (see Table 11). Our studies of the tandem CBM41 modules from *S. pneumoniae* and *S. pyogenes* revealed their contribution to virulence arose through interactions with glycogen in host lung tissue. Their tandem arrangement facilitates this interaction by forming a bivalent scaffold with two opposing binding sites optimized for interacting with opposing α-glucan chains within glycogen. This study is the first to identify glycogen as a target molecule in *Streptococcal* pathogenesis.

## 3.2 α-Glucan Recognition by a New Family of Carbohydrate-Binding Modules Found Primarily in Bacterial Pathogens

**Alicia Lammerts van Bueren, Ron Finn, Juan Ausió, and Alisdair B. Boraston**

Department of Biochemistry and Microbiology, University of Victoria, Victoria, Canada

Adapted from *Biochemistry*, **43** (49), 15633 -15642, 2004

Contributions to Research: Cloning, protein production, biochemical assays (except sedimentation), preparation of manuscript and figures.

### 3.2.1 Abstract

*Tm*Pul13, a family 13 glycoside hydrolase from *Thermotoga maritima*, is a four-module protein having pullulanase activity; the three N-terminal modules are of unknown function while the large C-terminal module is likely the catalytic module. Dissection of the functions of the three unknown modules revealed that the 100 amino acid module at the extreme N-terminus of *Tm*Pul13 comprises a new family of carbohydrate-binding modules (CBM) that a bioinformatic analysis shows are most frequently found in pullulanase-like sequences from bacterial pathogens. Detailed binding studies of this isolated CBM, here called *Tm*CBM41, reveals a preference for α-(1,4)-linked glucans, but occasional α-(1,6)-linked glucose residues, such as those found in pullulan, are tolerated. UV difference, isothermal titration calorimetry, and analytical ultracentrifugation binding studies suggest that maltooligosaccharides longer than four glucose residues are able to bind two *Tm*CBM41 molecules per oligosaccharide when sugar concentrations are below the CBM concentration. This is explained in terms of an equilibrium expression involving the formation of both a 1 to 1 sugar to CBM complex and a 1 to 2 sugar to CBM complex (i.e., a CBM dimer ligated by an oligosaccharide). The presence of an α-(1-6) linkage in the oligosaccharide appears to prevent this phenomenon.

### 3.2.2 Introduction

*Thermotoga maritima* is a hyperthermophilic eubacteria first discovered in geothermal heated marine sediment [141]. It has an optimal growth temperature of 80 degrees C and is one of the most thermophilic bacteria known. It produces a number of thermostable enzymes for polysaccharide depolymerization. One of these is a pullulanase, *Tm*Pul13, encoded by *pulA*, whose enzymatic activity has previously been studied [142; 143]. This enzyme appears to be specific for the α-(1,6)-glycosidic linkage in pullulan; however, it is not able to cleave the α-(1-6) bonds in amylopectin or α-(1-4) bonds in α-glucans. *Tm*Pul13 is a relatively large enzyme (843 amino acids) and is likely comprised of a number of modules. It is a member of the Glycoside hydrolase family 13 which is, based on the number of amino acid sequence entries, one of the largest glycoside hydrolase families (http://afmb.cnrs-mrs.fr/CAZY/index. html). This family comprises a number of α-glucan active enzymes, including amylases, pullulanases, cyclomaltodextrin transferases, dextranases, and α-glucosidases, which are found in a number of biological locales where they are involved in the depolymerization, modification, or synthesis of α-glucans. These enzymes are usually highly modular, comprising a catalytic module along with accessory modules, most frequently noncatalytic carbohydrate-binding modules (CBM) that mediate the tight association of the enzymes with their substrates [56]. Currently, there are 52 families of CBMs defined on the basis of sequence similarity. Members of eight of these families have been observed to bind to granular starch and/or α-glucooligosaccharides; however, previous to this research, only one fungal member of one family has been studied in detail [144; 145].

In this study, we dissect the modular structure of *Tm*Pul13 by heterologous production of its individual modular components in *Escherichia coli*. We report the α-glucan binding affinity and specificity of these modules and, on the basis of this information, propose a new CBM family, family 41. In addition, we dissect the kinetic and thermodynamic mechanism(s) of ligand recognition by this new CBM. This is the first study of a CBM from a bacterial pullulanase, which is made more unique by its hyperthermophilic source. Furthermore, examination of this new family of CBMs reveals its distribution in primarily pathogenic bacteria and suggests a carbohydrate-binding function for these modules in pathogenic microbes.

### 3.2.3 Materials and Methods

*Carbohydrates and Polysaccharides.* $6^3$-α-D-Glucosylmaltotriose (GM3) and $6^3$-α-D-glucosylmaltotriosylmaltotriose (GM3M3) were purchased from Megazyme Ltd. (Bray, C. Wicklow, Ireland). Maltose (M2), maltotriose (M3), maltotetraose (M4), maltopentaose (M5), maltohexaose (M6), isomaltose, isomaltotriose, panose, pullulan, and amylopectin (starch) were from Sigma (St. Louis, MO).

*Cloning, Expression, and Purification of CBMs.* The DNA fragments encoding the modules X28, X45, X20, X28/45/20, and *Tm*Pul13 (see Figure 1) were amplified by PCR from *T. maritima* genomic DNA (strain MS8B, ATCC 43589D) by a procedure described previously [68]. The 5' oligonucleotide primers were 5'-CACCGAAACCACCATCGTAGTC-3' (X28, X28/45/20, and Pul13), 5'-CACCGACACATCTCCCAGAATC-3' (X45), and 5'-CACCGGAGAGCTCGGAGCCGTA-3' (X20). The 3' oligonucleotide primers were 5'-

CTT**TTA**TGGTTTTTCGTAGAAAAA-3' (X28), 5'-CTT**TTA**ATCGTAATAGTAGTCGTC-3' (X45), 5'-CTT**TTA**TTCGTATCCTTCGATTTT-3' (X20 and X28/45/20), and 5'-CTT**TTA**CTCTCTGTACAGAACGTA-3' (Pul13) (the stop codon is in bold). These allowed for the amplification of nucleotide sequences from *pulA* encoding amino acids 20-120 (X28), 121-222 (X45), 223-339 (X20), 20-339 (X28/45/20), and 20-843 (Pul13) of *Tm*Pul13. The 5' CACC in the 5' oligonucleotide primers was essential for inserting purified fragments into pET Directional TOPO Expression Kits (Invitrogen, Carlsbad, CA) using pET150 to give pET150-X28 and pET150-Pul13 and using pET100 to give pET100-X45, pET100-X20, and pET100-X28/45/20. They were transformed into *E. coli* BL21(DE3) for polypeptide production. All polypeptides comprised a $His_6$ tag fused to the N-terminus by an enterokinase cleavage site. The fidelity of the cloned inserts was confirmed by bidirectional DNA sequencing.

Six liters of LB medium inoculated with *E. coli* BL21/DE3* harboring each expression vector was incubated at 37 deg C in shaking flasks to an $OD_{600nm}$ of ~0.6. Isopropyl β-D-thiogalactopyranoside (IPTG) was added to a final concentration of 0.1 mM for pET150-Pul13 and incubation continued for ~4 h. For pET150-X28, pET100-X28/45/20, pET100-X45, and pET100-X20, the flasks were incubated at 37 deg C with shaking to an $OD_{600nm}$ of ~0.6, at which point the temperature was cooled to room temperature (~21 deg C) and incubation continued for ~16 h. The addition of IPTG was not necessary. The cells were harvested by centrifugation at 5000 rpm for 10 min, resuspended in 90 mL of 20 mM Tris, pH 8.0, containing 0.5 M NaCl, and lysed by French press. The pET150-Pul13 cells were resuspended in BugBuster protein extraction reagent (Novagen, Madison, WI) and

lysed according to manufacturer's protocols. After centrifugation at 15000 rpm for 45 min, the supernatant was collected, and the polypeptides were purified by immobilized metal affinity chromatography (IMAC) with HIS-Select nickel affinity gel (Sigma, St. Louis, MO) according to manufacturer's protocols. Purified polypeptides were concentrated in a stirred ultrafiltration unit on a 5K molecular weight cutoff filter and dialyzed extensively against 50 mM Tris, pH 7.5, using regenerated cellulose dialysis tubing with a 3K MWCO. Purity was greater than 95%, as assessed by SDS-PAGE. Yields were typically 50 mg/L of culture or more.

*Determination of Protein Concentration.* The concentration of purified proteins was determined by UV absorbance at 280 nm using the following calculated extinction coefficients [109]: 34850 $M^{-1}$ $cm^{-1}$ for *Tm*CBM41 (X28), 7680 $M^{-1}$ $cm^{-1}$ for X45, 35560 $M^{-1}$ $cm^{-1}$ for X20, 76810 $M^{-1}$ $cm^{-1}$ for X28/45/20, and 155730 $M^{-1}$ $cm^{-1}$ for *Tm*Pul13.

*Affinity Electrophoresis.* The binding of all polypeptides was assessed by affinity electrophoresis [146] in 10% native polyacrylamide gels polymerized without polysaccharide or in the presence of 0.5% amylopectin, pullulan, amylose, or dextran. Electrophoresis was performed for 2 h at room temperature with native running buffer (25 mM Tris-base, 0.2 M glycine) in an XCell SureLock Mini-Cell system (Invitrogen, Carlsbad, CA) at a constant voltage of 100.

*Macroarrays.* BSA, *Cc*CBM17, *Tm*CBM4-2, and *Tm*CBM41 were labeled with Alexa Fluor 680 succinimidyl ester (Molecular Probes, Eugene, OR) according to the manufacturer's recommendations. Free label removal and buffer exhange were achieved by gel filtration using Sephadex G-25 (Amersham Biosciences). Macroarrays were

prepared by spotting 1 μL of a 10 or 1 mg/mL solution of polysaccharide onto a nitrocellulose membrane. The membranes were allowed to dry for at least 2 h. Membranes were then blocked for 1 h at 10 deg C with a solution of PBS containing 1% BSA and 0.05% Tween 20. These were probed by incubation at 10 deg C overnight with 10 mL of blocking buffer containing ~10 mg/mL Alexa Fluor 680 labeled protein. Probed membranes were briefly washed twice with PBS and then laser scanned with an excitation wavelength of 700 nm using a LICOR Odyssey infrared scanner (LICOR, Lincoln, NE). Binding was visualized by the presence of fluorescence.

*Isothermal Titration Calorimetry.* Isothermal titration calorimetry (ITC) was performed as described previously [68] using an VP-ITC (MicroCal, Northampton, MA). All polypeptides were dialyzed into 50 mM Tris buffer (pH 7.5). Carbohydrate solutions were prepared by mass using dialysate buffer saved from the protein dialysis. Protein and carbohydrate solutions were filtered (0.2 μm) and degassed prior to use. Experiments were performed at 25 deg C. Titrations were performed by injecting 10 μL samples of oligosaccharide solutions at 1-5 mM into the ITC sample cell containing 50-150 μM *Tm*CBM41. The concentrations of protein were chosen such they were in 5-fold or greater excess of the dissociation constants. Heats of dilution upon titration of buffer into carbohydrate or buffer into buffer were negligible. Reverse titrations were performed by injecting 4-10 μL samples of concentrated *Tm*CBM41 (1.385-2.22 mM) into solutions containing M4, M5, M6, pullulan, or amylopectin. These binding data were corrected for the heats of dilution determined by titrating *Tm*CBM41 into buffer. Binding stoichiometries, enthalpies, and equilibrium association constants were determined by fitting the corrected data to the appropriate binding model (see text) with MicroCal

Origin 7. In the cases of pullulan and amylopectin, the carbohydrate concentrations were expressed as equivalents of a tetrasaccharide in all of the calculations. Using a method adapted from Sigurskjold et al. [144], the acceptor concentration (i.e., the concentration of binding sites in the polysaccharides in terms of tetrasaccharide equivalents) was a regressed parameter. All data show the average and standard deviation of three independent titrations.

*UV Difference Titrations*. High-precision, automated UV difference titrations were performed using equipment essentially the same as that described previously [70; 111]. Protein and carbohydrate samples were prepared identically to those for ITC, filtered, and degassed prior to use. All experiments were held at 25 deg C in a Peltier thermostated cuvette holder. Samples of carbohydrate (1 mM) were added to 2 mL of protein (~33 μM) and allowed to equilibrate for 80 s with stirring. One thousand scans (10 ms integration time) collected from 225 to 380 nm were averaged for each carbohydrate addition. The protein concentration used for each titration was calculated from the absorbance at 280 nm extracted from the spectrum at zero carbohydrate concentration. Each spectrum was corrected for dilution due to the addition of ligand, and the difference spectra were calculated by subtraction of the absorbance spectrum at zero carbohydrate concentration. Difference spectra were examined for peak and trough wavelengths and values at the appropriate wavelengths extracted for further analysis. The peak-to-trough heights at the wavelength pairs 292.72/288.82 nm, 285.06/288.82 nm, and 285.06/278.24 nm were calculated by subtraction of the trough values from the peak values and the data plotted against total carbohydrate concentration. The resulting isotherms were analyzed with DynaFit [147] using the model equilibria as discussed in the text. Data for the three

wavelength pairs were fit simultaneously to improve the precision of the regressed parameters. The parameters determined were the $K_a$ and molar difference absorbance response ($\Delta A$). Though the CBM concentration ([M]) was measured during the experiment, because the experiments were performed under pseudo-first-order conditions (acceptor concentration >10-fold in excess of $K_d$), this parameter was also allowed to float in the analysis to give a regressed value of [M] (called [$M_{fit}$]). Under these conditions, [$M_{fit}$] actually represents an experimental estimate of the concentration of total binding sites, rather than just the macromolecule concentration. Assuming that the ligand concentration is known with accuracy as it is prepared by mass from pure lyophilized material, the ratio [$M_{fit}$]/[M] (i.e., the experimentally determined concentration of total binding sites divided by the known concentration of CBM) gives an estimate of the number of ligand binding sites on M, i.e., the stoichiometry or *n* value. The data reported are the averages and standard deviations of four independent titrations.

*Analytical Ultracentrifugation.* Sedimentation equilibrium experiments were carried out using a Beckman XL-I analytical ultracentrifuge using an An-60 Ti (titanium) rotor. The samples were loaded on six-hole charcoal-filled Epon 12 mm cells. All runs were carried out at 20 deg C. The scans were analyzed using XL-A UltraScan II version 6.0 sedimentation data analysis software (Borries Demeler, Missoula, MT) using a global nonlinear, least squares, curve fitting [148; 149]. The protein samples at different concentrations ranging from an $OD_{230}$ of 0.3-0.8 to an $OD_{280}$ of 0.3-0.8 were extensively dialyzed against 50 mM Tris-HCl (pH 7.5) buffer and were analyzed for equilibrium conditions achieved at different rotor speeds (see legend to Figure 29 for more details).The partial specific volume of the *Tm*CBM41 protein (0.740 $cm^3 \cdot g^{-1}$) was

calculated from its amino acid composition according to ref [150] using the partial specific volumes provided in ref [151].

### 3.2.4 Results and Discussion

*Modular Properties of TmPul13 and Identification of Its Carbohydrate-Binding Module.* Pullulanase activity has been demonstrated for *Tm*Pul13, but the modular arrangement and the functions of the accessory modules have not been reported. On the basis of PSI-BLAST amino acid sequence alignments [130] of the entire *Tm*Pul13 sequence, this protein appeared to comprise four modules (Figure 25). The three N-terminal modules have undetermined function and fall into the X-module families X28, X45, and X20 [22]. The large C-terminal module shows sequence homology to glycoside hydrolases from family 13, so it is therefore highly likely that this is the catalytic module responsible for *Tm*Pul13's pullulanase activity. To determine if any of these accessory modules are CBMs, each individual module (called X28, X45, and X20, respectively), a triple module consisting of all the X domains (called X28/45/20), and the full-length enzyme (*Tm*Pul13) were cloned and produced independently in *E. coli* (see Figure 25). All of the polypeptides could be produced at high levels in the cytoplasm and purified to homogeneity by IMAC.

X28, X28/45/20, and *Tm*Pul13 bound to amylopectin, pullulan, and amylose but not to dextran [linear α-(1,6)-linked glucose] as assessed by affinity gel electrophoresis (AGE) binding experiments [146] (Table 8). Neither, X45 nor X20 bound to any of these polysaccharides. These results clearly indicated that X28 is a CBM having affinity for α-glucans with α-(1,4) or mixed α-(1,4)(1,6) linkages but has no affinity for α-glucans of

Figure 25: Modular organization of *Tm*Pul13. Amino acid numbers corresponding to the module boundaries are shown above the schematic. The individual module constructs used in this study are also shown.

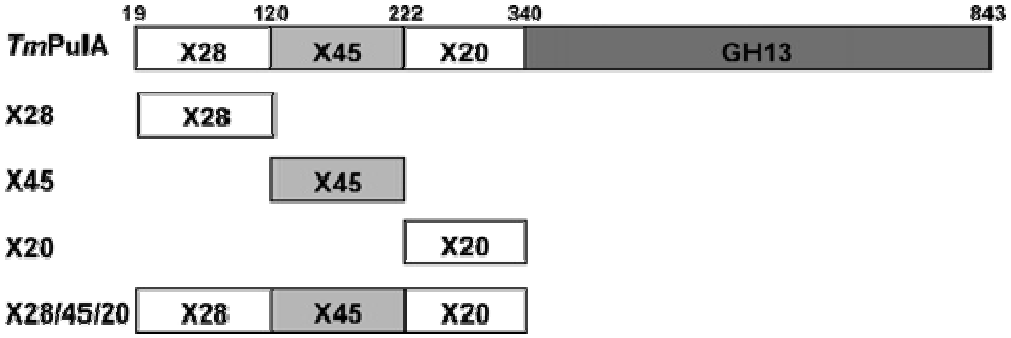


Table 8: Qualitative Assessment of Binding of *Tm*Pul13 and Its Modules to α-Glucans Determined by Affinity Electrophoresis

| Protein | Amylopectin | Amylose | Pullulan | Dextran |
|---|---|---|---|---|
| TmPul13 | +a | + | + | - b |
| X28/X45/X20 | + | + | + | - |
| TmCBM41 (X28) | + | + | + | - |
| X45 | - | - | - | - |
| X20 | - | - | - | - |

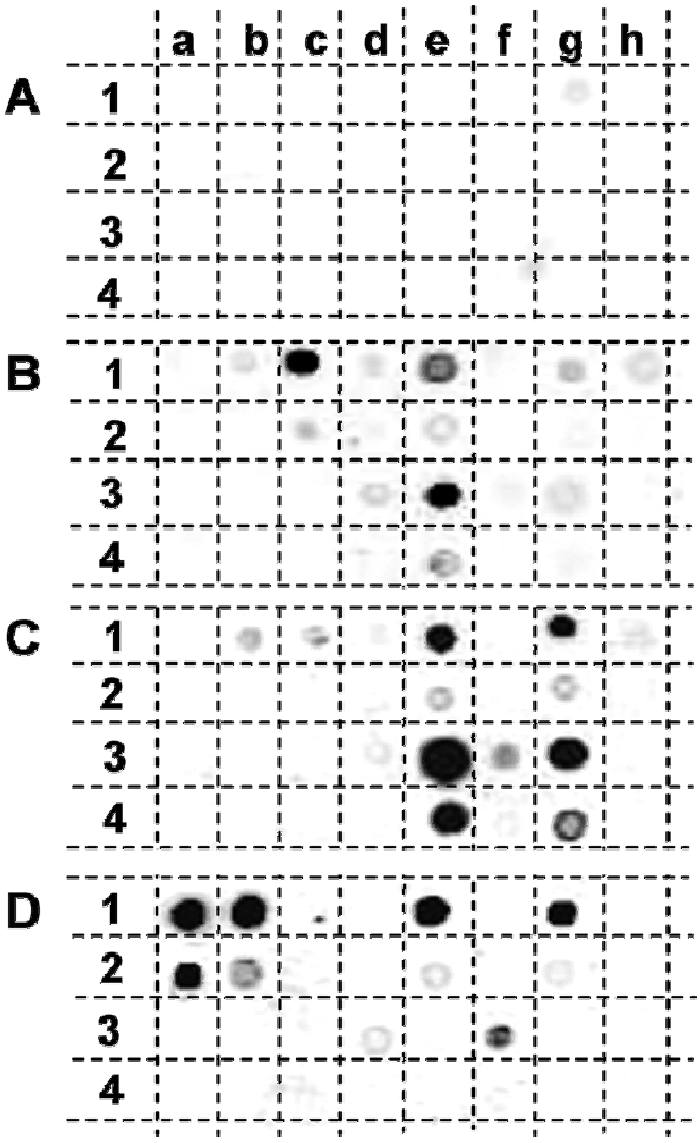
a + indicates binding

b - indicates no binding.

α-(1,6)-linked glucose. On the basis of amino acid sequence comparisons, X28 and its homologues comprise a new family of CBMs now classified as CBM family 41. Thus, the X28 module from *Tm*Pul13 will be referred to hereafter as *Tm*CBM41. The functions of the X45 and X20 modules remain to be determined. A more complete analysis of *Tm*CBM41's specificity to polysaccharides was performed by a macroarray binding assay (Figure 26). This confirmed binding to amylopectin and pullulan and the lack of binding to dextran. *Tm*CBM41 also displayed binding to pectin galactan, wheat arabinoxylan, and birchwood xylan, revealing cross-specificity for other polysaccharides. However, the binding to these polysaccharides was sufficiently weak to be unquantifiable by ITC. Thus, *Tm*CBM41 has a clear preference for α-glucans. Curiously, *Tm*CBM4-2 did not appear to bind to laminarin (a β-(1,3)-glucan), for which *Tm*CBM4-2 has a very high affinity, unfortunately suggesting that this relatively short and soluble polysaccharide was not effectively immobilized on the membrane.

*A Kinetic Model of Oligosaccharide Recognition by TmCBM41*. The addition of maltooligosaccharides to *Tm*CBM41 resulted in relatively large changes in the UV spectra as shown in the UV difference spectra (not shown). Large peaks at ~293, 285, and 275 nm and troughs at ~289 and 278 nm are distinctive of the involvement of tryptophan side chains in binding [70; 120]. The lack of the distinctive tyrosine peak at 286.5 nm suggested no involvement of tyrosine residues in binding; however, any tyrosine signal may have been masked by the large tryptophan signal. The peak-to-trough heights at wavelength pairs 292.72/288.82 nm, 285.06/288.82 nm, and 285.06/278.24 nm were dependent on the concentration of added carbohydrate and were used to quantitatively assess the binding interactions (Figure 27). Preliminary fitting of the binding isotherms to

Figure 26: Polysaccharide macroarray binding analysis of Alexa Fluor 680 labeled *Tm*CBM41. (A) BSA (negative control), (B) *Cc*CBM17 (positive control), (C) *Tm*CBM4-2 (positive control), and (D) *Tm*CBM41. Proteins were fluorescently labeled, and blots were prepared and probed as described in Materials and Methods. Polysaccharides tested include (1a) 1% amylopectin, (1b) 1% pullulan, (1c) 1% glucomannan, (1d) 1% galactomannan, (1e) 1% pectic galactan, (1f) 1% arabinogalactan, (1g) 1% wheat arabinoxylan, (1h) 1% xyloglucan; (2a-2h) polysaccharides in the same order except at 0.1%; (3a) blank, (3b) 1% dextran, (3c) 1% pachyman, (3d) 1% laminarin, (3e) 1% oat β-glucan, (3f) 1% birchwood xylan, (3g) 1% methylcellulose, (3h) 0.1% glycerol; (4a-4g) polysaccharides in the same order except at 0.1%, (4h) blank. Both positive controls displayed the expected binding specificity based on previous studies [70; 120] except that *Tm*CBM4-2 did not appear to bind laminarin, suggesting that this short and soluble polysaccharide was not effectively immobilized on the membrane.



a binding model assuming 1:1 interactions (equilibrium 1 in Figure 28) yielded two observations. First, the affinities for maltotriose (M3) up to maltoheptaose (M7) were in the vicinity of $5 \times 10^5$ to $2 \times 10^6$ $M^{-1}$. Second, the stoichiometry of binding appeared to steadily decrease from a 1:1 carbohydrate:protein ratio in the case of M3 and M4 to ~0.5:1 for M7. Because the concentration of protein used for these experiments was 16-70-fold in excess of the dissociation constants, the binding isotherms at total ligand concentrations less than the CBM concentration (~33 μM) should approximate pseudo-first-order reactions. Thus, the change in stoichiometry with ligand length was visually evident in the different initial slopes of the binding isotherms (Figure 3). This indicated that some of the interactions were not simple 1:1 interactions. The 1:2 carbohydrate:protein stoichiometry for M7 binding is reminiscent of observations made for the family 27 CBM from *T. maritima* binding to mannohexaose [123] and a mutant of the family 29 CBM from *Pyromyces equi* binding to cellohexaose [67]. In both cases the hexasaccharides were able to function as divalent acceptors, as in equilibrium 2 (see Figure 28). Such a model could explain the interaction of *Tm*CBM41 with maltooligosaccharides. Another possibility is that *Tm*CBM41 preassociates to form a dimer prior to binding carbohydrate, resulting in a 1:2 stoichiometry (equilibrium 3 in Figure 28).

Figure 27: Quantitative UV difference analysis of *Tm*CBM41 binding to α-glucooligosaccharides. Panel A: Isotherms of M3 (closed triangles), M4 (open circles), M5 (closed circles), M6 (open squares), and M7 (closed squares) titrated into *Tm*CBM41. The absorbance difference shown is the peak to trough height at 292.72 and 288.82 nm, respectively. Panel B: Isotherm of M7 titrated into *Tm*CBM41. The curves show the data at the wavelength pairs of 292.72/288.82 nm (closed circles), 285.06/288.82 nm (open circles), and 285.06/278.24 nm (closed squares). Solid lines show the fits resulting from the application of equilibrium 2 (see Figure 28). Panels C and D show the residuals resulting from the fits of equilibrium 2 and equilibrium 1, respectively.

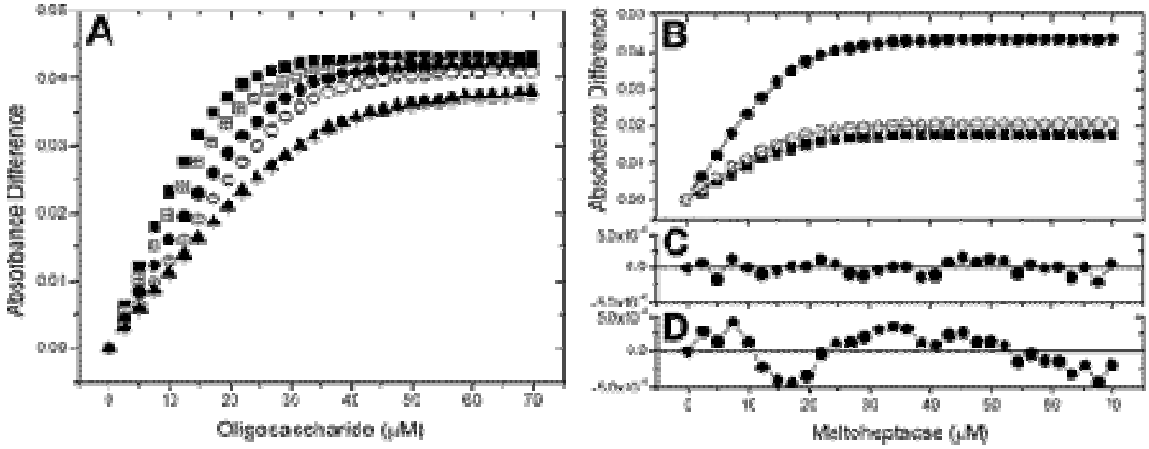


Figure 28: Equilibria used to model the interactions of *Tm*CBM41 with α-glucooligosaccharides. M represents *Tm*CBM41 and L the α-glucooligosaccharide. LM represents the 1:1 complex of *Tm*CBM41 and the α-glucooligosaccharide while $LM_2$ represents the α-glucooligosaccharide with two bound *Tm*CBM41 molecules. $M_2$ represents a *Tm*CBM41 dimer.

Equilibrium 1:

$$M + L \rightleftharpoons LM$$

Equilibrium 2:

$$M + L \rightleftharpoons LM$$

$$LM + M \rightleftharpoons LM_2$$

Equilibrium 3:

$$M + M \rightleftharpoons M_2$$

$$M_2 + L \rightleftharpoons LM_2$$

The fits of the data to three binding models (equilibria 1-3 in Figure 28) were evaluated with the program DynaFit [147] in order to discriminate between the possible binding mechanisms. The fits of all of the data using equilibrium 3 were extremely poor, and the model was immediately rejected on a statistical basis (large sum-of-squares values and clear systematic deviations between the model data and the experimental data; not shown). The data for $6^3$-α-D-glucosylmaltotriose (GM3; glucose linked α-(1-4) to maltotriose), $6^3$-α-D-glucosylmaltotriosylmaltotriose (GM3M3; maltotriose linked α-(1-4) to GM3), maltose (M2), maltotriose (M3), and M4 were adequately described by the bimolecular model of equilibrium 1. Trial of equilibrium 2 failed to reduce the sum of the squares of the fits. Furthermore, *P*-values of fit comparisons indicated a greater than 99% chance that the single binding site model (equilibrium 1) was more appropriate for these carbohydrate ligands. The near unitary stoichiometries obtained from the fits were consistent with this (Table 9). In contrast, the M5-M7 isotherms were not well modeled by equilibrium 1. The application of equilibrium 2 greatly reduced the sum of the squares of the fits. The residuals of the fits to equilibrium 2 (Figure 27C) and equilibrium 1 (Figure 27D) show random scatter for equilibrium 2 but large systematic deviations for equilibrium 1. Statistical comparison of equilibrium 1 vs equilibrium 2 indicated an insignificantly small probability (<1%) that equilibrium 1 was the more appropriate model. Thus, equilibrium 2 was deemed the best description of the interaction of *Tm*CBM41 with M5, M6, and M7.

Experiments performed by isothermal titration calorimetry (ITC) yielded the same conclusions. Titrations performed by titrating GM3, GM3M3, M2, M3, and M4 into *Tm*CBM41 gave isotherms of the expected sigmoidal shape (Figure 29A). These

isotherms could be analyzed in a symmetrical manner. That is, when the same isotherm was fitted using a one-site binding model (equilibrium 1, Figure 28), the same results were obtained if the protein was treated as the macromolecule (M) and the carbohydrate as the ligand (L) and when reversing these assignments (results not shown). Furthermore, a reverse titration of *Tm*CBM41 into M4 gave essentially the same results as the titration of M4 into protein (Table 9, Figure 29B), confirming the symmetry of the system and the suitability of using a single binding site model (equilibrium 1). As would be expected, analysis of the experiments using these ligands gave stoichiometries near unity (Table 10). The isotherms of M5, M6, and M7 titrated into *Tm*CBM41 were not sigmoidal (Figure 29C), and analyses using equilibrium 1 gave extremely poor *P*-values (<0.005) from run test analysis of the residuals, indicating strong systematic deviations between the data and model. The reverse titrations of *Tm*CBM41 into M5 (Figure 29D) and M6 were essentially sigmoidal but also did not fit a single site binding model as judged by the same statistical tests. Analyses that treated *Tm*CBM41 as having a single binding site and the carbohydrate as having two different and independent binding sites (i.e., equilibrium 2) gave good agreement between the model and the data. Thus, the ITC results were in accord with the UV difference binding studies. The binding of β-cyclodextrin (cycloheptaamylose), however, was a more ambiguous case. β-Cyclodextrin clearly presented two binding sites, each accommodating a single CBM. However, these binding sites appeared to be so similar in affinity that the individual parameters for binding these sites could not be resolved in these experiments. Thus, the two binding sites were treated as independent but, unlike with M5-M7, identical in the analysis.

Table 9: Parameters of *Tm*CBM41 Binding to Maltooligosaccharides Determined by UV Difference Titrations at 25 °C in 50 mM Tris, pH 7.5

| Ligand[a] | Species Formed[b] | $K_a$ ($x10^5$ $M^{-1}$) | $\Delta G$ (kcal/mole) | $\Delta A^c$ ($x10^3$ units/μmole/ L) | $n^d$ |
|---|---|---|---|---|---|
| M2 | LM | 0.53 (± 0.01) | -6.45 (± 0.03) | 1.15 (± 0.02) | 1.00 |
| M3 | LM | 4.26 (± 0.20) | -7.68 (± 0.03) | 1.23 (± 0.03) | 1.00 (± 0.02) |
| M4 | LM | 11.18 (± 1.36) | -8.25 (± 0.07) | 1.24 (± 0.05) | 0.92 (± 0.02) |
| M5 | LM | 22.72 (± 4.56) | -8.40 (± 0.12) | 1.25 (± 0.07) | 1.02 (± 0.05) |
| | $LM_2$ | 0.25 (± 0.13) | -6.02 (± 0.33) | 1.26 (± 0. 77) | |
| M6 | LM | 22.59 (± 6.81) | -8.29 (± 0.41) | 1.20 (± 0.04) | 1.05 (± 0.03) |
| | $LM_2$ | 0.49 (± 0.21) | -6.35 (± 0.40) | 1.68 (± 0.51) | |
| M7 | LM | 14.58 (± 12.99) | -8.01 (± 0.47) | 1.25 (± 0.04) | 1.02 (± 0.05) |
| | $LM_2$ | 1.22 (± 0.07) | -6.73 (± 0.04) | 1.64 (± 0.11) | |
| GM3 | LM | 0.27 ± (0.01) | -6.04 (± 0.01) | 1.25 (± 0.03) | 1.00 |
| GM3M3 | LM | 1.79 (± 0.14) | -7.16 (± 0.05) | 1.20 (± 0.02) | 0.85 (± 0.01) |

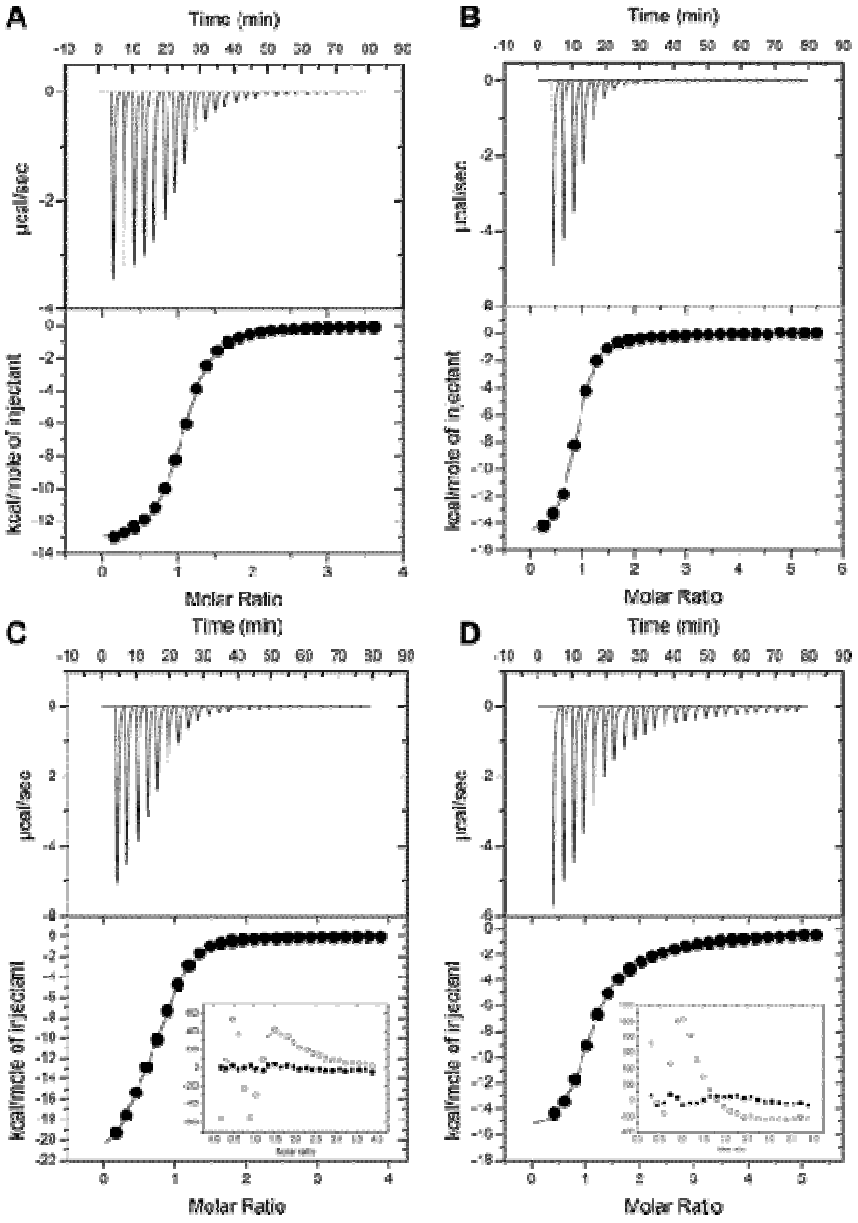
[a] Abbreviations: M2, maltose; M3, maltotriose; M4, maltotetraose; M5, maltopentaose; M6, maltohexaose; M7, maltoheptaose; GM3, $6^3$-α-D-glucosylmaltotriose; GM3M3, $6^3$-α-D-glucosylmaltotriosyl maltotriose.[b] GM3, GM3M3, and M2-M4 were fit to a one-site binding model (see equilibrium 1 in Figure 3.2.4); M5-M7 were fit to a binding model treating the carbohydrate as divalent (see equilibrium 2 in Figure 28). Notation corresponds to the species shown in the schemes representing the respective equilibria.[c] $\Delta A$ = molar UV difference response using the peak-to-trough height at 292 and 289 nm.[d] Binding stoichiometry. Where no error is reported, the value was fixed as a constant during the nonlinear fitting. The values for M5-M7 represent the number of binding sites per CBM.

Table 10: Parameters of *Tm*CBM41 Binding α-Glucans Determined by Isothermal Titration Calorimetry at 25 °C in 50 mM Tris, pH 7.5

| Ligand[a] | Species Formed[b] | $K_a$ ($x10^5$ $M^{-1}$) | ΔG (kcal/mole) | ΔH (kcal/mole) | ΔS (cal/mole/K) | *n*[c] |
|---|---|---|---|---|---|---|
| M2 | LM | 0.33 (± 0.00) | -5.97 (± 0.01) | -13.07 (± 0.03) | -23.15 (± 0.07) | 1.06 (± 0.01) |
| M3 | LM | 2.76 (± 0.04) | -7.20 (± 0.01) | -14.80 (± 0.23) | -24.75 (± 0.78) | 0.94 (± 0.01) |
| M4 | LM | 5.42 (± 0.08) | -7.59 (± 0.01) | -13.73 (± 0.44) | -19.80 (± 1.41) | 1.00 (± 0.05) |
| M4 (rev) | LM | 7.46 (± 0.19) | -7.77 (± 0.01) | -13.42 (± 0.17) | -18.15 (± 0.49) | 0.83 (± 0.00) |
| M5 | LM | 15.75 (± 2.62) | -8.20 (± 0.10) | -14.76 (± 0.18) | -21.15 (± 0.92) | 1.06 (± 0.10) |
| | $LM_2$ | 0.40 (± 0.11) | -6.08 (± 0.16) | -7.58 (± 0.25) | -4.41 (± 1.41) | 1.16 (± 0.04) |
| M5 (rev) | LM | 5.32 (± 0.32) | -7.57 (± 0.03) | -15.99 (± 0.32) | -27.40 (± 0.57) | 0.98 (± 0.04) |
| | $LM_2$ | 0.14 (± 0.01) | -5.47 (± 0.01) | -8.77 (± 2.59) | -10.49 (± 8.92) | 1.00 (± 0.23) |
| M6 | LM | 10.80 (± 0.73) | -7.98 (± 0.06) | -13.70 (± 0.05) | -18.20 (± 0.26) | 1.07 (± 0.05) |
| | $LM_2$ | 0.70 (± 0.03) | -6.41 (± 0.04) | -11.30 (± 0.86) | -15.70 (± 2.90) | 1.10 (± 0.01) |
| M6 (rev) | LM | 20.70 (± 1.84) | -8.35 (± 0.05) | -14.39 (± 0.03) | -19.35 (± 0.07) | 0.76 (± 0.00) |
| | $LM_2$ | 1.04 (± 0.04) | -6.63 (± 0.02) | -11.61 (± 0.13) | -15.95 (± 0.50) | 0.77 (± 0.01) |
| M7 | LM | 9.24 (± 2.50) | -7.87 (± 0.23) | -14.00 (± 0.16) | -19.60 (± 0.94) | 1.15 (± 0.08) |
| | $LM_2$ | 0.69 (± 0.18) | -6.38 (± 0.22) | -6.65 (± 4.33) | -0.18 (± 14.50) | 0.67 (± 0.22) |
| GM3 | LM | 0.26 ± (0.01) | -6.02 (± 0.01) | -12.40 (± 0.20) | -21.50 (± 0.71) | 1.11 (± 0.04) |
| GM3M3 | LM | 1.50 (± 0.07) | -7.06 (± 0.00) | -17.2 (± 0.38) | -34.1 (± 1.27) | 1.05 (± 0.01) |
| β-CD | LM & $LM_2$ | 4.20 (± 0.13) | -7.44 (± 0.03) | -12.70 (± 0.04) | -16.70 (± 0.17) | 1.90 (± 0.05) |
| β-CD (rev) | LM & $LM_2$ | 3.46 (± 0.00) | -7.33 (± 0.00) | -10.71 (± 0.05) | -10.55 (± 0.21) | 1.94 (± 0.05) |
| AmPec (rev) | LM | 1.29 (± 0.37) | -6.75 (± 0.17) | -15.17 (± 1.48) | -27.50 (± 4.38) | 0.32[d] (± 0.04) |
| Pullulan (rev) | LM | 1.01 (± 0.05) | -6.62 (± 0.27) | -16.66 (± 0.12) | -32.95 (± 0.33) | 0.23[d] (± 0.01) |

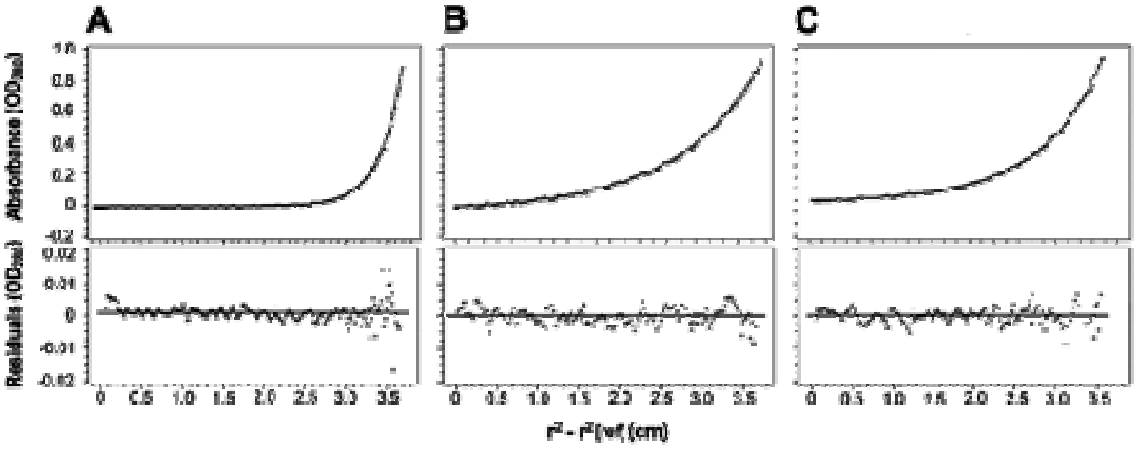
[a] M2, maltose; M3, maltotriose; M4, maltotetraose; M5 maltopentaose; M6, maltohexaose; M7, maltoheptaose; GM3, $6^3$-α-D-glucosylmaltotriose; GM3M3, $6^3$-α-D-glucosylmaltotriosyl maltotriose; β-CD, β-cylodextrin; AmPec, amylopectin. Experiments performed by titrating protein into carbohydrate (reverse titrations) are indicated by (rev).[b] GM3, GM3M3, AmPec, pullulan, β-CD and M2-M4 were fit to a one-site binding model (see Equilibrium 1 in Figure 5); M5-M7 were fit to a binding model treating the carbohydrate as divalent (see Equilibrium 2 in Figure 28). Notation corresponds to the species shown in the schemes representing the respective equilibria. [c] binding stoichiometry. The values for M5-M6 represent the number of binding sites per CBM.[d] amylopectin and pullulan are polysaccharides of unknown length. These stoichiometries are expressed as moles of CBM per tetraose equivalent (see Materials and Methods).

Figure 29: Isotherms of *Tm*CBM41 binding to α-glucooligosaccharides produced by ITC. Panels A and B show the isotherms of *Tm*CBM41 interacting with M4; panels C and D show the isotherms of *Tm*CBM41 interacting with M5. The isotherms in panels A and C were produced by titrating M4 or M5, respectively, into *Tm*CBM41. The isotherms in panels B and D were produced by titrating *Tm*CBM41 into M4 or M5, respectively. The solid lines in the lower panes of panels A and B show the fits of a one-site binding model (equilibrium 1). The solid lines in the lower panes of panels C and D show the fits of a two-site binding model where the carbohydrate was treated as the divalent species (equilibrium 2). Insets in panels C and D show plots of the residuals for the model of equilibrium 1 (open circles) and equilibrium 2 (closed circles).



The recognition of all of the oligosaccharides tested, other than M5-M7, appears to adhere to equilibrium 1 whereas the recognition of M5-M7 appears to go by equilibrium 2. Should this be the case, then at subsaturating concentrations of M5-M7 there should be a detectable concentration of dimeric species, but at excess M5-M7 concentrations, where the singly ligated high-affinity species is preferred, there should be no dimeric species. Likewise, in the absence of ligand or with any concentration of the other oligosaccharides (i.e., M2-M4, GM3, and GM3M3) no dimers should be formed. Sedimentation equilibrium analytical ultracentrifugation analysis of *Tm*CBM41 revealed only monomeric species in the absence of ligand, with a 1:2 molar ratio of M3 to protein, with a 5:1 molar ratio of M3 to protein, and with a 5:1 molar ratio of M6 to protein (Figure 30). The determined molecular masses were 15670, 15780, 16310, and 16700 Da for these species, respectively. These values are in reasonably good agreement with the expected molecular mass of 15391 Da for monomeric *Tm*CBM41, 15895 Da for monomeric *Tm*CBM41 + M3, and 16382 Da for monomeric *Tm*CBM41 + M6. In contrast, the experiments performed with a 1:2 molar ratio of M6 to protein yielded data that was inconsistent with the presence of a single monomeric molecular species. These data could be satisfactorily modeled using a two-component system representing a monomer-dimer equilibrium of a 16370 Da species (32740 Da for the dimer), suggesting an association constant of dimerization estimated to be (2-4) $\times 10^4$ $M^{-1}$, in acceptable agreement with the association constants estimated by UV difference and ITC for the formation of the maltohexaose ligated dimeric species [i.e., (~5-10) $\times 10^4$ $M^{-1}$; Tables 9 and 10]. Thus, these results are entirely consistent with the proposed kinetic model of maltooligosaccharide binding.

Figure 30: Sedimentation equilibrium analysis of *Tm*CBM41 (4.3 μM) in the absence (A) and in the presence of either 2.15 μM maltotriose (B) or 2.15 μM maltohexose (C). The upper plots show the absorbance at 280 nm as a function of the square of the radial distance of the sample at any position within the cell ($r$) minus the square of the radial position at a reference point ($r$[ref]) ($r^2$ - $r^2$[ref]). The continuous lines in the upper plots were obtained by fitting the experimental data (circles) to a single ideal component model of $M_r$ 15670 (A), to a single ideal component of $M_r$ 16310 (B), and to a monomer-dimer model of monomer $M_r$ 16370 (C). The lower plots show $\chi^2$ residuals as a function of $r^2$ - $r_0^2$ for the best fit (solid line). The rotor speed of the equilibrium plots shown in this figure was taken at 45000 rpm at 20 degrees C. The buffer composition was 50 mM Tris-HCl (pH 7.5). The $A_{280}$ of the starting sample was approximately 0.3.



On the basis of the dependence of $K_a$ on the length of the oligosaccharides, *Tm*CBM41 requires four to five sugars for maximal binding. Thus, M5, M6, and M7 must comprise up to two, three, and four overlapping and identical binding sites, respectively. Initial occupation of the first of these overlapping binding sites is likely to be the high-affinity interaction. The second binding site must then be some fragment of the maximal four to five sugar binding site, most likely an overhang at a terminus of the oligosaccharide. In strict terms, the two binding sites on the oligosaccharides are nonidentical but dependent: formation of the low-affinity site first depends on occupation of the high-affinity site and possible rearrangement to accommodate binding of the second CBM. The discrepancy between the M5-M7 data obtained in both titration modes likely reflects the inability to account for negative cooperativity. In the reverse titration mode (i.e., protein titrated into sugar) the occupation of the binding sites is sequential due to the initial occupation of the highly favored high-affinity sites, and thus, the approach that we employed to analyze the reverse titration data, which assumed nonidentical but independent binding sites, still provides reasonable approximations. However, at low sugar to protein molar ratios when sugar was titrated into protein, both high-affinity and low-affinity interactions would occur simultaneously, resulting in possibly large negatively cooperative effects. Failure to account for this would cause inaccuracies in determining the binding parameters. This mechanism of binding maltooligosaccharides longer than M4 is the same as the mechanism proposed for *Tm*CBM27 binding to mannohexaose [123]. A similar mechanism likely also explains the ligand-mediated dimerization of a mutant of *Pe*CBM29 [67]. However, in the case of the *Pe*CBM29 mutant, the formation of the species with one CBM bound to one sugar occurs with a low affinity. Binding of the second CBM to this

complex occurs with a high affinity and stabilizes the doubly ligated species of two CBMs per sugar. Unlike the *Tm*CBM41 scenario, this must involve positive cooperativity whereby binding of the first CBM creates a new binding site of greater affinity for another CBM, either by presentation of a sugar with an optimal binding conformation or by the formation of additional CBM-CBM interactions.

The potential biological significance of these dimerization events is unclear. *Tm*CBM41 does not appear to dimerize upon binding to complex sugars that more accurately represent the biological substrate of *Tm*Pul13 (see below), leaving the potential function of this phenomenon ambiguous. However, manipulation of this occurrence may lead to applications that require tunable dimerization.

*Recognition of Complex Oligosaccharides.* Pullulan is a repeating unit of α-(1,6)-linked M3. To assess the impact of the α-(1,6) linkages on the ability of *Tm*CBM41 to bind α-glucans, we studied the interaction of *Tm*CBM41 with GM3, GM3M3, isomaltose [disaccharide of α-(1,6)-linked glucose], isomaltotriose [trisaccharide of α-(1,6)-linked glucose], and panose [a trisaccharide of glucose with an α-(1,6) linkage followed by an α-(1,4) linkage]. *Tm*CBM41 did not bind to the latter three sugars. The affinity of *Tm*CBM41 for GM3 was similar to that of M2 and an order of magnitude weaker than for M3 and longer maltooligosaccharides (Tables 2 and 3). GM3M3 was only bound as tightly as M3 and comprised a single binding site, unlike similarly sized maltooligosaccharides. These results indicate that *Tm*CBM41 tolerates α-(1,6) linkages, but not particularly well, and only in the context of a sufficient number of α-(1,4)-linked glucose residues. Given the four to five sugar footprint of *Tm*CBM41, this CBM must

likely accommodate an internal α-(1,6) linkage giving *Tm*CBM41 the ability to bind to pullulan. However, its preference for α-(1,4) linkages would target *Tm*CBM41 to regions rich in this linkage, possibly leaving the α-(1,6) linkages, the preferred substrate for *Tm*Pul13, exposed and susceptible to cleavage.

*Recognition of Polysaccharides*. Pullulan and amylopectin ITC binding isotherms were obtained by titration of *Tm*CBM41 into the polysaccharide (not shown). The data for both polysaccharides could be suitably fit using a bimolecular interaction model (equilibrium 1 in Figure 28) where the concentration of the polysaccharide was expressed as equivalents of M4 (see Materials and Methods). Furthermore, Scatchard plots of the data were linear (not shown), further indicating a simple mode of interaction with a single class of binding site. The stoichiometry obtained for pullulan (Table 10) equated to one *Tm*CBM41 module binding every ~17 glucose residues, supporting the idea of well-spaced, independent binding sites. This was curiously at odds with the results obtained with GM3M3. One GM3M3 molecule accommodates a single CBM (Tables 9 and 10), suggesting that approximately every seven sugars in pullulan should constitute a binding site. Since the tertiary structure of pullulan in solution is not known, perhaps the decreased number of binding sites in pullulan is due to a conformation adopted by the polysaccharide that prevents binding of *Tm*CBM41 to some of the theoretical binding sites on the polysaccharide. The stoichiometry for amylopectin equated to one *Tm*CBM41 molecule binding every ~13 glucose residues (Table 10). The association constants for pullulan and amylopectin were ~$10^5$ $M^{-1}$ (Table 10), similar to that of M3 and GM3M3 binding. Considering the composition of pullulan, the similarity in affinities is, perhaps, expected. However, amylopectin contains infrequent α-(1,6) branches and, therefore,

contains longer stretches of linear α-(1,4)-linked glucose, leading one to expect affinities upward of those found for M4-M7. Again, this decrease in affinity may be due to a conformation adopted by amylopectin in solution that is nonoptimal for *Tm*CBM41 binding.

*Thermodynamic Mechanism of α-Glucan Binding.* All binding reactions between *Tm*CBM41 and α-glucan substrates reported were favored by a negative change in enthalpy (Δ*H*) (Table 10). In contrast, all changes in entropy (Δ*S*) were unfavorable. This signature is common to protein-carbohydrate interactions, CBMs in particular. Though Gibb's free energy of binding (Δ*G*) generally increased with oligosaccharide length, there was no correlation between oligosaccharide length and changes in Δ*H* and Δ*S,* suggesting different contributions to the thermodynamics from interactions at the different subsites in the CBM binding site.

The interaction of *Tm*CBM41 with GM3M3 was enthalpically more favorable than with other maltooligosaccharides (Table 10); however, the change in entropy for this interaction was comparably worse, resulting in a less favorable Δ*G*. A similar observation can be made for the interactions with amylopectin and pullulan in comparison with maltooligosaccharides. This latter phenomenon, whereby soluble polysaccharides are bound more weakly than oligosaccharides due to entropic penalties, appears to be common with CBMs (see refs [66; 101; 116; 121] for examples).

*A New Family of CBMs. Tm*CBM41 from *T. maritima* pullulanase Pul13 is a CBM that binds tightly to α-glucans. Modules with significant amino acid sequence identity are

found in no fewer than 34 proteins from 22 different bacterial species (Table 11). These modules frequently occur as repeats, almost exclusively in pullulanases. Interestingly, a great number of these pullulanases are from human pathogens. Recently, a membrane-associated pullulanase (GenPept accession number CAD32942) from *Streptococcus pyogenes* strain NZ131rgg was identified as a glycoprotein-specific adhesin [152]. In light of our study, it appears likely that the tandem N-terminal CBM41 modules of the *S. pyogenes* pullulanase are responsible for its glycoprotein binding activity. Thus, *Tm*CBM41 from *T. maritima* Pul13 is the first well-characterized member of a new, relatively large, and apparently diversely functioning family of carbohydrate-binding modules. This family has now been classified as CBM family 41 (carbohydrate-active enzyme families and CBM families can be accessed through the CAZy server at www.cazy.org).

Table 11: Proteins Containing Modules Similar to *Tm*CBM41

| Organism | Protein | GenPept accession number | Number of modules | % Identity with *Tm*CBM41 |
|---|---|---|---|---|
| *Thermotoga maritima* | pullulanase | NP_229641 | 1 | 100 |
| *Fervidobacterium pennivorans* | pullulanase type I | AAD30387 | 1 | 72 |
| *Bacillus cereus* | Pullulanase | NP_832487 | 1 | 49 |
| *Bacillus anthracis* | pullulanase | NP_845079 | 1 | 46 |
| *Bacillus sp.* | alkaline amylopullulanase | BAA11332 | 3 | 30, 21, 35 |
| *Streptomyces coelicolor* | α-amylase/dextrinase | NP_626477 | 2 | 39, 32 |
| *Micrococcus sp.* | α-amylase | A60999 | 2 | 23, 24 |
| *Bacillus halodurans* | alkaline amylopullulanase | NP_244564 | 2 | 33, 27 |
| *Streptomyces avermitilis* | pullulanase | NP_827159 | 2 | 28, 22 |
| *Streptococcus agalactiae* | pullulanase | NP_687867 | 1 | 31 |
| *Streptococcus agalactiae* | Unknown | NP_735320 | 1 | 31 |
| *Actinoplanes sp.* | α-amylase | CAC02970 | 2 | 31, 32 |
| *Bacillus sp.* KSM-1876 | alkaline pullulanase | BAB47586 | 2 | 24, 22 |
| *Streptomyces lividans* | 1,4- α-D-glucan glucanohydrolase, α-amylase precursor | Q05884 | 2 | 31, 40 |
| *Streptococcus pneumoniae* | pullulanase | AAG33958 | 2 | 21, 23 |
| *Streptococcus pneumoniae TIGR4* | alkaline amylopullulanase | NP_344806 | 2 | 21, 23 |
| *Streptococcus pneumoniae R6* | Alkaline amylopullulanase | NP_357841 | 2 | 21, 23 |
| *Streptococcus pyogenes* | pullulanase | NP_269940 | 2 | 15, 20 |
| *Streptococcus pyogenes* | pullulanase | CAD32942 | 2 | 15, 20 |
| *Streptococcus pyogenes* | pullulanase | NP_665498 | 2 | 15, 20 |
| *Streptococcus pyogenes* | pullulanase | NP_608014 | 2 | 15, 20 |
| *Streptococcus agalactiae* | Unknown | NP_735732 | 2 | 12, 18 |
| *Streptococcus agalactiae* | pullulanase | NP_688225 | 2 | 11, 17 |
| *Microbulbifer degradans* | hypothetical protein | ZP_00065715 | 1 | 22 |
| *Vibrio parahaemolyticus* | pullulanase precursor | NP_801148 | 1 | 19 |
| *Streptococcus mutans* | pullulanase | NP_721884 | 1 | 30 |
| *Streptococcus pneumoniae* | Thermostable pullulanase | NP_358619 | 1 | 23 |
| *Klebsiella pneumoniae* | α-dextrin endo-1,6- α-glucosidase | P07811 | 1 | 22 |
| *Klebsiella pneumoniae* | pullulanase precursor | P07206 | 1 | 19 |
| *Vibrio vulnificus* | Type II secretory pathway protein | NP_763140 | 2 | 22, 13 |
| *Klebsiella aerogenes* | pullulanase precursor | AAA25124 | 1 | 21 |
| *Saccharophagus degradans* | hypothetical protein | ZP_00065687 | 1 | 16 |
| *Klebsiella pneumoniae* | pullulanase precursor | S38801 | 1 | 20 |
| *Deinococcus radiodurans* | α-dextran endo-1,6- α-glucosidase | NP_294128 | 1 | 24 |

## 3.3 The Structural Basis of α-Glucan Recognition by a Family 41 Carbohydrate-binding Module from *Thermotoga maritima*

Alicia Lammerts van Bueren and Alisdair B. Boraston

Adapted from Journal of Molecular Biology 365 (3) 555-560 published 19 January 2007.

Contributions to research: Crystallization, data collection, structure refinement, manuscript and figure preparation.

### 3.3.1 Abstract

The N-terminal family 41 CBM, *Tm*CBM41 (from pullulanase PulA secreted by *Thermotoga maritima*) that we previously found to have α-glucan binding activity was shown to have specificity for α-1,4-glucans but was able to tolerate the α-1,6-linkages found roughly every three or four glucose units in pullulan. Using X-ray crystallography, the structures were solved for *Tm*CBM41 in an uncomplexed form and in complex with maltotetraose and $6^3$-α -D-glucosyl-maltotriose (GM3). Ligand binding was facilitated by stacking interactions between the α-faces of the glucose residues and two tryptophan side-chains in the two main subsites of the carbohydrate-binding site. Overall, this mode of starch binding is quite well conserved by other starch-binding modules. The structure in complex with GM3 revealed a third binding subsite with the flexibility to accommodate an α-1,4- or an α-1,6-linked glucose.

### 3.3.2 Introduction

Carbohydrate polymers are important in many biological processes. Among their many functions, they can serve structural roles, such as the cellulose in plant cell walls or chitin in crab and insect exoskeletons, or serve as storage molecules, such as the α-glucans starch and glycogen. The depolymerization of these polysaccharides, and others, is achieved primarily by glycoside hydrolases. Due to the recalcitrance of these polysaccharides to enzymatic degradation, many glycoside hydrolases contain carbohydrate-binding modules (CBMs), which serve to concentrate the enzyme onto its substrate thereby facilitating its action [56]. To date, there have been studies on seven CBM families that recognize α-glucans: CBM20 [98; 145] CBM21 [138; 153; 154],CBM25 and CBM26 [63], CBM34 [140], CBM41 [85], and CBM45 [155], with structures representing five of these families (20, 25, 26, 34 and 41).

*Thermotoga maritima* produces an 843 amino acid residue enzyme, *Tm*Pul13, that is active on the α-1,6-glycosidic linkages found in pullulan but has no activity on the α-1,6-linkages in amylopectin or on pure α-1,4-glucans [142; 143]. Previously, we dissected the modular structure of this enzyme and determined that the N-terminal ~100 residues defined a new family of α-glucan-specific CBMs (See section 3.2 and [85]. This module, *Tm*CBM41, bound tightly to pure α-1,4-glucans, α-1,4-glucooligosaccharides, mixed α-(1,4)(1,6)-glucooligosaccharides and pullulan. The ability to bind the latter ligands suggested an ability to tolerate the α-1,6-linkages found in pullulan. Here, we explore the structural basis of this by determining the high-resolution X-ray crystal structures of *Tm*CBM41 in native, maltotetraose (M4) bound, and $6^3$-α-D-glucosyl-maltotriose (GM3) bound forms.

### 3.3.3 Materials and Methods

*Tm*CBM41 was produced and purified as described (Section 3.2.3). TmCBM41 was treated for four days with recombinant enterokinase at room temperature and the cleaved His6 tag was removed by passing the cleavage reaction through an IMAC Ni2+-Sepharose mini column (Novagen, Madison WI) and collecting the flow-through containing cleaved protein product. TmCBM41 was buffer-exchanged in to 5 mM Tris–HCI (pH 7.5) in a 10 ml stirred ultrafiltration device using a 5 kDa cutoff membrane. Crystallization experiments were carried out using the vapor-diffusion, hanging-drop method at 18 oC. Crystals of TmCBM41 (20 mg/ml) were grown overnight in 1.6 M ammonium sulfate, 0.1 M Mes (pH 6.5) and 5% (v/v) isopropanol. An iodide derivative was obtained by growing crystals of TmCBM41 in mother liquor supplemented with 0.1 M NaI. The native and iodide derivative crystals were in the space group H32 with approximately 65% (v/v) solvent. TmCBM41 was co-crystallized with an excess of maltotetraose and GM3 in 30% (w/v) polyethylene glycol 1500 as the mother liquor. These crystals were in the space group C2 with roughly 36% (v/v) solvent. Crystals were frozen at 113 K directly in the cryo-stream after a short soak in mother liquor supplemented with 15–20% (v/v) ethylene glycol. Data were collected with a Rigaku R-Axis 4++ area detector coupled to an MM-002 X-ray generator with Osmic Blue optics and an Oxford Cryostream 700. Data were processed using Crystal Clear/d*trek [129]. Structure determination and refinement: All computing was carried out with the CCP4 suite unless stated otherwise [117]. Using only the anomalous signal, ShelxD was able to locate the one iodide ion incorporated in the asymmetric unit by co-crystallization with NaI [156]. Initial phasing to 3.0 Å was done by SIRAS using the program SHARP,[157] and

yielded isomorphous phasing powers of 0.5 for centric and acentric reflections, and an anomalous phasing power of 0.4. The figures of merit (FOM) for centric and acentric reflections were 0.19 and 0.24, respectively. At this stage the electron density maps were uninterpretable. The high solvent content of this crystal form was exploited through solvent flattening and the phases were extended to 1.69 Å with the program DM,[158] which ultimately yielded an FOM of 0.67 and easily interpretable maps (Figure 31), despite the apparently poor phasing power of this derivative. ARP/wARP [159] was able to build a nearly complete model that required minimal manual correction using COOT [160]. Refinement steps were carried out with REFMAC [115]. The unliganded model was used to solve the complexed structures by molecular replacement using MOLREP [113]. Carbohydrate ligands were added manually using COOT and the structures refined as described above. Water molecules were added using the REFMAC implementation of ARP/wARP and inspected visually before deposition. In all data sets, 5% of the observations were flagged as "free" and used to monitor refinement procedures [112]. All final model statistics are given in Table 12.

### 3.3.4: Results and Discussion

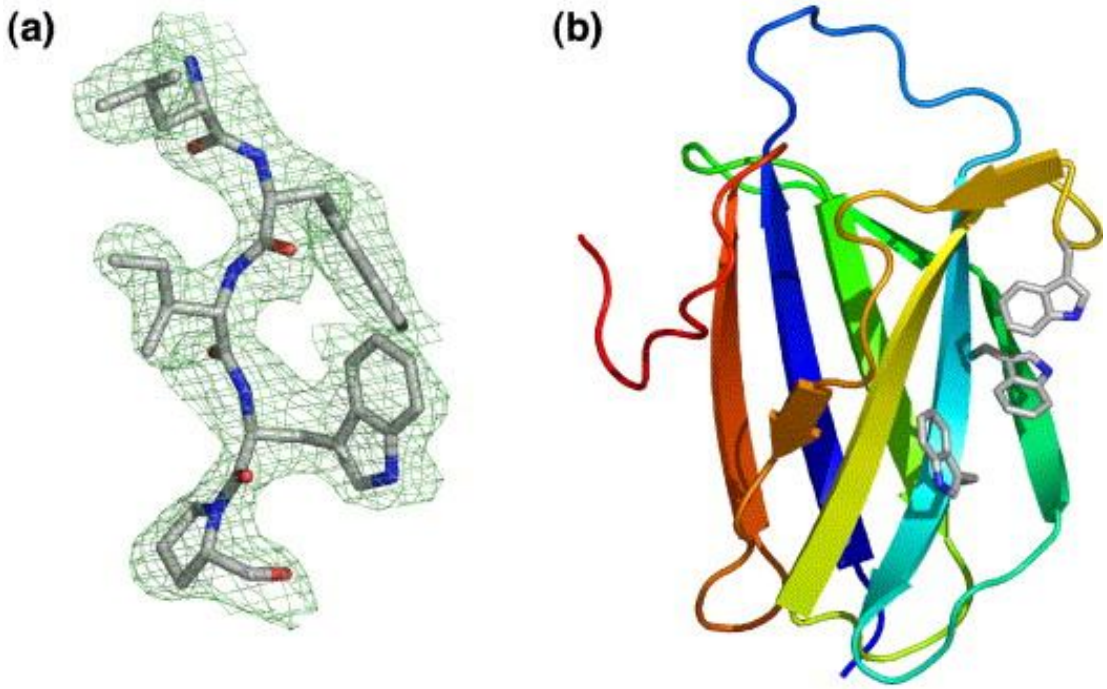
*Structure and α-glucooligosaccharide recognition: Tm*CBM41 exhibits a β-sandwich fold with four β-strands overlapping three β-strands in an immunoglobulin-like topology. Face 1 is comprised of β strands 1, 3, 4 and 7, while face 2 is comprised of β strands 2, 5 and 6 (Figure 31). W27, W29 and W73 are solvent-exposed and comprise the binding site of this molecule (Figure 31; see below).

Table 12: Data collection and model statistics for *Tm*CBM41

| | Native | Iodine Derrivative | Maltotetraose (M4) | Glucosyl-maltotriose (GM3) |
|---|---|---|---|---|
| **Data collection** | | | | |
| Space group | *H*32 | *H*32 | C2 | C2 |
| Cell dimensions | | | | |
| *a*, *b*, *c* (Å) | 112.5, 112.5, 83.7 | 113.9, 113.9, 81.8 | 106.4, 37.2, 57.2 | 106.5, 37.0, 54.9 |
| α, β, γ (°) | 90, 90 , 90 | 90, 90, 90 | 90, 112.4, 90 | 90, 112.0, 90 |
| Resolution (Å) | 19.81–1.69 (1.75–1.69)* | 19.81–2.95 (3.06–2.95) | 19.81–1.49 (1.54–1.49) | 19.80–1.40 (1.45–1.40) |
| $R_{merge}$ | 0.036 (0.384) | 0.086 (0.385) | 0.060 (0.402) | 0.044 (0.257) |
| $I / \sigma I$ | 14.6 (2.8) | 10.7 (2.5) | 11.9 (2.8) | 17.8 (2.8) |
| Completeness (%) | 99.4 (100.0) | 100.0 (100.0) | 97.5 (94.5) | 91.2 (68.4) |
| Redundancy | 6.84 (5.52) | 8.64 (8.76) | 3.97 (3.37) | 3.65 (2.74) |
| | | | | |
| **Refinement** | | | | |
| $R_{work}$ / $R_{free}$ | 0.203/0.240 | | 0.163/0.230 | 0.155/0.207 |
| No. residues | | | | |
| Protein | 101 | | 203 | 203 |
| Ligand/ion atoms | N/A | | 90 | 79 |
| Water | 135 | | 375 | 338 |
| *B*-factors | | | | |
| Protein | 39.8 | | 17.6/20.1 | 17.2/16.9 |
| Ligand/ion | N/A | | 26.8 | 23.0 |
| Water | 53.5 | | 33.7 | 30.3 |
| R.m.s deviations | | | | |
| Bond lengths (Å) | 0.021 | | 0.020 | 0.021 |
| Bond angles (°) | 1.883 | | 1.938 | 2.030 |

*Highest resolution shell is shown in parenthesis.

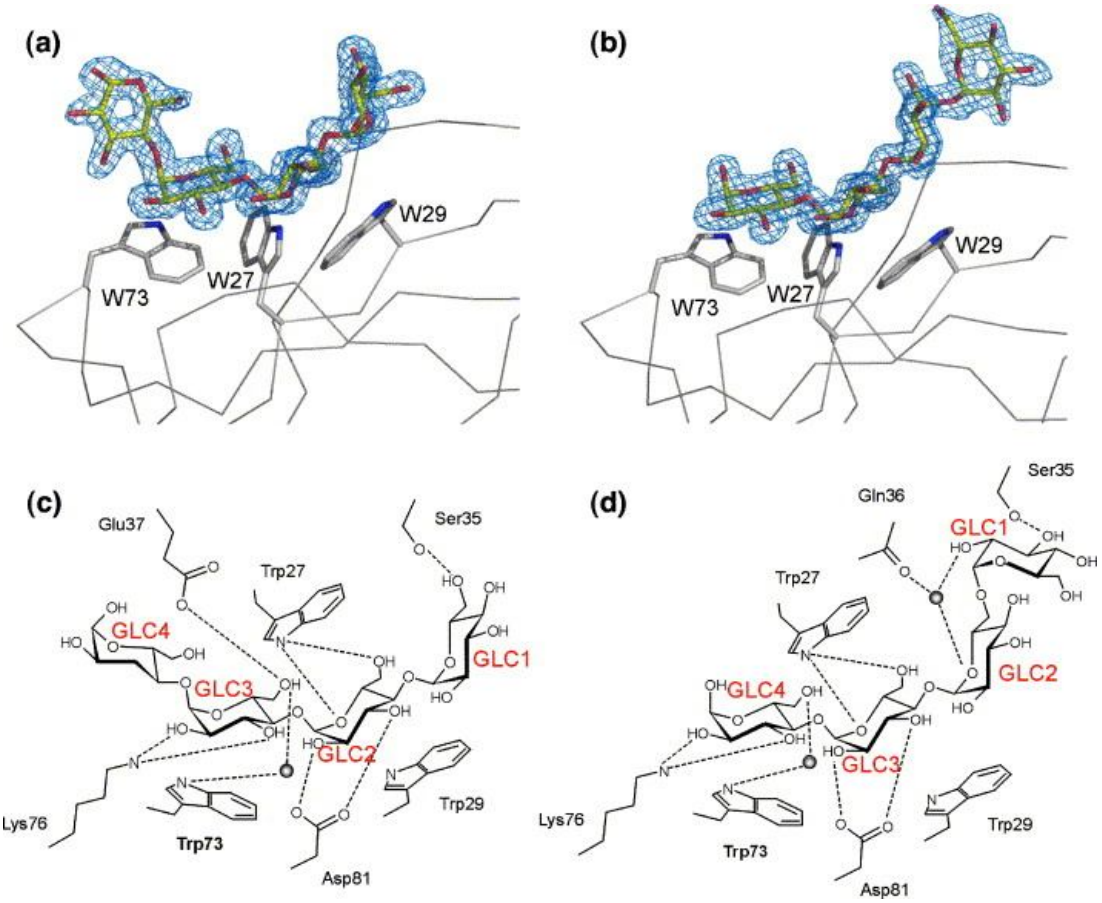
Figure 31: Structure of *Tm*CBM41. (a) The experimental electron density map (contoured at 1σ = 0.24 $e^{-}/Å^{3}$) produced by iodide SIRAS phasing contoured around the refined model of *Tm*CBM41. (b) The overall fold of *Tm*CBM41 showing tryptophan side-chains involved in ligand binding in stick representation.



Crystals of *Tm*CBM41 in complex with M4 and GM3 yielded excellent data and clear electron density for ligand bound to each of the two *Tm*CBM41 molecules in the asymmetric unit, allowing unambiguous modeling of all of the residues in these carbohydrates (Figure 32). The central glucose residues of M4, GLC2 and GLC3 make the majority of the interactions with the protein, most noticeably through what is often referred to as stacking interactions between the α-faces of adjacent pyranose rings and the indole rings of W29 and W73. GLC1 makes one potential direct hydrogen bond with TmCBM41, GLC2 makes four, and GLC3 makes three (Figure 32) for a total complement of eight direct hydrogen bonds. W27 contributes a single water-mediated hydrogen bond to the O6 of GLC3. GLC4 makes no apparent interaction and is likely visible in the structure only due to interactions with symmetry-related molecules.

The structure of *Tm*CBM41 in complex with GM3 revealed how this might tolerate the α-1,6-linkages found in pullulan. The directionality of GM3 in the binding site is maintained with M4. However, last three glucose residues of GM3 occupied the same subsites as the first three glucose residues of M4 (Figure 32). The non-reducing end α-1,6-linked glucose of GM3 occupied a new subsite. The protein–carbohydrate interactions at the subsites that are conserved with M4 were expected to be identical. Surprisingly, the side-chain of E37, which was well ordered in the M4 complex, was disordered in the GM3 complex and could not be modeled. The potential hydrogen bond contributed by this side-chain is thus lost in the GM3 complex, though we cannot explain why this is the case. GLC2 of GM3 is slightly displaced relative to the structurally analogous GLC1 of M4 due to the necessity of accommodating the terminal α-1,6-linked GLC1 of GM3. As a result, the side-chain of S35 no longer hydrogen bonds with the O6

Figure 32: *Tm*CBM41 in complex with (a) M4 and (b) GM3 showing the binding site tryptophan residues in stick representation. Maximum-likelihood[115] /$\sigma_A$[119] weighted $2F_o - F_c$ electron density maps contoured at 1σ (0.38 $e^-/Å^3$ and 0.43 $e^-/Å^3$ for M4 and GM3, respectively) are shown surrounding each ligand. The protein backbone is shown as $C^\alpha$ traces. Corresponding binding site architecture with hydrogen bonding patterns are shown for (c) M4 and (d) GM3.



of GLC2 in GM3 (i.e. GLC1 of M4). However, in the GM3 complex, the side-chain of S35 swings to make a putative hydrogen bond with O3 of the terminal α-1,6-linked GLC1 of GM3. Thus, relative to the M4 complex, two direct hydrogen bonds are lost in the GM3 interaction but one new hydrogen bond and two water-mediated hydrogen bonds are added, with the other interactions being conserved (Figure 32).

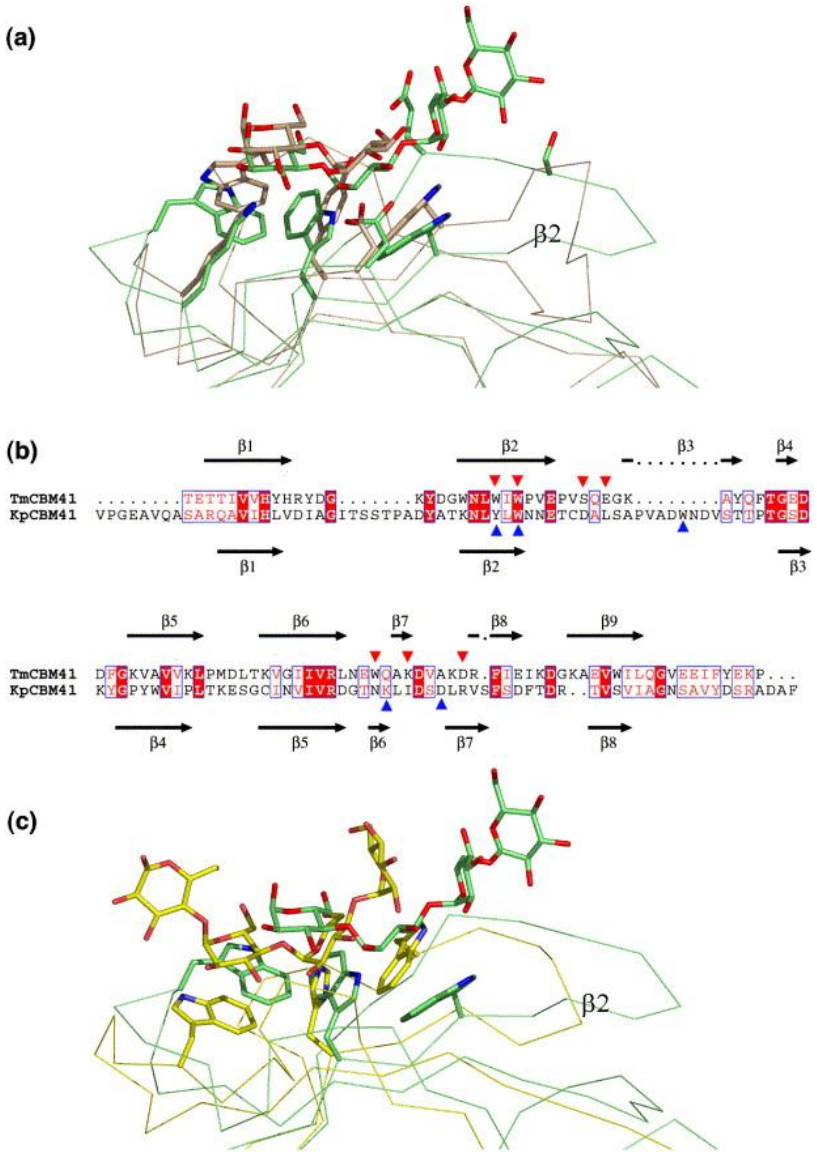
In total, both M4 and GM3 make eight and seven direct hydrogen bond contacts, respectively, with *Tm*CBM41 at three binding subsites. With M4, the calculated polar surface area of the protein buried upon ligand binding is 282.7 $Å^2$; similarly, 293.3 $Å^2$ of polar surface area is buried by GM3. Therefore, the hydrogen bond densities exhibited by these two ligands are very similar, with 2.8 and 2.7 hydrogen bonds per 100 $Å^2$ for M4 and GM3, respectively. By definition, type B CBMs have approximately 2.1 hydrogen bonds per 100 $Å^2$, while lectins have 3.4 hydrogen bonds per 100 $Å^2$ [56]. The classification of *Tm*CBM41 as either a type B or type C CBM is uncertain, as its ability to bind longer oligosaccharides with increasing affinities and four binding subsites would classify this CBM as type B; however, the hydrogen bond density observed in binding the ligand correlates with a lectin-like CBM (type C). Thus, as noted for other α-glucan-binding CBMs,[7] *Tm*CBM41 has properties that are intermediate between type B and C CBMs.

*Implications in polysaccharide recognition: Tm*CBM41 clearly accommodates an α-1,6-linked glucose on the non-reducing end of the maltotriose recognition sequence, and even makes specific contacts with this residue. The additional α-1,4-linked glucose units at the reducing end in the polysaccharide would extend into the solvent. Likewise, the α-1,6-linked glucose at the reducing end of the maltotriose would introduce a kink in the

polysaccharide that would extend the polysaccharide into the solvent. Thus, the recognition determinant in pullulan would appear to be the GM3 unit. Curiously, the oligosaccharide GM3 is bound tenfold less tightly than M3, while GM3M3 and pullulan, both multimers of GM3, are bound with affinities similar to that of M3.[9] On the basis of this observation, it would appear that a terminal α-1,6-linkage is a strong detriment to binding but internal α-1,6-linkages are better tolerated. Precisely why this is so is unclear.

*Comparison with other starch-binding CBMs:* The X-ray crystal structure of the CBM41, *Kp*CBM41, from the Klebsiella pneumoniae pullulanase was solved recently in the context of the entire enzyme [161]. This CBM shares~ 22% sequence identity with *Tm*CBM41, and has an RMSD of~ 1.9 Å, indicating that these CBMs are quite distantly related CBM family members. The architectures of the α-glucan-binding sites are structurally well conserved (Figure 33(a)). Two of the binding site aromatic residues are conserved at both the sequence and structure levels, while the third is only structurally conserved (Figure 33(a) and (b)). Many of the predicted hydrogen bonding residues are also conserved (Figure 33(a) and (b)). *Tm*CBM41 has 6–10% sequence identity and RMSDs of 2.0–2.8 Å with CBMs from families 20, 25, 26, and 34, revealing the structural similarity with these α-glucan-binding proteins. We reported previously that the starch-binding CBMs from these families showed a conserved mode of α-glucan recognition [63]. The α-glucan-binding sites of *Tm*CBM41 and *Kp*CBM41 share the same general location and architecture as the starch-binding members of families 20, 25, and 26 (Figure 33(c)). However, relative to the CBMs in these families, the binding sites of *Tm*CBM41 and *Kp*CBM41 are shifted towards the loop that separates β-strands 2 and 3 (Figure 33(c)). In *Tm*CBM41, this allows interactions between the α-1,6-linked glucose

Figure 33: A comparison of *Tm*CBM41 with other α-glucan-binding modules. (a) Overlap of the binding sites of *Tm*CBM41 (green) with *Kp*CBM41 (tan; PDB code 2FHF)[161] showing binding site aromatic side-chains, hydrogen bonding residues, and ligands in stick representation. (b) Amino acid sequence alignment of *Tm*CBM41 with *Kp*CBM41 showing the secondary structures above and below, respectively. Residues involved in binding are indicated above and below the sequences with red and blue triangles for *Tm*CBM41 and *Kp*CBM41, respectively. (c) Overlap of the binding sites of *Tm*CBM41 (green) with the *Bacillus halodurans* CBM25 (yellow; PDB code 2C3W)[63] showing binding site aromatic side-chains, hydrogen bonding residues and bound ligands in stick representation. In (a) and (c), the protein backbone is shown as $C^{\alpha}$ traces and the β2 strand is labelled for reference; (b) was prepared with ESPRIPT [162].



and residues in this loop (Figure 33(a) and (c)). This may be an adaptation of CBMs found in pullulanases. Indeed, CBMs from families 20, 25, and 26 are found mainly in predicted α-1,4-glucan-specific enzymes, while family 41 CBMs are found mainly in predicted pullulanases.

## 3.4 Identification and structural basis of binding to host lung glycogen by streptococcal virulence factors

Alicia Lammerts van Bueren, Melanie Higgins, Diana Wang, Robert D. Burke and Alisdair B. Boraston

Adapted from *Nature Structural & Molecular Biology* **14**, 76 - 84 Published January 2007

Contributions to Research: Cloning, protein production, mutagenesis, crystallization, data collection, structure solving and structure refinement, protein labelling, manuscript and figure preparation

### 3.4.1 Abstract

The ability of pathogenic bacteria to recognize host glycans is often essential to their virulence. Here we report structure-function studies of previously uncharacterized glycogen-binding modules in the surface-anchored pullulanases from *Streptococcus pneumoniae* (SpuA) and *Streptococcus pyogenes* (PulA). Multivalent binding to glycogen leads to a strong interaction with alveolar type II cells in mouse lung tissue. X-ray crystal structures of the binding modules reveal a novel fusion of tandem modules into single, bivalent functional domains. In addition to indicating a structural basis for multivalent attachment, the structure of the SpuA modules in complex with carbohydrate provides insight into the molecular basis for glycogen specificity. This report provides the first evidence that intracellular lung glycogen may be a novel target of pathogenic streptococci and thus provides a rationale for the identification of the streptococcal α-glucan–metabolizing machinery as virulence factors.

### 3.4.2 Introduction

Glycoside hydrolases are increasingly being found as virulence factors in bacterial pathogens. For example, sialidase, a glycoside hydrolase that cleaves the bonds between sialic acid residues, has long been known to act as a virulence factor in *S. pneumoniae* through its ability to unmask receptors and a more recently implicated role in biofilming [163; 164]. Recent signature-tagged mutagenesis studies performed with *S. pneumoniae* have indicated that a large number of known or putative glycoside hydrolases are necessary for full virulence in a mouse-lung model of *S. pneumoniae* infection[53]. An underappreciated feature of these and many other glycoside hydrolases is their modularity. Many glycoside hydrolases have, in addition to a domain harboring the catalytic activity, one or more accessory modules. For example, the *S. pneumoniae* sialidase has a ~200-residue N-terminal region with strong amino acid sequence identity to a family of sialic acid–binding modules classified as family 40 carbohydrate-binding modules (CBMs). Indeed, the type of accessory module most commonly found in glycoside hydrolases is the CBM. CBMs are currently classified into 52 families on the basis of amino acid sequence (see www.cazy.org). The primary function of these accessory modules seems to be targeting an enzyme to a particular substrate [56]. Thus, for example, sialidases may have sialic acid–specific CBMs. Family 41 in the CBM classification is a relatively recently classified family of CBMs whose only functionally characterized member is an α-glucan–specific CBM from an enzyme of the marine bacterium *Thermotoga maritima* [85]. This particular family is remarkable for its relatively high proportion of entries originating from bacterial pathogens, notably the streptococci *S. pneumoniae* and *S. pyogenes*, which are the bacteria relevant to this work.

A number of streptococci are found as human commensals that colonize a variety of environments in the body. Too often, these bacteria slip out of a passive role and into one of accomplished pathogenicity. One such streptococcus, *S. pneumoniae*, kills more people in the USA each year than all other vaccine-preventable diseases combined. Infection can result in acute respiratory disease (pneumonia), meningitis, septicaemia and otitis media[165]. A second notable streptococcus, *S. pyogenes* (group A streptococcus or GAS), is the causative agent of diseases such as pharyngitis (strep throat), scarlet fever (rash), impetigo (infection of the superficial layers of the skin) and cellulitis (infection of the deep layers of the skin). More invasive infections can result in necrotizing fasciitis, myositis and streptococcal toxic shock syndrome [166]. Frequently associated with 'flesh-eating disease', *S. pyogenes* is also an infrequent but particularly serious cause of community-acquired pneumoniae [167; 168]. Both *S. pneumoniae* and *S. pyogenes* are ever-present menaces to human health that are made that much more dangerous by the increasing prevalence of antibiotic resistance [165; 166].

As human commensals with no known environmental niches, *S. pneumoniae* and *S. pyogenes* must contribute a substantial portion of their genome toward maintaining the host-bacterium relationship. Therefore, it is somewhat surprising that both organisms host a battery of genes that constitute the machinery required for the metabolism of highly polymerized α-glucans, such as starch and glycogen, which are most likely to be found in environmental niches. Indeed, *S. pneumoniae* signature-tagged mutagenesis studies have implicated a number of α-glucan–active enzymes as virulence factors [53]. Likewise, the machinery involved in the transport of α-glucooligosaccharides has more recently been shown to be a virulence factor in *S. pyogenes* [169]. Key extracellular components in the α-

glucan–metabolizing machinery of both of these bacteria are the closely related *S. pneumoniae* SpuA and *S. pyogenes* PulA proteins. Both enzymes are anchored to the cell wall at their C termini and both can hydrolyze the α-glucan pullulan (comprising α-(1,4)-glucotriose joined by α-(1,6) linkages) and are thus classified as pullulanases [152; 170]. SpuA has been shown, through signature-tagged mutagenesis studies, to be necessary for full virulence in a mouse-lung model of *S. pneumoniae* infection [53], and PulA has been suggested to have glycoprotein-binding activity and may contribute to virulence through this strepadhesin activity [152; 171]. SpuA and PulA share nearly identical multimodular architectures comprising tandem putative family 41 CBMs separated from a family 13 glycoside hydrolase catalytic module by a region of unknown function (Fig. 34).

The importance of these α-glucan–specific enzymes to the virulence of streptococci currently lacks an adequate explanation. We approached this problem by dissecting the modular structures of SpuA and PulA and focusing structure-function studies on the putative N-terminal CBMs. The results reveal that these modules are α-glucan–binding modules that bind through a tight multivalent interaction with starch and glycogen. The X-ray crystal structures of the modules show the structural basis of their binding specificity, and fluorescence-microscopy experiments demonstrate that they have notable specificity for surfactant-producing alveolar type II pneumocytes. Thus, the N-terminal modules of SpuA and PulA represent a previously uncharacterized kind of bacterial carbohydrate-binding protein that targets glycogen in the lung and possibly in other tissues. Furthermore, on the basis of the carbohydrate-specific binding and catalytic function of the streptococcal pullulanases, we hypothesize a role for these proteins, as

well as other α-glucan–specific enzymes from *S. pneumoniae* and *S. pyogenes*, in respiratory infections.

### 3.4.3: Materials and Methods

*Carbohydrates:* All carbohydrates and glycoproteins were purchased from Sigma.

*Cloning:* The gene fragment encoding the tandem CBM41 modules and the X-module of SpuA was amplified by PCR using *S. pneumoniae* TIGR4 genomic DNA (American Type Culture Collection BAA-334D) as a template. Similarly, the gene fragment encoding the tandem CBM41 modules and the X-module of PulA was amplified from *S. pyogenes* genomic DNA (American Type Culture Collection 12344D). The amplified DNA fragments were cloned into pET 28a via engineered NheI and XhoI restriction sites to yield the plasmids pSpnDX and pSpyDX. Amino acid substitutions were introduced into SpyDX by site-directed mutagenesis using a 'megaprimer' PCR method [172] with wild-type pSpyDX plasmid DNA as a template. The DNA sequence fidelity of all constructs was verified by bidirectional sequencing. These constructs encode the respective wild-type and mutant CBM tandem modules separated from N-terminal $His_6$ tags by thrombin protease cleavage sites.

*Protein production and purification:* Appropriate plasmids were transformed into chemically competent *E. coli* BL21 STAR (DE3) cells (Novagen). Cultures were grown up in LB medium supplemented with 50 μg $ml^{-1}$ kanamycin at 37 °C to an absorbance at 600 nm (A600) of 0.6. IPTG was then added to a concentration of 1 mM and growth allowed to continue at 37 °C for an additional 4 h. Cells were harvested by centrifugation and resuspended in 40 ml of 20 mM Tris (pH 8.0) and 0.5 M NaCl (buffer A)

supplemented with protease inhibitor cocktail (Roche) and ruptured using a French pressure cell. Cell debris was removed by centrifugation and the target polypeptides present in the clarified supernatants purified by $Ni^{2+}$ immobilized metal affinity chromatography as described previously [68]. Purity of the fractions was assessed using SDS-PAGE, and those fractions with pure protein were pooled and concentrated in a stirred ultrafiltration device with a 5,000-Da MWCO membrane under pressurized nitrogen. The protein was then dialyzed into 20 mM Tris (pH 8.0) and 0.5 M NaCl. Protein prepared for crystallization was further purified by size-exclusion chromatography using a Sephacryl S-200 column equilibrated in 20 mM Tris (pH 8). Selenomethionine (SeMet)-labeled SpyDX was produced using procedures described [173] and purified as above with 1 mM DTT included in all buffers.

Protein concentration was determined from $A_{280}$ using calculated molar extinction coefficients for wild-type (SpyDX = 0.05718 $cm^{-1}$ $\mu M^{-1}$, SpnDX = 0.06287 $cm^{-1}$ $\mu M^{-1}$) and mutant proteins (SpyDX$\Delta$1 = 0.04882 $cm^{-1}$ $\mu M^{-1}$, SpyDX$\Delta$2 = 0.05432 $cm^{-1}$ $\mu M^{-1}$, SpyDX$\Delta$12 = 0.04332 $cm^{-1}$ $\mu M^{-1}$) [109].

*Affinity electrophoresis and carbohydrate macroarray:* Affinity electrophoresis was performed as described [146] using 10% (w/v) polyacrylamide gels polymerized with and without the inclusion of 0.5% (w/v) polysaccharide (amylose, amylopectin, pullulan, dextran and type IV glycogen from bovine liver). A 20-μg sample of each protein was loaded on to a gel and samples were run at 100 V for 150 min in an Invitrogen Mini Cell Sure lock system. Gels were stained with Coomassie blue R250 in 25% (v/v) methanol and 10% (v/v) acetic acid and destained with 25% (v/v) methanol and 10% (v/v) acetic

acid. Binding to polysaccharide was visualized as reduced mobility of the protein on these gels relative to the nondenaturing gel lacking polysaccharide.

Carbohydrate-glycoprotein macroarrays were prepared by spotting 1 μl of 1% (w/v) carbohydrate or glycoprotein solutions onto a nitrocellulose membrane. The membranes were then blocked for 1 h at 10 °C in 20 mM Tris (pH 8.0) containing 0.5 M NaCl and 1% (w/v) BSA (blocking buffer). To individual membranes, 200 μg of each protein, containing an N-terminal $His_6$ tag, was added in a 5-ml solution of blocking buffer. The membranes were incubated with shaking for 1 h at 10 °C. The membranes were then washed with 3 x5 ml of blocking buffer and put into a fresh 5 ml of blocking buffer, to which 10 μl of 0.5 mg $ml^{-1}$ AlexaFluor 680–labeled streptavidin, previously complexed with biotin-NTA $Ni^{2+}$, was added (see below). The membranes were incubated for 30 min at 10 °C to ensure complete binding of the streptavidin/biotin–nitrilotriacetic acid (NTA) label with the His tags of each protein. Membranes were then washed three times with 5 ml of 20 mM Tris (pH 8.0) and 0.5 M NaCl with 0.5% (v/v) Tween-20 to remove any unbound label. Binding was visualized on a LICOR Odyssey system (LICOR Biosciences) at 700 nm.

The biotin-NTA–streptavidin complex was made by adding 1 mg of AlexaFluor 680–labeled streptavidin (Molecular Probes–Invitrogen) and 1.5 mg of biotin-NTA (Molecular Probes) to 1 ml of 5 mM Tris (pH 8.0) plus 50 μl of a 10 mg $ml^{-1}$ solution of NiSO4 (500 μg). The solution was then incubated in the dark for 10 min at room temperature. Free nickel and biotin-NTA were separated from the mixture by size-exclusion

chromatography using Sephadex G-25 (GE Healthcare) pre-equilibrated with 5 mM Tris (pH 8.0). The collected fractions were pooled and stored at -20 °C in a dark box.

*Quantitative binding studies:* Adsorption isotherms using granular corn starch were performed and analyzed using methods identical to those described [63]. ITC was performed as described [68] using a VP-ITC (MicroCal) in 20 mM Tris (pH 8.0) and 0.5 M NaCl at 25 °C. Carbohydrate solutions were prepared by dissolving a known mass of lyophilized carbohydrate in the buffer saved from concentrating down protein in a stirred ultrafiltration device with a 5,000 MWCO membrane. A solution of 2 mM SpnDX or SpyDX was titrated into type IV glycogen from bovine liver (0.84 mg $ml^{-1}$). Using a well-established method [85; 144], the concentration of receptor (glycogen) was left as a floating parameter during the fitting to a one-site binding model using the concentration of glycogen converted to molar equivalents of glucose. Thus, the final stoichiometry is expressed as number of glucose units per binding site. For oligosaccharides, 13.5 mM maltotetraose was titrated into ~400 μM SpyDX or SpnDX. The concentrations of protein for all titrations were chosen so that they were in five-fold or greater excess of the dissociation constants. All data described are the average and standard deviation of three independent titrations, except for SpyDX with maltotetraose, where the values and errors are reported from a single titration.

*Crystallization of SpnDX and SpyDX: SpyDX* (50 mg) was treated overnight with restriction-grade thrombin (Novagen) to remove the His tag. SeMet-labeled SpyDX required 3 days of thrombin treatment because of the presence of DTT. The cleavage reactions were cleaned up by size-exclusion chromatography using a Sephacryl-S200 column. Crystals of both native and SeMet-containing SpyDX were obtained at 20 mg

$ml^{-1}$ in 11% (w/v) PEG 20,000, 0.1 M Tris (pH 7.0), 7.5% (v/v) ethylene glycol and 0.2 M $NaH_2PO_4$ by the vapor-diffusion hanging drop method. SpnDX was treated in the same manner as SpyDX to remove the His tag and then incubated at room temperature for 2 weeks to promote the formation of a stable 28-kDa degradation product. This stable product was isolated by size-exclusion chromatography as above. Fractions containing protein with a single band at 28 kDa were pooled and concentrated to 17 mg $ml^{-1}$. This 28-kDa fragment was cocrystallized with an excess of maltotetraose at 17 mg $ml^{-1}$ in a solution of 18% (w/v) PEG 4,000, 0.2 M zinc acetate and 0.1 M sodium cacodylate (pH 6.4). Crystals were flash-frozen in liquid nitrogen in mother liquor supplemented with 20%–25% (v/v) ethylene glycol.

*Data collection, structure determination and refinement*: Diffraction data for SeMet-labeled SpyDX were collected at the National Synchrotron Light Source on beamline X8-C. Diffraction data for the SpnDX–maltotetraose complex were collected with a Rigaku R-AXIS 4++ area detector coupled to a MM-002 X-ray generator with Osmic 'blue' optics and an Oxford Cryostream 700. All data were processed using Crystal Clear/d*trek [129]. Data collection and processing statistics are given in Table 13.

The structure of SpyDX was solved by the SAD method using crystals of SeMet-labeled protein and data collected at the selenium edge (0.9797 Å, f' = -7.83, f'' = 5.56), which was determined by a fluorescence scan. Six selenium sites (ten were expected, five per monomer in the asymmetric unit) were found by ShelxD [156], using data extending to 3.0 Å that was prepared for input into this program by ShelxC. Refinement of heavy atom parameters and initial phasing with SHARP [157] yielded an overall anomalous phasing power of 1.2 and figures of merit of 0.4 and 0.2 for acentric and centric reflections,

Table 13: Data collection and refinement statistics for SpyDX and SpnDX

| | SpyDX SeMet | SpnDX maltotetraose |
|---|---|---|
| **Data collection** | | |
| Space group | $P2_1$ | $P2_12_12$ |
| Cell dimensions | | |
| *a*, *b*, *c* (Å) | 37.62, 71.49, 101.04 | 50.56, 89.12, 60.00 |
| α, β, γ (°) | 90.0, 90.02, 90.0 | 90.00, 90.00, 90.00 |
| Resolution (Å) | 37.62-1.60 (1.66-1.60) * | 19.51-2.10 (2.17-2.10) |
| $R_{merge}$ | 0.087 (0.478) | 0.072 (0.367) |
| $I / \sigma I$ | 9.2 (3.1) | 10.3 (3.1) |
| Completeness (%) | 96.8 (95.1) | 96.9 (99.1) |
| Redundancy | 5.09 (5.14) | 4.40 (4.31) |
| **Refinement** | | |
| Resolution (Å) | 1.60 | 2.10 |
| No. reflections | 348383 (68393 unique) | 69812 (15873 unique) |
| $R_{work}$ / $R_{free}$ | 0.179/0.216 | 0.221/0.304 |
| No. atoms | | |
| Protein | 1819 (A) 1779 (B) | 1801 |
| Ligand/ion | N/A | 79/3 |
| Water | 497 | 189 |
| *B*-factors | | |
| Protein | 20.830 (A) 18.917 (B) | 41.877 |
| Ligand/ion | N/A | 57.548/80.613 |
| Water | 29.450 | 56.646 |
| R.m.s deviations | | |
| Bond lengths (Å) | 0.021 | 0.011 |
| Bond angles (°) | 1.707 | 1.554 |

*Highest resolution shell is shown in parenthesis.

respectively, over the full 1.6-Å resolution range. Density improvement by solvent flattening with DM [158] yielded a final figure of merit of ~0.85 and electron density maps of sufficient quality to allow nearly complete automatic building of the model with ARP/wARP [159]. The initial model was corrected and completed manually by successive rounds of building using COOT [160] and refinement with REFMAC [115]. The coordinates of SpyDX were used to solve the structure of SpnDX by molecular replacement using MOLREP [113] to find the single molecule of SpnDX in the asymmetric unit. The initial model of SpnDX was manually corrected and refined as above. Water molecules were added using the REFMAC implementation of ARP/wARP and inspected visually before deposition. In both data sets, 5% of the observations were flagged as 'free' [112] and used to monitor refinement procedures. All final model statistics are given in Table 13. (SpyDX Ramachandran plot statistics: 86.3% residues in most favored region, 12.1% residues in additional allowed regions, 1.1% residues in generously allowed regions, 0.5% in disallowed regions. SpnDX: 77.5% residues in most favored region, 16.8% residues in additional allowed regions, 2.1% residues in generously allowed regions, 3.7% in disallowed regions.) Structure images were prepared with PyMOL (http://pymol.sourceforge.net/).

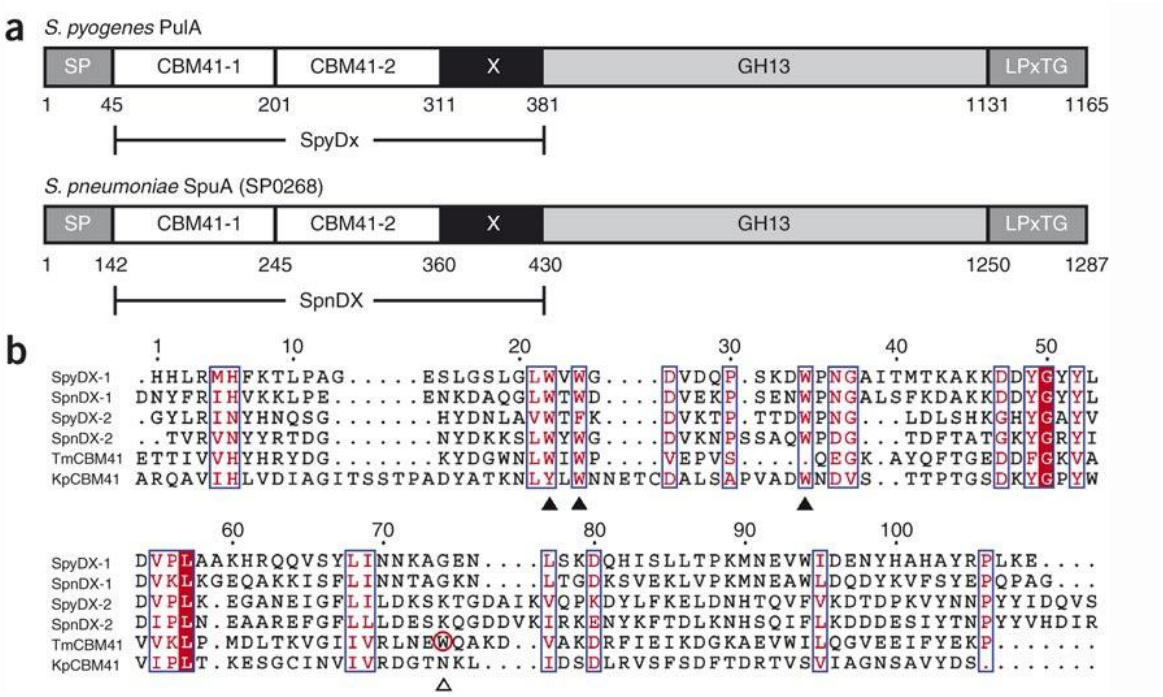
*Indirect immunofluorescence:* Proteins were labeled with FITC as per the manufacturer's protocol (Invitrogen). Free label was separated from the labeled protein by gel-filtration using Sephadex G-75. Freshly dissected, normal mouse lungs were infiltrated at room temperature with optimal–cutting temperature compound and frozen at -80.0 °C. Cryostat sections (10–15 μm) were collected on gelatin-coated glass slides. Some sections were post-fixed with 4% (v/v) paraformaldehyde in PBS (5 min) followed by 100% (v/v)

methanol (1 min), whereas other sections were not fixed. The sections were rinsed and blocked in 5% (v/v) lamb serum in PBS Tween-20 for 45 min at room temperature. Fluorescently labeled carbohydrate-affinity proteins were diluted to 200–250 pg μl-1 in 5% (v/v) lamb serum in PBS and applied directly to sections for overnight incubation (4 °C). For double labeling experiments, primary antibody (anti–proSurfactant protein C, Chemicon) was diluted in 5% (v/v) lamb serum in PBS (1:1,000) and applied to sections overnight. Sections were rinsed three times in PBS and the secondary antibody Alexa 568 conjugated goat anti-mouse IgG (Molecular Probes) diluted as 1:800 or 1:1,500 in 5% (v/v) lamb serum. PBS was applied to each section. After incubation (2 h), sections were rinsed with PBS. In some preparations, section was counter-stained with Hoechst 33342 (Molecular Probes). Sections were mounted with coverslips and examined with a Leica DM-6000 and images captured with a Hamamatsu Orca wide-field camera controlled with Openlab software (v. 4.04 from Improvison). Single optical sections were imaged using a Nikon C1 plus confocal laser scanning microscope. Contrast and brightness were adjusted and images cropped and assembled using Adobe Photoshop.

### 3.4.4: Results and Discussion

*α-glucan binding modules in streptococcal enzymes:* The large, surface-anchored pullulanases of *S. pneumoniae* and *S. pyogenes* have repeated ~100-residue sequences at their N termini (Fig. 34a). Each of the repeated modules bears ~44% amino acid sequence identity to the others and ~20% identity to *T. maritima* CBM41, a family 41 CBM that binds tightly to α-glucans comprising primarily α-(1,4) linkages [85]and is found in a secreted *T. maritima* pullulanase [142] (Fig. 34b). The *S. pneumoniae* and *S. pyogenes*

Figure 34: (A) Modular arrangement of streptococcal pullulanases PulA and SpuA. Each enzyme has an N-terminal leader signal peptide (SP) followed by two family 41 CBMs. An X-module with unknown function separates the tandem CBM41s (1 and 2) from the C-terminal catalytic module, a family 13 glycoside hydrolase (GH13). An LPxTG motif (where x is any amino acid) is present, indicating that the enzyme is anchored to the cell wall at the C terminus. Amplified SpyDX and SpnDX gene fragments for our study are indicated. (B) Amino acid sequence alignments of the individual CBM41s from SpyDX and SpnDX with CBM41 from T. maritima pullulanase PulA (*Tm*CBM41), and K. pneumoniae pullulanases (*Kp*CBM41). Red letters, conserved residues; red highlight, identical residues; black triangles, conserved binding site residues; open triangle and red circle, structurally conserved binding site tryptophan residue in T. maritima.



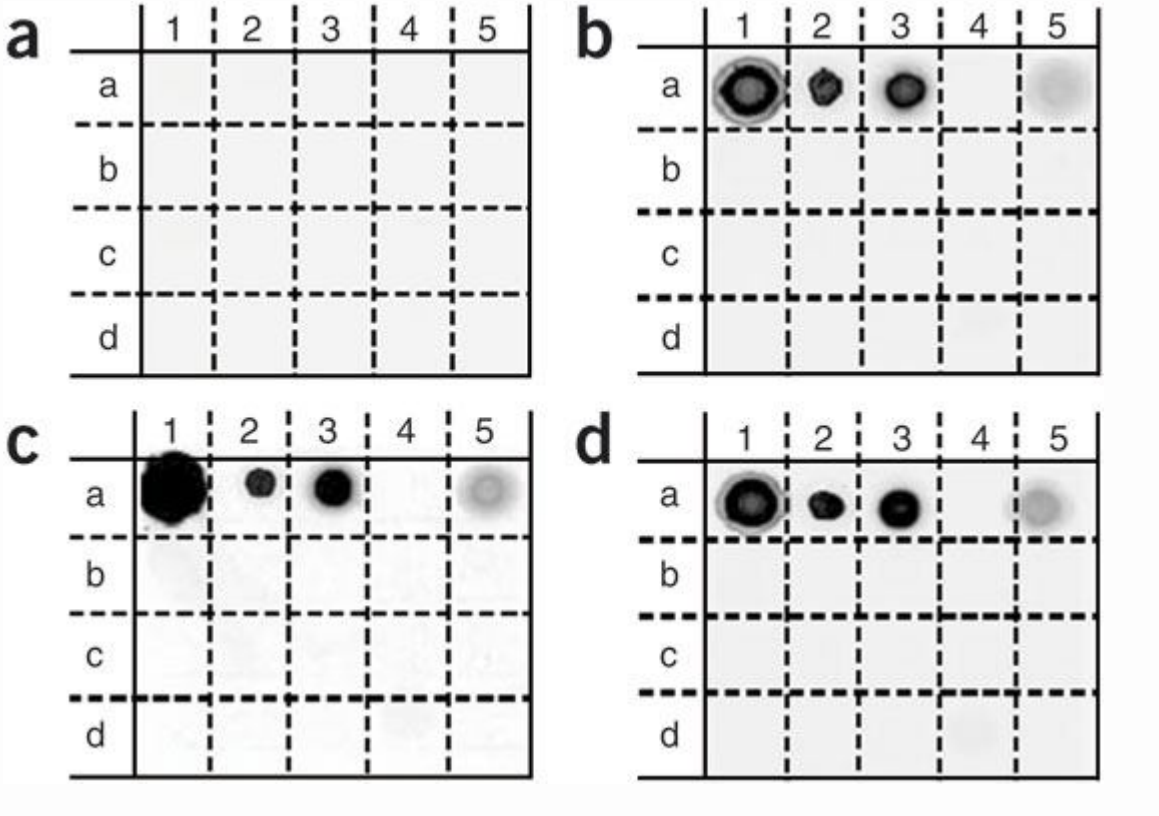
N-terminal modules also have ~17% amino acid sequence identity to an N-terminal module in the *Klebsiella pneumoniae* pullulanase (Fig. 34b). This module has recently been suggested to have α-glucan–binding activity [161]. The presence of the repeated modules in streptococcal pullulanases and their similarity to known carbohydrate-binding proteins suggested a carbohydrate-binding function for these modules. Indeed, the entire pullulanase of *S. pyogenes* has been implicated previously in carbohydrate-mediated strepadhesin activity [152; 171], lending greater support to the hypothesized function of the repeated streptococcal modules. However, a comparison of *T. maritima* CBM41 and *Klebsiella* CBM41 with the putative streptococcal CBMs provides very little insight into the function of the putative streptococcal CBMs because of their extremely low amino acid sequence identities. To investigate the functions of the putative family 41 CBMs in the streptococcal pullulanases, we dissected these proteins at the genetic level, producing recombinant soluble modules whose function could be studied in isolation.

We were unable to produce recombinant individual streptococcal CBM41 modules or tandem modules as soluble protein in *Escherichia coli*. However, when we included in the tandem constructs the domain of unknown function that separates the catalytic domain and putative CBMs (called the X-module; Figure 34a), excellent yields of soluble protein were obtained. Purified SpyDX, the recombinant tandem CBM41–X-module construct from the *S. pyogenes* pullulanase, was stable for several weeks when stored at room temperature or at 4 °C (data not shown). Purified SpnDX, the recombinant tandem CBM41–X-module construct from the *S. pneumoniae* pullulanase, was stable for several weeks when stored at 4 °C but slowly degraded over several weeks to a stable 28-kDa fragment when stored at room temperature (data not shown). We initially investigated the

ability of SpyDX and SpnDX to bind a variety of glycoproteins and polysaccharides by a macroarray binding analysis. This yielded identical results for SpyDX and SpnDX, indicating that they specifically recognize amylose (pure α-(1,4)-linked glucose), amylopectin (α-(1,4)-linked glucose with α-(1,6) branch points), pullulan (linear polymer of mixed α-(1,4)- and α-(1,6)-linked glucose) and glycogen (similar to amylopectin with more frequent α-(1,6) branch points) (Figure 35). Binding to these polysaccharides and lack of binding to dextran (a linear polymer of α-(1,6)-linked glucose) was confirmed by affinity electrophoresis (data not shown). Thus, the module pairs do seem to be α-glucan–specific carbohydrate-binding proteins, consistent with their similarity to known α-glucan–binding proteins. The proteins showed no cross-reactivity with glycoproteins that have previously been identified as ligands of the intact *S. pyogenes* pullulanase (thyroglobulin and mucin) [152].

*Tight binding via a multivalent interaction:* We initially quantified the binding parameters for SpyDX and SpnDX by a solid-state depletion isotherm method using the ligand cornstarch, which is a composite of amylose and amylopectin that is insoluble at room temperature. By this method, the affinities ($K_a$) of SpyDX and SpnDX for cornstarch were determined to be 2.2 ±0.4 x$10^5$ $M^{-1}$ and 1.1 ±0.5 x$10^6$ $M^{-1}$ (Figure 36a,b), and the binding capacities ($N_o$) were determined to be 5.9 ±0.3 and 3.0 ±0.3 μmoles per 10 g starch, respectively. Commercial preparations of glycogen, a highly polymerized branched α-glucan that is similar to amylopectin, are soluble, necessitating the use of isothermal titration calorimetry (ITC) to quantify binding to this ligand. Using ITC, the Ka, change in enthalpy (ΔH) and change in entropy (ΔS) upon the binding of

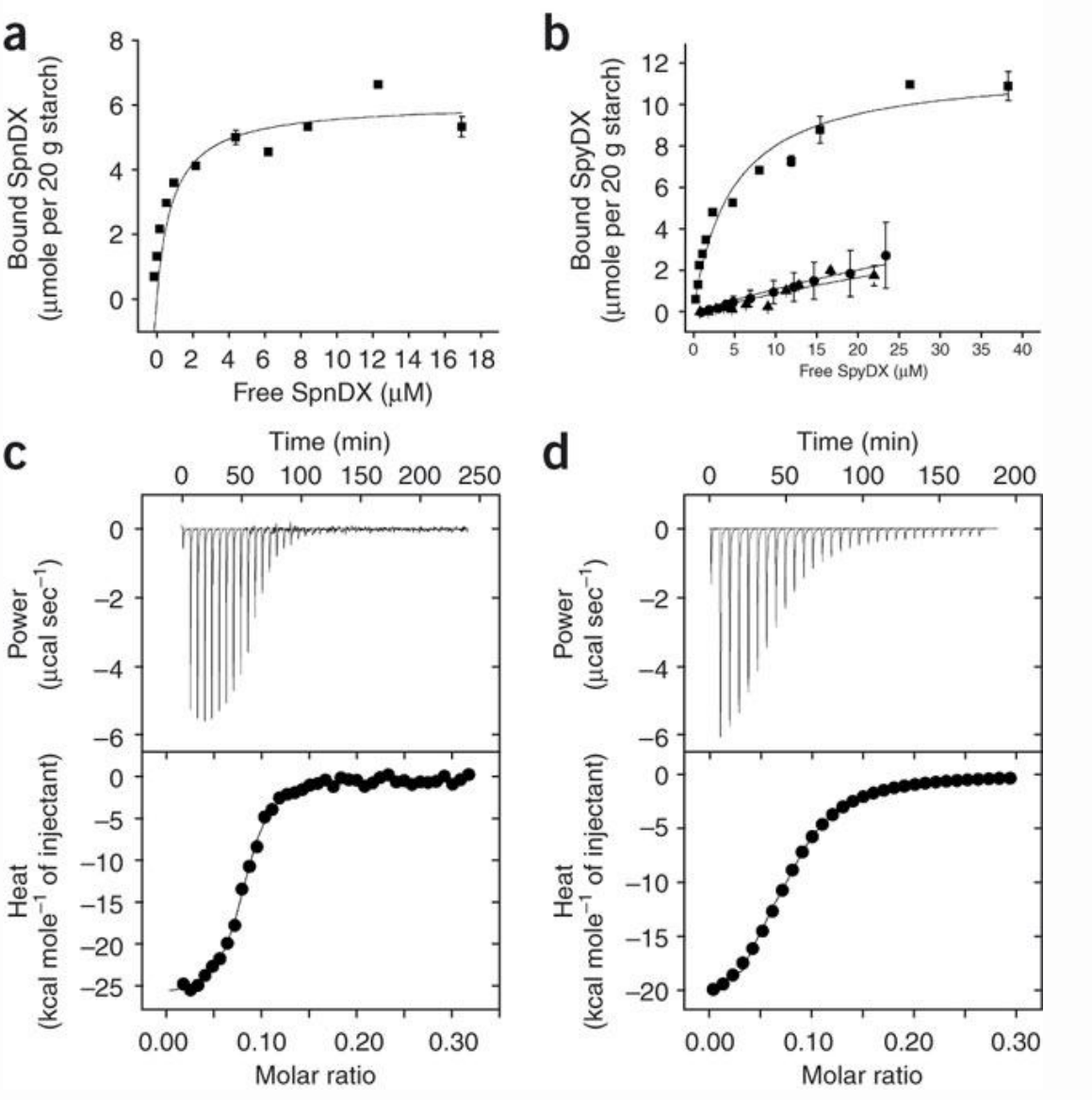
Figure 35: (a) *T. maritima* CBM27. (b) *T. maritima* CBM41. (c) SpyDX. (d) SpnDX. The following solutions were spotted on a nitrocellulose membrane: a1, amylose; a2, amylopectin; a3, pullulan; a4, dextran; a5, glycogen; b1, oat spelt xylan; b2, arabinogalactan; b3, pachyman; b4, glycerol; b5, thyroglobulin; c1, fetuin; c2, asialofetuin; c3, mucin type IS; c4, mucin type II; c5, mucin type III; d1, heparin; d2, chondroitin sulfate; d3, hyaluronic acid; d4, aggrecan; d5, proteoglycan. The ability of glycogen to interact with the nitrocellulose membrane is unknown, which may account for the weak signal produced in this experiment.



SpyDX to bovine liver glycogen were determined to be 1.4 ± 0.1 x$10^5$ $M^{-1}$, -20.4 ±0.6 kcal $mole^{-1}$ and -44.8 ±2.1 cal $mole^{-1}$ $K^{-1}$, respectively (Fig. 36 c,d). The stoichiometry (n) of binding was found to be 1 molecule of SpyDX for every ~91 glucose residues in glycogen. The corresponding Ka, ΔH, ΔS and n values for SpnDX were 5.1 ±0.4 x$10^5$ $M^{-1}$, -26.9 ±0.2 kcal $mole^{-1}$, -64.2 ±0.5 cal $mole^{-1}$ $K^{-1}$ and ~1:88 protein/glucose. Thus, the affinities were in good agreement with those determined for starch binding, and the stoichiometries suggested a sparse distribution of binding sites in glycogen.

Most if not all α-glucan–binding CBMs, including *T. maritima* CBM41, also bind smaller oligosaccharide fragments of their target polysaccharide ligands [63; 85; 98]. Both SpyDX and SpnDX bound maltotetraose (α-(1,4)-glucotetraose), as determined by ITC. The $K_a$, Δ*H* and Δ*S* upon the binding of SpyDX to maltotetraose were determined to be 2.6 ±0.2 x$10^3$ $M^{-1}$, -14.0 ±0.1 kcal $mole^{-1}$ and -31.2 cal $mole^{-1}$ $K^{-1}$ (data not shown). The stoichiometry (*n*) of binding was found to be 1.92 ±0.01. The corresponding $K_a$, Δ*H*, Δ*S* and *n* values for SpnDX were 9.2 ±0.7 x$10^3$ $M^{-1}$, -13.6 ±0.0 kcal $mole^{-1}$ and -27.4 ±0.0 cal $mole^{-1}$ $K^{-1}$ and 2.28 ±0.00 maltotetraose/protein. The stoichiometries indicated binding of two maltotetraose molecules to one molecule of SpyDX or SpnDX, which is consistent with the presence of two modules, each with a functional binding site, in both of the proteins. Notably, the measured affinities for maltotetraose were considerably lower than those measured for cornstarch or glycogen. We rationalized this on the basis of SpyDX and SpnDX having two functional binding sites (that is, they are bivalent) interacting with polysaccharides, which can be considered as polyvalent ligands. This type of avidity effect is seen frequently with tandem constructs of polysaccharide-binding CBMs[63; 174].

Figure 36: (**a**,**b**) Depletion binding isotherm of SpnDX (**a**) and SpyDX (**b**, solid squares) with binding site mutants SpyDXΔ1 (triangles ) and SpyDXΔ2 (circles) on granular cornstarch. Solid lines, best fits to a one-site binding model. (**c**,**d**) Isotherms of SpnDX (**c**) and SpyDX (**d**) binding to type IV bovine liver glycogen, produced by ITC. The isotherms (upper charts) were produced by titrating SpnDX and SpyDX into glycogen. Solid lines in lower charts show fit of a one-site binding model to the data. All error bars represent the s.d. of measurements made in triplicate.



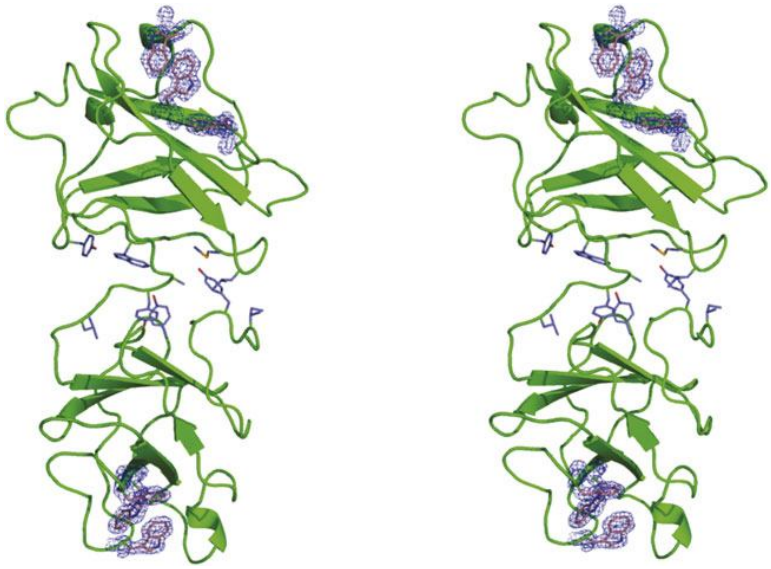
We tested this hypothesis by introducing mutations into the individual binding sites of SpyDX and assessing the binding of the mutants to starch. Three mutant SpyDX variants were created: a W28A W30A double mutant in the N-terminal module-binding site, called SpyDXΔ1; a W135A F137A double mutant in the C-terminal module binding site, called SpyDXΔ2; and a W28A W30A W135A F137A quadruple mutant, called SpyDXΔ12. These mutations were chosen on the basis of guidance from the structure of *T. maritima* CBM41 [175], which implicated these residues in binding (Figure 34b). ITC performed with SpyDXΔ1 and SpyDXΔ2 using maltotetraose as a ligand yielded stoichiometries of 1.05 ±0.12 and 0.84 ±0.11, respectively, and $K_a$s of 2.26 ±0.21 x$10^3$ $M^{-1}$ and 4.06 ±0.56 x$10^3$ $M^{-1}$, respectively. The stoichiometries of near unity, combined with the nearly wild-type affinity of the single binding site, indicated that the mutations did indeed successfully destroy the function of one site without compromising the other. Depletion isotherm studies of these mutants revealed no binding of SpyDXΔ12 (data not shown) and greatly reduced binding of SpyDXΔ1 and SpyDXΔ2 to cornstarch (Figure 36b). The low affinities of SpyDXΔ1 and SpyDXΔ2 prevented the isotherm from reaching saturation, precluding accurate deconvolution of the binding constants. However, fixing the $N_o$ value to that of the wild-type protein (5.9 μmole per 10 g starch) allowed us to determine estimates of the $K_a$s for SpyDXΔ1 (1.0 ±0.6 x$10^4$ $M^{-1}$) and SpyDXΔ2 (8.7 ±0.9 x$10^3$ $M^{-1}$), which indicated roughly 20- and 25-fold decreases in affinity relative to wild-type for the mutants, respectively. Overall, the results showed that SpyDX and SpnDX interact most tightly with highly polymerized α-glucans. Furthermore, SpyDX—and probably SpnDX, because of its amino acid sequence identity with SpyDX—requires two functional binding sites for the tight interaction

with polysaccharide. This strongly supports the hypothesis that a multivalent interaction between the proteins and α-glucan polymers is necessary for optimal affinity.

*The structural basis of α-glucan recognition:* Insights into the molecular determinants of α-glucan recognition and the basis of SpyDX and SpnDX multivalency were obtained by determining the X-ray crystal structures of these proteins. SpyDX crystallized overnight in the space group $P2_1$ with two SpyDX monomers per asymmetric unit. The final model of SpyDX lacked the C-terminal 76 amino acid residues comprising the X-module. SDS-PAGE analysis of SpyDX crystals in comparison to the protein sample that was used to set up the crystals revealed that the molecular weight of the protein that crystallized was ~7.5 kDa less than the input protein (data not shown). This is consistent with loss of the X-module by protein degradation in the crystal drop. SpnDX could be crystallized in the presence of maltotetraose only after we promoted the degradation of this protein to a stable 28-kDa product by storing the protein at room temperature for several weeks. The structure of this protein was then solved to 2.1 Å by molecular replacement, using the coordinates of SpyDX as a search model. Like SpyDX, the crystallized fragment of SpnDX lacked the amino acid residues comprising the X-module.

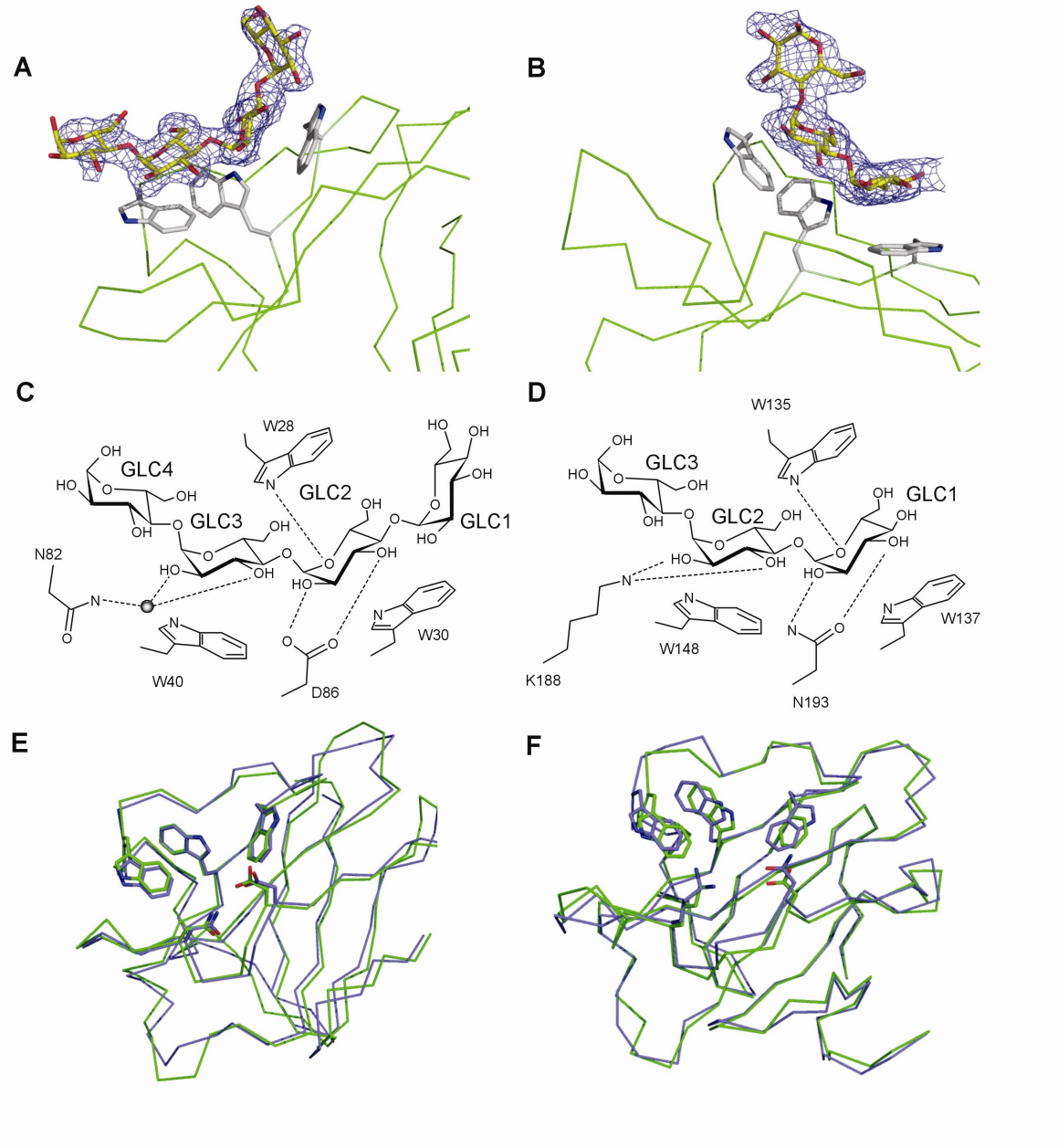
Consistent with amino acid sequence analyses of SpyDX and SpnDX, these proteins comprise two modules, each having the β-sandwich fold and Ig-like topology common to starch-binding proteins (Figure 37) [63; 98]. The two modules in these proteins are joined by a well-ordered linker and associate via an interface rich in hydrophobic amino acid side chains (Figure 37). The intimate structural association of the tandem modules, combined

Figure 37: Secondary structure of the tandem CBM41s is shown in 'wall-eyed' stereo. Binding site residues are shown as pink sticks inside blue mesh. Blue mesh is a representative electron density map, shown as a maximum likelihood[115] /$\sigma_A$-weighted [119] $2F_o - F_c$ map contoured at 1 $\sigma$(0.33 $e^-$ $Å^{-3}$), with phases and $F_c$s derived from the final refined model. The two CBM41s form a rigid structure via interacting hydrophobic interfaces, shown as blue sticks.



with our observed inability to produce functional individual modules, suggest that SpyDX and SpnDX form distinct rigid functional units. Structural overlap of SpyDX chain A with SpyDX chain B yielded an r.m.s. deviation of 0.25 Å. Overlap of SpyDX (either chain A or B) and SpnDX, which overall share 46% sequence identity, yielded an r.m.s. deviation of 1.4 Å. Comparisons between SpyDX (either chain A or B in the asymmetric unit) and SpnDX using only the N- or C-terminal modules gave r.m.s. deviations of 1.1 Å or 0.9 Å, respectively. These are comparable to the overall r.m.s. deviation, suggesting that the relative positions of the two modules in SpyDX and in SpnDX are similar. This supports the notion that the two modules in SpyDX and SpnDX are required for the formation of a single well-ordered structural unit that comprises the two modules. The electron density maps of SpnDX cocrystallized with maltotetraose revealed density for carbohydrates bound to both the N- and C-terminal modules, allowing the modeling of maltotetraose and maltotriose molecules, respectively (Fig. 38a,b). The observation of two functional binding sites was consistent with the experimental stoichiometry determined by ITC. The protein-carbohydrate interactions at the two sites were similar and dominated mainly by stacking interactions between the A-faces of the pyranose rings and aromatic amino acid side chains, but they did involve a few potential hydrogen bonds (Fig. 38c,d). There are two sub-sites apparent in each binding site, and each is occupied by a glucose molecule. In the N-terminal module, SpnDX-1, glucose 2 (GLC2; glucoses are numbered from the nonreducing end) stacks against Trp30 while Trp28 hydrogen bonds with the GLC2 O5. Asp86 makes two hydrogen bonds to the C2 and C3 hydroxyl groups of GLC2. GLC3 also stacks against Trp40 while two water-mediated hydrogen bonds are made between the O2 and O3

Figure 38: (**a**,**b**) SpnDX-1 with maltotetraose (**a**) and SpnDX-2 with maltotriose modeled (**b**). Representative electron density map (blue mesh) is a maximum likelihood [115] $\mathfrak{T}_A$-weighted[119] $2F_o - F_c$ map contoured at 1.5 σ(0.30 $e^-$ $Å^{-3}$) with phases and $F_c$s derived from the final refined model. Ligand (yellow) and binding site residues (gray) are shown as sticks. (**c**,**d**) Schematics of ligand interactions with SpnDX-1 (**c**) and SpnDX-2 (**d**). (**e**,**f**) Structural overlap of the individual CBM41s, CBM41-1 (**e**) and CBM41-2 (**f**), from SpnDX (blue) and SpyDX (green). Modules are shown as $C_\alpha$ traces and conserved binding site residues are shown as sticks.



hydroxyl groups and the side chain of Asn82. Analogous contacts are observed in the C-terminal module, SpnDX-2; however, the two water-mediated hydrogen bonds are replaced with direct hydrogen bonds between Lys188 and the O2 and O3 hydroxyl groups of GLC2. As observed with other α-glucan–binding CBMs, the convex A-face of α-glucan chains is complemented nicely by the concave shape of the SpnDX binding sites [63] (Fig. 38a,b). The amino acids involved in ligand binding by SpnDX are well conserved in SpyDX, as is the binding site architecture (Fig. 38e,f), suggesting that a nearly identical mode of interaction between α-glucans and SpyDX can be expected.

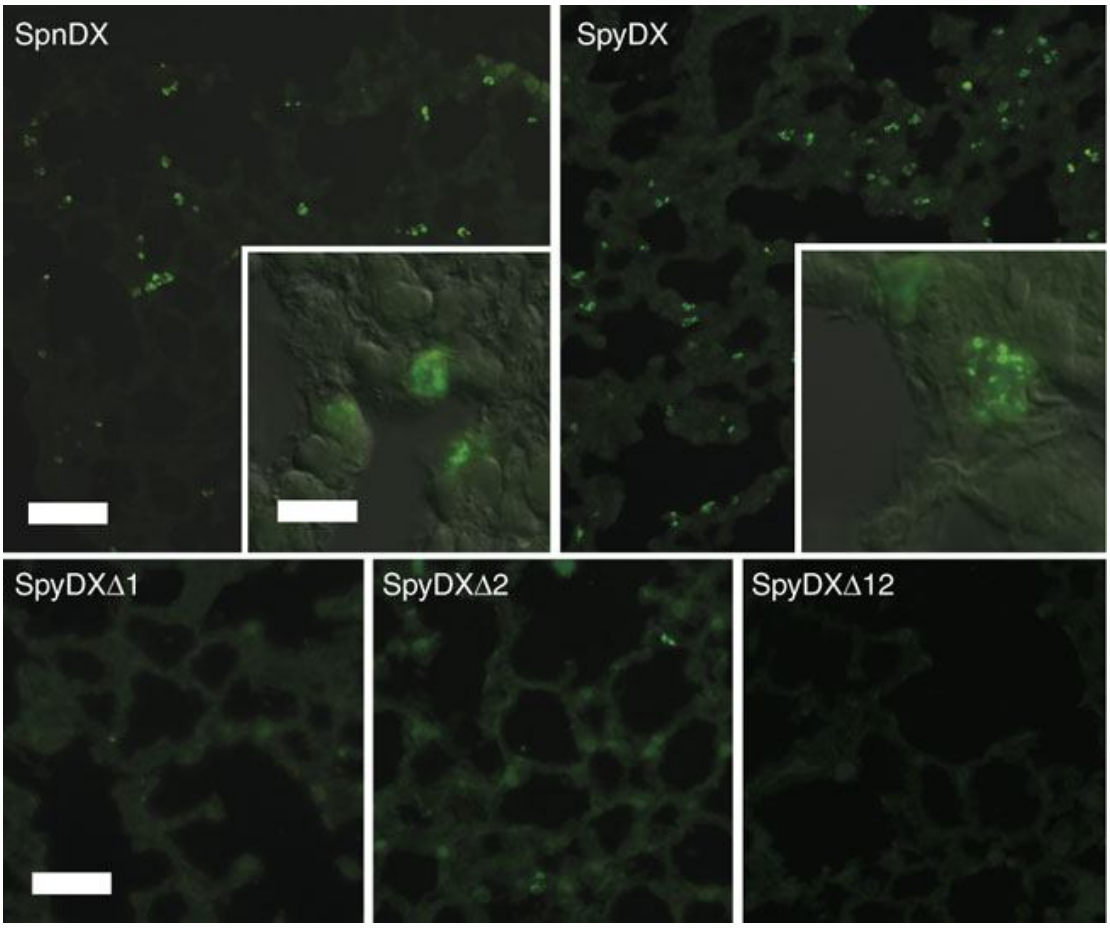
The overall arrangement of the modules in SpyDX and SpnDX is such that the two binding sites of these proteins are diametrically opposed and face in opposite directions (Figure 37). This suggests that the enhanced affinity (that is, the avidity) of these proteins for α-glucan polysaccharides is unlikely to result from the two binding sites interacting with the same glucan chain, as this would require a 180° bend in the polysaccharide chain. Rather, the enhanced affinity must result from the simultaneous interaction of the two binding sites with separate but tethered chains. This is relatively simple to envision with an insoluble polysaccharide, such as starch, where sugar chains may be held in proximity owing to their aggregated state. Consequently, one binding site on the protein interacts with one polysaccharide chain, while the second binding site on the protein interacts with a site on a separate but proximal polysaccharide chain that is anchored in the same aggregate. This scenario has been described extensively with respect to insoluble cellulose recognition[174]. This, of course, cannot occur with a soluble polysaccharide, such as the glycogen used in this study, where the binding to separate soluble chains is akin to the recognition of separate, independent ligands and would not

result in an avidity effect. Hence, the separate binding sites in SpyDX and SpnDX proteins are probably binding individual chains, but ones that are joined by a relatively close branch point, such that the binding 'region' is a somewhat conformationally restricted presentation of two binding sites.

Though the Ig-like fold of the individual modules comprising SpnDX and SpyDX is common to α-glucan–binding modules, the fusion of two modules into a single functional domain has not been observed before for any carbohydrate-binding modules. With the exception of certain family 3 CBMs that form part of the active sites of some enzymes that hydrolyze cellulose [176], all currently characterized CBMs comprise independently folded modules that function in the absence of additional accessory modules. Furthermore, to the best of our knowledge, this bilobed architecture built of two Ig-like folds to create a bivalent carbohydrate-binding protein with divergently opposed binding sites is unique among carbohydrate-binding proteins in general.

*Recognition of a ligand in the lung:* Because the *S. pneumoniae* pullulanase is a virulence factor in a murine-lung model of respiratory infection [53], we investigated whether the lung contains potential target ligands of SpyDX and SpnDX by fluorescence staining of mouse lung sections using fluorescein isothiocyanate (FITC)-labeled versions of these proteins as affinity probes. Both SpyDX and SpnDX interacted very specifically with a subpopulation of alveolar cells in both fixed (Fig. 39, top images) and unfixed (data not shown) tissue sections. In agreement with the binding data, SpyDXΔ1 and SpyDXΔ2 showed reduced binding, and SpyDXΔ12 did not appear to bind at all (Fig. 39, bottom images). The subpopulation of alveolar cells to which SpyDX and SpnDX bound was identified as alveolar type II cells by the strict colocalization of these proteins with

Figure 39: Top images, binding of wild-type modules to lung tissue, shown at x20; scale bar, 100 μM. Insets, closeups of single labeled cells, shown at x100; scale bar, 20 μM. Bottom images, decreased binding of SpyDX mutants SpyDXΔ1, SpyDXΔ2 and SpyDX Δ12, shown at x40; scale bar, 50 μM.



prosurfactant C protein (proSP-C), a marker of this specific cell type[177; 178] (Fig. 40). However, the fluorescence signal from CBM binding and anti–proSP-C binding localized to different subcellular structures (Fig. 40). The regions most intensely stained by SpyDX and SpnDX were consistent with recognition of structures in the cytoplasm of the cells (Fig. 41). However, owing to the diffuse staining at the peripheries of some alveolar type II cells by both CBMs, we cannot completely discount the possibility that SpyDX and SpnDX recognize a ligand associated with the membrane. Alveolar type II cells are relatively rich in glycogen stores; given the *in vitro* specificity of SpyDX and SpnDX for α-glucans, it is most likely that these pullulanase-derived streptococcal α-glucan–binding proteins specifically recognize the glycogen in alveolar type II cells.

*Conclusion:* Several lines of evidence are now indicating that degradation of host glycogen is an important factor in bacterial pathogenesis. First, four α-glucan–specific enzymes have been identified as virulence factors in a mouse model of *S. pneumoniae* lung infection: two predicted pullulanases (locus tags SP0268 and SP1118), a predicted α-amylase (SP1382) and a predicted glucanotransferase (SP1121) [53]. One of the pullulanases, SpuA of this study, is known to be anchored to the cell wall, whereas the α-amylase is predicted to be extracellular. The other two enzymes are predicted to be cytoplasmic. In addition to these enzymes, a malto-oligosaccharide/maltodextrin–binding component (SP2108; malX) of the malto-oligosaccharide ABC transporter has been identified as a virulence factor [53]. Overall, the α-glucan–metabolizing machinery is well conserved between *S. pneumoniae* and *S. pyogenes*, though of this machinery only the malto-oligosaccharide/maltodextrin–binding component (malE; SPY1058) of the *S.*

Figure 40: Lung tissue costained with FITC-labeled SpyDX (green, top), an antibody to ProSP-C detected with goat anti-mouse Alexa 568 (red, middle) and DAPI (blue, bottom) shown at x100; scale bar, 20 μM.

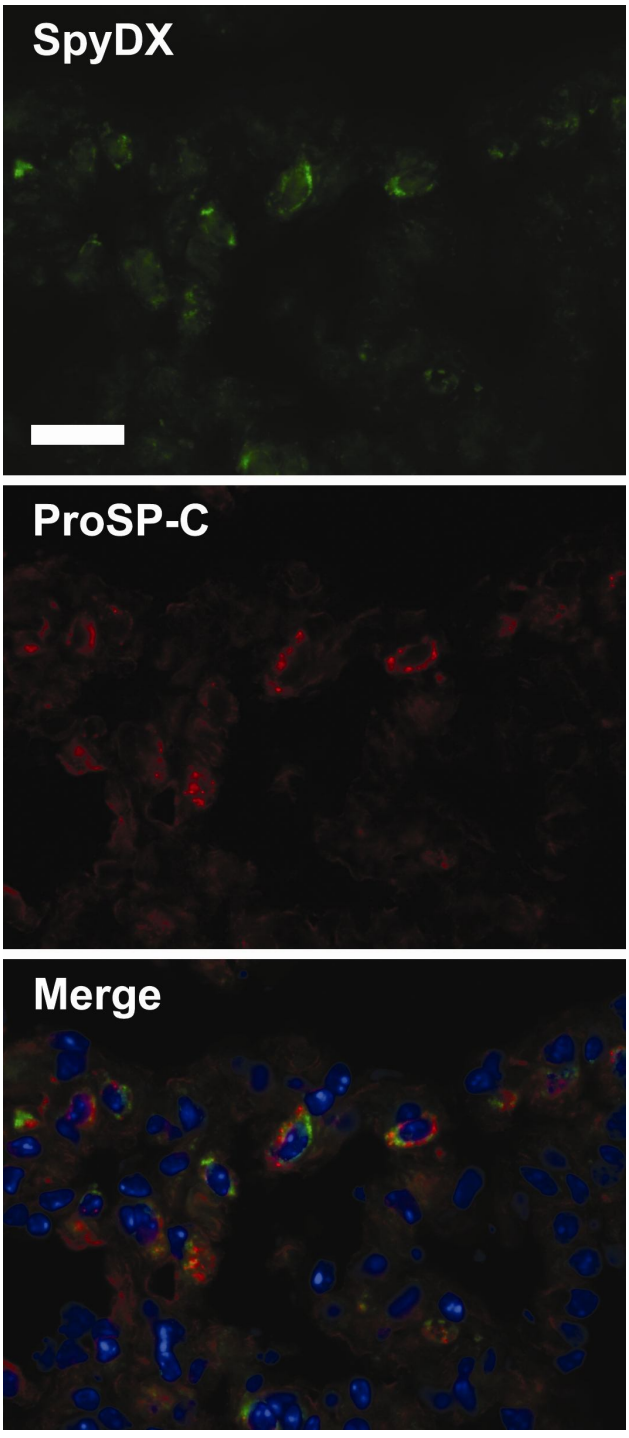
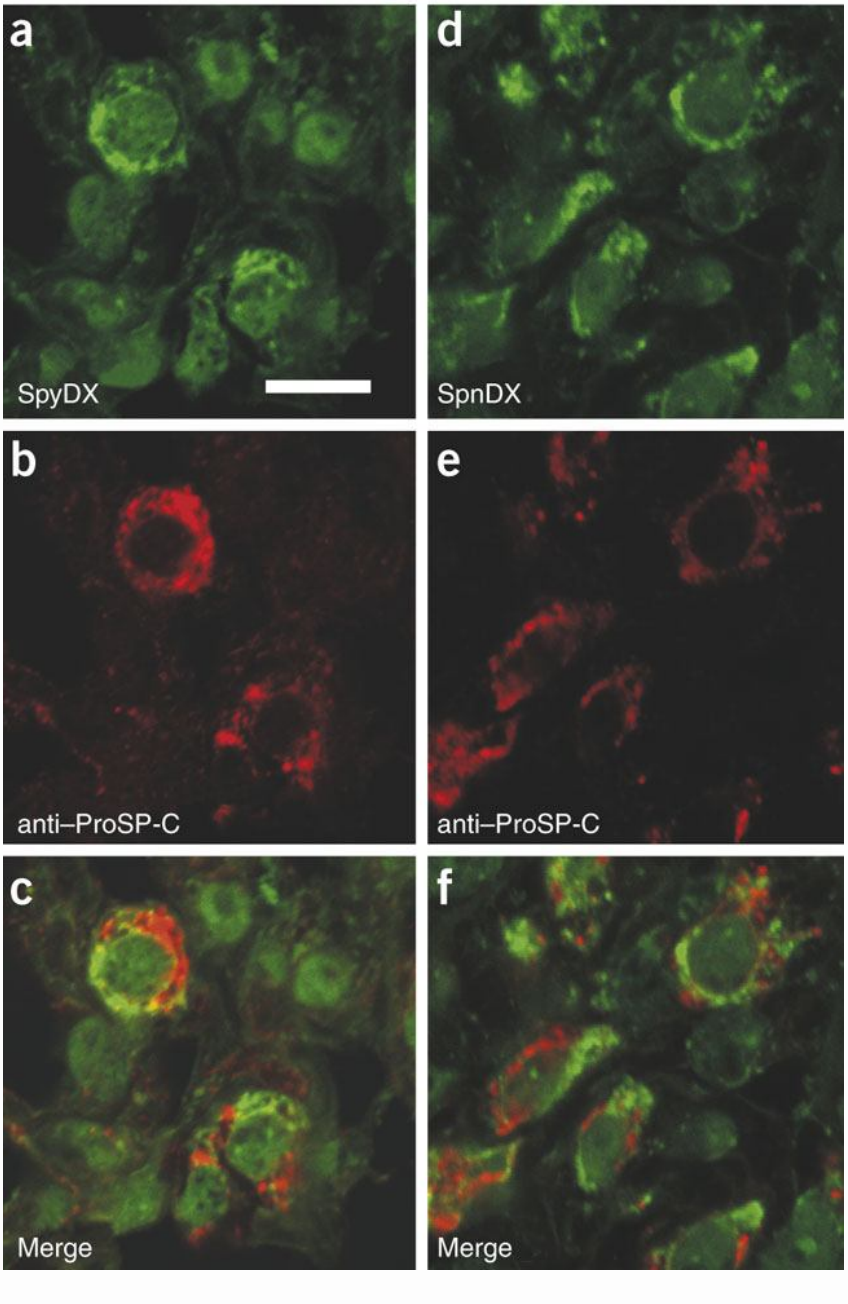


Figure 41: Shown are confocal images of lung tissue doubly stained with FITC-labeled SpyDX (**a**–**c**) or FITC-labeled SpnDX (**d**–**f**) and an antibody to ProSP-C, a marker for type II alveolar cells. **a** and **d** show staining with FITC-labeled SpyDX and SpnDX, respectively (green). **b** and **e** show staining with a specific ProSP-C antibody detected with goat anti-mouse Alexa 568 (red). **b** and **f** are merged CBM-FITC and anti–ProSP-C images. All images are shown at x60; numerical aperture, 1.4; scale bar, 25 μM.



*pyogenes* malto-oligosaccharide ABC transporter has been shown to be directly involved in virulence [169], whereas PulA is only implicated in virulence [152]. Here we have demonstrated the strong preference of the N-terminal carbohydrate-binding modules (SpyDX and SpnDX) of the cell wall–anchored streptococcal pullulanases for glycogen *in vitro*. These seem to directly target the glycogen in alveolar type II cells. This evidence, taken together, strongly points to a role of α-glucan degradation and/or metabolism in the success of *S. pneumoniae*, and probably *S. pyogenes*, as pathogens.

Alveolar type II cells have previously been suggested to be a primary target for carbohydrate-dependent adherence of *S. pneumoniae* [179]. This seems to be mediated by non–glucose-based extracellular glycans, indicating that α-glucans, and thus the CBM41s, are probably not a factor in adhesion to the surface of this type of lung cell [179]. However, *S. pneumoniae* is an invasive pathogen that can gain entry into cells without killing them [180]. Indeed, extensive tissue infiltration in the early stages of invasive infection seems to occur via 'transcellular migration', without causing cell death, rather than by a route that follows the extracellular spaces [180]. This suggests that the bacterium does not rely on cell death and consequent lysis to access glycogen in lung tissue. Our proposed model is that the bacterium first penetrates alveolar cells through the recognition of specific extracellular glycans or other non–carbohydrate-based ligands. Once inside, the bacterium is able to target and adhere to intracellular glycogen stores via the pullulanase-associated CBM41 modules, then degrade this polysaccharide.

The primary reason for attacking host glycogen, which is concentrated in a number of cell types throughout the body where it is needed as an easily mobilized energy source,

would be to liberate the energy from this carbohydrate. However, there is a plausible secondary purpose of targeting this polysaccharide in the lungs, and that is to undermine the initial defenses of the immune system. As mentioned, glycogen is abundant in alveolar type II cells, which are responsible for surfactant production in the lung [181]. This process can involve the recycling of existing surfactant or synthesis of new surfactant molecules for which glycogen is a precursor. In neonates, glycogen stores in alveolar type II cells are particularly abundant, as surfactant synthesis is especially vigorous owing to its importance in the establishment of normal lung function after birth [182]. In addition to its role in maintaining proper surface tension in the airways, surfactant is part of the innate immune system, as it creates a physiochemical barrier that contains a number of proteins involved in host defense, such as surfactant proteins A to D [183]. Indeed, surfactant protein D has been shown to have a role in the clearance of *S. pneumoniae* at the early stages of infection [184]. By targeting alveolar type II cells and depleting their glycogen stores, pathogens such as *S. pneumoniae*, and possibly *S. pyogenes*, may also negatively influence the synthesis of surfactant, thus compromising this first-line host defense mechanism and promoting fulminant infection.

In conclusion, the pullulanase-like surface proteins SpuA and PulA, from *S. pneumoniae* and *S. pyogenes*, respectively, harbor glycogen-binding modules with a unique fused architecture near their N termini. These modules may perform the classic function of CBMs, which is to hold the enzyme in proximity to the preferred substrate. As SpuA and PulA are surface attached, this function probably adheres entire cells to glycogen granules, thus promoting the efficient breakdown of glycogen. We made the crucial and novel observation that these modules adhere tightly and specifically to glycogen in the

context of type II alveolar cells in the lung. Such a function is quite unlike any described previously for a CBM and, perhaps more importantly, strongly suggests that this intracellular polysaccharide is a target of streptococci during invasive lung infections. This provides, for the first time, plausible rationales for the existence of complete α-glucan–metabolizing machineries in streptococci and the observation that many of these proteins are indeed virulence factors. Ultimately, our results suggest the potential for new therapeutic strategies that target the disruption of glycogen adherence, breakdown or both as a means of treating streptococcal lung infections.

## 3.5 Discussion on Family 41 CBMs

### 3.5.1 Comparison of family 41 CBMs

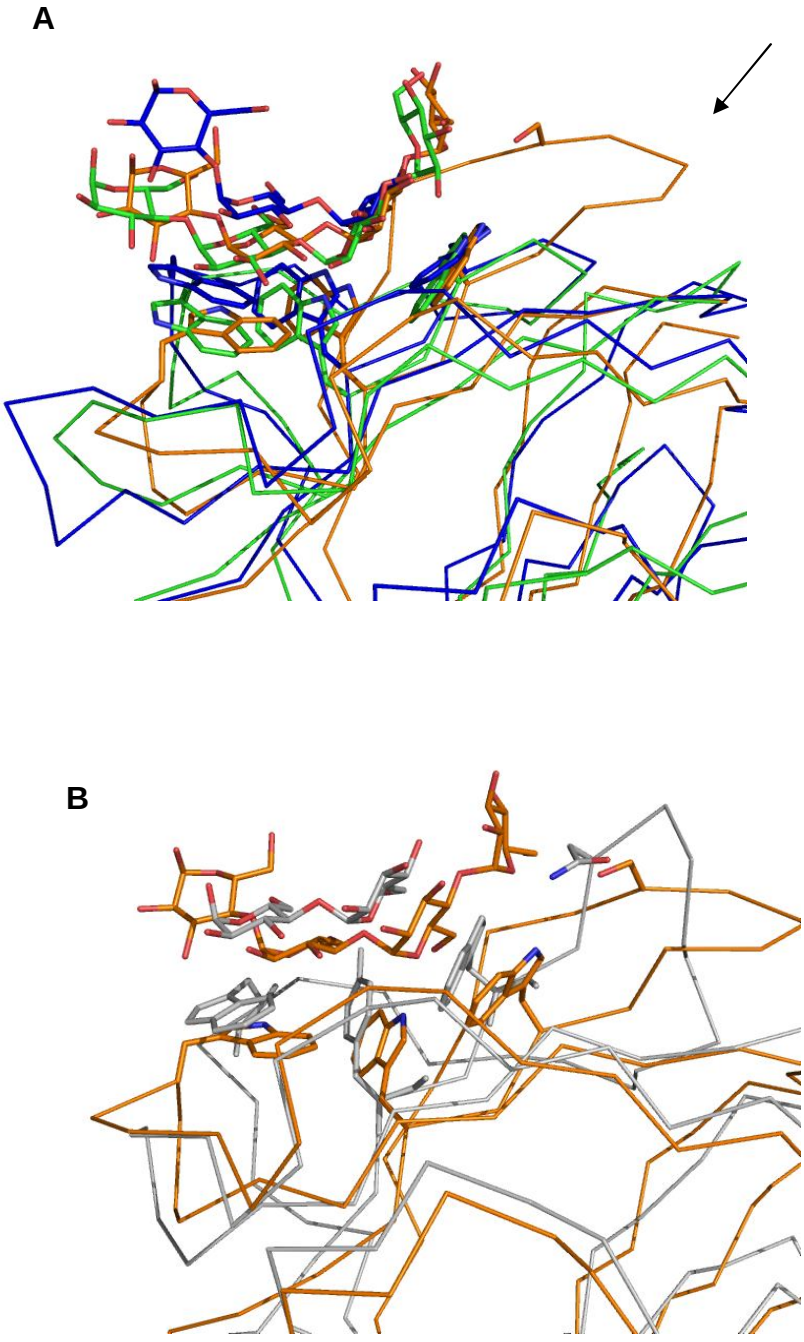
The objective was to determine how the family 41 CBMs would be able to accommodate an α-1,6-linked glucose found in pullulan. *Tm*CBM41 is able to achieve this with an additional loop region containing a serine residue. The serine side chain provides an additional binding subsite which can either interact with an α-1,4-linked glucose or an α-1,6-linked glucose (Figure 42A). This loop is absent in the SpnDX and SpyDX modules, however, this region is solvent exposed and in theory would allow for an α-1,6-linked glucose to be positioned at this location without making any additional contacts with the protein. Therefore it appears as though only *Tm*CBM41 has evolved an additional subsite for interacting directly with an α-1,6-linked glucose. CBM41 from *K. pneumoniae* PulA also has a loop region similar to the loop region in *Tm*CBM41 but has a structurally aligned asparagine residue instead of a serine (Figure 42B); however, it remains unknown whether this asparagine may interact with an α-1,6-linked glucose. The direct interaction of *Tm*CBM41 with an α-1,6-linked glucose can be rationalized because the *T. maritima* pullulanase only has one CBM41 module and perhaps in order to increase its enzymatic efficiency has evolved a third subsite to directly interact with α-1,6-linked glucose, thus increasing the likelihood it will bind to any location in starch and pullulan. *Kp*CBM41 interaction with α-1,6-linked glucose remains unknown; however, because this enzyme has only on CBM41 it may also have an additional subsite for interacting with α-1,6-linkages. The streptococcal pullulanases contain two tandem CBM41s for the high affinity interaction with chains of glycogen. Because both modules cooperatively bind α-

glucans, perhaps the modules do not require the additional binding subsite for their interaction with α-1,6-linked glucans.

Another rationale is that when we look at the origin of CBM41 modules, which are only appended to pullulanases, they are mainly found in pathogenic bacteria that would be unlikely to encounter pullulan as a substrate. The more likely target is glycogen which we have shown interacts with SpyDX and SpnDX modules from streptococcal pullulanases. Because the frequency of α-1,6-linkages is much less in glycogen than pullulan, it is likely unnecessary for these modules to require an additional subsite for interacting with an α-1,6-linkage. Additionally, the bivalent architecture of the double CBM41 modules in SpyDX and SpnDX specifically targets α-1,4-linked glucan chains within starch and glycogen granules, which decreases the probability that they would interact with α-1,6-branches. This feature also provides specificity for intact substrate rather than products formed from the debranching activity of the enzyme.

Perhaps only *Tm*CBM41 would require the additional subsite provided by S35 for interacting with pullulan, or, since *T. maritima* lives at temperatures around 80$^{o}$C in deep sea vents, it has evolved this additional subsite for a more stabile interaction with starch at high temperatures.

Figure 42: (A) Structural overlap of SpnDX modules (SpnDX-1 green, SpnDX-2 blue PDB code 2J44) and TmCBM41 (orange PDB code 2J72) in complex with maltotetraose. Loop region in *Tm*CBM41 with serine indicated with arrow. (B) Overlap of *Tm*CBM41 (orange) and *Kp*CBM41 (gray PDB code 2FHF).



### 3.5.2 Comparison of CBM41s with Starch-binding modules from different CBM families

There are 9 families of alpha-glucan-binding CBMs with structures representing 7 families (20, 21 (NMR only), 25, 26, 34, 41, and 48) . Structures of CBM in complex with α-1,4-linked glucose ligands are known for families 20, 25, 26, 34 and 41. A structural alignment with these CBMs in complex with malto-oligosaccharides shows extremely similar overall three-dimensional structures and location for ligand interaction (Figure 43). All have two regions, region a and b (Figure 44A) with tryptophan residues providing two binding subsites that serve as platforms for interacting with the A-faces of two α-1,4-linked glucose molecules (Figure 44A). They are positioned such that they form a concave binding groove to accommodate the convex three-dimensional shape of α-1,4-linked glucose. Region c contains a planar hydrophobic amino acid side chain (Try or Tyr), or a histidine as is the case for *Bh*CBM25 that hydrogen bonds with the glucose positioned in region b (Figure 44B). *Bc*CBM20 is an exception as it has a threonine in this region but the side chain hydrogen bonds with glucose in the same manner.

When we look at the affinities of these CBMs for maltooligosaccharides for which there is data, the affinities all fall in the $10^{-4}$ $M^{-1}$ range except for *Tm*CBM41 ($10^{-6}$ $M^{-1}$), however the data was taken at 25 deg C. At 80 $^{o}$C, the temperature in which *T. maritima* lives, the affinity would likely be similar. Thus, we can say that the forces driving starch-protein interaction (stacking interactions, van der waals, hydrogen bond) are all very similar. All of these CBMs bind starch/maltooligosaccharides in a nearly identical manner and have similar three dimensional structures with slight variability apparent in the loop regions between β-strands. However CBMs outside a given CBM

family have very different amino acid sequences, sometimes having less than 5% amino acid sequence identity. We can attribute this to the common three-dimensional structure of the α-1,4-linked glucose ligand shared by all of these CBM families. Even though the amino acid sequences vary outside CBM families, all have evolved a binding site best suited to interacting with maltooligosaccharides.

Figure 43: (A) Structural Overlap of all ligand-bound starch-binding CBMs showing the face where the binding sites are located. SpnDX-1 (green) and SpnDX-2 (blue) (PDB code 2J44), TmCBM41 (orange, PDB code 2J72), KpCBM41 (grey, PDB code 1FHF), BhCBM25 (cyan, PDB code 2C3X), BhCBM26 (magenta, PDB code 2C3H) and BcCBM20 (red, PDB code 1VEO). (B) Overlap of binding site residues. *Tm*CBM41 backbone shown for reference.

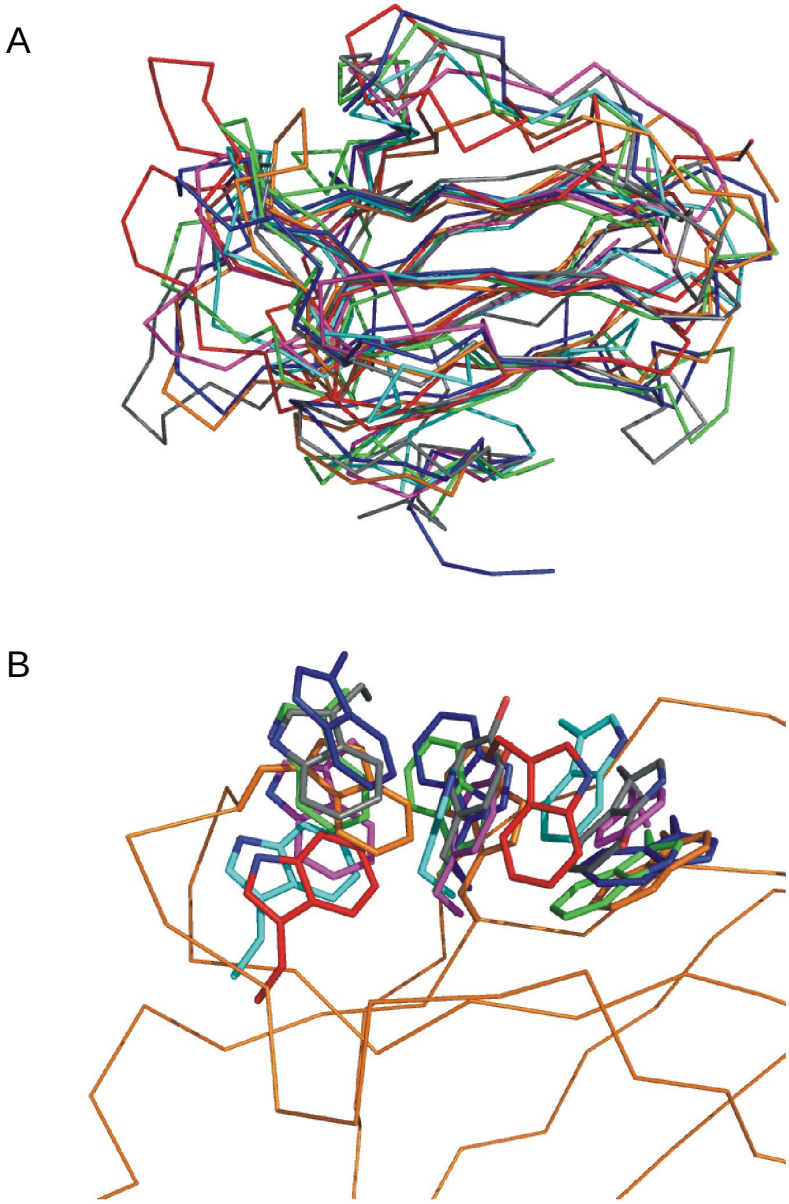
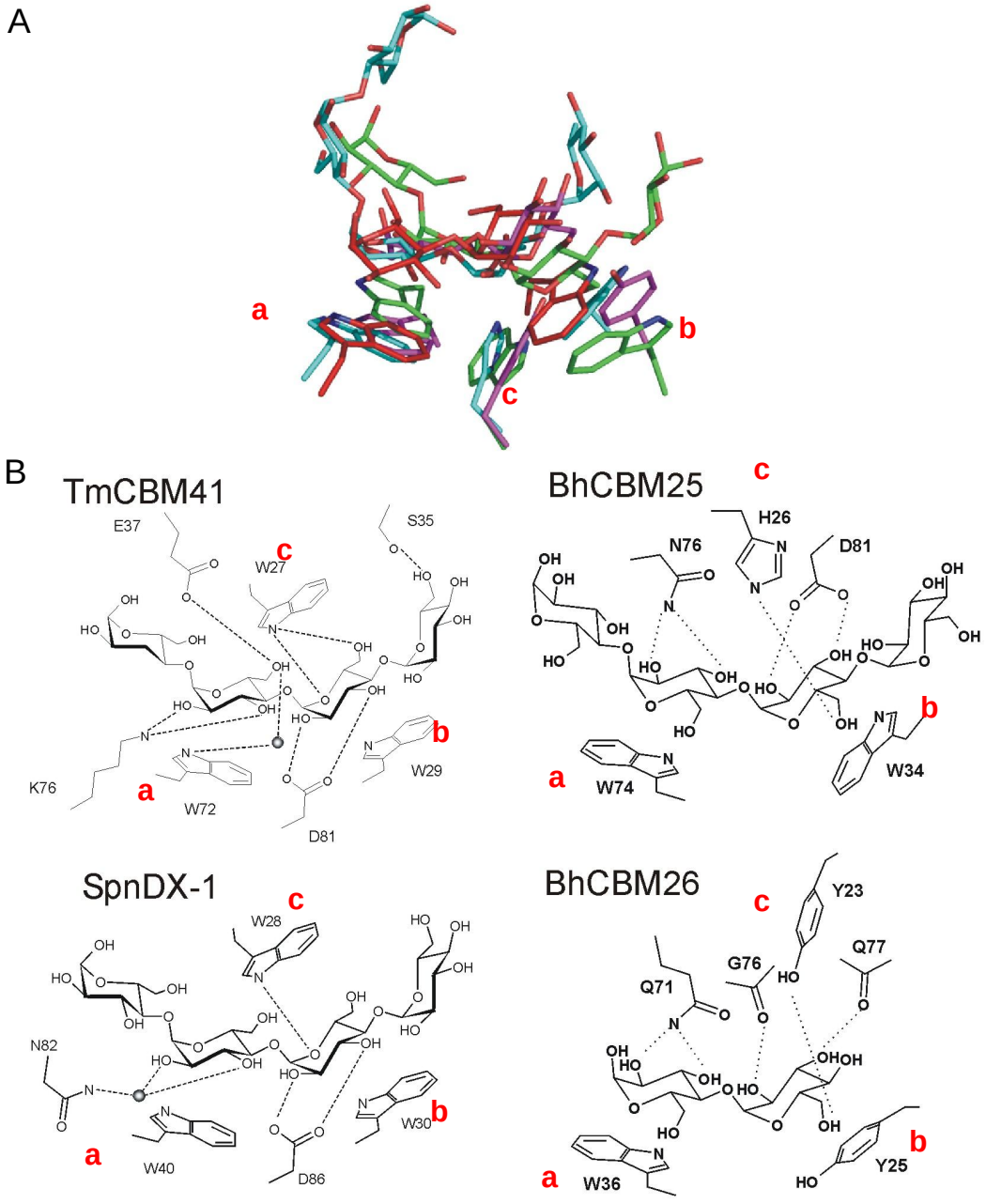


Figure 44: (A) Representative structures families of starch-binding CBMs bound to maltooligosaccharides: SpnDX-1 (green PDB code 2J44), *Bh*CBM25 (cyan PDB code 2C3X), *Bh*CBM26 (magenta PDB code 2C3H) and *Bc*CBM20 (red PDB code 1VEO).(B) Hydrogen bonding schemes from various starch-binding CBMs to show the conserved mode of α-glucan binding.



# Chapter 4: Glycogen Degradation by SpuA, a Streptococcal Virulence Factor

Alicia Lammerts van Bueren, Mirijam Czjzek, and Alisdair, B. Boraston

Unpublished Data

Contributions to research: Cloning, mutagenesis, protein production, enzymatic activity characterization, crystallization, data collection, structure solving, structure refinement, writing.

## 4.1 Abstract

SpuA is a cell surface-associated enzyme produced by *Streptococcus pneumoniae* and a known virulence factor essential for full virulence of the organism. In our studies with the N-terminal CBM41 modules SpnDX, we have identified a potential target of SpuA as glycogen granules within type II alveolar cells in lung tissue (See Section 3.4). In this study we demonstrate activity of SpuA on glycogen and show that it forms maltooligosaccharide products of varying lengths. The structure of SpuA in complex with maltotetraose was solved by X-ray crystallography and the structures of native SpuA and SpuA with maltotetraose in solution were solved using small angle X-ray scattering (SAXS) experiments. These structures reveal a flexible linker region which facilitates the positioning of the catalytic module at an α-1,6-branch point mediated by SpnDX interaction with opposing glycogen chains. Inhibition studies show that common glucanase inhibitors are ineffective against SpuA, likely due to SpuA's unique active site which accommodates branched glucans. Since SpuA is a target for developing therapeutic compounds that would impede the progression of *S. pneumoniae* infection, these studies may aid in designing selective inhibitors against SpuA without affecting human α-glucosidases.

## 4.2 Introduction

*Streptococcus pneumoniae* is a gram positive pathogenic bacterium that kills more people in the US each year than all other vaccine preventable disease combined. Approximately 20% of the population harbors *S. pneumoniae* in their normal flora but when an individual is immuno-compromised, the bacterium can cause serious infections due to its opportunistic nature. It mainly affects individuals such as the elderly and children. According to the World Health Organization *S. pneumoniae* kills approximately one million children under the age of five worldwide annually. Virtually every child in developed countries becomes a carrier of the bacteria leading most commonly to otitis media (ear infection) but often to more serious and sometimes deadly infections such as septicaemia, pneumonia and bacterial meningitis [165]. It is an invasive intracellular pathogen that infects patients through the nasopharynx and bronchio-epithelial cells of the lower respiratory system [180]. Attachment and internalization is mediated by capsular polysaccharide at mucosal surfaces [4] and bacteria slowly infiltrate through epithelial cell layers until they reach the bloodstream where they can then spread rapidly throughout the body [180].

The primary virulence factor of *S. pneumoniae* is capsular polysaccharide (CPS) which forms a sugar envelope around the bacterium for protection from the immune system and to facilitate cellular attachment [185]. There are over 90 different serotypes which are based on different CPS compositions; however, the 11 most common serotypes account for >75% of infections. Prevnar ® (Wyeth) is a 7 conjugate vaccine containing the capsular polysaccharide from the 7 most common disease causing serotypes of *S. pneumoniae* [186]. The vaccine is given to children starting at 2 months of age and greatly

reduces the rate of pneumococcal infection; however, immunity is short-lived and requires many boosters to extend protection. Due to protection from these 7 common serotypes, a phenomenon called serotype replacement is occurring, which has allowed other serotypes to enter the population as major disease causing strains against which individuals have no immunity [187]. Current vaccines offer little or no protection against these replacement serotypes [188].

The discovery of antibiotic resistant strains of *S. pneumoniae* has also made this bacterium a threat to the population [189; 190]. Emerging strains have been found to be resistant to all penicillins, cephalosporins and macrolides. Recently children have been infected with *S. pneumoniae* resistant to all known FDA approved antibiotics for children [191]. The imminent threat of emerging serotypes and new antibiotic-resistant strains of *S. pneumoniae* has initiated further research into finding new targets to stop the spread of infection as well as the development of extended conjugate vaccines and new vaccine candidates to prevent *S. pneumoniae* infection.

A large scale analysis was completed on *S. pneumoniae* serotype 4 that used signature-tagged mutagenesis (STM) to identify ~230 genes that are essential for lung infection [53]. Approximately 19 of these genes encode for glycoside hydrolases. Based on cell localization predictions, eight of these are anchored to the bacterial surface, 10 are considered cytoplasmic, and one, a putative endo-β-galactosidase active on blood group antigens, is fully secreted. One of these surface-associated enzymes is an enzyme from GH family 13 called SpuA that has been functionally characterized as a pullulanase [170]. Pullulanases are enzymes that hydrolyze the α-1,6-linkages in pullulan and are sometimes referred to as starch-debranching enzymes. SpuA has previously been shown

to be active on pullulan and starch but reported to be inactive on glycogen [170]. When we look further into the STM studies we find that there are other enzymes involved in α-glucan metabolism that are essential for *S. pneumoniae* virulence [53] (see Figure 4.1.1). *S. pneumoniae* is a human pathogen with no known environmental niche so the question asked is why is SpuA essential for *S. pneumoniae* virulence and furthermore, why is α-glucan metabolism necessary for *S. pneumoniae* infection in lung tissue?

Recently the N-terminal CBM41 modules from SpuA were shown to interact with α-1,4-glucans in starch and glycogen and with glycogen granules in type II alveolar cells within mouse lung tissue [192]. ***Our hypothesis is that the target substrate for SpuA is glycogen where it cleaves the α-1,6-linkages, releasing α-1,4-glucan products.*** The products of glycogen metabolism would then be used by the bacteria as a nutritional source via the α-glucan metabolism pathway contributing to the overall pathogenesis of the organism. To test this hypothesis, we carried out several biochemical analyses of SpuA to show activity of SpuA on glycogen and determine the products of glycogen degradation. We solved the structure of SpuA in complex with maltotetraose and identified the potential catalytic residues by site-directed mutagenesis to elucidate the mechanism of glycogen hydrolysis. We also show using SAXS that the region separating the catalytic module from the CBMs acts as a flexible linker which facilitates positioning of the catalytic module to the branch points in glycogen for effective hydrolysis of the α-1,6-linkages. X-Ray crystallography data shows that the first CBM41 module makes up a portion of the active site, demonstrating a novel function for CBMs. This is the first study revealing how glycogen degradation might contribute to pathogenesis of *Streptococcus pneumoniae* in lung tissue.

## 4.3 Materials and Methods

*Carbohydrates and Polysaccharides* – Type IV bovine liver glycogen, maltooligosaccharides, pullulan and amylopectin (starch) were from Sigma (St. Louis, MO). Red-pullulan and Red-starch were from Megazyme (Bray, C. Wicklow, Ireland).

*Cloning, Expression, and Purification of full length SpuA and catalytic module* . DNA fragments encoding full length SpuA without the N-terminal signal peptide and C-LPXTG motif and the catalytic GH13 module (SpuAΔCBM) without the C-terminal LPXTG motif were amplified from *S. pneumoniae* genomic DNA (TIGR strain BAA-334D). SpuA 5' oligonucleotide primer was 5'-CACCCATATG*GCTAGC*GATAACTACTTCCGTATC-3'. SpuAΔCBM 5' oligonucleotide primer was 5'-CATATG*GCTAGC*ACTGTTAGCTACAATTCCGACCAATTC-3' with NheI restriction sites italicized. The 3'oligonucleotide primer for both SpuA and SpuAΔCBM was 5'-GGATCC*CTCGAG***TTA**TTCAGCTTGTTTATCTGGGGTTGC-3' with stop codon in bold and XhoI restriction site italicized. Amplified DNA yielded products of 3350 bp for SpuA and 2500 bp for SpuAΔCBM, which were subsequently cloned into pET28a plasmid vector with NheI and XhoI restriction sites giving pET28SpuA and pET28 SpuAΔCBM. Alanine mutants of SpuA were generated by site-directed mutagenesis using a modified PCR primer method [172] using the mutagenic primer 5'GGCTTCCGTTTC**GCT**ATGATGG<u>G*C*GACC</u>ATGACGCCGCT3' to generate SpuA D323A (mutation from D to A in bold, silent site removing a BsaI restriction site in

italics and underlined) and the mutagenic primer

5’CTCATCATGCTTGGT**GCC**GG*C*TGGAGAACCTATGCC3’

to generate SpuA E352A (mutation from E to A in bold, silence site adding an NaeI restriction site in italics and underlined). The fidelity of the cloned inserts was determined by DNA sequencing.

*Protein Production and Purification.* Vectors containing insert were transformed into *Escherichia coli* BL21(STAR) DE3 cells (Novagen) for polypeptide production. All polypeptides contained an N-terminal 6-His tag fused to the protein of interest by a thrombin-cleavage site. Three litres of LB medium supplemented with 50 μg/L kanamycin were inoculated with *E. coli* BL21(STAR)DE3 cells harbouring each plasmid and incubated with shaking at 37$^{o}$C until an optical density at 600 nm of 0.8-1.0 was reached at which point the media was supplemented with 1.0 mM IPTG to induce protein production. Further growth continued for ~4 hours at which point the cells were harvested and cell pellets were frozen overnight at -20$^{o}$C. Cells were then thawed and resuspended in 100 ml of 20 mM Tris-HCl pH8.0, 0.5 M NaCl supplemented with Protease Inhibitor Cocktail (Roche) and lysed by French press. After centrifugation at 15000 rpm for 45 min, the supernatant was collected, and the polypeptides were purified by immobilized metal affinity chromatography (IMAC) with Ni-Sepharose resin (GE Healthcare) according to manufacturer's protocols. Purified polypeptides were concentrated in a stirred ultrafiltration unit on a 10K molecular weight cutoff filter and dialyzed against 20 mM Tris-HCl, pH 8.0, using regenerated cellulose dialysis tubing with a 3K MWCO. Purity was greater than 95%, as assessed by SDS-PAGE. Yields were typically 20-50 mg/L of culture.

*Determination of Protein Concentration.* The concentration of purified SpuA and SpuAΔCBM were determined by UV absorbance at 280 nm using the calculated extinction coefficients 172930 $M^{-1}$ $cm^{-1}$ for SpuA and 106120 $M^{-1}$ $cm^{-1}$ for SpuAΔCBM [109].

*Zymograms.* Native polyacrylamide gels containing 0.1% SDS were polymerized with 0.5% type IV bovine liver glycogen, red-pullulan or red-starch. 20 μg of SpuAΔCBM or SpuA were loaded into individual lanes and electrophoresed at 150V for 80 minutes in a XCell SureLock Mini-Cell system (Invitrogen). Gels were then incubated in 50 mM Tris-HCl pH 7.4 at 37°C for 1 hour changing the buffer every 20 minutes. Activity on red-starch and red-pullulan was observed by a clearing in the gel. Activity on glycogen was observed by a clearing in the gel after staining the gel with iodine solution (Sigma).

*Thin Layer Chromatography.* 30 μg of SpuAΔCBM was added to 1 ml solutions of 1% amylopectin, amylose, pullulan, glycogen and dextran in 20 mM Tris-HCl, pH 7.4, and incubated for 2 hours at 37°C. 3 μl of each reaction and 1 μl of maltooligosaccharide standards were spotted onto a silica gel-coated plate and allowed to dry completely at room temperature. The plate was placed in a sealed glass container preequilibrated for 30 minutes with 100 ml of a solution containing 7/2/1 ratio of n-propanol/ deionized water/ ethanol. The plate remained in solution until the solvent front was ¾ of the way up the plate (approximately 4 hours). Samples were visualized by dipping the plate in a solution of 95% ethanol/5% $H_2SO_4$ and baked at 110°C for 20 minutes.

*Fluorophore-Assisted Carbohydrate Electrophoresis.* The procedure for FACE was adapted from Jackson *et. al.* [193]. Enzyme reactions were carried out as per TLC reactions

(above) for both SpuAΔCBM and SpuA with glycogen and pullulan. Immediately following the reactions, 10 μl (100 μg) of polysaccharide was removed from each tube and dried in a Speedvac for 45 minutes at 50$^{o}$C. Labeling of the sugar products was carried out by adding 5μl of a solution of 0.15% ANTS in 15% acetic acid and 5μl of 1 M $NaCNBH_3$ in DMSO to the dried samples. The reaction was incubated overnight in the dark at 37$^{o}$C. The ANTS-labeled products were then resuspended in 50 μl deionized water plus 50 μl of 0.01% Thorin I loading dye (Sigma) in 20% glycerol. Approximately 1-3 μg of ANTS-labeled product were loaded onto a 28% polyacrylamide (19:1) gel with a 10% stacking gel and electrophoresed at a constant 15 mA for 105 minutes at 4$^{o}$C in native running buffer (25 mM Tris-HCl, 0.2M glycine) in a XCell SureLock Mini-Cell system (Invitrogen). Gels were immediately visualized under UV light.

*Crystallization of* SpuAΔCBM *and SpuA.* 10 mg of SpuA and SpuAΔCBM were further purified by size exclusion chromatography using a Sephacryl S-200 column (GE Healthcare) and separated in a buffer containing 20 mM Tris-HCl pH 8.0. Fractions were assessed by SDS-PAGE and those showing a single band were pooled and concentrated for crystallization using the vapour diffusion hanging drop method with drops containing a ratio of 1:1 protein:mother liquor. Crystals of SpuAΔCBM grew after 1 month at 5 mg/ml in 17.5% PEG 3350, 0.2 M $MgCl_2$, 0.1 M Na Acetate pH 5.5 and 3% glycerol. Crystals of SpuA in complex with maltotetraose grew overnight at 15 mg/ml in 2 M $(NH_4)_2SO_4$, 0.2 M Na/K Tartrate, 0.1 M tri-Na citrate pH 5.6 supplemented with 5% MPD or 3% Glycerol. Crystals were flash frozen in liquid nitrogen using a cryoprotectant

comprising mother liquor supplemented with 20% glycerol for SpuAΔCBM and 15% glycerol for SpuA.

*Data Collection, structure determination and refinement.* Diffraction data for both SpuAΔCBM and SpuA were collected with a Rigaku R-AXIS 4++ area detector coupled to a MM-002 X-ray generator with Osmic 'blue' optics and an Oxford Cryostream 700. All data were processed using Crystal Clear/d*trek [129]. Data collection and processing statistics are given in Table 14. The structure of SpuAΔCBM was solved by molecular replacement using MOLREP [113] with *the Klebsiella pneumoniae* pullulanase catalytic module coordinates as a search model (PDB code 2FHF). The initial model was corrected and completed manually by successive rounds of building using COOT [160] and refinement with REFMAC [115]. SpuA was solved by molecular replacement with MOLREP [113] using the coordinates of SpuAΔCBM and SpnDX (PDB code 2J44) as search models. The initial model of SpuA was manually corrected and refined as above. Water molecules were added using the REFMAC implementation of ARP/wARP and inspected visually. In both data sets, 5% of the observations were flagged as 'free' [112] and used to monitor refinement procedures. All final model statistics are given in Table 14.

*Isothermal Titration Calorimetry.* Isothermal titration calorimetry (ITC) was carried out as described previously [68] with a VP-ITC (MicroCal). Protein was buffer exchanged into 20 mM Tris-HCl pH 8.0 extensively by dialysis and the final buffer was saved for resuspending the inhibitors for the ITC experiments. ~1 mM inhibitor was titrated into

Table 14: Data Collection and structure statistics for SpuA

| | GH13 | SpuA with maltotetraose |
|---|---|---|
| **Data collection** | | |
| Space group | $P2_1$ | $P2_12_12_1$ |
| Cell dimensions | | |
| *a*, *b*, *c* (Å) | 57.57 75.20 87.04 | 80.14 86.43 193.22 |
| α, β, γ (°) | 90.00 97.31 90.00 | 90.00, 90.00, 90.00 |
| Resolution (Å) | 19.85 - 1.85 (1.92 - 1.85)* | 19.93 - 2.25 (2.33 - 2.25) |
| $R_{merge}$ | 0.057 (0.403) | 0.086 (0.396) |
| $I / \sigma I$ | 10.2 (2.3) | 9.4 (2.8) |
| Completeness (%) | 96.7 (94.1) | 92.6 (94.8) |
| Redundancy | 3.73 (3.64) | 4.25 (4.18) |
| **Refinement** | | |
| No. reflections | 226617 (60816 unique) | 253465 (59606 unique) |
| $R_{work}$ / $R_{free}$ | 0.176/0.234 | 0.206/0.274 |
| No. atoms | | |
| Protein | 5712 | 7561/140 (linker region) |
| Ligand/ion | n/a | 124 (GLC)/4 (Na) |
| Water | 940 | 806 |
| *B*-factors | | |
| Protein | 26.533 | 33.229/ 67.594 (linker region) |
| Ligand/ion | n/a | 36.493 (GLC)/ 40.849 (Na) |
| Water | 39.216 | 38.175 |
| R.m.s deviations | | |
| Bond lengths (Å) | 0.018 | 0.017 |
| Bond angles (°) | 1.603 | 1.999 |

*Highest resolution shell is shown in parenthesis.

~100 μM SpuAΔCBM or SpuA to sub-saturation levels due to the low affinity of protein for the inhibitors. Inhibitor concentrations were calculated such that they were in 5 times molar excess of the theoretical dissociation constants. All data generated are from single titrations.

*Small Angle X-Ray Scattering Experiments* - Synchrotron X-ray scattering data from solutions of native SpuA and SpuA with maltotetraose were collected at the X33 beamline of the EMBL (DESY, Hamburg) [194] using a MAR345 image plate detector by Dr. Mirijam Czjzek. Dr. Czjzek performed all of the data analysis for these experiments and provided the description for this Materials and Methods section. A 4.2 mg/ml solution of BSA was measured as a reference and for calibration. The scattering patterns were measured with an exposure time of 2 min at 288 K. The wavelength was 1.5Å. The sample-to-detector distance was set at 2.4m, leading to scattering vectors q ranging from $0.06Å^{-1}$ to $0.5Å^{-1}$. The scattering vector is defined as $q=4\pi/\lambda \sin\theta$, where $2\theta$ is the scattering angle. The concentration ranged from 11.4 mg/ml to 1.44 mg/ml for SpuA and from 11.11 mg/ml to 1.39 mg/ml for SpuA-maltotetraose. Background scattering was measured after each protein sample using the buffer solution and then subtracted from the protein scattering patterns after proper normalisation and correction from detector response.

*Scattering Data Analysis.* The values of radii of gyration ($R_g$) were derived from the Guinier approximation [195]: $I(q) = I(0) \exp(-q^2R_g^2/3)$, where $I(q)$ is the scattered intensity and $I(0)$ is the forward scattered intensity. The radius of gyration and $I(0)$ are inferred respectively from the slope and the intercept of the linear fit of $Ln[I(q)]$ *vs* $q^2$ in the *q*-

range *q.*Rg<1.12. The distance distribution function *P(r)* was calculated on the merged curve by the Fourier inversion of the scattering intensity *I(q)* using GNOM [196] and GIFT [197].

*Overall shape of SpuA and SpuA-M4 and compared to the crystal structure:* The low-resolution shape of SpuA fullength as well as its substrate complex SpuA-M4 were determined *ab initio* from the scattering curve using the program GASBOR [198]. This program restores low-resolution shapes of proteins and calculates a volume filled with densely packed spheres (dummy residues of 3.8Å diameter) fitting the experimental scattering curve by a simulated annealing minimisation procedure with a nearest-neighbour distribution constraint. Several independent fits were run with no symmetry restriction and the stability of the solution was checked. The atomic crystallographic structures of the individual modules were then positioned in the envelope using TURBO-FRODO [199] and PyMOL. The P(r) function and the Rg values taking into account the whole scattering curve were calculated using GNOM [196]. For each structural model obtained the theoretical SAXS profile, the $R_G$ and the corresponding fit to the experimental data were calculated using the program CRYSOL [200] (Appendix A).

## 4.4 Results and Discussion:

*SpuA and S. pneumoniae*

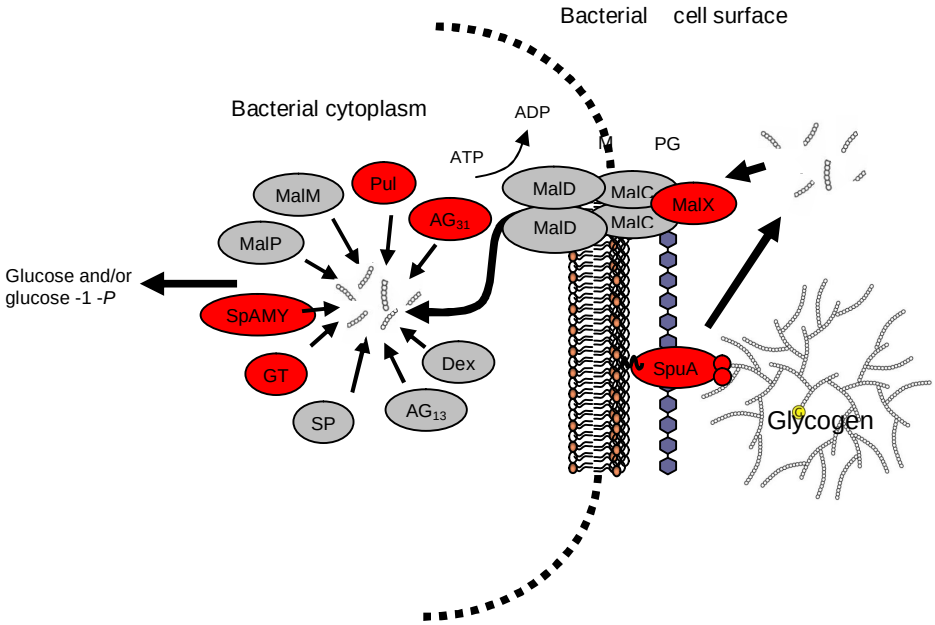
SpuA belongs to glycoside hydrolase family 13, the largest sequence-based glycoside hydrolase family with over 2500 entries from archea, bacteria and eukaryotes. It is known as the α-amylase family but it also includes other α-glucan metabolizing enzymes such as isoamylases, pullulanases, α-glucosidases, sucrases, and cyclomaltodextrinases. Due

to the large number of members of GH13, subfamilies have been proposed within the family to better distinguish members based on sequences and enzymatic properties [201]. SpuA is classified within subfamily 12 which includes pullulanases from firmicutes (bacteria with Gram positive cell wall structure).

SpuA (SP0268) is one of five α-glucan metabolizing proteins known to be essential for the full virulence of *S. pneumoniae*. Others include a second pullulanase (SP1118), a predicted α-amylase (SP1382), a predicted glucanotransferase (SP1121) and a protein involved in maltose transport (MalX). These proteins are part of a *S. pneumoniae* α-glucan metabolism pathway (Figure 45). The genes for maltose transport (*malXCD*) are found on the maltodextrin uptake locus which is induced by the presence of maltose [202] and regulated by the repressor MalR [203]. MalX mediates the transport of maltose; the equivalent in Group A *Streptococcus,* MalE, was shown to be necessary for bacterial colonization of the oropharynx [169]. The MalCD maltose transporter is a member of the ATP binding cassette superfamily similar to that found in *E. coli* [204]. The transport of glucose and maltooligosaccharides by *Streptococcal* species is requires for growth of the organisms in human hosts [169; 205].

SpuA is a known surface associated enzyme as it contains an N-terminal signal peptide and a C-terminal LP*X*TG motif, a motif which is implicated in sortase-mediated attachment of proteins to the peptidoglycan layer [206]. It also has been identified as an immunogenic surface protein in a genomic expression library probed with human convalescent-phase serum and surface localization has been confirmed by immunofluorescence with anti-SpuA [170]. Previous studies of the N-terminal CBM41

Figure 45: α-glucan metabolizing pathway harbored by *S. pneumoniae*. Bioinformatics was used to determine cellular localization of each protein. STM studies showed that enzymes in red are essential for virulence in a mouse lung model [53].



modules from SpuA demonstrated that they effectively bind to chains of glycogen with high affinity ($K_a$ 2 x$10^6$ $M^{-1}$) in a bivalent interaction with two opposing binding sites (see Section 3.3 and [192]). The target cells were shown to be type II alveolar cells in mouse lung tissue that harbor glycogen granules for the production of lung surfactant. ***Our hypothesis is that SpuA breaks down glycogen in these cells by hydrolyzing the α-1,6-linkages where products are then transported into the cell for use as a nutritional source by the bacteria.*** Since SpuA is essential for virulence, the nutrients provided by glycogen must be essential for viability of the organism.

*Glycogen Hydrolysis by SpuA*

Full length SpuA and the catalytic module, SpuAΔCBM, were cloned and produced as recombinant proteins in *E. coli* and the purified protein products were used to identify activity on glycogen. Zymograms were initially employed to observe whether SpuA and SpuAΔCBM demonstrated catalytic activity on α-glucans (Figure 46). The clearing observed in the gels containing polysaccharide substrate indicated that both SpuA and SpuAΔCBM are active on the α-glucans starch, pullulan and glycogen. Subsequent experiments were then performed to assess the products of α-glucan hydrolysis by SpuA. Separation of α-glucan products after a one hour incubation with SpuA by thin layer chromatography (TLC) show that glycogen hydrolysis resulted in a smearing when compared to the no enzyme control (Figure 47A lane 11 and 12), suggesting the formation of multiple products that cannot be resolved using this method. SpuA activity on the control α-glucan, pullulan, resulted in the formation of maltotriose (Glc-α-1,4-

Figure 46: Zymograms of SpuA and SpuAΔCBM. (A) red-pullulan and (B) Glycogen. Lane 1: SpuAΔCBM, Lane 2: SpuA

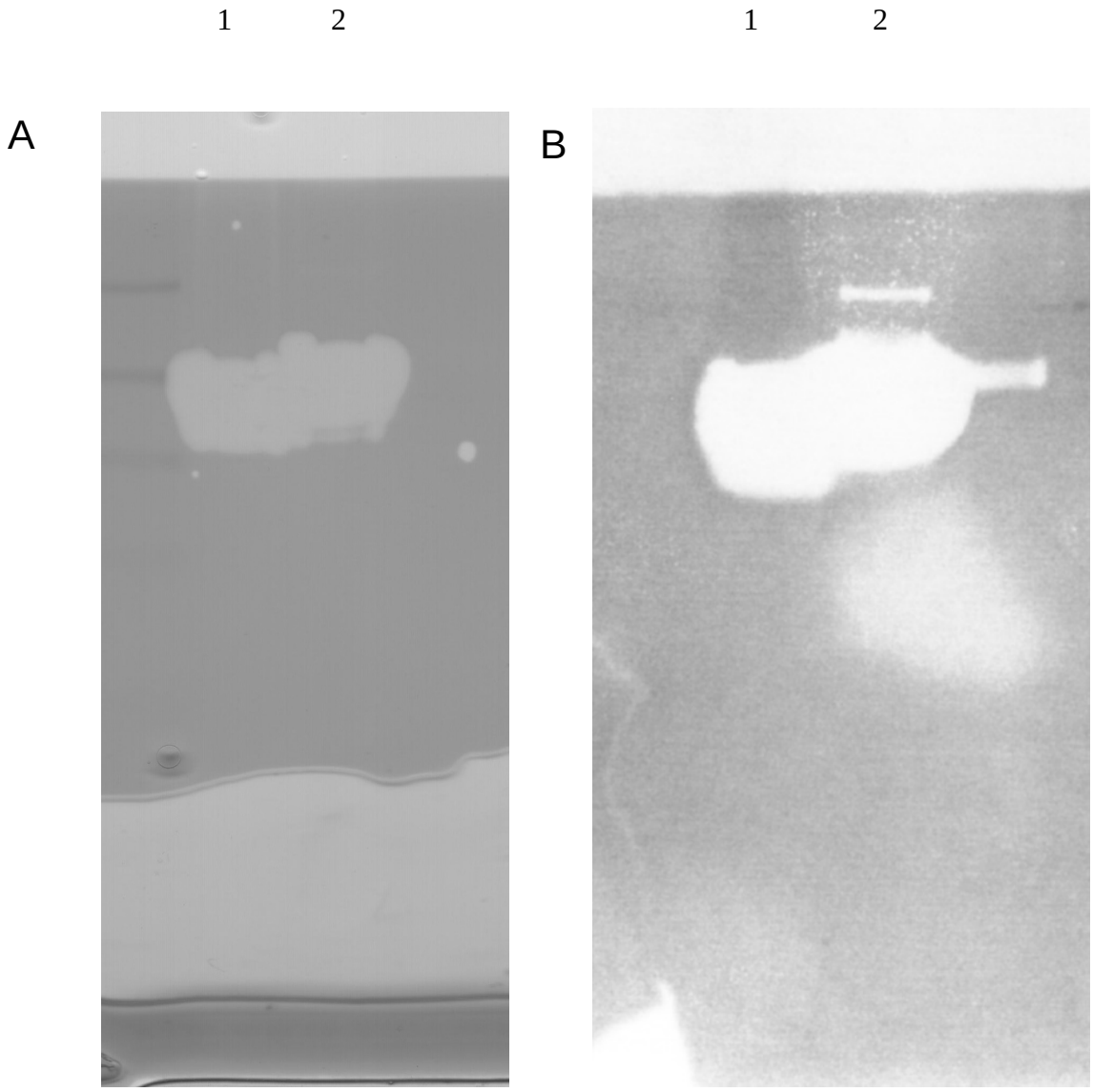
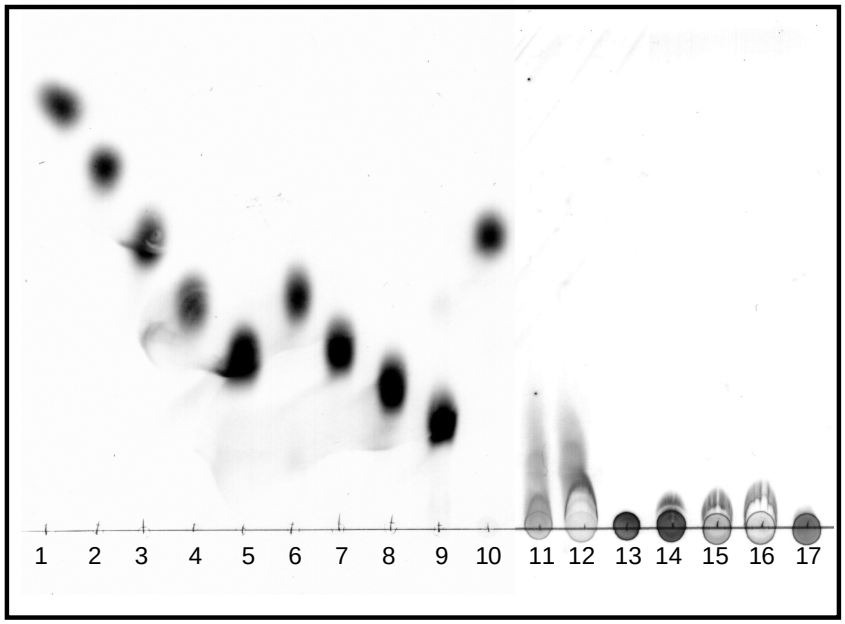
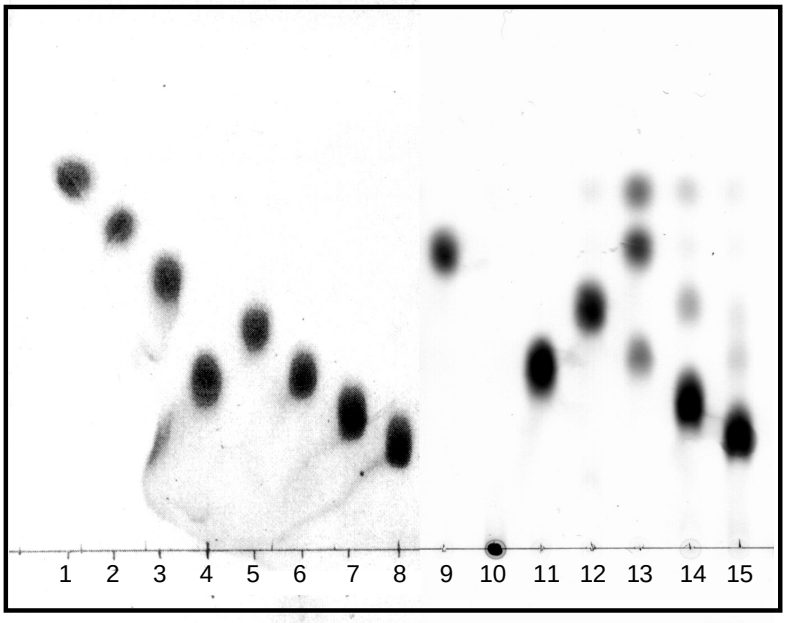


Figure 47: (A) Thin Layer chromatography of SpuA products of α-glucan hydrolysis. 1:glucose; 2: G2; 3: G3; 4: panose; 5: isoG3; 6: G4; 7: G5; 8: G6; 9: G7; 10: SpuA + pullulan; 11 & 12: SpuA + glycogen; 13: SpuA + dextran; 14-17: same as 10-13 without enzyme. (B) TLC of SpuA on maltooligosaccharides: 1:Glucose, 2:G2, 3:G3, 4:isoG3, 5:G4, 6:G5, 7:G6, 8:G7, 9: Pullulan + SpuA, 10: amylose + SpuA, 11: isoG2, 12:G4 + SpuA, 13: G5 + SpuA, 14: G6 + SpuA, 15: G7 + SpuA.

**A**




**B**



Glc-α-1,4-Glc) (Figure 47A). This result indicates the likelihood that SpuA is hydrolyzing α-1,6-linkages in pullulan, a linear polymer of α-1,6-linked maltotriose, since we do not see any products that co-migrate with α-1,6-containing gluco-oligosaccharides such as isomaltose (Glc-α-1,6-Glc) and panose (Glc-α-1,4-Glc-α1,6-Glc). SpuA is not active on amylose (pure α-1,4-linked glucose) or dextran (pure α-1,6-glucose) but had some activity on α-1,4-linked glucooligosaccharides (Figure 47B).

Fluorophore-assisted carbohydrate electrophoresis (FACE) is a high resolution polyacrylamide electrophoresis method which is able to separate oligosaccharides based on size. We were able to resolve the products of glycogen hydrolysis by SpuA and SpuAΔCBM using FACE as a ladder of α-1,4-linked glucans increasing in length by one glucose unit (Figure 48). Comparing the results of TLC and FACE, we believe that SpuA is hydrolyzing the α-1,6-branches in glycogen, producing α-1,4-gluco-oligosaccharides of varying length.

*Structural basis of Glycogen Degradation by SpuA*

The structure of the catalytic module, SpuAΔCBM, was solved to 1.85 Å by molecular replacement using a truncated form of the catalytic module from *Klebsiella pneumoniae* PulA (PDB code 2FHF) as the search model. The structure of SpuAΔCBM and SpnDX (PDB code 2J44) was then used to solve the structure of full length SpuA in complex with maltotetraose to 2.25 Å. SpuA is a large enzyme (140 KDa) with overall dimensions of 108 Å x 90 Å x 54 Å. It has 6 domains comprised of two N-terminal CBM41s (SpnDX-1 and SpnDX-2), a linker region, and the catalytic module (Cat) with flanking N

Figure 48: Products of glycogen breakdown by SpuA resolved by FACE. Lanes 1 & 8: Glucose standards; lane 2: Glycogen untreated; lane 3: glycogen treated with SpuA; lane 4: glycogen treated with SpuAΔCBM; lane 5: pullulan untreated; lane 6: pullulan treated with SpuA; lane 7: pullulan treated with SpuAΔCBM.

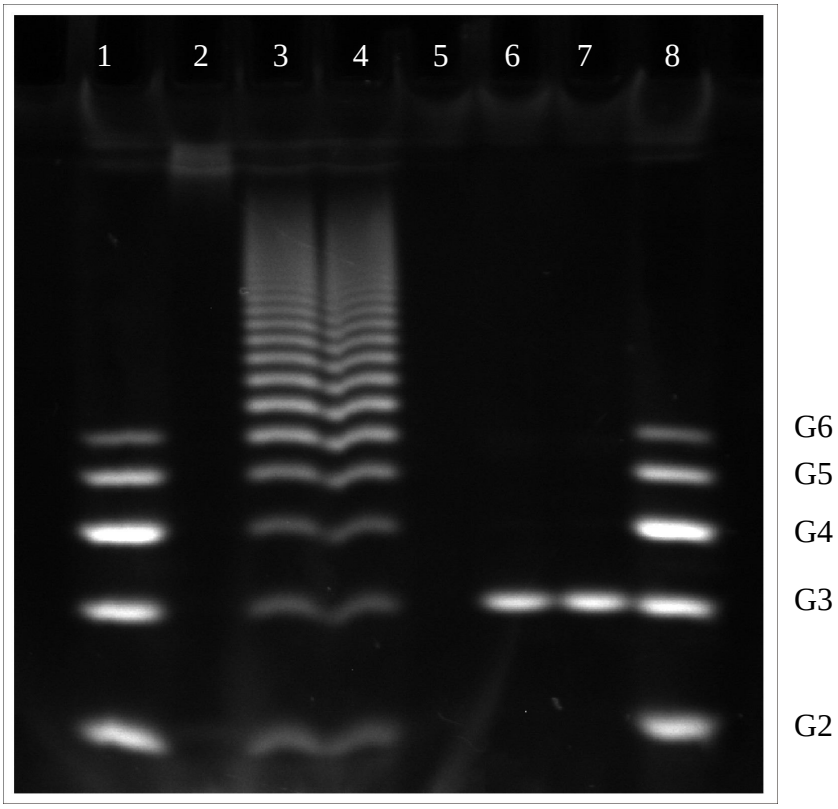
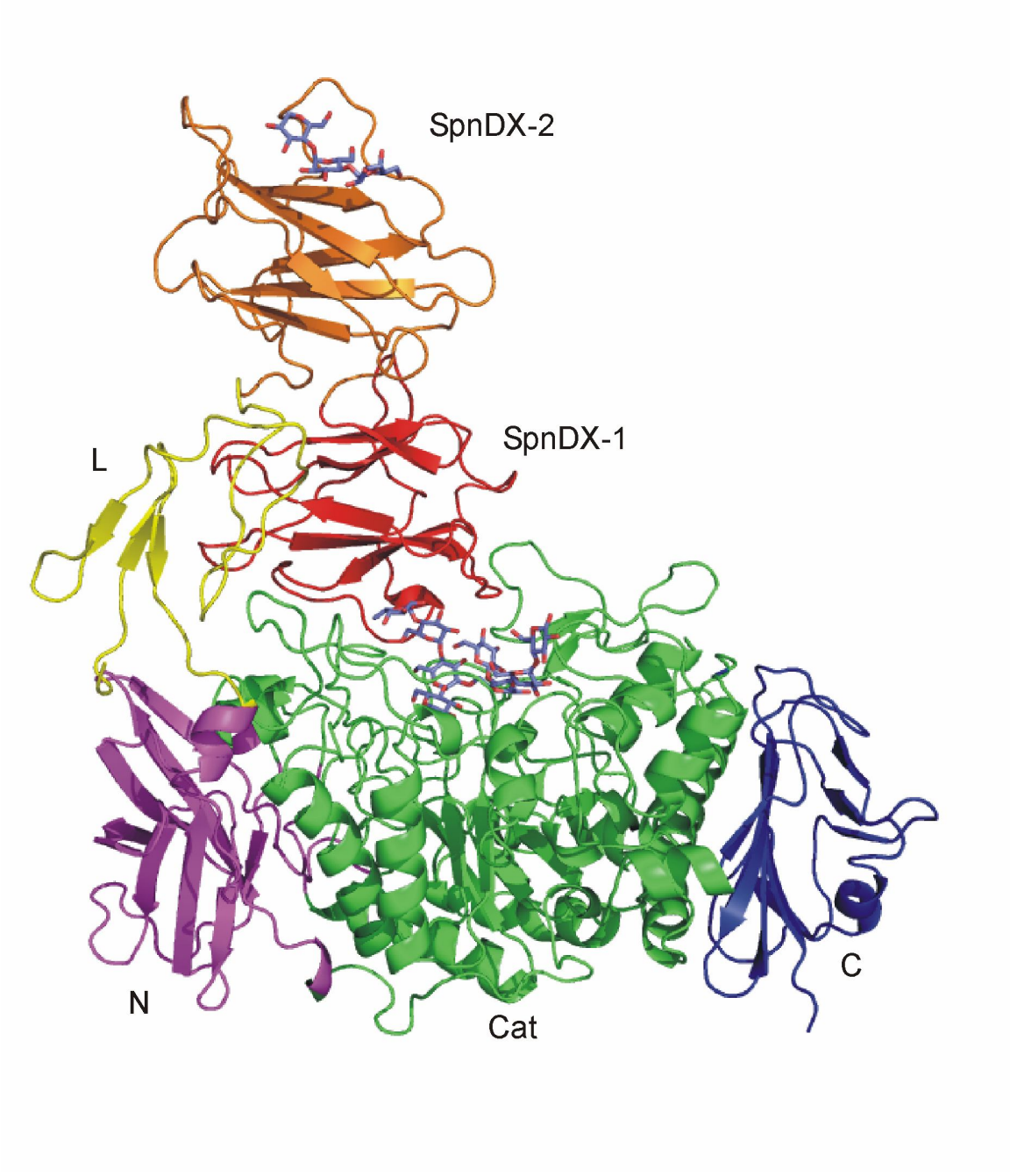


Figure 49: (A) Secondary structure representation of SpuA. N-terminal CBM41s (SpnDX-1, red, and SpnDX-2, orange), linker (L, yellow), N-terminal GH13 β-sandwich domain (N, magenta), catalytic domain (Cat, green) and C-terminal β-sandwich domain (C, deep blue). Maltotetraose in active site and maltotriose in SpnDX-2 shown as sticks (blue)



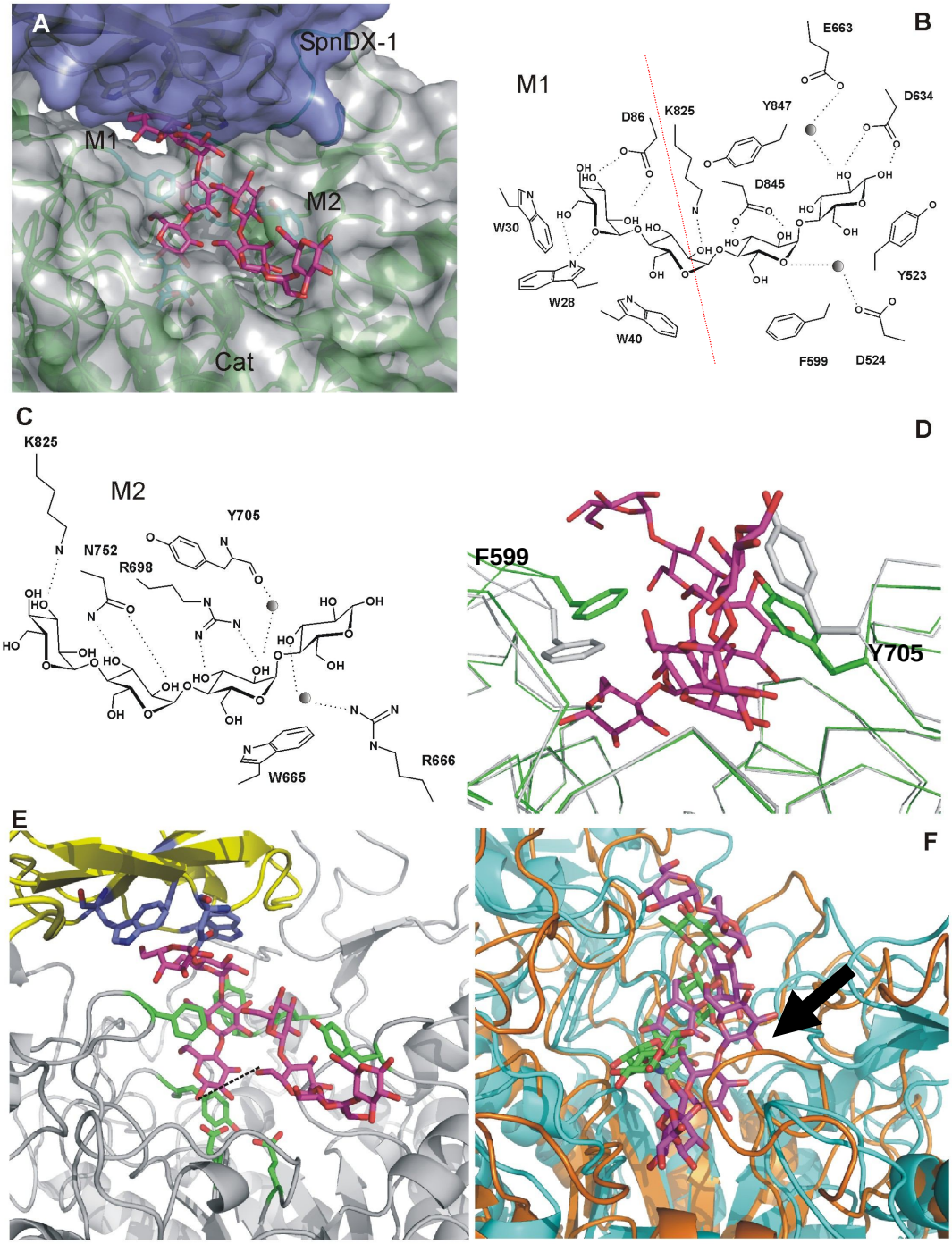
and C-terminal β-sandwich domains (Figure 49). The catalytic module has an $(\alpha\beta)_8$ barrel fold common to family 13 GHs, and the N and C domains are 4 β-strands overlapping 4 β-strands and 4 β-strands overlapping 3 β-strands respectively with an Ig-like topology. The C-terminal domain is common to other enzymes within family 13 GHs, however, the N-terminal GH13 domain appears to be unique to pullulanases and their functions are unknown. SpnDX-1 and SpnDX-2 are family 41 CBMs and have previously been shown to interact with α-1,4-linked glucans and together form a rigid structure optimal for binding chains of glycogen [192]. Family 41 CBMs are uniquely associated with pullulanases within the family 13 GHs.

Within the active site of SpuA are two maltotetraose sugars where notably the first two glucose residues of maltotetraose 1 (M1) is bound by SpnDX1 and the catalytic module (Figure 50A); hydrophobic stacking interactions occur between GLC1 and SpnDX-1 W30 and GLC2 and SpnDX-1 W40, while O2 and O3 of GLC1 hydrogen bond with SpnDX-1 D86 (Figure 50B). In the active site of the catalytic module GLC3 is sandwiched between F599 and Y847 and forms various direct and indirect hydrogen bonds while GLC4 interacts with the side chain of Y523. Maltotetraose 2 (M2) interacts solely with the catalytic module and forms stacking interactions and direct hydrogen bonds with the first 3 glucose sugars (GLC5, GLC6 and GLC7) and only a water mediated hydrogen bond with O3 of GLC8 (Figure 50C). K825 is the only side chain that interacts with both M1 and M2, forming direct hydrogen bonds with GLC2 O2 and GLC5 O3 hydroxyl groups. When we compare the native GH13 module with SpuA in complex with maltotetraose, we see a shift in loops where F599 moves ~3 Å and Y708 moves ~3 Å to accommodate the sugar residues (Figure 50D). Both maltotetraose units

lie parallel to each other such that O1 of the reducing sugar GLC4 would form an α-1,6-linkage with O6 of GLC6 as would be found in glycogen (Figure 50E) such that the SpuA active site is designed to accommodate the branched sugar residues. If we compare the structure of SpuA GH13 with human pancreatic α-amylase (HPA) in complex with acarbose, we see that HPA has an extended loop comprised of amino acids 303-310 which occupy the space of M2 in SpuA (Figure 50F). This differentiates SpuA from other family 13 α-glucan degrading enzymes such as amylases and isoamylases.

The linker region joining the SpnDX modules and the catalytic GH13 module was difficult to model due to flexibility within this region as indicated by high B-factors. In the crystal structure of the pullulanase from *K. pneumoniae* in complex with maltotetraose, the linker region also had high B-factors and the CBM41 module was actually bound to maltotetraose occupying site 1 of the adjacent symmetry-related molecule [161]. It is likely this flexibility that interfered with our inability to crystallize native SpuA. It is therefore possible that the native SpuA enzyme may take on a different conformation than SpuA in complex with maltotetraose. In order to look at the overall shape of native SpuA, small angle X-Ray scattering (SAXS) experiments in the absence and presence of maltotetraose were performed in collaboration with Dr. Mirjam Czjzek from Station biologique de Roscoff. SAXS is a medium resolution technique that generates an envelope corresponding to the overall shape of a protein at ~15-20Å. The benefit of SAXS is that it uses protein in solution and does not require protein crystals.

Figure 50: (next page) (A) Space filling model of maltotetraose in the active site bound by SpnDX-1 (blue) and catalytic site of GH13 (Cat, gray) with M4 in magenta. (B and C) Graphical representation of two maltotetraose units in the active site, K825 shared by both units. (B) M1, where red dashed line shows the contribution of SpnDX-1(left) and GH13 (right), (C) M2. (D) Native (gray) versus complexed (green) GH13 active site showing shift in loops upon substrate binding. (E) Potential α-1,6-linkage shown by a dashed line in active site of GH13 (F) Close-up of HPA and SpuA active sites. SpuA cyan with G4 in magenta, HPA in orange with acarbose in green. Extended loop region in HPA indicate with arrow indicating the loop region in HPA that occupies the site of M2 in SpuA.


A
SpnDX-1
M1
M2
Cat
B
M1
E663
D634
Y847
D86
K825
D845
W30
W28
W40
Y523
F599
D524
C
K825
M2
Y705
N752
R698
R666
W665
D
F599
Y705
E
F

Once the overall shape of the protein is known, we can then fit the known crystal structure of SpuA into the envelope (Figure 51) and observe if there are any changes in overall shape of the protein in the absence of substrate. The SAXS results show that the overall surface structure of the native SpuA had a slightly altered shape when compared to SpuA in complex with maltotetraose (Figure 51 and see Appendix A). There appears to be a shift in the linker region towards the active site (Figure 51 c and f) suggesting that the linker may be involved in substrate binding. Because the active site must accommodate glucan chains of varying length found in glycogen granules, it is possible that the slight flexibility of the linker region may assist SpuA in this process for efficient hydrolysis the α-1,6-branches within glycogen.

According to GH nomenclature, M1 occupies subsites -4 to -1 and GLC6, 7 and 8 of M2 occupies subsites +1 to +3, where cleavage of the α-1,6-linkage would occur between -1 and +1 (GLC4 and GLC6). A notable feature is that the -4 and -3 subsites are provided by SpnDX-1. At the cleavage site, the residues D634 and E663 are present which are the conserved catalytic residues in all other GH13 family members (Figure 52) [207 161]. D634, the catalytic nucleophile, forms direct hydrogen bonds with O1 and O2 of GLC4 in the -1 subsite and E663, the catalytic acid/base, forms a water mediated hydrogen bond with GLC4 O2 and GLC6 cyclic oxygen in the +1 subsite. SpuA likely cleaves the α-1,6-linkage via a retaining mechanism. To study the importance of D634 and E663 in glycogen hydrolysis we generated alanine substitution mutants in full length SpuA to generate SpuAD634A and SpuAE663A. Both residues D634 and E663 were shown to be essential for glycogen breakdown using FACE analysis for full length SpuA (Figure 53).

Figure 51: Averaged surfaces obtained by different GASBOR runs for SpuA (a,b,c) and SpuA-M4 (d,e,f).

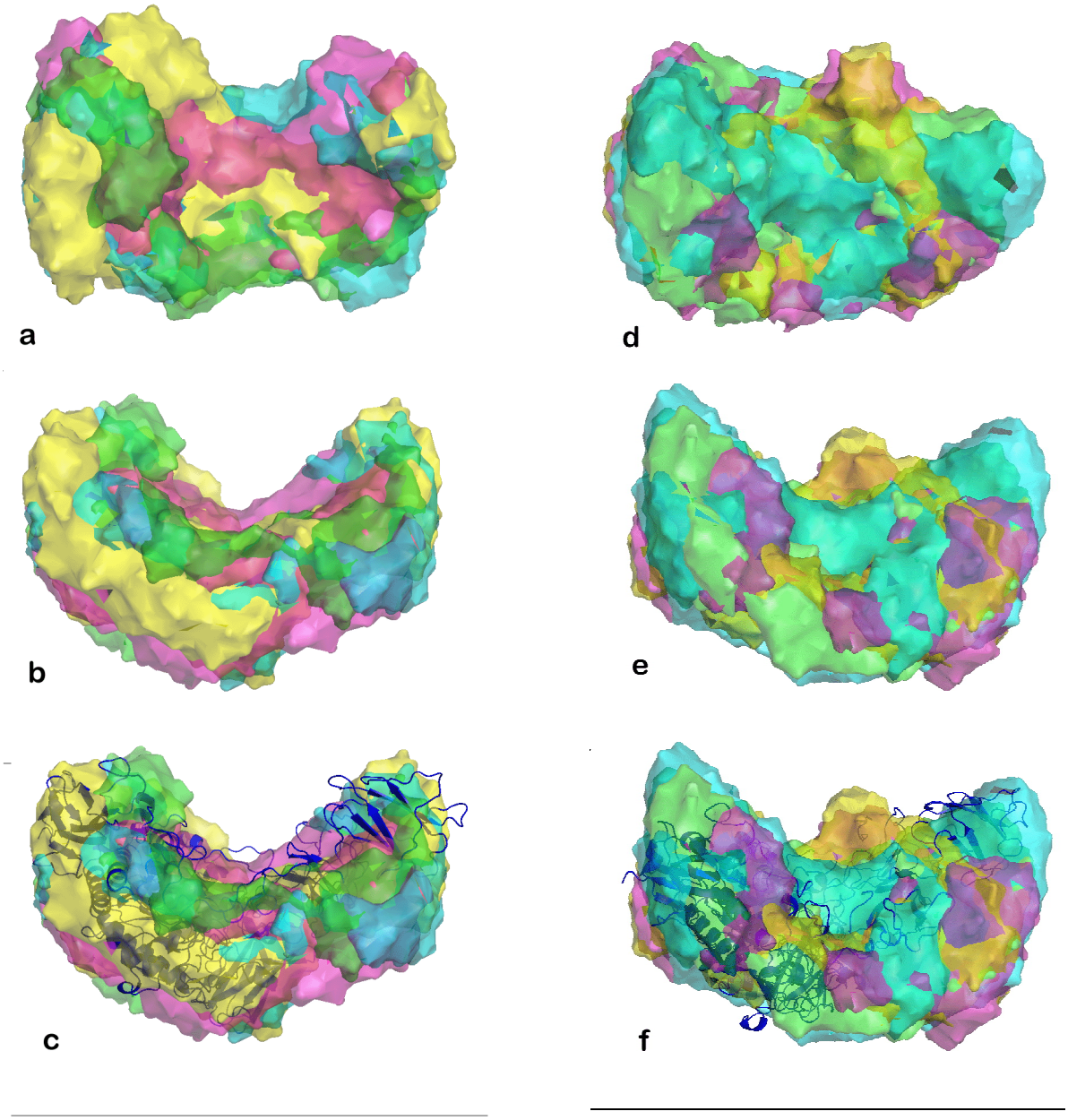


Figure 52: Amino acid sequence alignments of the SpuA catalytic module with other family 13 GHs whose structures are known and catalytic residues have been identified. Catalytic residues are indicated with a blue arrow.

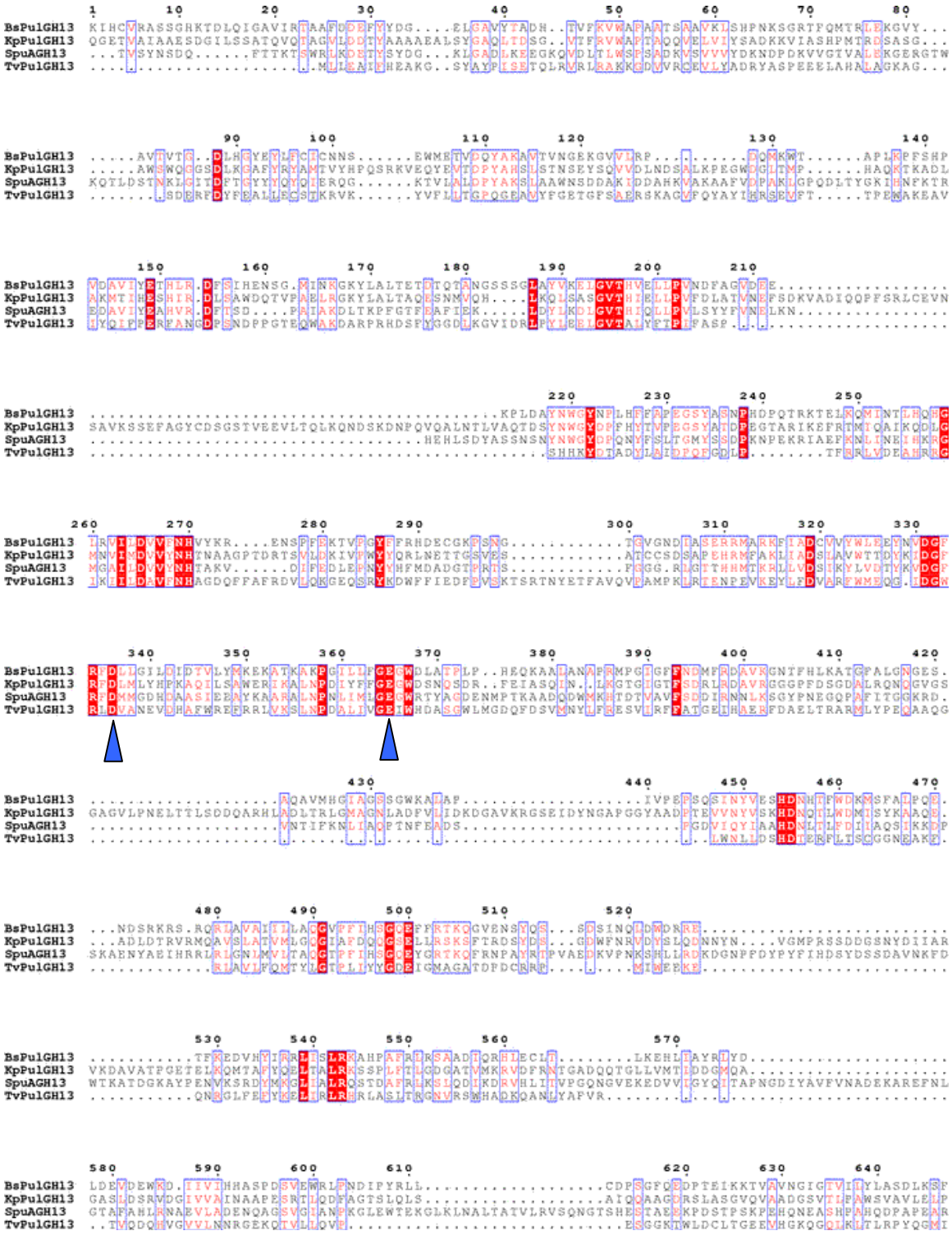
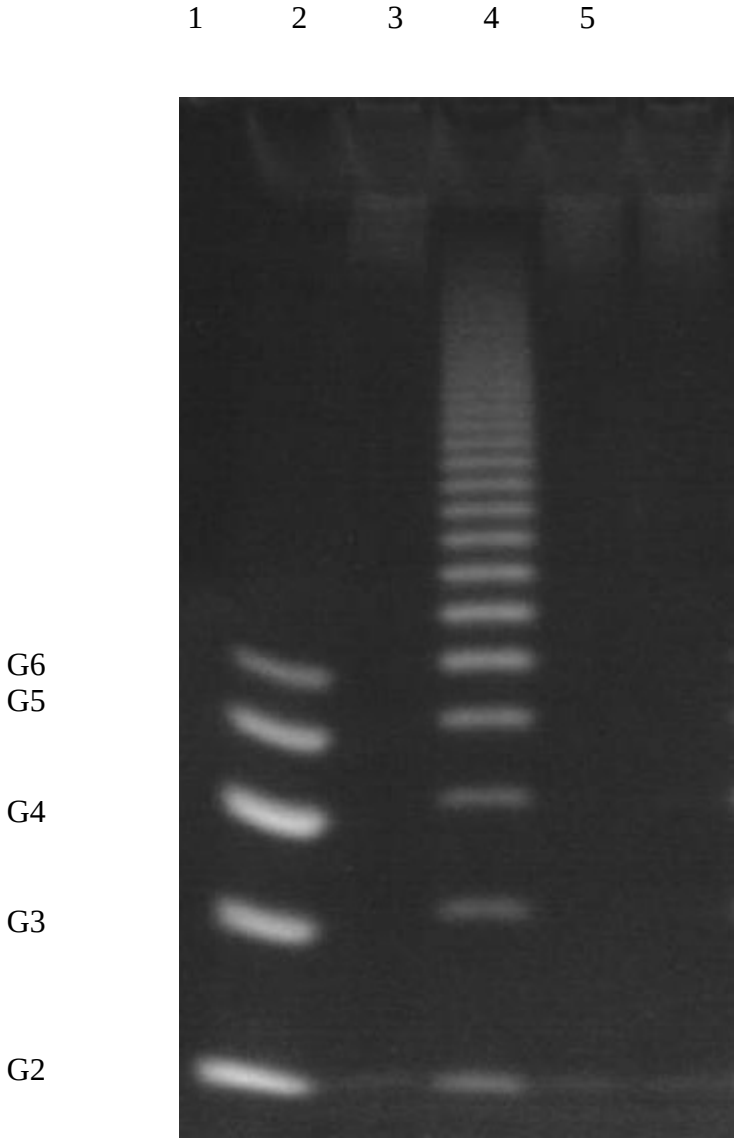


Figure 53: FACE of SpuA catalytic mutants D634A (catalytic nucleophile) and E663A (catalytic Acid/Base) on glycogen. Lane 1: glucose Standards, lane 2: no enzyme control, lane 3: wildtype SpuA, lane 4: SpuA D634A, lane 5: SpuA E663A.



*Comparison with other pullulanase structures*

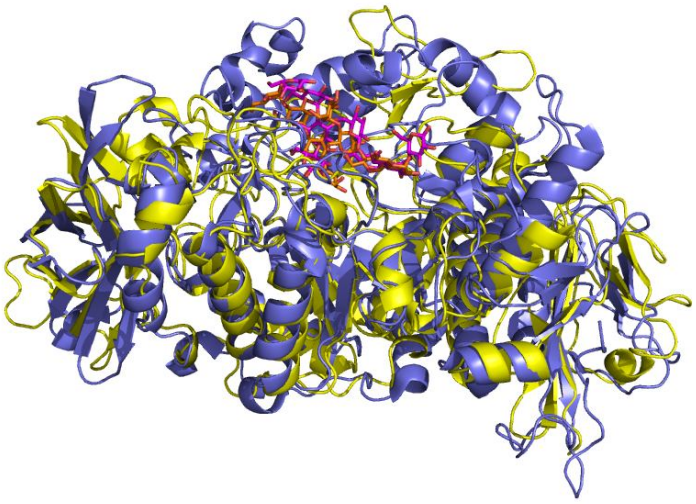
SpuA is very similar to the pullulanase PulA from *Klebsiella pneumoniae*, *Kp*PulA, whose structure was recently solved [161]. The GH13 modules share 29.9% sequence identity and have an RMSD of 1.89 Å over 593 matched Cα atoms (Figure 52 and Figure 54). *Kp*PulA only has one N-terminal CBM41 which shares 16% and 17.5% sequence identities and RMSDs of 1.97 Å and 2.11 Å with *Spn*DX-1 and *Spn*DX-2, respectively. A novel feature of this enzyme is *Spn*DX-1 faces the active site of the adjoining catalytic module, participating in positioning the substrate in the active site (Figure 51A). This is the first example of a CBM that participates in binding substrate within the active site. The CBM41 from *Kp*PulA faces the active site of the catalytic module in a symmetry related molecule within the crystal structure, however it appears to be an artifact of crystal packing. Interestingly, KpPulA utilizes one N-terminal CBM41 while SpuA has evolved the use of a second CBM41 which together have formed a rigid entity with opposing binding sites. This second CBM41 in SpuA is likely an evolutionary advantage towards binding glycogen branches over linear α-glucans. Also, because of its bivalent architecture, the tandem CBM41 modules have a much higher affinity for glycogen chains over the α-1,4-glucan products that would be formed by the debranching activity of the SpuA enzyme.

*Inhibition studies of SpuA*

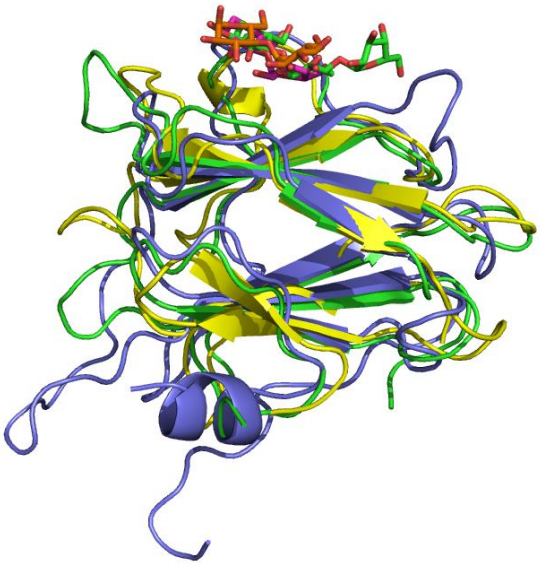
Preliminary inhibition studies on SpuAΔCBM were performed with known α-amylase/glucosidase inhibitors including acarbose, miglitol, voglibose and glucopyranosyl-moranoline (GPM) (Fig 55). Acarbose, miglitol and voglibose are used in

Figure 54: (A) *Spu*A (GH13 in yellow and G4 in magenta) and *Kp*PulA (GH13 in light blue, G4 in orange).(B) CBM41 overlap. *Kp*CBM41 (light blue,G2 in orange), *Spn*DX-1(yellow, G2 in magenta) and *Spn*DX-2 (green, G3 in green).

A



B



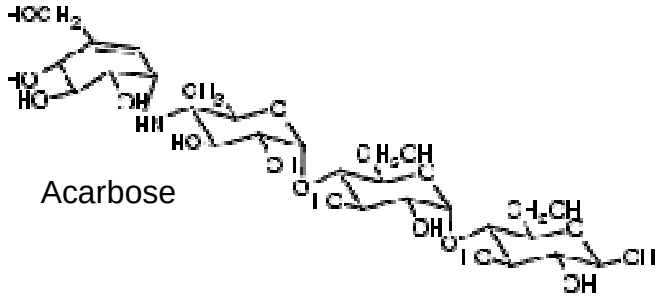
the treatment of type II diabetes and work by competitively inhibiting α-glucosidases and α-amylases in the GI tract to prevent the breakdown of sugars and slow the appearance of glucose in the bloodstream [208]. GPM is similar to the glucosidase inhibitor deoxynojirimycin with an additional non-reducing α-1,4-linked glucose. Table 15 shows the results of these inhibitors on SpuAΔCBM where miglitol and voglibose showed no specificity for SpuAΔCBM while acarbose and GPM had $K_D$'s in the mM range ($10^{-4}$ M). This demonstrates that inhibitors consisting of at least a disaccharide are required to interact with the active site of the catalytic module. However because of the low affinity interaction between SpuAΔCBM and either GPM or acarbose, none of these inhibitors are considered effective. The reason these are not effective inhibitors can be explained by looking at the active sites of SpuA and human pancreatic α-amylase (HPA) in complex with acarbose (Figure 50F). Structural overlap of SpuA and HPA shows that a loop region in HPA creates a linear-shaped active site for α-1,4-glucans. This loop is absent in SpuA where instead a second site is occupied by M2, creating an active site that is able to accommodate branched substrates. This region is what differentiates amylases from pullulanases and explains why alternative branched inhibitors may be more suitable for inhibiting the activity of SpuA.

Because SpuA presents a substrate binding surface for branched glucans, we then tried a branched hemithiomaltodextrin inhibitor (HTMD, Figure 55), a potential inhibitor for starch-debranching enzymes [209]. The interaction of HTMD with GH13 has a 100-fold increase in $K_D$ for SpuAΔCBM ($10^{-6}$ M). Therefore it appears that we are able to more effectively inhibit SpuA with branched inhibitors such as HTMD. Because inhibitors like

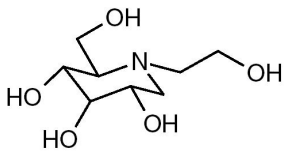
acarbose are very effective at inhibiting α-glucosidases, such as human pancreatic α-amylase, but not SpuA, this could serve as a basis for designing more effective branched inhibitors to selectively target SpuA without targeting other classes of GH13-containing enzymes.

All of the inhibitors discussed are sugar analogs that mimic the shape and charge of the transition state and act as competitive inhibitors such that it blocks the active site, preventing the enzyme from interacting with its intended substrate. Another possible method is to use peptide-based competitive inhibitors as a means of blocking enzyme activity. An interesting feature of the SpuAΔCBM native crystal structure was that crystal packing was facilitated by the C-terminal tail which interacts with the active site of the adjacent molecule. Closer inspection of the structure shows that the peptide VSENGTSHESTA interacts with the same amino acid side chains as the maltotetraose sugars (see Figure 50 and Figure 56A). Notably, the catalytic residues D634 and E663 hydrogen bond with the imidazole ring of H8 and R698 hydrogen bonds with E9. Both H8 and E9 fit nicely into the pockets of the binding site formed by F599 and Y705 normally occupied by both maltotetraose sugars, as seen in a model showing solvent-accessible surface area (Figure 56B). This phenomenon in SpuAΔCBM may assist in the development of potential polypeptide-based inhibitors which, to our knowledge, has yet to be explored as a method of inhibition in glycoside hydrolases. Peptide-based carbohydrate mimicry has been successful in inhibiting the T-cell proliferative activity of the lectin ConA [210], thus the potential for GH peptide-carbohydrate mimicry is feasible.

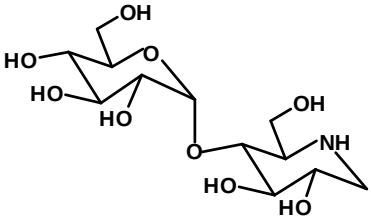
Figure 55: Structure of the transition state α-glucosidase inhibitors acarbose, miglitol, voglibose, GPM and the branched inhibitor HTMD.



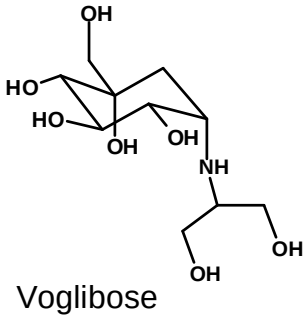
Acarbose



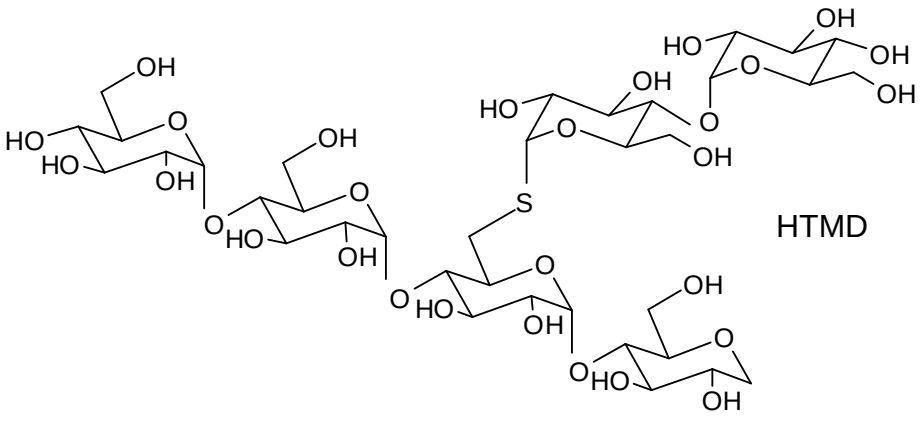
Miglitol



Glucopyranosylmoranoline



Voglibose



HTMD

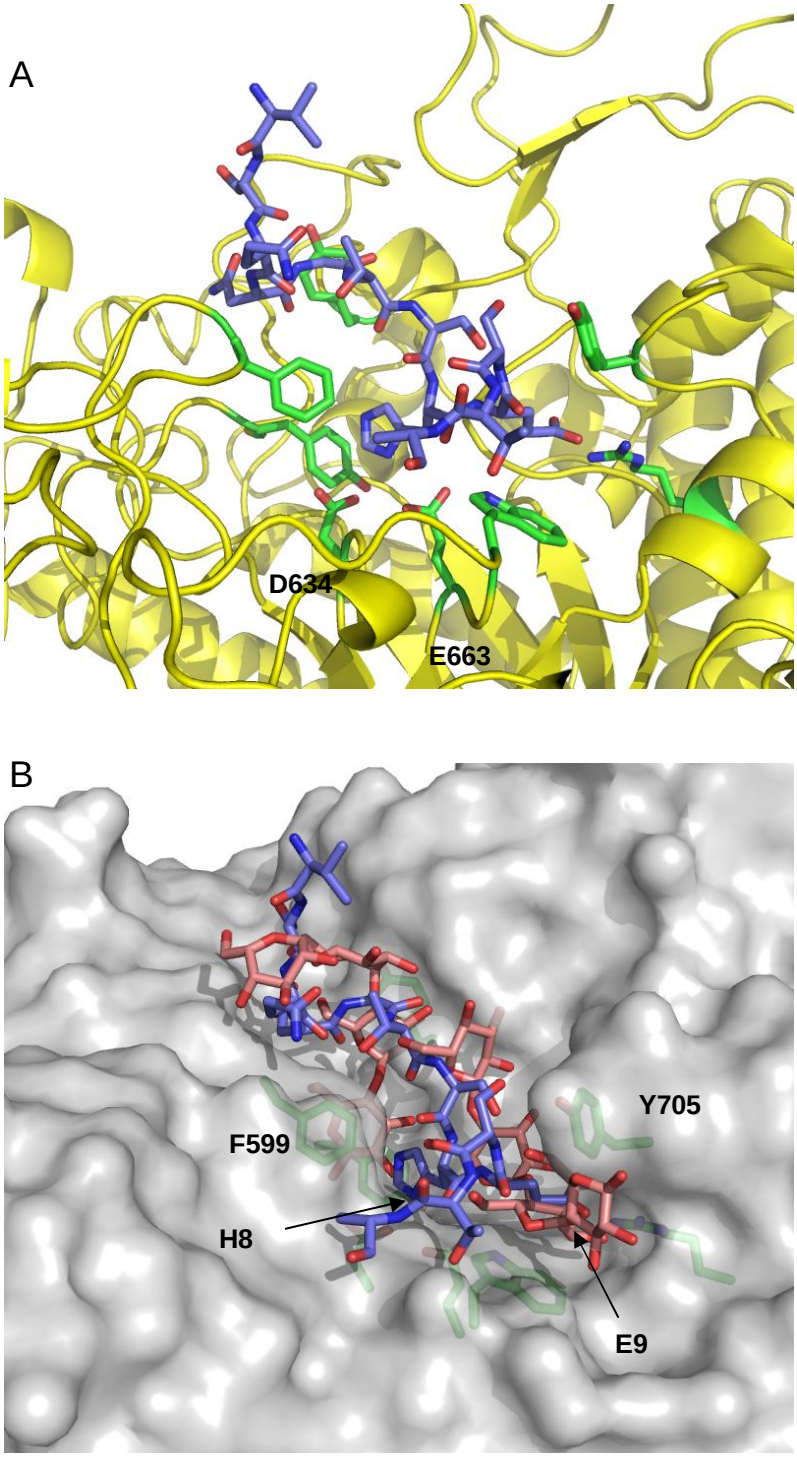
Table 15: ITC results of inhibitors on SpuA GH13

| | n | $K_D$ (M) | ΔH (cal/mol) | ΔS |
|---|---|---|---|---|
| miglitol | *nd** | | | |
| voglibose | *nd* | | | |
| acarbose | 1** | $2.87 \times 10^{-4}$ | -8674 | -12.9 |
| glucopyranosyl-moranoline (GPM) | 1** | $2.50 \times 10^{-4}$ | -5361 | -1.50 |
| hemithiomalto-dextrin (HTMD) | 1.31 (+/- 0.13) | $7.58 \times 10^{-6}$ | -11290 | -14.4 |

* *nd* no interaction detected

** n value was set as 1 in data analysis

Figure 56: Prospect of peptide-based inhibitors based on structure of native SpuAΔCBM. (A) C-terminal tail of SpuAΔCBM makes same contacts with amino acid side chains as maltotetraose (See figure 51A) (B) Solvent accessible surface area representation of GH13 with G4 in SpuA complexed structure (pink) overlapped with peptide from GH13 native structure (blue). Surface representation of active site in gray with residues shown in green.



*Conclusion:*

Previous studies demonstrated that the homologous enzyme PulA from *S. pyogenes* was inactive on glycogen [152]. Our results show that SpuA is active on glycogen in addition to pullulan. Because of the similarity between the two enzymes (73% sequence similarity, 57% sequence identity) one would expect that PulA would also be active on glycogen suggesting its lack of activity on glycogen may be due to the detection method used by Hytonen *et. al.* Due to the host environment in which SpuA is located, it is likely that its target substrate is glycogen and not pullulan Thus SpuA and other related pullulanases such as *S. pyogenes* PulA and those from other pathogenic bacteria, should be known as glycogenases. Like pullulases, glycogenases are active on α-1,6-branches; however, unlike pullanases, glycogenases mainly target branches in glycogen granules. Interestingly, these pullulanases from pathogenic bacteria cluster into the same subfamily in a phylogenetic tree analysis, suggesting a similar activity for these enzymes [201]

SpuA is a potential target for inhibition as an alternative to antibiotic treatment in *S. pneumoniae* infections because it is a virulence factor that is known to be essential for viability of the bacteria in its host [53]. It is also continuously present during all growth phases of *S. pneumoniae* and is identifiable in 41 strains of 17 distinct serotypes making SpuA serologically highly conserved [170]. This study on the structure and activity of SpuA provides valuable information in identifying its role in pneumococcal pathogenesis. We previously identified glycogen granules within type II alveolar cells as a target for the N-terminal CBM41 modules (SpnDX) from SpuA. SpuA then breaks down glycogen in to smaller maltooligosaccharide fragments by hydrolyzing the α-1,6-branch points and also

has residual activity on α-1,4-glucans. The smaller maltooligosaccharides then likely interact with the extracellular MalX for transport into the cell via the ATP-dependent MalC/MalD transporter [169; 205]. Within the bacterial cell it is further broken down into glucose to be used as a nutritional source by the bacteria which is essential for growth of the organism [205]. The lung is a main point of entry for *S. pneumoniae* and the nutrition provided by glycogen within lung tissue must be essential for cell viability as indicated by the importance of SpuA and other α-glucan metabolizing enzymes in bacterial virulence. Furthermore, by targeting glycogen in the lung, SpuA would inhibit the process of lung surfactant production. Lung surfactant is composed mainly of phosphatydilcholine which is synthesized from stores of glycogen granules housed in type II alveolar cells. Surfactant serves as a protective coating on the surface of lung epithelium to prevent collapse of the alveoli during respiration. It also houses many of the protective cells involved in the innate immune response, such as neutrophils and macrophages, since the respiratory tract is a main point of entry for many pathogenic bacteria, viruses and other foreign objects. By preventing lung surfactant synthesis, *S. pneumoniae* would promote its evasion of this first line of defense, allowing the organism to infect lung tissue. Inhibiting SpuA activity may be useful as a means of treatment to stop or slow the progression of infection. Because of its unique catalytic site architecture, we may be able to design inhibitors that selectively inhibit SpuA activity without compromising human α-glucosidases. The ability to combat such infections by alternative therapies, such as targeting SpuA, is increasingly critical with the rise in antibiotic resistant strains.

# Chapter 5: Global Conclusions

The overall goal set out in this thesis was to advance our understanding of the molecular determinants of carbohydrate recognition by carbohydrate-binding modules that belong to the same amino acid sequence based family. By studying multiple members within a CBM family we have been able closely examine the forces that drive carbohydrate recognition. In general, carbohydrates have a predetermined three-dimensional shape based on the configuration of the sugar and the type of glycosidic linkage joining the sugar molecules. CBMs have evolved binding sites that are complementary to the three-dimensional shape of the interacting sugar. Most importantly, these binding sites are preformed, that is, they do not undergo any conformational changes upon ligand binding. This preformed binding site is critical for all known protein-carbohydrate interactions, including those involving antibodies and lectins ([211; 212]), and is what allows for high affinity interactions with specific carbohydrate ligands. The preformed binding pocket also allows for localization of the protein at specific regions within carbohydrates. Looking closer into the preformed binding pocket, the main driving force for CBM-carbohydrate recogntion is through aromatic amino acids within the binding pocket. They are important for forming hydrophobic stacking interactions with sugar monomers Almost all protein-carbohydrate interactions involve at least one aromatic amino acid side chain, which demonstrates the importance of these residues to protein-carbohydrate recognition. Other key interactions include direct and water-mediated hydrogen bonding between amino acids in the binding site and sugar hydroxyl groups and/or the free electron pairs of the cyclic oxygen.

Specific research on polysaccharide recognition by family 6 CBMs showed us that members utilize amino acid side chains in order to alter the topology of the binding site, thus allowing for the recognition of a diverse range of plant polysaccharides. The dominant driving forces were hydrophobic stacking interactions between the primary sugar and aromatic amino acid side chains. Further molecular determinants were based on the length of the interacting sugar and whether binding occurrs at terminal sugars or within a polysaccharide chain which were facilitated by amino acid "hotspots" within the binding site. The research on family 6 CBMs may be useful in predicting the interaction of an unknown CBM6s with its potential ligand based on its amino acid sequence. It also has potential applications in industry for designing CBMs with enhanced binding affinities for plant biomass. This may be important as the need for alternative fuel sources increases. Currently work is being done to synthesize "cellulosic ethanol", fuel from plant material waste, such as corn stalks, wheat straw and forest trimmings [213] which contains a diverse amalgam of cellulose, hemicellulose and lignin which need to be efficiently hydrolyzed. By enhancing degradative enzymes with CBMs, we may create more efficient processing of cellulosic material for ethanol production.

α-Glucan recognition by family 41 CBMs are driven by a preformed concave binding surface composed of aromatic amino acid residues to accommodate the helical convex structure of α-1,4-linked glucose. *Tm*CBM41 has an additional binding subsite for interacting with α-1,6-linked glucose found within pullulan. This separated it from other starch binding CBM families which do not interact with α-1,6-linkages. Furthermore, SpnDX and SpyDX tandem CBM41 modules adopt a bivalent scaffold mediated by a hydrophobic interface to specifically interact with opposing α-glucan

chains in glycogen. This novel feature serves as an advantage towards binding intact substrate over the products formed by the glycogen debranching activity of the enzyme. Research on CBM41s may be potentially beneficial in the pharmaceutical industry. Because CBM41 modules are only found in pathogenic bacteria, these modules may serve as potential vaccine candidates in preventing many harmful bacterial infections, including those caused by *Streptococcal, Klebsiella* and *Streptomyces* species.

Another goal of this thesis was to observe branched α-glucan binding by a bacterial pullulanase. Similar to the preformed binding sites provided by CBMs, SpuA has a preformed active site specific for a branched α-glucan substrates with two concave subsites that allow for interaction with two α-1,4-linked glucan chains. Subsite 1 interacts with the non-reducing end of the glucan chain, blocked by an aspartic acid residue, which is O6-linked to the α-glucan chain in subsite 2. Again, we see that aromatic amino acid side chains are important for interacting with α-glucans. Although the active site is preformed to interact with branched substrates, we saw that slight conformational changes occur in aromatic amino acids involved in interacting with the concave face of the α-1,4-linked-glucans. A unique feature revealed in the structure of SpuA was the direct participation of SpnDX-1 in binding substrate within subsite 1 of the active site. This feature has never before been identified in CBMs and possibly allows for this class of enzyme to efficiently stabilize the branched substrate within the active site. Studies on SpuA identifies a potential target to combat antibiotic resistance in *Streptococcus pneumoniae* and possibly other pathogens that also harbor a surface associated pullulanase. By inhibiting the activity of SpuA, we may impede the progression of

infection of antibiotic resistant strains and providing time to treat the infection by other means.

## References


1. Duchesne, L. C., Larson, D.W. (1989). Cellulose and the Evolution of Plant Life. *BioScience* **39 (4)**, 238-241.

2. Besra, G. S. & Brennan, P. J. (1997). The mycobacterial cell wall: biosynthesis of arabinogalactan and lipoarabinomannan. *Biochem Soc Trans* **25**, 845-50.

3. Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., Deboy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., Brinkac, L. M., Dodson, R. J., Rosovitz, M. J., Sullivan, S. A., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, M. L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K. J., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Kasper, D. L., Telford, J. L., Wessels, M. R., Rappuoli, R. & Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**, 13950-5.

4. Hammerschmidt, S., Wolff, S., Hocke, A., Rosseau, S., Muller, E. & Rohde, M. (2005). Illustration of pneumococcal polysaccharide capsule during adherence and invasion of epithelial cells. *Infect Immun* **73**, 4653-67.

5. van de Rijn, I., Bernish, B. & Crater, D. L. (1997). Analysis of hyaluronic acid capsule expression in group A streptococci. *Adv Exp Med Biol* **418**, 965-9.

6. Husmann, L. K., Yung, D. L., Hollingshead, S. K. & Scott, J. R. (1997). Role of putative virulence factors of Streptococcus pyogenes in mouse models of long-term throat colonization and pneumonia. *Infect Immun* **65**, 1422-30.

7. Meisen, S., Wingender, J. & Telgheder, U. (2008). Analysis of microbial extracellular polysaccharides in biofilms by HPLC. Part I: development of the analytical method using two complementary stationary phases. *Anal Bioanal Chem* **391**, 993-1002.

8. Garcia-Medina, R., Dunne, W. M., Singh, P. K. & Brody, S. L. (2005). Pseudomonas aeruginosa acquires biofilm-like properties within airway epithelial cells. *Infect Immun* **73**, 8298-305.

9. Banas, J. A. & Vickerman, M. M. (2003). Glucan-binding proteins of the oral streptococci. *Crit Rev Oral Biol Med* **14**, 89-99.

10. Imberty, A. & Perez, S. (1989). Conformational analysis and molecular modelling of the branching point of amylopectin. *Int J Biol Macromol* **11**, 177-85.

11. Campbell, B. S., Siddique, A. B., McDougall, B. M. & Seviour, R. J. (2004). Which morphological forms of the fungus Aureobasidium pullulans are responsible for pullulan production? *FEMS Microbiol Lett* **232**, 225-8.

12. Igarashi, T., Morisaki, H., Yamamoto, A. & Goto, N. (2002). An essential amino acid residue for catalytic activity of the dextranase of Streptococcus mutans. *Oral Microbiol Immunol* **17**, 193-6.

13. Wolski, E., Maldonado, S., Daleo, G. & Andreu, A. (2007). Cell wall alpha-1,3-glucans from a biocontrol isolate of Rhizoctonia: immunocytolocalization and relationship with alpha-glucanase activity from potato sprouts. *Mycol Res* **111**, 976-84.

14. Dube, D. H. & Bertozzi, C. R. (2005). Glycans in cancer and inflammation--potential for therapeutics and diagnostics. *Nat Rev Drug Discov* **4**, 477-88.

15. Slawson, C., Housley, M. P. & Hart, G. W. (2006). O-GlcNAc cycling: how a single sugar post-translational modification is changing the way we think about signaling networks. *J Cell Biochem* **97**, 71-83.

16. Slawson, C. & Hart, G. W. (2003). Dynamic interplay between O-GlcNAc and O-phosphate: the sweet side of protein regulation. *Curr Opin Struct Biol* **13**, 631-6.

17. Kamemura, K. & Hart, G. W. (2003). Dynamic interplay between O-glycosylation and O-phosphorylation of nucleocytoplasmic proteins: a new paradigm for metabolic control of signal transduction and transcription. *Prog Nucleic Acid Res Mol Biol* **73**, 107-36.

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

19. Akimoto, Y., Hart, G. W., Hirano, H. & Kawakami, H. (2005). O-GlcNAc modification of nucleocytoplasmic proteins and diabetes. *Med Mol Morphol* **38**, 84-91.

20. Liu, F., Iqbal, K., Grundke-Iqbal, I., Hart, G. W. & Gong, C. X. (2004). O-GlcNAcylation regulates phosphorylation of tau: a mechanism involved in Alzheimer's disease. *Proc Natl Acad Sci U S A* **101**, 10804-9.

21. Chou, T. Y. & Hart, G. W. (2001). O-linked N-acetylglucosamine and cancer: messages from the glycosylation of c-Myc. *Adv Exp Med Biol* **491**, 413-8.

22. Coutinho, P. M. a. H., B., Ed. (1999). Carbohydrate-active enzymes: an integrated database approach. Recent Advances in Carbohydrate Bioengineering. Edited by H.J. Gilbert, G. D., B. Henrissat and B. Svensson eds. Cambridge: The Royal Society of Chemistry.

23. Davies, G. J., Gloster, T. M. & Henrissat, B. (2005). Recent structural insights into the expanding world of carbohydrate-active enzymes. *Curr Opin Struct Biol* **15**, 637-45.

24. Campbell, J. A., Davies, G. J., Bulone, V. & Henrissat, B. (1997). A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem J* **326 ( Pt 3)**, 929-39.

25. Coutinho, P. M., Deleury, E., Davies, G. J. & Henrissat, B. (2003). An evolving hierarchical family classification for glycosyltransferases. *J Mol Biol* **328**, 307-17.

26. Vrielink, A., Ruger, W., Driessen, H. P. & Freemont, P. S. (1994). Crystal structure of the DNA modifying enzyme beta-glucosyltransferase in the presence and absence of the substrate uridine diphosphoglucose. *EMBO J* **13**, 3413-22.

27. Charnock, S. J. & Davies, G. J. (1999). Structure of the nucleotide-diphospho-sugar transferase, SpsA from Bacillus subtilis, in native and nucleotide-complexed forms. *Biochemistry* **38**, 6380-5.

28. Chiu, C. P., Watts, A. G., Lairson, L. L., Gilbert, M., Lim, D., Wakarchuk, W. W., Withers, S. G. & Strynadka, N. C. (2004). Structural analysis of the sialyltransferase CstII from Campylobacter jejuni in complex with a substrate analog. *Nat Struct Mol Biol* **11**, 163-70.

29. Correia, M. A., Prates, J. A., Bras, J., Fontes, C. M., Newman, J. A., Lewis, R. J., Gilbert, H. J. & Flint, J. E. (2008). Crystal structure of a cellulosomal family 3 carbohydrate esterase from Clostridium thermocellum provides insights into the mechanism of substrate recognition. *J Mol Biol* **379**, 64-72.

30. Coggins, B. E., Li, X., McClerren, A. L., Hindsgaul, O., Raetz, C. R. & Zhou, P. (2003). Structure of the LpxC deacetylase with a bound substrate-analog inhibitor. *Nat Struct Biol* **10**, 645-51.

31. John, M., Rohrig, H., Schmidt, J., Wieneke, U. & Schell, J. (1993). Rhizobium NodB protein involved in nodulation signal synthesis is a chitooligosaccharide deacetylase. *Proc Natl Acad Sci U S A* **90**, 625-9.

32. Blair, D. E., Schuttelkopf, A. W., MacRae, J. I. & van Aalten, D. M. (2005). Structure and metal-dependent mechanism of peptidoglycan deacetylase, a streptococcal virulence factor. *Proc Natl Acad Sci U S A* **102**, 15429-34.

33. Taylor, E. J., Gloster, T. M., Turkenburg, J. P., Vincent, F., Brzozowski, A. M., Dupont, C., Shareck, F., Centeno, M. S., Prates, J. A., Puchart, V., Ferreira, L. M., Fontes, C. M., Biely, P. & Davies, G. J. (2006). Structure and activity of two metal ion-dependent acetylxylan esterases involved in plant cell wall degradation reveals a close similarity to peptidoglycan deacetylases. *J Biol Chem* **281**, 10968-75.

34. Creze, C., Castang, S., Derivery, E., Haser, R., Hugouvieux-Cotte-Pattat, N., Shevchik, V. E. & Gouet, P. (2008). The crystal structure of pectate lyase PelI from soft rot pathogen Erwinia chrysanthemi in complex with its substrate. *J Biol Chem*.

35. Li, S., Kelly, S. J., Lamani, E., Ferraroni, M. & Jedrzejas, M. J. (2000). Structural basis of hyaluronan degradation by Streptococcus pneumoniae hyaluronate lyase. *EMBO J* **19**, 1228-40.

36. Rodionov, D. A., Gelfand, M. S. & Hugouvieux-Cotte-Pattat, N. (2004). Comparative genomics of the KdgR regulon in Erwinia chrysanthemi 3937 and other gamma-proteobacteria. *Microbiology* **150**, 3571-90.

37. Abbott, D. W. & Boraston, A. B. (2007). A family 2 pectate lyase displays a rare fold and transition metal-assisted beta-elimination. *J Biol Chem* **282**, 35328-36.

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

39. Rye, C. S. & Withers, S. G. (2000). Glycosidase mechanisms. *Curr Opin Chem Biol* **4**, 573-80.

40. Henrissat, B. & Bairoch, A. (1996). Updating the sequence-based classification of glycosyl hydrolases. *Biochem J* **316 ( Pt 2)**, 695-6.

41. Ficko-Blean, E., Stubbs, K. A., Nemirovsky, O., Vocadlo, D. J. & Boraston, A. B. (2008). Structural and mechanistic insight into the basis of mucopolysaccharidosis IIIB. *Proc Natl Acad Sci U S A* **105**, 6560-5.

42. Itoh, T., Akao, S., Hashimoto, W., Mikami, B. & Murata, K. (2004). Crystal structure of unsaturated glucuronyl hydrolase, responsible for the degradation of glycosaminoglycan, from Bacillus sp. GL1 at 1.8 A resolution. *J Biol Chem* **279**, 31804-12.

43. Divne, C., Stahlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J. K., Teeri, T. T. & Jones, T. A. (1994). The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from Trichoderma reesei. *Science* **265**, 524-8.

44. Abbott, D. W. & Boraston, A. B. (2007). The structural basis for exopolygalacturonase activity in a family 28 glycoside hydrolase. *J Mol Biol* **368**, 1215-22.

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

46. Martinez-Fleites, C., Macauley, M. S., He, Y., D, L. S., D, J. V. & Davies, G. J. (2008). Structure of an O-GlcNAc transferase homolog provides insight into intracellular glycosylation. *Nat Struct Mol Biol*.

47. Zachara, N. E., Cole, R. N., Hart, G. W. & Gao, Y. (2001). Detection and analysis of proteins modified by O-linked N-acetylglucosamine. *Curr Protoc Protein Sci* **Chapter 12**, Unit 12 8.

48. Levy, I., Shani, Z. & Shoseyov, O. (2002). Modification of polysaccharides and plant cell wall by endo-1,4-beta-glucanase and cellulose-binding domains. *Biomol Eng* **19**, 17-30.

49. Gilbert, H. J. (2007). Cellulosomes: microbial nanomachines that display plasticity in quaternary structure. *Mol Microbiol* **63**, 1568-76.

50. Carvalho, A. L., Dias, F. M., Nagy, T., Prates, J. A., Proctor, M. R., Smith, N., Bayer, E. A., Davies, G. J., Ferreira, L. M., Romao, M. J., Fontes, C. M. & Gilbert, H. J. (2007). Evidence for a dual binding mode of dockerin modules to cohesins. *Proc Natl Acad Sci U S A* **104**, 3089-94.

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

52. Chitayat, S., Gregg, K., Adams, J. J., Ficko-Blean, E., Bayer, E. A., Boraston, A. B. & 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**, 20-8.

53. Hava, D. L. & Camilli, A. (2002). Large-scale identification of serotype 4 Streptococcus pneumoniae virulence factors. *Mol Microbiol* **45**, 1389-406.

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

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

56. Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. (2004). Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* **382**, 769-81.

57. McCartney, L., Gilbert, H. J., Bolam, D. N., Boraston, A. B. & Knox, J. P. (2004). Glycoside hydrolase carbohydrate-binding modules as molecular probes for the analysis of plant cell wall polymers. *Anal Biochem* **326**, 49-54.

58. Kwan, E. M., Boraston, A. B., McLean, B. W., Kilburn, D. G. & Warren, R. A. (2005). N-Glycosidase-carbohydrate-binding module fusion proteins as immobilized enzymes for protein deglycosylation. *Protein Eng Des Sel* **18**, 497-501.

59. Boraston, A. B., McLean, B. W., Guarna, M. M., Amandaron-Akow, E. & Kilburn, D. G. (2001). A family 2a carbohydrate-binding module suitable as an affinity tag for proteins produced in Pichia pastoris. *Protein Expr Purif* **21**, 417-23.

60. Levy, I., Paldi, T. & Shoseyov, O. (2004). Engineering a bifunctional starch-cellulose cross-bridge protein. *Biomaterials* **25**, 1841-9.

61. Vaaje-Kolstad, G., Houston, D. R., Riemen, A. H., Eijsink, V. G. & van Aalten, D. M. (2005). Crystal structure and binding properties of the Serratia marcescens chitin-binding protein CBP21. *J Biol Chem* **280**, 11313-9.

62. Boraston, A. B., McLean, B. W., Chen, G., Li, A., Warren, R. A. & Kilburn, D. G. (2002). Co-operative binding of triplicate carbohydrate-binding modules from a thermophilic xylanase. *Mol Microbiol* **43**, 187-94.

63. Boraston, A. B., Healey, M., Klassen, J., Ficko-Blean, E., Lammerts van Bueren, A. & Law, V. (2006). A structural and functional analysis of alpha-glucan recognition by family 25 and 26 carbohydrate-binding modules reveals a conserved mode of starch recognition. *J Biol Chem* **281**, 587-98.

64. Linder, M., Salovuori, I., Ruohonen, L. & Teeri, T. T. (1996). Characterization of a double cellulose-binding domain. Synergistic high affinity binding to crystalline cellulose. *J Biol Chem* **271**, 21268-72.

65. Bolam, D. N., Xie, H., White, P., Simpson, P. J., Hancock, S. M., Williamson, M. P. & Gilbert, H. J. (2001). Evidence for synergy between family 2b carbohydrate binding modules in Cellulomonas fimi xylanase 11A. *Biochemistry* **40**, 2468-77.

66. Charnock, S. J., Bolam, D. N., Nurizzo, D., Szabo, L., McKie, V. A., Gilbert, H. J. & Davies, G. J. (2002). Promiscuity in ligand-binding: The three-dimensional structure of a Piromyces carbohydrate-binding module, CBM29-2, in complex with cello- and mannohexaose. *Proc Natl Acad Sci U S A* **99**, 14077-82.

67. Flint, J., Nurizzo, D., Harding, S. E., Longman, E., Davies, G. J., Gilbert, H. J. & Bolam, D. N. (2004). Ligand-mediated dimerization of a carbohydrate-binding molecule reveals a novel mechanism for protein-carbohydrate recognition. *J Mol Biol* **337**, 417-26.

68. Boraston, A. B., Creagh, A. L., Alam, M. M., Kormos, J. M., Tomme, P., Haynes, C. A., Warren, R. A. & Kilburn, D. G. (2001). Binding specificity and thermodynamics of a family 9 carbohydrate-binding module from Thermotoga maritima xylanase 10A. *Biochemistry* **40**, 6240-7.

69. Notenboom, V., Boraston, A. B., Kilburn, D. G. & Rose, D. R. (2001). Crystal structures of the family 9 carbohydrate-binding module from Thermotoga maritima xylanase 10A in native and ligand-bound forms. *Biochemistry* **40**, 6248-56.

70. Boraston, A. B., Warren, R. A. & Kilburn, D. G. (2001). beta-1,3-Glucan binding by a thermostable carbohydrate-binding module from Thermotoga maritima. *Biochemistry* **40**, 14679-85.

71. Abou Hachem, M., Nordberg Karlsson, E., Bartonek-Roxa, E., Raghothama, S., Simpson, P. J., Gilbert, H. J., Williamson, M. P. & Holst, O. (2000). Carbohydrate-binding modules from a thermostable Rhodothermus marinus xylanase: cloning, expression and binding studies. *Biochem J* **345 Pt 1**, 53-60.

72. Boraston, A. B., Nurizzo, D., Notenboom, V., Ducros, V., Rose, D. R., Kilburn, D. G. & Davies, G. J. (2002). Differential oligosaccharide recognition by evolutionarily-related beta-1,4 and beta-1,3 glucan-binding modules. *J Mol Biol* **319**, 1143-56.

73. Fritz, T. A., Raman, J. & Tabak, L. A. (2006). Dynamic association between the catalytic and lectin domains of human UDP-GalNAc:polypeptide alpha-N-acetylgalactosaminyltransferase-2. *J Biol Chem* **281**, 8613-9.

74. van den Burg, H. A., Harrison, S. J., Joosten, M. H., Vervoort, J. & de Wit, P. J. (2006). Cladosporium fulvum Avr4 protects fungal cell walls against hydrolysis by plant chitinases accumulating during infection. *Mol Plant Microbe Interact* **19**, 1420-30.

75. Abbott, D. W., Hrynuik, S. & Boraston, A. B. (2007). Identification and characterization of a novel periplasmic polygalacturonic acid binding protein from Yersinia enterolitica. *J Mol Biol* **367**, 1023-33.

76. Jamal-Talabani, S., Boraston, A. B., Turkenburg, J. P., Tarbouriech, N., Ducros, V. M. & Davies, G. J. (2004). Ab initio structure determination and functional characterization of CBM36; a new family of calcium-dependent carbohydrate binding modules. *Structure (Camb)* **12**, 1177-87.

77. Andrews, S. R., Taylor, E. J., Pell, G., Vincent, F., Ducros, V. M., Davies, G. J., Lakey, J. H. & Gilbert, H. J. (2004). The use of forced protein evolution to investigate and improve stability of family 10 xylanases. The production of Ca2+-independent stable xylanases. *J Biol Chem* **279**, 54369-79.

78. Tormo, J., Lamed, R., Chirino, A. J., Morag, E., Bayer, E. A., Shoham, Y. & Steitz, T. A. (1996). Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. *EMBO J* **15**, 5739-51.

79. McLean, B. W., Bray, M. R., Boraston, A. B., Gilkes, N. R., Haynes, C. A. & Kilburn, D. G. (2000). Analysis of binding of the family 2a carbohydrate-binding module from Cellulomonas fimi xylanase 10A to cellulose: specificity and identification of functionally important amino acid residues. *Protein Eng* **13**, 801-9.

80. Boraston, A. B., Wang, D. & Burke, R. D. (2006). Blood group antigen recognition by a Streptococcus pneumoniae virulence factor. *J Biol Chem* **281**, 35263-71.

81. Gregg, K. J., Finn, R., Abbott, D. W. & Boraston, A. B. (2008). Divergent Modes of Glycan Recognition by a New Family of Carbohydrate-binding Modules. *J Biol Chem* **283**, 12604-13.

82. Kraulis, J., Clore, G. M., Nilges, M., Jones, T. A., Pettersson, G., Knowles, J. & Gronenborn, A. M. (1989). Determination of the three-dimensional solution structure of the C-terminal domain of cellobiohydrolase I from Trichoderma reesei. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. *Biochemistry* **28**, 7241-57.

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

84. Notenboom, V., Boraston, A. B., Williams, S. J., Kilburn, D. G. & Rose, D. R. (2002). High-resolution crystal structures of the lectin-like xylan binding domain

from Streptomyces lividans xylanase 10A with bound substrates reveal a novel mode of xylan binding. *Biochemistry* **41**, 4246-54.

85. Lammerts van Bueren, A., Finn, R., Ausio, J. & Boraston, A. B. (2004). Alpha-glucan recognition by a new family of carbohydrate-binding modules found primarily in bacterial pathogens. *Biochemistry* **43**, 15633-42.

86. Abuja, P. M., Pilz, I., Claeyssens, M. & Tomme, P. (1988). Domain structure of cellobiohydrolase II as studied by small angle X-ray scattering: close resemblance to cellobiohydrolase I. *Biochem Biophys Res Commun* **156**, 180-5.

87. Abuja, P. M., Pilz, I., Tomme, P. & Claeyssens, M. (1989). Structural changes in cellobiohydrolase I upon binding of a macromolecular ligand as evident by SAXS investigations. *Biochem Biophys Res Commun* **165**, 615-23.

88. Tomme, P., Warren, R. A. & Gilkes, N. R. (1995). Cellulose hydrolysis by bacteria and fungi. *Adv Microb Physiol* **37**, 1-81.

89. Black, G. W., Rixon, J. E., Clarke, J. H., Hazlewood, G. P., Theodorou, M. K., Morris, P. & Gilbert, H. J. (1996). Evidence that linker sequences and cellulose-binding domains enhance the activity of hemicellulases against complex substrates. *Biochem J* **319 ( Pt 2)**, 515-20.

90. Millward-Sadler, S. J., Poole, D. M., Henrissat, B., Hazlewood, G. P., Clarke, J. H. & Gilbert, H. J. (1994). Evidence for a general role for high-affinity non-catalytic cellulose binding domains in microbial plant cell wall hydrolases. *Mol Microbiol* **11**, 375-82.

91. Sheldon, W. L., Macauley, M. S., Taylor, E. J., Robinson, C. E., Charnock, S. J., Davies, G. J., Vocadlo, D. J. & Black, G. W. (2006). Functional analysis of a group A streptococcal glycoside hydrolase Spy1600 from family 84 reveals it is a beta-N-acetylglucosaminidase and not a hyaluronidase. *Biochem J* **399**, 241-7.

92. Langley, D. B., Harty, D. W., Jacques, N. A., Hunter, N., Guss, J. M. & Collyer, C. A. (2008). Structure of N-acetyl-beta-D-glucosaminidase (GcnA) from the endocarditis pathogen Streptococcus gordonii and its complex with the mechanism-based inhibitor NAG-thiazoline. *J Mol Biol* **377**, 104-16.

93. Burnaugh, A. M., Frantz, L. J. & King, S. J. (2008). Growth of Streptococcus pneumoniae on human glycoconjugates is dependent upon the sequential activity of bacterial exoglycosidases. *J Bacteriol* **190**, 221-30.

94. Newstead, S. L., Potter, J. A., Wilson, J. C., Xu, G., Chien, C. H., Watts, A. G., Withers, S. G. & Taylor, G. L. (2008). The structure of Clostridium perfringens NanI sialidase and its catalytic intermediates. *J Biol Chem* **283**, 9080-8.

95. Boraston, A. B., Ficko-Blean, E. & Healey, M. (2007). Carbohydrate recognition by a large sialidase toxin from Clostridium perfringens. *Biochemistry* **46**, 11352-60.

96. Rutenber, E. & Robertus, J. D. (1991). Structure of ricin B-chain at 2.5 A resolution. *Proteins* **10**, 260-9.

97. Bains, G., Lee, R. T., Lee, Y. C. & Freire, E. (1992). Microcalorimetric study of wheat germ agglutinin binding to N-acetylglucosamine and its oligomers. *Biochemistry* **31**, 12624-8.

98. Sorimachi, K., Le Gal-Coeffet, M. F., Williamson, G., Archer, D. B. & Williamson, M. P. (1997). Solution structure of the granular starch binding domain of Aspergillus niger glucoamylase bound to beta-cyclodextrin. *Structure* **5**, 647-61.

99. Atkins, E. D. T. (1992). *Xylan and Xylanases: Progress in Biotechnology* (Visser, J., Beldman, G. , van Kusters, S. , and Voragen, A. G. L., eds, Ed.), 7, Elsevier Science Publishers B.V., Amsterdam.

100. Hayashi, H., Takehara, M., Hattori, T., Kimura, T., Karita, S., Sakka, K. & Ohmiya, K. (1999). Nucleotide sequences of two contiguous and highly homologous xylanase genes xynA and xynB and characterization of XynA from Clostridium thermocellum. *Appl Microbiol Biotechnol* **51**, 348-57.

101. Czjzek, M., Bolam, D. N., Mosbah, A., Allouch, J., Fontes, C. M., Ferreira, L. M., Bornet, O., Zamboni, V., Darbon, H., Smith, N. L., Black, G. W., Henrissat, B. & Gilbert, H. J. (2001). The location of the ligand-binding site of carbohydrate-binding modules that have evolved from a common sequence is not conserved. *J Biol Chem* **276**, 48580-7.

102. Boraston, A. B., Notenboom, V., Warren, R. A., Kilburn, D. G., Rose, D. R. & Davies, G. (2003). Structure and ligand binding of carbohydrate-binding module CsCBM6-3 reveals similarities with fucose-specific lectins and "galactose-binding" domains. *J Mol Biol* **327**, 659-69.

103. Henshaw, J. L., Bolam, D. N., Pires, V. M., Czjzek, M., Henrissat, B., Ferreira, L. M., Fontes, C. M. & Gilbert, H. J. (2004). The family 6 carbohydrate binding module CmCBM6-2 contains two ligand-binding sites with distinct specificities. *J Biol Chem* **279**, 21552-9.

104. Pires, V. M., Henshaw, J. L., Prates, J. A., Bolam, D. N., Ferreira, L. M., Fontes, C. M., Henrissat, B., Planas, A., Gilbert, H. J. & Czjzek, M. (2004). The crystal structure of the family 6 carbohydrate binding module from Cellvibrio mixtus endoglucanase 5a in complex with oligosaccharides reveals two distinct binding sites with different ligand specificities. *J Biol Chem* **279**, 21560-8.

105. Notenboom, V., Boraston, A. B., Chiu, P., Freelove, A. C., Kilburn, D. G. & Rose, D. R. (2001). Recognition of cello-oligosaccharides by a family 17 carbohydrate-binding module: an X-ray crystallographic, thermodynamic and mutagenic study. *J Mol Biol* **314**, 797-806.

106. Pell, G., Williamson, M. P., Walters, C., Du, H., Gilbert, H. J. & Bolam, D. N. (2003). Importance of hydrophobic and polar residues in ligand binding in the family 15 carbohydrate-binding module from Cellvibrio japonicus Xyn10C. *Biochemistry* **42**, 9316-23.

107. Garcia-Hernandez, E., Zubillaga, R. A., Chavelas-Adame, E. A., Vazquez-Contreras, E., Rojo-Dominguez, A. & Costas, M. (2003). Structural energetics of protein-carbohydrate interactions: Insights derived from the study of lysozyme binding to its natural saccharide inhibitors. *Protein Sci* **12**, 135-42.

108. Chervenak, M. C. & Toone, E. J. (1995). Calorimetric analysis of the binding of lectins with overlapping carbohydrate-binding ligand specificities. *Biochemistry* **34**, 5685-95.

109. Mach, H., Middaugh, C. R. & 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.

110. Turnbull, W. B. & Daranas, A. H. (2003). On the value of c: can low affinity systems be studied by isothermal titration calorimetry? *J Am Chem Soc* **125**, 14859-66.

111. Boraston, A. B., Ghaffari, M., Warren, R. A. & Kilburn, D. G. (2002). Identification and glucan-binding properties of a new carbohydrate-binding module family. *Biochem J* **361**, 35-40.

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

113. Vagin, A. & Teplyakov, A. (2000). An approach to multi-copy search in molecular replacement. *Acta Crystallogr D Biol Crystallogr* **56**, 1622-4.

114. McRee, D. E. (1999). XtalView/Xfit--A versatile program for manipulating atomic coordinates and electron density. *J Struct Biol* **125**, 156-65.

115. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53**, 240-55.

116. Szabo, L., Jamal, S., Xie, H., Charnock, S. J., Bolam, D. N., Gilbert, H. J. & Davies, G. J. (2001). Structure of a family 15 carbohydrate-binding module in complex with xylopentaose. Evidence that xylan binds in an approximate 3-fold helical conformation. *J Biol Chem* **276**, 49061-5.

117. Bailey, S. (1994). The Ccp4 Suite - Programs for Protein Crystallography. *Acta Crystallographica Section D-Biological Crystallography* **50**, 760-763.

118. Fraczkiewicz, R. & Braun, W. (1998). Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *Journal of Computational Chemistry* **19**, 319-333.

119. Read, R. J. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr A* **42**, 140-149.

120. Boraston, A. B., Chiu, P., Warren, R. A. & Kilburn, D. G. (2000). Specificity and affinity of substrate binding by a family 17 carbohydrate-binding module from Clostridium cellulovorans cellulase 5A. *Biochemistry* **39**, 11129-36.

121. Charnock, S. J., Bolam, D. N., Turkenburg, J. P., Gilbert, H. J., Ferreira, L. M., Davies, G. J. & Fontes, C. M. (2000). The X6 "thermostabilizing" domains of xylanases are carbohydrate-binding modules: structure and biochemistry of the Clostridium thermocellum X6b domain. *Biochemistry* **39**, 5013-21.

122. Garcia-Hernandez, E. & Hernandez-Arana, A. (1999). Structural bases of lectin-carbohydrate affinities: comparison with protein-folding energetics. *Protein Sci* **8**, 1075-86.

123. Boraston, A. B., Revett, T. J., Boraston, C. M., Nurizzo, D. & Davies, G. J. (2003). Structural and thermodynamic dissection of specific mannan recognition by a carbohydrate binding module, TmCBM27. *Structure (Camb)* **11**, 665-75.

124. Tomme, P., Creagh, A. L., Kilburn, D. G. & Haynes, C. A. (1996). Interaction of polysaccharides with the N-terminal cellulose-binding domain of Cellulomonas fimi CenC. 1. Binding specificity and calorimetric analysis. *Biochemistry* **35**, 13885-94.

125. Schwarz, F. P., Ahmed, H., Bianchet, M. A., Amzel, L. M. & Vasta, G. R. (1998). Thermodynamics of bovine spleen galectin-1 binding to disaccharides: correlation with structure and its effect on oligomerization at the denaturation temperature. *Biochemistry* **37**, 5867-77.

126. Freelove, A. C., Bolam, D. N., White, P., Hazlewood, G. P. & Gilbert, H. J. (2001). A novel carbohydrate-binding protein is a component of the plant cell wall-degrading complex of Piromyces equi. *J Biol Chem* **276**, 43010-7.

127. Miller, G. L. (1959). Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar. *Anal. Chem.* **31**, 426-428.

128. Read, S. M., Currie, G. & Bacic, A. (1996). Analysis of the structural heterogeneity of laminarin by electrospray-ionisation-mass spectrometry. *Carbohydr Res* **281**, 187-201.

129. Pflugrath, J. W. (1999). The finer things in X-ray diffraction data collection. *Acta Crystallogr D Biol Crystallogr* **55**, 1718-25.

130. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.

131. Baladron, V., Ufano, S., Duenas, E., Martin-Cuadrado, A. B., del Rey, F. & Vazquez de Aldana, C. R. (2002). Eng1p, an endo-1,3-beta-glucanase localized at the daughter side of the septum, is involved in cell separation in Saccharomyces cerevisiae. *Eukaryot Cell* **1**, 774-86.

132. Martin-Cuadrado, A. B., Duenas, E., Sipiczki, M., Vazquez de Aldana, C. R. & del Rey, F. (2003). The endo-beta-1,3-glucanase eng1p is required for dissolution of the primary septum during cell separation in Schizosaccharomyces pombe. *J Cell Sci* **116**, 1689-98.

133. Fliegmann, J., Mithofer, A., Wanner, G. & Ebel, J. (2004). An ancient enzyme domain hidden in the putative beta-glucan elicitor receptor of soybean may play an active part in the perception of pathogen-associated molecular patterns during broad host resistance. *J Biol Chem* **279**, 1132-40.

134. Zverlov, V. V., Volkov, I. Y., Velikodvorskaya, G. A. & Schwarz, W. H. (2001). The binding pattern of two carbohydrate-binding modules of laminarinase Lam16A from Thermotoga neapolitana: differences in beta-glucan binding within family CBM4. *Microbiology* **147**, 621-9.

135. Frecer, V., Rizzo, R. & Miertus, S. (2000). Molecular dynamics study on the conformational stability of laminaran oligomers in various solvents. *Biomacromolecules* **1**, 91-9.

136. van Bueren, A. L., Morland, C., Gilbert, H. J. & Boraston, A. B. (2005). Family 6 carbohydrate binding modules recognize the non-reducing end of beta-1,3-linked glucans by presenting a unique ligand binding surface. *J Biol Chem* **280**, 530-7.

137. Henshaw, J., Horne-Bitschy, A., van Bueren, A. L., Money, V. A., Bolam, D. N., Czjzek, M., Ekborg, N. A., Weiner, R. M., Hutcheson, S. W., Davies, G. J., Boraston, A. B. & Gilbert, H. J. (2006). Family 6 carbohydrate binding modules in beta-agarases display exquisite selectivity for the non-reducing termini of agarose chains. *J Biol Chem* **281**, 17099-107.

138. Svensson, B., Jespersen, H., Sierks, M. R. & MacGregor, E. A. (1989). Sequence homology between putative raw-starch binding domains from different starch-degrading enzymes. *Biochem J* **264**, 309-11.

139. Takahashi, T., Kato, K., Ikegami, Y. & Irie, M. (1985). Different behavior towards raw starch of three forms of glucoamylase from a Rhizopus sp. *J Biochem* **98**, 663-71.

140. Abe, A., Tonozuka, T., Sakano, Y. & Kamitori, S. (2004). Complex structures of Thermoactinomyces vulgaris R-47 alpha-amylase 1 with malto-oligosaccharides demonstrate the role of domain N acting as a starch-binding domain. *J Mol Biol* **335**, 811-22.

141. Huber, R., Langworthy, T.A., Konig, H., Thomm, M., Woese, C.R., Sleytr, U.B., and Setter, K.O. (1986). Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 degrees celcius. *Arch. Microbiol.* **144**, 324-333.

142. Kriegshauser, G. & Liebl, W. (2000). Pullulanase from the hyperthermophilic bacterium Thermotoga maritima: purification by beta-cyclodextrin affinity chromatography. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **737**, 245-251.

143. Bibel, M., Brettl, C., Gosslar, U., Kriegshauser, G. & Liebl, W. (1998). Isolation and analysis of genes for amylolytic enzymes of the hyperthermophilic bacterium Thermotoga maritima. *FEMS Microbiol Lett* **158**, 9-15.

144. Sigurskjold, B. W., Altman, E. & Bundle, D. R. (1991). Sensitive titration microcalorimetric study of the binding of Salmonella O-antigenic oligosaccharides by a monoclonal antibody. *Eur J Biochem* **197**, 239-46.

145. Sorimachi, K., Jacks, A. J., Le Gal-Coeffet, M. F., Williamson, G., Archer, D. B. & Williamson, M. P. (1996). Solution structure of the granular starch binding domain of glucoamylase from Aspergillus niger by nuclear magnetic resonance spectroscopy. *J Mol Biol* **259**, 970-87.

146. Tomme, P., Boraston, A., Kormos, J. M., Warren, R. A. & Kilburn, D. G. (2000). Affinity electrophoresis for the identification and characterization of soluble sugar binding by carbohydrate-binding modules. *Enzyme Microb Technol* **27**, 453-458.

147. Kuzmic, P. (1996). Program DYNAFIT for the analysis of enzyme kinetic data: application to HIV proteinase. *Anal Biochem* **237**, 260-73.

148. Straume, M. & Johnson, M. L. (1992). Analysis of residuals: criteria for determining goodness-of-fit. *Methods Enzymol* **210**, 87-105.

149. Straume, M. & Johnson, M. L. (1992). Monte Carlo method for determining complete confidence probability distributions of estimated model parameters. *Methods Enzymol* **210**, 117-29.

150. Cohn, E. J. a. E., J.T. (1942). In *Proteins, Amino Acids and Peptides*, Reinhold, New York.

151. Perkins, S. J. (1986). Protein volumes and hydration effects. The calculations of partial specific volumes, neutron scattering matchpoints and 280-nm absorption coefficients for proteins and glycoproteins from amino acid sequences. *Eur J Biochem* **157**, 169-80.

152. Hytonen, J., Haataja, S. & Finne, J. (2003). Streptococcus pyogenes glycoprotein-binding strepadhesin activity is mediated by a surface-associated carbohydrate-degrading enzyme, pullulanase. *Infect Immun* **71**, 784-93.

153. Machovic, M., Svensson, B., MacGregor, E. A. & Janecek, S. (2005). A new clan of CBM families based on bioinformatics of starch-binding domains from families CBM20 and CBM21. *Febs J* **272**, 5497-513.

154. Chou, W. I., Pai, T. W., Liu, S. H., Hsiung, B. K. & Chang, M. D. (2006). The family 21 carbohydrate-binding module of glucoamylase from Rhizopus oryzae consists of two sites playing distinct roles in ligand binding. *Biochem J* **396**, 469-77.

155. Mikkelsen, R., Suszkiewicz, K. & Blennow, A. (2006). A novel type carbohydrate-binding module identified in alpha-glucan, water dikinases is specific for regulated plastidial starch metabolism. *Biochemistry* **45**, 4674-82.

156. Schneider, T. R. & Sheldrick, G. M. (2002). Substructure solution with SHELXD. *Acta Crystallogr D Biol Crystallogr* **58**, 1772-9.

157. Evans, G. & Bricogne, G. (2002). Triiodide derivatization and combinatorial counter-ion replacement: two methods for enhancing phasing signal using laboratory Cu Kalpha X-ray equipment. *Acta Crystallogr D Biol Crystallogr* **58**, 976-91.

158. Cowtan, K. D. & Zhang, K. Y. (1999). Density modification for macromolecular phase improvement. *Prog Biophys Mol Biol* **72**, 245-70.

159. Perrakis, A., Morris, R. & Lamzin, V. S. (1999). Automated protein model building combined with iterative structure refinement. *Nat Struct Biol* **6**, 458-63.

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

161. Mikami, B., Iwamoto, H., Malle, D., Yoon, H. J., Demirkan-Sarikaya, E., Mezaki, Y. & Katsuya, Y. (2006). Crystal structure of pullulanase: evidence for parallel binding of oligosaccharides in the active site. *J Mol Biol* **359**, 690-707.

162. Gouet, P., Robert, X. & Courcelle, E. (2003). ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res* **31**, 3320-3.

163. Manco, S., Hernon, F., Yesilkaya, H., Paton, J. C., Andrew, P. W. & Kadioglu, A. (2006). Pneumococcal neuraminidases A and B both have essential roles during infection of the respiratory tract and sepsis. *Infect Immun* **74**, 4014-20.

164. Soong, G., Muir, A., Gomez, M. I., Waks, J., Reddy, B., Planet, P., Singh, P. K., Kanetko, Y., Wolfgang, M. C., Hsiao, Y. S., Tong, L. & Prince, A. (2006). Bacterial neuraminidase facilitates mucosal infection by participating in biofilm production. *J Clin Invest*.

165. Tomasz, A. (1999). New faces of an old pathogen: emergence and spread of multidrug-resistant Streptococcus pneumoniae. *Am J Med* **107**, 55S-62S.

166. Cunningham, M. W. (2000). Pathogenesis of group A streptococcal infections. *Clin Microbiol Rev* **13**, 470-511.

167. Cengiz, A. B., Kanra, G., Caglar, M., Kara, A., Gucer, S. & Ince, T. (2004). Fatal necrotizing pneumonia caused by group A streptococcus. *J Paediatr Child Health* **40**, 69-71.

168. Morozumi, M., Nakayama, E., Iwata, S., Aoki, Y., Hasegawa, K., Kobayashi, R., Chiba, N., Tajima, T. & Ubukata, K. (2006). Simultaneous detection of pathogens in clinical samples from patients with community-acquired pneumonia by real-time PCR with pathogen-specific molecular beacon probes. *J Clin Microbiol* **44**, 1440-6.

169. Shelburne, S. A., 3rd, Sumby, P., Sitkiewicz, I., Okorafor, N., Granville, C., Patel, P., Voyich, J., Hull, R., DeLeo, F. R. & Musser, J. M. (2006). Maltodextrin utilization plays a key role in the ability of group A Streptococcus to colonize the oropharynx. *Infect Immun* **74**, 4605-14.

170. Bongaerts, R. J., Heinz, H. P., Hadding, U. & Zysk, G. (2000). Antigenicity, expression, and molecular characterization of surface-located pullulanase of Streptococcus pneumoniae. *Infect Immun* **68**, 7141-3.

171. Hytonen, J., Haataja, S. & Finne, J. (2006). Use of flow cytometry for the adhesion analysis of Streptococcus pyogenes mutant strains to epithelial cells: investigation of the possible role of surface pullulanase and cysteine protease, and the transcriptional regulator Rgg. *BMC Microbiol* **6**, 18.

172. Barik, S. (1996). Site-directed mutagenesis in vitro by megaprimer PCR. *Methods Mol Biol* **57**, 203-15.

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

174. Boraston, A. B., Kwan, E., Chiu, P., Warren, R. A. & Kilburn, D. G. (2003). Recognition and hydrolysis of noncrystalline cellulose. *J Biol Chem* **278**, 6120-7.

175. van Bueren, A. L. & Boraston, A. B. (2007). The structural basis of alpha-glucan recognition by a family 41 carbohydrate-binding module from Thermotoga maritima. *J Mol Biol* **365**, 555-60.

176. Sakon, J., Irwin, D., Wilson, D. B. & Karplus, P. A. (1997). Structure and mechanism of endo/exocellulase E4 from Thermomonospora fusca. *Nat Struct Biol* **4**, 810-8.

177. Glasser, S. W., Detmer, E. A., Ikegami, M., Na, C. L., Stahlman, M. T. & Whitsett, J. A. (2003). Pneumonitis and emphysema in sp-C gene targeted mice. *J Biol Chem* **278**, 14291-8.

178. Weaver, T. E. & Conkright, J. J. (2001). Function of surfactant proteins B and C. *Annu Rev Physiol* **63**, 555-78.

179. Cundell, D. R. & Tuomanen, E. I. (1994). Receptor specificity of adherence of Streptococcus pneumoniae to human type-II pneumocytes and vascular endothelial cells in vitro. *Microb Pathog* **17**, 361-74.

180. Kadioglu, A., Sharpe, J. A., Lazou, I., Svanborg, C., Ockleford, C., Mitchell, T. J. & Andrew, P. W. (2001). Use of green fluorescent protein in visualisation of pneumococcal invasion of broncho-epithelial cells in vivo. *FEMS Microbiol Lett* **194**, 105-10.

181. Ridsdale, R. & Post, M. (2004). Surfactant lipid synthesis and lamellar body formation in glycogen-laden type II cells. *Am J Physiol Lung Cell Mol Physiol* **287**, L743-51.

182. Rannels, S. R., Rannels, S. L., Sneyd, J. G. & Loten, E. G. (1991). Fetal lung development in rats with a glycogen storage disorder. *Am J Physiol* **260**, L419-27.

183. Rooney, S. A. (2001). Regulation of surfactant secretion. *Comp Biochem Physiol A Mol Integr Physiol* **129**, 233-43.

184. Jounblat, R., Kadioglu, A., Iannelli, F., Pozzi, G., Eggleton, P. & Andrew, P. W. (2004). Binding and agglutination of Streptococcus pneumoniae by human surfactant protein D (SP-D) vary between strains, but SP-D fails to enhance killing by neutrophils. *Infect Immun* **72**, 709-16.

185. Morona, J. K., Morona, R. & Paton, J. C. (2006). Attachment of capsular polysaccharide to the cell wall of Streptococcus pneumoniae type 2 is required for invasive disease. *Proc Natl Acad Sci U S A* **103**, 8505-10.

186. Darkes, M. J. & Plosker, G. L. (2002). Pneumococcal conjugate vaccine (Prevnar; PNCRM7): a review of its use in the prevention of Streptococcus pneumoniae infection. *Paediatr Drugs* **4**, 609-30.

187. Munoz-Almagro, C., Jordan, I., Gene, A., Latorre, C., Garcia-Garcia, J. J. & Pallares, R. (2008). Emergence of invasive pneumococcal disease caused by nonvaccine serotypes in the era of 7-valent conjugate vaccine. *Clin Infect Dis* **46**, 174-82.

188. Balmer, P., Borrow, R. & Arkwright, P. D. (2007). The 23-valent pneumococcal polysaccharide vaccine does not provide additional serotype antibody protection

in children who have been primed with two doses of heptavalent pneumococcal conjugate vaccine. *Vaccine* **25**, 6321-5.

189. File, T. M., Jr. (2006). Clinical implications and treatment of multiresistant Streptococcus pneumoniae pneumonia. *Clin Microbiol Infect* **12 Suppl 3**, 31-41.

190. File, T. M., Jr., Tan, J. S. & Boex, J. R. (2006). The clinical relevance of penicillin-resistant Streptococcus pneumoniae: a new perspective. *Clin Infect Dis* **42**, 798-800.

191. Pichichero, M. E. & Casey, J. R. (2007). Emergence of a multiresistant serotype 19A pneumococcal strain not included in the 7-valent conjugate vaccine as an otopathogen in children. *JAMA* **298**, 1772-8.

192. van Bueren, A. L., Higgins, M., Wang, D., Burke, R. D. & Boraston, A. B. (2007). Identification and structural basis of binding to host lung glycogen by streptococcal virulence factors. *Nat Struct Mol Biol* **14**, 76-84.

193. Jackson, P. (1990). The use of polyacrylamide-gel electrophoresis for the high-resolution separation of reducing saccharides labelled with the fluorophore 8-aminonaphthalene-1,3,6-trisulphonic acid. Detection of picomolar quantities by an imaging system based on a cooled charge-coupled device. *Biochem J* **270**, 705-13.

194. Koch, M. H. J., Bordas, J. (1983). X-ray diffraction and scattering on disordered systems using synchrotron radiation. *Nuclear Instruments and Methods in Physics Research* **208**, 461-469

195. Guinier, A., Fournet, F. . (1955). *Small angle scattering of X-rays*, Wiley Interscience, New York.

196. Svergun, D. I. (1992). Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *Journal of Applied Crystallography* **25**, 504-513.

197. Bergmann, A., Fritz, G., Glatter, O. (2000). Solving the indirect Fourier transformation including the structure factor requires a non-linear least-squares approach. The Boltzmann simplex simulated annealing proves to be very efficient for this task. *Journal of Applied Crystallography* **33**, 1212-1216.

198. Svergun, D. I., Petoukhov, M. V. & Koch, M. H. (2001). Determination of domain structure of proteins from X-ray solution scattering. *Biophys J* **80**, 2946-53.

199. Roussel, A., Cambillau, C. . (1989). *Silicon Graphics Geometry Partner Directory* (Graphics, S., Ed.), Silicon Graphics, Mountain View, CA.

200. Petoukhov, M. V., Eady, N. A., Brown, K. A. & Svergun, D. I. (2002). Addition of missing loops and domains to protein models by x-ray solution scattering. *Biophys J* **83**, 3113-25.

201. Stam, M. R., Danchin, E. G., Rancurel, C., Coutinho, P. M. & Henrissat, B. (2006). Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Eng Des Sel* **19**, 555-62.

202. Puyet, A. & Espinosa, M. (1993). Structure of the maltodextrin-uptake locus of Streptococcus pneumoniae. Correlation to the Escherichia coli maltose regulon. *J Mol Biol* **230**, 800-11.

203. Nieto, C., Espinosa, M. & Puyet, A. (1997). The maltose/maltodextrin regulon of Streptococcus pneumoniae. Differential promoter regulation by the transcriptional repressor MalR. *J Biol Chem* **272**, 30860-5.

204. Oldham, M. L., Khare, D., Quiocho, F. A., Davidson, A. L. & Chen, J. (2007). Crystal structure of a catalytic intermediate of the maltose transporter. *Nature* **450**, 515-21.

205. Shelburne, S. A., 3rd, Keith, D. B., Davenport, M. T., Horstmann, N., Brennan, R. G. & Musser, J. M. (2008). Molecular characterization of group A Streptococcus maltodextrin catabolism and its role in pharyngitis. *Mol Microbiol*.

206. Ton-That, H., Marraffini, L. A. & Schneewind, O. (2004). Protein sorting to the cell wall envelope of Gram-positive bacteria. *Biochim Biophys Acta* **1694**, 269-78.

207. Klein, C., Hollender, J., Bender, H. & Schulz, G. E. (1992). Catalytic center of cyclodextrin glycosyltransferase derived from X-ray structure analysis combined with site-directed mutagenesis. *Biochemistry* **31**, 8740-6.

208. Scheen, A. J. (2003). Is there a role for alpha-glucosidase inhibitors in the prevention of type 2 diabetes mellitus? *Drugs* **63**, 933-51.

209. Greffe, L., Jensen, M. T., Chang-Pi-Hin, F., Fruchard, S., O'Donohue, M. J., Svensson, B. & Driguez, H. (2002). Chemoenzymatic syntheses of linear and branched hemithiomaltodextrins as potential inhibitors for starch-debranching enzymes. *Chemistry* **8**, 5447-55.

210. Jain, D., Kaur, K., Sundaravadivel, B. & Salunke, D. M. (2000). Structural and functional consequences of peptide-carbohydrate mimicry. Crystal structure of a carbohydrate-mimicking peptide bound to concanavalin A. *J Biol Chem* **275**, 16098-102.

211. Evans, S. V., Sigurskjold, B. W., Jennings, H. J., Brisson, J. R., To, R., Tse, W. C., Altman, E., Frosch, M., Weisgerber, C., Kratzin, H. D. & et al. (1995). Evidence for the extended helical nature of polysaccharide epitopes. The 2.8 A resolution structure and thermodynamics of ligand binding of an antigen binding fragment specific for alpha-(2-->8)-polysialic acid. *Biochemistry* **34**, 6737-44.

212. Naismith, J. H. & Field, R. A. (1996). Structural basis of trimannoside recognition by concanavalin A. *J Biol Chem* **271**, 972-6.

213. Service, R. F. (2007). Cellulosic ethanol. Biofuel researchers prepare to reap a new harvest. *Science* **315**, 1488-91.

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

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

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

# Appendix A
# SpuA structure determination by SAXS

(Provided by Dr. Mirjam Czjzek, from the Station biologique de Roscoff)

At low scattering angles, the scattering of the samples with highest concentrations (i.e. 11 and 7 mg/ml) displayed an increase of intensity due to attractive inter-particle interactions. In order to exclude the influence of these interactions from shape calculations, a high and a low concentration measurement were merged both for SpuA and SpuA-M4 in such a manner that high q-values (0.5-4.0 Å) are taken from the curve at 7.74 and 5.56 mg/ml concentration for SpuA and SpuA-M4, respectively (good statistics of the high resolution data), while the low q-range corresponds to the 2.95 and 1.39 mg/ml measurements (no attractive inter-particle interaction).

The average size of the different constructs is estimated by the measure of their respective radius of gyration at 288 K. At low angles, the scattered intensities are very well approximated by the Guinier law [195], and reveal some attractive interparticle interactions at high concentrations. All scattering curves were indicative of monomeric states of the molecules in solution. The experimental scattering curves are shown in Figure A1. The radii of gyration extrapolated at zero concentration are reported in Table A1. The theoretically calculated RG value, corresponding to the SpuA fullength crystal structure is 33.04 Å. The distance distribution functions of SpuA compared to SpuA-M4 are illustrated in Figure A2, and clearly deviate from those typical for globular proteins, leading to Dmax values reported in Table A1. The RG and Dmax values for SpuA and SpuA-M4 are similar and indicative of rather compact molecules, (which is in agreement

with the crystal structure of SpuA fullength), however the overall shape and P(r) distribution are indicative of structural rearrangements.

Table A1. Structural parameters of SpuA and SpuA-M4 obtained by SAXS

| Sample | Conc. [mg/ml] | $R_G$ (Å) Guinier approx. | $R_G$ (Å) whole curve(GNOM) | $D_{max}$ (Å) | ab initio modeling $^{d}\chi^{(GASB)}$ | ab initio modeling $\chi^{(Model)}$ |
|---|---|---|---|---|---|---|
| SpuA -1 | 11.4 | 35.5±2 | 35.6 | 110±7 | 3.38 | 8.11 |
| SpuA - 2 | 7.74 | 33.8±4 | 35.1 | 105±5 | 3.52 | 5.89 |
| SpuA - 3 | 2.95 | 38.8±4 | 36.7 | 105±5 | 1.73 | 3.50 |
| SpuA | (merged) | 38.8±4 | 36.5 | 103±3 | 2.45 | 5.57 |
| SpuA - M4 -1 | 11.11 | 37.7 | 36.1 | 107±7 | nd | 8.71 |
| SpuA - M4 - 2 | 5.56 | 38.4 | 36.1 | 102±5 | 3.68 | 6.92 |
| SpuA - M4 - 3 | 2.78 | 40.3 | 37.6 | 110±5 | 2.07 | 3.68 |
| SpuA – M4 | (merged) | 38.4 | 35.8 | 110±3 | 2.18 | 3.95 |
| BSA | 4.5 | 30.7±1 | 30.2 | 90±2 | nd | nd |

$^{d}\chi^{(GASB)}$ and $\chi^{(Model)}$: discrepancies between the experimental SAXS profile and respectively the fits for the overall shapes-models calculated by program GASBOR and the average discrepancy of the best atomic models estimated with the program CRYSOL. Discrepancy was defined according to Konarev et al. [214].

For each construct, several ab initio GASBOR calculations were performed, and subsequently compared with the program DAMAVER [215] that computes the normalized spatial discrepancy (NSD) value for the various obtained shapes [216]. In all cases, the various calculations led to highly similar forms with NSD values are of roughly 1.2 for SpuA, and ranging from 1.2 to 1.5 for SpuA-M4, which is indicative of rather restricted conformational variability of the different modules in solution. The superimposition of different shapes obtained for SpuA and SpuA-M4 are illustrated in Figure 51a, b. The superimposition of the “best” overall ab initio shapes onto the adequately positioned crystal structure of SpuA fullength are illustrated in Figure 51e, f.

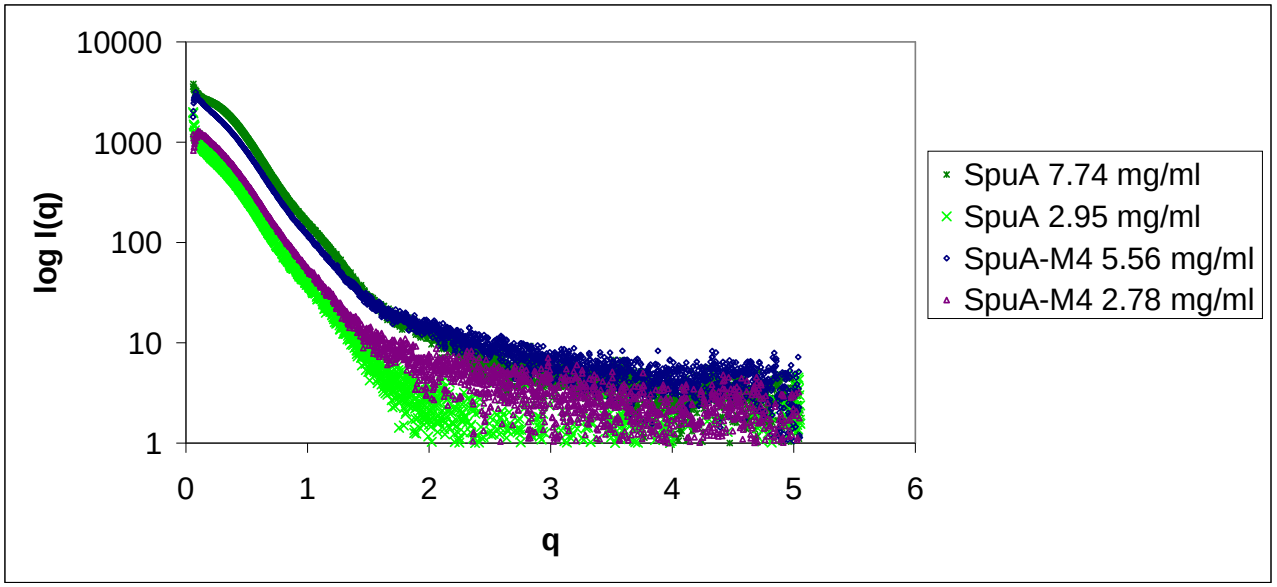



**Figure A1.** Experimental scattering curve at two different concentrations of SpuA (green curves) and SpuA-M4 (blue and purple curves).

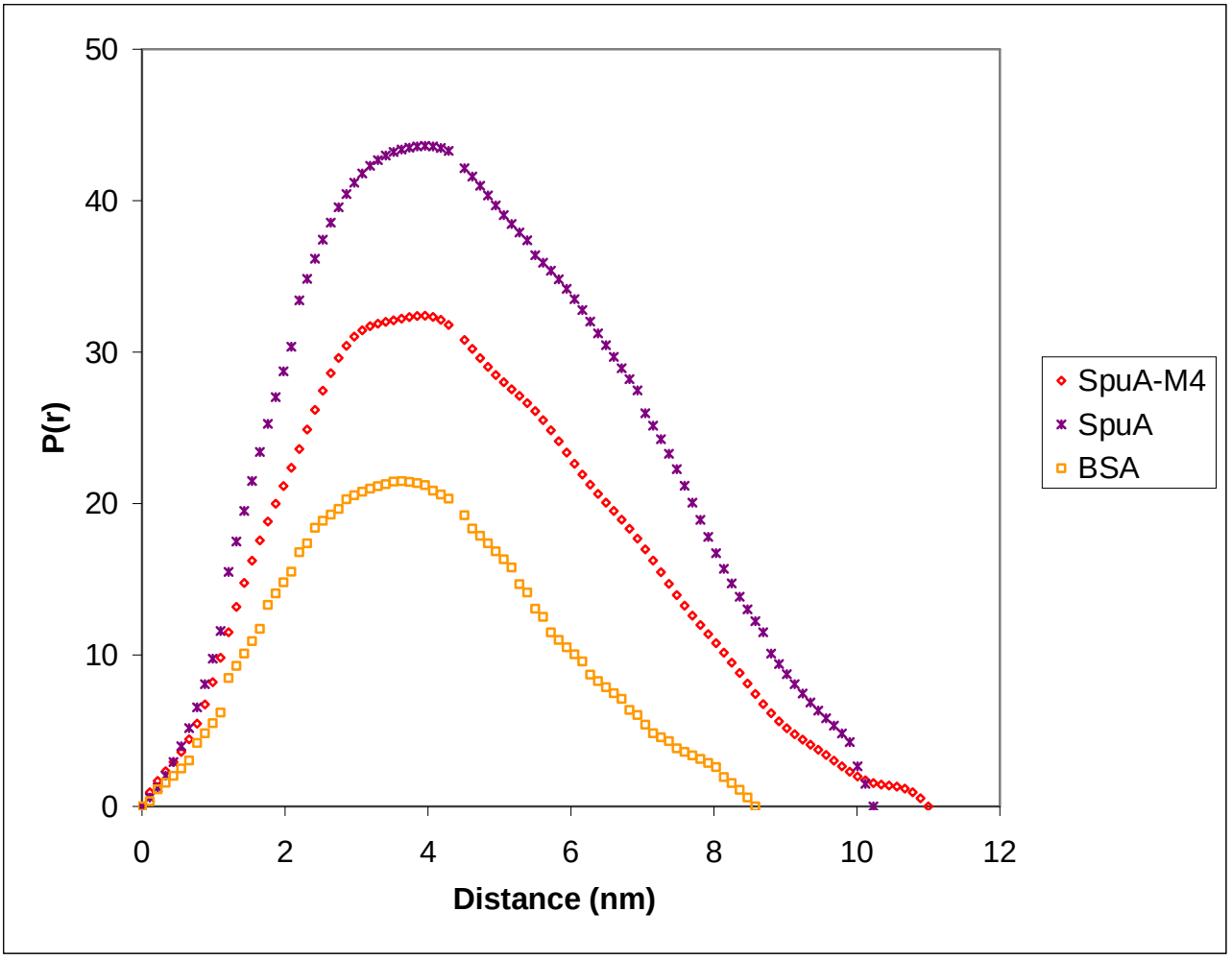



**Figure A2**. P(r) distance distribution of SpuA (purple stars) and SpuA-M4 (red circles) and the reference measurement of BSA (yellow squares), as derived from the experimental scattering in solution.