Biochemical Features of Resistant Starch Degradation by *Ruminococcus bromii*

by

Filipe M. Cerqueira

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Microbiology and Immunology)
in The University of Michigan
2022

Doctoral Committee:

Associate Professor Nicole Koropatkin, Chair
Professor Vernon Carruthers
Professor Matthew Chapman
Clinical Assistant Professor Mary Riwes

Filipe M. Cerqueira

filipehu@umich.edu

ORCID iD: 0000-0003-4438-7959

ACKNOWLEDGMENTS

I have been blessed to have a strong support system of mentors, formal and informal alike, the star among them being Dr. Nicole Koropatkin. Aside from the typical characteristics of a good mentor, she possesses rare traits that makes her exceptional. She adjusted her mentorship to what I needed at the time. She listened and admitted when she was wrong. She grew and learned alongside me. As the only racially underrepresented person in her lab, she went above and beyond to make an inclusive training environment for me. Instead of forcing me to assimilate into the current academic environment, she used her authority to *create* space for me to flourish as a scientist through the lens of my own identities and culture. She modelled how kindness, humility, and positivity are essential pillars on the making and practice of being a great scientist and leader. She has left an everlasting impact on me as a scientist and, more importantly, as a human being. Thank you, Nicole.

I would like to thank Dana Thomas from the Futures in Public Health Leaders Program at University of Michigan for believing in me when all I had was interest and enthusiasm. Her decision to invest in my career was the first major blow to my imposter syndrome and gave me the courage to continue to apply for other competitive positions. I would like to thank Leon Golson at UNIFIED Health, previously HIV/AIDS Resource Center, who taught me to always remember who I am and support those who need it most. Thank you to Dr. Jason Upton, who was the first person to give me a chance to get experience in the microbiology research laboratory and mentor me to think critically about science. I would also like to thank Dr. Bruce

Birren for seeing my potential as a biomedical researcher and supporting my quest for graduate school. Thank you to the rest of my thesis committee, Drs. Vernon Carruthers, Mary Riwes, and Matthew Chapman for their help in my professional development and constructive feedback on my science which were critical to the production and polishing of my work and career. Thank you to Drs. Mary O'Riordan and Scott Barolo for helping me navigate the hidden curriculum in academic medicine. A huge thank you goes to the incredible members of The DEI Taskforce, especially to my co-founder Dr. Lindy Jensen, who have been both my greatest friends and partners in the fight to make biomedical research inclusive.

I want to thank my mom, Camila Muniz, for pushing me to my limits and never letting me believe I wasn't worthy or smart enough to get an education. Because you never gave up on me, I never gave up on myself.

Lastly, I want to thank my life partner, Joseph Alcocer. Your support, sacrifice, love, and commitment that allowed us to move to the Michigan Tundra for my career. From you I learned that I am smart, I am talented, and I can do hard things.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The human gut microbial community influences many aspects of human physiology via the output of short chain fatty acids from the fermentation of dietary carbohydrates. *Ruminococcus bromii*, a keystone species in the human gut, degrades dietary resistant starch. The byproducts of this degradation cross-feed other gut bacteria that produce butyrate, a short chain fatty acid with potent pro-health properties. However, the molecular determinants of resistant starch degradation that initiate the metabolic cascade leading to increased colonic butyrate are unknown. *R. bromii* exports a unique suite of starch-active proteins that work in concert within larger complexes, amylosomes, that allow this bacterium to bind and degrade resistant starch. Although many gut bacteria have starch-binding and starch-degrading proteins, few can utilize resistant starch as the sole carbon source. The work presented here addresses how individual starch-active proteins from *R. bromii* facilitate the recognition of resistant starch, allowing its breakdown.

In this thesis, I present a structure-function study of the maltogenic α-amylase, Amy5, and two starch-binding proteins of the amylosome system, Sas20 and Sca5. Amy5 is not predicted to be a part of the amylosome, however, it is secreted into the extracellular space where it can encounter its preferred substrate, amylose. Sas20 has two starch-binding domains, the first with a preference for the non-reducing end of starch glycans and the second for starch glycan chains longer than seven glucose residues with helical character. Sca5 has two cohesin modules which have a typical jelly-roll fold and two starch-binding modules, one of which has high sequence homology to domain 2 of Sas20.

We propose two cooperative models to describe how Amy5, Sas20, and Sca5 work together to degrade resistant starch. On the surface of the starch granule, Amy5 may attack amylose and open the starch granule so that Sas20 and Sca5 within the amylosome can bind, bringing its starch-active glycoside hydrolase family 13 (GH13) enzymes to the granule for further degradation. Conversely, amylosome components such as Sas20 and Sca5 may anchor the cell surface to the starch granule where amylosome GH13 enzymes begin starch digestion before Amy5 can “clean up” and degrade leftover amylose. These studies of how Amy5, Sas20, and Sca5 may cooperate at the interface between bacterium and starch granule helps us to understand broadly how *R. bromii* digests resistant starch for growth.

CHAPTER 1:

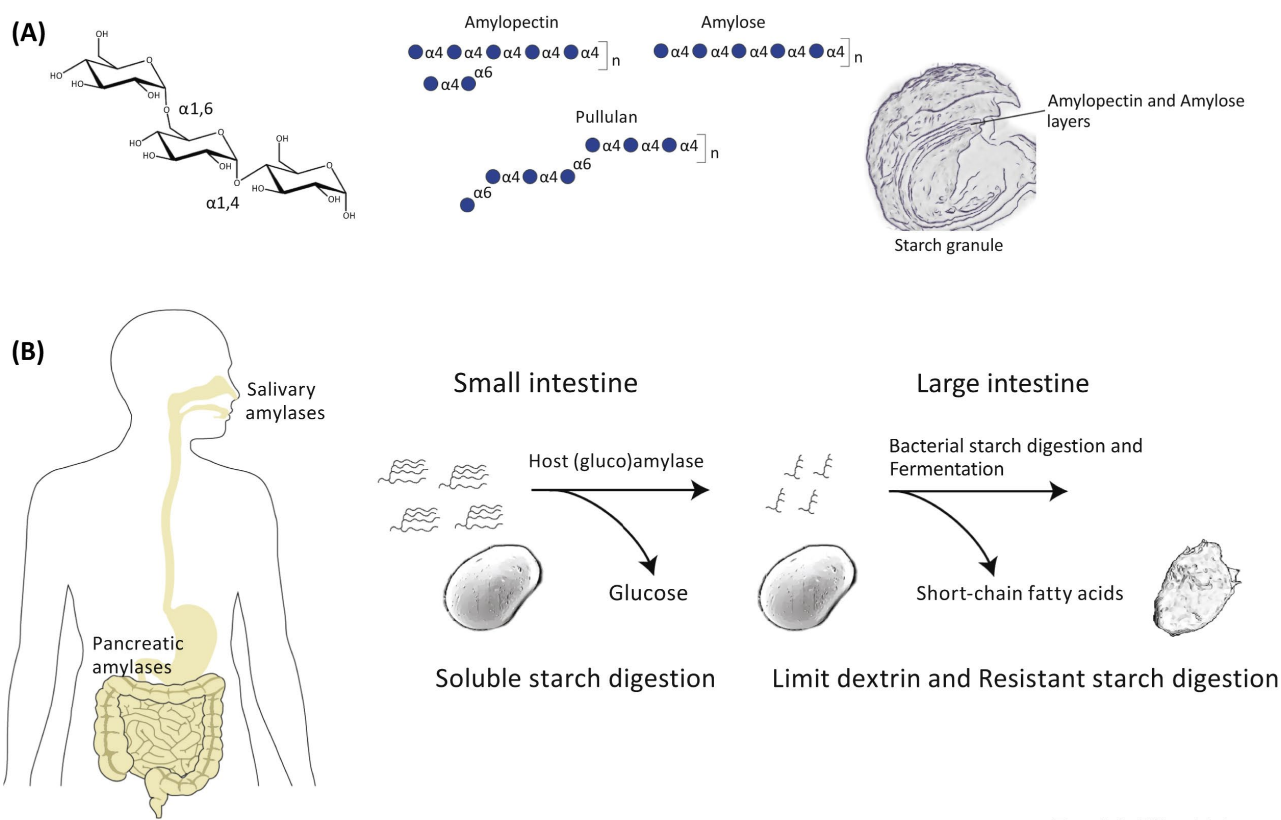
Introduction

The contents of this chapter have been published as: Cerqueira FM, Photenhauer AL, Pollet RM, Brown HA, Koropatkin NM. Starch digestion by gut bacteria: crowdsourcing for carbs. *Trends Microbiol*. 2020;28(2):95-108.

Starch in the Human Diet

The human gut microbial community shapes our health by driving immune development [1, 2] and intestinal maturation [3-5], protecting us from enteric pathogens [6, 7], and unlocking nutrients from our diet via the digestion of complex carbohydrates [8-10]. Assembly of this bacterial community begins at birth [11, 12] and is heavily influenced by the glycan landscape of the gut, comprised of host endogenous and dietary carbohydrates [13]. As the host diet matures, the composition of the bacterial community changes to include organisms that can degrade more complex dietary carbohydrates [11, 14, 15]. Starch is one of the most abundant carbohydrate components of a Western diet and one of the first plant-derived carbohydrate structures introduced in infant diets [16, 17]. Not surprisingly, bacteria with the potential to degrade starch are some of the first to appear in the gut [11, 14].

Starch, a polymer of glucose and a type of α-glucan, is abundant in many natural foods like potatoes, bananas, rice, and cereal grains. Amylose is a starch component comprised exclusive of α1,4-linked glucose, and this glycosidic linkage imparts a helical turn to the polymer [18]. These helices can pack tightly to form a crystalline lattice, making it resistant to enzymatic degradation. In contrast, amylopectin includes α1,6 linkages that form branch points along the α1,4-linked polymer [19] (**Figure 1.1A**). These branches enhance the solubility of the polymer. Starch granules are comprised of both amylose and amylopectin (**Figure 1.1A**), and the ratio of these two molecules affects digestibility by host and bacterial enzymes. Unlike other complex dietary carbohydrates, starch is one of few that is processed by host enzymes [20]. Host starch digestion begins with salivary amylase and proceeds through the small intestine where pancreatic amylases continue to digest the polysaccharide. At the brush border of the small intestine, maltase-glucoamylase and sucrase-isomaltase enzyme complexes liberate glucose from α-limit dextrins and oligosaccharides [21] (**Figure 1.1B**).



Trends in Microbiology

Figure 1.1 Starch Structure and Its Digestion along the Alimentary Canal
(A) Starch is composed of α1,4 linear and α1,6 branched glucose linkages. Starch granules, often resistant to host digestion, are composed of concentric rings of amorphous amylopectin and tightly packed helices of amylose. (B) Starch degradation is initiated in the oral cavity by the action of salivary amylases. Digestion continues via pancreatic amylases and small intestinal brush-border glucoamylases. The liberated glucose is absorbed through enterocytes in the small intestine. The remaining resistant starch (RS), as alpha limit dextrin or starch granules, is degraded in the colon by specialized bacterial species which release short-chain fatty acids for the host.

While much of the processed starch in our diet is degraded by host or bacterial enzymes in the small intestine, some dietary starch cannot be processed by host enzymes and transits the colon as resistant starch (RS), where it is digested by specialized members of the microbiota [22, 23]. Material left behind after the host has processed the polysaccharide includes α-limit dextrins, branched oligosaccharides enriched in α1,6 linkages, and RS [21, 24-27]. The precise amount of dietary starch that reaches the colon is highly variable and depends upon the origin of the starch consumed (e.g., potato or corn), its level of processing or cooking prior to consumption, and genetic factors such as the copy number of amylase genes in the host genome [28]. One study that directly measured the amount of dietary starch that passes into the human colon found that 2–20% of dietary starch escapes digestion in the small intestine [29]. In other studies, it has been determined that about 50% of granular or minimally processed starch escapes host digestion and transits to the colon [24, 30].

There are five different types of RS classified by their mechanism of resistance to host enzymes [31] (**Table 1.1)**. RSI and RSII describe starch that is naturally resistant to digestion, either due to inaccessibility within a seed or by virtue of the granular structure from tightly packed amylose and amylopectin layers. In contrast, RSIII, RSIV, and RSV describe starch that is resistant to host enzymes because of heat/cold or chemical treatment. These physical properties dictate accessibility and digestion by the host and the gut microbial community [23, 32]. In this review, we outline what is known about the biochemistry of starch degradation by different human colonic bacteria, how these microorganisms may synergize to utilize starch, and how this metabolism influences human health.

Table 1.1 Five Types of Resistant Starch

Type	Description	Example
RSI	Physically enclosed by kernel or cereal grains	Unmilled seeds
RSII	Tightly packed amylose and amylopectin helices within granules	Uncooked potato, corn, and green-banana flours
RSIII	Retrograded starch– starch that is heated then cooled (like cooking and refrigerating), causing starch to gelatinize or crystallize and become insoluble	Potato salad
RSIV	Chemically modified starch	FibersymRW (MGP Ingredients, Atchison, Kansas)
RSV	Lipid-modified starches from processing or cooking in oil	Fried rice

Starch as a Prebiotic Tool for Health

Because of the important interaction between RS and the gut microbiota, RS has been used as a prebiotic in human volunteer studies to manipulate the structure of the gut community. These studies found that the type (e.g., RSII vs RSIII) and source of RS (e.g., potato vs corn) drive different changes in the gut community. For example, RSII administration leads to increases in *Ruminococcus bromii* and Bifidobacteria, Gram-positive organisms that act as primary degraders of RS [33-36]. As these organisms hydrolyze starch granules, they cross-feed other bacteria, called secondary degraders, that scavenge the partially degraded granules or the solubilized starch fragments [35, 37, 38]. The collective metabolism of both primary and secondary degraders dictates the profile of fermentation by-products such as short-chain fatty acids associated with host health [35, 39].

Because of the crowdsourcing that occurs with glycan degradation and fermentation in the gut, the precise constellation of bacteria and metabolites that result from RS administration varies across feeding studies. However, some consistent observations across studies utilizing

RSII have included an increase in primary RS-degrading species, notably Bifidobacterium sp. and *R. bromii*, and an increase in butyrate-producing species including members of the Clostridium cluster XIVa [33-35, 39-41]. Interestingly, the source of RSII can dictate the overall response observed. Most recently, a study examining potato versus corn RSII consumption noted an increase in *R. bromii* with corn starch and an increase in Bifidobacteria with potato starch [35]. The primary degrader response in turn influenced the pattern of secondary degraders observed. *R. bromii*, which has an elaborate multiprotein system for degrading starch as discussed later, is also notably more abundant in human volunteer studies when RSIII is utilized and grows well *in vitro* on this substrate [37, 39, 42].

Dietary glycans such as RS are processed and fermented by gut bacteria into short-chain fatty acids, host-absorbable molecules that elicit several physiological effects [43, 44]. RS consumption typically increases the concentration of the short-chain fatty acid butyrate [45], which has profound anti-inflammatory and antitumorigenic effects [46]. Butyrate is the preferred energy source of colonocytes [47], and it strengthens the intestinal barrier against gastrointestinal pathogens through the suppression of virulence factors in *Salmonella* Typhimurium [48] and *Listeria monocytogenes* [49]. Butyrate induces differentiation of regulatory T cells in the colon [50], an immune subtype essential for commensal bacteria tolerance. In animal studies, butyrogenesis from RS feeding protects against chemically induced colorectal cancer [51, 52]. Most recently, a study revealed that a butyrate gavage of mice undergoing allogeneic bone marrow transplantation prevented the symptoms and severity of graft-versus-host disease post transplantation [53]. There is currently a human clinical trial underway to determine the efficacy of potato RS in preventing graft-versus-host disease following transplantation surgery (https://clinicaltrials.gov/ct2/show/NCT02763033). This suggests that this functional food has translational potential as a tool to improve health and treat disease.

Glucoamylase Inhibitors Enhance Starch Fermentation in the Gut

While RS consumption is a user-friendly, cheap, and noninvasive way to enhance starch transit to the colon, an alternative to RS consumption may be to shunt soluble starch to the distal gut with host glucoamylase inhibitors such as acarbose or voglibose [54]. These medications are FDA-approved for the treatment of prediabetes and type II diabetes as less glucose absorption in the small intestine effectively lowers post-meal blood glucose levels [54]. Beyond improvement in blood glucose levels from decreased glucose absorption, patients have enhanced butyrate output as a result of starch fermentation by the gut community [55-57]. In addition, individuals demonstrate decreased levels of proinflammatory cytokines that are independent of the blood-glucose-lowering effects of acarbose [58].

Whereas acarbose positively benefits type II diabetics, administration of this drug in mice consistently extends lifespan in a sex-dependent manner that is distinct from the effects of calorie restriction [59]. In fact, in a study across three different mouse facilities, diet supplementation with acarbose resulted in microbiota restructuring and increased butyrate and propionate output. Notably, while controlling for other factors, fecal concentrations of short-chain fatty acids were predictive of lifespan [60]. A similar study in mice that examined acarbose administration in the context of a low versus high fiber diet demonstrated increased short-chain fatty acid production, especially butyrate, in acarbose-treated mice on either diet [61]. The exact mechanism by which an increase in intestinal soluble starch alters the gut consortia is unclear; however, the change may not be exerted by the polysaccharide alone. Acarbose itself is minimally absorbed in the small intestine and thus may transit to the colon. Like humans, bacteria utilize glucoamylases to process starch, and these enzymes may also be subject to inhibition [62]. Given the established butyrogenic effects of acarbose, its use as a therapeutic beyond diabetes seems attractive. However, the possible starch-independent manner by which the molecule may restructure the gut community is due consideration given the numerous examples of both positive and negative effects of microbiota-mediated drug metabolism [63].

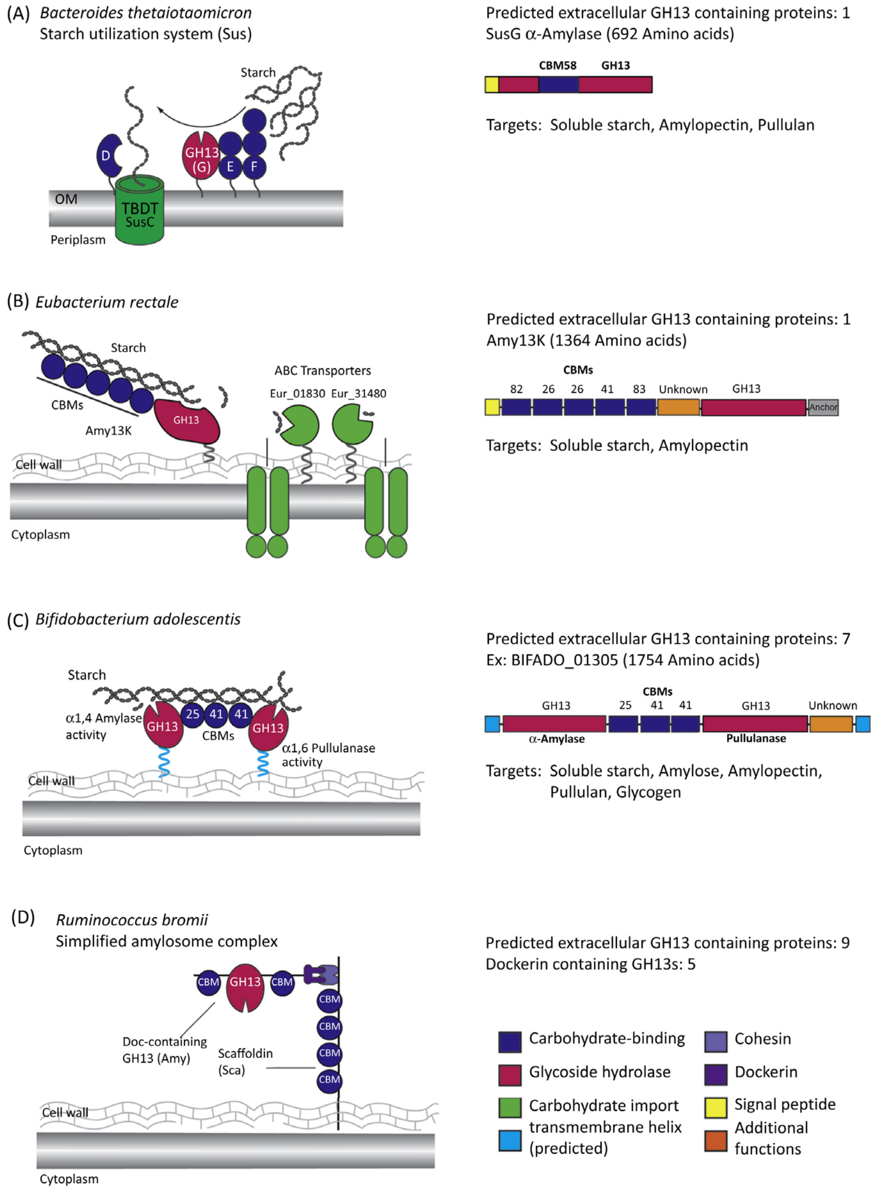
The Mechanics of Starch Breakdown by Bacteria

Because starch is a large polymer (10^5–10^8 Da), its initial deconstruction takes place at the cell surface by enzymes that can cleave the polysaccharide into glucose, maltose, or longer maltooligosaccharides for import into the bacterium [64]. Enzymes that target α-glucans fall into several glycoside hydrolase (GH) families based on sequence identity. A continuously updated database of known GHs and associated proteins that target carbohydrates is curated in the Carbohydrate Active Enzymes (CAZy) database (http://www.cazy.org/) [65]. The GH families 3, 13, 14, 15, 57, 119, and 126 are associated with starch degradation [66]. The GH13 enzyme family is one of the most abundant enzyme families found within gut bacteria and is most often associated with the initial bacterial processing of starch [20]. These are endo-acting enzymes that recognize internal regions of starch, rather than an end of the polymer, and can release a variety of maltooligosaccharides [66, 67].

The GH13 enzyme family is one of the largest and most well-studied carbohydrate-active enzyme families as starch-degrading enzymes are widespread among bacteria, fungi, animals, and plants. To date, there are 41 subfamilies for GH13 enzymes, which distinguish both the sequence and activity of the enzymes [68, 69]. The enzymatic and structural features of these subfamilies are the subject of several excellent reviews [66, 68-71]. Regardless of the precise α-glucan target of the enzyme, GH13s are generally comprised of a TIM barrel fold with one or more subdomains that may dictate specificity, oligomerization, or stability [70, 72, 73]. For simplicity, I will limit the mechanistic discussion to two general subtypes of GH13 enzymes: α-amylases and pullulanases. The α-amylases preferentially target α1,4 linkages within starch and a-glucans, while pullulanases target α1,6 linkages within starch and pullulan, a fungal polysaccharide composed of α1,6-linked maltotriose repeats [66, 71] (**Figure 1.1**). Some gut bacteria utilize both types of GH13 in concert, while others possess only one type that liberates oligosaccharides for import [74].

Some enzymes that attack starch harbor a GH13 catalytic domain in tandem with one or more carbohydrate-binding modules (CBMs) that allow the enzymes to dock to the substrate [75-81]. CBMs enhance catalytic efficiency via a proximity effect to position the substrate near the catalytic site [81, 82]. In addition, CBMs may aid in selecting regions of the starch structure that are more accessible to degradation [83]. To date, starch-binding CBM families in the CAZy database include CBM 20, 21,25, 26, 34, 41, 45, 48, 53, 58, 68, 69, 74, 82, and 83 [65]. While the families have distinct sequences, these domains generally display a β-sandwich-like fold and recognize the helical conformation of the α1,4-linked glucose polymer via an arc of at least two aromatic residues [77, 80, 84, 85].

The mechanism of starch digestion by gut bacteria is distinct depending on the physiology of the cell (Gram-negative or Gram-positive) and the type of starch the bacterium can access. In the following sections, we discuss the structure and organization of starch-active enzymes across gut bacteria phyla and how these features drive starch metabolism within the gut community.


(A) Bacteroides thetaiotaomicron
Starch utilization system (Sus)
Starch
D
GH13 (G)
E
F
OM
TBDT SusC
Periplasm
Predicted extracellular GH13 containing proteins: 1
SusG α-Amylase (692 Amino acids)
CBM58
GH13
Targets: Soluble starch, Amylopectin, Pullulan
(B) Eubacterium rectale
Starch
CBMs
Amy13K
GH13
ABC Transporters
Eur_01830
Eur_31480
Cell wall
Cytoplasm
Predicted extracellular GH13 containing proteins: 1
Amy13K (1364 Amino acids)
CBMs
82
26
26
41
83
Unknown
GH13
Anchor
Targets: Soluble starch, Amylopectin
(C) Bifidobacterium adolescentis
Starch
α1,4 Amylase activity
GH13
25
41
41
CBMs
GH13
α1,6 Pullulanase activity
Cell wall
Cytoplasm
Predicted extracellular GH13 containing proteins: 7
Ex: BIFADO_01305 (1754 Amino acids)
CBMs
GH13
25
41
41
GH13
Unknown
α-Amylase
Pullulanase
Targets: Soluble starch, Amylose, Amylopectin, Pullulan, Glycogen
(D) Ruminococcus bromii
Simplified amylosome complex
CBM
GH13
CBM
CBM
CBM
CBM
CBM
Doc-containing GH13 (Amy)
Scaffoldin (Sca)
Cell wall
Cytoplasm
Predicted extracellular GH13 containing proteins: 9
Dockerin containing GH13s: 5
Carbohydrate-binding
Glycoside hydrolase
Carbohydrate import
transmembrane helix (predicted)
Cohesin
Dockerin
Signal peptide
Additional functions

Trends in Microbiology

Figure 1.2: Examples of the Diverse Molecular Strategies Employed by Gut Bacteria for Starch Degradation (A) The starch utilization system (Sus) paradigm of Gram-negative *Bacteroides thetaiotaomicron* (left). several starch-binding proteins assemble around a TonB-dependent transporter (SusC) on the outer membrane (OM) surface to bind, hydrolyze, and import starch. The glycoside hydrolase (GH)13-containing protein, SusG, harbors a starch-binding carbohydrate-binding module (CBM)58 integrated within the catalytic domain primary sequence (right). (B) Amy13K from *Clostridium* cluster XIVa member *Eubacterium rectale* wherein multiple tandem CBMs cooperate to bind starch. ATP-binding cassette (ABC) transporter solute-binding proteins Eur_01830 and Eur_31480 capture released oligosaccharides for transport. (left). The linear architecture of Amy13K (right). (C) The amylopullulanase ApuB homolog from *Bifidobacterium adolescentis* has the potential to target both α-1,4 and α-1,6 linkages in starch (left). Unlike *B. thetaiotaomicron* and *E. rectale, B. adolescentis* has seven predicted extracellular GH13s. The linear arrangement of the ApuB amylopullulanase (right). (D) *Ruminococcus bromii* amylosome model in which CBMs and GH13 domains are expressed on different dockerin-containing polypeptides and assemble via binding to the cohesin domain on scaffoldin proteins.

Bacteroides Starch Utilization System (Sus)

Bacteroidetes, the dominant Gram-negative phylum of bacteria within the mammalian gut, organize genes encoding carbohydrate-active enzymes within polysaccharide utilization loci (PUL) [86, 87]. PULs encode all the necessary machinery for the initial cell-surface deconstruction of a polysaccharide as well as the import machinery. PULs are typically comprised of genes encoding cell surface proteins with discrete functions such as enzymatic activity or glycan-binding [88-90].

One of the first identified PULs was the starch utilization system (Sus) of *Bacteroides thetaiotaomicron* [91, 92]. The *B. thetaiotaomicron* Sus includes several cell-surface proteins (SusDEF), a TonB-dependent transporter (SusC), and three enzymes (SusABG) (**Figure 1.2A**) [93]. The Sus proteins work together to capture and degrade starch at the cell surface and subsequently import the liberated maltooligosaccharides into the periplasm for further digestion [94, 95].

B. thetaiotaomicron accesses starch in the environment via the cell-surface proteins SusCDEFG. SusE and SusF are multidomain starch-binding proteins that are not essential for growth on starch. SusE and SusF collectively contain five starch-binding domains which display a canonical Ig-like/β-sandwich fold found in CBMs appended to some starch-active enzymes. Despite this structural similarity, each site has distinct affinity for unbranched α1,4 maltooligosaccharides. SusE enhances the ability of the cell to import maltooligosaccharides of 17–30 glucose residues long, whereas cells lacking this protein access smaller oligosaccharides. Importing multiple glucose equivalents in a single event may be energetically beneficial for the organism.

While the surface-binding and enzymatic proteins such as SusEFG differ across PULs, SusC and SusD homologs are nearly ubiquitous within Bacteroidetes PULs. SusD contains a single starch-binding site and is essential for growth of *B. thetaiotaomicron* on starch polymers longer than five glucose units. The SusD binding site is critical for the optimal transcription of sus in response to malto-oligosaccharides but a binding-deficient mutant of SusD can grow on starch in the presence of maltose to upregulate the *Sus* locus. The recent crystal structures of two homologous SusCD-like transport systems in *B. thetaiotaomicron* revealed that the SusD-like protein is positioned over the top of the SusC-like TonB-dependent transporter in both structures. These structures coupled with molecular dynamics and in vitro experiments with the reconstituted transporters support a ’pedal-bin’ mechanism of transport whereby the SusD captures substrate and then closes over the top of SusC to facilitate import. The structure of the SusCD-like complex is likely to be highly conserved across the Bacteroidetes PULs, though we hypothesize that the unique features of the accompanying surface glycan proteins, such as SusEF, and the surface enzyme, such as SusG, will provide nuance to each system and tailor the precise assembly based upon the substrate. In the case of the *B. thetaiotaomicron* Sus, single-molecule fluorescence imaging of SusE, SusF, and SusG suggests a dynamic assembly of these components during the catabolism of starch.

SusG is the sole outer membrane enzyme of the system and it cleaves starch into maltooligosaccharides that can be imported through SusC [85]. SusG is comprised of a GH13 domain and a starch binding CBM58 domain that is inserted into the middle of the GH13 polypeptide sequence [85]. Deletion of CBM58 from SusG does not affect *B. thetaiotaomicron* growth on starch [94]. In fact, in vitro assays with the recombinant SusGΔCBM58 mutant showed higher activity against soluble starch and amylopectin than the wild-type enzyme, suggesting that CBM58 may be more important for accessing insoluble substrates [85]. The GH13 domain selectively cleaves α1,4-glycosidic linkages but has the somewhat unique distinction of accommodating α1,6 branch points within maltooligosaccharides and α1,6-linkages within the unbranched polysaccharide pullulan [96]. This flexible recognition of limit dextrins or α1,6 branch points within starch may be beneficial as these products are left behind by host enzymes and transit the gut [25, 26].

Many sequenced human gut Bacteroidetes possess a PUL that has some synteny with the *B. thetaiotaomicron* Sus, although the number of surface glycan-binding proteins like SusEF and the homology of these proteins differ, as does the size of the predicted SusG [97]. An alternative Sus has been identified in *B. thetaiotaomicron* that may target maltooligosaccharides [98]. This PUL, unlike Sus, does not possess a predicted cell-surface enzyme, and instead has periplasmic GH31 and GH97 α-glucosidases that target smaller oligosaccharides (approximately three to seven glucose units) [98]. While the *B. thetaiotaomicron* Sus is the only PUL in this organism required for starch degradation, the alternative Sus may facilitate the uptake of smaller oligosaccharides that could be liberated by starch-degrading organisms within the gut environment.

Although gut Bacteroidetes are prolific degraders of complex carbohydrates from the plant cell wall or endogenous mucosal glycans, they have not been observed to directly degrade RS without the help of other organisms. One study in which mice were fed stable-isotope (U13C)-labeled native potato starch revealed that gut Bacteroidetes incorporated the isotope into their RNA, supporting the notion that these organisms are a part of the metabolic cascade that results from RS degradation [99]. Members of the Bacteroidetes likely access more soluble forms of starch from incomplete digestion in the small intestine or soluble starch and oligosaccharides released by primary RS-degraders. Parabacteroides species have been observed to increase in abundance when the host diet is supplemented with RSIV [36, 100], but there is no direct evidence that these bacteria degrade this fiber.

The Sus paradigm as described is restricted to the Bacteroidetes phylum. Gram-positive organisms such as the Firmicutes and Actinobacteria may also encode polysaccharide utilization loci, named gpPUL. These encode GHs for degradation of a large glycan at the cell surface, as well as an ATP binding cassette (ABC) transporter that is specific for the oligosaccharides liberated by the surface enzyme [101]. Unlike the Bacteroidetes PUL, there is more variation in the organization of genes encoding the enzymes and transporters across Gram-positive species. There is also significant variation in gpPUL presence and in the starch-degradation strategies among Gram-positive human gut species. Below, we discuss three different themes in starch digestion by Gram-positive organisms that contribute significantly to RS digestion and its downstream effect on the microbiota and host health.

***Clostridium* Cluster XIVa Starch-Scavenging Enzymes**

Butyrate-producing organisms of the *Clostridium* cluster XIVa often show an increase in abundance in human RS feeding trials and *in vitro* growth of gut bacterial communities on RS [39, 102]. Because of this and the physical association of these bacteria with starch, it was believed that these organisms were primary degraders of RS [103-105]. However, a landmark study demonstrated that *Eubacterium rectale*, an abundant butyrate-producing Clostridia, did not directly degrade RSII or RSIII but rather was adept at scavenging the by-products of RS degradation that was initiated by *R. bromii* (discussed later) [37]. The high correlation of the *Clostridium* cluster XIVa with RS feeding suggests that these organisms play an important role in complete starch degradation in the gut and are responsible for the increase in butyrate.

Clostridium cluster XIVa organisms, including *Roseburia inulinovorans*, *Butyrivibrio fibrisolvens,* and *E. rectale*, have starch-active enzymes and ABC transporter genes in gpPUL that are expressed when starch is present in the environment [106-108]. These bacteria each possess a cell-wall anchored GH13 amylase that includes one or more CBM26 domains and Ig-like domains that may be unidentified starch-specific CBMs or play a structural role in starch binding and degradation (**Figure 1.2B**) [108]. Interestingly, the arrangement of the CBM and catalytic domains is distinct between these organisms. In *B. fibrisolvens*, the catalytic domain is located at the N terminus and followed by CBMs, while in *R. inulinovorans*, the catalytic domain is located at the C terminus, preceded by an array of CBM domains [108]. Whether these different arrangements affect starch hydrolysis is unknown. The specificity and affinity of individual CBMs for varying types of starch was reported for *E. rectale* (**Figure 1.2B**). This prominent butyrate-producing organism expresses a cell wall-anchored GH13 α-amylase, Amy13K, that contains five starch-specific CBMs [80]. Interestingly, these CBMs are important for enzyme docking to corn but not potato starch [80]. Recombinant Amy13K lacking these CBMs retained activity on soluble starch substrates, but degradation of corn RS granules was reduced [80]. Although *E. rectale* is not an RS-degrading bacterium, the CBMs of Amy13K, and similar enzymes from related Clostridia, may allow bacteria to localize to starch particles that have been partially processed by RS degraders. This may provide a benefit to these non-RS degrading species by allowing them access to a privileged nutrient niche.

Bifidobacterial Multimodular GH13s

Bifidobacteria are among the initial colonizers of the human gastrointestinal tract [109] driven primarily by their ability to utilize host-produced glycans such as human milk oligosaccharides and intestinal mucins [110]. For many dietary glycans, such as galacto- and arabinoxylo-oligosaccharides, these organisms deploy selective and high-affinity solute-binding proteins as part of ABC transporters that allow them to scavenge these nutrients in the gut ecosystem [111]. Bifidobacteria are primary RS degraders and contribute significantly to the starch-degrading pathway leading to butyrate [37, 112].

Bifidobacteria that degrade RS appear to have multiple predicted cell-surface GH13s [112, 113]. For example, *Bifidobacterium adolescentis* L2-32 encodes seven extracellular starch-specific GH13 enzymes (**Figure 1.2C**). These multimodular GH13 enzymes include multiple CBMs, and at least one of these enzymes includes a CBM74, a family first discovered in *Microbacterium aurum* [79]. The *M. aurum* CBM74 domain binds to RSII, RSIII, soluble potato starch, amylose, and amylopectin. A sequence homology search to identify CBM74 homologs within other bacterial genomes revealed 46 such domains in *Bifidobacterium* species. Of interest, there is one protein from *B. adolescentis* and *R. bromii* that contains a CBM74 domain. The CBM74 domain is almost always encoded on the same polypeptide with the starch-binding domains CBM25 or CBM26 [114]. As demonstrated in a number of starch-hydrolyzing enzymes, including those from *Bacillus halodurans* [77] and *Streptococcus pneumoniae* [75], tandemly arranged CBMs from families 25, 26, and 41 provide an avidity effect that enhances starch hydrolysis. It is likely that CBMs from families like 25, 26, and 74 work synergistically to dock to starch granules.

Complete starch degradation is thought to require the cooperative activity of catalytic modules that can hydrolyze both the α1,4 and α1,6 linkages that compose starch. A dual amylopullulanase enzyme, ApuB, from *Bifidobacterium breve* UCC2003 [115] shows activity on potato starch, amylopectin, glycogen, and pullulan [116], and inactivation of apuB eliminates growth on these substrates. ApuB is composed of an N terminal α-amylase and a C-terminal pullulanase separated by a CBM25 and two CBM41 domains. The recombinantly expressed α-amylase domain with the CBM25 and CBM41 hydrolyzes soluble starch, amylopectin, and glycogen while the pullulanase domain with one CBM41 hydrolyzes the α1,6 linkages of pullulan. The proximity of these catalytic activities in a single polypeptide chain likely facilitates their cooperation during starch hydrolysis [116]. To date, these multidomain amylopullulanases have only been identified in RS-degrading gut bacteria and soil isolates.

Ruminococcus bromii **Amylosome**

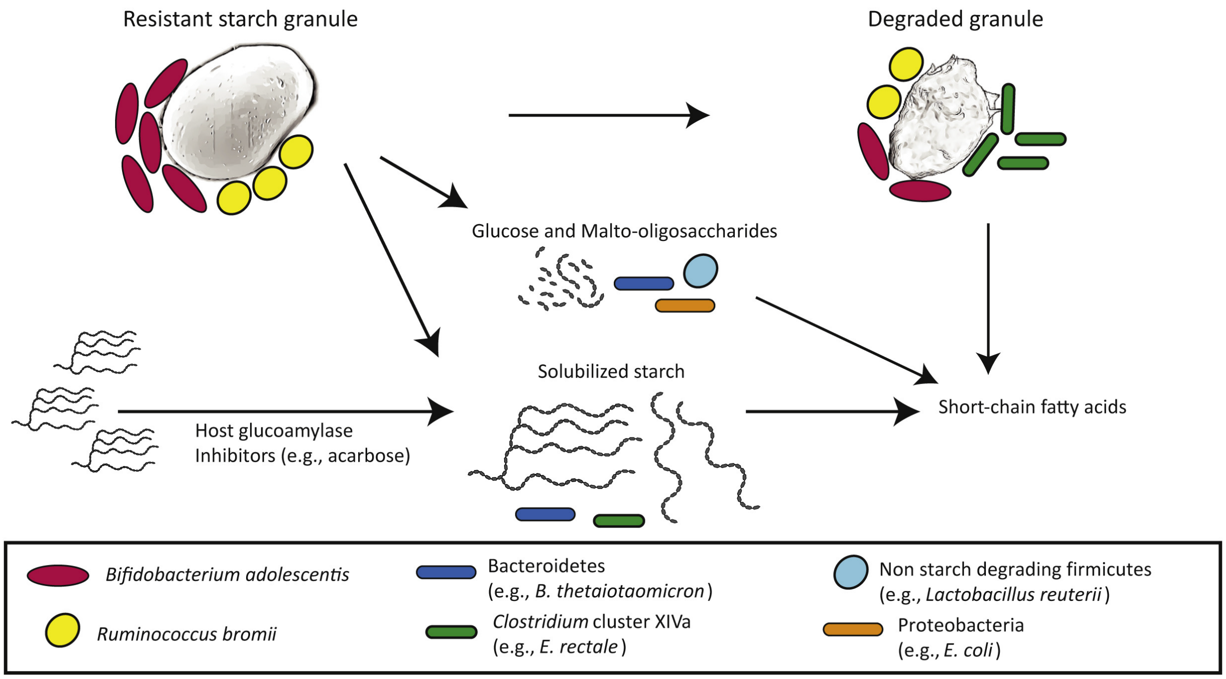
In vitro studies of RS digestion by *R. bromii* in monoculture demonstrate an abundance of starch degradation by-products such as glucose, maltose, and maltotriose in spent media [37]. This suggests that the efficiency of starch digestion by the bacterium exceeds what is required to support its growth. Indeed, secondary degraders such as *B. thetaiotaomicron* and *E. rectale* thrive in cell free media from *R. bromii* monocultures grown in RS and in coculture studies [37]. In a five-membered bacterial consortium grown on RSIII, *R. bromii* increased the abundance of *E. rectale* as well as butyrate levels [37]. These data suggest that *R. bromii* degrades RS and provides by-products to other bacteria, in contrast to *B. adolescentis* that seems to utilize RS more completely [35, 37].

A detailed analysis of the *R. bromii* genome identified 21 predicted GHs, including 17 that belong to the GH13 family, suggesting specialization for starch degradation [42]. A unique feature of five of these GH13s is the inclusion of a C-terminal dockerin domain [37, 117]. Dockerins bind to cohesin domains which are typically found in structural proteins called scaffoldins [118]. Cohesin–dockerin interactions were first discovered within multiprotein cellulose-degrading complexes called cellulosomes of cellulolytic bacteria [119]. In cellulosomes, GHs and binding proteins come together like molecular legos via cohesin–dockerin binding, and this close proximity facilitates cellulose degradation.

The cohesin–dockerin interaction is one of the strongest protein–protein interactions observed in nature, with the highest recorded K_a constants on the order of 10^{11} M [120-122]. Additionally, cohesin–dockerin binding increases the mechanical stability of the individual proteins, which is thought to be required for the degradation of crystalline substrates [16]. In *R. bromii*, dockerin-containing enzymes and scaffoldins comprised of CBMs and cohesin domains have been identified, suggesting that RS degradation is mediated by a multiprotein complex analogous to the cellulosome, termed the amylosome (**Figure 1.2D**) [37, 42].

Phylogenetic studies among several *R. bromii* isolates have confirmed that amylosome components are a conserved feature of this bacterium [42]. Studies on recombinant dockerins and cohesins from amylosome proteins has revealed that some dockerins, like those of amylases Amy4 and Amy9, bind several different cohesins, while dockerins from other proteins, like the pullulanase Amy12, are more specific [42, 117]. Although the number and type of amylosome complexes made during growth on RS is unknown, amylosome assembly on the cell surface is calcium-dependent, as cohesin–dockerin binding requires calcium. The amylosome complex likely provides the ideal juxtaposition of CBMs and enzymes for highly efficient degradation of RS.

The first detailed structure-function study two predicted pullulanases incorporated into the amylosome, Amy10 and Amy12, demonstrated how the amylosome utilizes multiple proteins to optimize RS degradation [123]. While they possess similar active sites, Amy10 is hyperspecialized at breaking α1,6 bonds within starch while Amy12 can catalyze cleave both α1,4 and α1,6 linkages with a preference for the latter. This suggests that these proteins are not redundant but may have different roles in the breakdown of RS. Both pullulanases have starch-binding CBMs and MucBP modules which likely facilitate binding to substrate and anchoring of the bacterium to the lining of the colon, respectively. However, the amylosome is not the only means by which *R. bromii* cells can bind and degrade starch. *R. bromii* expresses predicted extracellular GH13s, such as the recently reported Amy5, that lack cohesins or dockerins, and these enzymes are also likely to contribute to RS degradation [124].



Trends in Microbiology

Figure 1.3: Proposed Model of Resistant Starch (RS) Utilization by the Gut Community. Specialized primary degraders such as *Ruminococcus bromii* and Bifidobacterium adolescentis initiate the breakdown of RS and subsequently release soluble substrates that can be used by a number of species. Smaller sugars, including glucose, released by this process could support the growth of species like *Escherichia coli* and *Lactobacillus reuteri* that cannot access starch. Solubilized starch and longer oligosaccharides can be accessed by Clostridia like *Eubacterium rectale* or Bacteroidetes such as *Bacteroides thetaiotaomicron*. Digestion of RS by primary degraders may increase the accessible surface area of the granules, allowing other starch-degrading species to bind and access this nutrient niche.

Conclusion

The compete metabolism of starch by intestinal bacteria is a team effort mediated by the synergy between primary RS degraders and secondary starch scavengers (**Figure 1.3**). As RS is degraded by *R. bromii* and *B. adolescentis*, the physical structure of the granule must change. We hypothesize that these changes may make the structure more amenable to docking by the Clostridium cluster XIVa, explaining their attachment to digested starch particles. At the same time, it is likely that maltooligosaccharides of varying lengths are released from the granule and cross-feed several species, including but not limited to *Lactobacillus reuteri* and *Escherichia coli* that cannot degrade starch but can scavenge maltooligosaccharides [125, 126]. The precise metabolic interactions within this food web are likely dictated by the type of starch and the glycan degradation strategy employed by individual bacteria. Furthermore, additional bacterial metabolites, including fermentation end products, likely drive changes in the community structure irrespective of starch digestion [35].

How non-bacterial members of the gut microbiota, such as fungi and protists, contribute to primary RS degradation or this cycle of starch catabolism is largely unknown. Thus far we are limited in our ability to predict RS degradation from genomic data, as the sequence-specific features within GH13-containing enzymes and their associated CBMs that dictate RS degradation are not well established. Moreover, it is likely that the assembly of these features together (e.g., in the amylosome) on the cell surface facilitates starch recognition and digestion. One recent human RS feeding study reported an increase in the abundance of an uncultured species related to *Clostridium chartatabidum*, a ruminal fiber-degrading species [127], noting that this organism displayed similar response dynamics as known primary RS degraders [35]. Indeed, it is very likely that additional RS-degrading microbes contribute beyond the specific bacteria discussed here. As we better understand the mechanics of starch degradation and community features that support its digestion, we will be able to better predict how all members of the community contribute to this process.

Diet is a noninvasive way to change the microbiota towards improved health. Starch, particularly RS, is one prebiotic fiber that seems to consistently enhance beneficial butyrate output. As part of the next era of microbiome research it will be important to delve further into the molecular details driving these changes to prescribe and predict microbiome responses to functional foods.

Chapter Outline

The overall goal of this thesis work aims to uncover biochemical features of starch-active proteins expressed by RS degraders. By contrasting these features to those found in starch-active proteins from bacteria that cannot utilize RS as a sole carbon source, we can then understand the typical biochemical features of proteins that may confer RS degradation.

In Chapter 2, we describe the structure and function of ErAmy13B from *E. rectale* and RbAmy5 from *R. bromii*, two GH13 enzymes from subfamily 36 (GH13_36). This subfamily of enzymes hydrolyzes α1,4 linkages exclusively but can accommodate α1,6 linkages, and their major breakdown products are glucose and maltose. Generally, GH13_36 enzymes prefer longer substrates (longer than three glucose residues). We found that their limited binding pocket extending from -2 through +2 subsites likely contributes to maltose being their main product. Despite binning into the same GH13 subfamily and sharing structural homology, RbAmy5 has a typical GH13_36 substrate preference with its highest catalytic efficiency on amylose, while ErAmy13B is most efficient at degrading maltohexaose. We then show a broader comparison between GH13_36 and other maltogenic amylase subfamilies to explain how their activity profiles are influenced by their structures.

In Chapter 3, we perform an in-depth structure-function characterization of Sas20, a putative member of the *R. bromii* amylosome system. We pair functional assays with x-ray crystallography and small-angle x-ray scattering to reveal that Sas20 is a highly flexible starch-binding protein that helps direct the amylosome to more soluble regions of the starch granule.

Finally, in Chapter 4 we compare the crystal structures of the two cohesin modules of the scaffoldin protein, Sca5, from the *R. bromii* amylosome system. While the Sca5 cohesins display a typical cohesin module jelly-roll fold, they do not bin into canonical type-I or II binding paradigms. We then used AlphaFold-Multimer to predict how the dockerin module from Sas20 binds to the second cohesin module from Sca5.

In Chapter 5, we propose a model of how Sas20, Sca5, and Amy work together at the *R. bromii*-starch granule interface to efficiently degrade RS, and we suggest future work to clarify and resolve this model.

CHAPTER 2:

The Structures of the GH13_36 Amylases from *Eubacterium rectale* and *Ruminococcus bromii* Reveal Subsite Architectures that Favor Maltose Production

The contents of this chapter have been published as: Cockburn, D, Cerqueira FM, Bahr, C, Koropatkin, N. (2020) The structures of the GH13_36 amylases from *Eubacterium rectale* and *Ruminococcus bromii* reveal subsite architectures that favor maltose production. *Amylase.*

ABSTRACT

Bacteria in the human gut including *Ruminococcus bromii* and *Eubacterium rectale* encode starch-active enzymes that dictate how these bacteria interact with starch to initiate a metabolic cascade that leads to increased butyrate. Here, we determined the structures of two predicted secreted glycoside hydrolase 13 subfamily 36 (GH13_36) enzymes: ErAmy13B complexed with maltotetraose from *E. rectale* and RbAmy5 from *R. bromii.* The structures show a limited binding pocket extending from –2 through +2 subsites with limited possibilities for substrate interaction beyond this, which contributes to the propensity for members of this family to produce maltose as their main product. The enzyme structures reveal subtle differences in the +1/+2 subsites that may restrict the recognition of larger starch polymers by ErAmy13B. Our bioinformatic analysis of the biochemically characterized members of the GH13_36 subfamily,

which includes the cell-surface GH13 SusG from *Bacteroides thetaiotaomicron*, suggests that these maltogenic amylases (EC 3.2.1.133) are usually localized to the outside of the cell, display a range of substrate preferences, and most likely contribute to maltose liberation at the cell surface during growth on starch. A broader comparison between GH13_36 and other maltogenic amylase subfamilies explain how the activity profiles of these enzymes are influenced by their structures.

Introduction

Eubacterium rectale is an important member of the healthy human gut microbiota, producing butyrate as one of its primary fermentation products during growth on carbohydrates [128]. Butyrate production by *E. rectale* and other gut Firmicutes is associated with a number of health benefits, including reduced risk of colon cancer, decreased inflammation, and improved gut barrier function, among others [129]. Indeed, *E. rectale* and other butyrate-producing organisms are often found in reduced relative abundance in a variety of disease states, such as in colon cancer [130], inflammatory bowel disease [131] or diabetes [132]. Thus, understanding how these bacteria persist and acquire nutrients within the host provides a path towards manipulating the microbiome to improve their growth and either maintain health or help treat diseases.

E. rectale utilizes a variety of carbohydrates released from the breakdown of complex dietary fiber substrates found in the diet such as resistant starch (RS) [35, 37]. RS is not efficiently degraded by host amylases and therefore travels along the gastrointestinal tract to the colon where it can be degraded by the few gut bacterial species that are equipped for its direct degradation. Five different types of dietary RS have been defined according to the physical or chemical structure of the polysaccharide [31]. A diet supplemented with RS2 (naturally granular starch) or RS4 (chemically cross-linked starch) tends to result in an increase in colonic butyrate levels. Moreover, dietary supplementation with RS2, RS4 or RS3 (retrograded starch) can result in an increase in the abundance of primary RS degraders such as *Ruminococcus bromii*, and butyrate producers such as *E. rectale* [33, 35, 39]. *R. bromii* can directly degrade RS, while *E. rectale* cannot [37], which has been largely attributed to *R. bromii* encoding an amylosome, a multi-protein complex comprised of starch-binding proteins and starch-degrading enzymes [42, 117]. *R. bromii* degrades RS and releases mostly mono-, di- and maltooligosaccharides and acetate, which *E. rectale* can then use to produce butyrate [37]. However, the identities and biochemical characteristics of all proteins involved in this metabolic cascade are unclear.

Like many other Firmicutes, *E. rectale* makes use of a growth strategy dependent on scavenging the products of fiber digestion by primary fiber-degrading organisms through a suite of ATP-binding cassette (ABC) transporters targeting a variety of these carbohydrate products [74]. One of the few polysaccharides that *E. rectale* can degrade is soluble starch, an activity likely driven by the large cell wall-anchored amylase ErAmy13K [107]. ErAmy13K contains at least five carbohydrate binding modules (CBMs), two of which are from CBM families 82 and 83 that seem to be restricted to *E. rectale* and a few closely related *Roseburia* species [80]. We have previously demonstrated that these CBMs dictate the specificity of ErAmy13K and thus the types of starch that *E. rectale* utilizes for growth. Our proteomics work revealed that along with ErAmy13K, a second amylase, ErAmy13B, was upregulated in the cell wall/cell membrane fraction of *E. rectale* during growth on starch in comparison to growth on glucose [107]. ErAmy13B is part of a putative operon with an ABC transporter specific for longer oligosaccharides. Our activity profiling revealed that ErAmy13B acts as a maltogenic amylase (EC 3.2.1.133) producing maltose from both poly- and oligosaccharides.

The activity of ErAmy13B was surprising based upon its classification within the GH13_36 subfamily in the Carbohydrate-Active EnZymes database (http://www. cazy.org/) since most known maltogenic amylases reside within the GH13_20, GH13_21 and GH13_2 subfamilies. The GH13_20 and GH13_21 maltogenic amylases in particular contain an additional N-terminal domain (CBM34) [69] that is thought to be important for their product specificity through restriction of the active site cleft [133]. There are no CBM34 domains in recently assigned GH13_36 enzymes including ErAmy13B, SusG of *Bacteroides thetaiotaomicron* (BtSusG), and the secreted amylase Amy5 from *Ruminococcus bromii* (RbAmy5), the activity of which was recently reported [124]. Despite classification into the same subfamily, ErAmy13B, BtSusG and RbAmy5 display different activity profiles. RbAmy5 displays its maximum specific activity on amylose and amylopectin with weaker activity against cyclodextrins (CDs), similar to BtSusG, while ErAmy13B displays maximal activity on β-CD and is weakly active on amylose and amylopectin (5% and 23% maximum specific activity, respectively). The BtSusG structure was previously determined and is larger than both ErAmy13B and RbAmy5 due to the insertion of a CBM58 within the B-domain of the amylase fold [85]. However, ErAmy13B (564 amino acids) and RbAmy5 (551 amino acids) share ~36% sequence identity and do not have additional CBMs, making for a potentially interesting structural comparison of the salient structural features within this subfamily.

Here we use a combination of X-ray protein crystallography, enzymology, and sequence analysis to understand the structural basis for the maltogenic activity of ErAmy13B and resolve its respective role in starch scavenging by *E. rectale*. We also determined the X-ray crystal structure of RbAmy5 to identify the molecular features that may impart the difference in activities of the enzymes within this GH13 subfamily.

RESULTS

Activity analysis of Amy13B via ITC

We previously demonstrated via reducing sugar assay that ErAmy13B displays its highest specific activity against β-CD (6.14 ± 0.51 μmol min^{-1} mg^{-1}) followed by maltotriose (4.11 ± 0.46 μmol min^{-1} mg^{-1}) then glycogen (2.60 ± 0.21 μmol min^{-1} mg^{-1}) and amylopectin (1.43 ± 0.15 μmol min^{-1} mg^{-1}) [107]. Minimal activity was observed against amylose (0.29 ± 0.03 μmol min^{-1} mg^{-1}), pullulan (0.58 ± 0.49 μmol min^{-1} mg^{-1}), and trace activity against corn starch or α-CD [107]. Because these data suggest a preference for oligosaccharides, we sought to define the optimal length of substrates for this enzyme. Assays on non-labelled oligosaccharides can be challenging as a method of detection is required that is not overwhelmed by the relatively high molar concentrations of the substrate required. Reducing sugar assays are usually unsuitable for this reason, unless reduced variants are available, which is the case for maltotriose, but not for longer maltooligosaccharides. HPLC-based methods can be used but require a carbohydrate detection method, such as electrochemical detection or mass spectrometry. Thus, we developed an assay using ITC to derive the kinetic parameters of ErAmy13B. This has previously been performed for a few enzymatic activities, including urease, protease, and hexokinase [134, 135] as well as for chitinases [136], but this is the first instance that we are aware of where kinetic parameters of an amylase have been measured by ITC.

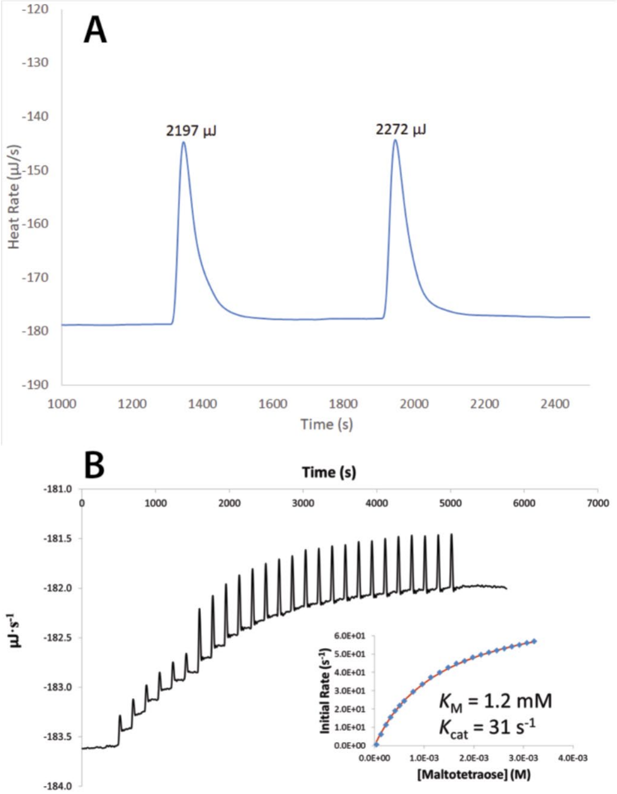


Figure 2.1: Example of ITC data for enzyme kinetics determination. (A) Determination of molar enthalpy by measuring the complete conversion of maltotetraose to maltose by a large excess of ErAmy13B. (B) ITC trace from a kinetics experiment with ErAmy13B and maltotetraose. Note that it is not the peak areas, but rather the change in baseline between injections that is measured, indicating increased heat released as the reaction rate increases substrate concentration. Heat rates are converted to reaction rates using the previously determined molar enthalpy of the reaction. Inset is the fitting of these calculated reaction rates and substrate concentration to the Michaelis-Menten equation, deriving values for k_{cat} and K_M

In this assay, only the energy released upon hydrolysis is detected and plotted, thus sensitivity is tied to the amount of energy released per reaction. An initial experiment with a large excess of ErAmy13B and maltotetraose, which was allowed to go to completion (**Figure 2.1A**), produced only maltose as previously reported and revealed that the energy released was 4.47 kJ/mol, similar to what had been determined with α-glucosidase to measure enthalpy change upon maltooligosaccharide hydrolysis [137]. Here we are measuring and plotting the change in baseline in between injections, which reflects the initial rate of the reaction for determining kinetic parameters (**Figure 2.1B**). While the energy released from glycosidic bonds is not as large as that for other reactions that have been studied by ITC in the past, we determined that it should be sufficient for determining enzymatic activity in a continuous manner with substrate concentration increasing with each injection. One challenge with this methodology is keeping the total progress of the reaction low to ensure that measurements are taking place within the initial reaction rate range. Ideally, total substrate conversion should remain below 1% and while this was achieved for many of the reactions, it was not always possible at the lower substrate concentrations of low K_M substrates, where there is a tradeoff between conversion amount and having sufficient signal to measure the activity, and so a revised upper limit of 5% conversion was used. Using this method, we determined that maltohexaose was the best substrate for ErAmy13B as measured by catalytic efficiency (k_{cat}/K_M), with activity decreasing at higher and lower length substrates (**Table 2.1**).

Table 2.1: Activity of ErAmy13B towards maltooligosaccharides as determined by ITC.

Substrate	k_{cat} **(s^{-1})**	K_M **(mM)**	k_{cat}/K_M
Maltotriose	2.96 ± 0.96	3.42 ± 1.50	0.87
Maltotetraose	31.33 ± 3.79	1.22 ± 0.27	25.68
Maltopentaose	46.00 ± 3.61	1.73 ± 0.25	26.59
Maltohexaose	91.00 ± 18.25	1.64 ± 0.20	55.49
Maltoheptaose	126.5 ± 45.96	3.05 ± 0.35	41.48

However, activity was similar for maltotetraose through maltoheptaose, with a large drop off occurring for maltotriose. The best substrate based upon k_{cat}/K_M is maltohexaose. In our previous work, we demonstrated that the hydrolysis of maltohexaose by ErAmy13B results in maltose, maltotriose and maltotetraose [107]. Similar, the hydrolysis of maltotetraose through maltoheptaose yields maltose as the smallest product, underscoring the strict requirement for occupancy of the +2 subsite [107]. Interestingly, the substrate profile for ErAmy13B is similar to the binding profile displayed by EUR_01830 [107], the solute binding protein of the ABC transporter located in the same genetic cluster as ErAmy13B.

Crystal Structure of ErAmy13B in Complex with Maltotetraose

To obtain a complex of ErAmy13B productively bound to substrate, we created the D265A inactive mutant enzyme via site-directed mutagenesis of the catalytic nucleophile. Initial attempts to crystallize wild-type or D265A ErAmy13B produced only poorly diffracting crystals and thus we employed surface entropy reducing mutations, replacing predicted neighboring lysines K543 and K544 with alanine. The resulting D265A/K543A/K544A ErAmy13B co-crystallized with maltotetraose produced crystals with strong diffraction and the structure was determined to a resolution of 2.25 Å (R_{work}= 16.2%, R_{free}= 19.8%; **Table 2.2**).

Table 2.2: X-ray data collection and refinement statistics. Values in parentheses indicate highest resolution shell.

	ErAmy13B with maltotetraose	**RbAmy5**
PDB ID	7JJN	7JJT
Wavelength (Å)	0.979	0.979
Resolution range (Å)	38.07–2.25 (2.33–2.25)	40.17–1.66 (1.719–1.66)
Space group	P 1 21 1	P 41 21 2
Unit cell (Å)	*a*=84.2, *b*=82.8, *c*=85.9, b=92.9°	*a*=*b*=98.5, c=196.2 α,β,γ =90°
Total reflections	388822 (39761)	1116162 (110864)
Unique reflections	55448 (5536)	114177 (11264)
Multiplicity	7.0 (7.2)	9.8 (9.8)
Completeness (%)	98.9 (99.5)	99.9 (99.8)
Mean I/σ(I)	9.9 (2.3)	10.6 (1.2)
Wilson B-factor	32.2	23.8
R-merge	0.15 (0.84)	0.12 (0.19)
CC1/2	0.99 (0.77)	0.99 (0.69)
CC*	0.99 (0.93)	0.99 (0.91)
Reflections used for R_{work}	52665 (3835)	114119 (11250)
Reflections used for R_{free}	2783 (239)	5739 (547)
Number of non-hydrogen atoms	8787	4946
R_{work}	0.16 (0.25)	0.16 (0.30)
R_{free}	0.20 (0.27)	0.18 (0.29)
macromolecules	8144	4074
ligands	100	94
water	543	778
Protein residues	1023	531
RMS bonds (Å)	0.006	0.009
RMS angles (degrees)	1.1	1.3
Ramachandran favored (%)	95	97.7
Ramachandran allowed (%)	4.5	2.3
Ramachandran outliers (%)	0.5	0
Clashscore	1.75	1.46
Average B-factor	35.8	30
macromolecules	35.5	27.4
ligands	41.8	45.8
solvent	39.1	41.4

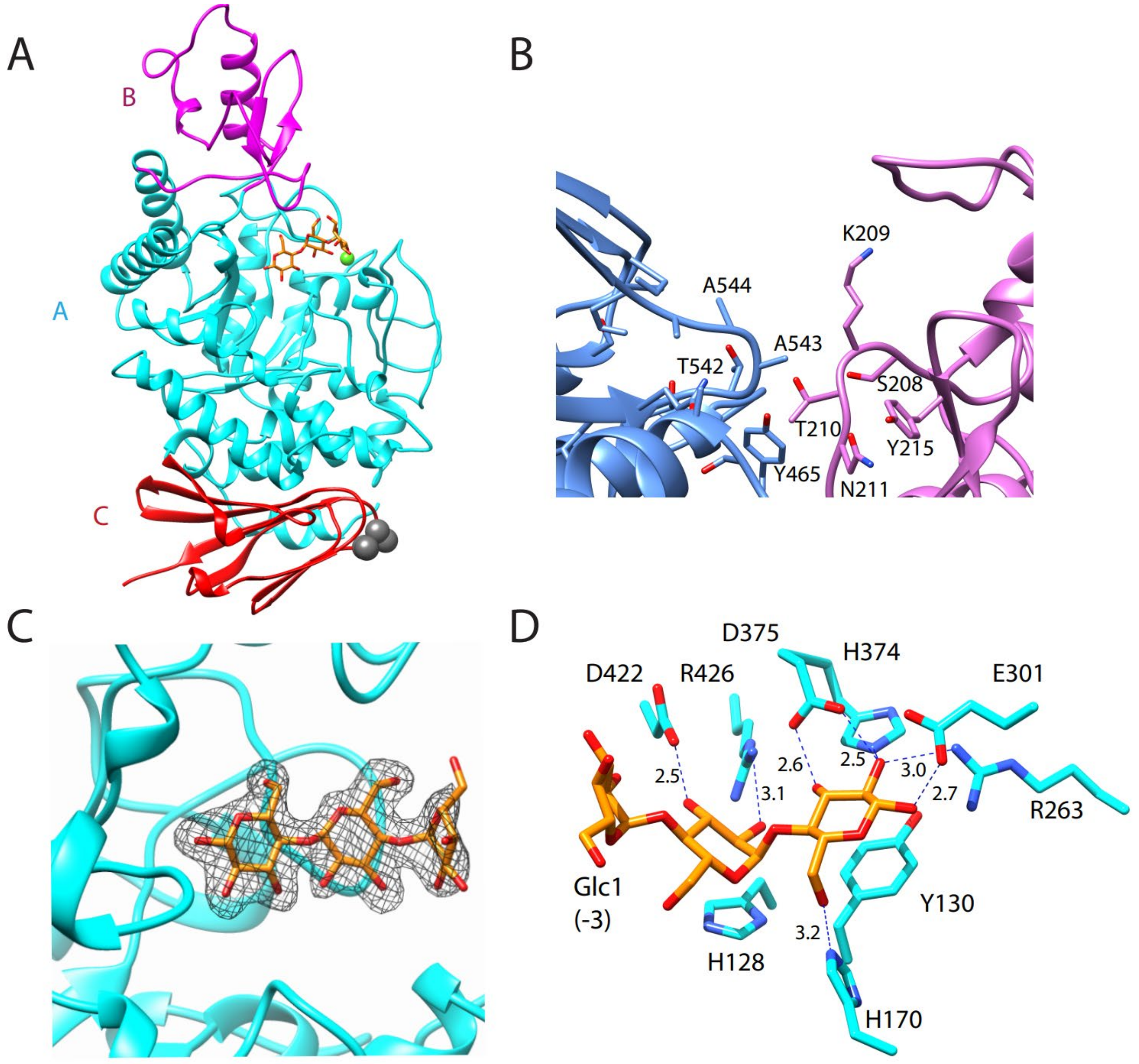


Figure 2.2: Molecular structure of ErAmy13B with maltotriose (A) Ribbon diagram of ErAmy13B, colored by domain: the A domain (residues 55–169 and 240–485) is in cyan, the B-domain (residues 170–239) is in pink, and the C domain (residues 487–567) is in red. The calcium ion is displayed as a green sphere, the maltotriose at the active site is displayed by orange sticks and the K543A/K544A mutations are displayed as gray spheres. (B) Chain B of ErAmy13B (blue) with neighboring monomer in the unit cell (pink) forming crystal contacts with mutated residues K543A/K544A for surface entropy reduction. (C) Electron density from an omit map of the bound maltotriose with the electron density contoured at 3.5σ. (D) Close-up of maltotriose bound at the active site with the potential hydrogen bonds denoted by dashed lines and distances in Å. The non-reducing end Glc1 at subsite –3 has been labeled for reference. F229 has been omitted for clarity.

The structure of ErAmy13B exhibits the standard $(\beta/\alpha)_8$ fold common to members of the GH13 family [70] (**Figure 2.2A**). The structure shows that in the A chain the introduced alanines occur at a crystal contact (**Figure 2.2B**), suggesting the introduced mutations behaved as expected by creating a hydrophobic patch on the surface of the protein to encourage monomer interactions. As expected based on the sequence, the structure lacks the N-terminal domain (CBM34) seen in other maltogenic amylases [68, 69, 133], suggesting other structural features must be responsible for this product specificity in ErAmy13B. The positioning of two copies of the protein within the unit cell does not suggest that dimerization can create limits on the active site, and we do not see any evidence that higher order oligomers might be formed. This observation was validated by analysis of the coordinates via PISA (https://www.ebi.ac.uk/pdbe/pisa/pistart.html) [138], which failed to identify evidence of multimerization.

One molecule of the co-crystallized maltotetraose is bound in the active site of each protein chain, though only three of the glucose units can be resolved within subsites –3 through –1 (**Figure 2.2C**). The reducing end glucose (Glc3) at subsite –1 has its anomeric hydroxyl in an equatorial configuration characteristic of the β-anomer of glucose and most closely resembles the covalent intermediate of the reaction mechanism (**Figure 2.2D**). The Glc3 O1 is stabilized by hydrogen bonds to E301 (the catalytic acid/base) and R263, with the latter also in a position to hydrogen bond O2. The D265A mutation of the catalytic nucleophile likely encouraged the trapping of the β-anomer, as the carboxylic acid of D265 would be directed towards C1 providing steric hindrance against the β-conformation; the O1 of the α-anomer would then only participate in hydrogen bonding with E301 (**Figure 2.3**). Moreover, the β-configuration of glucose is slightly preferred at equilibrium, and these crystals were obtained after >6 months of incubation at room temperature. The β-anomer of the reducing end has been observed in other GHs including amylases, as a consequence of substitution of the nucleophilic base to alanine [139]. Therefore, this structure with the nucleophile mutant is likely not fully representative of the wild-type enzyme-substrate complex. With an intact carboxylic acid side chain on D265, it is possible that the carbohydrate chain would be shifted towards the presumptive –2 binding site, providing a clearer path for the extension of the chain into the +1 and +2 binding sites.

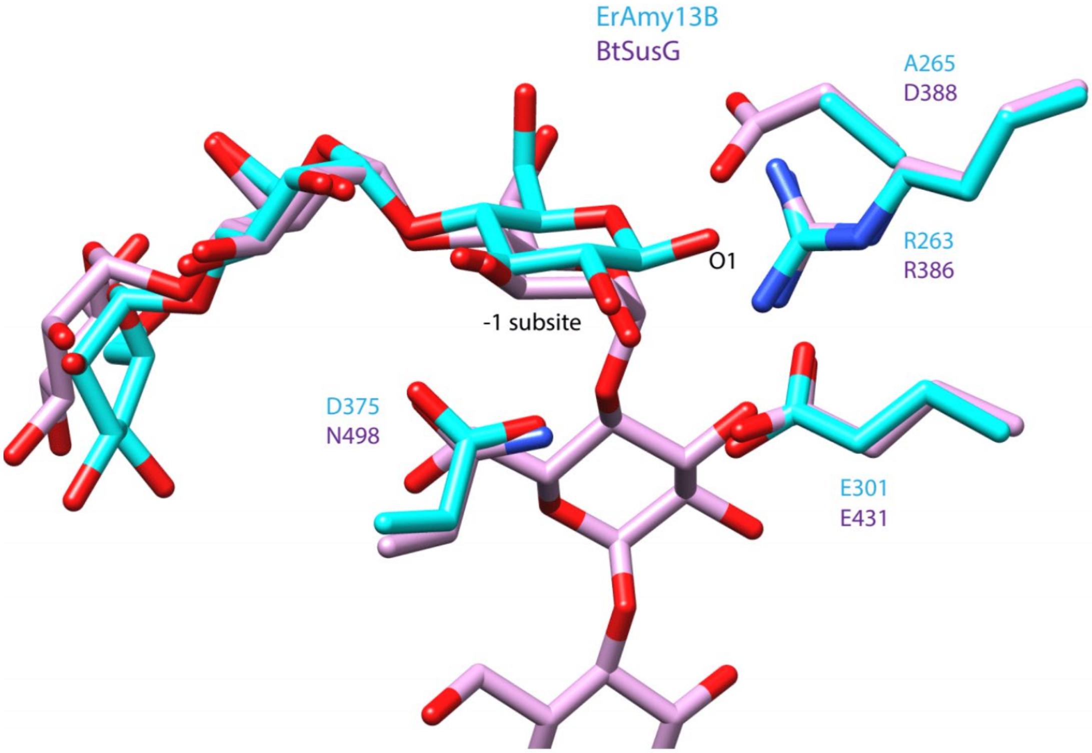


Figure 2.3: Close-up of the β-Glc at the –1 subsite of ErAmy13B. The structure of ErAmy13B (cyan) was overlaid with that of BtSusG with maltoheptaose (lilac, PDB 3K8L). The capture of the β-configuration at the –1 subsite glucose in ErAmy13B was facilitated by the D265A mutation, as the native carboxylate residue at this position would sterically clash with this anomer.

Both H128 and Y130 form a binding platform at the –2 and –1 subsites, respectively. At the –1 subsite the carboxylic acid of D375, which likely acts to stabilize the transition state during catalysis, is within hydrogen bonding distance of both O2 and O3 of Glc3, as is H374. The O6 of Glc3 is within 3.2 Å of N3 of H170. Glc2 in the –2 subsite is bound at O2 by R426 and at O3 by D422. At the non-reducing end, Glc1 has little interaction with the protein, but is coordinated at O2 by the side chain of D422, and at O6 by the peptidic O of F229, the latter interaction only present in one chain within the asymmetric unit.

Structure Comparison of ErAmy13B and RbAmy5

We determined the 1.7 Å structure of RbAmy5 (R_{work}= 16.0%, R_{free}= 18.1%; **Figure 2.4A**). The structures of ErAmy13B and RbAmy5 can be superimposed with a RMSD of 1.0 Å for 348 Cα pairs (**Figure 2.4B**). Looking more closely at the active sites of these enzymes, the –3 to –1 binding sites between RbAmy5 and ErAmy13B are very similar (**Figure 2.4C**). There is a minor deviation in the orientation of H170 in ErAmy13B that coordinates the O6 of Glc3; the equivalent position in RbAmy5 is H158.

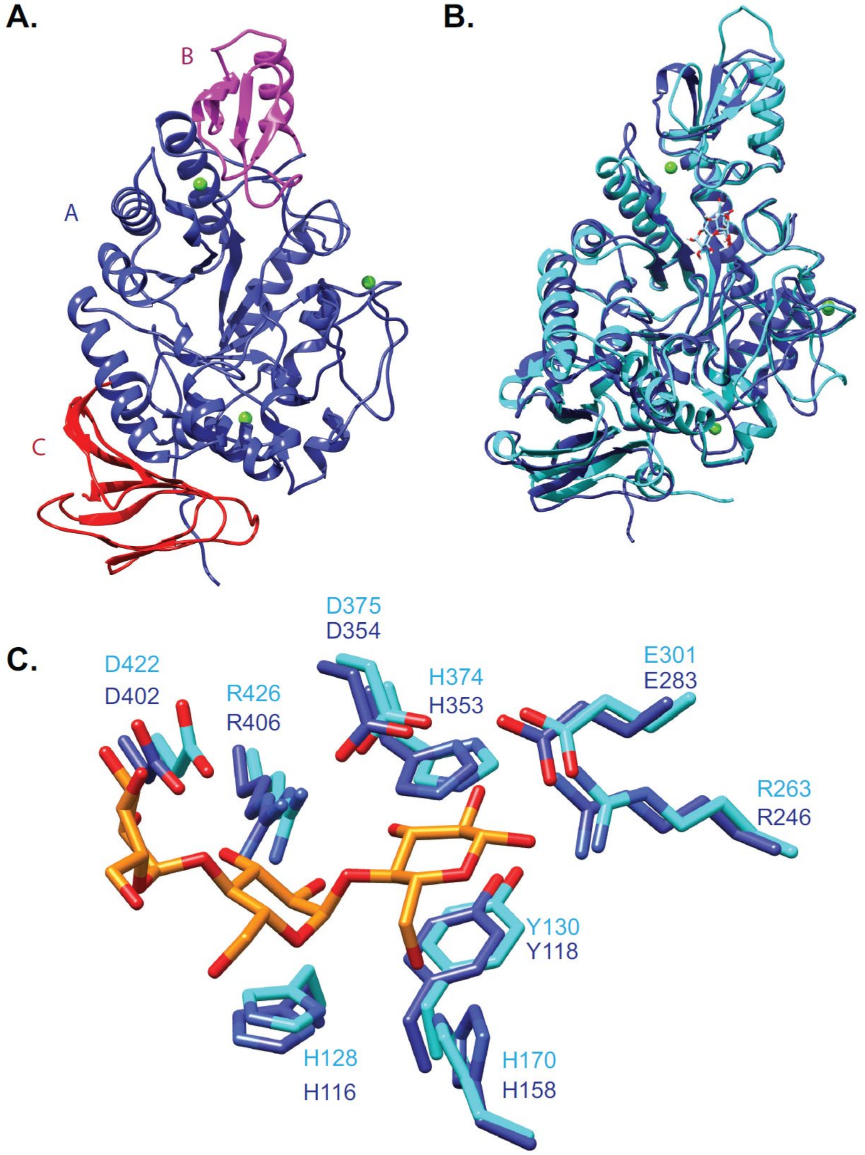


Figure 2.4: The structure of RbAmy5 and comparison with the active site of ErAmy13B. (A) Ribbon diagram of RbAmy5, colored by domain: the A domain (residues 12–145 and 190–438) is in blue, the B-domain (residues 136–189) is in pink, and the C-domain (residues 439–524) is in red. The calcium ions are displayed as green spheres. (B) Ribbon diagram overlay of RbAmy5 (blue) with ErAmy13B (cyan). (C) Close-up of the active site of RbAmy5 overlaid with that of ErAmy13B with maltotriose. The presumptive catalytic nucleophile D248 of RbAmy5 was omitted for clarity but is conserved with the position of the D265A residue of ErAmy13B.

RbAmy5 and ErAmy13B have three and two calcium binding sites, respectively, based upon the observed octahedral coordination spheres with distances of 2.3-2.7 Å, and their conservation in other GH13 enzymes [85, 140, 141]. Both enzymes have two structurally homologous calcium binding sites located in the A domain of these enzymes. Site 1 is coordinated by the side chains of three aspartic acids, one water, and the peptidic O of T75 and I95 in RbAmy5 and ErAmy13B, respectively (**Figure 2.5A**). A sixth coordinating residue is provided by N71 in RbAmy5 and D91 in ErAmy13B. This site has been observed in a number of amylases, as part of conserved sequence regions 1, 2 and 5 [68]. At a second site, calcium was captured in the RbAmy5 structure, coordinated by the backbone O of Y252, the side chains of N157 and E218, and four water molecules (**Figure 2.5B**). While this coordination sphere is conserved in ErAmy13B, we did not observe electron density for an ion perhaps as a result of the crystallization condition. Calcium here may lend additional structural support to the enzyme and be required for optimal enzymatic activity; to this end we did include calcium in all ErAmy13B enzymatic assays. Site 2 is closest to the active site (~10 Å from the putative catalytic nucleophile D248 of RbAmy5), and has been observed in the BtSusG, Taka-amylase A and barley α-amylase Amy2 structures, among others [85, 140, 141]. A third calcium binding site within a loop defined by residues 89-102 in domain A of RbAmy5 is not present in ErAmy13B, BtSusG or the HoAmyA, the other GH13_36 family members with known structures (**Figure 2.5C**) [85, 142]. Its coordination sphere includes two waters, the side chains of D98 and D90, and the peptidic oxygen of S95.

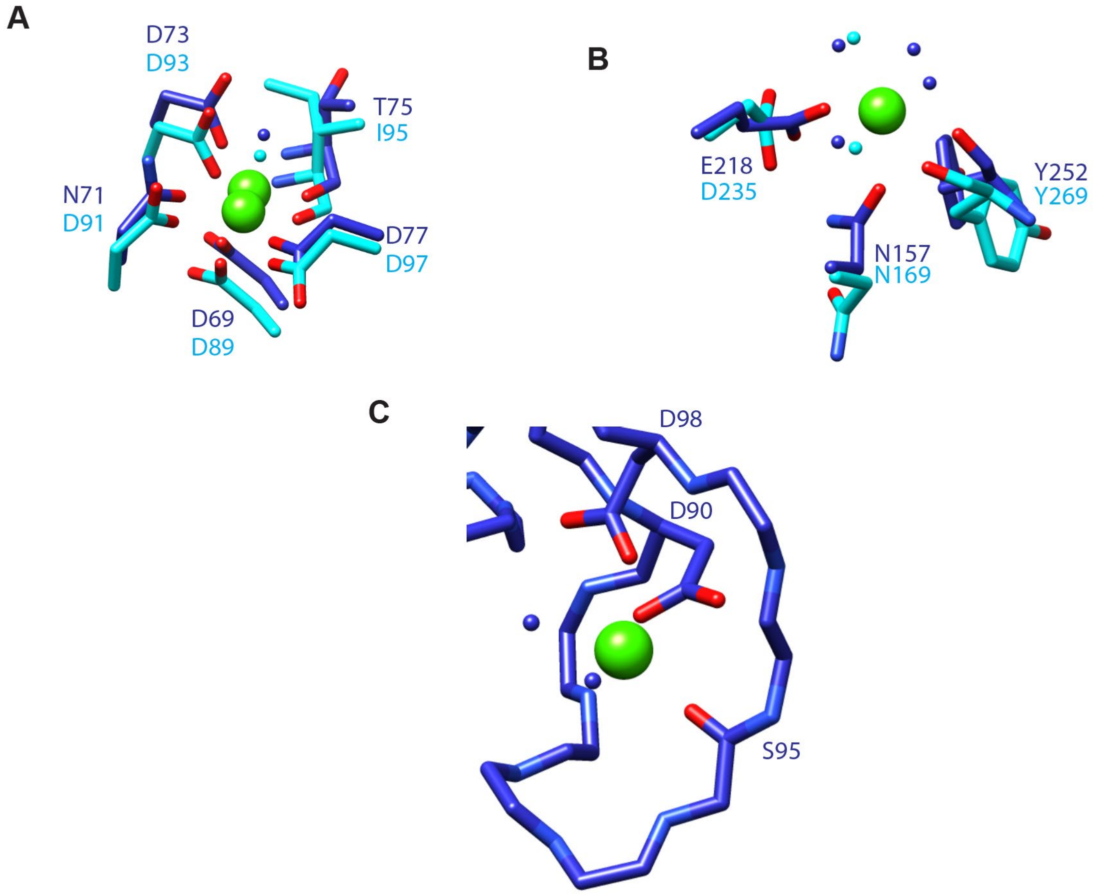
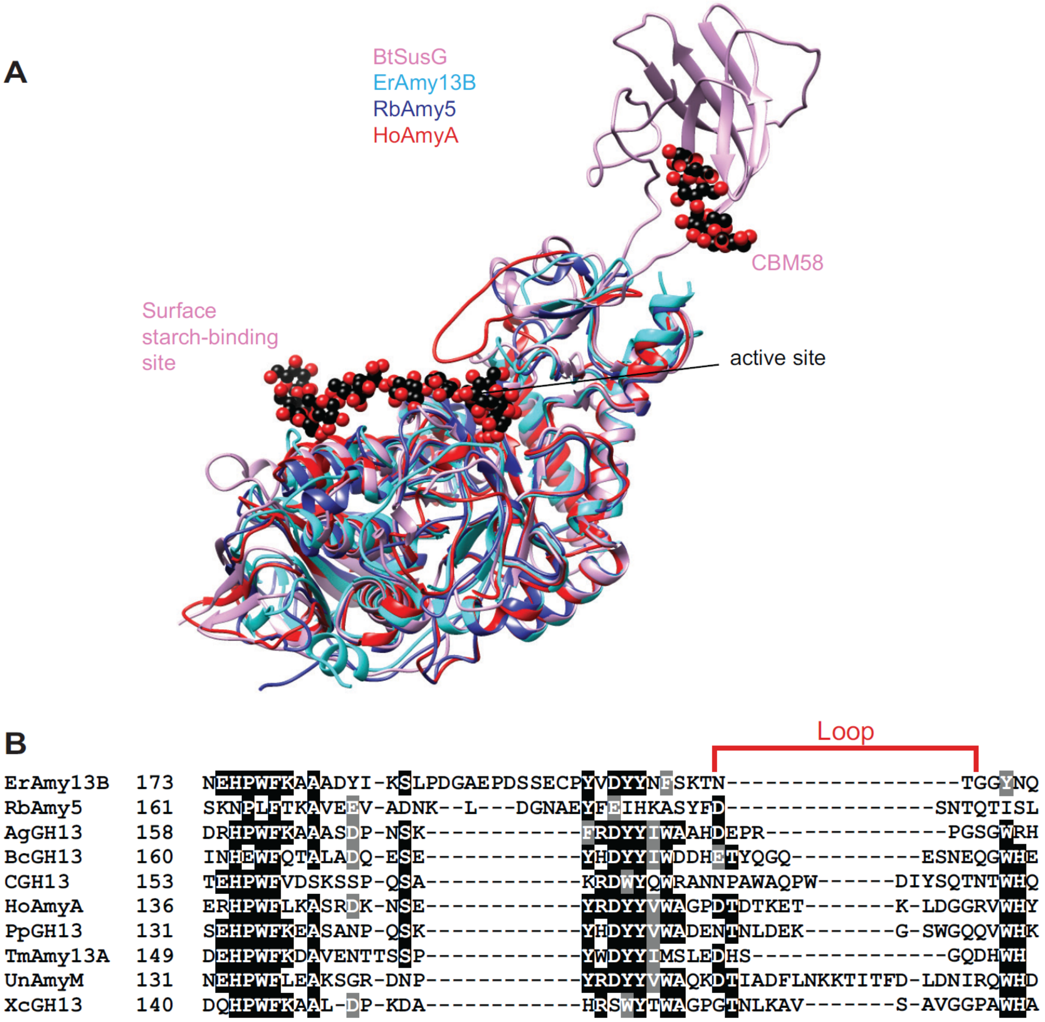


Figure 2.5: Close-up of calcium sites within ErAmy13B and RbAmy5. For all panels, ErAmy13B is displayed in cyan, including waters, RbAmy5 is displayed in blue, including waters, and calcium ions are shown as green spheres. (A) Conserved calcium site in both enzymes. (B) Calcium was captured in RbAmy5, but not ErAmy13B though the coordination sphere is conserved. (C) Calcium was captured within a surface loop of RbAmy5 that is not observed in ErAmy13B or BtSusG.

To visualize the structural differences that influence activity within the GH13_36 subfamily, we compared our structures with that of HoAmyA and BtSusG, and the sequence of our enzymes with other characterized members of this subfamily. The *H. orenii* enzyme has a loop (D164-R174) that creates a tunnel-like entrance to the active site of the enzyme, which is truncated and oriented away from the active site in ErAmy13B, RbAmy5 and BtSusG (**Figure 2.6A**) [142]. Looking at an alignment of the characterized members of GH13_36 shows that, while the sequence of this loop in HoAmyA is not conserved, its full truncation is unique to ErAmy13B and RbAmy5 (**Figure 2.6B**); while in BtSusG the insertion of the CBM58 domain shifts the orientation of this loop away from the active site. In alignments using the entire subfamily (**Figure S4, see online publication for full alignment**) the loss of this loop is consistent among members of the Lachnospiraceae. The BtSusG structure shows a unique insertion of a CBM58 module in the B-domain as well as a surface starch-binding site that is on the catalytic module adjacent to the active site (**Figure 2.6A**). Neither of these features is conserved in ErAmy13B or RbAmy5.



```
ErAmy13B   173   NEHPWFKAAADYI-KSLPDGAEPDSSECPYVDYYNFSKTN------------------TGGYNQ
RbAmy5     161   SKNPLFTKAVEEV-ADNK--L---DGNAEYFEIHKASYFD----------------SNTQTISL
AgGH13     158   DRHPWFKAAASDP-NSK------------FRDYYIWAAHDEPR--------------PGSGWRH
BcGH13     160   INHEWFQTALADQ-ESE------------YHDYYIWDDHETYQGQ----------ESNEQGWHE
CGH13      153   TEHPWFVDSKSSP-QSA------------KRDWYQWRANNPAWAQPW------DIYSQTNTWHQ
HoAmyA     136   ERHPWFLKASRDK-NSE------------YRDYYVWAGPDTDTKET-------K-LDGGRVWHY
PpGH13     131   SEHPWFKEASANP-QSK------------YHDYYVWADENTNLDEK-------G-SWGQQVWHK
TmAmy13A   149   DEHPWFKDAVENTTSSP------------YWDYYIMSLEDHS---------------GQDHWH
UnAmyM     131   NEHPWFLEAKSGR-DNP------------YRDYYVWAQKDTIADFLNKKTITFD-LDNIRQWHD
XcGH13     140   DQHPWFKAAL-DP-KDA------------HRSWYTWAGPGTNLKAV-------S-AVGGPAWHA
```

Figure 2.6: Comparison of GH13_36 structures. (A) Overlay of GH13_36 subfamily members BtSusG (lilac; PDB 3K8L), HoAmyA (red; PDB 1WZA) with ErAmy13B (cyan) and RbAmy5 (blue). The CBM58 and surface starch-binding site of BtSusG are labeled and bound maltooligosaccharide to BtSusG is displayed in black and red spheres. (B) Excerpt of sequence alignment among the characterized members of the GH13_36 subfamily. BtSusG has been omitted due to the insertion of CBM58. The loop that closes over the active site of the HoAmyA structure is noted.

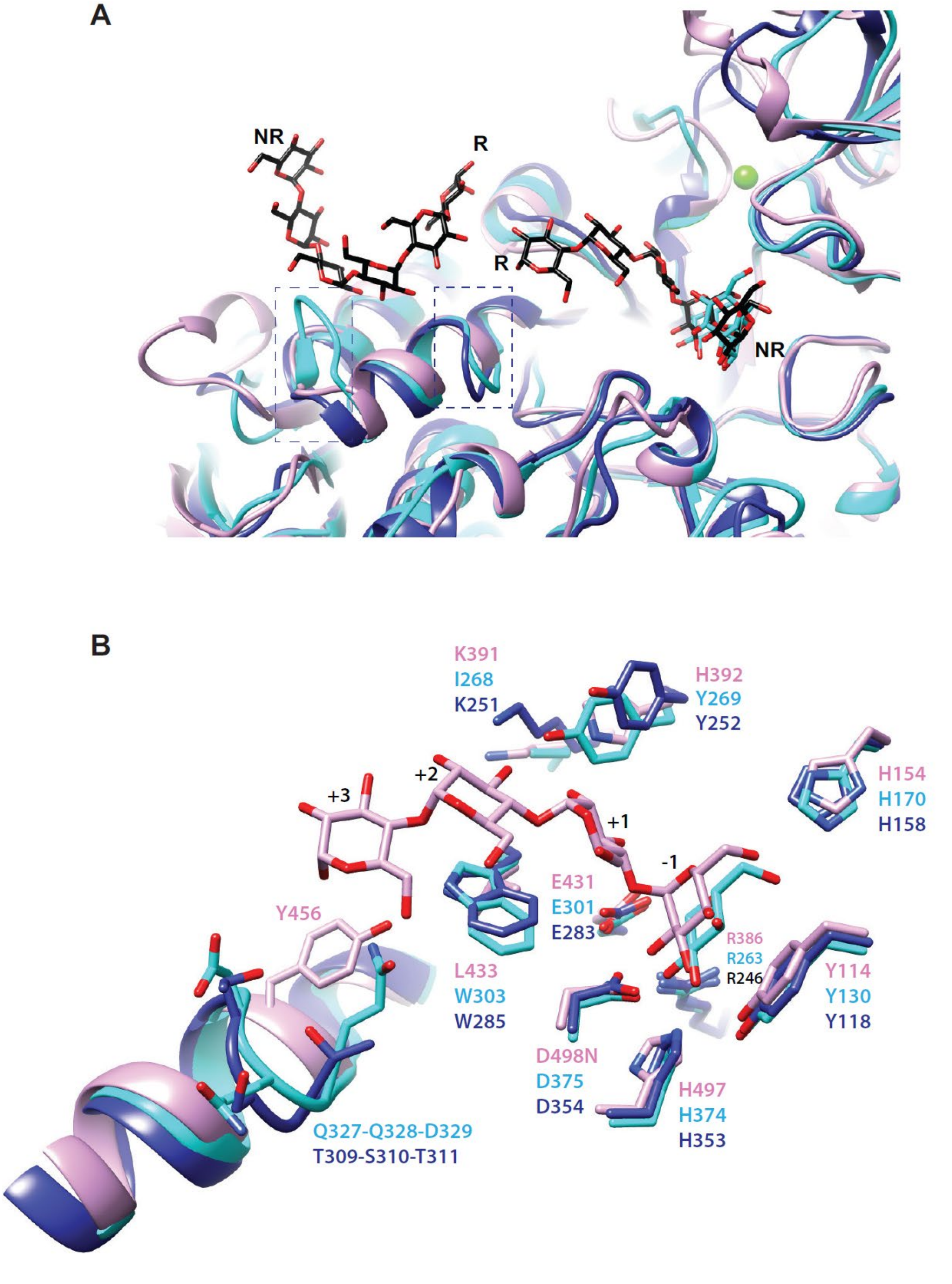


Figure 2.7: Close-up of the active sites among BtSusG, ErAmy13B and RbAmy5. (A) Close-up of the active site of the three enzymes with the maltoheptaose of BtSusG at the surface starch-binding and active site. Structures are colored as in Figure 3. Dashed lines indicate the only two areas where the structures deviate: at loops defined by Q329-D329/T309-T311 of Amy13B/Amy5 within the active site and G338-G343 of Amy13B at the surface starch-binding site of BtSusG. (B) Close-up of the –1 to +3 subsites of the three enzymes. Residues involved in coordinating substrate in the BtSusG structure, and their equivalents in ErAmy13B and RbAmy5, are displayed. R and NR indicate the reducing and non-reducing ends of the maltooligosaccharides, respectively.

Because the structure of BtSusG was determined with maltoheptaose spanning subsites –4 to +3, we superimposed this structure with RbAmy5 and ErAmy13B in order to discern differences among these enzymes, particularly at the positive subsites as the negative subsites are largely conserved and there is no apparent restriction for the accommodation of a longer α-glucan chain at the non-reducing end (**Figure 2.7**). The active sites of these enzymes are well conserved, though some differences exist, primarily near the positive subsites. There are small deviations in an α-helix that extends from the +3 subsite of BtSusG, including a loop created by Q327- D329 in ErAmy13B and T309-T311 in RbAmy5 that is absent in BtSusG (**Figure 2.7 A, B**). The +3 subsite of BtSusG, missing in ErAmy13B and RbAmy5, has minimal interaction with the substrate, featuring only Y456 that may be involved in van der Waals stacking with the glucose. The short additional loop in ErAmy13B features the sequence QQD compared to RbAmy5 that features the sequence TST. The longer side chains within the loop of ErAmy13B may potentially restrict optimal binding of longer α-glucans beyond the +2 subsite, though this is speculation. The only other striking difference between ErAmy13B and RbAmy5 lies in the region adjacent to the active site that houses the surface starch-binding site of BtSusG. This surface starch-binding site is lacking in other characterized members of the GH13_36 subfamily including ErAmy13B and RbAmy5. In its place, ErAmy13B has a protruding loop from G338-H342, though presumably this is far enough from the active site that it would not alter substrate binding (**Figure 2.7A**).

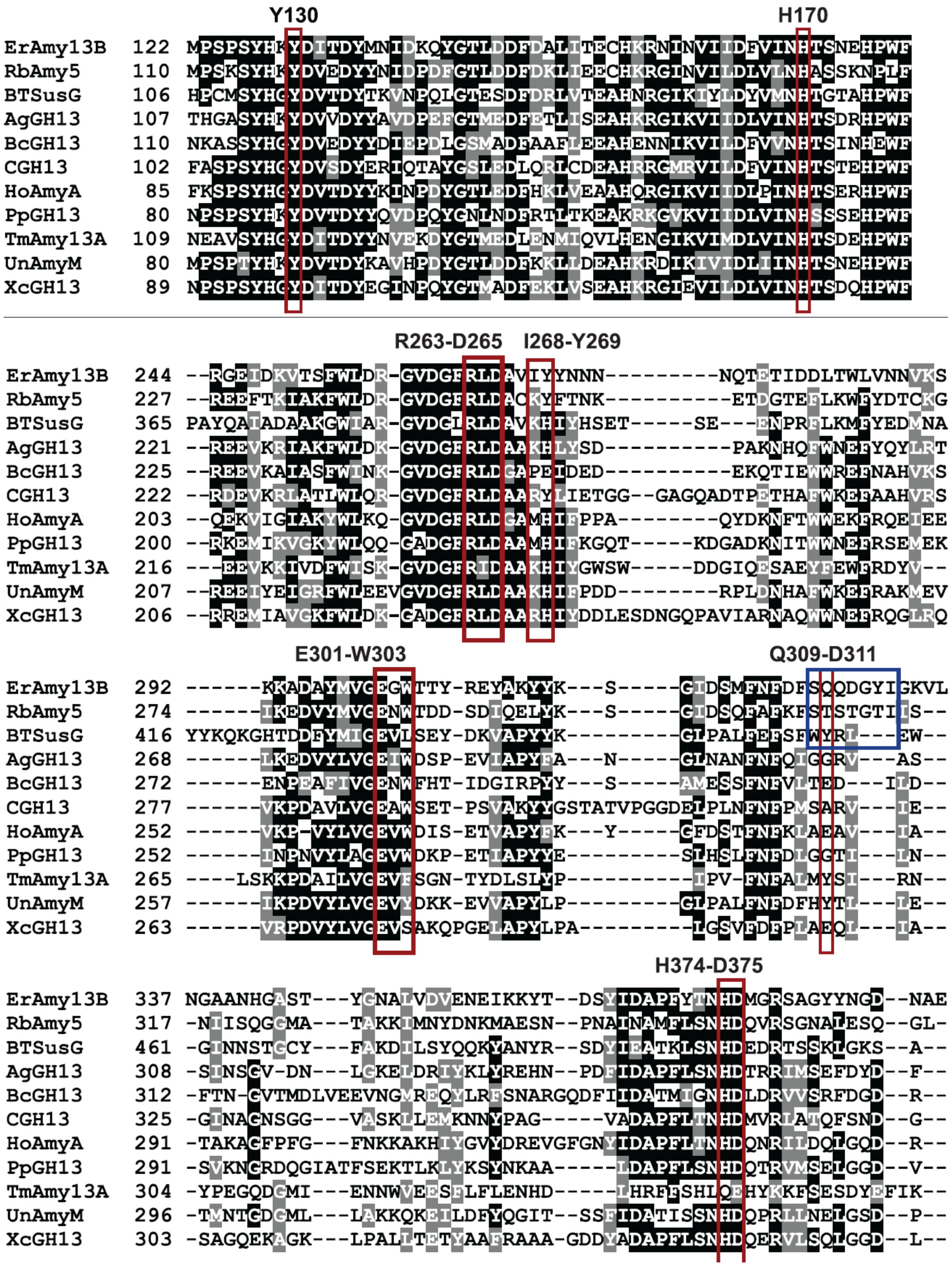


Figure 2.8: Alignment of characterized members of the GH13_36 subfamily. Residues that are involved in substrate coordination within the active site in the BtSusG, ErAmy13B and RbAmy5 are boxed in red. Numbered residues above boxes are in reference to the ErAmy13B sequence. The blue boxed region denotes the small loop insertion within ErAmy13B and RbAmy5 proximal to the +3 subsite in the BtSusG structure. The horizontal line indicates a break in the sequence as the region corresponding to the CBM58 insertion of BtSusG is not shown.

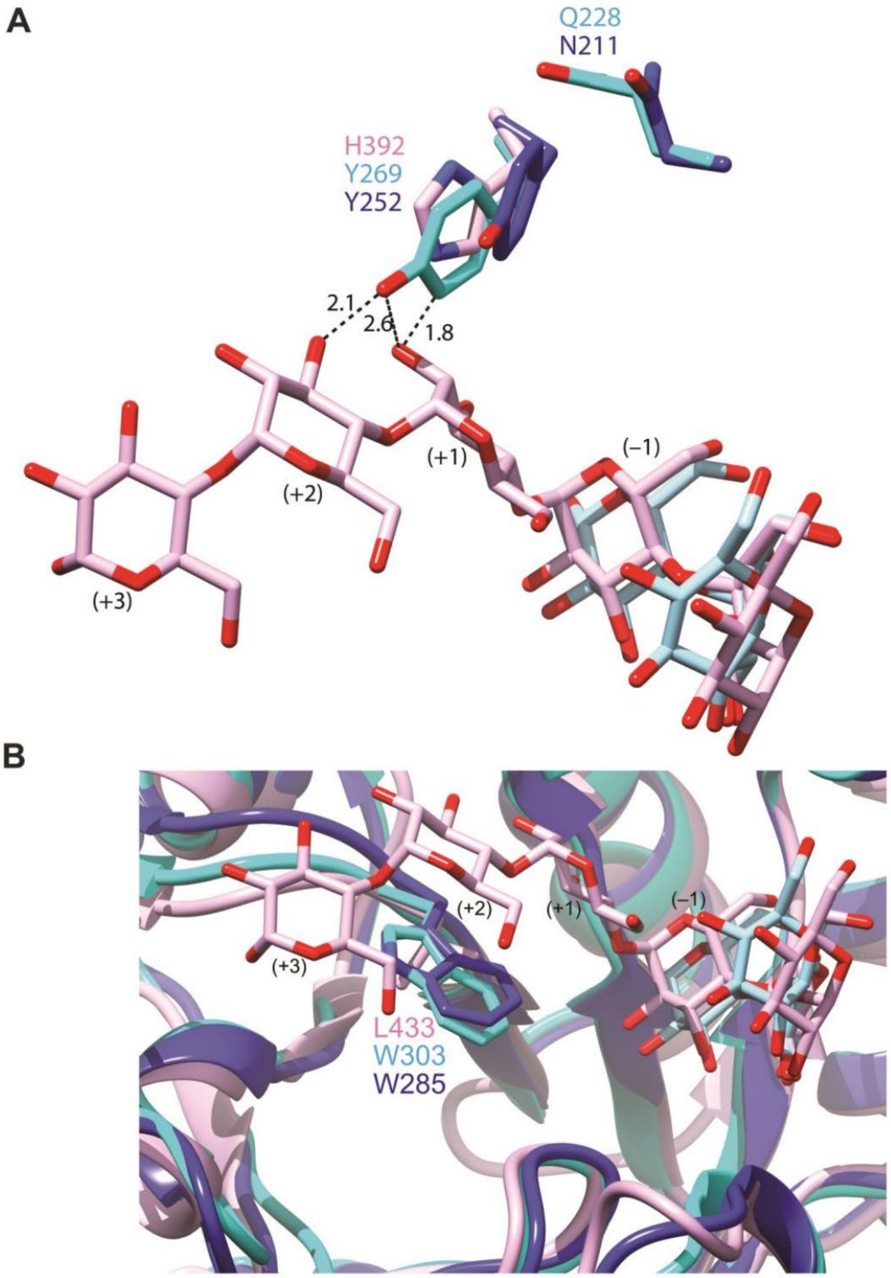


Figure 2.9: Close-up of subtle differences in the +1/+2 subsites of BtSusG, ErAmy13B and RbAmy5. (A) Close-up of Y269 within ErAmy13B and its potential limited range of movement due to the placement of Q228. The equivalent residue in RbAmy5, Y252, may be less limited as Q228 is replaced with N211. Distances shown in Å based upon the position of maltoheptaose bound to BtSusG. (B) Close-up of the active site of the three enzymes with the maltoheptaose of BtSusG at the +2 subsite demonstrating conservation of the subsite tryptophan between ErAmy13B and RbAmy5.

Overall, many of the residues involved in substrate binding are conserved among the characterized members of the GH13_36 subfamily (**Figure 2.8**). The difference between the active sites of ErAmy13B, RbAmy5 and BtSusG are more subtle and lie within the +1 to +2 subsite transition, particularly the orientation of Y269 in ErAmy13B. Both ErAmy13B and RbAmy5 feature a tyrosine, Y269 and Y252, respectively, at the +1 subsite, whereas most GH13_36 members have a histidine at this position (**Figure 2.8, S4**) as is the case in HoAmyA and BtSusG. The Y252 in RbAmy5 is oriented away from +1 subsite but the Y269 of ErAmy13B protrudes towards the +1/+2 Glc subsites, as suggested in the overlay with BtSusG. Most significantly, the aromatic ring in Y269 would be ~1.8 Å away from the O2 of the BtSusG Glc at +1 (**Figure 2.9A**). The Y269 of ErAmy13B is somewhat restrained in position by Q228, which prevents its shift away from the active site and may hold this residue in place. In RbAmy5, Q228 is replaced by N184, which may allow for more freedom of movement for the tyrosine, perhaps allowing more flexibility in substrate binding (Figure S5A). However, the putative +1 and +2 subsites of ErAmy13B and RbAmy5 are identical, including a conserved tryptophan (W303 in ErAmy13B and W285 in RbAmy5) that is maintained in a number of GH13 subfamilies, including GH13_36 [143] (**Figures 2.7B, S4, 2.8, 2.9B**). It typically serves as the +2 binding site (in some cases +1), and we can presume that it does in this case as well. Interestingly, this +2 subsite tryptophan is not conserved in BtSusG and rather replaced by L433 (**Figure 2.7B, 2.9B**). While maltoheptaose is captured in the BtSusG structure from the –4 to +3 subsites, this enzyme releases both maltose and glucose as the reaction achieves completion, supporting that interactions beyond the +1 subsite are not required for catalysis [85]. BtSusG has a more open active site architecture that readily accommodates α-1,6-branchpoints, even directly adjacent to the site of catalysis [96], allowing for its neopullulanase activity, which is less prominent in ErAmy13B, absent from HoAmyA, and intermediate in RbAmy5 (see Discussion for activity profiles in GH13_36). Therefore, we speculate that these changes in active site architecture may help to explain some of the substrate and product profiles differences between these enzymes.

Discussion

GH13_36 Activity and Product Profiles

With the ErAmy13B and RbAmy5 structures and the recent move of BtSusG into the GH13_36 subfamily, we now have four structures and fourteen members of this subfamily that have been biochemically characterized and eleven with information on the relative activity on a range of substrates (**Table 2.3**).

Table 2.3: Relative activities of GH13_36 family enzymes normalized to amylopectin (%).

Enzyme	GenBank	AM	AP	SS	Pul	α-CD	β-CD	γ-CD	Ref.
Anaerobranca gottschalkii (AgGH13)	AAW32491.1	245	100	156	27	20	92	91	[47]
Bacillus clarkii (BcGH13)	BAH14969.1	298	100	208	0	9	35	9058	[48]
Bacteroides thetaiotaomicron (BtSusG)	AAO78803.1	+[c]	100	192	90	3	15	ND	[19]
Corallococcus EGB (CGH13)	AVC05420.1	292	100	833	117	0	0	167	[49]
Eubacterium rectale (ErAmy13B)	CBK89424.1	20	100	ND	41	9	429	ND	[14]
Halothermothrix orenii (HoAmyA)	AAN52525.1	187	100	192	0	0	0	0	[52]
Paenibacillus polymyxa (PpGH13)	AAD05199.1	0	100	120	65	0	ND	ND	[53]
Ruminococcus bromii (RbAmy5)	SPE91476.1	839	100	556	161	0	2	33	[18]
Thermotoga maritima (TmAmy13a)	CAA72194.1	210	100	116	5	ND	ND	ND	[56]
Uncultured (UnAmyM)[b]	AAQ89599.1	ND	ND	100	6.5	0.4	11	34	[51]
Xanthomonas campestris (XcGH13)	BAA07401.1	156	100	169	112	161	183	185	[55]

AM, amylose; **AP**, amylopectin; **SS,** soluble starch; **Pul**, pullulan; **α-CD**, α-cyclodextrin; **β-CD,** β-cyclodextrin; **γ-CD**, γ-cyclodextrin; ND, not determined.

[b] Activity towards amylopectin was not determined, so activities are shown normalized to soluble starch.

[c] Activity tested and found to be present, but not quantified.

Universally these enzymes have demonstrated only α-1,4-hydrolytic activity, although several members have some transglycosylase activity [144-148]. We presently have no data that suggests or excludes the possibility that ErAmy13B and RbAmy5 have transglycosylase activity. Many of these enzymes have activity against pullulan (as neopullulanse activity generating panose) and against CDs, though with some variability in the relative amounts of activity towards these substrates. The *Bacillus clarkii* GH13 (BcGH13) is particularly odd in this regard as it has extremely high activity against γ-CD and lower levels of activity against other CDs or soluble starches [147]. Interestingly, BcGH13 lacks any activity towards pullulan, with the HoAmyA being the only other characterized member of this subfamily that lacks this activity

[149]. In the case of HoAmyA, we speculate that the tunnel-like architecture of the active site precludes the binding of substrates containing α-1,6-bonds, or CDs [149], while other enzymes with more open active sites like BtSusG, RbAmy5 and ErAmy13B can accommodate these features. While it was believed that the replacement of a conserved active site alanine at position 247 with glycine in BcGH13 might help accommodate CDs in this enzyme [147], the only other characterized member of the subfamily that has glycine at this position is HoAmyA, which seems to lack cyclodextrinase activity [149]. It is possible that it is the active site covering loop that prevents this activity in HoAmyA and that the equivalent (slightly shorter) loop in *B. clarkii* (**Figure 2.6B**) does not block CD binding in the same way. In another extreme example in this subfamily, the *Paenibacillus polymyxa* enzyme (PpGH13) has strong activity against soluble starch, amylopectin and pullulan, but lacks activity against amylose, glycogen or maltodecaose [150]. At first glance this may appear to indicate that the enzyme is in fact a pullulanase, but product analysis demonstrated only panose as the final end product of pullulan digestion [151], indicating that it is a neopullulanase, although one that apparently requires the presence of α-1,6-bonds for activity on neighboring α-1,4-bonds.

Members of the GH13_36 subfamily seem to, like RbAmy5 and ErAmy13B, generate glucose and maltose as their final reaction products after extended incubation. The exception to this is the *Corallococcus* EGB enzyme, which seems to generate only maltose due to transglycosylation reactions that consume glucose [146]. In the cases where it has been measured, maltotriose is a substrate for GH13_36 enzymes, though significantly poorer than longer oligosaccharides (**Table 2.1**) [152]. This is consistent with the structures of ErAmy13B and RbAmy5 where the strongest binding is likely to take place at the aromatic platforms in the +2 and –2 binding sites, indicating that substrates of at least the length of maltotetraose are needed for fully efficient binding but without evidence of any additional strong binding sites in either direction. Indeed, while the highest catalytic efficiency for ErAmy13B occurs with maltoheptaose, the lowest K_{M} occurs with maltotetraose (**Table 2.1**). This is somewhat in contrast to the *Xanthomonas campestris* GH13_36 enzyme (XcGH13), which exhibits a trend of decreasing K_{M} from maltotriose through maltoheptaose [152]. However, like for ErAmy13B, amylose is a worse substrate than the maltooligosaccharides for XcGH13 indicating a similar preference for oligosaccharides, albeit perhaps with a larger active site. For a few of the GH13_36 enzymes product profiles have been examined during the early stages of polysaccharide digest. The *Thermotoga maritima* enzyme produces a wide range of oligosaccharides in the initial phases of digestion before breaking these products down further to the glucose and maltose product [153] characteristic of this subfamily and consistent with typical endo-acting amylase activity. However, ErAmy13B [107] and RbAmy5 [124] produce only short oligosaccharides even in the earlier stages of digestion of oligosaccharides. The tunnel-like architecture of HoAmyA would support an exo-mode of action analogous to β-amylases to generate maltose, but this seems to be absent in ErAmy13B and RbAmy5. The strong preference for oligosaccharides vs polysaccharides exhibited by ErAmy13B could similarly drive this product profile, but it is not obvious from the structure why this preference should be so strong beyond the relatively limited number of binding subsites.

The BtSusG active site is remarkably open, and it is able to accommodate one or more branchpoints [96]. Initial product formation on starch has not been measured in BtSusG, however, it accumulates a higher proportion of glucose after extended incubation than most enzymes in this family [69, 96], suggesting that binding to the +2 subsite in this case is not as critical in this enzyme as for other subfamily members. This may be due to the lack of the +2 subsite tryptophan that is conserved in most other enzymes (**Figures 2.7B, 2.9B**). The only other characterized member of GH13_36 that lacks this +2 tryptophan or other aromatic residue at this position is XcGH13, which also produces large amounts of glucose [152]. Thus, this family is characterized by relatively few binding subsites overall that contributes to the formation of shorter products, but also contains other variable structural features that can limit product size.

Comparison of ErAmy13B and RbAmy5 to Maltogenic Amylases in Other Subfamilies

Besides GH13_36, maltogenic amylases are found in the GH13_2, GH13_20 and GH13_21 subfamilies [67]. Both the GH13_20 and GH13_21 subfamilies are part of what was originally proposed as the neopullulanase section of the GH13 phylogenetic tree [154] and were partially defined by the presence of a common N-terminal domain, which has since been identified as a CBM34 family domain (**Figure 2.10A**) [68, 155]. These enzymes are maltogenic in nature and tend to produce short oligosaccharide products, displaying activity against CDs and pullulan [156]. One area where they differ from the GH13_36 enzymes is that they tend to display both α-1,4- and α-1,6- (though weaker) hydrolyzing activity. The GH13_20 and GH13_21 seem to perform as dimers with the CBM34 from one molecule helping form part of the active site in the second molecule in a symmetric fashion (**Figure 2.10B**) [133]. As revealed by the crystal structure of the *Bacillus stearothermophilus* neopullulanase, this causes a narrowing of the active site and partially blocks it at one end, which is thought to contribute to the unique properties of these enzymes, including the restricted nature of the product profile[133].

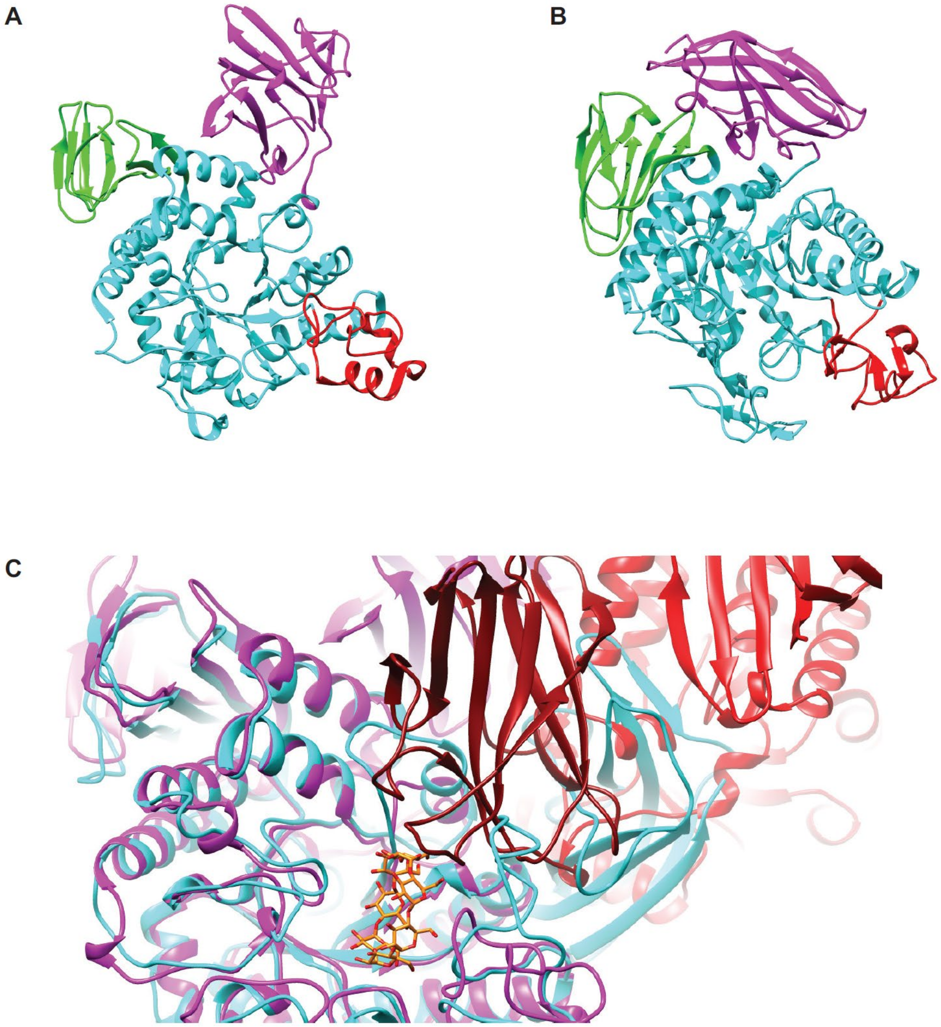


Figure 2.10: Comparison of GH13 subfamilies 2, 20, 21 structures and active sites. Ribbon diagram of (A) *Bacillus stearothermophilus* TRS40 (PDB 1J0H) as a representative GH13_20 member and (B) *Thermoactinomyces vulgaris* TVAI (PDB 1JI1) as a representative GH13_21 member. Domains coloring are as follows: N terminal CBM34 (magenta), domain A (cyan), domain B (red), domain C (green). (C) Overlay of *Bacillus stearothermophilus* TRS40 (PDB 1J0H; chain A red, chain B magenta) and cyclodextrin glycosyltransferase from *Bacillus circulans* (PDB 3CGT; cyan) as a GH13_2 representative. The GH13_2 (cyan) has a more open active site architecture compared to the GH13_20 whereby the CBM34 from the dimer pair restricts the active site.

The CBM34 family is one of the largest CBM families [68], and its role in shaping the active site of neopullulanases has been documented in a number of crystals structures [156, 157]. Contrastingly, GH13_2 is active as monomers and has a deep pocket for an active site (**Figure 2.10B**). This may facilitate the other activity that this subfamily is known, for which is cyclomaltodextrin glucanotransferase that leads to CD formation [158], while also limiting the substrate/product size for hydrolysis. It does not appear that either of these mechanisms are responsible for the product profiles seen in ErAmy13B and RbAmy5, though possible quaternary structure formation has not been extensively studied. It should be noted that HoAmyA has been found to oligomerize, but this may be related to its evolutionary adaptation to the lifestyle of *Halothermothrix orenii*, an obligately anaerobic, halophilic thermophile [142].

Localization and Role of GH13_36 Enzymes

Of the characterized enzymes in GH13_36, most are predicted to be lipidated and exposed on the cell surface (**Table 2.4**). There are exceptions to this, however. The *Xanthomonas campestris* XcGH13 is predicted to be periplasmic as it was initially isolated from the periplasm of the organism [152]. The *Bacillus megaterium* BMW-amylase and the *Paenibacillus polymyxa* enzyme are predicted to be extracellular in nature, containing cleavable signal sequences without a lipidation site or any detectable sortase signal. None of the characterized members of this subfamily are predicted to be cytoplasmic. Upon examination of the genomes of GH13_36 producing organisms, there is at least one additional predicted surface/extracellular amylase present. This can be a single additional enzyme as in *E. rectale* and *H. orenii* or a suite of additional amylases as in *R. bromii* and *P. polymyxa*, perhaps suggesting that the main function of these enzymes, in many cases, is the production of smaller oligosaccharides that are more easily imported and utilized by the organism. However, for *Bacteroides thetaiotaomicron* and *Bacillus clarkii* their GH13_36 enzymes are their sole extracellular amylases. In each of these cases, the enzyme is modified by the addition of a CBM, a CBM58 in BtSusG and a CBM20 in the case of BcGH13 likely to improve efficiency towards polysaccharide substrate reflecting the primary role of CBMs in starch degradation in these organisms [68].

Table 2.4: Signal sequences and predicted localization of characterized GH13_36 enzymes.

Enzymes	GenBank	Signal Type[a]	Probability (%)[b]	Predicted localization[c]
Anaerobranca gottschalkii (AgGH13)	AAW32491.1	LIPO	99	GP membrane
Bacillus clarkii (BcGH13)	BAH14969.1	LIPO	99	GP membrane
Bacteroides thetaiotaomicron (BtSusG)	AAO78803.1	LIPO	99	GN outer membrane
Corallococcus EGB (CGH13)	AVC05420.1	LIPO	99	GN outer membrane
Eubacterium rectale (ErAmy13B)	CBK89424.1	LIPO	93	GP membrane
Halothermothrix orenii (HoAmyA)	AAN52525.1	SP	99	GP Extracellular
Paenibacillus polymyxa (PpGH13)	AAD05199.1	OTHER	79	GP Extracellular
Ruminococcus bromii (RbAmy5)	SPE91476.1	LIPO	99	GP membrane
Thermotoga maritima (TmAmy13a)	CAA72194.1	LIPO	99	GN outer membrane
Uncultured (UnAmyM)	AAQ89599.1	LIPO	98	GP membrane
Xanthomonas campestris (XcGH13)	BAA07401.1	SP	95	GN periplasm

a Predicted signal type from SignalP-5.0. LIPO indicates secretion through the Sec translocon followed by cleavage by Signal Peptidase II and lipidation (membrane localization), SP indicates secretion through the Sec translocon followed by cleavage by Signal Peptidase I (secreted across plasma membrane), OTHER indicates a likely signal sequence, but type uncertain.
b Probability of the given signal type from SignalP-5.0.
c Prediction from PSORTb-3.02, GP is Gram positive, GN is Gram negative.

In addition to CBMs, starch-active enzymes, and many other carbohydrate-active enzymes, can contain surface binding sites (SBS) that are carbohydrate binding sites on the surface of the catalytic module but outside of the active site [159, 160]. These SBS perform a number of important functions in the enzymes, in which they are present but are often poorly conserved between enzymes [161]. BtSusG contains such an SBS **(Figure 2.6A**) that enhances activity on soluble starch [85]. Examining this SBS in GH13_36 reveals that the key aromatic residues of the BtSusG SBS are not conserved among the characterized members of this subfamily (**Figure 2.8**) and, indeed, no SBS is evident in our structures of ErAmy13B or RbAmy5. Looking at the entire subfamily it seems that the BtSusG SBS residues are only conserved in closely related *Bacteroides*-derived enzymes (**Figure S4**), suggesting that it may be a particular adaptation of the enzymes that have a specific role for these gut microorganisms. None of the other characterized GH13_36 enzymes possess CBMs or show evidence of SBS, suggesting that polysaccharide degradation is not their primary role, consistent with the observed superior activity towards oligosaccharides.

In complement to its surface amylases, *E. rectale* encodes two maltooligosaccharide transporters, one with a solute binding protein EUR_31480 that has affinity for maltose and maltotriose, and the other with the solute binding protein EUR_01830, which has affinity for maltooligosaccharides of the length of maltotetraose or longer [107]. The other major amylase on the *E. rectale* surface, ErAmy13K, produces primarily products in the size range of EUR_01830, which would then overlap with the preferred substrate range of ErAmy13B. The dissociation constant of EUR_01830 towards maltooligosaccharides is several orders of magnitude lower than the K_M for ErAmy13B towards these substrates, suggesting it would easily outcompete ErAmy13B for these substrates. However, it is possible that ErAmy13B plays a role when the EUR_01830 transporter is saturated, such as may occur when there are locally high oligosaccharide concentrations as is likely the case when *E. rectale* is localized to starch particles along with other starch degrading organisms [103].

Furthermore, it may be the rate of maltooligosaccharide transport that is limiting rather than affinity. While maltooligosaccharides are imported through ABC transporters, this has not been thoroughly investigated. When maltose vs maltotriose transport was compared, a large decrease in the rate for maltotriose transport was found despite similar affinities of the solute binding protein for maltose and maltotriose [162]. It is possible that longer substrates would have even larger decreases in rates. Therefore, it may be that the role of ErAmy13B is a load-balancing function where it breaks down oligosaccharides released from ErAmy13K or other maltooligosaccharides in the local environment to maltose. That maltose can then be funneled to the maltose transporter EUR_31480, so that the EUR_01830 transporter is not overloaded and allowing oligosaccharides to escape to competing organisms. This role could allow for more efficient use of the transporters to maximize uptake in the highly competitive gut environment during periods when viable substrates are plentiful. For other organisms that have maltose transporters, but lack those necessary for longer oligosaccharide uptake, GH13_36 enzymes could play an important role in trimming the oligosaccharides down to a size that the organism's transporters can handle. In contrast, BtSusG seems to have diverged from other GH13_36 members and has adapted to play a role similar to other known surface amylases such as ErAmy13K, handling mostly polymeric substrates to cut them down to size for the SusC outer membrane transporter, which is followed by further periplasmic processing.

SignalP analysis of the RbAmy5 N-terminal sequence suggests this protein, like ErAmy13B, is lipidated and localized to the plasma membrane as it features a putative SPII secretion signal. *Ruminococcus bromii* encodes a number of starch-binding and starch-hydrolyzing enzymes that possess cohesin and/or dockerin domains that adhere strongly to each other and facilitate the assembly of a larger starch-digesting complex called the amylosome [117]. RbAmy5 is not a recognized component of the amylosome due to its lack of cohesins and dockerins, nor has it been identified proteomically as a major component of the cell-free supernatant or the cell pellet from stationary phase *R. bromii* cells cultured on type RSIII or fructose [117]. The gene encoding RbAmy5 is found within a putative operon also encoding a homolog of maltose binding protein, MalE, and a putative maltose ABC transporter, MalF/G, similar to ErAmy13B. Operons encoding both cell-surface glycosidases and an ABC transporter have been observed in a number of Gram-positive microbes and have been termed Gram positive polysaccharide utilization locus (gpPUL) [101, 163]. These are analogous to the polysaccharide utilization loci (PUL) described in the Gram negative Bacteroidetes, exemplified by the starch utilization system, of which BtSusG is an integral part [164]. It is possible that the MalEFG works alongside RbAmy5 on the cell surface to degrade starch then bind and import maltose in a manner similar to PUL and gpPul in other organisms. Perhaps the efficiency of the production and transport of maltose can partially explain the phenomenon that *R. bromii* cannot grow on glucose as a sole carbon source *in vitro* but instead favors maltose for growth [37].

Conclusion

ErAmy13B and RbAmy5 are maltogenic amylases belonging to the GH13_36 subfamily. Comparison of the X-ray crystal structures of these enzymes has revealed active sites with a limited number of subsites that favor the binding of shorter substrates which helps to drive a narrow product profile that is heavily biased towards maltose. We hypothesize that this is tied integrally to the role of ErAmy13B in the organism, playing a load-balancing function, distributing the products of the large cell-wall anchored amylase ErAmy13K between the maltose and maltooligosaccharide transporters that *E. rectale* upregulates in response to growth on starch. In other organisms, this functionality may serve to make these substrates accessible when the organism lacks a longer maltooligosaccharide transporter. This type of restricted substrate/product profile seems to be widespread among GH13_36 enzymes as well as other maltogenic amylase containing subfamilies, though there are multiple mechanisms of achieving this restriction. These differing mechanisms then in turn have consequences for other properties of the enzyme, such as tolerance for α-1,6-branchpoints, or production of glucose, which typically has one or more additional dedicated transporters in many organisms. Thus, these enzymes appear to play a central role in the ability of many organisms to utilize starch degradation products.

Materials and Methods

Materials

Crystallization screening kits were obtained from Hampton Research and Molecular Dimensions. Unless otherwise noted all other chemicals were obtained from Sigma.

Table 2.5: Primers used in this study. Mutations are shown as bold capital letters

Primer Name	**Sequence**
Amy13B_D265A_F[a]	aga ctg g**CC** gct gtg att tat tac aat aat aat aac cag acc gag
Amy13B_D265A_R[a]	Ata aat cac agc **GG**c cag tct aaa gcc atc tac gcc
Amy13B_K243A-K244A_F[a]	aac aat gtg aag agc **GC**a **GC**g gcc gat gct tat atg g
Amy5_F	cat cat cac cac cat cac gag aac ctg tac ttc cag ggc tca aaa tca gat tca tcc gac gga aa
Amy5_R	gtg gcg gcc gct cta tta ttc agc aga ctt aat gat tac cgt tga

Mutagenesis, Expression and Purification of ErAmy13B

To generate an inactive nucleophile mutant of ErAmy13B (D265A), site directed mutagenesis was performed on a wild-type construct in the pETite vector, which had been created previously [107]. The primers ErAmy13B_D265A_F and Amy13B_D265A_R (**Table 2.5**) were used at a 10 µM concentration along with 1.5 ng/µL of plasmid template DNA with the Phusion Flash high fidelity master mix (Thermo) in a 50 µL reaction. This was subjected to an initial denaturation of 15 s at 95 °C, followed by 30 cycles of 95 °C for 15 s, 65 °C for 30 s, 72 °C for 6 min followed by a final extension of 72 °C for 10 min. The resulting product was digested with *Dpn*I to remove template DNA and the reaction cleaned using the QIAquick PCR purification kit (Qiagen). The purified PCR product was then electroporated into electrocompetent *E. coli* S17 cells and plated on LB supplemented with 50 µg/mL kanamycin (Kan). Individual colonies were selected and inoculated into 5 mL of LB with 50 µg/mL Kan and the plasmid was isolated from overnight cultures using the Qiagen Plasmid Mini kit (Qiagen).

Plasmids were sent for Sanger sequencing at the University of Michigan DNA sequencing facility using the T7_F and T7_R sequencing primers (**Table 2.5**) to confirm the presence of the desired mutation and the absence of any additional mutations. Surface entropy reduction (SER) mutations were predicted using the SERp server (http://services.mbi.ucla.edu/SER) [165], which suggested a K543A/K544A double mutant would improve crystallization of the protein. These mutations were then introduced using the QuikChange Lightning Multi site directed mutagenesis kit with primer ErAmy13B_K543A-K544A_F (**Table 2.5**) according to the manufacturer's directions. The resulting plasmid, pErAmy13B_D265A/ K543A/K544A was sequenced as above. The triple mutant plasmid was transformed into chemically competent *E. coli* Rosetta II cells, which were then used to express ErAmy13B and the protein was purified as previously described [107].

Cloning and Expression of RbAmy5

The gene encoding wild-type RbAmy5 (GenBank Accession No.: SPE91476.1) was cloned similarly to wild-type ErAmy13B. Briefly, RbAmy5 was PCR amplified from genomic DNA for ligation-independent cloning into the pETite N-His vector (Lucigen Madison, WI, USA) according to the manufacturer's instructions and using the primers RbAmy5f and RbAmy5r (Table S1). The N-terminal primer contained a tobacco etch virus (TEV) protease cleavage site immediately downstream of the complementary 18 bp overlap (encoding the His tag) to create a TEV protease-cleavable His-tagged protein. The resulting pETite plasmid was transformed into Rosetta(DE3) pLysS cells, plated on LB agar containing 50 µg/mL Kan and incubated overnight at 37 °C. These plates were used to inoculate 1 L of terrific broth media supplemented with 50 µg/mL Kan and 30 µg/mL of chloramphenicol for protein expression. Cells were grown at 37 °C to an OD600 of ~0.6, isopropyl β-d-1-thiogalactopyranoside (IPTG) was added to a final concentration of 0.5 mM, and the cells were grown at 19 °C for an additional 16 h. Cells were then harvested by centrifugation and immediately used for protein purification.

Purification of Recombinantly Expressed RbAmy5

Cell pellets were resuspended in His buffer (25 mM NaH2PO4, 500 mM NaCl, 20 mM imidazole, pH 7.4), sonicated to lyse cells and centrifuged at 30,000 × *g* for 30 minutes. RbAmy5 was purified from the cell-free supernatant fraction on a 5 mL Hi-Trap metal affinity cartridge (GE Healthcare) column in His buffer and proteins were eluted with an imidazole (20-300 mM) gradient. The His-tag was removed by incubation with TEV protease (1:100 Molar ratio relative to protein) at room temperature for 3 h, then overnight at 4 °C while dialyzing against His buffer. The cleaved protein was then re-purified on the 5 mL Ni column to remove undigested target protein, the cleaved His-tag and His-tagged TEV protease. Purified RbAmy5 was dialyzed against 50 mM Tris buffer (pH 7.5) prior to crystallization and concentrated using Vivaspin 15 (10 000 MWCO) centrifugal concentrators (Vivaproducts).

Crystallization and Structure Determination of ErAmy13B

The D265A/K543A/K544A ErAmy13B (19 mg/mL) plus (15 mM) maltotetraose was subjected to a series of 96-well hanging drop sparse matrix screens to identify crystallization conditions. Crystals were obtained via hanging drop vapor diffusion at room temperature against 0.17 M ammonium sulfate, 25.5% PEG 4000, 15% glycerol (JCSG + screen, Molecular Dimensions). Crystals were obtained after several months and attempts to reproduce crystals of ErAmy13B failed. Prior to data collection crystals were cryoprotected by a swiping through a solution of 80% mother liquor supplemented with 20% ethylene glycol then plunged into liquid nitrogen. X-ray data were collected at the Life Sciences Collaborative Access Team (LSCAT) beamline ID-D of the Advanced Photon Source at Argonne National Laboratory. Data were integrated and scaled within XIA2 [166] using XDS [167] from the CCP4 package [168]. Phases were solved using molecular replacement with the structure of α-amylase from *Halothermothrix orenii* (HoAmyA; PDB 1WZA) [142] using Phaser-MR [169] within Phenix [170]. The structure was manually adjusted in Coot [171] then refined using Phenix.refine [172] and Refmac [173] and the conformation of bound carbohydrates was validated using Privateer from the CCP4 package [174].

Crystallization and Structure Determination of RbAmy5

RbAmy5 crystallization experiments were performed using a Crystal Gryphon (Art Robbins) in 96 well trays using a sitting drop format. Diffraction quality crystals of RbAmy5 were obtained by mixing 35 mg/mL protein 1:1 (vol/vol) with the crystallization solution containing 0.02 M ZnCl2 and 20% PEG 3350. Crystals were cryoprotected by a brief soak in 80% crystallization solution / 20% ethylene glycol and plunged into liquid nitrogen. X-ray data were collected at the Life Sciences Collaborative Access Team (LSCAT) beamline ID-G of the Advanced Photon Source at Argonne National Laboratory. Data were indexed and scaled using XIA2 [166] and XDS [167] from the CCP4 package [168]. The structure of RbAmy5 was determined by molecular replacement using the structure of HoAmyA (PDB 1WZA) [142]within the program MORDA from CCP4. The initial model was built using AutoBuild [175]within Phenix. The final model was constructed by alternate cycles of manual model building in Coot [171] and refinement in Phenix [172].

Isothermal Titration Calorimetry (ITC)

Activity of ErAmy13B against maltooligosaccharides was measured on a standard volume Nano ITC (TA Instruments). Enzymes and substrates were all dissolved in 10 mM

HEPES buffer at pH 6.5 with 150 mM NaCl and 2 mM CaCl2. The enzyme (150 nM for maltotriose measurements, 7.5 nM for all others) was placed in the cell (1.3 mL) and allowed to equilibrate to 37 °C. The substrate (100 mM maltotriose, 40 mM maltotetraose, 40 mM maltopentaose, 30 mM maltohexaose or 30 mM maltoheptaose) was placed in a syringe (250 μL) for injection into the ITC.

Stirring was initiated at 350 RPM and injections began after the instrument reached full equilibrium defined as 0.1 μW/h slope and 0.01 standard deviation. The method started with a 300 s baseline collection period followed by an injection program of a 2 μL injection, followed by 2× 5 μL injections, 3× 10 μL injections, 5× 15 μL injections and 6× 20 μL injections, each separated by 165 s intervals. Enzyme activity was measured by averaging the signal of the last 45 s of each injection period after the signal had reached a new steady state and subtracting the average of the baseline signal, giving activity in μJ/s. This was then converted to initial reaction velocity (s^{-1}) by dividing by the reaction enthalpy (determined to be 4.47 kJ/mol) and the enzyme concentration (corrected for dilution after each injection). The determination of the reaction enthalpy was performed by incubating a large excess of ErAmy13B (8.7 mg/mL) with maltotetraose (10 mM), allowing the reaction to go to completion and calculating the area under the curve. This was repeated and the averaged results were found to be within range of previously reported literature values for α-1,4-linked glucose hydrolysis [137]. This initial velocity and the substrate concentration at each injection (the amount injected up until that point with the calculated amount consumed subtracted) were then used in a Michaelis-Menten plot, with *k*cat and K_M solved via non-linear regression analysis in Microsoft Excel.

Sequence Analysis

The presence and type of signal peptides on enzymes was determined using SignalP 5.0 (http://www.cbs.dtu. dk/services/SignalP/) [176] with default settings for Gram-positive and Gram-negative organisms. Cellular localization of enzymes was predicted via PSORTb (https://www.psort. org/psortb/) [177] with default settings for Gram positive and Gram-negative organisms. Sequence alignments were performed with T-Coffee (https://www.ebi.ac.uk/Tools/ msa/tcoffee/) [178] with default parameters. Sequences for the characterized members of the GH13_36 subfamily and for the full subfamily were obtained from the CAZy database (http://www.cazy.org/) on 8/29/2019.

CHAPTER 3:

Sas20 is a Highly Flexible Starch-binding Protein in the *Ruminococcus bromii* Cell-surface Amylosome

The contents of this chapter have been published as: Cerqueira FM, Photenhauer AL, Doden HL, et al. Sas20 is a highly flexible starch-binding protein in the *Ruminococcus bromii* cell-surface amylosome. *J Biol Chem*. 2022;298(5):101896.

Abstract

Ruminococcus bromii is a keystone species in the human gut that has the rare ability to degrade dietary resistant starch (RS). This bacterium secretes a suite of starch-active proteins that work together within larger complexes called amylosomes that allow *R. bromii* to adhere to and degrade RS. Sas20 is one of the more abundant proteins assembled within amylosomes, but little could be predicted about its molecular features based upon amino acid sequence. Here, we perform a structure-function analysis of Sas20 which features two discrete starch-binding domains separated by a flexible linker. Sas20 domain 1 has an N-terminal β-sandwich followed by a cluster of α-helices and captures the non-reducing end of maltooligosaccharides between these structural features. The crystal structure of a close homolog of Sas20 domain 2 revealed a unique bilobed starch-binding groove that targets the helical α1,4-linked glycan chains found in amorphous regions of amylopectin and crystalline regions of amylose within starch granules. Affinity PAGE and isothermal titration calorimetry demonstrate both domains bind maltoheptaose and soluble starch with relatively high affinity ($K_d \leq 20$ μM) but exhibit limited or no binding to cyclodextrins. Small angle x-ray scattering analysis of the individual and combined domains support that these structures are highly flexible, which may allow the protein to adopt conformations that enhance its starch-targeting efficiency.

Introduction

The human gut microbiota, the dense and heterogeneous consortium of bacteria that reside in the intestinal tract, has a profound influence on host health and disease [179, 180]. Dietary fiber feeds this community and dictates the bacterial fermentation profile of short chain fatty acids that mediate several host responses [181]. Resistant starch (RS) is one such dietary fiber that tends to shift our gut bacterial community to one that promotes health [31]. While much of the processed starch in our diet is degraded by host or bacterial enzymes in the small intestine, a fraction of dietary starch resists enzymatic degradation and transits the large intestine. In the distal part of the gut, few specialized members of the microbiota can utilize RS [37, 182]. There are different types of RS classified according to the mechanism by which they are resistant to host intestinal enzymatic processing [183]. While not all RS has similar effects on our microbiome [36], RS consumption tends to increase colonic butyrate, a microbial short chain fatty acid that strengthens the gut barrier and has anti-inflammatory and anti-tumorigenic properties [33, 184-186].

Ruminococcus bromii is a primary degrader of RS and is considered a keystone species by cross-feeding starch breakdown products to other bacteria in the gut [37]. *R. bromii* organizes its starch-binding and starch-degrading proteins into one or more extracellular complexes called amylosomes [42, 117]. Akin to multiprotein cellulosome complexes synthesized by Gram-positive organisms for the degradation of cellulose, amylosomes are assembled via calcium-dependent protein-protein interactions [187, 188]. Like cellulosomes, amylosomes are built around a structural protein called a scaffoldin that possesses one or more cohesin modules. These cohesin modules bind to dockerin modules on secreted starch-targeting enzymes and binding proteins, creating a complex that hydrolyzes starch [42, 117, 182]. Biochemical studies on the recombinantly expressed cohesin and dockerin domains have revealed that there is a number of potential interactions among putative amylosome proteins [42, 117]. This suggests that there may be more than one type of amylosome synthesized, perhaps allowing the cell to respond to different environmental conditions, as has been observed for cellulosomes [189, 190].

A key feature of enzymes that degrade insoluble fibers like RS is the presence of carbohydrate binding modules (CBMs) [68]. CBMs are auxiliary modules of ~100 amino acids that bind to substrate and thus enhance enzymatic efficiency [81, 82]. CBMs are classified by amino acid sequence and there are currently 15 CBM families that target starch [182, 191]. While the precise molecular recognition varies, starch CBMs generally have a curved aromatic platform that complements the natural helical turn of the α1,4 glycosidic bond [68]. This molecular feature is also observed within the proteins of the starch utilization system (Sus) from the Gram-negative human gut bacterium *Bacteroides thetaiotaomicron*. The Sus features three cell-surface exposed starch-binding lipoproteins (SusDEF) and a single glycoside hydrolase 13 enzyme (SusG) that targets a-glucans such that starch-binding and hydrolysis are split across the four proteins [97]. Numerous examples of Sus-like complexes, comprised of glycan-binding proteins and enzymes that target many other carbohydrates, have been studied in detail in several Bacteroides species [192-195]. Other examples of bacterial complexes that include both non-catalytic carbohydrate-binding proteins and enzymes include cellulosomes from Gram-positive bacteria, in which both enzymes and carbohydrate-binding proteins dock to the scaffoldin, which may also feature carbohydrate-binding domains for docking to cellulose [196, 197].

Bioinformatic analysis of the *R. bromii* genome identified five scaffoldin proteins with cohesin domains (Sca1-5) and 27 proteins with dockerin domains [42, 117]. Only five of these dockerin-containing proteins have predicted glycoside hydrolase family 13 (GH13) catalytic modules that are specific for a-glucan degradation. This leaves 22 proteins, originally called "Doc" proteins 1-22, that may be incorporated into the amylosome. Many of these proteins likely bind starch, creating a system of starch adhering proteins that help tether the bacterium to RS granules. Here, we extend our previous work on the amylosome by characterizing one such dockerin-containing protein that assembles into this complex that we have named Sas20 for starch adherence system protein 20. Using a combination of x-ray crystallography, small angle x-ray scattering and isothermal titration calorimetry, we demonstrate that Sas20 is a highly flexible, starch-binding protein comprised of two domains with different starch-binding features. These data extend our molecular understanding of how a keystone human gut bacterium targets resistant starch in the gut.

Results

Sas20 is a Component of Cell-surface Amylosomes

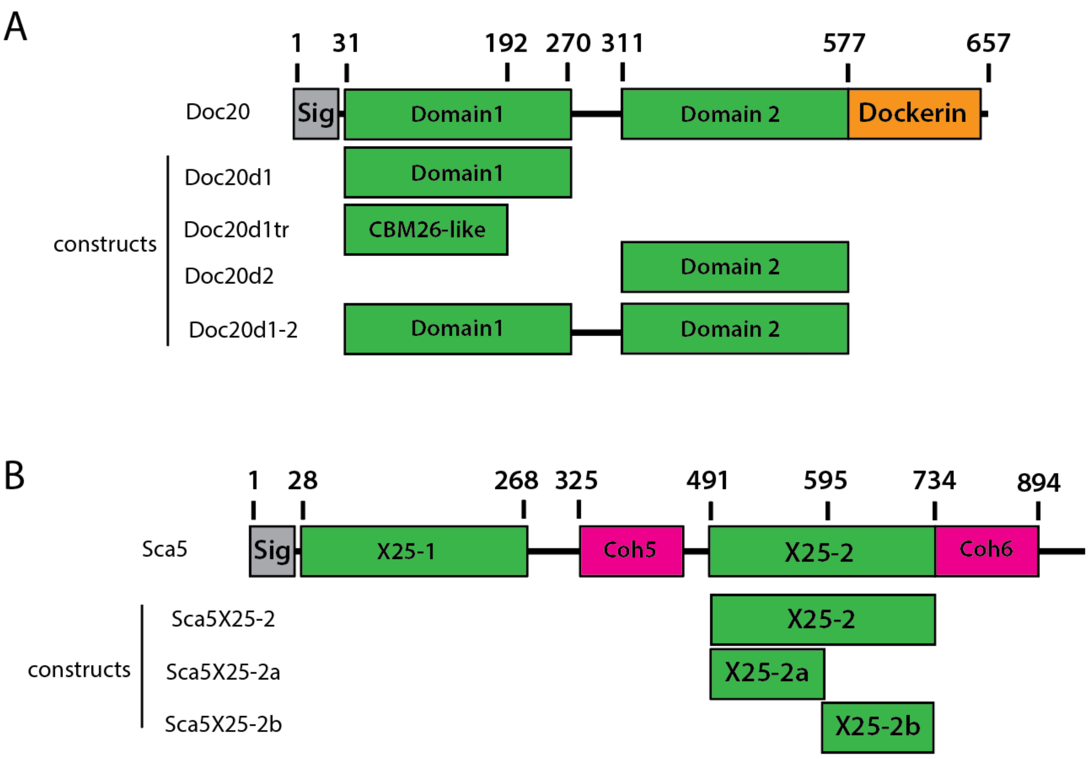
Previous work using the cohesin domain from Amy4, a cell-surface amylosome protein, as a probe to capture amylosome proteins from fractionated *R. bromii* cells identified Sas20 (previously named Doc20), as one of the more abundant proteins [117]. In the same study, Sas20 was also identified as one of the major proteins found in the cell pellet and cell culture supernatant of *R. bromii* cells grown on soluble starch. Following on these results, we sought to identify proteins that make up the cell-surface amylosome network by leveraging the calcium-dependent nature of cohesin-dockerin assembly [198, 199]. *R. bromii* cells were grown in either galactose or autoclaved potato amylopectin to early stationary phase, washed with PBS, then incubated in PBS with or without 10 mM EDTA to disrupt cohesin-dockerin interactions (see Materials and Methods) [42]. Proteomic analysis of the washed cells revealed many peptide spectral matches (PSMs) to predicted amylosome proteins, with an enrichment of these proteins in the EDTA-treated sample (**Table 3.1, all data in Table S1 of published manuscript**). Amy4, an amylase with both a cohesin and dockerin module, had the highest number of PSM matches in the EDTA samples. Interestingly, Amy1 and Amy2, secreted amylases that lack predicted cohesin or dockerin modules, were also higher in the EDTA wash. This may suggest that not all amylosome proteins interact via cohesin-dockerin interactions. Sca2 and Sca5, scaffoldin proteins that encode sortase recognition sequences, represented a negligible amount of the peptide repertoire in the PBS- or EDTA-wash conditions. Sas20 was also a protein for which there were more PSM assignments from the EDTA wash compared to the PBS wash in cells grown in either galactose or potato amylopectin. Intrigued by the recurring presence of Sas20 as an amylosome component across studies and its low sequence homology to characterized proteins, we performed a structure-function study of Sas20 to determine its role in the *R. bromii* amylosome.

Table 3.1: Highest abundant proteins from EDTA elution.

Locus tag	Name	# Amino Acids	PBS Gal PS M	AVG EDTA Gal PSM	PBS Amylo PSM	AVG EDTA Amylo PSM	Domain Architecture
L2-63_00682	Amy4	1356	19	107 ± 11.3	29	210.5 ± 2.1	SP GH13 CBM26 CBM26 Coh Doc
L2-63_00496	Amy2	751	17	76 + 9.9	29	128.5 ± 16.3	SP CBM26 GH13
L2-63_00433	Amy1	804	28	76.5 + 4.9	31	117.5 ± 7.8	SP CBM26 GH13
L2-63_01094	Amy10	1233	5	77 ± 14.1	2	115.5 ± 2.1	SP CBM48 GH13 MucBP MucBP CBM26 MucBP Doc CBM26
L2-63_01654	Amy16	876	11	68.5 ± 9.2	18	89.5 ± 4.9	SP GH13 CBM26 Doc CBM26
L2-63_00434	Doc22	548	12	16.5 ± 2.1	6	53 ± 1.4	SP CBM26 CBM26 DUF Doc
L2-63_00125	Sas20	630	6	40.5 ± 4.9	15	49 ± 1.4	SP Sas20d1 Sas20d2 Doc
L2-63_01357	Amy12	1059	0	23 ± 0.0	1	32.5 ± 3.5	SP CBM48 GH13 MucBP Doc MucBP CBM26
L2-63_02041	Amy9	1056	8	14 ± 1.4	25	30 ± 1.4	SP GH13 CBM26 Doc
L2-63_01861	Doc8	245	19	17 ± 0.0	14	22.5 ± 2.1	SP DUF Doc
L2-63_00436	Doc14	550	0	22.5 ± 0.7	0	21 ± 1.4	SP PEP A-S Doc
L2-63_00285	Doc1	549	2	16.5 ± 2.1	1	20.5 ± 2.1	SP LRR LRR Doc
L2-63_01443	Doc6	734	2	11.5 ± 3.5	1	17.5 ± 0.7	SP DUF Doc
L2-63_00287	Doc2	471	2	13.5 ± 2.1	2	15 ± 1.4	SP LRR Doc
L2-63_00780	Amy5	551	4	4.5 ± 0.7	4	10 ± 2.8	SP GH13

Common contaminants and cytoplasmic proteins were omitted. PBS samples n=1. EDTA samples are average of n=2. **PSM** Peptide Spectral Matches, **AVG** average, **Gal** galactose-grown cells. **Amylo** autoclaved potato amylopectin-grown cells. **SP** signal peptide, **PEP** Peptidase, **GH13** Glycoside Hydrolase family 13, **CBM** Carbohydrate Binding Module, **LRR** Leucine rich repeat, **DUF** domain of unknown function, **SP** signal peptide, **A-S** NAD(P)+-dependent aldehyde dehydrogenase superfamily.

Sas20 is a protein of 657 amino acids that has an N-terminal secretion signal, two predicted globular domains and C-terminal dockerin domain (**Figure 3.1A**). Domain 1 of Sas20 (Sas20d1) has no significant sequence homology to any proteins in the PDB and no sequence similarity (E-value < 0.05) to characterized proteins. Domain 2 of Sas20 (Sas20d2) has distant homology to the X25_BaPul-like family of starch-binding domains (E-value = 10^{-6}) [200]. A linker of 41 amino acids rich in Thr/Pro separates Sas20d1 and Sas20d2. Interestingly, Sas20d2 shares 81% sequence identity with residues 491-734 of Sca5, hereafter referred to as Sca5X25-2 as it is the second X25-containing domain in the sequence. Therefore, we included this domain in our analysis (**Figure 3.1B**). Sca5 is an 894 amino acid scaffoldin protein that also has an N-terminal secretion signal, two X25 modules, two cohesin domains, and a C-terminal sortase sequence [42].



C

	Doc20d1	Doc20d2	Sca5X25-2	Doc20d1-2
Corn Amylopectin	NB	+++	+++	+++
Potato Amylopectin	NB	+++	+++	+++
Glycogen	+	+	++	++
Pullulan	NB	++	+	++
Dextran	NB	NB	NB	NB

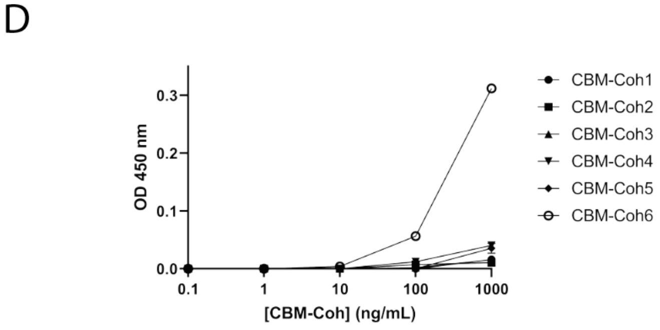


Figure 3.1: Protein constructs and affinity PAGE results. A) Sas20 and B) Sca5 constructs used in this study. C) Summary of affinity PAGE results for select polysaccharides; gels presented in Figure S1. NB= no binding. D) Functionality of the Sas20 dockerin as measured by ELISA. A microtiter plate was coated with Xyn-Sas20. Positive interaction of the Sas20 dockerin was observed with Coh6. Error bars indicate SD from the mean of duplicate samples from one experiment.

We created the construct Sas20d1-2 that lacks the dockerin domain and secretion signal, as well as the individual domains Sas20d1, Sas20d2 and Sca5X25-2 to determine their potential for starch binding via affinity PAGE (**Figures 3.1C, 3.2**) [201, 202]. In this method, protein binding is qualitatively assessed by a decrease in mobility through non-denaturing gel upon interaction with polysaccharide. For this analysis, we tested the soluble polysaccharides amylopectin, glycogen, pullulan, and dextran. Amylopectin is one of the two polysaccharides within starch granules and contains both α1,4 and α1,6 linkages, while glycogen, found in animals and bacteria, has a higher proportion of α1,6 branches [18, 19]. Pullulan is found in fungal cell walls and is a linear polysaccharide of maltotriose linked by α1,6 linkages [66, 71]. Sas20d2, Sca5X25-2, and Sas20d1-2 bind to corn and potato amylopectin with relatively high affinity as suggested by their retention at the top of the gels but demonstrated more moderate binding to glycogen and pullulan (**Figure 3.1C**). These data suggest that Sas20d2 and Sca5X25-2 accommodate α1,6 linkages but that binding is likely driven by binding to α1,4 glucan regions. While Sas20d1 only showed modest affinity to glycogen in this assay, we could quantify its binding to amylopectin via isothermal titration calorimetry (described later). We speculate that our inability to observe binding by Sas20d1 in this assay may be due to incompatibility of the protein with the electrophoresis conditions, as some aggregation may occur in the non-denaturing gel. None of the constructs bound dextran, an α1,6-linked glucan, underscoring the specificity of the Sas20 and Sca5 domains for α1,4-linked starch components.

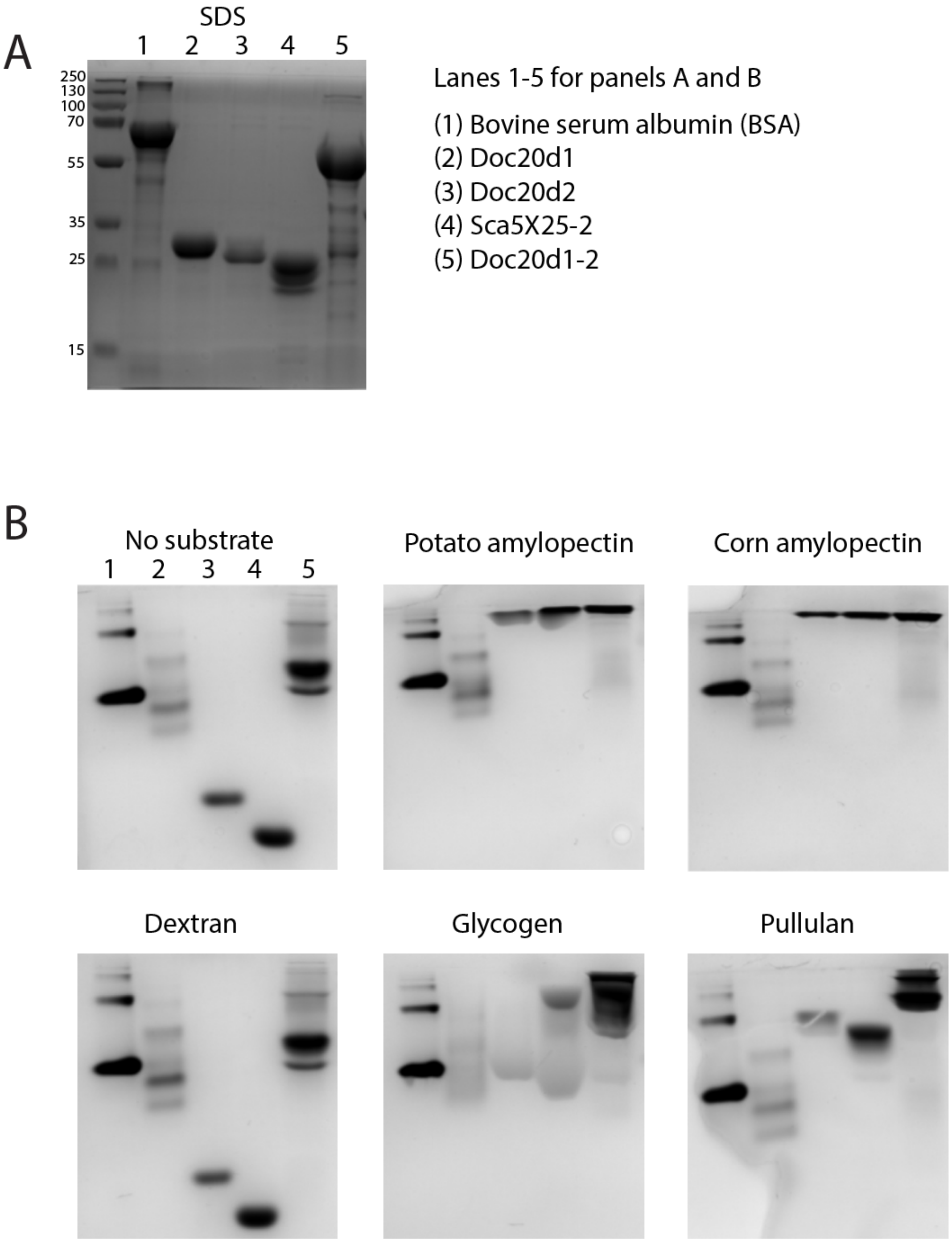


Figure 3.2: Sas20d1, Sas20d2 and X25-2 from Sca5 polysaccharide binding profile: Lanes are of the following purified proteins: (1) Bovine serum albumin (BSA) (2) Sas20d1 (3) Sas20d2 (4) Sca5X25-2 (5) Sas20d1-2. A) 10% SDS PAGE for purity. B) Proteins are separated in 10% acrylamide gel with or without 0.4% of the indicated polysaccharide: potato amylopectin, corn amylopectin, dextran from *Leuconostoc* ssp., pullulan from *Aureobasidium pullulans*, and glycogen from bovine liver.

To determine how Sas20 is assembled into the amylosome system, a standard affinity-based ELISA procedure was performed by using a fusion construct including the dockerin module from Sas20 [203]. We tested binding to the six known cohesin modules in the *R. bromii* genome (CBM-Coh1-6) and discovered that the Sas20 dockerin module interacts specifically with CBM-Coh6, the second cohesin of the anchoring scaffoldin Sca5 (**Figure 3.1D**). These data support the results of our proteomic experiments and suggest that Sas20 is a component of the cell-surface amylosome via its interaction with Sca5 and likely aids in the docking of *R. bromii* to starch granules.

Sas20 Domain 1 Structure

We solved the crystal structure of Sas20d1 via sulfur SAD phasing (2.1Å, R_w=17.7%, R_f=21.4%) and then used this as a model to determine the structure with maltotriose (1.5Å, R_w=17.5%, R_f= 19.7%; **Table 3.2**).

Table 3.2: X-Ray data collection and refinement statistics

	Sas20 native	**Sas20 maltotriose**	**Sca5X25-2 maltotriose**
PDB accession	7RAW	7RFT	7RPY
Wavelength (Å)	0.979	0.979	0.979
Resolution range (Å)	41.13-2.10 (2.15-2.10)	30.00-1.53 (1.56 -1.53)	39.27 - 1.67 (1.73 - 1.67)
Space group	I 21 3	C 1 2 1	P 32 2 1
Unit cell (Å)	*a*=*b*=*c*= 130.0	*a*=121.8, *b*=*c*= 64.7, β=102.8	*a*=*b*=100.8, *c*=87.9
Total reflections	319452 (13663)	339801 (14796)	556138 (53864)
Unique reflections	21541 (1051)	74182 (3699)	60154 (5957)
Multiplicity	14.8 (13.0)	4.6 (4.0)	9.2 (9.0)
Completeness (%)	100.0 (100.0)	100.0 (99.9)	100.0 (100.00)
Mean I/sigma(I)	40.5 (1.2)	32.5 (1.0)	17.1 (1.3)
R-merge	0.047 (2.31)	0.047 (1.44)	0.074 (1.77)
R-meas	0.074 (2.41)	0.053 (1.67)	0.078(1.87)
R-pim	0.019 (0.67)	0.025 (0.83)	0.026 (0.62)
CC1/2 in highest resolution shell	0.43	0.36	0.48
CC* in highest resolution shell	0.78	0.73	0.81
Reflections used in refinement	21522 (1388)	70481 (5079)	60153 (5958)
Reflections used for R-free	1995(144)	3699 (251)	3048 (331)
R-work	0.177 (0.281)	0.175 (0.319)	0.191(0.309)
R-free	0.214 (0.324)	0.197 (0.328)	0.203(0.309)
Number of non-hydrogen atoms	1921	4290	2255
macromolecules	1793	3641	1877
ligands	41	84	74
solvent	111	561	304
ions	n/a	4	n/a
Protein residues	233	464	241
RMS(bonds)	0.008	0.013	0.013
RMS(angles)	1.0	1.6	1.7
Ramachandran favored (%)	97.4	99.8	97.9
Ramachandran allowed (%)	2.6	0.2	2.1
Ramachandran outliers (%)	0	0	0
Rotamer outliers (%)	0.52	0	0
Clashscore	9.33	0.82	1.58
Average B-factor	66.58	24.3	25.0
macromolecules	65.54	25.4	22.6
ligands	98.8	24.0	34.6
solvent	56.0	36.1	36.9
ions	n/a	21.9	n/a

Sas20d1 has a canonical β-sandwich carbohydrate-binding module (CBM) fold at the N-terminus with a bundle of three α-helices at the C-terminus, with maltotriose accommodated between these features (**Figure 3.3A-C**). The N-terminal β-sandwich most closely resembles a CBM26 domain which can be found adjacent to catalytic domains on α-amylases and typically binds maltoheptaose and β-cyclodextrin [68, 77, 204, 205]. A search on the DALI server showed that CBM26 from the *Eubacterium rectale* α-amylase Amy13K (ErCBM26) had the highest structural homology to Sas20d1 and aligns with an RMSD of ~2.3Å over 85 Cα atoms (**Figure 3.3D**) [80, 206]. While ErCBM26 and Sas20d1 share a conserved β-sandwich fold, two long loops formed by residues 146-161 (Loop A) and 169-189 (Loop B) protrude from Sas20d1 and are not found in ErCBM26. These two loops are near the maltooligosaccharide-binding interface and residues of Loop A provide a hydrogen-bonding network for the O2 and O3 hydroxyls of the ligand. (**Figure 3.3D, E**). Maltotriose is primarily bound at the β-sandwich surface of Sas20d1 via the aromatic platform created by Y60 and W72. The non-reducing end O4 is directed towards the solvent-filled cavity between the β-sandwich and the α-helical bundle and does not directly interact with the protein (**Figure 3.3B, E**). The O2 of Glc1 is positioned 2.6Å and 2.9Å away from the side chains of T152 and N130, respectively. Q127 makes hydrogen bonds with Glc2 O2 and O3 while the side chain of N151 is located 3.1Å from Glc2 O2. At the reducing end, Glc3 has little direct interaction with the protein, with O2 positioned 3.0Å and 2.7Å away from the side chains of K157 and D154, respectively. While we later show that Sas20d1 binds maltoheptaose with enhanced affinity over maltotriose, our attempts at co-crystallization with maltoheptaose failed to demonstrate additional density at the non-reducing end, and only disordered density for an extra glucose at the reducing end, likely due to lack of productive interaction with the protein (data not shown). The φ (O5 -C1 -O4' -C4'), Ψ (C1 -O4' -C4' -C5') angles of maltotriose in our structure (φ = 102.4°, Ψ = -137.3°; φ = 103.8°, Ψ = -137.9°) are more obtuse than those found in double-helical amylose (φ = 91.8°, Ψ = -153.2°; φ = 85.7°, Ψ = -145.3°; φ = 91.8°, Ψ = -151.3°) [207]. Therefore, we think this domain targets more amorphous and less helical regions of starch at the non-reducing end of the α-glucan chain.

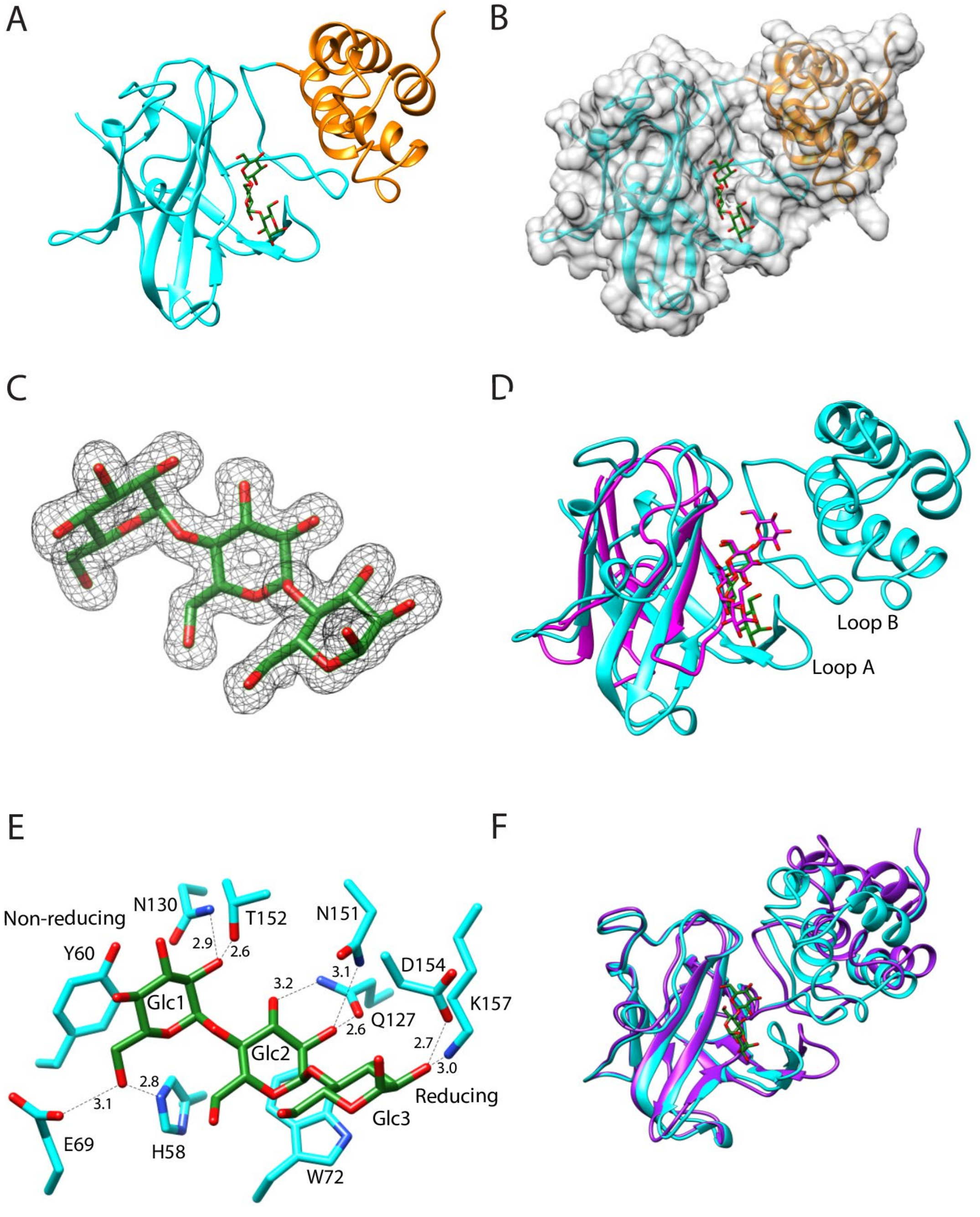


Figure 3.3: Sas20 domain 1 structure. A) Cartoon of Sas20d1 with maltotriose (green) with the β-sandwich (residues 34-190) in cyan and α-helical bundle (residues 191-268) in orange, B) Surface rendering of Sas20d1 structure demonstrating capture of maltotriose between the β-sandwich and helices. C) Omit map of maltotriose, σ=3.0 D) Structural alignment of Sas20d1 with maltotriose (cyan) and CBM26 (residues 279-387) with maltotetraose from Amy13k (ErCBM26, PDB:6B15, magenta). Residues 146-161 make up Loop A; residues 169-189 make up Loop B of Sas20d1. E) Close up view of maltotriose-binding site in Sas20d1 as colored in A. Hydrogen bonds are depicted as black dashed lines and with distances in angstroms. F) Overlay of the Sas20d1 native (purple) and maltotriose-bound (cyan) structures.

When comparing the native and maltotriose-bound Sas20d1 crystal structures, the CBM26-like fold at the N-terminus is nearly identical (**Figure 3.3F**). In the native structure, the α-helices at the C-terminus of Sas20d1 are somewhat disordered with elevated B-factors compared to the rest of the structure, but in the maltotriose-bound structure, this region is well ordered (**Figure 3.4A**). The Sas20d1 crystals with maltotriose (space group C2) have 45% solvent content and a tightly packed arrangement, with a crystal contact at the helical bundle. In each monomer, the helices (residues 237-257) are sandwiched between the same helical region (residues 237-257) and two β-strands (residues 58-70) of the neighboring monomer within the asymmetric unit and a loop (residues 93-104) of a symmetry-related monomer (**Figure 3.4B**). This arrangement is in stark contrast to the native crystals which were of the cubic space group I213 and have ~62% solvent. In these crystals, there are no crystal contacts in the region surrounding the helical bundle which in part explains the elevated B-factors.

In the maltotriose-bound structure, the helices move towards the ligand-binding site with a maximum displacement of ~8Å, although no part of this bundle directly interacts with maltotriose in our structure (**Figure 3.3F**). In solution, this flexibility may allow the protein to accommodate larger ligands and facilitate the capture of non-reducing ends between the β-sandwich and the helical bundle. We used CASTp to determine the size and volume of the solvent-accessible pocket created between the β-sandwich and α-helical bundle in both structures [208]. Not surprisingly, the pocket of the native structure has an area of ~783Å^2 and volume of ~1350 Å^3, while this space constricts to ~521Å^2 and a volume of ~848Å^3 in the maltotriose-bound structure (**Figure 3.4C**).

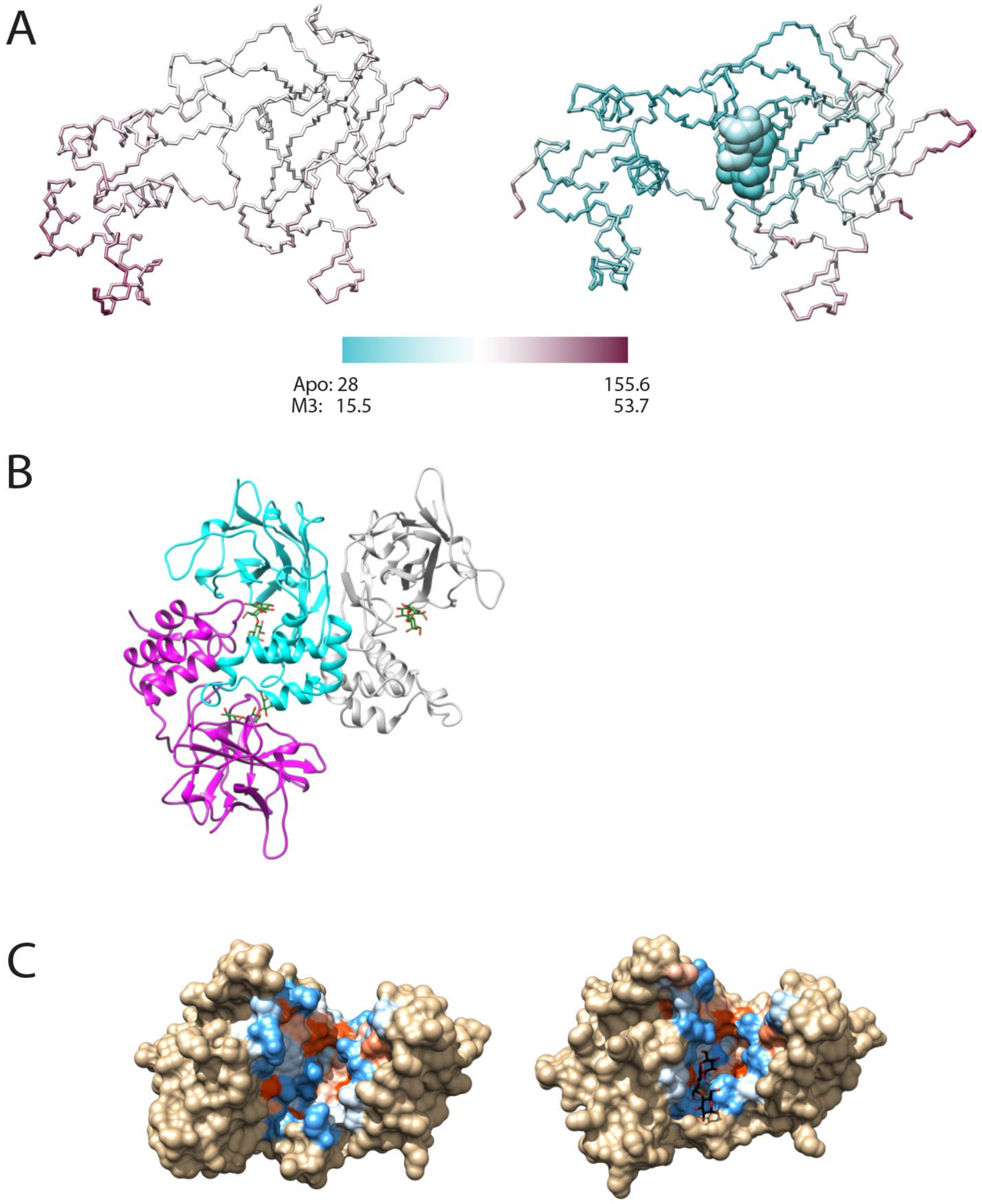


Figure 3.4: **Sas20d1 crystal structure analysis.** A) Ribbon depiction of native (left) and maltotriose-bound (right) Sas20d1 structures color ramped cyan to magenta from lower to higher B-factors. B-factor range for both structures is shown on the heat map legend. B) One of the crystallographic symmetry interactions between adjacent monomers in the Sas20d1 maltotriose crystals. Chain A (cyan) and chain B (gray) of an asymmetric unit with a symmetry-related chain (pink) that may effectively restrict the flexibility between the β-sandwich and helical bundle. Maltotriose is shown in green. C) CASTp results of the Sas20d1 native (left) and maltotriose (right) structures. The solvent exposed surface area is colored by electrostatic feature (red = negative charge, blue = positive charge) with default settings in Chimera.

Sas20 Domain 2 Homolog Structure

We could not obtain crystals of Sas20d2 but were successful in determining the structure of the Sca5X25-2 domain (residues 491-734) that is 81% identical in sequence (**Figures 3.1B, 3.5**). The Sca5X25-2 crystal structure with maltotriose was determined by SAD phasing with selenomethionine-substituted protein (1.7Å, R_w = 19.1%, R_f= 20.3%; **Table 3.2**). The Sca5X25-2 structure with maltotriose revealed two X25 modules in tandem, Sca5X25-2a and Sca5X25-2b (**Figure 3.6A**). X25 modules fold as a β-sandwich of ~120 amino acids and are found in tandem in the starch-binding proteins SusE and SusF from *Bacteroides thetaiotamicron* [66] and are features of some GH13 enzymes, such as the *Bacillus acidopullyticus* pullulanase [90]. Interestingly, both the *R. bromii* scaffoldins Sca3 and Sca5 have multiple predicted X25 modules [42]. Sas20d2 and Sca5X25-2 are roughly twice the size of a single X25 domain, so we predicted two X25 modules in tandem, each with its own starch-binding site (**Figure 3.1B)**. However, a single maltotriose molecule was captured between these modules with amino acids from both lobes coordinating the ligand (**Figure 3.6A, B**). The aromatic ring of W509 in Sca5X25-2a interacts via van der Waals forces with the hexose ring of Glc3 at the reducing end. The O2 and O3 of Glc3 is stabilized by hydrogen bonding to the side chains of Sca5X25-2a N564 and Sca5X25-2b N684. The aromatic rings of W661 and side chain of K654 in Sca5X25-2b interact with the aglycone face and O2 of Glc2, respectively. The O6 of Glc2 is within 2.5Å of the side chain of Sca5X25-2a E508. Glc1 interacts with W620, and its O2 and O3 coordinate with the side chain of N687. A sequence alignment between Sca5X25-2 and Sas20d2 shows that these residues within the ligand-binding cleft are conserved in the Sas20d2 sequence, suggesting that starch-binding sites in Sca5X25-2 and Sas20d2 are similar (**Figure 3.5**). Sca5X25-1 also shares conservation of these residues suggesting that there are multiple starch-binding sites within Sca5.

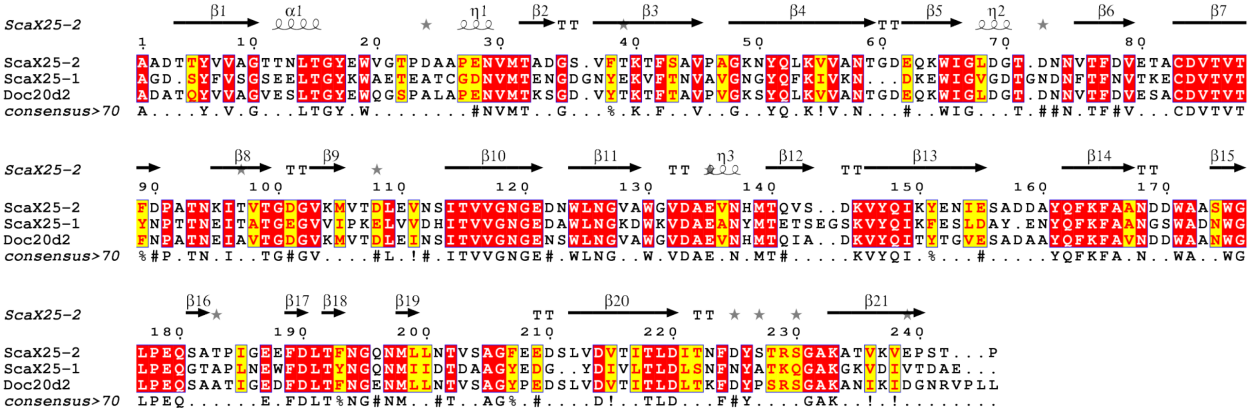


Figure 3.5: Sequence analysis of Sca5X25-1, Sca5X25-2, and Sas20d2. Alignment of the first and second X25 modules of Sca5 along with Sas20d2. Alignment was performed with T-Coffee and rendered in ESPript with secondary structure annotations referenced by Sca5X25-2 crystal structure.

Sca5X25-2a and Sca5X25-2b overlay with an RMSD of 1.0Å over 49 Ca atoms and demonstrate a conserved binding platform; when maltotriose is included in this overlay, the ligand displays the same polarity. A search on the DALI server revealed that the Sca5X25-2a and 2b folds share homology with the X25 domain in the *Bacteroides thetaiotamicron* starch-binding protein SusF (PDB 4FE9, Z-score= 7.8, RMSD 2.5Å; **Figure 3.6C, D**), including a conserved starch-binding site. W620 and W661 of Sca5X25-2a are conserved with W509 and W555 of Sca5X25-2b, although W555 was not involved in maltotriose binding in our structure. The position of W555 suggests that the binding platform shared between both lobes of Sca5X25-2 is extensive and can either accommodate longer maltooligosaccharides or allow each lobe to bind maltooligosaccharide independently. SusF has three X25 modules akin to Sca5X25-2a/b, and each recognizes maltooligosaccharides with K_d s of ~300µM [209]. However, for both Sca5X25-2a and Sca5X25-2b to bind individual maltooligosaccharides there would have to be significant opening of the cleft between these lobes. The φ (O5 -C1 -O4' -C4'), Ψ (C1 -O4' -C4' -C5') angles of maltotriose in our structure are φ = 107.5°, Ψ = -144.3°; φ = 90.8°, Ψ = -153.7°. The first φ/Ψ angles that is near the end of the chain is more obtuse, while the φ/Ψ angles cloistered within the binding cleft are similar to those found in double-helical amylose [207]. In contrast to Sas20d1, the architecture of the Sas20d2 binding site suggests to us a preference for helical regions within α-glucan.

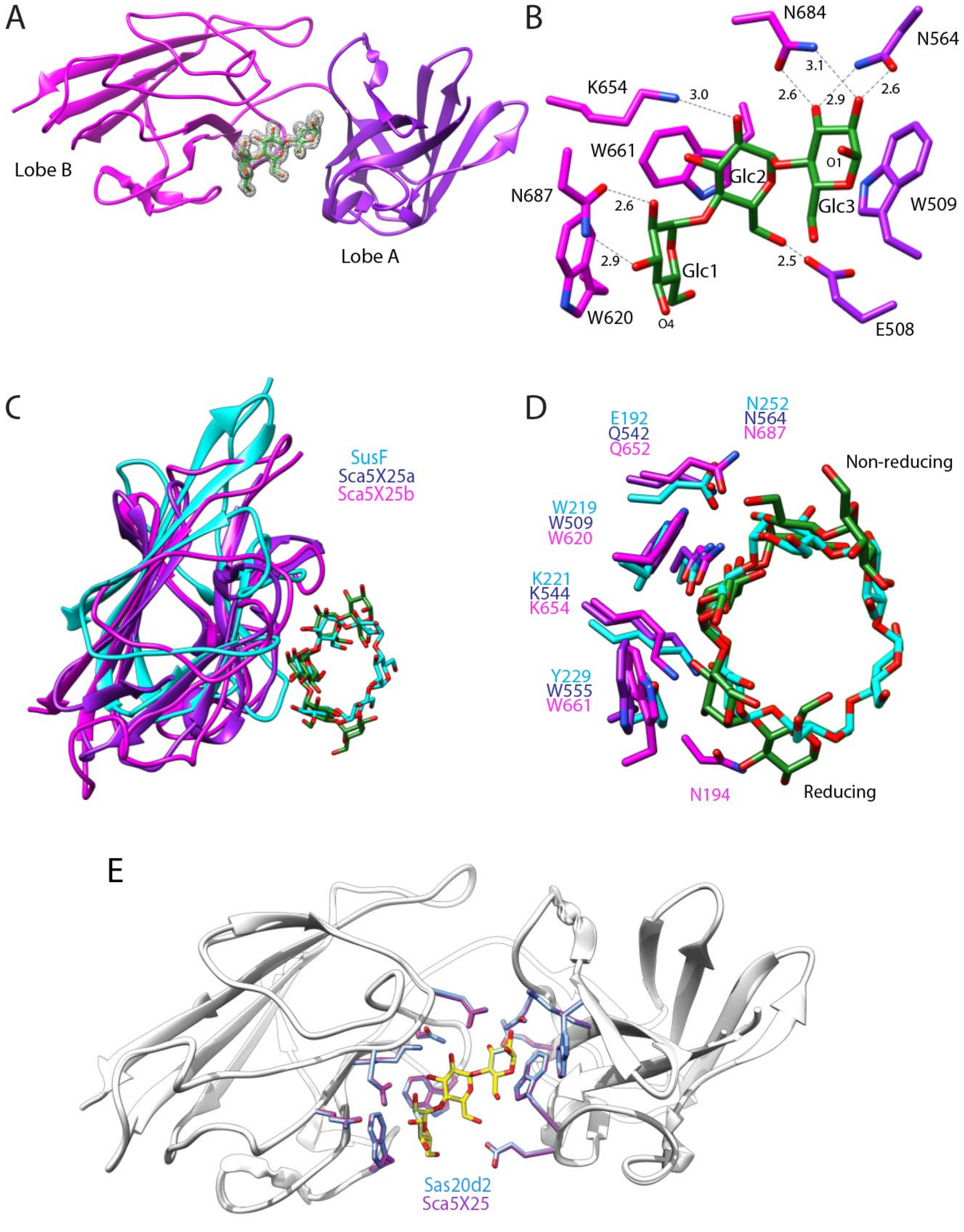


Figure 3.6: Sca5X25-2 structure. A) Cartoon of Sca5X25-2, with Sca5X25-2a (residues 491-595) in purple, Sca5X25-2b (residues 596-734) in pink. Omit map of maltotriose, σ=5.0. B) Close up of the maltotriose binding site colored as in panel A. Hydrogen bonds are depicted as black dashed lines and their distances are noted in angstroms. C) Overlay of Sca5X25-2a (purple), Sca5X25-2b (pink), and residues 170-272 from α-cyclodextrin-bound SusF (PDB:4FE9, cyan) D) Close up of binding site from the overlay in panel C demonstrating the conserved starch binding site. E) Phyre2 model Sas20d2 (gray ribbon, blue residues) overlaid on Sca5X25-2 (white ribbon, pink and purple residues as in panel B) The RMSD is 0.4Å for 240 Cα. The four conserved tryptophans are numbered according to the Sas20d sequence.

Domain 1 of Sas20 Binds to Extended α-glucan Structures

We used ITC to quantify the affinity of maltotriose, maltoheptaose, and solubilized corn and potato amylopectin binding to the domains of Sas20 and the Sca5X25-2 (**Table 3.3, Figures 3.7-11**). Sas20d1 binds to maltoheptaose (K_d= 1.5 ± 0.3 μM) with a K_d nearly two orders of magnitude stronger than maltotriose (K_d= 187.9 ± 58.1 μM). While the crystal structure revealed a short binding platform for three glucose residues, the enhanced affinity of maltoheptaose suggests that our crystal structure does not capture all possible interactions between the protein and ligand [77]. As mentioned earlier, we determined a crystal structure of Sas20d1 with maltoheptaose, but did not observe additional density at the non-reducing end beyond that of the maltotriose structure. We did note some fading density towards the reducing end that is directed outside of the binding cleft, supporting a lack of specific interaction with the protein at this end. Manual inspection and modeling of an additional glucose at the non-reducing end that is tucked within the binding cleft revealed that Sas20d1 can accommodate a longer ligand here, though there is somewhat more space if modeled in the native structure (**Figures 3.12A, B, C).** We did not observe an additional aromatic residue within this cleft, however, that might provide a platform for an additional glucose. An intermediate conformation of the helices between the maltotriose-bound and native Sas20d1 structures may lead to additional protein-ligand interactions that support maltoheptaose binding, although we could not capture this binding *in crystallo*. Regardless, the structure with maltotriose suggested that this domain has some specific preference for binding at the non-reducing ends of starch and maltooligosaccharides. This may in part account for the apparent lack of binding in affinity PAGE with amylopectin, as there is a very low concentration of polymer ends in a high molecular weight polysaccharide (MW ~10^8 Da) [210]. However, we found that Sas20d1 binds to both corn (K_d= 10.0 ± 1.7 μM) and potato amylopectin (K_d= 17.6 ±

7.2 μM), demonstrating a slight preference for corn amylopectin **(Table 3.3)**. Therefore, it is likely that some aspect of the affinity PAGE assay was incompatible with Sas20d1 starch binding.

Table 3.3: **Affinity of Sas20 and Sca5 constructs for starch substrates determined by ITC**.

Protein	Ligand	N (**binding sites**)	*K_d* (**μM**)
Sas20d1	Maltotriose	1.14 ± 0.28	187.9 ± 58.1
	Maltoheptaose	0.893 ± 0.38	1.53 ± 0.341
	β-cyclodextrin	NB	NB
	α-cyclodextrin	NB	NB
	PNP-M6	1.15 ± 0.07	0.870 ± 0.48
	B-PNP-M7	1.28 ± 0.29	7.12 ± 1.53
	Corn amylopectin	1*	10.0 ± 1.74
	Potato amylopectin	1*	17.6 ± 7.18
Sas20d1 Y60A	Maltotriose	NB	NB
	Maltoheptaose	1.55 ± 0.18	8.29 ± 0.513
Sas20d1 W72A	Maltotriose	NB	NB
	Maltoheptaose	NB	NB
Sas20d1tr	Maltotriose	1*	>1000*
	Maltoheptaose	1.45 ± 0.27	154.9 ± 63.0
	β-cyclodextrin	1*	1050 ± 168
	α-cyclodextrin	NB	NB
Sas20 domain 2	Maltotriose	1.18 ± 0.05	912.4 ± 110
	Maltoheptaose	1.15 ± 0.15	0.61 ± 0.027
	Corn amylopectin	1*	7.86 ± 1.4
	Potato amylopectin	1*	5.68 ± 1.5
	β-cyclodextrin	0.983± 0.09	532.7± 16.27
	α-cyclodextrin	NB	NB
Sas20d2 W329A	Maltotriose	NB	NB
	Maltoheptaose	1.33 ± 0.13	90.84 ± 25.7
Sas20d2 W375A	Maltotriose	NB	NB
	Maltoheptaose	1.12 ± 0.41	88.07 ± 36.0
Sas20d2 W440A	Maltotriose	NB	NB
	Maltoheptaose	1.39 ± 0.37	89.99 ± 7.72
Sas20d2 W481A	Maltotriose	NB	NB
	Maltoheptaose	NB	NB
Sca5X25-2	Maltotriose	1.02 ± 0.62	595.8 ± 51.4
	Maltoheptaose	0.81 ± 0.09	0.21 ± 0.029
	β-cyclodextrin	0.958 ± 0.01	346.4 ± 78.8
	α-cyclodextrin	NB	NB
Sca5X25-2a	Maltotriose	NB	NB
	Maltoheptaose	NB	NB
Sca5X25-2b	Maltotriose	NB	NB
	Maltoheptaose	NB	NB

Sca5X25-2 tryptophans that correspond to Sas20d2 tryptophans that were mutated are in parenthesis. **NB**: no binding detected. **PNP-M6**: PNP-α-maltohexaose. **B-PNP-M7**: PNP-α-maltoheptaose with a 4,6-linked-O-benzylidine group at the non-reducing end. Asterisk denotes fixed N or K_d. Each N and K_d are average of three replicates. N was set to 1 for amylopectin interactions as molarity cannot be defined for polysaccharides and both the crystal structures and oligosaccharide N is approximately 1. The Kd reported here is for the concentration of binding sites per gram of substrate.

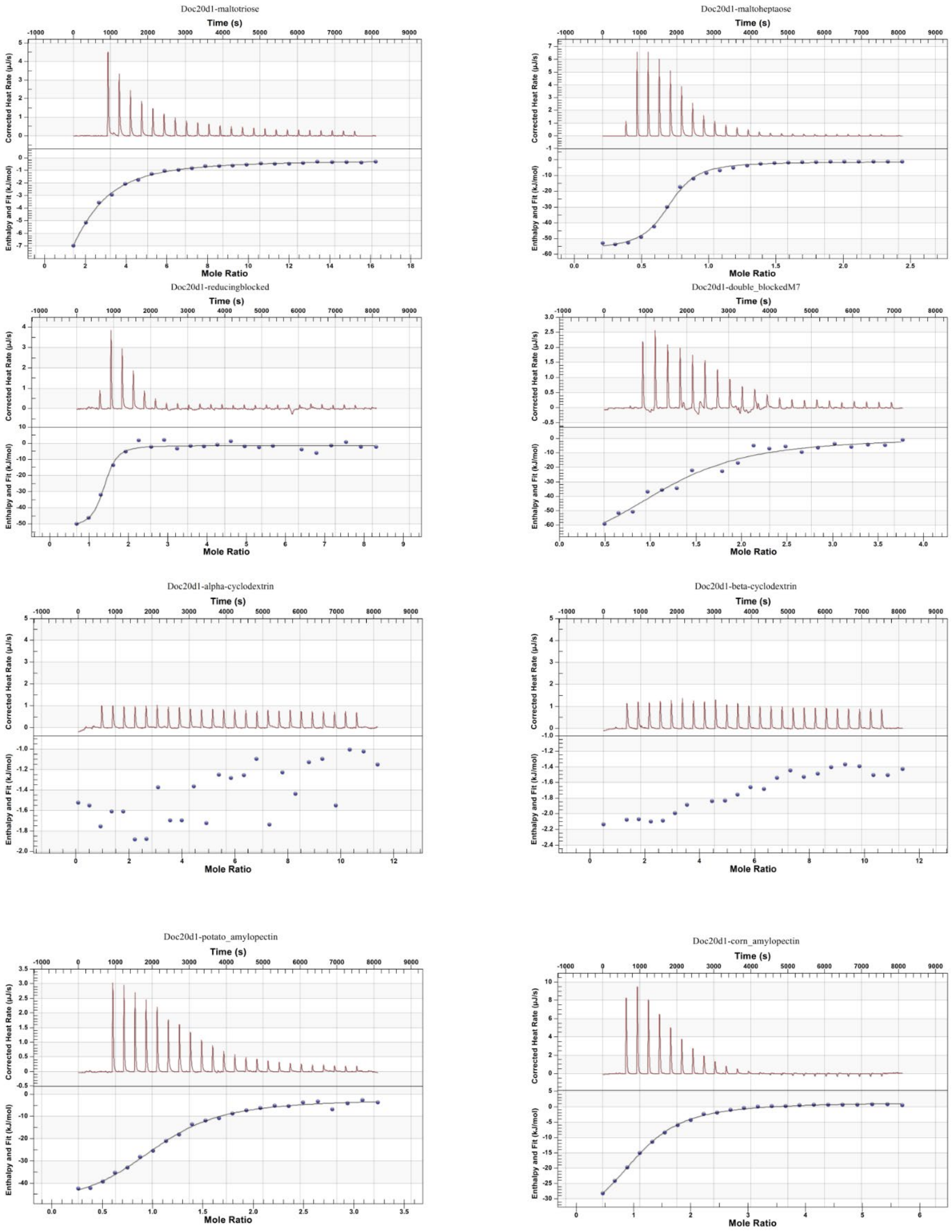


Figure 3.7: Representative ITC curves for Sas20d1 analysis.

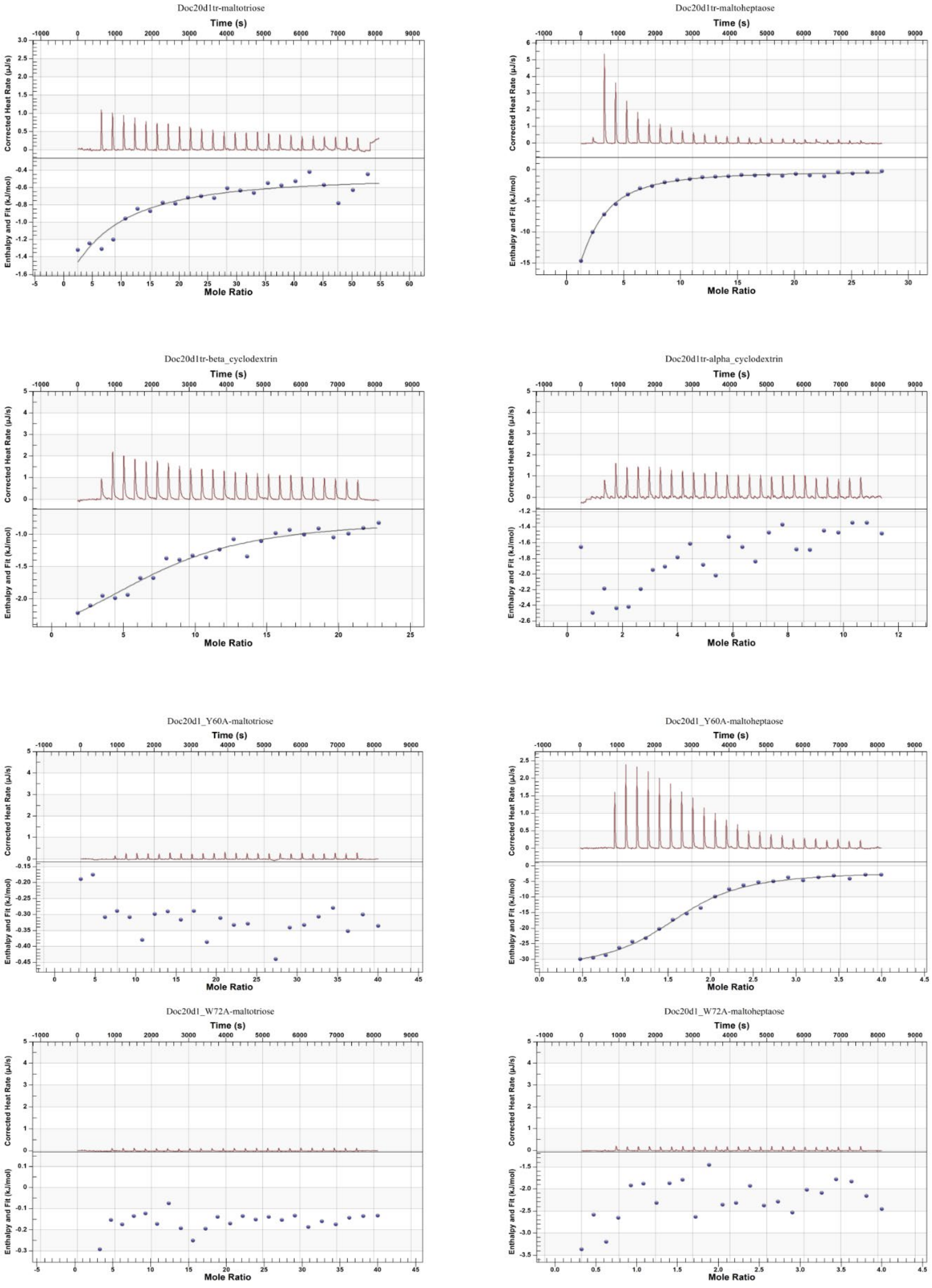
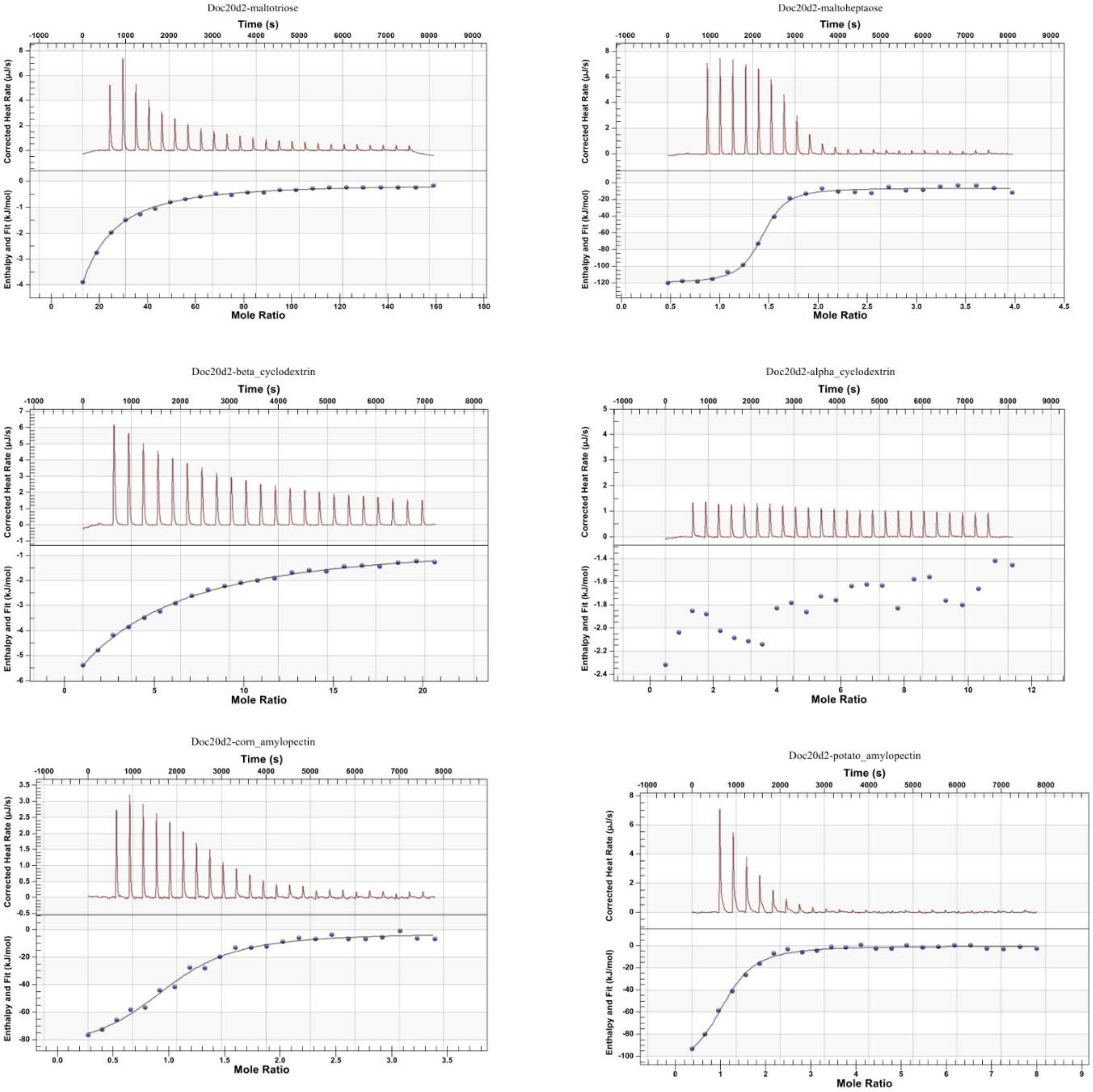


Figure 3.8: Representative ITC curves for mutant Sas20d1 analysis.



Doc20d2-maltotriose

Doc20d2-maltoheptaose

Doc20d2-beta_cyclodextrin

Doc20d2-alpha_cyclodextrin

Doc20d2-corn_amylopectin

Doc20d2-potato_amylopectin

Figure 3.9: Representative ITC curves for Sas20d2 analysis.

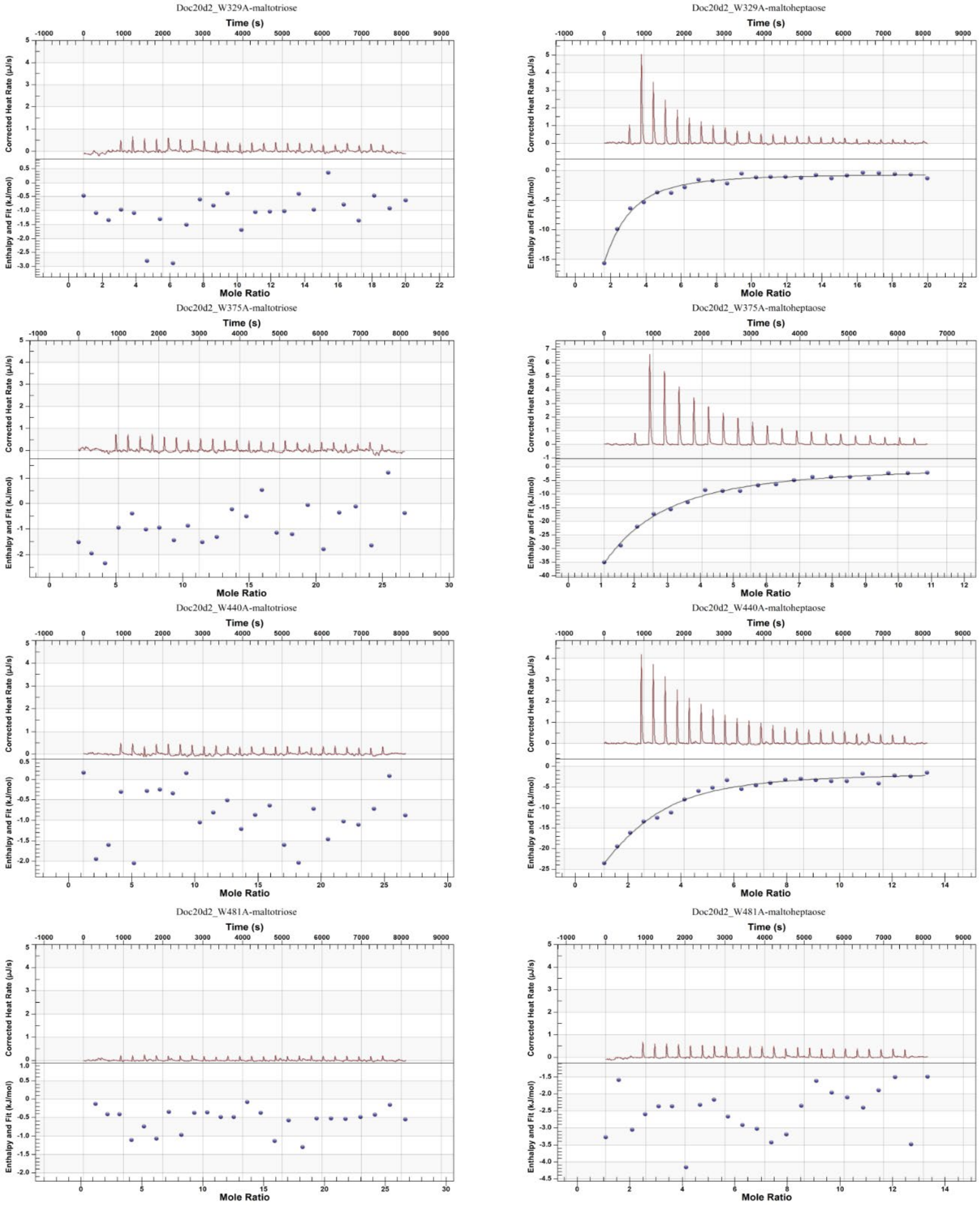


Figure 3.10: Representative ITC curves for mutant Sas20d2 analysis.

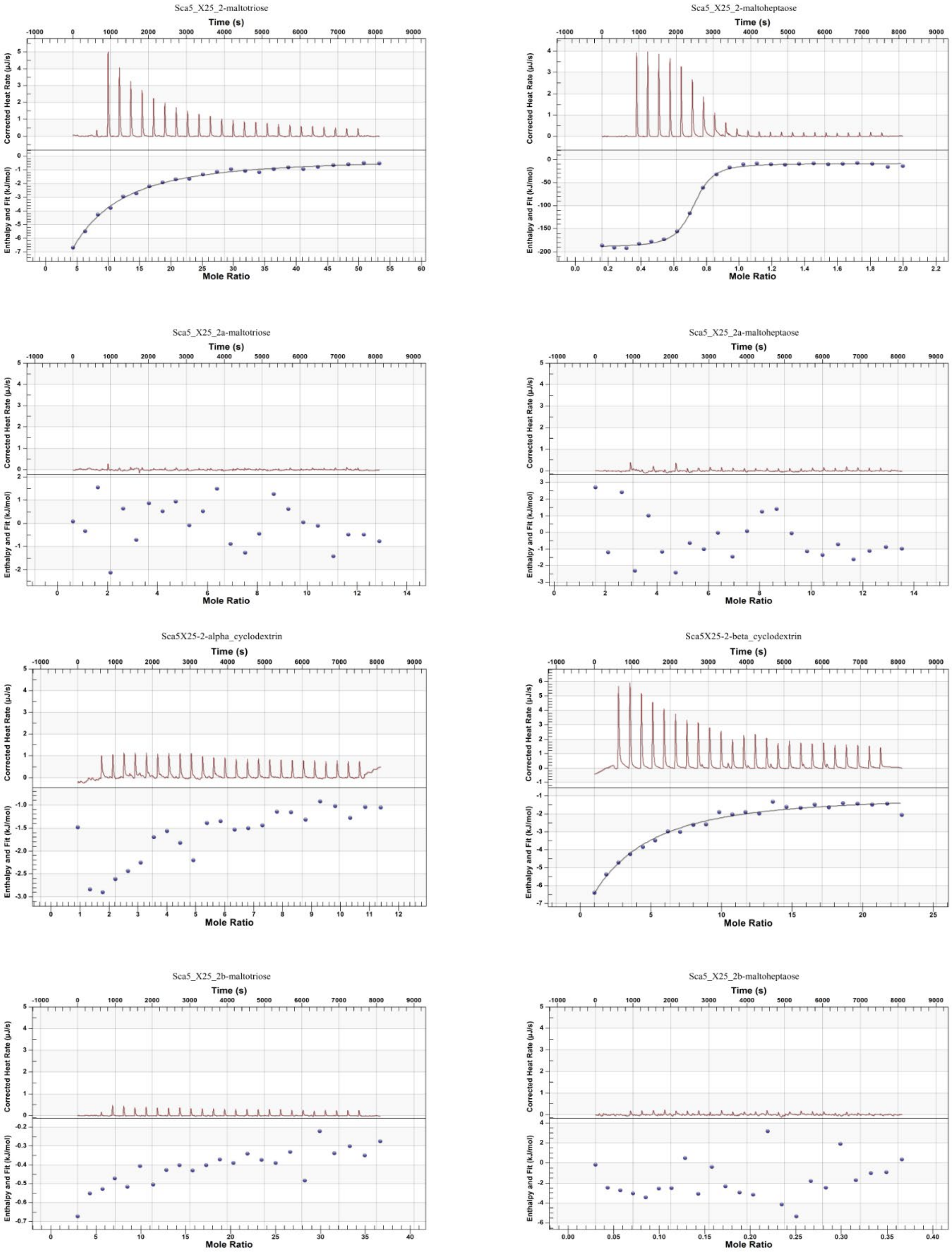


Figure 3.11: Representative ITC curves for Sca5X25-2 analysis.

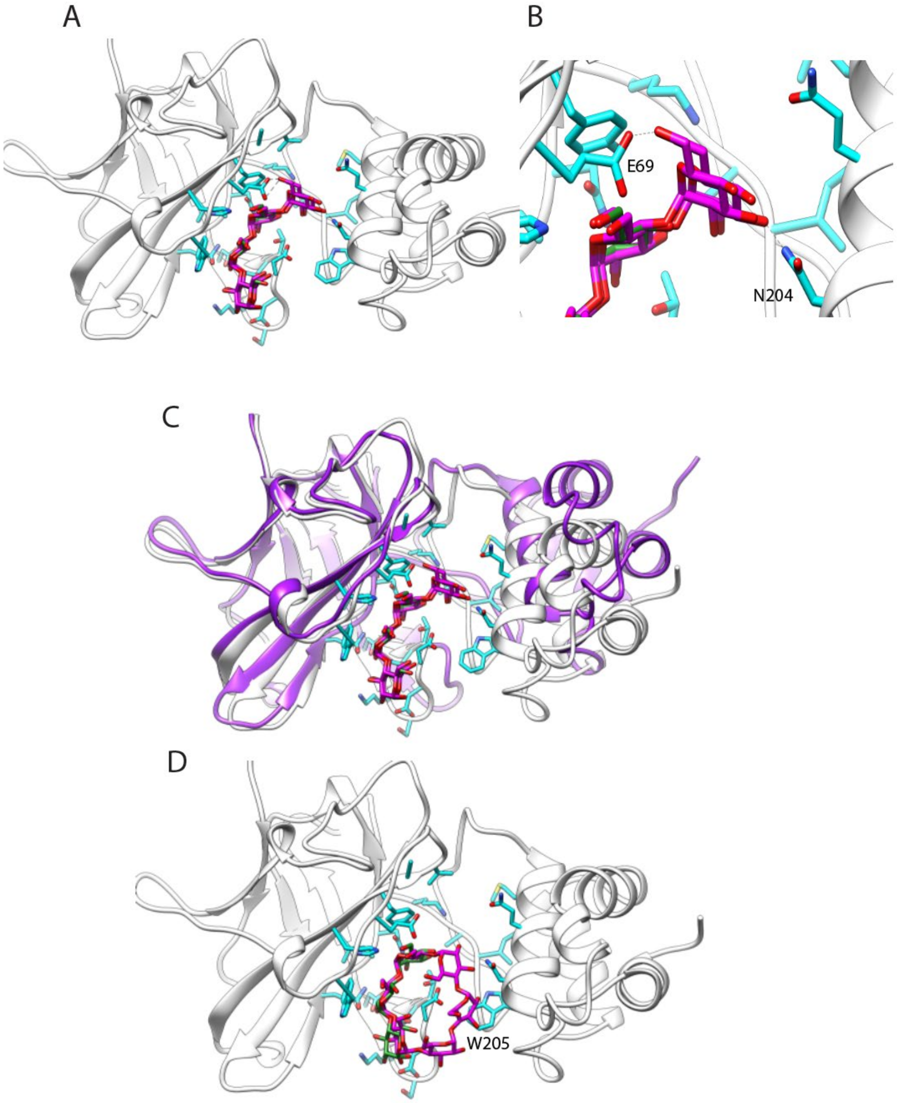


Figure 3.12: The Sas20d1 binding site can accommodate an additional glucose at the non-reducing end. A) Sas20 structure with maltotriose is displayed as white ribbon and maltotriose as green and red sticks. The maltotetraose (magenta) from the ErCBM26 structure (PDB 6B3P) was manually overlaid on top of maltotriose in two similar poses to extend the non-reducing end of the ligand within the binding cavity. All residues of Sas20 within 5.5Å of the modeled maltotetraose are displayed in cyan. The closest contacts with the modeled non-reducing glucose are from E69 (2.8Å) and N204 (2.0Å), the latter of which could adopt an alternative conformation. B) Close-up view of the modeled non-reducing end glucose from panel A. C) Same view as panel A with the native structure of Sas20d1 in purple ribbon, demonstrating the movement of the helices away from the binding cavity. D) Sas20 structure with maltotriose is displayed as white ribbon and maltotriose as green and red sticks. The α-cyclodextrin (magenta) from PDB 3CK7 was manually overlaid on top of maltotriose, demonstrating how the cyclic molecule would sterically clashes with W205.

Sas20d1 failed to bind α- or β-cyclodextrin supporting our observation that binding is restricted to chain ends. Indeed, when we attempted to model α-cyclodextrin on top of the maltotriose in our structure there was steric clashing with W205 from the helical bundle **(Figure 3.12D)**. To test whether the non-reducing ends of maltooligosaccharides are required for binding, we tested binding to benzylidene-blocked para-nitrophenyl maltoheptaoside (B-PNP-M7) which has a para-nitrophenyl (PNP) group at the reducing end and 4,6-linked-O-benzylidine at the non-reducing end. We also tested a PNP-α-maltohexaose which has an exposed O4 at the non-reducing end. Surprisingly, Sas20d1 bound both ligands with a similar K_d as maltoheptaose, though B-PNP-M7 bound with slightly less affinity (**Table 3.3**). Therefore, while our structural and biochemical data support that binding by Sas20d1 is likely limited to chain ends, there is indeed some flexibility within the binding cleft to accommodate a blocked nonreducing end. Specific recognition of the non-reducing end O4 by Sas20d1 is not required for binding.

To further examine the nature of Sas20d1 binding, we created single mutants Y60A and W72A. The Y60A Sas20d1 mutant binds to maltoheptaose but not maltotriose, while the W72A mutant did not bind either ligand. This suggests that W72, which is positioned at the reducing end of the binding platform, is required to anchor maltooligosaccharides and perhaps aids in guiding the nonreducing end of the ligand into place. Y60 creates a platform for binding the aglycone face of the nonreducing end glucose and is clearly essential for shorter oligosaccharides, perhaps because these are wedged further within the binding cleft and therefore are not stabilized by interaction with W72. That Y60 is not required for maltoheptaose binding further suggests that there may be additional interactions between ligand and protein that extend beyond the non-reducing end of maltotriose in our structure, but they are difficult to predict from the current models (**Figure 3.12**).

C-terminal Helices are Important for Substrate Binding in Sas20d1

Although the helical bundle at the C-terminus of Sas20d1 does not directly interact with maltooligosaccharide, we hypothesized that its presence is an important feature that either lends structural stability to the binding pocket or restricts the binding of cyclodextrins. A truncated version of Sas20d1 lacking these helices (Sas20d1tr, **Figure 3.1A**) displayed dramatically reduced binding for maltotriose that could not be quantified via ITC, while binding for maltoheptaose decreased by ~100-fold (**Table 3.3**). This truncation did not facilitate binding of α- or β-cyclodextrin at relevant biological levels (K_d >1mM). We therefore speculate that these helices support competent binding by providing stability to loops A and B (**Figure 3.3D**).

Table 3.4: Calculated secondary structure of Sas20d1 with or without substrate bound.

	Secondary Structure (%)			
Condition	α-Helix	β-Strand	β-Turn	Unordered
Sas20d1	11.7 ± 1.5	33.3 ± 1.5	23.3 ± 0.6	31.7 ± 0.6
Sas20d1 + M3	9.7 ± 0.6	34.7 ± 0.6	23.3 ± 0.6	31.3 ± 1.0
Sas20d1 + M7	13.0 ± 1.7	30.0 ± 1.7	25.0 ± 0.0	32.0 ± 0.0

Values represent the means ± SD based on three replications **M3=** maltotriose, **M7=** maltoheptaose

To test if the helices have more order in solution when Sas20d1 is bound to substrate, circular dichroism (CD) was performed on Sas20d1 alone or with maltotriose or maltoheptaose (**Table 3.4, Figure 3.13A**). However, there was no significant shift in secondary structure in the presence or absence of substrate. We then tested if WT Sas20d1 could resist thermal unfolding compared to the Sas20d1tr construct (**Table 3.5, Figure 3.13B, C**). As expected, we observed a marked decrease in α-helical quality in Sas20d1tr compared to the full-length domain. However, the percentage of unordered region remained the same across both Sas20d1 and Sas20d1tr at all temperatures suggesting that the C-terminal helices in Sas20d1 contribute marginally to the stability of this domain.

Table 3.5: Calculated secondary structure of Sas20d1 and Sas20d1tr at temperature intervals.

	Secondary Structure (%)			
Condition	α-Helix	β-Strand	β-Turn	Unordered
Sas20d1 25ºC	11.0 ± 1.0	33.3 ± 0.6	23.0 ± 0.0	32.0 ± 1.0
Sas20d1 55ºC	19.7 ± 1.2	22.33 ± 1.5	26.3 ± 1.5	31.7 ± 0.6
Sas20d1 85ºC	11.3 ± 1.5	28.7 ± 1.5	25.0 ± 0.0	34.7 ± 0.6
Sas20d1tr 25ºC	2.7 ± 0.6	39.3 ± 0.6	23.0 ± 1.0	32.3 ± 0.6
Sas20d1tr 85ºC	8.7 ± 2.1	31.7 ± 2.3	24.7 ± 0.6	34.7 ± 0.6

Values represent the means ± SD based on three replications.

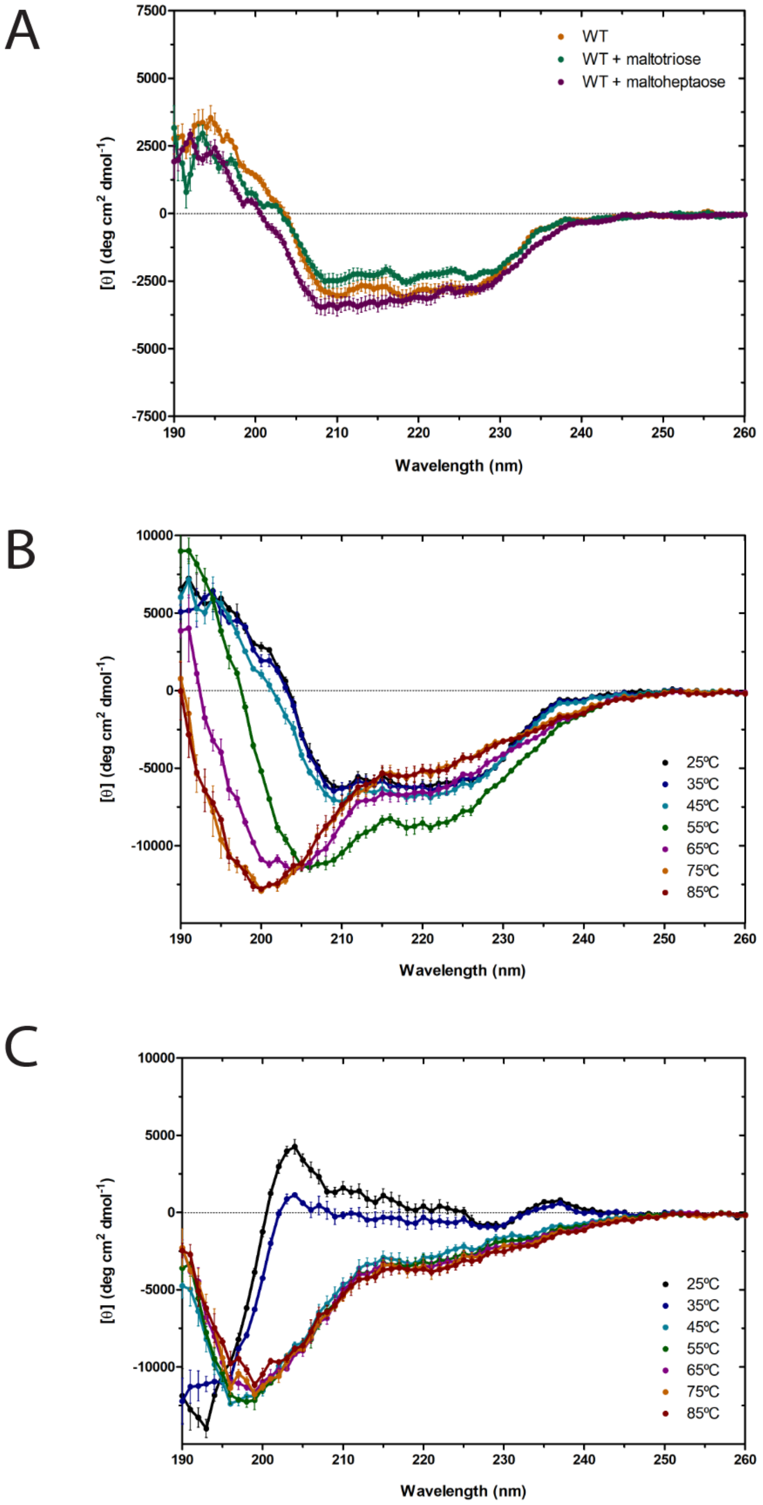


Figure 3.13: Circular dichroism on Sas20 constructs. A) Sas20d1 with no substrate, maltotriose, or maltoheptaose. CD melting curves for B) Sas20d1 and C) Sas20d1tr. **WT=** Wildtype Sas20d1

Domain 2 of Sas20 Binds to Starch

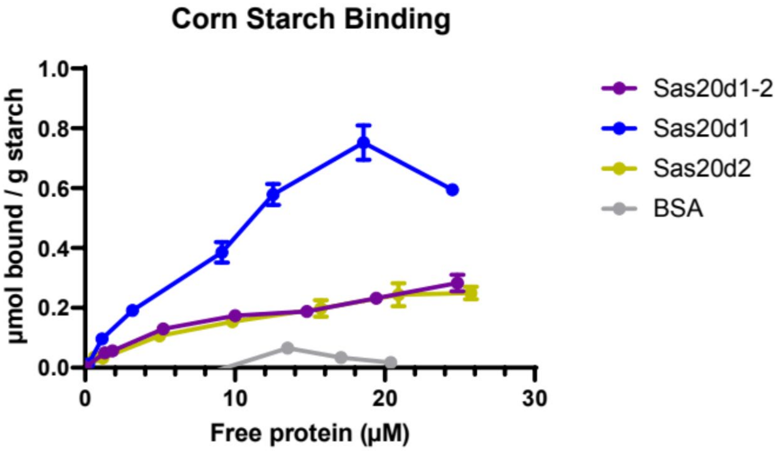
Like Sas20d1, Sas20d2 binds to maltoheptaose (K_d = 0.6 ± 0.02 μM) with greatly enhanced affinity over maltotriose (K_d = 912.4 ± 110 μM), suggesting that the domain utilizes the extensive binding platform between both X25 lobes. Sca5X25-2 shows a nearly identical trend, although the binding for each ligand is modestly better compared to Sas20d2. The number of binding sites (N) for these interactions is ~1 suggesting that there is only one extended ligand-binding site as observed in the Sca5X25-2 crystal structure. Although each module of Sca5X25-2 resembles a fully competent starch-binding site akin to those found within SusF (**Figure 3.6**), individual constructs of Sca5X25-2a and 2b (**Figure 3.1B**) failed to bind either maltotriose or maltoheptaose underscoring the need for the extended platform comprised of four tryptophan residues between both X25s for the high-affinity binding as observed with maltoheptaose.

Neither Sas20d2 nor Sca5X25-2 bound to α-cyclodextrin, but they did bind β-cyclodextrin, albeit with low affinity (~100-fold higher K_d compared to maltoheptaose), likely due to the increased ability of β-cyclodextrin to contort to a favorable binding geometry (**Table 3.3**). Cyclodextrins are often used as a proxy for the recognition of internal regions of a starch polymer, and many starch-binding CBMs recognize cyclodextrins and starch via a shallow cleft comprised of two aromatic residues that mimic the curvature of the α1,4-glucan bond [211, 212]. While the volume of the Sas20d2 binding site is large enough to accommodate α-cyclodextrin, the helical arrangement of the aromatic platform likely prevents productive binding of the ligand. We quantified our affinity PAGE results (**Figure 3.1, 3.2**) by ITC (**Table 3.3**) and determined that Sas20d2 binds to both corn (K_d= 7.9 ± 1.4 μM) and potato amylopectin (K_d= 5.7 ± 1.5 μM) with similar affinity. Sas20d2 binds only modestly better to these polysaccharides compared to Sas20d1.

As with Sas20d1, we mutated the four Trp residues (W329A, W375A, W440A, W481A) in Sas20d2 that corresponded to the aromatic platform observed within the Sca5X25-2 structure (**Figures 3.6E, 3.5**). A consistent trend for each mutation was the loss of binding for maltotriose. This was true for both W440A and W375A, equivalent to W620 and W555 of Sca5X25-2, positioned at the edges of the binding pocket, which we thought might be unnecessary for the smaller ligand. In fact, W555 of Sca5X25-2 (W375 of Sas20d2) did not participate in binding in our crystal structure. W481 of Sas20d2 (W661 of Sca5X25-2) is positioned towards the interior of the binding cavity and mutation eliminated binding to both maltotriose and maltoheptaose, while the W329A, W375A and W440A mutants retained binding to maltoheptaose but displayed ~100-fold increase in the K_d compared to WT Sas20d2. Notably, despite the symmetry within the binding pocket, mutations within each lobe had unique phenotypes. Particularly, W481 of the second X25 module seems to be most essential for anchoring maltooligosaccharides. Together, these data underscore that this domain is tuned to recognize longer helical regions of α-glucan. However, it cannot be excluded that there can be binding to crystalline regions in starch granules.

Sas20 Domains Bind to Insoluble Corn Starch

The ITC results allowed us to make conclusions on the binding profile of soluble substrates, but since *R. bromii* degrades RS, we investigated insoluble starch binding of Sas20 to corn starch. Sas20d1, Sas20d2, and Sas20d1-2 had similar K_d values ranging from 10-20 μM (**Figure 3.14**). However, Sas20d1 had a B_{max} that is nearly triple that of Sas20d2 or Sas20d1-2. This suggests that Sas20d1 can access more binding sites on the corn starch granule. Interestingly, we did not observe synergy or enhanced binding of the protein when both domains were present. This could be because the Sas20d1-2 construct is bulkier, and since each binding site is tuned to recognize different aspects of the polysaccharide, the larger protein makes fewer productive interactions with the granule. Therefore, the sequential position of both domains appears to not display avidity with respect to binding to ligand.



Sas20d1	Kd = 11.9 ± 3.5 μM,	Bmax = 1.0 ± 0.14
Sas20d2	Kd = 13.6 ± 3.1 μM,	Bmax = 0.38 ± 0.04
Sas20d1-2	Kd = 11.8 ± 2.2 μM,	Bmax = 0.38 ± 0.03

Figure 3.14: **Isothermal depletion for corn starch.** Affinity by indicated protein constructs on insoluble corn starch. All data fit to a one-site specific binding isotherm model. R^2 of fit for Sas20d1, Sas20d2, and Sas20d1-2 is 94.0%, 96.1%, and 96.5%, respectively.

Sas20 Domains are Flexible and Extended in Solution

To better connect how our crystal structures correlate to the substrate preferences we observe in solution, we used size-exclusion chromatography coupled with small angle x-ray scattering (SEC-SAXS) on Sas20d1, Sas20d2, and Sas20d1-2 with and without 5mM maltoheptaose (**Table 3.6**). Since Sas20d2 could not be crystallized, we used Phyre2 to generate a Sas20d2 model (100% confidence) using the Sca5X25-2 crystal structure for fitting the solution data [213].

Table 3.6A: Sample Parameters for SAXS Data Collection

	Sas20d1	Sas20d2	Sas20d1-2
Organism	*Ruminococcus bromii* L2-63		
Source	*E. coli* Rosetta (DE3) pLyS		
UnitProt sequence ID (residues in construct)	R5DX05 (31-270)	R5DX05 (311-577)	R5DX05 (31-577)
Extinction coefficient [A_{280}, 0.1%(w/v)]	1.674	1.757	1.561
Partial specific volume from chemical composition (cm^3 g^{-1})	0.726	0.73	0.727
Particle contrast from sequence and solvent constituents, (ρprotein − ρsolvent; 10^{10} cm^{-2})	6.453 (9.457-3.004)	6.523 (9.457-2.34)	7.051 (9.525-2.915)
Mass from chemical composition (Da)	26,000	26,400	56,000
Superdex 200 10/300 Increase			
Loading concentration (mg/mL)	36	17	36
Injection volume (μL)	150	200	200
Flow rate (mL/min)	0.6	0.6	0.6
Solvent	1X PBS, 1mM TCEP, pH=7		

Table 3.6B: Instrumentation and Data Collection Protocols for SAXS

Instrument	BioCAT facility at the Advanced Photon Source beamline 18ID with Pilatus3 X IM (Dectris) detector
Wavelength (Å)	1.033
Beam size (μm^2)	150 (h) x 25 (v) focused at the detector
Camera length (m)	3.629
q-measurement range (Å-1)	.0042-.36
Absolute scaling method	Glassy Carbon, NIST SRM 3600
Basis for normalization to constant counts	To transmitted intensity by beam-stop counter
Method for monitoring radiation damage	Automated frame-by-frame comparison of relevant regions using CORMAP implemented in BioXTAS RAW
Exposure time, number of exposures	0.5 s exposure time with a 1 s total exposure period (0.5 s on, 0.5 s off) of entire SEC elution
Sample configuration	SEC-SAXS with sheath -flow cell, effective path length 0.542 mm. Size separation by an AKTA Pure with a Superdex 200 10/300 Increase column
Sample temperature (ºC)	23

Table 3.6C: Software employed for SAS data reduction, analysis, and interpretation

SAXS data reduction	Radial averaging; frame comparison, averaging, and subtraction done using BioXTAS RAW 2.1.0
Basic analysis: Guinier, M.W., P(r)	Guinier fit and M.W. using BioXTAS RAW, P(r) function using GNOM (Svergun, 1992). RAW uses MoW and Vc M.W. methods (Rambo & Tainer, 2013; Piiadov et al., 2018)

The SEC-SAXS experiments for Sas20d1 and Sas20d2 with and without maltoheptaose were monodisperse and the radius of gyration (R_g) across the eluted peak was relatively constant (**Table 3.7**, **Figure 3.15A-D**). The Guinier fit for the R_g and I(0) values confirmed these samples were monodisperse (**Figure 3.16A- D**). The molecular weights of Sas20d1 and Sas20d2 with and without maltoheptaose was calculated to be ~26 kDa, which corroborates the predicted monomeric molecular weight based on their sequences (**Table 3.7**). The D_{max} from the P(r) function for Sas20d1 without and with maltoheptaose are 103Å and 78Å, respectively, and for Sas20d2 without and with maltoheptaose are 78Å and 74Å, respectively, while the maximum dimension in the crystal structure or model for both proteins are approximately 66Å (**Table 3.7**, **Figures 3.17A, B, 3.18A-D**). Together, this suggests that Sas20d1 undergoes a contraction upon the addition of ligand, while only a marginal contraction occurs with Sas20d2. Additionally, the calculated D_{max} indicates Sas20d1 and Sca5X25-2 were crystallized in a relatively compact conformation in contrast to their average conformation in solution.

Table 3.7: Small Angle X-ray Data

Protein	I(0)	R_g (Å) SAXS	D_{max} (Å) crystal	D_{max} (Å) solution	Sequence MW (kDa)	SAXS MW (kDa)
Sas20d1	1.5 x 10^{-6} ± 6.0 x 10^{-10}	21.1 ± 0.02	64.3	103	25.9	25.6
Sas20d1 +maltoheptaose	0.05 ± 2.3 x 10^{-5}	20.4 ± 0.03	60.6	78		24.3
Sas20d2	8.3 x 10^{-7} ± 5.3 x 10^{-10}	23.1 ± 0.04		78	26.5	25.6
Sas20d2 +maltoheptaose	0.03 ± 2.6 x 10^{-5}	20.8 ± 0.04	67.5	74		25.9
Sas20d1-2	0.04 ± 7.9 x 10^{-5}	53.9 ± 0.26		203	57.2	46.6
Sas20d1-2 +maltoheptaose	0.04 ± 6.4 x 10^{-5}	51.8 ± 0.17		190		53.1

I(0) and R_g were determined from Guinier analysis. D_{max} in solution was determined IFT using GNOM. To calculate D_{max} *in crystallo*, we calculated the farthest distance between two amino acids in one peptide in the crystal structures for native Sas20d1, maltotriose-bound Sas20d1, and Phyre 2.0 generated model for Sas20d2. The Bayes method of molecular weight calculation from SAXS data is presented here.

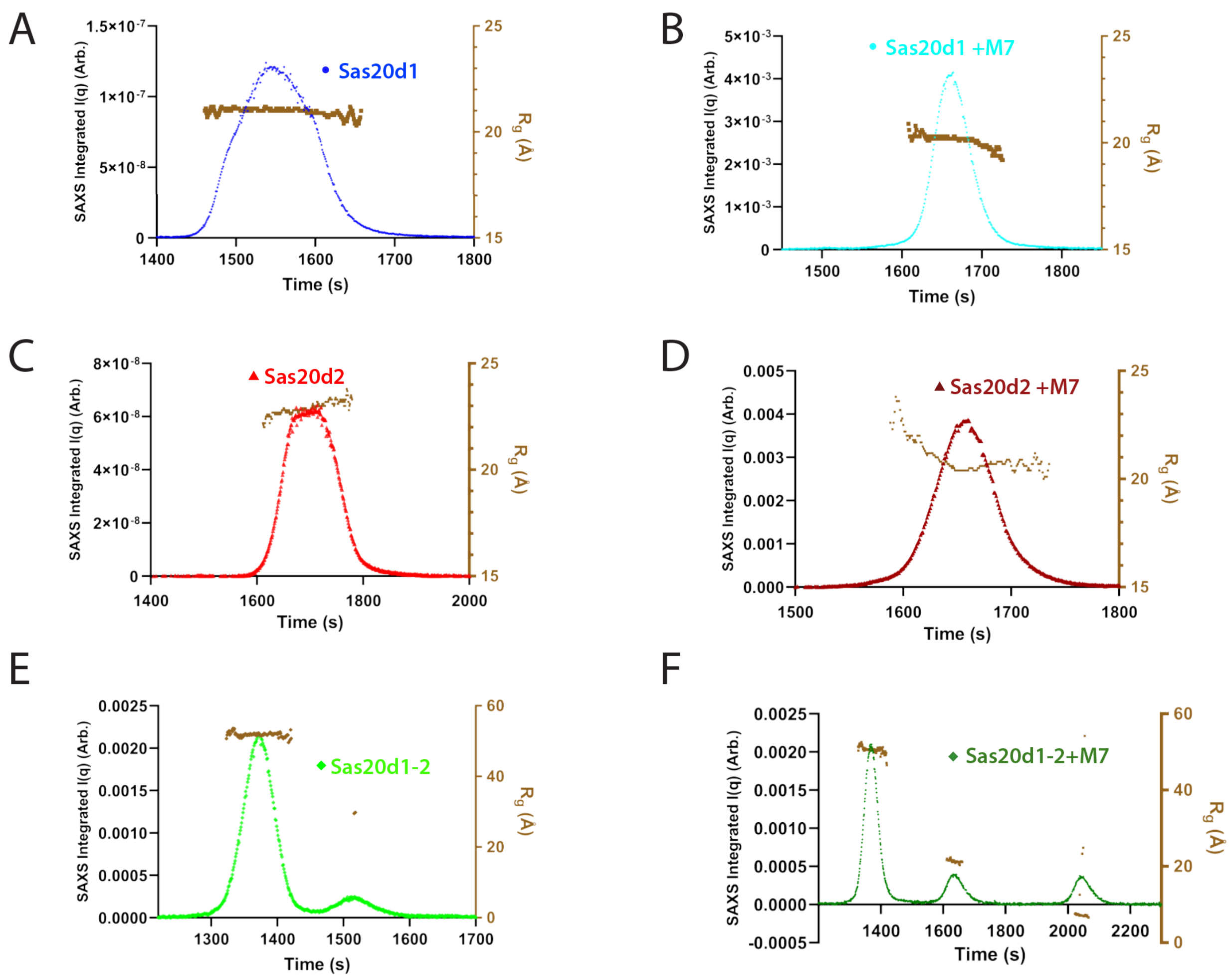


Figure 3.15: SEC-SAXS elution profiles. Total subtracted scattering intensity (left y axis) and R_g (right y axis) as a function of time for the SEC-SAXS elutions of A) Sas20d1, B) Sas20d1 with 5mM maltoheptaose (M7), C) Sas20d2 and D) Sas20d2 with 5mM M7, E) Sas20d1-2 and F) Sas20d1-2 with 5mM M7

The overall shape of the P(r) function for Sas20d1 and Sas20d2, calculated by indirect Fourier transform (IFT) using GNOM [214], has a relatively Gaussian shape that is characteristic of a globular compact particle (**Figure 3.17A, B**). Upon the addition of ligand, the P(r) function demonstrates that Sas20d1 undergoes a contraction in solution, but the overall shape of the P(r) function, and thus the protein itself, remains relatively constant. There is a truncation in the tail of the function which can be interpreted as a decrease in flexibility upon binding to ligand. However, the P(r) function for Sas20d2 without ligand shows a clear shoulder near r = 40Å which is characteristic of a protein with two structural motifs. This right shoulder is not found in the presence of ligand, which suggests that the two lobes seen in Sas20d2 associate more tightly upon binding to ligand while retaining the overall size of the protein.

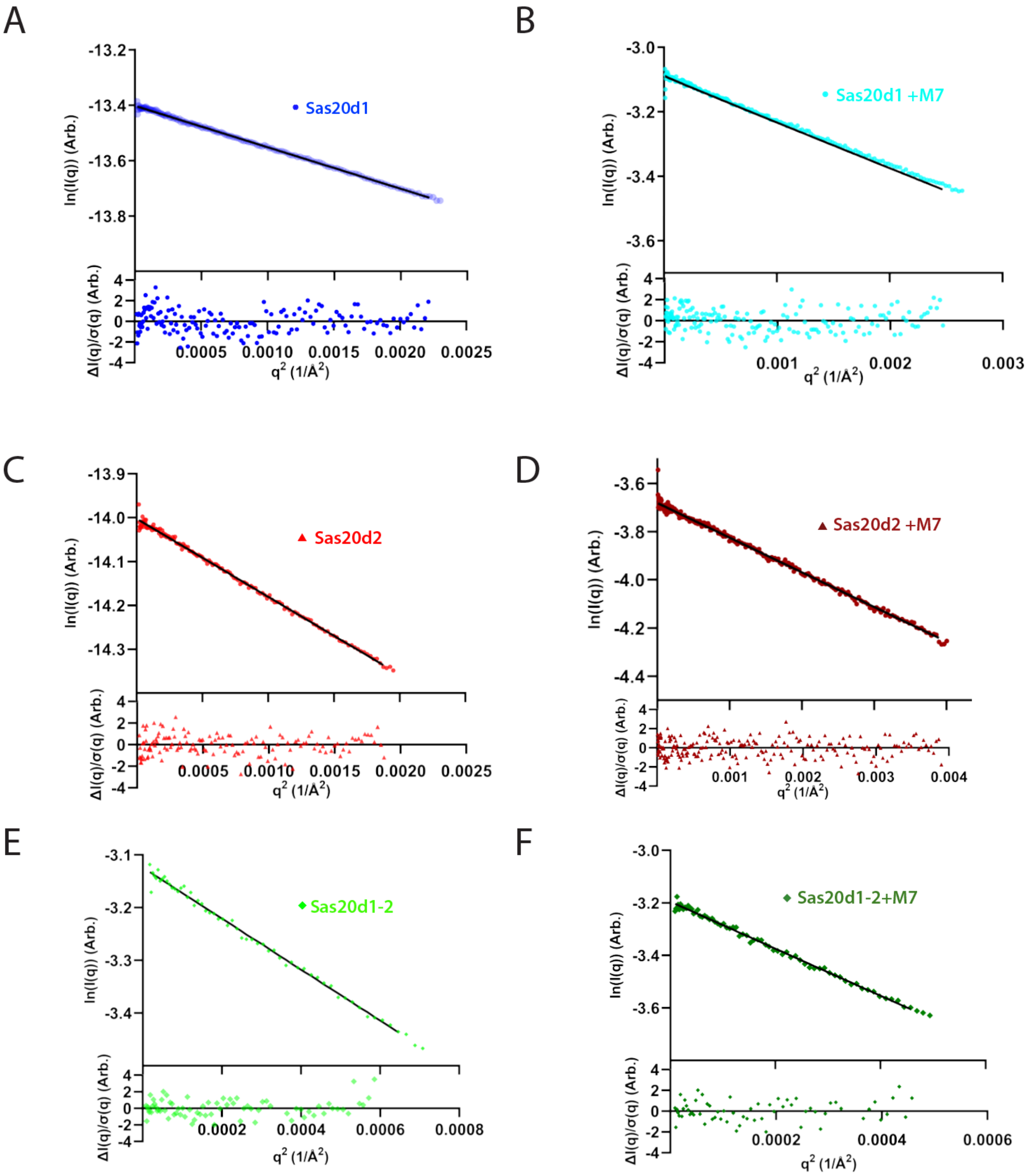


Figure 3.16: Guinier fit. Guinier fit for A) Sas20d1, B) Sas20d1 with 5mM maltoheptaose (M7), C) Sas20d2 and D) Sas20d2 with 5mM M7, E) Sas20d1-2 and F) Sas20d1-2 with 5mM M7 with normalized residual shown in the bottom panel.

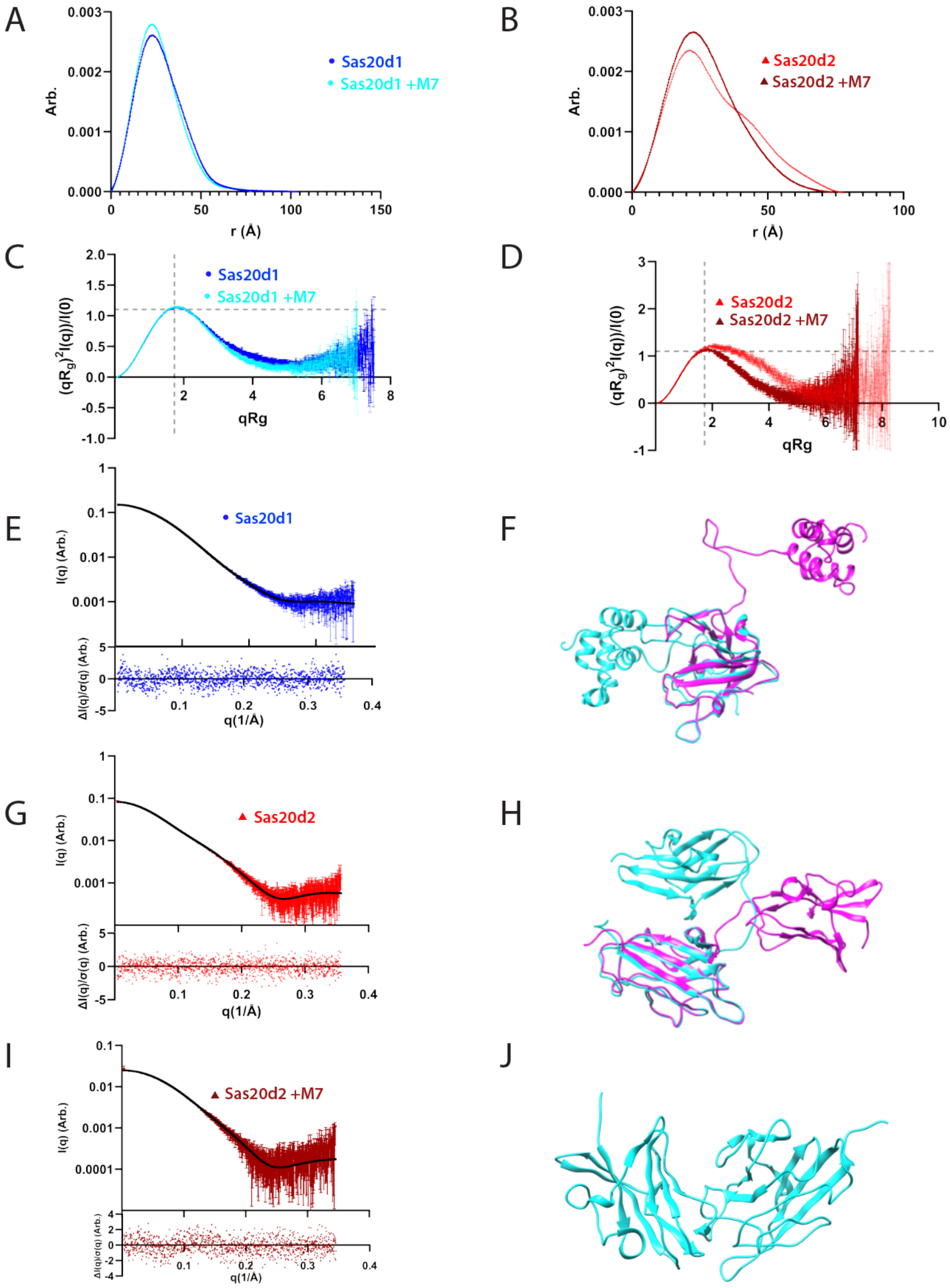


Figure 3.17: **Experimental SAXS and MultiFoXS results for Sas20d1 and Sas20d2.** Sas20d1 is in blue circles, Sas20d2 in red triangles. P(r) versus r for A) Sas20d1 and B) Sas20d2 with and without maltoheptaose normalized by I(0). Dimensionless Kratky plot for C) Sas20d1 and D) Sas20d2 with and without maltoheptaose; y=3/*e* and x=$\sqrt{3}$ as dashed gray lines to indicate where a globular protein would peak. E) SAXS scattering profile (points) and MultiFoXS fit (black line) for Sas20d1 (χ^2=1.19) The bottom panel shows the normalized fit residual. F) MultiFoXS 2-state model results for Sas20d1 with compact (cyan, R_g=19Å, weight=86%) and extended (magenta, R_g=25Å, weight=14%) conformations. Models aligned to residues 32-163 and were slightly offset for clarity. SAXS scattering profile (points) and MultiFoXS fit (black line) for G) Sas20d2 (χ^2=0.97) and I) Sas20d2 with 5mM maltoheptaose (χ^2=1.01). The bottom panel shows the normalized fit residual. H) MultiFoXS 2-state model results for Sas20d2 with compact (cyan, R_g=20Å, weight=36%) and extended (magenta, R_g=24Å, weight=64%) conformation. J) MultiFoXS 1-state model for Sas20d2 with maltoheptaose (R_g=19.5Å)

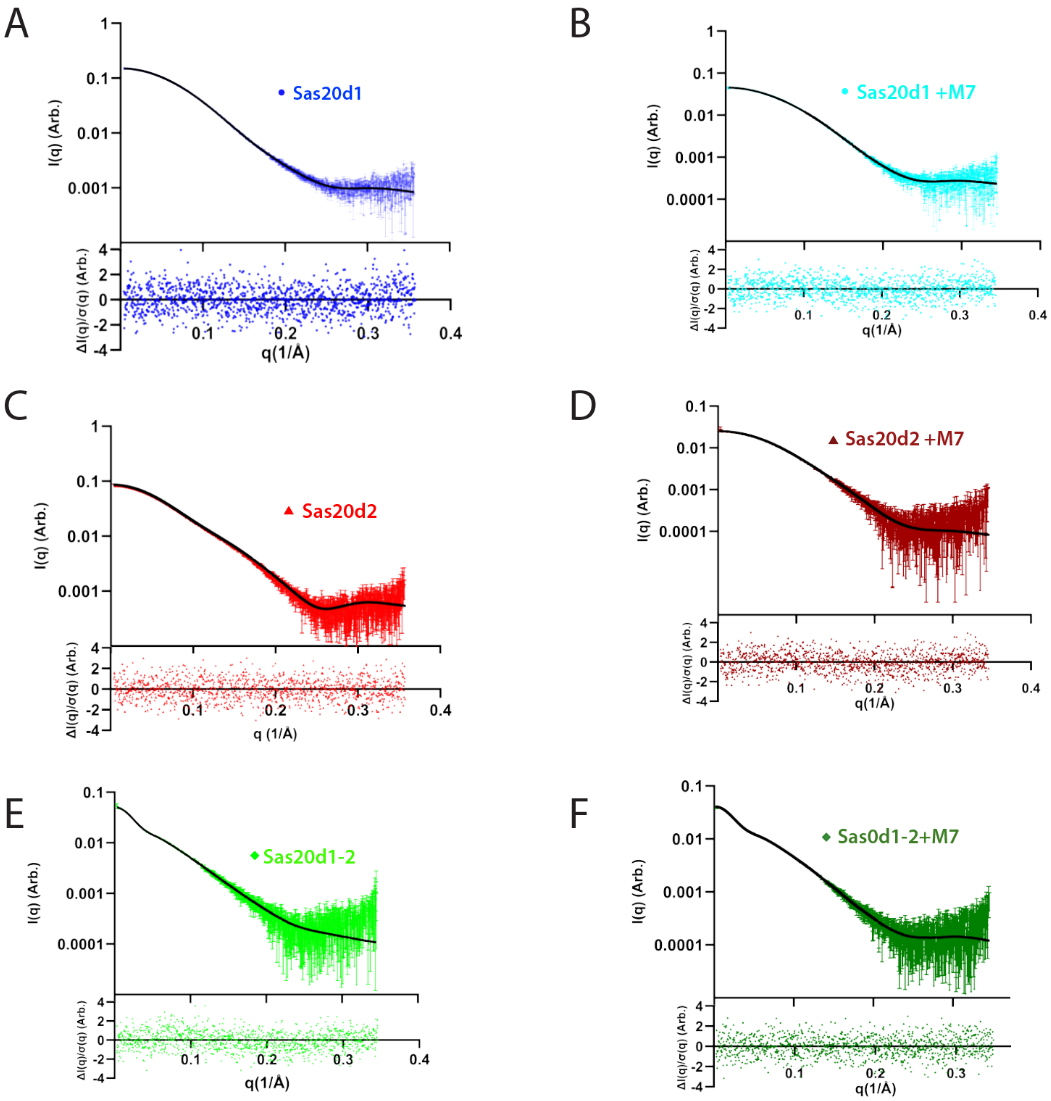


Figure 3.18 P(r) fit. P(r) fit A) Sas20d1, B) Sas20d1 with 5mM maltoheptaose (M7), C) Sas20d2 and D) Sas20d2 with 5mM M7, E) Sas20d1-2 and F) Sas20d1-2 with 5mM M7 with normalized residual shown in the bottom panel.

The dimensionless Kratky plot maxima for Sas20d1 and Sas20d2 are where typical rigid globular proteins would peak (**Figure 3.17C, D**). Upon addition of maltoheptaose, Sas20d1 shows a small but significant decrease in the mid to high q region, around qRg = 4, which indicates the ligand made this protein more compact and globular in solution. In the Sas20d2 analysis, the small plateau in the mid-to-high q region, around $qR_g = 4$ in the dimensionless Kratky plot, indicates some extension or flexibility in the system, likely associated with the two structural motifs visible via the P(r) plot. This plateau vanishes in the presence of maltoheptaose, and the resulting dimensionless Kratky plot shows that the protein with ligand is a more compact globular shape. Thus, the SAXS shows that ligand binding results in a more compact, globular shape of Sas20d2.

To fit our high-resolution structures to the SAXS data, we used MultiFoXS to generate a set of possible conformations in solution and selected the ensemble with the best fit [215]. For Sas20d1, we assigned the linker between the CBM26-like structure and bundle of helices (residues 164-191) as flexible. Since the differences in the basic SAXS analysis were subtle, MultiFoXS modelling was only done for Sas20d1 without ligand. MultiFoXS found the best fit solution was with two states, one compact and one extended with a $\chi^2 = 1.19$ (**Figure 3.17E, F**). Sas20d1 only exists in the extended conformation ~14% of the time in solution, which agrees with the compactness and minimal flexibility indicated by the P(r) distribution and dimensionless Kratky plot.

Since the differences in the basic SAXS analysis indicated there was a significant change in shape upon addition of ligand to Sas20d2, MultiFoXS modelling was done for both Sas20d2 with and without ligand. We assigned the linker between the two X25-like lobes (residues 415-423) as flexible. For Sas20 without ligand, MultiFoXS found the best-fit solution was also with two states, one compact and one extended with a $\chi^2= 1.01$ (**Figure 3.17G**). In contrast to Sas20d1, Sas20d2 without ligand exists in the extended state ~64% of the time in solution **(Figure 3.17H**). When ligand is present, MultiFoXS found the best-fit solution was a 1-state model that resembles the compact conformation (**Figure 3.17I, J**). Both ensembles corroborate the shapes indicated by the P(r) function and Kratky plots. However, because there is flexibility in the system, the displayed states in **Figures 3.17F, H,** and **J** are representative of these extended and compact conformations but should not be taken as prescriptive; that is, there are likely many similar states with the same overall size and extension but slightly different relative positions of the two folded motifs.

We then performed SEC-SAXS on Sas20d1-2 with and without 5mM maltoheptaose to discern how the two domains are oriented in solution and if this protein possesses notable flexibility. The elution profiles revealed that the SEC column separated a minor contaminant (peak 1520 s) in the Sas20d1-2 run and two minor contaminants (peaks 1650 and 2050 s) from the Sas20d1-2 with maltoheptaose run from our protein of interest (peak 1370 s) (**Figure 3.15E, F**). The R_g across the eluted peaks was relatively constant. The Guinier fit for the R_g and I(0) values confirmed that Sas20d1-2 with and without maltoheptaose were monodisperse (**Figure 3.16E, F**). The calculated molecular weight from the scattering profile, 53.7 kDa, agreed with the predicted monomeric molecular weight by sequence (**Table 3.7).** The right shoulder in the P(r) plot is characteristic of a second domain with significant (~100Å) separation from the first and is consistent with some flexibility given the long tail down to the maximum dimension of ~200Å. (**Figure 3.19A, 3.18E, F**). The shape of the dimensionless Kratky plot for Sas20d1-2 shows significant deviation from where we expect globular proteins to peak (**Figure 3.19B**). In particular, the peak near qRg of 5 is above 2 which indicates a highly extended molecule, and the plateau at higher qRg also indicates some flexibility in the system. As with Sas20d1, addition of maltoheptaose to Sas20d1-2 had a subtle effect on the overall shape of the protein but induced a more globular shape and decrease in flexibility.

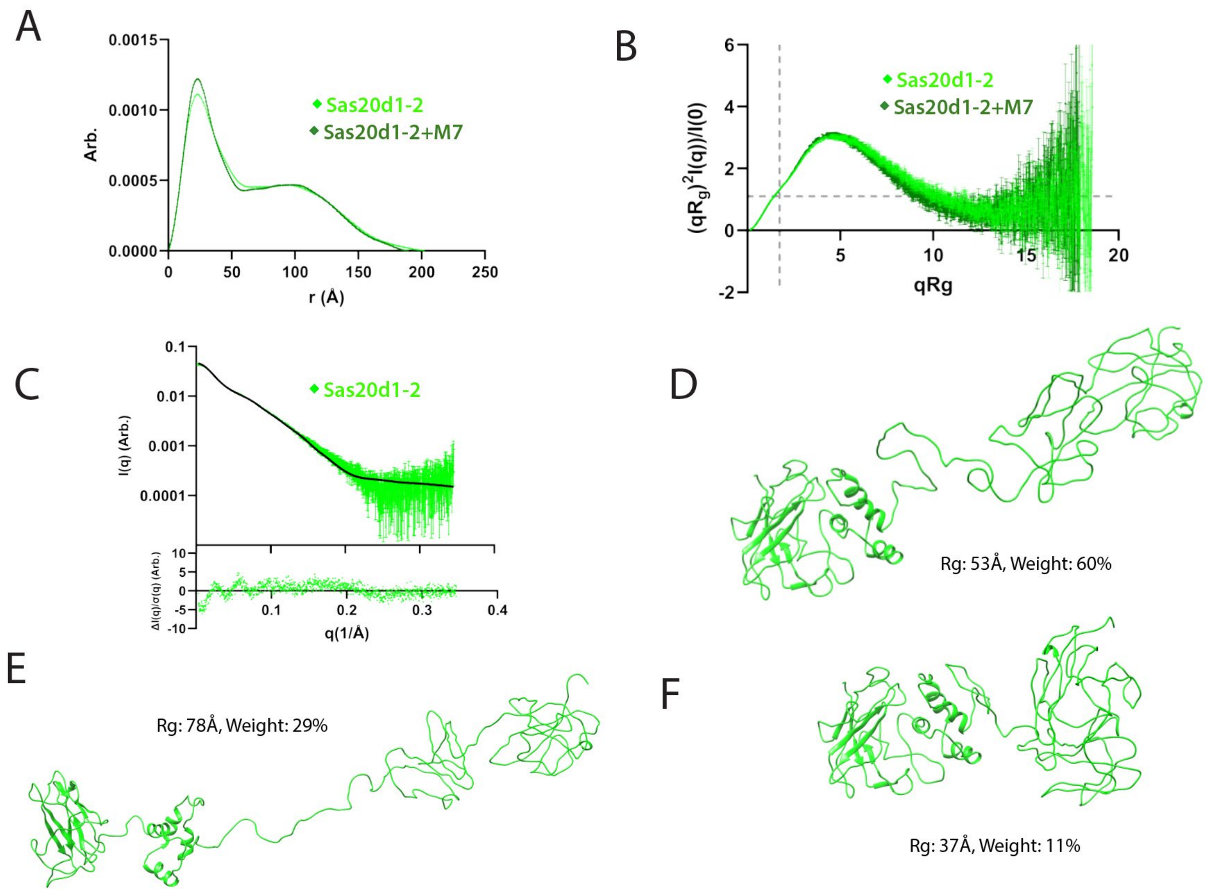


Figure 3.19: **Experimental SAXS and MultiFoXS results for Sas20d1-2**. Sas20d1-2 in green diamonds. A) P(r) versus r for Sas20d1-2 and Sas20d1-2 with maltoheptaose normalized by I(0). B) Dimensionless Kratky plot with y=3/*e* and x=$\sqrt{3}$ as dashed gray lines to indicate where a globular protein would peak. C) The SAXS scattering profile (green points) and MultiFoXS fit (black line) for Sas20d1-2 (χ^2= 2.65). The bottom panel shows the normalized fit residual. F-H) MultiFoXS 3-state results for Sas20d1-2 with their associated R_g and weight.

We then used MultiFoXS with our high-resolution structure of the Sas20d1 domain and model of Sas20d2 in isolation to investigate how the domains are positioned relative to each other. The best model fit was a 3-state ensemble with an acceptable χ^2 = 2.65, but the residual from this fit to the SAXS scattering profile is not randomly distributed, particularly in the low q range (**Figure 3.19C)**. Here we see that Sas20d1-2 shows a range of conformations from very compact to very extended, where this protein exists in the most compact state only ~11% of the time (**Figure 3.19D-F**). This agrees with the observations from the P(r) function and dimensionless Kratky plot, which showed highly extended, flexible systems with well separated domains. Also, no single solution, compact or extended, fits the data well, as the best single model fit has a $\chi^2 = 8.2$, further indicating a flexible system that exists in a continuum of states in solution. In conclusion, while the precise number and extent of conformations adopted by Sas20d1-2 in solution is unclear, both the MultiFoXS and basic SAXS analysis indicate that Sas20d1-2 is highly flexible and extended in solution.

Sas20 Domain Homology

Sas20 has two distinct domains that recognize different aspects of the starch substructure. To determine if the Sas20 domains occur in other bacteria, we performed a BLAST analysis of each Sas20 domain [216]. Using an E-value < 0.01, we found 101 sequences for the first domain, and the vast majority of these are found within *Ruminococcus* species, suggesting an extremely narrow phylogenetic distribution (**Figure 3.20**). Among these sequences, many possess homology to domain 1 and domain 2 of Sas20. Interestingly, we discovered that *R. bromii* has a second Sas20d1-like protein. The protein encoded within locus tag RBR_02940 (L2-63_00923) of *Ruminococcus bromii* L2-63 is a predicted cell wall-anchored protein and shares 31% sequence identity with Sas20 domain 1 along the length of the β-sandwich and including part of the α-helical bundle. Using JPred4 for secondary structure prediction, RBR_09240 is expected to possess four helices that are C-terminal to the β-sandwich, and followed by a Gly-Ser-Asn rich linker and sortase motif (**Figure 3.21**) [217]. Most of the maltotriose-binding platform observed in the Sas20d1 structure is conserved in RBR_09240, except for Y60 (substituted conservatively as tryptophan) and T152 (substituted for proline). Therefore, we predict that RBR_09240 is a starch-binding cell surface-anchored protein but is unlikely to be incorporated into an amylosome complex due to its apparent lack of a dockerin or cohesin domain. Interestingly, the genomic context for this protein does not further imply function, as the gene is sandwiched between a predicted alanine-tRNA ligase and probable endonuclease.

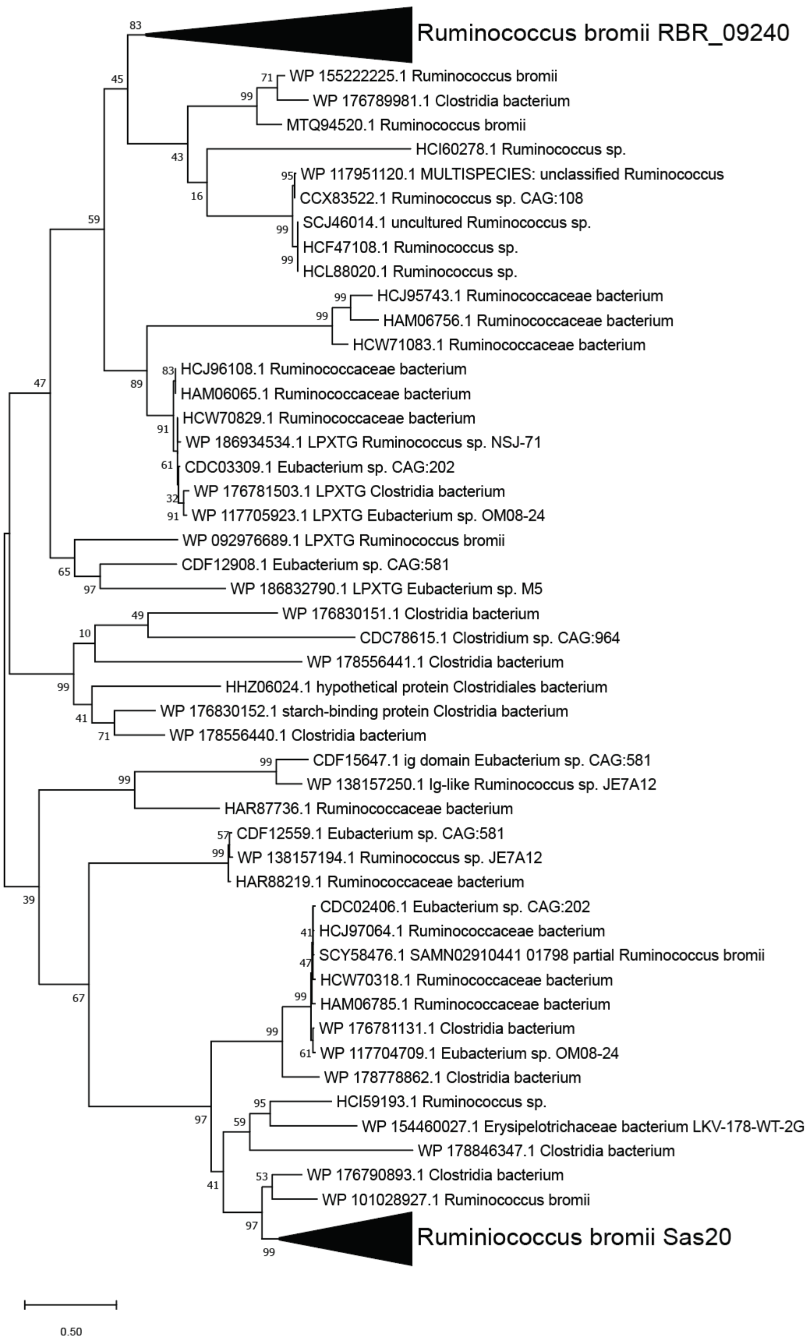
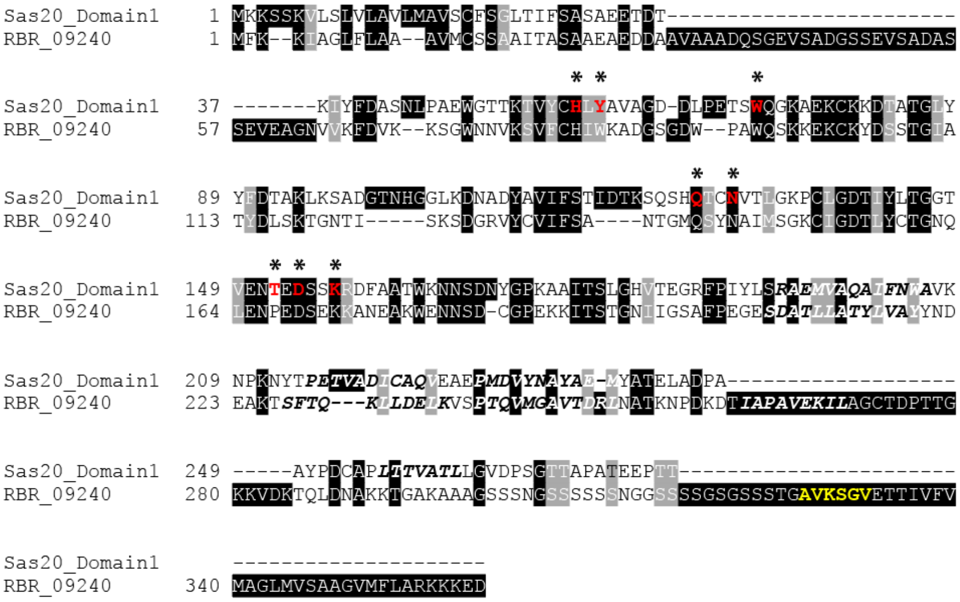

83
Ruminococcus bromii RBR_09240
45
71
WP 155222225.1 Ruminococcus bromii
99
WP 176789981.1 Clostridia bacterium
MTQ94520.1 Ruminococcus bromii
43
HCI60278.1 Ruminococcus sp.
95
WP 117951120.1 MULTISPECIES: unclassified Ruminococcus
16
CCX83522.1 Ruminococcus sp. CAG:108
59
SCJ46014.1 uncultured Ruminococcus sp.
99
HCF47108.1 Ruminococcus sp.
99
HCL88020.1 Ruminococcus sp.
99
HCJ95743.1 Ruminococcaceae bacterium
99
HAM06756.1 Ruminococcaceae bacterium
HCW71083.1 Ruminococcaceae bacterium
83
HCJ96108.1 Ruminococcaceae bacterium
89
HAM06065.1 Ruminococcaceae bacterium
47
HCW70829.1 Ruminococcaceae bacterium
91
WP 186934534.1 LPXTG Ruminococcus sp. NSJ-71
61
CDC03309.1 Eubacterium sp. CAG:202
32
WP 176781503.1 LPXTG Clostridia bacterium
91
WP 117705923.1 LPXTG Eubacterium sp. OM08-24
WP 092976689.1 LPXTG Ruminococcus bromii
65
CDF12908.1 Eubacterium sp. CAG:581
97
WP 186832790.1 LPXTG Eubacterium sp. M5
49
WP 176830151.1 Clostridia bacterium
10
CDC78615.1 Clostridium sp. CAG:964
WP 178556441.1 Clostridia bacterium
99
HHZ06024.1 hypothetical protein Clostridiales bacterium
41
WP 176830152.1 starch-binding protein Clostridia bacterium
71
WP 178556440.1 Clostridia bacterium
99
CDF15647.1 ig domain Eubacterium sp. CAG:581
99
WP 138157250.1 Ig-like Ruminococcus sp. JE7A12
HAR87736.1 Ruminococcaceae bacterium
57
CDF12559.1 Eubacterium sp. CAG:581
99
WP 138157194.1 Ruminococcus sp. JE7A12
HAR88219.1 Ruminococcaceae bacterium
39
CDC02406.1 Eubacterium sp. CAG:202
41
HCJ97064.1 Ruminococcaceae bacterium
47
SCY58476.1 SAMN02910441 01798 partial Ruminococcus bromii
HCW70318.1 Ruminococcaceae bacterium
67
99
HAM06785.1 Ruminococcaceae bacterium
99
WP 176781131.1 Clostridia bacterium
61
WP 117704709.1 Eubacterium sp. OM08-24
WP 178778862.1 Clostridia bacterium
95
HCI59193.1 Ruminococcus sp.
97
59
WP 154460027.1 Erysipelotrichaceae bacterium LKV-178-WT-2G
WP 178846347.1 Clostridia bacterium
53
WP 176790893.1 Clostridia bacterium
41
WP 101028927.1 Ruminococcus bromii
97
Ruminiococcus bromii Sas20
99
0.50

Figure 3.20 Evolutionary analysis of Sas20d1 by maximum likelihood analysis. The evolutionary history was inferred by using the Maximum Likelihood method and Whelan and Goldman + Freq. model. The tree with the highest log likelihood (-32347.71) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 100 amino acid sequences, with E value > 0.001 for Sas20d1 only. Evolutionary analyses were conducted in MEGAX.



```
Sas20_Domain1     1 MKKSSKVLSLVLAVLMAVSCFSGLTIFSASAEETDT------------------------
RBR_09240         1 MFK--KIAGLFLAA--AVMCSSAAITASAAEAEDDAAVAAADQSGEVSADGSSEVSADAS

                                                   * *            *
Sas20_Domain1    37 -------KIYFDASNLPAEWGTTKTVYCHLYAVAGD-DLPETSWQGKAEKCKKDTATGLY
RBR_09240        57 SEVEAGNVVKFDVK--KSGWNNVKSVFCHIWKADGSGDW--PAWQSKKEKCKYDSSTGIA

                                                          *  *
Sas20_Domain1    89 YFDTAKLKSADGTNHGGLKDNADYAVIFSTIDTKSQSHQTCNVTLGKPCLGDTIYLTGGT
RBR_09240       113 TYDLSKTGNTI-----SKSDGRVYCVIFSA----NTGMQSYNAIMSGKCIGDTLYCTGNQ

                       * *  *
Sas20_Domain1   149 VENTEDSSKRDFAATWKNNSDNYGPKAAITSLGHVTEGRFPIYLSRAEMVAQAIFNWAVK
RBR_09240       164 LENPEDSEKKANEAKWENNSD-CGPEKKITSTGNIIGSAFPEGESDATLLATYLVAYYND

Sas20_Domain1   209 NPKNYTPETVADICAQVEAEPMDVYNAYAE-MYATELADPA-------------------
RBR_09240       223 EAKTSFTQ---KLLDELKVSPTQVMGAVTDRLNATKNPDKDTIAPAVEKILAGCTDPTTG

Sas20_Domain1   249 -----AYPDCAPLTTVATLLGVDPSGTTAPATEEPTT-----------------------
RBR_09240       280 KKVDKTQLDNAKKTGAKAAAGSSSNGSSSSSSNGGSSSSGSGSSSTGAVKSGVETTIVFV

Sas20_Domain1       ---------------------
RBR_09240       340 MAGLMVSAAGVMFLARKKKED
```

Figure 3.21: Sequence alignment of Sas20d1 and locus tag RBR_09240 of *R. bromii* L2-63. Alignment was performed in T-Coffee and rendered in Boxshade. Residues that are involved in maltotriose binding in Sas20d1 are displayed in red (H58, Y60, W72, T152, Q127, N130, D154, K157). Residues within α-helical segments are in bold italics and the putative sortase motif is displayed in yellow.

Like Sas20d1, Sas20d2 is fairly restricted in its phylogenetic distribution. We found 328 sequences with homology to Sas20d2 via BLAST (E-value < 0.0001), of which 206 were from *Ruminococcus*, 24 from the CFB bacteria (Cytophaga-Fusobacterium-Bacteroidetes), and the remainder within the Firmicutes, many in the Oscillspiracaea, which includes *Ruminococcus*. Of the 328 sequences, only 19 were identified by the DBCan server as sharing homology with a known CBM or glycoside hydrolase family; 12 of these proteins appear to possess multiple starch-targeting CBMs and/or a GH13 in addition to a domain with homology to Sas20d2 (**Figure 3.22**) [218]. Most of these sequences retain the residues found in Sca5X25-2 that are involved in capturing maltooligosaccharide (**Figure 3.23**). Beyond Sca5 and Sas20, the scaffoldin protein Sca3 of *R. bromii* L2-63 is predicted to consist of four X25-like modules [117]. However, a sequence alignment of the Sca3 domains with the X25s within Sca5 and

Sas20 suggests that only one tryptophan is conserved (**Figure 3.24**). Sca3 may bind starch, but the sequence diverges from what is seen in Sca5 and Sas20.

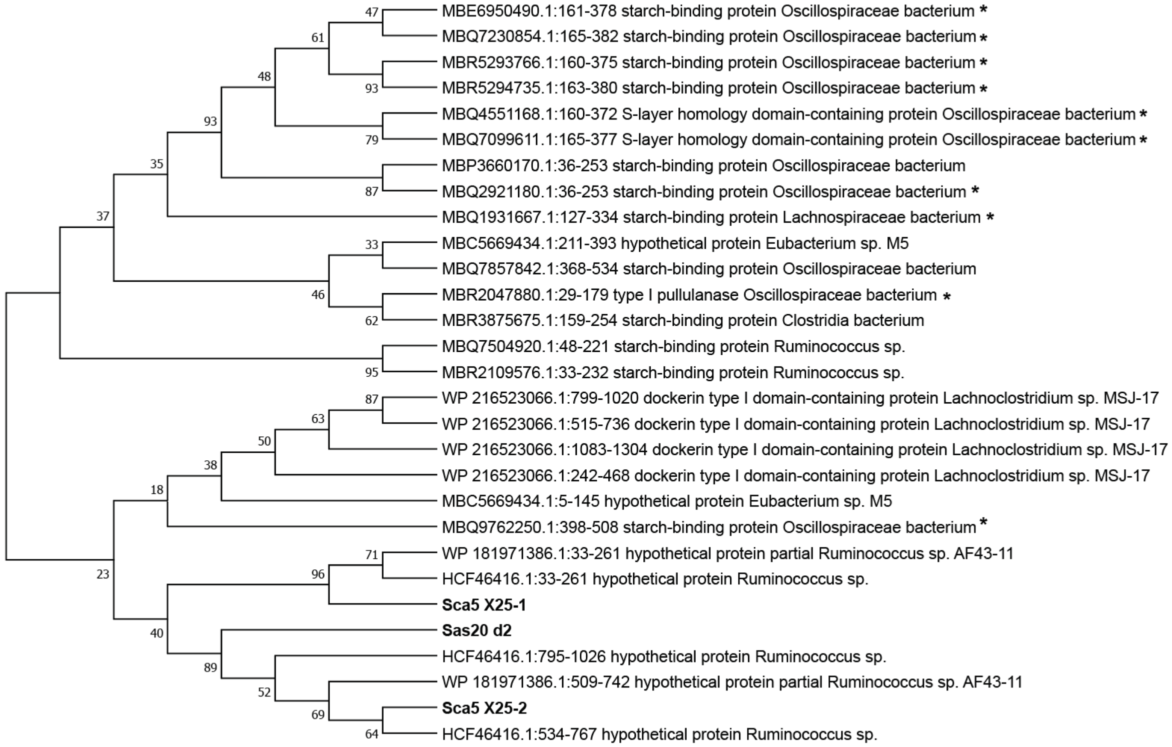

47
61
MBE6950490.1:161-378 starch-binding protein Oscillospiraceae bacterium *
MBQ7230854.1:165-382 starch-binding protein Oscillospiraceae bacterium *
48
MBR5293766.1:160-375 starch-binding protein Oscillospiraceae bacterium *
93
MBR5294735.1:163-380 starch-binding protein Oscillospiraceae bacterium *
93
MBQ4551168.1:160-372 S-layer homology domain-containing protein Oscillospiraceae bacterium *
79
MBQ7099611.1:165-377 S-layer homology domain-containing protein Oscillospiraceae bacterium *
35
MBP3660170.1:36-253 starch-binding protein Oscillospiraceae bacterium
87
MBQ2921180.1:36-253 starch-binding protein Oscillospiraceae bacterium *
37
MBQ1931667.1:127-334 starch-binding protein Lachnospiraceae bacterium *
33
MBC5669434.1:211-393 hypothetical protein Eubacterium sp. M5
MBQ7857842.1:368-534 starch-binding protein Oscillospiraceae bacterium
46
MBR2047880.1:29-179 type I pullulanase Oscillospiraceae bacterium *
62
MBR3875675.1:159-254 starch-binding protein Clostridia bacterium
MBQ7504920.1:48-221 starch-binding protein Ruminococcus sp.
95
MBR2109576.1:33-232 starch-binding protein Ruminococcus sp.
87
WP 216523066.1:799-1020 dockerin type I domain-containing protein Lachnoclostridium sp. MSJ-17
63
WP 216523066.1:515-736 dockerin type I domain-containing protein Lachnoclostridium sp. MSJ-17
50
WP 216523066.1:1083-1304 dockerin type I domain-containing protein Lachnoclostridium sp. MSJ-17
38
WP 216523066.1:242-468 dockerin type I domain-containing protein Lachnoclostridium sp. MSJ-17
18
MBC5669434.1:5-145 hypothetical protein Eubacterium sp. M5
MBQ9762250.1:398-508 starch-binding protein Oscillospiraceae bacterium *
71
WP 181971386.1:33-261 hypothetical protein partial Ruminococcus sp. AF43-11
23
96
HCF46416.1:33-261 hypothetical protein Ruminococcus sp.
Sca5 X25-1
40
Sas20 d2
89
HCF46416.1:795-1026 hypothetical protein Ruminococcus sp.
52
WP 181971386.1:509-742 hypothetical protein partial Ruminococcus sp. AF43-11
69
Sca5 X25-2
64
HCF46416.1:534-767 hypothetical protein Ruminococcus sp.

Figure 3.22: Evolutionary analysis of Sas20d2 by Maximum Likelihood analysis. The evolutionary history was inferred by using the Maximum Likelihood method and Whelan and Goldman + Freq. model. The bootstrap consensus tree inferred from 100 replicates is taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+*G*, parameter = 3.5644)). This analysis involved 29 amino acid sequences. All positions with less than 95% site coverage were eliminated, i.e., fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position (partial deletion option). There was a total of 61 positions in the final dataset. Evolutionary analyses were conducted in MEGAX.

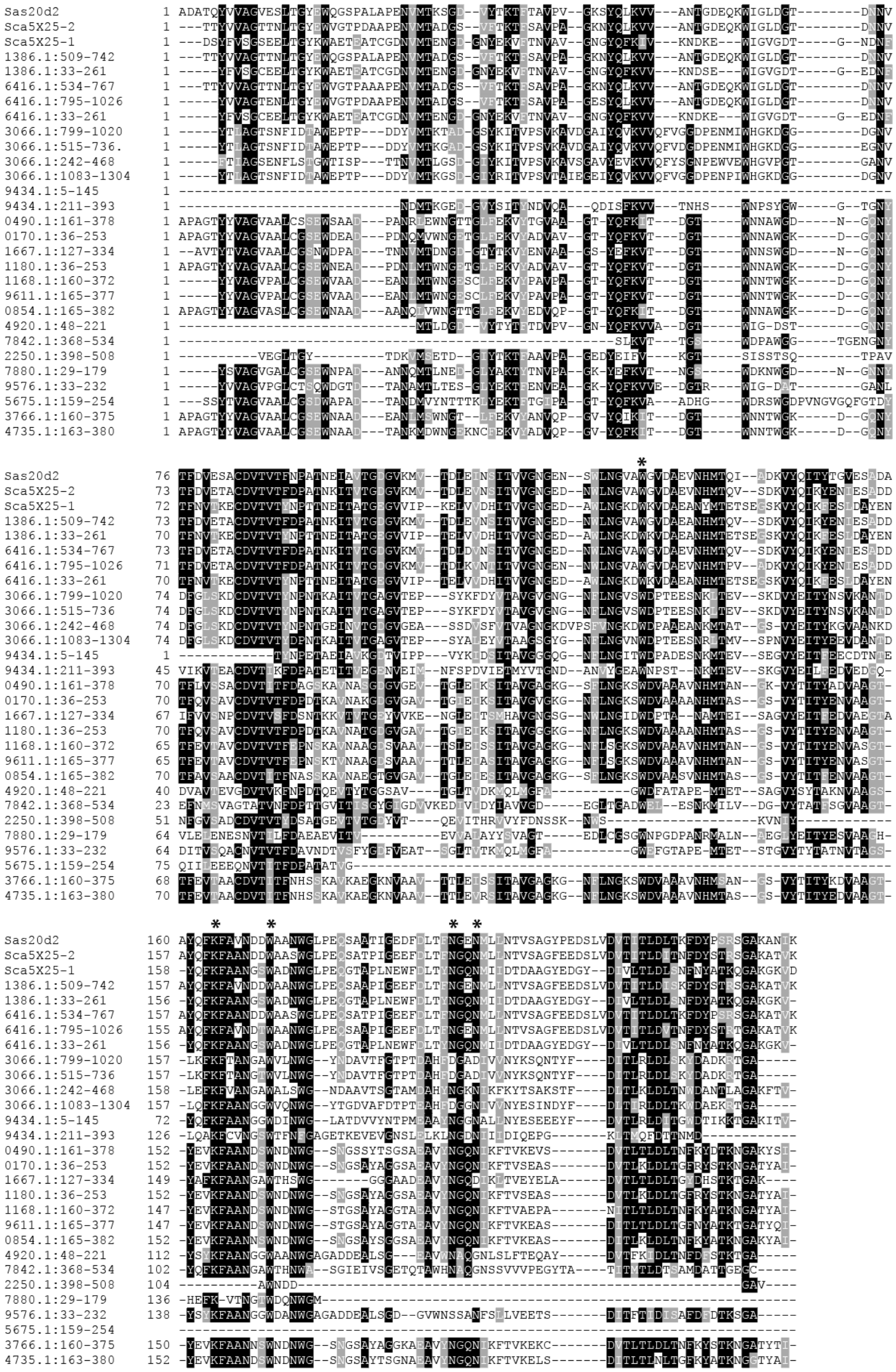
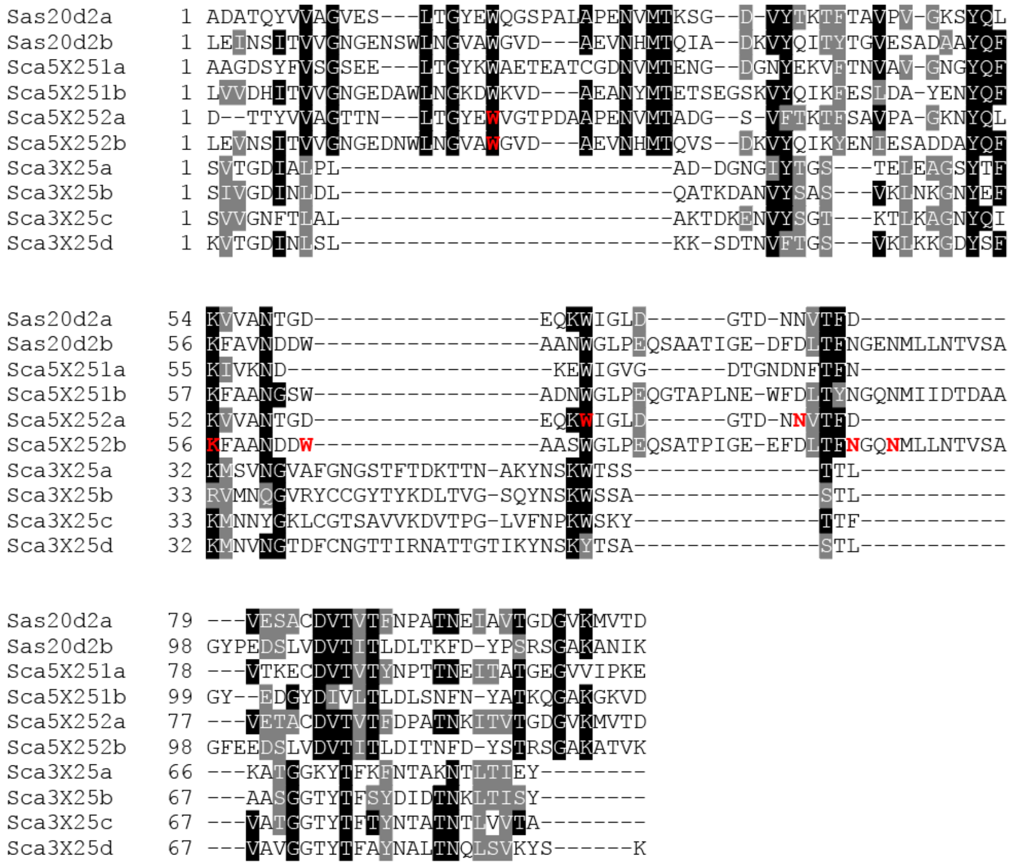


Figure 3.23: Sequence Alignment of Sas20d2, Sca5X25 domains and homologs. Sequences were aligned using the MUSCLE default parameters in MEGAX and rendered in Boxshade. The asterisks indicate residues involved in maltooligosaccharide binding in Sas20 and Sca5X25-2.



```
Sas20d2a      1 ADATQYVVAGVES---LTGYEWQGSPALAPENVMTKSG--D-VYTKTFTAVPV-GKSYQL
Sas20d2b      1 LEINSITVVGNGENSWLNGVAWGVD---AEVNHMTQIA--DKVYQITYTGVESADAAYQF
Sca5X251a     1 AAGDSYFVSGSEE---LTGYKWAETEATCGDNVMTENG--DGNYEKVFTNVAV-GNGYQF
Sca5X251b     1 LVVDHITVVGNGEDAWLNGKDWKVD---AEANYMTETSEGSKVYQIKFESLDA-YENYQF
Sca5X252a     1 D--TTYVVAGTTN---LTGYEWVGTPDAAPENVMTADG--S-VFTKTFSAVPA-GKNYQL
Sca5X252b     1 LEVNSITVVGNGEDNWLNGVAWGVD---AEVNHMTQVS--DKVYQIKYENIESADDAYQF
Sca3X25a      1 SVTGDIALPL-------------------------AD-DGNGIYTGS---TELEAGSYTF
Sca3X25b      1 SIVGDINLDL-------------------------QATKDANVYSAS---VKLNKGNYEF
Sca3X25c      1 SVVGNFTLAL-------------------------AKTDKENVYSGT---KTLKAGNYQI
Sca3X25d      1 KVTGDINLSL-------------------------KK-SDTNVFTGS---VKLKKGDYSF

Sas20d2a     54 KVVANTGD-----------------EQKWIGLD------GTD-NNVTFD-----------
Sas20d2b     56 KFAVNDDW-----------------AANWGLPEQSAATIGE-DFDLTFNGENMLLNTVSA
Sca5X251a    55 KIVKND-------------------KEWIGVG------DTGNDNFTFN-----------
Sca5X251b    57 KFAANGSW-----------------ADNWGLPEQGTAPLNE-WFDLTYNGQNMIIDTDAA
Sca5X252a    52 KVVANTGD-----------------EQKWIGLD------GTD-NNVTFD-----------
Sca5X252b    56 KFAANDDW-----------------AASWGLPEQSATPIGE-EFDLTFNGQNMLLNTVSA
Sca3X25a     32 KMSVNGVAFGNGSTFTDKTTN-AKYNSKWTSS--------------TTL-----------
Sca3X25b     33 RVMNQGVRYCCGYTYKDLTVG-SQYNSKWSSA--------------STL-----------
Sca3X25c     33 KMNNYGKLCGTSAVVKDVTPG-LVFNPKWSKY--------------TTF-----------
Sca3X25d     32 KMNVNGTDFCNGTTIRNATTGTIKYNSKYTSA--------------STL-----------

Sas20d2a     79 ---VESACDVTVTFNPATNEIAVTGDGVKMVTD
Sas20d2b     98 GYPEDSLVDVTITLDLTKFD-YPSRSGAKANIK
Sca5X251a    78 ---VTKECDVTVTYNPTTNEITATGEGVVIPKE
Sca5X251b    99 GY--EDGYDIVLTLDLSNFN-YATKQGAKGKVD
Sca5X252a    77 ---VETACDVTVTFDPATNKITVTGDGVKMVTD
Sca5X252b    98 GFEEDSLVDVTITLDITNFD-YSTRSGAKATVK
Sca3X25a     66 ---KATGGKYTFKFNTAKNTLTIEY--------
Sca3X25b     67 ---AASGGTYTFSYDIDTNKLTISY--------
Sca3X25c     67 ---VATGGTYTFTYNTATNTLVVTA--------
Sca3X25d     67 ---VAVGGTYTFAYNALTNQLSVKYS------K
```

Figure 3.24: Sequence alignment of the individual X25 modules within Sca5, Sas20, and Sca3. Alignment was performed in T-Coffee and rendered in Boxshade. Residues involved in maltooligosaccharide binding in Sca5X25-2 are colored red.

Discussion

We harnessed a diverse array of biophysical and biochemical techniques to perform a structure-function characterization of Sas20, a multi-domain starch-binding amylosome protein in *R. bromii*. Our data revealed that one of these domains, Sas20d1, has a binding preference for the non-reducing ends of starch chains. In plants, starch granules are synthesized as a series of concentric layers of amorphous and semi-crystalline regions of amylose and amylopectin, from the reducing to the non-reducing end. The reducing ends of the α-glucan chains in amylopectin are less accessible as they are involved in the α1,6 glycosidic linkage that creates the branch points in amylopectin, whereas the non-reducing ends are much more abundant within these layers [219]. Due to the way starch is synthesized in plants, non-reducing ends may be more enriched towards the surface of the granules, and Sas20d1 may aid in anchoring *R. bromii* to the starch granule surface [219-221]. The Sas20d1 with maltotriose crystal structure showed a closing in of the bundle of two loops and α-helices over the ligand (**Figure 3.3D, F**), representative of the more compact states of Sas20d1, compared to the more extended states observed via SAXS (**Figure 3.16**). It is possible that the apparent ability of the Sas20d1 site to open facilitates the capture of the ends of the α-glucan chains within starch granules. The geometry of this binding site, based upon the orientation of maltotriose in the crystal structure, seems to not only target the non-reducing end of the α-glucan but favors a somewhat less helical α1,4-linked chain as might be more thermodynamically feasible at the chain end.

In contrast to Sas20d1, Sas20d2 has an elongated binding platform created by two X25 modules in tandem, which create a clamshell type structure that can recognize the helical turn of the α1,4 glycosidic bond. This binding site features four tryptophan residues, which is more extensive than the typical dual aromatic amino acid motif found in most structurally characterized starch-binding CBMs [68]. While the individual X25 modules of proteins such as SusE and SusF, which have two and three X25s respectively, bind maltooligosaccharides, our constructs of the individual X25 domains of the Sas20d2 failed to demonstrate maltooligosaccharide binding [209]. Sca5X25-2 and Sas20d2 demonstrate a ~1500x lower K_d for maltoheptaose over maltotriose, a modest preference for the longer sugar, similar to what we observed with Sas20d1 binding for these same substrates. For Sas20d2, the participation of both X25 modules in binding may be required to close the protein around the helical ligand, as suggested by the SAXS analysis of the domain with and without ligand. Sas20d2 failed to bind α-cyclodextrin and demonstrated weak binding for β-cyclodextrin which supports that the specific helical geometry of starch is indeed recognized, likely imposed by arrangement of the elongated binding platform.

In our isothermal depletion experiments, all constructs had similar affinities to starch granules, underscoring that both domains, despite the differences in their architectures, contribute to starch binding. We were somewhat surprised that Sas20d1-2 had a lower B_{max} than Sas20d1 on insoluble corn starch, as we speculated that additional binding modules may allow the protein to find more binding sites on the granule. It seems that instead the larger two-domain construct binds to fewer places on the granule, perhaps because the two domains recognize different structural motifs and/or the larger protein is more sterically restricted from adopting a range of binding orientations with the granule. Sas20, as part of cell-surface amylosomes, may provide the flexible recognition of different aspects of the starch structure that are revealed during RS degradation. The ability to recognize different parts of starch may be important for efficient RS degradation and may be one reason why there are several genes encoding putative starch-binding/dockerin-containing proteins in the *R. bromii* genome [42].

The SAXS data revealed that both Sas20 domains are flexible and less compact in solution compared to the crystal structure and homology model. However, contraction was observed in all samples in solution upon binding to ligand, especially Sas20d2. Because each individual domain displays a significant amount of flexibility it is difficult to determine how the linker contributes to this in the full-length construct, though presumably this linker adds to the potential range of conformations of the protein in solution which may enhance the ability of the protein to find starch motifs. Linkers between cellulose-active domains in the cellulosome have significant impacts on the higher-order structure of these complexes. Modifications and characteristics like heavy glycosylation, increased concentration of glycines or negative charged amino acids, and even short disulfide bridged loops may contribute to the extension of these complexes [222-225]. The linker between Sas20d1 and Sas20d2 is threonine-rich and may be a target of *O*-glycosylation. Since our recombinant protein work was expressed in *E. coli* which lacks the machinery required for *O*-glycosylation of proteins, it is still unclear if this linker is indeed glycosylated and how that modification affects the extension of Sas20.

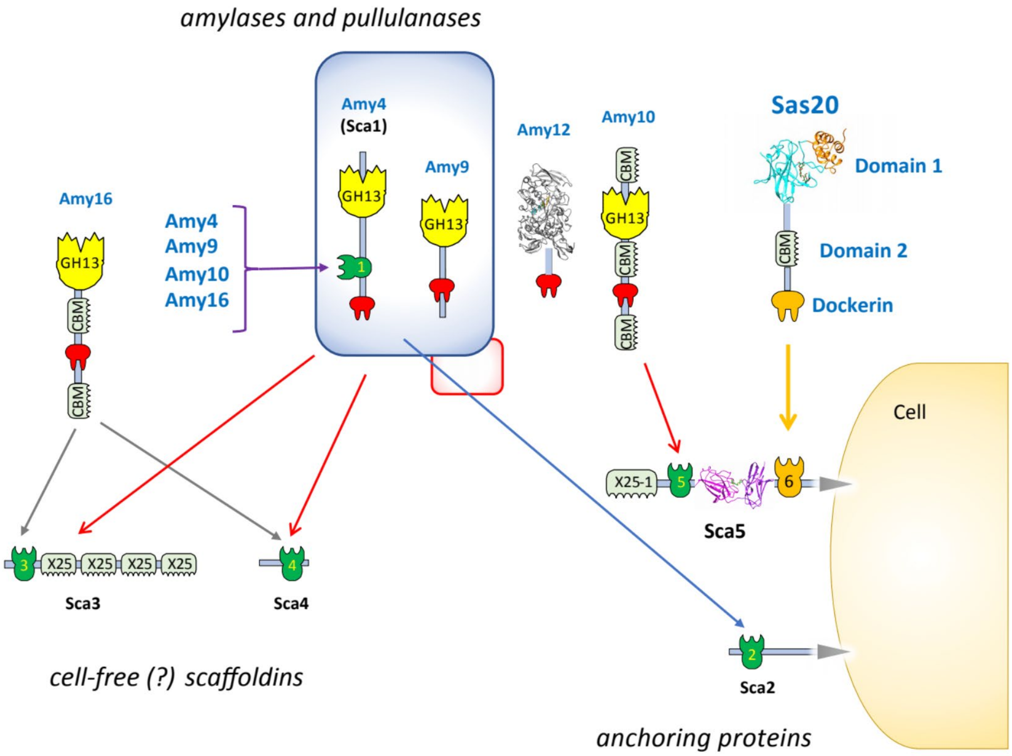


Figure 3.25: Updated model for cell-bound and cell-free amylosome complexes in *R. bromii* L2–63. We have added our newly found cohesin-dockerin interaction between the Sas20 dockerin and second cohesin of Sca5 to the most recent model of the amylosome system in *R. bromii*, adapted from Mukhopadhya et al. [42]. The crystal structures solved of amylosome protein domains in Sca5, Sas20, and Amy12 (PDB 7LSA) are shown.

With our data on Sas20 we present an updated model of the known cohesin-dockerin interactions that make the amylosome system (**Figure 3.25**) [42, 117]. Previous work and our EDTA elution experiment highlight that there are many other dockerin-containing amylosome proteins that are worthy of biochemical and/or structural characterization (**Table 3.1, Table S1**)

[42, 123]. Equally important to the biochemical properties of the starch-active portions of these proteins are their mechanisms of assembly into their respective amylosomes complexes. In the cellulosome system, cohesin-dockerin interactions are important in dictating the final architecture of the complex and even ligand preferences therein [226]. Each cohesin-dockerin complex differs in their binding interface, and this interface relates to their role in the cellulosome [227]. Moderate-affinity cohesin-dockerin interactions can permit the exchange of dockerin-containing enzymes in the cellulosome depending upon the substrates in the environment [228]. This allows enzymes with different substrate preferences to be incorporated into the cellulosome when the cell detects a change in the environmental polysaccharide. However, there is little evidence that genes encoding amylosome proteins are differentially regulated by exposure to different monosaccharides or different forms of starch [117, 229]. It is possible that at different phases in *R. bromii* growth, there are subtle changes in amylosome protein composition that may affect the types of amylosomes that are assembled. Therefore, further studies on the Sas20 dockerin and its interaction with the second cohesin of Sca5 are important for understanding the full role of Sas20 in *R. bromii*.

Materials and Methods

Growth and Proteomic Analysis of *R. bromii*

Freezer stocks of *Ruminococcus bromii* L2-63 were inoculated into 2 x 10 mL RUM medium as described in [117] supplemented with 1% galactose or autoclaved potato amylopectin in an anaerobic chamber (85% N_2, 10% H_2, 5% CO_2) and grown until they reached an OD_{600} of 0.5 (~48 h). Aliquots totaling 20 ml from each condition were harvested by centrifugation (4,500 x g for 5 min). Cells were resuspended in 1 mL of PBS (137mM NaCl, 2.7mM KCl, 10mM Na_2HPO_4, 1.8mM KH_2PO_4 pH = 7.4). The cells were again subjected to centrifugation and resuspended in 400 μL of PBS or PBS with 10mM EDTA and left to incubate at room temperature for 20 min. The cells were centrifuged again, and the supernatant was stored at -80°C before proteomic analysis.

Proteomics Analysis

R. bromii proteomic analysis was performed at the University of Michigan Proteomics Resource Facility. Cysteines were reduced by adding 10 mM DTT and incubating at 45°C for 30 min. Samples were then cooled to room temperature, and alkylation of cysteines was achieved by incubating with 65 mM 2-Chloroacetamide, under darkness, for 30 min at room temperature. An overnight digestion with sequencing grade modified trypsin (Enzyme:Substrate ratio of 1:50) was carried out at 37°C with constant shaking in a Thermomixer. Digestion was stopped by acidification, and peptides were desalted using SepPak C18 cartridges using the manufacturer's protocol (Waters). Samples were completely dried using a vacufuge, and resulting peptides were dissolved in an appropriate volume of 0.1% formic acid/2% acetonitrile solution to achieve ~500 ng peptide/ml. 2 mls of the peptide solution were resolved on a nano-capillary reverse phase column (Acclaim PepMap C18, 2 micron, 50 cm, ThermoScientific) using a 0.1% formic acid/2% acetonitrile (Buffer A) and 0.1% formic acid/95% acetonitrile (Buffer B) gradient at 300 nl/min over a period of 90 min (2-25% buffer B in 45 min, 25-40% in 5 min, 40-90% in 5 min followed by holding at 90% buffer B for 5 min and equilibration with buffer A for 30 min). Eluent was directly introduced into an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific, San Jose CA) using an EasySpray source. MS1 scans were acquired at 120K resolution (AGC target=1x10^6; max IT=50 ms). Data-dependent collision induced dissociation MS/MS spectra were acquired using the Top speed method (3 s) following each MS1 scan (NCE ~32%; AGC target 1x10^5; max IT 45 ms). Proteins were identified by searching the data against *the R. bromii* L2-63 protein database with 2111 entries, provided by Dr. Paul Sheridan at the Rowett Institute, using Proteome Discoverer (v2.4, Thermo Scientific). Search parameters included MS1 mass tolerance of 10 ppm and fragment tolerance of 0.1 Da; two missed cleavages were allowed; carbamidimethylation of cysteine was considered as fixed; oxidation of methionine, and deamidation of asparagine and glutamine, were considered as potential modifications. False discovery rate (FDR) was determined using Percolator, and proteins/peptides with an FDR of ≤1% were retained for further analysis.

Cloning, Protein Expression, and Purification

All genes and gene fragments were amplified from *R. bromii* genomic DNA using the Phusion^{TM} Flash polymerase (Thermo Fisher Scientific) according to the manufacturer's instructions for ligand-independent cloning with the Expresso T7 Cloning system using the pETite N-His vector kit (Lucigen Madison, WI, USA) according to the manufacturer's instructions. Primer sequences are listed in **Table 3.8** wherein the N-terminus contained a tobacco etch virus protease (TEV) cleavage site immediately downstream of the complementary 15 bp overlap (encoding the His tag) to create a TEV-cleavable His-tagged protein. Site directed mutagenesis was performed using the Agilent Technologies QuikChange Lightning Site-Directed Mutagenesis Kit according to the manufacturer's instructions.

Table 3.8: Primer Table.

Name	Sequence	Purpose
Sas20d1f	**CATCATCACCACCATCACGAGAACCTGTACTTCCAGGGC** TTCTCTGCATCAGCTGAAGAAACC	Forward primer for Sas20d1 and Sas20d1-2 constructs
Sas20d1Ar	**GTGGCGGCCGCTCTATTA** TCTGCTCTTGAAAGGTAA	Reverse primer for Sas20d1A construct
Sas20d1r	**GTGGCGGCCGCTCTATTA** AGGAGCTGTTGTACCTGAAGG	Reverse primer for Sas20d1 construct
Sas20d2f	**CATCATCACCACCATCACGAGAACCTGTACTTCCAGGGC** GAGCCTGCTGACGCAACACAG	Forward primer for Sas20d2 construct
Sas20d2r	**GTGGCGGCCGCTCTATTA** AGCATCGCCGAGAAGCGGAACACG	Reverse primer for Sas20d2 and Sas20d1-2 constructs
Sca5X25-2f	**CATCATCACCACCATCACGAGAACCTGTACTTCCAGGGC** GCTGCCGATACTACATATGTA	Forward primer for Sca5X25-2 and Sca5X25-2a constructs
Sca5X25-2ar	**GTGGCGGCCGCTCTATTA** GTTAACTTCAAGGTCTGTTAC	Reverse primer for Sca5X25-2a
Sca5X25-2bf	**CATCATCACCACCATCACGAGAACCTGTACTTCCAGGGC** ACAGACCTTGAAGTTAAC	Forward primer for Sca5X25b
Sca5X25-2r	**GTGGCGGCCGCTCTATTA** AGTTGAAGGCTCAACCTTAAC	Reverse primer for Sca5X25-2 and Sca5X25-2b
Sas20d1Y60Af	GTATATTGTCACCTTgccGCTGTAGCTGGCGAT	Forward primer for Sas20d1 Y60A mutant
Sas20d1Y60Ar	ATCGCCAGCTACAGCggcAAGGTGACAATATAC	Reverse primer for Sas20d1 Y60A mutant
Sas20d1W72Af	TTACCTGAAACATCTgcgCAGGGTAAGGCAGAG	Forward primer for Sas20d1 W72A mutant
Sas20d1W72Ar	CTCTGCCTTACCCTGcgcAGATGTTTCAGGTAA	Reverse primer for Sas20d1 W72A mutant
Sas20d2W329Af	CTCACAGGTTATGAAgcgCAGGGTTCTCCTGCA	Forward primer for Sas20d2 W329A mutant
Sas20d2W329Ar	TGCAGGAGAACCCTGcgcTTCATAACCTGTGAG	Reverse primer for Sas20d2 W329A mutant
Sas20d2W375Af	GGCGACGAGCAGAAGgcgATCGGTCTTGACGGT	Forward primer for Sas20d2 W375A mutant
Sas20d2W375Ar	ACCGTCAAGACCGATcgcCTTCTGCTCGTCGCC	Reverse primer for Sas20d2 W375A mutant

Sas20d2W440Af	CTTAACGGTGTAGCAgcgGGCGTTGACGCTGAA	Forward primer for Sas20d2 W440A mutant
Sas20d2W440Ar	TTCAGCGTCAACGCCcgcTGCTACACCGTTAAG	Reverse primer for Sas20d2 W440A mutant
Sas20d2W481Af	GCAGTTAACGACGATgcgGCTGCTAACTGGGGT	Forward primer for Sas20d2 W481A mutant
Sas20d2W481Ar	ACCCCAGTTAGCAGCcgcATCGTCGTTAACTGC	Reverse primer for Sas20d2 W481A mutant

Bold text denotes regions of homology for cloning into pETite vectors and engineered TEV cleavage site, lower case text denotes mutagenized region.

For Sas20 dockerin-cohesin interaction studies, the PCR product was digested with KpnI and BamHI restriction enzymes (New England Biolabs, Inc., Ipswich, MA) and inserted into the restricted pET28a, containing *Geobacillus stearothermophilus* xylanase T-6 [203]. CBM-fused cohesins (CBM-Cohs) were cloned as described previously [42, 117]. All plasmids insert sequences were verified by Sanger sequencing conducted by Eurofins Scientific. Xyn-Sas20 and the CBM-Coh fusion proteins were expressed in *E. coli* BL21 pLysS (DE3) and purified as described by Ben David et al. To determine potential Sas20 interactions to *R. bromii* cohesins, the standard affinity-based ELISA procedure of Barak was performed [203].

Expression plasmids were transformed into *E. coli* Rosetta (DE3) pLysS cells, expressed and purified as previously described. Selenomethionine-substituted Sca5X25-2 was produced by first transforming the plasmid into *E. coli* Rosetta (DE3) pLysS and plating onto LB, supplemented with kanamycin (50 mg/ml) and chloramphenicol (20 mg/ml). The bacteria were grown for 16 h at 37°C, and then colonies were harvested from the plate to inoculate 100 ml of M9 minimal medium supplemented with the same antibiotics. After 16 h of incubation at 37°C, this starter culture was used to inoculate a 2-liter baffled flask containing 1 liter of Molecular

Dimensions Seleno-Met premade medium, supplemented with 50 ml of the recommended sterile nutrient mix, chloramphenicol, and kanamycin. Cultures were incubated at 37°C to an OD_{600} of 0.5, the temperature was adjusted to 23°C and each flask was supplemented with 100 mg each of L-lysine, L-threonine and L-phenylalanine and 50 mg each of L-leucine, L-isoleucine, L-valine and L-selenomethionine [230]. After 20 min of further incubation, protein expression was induced by the addition of 0.5 mM IPTG, and cultures were allowed to grow for an additional 48 h before harvest by centrifugation. Cells were then lysed by sonication, and the protein purified as previously described via nickel affinity chromatography [107].

Affinity PAGE

Native 10% polyacrylamide gels with and without 0.1% added polysaccharide (glycogen, pullulan, autoclaved potato and corn amylopectin, and dextran) were cast with 0.375 M Tris-HCl pH 8.8 as described in [231, 232]. Gels were subjected to 100 V for 4 h and then stained for 2 h with 0.1% Coomassie Brilliant Blue R-250 in 10% acetic acid, 50% methanol, 40% water before destaining with solution lacking dye overnight with one change of solution. Binding was considered positive if the migration of the protein in the polysaccharide gel relative to a noninteracting protein (bovine serum albumin) was significantly slower (<0.85 relative mobility) compared to that in the control gel.

Crystallization and X-ray Structure Determination

Sas20d1 crystallization experiments were performed using a Crystal Gryphon (Art Robbins) in 96-well trays using a sitting drop format. Diffraction quality crystals of native Sas20d1 were obtained by mixing 35 mg/ml protein 1:1 (vol/vol) with the crystallization solution containing 0.024 M 1,6-Hexanediol; 0.024 M 1-Butanol, 0.024 M 1,2-Propanediol; 0.024 M 2-

Propanol; 0.024 M 1,4-Butanediol; 0.024 M 1,3-Propanediol; 0.1 M Imidazole; 0.1 M MES monohydrate pH = 7.5; 20% PEG 500 MME; and 10% PEG 20000. Native Sas20d1 crystals were plunged directly from the well into liquid nitrogen for x-ray data collection. Sas20d1 (32 mg/ml) plus (10 mM) maltotriose was subjected to a series of 24-well hanging drop sparse matrix screens to identify crystallization conditions. Crystals were obtained via hanging drop vapor diffusion at room temperature against 27% PEG 4000, 0.2 M $MgCl_2$, 0.1 M Tris pH=7.5. Prior to data collection, crystals were cryoprotected by a swiping through a solution of 80% mother liquor supplemented with 20% ethylene glycol then plunged into liquid nitrogen. Selenomethionine-substituted Sca5X25-2 (40 mg/ml) plus (10 mM) maltotriose was subjected to a series of 96-well hanging drop sparse matrix screens to identify crystallization conditions. Crystals were obtained via hanging drop vapor diffusion at room temperature against 2 M ammonium sulfate, 0.1 M sodium acetate pH 4.6. Prior to data collection, crystals were cryoprotected by a swiping through a solution of 70% mother liquor supplemented with 30% glycerol then plunged into liquid nitrogen.

X-ray data from Sas20d1 crystals were collected at the Life Sciences Collaborative Access Team (LSCAT) beamline ID-F of the Advanced Photon Source at Argonne National Laboratory, and data from Sca5X25-2 crystals were collected at beamline ID-G from the same source. The Sas20d1 structure was determined via sulfur single anomalous dispersion (SAD) phasing using multiple data sets, processed and merged within HKL2000 and Scalepack [233], and the maltotriose-bound Sas20d1 structure was phased by molecular replacement with the native Sas20d1 dataset. The Sca5X25-2 with maltotriose structure was phased by seleno-methionine substitution. Phasing was performed using AutoSol in Phenix [234]. The protein models were finalized via alternating cycles of manual model building in Coot and refinement in Phenix.refine and/or Refmac5 from the CCP4 suite [168, 235, 236].

Isothermal Titration Calorimetry

ITC measurements were carried out using a TA Instruments Nano ITC. Proteins were dialyzed into 50 mM HEPES pH = 7.0, and oligosaccharides were prepared using the dialysis buffer. Protein (25-75 µM) was placed in the sample cell, and the reference cell was filled with water. After the temperature was equilibrated to 25°C, a first injection of 2 µL was performed, followed by 29 subsequent injections of 10 µL of 2-10 mM maltotriose, maltoheptaose, or 0.025% autoclaved corn and potato amylopectin. For polysaccharide titrations, the concentration of ligand was adjusted to fit a one-site binding model with n=1; this sets the concentration of the ligand to the concentration of binding sites for the protein within the polysaccharide, as previously described [232]. The solution was stirred at 250 rpm and the resulting heat of reaction was measured. Data were analyzed using the TA Instruments NanoAnalyze software package fitting to a one-site binding model. Isotherms are displayed in **Figures 3.7-11**.

Isothermal Depletion Assay

Recombinantly expressed protein binding to raw corn starch (National Starch Food Innovation 9735) was determined by adsorption as previously described [209, 232]. Raw starch was prepared by washing with sterile PBS three times by resuspension and centrifugation. Aliquots (150 µL) of 10% w/v starch were aliquoted into 0.2 mL tubes, pelleted by centrifugation (2,000 x g), and the supernatant fluids were removed leaving 15 mg of raw starch per tube in triplicate for each concentration. Aliquots (150 µl) of protein (0-1.0 mg/mL) in 100 mM NaCl, 20 mM pH = 7.0 HEPES buffer were added to the starch for a final 10% w/v of starch. Triplicate reactions were agitated by inversion for 1 h at 23°C then pelleted (2,000 x g), and the protein concentration remaining in the supernatant was measured by Pierce BCA Assay, using free protein concentrations to create a standard curve for each construct. The results were validated by measuring A_{280} on a NanodropC with the theoretical molecular weight and extinction coefficient for each protein. The µmol protein bound was determined by subtracting the bound protein measurement from the free protein value and normalized to the amount of starch as µmol bound per gram of starch. Bovine serum albumin was used as a non-binding negative control. A one-site specific binding model was used to determine K_d and B_{max} in GraphPad Prism.

Circular Dichroism

Determination of circular dichroism (CD) spectra for both wild-type and the truncation mutant was carried out with a J-815 circular dichroism spectropolarimeter (Jasco, Tokyo, Japan). A protein concentration of 0.1 mg/ml was prepared in 10 mM KH_2PO_4 buffer (pH= 7.5). Substrate was added to a concentration of 1 mM and incubated for 24 h with protein before performing CD. A quartz cell with a path length of 0.1 cm was used. Three CD scan replicates per condition were carried out at 25°C from 190 nm to 260 nm at a speed of 50 nm/min with a 0.5-nm wavelength pitch. Data files were analyzed with the DICHROWEB online server (http://dichroweb.cryst.bbk.ac.uk/html/process.shtml) using the CDSSTR algorithm with reference set 4, which is optimized for analysis of data recorded in the 190 nm to 240 nm range. Mean residue ellipticity was calculated using millidegrees recorded, molecular weight, number of amino acids and concentration of protein. Temperature interval experiments were performed in triplicate with a protein concentration of 0.1 mg/ml prepared in 10 mM KH_2PO_4 buffer (pH= 7.5). CD scans were collected from 190 to 260 nm at a speed of 50 nm/min with a 1-nm wavelength pitch at temperature intervals of 10°C between 25°C to 95°C.

SEC-SAXS Experiments

SAXS was performed at BioCAT beamline 18ID at the Advanced Photon Source at Argonne National Labs using in-line size exclusion chromatography (SEC-SAXS) to separate sample from aggregates and other contaminants. Sample was loaded onto a Superdex 200 Increase 10/300 GL column (Cytiva), which was run at 0.6 ml/min by an AKTA Pure FPLC (GE), and the eluate after it passed through the UV monitor was flown through the SAXS flow cell. The flow cell consists of a 1.0 mm ID quartz capillary with ~20 µm walls. A coflowing buffer sheath was used to separate sample from the capillary walls, helping prevent radiation damage. Scattering intensity was recorded using a Pilatus3 X 1M (Dectris) detector which was placed 3.6 m from the sample providing a q-range of 0.005Å-1 to 0.35Å-1. Exposures of 0.5 s were acquired every 1 s during elution, and data were reduced using BioXTAS RAW 2.1.0 [237]. Buffer blanks were created by averaging regions flanking the elution peak and subtracted from exposures selected from the elution peak to create the I(q) vs q curves used for subsequent analyses. The Bayes method was used to calculate molecular weights. MultiFoXS was used to generate ensembles using the SAXS data and high-resolution crystal structures or models [215].

Data Availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [238] partner repository with the dataset identifier PXD032013. The X-ray structures and diffraction data reported in this paper have been deposited in the Protein Data Bank under the accession codes 7RPY, 7RFT, and 7RAW. The SAXS data are deposited in the small angle x-ray scattering database (SASDB) under the accession codes SASDMX9, SASDMY9, SASDMZ9, SASDN22, SASDN32, SASDN42 [239].

CHAPTER 4:

Molecular Structure of the Cohesin Modules of Sca5, a Scaffoldin Protein from the *Ruminococcus bromii* Amylosome

Abstract

Ruminococcus bromii is a common member of the human gut microbiota and has the rare ability to degrade resistant starch (RS). *R. bromii* expresses one or more assemblies of starch-active proteins, called amylosomes, that bind and degrade starch. While recent studies have elucidated the biochemical features of individual amylosome proteins, our knowledge of how these components are assembled is limited. Cohesins and dockerins are protein-protein interaction modules that drive amylosome assembly, yet little is known about the structure of these modules from *R. bromii*. In this study, we present the first crystal structures of two cohesin modules from the amylosome scaffoldin protein Sca5. Both cohesin modules displayed a canonical jelly-roll fold comprised almost entirely of β-sheets. We used AlphaFold-Multimer to predict the cohesin-dockerin complex structure of the second Sca5 cohesin module, Sca5Coh6, to its cognate dockerin module from Sas20, Sas20doc. The prediction displayed a typical cohesin-dockerin complex structure with hydrogen bonding and hydrophobic interactions at the interface. Even though amylosomes and cellulosomes are comprised of complexes that target

markedly different substrates, this study supports that cohesin modules from both systems are structurally conserved.

Introduction

The unique ability of *Ruminococcus bromii* to dock onto and degrade starch granules is attributed to its production of amylosomes, complexes of starch-binding and starch-degrading proteins that assemble via complementary cohesin-dockerin modules. [42, 117]. Cohesins and dockerins are well studied components of cellulosomes, complexes of cellulose-degrading proteins synthesized by Gram positive rumen and environmental bacteria [226]. Bioinformatic analysis of the *R. bromii* proteome revealed several amylases and pullulanases with dockerin domains, and scaffoldin proteins harboring cohesin domains [42]. Both α-amylases and pullulanases are glycoside hydrolase family 13 (GH13) enzymes that cleave α-1,4 linkages and α-1,6 linkages, respectively, the two glycosidic bonds that dictate starch structure [73]. These enzymes are found in many gut bacterial species, including those without the ability to grow on RS [20]. Scaffoldin proteins are non-catalytic structural proteins that possess one or more cohesins, domains that interact strongly with dockerins [240]. Scaffoldins are central to the assembly of dockerin-containing multiprotein complexes and can be anchored to the cell wall, or freely secreted. While the catalytic domains of GH13 enzymes in the amylosome are similar to those from non-RS degrading bacteria, their assembly into amylosome complexes facilitates enzymatic synergy which allows for efficient degradation of starch granules. [42, 117, 123].

Several crystal structures of cohesin-dockerin complexes from cellulosomes have revealed that cohesins are typically β-strand-rich structures and that dockerins possess a helix-loop-helix motif and an EF-hand calcium-binding loop [241-244]. Each cohesin-dockerin complex differs in the composition of the binding interface, and this relates to their role in the cellulosome [227]. Cohesin-dockerin interactions are usually driven by either a small hydrogen-bonding interface or an extended hydrophobic interface, leading to moderate affinity (K_a~10^8 M) or ultra-high affinity (K_a ~10^{11}) binding, respectively [245].

The *R. bromii* genome encodes five putative cohesin-containing scaffoldin proteins (Sca1-5). Sca2 and Sca5 have sortase recognition sequences at their C-terminus, making them putative cell wall-anchored scaffoldins [42]. As there have been no cohesin structures solved from the amylosome system, our first goal was to gain a structural understanding of these cohesin modules. Sca5 has two starch-binding domains, each comprised of two X25 family folds in addition to two cohesin modules (**Figure 4.1A**). Sca5 is the only scaffoldin in *R. bromii* with more than one cohesin module, and its first cohesin, Sca5Coh5, binds promiscuously to several dockerin modules, including those from the amylosome enzymes Amy4, 9, 10, 12, and 16 [42]. We discovered in Chapter Three that the Sas20 dockerin module, Sas20doc, preferentially binds to the second cohesin module of Sca5, Sca5Coh6. The structural basis for specificity and molecular features of the Sca5-Sas20 cohesin-dockerin interaction that drive assembly of the cell surface amylosome is still unclear. In this study, we present the crystal structures of the two cohesin modules of Sca5, Sca5Coh5 and Sca5Coh6, and a model generated through AlphaFold-Multimer to predict the Sca5Coh6-Sas20doc complex structure. A thorough understanding of how the amylosome is assembled may allow us to predict additional interacting partners for Sca5, or rationally design efficient starch-degrading protein complexes for application in the food, paper, or detergent industries [246].

Results

The Sca5Coh5 and Sca5Coh6 Structures are Conserved

We determined the structures of Sca5Coh5 and Sca5Coh6 to 1.8 Å and 1.0 Å resolution, respectively (**Table 4.1**). The structures can be superimposed with a RMSD of 1.0 Å for 96 Cα pairs (**Figures 4.1B, C, D**). Overall, these structures are surprisingly homologous considering they only possess 44% similarity across 85% of their sequences with 29% identity. Both structures have ten β-strands showing an elongated elliptical jelly-roll fold typical of a cohesin structure [247]. The first and last β-strands align parallel to each other and function to close the jelly-roll, while the rest are antiparallel. As we were particularly interested in understanding the specificity of the Sca5Coh6-Sas20doc interface, determined via ELISA in our previous work, we focused our structural analysis on Sca5Coh6.

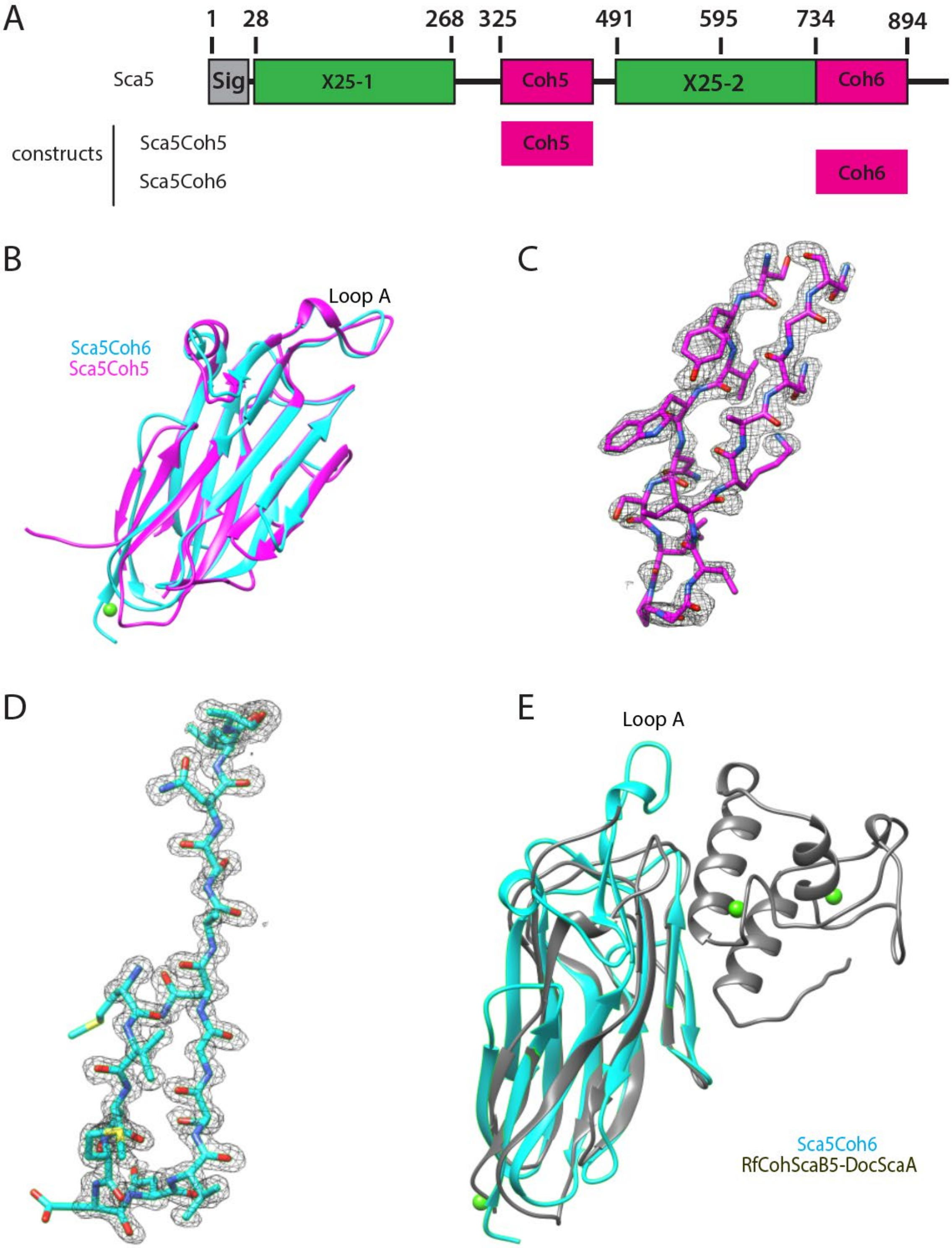


Figure 4.1: Sca5Coh5 and Sca5Coh6 structures. A) Cartoon depiction of constructs used for crystallization. B) Alignment of Sca5Coh5 and Sca5Coh6 structures. Omit map of residues C) 279-294 from Sca5Coh5 and D) 803-817 from Sca5Coh6, σ=2.5. E) Alignment of Sca5Coh6 with structural homolog from *R. flavefaciens* (PDB 5N5P).

Sca5Coh6-Sas20doc Complex Structure Prediction

A search on the DALI server showed that the cohesin module from the *Rf*CohScaB5-DocScaA cohesin-dockerin complex structure from *Ruminococcus flavefaciens* (PDB: 5N5P) was the closest structural homolog to Sca5Coh6, and they aligned with an RMSD of 1.0 Å for 49 Cα pairs [248] (**Figure 4.1E**). The most notable difference between Sca5Coh6 and *Rf*CohScaB5 is that Sca5Coh6 has a ~14 amino acid insertion (residues 851-865) that makes up a loop (Loop A) with an α-helix turn. Unfortunately, we were unable to crystallize the Sca5Coh6-Sas20doc complex. Therefore, we used AlphaFold-Multimer to generate a model of this complex and compared it to the *Rf*CohScaB5-DocScaA cohesin-dockerin complex structure [249].

AlphaFold-Multimer defines a high-confidence model as having a confidence score of greater than 80%; the confidence score for the Sca5Coh6-Sas20doc complex was over 90% across most of the sequence while the N-terminus of Sca5Coh6 and C-terminus of Sas20doc were regions of low-confidence **(Figure 4.2A**) [249]. Additionally, the cohesin portion of the prediction was nearly identical to the crystal structure of Sca5Coh6 as these two models aligned with an RMSD of 0.5 Å for 155 Cα pairs. The planar surface formed by β-sheets 2-1-9-4-7-6 from both cohesin modules form many interactions with the first and third α-helices of the dockerin modules. The opposite face of the cohesin module, comprised of β-sheets 10-3-8-5, are solvent exposed and do participate in dockerin binding.

Overall, the Sca5-Sas20 prediction aligned with the cohesin of *Rf*CohScaB5-DocScaA cohesin-dockerin complex structure from *Ruminococcus flavefaciens* with an RMSD of 1.0 Å for 50 Cα pairs and has the same general fold (**Figure 4.2B**). The main difference is the order of the three α helices that comprise the dockerin modules. The first α helix encoded by *Rf*DocScaA (residues 661-674) aligns with the third α helix of Sas20doc (residues 604-616). The third α helix encoded by *Rf*DocScaA (residues 710-721) aligns with first α helix (residues 566-578) encoded by Sas20doc. However, this could be due to the use of prediction software, as the dockerin is made of two repeated structural motifs with 50% sequence identity. Nevertheless, the first and third helices are antiparallel and form the planar surface by which their cognate cohesin module may bind. The two repeated helices, each resembling an EF-hand motif with one Ca^{2+}, are typical of dockerin module structures. Since the order of the helices between the *R. flavefaciens* and *R. bromii* dockerins seem to be inverted, the second helices are not conserved and are oriented inversely of each other.

Since dockerin modules are made up of two repeated structurally conserved motifs, they can sometimes bind to their cognate cohesin in two different orientations if structural symmetry is conserved [120, 250]. These cohesin-dockerin interactions are said to possess a dual-binding mode of binding as the dockerin module can bind to its cognate cohesin in two different orientations. To explore if this interaction is single or dual-binding mode, we aligned the N-terminal and C-terminal EF-hand motif of the Sas20doc module structure prediction (**Figure 4.2C)**. While the helix fold was conserved, the motifs aligned in opposite orientations of each other. Two key residues in the cohesin-dockerin interface, I604 and I578, are positioned on opposite sides of the helix. This suggests that both helices are likely not competent for binding, and this interface may display a single-binding mode.

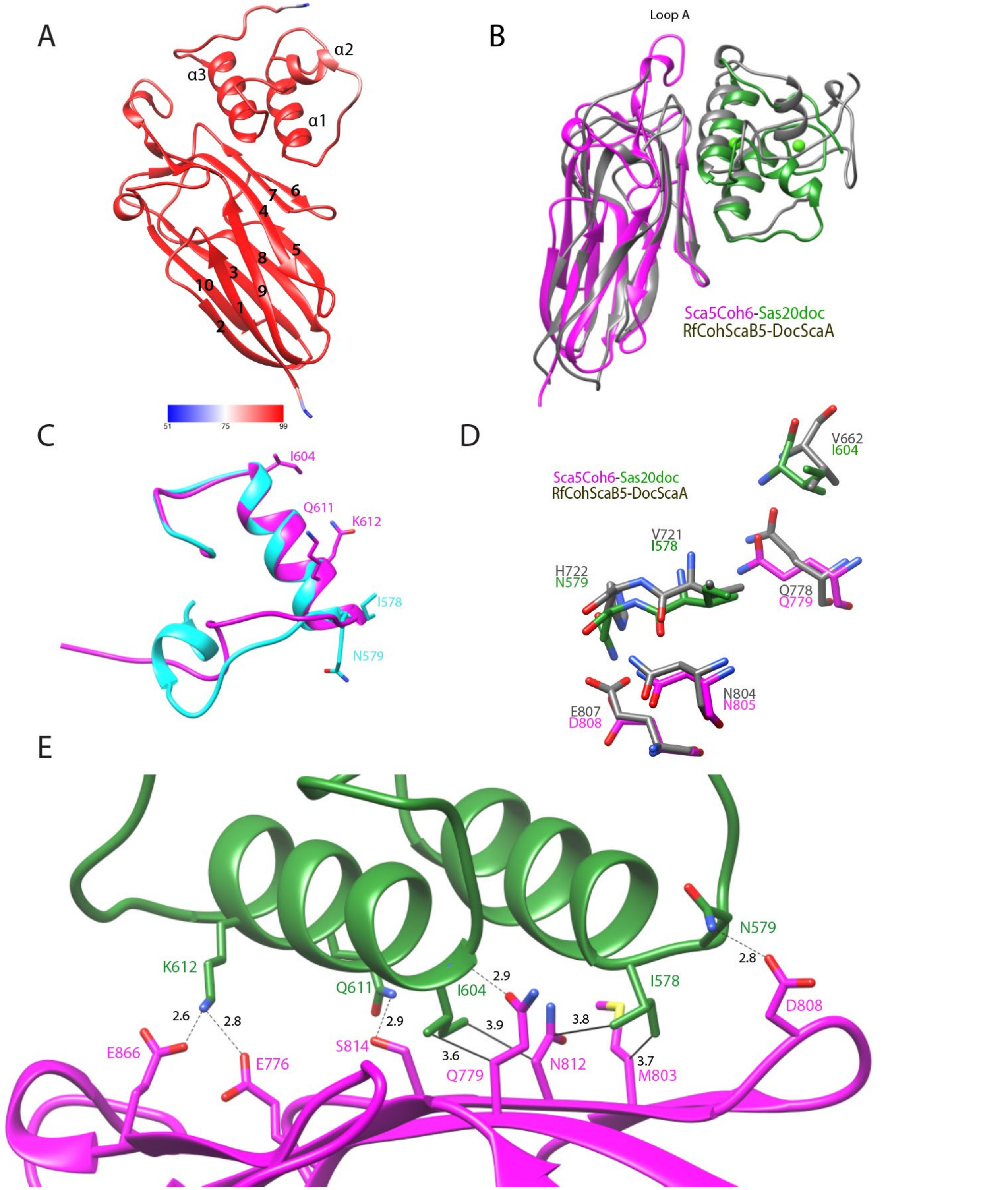


Figure 4.2: Structural analysis of the Sca5Coh6-Sas20doc interface. A) AlphaFold-Multimer prediction colored by confidence level. 10 β-sheets from Sca5Coh6 and 3 α-helices in Sas20doc are numbered B) Alignment of Sca5Coh6-Sas20doc with RfCohSca5B-DocScaA (PDB 5N5P). Protein is depicted in ribbons, and calciums are depicted as green spheres. C) Alignment of Sas20 N-terminus (residues 557-593) and C-terminus (residues 594-630) with key residues in stick diagram D) Conserved residues at the cohesin-dockerin interface of Sca5Coh6-Sas20doc and RfCohSca5B-DocScaA in stick diagrams. E) Zoom in of Sca5Coh6-Sas20doc interface with key residues in stick diagram. Dotted lines represent hydrogen bonds and solid lines represent hydrophobic interactions.

Q779, D808, and N805 from Sca5Coh6 and N579, I604, and I578 from Sas20doc and their pattern of interactions are conserved with *Rf*CohScaB5-DocScaA at the cohesin-dockerin interface (**Figure 4.2D)**. Loop A from Sca5Coh6 does not seem to participate via hydrogen or nonpolar interactions at the cohesin-dockerin interface. E866 and E776 of Sca5Coh6 can potentially form hydrogen bonds with the N^{ε} of K612 Sas20doc (**Figure 4.2E)**. In this model, hydrogen bonding occurs with Q611 and N579 of Sas20doc and S814 and D808 of Sca5Coh6, respectively. Likewise, the side chain amide O of Q779 from this Sca5Coh6 structure prediction hydrogen bonds with the backbone N of I604 of Sas20doc. The first carbon of the Q779 side chain may also contribute to a group of nonpolar interactions alongside N812 and M803 of Sca5Coh6 with the side chains of I604 and I578 of Sas20doc.

Table 4.1: X-Ray data collection and refinement statistics

	Sca5Coh5 with NaI	**Sca5Coh6 with NaI**	**Sca5Coh6 native**
PDB accession	7UVG		7URP
Wavelength (Å)	0.979	0.979	0.729
Resolution range (Å)	22.04-1.77 (1.81-1.77)	24.93-1.18 (1.21 -1.18)	30.05 - 1.03 (1.06 - 1.03)
Space group	C 2 2 21	C 1 2 1	C 1 2 1
Unit cell (Å)	*a*=31.0, *b*=137.8 *c*= 85.6	*a*=121.1, *b*=30 *c*= 50.0 β=94.7	*a*=120.6, *b*=30 *c*= 50.0 β=94.5
Total reflections	67456 (6942)	372113 (14384)	1166863 (55466)
Unique reflections	17738 (1749)	53479 (3699)	87298 (6377)
Multiplicity	3.8 (4.0)	7.0 (4.0)	13.4 (8.7)
Completeness (%)	94.5 (92.0)	83.5 (99.9)	98.6 (98.12)
Mean I/sigma(I)	11.47 (1.36)	9.7 (1.0)	12.0 (0.7)
R-merge	0.050 (0.959)	0.084 (1.44)	0.077 (2.89)
R-meas	0.058 (1.106)	0.091 (1.57)	0.083(3.26)
R-pim	0.029 (0.543)	0.034 (0.62)	0.031 (1.45)
CC1/2 in highest resolution shell	0.69	0.40	0.42
CC* in highest resolution shell	0.90	0.76	0.77
Reflections used in refinement	17438 (1677)		82774 (6046)
Reflections used for R-free	2114 (184)		4476 (331)
R-work	0.216 (0.290)		0.143(0.317)
R-free	0.277 (0.320)		0.165(0.319)
Number of non-hydrogen atoms	1385		2255
macromolecules	1301		1877
ligands	12		74
solvent	72		304
ions	n/a		n/a
Protein residues	166		241
RMS(bonds)	0.013		0.014
RMS(angles)	1.7		1.5
Ramachandran favored (%)	95.1		98.1
Ramachandran allowed (%)	4.9		1.9
Ramachandran outliers (%)	0		0
Rotamer outliers (%)	1.4		0
Clashscore	1.55		1.26
Average B-factor	41.2		18.1
macromolecules	40.5		16.4
ligands	48.4		40.7
solvent	52.5		31.7
ions	n/a		13.0

Discussion

Typically, high affinity (K_d ~10^{-11} M) cohesin-dockerin binding represent “type-II” interactions and are crucial to the stability and anchorage of cellulosomes to the cell surface [251]. Type-II cohesin-dockerins are typically bound by highly extensive networks of hydrophobic interactions. There are no cohesin-dockerin interactions predicted to be critical to anchoring amylosomes to the cell surface analogous to cellulosomes. *R. bromii* uses the alternative method of covalently anchoring its cell-surface amylosomes through Sca2 and Sca5 via C-terminal sortase recognition[42].

Moderate affinity interactions (K_d ~10^{-8} M) that represent “type-I” interactions allow for flexibility of the enzyme repertoire that a cellulosome complex possesses [252]. This is thought to give these complexes the ability to modify and fine-tune the available motifs within cellulose that cellulosome can attack. According to BLAST, the Sas20doc sequence is most similar to type-I dockerins [253]. However, a key feature of the type-I interaction is a Ser/Thr “recognition” residue near position 11 of the dockerin module of which Sas20 lacks (**Figure 4.3**) [120, 254].

```
Doc22  LRGDVDGDGEVTVMDSTYLQKSIVGISGVPALETLDLKICDLDGDGT-ISVKDATIIQKIVVNLE-------------
Sas20  LLGDADGDYSITVVDATTIQKIAINLMSIA-ADDANAFKACDANEDGRISIKDATIVQKYIVGGYETGNVGSPISVE-
Doc19  EINDVNTDGSIDILDATQVQKYLAKIDS---PTYVQNKLADCNGDGV-INVRDATYIQKTIVKIPVYNLKNS------
Amy10  LIGDVNLNGAIDIVDTTAVQKYIVKLITLS-KALIAAARCDADGENDIVSVKDATYIQMYVAKLDGHGNVGTYYESEV
Amy4   TLGDVNMDGDITVVDATEVQRYVAQLVA---FTNDQFTAADVDHDGT-ITVKDATTIQKFVVDLISNF----------
Doc5   VLGDVNGDGVVSVVDATLVQKYIVGEVD---FNCAQKLRARVNWNAFPITVKESTTIQKYIVGCTDVYFYTGGIVENE
Doc8   KIGDIDADDRITVKDATEVQKYCASDIE---FSELEKSLADVNGDGT-VNVIDATEIQKIAINAK-------------
Doc14  LNGDVDLDGDIAVKDATLVQKYIVKLEQ---FDNTQLCNADCDGDGD-ITVADATKIQKIVVGIN-------------
          *          * *  *                                  *  *
```

Figure 4.3: Sequence alignment of related dockerin modules from *R. bromii* amylosome proteins. Conserved residues are marked in red and denoted with an asterisk. The recognition residue is marked in blue. Residues in Sas20doc predicted to contribute to the cohesin-dockerin interaction are marked in green. Alignment is reformatted from [254].

Not all cohesin-dockerin complexes can be categorized into Type I or II [255]. Both the putative Sca5Coh6-Sas20doc interaction from *R. bromii* and the *Rf*CohScaB5-DocScaA interaction from *R. flavefaciens* display hydrogen bonding and hydrophobic interactions. The

*Rf*CohScaB5-DocScaA interaction has a moderate affinity (K_d ~4 x 10^{-8} M) [248]. The Sca5Coh6-Sas20doc interaction was discovered by assaying Sas20doc binding to all six cohesin modules in *R. bromii* via ELISA. In this assay, Sas20doc clearly had higher binding sensitivity to Sca5Coh6 and showed little to no signal for any of the other five *R. bromii* cohesins (Chapter Three). However, we have so far been unable to confirm the Sca5Coh6-Sas20doc interaction using isothermal titration calorimetry or biolayer interferometry. One reason for this might be that the Sca5Coh6-Sas20doc interaction is low affinity. This low affinity could be detected by the high sensitivity of the ELISA assay while it may be below the threshold of detection for isothermal titration calorimetry or biolayer interferometry techniques.

Our colleague Yonit Ben David (Weizmann Institute, Israel) performed a sequence analysis of the dockerins in *R. bromii* (**Figure 4.3**) [254]. Here, dockerins were aligned and clustered into groups with stronger weight given to the sequence defining the calcium-binding loops and putative helix regions. The Sas20 dockerin was grouped with seven other dockerin-containing proteins that have conserved residues at these motifs. Q611 was the only key residue at the predicted binding interface from Sca5Coh6-Sas20doc that was conserved among all similar *R. bromii* dockerins. K612 was mostly conserved with the only exception being that the dockerin module from Amy10 has a Met. Additionally, at position I604 from Sas20doc, all other dockerin modules in this group have a Val suggesting that the Sca5Coh6-Sas20doc binding interface is relatively unique. Due to the sparse conservation of residues contributing to the Sca5Coh6-Sas20doc interface, it is unclear which additional dockerin modules from *R. bromii*, if any, bind to Sca5Coh6.

The *Rf*CohScaB5-DocScaA interaction assembles two cohesin-containing scaffoldins, ScaB5 and ScaA, the latter of which also has a dockerin module, on the cell-surface of *R.*

flavefaciens. This interaction gives this bacterium the ability to form higher-order cellulosome complexes. In some circumstances, the ability to assemble complex cellulosomes with many enzymes that can attack diverse glycosidic linkages within the plant cell wall is beneficial. However, in environments with relatively simple cellulose substrates, assembling smaller cellulosomes is more energetically favorable [256, 257]. The only amylosome protein that has both a cohesin and dockerin module is Amy4, and we know that the cohesin module from this protein can bind to its own dockerin as well as the dockerin module from Amy9 [117]. Sca5Coh5 pairs with several amylosome protein dockerins, so it is plausible this promiscuity allows for continuous exchange of different dockerins [42]. The exchange of dockerin-containing proteins could help *R. bromii* navigate the complex and dynamic nature of nutrient availability in the colon. However, this exchange can only happen if the affinity is moderate and not ultra-high, and we have yet to determine the K_d of any cohesin-dockerin interactions from the amylosome. Further cohesin-dockerin studies that determine the structure and precise affinities of interactions are needed to fully elucidate the mechanism of amylosome assembly and its impact on RS degradation.

Materials and Methods

Cloning, Protein Expression, and Purification

All genes and gene fragments were amplified from *R. bromii* genomic DNA using the Phusion^{TM} Flash polymerase (Thermo Fisher Scientific) according to the manufacturer's instructions. Amplified fragments were inserted into the Expresso T7 pETite N-his vector (Lucigen Madison, WI, USA) via ligation-independent cloning according to the manufacturer's instructions. Primer sequences are listed in **Table 4.2** wherein the N-terminus contained a tobacco etch virus protease

(TEV) cleavage site immediately downstream of the complementary 15 bp overlap (encoding the His tag) to create a TEV-cleavable His-tagged protein.

Table 4.2: Primers used in study

Name	Sequence	Purpose
Coh5F	*CATCATCACCACCATCACGAGAACCTGTACTTCCAGGGC* GTTACAGCTACTTCAAAC	Forward primer for Sca5Coh5 construct
Coh5R	*GTGGCGGCCGCTCTATTA* CTCTTCTGAACCGTCGGGATC	Reverse primer for Sca5Coh5 construct
Coh6F	*CATCATCACCACCATCACGAGAACCTGTACTTCCAGGGC* GCAGTTGATAATTTAACAATC	Forward primer for Sca5Coh6 construct
Coh6R	*GTGGCGGCCGCTCTATTA* CTCAACATATGCCTCAACCTT	Reverse primer for Sca5Coh6 construct

Italicized sequence is responsible for cloning His tag and TEV cleavage site.

Protein Expression and Purification

Expression plasmids were transformed into *E. coli* Rosetta (DE3) pLysS and colonies selected on LB supplemented with kanamycin (50 mg/ml) and chloramphenicol (20 mg/ml). The bacteria were grown for 16 h at 37°C, and then colonies were harvested from the plate to inoculate 50 ml of LB supplemented with the same antibiotics. Bacteria were grown for 16 h at 37°C before inoculating a 2-liter baffled flask containing 1 liter of Terrific Broth. Cultures were incubated at 37°C to an OD_{600} of 0.6, then the temperature was adjusted to 23°C for 20 min. Protein expression was then induced by the addition of 0.5 mM IPTG, and cultures were allowed to grow for an additional 24 h before harvest by centrifugation. Cells were then lysed by sonication, and the protein purified as previously described via nickel affinity chromatography [107].

Crystallization and X-ray Structure Determination

Crystallization experiments were performed using a Crystal Gryphon (Art Robbins) in 96-well trays using a sitting drop format. Diffraction quality crystals of native Sca5Coh5 were obtained by mixing 35 mg/ml protein 1:1 (vol/vol) with the crystallization solution containing 0.1 M sodium cacodylate; 0.2 M zinc chloride, 18% PEG 8000. Native Sca5Coh5 crystals were soaked for 30 minutes in the well solution with 200mM sodium iodide. The crystals were then swiped across the mother liquor solution supplemented with 200mM sodium iodide and 20% ethylene glycol and plunged directly into liquid nitrogen for x-ray data collection.

Diffraction quality crystals of native Sca5Coh6 were obtained by mixing 45 mg/ml protein 1:1 (vol/vol) with the crystallization solution containing PEG 20000, 0.06M calcium chloride, 0.06M magnesium chloride, 0.1M imidazole, 0.1M 2-(N-morpholino)ethanesulfonic acid pH=6.5. Prior to data collection, crystals were cryoprotected by a swiping through a solution of 80% mother liquor supplemented with 20% ethylene glycol then plunged into liquid nitrogen. For phasing, data was collected on crystals that were soaked in the well solution supplemented with 200mM sodium iodide, but native crystals showed data with more completeness and higher resolution, therefore these data were used for the final model refinement.

X-ray data from crystals were collected at the Life Sciences Collaborative Access Team (LSCAT) beamline ID-D of the Advanced Photon Source at Argonne National Laboratory. Both crystal structures were determined via single anomalous dispersion (SAD) of incorporated iodine atoms. Multiple data sets were processed and merged with Xia2. The location of iodine and initial protein model building was performed with AutoSol in Phenix[258, 259]. The Sca5Coh6 native structure was phased by molecular replacement with the iodide soaked Sca5Coh6 dataset. The protein models were finalized via alternating cycles of manual model building in Coot and refinement in Phenix.refine and/or Refmac5 from the CCP4 suite [168, 235, 236].

AlphaFold-Multimer Model Building

Residues 734-893 from Sca5 (L2-63_01064) and 556-629 from Sas20 (L2-63_00125) were input into AlphaFold-Multimer [249].

CHAPTER 5:

Implications, Future Directions, and Conclusions

Summary of Results

The main goal of this thesis work is to better understand the molecular basis of starch recognition by the human gut symbiont *Ruminococcus bromii*. While *R. bromii* encodes over 30 proteins believed to be involved in this process, my work examined the specific roles of three proteins: Sas20, Sca5 and Amy5. Sas20 and Sca5 are putative members of the amylosome, a multiprotein extracellular complex that likely imparts *R. bromii* with the ability to degrade resistant starch (RS) [42, 117]. Amy5 is an extracellular amylase that is not predicted to be incorporated into the amylosome but has a prolific ability to degrade starch [124]. All three proteins likely contribute to the full digestion of RS in the human gut by *R. bromii*.

Amy5 is a maltogenic α-amylase that displays its highest catalytic efficiency on amylose, in contrast to the *Eubacterium rectale* maltogenic α-amylase Amy13B, that is most efficient on maltohexaose (Chapter Two). Both enzymes are part of the same glycoside hydrolase subfamily (GH13_36) via sequence homology, and therefore we expected to observe similar preferences from both enzymes for longer α-glucan polysaccharides. We solved the crystal structures of Amy13B and Amy5 with the hopes of understanding the structural basis of this substrate bias. Alignment of the active sites from Amy13B and Amy5 reveal their active sites have high structural homology and neither Amy13B nor Amy5 bind starch beyond the active site (e.g., lacking surface binding sites or a carbohydrate-binding module). However, a deeper look at the secondary coordination sphere around the +1/+2 subsite in the catalytic cleft of Amy13B revealed that a loop having an amino acid sequence with longer side chains (QQD) compared to the same loop in Amy5 (TST) may limit the binding of longer α-glucans in Amy13B. This demonstrates that biologically significant changes in activity may manifest from subtle differences in sequence which only detailed structure-function or biochemical characterization studies can elucidate.

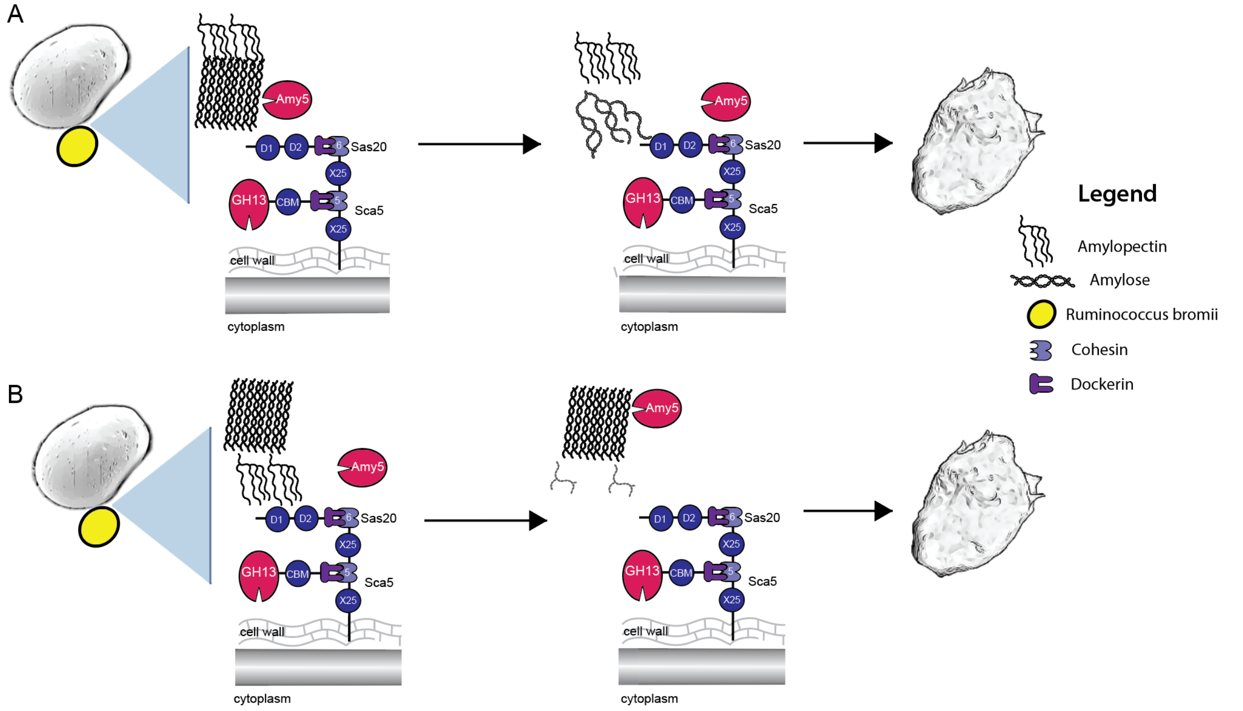


Figure 5.1: Two models of the amylosome system and Amy5 cooperation for starch granule degradation. A) On the surface of the starch granule, Amy5 expands pores by hydrolyzing linkages within amylose motifs. This opens the glycan architecture so that Sas20 and Sca5 of the amylosome system can bind and its associated GH13 enzymes may further degrade the granule. B) Alternatively, amylosome components such as Sas20 and Sca5 anchor the cell surface to the starch granule and its associated GH13 enzymes begin starch digestion. This then allows access to amylose by Amy5.

Together, this means that while Amy13B and Amy5 have similar structures, Amy5 can access amylose regions of RS granules that the amylosome system may have less access to or be inefficient at degrading. While we believe the amylosome system allows *R. bromii* to utilize RS degradation, complete breakdown of RS granules by amylosome components are not required for *R. bromii* growth on this substrate *in vitro* or in humans [37, 39]. One potential reason for the incomplete digestion of RS granules is amylosome complex size. Compared to a single enzyme such as Amy5, an amylosome complex may be made up of multiple scaffoldins and dockerin-containing proteins. While this may give these complexes the ability to degrade a diverse range of motifs within starch (α1,4 and α1,6 linkages, chain ends, amorphous regions of starch, etc), assembling multiple proteins may sterically hinder access to regions within the starch granule.

Because Amy5 lacks a cohesin or dockerin module, protein interaction modules believed to drive amylosome assembly, it may be a freely secreted enzyme that is able to penetrate and expand pores on the surface of the starch granule, priming the granule for hydrolysis by the amylosome system (**Figure 5.1A**). Amy5 may expose glycan motifs that can be recognized by the amylosome. Alternatively, this cooperation could happen inversely, meaning amylosome components could initiate binding and degradation of the starch granule, uncovering amylose motifs for which Amy5 has enzymatic preference (**Figure 5.1B**). It is likely that both models of Amy5 and amylosome synergy occur in the human gut simultaneously. However, it will be difficult to discretely test these until we can either genetically manipulate *R. bromii*, or accurately reconstitute amylosomes *in vitro*.

Sas20 is a non-catalytic starch-binding protein that is incorporated into the amylosome via its C-terminal dockerin module. Sas20 has two starch-binding domains that may help direct the amylosome to less helical and therefore more accessible or soluble regions within the starch granule. The N-terminus of Sas20 domain 1 (Sas20d1) is structurally homologous to the well-characterized CBM26 family, with a bundle of α-helices at its C-terminus that aids in substrate binding. The starch-binding profile of the Sas20 CBM26-like module differs from other CBM26 family members in that it does not bind to cyclic maltooligosaccharides like β-cyclodextrin [77]. These cyclic ligands are often used as a proxy for the helical curves found in crystalline amylose. It does, however, bind very tightly to longer maltooligosaccharides (seven glucose residues and longer), which may be more structurally similar to unwound amorphous regions of the starch granule. Additionally, we discovered that it likely has a structural, but not biochemical, preference for the non-reducing ends of glycan chains. While this domain does not require the free 4' hydroxyl group found at the non-reducing end, we believe this domain is particularly apt at binding to these chain ends enriched at the starch granule surface. Isothermal depletion experiments corroborate this as Sas20d1 had the highest B_{max} amongst the three proteins tested, Sas20d1, Sas20 domain 2 (Sas20d2), and both domains together (Sas20d1-2), for insoluble corn starch. We believe that Sas20d1 binds to more places on the corn granule surface compared to the Sas20d1-2 construct because the latter is bulkier, and the two domains may compete rather than synergize in binding to starch granules.

Like Sas20d1, Sas20d2 also prefers longer (seven glucose residues or more) maltooligosaccharides. The structure of Sca5X25-2, a close sequence homolog to Sas20d2, reveals a unique bilobed structure of starch-binding modules that coordinate binding to one maltooligosaccharide chain with helical characteristics. We therefore found that the motifs within starch granules that Sas20 binds are non-reducing ends and long stretches of loosely packed, unwound amorphous regions of glucan chains. These regions of starch are vulnerable to enzymatic attack as starch-active enzymes can more readily access them compared to tightly packed amylose helices. Additionally, small-angle x-ray scattering (SAXS) experiments revealed that Sas20 is a flexible protein that can adopt multiple conformations in solution. This is likely advantageous to the amylosome system as it can then contort its structure and therefore better attach to vulnerable glycosidic linkages exposed on the granule surface.

Sas20 is assembled into the amylosome via its C-terminal dockerin module which binds to Sca5Coh6, the second cohesin module of Sca5. We then wanted to compare the structures of the two cohesin modules in Sca5 to better understand this interaction. The crystal structures of Sca5Coh5 and Sca5Coh6 are similar and resemble a typical cohesin module jelly-roll fold. The Sca5Coh6 structure is most similar to the cohesin from the *Rf*CohScaB5-DocScaA complex structure from *Ruminococcus flavefaciens*. AlphaFold-Multimer modelling prediction of the Sca5Coh6-Sas20doc complex shows that both hydrophobic interactions and hydrogen bonding may be important for this interaction. Unfortunately, we could not quantify the affinity of this complex. This information is important for understanding amylosome assembly since if Sca5 and Sas20 readily disassociate, this might allow greater turnover within the amylosome as other dockerin-containing proteins could bind and contribute their activity against starch. Conversely, if this interaction does not readily disassociate, it might be more essential in shaping amylosome architecture.

In summary, my studies support a model whereby Sas20 docks to the cell-surface scaffoldin Sca5, and directs amylosome proteins to regions of starch granules that are most vulnerable to catalytic attack. A GH13_36 enzyme that is not incorporated into the amylosome system like Amy5 is advantageous to *R. bromii* as it can trim amylose that is blocking efficient degradation of RS by the amylosome or hydrolyze exposed amylose after degradation by the amylosome.

Regulation of Amylosome components

It is tempting to extrapolate the function and regulation of amylosomes from what we know about cellulosomes, cellulose-degrading enzyme complexes from soil and rumen bacteria. In the presence of cellulose or similar substrates, these bacteria upregulate many of the genes encoding cellulosome components [260]. Much of the work on cellulosome assembly has come from biochemical studies on cellulosomes that have been shed or released from cells late in stationary phase. One method for cellulosome isolation is to concentrate the stationary phase spent media from the monoculture of a cellulose-degrading bacterium, separate this consortium of proteins via size exclusion chromatography and test fractions for activity against cellulose [261]. The amylosome system seems to differ from cellulosomes both in the expression of the individual genes and in secretion of the complexes. Genes encoding amylosome proteins seem to be constitutively expressed across different growth conditions, including soluble or resistant starch and fructose [117, 229]. My data support these findings as in Chapter Three, I found that there were similar levels of amylosome proteins eluted from cells grown in fructose or autoclaved potato amylopectin. Furthermore, while it cannot be concluded that amylosomes are not shed or released from the cell during growth, my attempts to harvest amylosomes, as detailed in a later section, suggested they are not abundant in spent growth media.

Polysaccharide utilization loci (see Chapter One) have been observed in Gram-positive bacteria and are thusly named Gram-positive polysaccharide utilization loci (gpPUL) with notable cases in the human gut bacterium *E. rectale* [80, 262]. The definition of a gpPUL is a locus encoding, at minimum, one polysaccharide-degrading enzyme, a carbohydrate transport system, and a transcriptional regulator [101]. gpPULs also tend to be upregulated when the cell encounters the target substrate. The gene neighborhood of *sas20* (RBR_01410) includes genes that may be essential for cell growth and maintenance such as tryptophan turnover (RBR_01420), lysine synthesis (RBR_01430), and an anaerobic thioredoxin reductase (RBR_01510). Since it is likely that *sas20* is always "on", it makes sense that it is near housekeeping genes. This is not unprecedented as some cellulosome components are constitutively expressed [263, 264]. The genes encoding the amylosome proteins Doc22, Doc14, and Amy1 are in the same gene neighborhood, but since there is no import or regulation machinery, it cannot be called a gpPUL. Upon reviewing the other amylosome genes, there are no apparent starch or maltose-active genes nearby in the genome which suggests that the amylosome system does not rely on gpPULs to co-express starch-active genes.

The *R. bromii* starch-active protein repertoire extends beyond the amylosome system. Amy5 (RBR_07800) appears to be a part of a gpPUL. RBR_07790 is a predicted maltose permease, RBR_07780 is a predicted carbohydrate ABC transporter, and RBR_07770 is a predicted maltose-binding protein. These genes within this gpPUL are not part of the amylosome complex as they have no cohesin or dockerin modules in their predicted sequence, but they are co-expressed [229]. Contrastingly the Amy13B gpPUL in *E. rectale* is upregulated with exposure to maltose [107]. While *E. rectale* has the general strategy of conserving protein synthesis to utilize polysaccharides that it detects in its environment, *R. bromii* seems to constitutively express its starch-active proteins and benefits in the presence of starch that has escaped digestion by other gut bacteria. This reflects a fundamental difference between *E. rectale* and *R. bromii*. While *E. rectale* is more of a generalist that can grow on non-starch polysaccharides like arabinoxylan and has a diverse glycoside hydrolase repertoire, *R. bromii* is a starch specialist that has fewer glycoside hydrolases that all target starch [37, 101, 117].

Looking Beyond CBMs and GH13s

There are many genes in the *R. bromii* L2-63 genome that encode for predicted GH13 and CBMs with appended cohesin and dockerin modules, and these starch-active proteins are conserved among the five reported strains of *R. bromii* [42]. Computational analysis of sequences is extremely useful as it can be high-throughput and bypass the often laborious task of cloning, expressing, and purifying a protein of interest. However, determining function from sequence homology is not always feasible as many cohesin and dockerin-appended proteins from *R. bromii* have sequences with unknown function. Additionally, sequence analysis can bias our understanding of the system we are studying. For example, Amy5 and Amy13B from Chapter Two both binned into the same GH13 subfamily, so it is tempting to assume that they would have similar activity profiles, but our functional assays do not support this. Without our structure-function studies of these two enzymes, we would have been limited in our understanding how Amy13B contributes to the starch-degradation potential of *E. rectale* since Amy13B has an atypical GH13_36 activity preference. This is a particularly important problem when dissecting the differences between the molecular details of starch-active proteins expressed by bacteria that can and cannot utilize RS.

There are at least 10 genes in the *R. bromii* genome encoding proteins with dockerin domains that have little to no known sequence-based predictions, so their role in the amylosome system is currently unknown. Biochemical characterization of all putative amylosome proteins is important as it is only these molecular details that can uncover the full mechanism of RS degradation by *R. bromii*. A great example of how biochemical characterization of a dockerin-containing protein of unknown function can lead to valuable insights is the work presented in Chapter Three. We started with the reasonable assumption that this protein would bind starch based on its assembly into the amylosome and the homology of its second domain to X25 modules. My work revealed a new structural context for the CBM26-like and X25 modules that impart unprecedented function for these folds.

There are also genes within the *R. bromii* genome with putative dockerin modules, annotated as Doc#, that have predicted functions outside of the direct binding and degradation of starch. Four predicted Docs have leucine-rich repeats (LRRs). While there is no universal function for this motif, many have been characterized to be involved in protein-protein interactions [265-267]. Although the amylosome is presumed to assemble through cohesin-dockerin interactions, this does not exclude the possibility that proteins are incorporated into the amylosome via alternative strategies. Doc14 and Doc16 encode putative cysteine peptidases, and Doc17 encodes a putative serine peptidase [268, 269]. These peptidases could be important for turnover of the amylosome proteins themselves. The exchange of enzymes in the cellulosome system promotes fine-tuning of the components based on nutrient conditions for efficient fiber breakdown [256, 270]. Degradation of amylosome components could facilitate the exchange of dockerin-containing proteins with different functions in starch-binding or hydrolysis. Depending on the abundance of RS in the environment, this could be advantageous by allowing *R. bromii* to access a broader range of glycan motifs on the starch surface.

Doc19 has a predicted cysteine/histidine-dependent amidohydrolase/peptidase (CHAP) domain which has been associated with peptidoglycan trimming/hydrolysis [271]. *Ruminococcus champanellensis* is a human gut isolate that expresses a cellulosome with a glycoside hydrolase family 25 (GH25) enzyme with hydrolytic activity against peptidoglycan [272]. Interestingly, this GH25 does not target its own cell wall peptidoglycan but does inhibit growth of common bacterial competitors in the human gut. Likewise, it is plausible that the role of Doc19 in the amylosome may be to repel or inhibit the growth of its starch-utilizing competition while not affecting its own cell wall. In conclusion, there are likely many supporting players in the amylosome whose roles are auxiliary but may be important for competing with other microbes and/or for the degradation of RS.

Experimental Shortcomings and Future Directions

Isolation of Amylosomes

Isolating native cellulosomes was key to understanding the biology of how this system contributed to the complete breakdown of cellulose. Upon reviewing the cellulosome literature, I used some of the methodologies to isolate amylosomes that may be expressed by *R. bromii*. To understand the different amylosome complexes that are assembled on the cell surface and secreted, it is critical to isolate them from cells to determine the identities of the components therein.

Many cellulosome-producing bacteria shed these complexes from the cell surface when the culture reaches stationary phase [273]. Many studies have taken advantage of this feature to isolate and study cellulosomes *in vitro* [274-276]. However, in a study by Ze et al that compared the cell-associated or secreted amylolytic activity of *R. bromii* culture [117], the authors observed that most amylolytic activity (~70%) remained in the cell pellet at both mid-exponential phase and stationary phase which suggests that most amylosomes are not released into the supernatant like cellulosomes. To study cell wall/membrane proteins from *R. bromii*, I attempted to adapt the methodology previously used to isolate cell wall/membrane proteins from *E. rectale* [107]. Briefly, I cultured *R. bromii* cells with autoclaved potato amylopectin as the sole carbon source and harvested the cell pellet at stationary phase by centrifugation. Cells were lysed using a French press, unbroken cells pelleted, and the cell wall/membrane proteins were harvested by ultracentrifugation at 200k g for 2 hrs. While there was a smear of proteins visible via Coomassie-stained SDS-PAGE, I was unable to detect amylosome components in this fraction by western blot. Ze et al showed that ~30% of the amylolytic activity was found to be in the spent media of the *R. bromii* monoculture [117]. I attempted to isolate the cell-free amylosomes responsible for this amylolytic activity. I isolated, filtered, and concentrated the cell-free culture supernatant to isolate shed amylosomes. Unfortunately, I did not detect amylolytic activity from the input sample or of fractions of supernatant that eluted off a size exclusion column.

As an alternative approach, I attempted substrate affinity isolation [274] to isolate amylosomes. Briefly, I cultured *R. bromii* and incubated the spent media with RS granules at 4° C overnight, a temperature at which GH13 enzymes are typically less efficient. After incubation, the RS granules were washed and left in a series of buffers with a pH ranging from 5-8 at 37° C, conditions that facilitate starch hydrolysis [277]. Unfortunately, after an overnight incubation at 37° C, there was no significant degradation of insoluble starch granules. This was surprising as cellulose suspension is cleared after 2-4 hours of degradation, and amylase reactions are quicker than cellulase breakdown reactions [274]. Fendri et al [276] incubated spent media containing cellulosomes at 4° C overnight then eluted cellulosome components with incubation of cellulose in water. However, when I subjected the RS granules that had been incubated with amylosome-containing spent media at 4° C overnight, I did not observe any proteins eluted with water via Coomassie-stained SDS PAGE.

There are many reasons why these experiments may have failed. Further troubleshooting of the experimental conditions such as the speed and time of ultracentrifugation in the cell wall membrane fractionation, the quantity of cells used, and buffer composition for each method could be key. Additionally, all these experiments were conducted *in vitro* in monoculture. Perhaps the presence of other amylolytic gut bacteria may boost the RS degradation potential of *R. bromii* by attacking starch granules in tandem. One way to model how starch degradation cooperation among multiple bacterial species may occur in the gut would be to supplement amylosome-containing *R. bromii* cell-surface proteins or spent media with GH13 enzymes expressed by other gut bacterial species. Even so, *R. bromii* may not be able to completely digest RS granules within the time feces remains in the large intestine. Nevertheless, the incomplete breakdown of RS granules observed in my experiments may highlight that complete breakdown of the RS granule is not required for *R. bromii* to grow or crossfeed other butyrogenic bacteria [278].

Linkers of the Amylosome

All predicted GH13-containing members of the amylosome system have 5-50 amino acid linkers separating their dockerin modules from other starch-active domains on the polypeptide. Linkers between functional domains and dockerin modules may be flexible or static [279]. Flexible linkers allow their connected domains to adjust for the global geometric requirements of the substrate [222]. Furthermore, this gives protein with flexible linkers the ability to survey the surrounding space and “grab” binding partners and substrate [280]. For some cohesin-dockerin pairs, linkers contract upon binding which decreases their flexibility [222]. These linkers are more resistant to proteolytic cleave and provide additional stability for critical cohesin-dockerin interactions such as those that anchor scaffoldins to the cell surface [281-283]. The contraction of the linkers upon cohesin-dockerin binding may also be important in bringing enzymes closer to the substrate for efficient degradation [222].

It is important to understand the dynamics of the intra-domain linker in Sas20 and its impact on starch binding in the amylosome. As a future direction, I would perform small angle x-ray scattering (SAXS) to discern the functional arrangement of the domains of Sas20 with and without Sca5Coh6 bound [24]. These data may reveal the arrangement of the Sas20 domains within the amylosome. I would also test the functional effects of Sas20-Sca5 binding by measuring the affinity of the complex for starch substrates. While my domain-specific structure-function studies were important in identifying the mechanistic details of starch-binding by Sas20, it is important to put these findings in the context of the multi-protein network of the amylosome. In cellulosomes, we know that there are key enzyme-CBM interactions, facilitated by cohesin-dockerin modules, that allow for cellulolytic synergy [118]. Comparing binding affinities of Sas20, Sca5, and its complex will allow us to understand the functional relevance of Sas20-Sca5 binding and reveal an important part of the mechanism by which a gut symbiont like *R. bromii* accesses resistant starch.

Bibliography

1. Mazmanian, S.K., et al., *An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system.* Cell, 2005. **122**(1): p. 107-18.
2. Round, J.L. and S.K. Mazmanian, *The gut microbiota shapes intestinal immune responses during health and disease.* Nat Rev Immunol, 2009. **9**(5): p. 313-23.
3. Stappenbeck, T.S., L.V. Hooper, and J.I. Gordon, *Developmental regulation of intestinal angiogenesis by indigenous microbes via Paneth cells.* Proc Natl Acad Sci U S A, 2002. **99**(24): p. 15451-5.
4. Hooper, L.V., et al., *A molecular sensor that allows a gut commensal to control its nutrient foundation in a competitive ecosystem.* Proc Natl Acad Sci U S A, 1999. **96**(17): p. 9833-8.
5. Schirbel, A., et al., *Pro-angiogenic activity of TLRs and NLRs: a novel link between gut microbiota and intestinal angiogenesis.* Gastroenterology, 2013. **144**(3): p. 613-623 e9.
6. Cameron, E.A. and V. Sperandio, *Frenemies: Signaling and nutritional integration in pathogen-microbiota-host interactions.* Cell Host Microbe, 2015. **18**(3): p. 275-84.
7. Kamada, N., et al., *Control of pathogens and pathobionts by the gut microbiota.* Nat Immunol, 2013. **14**(7): p. 685-90.
8. Backhed, F., et al., *Host-bacterial mutualism in the human intestine.* Science, 2005. **307**(5717): p. 1915-20.
9. Salyers, A.A., et al., *Fermentation of mucins and plant polysaccharides by anaerobic bacteria from the human colon.* Appl Environ Microbiol, 1977. **34**(5): p. 529-33.
10. Flint, H.J., et al., *Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis.* Nat Rev Microbiol, 2008. **6**(2): p. 121-31.
11. Koenig, J.E., et al., *Succession of microbial consortia in the developing infant gut microbiome.* Proc Natl Acad Sci U S A, 2010. **108 Suppl 1**: p. 4578-85.
12. Collado, M.C., et al., *Microbial ecology and host-microbiota interactions during early life stages.* Gut Microbes, 2012. **3**(4): p. 352-65.
13. Koropatkin, N.M., E.A. Cameron, and E.C. Martens, *How glycan metabolism shapes the human gut microbiota.* Nat Rev Microbiol, 2012. **10**(5): p. 323-35.
14. Backhed, F., et al., *Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life.* Cell Host Microbe, 2015. **17**(5): p. 690-703.
15. Laursen, M.F., et al., *Infant gut microbiota development is driven by transition to family foods independent of maternal obesity.* mSphere, 2016. **1**(1): p. pii: e00069-15.
16. Quann, E. and R. Carvalho, *Starch consumption patterns in infants and young children.* J Pediatr Gastroenterol Nutr, 2018. **66 Suppl 3**: p. S39-S41.

17. Cordain, L., et al., *Origins and evolution of the Western diet: health implications for the 21st century.* Am J Clin Nutr, 2005. **81**(2): p. 341-54.
18. Buleon, A., et al., *Starch granules: structure and biosynthesis.* Int J Biol Macromol, 1998. **23**(2): p. 85-112.
19. Jane, J., *Current understanding on starch granule structures.* J Appl Glycosci, 2006. **53**: p. 205-213.
20. El Kaoutari, A., et al., *The abundance and variety of carbohydrate-active enzymes in the human gut microbiota.* Nat Rev Microbiol, 2013. **11**(7): p. 497-504.
21. Brownlee, I.A., et al., *Starch digestion in the upper gastrointestinal tract of humans.* Starch, 2018. **70**.
22. Englyst, H.N. and J.H. Cummings, *Digestion of polysaccharides of potato in the small intestine of man.* Am J Clin Nutr, 1987. **45**(2): p. 423-31.
23. Englyst, H.N., S.M. Kingman, and J.H. Cummings, *Classification and measurement of nutritionally important starch fractions.* Eur J Clin Nutr, 1992. **46 Suppl 2**: p. S33-50.
24. Muir, J.G., et al., *Food processing and maize variety affects amounts of starch escaping digestion in the small intestine.* Am J Clin Nutr, 1995. **61**(1): p. 82-9.
25. Lee, B.H. and B.R. Hamaker, *Number of branch points in alpha-limit dextrins impact glucose generation rates by mammalian mucosal alpha-glucosidases.* Carbohydr Polym, 2017. **157**: p. 207-213.
26. Lee, B.H., et al., *Enzyme-synthesized highly branched maltodextrins have slow glucose generation at the mucosal alpha-glucosidase level and are slowly digestible in vivo.* PLoS One, 2013. **8**(4): p. e59745.
27. Heller, J. and M. Schramm, *α-Amylase limit dextrins of high molecular weight obtained from glycogen.* Biochimica et Biophysica Acta, 1964. **81**(1): p. 96-100.
28. Poole, A.C., et al., *Human salivary amylase gene copy number impacts oral and gut microbiomes.* Cell Host Microbe, 2019. **25**(4): p. 553-564 e7.
29. Stephen, A.M., A.C. Haddad, and S.F. Phillips, *Passage of carbohydrate into the colon. Direct measurements in humans.* Gastroenterology, 1983. **85**(3): p. 589-95.
30. Vonk, R.J., et al., *Digestion of so-called resistant starch sources in the human small intestine.* Am J Clin Nutr, 2000. **72**(2): p. 432-8.
31. Birt, D.F., et al., *Resistant starch: promise for improving human health.* Adv Nutr, 2013. **4**(6): p. 587-601.
32. Zhang, G. and B.R. Hamaker, *Slowly digestible starch: concept, mechanism, and proposed extended glycemic index.* Crit Rev Food Sci Nutr, 2009. **49**(10): p. 852-67.
33. Venkataraman, A., et al., *Variable responses of human microbiomes to dietary supplementation with resistant starch.* Microbiome, 2016. **4**(1): p. 33.
34. Maier, T.V., et al., *Impact of dietary resistant starch on the human gut microbiome, metaproteome, and metabolome.* MBio, 2017. **8**(5): p. pii: e01343-17.
35. Baxter, N.T., et al., *Dynamics of human gut microbiota and short-chain fatty acids in response to dietary interventions with three fermentable fibers.* MBio, 2019. **10:e02566-18**.
36. Martinez, I., et al., *Resistant starches types 2 and 4 have differential effects on the composition of the fecal microbiota in human subjects.* PLoS One, 2010. **5**(11): p. e15046.
37. Ze, X., et al., *Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon.* ISME J, 2012. **6**(8): p. 1535-43.

38. Herrmann, E., et al., *Determination of resistant starch assimilating bacteria in fecal samples of mice by in vitro RNA-based stable isotope probing.* Front Microbiol, 2017. **8**: p. 1331.
39. Walker, A.W., et al., *Dominant and diet-responsive groups of bacteria within the human colonic microbiota.* Isme J, 2011. **5**(2): p. 220-30.
40. Vital, M., et al., *Metagenomic insights into the degradation of resistant starch by human gut microbiota.* Appl Environ Microbiol, 2018. **84**(23): p. pii: e01562-18.
41. Abell, G.C., et al., *Phylotypes related to Ruminococcus bromii are abundant in the large bowel of humans and increase in response to a diet high in resistant starch.* FEMS Microbiol Ecol, 2008. **66**(3): p. 505-15.
42. Mukhopadhya, I., et al., *Sporulation capability and amylosome conservation among diverse human colonic and rumen isolates of the keystone starch-degrader Ruminococcus bromii.* Environ Microbiol, 2018. **20**(1): p. 324-336.
43. Wong, J.M., et al., *Colonic health: fermentation and short chain fatty acids.* J Clin Gastroenterol, 2006. **40**(3): p. 235-43.
44. Morrison, D.J. and T. Preston, *Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism.* Gut Microbes, 2016. **7**(3): p. 189-200.
45. McOrist, A.L., et al., *Fecal butyrate levels vary widely among individuals but are usually increased by a diet high in resistant starch.* J Nutr, 2011. **141**(5): p. 883-9.
46. Hamer, H.M., et al., *Review article: the role of butyrate on colonic function.* Aliment Pharmacol Ther, 2008. **27**(2): p. 104-19.
47. Roediger, W.E., *Utilization of nutrients by isolated epithelial cells of the rat colon.* Gastroenterology, 1982. **83**(2): p. 424-9.
48. Gantois, I., et al., *Butyrate specifically down-regulates salmonella pathogenicity island 1 gene expression.* Appl Environ Microbiol, 2006. **72**(1): p. 946-9.
49. Sun, Y., et al., *Fatty acids regulate stress resistance and virulence factor production for Listeria monocytogenes.* J Bacteriol, 2012. **194**(19): p. 5274-84.
50. Furusawa, Y., et al., *Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells.* Nature, 2013. **504**(7480): p. 446-50.
51. Clarke, J.M., et al., *Effects of high-amylose maize starch and butyrylated high-amylose maize starch on azoxymethane-induced intestinal cancer in rats.* Carcinogenesis, 2008. **29**(11): p. 2190-4.
52. Conlon, M.A., et al., *Resistant starches protect against colonic DNA damage and alter microbiota and gene expression in rats fed a Western diet.* J Nutr, 2012. **142**(5): p. 832-40.
53. Mathewson, N.D., et al., *Gut microbiome-derived metabolites modulate intestinal epithelial cell damage and mitigate graft-versus-host disease.* Nat Immunol, 2016. **17**(5): p. 505-13.
54. Moelands, S.V., et al., *Alpha-glucosidase inhibitors for prevention or delay of type 2 diabetes mellitus and its associated complications in people at increased risk of developing type 2 diabetes mellitus.* Cochrane Database Syst Rev, 2018. **12**: p. CD005061.
55. Weaver, G.A., et al., *Acarbose enhances human colonic butyrate production.* J Nutr, 1997. **127**(5): p. 717-23.
56. Holt, P.R., et al., *Effects of acarbose on fecal nutrients, colonic pH, and short-chain fatty acids and rectal proliferative indices.* Metabolism, 1996. **45**(9): p. 1179-87.

57. Wolever, T.M. and J.L. Chiasson, *Acarbose raises serum butyrate in human subjects with impaired glucose tolerance.* Br J Nutr, 2000. **84**(1): p. 57-61.
58. Su, B., et al., *Acarbose treatment affects the serum levels of inflammatory cytokines and the gut content of Bifidobacteria in Chinese patients with type 2 diabetes mellitus.* J Diabetes, 2015. **7**(5): p. 729-39.
59. Harrison, D.E., et al., *Acarbose, 17-alpha-estradiol, and nordihydroguaiaretic acid extend mouse lifespan preferentially in males.* Aging Cell, 2014. **13**(2): p. 273-82.
60. Smith, B.J., et al., *Changes in the gut microbiome and fermentation products concurrent with enhanced longevity in acarbose-treated mice.* BMC Microbiol, 2019. **19**(1): p. 130.
61. Baxter, N.T., et al., *The glucoamylase inhibitor acarbose has a diet-dependent and reversible effect on the murine gut microbiome.* mSphere, 2019. **4**(1): p. pii: e00347-19.
62. Santilli, A.D., et al., *Nonmicrobicidal small molecule inhibition of polysaccharide metabolism in human gut microbes: A potential therapeutic avenue.* ACS Chem Biol, 2018. **13**: p. 1165-1172.
63. Clarke, G., et al., *Gut reactions: Breaking down xenobiotic-microbiome interactions.* Pharmacol Rev, 2019. **71**(2): p. 198-224.
64. Bertoft, E., *Understanding Starch Structure: Recent Progress.* Agronomy, 2017. **7**(3): p. 56.
65. Cantarel, B.L., et al., *The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics.* Nucleic Acids Res, 2009. **37**(Database issue): p. D233-8.
66. Moller, M.S. and B. Svensson, *Structural biology of starch-degrading enzymes and their regulation.* Curr Opin Struct Biol, 2016. **40**: p. 33-42.
67. Stam, M.R., et al., *Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins.* Protein Eng Des Sel, 2006. **19**(12): p. 555-62.
68. Janecek, S., et al., *Starch-binding domains as CBM families-history, occurrence, structure, function and evolution.* Biotechnol Adv, 2019: p. 107451.
69. Kuchtova, A. and S. Janecek, *Domain evolution in enzymes of the neopullulanase subfamily.* Microbiology, 2016. **162**(12): p. 2099-2115.
70. Janecek, S., B. Svensson, and E.A. MacGregor, *alpha-Amylase: an enzyme specificity found in various families of glycoside hydrolases.* Cell Mol Life Sci, 2014. **71**(7): p. 1149-70.
71. Moller, M.S., A. Henriksen, and B. Svensson, *Structure and function of alpha-glucan debranching enzymes.* Cell Mol Life Sci, 2016. **73**(14): p. 2619-41.
72. Janecek, S., B. Svensson, and B. Henrissat, *Domain evolution in the alpha-amylase family.* J Mol Evol, 1997. **45**(3): p. 322-31.
73. van der Maarel, M.J., et al., *Properties and applications of starch-converting enzymes of the alpha-amylase family.* J Biotechnol, 2002. **94**(2): p. 137-55.
74. Cockburn, D. and N.M. Koropatkin, *Polysaccharide degradation by the intestinal microbiota and its influence on human health and disease.* J Mol Biol, 2016. **428**(16).
75. Abbott, D.W., et al., *The molecular basis of glycogen breakdown and transport in Streptococcus pneumoniae.* Mol Microbiol, 2010. **77**(1): p. 183-99.
76. Juge, N., et al., *The activity of barley alpha-amylase on starch granules is enhanced by fusion of a starch binding domain from Aspergillus niger glucoamylase.* Biochim Biophys Acta, 2006. **1764**(2): p. 275-84.

77. Boraston, A.B., et al., *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, 2006. **281**(1): p. 587-98.
78. van Bueren, A.L., et al., *Identification and structural basis of binding to host lung glycogen by streptococcal virulence factors.* Nat Struct Mol Biol, 2007. **14**(1): p. 76-84.
79. Valk, V., et al., *Carbohydrate-binding module 74 is a novel starch-binding domain associated with large and multidomain alpha-amylase enzymes.* FEBS J, 2016. **283**(12): p. 2354-68.
80. Cockburn, D.W., et al., *Novel carbohydrate binding modules in the surface anchored alpha-amylase of Eubacterium rectale provide a molecular rationale for the range of starches used by this organism in the human gut.* Mol Microbiol, 2018. **107**(2): p. 249-264.
81. Boraston, A.B., et al., *Carbohydrate-binding modules: fine-tuning polysaccharide recognition.* Biochem J, 2004. **382**(Pt 3): p. 769-81.
82. Guillen, D., S. Sanchez, and R. Rodriguez-Sanoja, *Carbohydrate-binding domains: multiplicity of biological roles.* Appl Microbiol Biotechnol, 2010. **85**(5): p. 1241-9.
83. Southall, S.M., et al., *The starch-binding domain from glucoamylase disrupts the structure of starch.* FEBS Lett, 1999. **447**(1): p. 58-60.
84. Lammerts van Bueren, A. and A.B. Boraston, *The structural basis of alpha-glucan recognition by a family 41 carbohydrate-binding module from Thermotoga maritima.* J Mol Biol, 2007. **365**(3): p. 555-60.
85. Koropatkin, N.M. and T.J. Smith, *SusG: A unique cell-membrane-associated alpha-amylase from a prominent human gut symbiont targets complex starch molecules.* Structure, 2010. **18**(2): p. 200-215.
86. Grondin, J.M., et al., *Polysaccharide Utilization Loci: Fueling microbial communities.* J Bacteriol, 2017. **199**(15): p. pii: e00860-16.
87. Martens, E.C., et al., *Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm.* J Biol Chem, 2009. **284**(37): p. 24673-7.
88. Tauzin, A.S., et al., *Molecular dissection of xyloglucan recognition in a prominent human gut symbiont* MBio, 2016. **7**(2): p. e02134-15.
89. Foley, M.H., E.C. Martens, and N.M. Koropatkin, *SusE facilitates starch uptake independent of starch binding in B. thetaiotaomicron.* Mol Microbiol, 2018. **108**(5): p. 551-566.
90. Tamura, K., et al., *Surface glycan-binding proteins are essential for cereal beta-glucan utilization by the human gut symbiont Bacteroides ovatus.* Cell Mol Life Sci, 2019.
91. Anderson, K.L. and A.A. Salyers, *Genetic evidence that outer membrane binding of starch is required for starch utilization by Bacteroides thetaiotaomicron.* J Bacteriol, 1989. **171**(6): p. 3199-204.
92. Anderson, K.L. and A.A. Salyers, *Biochemical evidence that starch breakdown by Bacteroides thetaiotaomicron involves outer membrane starch-binding sites and periplasmic starch-degrading enzymes.* J Bacteriol, 1989. **171**(6): p. 3192-8.
93. D'Elia, J.N. and A.A. Salyers, *Effect of regulatory protein levels on utilization of starch by Bacteroides thetaiotaomicron.* J Bacteriol, 1996. **178**(24): p. 7180-6.
94. Cameron, E.A., et al., *Multifunctional nutrient-binding proteins adapt human symbiotic bacteria for glycan competition in the gut by separately promoting enhanced sensing and catalysis.* MBio, 2014. **5**(5): p. e01441-14.

95. D'Elia, J.N. and A.A. Salyers, *Contribution of a neopullulanase, a pullulanase, and an alpha-glucosidase to growth of Bacteroides thetaiotaomicron on starch.* J Bacteriol, 1996. **178**(24): p. 7173-9.

96. Arnal, G., et al., *Structural basis for the flexible recognition of alpha-glucan substrates by Bacteroides thetaiotaomicron SusG.* Protein Sci, 2018. **27**(6): p. 1093-1101.

97. Foley, M.H., D.W. Cockburn, and N.M. Koropatkin, *The Sus operon: a model system for starch uptake by the human gut Bacteroidetes.* Cell Mol Life Sci, 2016. **73**(14): p. 2603-17.

98. Chaudet, M.M. and D.R. Rose, *Suggested alternative starch utilization system from the human gut bacterium Bacteroides thetaiotaomicron.* Biochem Cell Biol, 2016. **94**(3): p. 241-6.

99. Herrmann, E., et al., *In vivo assessment of resistant starch degradation by the caecal microbiota of mice using RNA-based stable isotope probing-a proof-of-principle study.* Nutrients, 2018. **10**(2).

100. Upadhyaya, B., et al., *Impact of dietary resistant starch type 4 on human gut microbiota and immunometabolic functions.* Sci Rep, 2016. **6**: p. 28797.

101. Sheridan, P.O., et al., *Polysaccharide utilization loci and nutritional specialization in a dominant group of butyrate-producing human colonic Firmicutes.* Microb Genom, 2016. **2**(2): p. e000043.

102. Kovatcheva-Datchary, P., et al., *Linking phylogenetic identities of bacteria to starch fermentation in an in vitro model of the large intestine by RNA-based stable isotope probing.* Environ Microbiol, 2009. **11**(4): p. 914-26.

103. Leitch, E.C., et al., *Selective colonization of insoluble substrates by human faecal bacteria.* Environ Microbiol, 2007. **9**(3): p. 667-79.

104. Walker, A.W., et al., *The species composition of the human intestinal microbiota differs between particle-associated and liquid phase communities.* Environ Microbiol, 2008. **10**(12): p. 3275-83.

105. Duncan, S.H., et al., *Effects of alternative dietary substrates on competition between human colonic bacteria in an anaerobic fermentor system.* Appl Environ Microbiol, 2003. **69**(2): p. 1136-42.

106. Scott, K.P., et al., *Substrate-driven gene expression in Roseburia inulinivorans: importance of inducible enzymes in the utilization of inulin and starch.* Proc Natl Acad Sci U S A, 2011. **108 Suppl 1**: p. 4672-9.

107. Cockburn, D.W., et al., *Molecular details of a starch utilization pathway in the human gut symbiont Eubacterium rectale.* Mol Microbiol, 2015. **95**(2): p. 209-30.

108. Ramsay, A.G., et al., *Cell-associated alpha-amylases of butyrate-producing Firmicute bacteria from the human colon.* Microbiology, 2006. **152**(Pt 11): p. 3281-90.

109. Underwood, M.A., et al., *Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut.* Pediatr Res, 2015. **77**(1-2): p. 229-35.

110. Zuniga, M., V. Monedero, and M.J. Yebra, *Utilization of host-derived glycans by intestinal Lactobacillus and Bifidobacterium species.* Front Microbiol, 2018. **9**: p. 1917.

111. Ejby, M., et al., *An ATP binding cassette transporter mediates the uptake of alpha-(1,6)-linked dietary oligosaccharides in Bifidobacterium and correlates with competitive growth on these substrates.* J Biol Chem, 2016. **291**(38): p. 20220-31.

112. Crittenden, R.G., et al., *Selection of a Bifidobacterium strain to complement resistant starch in a synbiotic yoghurt.* J Appl Microbiol, 2001. **90**(2): p. 268-78.

113. Duranti, S., et al., *Genomic characterization and transcriptional studies of the starch-utilizing strain Bifidobacterium adolescentis 22L.* Appl Environ Microbiol, 2014. **80**(19): p. 6080-90.

114. Rodriguez Sanoja, R., et al., *Comparative characterization of complete and truncated forms of Lactobacillus amylovorus alpha-amylase and role of the C-terminal direct repeats in raw-starch binding.* Appl Environ Microbiol, 2000. **66**(8): p. 3350-6.

115. Ryan, S.M., G.F. Fitzgerald, and D. van Sinderen, *Screening for and identification of starch-, amylopectin-, and pullulan-degrading activities in bifidobacterial strains.* Appl Environ Microbiol, 2006. **72**(8): p. 5289-96.

116. O'Connell Motherway, M., et al., *Characterization of ApuB, an extracellular type II amylopullulanase from Bifidobacterium breve UCC2003.* Appl Environ Microbiol, 2008. **74**(20): p. 6271-9.

117. Ze, X., et al., *Unique organization of extracellular amylases into amylosomes in the resistant starch-utilizing human colonic firmicutes bacterium Ruminococcus bromii.* MBio, 2015. **6**(5): p. e01058-15.

118. Bayer, E.A., et al., *The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides.* Annu Rev Microbiol, 2004. **58**: p. 521-54.

119. Alber, O., et al., *Cohesin diversity revealed by the crystal structure of the anchoring cohesin from Ruminococcus flavefaciens.* Proteins, 2009. **77**(3): p. 699-709.

120. Nash, M.A., et al., *Single versus dual-binding conformations in cellulosomal cohesin-dockerin complexes.* Curr Opin Struct Biol, 2016. **40**: p. 89-96.

121. Stahl, S.W., et al., *Single-molecule dissection of the high-affinity cohesin-dockerin complex.* Proc Natl Acad Sci U S A, 2012. **109**(50): p. 20431-6.

122. Schaeffer, F., et al., *Duplicated dockerin subdomains of Clostridium thermocellum endoglucanase CelD bind to a cohesin domain of the scaffolding protein CipA with distinct thermodynamic parameters and a negative cooperativity.* Biochemistry, 2002. **41**(7): p. 2106-14.

123. Cockburn, D.W., et al., *Structure and substrate recognition by the Ruminococcus bromii amylosome pullulanases.* J Struct Biol, 2021. **213**(3): p. 107765.

124. Jung, J.H., et al., *Characterization of a novel extracellular alpha-amylase from Ruminococcus bromii ATCC 27255 with neopullulanase-like activity.* Int J Biol Macromol, 2019. **130**: p. 605-614.

125. Boos, W. and H. Shuman, *Maltose/maltodextrin system of Escherichia coli: transport, metabolism, and regulation.* Microbiol Mol Biol Rev, 1998. **62**(1): p. 204-29.

126. Ganzle, M.G. and R. Follador, *Metabolism of oligosaccharides and starch in lactobacilli: a review.* Front Microbiol, 2012. **3**: p. 340.

127. Kelly, W.J., R.V. Asmundson, and D.H. Hopcroft, *Isolation and characterization of a strictly anaerobic, cellulolytic spore former: Clostridium chartatabidum sp. nov.* Arch Microbiol, 1987. **147**(2): p. 169-73.

128. Pryde, S.E., et al., *The microbiology of butyrate formation in the human colon.* FEMS Microbiol Lett, 2002. **217**(2): p. 133-9.

129. Guilloteau, P., et al., *From the gut to the peripheral tissues: the multiple effects of butyrate.* Nutr Res Rev, 2010. **23**(2): p. 366-84.

130. Zeller, G., et al., *Potential of fecal microbiota for early-stage detection of colorectal cancer.* Mol Syst Biol, 2014. **10**(11): p. 766.

131. Knoll, R.L., et al., *Gut microbiota differs between children with Inflammatory Bowel Disease and healthy siblings in taxonomic and functional composition: a metagenomic analysis.* Am J Physiol Gastrointest Liver Physiol, 2017. **312**(4): p. G327-g339.

132. Murri, M., et al., *Gut microbiota in children with type 1 diabetes differs from that in healthy children: a case-control study.* BMC Med, 2013. **11**: p. 46.

133. Hondoh, H., T. Kuriki, and Y. Matsuura, *Three-dimensional structure and substrate binding of Bacillus stearothermophilus neopullulanase.* J Mol Biol, 2003. **326**(1): p. 177-88.

134. Todd, M.J. and J. Gomez, *Enzyme kinetics determined using calorimetry: a general assay for enzyme activity?* Anal Biochem, 2001. **296**(2): p. 179-87.

135. Olsen, S.N., *Applications of isothermal titration calorimetry to measure enzyme kinetics and activity in complex solutions.* Thermochimica Acta, 2006. **448**(1): p. 12-18.

136. Lonhienne, T., et al., *Enzyme activity determination on macromolecular substrates by isothermal titration calorimetry: application to mesophilic and psychrophilic chitinases.* Biochim Biophys Acta, 2001. **1545**(1-2): p. 349-56.

137. Goldberg, R.N., et al., *Thermodynamics of hydrolysis of oligosaccharides.* Biophys Chem, 1991. **40**(1): p. 69-76.

138. Krissinel, E., *Stock-based detection of protein oligomeric states in jsPISA.* Nucleic Acids Res, 2015. **43**(W1): p. W314-9.

139. Uitdehaag, J.C., et al., *Structures of maltohexaose and maltoheptaose bound at the donor sites of cyclodextrin glycosyltransferase give insight into the mechanisms of transglycosylation activity and cyclodextrin size specificity.* Biochemistry, 2000. **39**(26): p. 7772-80.

140. Matsuura, Y., et al., *Structure and possible catalytic residues of Taka-amylase A.* J Biochem, 1984. **95**(3): p. 697-702.

141. Kadziola, A., et al., *Crystal and molecular structure of barley alpha-amylase.* J Mol Biol, 1994. **239**(1): p. 104-21.

142. Sivakumar, N., et al., *Crystal structure of AmyA lacks acidic surface and provide insights into protein stability at poly-extreme condition.* FEBS Lett, 2006. **580**(11): p. 2646-52.

143. Majzlova, K., Z. Pukajova, and S. Janecek, *Tracing the evolution of the alpha-amylase subfamily GH13_36 covering the amylolytic enzymes intermediate between oligo-1,6-glucosidases and neopullulanases.* Carbohydr Res, 2013. **367**: p. 48-57.

144. Yun, J., et al., *Characterization of a novel amylolytic enzyme encoded by a gene from a soil-derived metagenomic library.* Appl Environ Microbiol, 2004. **70**(12): p. 7229-35.

145. Damián-Almazo, J.Y., et al., *Enhancement of the alcoholytic activity of alpha-amylase AmyA from Thermotoga maritima MSB8 (DSM 3109) by site-directed mutagenesis.* Appl Environ Microbiol, 2008. **74**(16): p. 5168-77.

146. Zhou, J., et al., *Novel Maltogenic Amylase CoMA from Corallococcus sp. Strain EGB Catalyzes the Conversion of Maltooligosaccharides and Soluble Starch to Maltose.* Appl Environ Microbiol, 2018. **84**(14).

147. Nakagawa, Y., et al., *Gene cloning and enzymatic characteristics of a novel gamma-cyclodextrin-specific cyclodextrinase from alkalophilic Bacillus clarkii 7364.* Biochim Biophys Acta, 2008. **1784**(12): p. 2004-11.

148. Ballschmiter, M., et al., *AmyA, an alpha-amylase with beta-cyclodextrin-forming activity, and AmyB from the thermoalkaliphilic organism Anaerobranca gottschalkii: two alpha-*

amylases adapted to their different cellular localizations. Appl Environ Microbiol, 2005. **71**(7): p. 3709-15.

149. Mijts, B.N. and B.K.C. Patel, *Cloning, sequencing and expression of an alpha-amylase gene, amyA, from the thermophilic halophile Halothermothrix orenii and purification and biochemical characterization of the recombinant enzyme.* Microbiology (Reading), 2002. **148**(Pt 8): p. 2343-2349.

150. Yebra, M.J., A. Blasco, and P. Sanz, *Expression and secretion of Bacillus polymyxa neopullulanase in Saccharomyces cerevisiae.* FEMS Microbiol Lett, 1999. **170**(1): p. 41-9.

151. Yebra, M.J., et al., *Characterization of novel neopullulanase fromBacillus polymyxa.* Applied Biochemistry and Biotechnology, 1997. **68**(1): p. 113-120.

152. Abe, J., et al., *Purification and characterization of periplasmic alpha-amylase from Xanthomonas campestris K-11151.* Journal of Bacteriology, 1994. **176**(12): p. 3584-3588.

153. Liebl, W., I. Stemplinger, and P. Ruile, *Properties and gene structure of the Thermotoga maritima alpha-amylase AmyA, a putative lipoprotein of a hyperthermophilic bacterium.* J Bacteriol, 1997. **179**(3): p. 941-8.

154. Oslancová, A. and S. Janecek, *Oligo-1,6-glucosidase and neopullulanase enzyme subfamilies from the alpha-amylase family defined by the fifth conserved sequence region.* Cell Mol Life Sci, 2002. **59**(11): p. 1945-59.

155. Machovic, M. and S. Janecek, *Starch-binding domains in the post-genome era.* Cell Mol Life Sci, 2006. **63**(23): p. 2710-24.

156. Lee, H.S., et al., *Cyclomaltodextrinase, neopullulanase, and maltogenic amylase are nearly indistinguishable from each other.* J Biol Chem, 2002. **277**(24): p. 21891-7.

157. Abe, A., et al., *Complexes of Thermoactinomyces vulgaris R-47 alpha-amylase 1 and pullulan model oligossacharides provide new insight into the mechanism for recognizing substrates with alpha-(1,6) glycosidic linkages.* FEBS J, 2005. **272**(23): p. 6145-53.

158. van der Maarel, M.J.E.C. and H. Leemhuis, *Starch modification with microbial alpha-glucanotransferase enzymes.* Carbohydrate Polymers, 2013. **93**(1): p. 116-121.

159. Cockburn, D.W. and B. Svensson, *Surface binding sites in carbohydrate active enzymes: and emerging picture of structural and functional diversity* in *Carbohydrate Chemistry: Chemical and Biological Approaches*, T.K. Lindhorst and A.P. Rauter, Editors. 2013, Royal Society of Chemistry: Cambridge. p. 204-221.

160. Cockburn, D., et al., *Analysis of surface binding sites (SBSs) in carbohydrate active enzymes with focus on glycoside hydrolase families 13 and 77 – a mini-review.* Biologia, 2014. **69**(6): p. 705-712

161. Cockburn, D., et al., *Surface binding sites in amylase have distinct roles in recognition of starch structure motifs and degradation.* Int J Biol Macromol, 2015. **75**: p. 338-45.

162. Webb, A.J., K.A. Homer, and A.H.F. Hosie, *Two closely related ABC transporters in Streptococcus mutans are involved in disaccharide and/or oligosaccharide uptake.* Journal of bacteriology, 2008. **190**(1): p. 168-178.

163. La Rosa, S.L., et al., *The human gut Firmicute Roseburia intestinalis is a primary degrader of dietary β-mannans.* Nature Communications, 2019. **10**(1): p. 905.

164. Foley, M.H., D. Cockburn, and N.M. Koropatkin, *The Sus operon – a model system for starch uptake by the human gut Bacteroidetes.* Cellular and Molecular Life Sciences, 2016.

165. Goldschmidt, L., et al., *Toward rational protein crystallization: A Web server for the design of crystallizable protein variants.* Protein Sci, 2007. **16**(8): p. 1569-76.

166. Winter, G., C.M. Lobley, and S.M. Prince, *Decision making in xia2.* Acta Crystallogr D Biol Crystallogr, 2013. **69**(Pt 7): p. 1260-73.

167. Kabsch, W., *XDS.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 125-32.

168. Winn, M.D., et al., *Overview of the CCP4 suite and current developments.* Acta Crystallogr D Biol Crystallogr, 2011. **67**(Pt 4): p. 235-42.

169. McCoy, A.J., et al., *Phaser crystallographic software.* J Appl Crystallogr, 2007. **40**(Pt 4): p. 658-674.

170. Adams, P.D., et al., *PHENIX: building new software for automated crystallographic structure determination.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 11): p. 1948-54.

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

172. Afonine, P.V., et al., *Towards automated crystallographic structure refinement with phenix.refine.* Acta Crystallogr D Biol Crystallogr, 2012. **68**(Pt 4): p. 352-67.

173. Murshudov, G.N., A.A. Vagin, and E.J. Dodson, *Refinement of macromolecular structures by the maximum-likelihood method.* Acta Crystallogr D Biol Crystallogr, 1997. **53**(Pt 3): p. 240-55.

174. Agirre, J., et al., *Privateer: software for the conformational validation of carbohydrate structures.* Nat Struct Mol Biol, 2015. **22**(11): p. 833-4.

175. Terwilliger, T.C., et al., *Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard.* Acta Crystallographica Section D, 2008. **64**(1): p. 61-69.

176. Almagro Armenteros, J.J., et al., *SignalP 5.0 improves signal peptide predictions using deep neural networks.* Nat Biotechnol, 2019. **37**(4): p. 420-423.

177. Yu, N.Y., et al., *PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.* Bioinformatics, 2010. **26**(13): p. 1608-15.

178. Madeira, F., et al., *The EMBL-EBI search and sequence analysis tools APIs in 2019.* Nucleic Acids Res, 2019. **47**(W1): p. W636-w641.

179. Sekirov, I., et al., *Gut microbiota in health and disease.* Physiol Rev, 2010. **90**(3): p. 859-904.

180. Shreiner, A.B., J.Y. Kao, and V.B. Young, *The gut microbiome in health and in disease.* Curr Opin Gastroenterol, 2015. **31**(1): p. 69-75.

181. Lam, Y.Y., C. Zhang, and L. Zhao, *Causality in dietary interventions-building a case for gut microbiota.* Genome Med, 2018. **10**(1): p. 62.

182. Cerqueira, F.M., et al., *Starch Digestion by Gut Bacteria: Crowdsourcing for Carbs.* Trends Microbiol, 2019.

183. DeMartino, P. and D.W. Cockburn, *Resistant starch: impact on the gut microbiome and health.* Curr Opin Biotechnol, 2020. **61**: p. 66-71.

184. Canani, R.B., et al., *Potential beneficial effects of butyrate in intestinal and extraintestinal diseases.* World J Gastroenterol, 2011. **17**(12): p. 1519-28.

185. Zaman, S.A. and S.R. Sarbini, *The potential of resistant starch as a prebiotic.* Crit Rev Biotechnol, 2016. **36**(3): p. 578-84.
186. Koh, A., et al., *From Dietary Fiber to Host Physiology: Short-Chain Fatty Acids as Key Bacterial Metabolites.* Cell, 2016. **165**(6): p. 1332-1345.
187. Yaron, S., et al., *Expression, purification and subunit-binding properties of cohesins 2 and 3 of the Clostridium thermocellum cellulosome.* FEBS Lett, 1995. **360**(2): p. 121-4.
188. Pagès, S., et al., *Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: prediction of specificity determinants of the dockerin domain.* Proteins, 1997. **29**(4): p. 517-27.
189. Yoav, S., et al., *How does cellulosome composition influence deconstruction of lignocellulosic substrates in Clostridium (Ruminiclostridium) thermocellum DSM 1313?* Biotechnol Biofuels, 2017. **10**: p. 222.
190. Osiro, K.O., et al., *Characterization of Clostridium thermocellum (B8) secretome and purified cellulosomes for lignocellulosic biomass degradation.* Enzyme Microb Technol, 2017. **97**: p. 43-54.
191. Lombard, V., et al., *The carbohydrate-active enzymes database (CAZy) in 2013.* Nucleic Acids Res, 2014. **42**(Database issue): p. D490-5.
192. Tamura, K., et al., *Surface glycan-binding proteins are essential for cereal beta-glucan utilization by the human gut symbiont Bacteroides ovatus.* Cell Mol Life Sci, 2019: p. 10.1007/s00018-019-03115-3.
193. Cuskin, F., et al., *Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism.* Nature, 2015. **517**(7533): p. 165-9.
194. Rogowski, A., et al., *Glycan complexity dictates microbial resource allocation in the large intestine.* Nat Commun, 2015. **6**: p. 7481.
195. Glenwright, A.J., et al., *Structural basis for nutrient acquisition by dominant members of the human gut microbiota.* Nature, 2017. **541**(7637): p. 407-411.
196. Dassa, B., et al., *Genome-wide analysis of acetivibrio cellulolyticus provides a blueprint of an elaborate cellulosome system.* BMC Genomics, 2012. **13**: p. 210.
197. Artzi, L., E.A. Bayer, and S. Morais, *Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides.* Nat Rev Microbiol, 2017. **15**(2): p. 83-95.
198. Lytle, B.L., et al., *Secondary structure and calcium-induced folding of the Clostridium thermocellum dockerin domain determined by NMR spectroscopy.* Arch Biochem Biophys, 2000. **379**(2): p. 237-44.
199. Chen, C., et al., *Revisiting the NMR solution structure of the Cel48S type-I dockerin module from Clostridium thermocellum reveals a cohesin-primed conformation.* J Struct Biol, 2014. **188**(2): p. 188-93.
200. Turkenburg, J.P., et al., *Structure of a pullulanase from Bacillus acidopullulyticus.* Proteins, 2009. **76**(2): p. 516-9.
201. Takeo, K., *Affinity electrophoresis: Principles and applications.* ELECTROPHORESIS, 1984. **5**(4): p. 187-195.
202. Freelove, A.C., et al., *A novel carbohydrate-binding protein is a component of the plant cell wall-degrading complex of Piromyces equi.* J Biol Chem, 2001. **276**(46): p. 43010-7.
203. Barak, Y., et al., *Matching fusion protein systems for affinity analysis of two interacting families of proteins: the cohesin-dockerin interaction.* J Mol Recognit, 2005. **18**(6): p. 491-501.

204. Giraud, E. and G. Cuny, *Molecular characterization of the α-amylase genes of Lactobacillus plantarum A6 and Lactobacillus amylovorus reveals an unusual 3' end structure with direct tandem repeats and suggests a common evolutionary origin.* Gene, 1997. **198**(1): p. 149-157.

205. Morlon-Guyot, J., et al., *Characterization of the L. manihotivorans α-Amylase Gene.* DNA Sequence, 2001. **12**(1): p. 27-37.

206. Holm, L., *Using Dali for Protein Structure Comparison*, in *Structural Bioinformatics: Methods and Protocols*, Z. Gáspári, Editor. 2020, Springer US: New York, NY. p. 29-42.

207. Imberty, A., et al., *The double-helical nature of the crystalline part of A-starch.* J Mol Biol, 1988. **201**(2): p. 365-78.

208. Tian, W., et al., *CASTp 3.0: computed atlas of surface topography of proteins.* Nucleic Acids Res, 2018. **46**(W1): p. W363-w367.

209. Cameron, E.A., et al., *Multidomain carbohydrate-binding proteins involved in Bacteroides thetaiotaomicron starch metabolism.* J Biol Chem, 2012. **287**(41): p. 34614-25.

210. Les Copeland, J.B., Hayfa Salman, Mary Chiming Tang, *Form and functionality of starch.* Food Hydrocolloids, 2008: p. 1527–1534.

211. Atwood, J.L., J.E.D. Davies, and D.D. MacNicol, *Inclusion Compounds: Physical properties and applications*. Vol. 3. 1984: Academic Press.

212. Gessler, K., et al., *V-Amylose at atomic resolution: X-ray structure of a cycloamylose with 26 glucose residues (cyclomaltohexaicosaose).* Proc Natl Acad Sci U S A, 1999. **96**(8): p. 4246-51.

213. Kelley, L.A., et al., *The Phyre2 web portal for protein modeling, prediction and analysis.* Nat Protoc, 2015. **10**(6): p. 845-58.

214. Svergun, D., *Determination of the regularization parameter in indirect-transform methods using perceptual criteria.* Journal of Applied Crystallography, 1992. **25**(4): p. 495-503.

215. Schneidman-Duhovny, D., et al., *FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles.* Nucleic Acids Research, 2016. **44**(W1): p. W424-W429.

216. *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2016. **44**(D1): p. D7-19.

217. Drozdetskiy, A., et al., *JPred4: a protein secondary structure prediction server.* Nucleic Acids Research, 2015. **43**(W1): p. W389-W394.

218. Yin, Y., et al., *dbCAN: a web resource for automated carbohydrate-active enzyme annotation.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W445-51.

219. Zeeman, S.C., J. Kossmann, and A.M. Smith, *Starch: its metabolism, evolution, and biotechnological modification in plants.* Annu Rev Plant Biol, 2010. **61**: p. 209-34.

220. Pérez, S., P.M. Baldwin, and D.J. Gallant, *Chapter 5 - Structural Features of Starch Granules I*, in *Starch (Third Edition)*, J. BeMiller and R. Whistler, Editors. 2009, Academic Press: San Diego. p. 149-192.

221. Jane, J.-l., *Chapter 6 - Structural Features of Starch Granules II*, in *Starch (Third Edition)*, J. BeMiller and R. Whistler, Editors. 2009, Academic Press: San Diego. p. 193-236.

222. Hammel, M., et al., *Structural basis of cellulosome efficiency explored by small angle X-ray scattering.* J Biol Chem, 2005. **280**(46): p. 38562-8.

223. Ossowski, I., et al., *Protein disorder: Conformational distribution of the flexible linker in a chimeric double cellulase.* Biophys J, 2005. **88**: p. 2823-2832.

224. Violot, S., et al., *Structure of a full length psychrophilic cellulase from Pseudoalteromonas haloplanktis revealed by X-ray diffraction and small angle X-ray scattering.* J Mol Biol, 2005. **348**(5): p. 1211-24.

225. Receveur, V., et al., *Dimension, shape, and conformational flexibility of a two domain fungal cellulase in solution probed by small angle X-ray scattering.* J Biol Chem, 2002. **277**(43): p. 40887-92.

226. Artzi, L., E.A. Bayer, and S. Moraïs, *Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides.* Nat Rev Microbiol, 2017. **15**(2): p. 83-95.

227. Bule, P., et al., *Structure-function analyses generate novel specificities to assemble the components of multienzyme bacterial cellulosome complexes.* J Biol Chem, 2018. **293**(11): p. 4201-4212.

228. Ravachol, J., et al., *Combining free and aggregated cellulolytic systems in the cellulosome-producing bacterium Ruminiclostridium cellulolyticum.* Biotechnol Biofuels, 2015. **8**: p. 114.

229. Crost, E.H., et al., *Mechanistic Insights Into the Cross-Feeding of Ruminococcus gnavus and Ruminococcus bromii on Host and Dietary Carbohydrates.* Front Microbiol, 2018. **9**: p. 2558.

230. Van Duyne, G.D., et al., *Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin.* J Mol Biol, 1993. **229**(1): p. 105-24.

231. Cockburn, D., C. Wilkens, and B. Svensson, *Affinity Electrophoresis for Analysis of Catalytic Module-Carbohydrate Interactions.* Methods Mol Biol, 2017. **1588**: p. 119-127.

232. Abbott, D.W. and A.B. Boraston, *Quantitative approaches to the analysis of carbohydrate-binding module function.* Methods Enzymol, 2012. **510**: p. 211-31.

233. Otwinowski, Z. and W. Minor, *Processing of X-ray Diffraction Data Collected in Oscillation Mode.*, in *Methods in Enzymology*, C.W.J. Carter and R.M. R.M. Sweet, Editors. 1997, Academic Press. p. 307-326.

234. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 213-21.

235. Liebschner, D., et al., *Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix.* Acta Crystallogr D Struct Biol, 2019. **75**(Pt 10): p. 861-877.

236. Emsley, P., et al., *Features and development of Coot.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 4): p. 486-501.

237. Hopkins, J.B., R.E. Gillilan, and S. Skou, *BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis.* J Appl Crystallogr, 2017. **50**(Pt 5): p. 1545-1553.

238. Perez-Riverol, Y., et al., *The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences.* Nucleic Acids Res, 2022. **50**(D1): p. D543-d552.

239. Kikhney, A.G., et al., *SASBDB: Towards an automatically curated and validated repository for biological scattering data.* Protein Sci, 2020. **29**(1): p. 66-75.

240. Stern, J., et al., *Adaptor Scaffoldins: An Original Strategy for Extended Designer Cellulosomes, Inspired from Nature.* mBio, 2016. **7**(2): p. e00083.

241. Spinelli, S., et al., *Crystal structure of a cohesin module from Clostridium cellulolyticum: implications for dockerin recognition.* J Mol Biol, 2000. **304**(2): p. 189-200.

242. Cameron, K., et al., *Combined Crystal Structure of a Type I Cohesin: MUTATION AND AFFINITY BINDING STUDIES REVEAL STRUCTURAL DETERMINANTS OF COHESIN-DOCKERIN SPECIFICITIES.* J Biol Chem, 2015. **290**(26): p. 16215-25.

243. Tavares, G.A., P. Béguin, and P.M. Alzari, *The crystal structure of a type I cohesin domain at 1.7 A resolution.* J Mol Biol, 1997. **273**(3): p. 701-13.

244. Carvalho, A.L., et al., *Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex.* Proc Natl Acad Sci U S A, 2003. **100**(24): p. 13809-14.

245. Slutzki, M., et al., *Crucial roles of single residues in binding affinity, specificity, and promiscuity in the cellulosomal cohesin-dockerin interface.* J Biol Chem, 2015. **290**(22): p. 13654-66.

246. de Souza, P.M. and P. de Oliveira Magalhães, *Application of microbial α-amylase in industry - A review.* Brazilian journal of microbiology : [publication of the Brazilian Society for Microbiology], 2010. **41**(4): p. 850-861.

247. Shimon, L.J.W., et al., *A cohesin domain from Clostridium thermocellum: the crystal structure provides new insights into cellulosome assembly.* Structure, 1997. **5**(3): p. 381-390.

248. Bule, P., et al., *Higher order scaffoldin assembly in Ruminococcus flavefaciens cellulosome is coordinated by a discrete cohesin-dockerin interaction.* Sci Rep, 2018. **8**(1): p. 6987.

249. Evans, R., et al., *Protein complex prediction with AlphaFold-Multimer.* bioRxiv, 2022: p. 2021.10.04.463034.

250. Carvalho, A.L., et al., *Evidence for a dual binding mode of dockerin modules to cohesins.* Proceedings of the National Academy of Sciences, 2007. **104**(9): p. 3089-3094.

251. Adams, J.J., et al., *Mechanism of bacterial cell-surface attachment revealed by the structure of cellulosomal type II cohesin-dockerin complex.* Proc Natl Acad Sci U S A, 2006. **103**(2): p. 305-10.

252. Brás, J.L.A., et al., *Novel Clostridium thermocellum type I cohesin-dockerin complexes reveal a single binding mode.* The Journal of biological chemistry, 2012. **287**(53): p. 44394-44405.

253. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

254. David, Y.B., *Characterization of the cellulosome systems of two related human-gut bacteria, Ruminococcus champanellensis and Ruminococcus bromii.* 2013, Weizmann Institute of Science. p. 62.

255. Weinstein, J.Y., et al., *Insights into a type III cohesin-dockerin recognition interface from the cellulose-degrading bacterium Ruminococcus flavefaciens.* J Mol Recognit, 2015. **28**(3): p. 148-54.

256. Hirano, K., et al., *Enzymatic diversity of the Clostridium thermocellum cellulosome is crucial for the degradation of crystalline cellulose and plant biomass.* Scientific Reports, 2016. **6**(1): p. 35709.

257. Brás, J.L., et al., *Diverse specificity of cellulosome attachment to the bacterial cell surface.* Sci Rep, 2016. **6**: p. 38292.

258. Winter, G., et al., *DIALS: implementation and evaluation of a new integration package.* Acta Crystallographica Section D, 2018. **74**(2): p. 85-97.

259. Terwilliger, T.C., et al., *Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard.* Acta Crystallogr D Biol Crystallogr, 2009. **65**(Pt 6): p. 582-601.

260. Han, S.O., et al., *Regulation of expression of cellulosomes and noncellulosomal (hemi)cellulolytic enzymes in Clostridium cellulovorans during growth on different carbon sources.* Journal of bacteriology, 2004. **186**(13): p. 4218-4227.

261. Yoav, S., et al., *How does cellulosome composition influence deconstruction of lignocellulosic substrates in Clostridium (Ruminiclostridium) thermocellum DSM 1313?* Biotechnology for biofuels, 2017. **10**: p. 222-222.

262. Ndeh, D. and H.J. Gilbert, *Biochemistry of complex glycan depolymerisation by the human gut microbiota.* FEMS Microbiology Reviews, 2018. **42**(2): p. 146-164.

263. Matano, Y., et al., *Cellulose promotes extracellular assembly of Clostridium cellulovorans cellulosomes.* J Bacteriol, 1994. **176**(22): p. 6952-6.

264. Han, S.O., et al., *Regulation of expression of cellulosomal cellulase and hemicellulase genes in Clostridium cellulovorans.* J Bacteriol, 2003. **185**(20): p. 6067-75.

265. Kobe, B. and A.V. Kajava, *The leucine-rich repeat as a protein recognition motif.* Curr Opin Struct Biol, 2001. **11**(6): p. 725-32.

266. Ceulemans, H., et al., *A capping domain for LRR protein interaction modules.* FEBS Lett, 1999. **456**(3): p. 349-51.

267. Matsushima, N., et al., *Structural analysis of leucine-rich-repeat variants in proteins associated with human diseases.* Cell Mol Life Sci, 2005. **62**(23): p. 2771-91.

268. Page, M.J. and E. Di Cera, *Serine peptidases: classification, structure and function.* Cell Mol Life Sci, 2008. **65**(7-8): p. 1220-36.

269. Dickinson, D.P., *Cysteine peptidases of mammals: their biological roles and potential effects in the oral cavity and other tissues in health and disease.* Crit Rev Oral Biol Med, 2002. **13**(3): p. 238-75.

270. Phitsuwan, P., et al., *The Cellulosome Paradigm in An Extreme Alkaline Environment.* Microorganisms, 2019. **7**(9): p. 347.

271. Bateman, A. and N.D. Rawlings, *The CHAP domain: a large family of amidases including GSP amidase and peptidoglycan hydrolases.* Trends Biochem Sci, 2003. **28**(5): p. 234-7.

272. Morais, S., et al., *Lysozyme activity of the Ruminococcus champanellensis cellulosome.* Environ Microbiol, 2016. **18**(12): p. 5112-5122.

273. Artzi, L., et al., *Clostridium clariflavum: Key Cellulosome Players Are Revealed by Proteomic Analysis.* mBio, 2015. **6**(3): p. e00411-15.

274. Raman, B., et al., *Impact of pretreated Switchgrass and biomass carbohydrates on Clostridium thermocellum ATCC 27405 cellulosome composition: a quantitative proteomic analysis.* PLoS One, 2009. **4**(4): p. e5271.

275. Morisaka, H., et al., *Profile of native cellulosomal proteins of Clostridium cellulovorans adapted to various carbon sources.* AMB Express, 2012. **2**(1): p. 37.

276. Fendri, I., et al., *The cellulosomes from Clostridium cellulolyticum: identification of new components and synergies between complexes.* Febs j, 2009. **276**(11): p. 3076-86.

277. Yadav, J.K. and V. Prakash, *Stabilization of α-Amylase, the Key Enzyme in Carbohydrates Properties Alterations, at Low pH.* International Journal of Food Properties, 2011. **14**(6): p. 1182-1196.

278. Han, Y., et al., *Biochemical and structural insights into xylan utilization by the thermophilic bacterium Caldanaerobius polysaccharolyticus.* J Biol Chem, 2012. **287**(42): p. 34946-60.

279. Czjzek, M. and E. Ficko-Blean, *Probing the Complex Architecture of Multimodular Carbohydrate-Active Enzymes Using a Combination of Small Angle X-Ray Scattering and X-Ray Crystallography.* Methods Mol Biol, 2017. **1588**: p. 239-253.

280. Schwarz, W.H., *The cellulosome and cellulose degradation by anaerobic bacteria.* Appl Microbiol Biotechnol, 2001. **56**(5-6): p. 634-49.

281. Fierobe, H.P., et al., *Purification and characterization of endoglucanase C from Clostridium cellulolyticum. Catalytic comparison with endoglucanase A.* Eur J Biochem, 1993. **217**(2): p. 557-65.

282. Lamed, R., et al., *Nonproteolytic cleavage of aspartyl proline bonds in the cellulosomal scaffoldin subunit from Clostridium thermocellum.* Appl Biochem Biotechnol, 2001. **90**(1): p. 67-73.

283. Chen, H.M., C. Ford, and P.J. Reilly, *Identification and elimination by site-directed mutagenesis of thermolabile aspartyl bonds in Aspergillus awamori glucoamylase.* Protein Eng, 1995. **8**(6): p. 575-82.