INVESTIGATING STRUCTURE-FUNCTION RELATIONSHIPS IN FAMILY 7 CELLULASES BY MOLECULAR SIMULATION

By

Courtney Barnett Taylor

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical Engineering

August, 2012

Nashville, Tennessee

Approved:

Clare McCabe

Peter T. Cummings

Eugene LeBoeuf

Kenneth A. Debelak

To Mom, Dad, John, and Mal, unwavering in their support

and

To Trent, for 13 years of patience

ACKNOWLEDGEMENTS

This work has been possible through the financial support of the DOE Office of the Biomass Program for funding. Computational time for this research was provided under NSF Teragrid on the TACC Ranger cluster and the NICS Athena and Kraken clusters, and through the resources at NERSC.

I am grateful for the support and guidance I have received during my research from my advisor, Clare McCabe. I am especially grateful for the leap of faith taken when agreeing to let me work from Baton Rouge for nearly my entire graduate career. I would like to thank the members of my Dissertation Committee for their time contributed to this work. Thanks to Faiz Talib for his hard work as an undergrad. I cannot have an acknowledgements section without addressing the group at NREL, especially Dr. Gregg Beckham, who was in essence my co-advisor and instrumental to this work. I thank the rest of the group, Drs. Christy Payne, Mike Crowley, Yannick Bomble, Deanne Sammonds, Lintao Bu, James Matthews, and Mike Himmel, for their massive amounts of patience, assistance, and willingness to answer every rambling email sent their way. They make pretty good friends, too!

Finally, with limited space, I thank my family and friends for maintaining my sanity. To all those I met at Vanderbilt, you are great people. To my Nashville crew (you know who you are), I feel like Dorothy leaving Oz; you all were the best part of this experience, and thanks for walking the road with me. To my closest friends and family that I wish I could name individually, I cannot thank you all enough for your support. And finally to Trent, bless you bubs, you deserve sainthood at this point; thank you, I love you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Background

In 2010, the United States consumed 98 quadrillion BTU's of energy with a breakdown of 37% from petroleum, 21% from coal, 25% from natural gas and only 8% from renewable resources [1]. Approximately 70% of the petroleum consumed makes up 92% of the energy required by the transportation sector, with the total consumption in all sectors on a steady increase since 1950, as shown in Figure 1.1 [2].

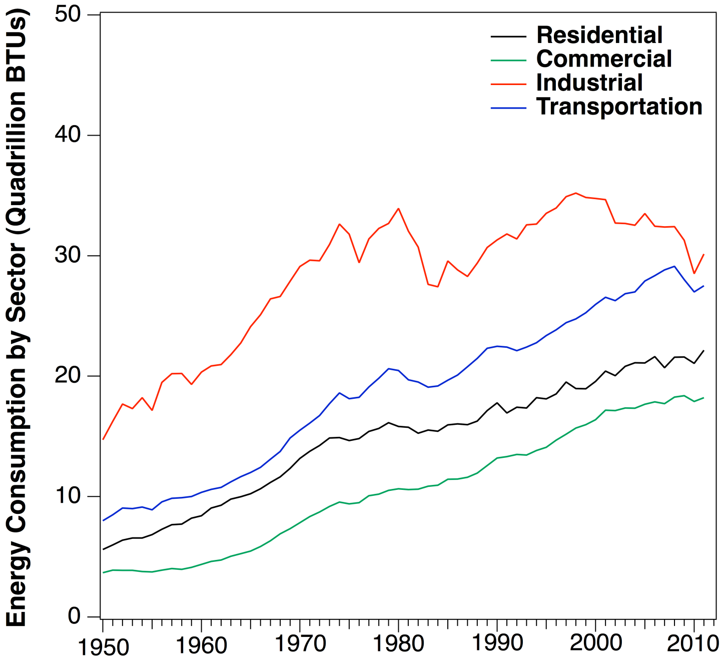


Figure 1.1: Energy Consumption in the United States in 2010 by sector (Quadrillion BTUs) [2].

Although the United States is the third largest producer of crude oil, approximately half of the petroleum used is imported, with 50% of imports coming from the Western Hemisphere and 18% originating in the Middle East [1]. As the United States faces an uncertain economy, continuing conflict in the Middle East, industrial disasters like the Deepwater Horizon oil spill, and other disruptions in oil and gas supply, the desire to relieve the dependence on traditional fossil fuel resources, while improving alternative production at home and thus creating new jobs, is becoming ever more prominent. In addition, the constant burning of fossil fuel provides continuing damage to the environment via emissions. All these factors reinforce the need for research into alternative, renewable fuel sources [3,4].

The Obama administration is committed to an "all-of-the-above" energy strategy that is focused on developing all sources of American energy with a focus on clean technologies, requesting $27.2 billion for the DOE budget in the 2013 Fiscal Year 2013 [5]. A primary goal of this strategy is to reduce dependence on oil by 30% by 2025, and the DOE Office of Energy Efficiency and Renewable Energy's (EERE's) Biomass Program focuses efforts towards achieving this goal in the advanced biofuel sector specifically [4,5].

While there are many potential alternatives to fossil fuels, including solar and wind energy, biomass shows significant potential for transportation purposes. The DOE has reported that the use of ethanol from cellulosic sources may reduce greenhouse emissions to 90% less than that of gasoline and replace up to a third of the current U.S. demand for traditional transportation fuels [6]. Cellulosic ethanol, or biofuels created from lignocellulosic biomass, including plant wastes, switchgrass, or poplar trees, have come under considerable research focus of late due to the growing debate of "food for fuel"; the dilemma raised when choosing between growing corn for food or ethanol fuel [3,6]. The use of lignocellulosic biomass allows corn and agricultural land to remain reserved for food production; additionally, utilization of lignocellulosic feedstock is less expensive and incorporates recycling of waste products. However, one of the issues stemming from using lignocellulosic biomass as a fuel source is inefficiency [7,8]; although cellulose is one of the most abundant polymers on the planet it is very resistant to hydrolysis into its monomer glucose, which can be fermented to ethanol [9].

The dry mass of lignocellulose is composed of polymers of cellulose, hemicellulose, pectin, lignin, and lignocellulose, a strong laminate that surrounds the cellulose fibers. A glycosidic bond is formed by the hemiacetal group of a sugar bonding to a hydroxyl group of another sugar, in this case two β1,4-D-glucose residues (Figure 1.2a), to form a polymer chain of cellobiose units (Figure 1.2b) [10]. Native crystalline cellulose exists in two primary forms in plants, Iα and Iβ [11], that contain enhanced networks of hydrogen bonds and hydrophobic interactions that contribute to its resistance to hydrolysis [10,12-14]. Plants, such as poplar or switchgrass, that are targeted for biomass conversion contain cellulose Iβ in their cell walls, which according to crystallographic studies [13], contains two cellobiose molecules (Figure 1.1a) per unit cell and linear combinations of these chains, connected by hydrogen bonding interactions between the glucosyl hydroxyl of one chain and the oxygen in the neighboring chain, form the cellulose sheets or layers (Figure 1.1b) [12,15]. Cellulose Iα and Iβ also exhibit intra-layer hydrogen bonding to form 36-chain microfibrils (Figure 1.2c), which in turn combine with hemicellulose and the other components listed above to form the entire dry mass of lignocellulosic material [10,13,14,16].

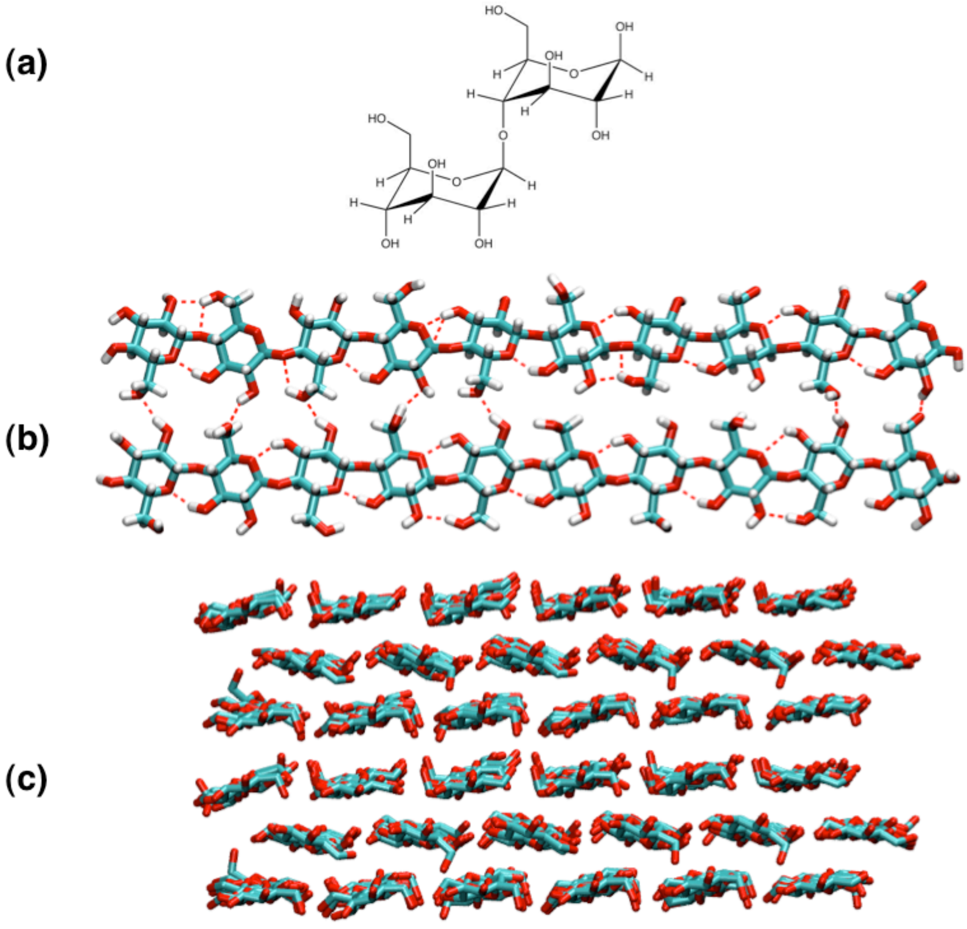


Figure 1.2: Cellobiose units (a) form cellobiose chains (b) containing hydrogen bonding networks, shown with red dotted lines. The chains and sheets combine to form 36-chain microfibril structures (c).

This complex lignocellulosic material must be broken down and separated via hydrolysis into fermentable sugars that can then be converted to ethanol [6]. The near-term solutions to overcoming recalcitrance are thermochemical conversion via pyrolysis and gasification or biochemical conversion [3]. Biochemical conversion has the advantage of high selectivity of end products and is comprised of a thermal-chemical, often acidic, pretreatment step followed by an enzymatic hydrolysis step to convert the treated biomass to simple sugars that are them fermented to fuels [3,7,10]. Fungi are commercially important because they can secrete large amounts of highly effective cocktails of glycoside hydrolase (GH) enzymes, which are responsible for breaking the glycosidic bonds in cellulose and include cellulases, hemicellulases, and accessory enzymes, such as those that assist in lignin degradation [3,9]. GHs are divided into families based on the amino acid sequence and structure, and cellulases are further divided into three categories based on catalytic action: processive exoglucanases, non-processive endoglucanases, and β-glucosidases [10,17,18] The enzymes operate synergistically, with endoglucanases locating and hydrolyzing cellulose surface sites at random to produce cellobiose or expose a terminus and exoglucanases acting in a processive motion, targeting the termini at the reducing (containing the terminal aldehyde group) and non-reducing ends of a cellulose chain and breaking the chain down to one cellobiose unit per event until the chain ends or the enzyme activity is lost [9,18]. β-glucosidases convert cellobiose to glucose, which helps to prevent product inhibition [18]. A simplified representation of enzymatic hydrolysis is provided in Figure 1.3 [18].

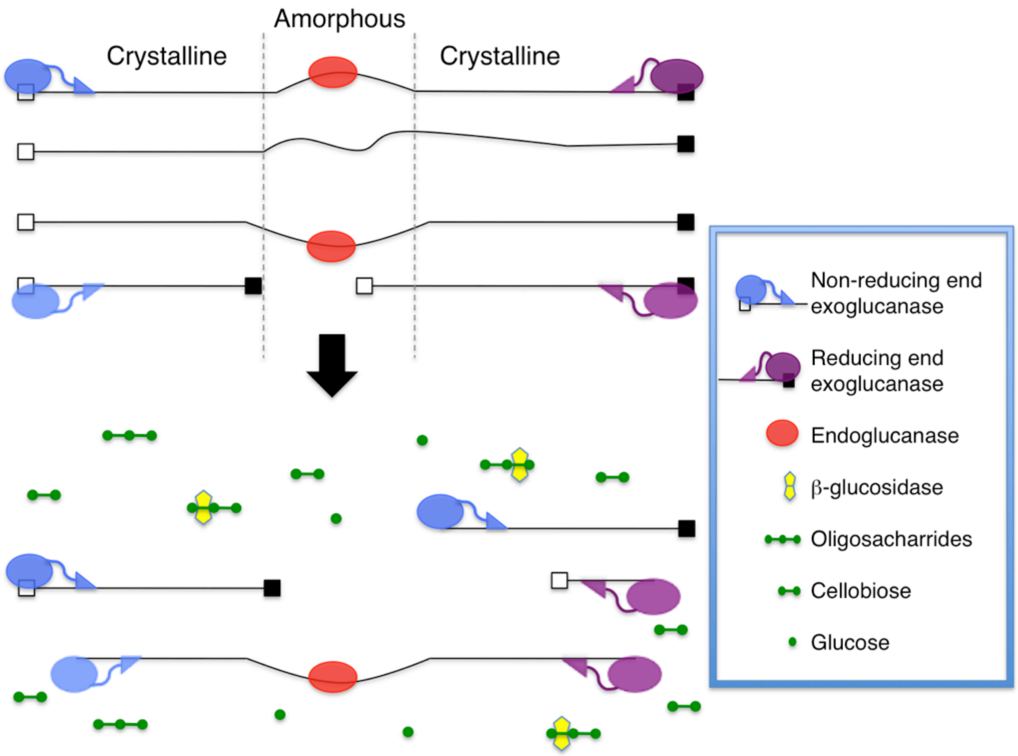


Figure 1.3: Simplified schematic of synergistic cellulase enzyme action on crystalline and amorphous cellulose as suggested by Lynd *et al.* [18] Non-reducing ends (white box) are drawn to the left and reducing ends (black box) are drawn to the right. In *Trichoderma reesei,* Cel7A is a reducing end exoglucanase, Cel6A is a non-reducing end exoglucanase, and Cel7B is an exoglucanase.

While cellulases are highly effective in nature, their turnover rates are not yet sufficient for large scale commercial biofuel production [7,8]. Understanding the enzymatic crystallization process, including how the enzymes access and act upon the cellulose substrate is crucial for providing fundamental knowledge about the cellulosic biofuel process as a whole [6] If the action of these enzymes were understood, more efficient enzymes could be engineered that would subsequently result in improvements to the production process [7] A primary focus of this work is thus to contribute to the understanding of the enzymatic mechanistic action on cellulose by focusing on how exo-

and endoglucanases from filamentous fungi identify and bind to crystalline cellulose, and how this may be improved to increase hydrolytic efficiency.

1.2 *Tricoderma reesei* Cellulases

Cellobiohydrolases are a type of exoglucanase, and the Family 7 cellobiohydrolase I (Cel7A or CBH-1) from the filamentous fungi *Trichoderma reesei* is one of the most active and most studied to date [3]. It is comprised of three parts, the large catalytic domain (CD) containing a 50 Å long tunnel to bind the cellulose chain, a small carbohydrate binding module (CBM), and a *O*-glycosylated connective linker peptide; experimental methods have determined the structure and sequence of the binding and catalytic domain, while only the linker peptide amino acid sequence and glycosylation pattern is known [17,19-22]. A representative picture of the enzyme on crystalline cellulose is shown in Figure 1.4 [9].

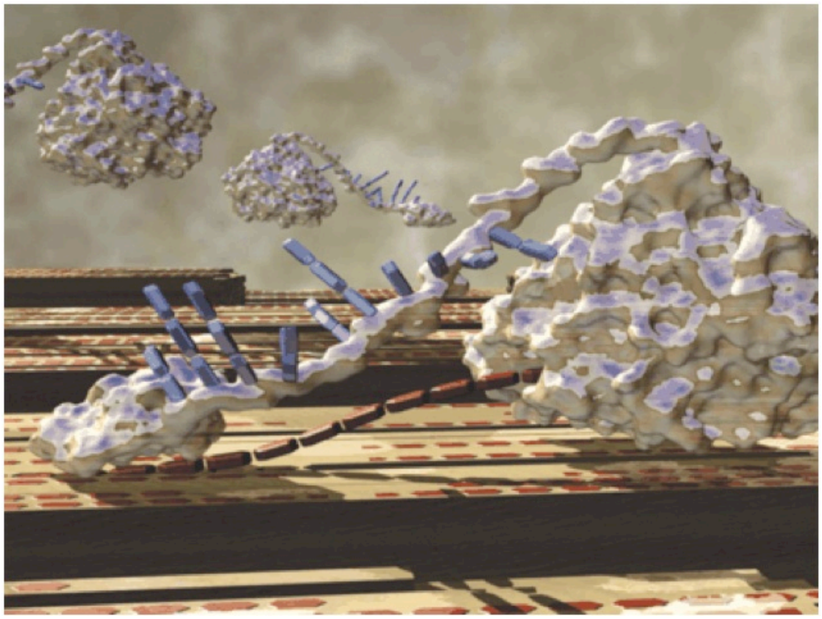


Figure 1.4: Rendering of the Cellobiohydrolase-I (Cel7A) enzyme in complex with a cellulose fibril. A cellodextrin chain (red) is threaded into the CD where hydrolysis will take place. The connective linker is heavily glycosylated with mannose sugars (light blue). Rendering taken from Himmel *et al*., 2007 [9].

Cel7A is believed to move down a cellulose microfibril in a processive motion, as other exoglucanases do, breaking the sequential glycosidic bonds to release cellobiose. While the exact catalytic cycle is unknown, it proposed that the following steps occur (Figure 1.5): the enzyme CBM recognizes the reducing end of a cellulose chain within a microfibril (Figure 1.5a-b), the chain is threaded into the CD tunnel (Figure 1.5c-d), then the chain is hydrolyzed to cellobiose and the cellobiose product is expelled (Figure 1.5e-f) [23-25].

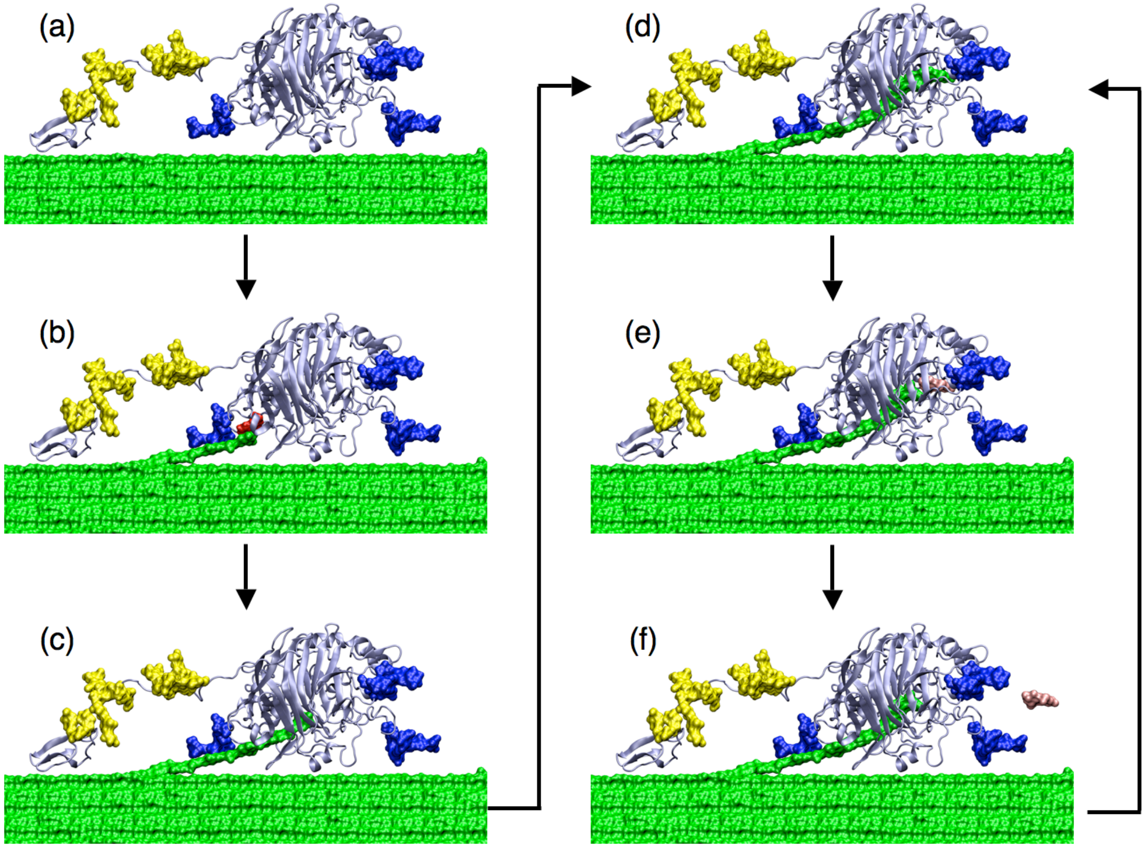


Figure 1.5: Proposed catalytic cycle of Cel7A on crystalline cellulose from Chundawat *et al.*, 2011 [3]. The yellow and blue space-filling representations are *O*-glycosylation on the linker and CBM and *N*-glycosylation on the CD, respectively. The light blue molecule is the Cel7A enzyme, and the green substrate is a cellulose microfibril. The cellobiose product is expelled in (f) and shown in pink.

Carbohydrate-Binding Module (CBM)

The first proposed step in the mechanistic action on cellulose is the recognition and binding of the CBM (Figure 1.5a). The primary role of the CBM is to promote the association of the enzyme complex with the cellulose substrate. CBMs exhibit ligand specificity, that is, they are able to target specific sugar structures, such as crystalline cellulose, and either bind solely to one ligand type or bind to a range of accepted ligand types. Because of this specificity, they are excellent models to study and identify the mechanism by which the CBM recognizes different carbohydrate molecules. Studies of the CBM to date have focused on the structural, functional, and biological components of the molecule and have determined that the CBM serves two roles with respect to the enzyme as a whole by maintaining proximity to the surface and targeting the specific ligands, and serves a third proposed, but unconfirmed, role of disrupting the cellulose surface so that cellulose chains can be fed to the CD [26]. Furthermore, experiments comparing Cel7A with and without the CBM confirm that the CBM is required not only for binding but also for efficient cellulose degradation [24]. CBMs are separated into families based on amino acid similarity and then types based on the structure of the cellulose they bind to; type A CBMs bind to surfaces of crystalline cellulose, type B bind to individual chains, and type C bind to small sugars [26]. CBM family designations are not to be confused with the GH family designation; the CD, and thus entire enzyme belongs to GH7, and the CBM belongs to CBM-Family 1 with a type A module that binds to the hydrophobic (1,0,0) face of crystalline cellulose [26,27].

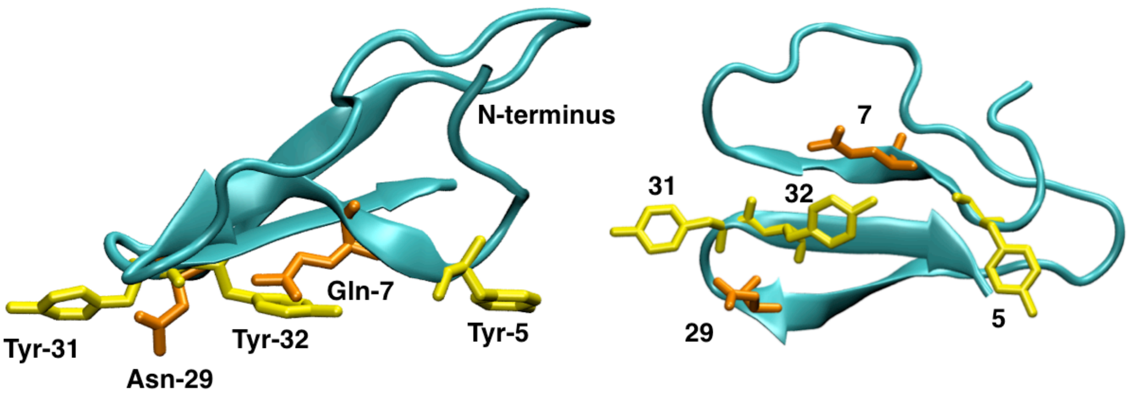


Figure 1.6: Cel7A Family 1 CBM with flat face residues shown. Aromatics (Tyr-5, Tyr-31, and Tyr-32) are shown in yellow and polar (Asn-29 and Gln-7) are shown in orange.

The structure of the Cel7A CBM is thought to contribute to its binding affinity because it contains a flat face, confirmed by NMR data (Figure 1.6), that complements the surface of the cellulose [24]. As can be seen from Figure 1.6, three tyrosines (Tyr or Y) at positions 5, 31, and 32, one glutamine (Gln or Q) at position 7, and one asparagine (Asn or N) at position 29 form the flat face of the cellulose surface. Experimental studies have calculated the change in binding affinity that results from mutation of various residues in the Cel7A CBM [28,29]. The aforementioned flat-face residues show high levels of homology or conservation in many Family 1 CBMs from various enzymes [24]. In proteins, homologous regions are derived from common ancestors and/or genetic sequences, while sequence identity and similarity measures the exact match of amino acids or resemblance of amino acid sequences, respectively. Conserved residues are those that occupy the same position in the amino acid chain across various types of CBMs, such as Cel7A and *T. reesei*'s homologous endoglucanase (Cel7B), and various fungi; examples of these sequences along with the wild type Cel7A CBM sequence are shown in Table 1.1.

Table 1.1: Sequence alignment of various CBMs for exoglucanases (Cel7A and Cel6A) and endoglucanases (Cel7B) in example fungi [24]. The flat-face residues of interest are highlighted in yellow. The native, or wild type, Cel7A amino acid sequence is shown on the first line.

Organism	Enzyme	N				5		7																						29		31	32				C
Trichoderma reesei	**Cel7A**	T	Q	S	H	Y	G	Q	C	G	G	I	G	Y	S	G	P	T	V	C	A	S	G	T	T	C	Q	V	L	N	P	Y	Y	S	Q	C	L
Aspergillus aculeatus	Cel7A	V	A	Q	L	Y	G	Q	C	G	G	Q	G	W	T	G	P	T	T	C	A	S	G	--	T	C	T	K	Q	N	D	Y	Y	S	Q	C	L
Fusarium oxysporum	Cel7A	S	V	D	Q	W	G	Q	C	G	G	Q	N	Y	S	G	P	T	T	C	K	S	P	F	T	C	K	K	I	N	D	F	Y	S	Q	C	C
Humicola grisea	Cel7A	K	A	G	R	W	Q	Q	C	G	G	I	G	F	T	G	P	T	Q	C	E	E	P	Y	T	C	T	K	L	N	D	W	Y	S	Q	C	L
Penicillium funiculosum	Cel7A	T	A	A	H	W	A	Q	C	G	G	I	G	Y	S	G	C	T	A	C	A	S	P	Y	T	C	Q	K	A	N	D	Y	Y	S	Q	C	L
Trichoderma reesei	Cel6A	C	S	S	V	W	G	Q	C	G	G	Q	N	W	S	G	P	T	C	C	A	S	G	S	T	C	V	Y	S	N	D	Y	Y	S	Q	C	L
Trichoderma reesei	**Cel7B**	T	Q	T	H	W	G	Q	C	G	G	I	G	Y	S	G	C	K	T	C	T	S	G	T	T	C	Q	Y	S	N	D	Y	Y	S	Q	C	L
Tricoderma longibrachiatum	Cel7B	T	Q	T	H	W	G	Q	C	G	G	I	G	Y	T	G	C	K	T	C	T	S	G	T	T	C	Q	Y	G	N	D	Y	Y	S	Q	C	L

To test how the highly conserved residues impact the CBM structure and function, Linder *et al.* and later Takashima *et al.* targeted mutations intended to disrupt the structure of the planar face [29,30]. Additionally, these experimental studies recognized the improved activity and binding affinity of Cel7B over the Cel7A, and studied variants of the native CBM (wild type) amino acid sequence (Table 1.1) to attempt to increase the binding affinity to that of Cel7B. Single amino acid residues utilize the three-letter naming convention followed by the residue number, and mutations utilize the wild type single letter, residue number, and mutant single letter amino acid naming convention, e.g., Tyr-5 represents tyrosine at residue 5, and Y5A denotes mutation of Tyr to alanine (Ala or A) at residue 5. Bound and unbound concentrations of enzyme above cellulose were calculated and used to generate adsorption isotherms for mutated and wild type Cel7A [28,29]. Assuming that all CBMs have the same number of binding sites, Linder, *et al.* calculated the difference in free energy of binding, $\Delta\Delta G$, using the following relationship [28,29]:

$$\Delta\Delta G = -RT\ln\left({K_{Mut}}/{K_{WT}} \right) \tag{1.1}$$

where K_{Mut} and K_{WT} are the partition coefficients of the mutant and wild type Cel7A, respectively, T is temperature, and R the gas constant. The partition coefficients were taken from the initial slopes of the Langmuir adsorption isotherms. All of the variants, except mutation of Tyr-5 to tryptophan (Trp or W), denoted Y5W, had lower affinity than the wild type Cel7A with $\Delta\Delta G$ values of -0.6 to 1.7 kcal/mol, and while Y5W affinity was increased it still did not meet that of Cel7B. Table 1.2 documents the binding free energy results from the Linder group studies, where the distribution of bound over free in the wild type CBM was measured at 1.7 L/g and not further normalized to 1 for this particular study [28].

Table 1.2: Linder *et al*. calculated partition coefficients and binding free energy for Cel7B over Cel7A wild type CBM and mutations in the Cel7A wild type CBM planar face [28,29]. Amino acid lettering: Y, tyrosine; W, tryptophan; N, asparagine; A, alanine.

	$K_{\mathrm{Mut}}/K_{\mathrm{WT}}$ (L/g)	$\Delta\Delta G$ (kcal/mol)
Cel7B	5.2	-0.57
Cel7A CBM (wild type)	1.7	--
CBM Y5W	2.8	-0.26
CBM N29A	0.6	0.57
CBM Y31A	0.07	1.7
CBM Y5A	0	--
CBM Y32A	0	--

Changes closest to the N-terminal region appear to be the most sensitive to mutations and while Y5W has a positive affect on binding [28], Y5A has a detrimental effect because it drastically changed the overall structure of the CBM [29]. Thus, Y5A has a flat Langmuir isotherm and thus a value of zero for $\Delta\Delta G$ in Table 1.2. It was also noted that a decreased affinity in the CBM decreased the overall catalytic activity of the enzyme complex [28], supporting the importance of the three roles of the CBM in the overall action of the enzyme. However, the physical basis for the differences was not obvious; while it was suggested that changes in hydrophobicity, charge distribution, and van der Waals interactions could be contributing factors, it was not possible to quantify these affects [28]. The planar-face mutations of the conserved Tyr-32 and Tyr-31 were also studied, and it was demonstrated that mutation of these two residues significantly reduced the binding affinity to nearly zero [29]. This finding was consistent with a further study of Tyr-31, showing that mutation here decreases binding and overall activity of Cel7A to nearly that of a Cel7A CD with no CBM [31].

Several molecular simulation studies have also been performed to probe the interaction of the CBM flat-face and cellulose substrate. Nimlos *et al*. were able to show that a CBM in close proximity to the hydrophobic cellulose surface was able to distort itself in a way conducive to hydrophobic interactions with the surface, and identified hydrogen bonding between residues 5, 7, 29, 31, and 32 and the cellulose surface [15]. Subsequently, Beckham *et al.,* using atomistic modeling of the CBM and a coarse grained model of the cellulose surface, generated potential energy surfaces by placing the CBM on an x-y grid over the cellulose and minimizing the system at each grid point [24]. The results showed that the CBM exhibited a downhill potential energy surface when moving away from a hydrolyzed cellulose molecule and also exhibited thermodynamically stable minima on the surface of cellulose Iβ on a length scale of a cellobiose molecule. These findings support the theory that the CBM aids in thermodynamically driving the enzyme across the cellulose surface [24].

The initial study of the CBM, found in Chapter III, uses molecular simulations to reproduce the experimental binding results from Linder *et al.* [28,29], and provides molecular level energetic and structural details associated with mutation at Tyr-5. In accordance with other molecular simulation studies, the studies of the CBM also seek to describe the interaction of the protein and cellulose surface on an atomstic level.

Connective Linker Peptide

The connective linker attaches to the N-terminus of the CBM and the C-terminus of the

CD. Although linkers are common in cellulases [32] and they are known to be vital for the secretion of other enzymes [33,34], little is known about their structural and functional properties [33]. The amino acid sequence is known and provided in Figure 1.7. Several experimental studies have been conducted to characterize fungal cellulase linker peptides [33,35-43] and it has been suggested that the linker may help to maintain the spatial orientations of the enzyme by acting as a hinge between the two globular domains [21] or by acting as a "torsional leash" or "spring" [21,33,38,41,44]. Additionally, the extensive glycosylation, or sugar attachment to the protein, seen on linker peptides is believed to protect the enzyme from proteolysis (Figure 1.7). Furthermore, mutational biochemistry studies have demonstrated the importance of the linker in the overall action of the enzyme, in that shortening or removal of the linker peptide results in the reduction or loss of activity [21,40].

In an effort to elucidate the role of the linker, molecular dynamics simulation studies have been performed to examine its structural flexibility as a function of molecular position. Zhao *et al.* calculated the free energy of the linker above cellulose in explicit water as a function of the linker end-to-end distance from atomistic molecular dynamics simulations, and while the results indicated that the compression and extension of the linker could play a role in movement of the enzyme across the cellulose surface [34], questions remained regarding whether or not the simulations were able to sample all of the possible protein configurations. A second atomistic simulation study of the entire Cel7A enzyme over cellulose showed the linker exhibiting flexibility between the two larger domains [45], but the simulation time of 1.5 ns was too short to reliably deduce any insight into the collective motion of the enzyme. Ting *et al.* constructed a kinetic model describing the motion of the Cel7A enzyme using mathematical models that treated the large domains as random walkers coupled by a spring-like linker and concluded that the linker stiffness and length both play a role in the interaction of the CBM and CD [46]. Finally, Beckham *et al.,* calculated the free energy of the linker as a function of the length using atomistic simulations with an implicit solvent model and specifically examined the effect of glycosylation on the linker flexibility; the results suggested that the linker is an intrinsically disordered protein both with and without *O*-glycosylation [47] and that the *O*-glycosylation does not change the stiffness of the linker but provides an extension to increase the operating range of Cel7A [47]. While the linker domain is not a part of this particular study, the glycosylation on the peptide was found to extend to the CBM [44], as shown in Figure 1.7, and the potential implications of this are discussed in Chapters III and IV.

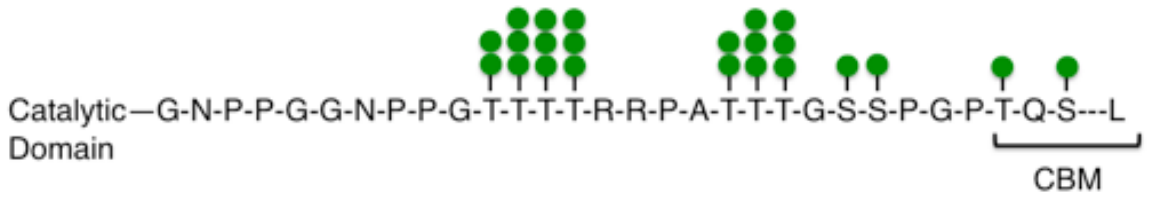

Catalytic—G-N-P-P-G-G-N-P-P-G-T-T-T-T-R-R-P-A-T-T-T-G-S-S-P-G-P-T-Q-S---L
Domain
CBM

Figure 1.7: *O*-glycosylation cites of Cel7A linker peptide as proposed by Harrison [44] and Nevalainen [19] with first three residues of CBM included. Mannose residues are represented by green circles. All of the Thr and Ser in Cel7A on the linker peptide are glycosylated with 1 to 3 mannose sugars with evidence of sulfation and phospohorylation also occurring in some strains [19,44,48].

Catalytic Domain

Once the CBM recognizes and binds to the surface, a cellodextrin strand must be bound into the CD for hydrolysis to cellobiose products (Figure 1.5c-e). Cel7A decreases the degree of polymerization slowly, so endoglucanases like Cel7B can aid by internally cleaving and exposing new chain ends by attacking non-crystalline, or amorphous, areas of the surface [18,49-52]. To more fully understand how natural cocktails of enzymes work synergistically to break down substrates, we must first understand the molecular-level mechanisms each individual enzyme employs to break down crystalline and amorphous surfaces [3,9,20,23,26,30,32,35,41,50,52-63]. Figure 1.5 is a representation of the processive exoglucanase Cel7A on crystalline cellulose, but as Figure 1.3 shows, Cel7A often works synergistically with other enzymes to break down cellulose. Understanding how enzymatic hydrolysis occurs using a mixture of enzymes is also key for designing enhanced cocktails for use in the biofuel production process [9,18,55]. Cel7A and Cel7B demonstrate high sequence homology with ~45% identity, but serve complementary functions when degrading cellulose, and also have differing structural components that may contribute to these differences [59,64]. Proteins are often classified by secondary structural elements, and a turn is an element where the backbone reverses direction; loops are turns comprised of longer, disordered segments of the protein with no hydrogen bonding between the residues [65]. Processive and non-processive cellulase enzymes are also broadly characterized by their respective structural aspects wherein processive enzymes typically exhibit tunnels or deep cleft geometries in the cellodextrin binding sites and non-processive enzymes exhibit more open, solvent-exposed clefts for substrate binding (Figure 1.8), [17,18,23,25,51,59]. Understanding the molecular-level basis for GH processivity is central to understanding carbohydrate metabolism in the biosphere [66] and is a key property in the design of enzymes for the burgeoning biofuels industry.

Processive and non-processive cellulases conduct work to decrystallize polymer chains from the surfaces of insoluble substrates, which is offset by the cellodextrin binding free energy of the substrate to the enzyme [10,67,68]. The magnitude of the cellodextrin binding free energy for a given enzyme, in turn, will dictate the locations wherein an enzyme can decrystallize and hydrolyze polymer chains from the surfaces of crystals. As mentioned, structural studies have suggested broadly that the respective functions of processive and non-processive enzymes are imparted by the formation of tunnels or deep clefts in processive enzymes relative to the more open substrate binding sites in non-processive enzymes. However, several studies have provided examples that such a clear structural delineation of GH processivity may not be as straightforward. For example, the GH Family 6 cellulase Cel6B from *Humicola insolens* is an endoglucanase, despite exhibiting a tunnel-like cellodextrin binding site in the crystal structure [69]. The GH Family 6 enzyme CBHA from *Thermomonospora fusca* was converted from a processive enzyme to a non-process enzyme with the removal of a single tunnel loop [70]. Additionally, a GH Family 18 chitinase (a GH enzyme that degrades chitin), ChiB, from *Serratia marscesens* was converted from exo- to endochitinase behavior with only a single mutation of a tryptophan residue to alanine [53,56]. Thus, the causal basis for GH processivity may stem from subtler differences in structural and dynamical features that impact the cellodextrin binding free energy. Here, we use molecular dynamics simulations and thermodynamic integration mutation cycles to probe the differences in cellodextrin interactions for a processive and non-processive cellulase from the same GH family.

To elucidate details of hydrolysis and binding, studies have been performed on GH6, GH7, and chitinase enzymes in attempts to highlight the functional differences in the CD, where hydrolysis occurs [17,25,50,51,56,57,59-61,64,71-74]. Overall, Cel7A and Cel7B exhibit a similar β-jelly roll structure (Figure 1.8). The Cel7A CD contains four surface loops that create a 50Å tunnel for cellodextrin binding [17,18,25,50,71]. In contrast, Cel7B has shorter surface loops that create a more open cleft structure instead of a tunnel [17,18,51,59,61]. Divne *et al.* determined that Cel7A exhibits up to ten binding sites for a cellodextrin chain, up to three on the product and seven on the substrate end, where the glycosidic cleavage occurs between the +1 and -1 sites, the "entrance" site refers to the -7 site, and the cellobiose product is bound in the +2 and +1 sites, with a putative third product site at +3 [25]. Kleywegt *et al.* considered nine of these sites, from -7 to +2, in their structural comparison of Cel7A and Cel7B [59]. The tunnel formation in exoglucanases such as Cel7A may block cellodextrin contact with the bulk solvent, and experimental studies speculate that specific water-mediated protein-cellodextrin interactions facilitate the gliding action of the cellodextrin through the tunnel in a processive enzyme [25,71]. Also, the tunnel loops may block any product or intermediate product binding potential [59]. Conversely, the exposure of the binding sites in endoglucanases such as Cel7B results in product inhibition and a higher degree of transglycosylation activity at high substrate concentrations [51,72]. A comparison of the two CD structures with a 9-mer cellodextrin chain present in each is shown in Figure 1A comparison of the two CD structures with a cellodextrin present in each is shown in Figure 1.8.

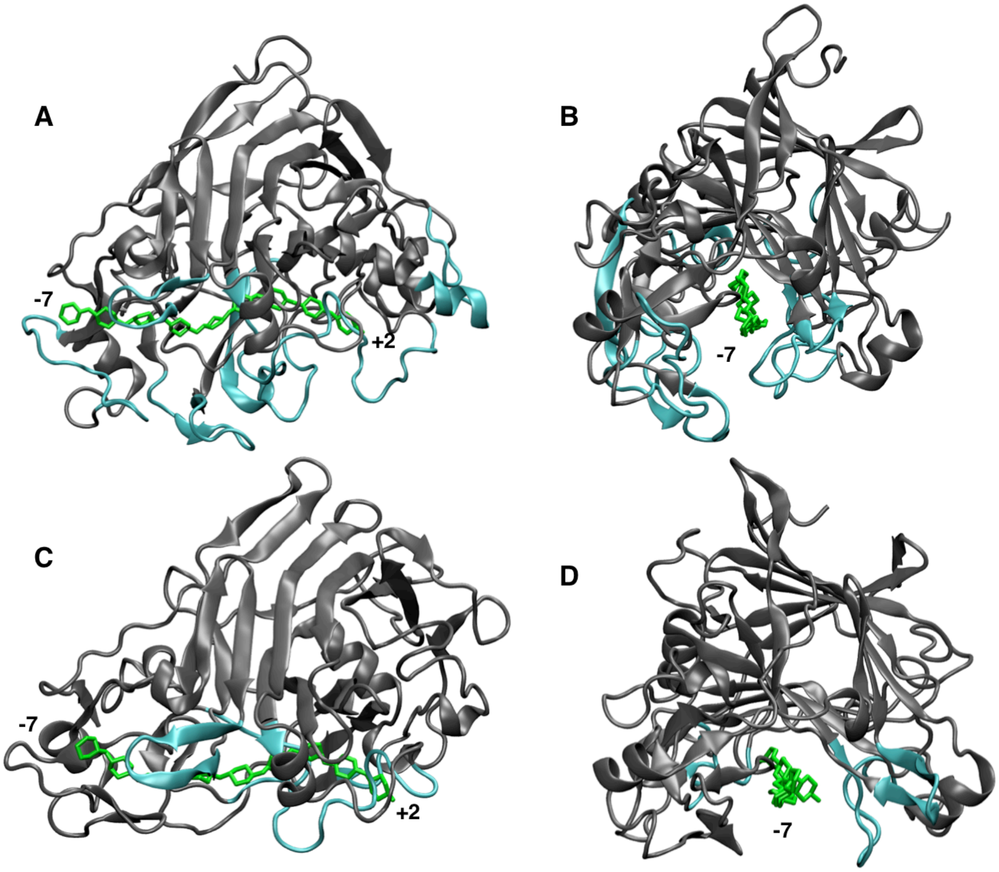


Figure 1.8: Comparison of Cel7A (A and B) and Cel7B (C and D) catalytic structures [25,59]. The entrance to the Cel7A tunnel is shown in panel B and the entrance to the Cel7B cleft is shown in panel D. The protein loop structures that form the tunnel in Cel7A and the shorter loop structures present in the Cel7B cleft are both shown in blue. The cellodextrin is shown in green.

Additionally, the binding tunnels and clefts in GHs are ubiquitously lined with aromatic amino acids to promote binding, processing, and stability in the cellodextrin [74]. The studies performed to date investigating the roles of these residues in GHs in addition to the current work's probe of these residues in GH7 Cel7A and Cel7B are further discussed in Chapter V.

The goal of the CD study is to examine structural and dynamical differences in the processive Cel7A and the non-processive Cel7B. Molecular-level simulations can offer insights to energetics and structure-function relationships necessary for catalytic binding and processivity in carbohydrate-active enzymes. An in-depth discussion can be found in Chapter V, wherein we attempt to define processivity in terms of cellodextrin and protein stability and geometries in the tunnel versus cleft conformation and investigate how the relative binding affinity and interactions between the cellodextrin and protein are impacted with aromatic mutations.

1.3 Role of Glycosylation in Cellulase Action

Glycosylation is an important and ubiquitous post-translational modification occurring not only in fungi, but in all kingdoms of life [75,76]. The most prevalent forms of protein glycans are *N*-glycosylation, where glycans attach to the β-amide group of an asparagine (N) residue in a N-X-Ser/Thr motif (where X is any amino acid except proline, Ser is serine, and Thr is threonine), and *O*-glycosylation, where glycans attach at the β-hydroxyl group of hydroxylysine, hydroxyproline, Ser, or Thr [77]. *N*-glycans, found in the CD of Cel7A are often larger carbohydrate moieties composed of an *N*-acetylglucosamine (GlcNAc) disaccharide directly attached to the protein side chain followed by branched carbohydrate oligomers. *O*-glycans, such as those found on the linker and CBM of Cel7A can range from smaller structures composed of mannose, glucose, galactose, or *N*-acetylgalactosamine to much longer branched structures of single or mixed carbohydrates [76,78]. Glycosylation is known to serve key biological roles in almost all biological processes such as signaling, protein secretion, protein stability, and proteolysis protection [41,77-89].

To our knowledge studies of the CBM to date have not examined the impact of natural or engineered *O*-glycosylation on binding affinity, only the roles of aromatic and polar residues [28-30,90]. There have been some experimental studies wherein *N*-glycosylation in the CBM or CD of various cellulases was altered with negative impacts to activity and binding affinity [91-93]. *O*-glycans were also added in a separate study but were too far away from the cellulose surface to impact any interactions [94]. Using molecular dynamics and thermodynamic integration methods, Chapter III investigates the impact that small *O*-glycan moieties in proximity to the cellulose surface, like those found by Harrison *et al.* [44] and shown in Figure 1.7, can have on binding affinity. The findings on the impact of glycans on CBMs can be more broadly applied to other organisms, and a more in-depth discussion follows in Chapters IV.

1.4 Summary and Outline of Thesis

In summary, cellulases can be utilized to achieve the benefits of biofuel production from lignocellulosic sources if throughput can be increased to commercial scale and current costs can be reduced [7,9,95]. Understanding the molecular-level actions cellulases use to deconstruct plant cell walls is a significant challenge that must be overcome to achieve efficiency in biofuel production [96], and this work contains two main thrusts to contribute to the efforts underway to address this challenge. First, we suggest a novel approach, manipulation of *O*-glycosylation, for improving binding affinity in the CBM, which could lead to overall activity improvements. By producing results consistent with prior experiments [29,30,97], we will show that thermodynamic integration is a useful tool for investigating the impact that amino acid mutations have on binding affinity. Since large aromatic amino acids such as tryptophan are similar in size to a mannose, we extended this method to mutated structures where simple *O*-glycans are added to the CBM. We will show that relative binding affinity can be enhanced by 3 to 6-fold over a non-glycosylated CBM with the addition of a single mannose or mannose disaccharide. Additionally, we found that adding additional glycans not identified by Harrision *et al.* [44] increased binding affinity by 140-fold relative to the non-glycosylated CBM.

These results moved us to then continue the CBM binding study to include glycoforms produced in other organisms such as yeasts, fungi, and even mammals. Fungi are often employed as model systems for studying the role of glycans on protein function because they impart a wide variety of glycosylation motifs between species. In Family 1 CBMs the role of glycans has not been fully elucidated. To probe glycan effects on CBM function across species, and thus different glycan motifs, we use simulation to identify potential structural or thermodynamic changes associated with the addition of glycan patterns observed in *Saccharomyces cerevisiae, T. reesei*, *Aspergillus niger*, and *Aspergillus awamori*. The changes in relative binding affinity to cellulose has also been calculated for each of the glycan patterns studied. We will show that again the results suggest that changing *O*-glycosylation patterns positively impacts binding affinity in Family 1 CBMs, with increases in the calculated binding affinity over a non-glycosylated CBM ranging from 1 to 4 kcal/mol. In our models, CBMs with linear glycans and sulfated glycans exhibit improved stability and higher binding over those with branched structures, with the nature of the carbohydrate moieties having a greater impact than covalent linkage. In some cases, the larger glycans affect the CBM structure, reducing initial gains to the binding affinity. Overall, this study highlights the fact that glycans should be carefully considered in protein engineering when host organism and external factors can affect glycosylation patterns. This work also demonstrates that simulation can provide molecular-level details important to glycoprotein structure, which is a useful step in rational engineering approaches.

The last project shifts focus to the catalytic domains in two *T. reesei* cellulases, Cel7A and Cel7B. In nature, cocktails of processive and non-processive cellulase enzymes convert cellulose to glucose. From structural studies, the consensus is that processive cellulases exhibit tunnels or deep clefts lined with aromatic and polar residues, whereas non-processive cellulases exhibit open clefts with fewer substrate contacts. However, removal of a single loop or the mutation of a single residue in the active binding area can dramatically affect enzyme processivity, suggesting that the molecular-level mechanism of processivity is not yet fully understood. To gain further insight into the differences between processive and non-processive cellulases, we examine the GH7 processive exoglucanase, Cel7A, and the non-processive endoglucanse, Cel7B from *T. reesei* with molecular simulation. We compare properties putatively related to processivity at each ligand binding site and the binding affinity changes for the mutation of four aromatic residues lining each tunnel to alanine. We observe similar energetic profiles and structural behavior in both systems at the catalytic sites, suggesting that the residues directly around the active site may not be directly associated with processivity in GH7s. However, the entrance and exits of the ligand binding sites exhibit significantly different interactions and binding affinities in Cel7A and Cel7B. In Cel7A, aromatic residue mutations at the entrance and center of the tunnel negatively impact binding and increase ligand fluctuations, indicating that these sites may impact ligand acquisition and potentially processivity. Conversely, weak protein-ligand interactions in the Cel7B cleft support the notion that lack of strong binding, especially at the cleft entrance may decrease processivity. Improvements in molecular-level understanding of differences within GH Families aid in efforts to determine how cocktails of cellulases have evolved to perform different functions in cellulosic degradation, the findings of which may be applied to improve enzymatic hydrolysis in biofuels production.

1.5 References

1. "U.S. Energy Facts Explained," US Energy Information Administration: **2010**; pp.

2. "Annual Energy Review 2010," U.S. Energy Information Administration: Washington, D.C., **2010**; pp.

3. S.P.S. Chundawat, G.T. Beckham, M.E. Himmel and B.E. Dale, "Deconstruction of Lignocellulosic Biomass to Fuels and Chemicals", *Annu. Rev. Chem. Biomol. Eng.* **2011**, 2, 6.1-6.25.

4. "Biomass Multi-Year Program Plan April 2012," U.S. Department of Energy: Washington, D.C. , **2012**; pp 206.

5. "Chu: President's 2013 Energy Budget Makes Critical Investments in Innovation, Clean Energy, and National Security," **2012** US Department of Energy: Place.

6. R. Schulz, B. Lindner, L. Petridis and J.C. Smith, "Scaling of Multimillion-Atom Biological Molecular Dynamics Simulation on a Petascale Supercomputer", *Journal of Chemical Theory and Computation* **2009**, 5 (10), 2798-2808.

7. M.E. Himmel, M.F. Ruth and C.E. Wyman, "Cellulase for commodity products from cellulosic biomass", *Current Opinion in Biotechnology* **1999**, 10 (4), 358-364.

8. Q. Xu, A. Singh and M.E. Himmel, "Perspectives and new directions for the production of bioethanol using consolidated bioprocessing of lignocellulose", *Current Opinion in Biotechnology* **2009**, 20 (3), 364-371.

9. M.E. Himmel, S.Y. Ding, D.K. Johnson, W.S. Adney, M.R. Nimlos, J.W. Brady and T.D. Foust, "Biomass recalcitrance: Engineering plants and enzymes for biofuels production", *Science* **2007**, 315 (5813), 804-807.

10. G.T. Beckham, J.F. Matthews, B. Peters, Y.J. Bomble, M.E. Himmel and M.F. Crowley, "Molecular-Level Origins of Biomass Recalcitrance: Decrystallization Free Energies for Four Common Cellulose Polymorphs", *Journal of Physical Chemistry B* **2011**, 115 (14), 4118-4127.

11. R.H. Atalla and D.L. Vanderhart, "NATIVE CELLULOSE - A COMPOSITE OF 2 DISTINCT CRYSTALLINE FORMS", *Science* **1984**, 223 (4633), 283-285.

12. J.F. Matthews, C.E. Skopec, P.E. Mason, P. Zuccato, R.W. Torget, J. Sugiyama, M.E. Himmel and J.W. Brady, "Computer simulations of microcrystalline cellulose Ib", *Carb. Res.* **2006**, 341, 138-152.

13. Y. Nishiyama, P. Langan and H. Chanzy, "Crystal structure and hydrogen-bonding system in cellulose 1 beta from synchrotron X-ray and neutron fiber diffraction", *J. Am. Chem. Soc.* **2002**, 124 (31), 9074-9082.

14. Y. Nishiyama, J. Sugiyama, H. Chanzy and P. Langan, "Crystal structure and hydrogen bonding system in cellulose 1 alpha", *J. Amer. Chem. Soc.* **2003**, 125, 14300-14306.

15. M.R. Nimlos, J.F. Matthews, M.F. Crowley, R.C. Walker, G. Chukkapalli, J.V. Brady, W.S. Adney, J.M. Clearyl, L.H. Zhong and M.E. Himmel, "Molecular modeling suggests induced fit of Family I carbohydrate-binding modules with a broken-chain cellulose surface", *Protein Eng. Des. Sel.* **2007**, 20 (4), 179-187.

16. Y. Nishiyama, "Structure and properties of the cellulose microfibril", *J. Wood Sci.* **2009**, 55 (4), 241-249.

17. C. Divne, J. Stahlberg, T. Reinikainen, L. Ruohonen, G. Pettersson, J.K.C. Knowles, T.T. Teeri and T.A. Jones, "The 3-dimensional crystal-structure of the catalytic core of Cellobiohydrolase-I from *Trichodermal reesei*", *Science* **1994**, 265 (5171), 524-528.

18. L.R. Lynd, P.J. Weimer, W.H. van Zyl and I.S. Pretorius, "Microbial cellulose utilization: Fundamentals and biotechnology", *Microbiol. Mol. Biol. Rev.* **2002**, 66 (3), 506-+.

19. H. Nevalainen, et al In Glycosylation of cellobiohydrolase I from Trichoderma reesei, TRICEL 97 Conference Carbohydrates from Trichoderma reesei and Other Microorganisms, Cambridge, UK, Ghent, Belgium, T.R.S.o. Cambridge, Ed. Cambridge, UK, Ghent, Belgium, **1997**.

20. C. Divne, I. Sinning, J. Stahlberg, G. Pettersson, M. Bailey, M. Siikaaho, E. Margollesclark, T. Teeri and T.A. Jones, "CRYSTALLIZATION AND PRELIMINARY-X-RAY STUDIES ON THE CORE PROTEINS OF CELLOBIOHYDROLASE-I AND ENDOGLUCANASE-I FROM TRICHODERMA-REESEI", *J. Mol. Biol.* **1993**, 234 (3), 905-907.

21. M. Srisodsuk, T. Reinikainen, M. Penttila and T.T. Teeri, "ROLE OF THE INTERDOMAIN LINKER PEPTIDE OF TRICHODERMA-REESEI CELLOBIOHYDROLASE-I IN ITS INTERACTION WITH CRYSTALLINE CELLULOSE", *J. Biol. Chem.* **1993**, 268 (28), 20756-20761.

22. T. Teeri, T. Reinikainen, L. Ruohonen, T.A. Jones and J.K.C. Knowles, "Domain function in *Trichoderma reesei* cellobiohydrolases", *J Biotechnol* **1992**, 24 (2), 169-176.

23. J. Stahlberg, G. Johansson and G. Pettersson, "A NEW MODEL FOR ENZYMATIC-HYDROLYSIS OF CELLULOSE BASED ON THE 2-DOMAIN STRUCTURE OF CELLOBIOHYDROLASE-I", *Bio-Technology* **1991**, 9 (3), 286-290.

24. G.T. Beckham, J.F. Matthews, Y.J. Bomble, L.T. Bu, W.S. Adney, M.E. Himmel, M.R. Nimlos and M.F. Crowley, "Identification of Amino Acids Responsible for Processivity in a Family 1 Carbohydrate-Binding Module from a Fungal Cellulase", *J. Phys. Chem. B* **2009**, 114 (3), 1447-1453.

25. C. Divne, J. Stahlberg, T.T. Teeri and T.A. Jones, "High-resolution crystal structures reveal how a cellulose chain is bound in the 50 angstrom long tunnel of cellobiohydrolase I from Trichoderma reesei", *J. Mol. Biol.* **1998**, 275 (2), 309-325.

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

27. J. Lehtio, J. Sugiyama, M. Gustavsson, L. Fransson, M. Linder and T.T. Teeri, "The binding specificity and affinity determinants of family 1 and family 3 cellulose binding modules", *Proc. Natl. Acad. Sci. U. S. A.* **2003**, 100 (2), 484-489.

28. M. Linder, G. Lindeberg, T. Reinikainen, T.T. Teeri and G. Pettersson, "The difference in affinity between 2 fungal cellulose-binding domains is dominated by a single amino-acid substitution", *FEBS Lett.* **1995**, 372 (1), 96-98.

29. M. Linder, M.L. Mattinen, M. Kontteli, G. Lindeberg, J. Stahlberg, T. Drakenberg, T. Reinikainen, G. Pettersson and A. Annila, "Identification of functionally important amino-acids in the cellulose-binding domain of *Trichoderma reesei* Cellobiohydrolase I", *Protein Sci.* **1995**, 4 (6), 1056-1064.

30. S. Takashima, M. Ohno, M. Hidaka, A. Nakamura and H. Masaki, "Correlation between cellulose binding and activity of cellulose-binding domain mutants of Humicola grisea cellobiohydrolase 1", *FEBS Lett.* **2007**, 581 (30), 5891-5896.

31. T. Reinikainen, L. Ruohonen, T. Nevanen, L. Laaksonen, P. Kraulis, T.A. Jones, J.K.C. Knowles and T.T. Teeri, "INVESTIGATION OF THE FUNCTION OF MUTATED CELLULOSE-BINDING DOMAINS OF TRICHODERMA-REESEI CELLOBIOHYDROLASE-I", *Proteins* **1992**, 14 (4), 475-482.

32. N.R. Gilkes, B. Henrissat, D.G. Kilburn, R.C. Miller and R.A.J. Warren, "DOMAINS IN MICROBIAL BETA-1,4-GLYCANASES - SEQUENCE CONSERVATION, FUNCTION, AND ENZYME FAMILIES", *Microbiological Reviews* **1991**, 55 (2), 303-315.

33. I. von Ossowski, J.T. Eaton, M. Czjzek, S.J. Perkins, T.P. Frandsen, M. Schulein, P. Panine, B. Henrissat and V. Receveur-Brechot, "Protein disorder: Conformational distribution of the flexible linker in a chimeric double cellulase", *Biophys. J.* **2005**, 88 (4), 2823-2832.

34. X. Zhao, T.R. Rignall, C. McCabe, W.S. Adney and M.E. Himmel, "Molecular simulation evidence for processive motion of Trichoderma reesei Cel7A during cellulose depolymerization", *Chem. Phys. Lett.* **2008**, 460 (1-3), 284-288.

35. T. Reinikainen, M. Srisodsuk, A. Jones and T.T. Teeri, "Enzymatic hydrolysis of crystalline cellulose by *Trichoderma reesei* cellobiohydrolase I", *Protein Eng* **1993**, 6, 49-49.

36. P.M. Abuja, I. Pilz, M. Claeyssens and P. Tomme, "DOMAIN-STRUCTURE OF CELLOBIOHYDROLASE-II AS STUDIED BY SMALL-ANGLE X-RAY-SCATTERING - CLOSE RESEMBLANCE TO CELLOBIOHYDROLASE-I", *Biochem. Biophys. Res. Commun.* **1988**, 156 (1), 180-185.

37. P.M. Abuja, M. Schmuck, I. Pilz, P. Tomme, M. Claeyssens and H. Esterbauer, "STRUCTURAL AND FUNCTIONAL DOMAINS OF CELLOBIOHYDROLASE-I FROM TRICHODERMA-REESEI - A SMALL-ANGLE X-RAY-SCATTERING STUDY OF THE INTACT ENZYME AND ITS CORE", *Eur. Biophys. J. Biophys. Lett.* **1988**, 15 (6), 339-342.

38. D.K.Y. Poon, S.G. Withers and L.P. McIntosh, "Direct demonstration of the flexibility of the glycosylated proline-threonine linker in the Cellulomonas fimi xylanase Cex through NMR spectroscopic analysis", *J. Biol. Chem.* **2007**, 282 (3), 2091-2100.

39. G.K. Sonan, V. Receveur-Brechot, C. Duez, N. Aghajari, M. Czjzek, R. Haser and C. Gerday, "The linker region plays a key role in the adaptation to cold of the cellulase from an Antarctic bacterium", *Biochem. J.* **2007**, 407, 293-302.

40. H. Shen, M. Schmuck, I. Pilz, N.R. Gilkes, D.G. Kilburn, R.C. Miller and R.A.J. Warren, "Deletion of the linker connection the catalytic and CBD of endoglucanase-A (CenA) of Cellulomonas fimi alters its conformation and catalytic activity", *J. Biol. Chem.* **1991**, 266 (17), 11335-11340.

41. V. Receveur, M. Czjzek, M. Schulein, P. Panine and B. Henrissat, "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), 40887-40892.

42. C. Boisset, R. Borsali, M. Schulein and B. Henrissat, "DYNAMIC LIGHT-SCATTERING STUDY OF THE 2-DOMAIN STRUCTURE OF HUMICOLA INSOLENS ENDOGLUCANASE-V", *FEBS Lett.* **1995**, 376 (1-2), 49-52.

43. I. Noach, F. Frolow, O. Alber, R. Lamed, L.J.W. Shimon and E.A. Bayer, "Intermodular Linker Flexibility Revealed from Crystal Structures of Adjacent Cellulosomal Cohesins of Acetivibrio celluloyticus", *J. Mol. Biol.* **2009**, 391 (1), 86-97.

44. M.J. Harrison, A.S. Nouwens, D.R. Jardine, N.E. Zachara, A.A. Gooley, H. Nevalainen and N.H. Packer, "Modified glycosylation of cellobiohydrolase I from a high cellulase-producing mutant strain of Trichoderma reesei", *Eur. J. Biochem.* **1998**, 256 (1), 119-127.

45. L. Zhong, J.F. Matthews, M.F. Crowley, T. Rignall, C. Talon, J.M. Cleary, R.C. Walker, G. Chukkapalli, C. McCabe, M.R. Nimlos, C.L. Brooks, M.E. Himmel and J.W. Brady, "Interactions of the complete cellobiohydrolase I from Trichodera reesei with microcrystalline cellulose I beta", *Cellulose* **2008**, 15 (2), 261-273.

46. C.L. Ting, D.E. Makarov and Z.G. Wang, "A Kinetic Model for the Enzymatic Action of Cellulase", *J. Phys. Chem. B* **2009**, 113 (14), 4970-4977.

47. G.T. Beckham, Y.J. Bomble, J.F. Matthews, C.B. Taylor, M.G. Resch, J.M. Yarbrough, S.R. Decker, L.T. Bu, X.C. Zhao, C. McCabe, J. Wohlert, M. Bergenstrahle, J.W. Brady, W.S. Adney, M.E. Himmel and M.F. Crowley, "The O-Glycosylated Linker from the Trichoderma reesei Family 7 Cellulase Is a Flexible, Disordered Protein", *Biophys. J.* **2010**, 99 (11), 3773-3781.

48. J.P.M. Hui, P. Lanthier, T.C. White, S.G. McHugh, M. Yaguchi, R. Roy and P. Thibault, "Characterization of cellobiohydrolase I (Cel7A) glycoforms from extracts of Trichoderma reesei using capillary isoelectric focusing and electrospray mass spectrometry", *J. Chromatogr. B* **2001**, 752 (2), 349-368.

49. B. Nidetzky, W. Steiner, M. Hayn and M. Claeyssens, "CELLULOSE HYDROLYSIS BY THE CELLULASES FROM TRICHODERMA-REESEI - A NEW MODEL FOR SYNERGISTIC INTERACTION", *Biochem. J.* **1994**, 298, 705-710.

50. P. Valjamae, V. Sild, A. Nutt, G. Pettersson and G. Johansson, "Acid hydrolosis of bacterial cellulose reveals different modes of synergistic action between cellobiohydrolase I and endoglucanase I", *Eur. J. Biochem.* **1999**, 266 (2), 327-334.

51. L.F. Mackenzie, G. Sulzenbacher, C. Divne, T.A. Jones, H.F. Woldike, M. Schulein, S.G. Withers and G.J. Davies, "Crystal structure of the family 7 endoglucanase I (Cel7B) from Humicola insolens at 2.2 angstrom resolution and identification of the catalytic nucleophile by trapping of the covalent glycosyl-enzyme intermediate", *Biochem. J.* **1998**, 335, 409-416.

52. B.K. Barr, Y.L. Hsieh, B. Ganem and D.B. Wilson, "Identification of two functionally different classes of exocellulases", *Biochemistry* **1996**, 35 (2), 586-592.

53. V.G.H. Eijsink, G. Vaaje-Kolstad, K.M. Varum and S.J. Horn, "Towards new enzymes for biofuels: lessons from chitinase research", *Trends Biotechnol.* **2008**, 26 (5), 228-235.

54. T. Eriksson, I. Stals, A. Collen, F. Tjerneld, M. Claeyssens, H. Stalbrand and H. Brumer, "Heterogeneity of homologously expressed Hypocrea jecorina (Trichoderma reesei) Cel7B catalytic module", *Eur. J. Biochem.* **2004**, 271 (7), 1266-1276.

55. M.E. Himmel and E.A. Bayer, "Lignocellulose conversion to biofuels: current challenges, global perspectives", *Current Opinion in Biotechnology* **2009**, 20 (3), 316-317.

56. S.J. Horn, P. Sikorski, J.B. Cederkvist, G. Vaaje-Kolstad, M. Sorlie, B. Synstad, G. Vriend, K.M. Varum and V.G.H. Eijsink, "Costs and benefits of processivity in enzymatic degradation of recalcitrant polysaccharides", *Proc. Natl. Acad. Sci.* **2006**, 103 (48), 18089-18094.

57. K. Igarashi, A. Koivula, M. Wada, S. Kimura, M. Penttila and M. Samejima, "High Speed Atomic Force Microscopy Visualizes Processive Movement of Trichoderma reesei Cellobiohydrolase I on Crystalline Cellulose", *J. Biol. Chem.* **2009**, 284 (52), 36186-36190.

58. J. Karlsson, M. Siika-aho, M. Tenkanen and F. Tjerneld, "Enzymatic properties of the low molecular mass endoglucanases Cel12A (EG III) and Cel45A (EG V) of Trichoderma reesei", *J. Biotechnol.* **2002**, 99 (1), 63-78.

59. G.J. Kleywegt, J.Y. Zou, C. Divne, G.J. Davies, I. Sinning, J. Stahlberg, T. Reinikainen, M. Srisodsuk, T.T. Teeri and T.A. Jones, "The crystal structure of the catalytic core domain of endoglucanase I from Trichoderma reesei at 3.6 angstrom resolution, and a comparison with related enzymes", *J. Mol. Biol.* **1997**, 272 (3), 383-397.

60. A. Koivula, T. Kinnari, V. Harjunpaa, L. Ruohonen, A. Teleman, T. Drakenberg, J. Rouvinen, T.A. Jones and T.T. Teeri, "Tryptophan 272: an essential determinant of crystalline cellulose degradation by Trichoderma reesei cellobiohydrolase Cel6A", *FEBS Lett.* **1998**, 429 (3), 341-346.

61. M. Penttila, P. Lehtovaara, H. Nevalainen, R. Bhikhabhai and J. Knowles, "HOMOLOGY BETWEEN CELLULASE GENES OF TRICHODERMA-REESEI - COMPLETE NUCLEOTIDE-SEQUENCE OF THE ENDOGLUCANASE-I GENE", *Gene* **1986**, 45 (3), 253-263.

62. A.J. Ragauskas, C.K. Williams, B.H. Davison, G. Britovsek, J. Cairney, C.A. Eckert, W.J. Frederick, J.P. Hallett, D.J. Leak, C.L. Liotta, J.R. Mielenz, R. Murphy, R. Templer and T. Tschaplinski, "The path forward for biofuels and biomaterials", *Science* **2006**, 311 (5760), 484-489.

63. T.T. Teeri, A. Koivula, M. Linder, G. Wohlfahrt, C. Divne and T.A. Jones, "Trichoderma reesei cellobiohydrolases: why so efficient on crystalline cellulose?", *Biochemical Society Transactions* **1998**, 26 (2), 173-178.

64. M. Claeyssens, H. Vantilbeurgh, J.P. Kamerling, J. Berg, M. Vrsanska and P. Biely, "STUDIES OF THE CELLULOLYTIC SYSTEM OF THE FILAMENTOUS FUNGUS TRICHODERMA-REESEI QM-9414 - SUBSTRATE-SPECIFICITY AND TRANSFER ACTIVITY OF ENDOGLUCANASE-I", *Biochem. J.* **1990**, 270 (1), 251-256.

65. G.D. Rose, L.M. Glerasch and J.A. Smith, "Turns in Peptides and Proteins," In Advances in Protein Chemistry; J.T.E. C.B. Anfinsen and M.R. Frederic, Ed.; Academic Press: **1985**; Vol. Volume 37; pp 1-109.

66. R. Stern and M.J. Jedrzejas, "Carbohydrate Polymers at the Center of Life's Origins: The Importance of Molecular Processivity", *Chemical Reviews* **2008**, 108 (12), 5061-5085.

67. G.T. Beckham and M.F. Crowley, "Examination of the alpha-Chitin Structure and Decrystallization Thermodynamics at the Nanoscale", *Journal of Physical Chemistry B* **2011**, 115 (15), 4516-4522.

68. C.M. Payne, M.E. Himmel, M.F. Crowley and G.T. Beckham, "Decrystallization of Oligosaccharides from the Cellulose I beta Surface with Molecular Simulation", *J. Phys. Chem. Lett.* **2011**, 2 (13), 1546-1550.

69. G.J. Davies, A.M. Brzozowski, M. Dauter, A. Varrot and M. Schulein, "Structure and function of Humicola insolens family 6 cellulases: structure of the endoglucanase, Cel6B, at 1.6 angstrom resolution", *Biochem. J.* **2000**, 348, 201-207.

70. A. Meinke, H.G. Damude, P. Tomme, E. Kwan, D.G. Kilburn, R.C. Miller, R.A.J. Warren and N.R. Gilkes, "ENHANCEMENT OF THE ENDO-BETA-1,4-GLUCANASE ACTIVITY OF AN EXOCELLOBIOHYDROLASE BY DELETION OF A SURFACE LOOP", *J. Biol. Chem.* **1995**, 270 (9), 4383-4386.

71. I. von Ossowski, J. Stahlberg, A. Koivula, K. Piens, D. Becker, H. Boer, R. Harle, M. Harris, C. Divne, S. Mahdi, Y.X. Zhao, H. Driguez, M. Claeyssens, M.L. Sinnott and T.T. Teeri, "Engineering the exo-loop of *Trichoderma reesei* cellobiohydrolase, Cel7A. A comparison with *Phanerochaete chrysosporium* Cel7D", *J. Mol. Biol.* **2003**, 333 (4), 817-829.

72. T. Parkkinen, A. Koivula, J. Vehmaanpera and J. Rouvinen, "Crystal structures of Melanocarpus albomyces cellobiohydrolase Ce17B in complex with cello-oligomers show high flexibility in the substrate binding", *Protein Sci.* **2008**, 17 (8), 1383-1394.

73. M. Kurasin and P. Valjamae, "Processivity of Cellobiohydrolases Is Limited by the Substrate", *J. Biol. Chem.* **2011**, 286 (1), 169-177.

74. C.M. Payne, Y. Bomble, C.B. Taylor, C. McCabe, M.E. Himmel, M.F. Crowley and G.T. Beckham, "Multiple Functions of Aromatic-Carbohydrate Interactions in a Processive Cellulase Examined with Molecular Simulation", *J. Biol. Chem.* **2011**, 286 (47), 41028-41035.

75. G.T. Beckham, Z. Dai, J.F. Matthews, M. Momany, C.M. Payne, W.S. Adney, S.E. Baker and M.E. Himmel, "Harnessing glycosylation to improve cellulase activity", *Current Opinion in Biotechnology* **2012**, (0).

76. A. Varki, *Essentials of Glycobiology*. Cold Springs Harbor Laboratory Press: Cold Springs Harbor, NY, USA, **2009**.

77. N. Deshpande, M.R. Wilkins, N. Packer and H. Nevalainen, "Protein glycosylation pathways in filamentous fungi", *Glycobiology* **2008**, 18 (8), 626-637.

78. K. Ohtsubo and J.D. Marth, "Glycosylation in cellular mechanisms of health and disease", *Cell* **2006**, 126 (5), 855-867.

79. M. Goto, "Protein O-glycosylation in fungi: Diverse structures and multiple functions", *Biosci. Biotechnol. Biochem.* **2007**, 71 (6), 1415-1427.

80. Y. Mazola, G. Chinea and A. Musacchio, "Integrating Bioinformatics Tools to Handle Glycosylation", *PLoS Comput. Biol.* **2011**, 7 (12), 7.

81. A.M. Sinclair and S. Elliott, "Glycoengineering: The effect of glycosylation on the properties of therapeutic proteins", *Journal of Pharmaceutical Sciences* **2005**, 94 (8), 1626-1635.

82. K. Drickamer and M.E. Taylor, "Evolving views of protein glycosylation", *Trends Biochem.Sci.* **1998**, 23 (9), 321-324.

83. G. Lauc and V. Zoldos, "Protein glycosylation-an evolutionary crossroad between genes and environment", *Mol. Biosyst.* **2010**, 6 (12), 2373-2379.

84. A. Varki, "BIOLOGICAL ROLES OF OLIGOSACCHARIDES - ALL OF THE THEORIES ARE CORRECT", *Glycobiology* **1993**, 3 (2), 97-130.

85. R.J. Sola and K. Griebenow, "Glycosylation of Therapeutic Proteins An Effective Strategy to Optimize Efficacy", *Biodrugs* **2010**, 24 (1), 9-21.

86. R.A. Dwek, "Glycobiology: More Functions for Oligosaccharides", *Science* **1995**, 269 (5228), 1234-1235.

87. P.M. Rudd, T. Elliott, P. Cresswell, I.A. Wilson and R.A. Dwek, "Glycosylation and the Immune System", *Carbohydrates and Glycobiology* **2001**, 291 (5512), 2370-2376.

88. G.W. Hart and R.J. Copeland, "Glycomics Hits the Big Time", *Cell* **2010**, 143 (5), 672-676.

89. S.I. Van Kasteren, H.B. Kramer, D.P. Gamblin and B.G. Davis, "Site-selective glycosylation of proteins: creating synthetic glycoproteins", *Nature Protocols* **2007**, 2 (12), 3185-3194.

90. A.L. Creagh, E. Ong, E. Jervis, D.G. Kilburn and C.A. Haynes, "Binding of the cellulose-binding domain of exoglucanase Cex from Cellulomonas fimi to insoluble microcrystalline cellulose is entropically driven", *Proc. Natl. Acad. Sci. U. S. A.* **1996**, 93 (22), 12229-12234.

91. A.B. Boraston, R.A.J. Warren and D.G. Kilburn, "Glycosylation by Pichia pastoris decreases the affinity of a family 2a carbohydrate-binding module from Cellulomonas fimi: a functional and mutational analysis", *Biochem. J.* **2001**, 358, 423-430.

92. W.S. Adney, T. Jeoh, G.T. Beckham, Y.C. Chou, J.O. Baker, W. Michener, R. Brunecky and M.E. Himmel, "Probing the role of N-linked glycans in the stability and activity of fungal cellobiohydrolases by mutational analysis", *Cellulose* **2009**, 16 (4), 699-709.

93. T. Jeoh, W. Michener, M.E. Himmel, S.R. Decker and W.S. Adney, "Implications of cellobiohydrolase glycosylation for use in biomass conversion", *Biotechnol. Biofuels* **2008**, 1, 12.

94. A.B. Boraston, L.E. Sandercock, R.A.J. Warren and D.G. Kilburn, "O-glycosylation of a recombinant carbohydrate-binding module mutant secreted by Pichia pastoris", *J. Mol. Microbiol. Biotechnol.* **2003**, 5 (1), 29-36.

95. C. Schubert, "Can biofuels finally take center stage?", *Nat. Biotechnol.* **2006**, 24 (7), 777-784.

96. G.T. Beckham, Y.J. Bomble, E.A. Bayer, M.E. Himmel and M.F. Crowley, "Applications of computational science for understanding enzymatic deconstruction of cellulose", *Current Opinion in Biotechnology* **2011**, 22 (2), 231-238.

97. M. Linder and T.T. Teeri, "The roles and function of cellulose-binding domains", *J. Biotechnol.* **1997**, 57 (1-3), 15-28.

CHAPTER 2

BACKGROUND AND THEORY

2.1 General Methodology: Molecular Dynamics

Molecular dynamics (MD) simulations have been performed to calculate thermodynamic and energetic properties that will further the understanding of the molecular-level actions that cellulases employ to bind with cellulose. In MD, Newton's equations of motion are solved numerically by calculating the forces on each atom in the system. The continuous equations are discretized and evaluated at regular time intervals, or timesteps. The force acting on each atom is derived as the gradient of the potential energy, $U(r_1 \ldots r_N)$ with respect to the displacement of the atoms, as shown in equation 2.1.

$$F_i = m_i a_i = m_i \frac{d^2 r_i}{dt^2} = -\nabla_{r_i} U(r_i \ldots, r_N) \tag{2.1}$$

Thus, a potential energy function, or force field, is defined to calculate U and its derivatives from the coordinates corresponding to a molecule's structure. The force field describes the energy of each conformation of the molecules and the interactions between them [1,2]. The mathematical form is typically separated into three intramolecular interactions, which include bond stretching, angle bending, and angle rotation, and intermolecular non-bonded interactions, which include, but are not limited to, electrostatic interactions and van der Waals pair potentials. The parameters used in the force field, such as the equilibrium bond length and angle, atomic charges, and radius between atom pairs, are generally developed using experimental values and detailed quantum-mechanics based calculations on small molecules or systems [2]. The summation of energetic changes associated with bonds and angles deviating from the "equilibrium" values determined by the parameters (bonded interactions) and atoms moving closer or farther apart (non-bonded interactions) forms the basis for the mathematical form of a force field [2].

Force fields vary in complexity depending on the type and size of the system of interest. There are three common types (1) all-atom force fields, which model each atom independently and explicitly, (2) united-atom force fields, which incorporate hydrogen atoms into the atoms they are bonded to, and (3) coarse-grained force fields, which group atoms together into larger components [2]. The goal when developing a force field is that it accurately reproduces thermodynamic and thermophysical properties of interest and contains flexible and transferable parameters for modeling related molecules (e.g., all α-carbons in a protein backbone), all while allowing for efficient computation [2].

All-atom potential force fields are used to describe the potential energy of the biological systems described in Chapter I. The general form of the CHARMM27 and CHARMM36 force field [1,3,4] used in this study is given by equation 2.2 [5]

$$\begin{aligned} U\left(\vec{R}\right) = & \sum_{\text{bonds}} K_b\left(b-b_0\right)^2 + \sum_{\text{angles}} K_\theta\left(\theta-\theta_0\right)^2 + \sum_{\text{dihedrals}} K_\varphi(1+\cos(n\varphi-\delta)) \\ & + \sum_{\text{Urey-Bradley}} K_{UB}\left(S-S_0\right)^2 + \sum_{\text{impropers}} K_\omega\left(\omega-\omega_0\right) + \sum_{\text{residues}} U_{CMAP}\left(\varphi,\psi\right) \\ & + \sum_{\substack{\text{non-bonded} \\ \text{pairs}}} \left\{ \varepsilon_{ij}^{\min}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} \right\} \end{aligned} \tag{2.2}$$

and includes six intramolecular terms: a bond stretching term (*b*), a bond angle bending term (*θ*), a dihedral angle term (*φ*) term, which is fit to bond rotational data and includes a phase shift (*δ*), Urey-Bradley (*UB*) and improper angle (*ω*) terms, which are fit to vibrational spectra and out of plane motion data, and the backbone torsional correction

(CMAP) term, a 2-D spline-based dihedral term based on *ab initio* quantum mechanics calculations and designed to correct systematic errors in the description of a protein backbone in the force field [1]. The K values represent the respective force constants and the variables with subscript 0 represent the equilibrium values. A Lennard-Jones 12-6 potential is used to describe non-bonded interactions, where ε_{ij}^{min} is the well depth, r_{ij} is the interatomic distance, and σ_{ij} is the Lennard Jones radii. The values for ε_{ij}^{min} and σ_{ij} for the interacting atoms are calculated by using the geometric mean and arithmetic mean, respectively, of the values of the individual atoms; thus $\varepsilon_{ij}^{min} = (\varepsilon_{ii}^{min}\ \varepsilon_{jj}^{min})^{1/2}$ and $\sigma_{ij} = (\sigma_{ii} + \sigma_{jj})/2$ [1]. Coulombic interactions are also included in the non-bonded terms, where q_i is the partial charge and ε and ε_0 the relative dielectric constant and permittivity of free space, respectively.

Simulations are set-up in a similar manner to real experiments where samples are prepared and properties of interest are measured once a system reaches steady state; first a sample system is selected, Newton's equation of motion are solved using the force field to describe the potential energy until the system reaches equilibrium, and then observables are measured [6]. Additional properties of interest can then be calculated from the atomistic trajectory information.

The molecular dynamics simulations reported herein are performed in the canonical ensemble in which the number of particles (N), volume (V), and temperature (T) are held fixed. The canonical ensemble partition function, $Q(N,V,T)$, and then the Helmholtz free energy are expressed by [10]

$$Q(N,V,T) = \sum_{U} \Omega(N,V,U) e^{-\beta U(T,V)} \tag{2.3}$$

$$A(N,V,T) = -kT\ln Q(N,V,T) \tag{2.4}$$

where *Ω(N,V,U)* is the degeneracy of the system and β is equal to $1/k_BT$ and where k_B is the Boltzmann constant.

The Gibbs free energy has natural independent variables in the isothermal-isobaric ensemble (NPT); however, the high value of the pressure tensor required for maintaining the system box size causes the cellulose surface to buckle during simulation, therefore the canonical ensemble in which the Helmholtz free energy has natural independent variables was used as the starting point to obtain the Gibbs free energy, *ΔΔG.* For a closed system, the following relations are used [11]

$$\begin{aligned} &A \equiv U - TS \\ &dA = dU - TdS \text{ (constant N,V,T)} \\ &G \equiv U + pV - TS = A + pV \\ &dG = dA + dp(V) \text{ (constant N,V,T)} \end{aligned} \tag{2.5-2.8}$$

For the systems of interest, the compressibility of water at 1 atm and 300 K is small compared to the contribution of *dA*, which allows the approximation: $dG = dA$. In statistical mechanics terms equation 2.7 can be written [11]:

$$G = A + \beta\left(\frac{\partial \ln Q}{\partial \ln V}\right)_{T,N} \tag{2.9}$$

The order of magnitude of β is 10^{-21} J, and the derivative of the natural log of the partition function, *Q,* is not large enough to overcome this, again demonstrating that *dG = dA* can be used in this study.

2.2 Thermodynamic Integration and Relative Binding Free Energy

Having established that the Gibbs free energy can be estimated from the Helmholtz free energy, the relative difference in the Gibbs free energy can be calculated using the potential energy of the system. Gibbs free energy is a state function and thus independent of the thermodynamic path; it is therefore possible to compute relative free energy changes for processes by defining a non-physical pathway that is tractable to simulation. The current study seeks to use this method to calculate relative free energy of binding using equation 2.10, illustrated by the simple schematic of bound and free enzymes in explicit water is shown in figure 2.1.

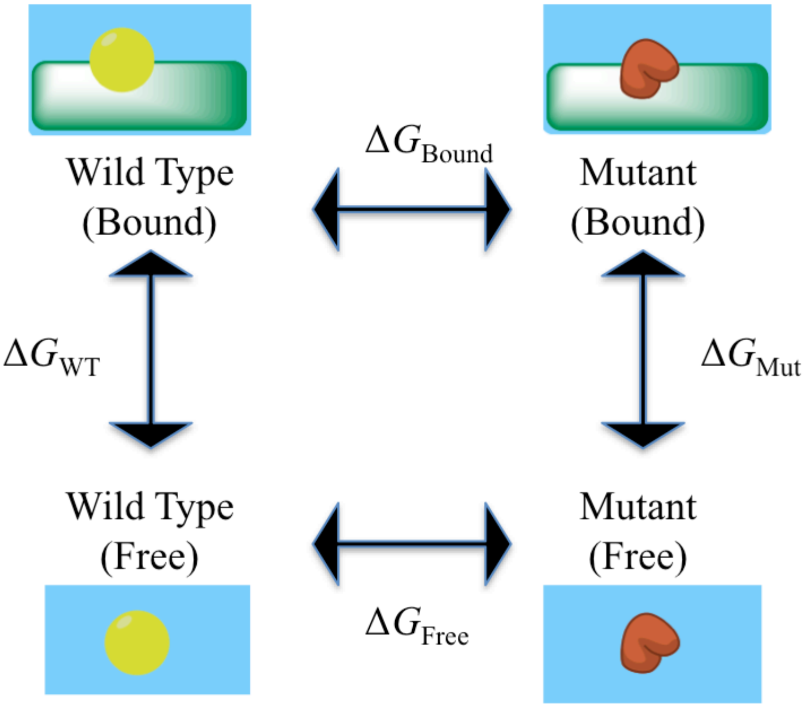


Figure 2.1: Graphical representation of the thermodynamic paths used to calculated $\Delta\Delta G$ in experiment (top to bottom) and alchemically via computation (right to left). *WT* represents the wild type and *Mut* represents the mutation.

While the goal is to calculate the free energy from the difference in the individual wild type (yellow) and mutant (red) free energy, ΔG_{Mut} - ΔG_{WT}, this is infeasible to accomplish directly, as the calculation would require the addition of waters in void space left by the disappearing surface, changing the overall composition of the system and the number of degrees of freedom. However, while the alchemical transformation of the bound wild type to mutant and the free wild type to mutant to obtain $\Delta G_{\mathrm{Bound}}$ and ΔG_{Free}, respectively, is unphysical, it is accessible via computation. Thus we can calculate the relative binding energy from the closed thermodynamic cycle:

$$\begin{aligned}\Delta\Delta G_{\mathrm{Binding}} = \Delta\Delta G_{\mathrm{TOT}} &= \left\{\Delta G_{\mathrm{Mut(Bound-Free)}} - \Delta G_{\mathrm{WT(Bound-Free)}}\right\}_{\mathrm{Actual}} \\ &= \left\{\Delta G_{\mathrm{Bound(Mut-WT)}} - \Delta G_{\mathrm{Free(Mut-WT)}}\right\}_{\mathrm{Alchemical}}\end{aligned} \tag{2.10}$$

A scaling factor must be introduced to convert from the starting state to the ending state; this calculation cannot be accurately or efficiently performed in one step due to the large number of factors that will be changing in the potential function. For a system with N particles and a potential energy U, a coupling parameter, λ, is introduced to relate U of the wild type and mutated systems calculated in MD simulations, where $\lambda = 0$ represents the wild type and $\lambda = 1$ represents the mutation. The potential energy of the system becomes [1,6]:

$$U(\lambda) = U_0 + (1-\lambda)U_{\mathrm{WT}} + \lambda U_{\mathrm{Mut}} \tag{2.11}$$

where U_0 represents the portion of the potential energy that does not change, U_{WT} represents the portion of the potential energy specific to the wild type, and U_{Mut} represents the portion of the potential energy specific to the mutation.

The partition function for this system is now also a function of λ and is determined by [6],

$$Q(N,V,T,\lambda) = \frac{1}{\Lambda^{3N} N!} \int d\bar{r}^N e^{-\beta U(\lambda)} \tag{2.12}$$

The derivative of the free energy in equation 2.13 above based on the potential energy becomes

$$\begin{aligned}\left(\frac{\partial G(\lambda)}{\partial \lambda}\right)_{N,P,T} \approx \left(\frac{\partial A(\lambda)}{\partial \lambda}\right)_{N,V,T} &= -\frac{1}{\beta}\frac{\partial}{\partial \lambda}\ln Q(N,V,T,\lambda) \\ &= -\frac{1}{\beta Q(N,V,T,\lambda)}\frac{\partial Q(N,V,T,\lambda)}{\partial \lambda} \\ &= \frac{\int d\bar{r}^N (\partial U(\lambda)/\partial\lambda) e^{-\beta U(\lambda)}}{\int d\bar{r}^N e^{-\beta U(\lambda)}} \\ &= \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda \end{aligned} \tag{2.13}$$

where $\left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda$ is the ensemble average of the system over all λ values. The free energy difference between the wild type and mutant systems can be obtained by integration of equation 2.13 [6]:

$$G_{\text{Mut}} - G_{\text{WT}} = \int_{\lambda=0}^{\lambda=1} d\lambda \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda \tag{2.14}$$

Incorporation of the coupling parameter into the force-field, or, potential energy function, is based on the definition of the wild type and mutant with relation to λ, i.e., each term in equation 2.2 is linearly related to λ via equation 2.11. In the case of the simple example from Figure 2.1, at λ=0 all internal, or bonded, terms (bond, angle bending, dihedral) will be "full" for the wild type and zero at λ=1. Incorporation of λ into the non-bonded terms often requires more effort than the bonded terms to ensure that the calculations remain stable and efficient [12]. Using a general form of the non-bonded pair terms in the potential energy function in equation 2.2 and consolidating the repulsive and attractive Lennard-Jones terms (ε_{ij}, σ_{ij}) to simple terms A and B, the coupling parameter can be implemented into the van der Waals potential function by [13]:

$$U_{\text{vdw}}(\lambda) = U_{\text{vdw},0} + (1-\lambda)\left(\frac{A_{\text{WT}}}{r_{ij}^{12}} - \frac{B_{\text{WT}}}{r_{ij}^{6}}\right) + \lambda\left(\frac{A_{\text{Mut}}}{r_{ij}^{12}} - \frac{B_{\text{Mut}}}{r_{ij}^{6}}\right) \qquad (2.15)$$

The electrostatic potential function in equation incorporates λ similarly:

$$U_{\text{elec}}(\lambda) = U_{\text{elec},0} + (1-\lambda)\left(\frac{q_{i,\text{WT}}q_{j,\text{WT}}}{r_{ij}}\right) + \lambda\left(\frac{q_{i,\text{Mut}}q_{j,\text{Mut}}}{r_{ij}}\right) \qquad (2.16)$$

Implementing the coupling parameter into the potential functions using a linear relation, while simple, can lead to issues at the endpoints where large changes in the forces can occur as a result of the steep repulsive terms in the Lennard-Jones parameters. Singularities can also arise at the endpoints where wild type and mutant atoms overlap, that is, where $r_{ij} \to 0$. To overcome the singularities and high repulsive terms, soft-core potential functions can be included for both van der Waals and electrostatic interactions. In the simulations reported herein, the partial-shift or separated shift soft core potential (PSSP) was implemented and adjusts the distance between two atoms by [1]:

$$r_{new} = \sqrt{r^2 + \alpha f(\lambda)} \qquad (2.17)$$

where α is an adjustable parameter, $f(\lambda) = \lambda$ for terms in the wild type state, and $f(\lambda) =$ $1\text{-}\lambda$ for terms in the mutant state. With this equation, the van der Waals and electrostatic interaction terms are now adjusted from linear to the following [1,13],

$$\begin{aligned} U_{\mathrm{VDW}}(\lambda) = U_{\mathrm{VDW},0} &+ (1-\lambda)\left(\frac{A_{\mathrm{WT}}}{\left(r_{ij}^{\ 2}+\alpha_v\lambda\right)^6} - \frac{B_{\mathrm{WT}}}{\left(r_{ij}^{\ 2}+\alpha_v\lambda\right)^3}\right) \\ &+ \lambda\left(\frac{A_{\mathrm{Mut}}}{\left(r_{ij}^{\ 2}+\alpha_v(1-\lambda)\right)^6} - \frac{B_{\mathrm{Mut}}}{\left(r_{ij}^{\ 2}+\alpha_v(1-\lambda)\right)^3}\right) \end{aligned} \tag{2.18}$$

$$U_{\mathrm{Elec}}(\lambda) = U_{\mathrm{Elec},0}(\lambda) + (1-\lambda)\left(\frac{q_{i,\mathrm{WT}}q_{j,\mathrm{WT}}}{\sqrt{r_{ij}^{\ 2}+\alpha_e\lambda}}\right) + \lambda\left(\frac{q_{i,\mathrm{Mut}}q_{j,\mathrm{Mut}}}{\sqrt{r_{ij}^{\ 2}+\alpha_e(1-\lambda)}}\right) \tag{2.19}$$

While the shift parameters, α_v and α_e, can be altered to optimize computational speed and efficiency, the commonly accepted value is 5 Å^2 [1,13]. Figure 2.2 shows the smoothing effect that partial shift has on the van der Waals potential function for annihilation of an atom [1,13].

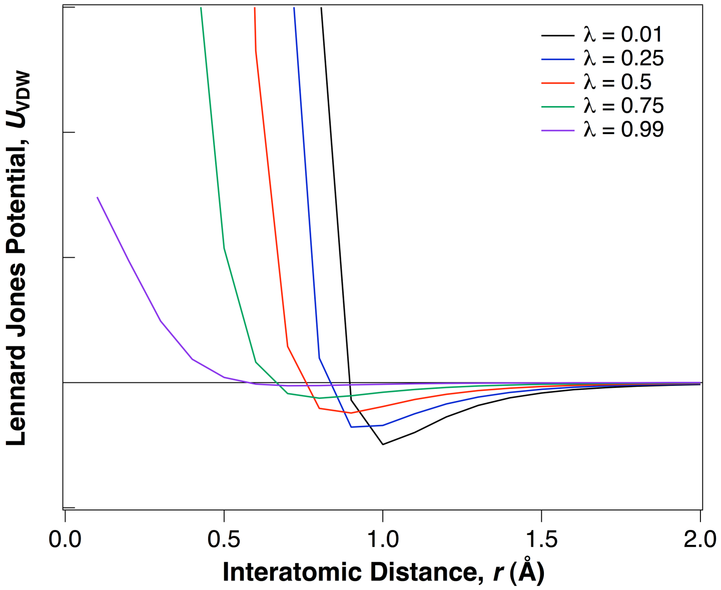


Figure 2.2: Example Lennard Jones (van der Waals) potential curves generated from equation 2.19 which incorporates the shift potential for annihilation of a wild type atom. Five values of λ were selected and U_0=0, A=1, B=2, and α=0.5.

Algorithms used to calculate $\partial U/\partial\lambda$ utilize a single-topology or a dual-topology hybrid potential energy function with a linear dependence on λ, as shown in equation 2.11. All of the studies were completed in NAMD [5] using a dual-topology function, and the non-bonded pair potentials are divided into three groups: (i) interactions within the unchanging environment, (ii) interactions with and within the wild type state, and (iii) interactions with and within the mutant state. The wild type and mutant states do not interact with each other. Since these simulations involve alchemical mutations, where the initial number of atoms does not equal the final number, a hybrid residue is defined to ensure that the number of degrees of freedom, and thus phase space of U_{Mut} and U_{WT}, remain the same throughout simulation. The hybrid is built based on the CHARMM topology data; electrostatic and atom type changes are included in the hybrid. For example, in a tyrosine-phenylalanine hybrid (Tyr-Phe, Figure 2.3 A), only the hydroxyl group changes to a methyl, but in a tyrosine-tryptophan hybrid (Tyr-Trp, Figure 2.3 B), the entire sidechain of the amino acid must be hybridized. During TI calculations, the interactions and charges associated with the wild type will be "switched off" while the interactions and charges associated with the mutation will be "switched on."

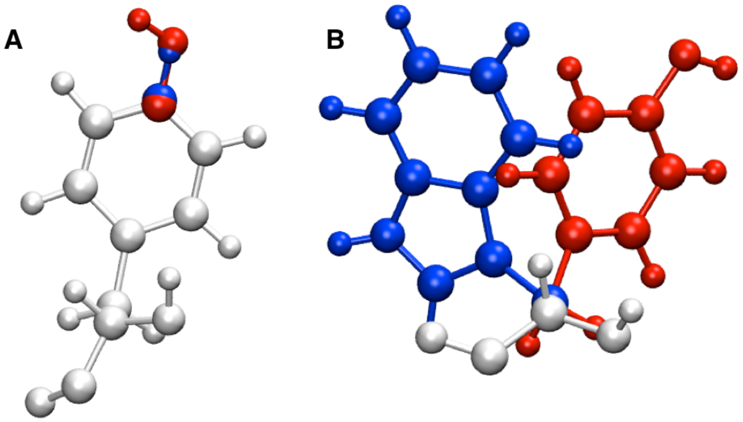


Figure 2.3: Hybrid Tyr-Phe (A) and Tyr-Trp (B) structures used in dual-topology TI calculations. The wild type structure, Tyr, is shown in red, and the mutations, Phe (A) or Trp (B) are shown in blue. Atoms that are identical for the wild type and mutant are shown in white. In TI calculations, the red and blue atoms do not "see" each other, and thus have no potential interaction.

Once the hybrid structures are constructed, TI simulations can be performed. The electrostatic and van der Waals simulations were decoupled to reduce computational effort and eliminate instabilities arising from large energy interactions [14]. Values of $\partial U/\partial\lambda$ are calculated at each timestep, and output to a data file for averaging. To ensure well-converged averages are obtained, the frequency histograms of the $\partial U/\partial\lambda$values at each window (Figure 2.4) were examined to ensure that there was sufficient overlap between the windows and that Gaussian distributions were achieved [14].

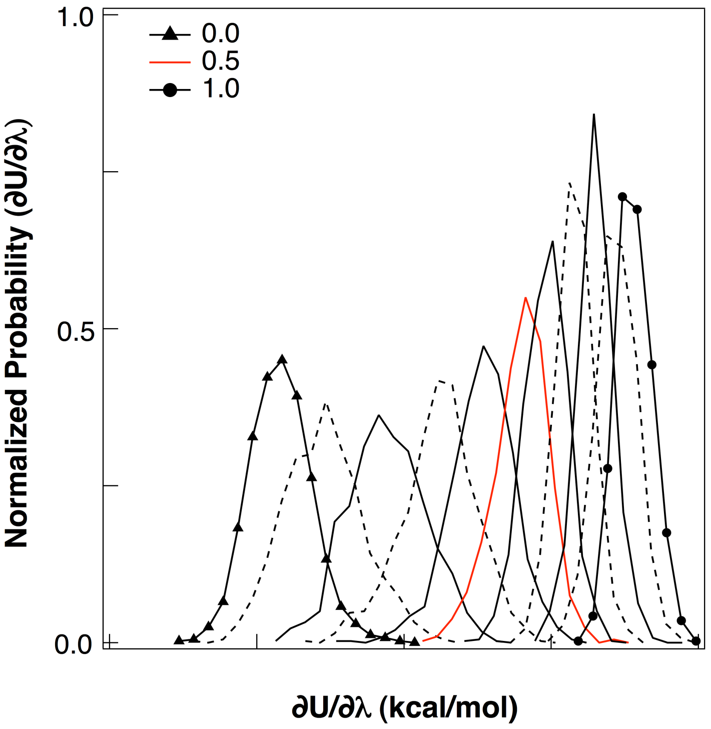


Figure 2.4: Characteristic acceptable histogram of $\partial U/\partial\lambda$ values generated in TI simulations with equidistant λ windows

For the biological systems of interest in this study, the electrostatic simulations, where atom charges are turned on or off, were separated into equidistant λ windows from 0 to 1. The van der Waals simulations, where atom interactions with the environment are turned on or off, required more windows, especially near the endpoints. A characteristic representation of a TI curve from these calculations is provided in Figure 2.5.

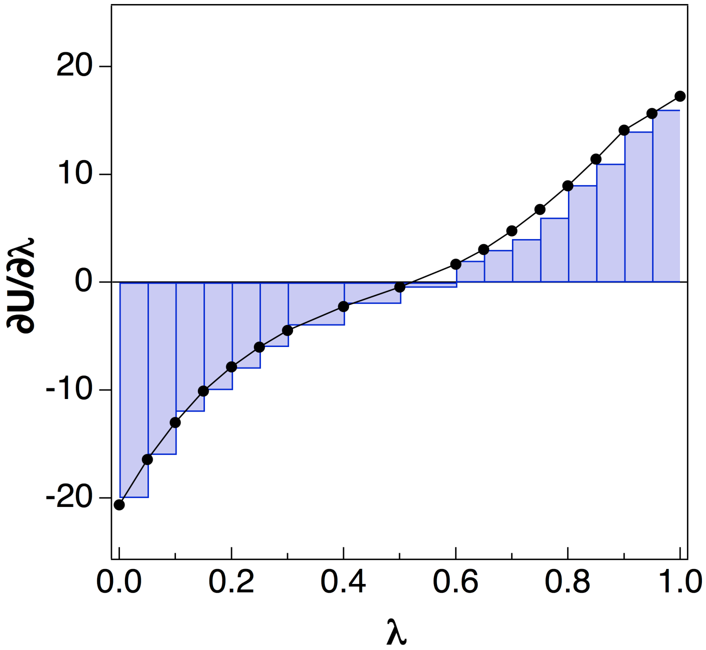


Figure 2.5: Typical $\partial U/\partial\lambda$ curve generated in TI calculations. The points represent the simulation output, and the bars represent the area under the curve to be calculated using numerical integration.

To solve for Δ*ΔG*, Simpson's rule was used. For equidistant nodes, Simpson's rule is written:

$$\int_a^b f(x)dx \approx \frac{\Delta x}{3}\left[f_1 + 2f_2 + 4f_3 + ... + 2f_{n-2} + 4f_{n-1} + f_n\right] \tag{2.20}$$

or, alternatively by implementing the trapezoid rule twice in:

$$\int_a^b f(x)dx \approx \frac{1}{3}\left[4T_n - T_{n/2+1}\right] \tag{2.21}$$

where T_n uses all nodes and $T_{n/2\ +1}$ uses every other node. For non-equidistant nodes, Simpson's rule is modified using the Brun generalization for arbitrary nodes [15], $x_i < x_{i+1} < x_{i+2}$:

$$\int_{x_i}^{x_{i+2}} f(x)dx \approx \frac{x_{i+2}-x_i}{6}[f_i + 4f_{i+1} + f_{i+2}] + \frac{x_i - 2x_{i+1} + x_{i+2}}{3}[f_{i+2} - f_i] \tag{2.22}$$

Equations 2.21 and 2.22 can eventually be combined to show the Simpson's rule with non-equidistant nodes as [15]:

$$\int_{x_i}^{x_{i+2}} f(x)dx \approx \frac{1}{3}[4T_3 - T_2] \tag{2.23}$$

where T_3 uses all nodes x_i, x_{i+1}, x_{i+2}, and T_2 uses nodes x_i and x_{i+2}. Equation 2.21 was implemented in this study where λ values were equally spaced, e.g., most of the electrostatic calculations, and Equation 2.23 was implemented where λ values were not all equal, e.g., all of the van der Waals calculations [15].

The error for each window, Δ_j, was calculated using a combination of the methods outlined by Steinbrecher *et al.* [16] and Paliwal and Shirts [17], though we note that several different methods for error analysis in TI have been reported in the literature [14,16-19]. Thus, in this work the standard deviation, σ_j, for each window j is related to the total window length, t_MD, and the autocorrelation time, τ, of $dU/d\lambda$ as [16],

$$\Delta_j(\partial U/\partial\lambda) = \sigma_j \Big/ \sqrt{t_\mathrm{MD}/2\tau} \tag{2.24}$$

The total error, Δ_tot, is then calculated by incorporating the squares of both the weight of each window and the individual window error to obtain equation 2.25 [17]:

$$\Delta_\mathrm{tot} = \sqrt{\sum_j \frac{1}{4}\left(\lambda_{j+1} - \lambda_{j-1}\right)^2 \Delta_j^2} \tag{2.25}$$

The autocorrelation of the potential, *c(t)*, used to calculate τ is given by [2,20]

$$c(t) = \frac{1}{c_0(N-t)} \sum_{j=1}^{N-t} \left(U_j - \overline{U}\right)\left(U_{j+t-1} - \overline{U}\right) \qquad (2.26)$$

where c_0 and $c(t)$ are the autocorrelation functions at time 0 and t, respectively, N represents the number of sample points, and U represents the potential energy, which is $\partial U/\partial \lambda$ in this case. Each window exhibited a single exponential decay and the window correlation time, *c(t)* was calculated by dividing the total window equal increments and averaging over the first 100-150 ps of each increment. The autocorrelation time, τ, used in equation 2.25 is calculated from the reciprocal of the curve power function of *c(t)*. The average autocorrelation times for the simulations in this study were similar, and an example data set is provided in Figure 2.6. The total error from λ = 0 to 1 was then calculated by weighting the error with the window span using Equation 2.25. The individual window error, Δ_j, was found to increase for larger mutations and was mitigated by increasing the number of windows to reduce the weight of each window.

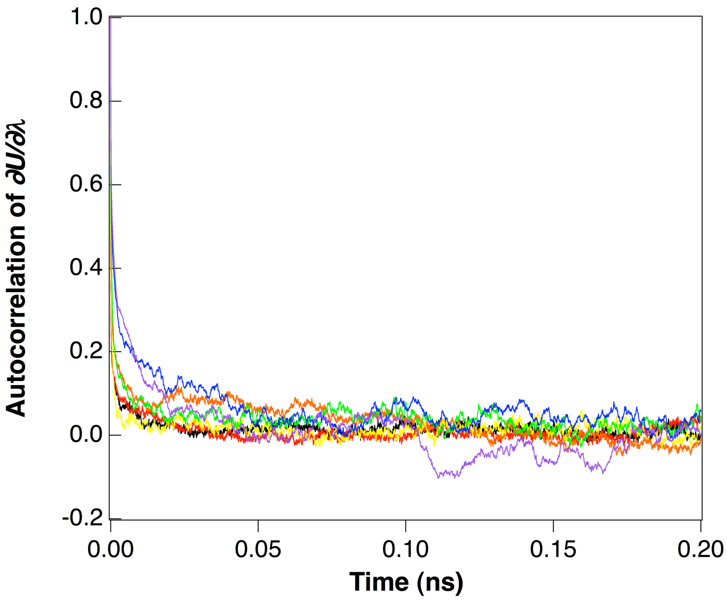


Figure 2.6: Autocorrelation functions for $\partial U/\partial\lambda$ for λ values ranging from 0 to 1. All of the values decay within 0.02-0.05 ns. τ is calculated from the reciprocal of the curve power function.

Finally, partition coefficients can be estimated using equation 1.1. The relative binding partition coefficient error, $\Delta_{\text{tot,}}$ was propagated to the relative binding affinity error, δ, using [21]:

$$\delta = \frac{K_{\text{Mut}}}{K_{\text{WT-NG}}}\frac{\Delta_{\text{tot}}}{RT} \tag{2.27}$$

The TI method just described was employed in the *T. reesei* Family 7 cellulase CBM and CD studies to calculate relative binding affinity changes with mutation. Other descriptions of calculated properties determined from longer MD production runs, such as root mean square deviation, protein-surface potential interaction energy, and hydrogen bonding potential are given in the subsequent chapters.

2.3 References

1. B.R. Brooks, C.L. Brooks, A.D. Mackerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, W. Yang, D.M. York and M. Karplus, "CHARMM: The Biomolecular Simulation Program", *J. Comput. Chem.* **2009**, 30 (10), 1545-1614.

2. A.R. Leach, *Molecular Modelling: Principles and Applications*. 2nd edition ed.; Pearson Education Limited: Harlow, England, **2001**.

3. A.D. MacKerell, N. Banavali and N. Foloppe, "Development and current status of the CHARMM force field for nucleic acids", *Biopolymers* **2000**, 56 (4), 257-265.

4. A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins", *J. Phys. Chem. B* **1998**, 102 (18), 3586-3616.

5. J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale and K. Schulten, "Scalable molecular dynamics with NAMD", *J. Comput. Chem.* **2005**, 26 (16), 1781-1802.

6. D.S. Frenkel, B. , *Understanding Molecular Simulation: From Algorithms to Applications*. 2nd ed ed.; Academic Press: San Diego, CA, **2002**.

7. M. Linder, G. Lindeberg, T. Reinikainen, T.T. Teeri and G. Pettersson, "The difference in affinity between 2 fungal cellulose-binding domains is dominated by a single amino-acid substitution", *FEBS Lett.* **1995**, 372 (1), 96-98.

8. M. Linder, M.L. Mattinen, M. Kontteli, G. Lindeberg, J. Stahlberg, T. Drakenberg, T. Reinikainen, G. Pettersson and A. Annila, "Identification of functionally important amino-acids in the cellulose-binding domain of *Trichoderma reesei* Cellobiohydrolase I", *Protein Sci.* **1995**, 4 (6), 1056-1064.

9. S. Takashima, M. Ohno, M. Hidaka, A. Nakamura and H. Masaki, "Correlation between cellulose binding and activity of cellulose-binding domain mutants of Humicola grisea cellobiohydrolase 1", *FEBS Lett.* **2007**, 581 (30), 5891-5896.

10. D.A. McQuarrie, *Statistical Mechanics*. University Science Books: Sausalito, CA, **2000**.

11. J.M. Prausnitz, Lichtenthaler, Rudiger N., andGomes de Avezedo, Edmundo, *Molecular Thermodynamics of Fluid-Phase Equilibria*. Prentice Hall: Upper Saddle River, NJ, **1999**.

12. T.C. Beutler, A.E. Mark, R.C. Vanschaik, P.R. Gerber and W.F. Vangunsteren, "AVOIDING SINGULARITIES AND NUMERICAL INSTABILITIES IN FREE-ENERGY CALCULATIONS BASED ON MOLECULAR SIMULATIONS", *Chemical Physics Letters* **1994**, 222 (6), 529-539.

13. M. Zacharias, T.P. Straatsma and J.A. McCammon, "SEPARATION-SHIFTED SCALING, A NEW SCALING METHOD FOR LENNARD-JONES INTERACTIONS IN THERMODYNAMIC INTEGRATION", *J. Chem. Phys.* **1994**, 100 (12), 9025-9031.

14. A. Pohorille, Jarzynski, C., & Chipot, C., "Good Practices in Free-Energy Calculations", *J. Phys. Chem. B* **2010**, (114), 10235-10253.

15. S. Bruckner and S. Boresch, "Efficiency of alchemical free energy simulations. II. Improvements for thermodynamic integration", *J. of Comp. Chem.* **2010**, 32 (7), 1320-1333.

16. T. Steinbrecher, D.L. Mobley and D.A. Case, "Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations", *J. Chem. Phys.* **2007**, 127 (21).

17. H. Paliwal and M.R. Shirts, "A Benchmark Test Set for Alchemical Free Energy Transformations and Its Use to Quantify Error in Common Free Energy Methods", *Journal of Chemical Theory and Computation* **2011**.

18. M.P.a.T. Allen, D.J., *Computer Simulation of Liquids*. Clarendon Press: Oxford, **1987**.

19. M.R. Shirts, J.W. Pitera, W.C. Swope and V.S. Pande, "Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins", *J. Chem. Phys.* **2003**, 119 (11), 5740-5761.

20. W. Smith, "Auto: CCP5 Summer School Autocorrelation Function," Daresbury Laboratory: Daresbury, Warrington, U.K., **1992**; pp.

21. "NIST/SEMATECH e-Handbook of Statistical Methods," NIST/SEMATECH: Gaithersburg, MD, **2010**; pp.

CHAPTER 3

COMPUTATIONAL INVESTIGATION OF AMINO ACID MUTATION AND GLYCOSYLATION EFFECTS ON THE CEL7A CBM

3.1 Introduction

Carbohydrate-binding modules (CBMs) represent a primary biological means for protein-carbohydrate recognition [1]. CBMs recognize and bind to a wide range of carbohydrates and are often components of multi-modular cellulase, hemicellulase, or chitinase enzymes, which are able to deconstruct plant, fungal, or algal cell wall carbohydrates [1-5]. It has been shown that CBMs serve two roles as components of multi-modular, carbohydrate-active enzymes: to target the specific structures of interest for catalytic action and to maintain proximity to a given carbohydrate surface [1,2]. A third hypothesized role is disruption of the carbohydrate crystal packing, making the chains easier to decrystallize for enzymatic attack [6-8], but the mechanism of this function remains unknown and evidence to support this role is limited [1].

Engineering enhanced cellulase enzymes is currently a topic of significant worldwide interest, a situation primarily driven by research to enable commercialization of renewable biofuels from lignocellulosic biomass [6-10]. Several routes exist for increasing cellulase performance, including rational and evolutionary methods to improve specific activity [8], screening for improved thermal tolerance [11,12], and the often overlooked strategy of increasing binding affinity to carbohydrates via CBM engineering [13-16]. In the lattermost case, several groups have shown that by increasing CBM binding affinity to cellulose via single point mutations or by swapping complete CBMs from other enzymes with higher binding affinity, cellulase enzyme activity can be improved [15,16]. This effect is likely due to higher enzyme concentrations on the cellulose surface, which potentially leads to a higher fraction of catalytically engaged enzymes. Higher binding affinity and thus higher activity, in turn, will lead to lower enzyme loadings and the eventual realization of more cost-effective biofuels processes [6].

To examine CBM binding at the molecular level, we investigate a novel strategy, namely the use of CBM glycosylation, to enhance the affinity of cellulases, which we predict will improve affinity more so than standard protein engineering strategies wherein amino acids are mutated to other residues. Specifically, using molecular simulation we quantify the impact of natural and engineered glycosylation on the binding affinity of the Family 1 CBM from the *T. reesei* Cel7A enzyme. A detailed visualization of the Cel7A enzyme taken from Figure 1.5 is shown in Figure 3.1 [17-19].

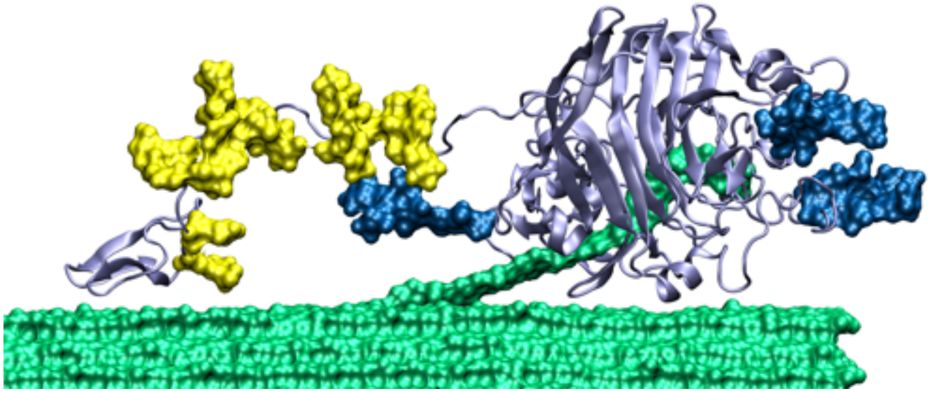


Figure 3.1: The catalytically-active complex of the *T. reesei* Family 7 cellobiohydrolase (light purple) on cellulose (green) [17-19]. The native *O*-glycosylation is shown in yellow and the native *N*-glycans are shown in dark blue. The CBM is the small protein domain on the left containing 2 native *O*-glycans, and the catalytic domain is the large protein domain on the right with a cellulose chain threaded into the tunnel.

As discussed in Chapter I, Family 1 CBMs are ~36 residue proteins that exhibit high sequence homology and are predominantly produced by fungi [20]. The Cel7A CBM exhibits a planar face with three tyrosine (Tyr) residues that are hypothesized to form the binding face to cellulose (Figure 3.2) [21,22]. The Cel7A CBM, linker, and catalytic domain (CD) have all been shown to be glycosylated when expressed in the native host [23]. Two residues on the CBM, threonine (Thr) 1 and serine (Ser) 3 (Figure 2.3), are natively glycosylated with at least one mannose residue and potentially up to three mannose residues each, but as Harrison *et al.* discuss, the *O*-glycosylation assignments on the Cel7A CBM and linker are at best an average extent of glycosylation [23]. An engineered (non-native) glycan was also chosen for our study and is shown on Ser-14 in Figure 3.2.

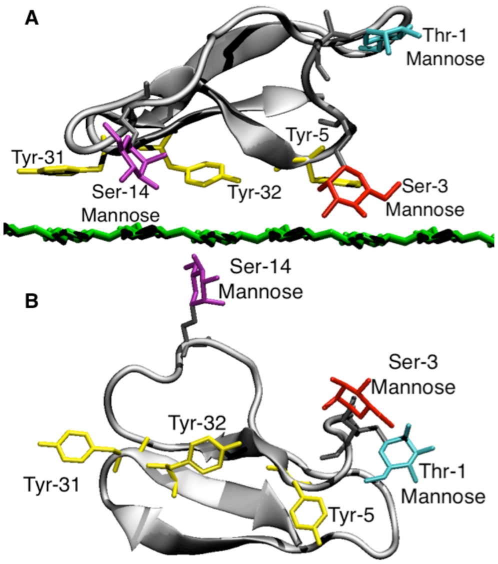


Figure 3.2: Side view (A) and top view (B) of the Cel7A CBM and the top layer of the cellulose slab. The tyrosine residues are shown in yellow. The native *O*-glycans considered here are shown on Thr-1 and Ser-3. The artificial glycan examined here is shown on Ser-14. The N-terminus and Thr-1 and Ser-3 glycans are located in the posterior region, and the Ser-14 glycan is located in the anterior region.

Experimental studies to date have examined the role of aromatic and polar residues on the Cel7A CBM binding affinity [13-15,24], but no work to our knowledge has considered the role of natural, recombinant, or engineered *O*-glycosylation on the Family 1 CBM binding affinity. Boraston *et al.* demonstrated that the addition of high mannan *N*-glycans ($GlcNAc_2Man_8$ and higher) to a Family 2 CBM (CBM2a) from *Cellulomonas fimi*, expressed by the yeast *Pichia pastoris*, was detrimental to binding affinity. This result was attributed to the size of the *N*-glycan inhibiting the ability of the CBM to interact with the carbohydrate substrate [25]. In a later study, Boraston *et al.* expressed a modified CBM2a with the *N*-glycosylation sites removed via mutations of asparagine to alanine in *P. pastoris*. The authors demonstrated that this CBM was *O*-glycosylated with one to four mannose residues on each of the glycosylated serine and threonine residues [26]. They concluded that the *O*-glycosylation did not impact binding affinity in this case, because the serine and threonines are located above the cellulose surface where the glycans could neither interact with cellulose nor inhibit the CBM-cellulose interaction as the *N*-glycan did previously [25,26].

In many filamentous fungi and yeasts, *O*-glycans often contain fewer carbohydrate moieties than *N*-glycans [26,27]. In the current work we examine natural or small artificial *O*-glycans, neither of which are expected to restrict access of the CBM to cellulose. Before studying the impact of *O*-glycans on binding affinity, to validate the simulation approach, we first study a number of amino acid mutations for which we can compare with the experimental results of Linder *et al.*, who showed that the non-glycosylated Cel7A CBM binding affinity is improved by the mutation of residue 5 from tyrosine to tryptophan (denoted Y5W) [13]. Experimentally bound and unbound concentrations of the non-glycosylated CBM variants were used to generate adsorption isotherms [13,14] from which partition coefficients can be determined and the relative binding free energy ($\Delta\Delta G$) calculated as:

$$\Delta\Delta G = - RT\ln(K_{\mathrm{Mut}}/K_{\mathrm{WT\text{-}NG}}) \tag{3.1}$$

where K_{Mut} and $K_{\mathrm{WT\text{-}NG}}$ are the partition coefficients for the mutant CBM of interest and the wild type CBM *without glycosylation* (NG = no glycosylation), respectively, T is temperature, and R is the gas constant. We note that the partition coefficients are defined between the free CBM in solution and bound to the hydrophobic face of cellulose [3]. All of the variants studied by Linder *et al.*, except the Y5W mutation, were found to exhibit lower binding affinity to cellulose than the wild type, non-glycosylated Cel7A CBM with $\Delta\Delta G$ values of +0.2 to +1.7 kcal/mol. The Y5W mutation exhibited an increase in K_{Mut} over $K_{\mathrm{WT\text{-}NG}}$ by a factor of 1.6 (-0.26 kcal/mol) [13,14]. In similar work, Takashima *et al.* examined the enzymatic activity of various *Humicola grisea* Cel7A enzymes with mutant CBMs [15]. The mutation of Y5W produced a favorable 33% increase in binding and a 12% increase in overall enzymatic activity, similar to the results reported by Linder *et al.* [13,15].

Although the effects on binding of *O*-glycosylation on the Cel7A CBM have not been studied, both *N*- and *O*-glycosylation in cellulases are known to vary with expression host and growth conditions and have the potential ability to affect the activity and stability of cellulases [25,28-31]. To understand the impact of glycosylation on CBM function and to potentially engineer enhanced protein-carbohydrate binding affinity, we study the impact of *O*-glycosylation on the Cel7A CBM binding affinity to cellulose using molecular dynamics (MD) free energy methods. When referring to “wild type” throughout the chapter, we are referring to the wild type protein sequence only, and we discuss the sequence with or without glycosylation depending on the simulation of interest. We refer to the “native” glycosylation as that defined by Harrison *et al.* [23], i.e., with one *O*-mannose residue on Thr-1 and one *O*-mannose residue on Ser-3. To validate the computational approach, the free energy for mutating a characteristic aromatic residue (Tyr-5) is calculated, and agreement with experimental results for non-glycosylated CBMs is achieved. We then predict that in the case of a single native *O*-glycan at Thr-1 (which does not interact directly with cellulose, and thus we assume will not significantly affect the CBM binding affinity), the addition of a single native *O*-glycan at Ser-3 on the Cel7A CBM can increase the binding affinity by over 3-fold. Furthermore, the addition of a glycan disaccharide at Ser-3, instead of the monomer, increases the binding affinity by 6-fold relative to the non-glycosylated wild type CBM. We also show that an engineered glycan at Ser-14, situated at the anterior of the CBM, has a more pronounced effect, increasing the binding affinity by 20-fold compared to the non-glycosylated wild type. When we combine the addition of the engineered Ser-14 *O*-mannose residue with the native glycosylation, the binding affinity increases by 140-fold relative to the non-glycosylated wild type. This work suggests a general strategy for engineering enhanced cellulases by increasing CBM binding affinity through introduction of artificial glycosylation sites through amino acid mutation or altering glycosylation via heterologous expression or manipulation of growth conditions.

3.2 Computational Methods

The systems studied by thermodynamic integration (TI) and long MD simulations are summarized with their respective results in Tables 1 to 3, along with the nomenclature that will be used throughout the chapter. All binding affinity comparisons, $K_{\mathrm{Mut}}/K_{\mathrm{WT\text{-}NG}}$, are performed as in Equation 3.1 by comparing to the wild type non-glycosylated CBM. The TI simulations [32,33] are designed to measure the relative binding free energy between the CBM in solution and the CBM on the hydrophobic face of cellulose as a function of amino acid mutation or addition of glycosylation. The thermodynamic cycle thus consists of TI calculations of the CBM in solution (without cellulose) and the CBM on the hydrophobic face of cellulose. The simulations performed are summarized in Figure 3.3. An *O*-mannose residue is present at Thr-1 for all glycosylated simulations and S3M1, but this mannose does not interact with the surface and thus was not included in a separate TI simulation. In Figure 3.3, for simulations where S3M1 is an intermediate step between the non-glycosylated wild type and the mutation, the total impact of the mutation is calculated by adding $\Delta\Delta G$(S3M1) and the $\Delta\Delta G$ of the current run, resulting in S3M1 + Y5W-G, etc., to the right of the equality.

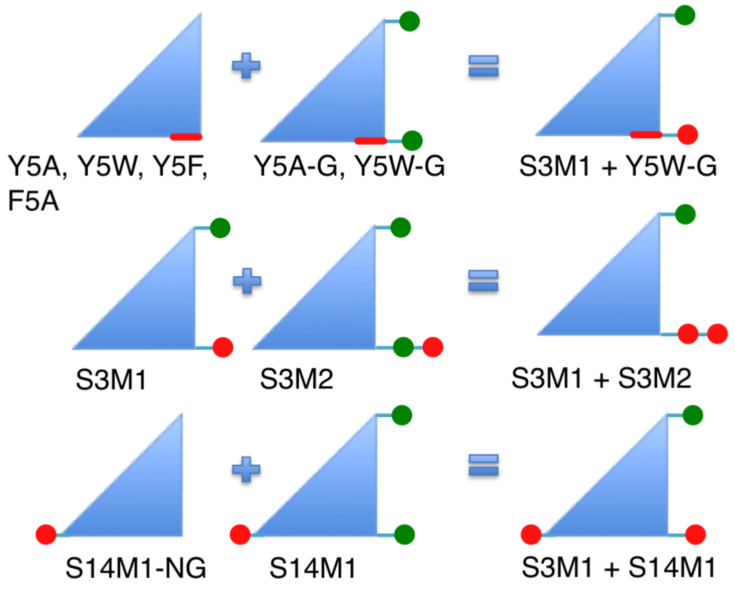


Figure 3.3: Illustration representing the Cel7A CBM TI calculations. Items in red are the mutations added during TI. The green circles represent mannose residues.

TI Simulations of Amino Acid Mutations

Because the Tyr-5 residue has been well characterized for the Cel7A CBM [13,14,21], we use this mutation as a validation of our computational approach. These simulations were run with no glycosylation at Thr-1 and Ser-3 for direct comparison to binding experiments where the CBMs were produced via solid-state peptide synthesis, and thus the CBMs have no glycosylation [13,14]. For Tyr-5, four TI calculations were performed: Y5A, Y5W, Y5F, and F5A (Figure 3.3). Y5A was chosen to modify the polarity of the Tyr-5 residue by mutation to Ala-5, which is one of the simplest non-polar amino acids, while Y5W replaces Tyr-5 with Trp-5, a larger aromatic residue with known higher binding affinity [13]. Y5F and F5A (Tyr-5 to Phe-5 and Phe-5 to Ala-5, respectively) were selected as a control to ensure internal consistency, such that:

$$\Delta\Delta G(\text{Y5A}) - \Delta\Delta G(\text{F5A}) \approx \Delta\Delta G(\text{Y5F}) \tag{3.2}$$

TI Simulations of Glycosylated CBMs

Previous work by Harrison *et al.* indicated the presence of at least one *O*-mannose residue at both threonine 1 (Thr-1) and serine 3 (Ser-3) on the Family 1 CBM expressed in a particular *T. reesei* strain [23]. Based on the NMR structure of the CBM [22], which suggests that the sugar on Thr-1 may be too far above the cellulose surface to interact with the cellulose directly, we have not examined mutations of the native *O*-glycosylation on Thr-1. An potential glycosylation site was also studied at Ser-14, which is positioned near the cellulose surface at the anterior of the CBM. The Cel7A CBM contains eight Thr and Ser residues, but only two, Ser-3 and Ser-14, are located near the surface. As previously discussed, Boraston *et al.* showed that *O*-glycans far above the surface do not impact binding affinity [26]; anticipating that glycosylation in Cel7A far above the surface would have a similar result, we thus focused our study on Ser-14. This TI simulation was conducted both with and without the native glycans at Thr-1 and Ser-3 (Figure 3.3). The four glycan mutations studied were used to test the impact of natural (S3M1 and S3M2) and potential (S14M1-NG, S14M1) glycosylation on binding affinity. Additionally, to test the impact of glycosylation on amino acid mutations, we repeated the Y5A and Y5W simulations with the native Thr-1 and Ser-3 mannan pattern, denoted Y5A-G and Y5W-G. This also provides a second check for internal consistency in that:

$$\Delta\Delta G(\text{Y5W}) + \Delta\Delta G(\text{S3M1}) \approx \Delta\Delta G(\text{S3M1+Y5W-G}) \qquad (3.3)$$

MD Simulations to Examine the Impact of Glycosylation on CBM Stability

In addition to the TI calculations, 100 ns MD simulations of the bound wild type CBM, the bound Y5A mutant CBM, and the bound Y5W mutant CBM were all performed without glycosylation. MD runs of bound CBMs using the glycan patterns found experimentally [23], were also conducted. Finally, simulations of the bound, native *O*-mannose residue at Thr-1 and disaccharide mannan at Ser-3 and the bound, native *O*-mannose residues at Thr-1 and Ser-3 with the engineered *O*-mannose residue at Ser-14 were also performed.

Simulation Details

CHARMM [34] was used to build the hybrid protein structures and initial coordinate files from the original wild type structure. NAMD [35] was used for all equilibrations and thermodynamic integration (TI) calculations and VMD [36] was used for visualization. The CHARMM27 force field with the CMAP correction [34,37,38] was used to describe the protein, while cellulose and the *O*-glycosylation were modeled using the CHARMM35 carbohydrate force field [39,40]. Water was modeled using the modified TIP3P force field [41,42].

The cellulose Iβ crystal structure was used to generate the cellulose slab [43]. The cellulose slab thickness, the CBM positioning above the surface, and overall dimensions, were taken from Beckham *et al.* [21]. The mannan disaccharide is linked α-1,2 and all *O* links to Ser and Thr are in the α-configuration [27,44]. The solvated, bound system contained approximately 18,100 atoms. Particle mesh Ewald [45] was used to describe the long-range electrostatic interactions with a sixth order b-spline interpolation, a Gaussian distribution with a width of 0.312 Å, and a mesh size of 60 x 60 x 45. A non-bonded interaction cutoff of 10 Å was used. The SHAKE algorithm [46] was employed to fix covalent bonds to hydrogen atoms. The CBM-cellulose system (referred to as the bound system), was minimized for 2,000 steps, and then equilibrated in the NVT ensemble at 300 K using a 2 fs timestep for 2 ns, at which time the RMSD of the protein backbone had stabilized. To calculate relative binding free energy, systems without cellulose (referred to as the free system), were also prepared. The wild type CBM and mutated CBMs structures were solvated in CHARMM with approximately 4,000 water molecules and simulations performed under the same conditions as those with the cellulose surface. The equilibrated, final coordinates of each system, bound and free, were used as the starting coordinates for the TI simulations.

In NAMD, TI was performed using the dual-topology method [33], implemented by equilibrating a single structure with a hybrid residue containing both the wild type and mutated atoms. The electrostatic and van der Waals calculations were decoupled, reducing computational effort and eliminating instabilities arising from large energy interactions [33]. The electrostatic calculations comprised 11 equidistant λ windows from 0 to 1, each equilibrated for 0.5 ns before 10 ns TI NVT runs. Two additional windows, 0.05 and 0.95 were added to the glycan electrostatic cases to improve the precision of the results. Van der Waals calculations required more windows and longer equilibrations, especially near the endpoints of $\lambda = 0$ and 1, as well as longer overall run lengths to reduce statistical error. For all systems studied, 11 to 18 windows with equilibration periods of 2 to 4 ns and window run lengths of 20 ns each were found to be sufficient to achieve the desired overlap and Gaussian distribution goals. For the simulations with the largest structural changes, 30 to 40 steps of minimization were added at the beginning of the $\lambda = 0$ to 0.5 van der Waals TI windows to avoid potential steric clashes of "appearing" and "disappearing" structures near $\lambda = 0$. The soft-core pair potential [33,47] was incorporated with a shift parameter, α, equal to 5 Å^2 for all simulations except the glycan mutations, where an α equal to 8.5 Å^2 exhibited increased simulation stability at λ = 0. A 2 fs timestep was used in all simulations except Y5W and the van der Waal portions of the glycan mutations, where a 1 fs timestep was used for stability. An example data set of bound and free $dU/d\lambda$ results from the Y5F simulation are shown in Figure 3.4. Discarding the minimization and pre-equilibration data, $dU/d\lambda$ was integrated over all λ values to calculate the final ΔG for bound and free systems (Equation 2.14) using Simpson's rule with Brun's extension for non-equidistant data [48].

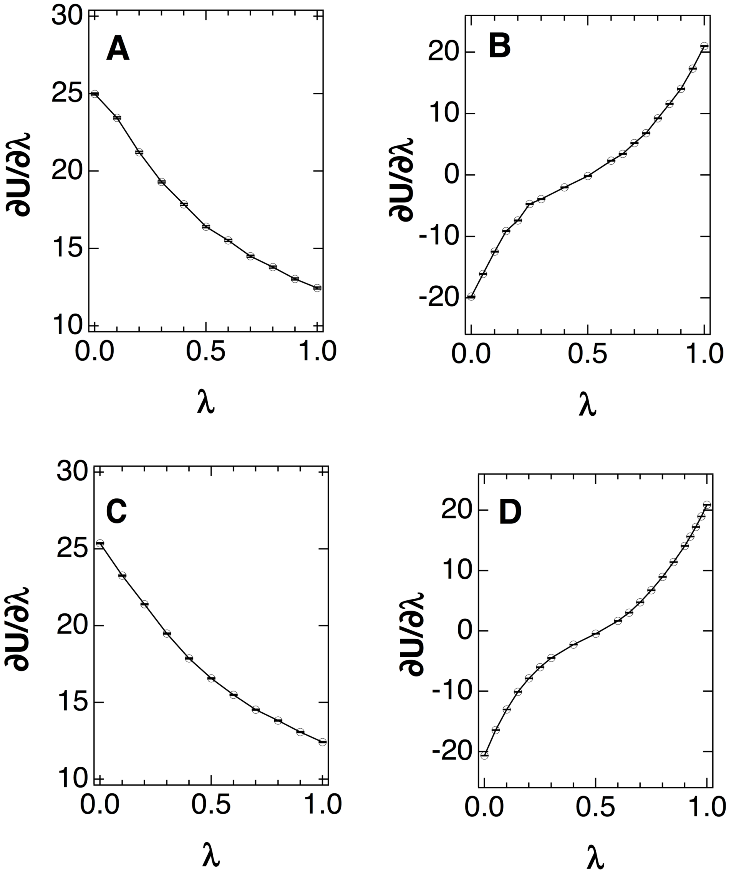


Figure 3.4: Y5F energy curves for (A) bound electrostatic, (B) bound van der Waals, (C) free electrostatic, and (D) free van der Waals systems. Associated errors are (A) 0.07 kcal/mol, (B) 0.06 kcal/mol, (C) 0.03 kcal/mol, and (D) 0.02 kcal/mol.

Additional windows were selected by examining the probability histograms of *dU/dλ* at each *λ* value [33,47]. Frequency histograms of the *dU/dλ* values at each window were generated to ensure that there was sufficient overlap between the windows and that Gaussian distributions were achieved [33]. The histograms from the Y5F simulation are provided in Figures 3.5 and 3.6 as an example. For the electrostatic interactions data were collected every 0.1 ns for a total of 9.5 ns of TI data/window, while for the van der Waals interactions data were collected over a total of 19.5 ns.

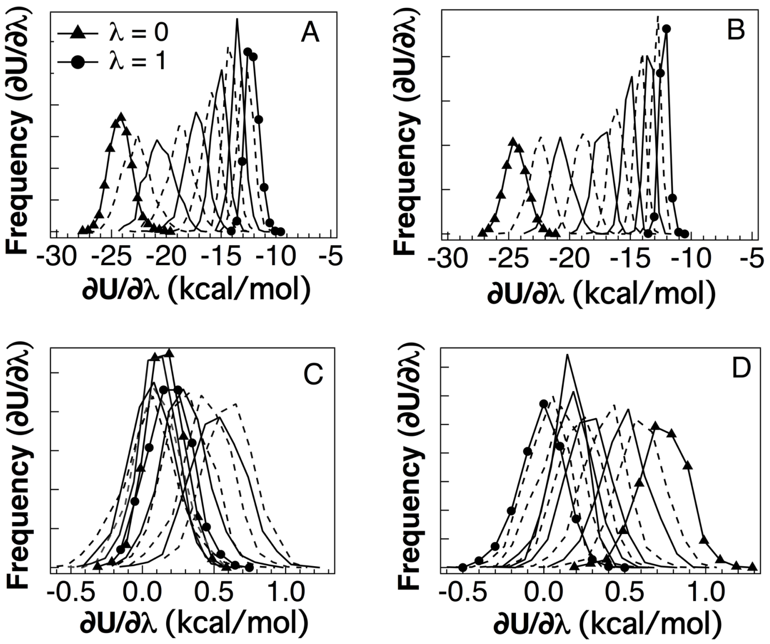


Figure 3.5:Y5F frequency histograms for Tyr-5 bound (A) and Tyr-5 free (B), Phe-5 bound (C), and Phe-5 free (D), electrostatic simulations.

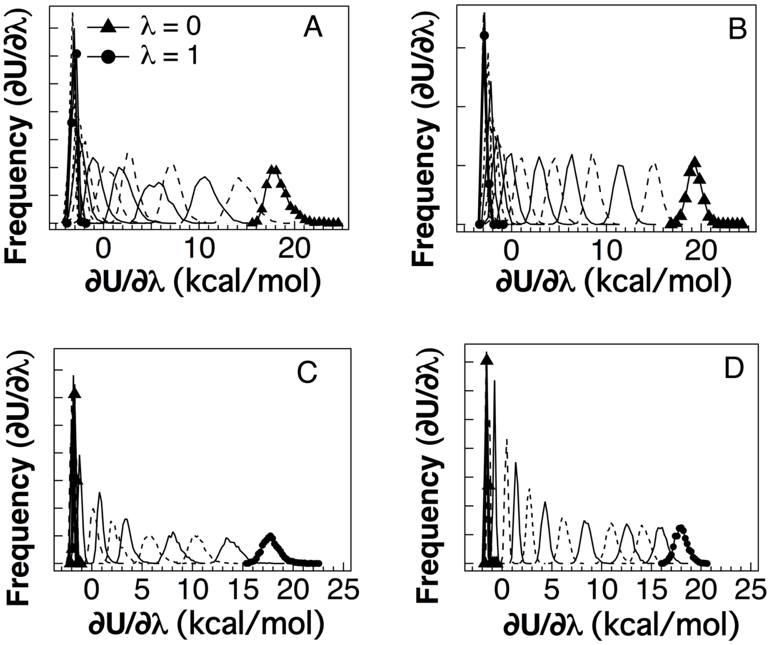


Figure 3.6: Y5F frequency histograms for Tyr-5 bound (A) and Tyr-5 free (B), Phe-5 bound (C), and Phe-5 free (D), van der Waals simulations.

The error for each window, Δ_j, was calculated using a combination of the methods outlined by Steinbrecher *et al.* [47] and Paliwal and Shirts [49], as described in Chapter II (Equations 2.24-2.27). Each window exhibited a single exponential decay and the window correlation time was calculated by dividing the total window into 1 ns increments and averaging τ over the first 100-150 ps of each increment. The total error from $\lambda = 0$ to 1 was then calculated by weighting the error with the window span using Equation 3.5. Individual window error, Δ_j, was found to increase for larger mutations, such as Y5W and the glycan mutations, and was mitigated by increasing the number of windows to reduce the weight of each window. The average autocorrelation time for all simulations is 0.025 ± 0.015 ns, keeping in mind that certain rare fluctuations could be missed due to the noisiness of the simulations, leading to slight underestimation of the error. However, with larger errors as we see here, the difference should be minor overall. The autocorrelation function for Y5F, bound, $Y_{ELEC,\lambda=0}$ (0.02 ns) and $Y_{ELEC,\lambda=1}$ (0.011 ns) is shown in Figure 3.7.

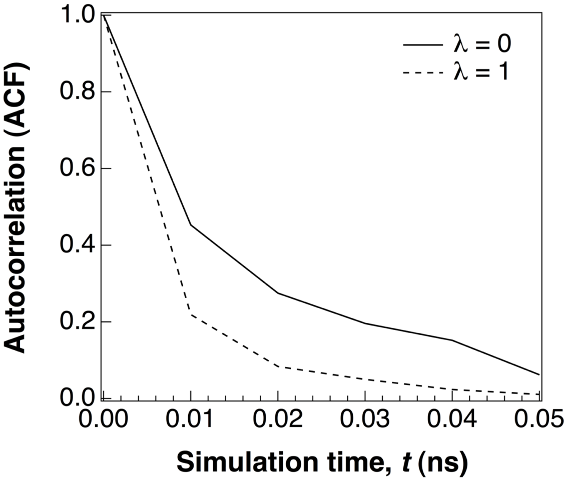


Figure 3.7: Average autocorrelation function of Y5F, bound electrostatic case at $\lambda = 0$ and $\lambda = 1$. Autocorrelation time, τ, used in equation S5 is calculated from the reciprocal of the curve power function.

As discussed above, long MD runs were also performed for selected systems. In these simulations, the system size and simulation parameters were identical to those used in the TI studies. Each system was run for 100 ns in the NVT ensemble at 300 K with a 2 fs timestep. The non-bonded interaction energy between the residues of interest and the surface was calculated in NAMD and the error for these values over the total run was determined using block averaging [50]. Hydrogen bonding potentials between residues of interest and the surface were calculated in VMD for comparison to the interaction energy values. RMSF and RMSD of the backbone were also calculated and the typical movement and rotation of the glycan sugars during the simulation determined.

3.3 Results

Relative Binding Free Energy from TI

The $\Delta\Delta G$ results from the TI calculations are provided in Tables 1 to 3. We constructed Langmuir isotherms for key mutations, as is typical with binding affinity measurements. The experimental and predicted changes in the mutant and wild type partition coefficients, K_{Mut} and $K_{\mathrm{WT\text{-}NG}}$ respectively, are shown in Figure 3.8. Since these simulations were performed with a single CBM in infinite dilution, or explicit water with periodic boundary conditions, we only calculate the initial slope of the Langmuir isotherm. For the simulations where S3M1 is the intermediate step between the non-glycosylated wild type and the mutation, the total impact of the mutation is calculated by adding $\Delta\Delta G$(S3M1) and the $\Delta\Delta G$ of the current run, resulting in S3M1+Y5A-G, S3M1+Y5W-G, S3M1+S3M2, and S3M1+S14M1 (Table 3.1 to 3.5). For simplicity in figures and text, the "end-state" name, e.g. Y5A-G, will be used rather than the cumulative names shown in Table 3.1 to 3.5.

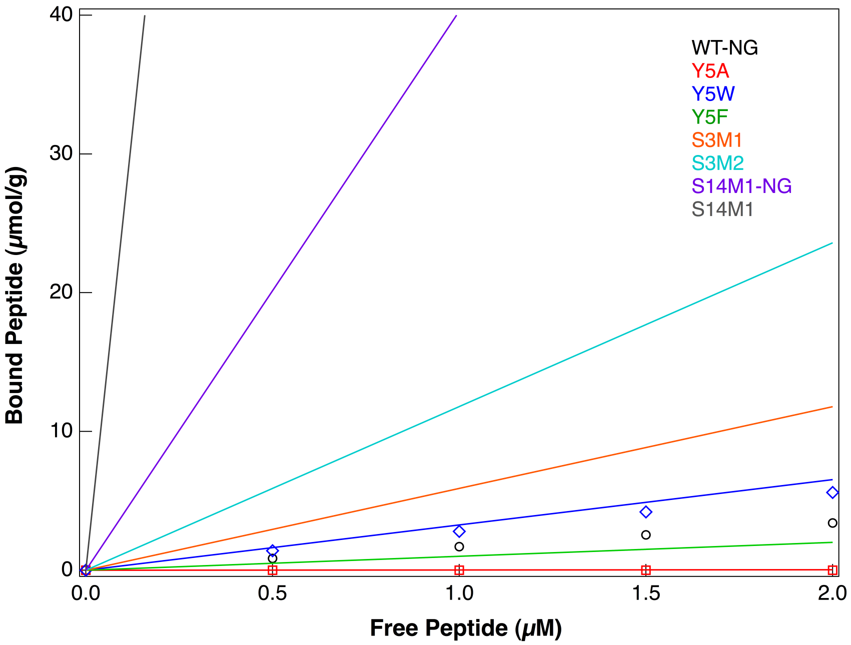


Figure 3.8: Langmuir isotherms of bound CBM as a function of free CBM as measured by Linder *et al.* (symbols) [13,14] and predicted by the free energy simulations conducted in this work (lines). Our results indicate improvement in binding affinity over that of just amino acid mutation with the incorporation of a single or disaccharide *O*-mannan at S3 (3 to 6-fold improvement in binding affinity) and a synthetic glycan site at Ser-14 with and without the native glycans (20 to 140-fold improvement in binding affinity, respectively).

The $\Delta\Delta G$ results for the amino acid TI calculations are shown in Table 3.1. We find agreement between the experimental data [13-15] and our computational predictions for the two large mutations: Y5W and Y5A, which provides direct quantitative evidence for the viability of our computational approach. Congruent with the results from Linder *et al.* [13,14] for the non-glycosylated systems, mutation of Tyr to Ala was found here to decrease binding affinity, whereas mutation to Trp increases binding affinity. The nearly 2-fold improvement in Y5W $K_{\mathrm{Mut}}/K_{\mathrm{WT\text{-}NG}}$ is also the same order of magnitude as the 1.3-fold improvement measured by Takashima *et al.* for a similar Family 1 CBM [15].

Table 3.1: Relative binding free energy ($\Delta\Delta G$, kcal/mol) from amino acid TI calculations, including the two containing native glycosylation (Y5A-G and Y5W-G), and associated change in partition coefficient ($K_{\mathrm{Mut}}/K_{\mathrm{WT\text{-}NG}}$). The partition coefficient change was not calculated for the intermediate steps, Y5A-G and Y5W-G. The S3M1+Y5A-G and S3M1+Y5W-G entries are the sum of the Y5A-G and Y5W-G and S3M1 (see Table 3.4) entries, respectively.

Mutation	$\Delta\Delta G$ (kcal/mol)	$K_{\mathrm{Mut}}/K_{\mathrm{WT\text{-}NG}}$
Y5A	2.6 ± 0.14	0.01 ± 0.00
Y5F	0.32 ± 0.04	0.59 ± 0.04
Y5W	-0.40 ± 0.14	1.9 ± 0.45
Y5A-G	2.26 ± 0.04	--
Y5W-G	-0.92 ± 0.06	--
S3M1 + Y5A-G	1.5 ± 0.05	0.08 ± 0.01
S3M1 + Y5W-G	-1.7 ± 0.07	16 ± 1.8

As discussed earlier, the F5A mutation was studied to validate the simulation approach, and Equation 3.2 holds nearly within error for the Y5A, F5A, and Y5F mutations. Since these calculations are computationally expensive, and this was a proof of concept simulation, the number of windows was limited to a spacing of 0.1 with only two additional windows at $\lambda = 0.05$ and $\lambda = 0.95$ for the van der Waals TI calculations. Removing the data from the extra windows for Y5F and Y5A and re-solving Equation 2.15, we can then compare estimated $\Delta\Delta G$(Y5F) from Equation 3.2 with actual $\Delta\Delta G$(Y5F) from Equation 2.14. The results, shown in Table 3.2, confirm that Equation 3.2 holds nearly within error, suggesting that our method is compatible with experiment.

Table 3.2: Relative binding free energy ($\Delta\Delta G$ (kcal/mol)) of Y5A, F5A, Y5F calculated (Equation 3.2) and Y5F actual from 11 electrostatic windows and 13 van der Waals windows.

Mutation	$\Delta\Delta G$ (kcal/mol)
Y5A	2.5 ± 0.17
F5A	0.84 ± 0.14
Y5F, Eqn 3.2	1.7 ± 0.31
Y5F actual	1.2 ± 0.05

Loss of the large binding surface area and a hydrogen bond in the Y5A case is clearly unfavorable for binding and is confirmed by the long MD simulations described later in this discussion, where a decrease in overall interaction energy with the surface is observed for the Y5A mutant compared to the wild type CBM or the Y5W mutant. The relative binding free energy for the Y5F mutant is between that of the Y5A and Y5W mutants, which is expected, given that loss of the hydroxyl group on Tyr removes a hydrogen bonding site but the retention of the aromatic ring maintains the shape of the planar face of the CBM. For the Y5W mutant, the loss of the Tyr hydroxyl is offset by an increase in the surface area of the side chain, which is corroborated by the long MD simulations in that the local Trp-5-surface non-bonded interaction is primarily mediated by van der Waals interactions.

We found no differences in the Y5A and Y5A-G mutant simulations; mutation to Ala-5 is very detrimental to binding with or without the native glycosylation pattern from Harrison *et al.* [23]. However, we find that the binding affinity improvement for the Y5W mutant increases 16-fold over the non-glycosylated wild type CBM with native glycosylation (S3M1+Y5W-G) and also confirmed that Equation 3.3 holds near within error for Y5W, S3M1, and Y5W-G. In accordance with equation 3.2, the results of the

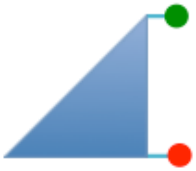
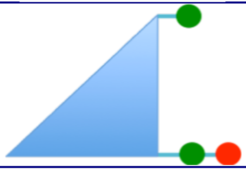
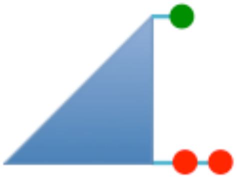
Y5W mutation and the glycosylation mutation should result in equation 3.3, as shown in Table 3.3.

Table 3.3: Relative binding free energy ($\Delta\Delta G$ (kcal/mol)) of Y5W, S3M1, Y5W-G calculated (Equation 3.3) and Y5W-G actual

Mutation	$\Delta\Delta G$ (kcal/mol)
Y5W	-0.40 ± 0.14
S3M1	-0.75 ± 0.03
Y5W-G, Eqn 3.3	-1.2 ± 0.14
Y5W-G + S3M1 actual	-1.7 ± 0.07

The glycosylation TI simulations were designed to understand the changes in the CBM binding affinity upon (i) the addition of a single, native O-mannose residue at Ser-3 (S3M1), (ii) the addition of a mannan disaccharide at Ser-3 (S3M1+S3M2), and (iii) the addition of a non-native *O*-mannose residue at Ser-14, with the lattermost examined both with and without the native glycosylation at Thr-1 and Ser-3 (S14M1 and S14M1-NG, respectively). The Thr-1 *O*-mannose residue does not impact binding affinity directly, so it was not considered part of the thermodynamic cycle, but it is present in all glycosylated simulations. For all the glycosylation simulations, we find that CBM binding from solution to cellulose improves with the addition of *O*-mannose residue, increasing both the potential for hydrogen bonding and the surface area for interaction. The $\Delta\Delta G$ results for the S3M1 and S3M2 TI simulations are given in Table 3.4. The partition coefficient change was not calculated for the intermediate step, S3M2. In the schematics, the green circles represent wild type *O*-mannose residues present in the reactant and product state, and the red circles represent the mannose residue(s) added to Ser-3 as the product of the thermodynamic cycle for the S3M1 and S3M2 cases. Errors were calculated as described in the simulation details section. The S3M1 glycan TI simulation demonstrates that the addition of the single native *O*-mannose residue at Ser-3 located near the posterior of the CBM improves the binding affinity by 3-fold. The addition of a second *O*-mannose residue at Ser-3 (S3M2), from a second TI calculation, improves the binding affinity by 2-fold increase relative to the S3M1 case, resulting in a total 6-fold improvement over the non-glycosylated wild type CBM in binding affinity for the addition of the disaccharide (S3M1+S3M2).

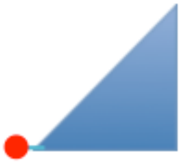
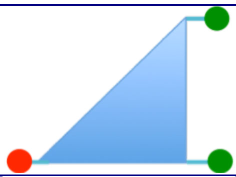
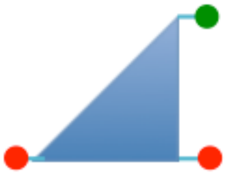
Table 3.4: Relative binding free energy ($\Delta\Delta G$, kcal/mol) from the native glycosylation TI calculations, for a single *O*-mannose residue at Ser-3 (S3M1) and a disaccharide mannan at Ser-3 (S3M2), and associated change in partition coefficient ($K_{\mathrm{Mut}}/K_{\mathrm{WT\text{-}NG}}$). The S3M1+S3M2 entry is the sum of the previous two entries.

Mutation		**$\Delta\Delta G$ (kcal/mol)**	**$K_{\mathrm{Mut}}/K_{\mathrm{WT\text{-}NG}}$**
S3M1		-0.75 ± 0.03	3.5 ± 0.17
S3M2		-0.42 ± 0.03	--
S3M1 + S3M2		-1.2 ± 0.04	6.9 ± 0.49

The TI results for the addition of the artificial glycan site on Ser-14, located towards the anterior of the CBM, both with and without native glycosylation, are shown in Table 3.5. Again, the partition coefficient was not calculated for the intermediate step, S14M1, and errors were calculated as described in the computational methods section.

The addition of the single glycan without the native glycans in the posterior region improves the binding affinity by 20-fold in $K_{\mathrm{Mut}}/K_{\mathrm{WT\text{-}NG}}$ (S14M1-NG). When the native glycosylation is present and the engineered Ser-14 *O*-mannose residue is added, the favorable $\Delta\Delta G$ increases to 3 kcal/mol, resulting in a 140-fold increase in K_{Mut} over $K_{\mathrm{WT\text{-}NG}}$ (S3M1+S14M1).

Table 3.5: Relative binding free energy ($\Delta\Delta G$, kcal/mol) from the engineered glycosylation TI calculations, with a single *O*-mannose residue at Ser-14 (S14M1-NG) with no glycosylation and a single *O*-mannose residue at Ser-14 with the native glycans present (S14M1), and associated change in partition coefficient ($K_{\mathrm{Mut}}/K_{\mathrm{WT\text{-}NG}}$). The S3M1+S14M1 entry is the sum of the S14M1 and S3M1 (see Table 3.4) entries.

Mutation	$\Delta\Delta G$ (kcal/mol)	$K_{\mathrm{Mut}}/K_{\mathrm{WT\text{-}NG}}$
S14M1-NG 	-1.9 ± 0.31	23 ± 12
S14M1 	-2.3 ± 0.09	--
S3M1 + S14M1 	-3.0 ± 0.10	149 ± 24

Interaction and H-bonding Between Residues of Interest and Cellulose

For the Ser-14 systems studied, the change in binding affinity is significantly higher than in the Ser-3 cases. To gain insights into the reasons for this increase in binding affinity, we examined thermodynamic and structural properties from the 100 ns MD simulations. First, we examine the interaction energies: Figure 3.9 shows the total interaction energies, comprising electrostatic and van der Waals energies, between the glycans and the cellulose surface, for the four scenarios studied: S3M1, S14M1-NG, S3M2, and S14M1. We also confirmed that the interaction between Thr-1 *O*-mannose residue and the surface was indeed zero using these simulations. If we consider these interactions with the residues of interest, the addition of a single glycan at Ser-3 (S3M1) or Ser-14 (S14M1-NG) adds an additional 4 to 7.5 kcal/mol of favorable interaction with the cellulose. Furthermore, we find that whereas the addition of the second mannose residue at Ser-3 increases the hydrogen bonding potential (Table 3.6), the total interaction energy with the cellulose surface does not change compared to the S3M1 case alone (Figure 3.9). In contrast, when a *O*-mannose residue is added at Ser-14 with the native glycans present (S14M1), the interaction energy between the Ser-3 *O*-mannose residue and cellulose remains constant, while the Ser-14 *O*-mannose residue adds an additional 3 kcal/mol of favorable interaction with the surface, increasing the total slightly over the non-glycosylated (S14M1-NG) case. In our simulations the ability of the critical CBM binding-face amino acids [14,21] to interact with the surface does not appear to be negatively impacted by the presence of the Ser-3 or Ser-14 *O*-mannose residues. The glycans also form hydrogen bonds with the surface in all cases, and when the Ser-14 *O*-mannose residue is present, the number of hydrogen bonds possible between the Ser-3 *O*-mannose residue and surface nearly doubles.

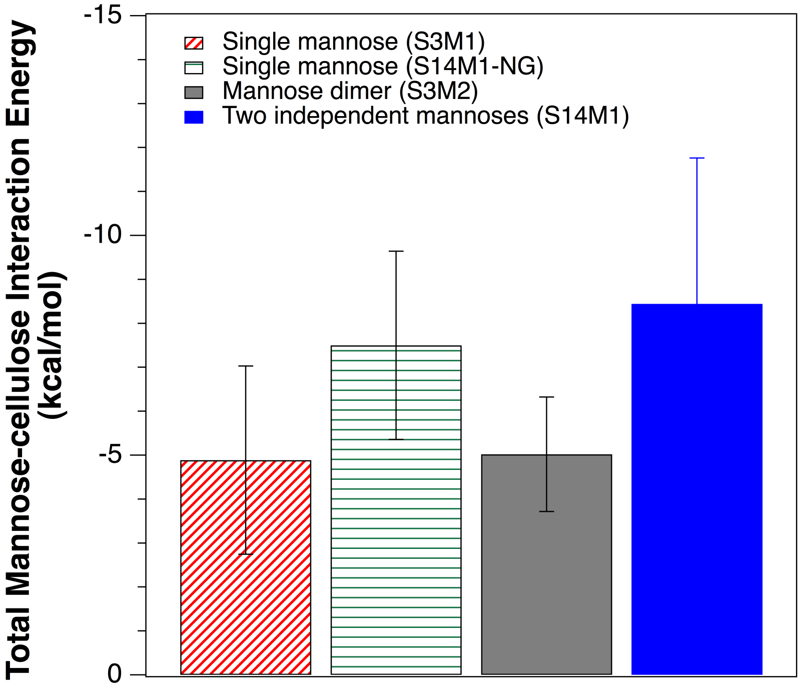


Figure 3.9: Comparison of the total interaction energy (Total Energy = Electrostatic + van der Waals) between the glycans and cellulose surface present in the S3M1, S14M1-NG, S3M2, and S14M1 100 ns MD simulations. The Thr-1 *O*-mannose residue has zero interaction energy with the surface in each simulation.

The interaction energy does not delineate the contributions of hydrogen bonding, which could also be a factor in binding affinity improvements [1,21,51]. While we cannot quantitatively delineate the contributions of mannan-cellulose hydrogen bonding relative to other enthalpic and entropic contributions to improvements in binding free energy, the data suggest that increased hydrogen bonding of the CBM-glycan system with the cellulose surface correlates with improved binding and stability of the CBM on the surface. While we also did not calculate the net increase in hydrogen bonding for the entire system, (*i.e.* water, glycans, protein, and cellulose surface), we did calculate number of potential unique (*i.e.* forming with a different cellulose primary alcohol or different *O*-mannose residue hydroxyl group) hydrogen bonds formed between the *O*-

mannose residues and the cellulose for 1 to 2 bonds forming during the 100 ns MD simulations (Table 3.6).

Table 3.6: Number of unique hydrogen bonds formed between the mannose residue of interest and the cellulose surface during the 100 ns MD simulations using a hydrogen bond cutoff of 3.0 Å and an angle criteria of 60° from linear. The typical duration of each bond was between 10 and 15% of the total run.

	S3M1	S3M2	S14M1-NG	S14M1
Ser-3 single mannose	7	12	NA	11
Ser-3 second mannose	NA	10	NA	NA
Ser-14 single mannose	NA	NA	8	8

The behavior of the Ser-3 *O*-mannose residue is similar for the native glycosylation [23] and our artificial glycosylation case in that we observe one to two stable hydrogen bonds between the Ser-3 *O*-mannose residue and cellulose, which contributes to enhanced protein-carbohydrate interactions. However, as noted in the previously, bonding was found to occur on both sides of the *O*-mannose residue ring in the S14M1 case, which could be indicative of enhanced ring stacking over the cellulose surface. Also, the Ser-3 *O*-mannose residue is closer to the surface with the Ser-14 *O*-mannose residue present, which increases the number of bonds possible in S14M1 over S3M1 (Table 3.6). In both Ser-14 MD simulations, the Ser-14 *O*-mannose residue can also form stable hydrogen bonds with the cellulose surface. Additionally, the bonding of the protein with the surface remains the same or shows slight improvement over the non-glycosylated wild type, which could indicate a net increase of hydrogen bonds for at least the CBM-glycan system and the cellulose substrate and the cellulose. These results support the hypothesis that the addition of glycans may be beneficial to binding, in that the sugars stabilize the CBM via increased hydrogen bonding potential and hydrophobic interactions.

We also calculated the interaction energies between the protein Tyr-5 mutations. From Figure 3.10 we can see that Trp-5 has the highest favorable interaction energy at 12.9 ± 0.82 kcal/mol, which is 2.5 kcal/mol more than Tyr-5 and nearly double that of Ala-5 or the Ser-3 *O*-mannose residue. Intermolecular van der Waals forces contribute 75-80% of the total interactions of Trp-5 with the cellulose surface, which can be attributed to the larger surface area of Trp over Tyr increasing the residue footprint on the cellulose surface [52,53]. For Ala-5, the Ser-3 *O*-mannose residue, and the Ser-14 *O*-mannose residue the electrostatic and van der Waals interactions contribute equally to the total interaction with the surface. We also note that the interaction of Tyr-5 does not change with the presence of a *O*-mannose residue on Ser-3 (data not shown), and that the favorable interaction of the Ser-3 *O*-mannose residue with the surface is approximately 5 kcal/mol for the S3M1 and S14M1 glycosylation simulations (only S3M1 case shown). The disaccharide mannan at Ser-3 does not add additional interaction with the surface over the single *O*-mannose residue, which is corroborated by our finding that the relative binding affinity only slightly increases, with the addition of a second mannose residue at Ser-3. The Tyr-5, Trp-5, and Ala-5 residues maintain a constant distance above the surface of 2 to 3 Å, which is reflected by stable interaction energy profiles. Importantly, the aromatic residues display little rotation away from a parallel configuration, making CH-π interactions possible [53,54]. We also observe a stable hydrogen bond between the hydroxyl group on Tyr-5 and a primary alcohol group on the cellulose surface, consistent with previous simulation findings [21]. Although hydrogen bonds exist between Trp-5 and the surface, no single hydrogen bond is persistent throughout the simulation.

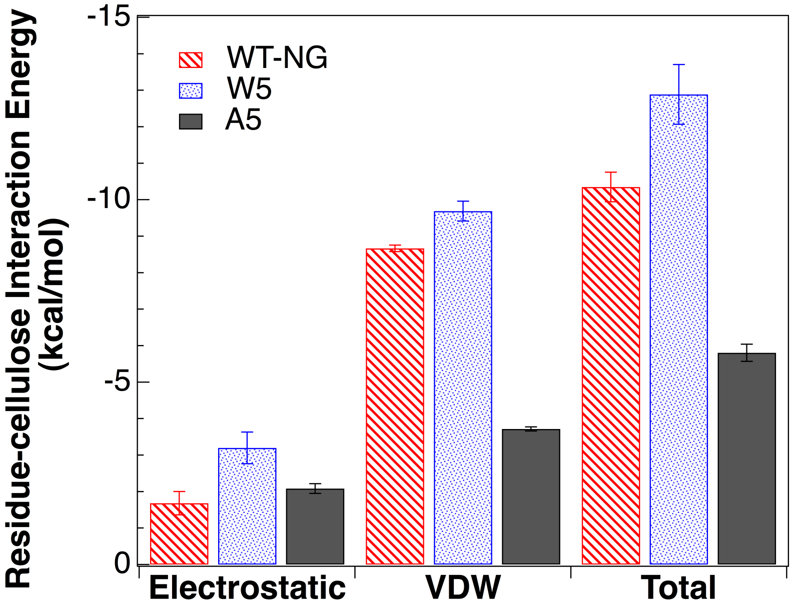


Figure 3.10: Interaction energy between individual residues of interest, Tyr-5 (WT), Trp-5 (W5), Ala-5 (A5) and the cellulose surface over 100 ns NVT simulations. Errors were calculated using block averaging [50].

Protein Backbone Fluctuations With and Without Glycosylation

Finally, we calculated the root mean square deviation (RMSD) and root mean square fluctuation (RMSF) for the extended MD simulations described previously, and found a slight improvement in stability of the CBM backbone over the cellulose surface for the native glycosylated versus non-glycosylated systems. As shown in Figure 3.11, the addition of sugars only impacts the fluctuations in the Ser-14 cases.

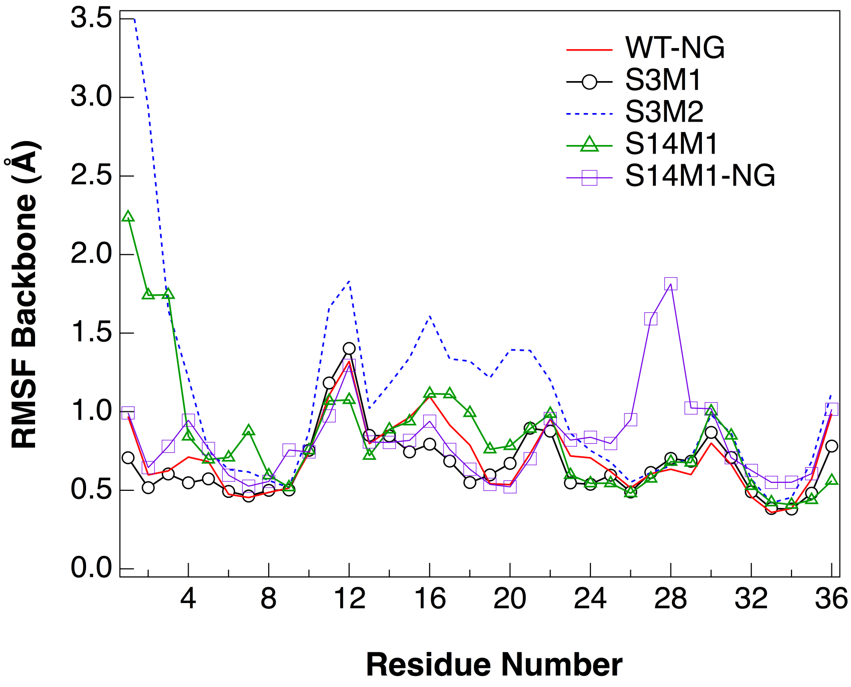


Figure 3.11: RMSF of the CBM backbone for the non-glycosylated wild type and glycosylated variants.

For the 100 ns NVT MD simulations, the RMSD of each CBM backbone was calculated. The curves were smoothed using a binomial smoothing method and the results are presented in Figure 3.12. Beginning with the non-glycosylated systems, Linder *et al.* attribute the decrease in binding for Y5A to a loss in structural compactness caused by changes in the loop between the β-1 and β-2 sheets based on two-dimensional NMR and evaluation of chemical shifts in the backbone [14]. In the simulations, we observe that unlike the RMSD curves for WT and W5, the RMSD of the A5 backbone (Figure 3.12A) shifts and re-stabilizes at a higher value after 40 ns, which corresponds to a shift in the loop from Ser-14 to Thr-24 between β-1 and β-2 sheets. Corresponding to the RMSF calculations, the S3M1 simulation with the native glycosylation pattern has the least fluctuation (Figure 3.12B), indicating that glycosylation could improve stability of the CBM above cellulose over the non-glycosylated wild type.

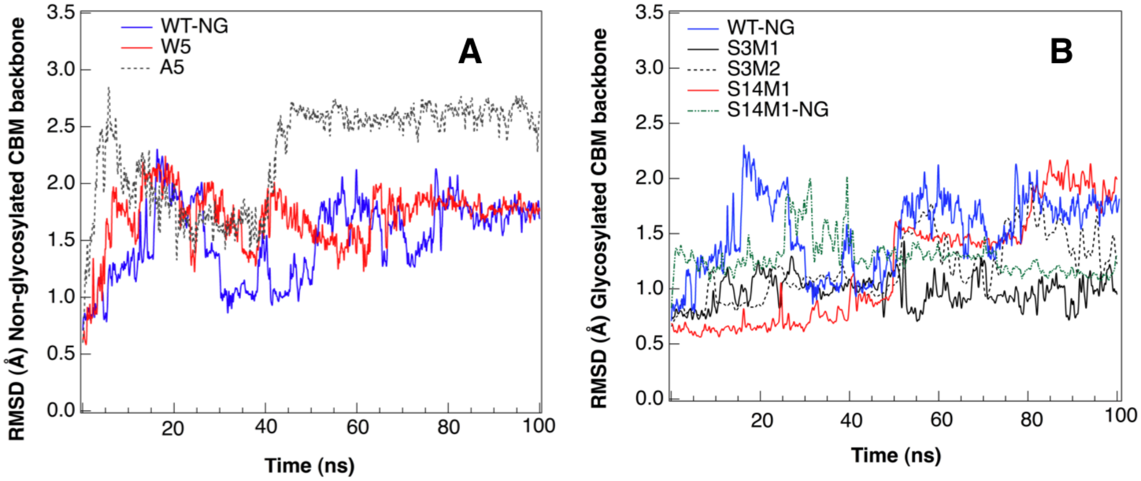


Figure 3.12: (A) RMSD of non-glycosylated systems with amino acid mutations and (B) RMSD of the non-glycosylated wild type CBM backbone compared to the glycosylated CBM backbone systems over 100 ns of NVT simulation.

3.4 Discussion

We have used TI simulations [32,33] to examine changes in the binding affinity of a Family 1 CBM with both amino acid mutations and native and artificial *O*-glycans. The well-characterized *T. reesei* Cel7A CBM, for which glycosylation has been quantified experimentally [23] and biochemical mutation data exist [13,14], was used as a model CBM. From these biochemical data, the computational approach was validated by demonstrating that we can achieve quantitative agreement in binding affinity changes for two large amino acid mutations, as shown in Figure 3. Specifically, we have shown that the Y5W mutation for the *T. reesei* Cel7A CBM improves the binding affinity by 2-fold, whereas the Y5A mutation is detrimental to binding as shown experimentally [13-15]. By producing results consistent with experiments for this system, we have demonstrated that TI can potentially be a useful screening tool for mutations that modify CBM binding affinity.

Following these initial validation simulations, we extended our approach to predict the change in binding affinity for both a native and an artificial *O*-glycan on specific regions of the Cel7A CBM. We predict that a single native *O*-glycan near the posterior of the CBM interacts directly with cellulose and can change the binding affinity by 3-fold, and that the addition of a second, independent, engineered glycan combined with the native glycosylation can change the binding affinity by 140-fold, which is a striking increase over an amino acid mutation alone. The results of the glycan simulations reported here are a promising demonstration of the potential for engineering improved cellulases via the introduction of non-native glycosylation, considering that only a 2-fold increase in binding affinity for the Y5W mutation relative to the wild type non-glycosylated Cel7A CBM resulted in higher activity for the enzyme with the mutant CBM [15]. Since even small glycosylation motifs (a single *O*-mannose residue) can impact the binding affinity, the addition of artificial glycosylation sites via site-directed mutagenesis to either *N*-glycan or *O*-glycan motifs is a potentially powerful strategy to improve the specific activity of glycoside hydrolase enzymes.

Modification of culture growth conditions and expression hosts for glycoproteins is generally known to affect the glycosylation pattern of a given secreted protein [55]. This previous finding has significant implications in light of the current study for comparing the binding affinity and activities of enzymes and enzyme cocktails for biomass conversion purposes. For example, if the expression host or growth conditions vary between protein cultures, changes can arise in experimental observables (e.g., binding and/or enzyme activity) from differences in the extent of glycosylation alone, which can alter the outcome of enzyme screening or directed evolution experiments.

Lastly, many groups are constructing quantitative, mesoscale models of cellulase action to predict enzyme synergy and similar phenomena with the aim to design enhanced cellulase cocktails for biomass conversion [56-58]. These models rely on accurate thermodynamic and kinetic measurements of cellulase-cellulose interactions and insights from advanced experimental techniques and simulation predictions related to the molecular-level mechanisms of cellulase action [10,59-61]. Measurement or simulations of the absolute or relative binding free energies (and partition coefficients) of CBMs to cellulose should account for native glycosylation patterns to obtain accurate measurements or predictions, respectively.

An important question from this study is how both the natural and artificial glycans affect the structure and function of the CBM. We show that the native glycosylation pattern [23] studied here stabilizes the CBM structure slightly, as demonstrated by the changes in hydrogen bonding, RMSF, and RSMD results (Table 3.7 and Figures 3.11 and 3.12, respectively). In terms of the CBM function, we note that it is commonly stated in the literature that aromatic amino acids on the binding faces of CBMs, like Tyr-5 in the Cel7A CBM, primarily interact with cellulose via hydrophobic interactions [13,14,22,24]. However, there are two types of functions to consider when discussing binding affinity: first, an absolute binding affinity wherein the CBM binds to cellulose from solution, and second, the function of the CBM after it is bound, which is typically thought of as translation or processivity along the surface. Here we have examined the effect of the former (CBM binding from solution) with TI calculations. It has yet to be definitively shown if the binding effect of aromatic residues in Family 1 CBMs is primarily due to enthalpic or entropic contributions, which could be demonstrated with isothermal titration calorimetry. For translation along the surface, which is likely relevant to CBM function as part of an enzyme [62], Beckham and coworkers have shown in a previous study [21] that the aromatic residues (Tyr-5 and Tyr-31) and several polar residues form hydrogen bonds approximately every 1 nm (or one cellobiose unit) on the hydrophobic surface of cellulose. These residues are likely critical for the function of Family 1 CBMs, and the importance of these hydrogen bonds is demonstrated in that either Tyr or Trp is usually preferred in these sequence positions in Family 1 CBMs over Phe [21]. In the case of adding *O*-mannose residue to Ser-3 or Ser-14 in the Cel7A CBM, the glycans could have a similar effect on CBM translation as the existing aromatic and polar residues, because *O*-mannose residue presents a large, planar face accompanied by the ability to form hydrogen bonds with primary alcohol groups on cellulose that are positioned uniformly along a given polymer chain (see the Supplemental Information). Therefore it is unlikely that the addition of *O*-mannose residue will significantly affect CBM translation once bound to cellulose, but rather serve as additional surface area for binding. This hypothesis is validated experimentally in that a single *O*-mannose residue exists on the CBM already in functional enzymes at Ser-3 [23]. Moreover, it is unknown for a consistent set of CBMs and whole cellulases, either experimentally or computationally, if the CBM processivity rate differs from the processivity rate (i.e., hydrolysis rate) of an engaged enzyme. Recently, Igarashi *et al.* used high-speed atomic force microscopy to measure the rate of the *T. reesei* Cel7A enzyme acting on cellulose [60], but the diffusion coefficient of Family 1 CBMs on cellulose has not been explicitly measured to our knowledge. If the processivity rate of a CBM is much faster than the combined hydrolysis and processivity rate of the whole enzyme, which it likely is, addition of glycans should not affect the ability of the CBM to translate on a biologically relevant timescale. Thus it is unlikely that the CBM will get "stuck" such that CBM translation becomes the rate-limiting step in processive hydrolysis.

Finally, we note that experimental validation of these computational results is of paramount importance. Expression of the Family 1 CBM in hosts that do not impart glycosylation or production via solid-state synthesis as was conducted previously can yield a non-glycosylated CBM [13,14]. Expression and purification of the CBM from *T. reesei,* or other expression hosts that impart glycosylation, or via chemical synthesis procedures in which glycosylation can be chemically added, could produce a CBM with the glycosylation patterns examined here and binding isotherms measured. However, from a computational standpoint, we note that carefully conducted TI simulations and free energy calculations in general for ligand binding have been shown to yield agreement with experiment within several kcal/mol [63-68]. Here we have obtained results consistent with available experimental data on amino acid mutations, which at the least, lends confidence to the qualitative nature of the predictions regarding glycosylation.

3.5 Conclusions

In summary, our results indicate that CBM glycosylation is a likely contributor to enzyme binding affinity. To our knowledge, this study is the first to apply TI calculations to test the effects of glycosylation on a cellulase enzyme and has broad applicability as many cellulases contain both *N*- and *O*-linked glycosylation. For the Cel7A CBM, the addition of glycans increases hydrogen bonding potential and hydrophobic stacking with cellulose via glycoprotein-carbohydrate interactions, stabilizing the CBM on the surface and improving binding affinity. Our results highlight the need for consideration of post-translational modifications when selecting expression hosts and growth conditions for these types of enzymes [25,28,30,55]. Glycosylation, or lack thereof, could have an impact on binding that can translate into effects on enzyme mechanistic action as a whole. Moreover, the manipulation of glycan sites via recombinant expression, by varying growth conditions, or via addition of artificial glycan sites could be used as a general protein engineering strategy to tune protein-carbohydrate binding affinity for improving cellulases.

3.6 References

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

2. A.W. Blake, L. McCartney, J.E. Flint, D.N. Bolam, A.B. Boraston, H.J. Gilbert and J.P. Knox, "Understanding the biological rationale for the diversity of cellulose-directed carbohydrate-binding modules in prokaryotic enzymes", *J. Biol. Chem.* **2006**, 281 (39), 29321-29329.

3. J. Lehtio, J. Sugiyama, M. Gustavsson, L. Fransson, M. Linder and T.T. Teeri, "The binding specificity and affinity determinants of family 1 and family 3 cellulose binding modules", *Proc. Natl. Acad. Sci. U. S. A.* **2003**, 100 (2), 484-489.

4. S.J. Horn, P. Sikorski, J.B. Cederkvist, G. Vaaje-Kolstad, M. Sorlie, B. Synstad, G. Vriend, K.M. Varum and V.G.H. Eijsink, "Costs and benefits of processivity in enzymatic degradation of recalcitrant polysaccharides", *Proc. Natl. Acad. Sci.* **2006**, 103 (48), 18089-18094.

5. D.M.F. van Aalten, B. Synstad, M.B. Brurberg, E. Hough, B.W. Riise, V.G.H. Eijsink and R.K. Wierenga, "Structure of a two-domain chitotriosidase from *Serratia marcescens* at 1.9 A resolution", *Proc. Natl. Acad. Sci.* **2000**, 97, 5842-5847.

6. M.E. Himmel, S.Y. Ding, D.K. Johnson, W.S. Adney, M.R. Nimlos, J.W. Brady and T.D. Foust, "Biomass recalcitrance: Engineering plants and enzymes for biofuels production", *Science* **2007**, 315 (5813), 804-807.

7. A.J. Ragauskas, C.K. Williams, B.H. Davison, G. Britovsek, J. Cairney, C.A. Eckert, W.J. Frederick, J.P. Hallett, D.J. Leak, C.L. Liotta, J.R. Mielenz, R. Murphy, R. Templer and T. Tschaplinski, "The path forward for biofuels and biomaterials", *Science* **2006**, 311 (5760), 484-489.

8. D.B. Wilson, "Cellulases and biofuels", *Curr. Opin. Biotech.* **2009**, 20, 295-299.

9. Y.H.P. Zhang, M.E. Himmel and J.R. Mielenz, "Outlook for cellulase improvement: Screening and selection strategies", *Biotechnol. Adv.* **2006**, 24 (5), 452-481.

10. S.P.S. Chundawat, G.T. Beckham, M.E. Himmel and B.E. Dale, "Deconstruction of Lignocellulosic Biomass to Fuels and Chemicals", *Annu. Rev. Chem. Biomol. Eng.* **2011**, 2, 6.1-6.25.

11. S.E. Lantz, F. Goedegebuur, R. Hommes, T. Kaper, B.R. Kelemen, C. Mitchinson, L. Wallace, J. Stahlberg and E.A. Larenas, "Hypocrea jecorina Cel6A protein engineering", *Biotechnol. Biofuels* **2010**, 3 (20).

12. P. Heinzelman, C.D. Snow, M.A. Smith, X.L. Yu, A. Kannan, K. Boulware, A. Villalobos, S. Govindarajan, J. Minshull and F.H. Arnold, "SCHEMA Recombination of a Fungal Cellulase Uncovers a Single Mutation That Contributes Markedly to Stability", *J. Biol. Chem.* **2009**, 284 (39), 26229-26233.

13. M. Linder, G. Lindeberg, T. Reinikainen, T.T. Teeri and G. Pettersson, "The difference in affinity between 2 fungal cellulose-binding domains is dominated by a single amino-acid substitution", *FEBS Lett.* **1995**, 372 (1), 96-98.

14. M. Linder, M.L. Mattinen, M. Kontteli, G. Lindeberg, J. Stahlberg, T. Drakenberg, T. Reinikainen, G. Pettersson and A. Annila, "Identification of functionally important amino-acids in the cellulose-binding domain of *Trichoderma reesei* Cellobiohydrolase I", *Protein Sci.* **1995**, 4 (6), 1056-1064.

15. S. Takashima, M. Ohno, M. Hidaka, A. Nakamura and H. Masaki, "Correlation between cellulose binding and activity of cellulose-binding domain mutants of Humicola grisea cellobiohydrolase 1", *FEBS Lett.* **2007**, 581 (30), 5891-5896.

16. T.W. Kim, H.A. Chokhawala, D. Nadler, H.W. Blanch and D.S. Clark, "Binding Modules Alter the Activity of Chimeric Cellulases: Effects of Biomass Pretreatment and Enzyme Source", *Biotechnol. Bioeng.* **2010**, 107 (4), 601-611.

17. C. Divne, J. Stahlberg, T. Reinikainen, L. Ruohonen, G. Pettersson, J.K.C. Knowles, T.T. Teeri and T.A. Jones, "The 3-dimensional crystal-structure of the catalytic core of Cellobiohydrolase-I from *Trichodermal reesei*", *Science* **1994**, 265 (5171), 524-528.

18. C. Divne, J. Stahlberg, T.T. Teeri and T.A. Jones, "High-resolution crystal structures reveal how a cellulose chain is bound in the 50 angstrom long tunnel of cellobiohydrolase I from Trichoderma reesei", *J. Mol. Biol.* **1998**, 275 (2), 309-325.

19. L.H. Zhong, J.F. Matthews, P.I. Hansen, M.F. Crowley, J.M. Cleary, R.C. Walker, M.R. Nimlos, C.L. Brooks, W.S. Adney, M.E. Himmel and J.W. Brady, "Computational simulations of the Trichoderma reesei cellobiohydrolase I acting on microcrystalline cellulose I beta: the enzyme-substrate complex", *Carbohydr. Res.* **2009**, 344 (15), 1984-1992.

20. B.L. Cantarel, P.M. Coutinho, C. Rancurel, T. Bernard, V. Lombard and B. Henrissat, "The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics", *Nucleic Acids Res.* **2009**, 37, D233-D238.

21. G.T. Beckham, J.F. Matthews, Y.J. Bomble, L.T. Bu, W.S. Adney, M.E. Himmel, M.R. Nimlos and M.F. Crowley, "Identification of Amino Acids Responsible for Processivity in a Family 1 Carbohydrate-Binding Module from a Fungal Cellulase", *J. Phys. Chem. B* **2009**, 114 (3), 1447-1453.

22. P.J. Kraulis, G.M. Clore, M. Nilges, T.A. Jones, G. Pettersson, J. Knowles and A.M. Gronenborn, "Determination of the 3-dimensional solution structure of the C-terminal domani of Cellobiohydrolase-I from Trichoderma reesei - A study using Nuclear Magnetic Resonance and hybrid distance geometry dynamical simulated annealing", *Biochemistry* **1989**, 28 (18), 7241-7257.

23. M.J. Harrison, A.S. Nouwens, D.R. Jardine, N.E. Zachara, A.A. Gooley, H. Nevalainen and N.H. Packer, "Modified glycosylation of cellobiohydrolase I from a high cellulase-producing mutant strain of Trichoderma reesei", *Eur. J. Biochem.* **1998**, 256 (1), 119-127.

24. A.L. Creagh, E. Ong, E. Jervis, D.G. Kilburn and C.A. Haynes, "Binding of the cellulose-binding domain of exoglucanase Cex from Cellulomonas fimi to insoluble microcrystalline cellulose is entropically driven", *Proc. Natl. Acad. Sci. U. S. A.* **1996**, 93 (22), 12229-12234.

25. A.B. Boraston, R.A.J. Warren and D.G. Kilburn, "Glycosylation by Pichia pastoris decreases the affinity of a family 2a carbohydrate-binding module from Cellulomonas fimi: a functional and mutational analysis", *Biochem. J.* **2001**, 358, 423-430.

26. A.B. Boraston, L.E. Sandercock, R.A.J. Warren and D.G. Kilburn, "O-glycosylation of a recombinant carbohydrate-binding module mutant secreted by Pichia pastoris", *J. Mol. Microbiol. Biotechnol.* **2003**, 5 (1), 29-36.

27. N. Deshpande, M.R. Wilkins, N. Packer and H. Nevalainen, "Protein glycosylation pathways in filamentous fungi", *Glycobiology* **2008**, 18 (8), 626-637.

28. W.S. Adney, T. Jeoh, G.T. Beckham, Y.C. Chou, J.O. Baker, W. Michener, R. Brunecky and M.E. Himmel, "Probing the role of N-linked glycans in the stability and activity of fungal cellobiohydrolases by mutational analysis", *Cellulose* **2009**, 16 (4), 699-709.

29. G.T. Beckham, Y.J. Bomble, J.F. Matthews, C.B. Taylor, M.G. Resch, J.M. Yarbrough, S.R. Decker, L.T. Bu, X.C. Zhao, C. McCabe, J. Wohlert, M. Bergenstrahle, J.W. Brady, W.S. Adney, M.E. Himmel and M.F. Crowley, "The O-Glycosylated Linker from the Trichoderma reesei Family 7 Cellulase Is a Flexible, Disordered Protein", *Biophys. J.* **2010**, 99 (11), 3773-3781.

30. T. Jeoh, W. Michener, M.E. Himmel, S.R. Decker and W.S. Adney, "Implications of cellobiohydrolase glycosylation for use in biomass conversion", *Biotechnol. Biofuels* **2008**, 1, 12.

31. T. Eriksson, I. Stals, A. Collen, F. Tjerneld, M. Claeyssens, H. Stalbrand and H. Brumer, "Heterogeneity of homologously expressed Hypocrea jecorina (Trichoderma reesei) Cel7B catalytic module", *Eur. J. Biochem.* **2004**, 271 (7), 1266-1276.

32. P. Kollman, "Free energy calculations: Applications to chemical and biological phenomena", *Chem. Rev.* **1993**, 93, 2395-2417.

33. A. Pohorille, Jarzynski, C., & Chipot, C., "Good Practices in Free-Energy Calculations", *J. Phys. Chem. B* **2010**, (114), 10235-10253.

34. B.R. Brooks, C.L. Brooks, A.D. Mackerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, W. Yang, D.M. York and M. Karplus, "CHARMM: The Biomolecular Simulation Program", *J. Comput. Chem.* **2009**, 30 (10), 1545-1614.

35. J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale and K. Schulten, "Scalable molecular dynamics with NAMD", *J. Comput. Chem.* **2005**, 26 (16), 1781-1802.

36. W. Humphrey, A. Dalke and K. Schulten, "VMD: Visual molecular dynamics", *J. Mol. Graph.* **1996**, 14 (1), 33-&.

37. A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins", *J. Phys. Chem. B* **1998**, 102 (18), 3586-3616.

38. A.D. Mackerell, M. Feig and C.L. Brooks, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations", *J. Comput. Chem.* **2004**, 25 (11), 1400-1415.

39. O. Guvench, E. Hatcher, R.M. Venable, R.W. Pastor and A.D. MacKerell, "CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses", *Journal of Chemical Theory and Computation* **2009**, 5 (9), 2353-2370.

40. O. Guvench, S.N. Greene, G. Kamath, J.W. Brady, R.M. Venable, R.W. Pastor and A.D. Mackerell, "Additive Empirical Force Field for Hexopyranose Monosaccharides", *Journal of Computational Chemistry* **2008**, 29 (15), 2543-2564.

41. W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey and M.L. Klein, "Comparison of simple potential functions for simulating liquid water", *J. Chem. Phys.* **1983**, 79 (2), 926-935.

42. S.R. Durell, B.R. Brooks and A. Bennaim, "Solvent-induced forces between 2 hydrophilic groups", *J. Phys. Chem.* **1994**, 98 (8), 2198-2202.

43. Y. Nishiyama, P. Langan and H. Chanzy, "Crystal structure and hydrogen-bonding system in cellulose 1 beta from synchrotron X-ray and neutron fiber diffraction", *J. Am. Chem. Soc.* **2002**, 124 (31), 9074-9082.

44. A. Varki, *Essentials of Glycobiology*. Cold Springs Harbor Laboratory Press: Cold Springs Harbor, NY, USA, **2009**.

45. U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee and L.G. Pedersen, "A Smooth Particle Mesh Ewald Method", *J. Chem. Phys.* **1995**, 103 (19), 8577-8593.

46. J.P. Ryckaert, G. Ciccotti and H.J.C. Berendsen, "Numerical-integration of cartesian equations of motion of a system with constraints - Molecular-dynamiccs of N-alkanes", *J. Comput. Phys.* **1977**, 23 (3), 327-341.

47. T. Steinbrecher, D.L. Mobley and D.A. Case, "Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations", *J. Chem. Phys.* **2007**, 127 (21).

48. S. Bruckner and S. Boresch, "Efficiency of alchemical free energy simulations. II. Improvements for thermodynamic integration", *J. of Comp. Chem.* **2010**, 32 (7), 1320-1333.

49. H. Paliwal and M.R. Shirts, "A Benchmark Test Set for Alchemical Free Energy Transformations and Its Use to Quantify Error in Common Free Energy Methods", *Journal of Chemical Theory and Computation* **2011**.

50. D.S. Frenkel, B. , *Understanding Molecular Simulation: From Algorithms to Applications*. 2nd ed ed.; Academic Press: San Diego, CA, **2002**.

51. M.R. Nimlos, J.F. Matthews, M.F. Crowley, R.C. Walker, G. Chukkapalli, J.V. Brady, W.S. Adney, J.M. Clearyl, L.H. Zhong and M.E. Himmel, "Molecular modeling suggests induced fit of Family I carbohydrate-binding modules with a broken-chain cellulose surface", *Protein Eng. Des. Sel.* **2007**, 20 (4), 179-187.

52. Z.R. Laughrey, S.E. Kiehna, A.J. Riemen and M.L. Waters, "Carbohydrate-pi Interactions: What Are They Worth?", *J. Am. Chem. Soc.* **2008**, 130 (44), 14625-14633.

53. M.S. Sujatha, Y.U. Sasidhar and P.V. Balaji, "Energetics of galactose- and glucose-aromatic amino acid interactions: Implications for binding in galactose-specific proteins", *Protein Sci.* **2004**, 13 (9), 2502-2514.

54. M.D. Fernandez, F.J. Canada, J. Jimenez-Barbero and G. Cuevas, "Molecular recognition of saccharides by proteins. Insights on the origin of the carbohydrate-aromatic interactions", *J. Am. Chem. Soc.* **2005**, 127 (20), 7379-7386.

55. I. Stals, K. Sandra, B. Devreese, J. Van Beeumen and M. Claeyssens, "Factors influencing glycosylation of Trichoderma reesei cellulases. II: N-glycosylation of Cel7A core protein isolated from different strains", *Glycobiology* **2004**, 14 (8), 725-737.

56. P. Bansal, M. Hall, M.J. Realff, J.H. Lee and A.S. Bommarius, "Modeling cellulase kinetics on lignocellulosic substrates", *Biotechnol. Adv.* **2009**, 27 (6), 833-848.

57. S.E. Levine, J.M. Fox, H.W. Blanch and D.S. Clark, "A Mechanistic Model of the Enzymatic Hydrolysis of Cellulose", *Biotechnol. Bioeng.* **2010**, 107 (1), 37-51.

58. W. Zhou, Z.Q. Hao, Y. Xu and H.B. Schuttler, "Cellulose Hydrolysis in Evolving Substrate Morphologies II: Numerical Results and Analysis", *Biotechnol. Bioeng.* **2009**, 104 (2), 275-289.

59. G.T. Beckham, Bomble, Y. J., Bayer, E.A., Himmel, M.E., and Crowley, M. F., "Applications of computational science for understanding enzymatic deconstruction of cellulose", *Curr. Opin. Biotechnol.* **2011**, 22, 1-8.

60. K. Igarashi, A. Koivula, M. Wada, S. Kimura, M. Penttila and M. Samejima, "High Speed Atomic Force Microscopy Visualizes Processive Movement of Trichoderma

reesei Cellobiohydrolase I on Crystalline Cellulose", *J. Biol. Chem.* **2009**, 284 (52), 36186-36190.

61. X. Zhao, T.R. Rignall, C. McCabe, W.S. Adney and M.E. Himmel, "Molecular simulation evidence for processive motion of Trichoderma reesei Cel7A during cellulose depolymerization", *Chem. Phys. Lett.* **2008**, 460 (1-3), 284-288.

62. L.T. Bu, G.T. Beckham, M.F. Crowley, C.H. Chang, J.F. Matthews, Y.J. Bomble, W.S. Adney, M.E. Himmel and M.R. Nimlos, "The Energy Landscape for the Interaction of the Family 1 Carbohydrate-Binding Module and the Cellulose Surface is Altered by Hydrolyzed Glycosidic Bonds", *J. Phys. Chem. B* **2009**, 113 (31), 10994-11002.

63. D.L. Mobley, A.P. Graves, J.D. Chodera, A.C. McReynolds, B.K. Shoichet and K.A. Dill, "Predicting absolute ligand binding free energies to a simple model site", *J. Mol. Biol.* **2007**, 371 (4), 1118-1134.

64. D.L. Mobley, J.D. Chodera and K.A. Dill, "Confine-and-release method: Obtaining correct binding free energies in the presence of protein conformational change", *Journal of Chemical Theory and Computation* **2007**, 3 (4), 1231-1235.

65. Y. Deng and B. Roux, "Computations of standard binding free energies with molecular dynamics simulations", *The journal of physical chemistry* **2009**, 113 (8), 2234-46.

66. J. Wang, Y. Deng and B. Roux, "Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials", *Biophys J* **2006**, 91 (8), 2798-814.

67. H. Fujitani, Y. Tanida, M. Ito, G. Jayachandran, C.D. Snow, M.R. Shirts, E.J. Sorin and V.S. Pande, "Direct calculation of the binding free energies of FKBP ligands", *J. Chem. Phys.* **2005**, 123 (8), 5.

68. S.E. Boyce, D.L. Mobley, G.J. Rocklin, A.P. Graves, K.A. Dill and B.K. Shoichet, "Predicting Ligand Binding Affinity with Alchemical Free Energy Methods in a Polar Model Binding Site", *J. Mol. Biol.* **2009**, 394 (4), 747-763.

CHAPTER 4

IMPACTS OF O-GLYCOSYLATION ON THE STRUCTURE AND FUNCTION OF CBMS IN FUNGI AND YEASTS

4.1 Introduction

As discussed in Chapters I and III, glycosylation, the attachment of carbohydrates to the sidechains of proteins and lipids, is a structurally diverse post-translational modification that is widely varied in nature as a result of the diversity of glycosyltransferase enzymes found across the kingdoms of life [1,2]. The most prevalent forms of glycans in protein are *N*-glycans, where the glycans attach to the β-amide group of an asparagine (N) residue in a N-X-Ser/Thr motif (where X is any amino acid except proline), and *O*-glycosylation, where glycans attach at the β-hydroxyl group of hydroxylysine, hydroxyproline, Ser, or Thr [3]. *N*-glycans are typically composed an *N*-acetylglucosamine (GlcNAc) disaccharide directly attached to the protein side chain followed by branched carbohydrate oligomers. In contrast, *O*-glycans can range from smaller structures composed of mannose, glucose, galactose, or *N*-acetylgalactosamine to much longer branched structures of single or mixed carbohydrates [1,4]. The outer termini modifications of both *N*- and *O*-glycans can also contain sialic acid, glycosaminoglycans, sulfate, and phosphate, adding to the diversity of possible glycoprotein structures [1,2].

Generally, glycosylation plays important roles in almost all biochemical and cellular processes such as signaling, protein secretion, protein stability, and proteolysis protection [2-17]. Protein pharmaceuticals are heavily influenced by glycosylation, with many currently approved therapeutics requiring correct glycosylation in order to maintain molecular integrity and exhibit optimal efficacy [4,8,12-14,18]. In addition, glycans can promote or inhibit both intra- and intermolecular interactions [1,2,4,9,19]. For example, the HIV virus produces an envelope protein that is decorated with a "shield" of glycans to prevent antibody recognition [20] and some mammalian glycans can be so large that they contribute a mass or charge to the glycoprotein that can inhibit protein-protein binding [4]. Additionally, binding sites to proteins or other structures for low-molecular weight glycans can achieve high binding selectivity via a combination of hydrogen bonding of the glycan hydroxyl groups and stacking of glycan rings against aromatic side chains, increasing van der Waals or CH-π interactions [1,4,13,18,21]. While general contributions are known, the study of glycosylation on protein structure and function represents an interesting challenge in biophysical chemistry due to microheterogeneity, with the added complexity that external growth conditions and other factors influencing expression in a given organism can impact these patterns [1,9,10]. This microheterogeneity often creates significant challenges in determining the impact glycans have on protein structure-function relationships, which can change when heterologous hosts and growth media are used in protein production [10,12,18,22]. However, technological advances in glycomics [1,2,23] and accompanying advances in genomics and proteomics have improved profiling of glycoproteins [2]. Additionally, improvements in glycan synthesis via chemical, genetic, and enzymatic pathways that achieve homogeneity in glycoforms have also provided systematic methods to study protein-glycan interactions and glycosylation-dependent processes [22,24-26].

In this chapter we examine the impact of *O*-glycosyation on carbohydrate-binding proteins. *O*-glycosylation is imparted to proteins via the endoplasmic reticulum-Golgi pathway, catalyzed by mannosyltranferases [5]. *O*-glycosylation pathways have been extensively studied in mammalian systems, in *Saccharomyces cerevisiae*, and in filamentous fungi, such as the *Aspergillus* and *Trichoderma* genera [1,3-5,27]. *S. cerevisiae*, or baker's yeast and by far the best-studied fungal organism, is a single-celled eukaryote producing an amazing assortment of different glycoproteins making it an excellent model system for the study of glycan pathways [1]. *Aspergillus* and *Trichoderma* are also well-studied genera that produce cellulase enzymes, which are of interest in this work. Filamentous fungi and yeast in particular are useful as heterologous production hosts and model organisms for studying the impacts of glycosylation as they secrete large amounts of proteins into the growth media, often with shorter processing times and simpler growth media than mammalian cells [1,3,5,27]. The most prevalent *O*-glycosylation patterns in fungi and yeasts, and thus the ones examined in this study (Figure 4.1), are α-1,2 linked mannose glycans, or mannans [1,3,5]. The genes encoding the transferases in *T. reesei* show a higher similarity to *S. cerevisiae* than to *Aspergillus* fungi, explaining why branched and glucosyl patterns are expressed only in *Aspergillus* [3]. Sulfate [28] and phosphate [29] groups have also been identified as terminally attached to mannan disaccharide in *T. reesei* but not in the filamentous *A. niger* or *A. awamori*. *N*-glycans in yeasts such as *S. cerevisiae* and *Pichia pastoris,* another yeast used often as an expression host, have been reported with mannose phosphate diester, while phosphorylated *O*-glycans in yeast are only rarely reported [17,30]. Though mammalian cell *O*-glycans are most often composed of the base *O*-GalNAc [1], they can also adopt many of the linear mannan patterns shown in Figure 4.1, and linear glycosaminoglycans linked to Ser/Thr are often heavily sulfated [4]. The range of glycan motifs chosen for study (shown in Figure 4.1) are small enough to be computationally tractable yet still provide a valuable molecular-level understanding of the impact of *O*-glycans on protein structure and function.

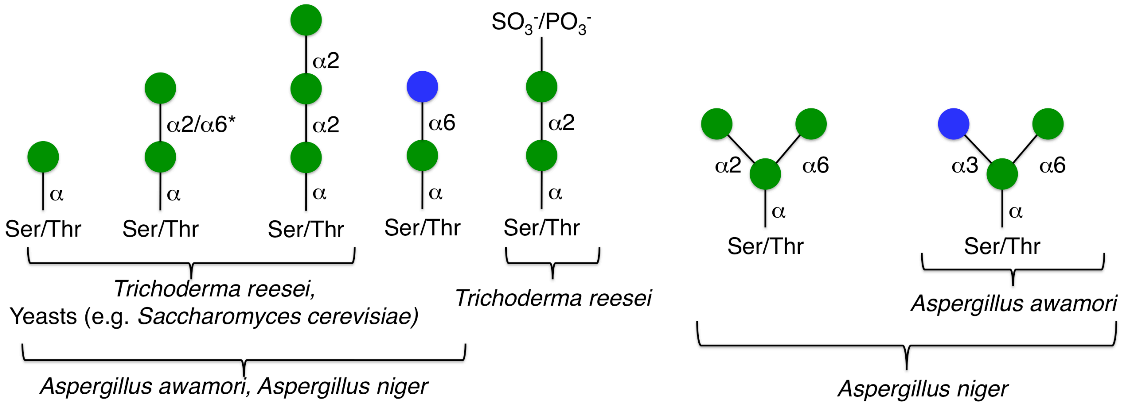


Figure 4.1: O-glycosylation patterns found in filamentous fungi and yeast. Mannose residues are shown in green and glucose residues are shown in blue. *The a-1,6 linear mannan disaccharide occurs in A. niger, not T. reesei or yeast [1,3,5,28,29].

Many biomass-degrading fungi produce glycoside hydrolase (GH) Family 6 and 7 processive cellobiohydrolases in abundance, which offer the majority of hydrolytic potential to degrade cellulose, and may also exhibit glycosylation [31]. Understanding cellobiohydrolase action on cellulose, including a detailed understanding of the role glycosylation plays, is critical to cellulase engineering [18,31,32] and driven by research to accelerate production of biofuels from lignocellulose [33-35]. Most fungal cellobiohydrolases are multi-modular enzymes compromised of a large catalytic domain and a Family 1 CBM that are connected by an *O*-glycosylated linker [28,29,31,36,37].

However, only Cel7A in *T. reesei* has been characterized in detail for *O*-glycosylation over the entire length of the protein. As mentioned in Chapter I, Harrison *et al.* identified 14-26 attached hexose sugars in single, di- and tri-mannan form on the Cel7A CBM and linker, with at least a single *O*-mannose residue at Thr-1 and Ser-3 on the CBM [28]. A sulfate was also detected [28] and Hui *et al.* identified a phosphate [29] attached to a mannan disaccharide, though the location of the sulfate or phosphate on the CBM or linker was not determined. The CBM serves as the primary means for protein-carbohydrate recognition, targeting specific faces of cellulose and for maintaining the proximity of the enzyme to the surface [38,39]. The CBM also has the potential to improve enzyme performance; previous studies demonstrated increasing affinity for cellulose improved activity [34,40,41]. The impacts of *O*-glycan structures on the binding affinity of the Cel7A CBM is an interesting model system for elucidating mechanisms by which glycans identify and bind with external proteins and substrates, and how their presence affects protein structure; the results may also be particularly relevant to systems beyond Family 1 CBMs given the significant roles glycans play in cell adhesion and intermolecular recognition [1,2,4,9,19].

Mannans in *T. reesei* are covalently-bonded to Thr or Ser by an α-O linkage with subsequent linear mannan chains formed via α-1,2 linkages [3]. As previously discussed in Chapter III, experimental studies have shown that altering aromaticity or polarity of key residues significantly alters binding affinity and thus activity [40-42]. Cellulase activity experiments have been performed where *N*-glycans or *O*-glycans were over-expressed in various fungal hosts, with negative or neutral impacts on activity [32,43-45], but experiments have not been performed on the impact of *O*-glycans on Family 1 CBM binding or Cel7A activity. In Chapter III, we demonstrated that the relative binding affinity changes of amino acid mutations can be accurately calculated using thermodynamic integration and molecular dynamics simulation. Also, the addition of single or disaccharide glycans at the sites identified by Harrison *et al.*, referred to herein as the native sites, was subsequently studied and a 3- to 6-fold increase in binding partition coefficient over a non-glycosylated wild type found [28,46]. In this chapter we seek to more generally investigate how the interaction and affinity of *O*-glycans and cellulose change with the variations in binding patterns that arise from different yeast and fungal species or growth conditions. The primary question we address with thermodynamic integration is if a single *O*-mannose residue can improve affinity, will larger glycoforms have an additive impact or is there a limit to the benefits? We also aim to determine whether or not glycoprotein engineering could be implemented as a strategy to improve binding affinity, and potentially performance, of cellulases, though the implications of these findings are not limited to Family 1 CBMs. Since glycans are key in biological recognition events [1,9,19], this study highlights the need for careful consideration of glycosylation in general binding affinity studies and attempts to identify how glycan chemistry and linkage impact binding interactions. Additionally, the inherent flexibility of glycosidic linkages makes single crystal structures of glycoforms difficult to solve, but simulation provides a means to generally visualize the size and shape of the glycans and any changes the glycans impart to the protein backbone [1,7]. While tuning of glycosylation is challenging [47], successful implementation could propose a potential new direction in protein engineering through manipulation of available glycosylation patterns via chemical synthesis and heterologous host or gene expression [1,2,15,22,24,25,48,49].

4.2 Computational Approach

Thermodynamic Integration

In experimental binding affinity studies, concentrations of the non-glycosylated CBM variants adsorbed to the surface and free in solution are used to generate adsorption isotherms.[41,42] From these isotherms, partition coefficients can be determined, and the relative binding free energy ($\Delta\Delta G$) calculated as:

$$\Delta\Delta G = \Delta G_{\text{Mut(Bound-Free)}} - \Delta G_{\text{WT-NG(Bound-Free)}} = -RT\ln\left(K_{\text{Mut}}/K_{\text{WT-NG}}\right) \quad (4.1)$$

where $\Delta G_{\text{Mut(Bound-Free)}}$ and $\Delta G_{\text{WT-NG(Bound-Free)}}$ are the free energies of binding and K_{Mut} and $K_{\text{WT-NG}}$ are the partition coefficients for the mutant CBM of interest and the wild type CBM without glycosylation (NG = no glycosylation), respectively, T is temperature, and R is the gas constant. Thermodynamic integration (TI) simulations [50-53] measure $\Delta\Delta G$, separated into electrostatic ($\Delta\Delta G_{\text{Elec}}$) and van der Waals ($\Delta\Delta G_{\text{VDW}}$) components, between the CBM in solution (i.e., no cellulose) and the CBM on the hydrophobic face of cellulose (i.e., CBM bound to cellulose) as a function of the addition of glycosylation; the thermodynamic cycle thus consists of separate TI calculations of the free CBM and the CBM on the hydrophobic face of cellulose. Large mutations, such as a non-glycosylated backbone to a backbone glycosylated with a mannan disaccharide, can cause simulation instability and high error [46,53] Thus, beginning with a non-glycosylated backbone, a single mannose or glucose residue was added per TI cycle (Figure 4.2) to build to the final, cumulative glycoforms shown in Figure 4.1 and Table 4.1 [1,3,5,28,29]. The wild type structure for S3M1 is glycosylated with a single *O*-mannose residue at Thr-1, but is assumed to be equivalent to a non-glycosylated structure since the Thr-1 *O*-mannose does not interact with the surface. For the remaining systems, the “wild type” is directly left of the system of interest in Figure 4.2, and the final results reported in Table 4.1 and Figure 4.3 are the cumulative results from a non-glycosylated wild type to the final structure, shown to the right of the equality sign.

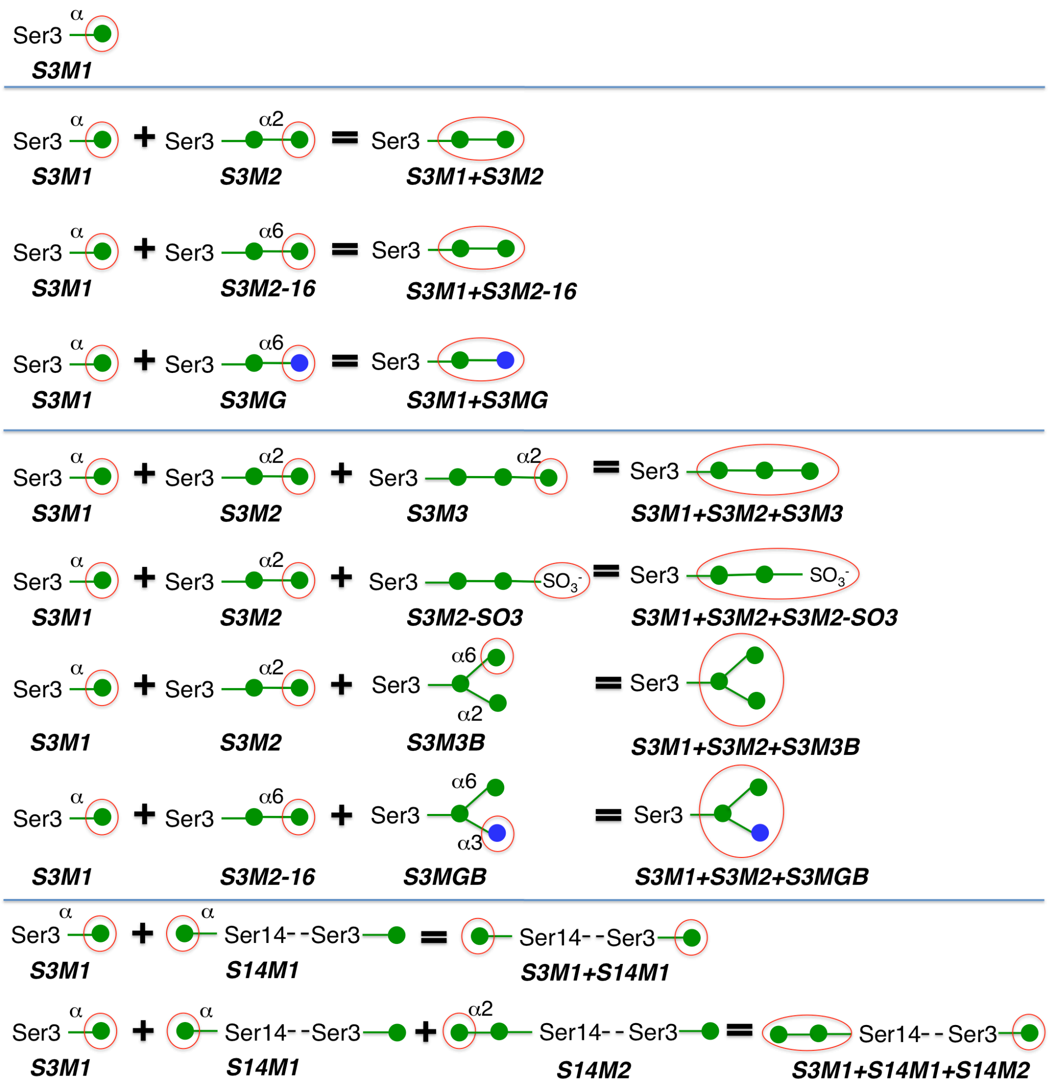


Figure 4.2: Simplified schematic of the simulations performed in the current study. Mannoses are shown in green and glucose in blue, and the mutated states are circled in red. A single O-mannose residue is present on Thr-1 in all cases [28]. The final states to the right of the equality signs are cumulative from a non-glycosylated wild type structure. The sulfate is attached at C2 on the second mannose residue [54]. S3M1, S3M2, and S14M1 are included from Chapter III for completedness.

CHARMM [55] was used to build the hybrid protein structures from the known NMR structure [56]. The cellulose Iβ crystal structure was used to generate the cellulose slab [57]. The cellulose slab thickness, the CBM positioning above the surface, and overall dimensions, were taken from Beckham *et al.* [31] All of the mannose residues linked to Thr-1, Ser-3, or Ser-14 were attached via an α-O linkage and the subsequent mannan and mannan-glucan patterns, shown in Figure 4.1, were linked as described by Desphande *et al.*, Goto *et al.* and Varki *et al.* [1,3,5]. NAMD [58] was used for all equilibration simulations and TI calculations and VMD [59] was used for visualization. The CHARMM27 force field with the CMAP correction [55,60,61] was used to describe the protein, while cellulose and *O*-glycosylation were modeled using the CHARMM35 carbohydrate force field [62,63]. The recently published CHARMM36 phosphate and sulfate force field was used to describe the mannan disaccharide-sulfate linkage [54]. Water was modeled using the modified TIP3P force field [64,65]. TI was performed using the dual-topology method [53], which is implemented by equilibrating a single structure with a hybrid residue containing both the wild type and mutated atoms.

The solvated, bound system of the CBM over a Iβ cellulose slab [31,57] contained approximately 18,100 atoms. Particle mesh Ewald [66] was used to describe the long-range electrostatic interactions with a sixth order b-spline interpolation, a Gaussian distribution with a width of 0.312 Å, and a mesh size of 60 x 60 x 45. A non-bonded interaction cutoff of 10 Å was used. The SHAKE algorithm [67] was employed to fix covalent bonds to hydrogen atoms. The CBM-cellulose system (referred to as the bound system), was minimized for 2,000 steps, and then equilibrated in the NVT ensemble at 300 K using a 2 fs timestep for 2 ns, at which time the RMSD of the protein backbone stabilized. To calculate relative binding free energy, systems without cellulose were also prepared. The wild type CBM and mutated CBMs structures were solvated in CHARMM with approximately 4,000 water molecules and simulations performed under the same conditions as those with the cellulose surface. The equilibrated, final coordinates of each system, bound and free, were used as the starting coordinates for the TI simulations.

In NAMD, the TI caculations were performed using the dual-topology method [53], implemented by equilibrating a single structure with a hybrid residue containing both the wild type and mutated atoms. The electrostatic and van der Waals calculations were decoupled, reducing computational effort and eliminating instabilities arising from large energy interactions [53]. Optimal window spacing, simulation time, and the error analysis method suitable for this system was determined in our study reviewed in Chapter III [46,53,68,69] and repeated here. The electrostatic and van der Waals calculations comprised 11 equidistant λ windows from 0 to 1 with two additional windows at 0.05 and 0.95 each. Using a 1 fs timestep, the electrostatic windows were each equilibrated for 0.5 ns before 10 ns TI NVT runs, and the van der Waals windows were each equilibrated for 4 ns before 20 ns TI NVT runs. Since the sulfated structure is charged, equilibration was extended to 1 ns for the electrostatics portion before TI data collection began.

The error for each window, Δ_j, was calculated using a combination of the methods outlined by Steinbrecher *et al.* [68] and Paliwal and Shirts [69], described in Chapter II, using Equations 2.24-2.27. Each window exhibited a single exponential decay and the window correlation time was calculated by dividing the total window into 1 ns increments and averaging τ over the first 100-150 ps of each increment. The total error from $\lambda = 0$ to 1 was then calculated by weighting the error with the window span using Equation 2.25. The additional windows near the endpoint were added to mitigate larger individual window error, Δ_j, associated with large (~15-22 atom) glycan mutations. As in Chapter III, the average autocorrelation time for all simulations is 0.025 ± 0.015 ns, and the total $\Delta\Delta G$ errors were less than 0.13 kcal/mol. The errors were propagated to the partition coefficients using Equation 2.27.

All-atom MD Simulations of Glycosylated CBMs

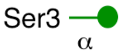
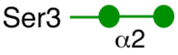
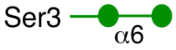
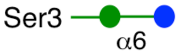
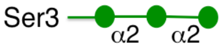
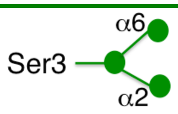
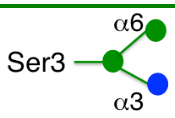
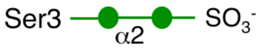
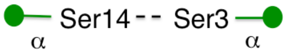
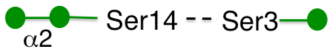
Long molecular dynamics simulations were performed to identify changes in the stability of the protein backbone, to identify changes in the hydrogen bonding potential of the CBM-glycan complex with the surface, and to calculate the carbohydrate-cellulose and protein-cellulose interaction energy. Three independent MD simulations of each system were performed in the NVT ensemble for 200 ns at 300 K with a 2 fs timestep using NAMD [58]. The two Ser-14 systems were run for 100 ns each. The system properties and simulation parameters were identical to those used in the TI studies. The non-bonded interaction energy between the residues of interest and the surface was calculated using CHARMM [55] and the error for these values over the total run was determined using block averaging [50]. Hydrogen bonding potentials between residues of interest and the surface were calculated in VMD [59] for comparison to the interaction energy values. RMSF and RMSD of the backbone were also calculated using CHARMM [55].

4.3 Results

Relative Binding Free Energy from TI Simulations

The cumulative $\Delta\Delta G$ results for all of the glycan patterns shown in Figure 4.1 are presented in Table 4.1 and Figure 4.3A. The changes reported are relative to the non-glycosylated wild type. A single *O*-mannose is present on Thr-1 in each simulation, but does not interact with the surface and thus is not included in the free energy calculation. The partition coefficients, $K_{\mathrm{Mut}}/K_{\mathrm{WT\text{-}NG}}$, were estimated using Equation 4.1, and the initial-slope Langmuir isotherms generated from these partition coefficients are presented in Figure 4.3B. The experimental values for the non-glycosylated wild type shown in Figure 4.3B are taken from Linder *et al*. [42]. The labels S3 and S14 refer to glycan attachment at Ser-3 and Ser-14, respectively. The labels M, G, and SO_3 denote mannose, glucose, and sulfate glycan moieties, respectively. The individual S3M1, S3M2, and S14M1 simulations are taken Chapter III for completeness. For simplicity in the remaining text when discussing relative free energy, bonding, and energetics, we describe each system using the end-state name rather than the cumulative name shown in Table 4.1 and Figure 4.2, e.g., S3M2 is used in place of S3M1+S3M2.

Table 4.1: Cumulative relative binding free energies as a result of changing glycan patterns compared to a non-glycosylated wild type CBM. The entire glycoform is the mutation from the non-glycosylated wild type, with green and blue representing mannose and glucose residues respectively.

Cumulative Name	End-state Glycoform	$\Delta\Delta G$, Cumulative (kcal/mol)	$K_{Mut}/K_{WT\text{-}NG}$
S3M1		-0.75 ± 0.03	3.5 ± 0.2
S3M1+S3M2		-1.2 ± 0.05	7.0 ± 0.6
S3M1+S3M2-16		-1.4 ± 0.10	11 ± 1.8
S3M1+S3MG		-3.4 ± 0.09	267 ± 40
S3M1+S3M2+S3M3		-3.8 ± 0.09	576 ± 91
S3M1+S3M2+S3M3B		-1.1 ± 0.10	6.5 ± 1.1
S3M1+S3M2-16+S3MGB		-2.5 ± 0.13	67 ± 14
S3M1+S3M2+S3M2-SO$_3$		-3.3 ± 0.08	238 ± 33
S3M1+S14M1		-3.0 ± 0.10	147 ± 25
S3M1+S14M1+S14M2		-3.1 ± 0.13	170 ± 37

The results of the individual simulations based on our alchemical paths chosen in Figure 4.2 are found in Table 4.2. For the addition of a single glycan to form the structure described by Harrison *et al.* [28] (S3M1), the electrostatic and VDW contributions contribute equally to the total relative binding affinity change. In cases with mannan disaccharide attachment (S3M2, S3M2-16), the unfavorable electrostatics are balanced out by the favorable VDW energy changes. For the mixed mannan-glucan structures (S3MG, S3MGB), the electrostatic contribution is relatively neutral compared to VDW contributions. This supports our observations that the glucose structures stack more efficiently over the cellulose surface, increasing relative binding affinity over mannan-only structures. The linear mannan trisaccharide (S3M3) also favors VDW contributions over electrostatic contributions, corresponding to slightly higher van der Waals interaction energy between the glycans and surface over electrostatic interaction energy (data not shown), and parallel alignment of the final mannose residue over the cellulose. Electrostatics and VDW offset each other in the branched mannan (S3M3B) and anterior mannan disaccharide (S14M2) cases, which may be attributed loss of stability in the actual CBM structures for these two systems. The sulfate (S3M2-SO_3) adds a charge to the system, resulting in a highly favorable electrostatic contribution to binding affinity.

Table 4.2: Individual TI simulation results corresponding to structures in Figure 4.2. The electrostatic and VDW values are calculated using Equation S1, i.e. $\Delta\Delta G_{\mathrm{Elec}} = \Delta G_{\mathrm{Bound}} - \Delta G_{\mathrm{Free}}$. Partition coefficient ratios were not calculated for individual simulations.

		Energy, ΔΔG (kcal/mol)	Error (kcal/mol)			Energy, ΔΔG (kcal/mol)	Error (kcal/mol)
S3M1	Electrostatics	-0.39	0.01	**S3M2-SO3**	Electrostatics	-1.33	0.04
	VDW	-0.36	0.02		VDW	-0.80	0.03
	ΔΔG (kcal/mol)	**-0.75**	**0.03**		**ΔΔG (kcal/mol)**	**-2.13**	**0.07**
S3M2	Electrostatics	2.21	0.01	**S3M3B**	Electrostatics	0.44	0.04
	VDW	-2.63	0.02		VDW	-0.39	0.05
	ΔΔG (kcal/mol)	**-0.42**	**0.03**		**ΔΔG (kcal/mol)**	**0.04**	**0.09**
S3M2-16	Electrostatics	1.07	0.04	**S3MGB**	Electrostatics	0.38	0.04
	VDW	-1.75	0.06		VDW	-1.49	0.04
	ΔΔG (kcal/mol)	**-0.68**	**0.10**		**ΔΔG (kcal/mol)**	**-1.11**	**0.08**
S3MG	Electrostatics	-0.14	0.04	**S14M1**	Electrostatics	-0.53	0.04
	VDW	-2.49	0.04		VDW	-1.74	0.05
	ΔΔG (kcal/mol)	**-2.63**	**0.08**		**ΔΔG (kcal/mol)**	**-2.27**	**0.09**
S3M3	Electrostatics	0.29	0.04	**S14M2**	Electrostatics	2.41	0.04
	VDW	-2.96	0.05		VDW	-2.50	0.05
	ΔΔG (kcal/mol)	**-2.67**	**0.08**		**ΔΔG (kcal/mol)**	**-0.09**	**0.08**

In all cases, we find that the addition of glycosylation improves binding affinity over a non-glycosylated wild type. Specifically, we observe two clusters of glycan structures, one of which imparts a change of -0.75 to -1.4 kcal/mol and the other which imparts a change in the range of -2.5 to -3.8 kcal/mol. The latter group includes glycan motifs that contain a sulfate or glucose (instead of mannose) or separation of the mannans to the posterior and anterior of the CBM. The former group contains single and disaccharide mannan motifs and a branched mannan trisaccharide motif. A linear mannan trisaccharide (S3M3) is found to improve the binding affinity over the non-glycoslyated wild type by 3.8 ± 0.09 kcal/mol, which yields a dramatic 500 plus-fold increase in the partition coefficient ratio. While the largest individual relative free energy gains are found for the final cycles of S3M3, S3MG, and S3M2-SO_3 (Figure 4.3B, Table 4.2), in general continuing to add sugars results in diminishing utility, with the maximum cumulative improvements in $\Delta\Delta G$ ranging from 3 - 4 kcal/mol, as shown in Figure 4.3A. The favorable total $\Delta\Delta G$ achieved with the addition of a sulfate to the mannan disaccharide (S3M2-SO_3) is driven by a change in the electrostatic contribution to $\Delta\Delta G$, caused by the negative charge on the sulfate; this differs from all of the other simulations where $\Delta\Delta G_{Elec}$ is nearly zero (Table 2). In the context of the simulations and the alchemical paths chosen [53], the fact that $\Delta\Delta G_{VDW}$ is the primary contributor the total $\Delta\Delta G$ for all but the sulfate case supports the notion that glycans enhance binding on the model cellulose surface primarily through enhanced van der Waals interactions [1,18].

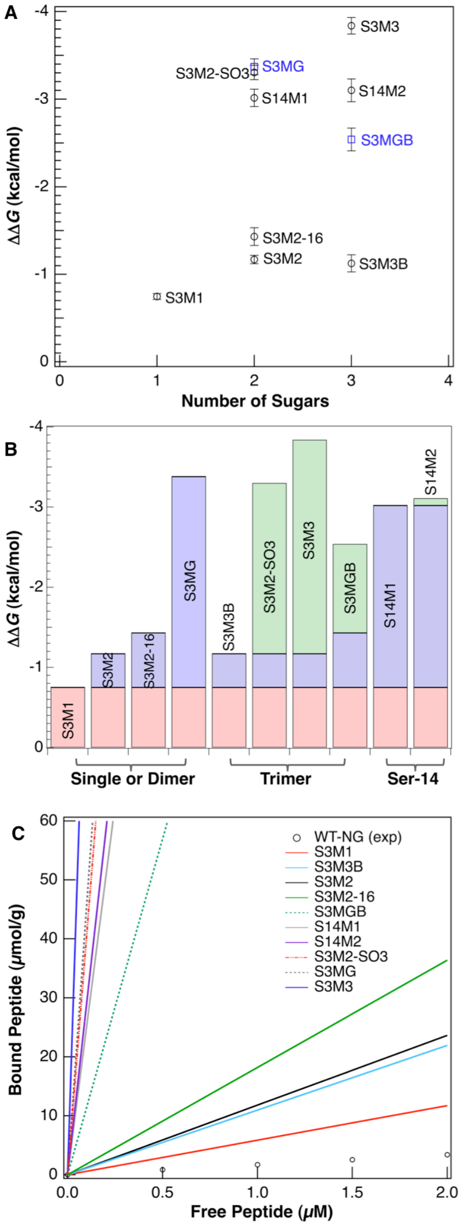


Figure 4.3: **T**I simulation results showing improved binding affinity for glycosylated CBMs over the non-glycosylated wild type CBM. (A) Cumulative relative binding free energy change over the non-glycosylated wild type as a function of the total number of sugars added. Mixed mannan-glucan systems are denoted with blue squares. (B) Relative binding results for individual simulations, with end-state labeled. Total errors are not shown in B, but are the same as (A). S3M1 forms the base simulation for all subsequent simulations. For S3M3B, a third mannose residue is present, but the individual $\Delta\Delta G$ value for the final simulation is zero, thus the third value in green is not visible. (C) Langmuir isotherms of bound CBM concentration as a function of free CBM concentration predicted by the TI simulations conducted in this work. The experimental values for the non-glycosylated wild type are taken from Linder et al. [42]

The cumulative $\Delta\Delta G$ values are almost within error for a mannan disaccharide with an α-1,2 linkage (S3M2) and a mannan disaccharide with an α-1,6 linkage (S3M2-16). In contrast, the calculated $\Delta\Delta G$ for a mannan-glucan disaccharide in an α-1,6 linkage (S3MG) is 2.5 times that of the analogous mannan motif (S3M2-16). Notably, both mixed mannan-glucan systems (S3MG and S3MGB) demonstrate improved binding over their mannan-only counterparts (S3M2-16 and S3M3B), resulting from protein stability and glycan interaction differences (discussed below). We also note that not all additions are favorable; the branched mannan trisaccharide (S3M3B) is less favorable than the mannan disaccharide, and adding a mannan disaccharide in the anterior position (S14M2) results in a marginal increase in relative affinity over the S14M1 case. Variations in relative binding affinity in each system can be attributed to structural and thermodynamic changes, which we discuss below.

CBM Glycoprotein-Cellulose Interaction Energy

The total interaction energy, composed of electrostatic and van der Waals contributions, between the cellulose, protein, and glycans as well as the protein backbone fluctuations (RMSD and RMSF) and hydrogen bonding (H-bonding) potential between the CBM and cellulose and the glycans and cellulose, was calculated from the 200 ns MD runs. While the Thr-1 *O*-mannose interacts with the surface in rare cases, its contributions are typically < 1-2 kcal/mol and thus are not included in these data. From this analysis, we find that neither interaction energy nor H-bonding alone, but a combination of the two, controls the binding affinity differences among the systems. Our results also suggest that maintaining or improving the stability of the protein backbone is key to achieving improvements in binding affinity. In all cases, glycans increase the total interaction energy, shown in Figure 4.4, and H-bonding potential (Table 4.3, Figures 4.7-4.9) of the glycoprotein with cellulose; the data logically correlate with the ΔΔ*G*, with systems with high interaction or improved H-bonding exhibiting favorable changes in binding affinity.

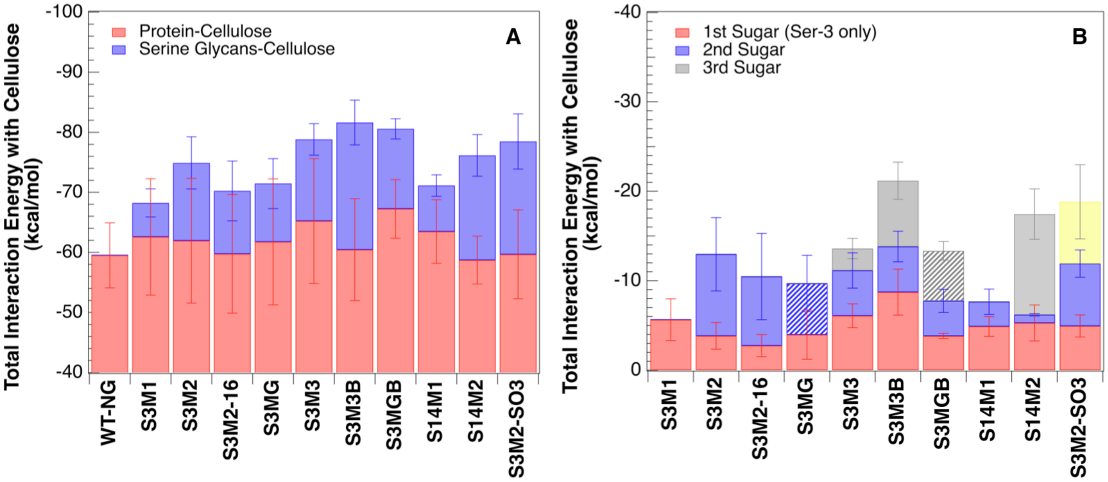


Figure 4.4: Interactions of the CBM-glycan complex with cellulose calculated over 200 ns MD simulations. Total interaction energy of (A) the CBM and glycan structures with the cellulose substrate and (B) individual sugar components with the cellulose substrate. Mannose, glucose, and sulfate moieties are solid, striped, and yellow, respectively.

While small variations in glycosylation motifs can have an impact on relative affinity or protein structure, the CBM-cellulose interactions are found to be approximately equal (within error) across all systems (Figure 4.4A). In each variant, the total electrostatic and van der Waals components of the glycan-cellulose interaction energy (total shown in Figure 4.4B) are within 1 kcal of each other with the exception of S3MGB, where the van der Waals interaction is approximately 3 kcal more than the electrostatic interaction (data not shown).

CBM Protein Backbone Root-mean Square Deviation (RMSD) and Root-mean Square Fluctuation (RMSF)

RMSD and RMSF of the protein backbone highlight systems in which the CBM stability is compromised, most notably the S3M3B motif. The RMSD curves for each system are shown in Figure 4.5. The curves across each 200 ns run for a specific system are fairly consistent. The RMSD curves are generally lower and more stable for mannan-glucan systems (bottom row, Figure 4.5) are than their mannan counterparts (S3M2, S3M2-16, and S3M3B). S3M3B is the only system to exhibit a RMSD value greater than 3 Å.

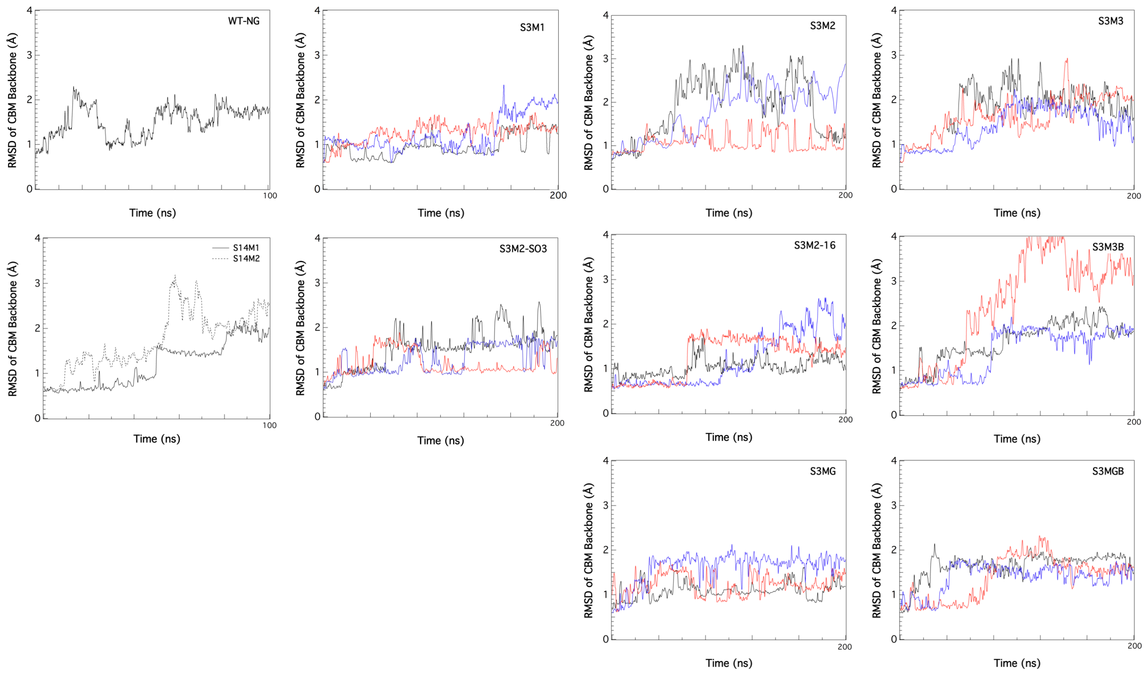


Figure 4.5: RMSD curves for the protein backbone. Each system was run three times for 200 ns each. These three runs are shown in black, red, and blue above. The WT-NG and S14M1 (included from Chapter III), and S14M2 systems were run for 100 ns (far left column). The mannan-only systems are in the top two rows, and the mannan-glucan systems are in the bottom row.

The RMSF of the protein backbone are shown in Figure 4.6. The value at each residue is the average over the three 200 ns runs, and the errors equal the standard deviation of these values. S3M3B and S14M2 show the highest fluctuations of all the systems in the N-terminus to residue 8 region, highlighting potential downsides to utilizing these structures to improve binding affinity. Fluctuations near residue 20 are indicative of glycan interaction with residues Ala-20 and Ser-21 in the top loop of the protein; the trisaccharide glycans are especially prone to these types of "inversions".

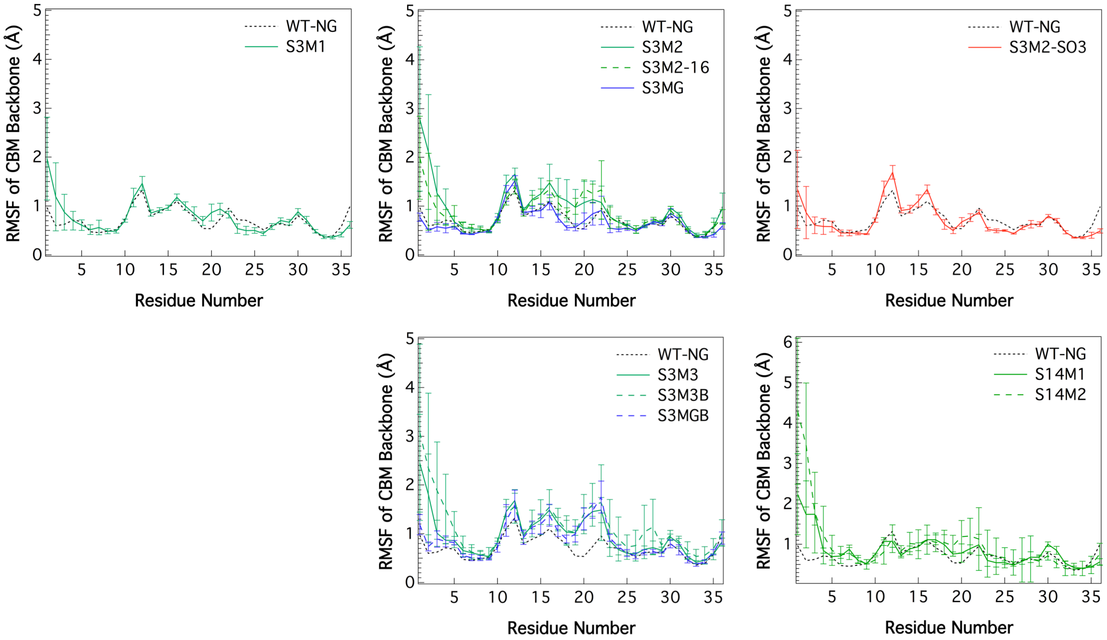


Figure 4.6: RMSF of the CBM backbone by residue number. The results from the 100 ns WT-NG simulation performed in Chapter III are included in each graph for comparison to the glycosylated structures.

CBM Glycoprotein-Cellulose Hydrogen Bond Analysis

The H-bond networks described in this study involve cellulose and the glycosylated CBM while neglecting water contributions; quantitatively separating the H-bonding contributions and the other enthalpic or entropic contributions to binding affinity is not feasible, though qualitative conclusions are possible [46]. The natural fluctuations of the glycans, indicated by RMSF fluctuations in the N-terminus and near residue 20 (Figure 4.6), also cause fluctuations in H-bonding potential data. However, we can use VMD [59] to calculate the number of potential bond formed and make general comparisons between the systems studied. We selected exemplary systems from each data set that exhibited stable RMSD curves and where glycan inversion and interaction with the top loop was infrequent to avoid differences associated with glycan inversions only. To ensure the protein was still forming bonds with the cellulose via the aromatic and polar "flat-face" residues we also calculated potential bonds between the CBM and cellulose. Since each system was run three times for 200 ns, we compared the counts for each system to check for drastic variations and found deviations in glycan-cellulose and protein-cellulose H-bond counts across the three runs were less than 10%. The duration of one, two, or over three potential H-bonds between the complete glycan structure and cellulose is calculated by the number of bonds possible (Figure 4.7) divided by the total number of observations (*i.e.* frames in the trajectory, which were output every 4 ps during the 200 ns simulations), and the results are shown in Table 4.3. The total number of bonds possible and thus duration increases as glycans are added. The mannan disaccharide with sulfate (S3M2-SO_3) and the mannan-glucan disaccharide (S3MG) form one to two bonds consistently, resulting in the highest total durations correlating with the fact that these two systems have two of the highest relative affinities.

Table 4.3: H-bond duration between the glycans and cellulose surface. H-bonds are calculated using a distance cutoff of 3.0 Å and an angle criteria of 60° from linear. The duration (% of run) is calculated by counting the number of bonds and dividing by the total number of observations. Refer to Figures 4.7 and 4.8 for graphical representations of the number of bonds.

		One H-bond (% of run)	Two H-bonds (% of run)	> 3 H-bonds (% of run)	Total Glycan-Cellulose (% of run)
Mannan	**S3M1**	13%	4%	1%	18%
	S3M2	38%	22%	7%	68%
	S3M2-16	28%	19%	9%	57%
	S3M2-SO$_3$	62%	21%	1%	83%
	S3M3	30%	14%	7%	51%
	S3M3B	40%	25%	7%	72%
Mannan-glucan	**S3MG**	52%	24%	9%	85%
	S3MGB	50%	6%	1%	56%
Anterior mannan	**S14M1**	20%	7%	1%	28%
	S14M2	47%	20%	5%	72%

To visualize the H-bond data, representations of the glycan hydrogen bonding data are shown in Figures 4.7 and 4.8, where the number of bonds shown divided by the total number of observations corresponds to the percentages shown in Table 4.3 from four representative systems, S3M1 and S3M2 (Figure 4.7) and S3M2, S3M2-16, and S3MG (Figure 4.8).

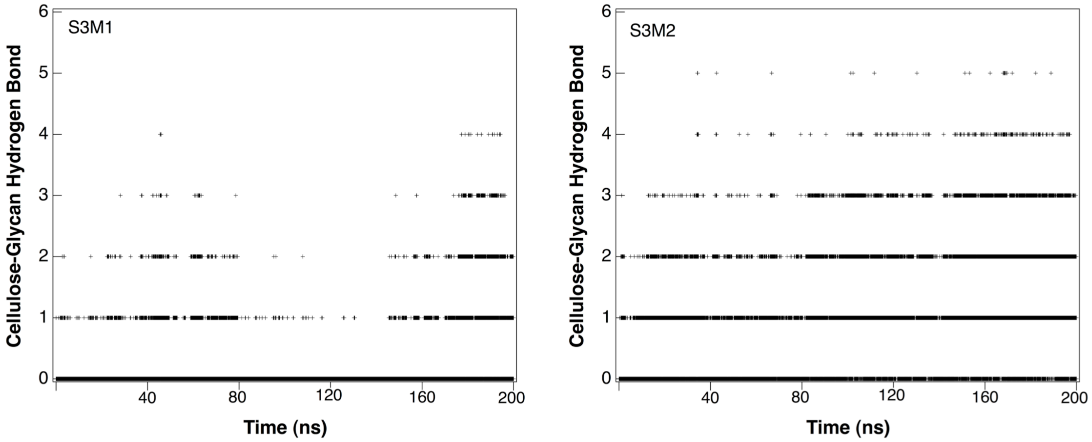


Figure 4.7: H-bond count for S3M1 and S3M2 systems. Total increase in number of bonds for S3M2 corresponds to a 60% increase in duration of H-bonds during the run (Table 4.3). H-bonds are calculated using a distance cutoff of 3.0 Å and an angle criteria of 60° from linear.

Figure 4.8 suggests that glycan chemistry has a greater impact than linkage type; while the S3M2 (α-1,2 linkage) and S3M2-16 (α-1,6) linkages are nearly identical, the S3MG (α-1,6) adds additional stable bonds (shown in blue, Figure 4.8), corresponding to the increase in duration for S3MG shown in Table 4.3.

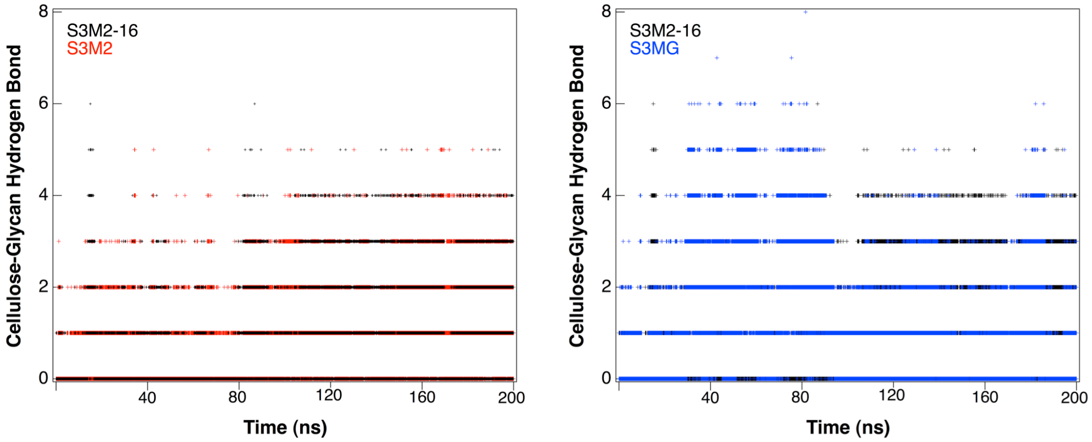


Figure 4.8: H-bond count between the cellulose and Ser-3 glycan comparing an a-1,2 mannan disaccharide (S3M2), an a-1,6 mannan disaccharide (S3M2-16), and an a-1,6 mannan-glucan disaccharide (S3MG). H-bonds are calculated using a distance cutoff of 3.0 Å and an angle criteria of 60° from linear.

Generally the protein-cellulose bonding potential increased for all glycan variants over the non-glycosylated wild type, although as mentioned in some cases the key flat-face residue bonds were disrupted (data not shown). This general increase may be attributed to increased stability over a portion of the surface, allowing for more interaction between the CBM and cellulose. There was little variation in the duration of bonds formed between the glycan systems, except for the branched mannan-glucan motif (S3MGB), where the duration of CBM bonds doubled over the duration of the non-glycosylated wild type and increased approximately 1.3 times over the average duration of the other glycosylated systems; a comparison of H-bonds formed in the non-glycosylated wild type and S3MGB is found in Figure 4.9. This improvement may help push the total ΔΔ*G*(S3MGB) to the favorable 2.5±0.13 kcal/mol increase over a non-glycosylated wild type.

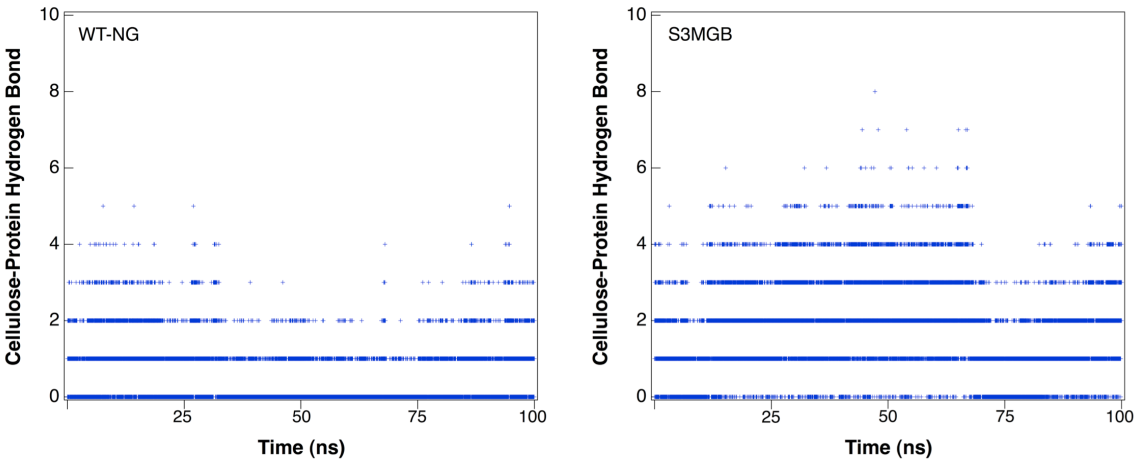


Figure 4.9: H-bond count over 100 ns between the cellulose and a non-glycosylated wild type CBM or a CBM with a mannan-glucan branch trisaccharide (S3MGB). The wild type data was calculated in Chapter III. H-bonds are calculated using a distance cutoff of 3.0 Å and an angle criteria of 60° from linear.

4.4 Discussion

Binding Affinity Differences in Dimer Glycoforms

The glycan-cellulose interaction energy can be separated into contributions from each sugar molecule, as shown in Figure 4.4B. For the mannan disaccharides (S3M2 and S3M2-16), both the interaction energy and H-bonding slightly favor the S3M2 configuration, but the protein stability is somewhat improved in S3M2-16 (RMSD Figure 4.5), resulting in nearly equal $\Delta\Delta G$ values. Whereas the mannan-glucan disaccharide (S3MG) has slightly lower glycan interaction than the analogous mannan disaccharide (S3M2-16), the ability of the first *O*-linked mannose to form a consistent H-bond with the surface is maintained, increasing the overall glycan-cellulose bonding potential over S3M2-16 (Table 4.3 and Figure 4.6). Additionally in S3MG, the mannose and glucose residues form H-bonds on both sides of their ring structures, indicating a more stacked conformation of the glycans and cellulose, which could impact binding affinity (i.e., $\Delta\Delta G$(S3MG) > $\Delta\Delta G$(S3M2-16) and $\Delta\Delta G$(S3M2)). The data generated from comparing these three disaccharide systems indicate that the nature of the carbohydrate moieties may have more of an impact than the covalent linkage in binding affinity or, more generally, cell recognition and adhesion. This conclusion can be tested experimentally by using chemical tagging methods to specifically alter linkage patterns [24] and then measuring changes in binding affinity or enzymatic activity.

Binding Affinity Differences in Branched and Linear Trimer Glycoforms

Our results show more favorable binding exhibited in linear trisaccharides over branched trisaccharides. In the trisaccharides motifs, the glycans fluctuate in the posterior region of the CBM, but the degree of fluctuation can impact the stability of the protein backbone and the glycan-cellulose interaction. The stability of the protein backbone, especially the N-terminus through residues 8-10, is key to maintaining contact between the CBM and cellulose and increasing relative binding affinity. Binding is negatively impacted if the structure of the N-terminus is significantly disrupted, as shown in experiment [42] and Chapter III. The branched mannan trisaccharide (S3M3B) has the highest total glycan interaction energy at approximately 21 kcal/mol, but there is a corresponding slight decrease in the CBM's total interaction compared to the other glycosylated systems (Figure 4.4A). This inverse relationship is indicative of a change in the CBM structure caused by the glycan, reflected also by the high RMSD of this system compared to others (Figure 4.5). A prior experimental study determined when larger *N*-glycan structures inhibit protein-cellulose contact, binding affinity was negatively impacted [44]; a similar situation may be present in this study with the S3M3B structure, where hydrogen bonding between the key CBM flat-face residues and cellulose is also reduced (data not shown), further explaining the lowered binding affinity relative to the other systems except S3M1.

The mannan-glucan branched motif, S3MGB, doubles the CBM affinity over the mannan branched trisaccharide motif, S3M3B. The S3MGB glycans, primarily driven by moiety changes, align parallel to the surface and exhibit lower fluctuations translating to stability in the protein backbone and leading to a more stable interaction energy (lower error in Figure 4.4) and more consistent glycan-cellulose and CBM-cellulose H-bonding (Table 4.3, Figure 4.9). Also as previously mentioned, the total van der Waals interaction energy between S3MGB and modeled I-β cellulose face is larger than the electrostatic interaction energy by 3 kcal/mol (data not shown), which supports the stacked conformation of the glycans observed when visualizing the trajectory. A visual comparison of the N-terminus fluctuations in S3M3B and S3MGB is provided in Figure 4.10, accompanying the RMSF comparison to a non-glycosylated wild type (WT-NG). As can be seen from the figure, the branched structures do fluctuate and interact with the top protein loop (Ala-20 and Ser-21), resulting in higher RMSF values from 15 to 22 for these two systems over the WT-NG system. Concurrent with the RMSD curves (Figure 4.5), the N-terminus and glycan area contacting the cellulose (the cellulose surface in the x-y plane is not shown) is less compact for S3M3B than S3MGB; the results of this CBM structure change are reflected in the lower $\Delta\Delta G$ for S3M3B. CBM stability in S3MGB may also lead to increased H-bonding of the protein with the cellulose, as shown in Figure 4.9.

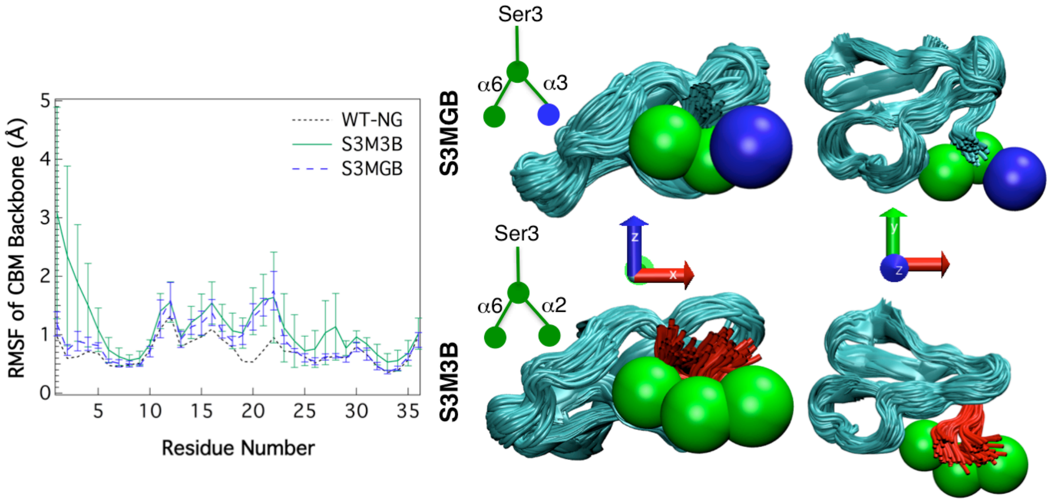


Figure 4.10: RMSF of S3M3B and S3MGB compared to a non-glycosylated wild type (WT-NG) and side and top-view cluster representations of the CBM backbone and Ser-3 glycans, S3MGB and S3M3B, over 200 ns. Typical mannose residue positions are shown in green and the typical glucose residue position in blue using the bead representation with bonds to Ser-3 not shown. The interaction of the mannan branched trisaccharide results in loss of in the compactness of the N-terminus for S3M3B (shown in red). Additionally, the glycans in S3MGB exhibit a more parallel alignment over the surface (x-y plane) than S3M3B.

In contrast to the branched structures, the linear trisaccharide motif (S3M3) maintains hydrogen bonding between the CBM and cellulose (data not shown), the total CBM-glycan interaction energy with the surface increases (Figure 4.4A), and there is little loss in CBM stability over the wild type (Figure 4.5 and 4.6). In the linear S3M2-SO_3 system, the individual component interaction is comparable to S3M3B (Figure 4.4B), but the data suggests the smaller size of the SO_3^- does not impact the stability of the protein backbone (Figures 4.5 and 4.6), resulting in a higher binding affinity for the sulfated system.

Impact of Glycosylation Location on Protein Structure and Affinity

Whereas both of the Ser-14 systems studied have favorable binding affinity, the disaccharide motif (S14M2) is found not to add any additional benefit over the single mannan motif (S14M1). These systems were compared using 100 ns MD simulations and the total interaction energy found to increase for S14M2 over S14M1 (Figure 4.4A), with the second mannose residue dominating the interaction at Ser-14 with the surface (Figure 4.4B). The presence of the disaccharide also begins to impact the CBM structure, primarily in the sensitive N-terminal region. The high interaction of the Ser-14 disaccharide with the surface (Figure 4.4A) induces a pulling on the anterior of the protein backbone, which in turn induces a constriction of residues 14-22 (Figure 4.6); the normal N-terminus interaction with the top loop at Ala-20 is disrupted and the entire N-terminus behavior, including the Ser-3 *O*-mannose, becomes more erratic. The RMSD (Figure 4.5), RMSF, and cluster visualizations (Figure 4.11) confirm the structural changes and support a postulation that while binding affinity is high, the Ser-14 disaccharide configuration may not be ideal for this Family 1 CBM.

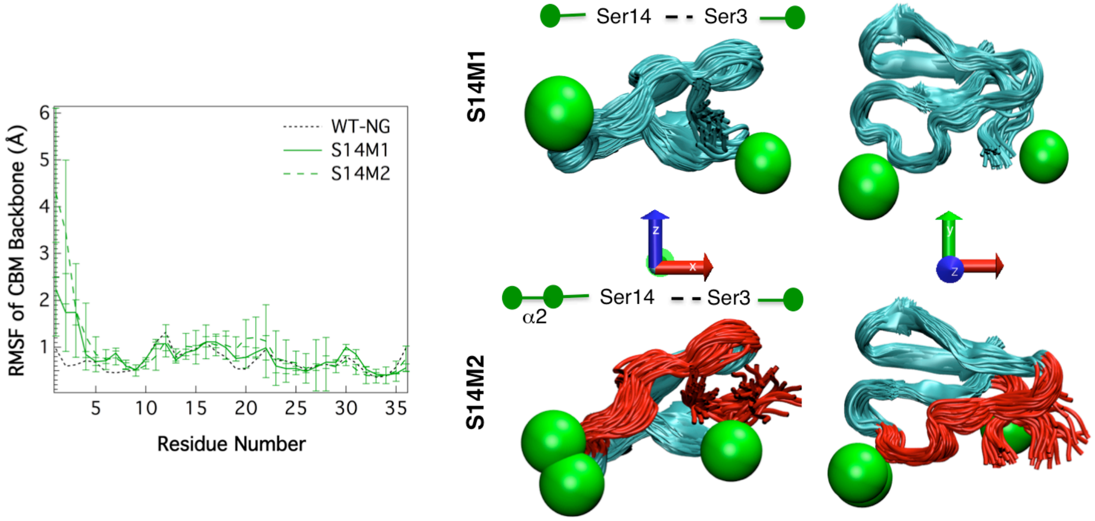


Figure 4.11: RMSF of S3M3B and S3MGB compared to a non-glycosylated wild type (WT-NG) and side and top-view cluster representations of the CBM backbone and Ser-3 and Ser-14 glycans, S14M1 and S14M2 over 100 ns. Typical mannose residue positions are shown in green using the bead representation with the bonds to Ser-3 and Ser-14 not shown. Constriction of the protein backbone from Ser-14 to Gly-22 causes the CBM N-terminus' fluctuation to increase (S14M2 in red) and the anterior of the CBM to be raised higher in the z plane than that of S14M1 or the WT-NG (cluster not shown).

Glycosylation As a Natural Means to Improve Binding Affinity

To determine whether or not fungi commonly employ glycosylation on CBMs as a natural means to improve binding affinity, we performed a multiple sequence alignment of Family 1 CBMs [70,71]. The LOGO [72] representation shown in Figure 4.12 shows that Thr and Ser are conserved at the native glycosylation positions 1 and 3 and also at the anterior of the CBM at positions 14 and 15. Serine is also conserved in residues 23, 25, 40, and 41, but these residues are not in direct contact with the surface, making improvements to binding via glycosylation unlikely [43]. Since over-glycosylation in the catalytic domain could be detrimental [32,43-45], and structural changes imparted by glycosylation should be carefully monitored, organism and protein-level modifications will likely be required to achieve maximum benefit to the enzyme as a whole.

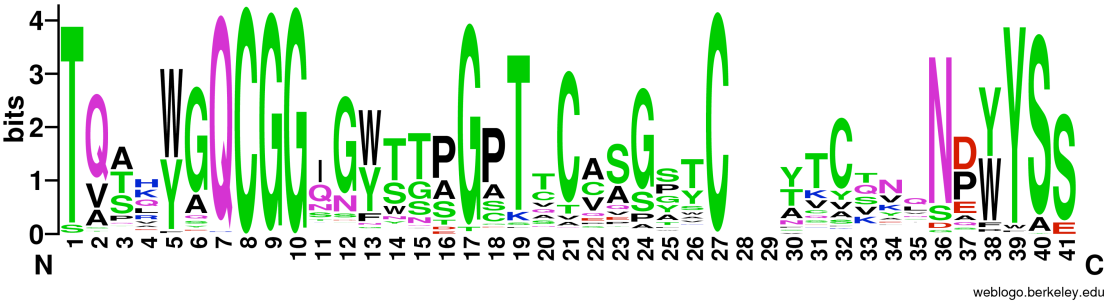


Figure 4.12: Sequence homology in Family 1 CBMs. The LOGO [72] displays conservation of O-glycosylation sites at residues 1, 3, and also 14 and 15. The size of the letter represents prevalence of the residue at the sequence position.

4.5 Conclusions

In this work we sought to understand the impact of various *O*-glycosylation motifs from different organisms on the structure and function of carbohydrate-binding proteins. We chose a commonly studied Family 1 CBM as a model system, which allowed us to investigate options for improving binding affinity, and thus potential activity, of fungal cellulases [40]. Using molecular simulation and TI calculations [50-53], we show that simple *O*-mannans or *O*-mannan-glucans can increase binding affinity up to 3.8 kcal/mol over a non-glycosylated wild type CBM. This is achieved through an increase in the interaction energy and H-bonding potential of the CBM-glycan complex with cellulose. The linear mannan trisaccharide produced in *T. reesei*, *A. niger*, *A. awamori,* and yeasts (Figure 4.1, Table 4.1) is found to yield the largest improvement over the non-glycosylated wild type, corresponding to a 20 kcal/mol increase in total interaction energy with the cellulose over the wild type and a stabilization of the protein backbone, increasing both the protein and glycan H-bonding potential with cellulose. In terms of glycan chemistry, the mannan-glucan structures examined here exhibit higher affinity than their mannan counterparts (e.g., S3MG as compared to S3M2 and S3M2-16), which we attribute to the observation that a mannan-glucan glycan orients more parallel to the surface than a mannan glycan, increasing H-bond potential and interaction energy. Additionally, the CBM backbone is more stable with systems containing a glucose residue, increasing the favorable enthalpic interactions of the CBM with cellulose. Lastly, our comparison of three disaccharide systems (S3M2, S3M2-16, and S3MG) with mannan and glucan α-1,2 and α-1,6 linkages indicates that linkage patterns in simple *O*-glycans do not appear to impact binding affinity and protein structure as much as the nature of the glycan moiety. However, the size of the protein itself must also be considered; the Cel7A CBM is relatively small (~3700 Da), so even branched mannan trisaccharides (~480 Da) begin to change the protein structure and impair further improvements to binding affinity over the mannan single or disaccharide cases. Thus an important finding of this study is that care must be taken to ensure glycosylation does not change the inherent structure of the protein, such as in the S3M3B or S14M2 cases, as this can negatively impact normal binding and function.

Given mounting evidence of the functional role of glycosylation in biochemical interactions and its prevalence as a post-translational modification, this study highlights the importance of conscientious design of experimental and computational studies that may be affected by changes imparted by glycosylation. Host organism and growth media selection can have a dramatic effect on glycosylation patterns [1,9,10] and failure to account for changes imparted by glycans could alter the outcomes of protein engineering efforts. Glycosylation is important in both intra- and inter-cellular recognition [1,9,19], and this work shows that *O*-glycosylation can impact binding affinity and protein structure in the model Family 1 CBM from Cel7A in *T. reesei*. While experimental conformation of this study is needed, our TI calculations suggest that broadening the *O*-glycan patterns available in *T. reesei* via expression of protein-*O*-mannosyltransferases and mannosyltranferases found in *Aspergillus* or yeast or via chemical synthesis could be used to tune binding affinity and overall activity of Cel7A and other cellulase enzymes. For cellulase production, glycans that change the inherent structure of the protein backbone should be avoided, however, and over-expression of glycans in other domains should also be considered when using glycosylation in protein engineering. Regardless of application, maintaining or improving structural integrity of proteins is key to optimizing enhancements via glycoprotein engineering. Finally, we demonstrate that computational studies are a valuable tool in rational approaches to glycoprotein engineering, allowing researchers to screen various mutations in binding studies, for example, and providing molecular-level details and general visualization of glycoproteins [7,46].

4.6 References

1. A. Varki, *Essentials of Glycobiology*. Cold Springs Harbor Laboratory Press: Cold Springs Harbor, NY, USA, **2009**.

2. G.W. Hart and R.J. Copeland, "Glycomics Hits the Big Time", *Cell* **2010**, 143 (5), 672-676.

3. N. Deshpande, M.R. Wilkins, N. Packer and H. Nevalainen, "Protein glycosylation pathways in filamentous fungi", *Glycobiology* **2008**, 18 (8), 626-637.

4. K. Ohtsubo and J.D. Marth, "Glycosylation in cellular mechanisms of health and disease", *Cell* **2006**, 126 (5), 855-867.

5. M. Goto, "Protein O-glycosylation in fungi: Diverse structures and multiple functions", *Biosci. Biotechnol. Biochem.* **2007**, 71 (6), 1415-1427.

6. V. Receveur, M. Czjzek, M. Schulein, P. Panine and B. Henrissat, "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), 40887-40892.

7. Y. Mazola, G. Chinea and A. Musacchio, "Integrating Bioinformatics Tools to Handle Glycosylation", *PLoS Comput. Biol.* **2011**, 7 (12), 7.

8. A.M. Sinclair and S. Elliott, "Glycoengineering: The effect of glycosylation on the properties of therapeutic proteins", *Journal of Pharmaceutical Sciences* **2005**, 94 (8), 1626-1635.

9. K. Drickamer and M.E. Taylor, "Evolving views of protein glycosylation", *Trends Biochem.Sci.* **1998**, 23 (9), 321-324.

10. G. Lauc and V. Zoldos, "Protein glycosylation-an evolutionary crossroad between genes and environment", *Mol. Biosyst.* **2010**, 6 (12), 2373-2379.

11. A. Varki, "BIOLOGICAL ROLES OF OLIGOSACCHARIDES - ALL OF THE THEORIES ARE CORRECT", *Glycobiology* **1993**, 3 (2), 97-130.

12. R.J. Sola and K. Griebenow, "Glycosylation of Therapeutic Proteins An Effective Strategy to Optimize Efficacy", *Biodrugs* **2010**, 24 (1), 9-21.

13. R.A. Dwek, "Glycobiology: More Functions for Oligosaccharides", *Science* **1995**, 269 (5228), 1234-1235.

14. P.M. Rudd, T. Elliott, P. Cresswell, I.A. Wilson and R.A. Dwek, "Glycosylation and the Immune System", *Carbohydrates and Glycobiology* **2001**, 291 (5512), 2370-2376.

15. S.I. Van Kasteren, H.B. Kramer, D.P. Gamblin and B.G. Davis, "Site-selective glycosylation of proteins: creating synthetic glycoproteins", *Nature Protocols* **2007**, 2 (12), 3185-3194.

16. L. Wang, U.K. Aryal, Z.Y. Dai, A.C. Mason, M.E. Monroe, Z.X. Tian, J.Y. Zhou, D. Su, K.K. Weitz, T. Liu, D.G. Camp, R.D. Smith, S.E. Baker and W.J. Qian, "Mapping N-Linked Glycosylation Sites in the Secretome and Whole Cells of Aspergillus niger Using Hydrazide Chemistry and Mass Spectrometry", *J. Proteome Res.* **2012**, 11 (1), 143-156.

17. M.U. Jars, S. Osborn, J. Forstrom and V.L. Mackay, "N-GLYCOSYLATION AND O-GLYCOSYLATION AND PHOSPHORYLATION OF THE BAR SECRETION LEADER DERIVED FROM THE BARRIER PROTEASE OF SACCHAROMYCES-CEREVISIAE", *J. Biol. Chem.* **1995**, 270 (42), 24810-24817.

18. G.T. Beckham, Z. Dai, J.F. Matthews, M. Momany, C.M. Payne, W.S. Adney, S.E. Baker and M.E. Himmel, "Harnessing glycosylation to improve cellulase activity", *Current Opinion in Biotechnology* **2012**, (0).

19. D. Spillmann and M.M. Burger, "Carbohydrate-carbohydrate interactions in adhesion", *J. Cell. Biochem.* **1996**, 61 (4), 562-568.

20. R. Pejchal, K.J. Doores, L.M. Walker, R. Khayat, P.S. Huang, S.K. Wang, R.L. Stanfield, J.P. Julien, A. Ramos, M. Crispin, R. Depetris, U. Katpally, A. Marozsan, A. Cupo, S. Maloveste, Y. Liu, R. McBride, Y. Ito, R.W. Sanders, C. Ogohara, J.C. Paulson, T. Feizi, C.N. Scanlan, C.H. Wong, J.P. Moore, W.C. Olson, A.B. Ward, P. Poignard, W.R. Schief, D.R. Burton and I.A. Wilson, "A Potent and Broad Neutralizing Antibody Recognizes and Penetrates the HIV Glycan Shield", *Science* **2011**, 334 (6059), 1097-1103.

21. E.K. Culyba, J.L. Price, S.R. Hanson, A. Dhar, C.H. Wong, M. Gruebele, E.T. Powers and J.W. Kelly, "Protein Native-State Stabilization by Placing Aromatic Side Chains in N-Glycosylated Reverse Turns", *Science* **2011**, 331 (6017), 571-575.

22. S.R. Hamilton, P. Bobrowicz, B. Bobrowicz, R.C. Davidson, H.J. Li, T. Mitchell, J.H. Nett, S. Rausch, T.A. Stadheim, H. Wischnewski, S. Wildt and T.U. Gerngross, "Production of complex human glycoproteins in yeast", *Science* **2003**, 301 (5637), 1244-1246.

23. J.A. Prescher and C.R. Bertozzi, "Chemical technologies for probing glycans", *Cell* **2006**, 126 (5), 851-854.

24. J.E. Hudak, H.H. Yu and C.R. Bertozzi, "Protein Glycoengineering Enabled by the Versatile Synthesis of Aminooxy Glycans and the Genetically Encoded Aldehyde Tag", *Journal of the American Chemical Society* **2011**, 133 (40), 16127-16135.

25. S.I. van Kasteren, H.B. Kramer, H.H. Jensen, S.J. Campbell, J. Kirkpatrick, N.J. Oldham, D.C. Anthony and B.G. Davis, "Expanding the diversity of chemical protein modification allows post-translational mimicry", *Nature* **2007**, 446 (7139), 1105-1109.

26. N.J. Agard and C.R. Bertozzi, "Chemical Approaches To Perturb, Profile, and Perceive Glycans", *Accounts Chem. Res.* **2009**, 42 (6), 788-797.

27. K. De Pourcq, K. De Schutter and N. Callewaert, "Engineering of glycosylation in yeast and other fungi: current state and perspectives", *Applied Microbiology and Biotechnology* **2010**, 87 (5), 1617-1631.

28. M.J. Harrison, A.S. Nouwens, D.R. Jardine, N.E. Zachara, A.A. Gooley, H. Nevalainen and N.H. Packer, "Modified glycosylation of cellobiohydrolase I from a high cellulase-producing mutant strain of Trichoderma reesei", *Eur. J. Biochem.* **1998**, 256 (1), 119-127.

29. J.P.M. Hui, P. Lanthier, T.C. White, S.G. McHugh, M. Yaguchi, R. Roy and P. Thibault, "Characterization of cellobiohydrolase I (Cel7A) glycoforms from extracts of Trichoderma reesei using capillary isoelectric focusing and electrospray mass spectrometry", *J. Chromatogr. B* **2001**, 752 (2), 349-368.

30. T.R. Gemmill and R.B. Trimble, "Overview of N- and O-linked oligosaccharide structures found in various yeast species", *Biochim. Biophys. Acta-Gen. Subj.* **1999**, 1426 (2), 227-237.

31. G.T. Beckham, J.F. Matthews, Y.J. Bomble, L.T. Bu, W.S. Adney, M.E. Himmel, M.R. Nimlos and M.F. Crowley, "Identification of Amino Acids Responsible for Processivity in a Family 1 Carbohydrate-Binding Module from a Fungal Cellulase", *J. Phys. Chem. B* **2009**, 114 (3), 1447-1453.

32. W.S. Adney, T. Jeoh, G.T. Beckham, Y.C. Chou, J.O. Baker, W. Michener, R. Brunecky and M.E. Himmel, "Probing the role of N-linked glycans in the stability and activity of fungal cellobiohydrolases by mutational analysis", *Cellulose* **2009**, 16 (4), 699-709.

33. S.P.S. Chundawat, G.T. Beckham, M.E. Himmel and B.E. Dale, "Deconstruction of Lignocellulosic Biomass to Fuels and Chemicals", *Annu. Rev. Chem. Biomol. Eng.* **2011**, 2, 6.1-6.25.

34. M.E. Himmel, S.Y. Ding, D.K. Johnson, W.S. Adney, M.R. Nimlos, J.W. Brady and T.D. Foust, "Biomass recalcitrance: Engineering plants and enzymes for biofuels production", *Science* **2007**, 315 (5813), 804-807.

35. A.J. Ragauskas, C.K. Williams, B.H. Davison, G. Britovsek, J. Cairney, C.A. Eckert, W.J. Frederick, J.P. Hallett, D.J. Leak, C.L. Liotta, J.R. Mielenz, R. Murphy, R. Templer and T. Tschaplinski, "The path forward for biofuels and biomaterials", *Science* **2006**, 311 (5760), 484-489.

36. C. Divne, J. Stahlberg, T. Reinikainen, L. Ruohonen, G. Pettersson, J.K.C. Knowles, T.T. Teeri and T.A. Jones, "The 3-dimensional crystal-structure of the catalytic core of Cellobiohydrolase-I from *Trichodermal reesei*", *Science* **1994**, 265 (5171), 524-528.

37. M. Srisodsuk, T. Reinikainen, M. Penttila and T.T. Teeri, "ROLE OF THE INTERDOMAIN LINKER PEPTIDE OF TRICHODERMA-REESEI CELLOBIOHYDROLASE-I IN ITS INTERACTION WITH CRYSTALLINE CELLULOSE", *J. Biol. Chem.* **1993**, 268 (28), 20756-20761.

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

39. A.W. Blake, L. McCartney, J.E. Flint, D.N. Bolam, A.B. Boraston, H.J. Gilbert and J.P. Knox, "Understanding the biological rationale for the diversity of cellulose-directed carbohydrate-binding modules in prokaryotic enzymes", *J. Biol. Chem.* **2006**, 281 (39), 29321-29329.

40. S. Takashima, M. Ohno, M. Hidaka, A. Nakamura and H. Masaki, "Correlation between cellulose binding and activity of cellulose-binding domain mutants of Humicola grisea cellobiohydrolase 1", *FEBS Lett.* **2007**, 581 (30), 5891-5896.

41. M. Linder, G. Lindeberg, T. Reinikainen, T.T. Teeri and G. Pettersson, "The difference in affinity between 2 fungal cellulose-binding domains is dominated by a single amino-acid substitution", *FEBS Lett.* **1995**, 372 (1), 96-98.

42. M. Linder, M.L. Mattinen, M. Kontteli, G. Lindeberg, J. Stahlberg, T. Drakenberg, T. Reinikainen, G. Pettersson and A. Annila, "Identification of functionally important amino-acids in the cellulose-binding domain of *Trichoderma reesei* Cellobiohydrolase I", *Protein Sci.* **1995**, 4 (6), 1056-1064.

43. A.B. Boraston, L.E. Sandercock, R.A.J. Warren and D.G. Kilburn, "O-glycosylation of a recombinant carbohydrate-binding module mutant secreted by Pichia pastoris", *J. Mol. Microbiol. Biotechnol.* **2003**, 5 (1), 29-36.

44. A.B. Boraston, R.A.J. Warren and D.G. Kilburn, "Glycosylation by Pichia pastoris decreases the affinity of a family 2a carbohydrate-binding module from Cellulomonas fimi: a functional and mutational analysis", *Biochem. J.* **2001**, 358, 423-430.

45. T. Jeoh, W. Michener, M.E. Himmel, S.R. Decker and W.S. Adney, "Implications of cellobiohydrolase glycosylation for use in biomass conversion", *Biotechnol. Biofuels* **2008**, 1, 12.

46. C.B. Taylor, M.F. Talib, C. McCabe, L. Bu, W.S. Adney, M.E. Himmel, M.F. Crowley and G.T. Beckham, "Computational investigation of glycosylation effects on a Family 1 carbohydrate-binding module", *J. Biol. Chem.* **2012**, (287), 3147-3155

47. M.N. Christiansen, D. Kolarich, H. Nevalainen, N.H. Packer and P.H. Jensen, "Challenges of Determining O-Glycopeptide Heterogeneity: A Fungal Glucanase Model System", *Anal. Chem.* **2010**, 82 (9), 3500-3509.

48. A. Diaz-Rodriguez and B.G. Davis, "Chemical modification in the creation of novel biocatalysts", *Current Opinion in Chemical Biology* **2011**, 15 (2), 211-219.

49. B.G. Davis, "Sugars and proteins: New strategies in synthetic biology", *Pure and Applied Chemistry* **2009**, 81 (2), 285-298.

50. D.S. Frenkel, B. , *Understanding Molecular Simulation: From Algorithms to Applications*. 2nd ed ed.; Academic Press: San Diego, CA, **2002**.

51. M. Karplus and G.A. Petsko, "MOLECULAR-DYNAMICS SIMULATIONS IN BIOLOGY", *Nature* **1990**, 347 (6294), 631-639.

52. P. Kollman, "Free energy calculations: Applications to chemical and biological phenomena", *Chem. Rev.* **1993**, 93, 2395-2417.

53. A. Pohorille, Jarzynski, C., & Chipot, C., "Good Practices in Free-Energy Calculations", *J. Phys. Chem. B* **2010**, (114), 10235-10253.

54. S.S. Mallajosyula, O. Guvench, E. Hatcher and A.D. MacKerell, "CHARMM Additive All-Atom Force Field for Phosphate and Sulfate Linked to Carbohydrates", *Journal of Chemical Theory and Computation* **2012**.

55. B.R. Brooks, C.L. Brooks, A.D. Mackerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, W. Yang, D.M. York and M. Karplus, "CHARMM: The Biomolecular Simulation Program", *J. Comput. Chem.* **2009**, 30 (10), 1545-1614.

56. P.J. Kraulis, G.M. Clore, M. Nilges, T.A. Jones, G. Pettersson, J. Knowles and A.M. Gronenborn, "Determination of the 3-dimensional solution structure of the C-terminal domani of Cellobiohydrolase-I from Trichoderma reesei - A study using Nuclear Magnetic Resonance and hybrid distance geometry dynamical simulated annealing", *Biochemistry* **1989**, 28 (18), 7241-7257.

57. Y. Nishiyama, P. Langan and H. Chanzy, "Crystal structure and hydrogen-bonding system in cellulose 1 beta from synchrotron X-ray and neutron fiber diffraction", *J. Am. Chem. Soc.* **2002**, 124 (31), 9074-9082.

58. J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale and K. Schulten, "Scalable molecular dynamics with NAMD", *J. Comput. Chem.* **2005**, 26 (16), 1781-1802.

59. W. Humphrey, A. Dalke and K. Schulten, "VMD: Visual molecular dynamics", *J. Mol. Graph.* **1996**, 14 (1), 33-&.

60. A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins", *J. Phys. Chem. B* **1998**, 102 (18), 3586-3616.

61. A.D. Mackerell, M. Feig and C.L. Brooks, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations", *J. Comput. Chem.* **2004**, 25 (11), 1400-1415.

62. O. Guvench, E. Hatcher, R.M. Venable, R.W. Pastor and A.D. MacKerell, "CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses", *Journal of Chemical Theory and Computation* **2009**, 5 (9), 2353-2370.

63. O. Guvench, S.N. Greene, G. Kamath, J.W. Brady, R.M. Venable, R.W. Pastor and A.D. Mackerell, "Additive Empirical Force Field for Hexopyranose Monosaccharides", *Journal of Computational Chemistry* **2008**, 29 (15), 2543-2564.

64. W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey and M.L. Klein, "Comparison of simple potential functions for simulating liquid water", *J. Chem. Phys.* **1983**, 79 (2), 926-935.

65. S.R. Durell, B.R. Brooks and A. Bennaim, "Solvent-induced forces between 2 hydrophilic groups", *J. Phys. Chem.* **1994**, 98 (8), 2198-2202.

66. U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee and L.G. Pedersen, "A Smooth Particle Mesh Ewald Method", *J. Chem. Phys.* **1995**, 103 (19), 8577-8593.

67. J.P. Ryckaert, G. Ciccotti and H.J.C. Berendsen, "Numerical-integration of cartesian equations of motion of a system with constraints - Molecular-dynamiccs of N-alkanes", *J. Comput. Phys.* **1977**, 23 (3), 327-341.

68. T. Steinbrecher, D.L. Mobley and D.A. Case, "Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations", *J. Chem. Phys.* **2007**, 127 (21).

69. H. Paliwal and M.R. Shirts, "A Benchmark Test Set for Alchemical Free Energy Transformations and Its Use to Quantify Error in Common Free Energy Methods", *Journal of Chemical Theory and Computation* **2011**.

70. S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, "BASIC LOCAL ALIGNMENT SEARCH TOOL", *J. Mol. Biol.* **1990**, 215 (3), 403-410.

71. S.F. Altschul, T.L. Madden, A.A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller and D.J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Research* **1997**, 25 (17), 3389-3402.

72. G.E. Crooks, G. Hon, J.M. Chandonia and S.E. Brenner, "WebLogo: A sequence logo generator", *Genome Res.* **2004**, 14 (6), 1188-1190.

CHAPTER 5

BINDING SITE DYNAMICS AND AROMATIC-CARBOHYDRATE INTERACTIONS IN PROCESSIVE AND NON-PROCESSIVE FAMILY 7 GLYCOSIDE HYDROLASES

5.1 Introduction

As discussed throughout this thesis, fungi and bacteria secrete highly effective cocktails of glycoside hydrolases (GH) and oxidative enzymes that can break down recalcitrant lignocellulosic material to soluble sugars [1-6]. Cellulases are GHs that specifically target cellulose. In this chapter we focus on the catalytic domains of two synergistic cellulases belonging to GH Family 7 from the fungi *T. reesei:* the processive Cel7A exoglucanase, which targets crystalline cellulose at reducing ends, and the non-processive Cel7B endoglucanase, which breaks glycosidic bonds at random substrate sites [1,2,7-10]. The primary difference between the two enzymes is their secondary structural elements; the Cel7A cellodextrin binding sites are enclosed in a tunnel whereas in Cel7B the binding sites are surrounded by an open cleft (Figure 1.8) [7-9,11,12]. As discussed in Chapter I and shown in Figure 1.8, the Cel7A tunnel is comprised of loop structures lining the underside of the CD, while the Cel7B cleft is either missing loops present in Cel7A or contains shorter, more open loops. Enzymes must perform work to break down cellulose [13], and the free energy barriers that a cellodextrin chain, or cellodextrin, must overcome to break free from the protein are higher in a tunnel than that in a cleft conformation. The tunnel loops in Cel7A can also reduce binding site exposure to solvent and intermediate products [9,14,15]. While the structural differences can impart specific functions, the actual basis for processivity may not be as simple as a conformational change, i.e., tunnel or cleft, but may actually come from more subtle molecular-level changes in structural features or dynamics. Using simulations, this study probes the molecular-level differences between the two CDs, which ultimately could help describe the molecular details of both processivity and the synergistic behavior between the two enzymes.

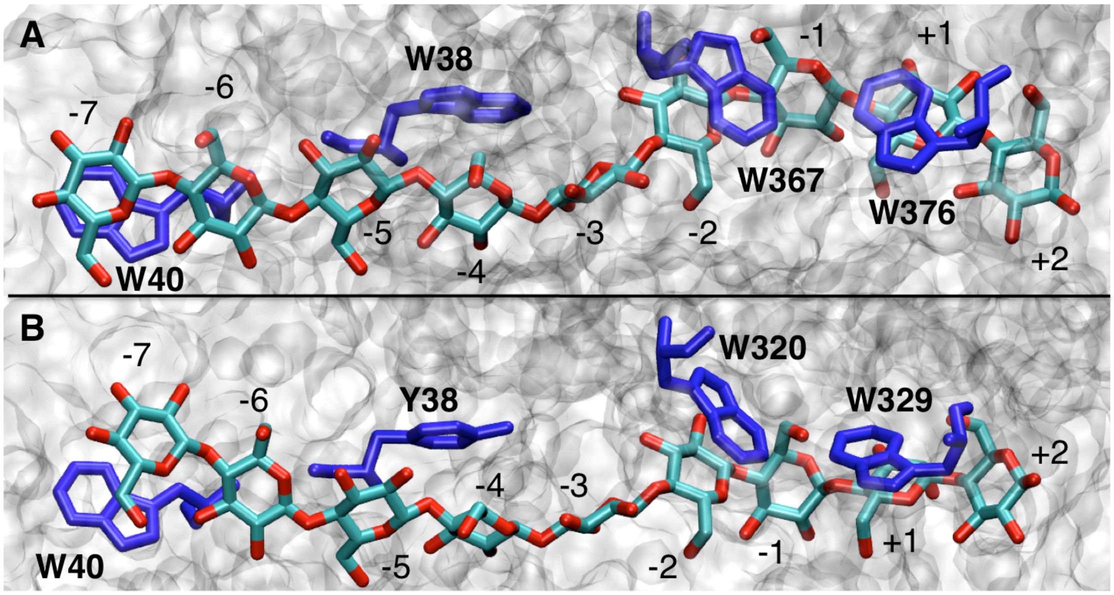


Figure 5.1: *T. reesei* Cel7A (A) and Cel7B (B) cellodextrin binding sites with aromatic residues of interest shown in blue. Binding sites are labeled from the entrance at -7 to the exit at +2. The glycosidic bond cleavage occurs between the -1 and +1 sites.

An additional structural feature of interest in both tunnels and clefts in GHs is the ubiquitous lining of the active sites with aromatic amino acids [3,7,9,16-20]. In Cel7A, four tryptophan (Trp) residues are distributed along the tunnel and participate in cellodextrin binding at the -7, -4, -2, and +1/+2 binding sites, as shown in Figure 5.1A [8]. In Cel7B, three Trp residues and one tyrosine (Tyr) residue are similarly distributed, impacting the same binding sites, as shown in Figure 5.1B [9]. Studies to date have investigated the role some of these residues play in cellodextrin binding and processivity. Koivula *et al.* investigated the role of Trp-272 at the tunnel entrance in the exoglucanase Cel6A from *T. reesei* and found mutagenesis at this site caused dramatic decrease in activity on crystalline cellulose but did not impact activity on amorphous cellulose or protein stability [19]. In a recent computational study on *T. reesei* Cel6A, Payne *et al.* showed that residues associated with cellodextrin acquisition (Trp-272) and product stabilization (Trp-135) had the greatest impact on the cellodextrin binding free energy [16]. Igarashi *et al.* recently used high-speed atomic force microscopy to study processivity in *T. reesei* Cel7A and found that the wild type exhibited a velocity of 3.5 nm/s when degrading a cellulose chain from the crystal surface, but mutation from Trp-40 to alanine (W40A) near the entrance binding site would result in an enzyme displaying repeated attachment and detachment from the surface, but without translation across the cellulose surface [21], suggesting that entrance residues are important for both substrate recognition and processivity. In an experimental study of the processive chitinase B from *Serratia marcescens*, Horn *et al.* found that mutation of two Trp residues near the tunnel entrance reduced processivity and degradation of crystalline chitin, while mutation of a Trp just upstream of the catalytic site (W97) reduced processivity but dramatically increased activity on chitosan [3]. On crystalline substrates, binding of the cellodextrin by aromatic residues is potentially favorable for maintaining chain detachment from the surface, but for soluble substrates, tight binding could reduce the dissociation rate, as illustrated for chitinase B wherein the Trp-97 mutation to alanine (W97A) converted the exochitinase to an endochitinase [3]. Payne *et al.* and von Ossowski *et al.* both suggest that conclusions for one GH family may not be universally applicable to all GH families, but that comparison within GH families might be of significant interest to characterize carbohydrate processivity [12,16].

The goal of the current study is to examine structural and dynamical differences in the processive cellulase Cel7A and the non-processive cellulase Cel7B from *T. reesei*. Molecular simulations can offer insights into the energetics and structure-function relationships related to processivity in carbohydrate-active enzymes [22,23]. Here, we examine structural and dynamical properties related to processivity in terms of cellodextrin and protein dynamics. Additionally, by mutating aromatic amino acids to alanine individually in a processive and non-processive enzyme from the same GH family, we also investigate how the relative binding affinity and interactions between the cellodextrin and protein are impacted. Improvements in molecular-level understanding of differences in exo- and endoglucanases aid in efforts to determine how cocktails of cellulases have evolved to perform different functions in cellulosic degradation, the findings of which may be applied to improve enzymatic hydrolysis in biofuels production [1,12,24].

5.2 Computational Procedures

Cel7A and Cel7B each exhibit at least nine binding sites numbered -7 (at the entrance of the tunnel or cleft near the substrate) to +2 (upstream of the catalytic site), as shown in Figure 5.1, with a putative tenth binding site at +3, not shown in the figure. We perform molecular dynamics (MD) simulations of the two wild type systems and MD simulations of both enzymes with each aromatic residue of interest mutated individually to alanine (Figure 5.1). Using thermodynamic integration (TI), we also calculate the relative binding affinity change for the aromatic mutations to alanine. The MD simulations are used to quantify energetic and dynamical differences between the two systems, both wild type and with aromatic-to-alanine mutations, whereas the TI data quantify how the binding affinity and cellodextrin structure are impacted by mutation in both systems.

All-Atom MD Simulations of the Cel7A and Cel7B Catalytic Domains

Using 250 ns MD simulations, we compare the fluctuations of the protein backbones, calculate the interaction energy between the protein and each cellodextrin-binding site, and measure the number of hydrogen bonds formed between the protein and cellodextrin. Additionally, we investigate changes in cellodextrin solvation and water residence times resulting from the tunnel and cleft formations. VMD was used to determine the number of waters within 3.5 Å of the cellodextrin as a measure of solvation [25]. We calculate the fraction of native contacts between each protein side chain within 6.5 Å of the cellodextrin for both Cel7A and Cel7B to determine the number of long-lived structural contacts between the wild type protein and cellodextrin [26]. Cross-correlation maps were generated for the residues of each wild type using principle component analysis in Amber12 [27,28]. By comparing the wild type and mutants for each enzyme, we attempt to gain insights into the differing roles of the aromatic residues as a function of position in the catalytic tunnel or cleft on cellodextrin binding and stability between a processive and a non-processive cellulase.

CHARMM [29] was used to build, solvate, and minimize the Cel7A and Cel7B protein structures from the original crystal structure PDBs, ID 8CEL and 1EG1, respectively [7,9]. The starting coordinates of the cellodextrin for both systems were taken from a cellodextrin bound Cel7A with the reducing end threaded in first, and the appropriate twist of the cellodextrin and the position of the -1 binding site considered, *i.e.,* the catalytic residues that cleave the glycosydic bond for Cel7A and Cel7B, Glu212 or Glu196 and Glu217 or 201, respectively are positioned ``below" and ``above" the glycosidic linkage to be cleaved, respectively, referred to as the productive binding mode [8]. The Cel7A protein consists of a large globular domain with dimensions of 60 Å x 50 Å x 40 Å containing 434 amino acids [7]. The Cel7B protein is also a globular protein with dimensions of 60 Å x 50 Å x 40 Å and containing 371 amino acids [9]. Approximately one-third of these two structures are arranged in antiparallel β sheets that stack together to form a β sandwich. The solvated systems each contain approximately 55,500 atoms.

NAMD [30] was used for the MD simulations and thermodynamic integration calculations. All simulations were performed in the NVT ensemble and VMD [25] was used for all visualizations. The CHARMM27 all-atom force field with the CMAP correction [29,31,32] was used to describe the proteins, and the cellodextrins were modeled using the most recent CHARMM35 carbohydrate force field [33]. The particle mesh Ewald (PME) method [34] was used for electrostatics with a sixth order b-spline interpolation, a Gaussian distribution with a width of 0.312 Å, and a mesh size of 90 x 90 x 90. The non-bonded interaction cutoff used was 10 Å. The SHAKE algorithm [35] was used to fix covalent bonds to hydrogen atoms. A timestep of 2 fs was used in all simulations. The PBC were 82 x 82 x 82 Å.

The Cel7B backbone was glycosylated at Asn56 and Asn182 in accordance with experimental structural data [9,36,37]. The *N*-glycan structures alone were minimized for 5000 steps using steepest descent in vacuo to improve the starting coordinates generated in CHARMM. Three bound Cel7B systems were built with a 5-mer, a 7-mer, and a 9-mer-glucose chain in the active site to test stability of various cellodextrin lengths in the cleft. Asp-198 and Glu-202 in the active site tunnel are protonated, and two sodium counterions were present in each system. The four systems, Cel7B free and the three test Cel7B bound, were solvated in a 82 Å^3 water box with 2 sodium ions, and then minimized by using the following protocol: minimize water and sodium for 1000 steps of steepest descent, then water and the protein for 1000 steps of steepest descent, and finally the entire system, water, protein, and sugars, for 1000 steps using steepest descent and 1000 steps using the adopted basis Newton-Raphson (ABNR) method. The systems were heated for 50 ps from 50 to 300 K in steps of 50K and equilibrated in the NPT ensemble at 300 K for 100 ps. The Cel7B free system was equilibrated for 100 ns in the NVT ensemble. A non-glycosylated Cel7B system with a cellodextrin was also built and equilibrated under the same conditions for 100 ns in the NVT ensemble. The *N*-glycosylation impacts neither the active site tunnel of Cel7B nor the protein backbone fluctuations (Figure 5.2) and so it was deemed unnecessary to add *N*-glyans to the Cel7A backbone as well for this particular study.

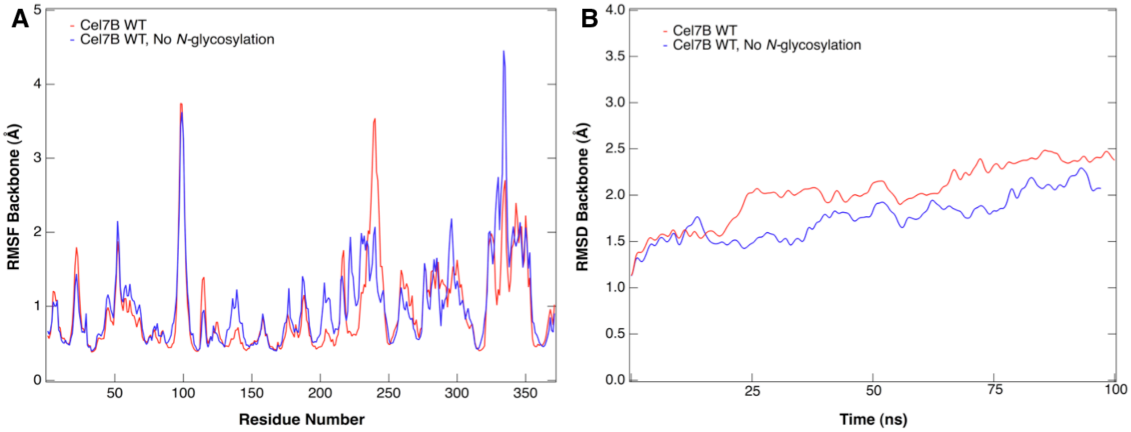


Figure 5.2: Comparison of the Cel7B CD RMSF (A) and RMSD (B) for a glycosylated and non-glycosylated backbone.

The three bound systems were equilibrated for 50 ns in the NVT ensemble after which the RMSF of the glucose chains relative to the Cel7B backbone were calculated. The 9-mer-glucose chain demonstrated the least fluctuation in RMSF (data not shown) and traverses the entire active site, so it was chosen as the current study's bound system and run for an additional 200 ns (for a total of 250 ns) in the NVT ensemble. The final coordinates of the free and bound system were used as the starting coordinates for the TI calculations. Based on the extensive Cel7B setup, the Cel7A systems were bound with a 9-mer-glucose chain, minimized and equilibrated following that of Cel7B, and then run for 250 ns in the NVT ensemble.

MD simulations of the Cel7A mutated systems (Ala-40/38/367/376) and Cel7B mutated systems (Ala-40/38/320/329) were also performed in the NVT ensemble for 250 ns at 300 K with a 2 fs timestep. In these simulations, the system size and simulation parameters were identical to those used in original wild type setups. Each system also followed the same minimization and heating scheme as described previously for the wild type systems.

Post analysis, such as the non-bonded interaction energy between the cellodextrin and protein, RMSF and root mean square deviations (RMSD) of the backbone and cellodextrin were calculated using scripts in CHARMM [29] and in-house codes. Where applicable, the error for these values over the total run was determined using block averaging [38]. Using VMD and a potential distance cutoff of 3.0 Å and 60º from the horizontal plane between a protein residue and binding site, the average number of bonds formed, the number of individual protein residues involved with H-bonding at each cellodextrin site, and the occupancy, or approximate duration, of each bond were determined over 250 ns in the WT and mutated systems. The errors for the number of H-bonds by site equal the standard deviation over the 250 ns simulation. Solvation of the cellodextrin was estimated using VMD [25] by determining the number of waters within 3.5 Å of each binding site over the trajectory. Principle component analysis (PCA), or the residue cross correlation maps for both Cel7A and Cel7B were generated using the ptraj module in Amber12 [27,28]. The 250 ns MD trajectories were first recentered and oriented to remove translational and rotational motions and then stripped of cellodextrin, solvent, and counterions leaving only the protein. Two-dimensional mass-weighted correlation of the alpha carbons was calculated using the matrix tool in ptraj. The fraction of native protein contacts (NC), a reaction coordinate that is useful for measuring deviations from the protein's native folded state, within 6.5 Å of the cellodextrin was calculated for both Cel7A and Cel7B to determine the number of structural contacts between the wild type protein and cellodextrin and associate native contacts with interaction energies [26]. The 250 ns MD trajectories were also first stripped of water and counter-ions and then re-centered and oriented to remove translational and rotational motions before NC analysis. The NC fraction is a measure of the fraction of the time that the contact exists during the entirety of the simulation. Electrostatic and van der Waals interaction energies of the individual residues were calculated as described above using CHARMM [29].

Thermodynamic Integration and Relative Cellodextrin Binding Free Energy

Concurrent with the MD simulations, TI simulations were performed to calculate the changes in cellodextrin binding free energy for removal of aromatic residues in the tunnel of Cel7A and the cleft of Cel7B [16,38-40]. The Cel7A catalytic domain tunnel residues (Trp-40, Trp-38, Trp-367, and Trp-376), and the Cel7B catalytic cleft residues (Trp-40, Tyr-38, Trp-320, and Trp-329), were all been individually mutated to alanine. The initial configurations for both the Cel7A and Cel7B TI calculations were taken from a snapshot following 100 ns of MD simulation described above. In NAMD, the TI calculations were performed using the dual-topology method [39], which is implemented by equilibrating a single structure with a hybrid residue containing both the wild type and mutated atoms. The electrostatic and van der Waals calculations were decoupled, reducing computational effort and eliminating instabilities arising from large energy interactions [39]. Window width selection and simulation times were based on our experiences in prior TI studies on aromatic mutations in the Cel7A CBM and the Cel6A CD and the guidelines outlined by Pohorille *et al.* [16,39,41]. Figure 5.3A shows an example of the probability distribution analysis demonstrating the requirement of sufficient window overlap is achieved. Figure 5.3B shows an example plot of the autocorrelation function versus time; the single exponential decay was fitted in order to determine τ for use in the error analysis (Eqn. 2.24).

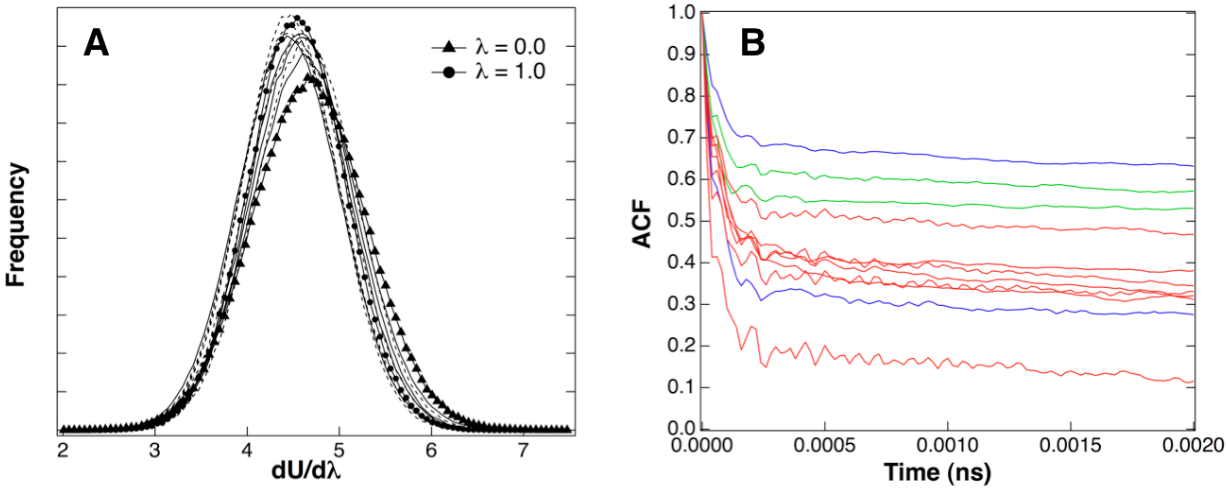


Figure 5.3: Example data set from the Cel7A/B CD TI simulations for window overalp and autocorrelation function. These data are taken from the production portion in the electrostatic simulation set from the Cel7B bound W320A calculation illustrating window overlap (A) and calculation of the autocorrelation time (B).

The electrostatic and van der Waals calculations comprised 11 equidistant λ windows from 0 to 1 with additional windows at 0.05, 0.15, 0.85, and 0.95, selected by examining the probability histograms of $dU/d\lambda$ at each λ value [39,42]. The electrostatic windows were each equilibrated for 1 ns before 10 ns of TI NVT runs, and the van der Waals windows were each equilibrated for 2 ns followed by a 15 ns TI NVT run. The error for each window, Δ_j, was calculated using the methods outlined by Steinbrecher *et al.* [42] and described in Chapter II (Equations 2.24-2.27), with a slight modification to error propagation, shown below.

$$\Delta_{tot} = \sum_i \frac{1}{2}\left(\lambda_{j+1} - \lambda_{j-1}\right)\Delta \tag{5.1}$$

This older method of error propagation does not impact the integrity of the error calculations; in fact, Eqn. 5.1 actually increases the estimated error over Eqn. 2.25. Each window exhibited a single exponential decay and the window correlation time was calculated by dividing the total window into 1 ns increments and averaging τ over the first 100-150 ps of each increment. In the Cel7A and Cel7B CD simulations, the average autocorrelation time for all simulations was 0.0002 ± 0.0003 ns, and the total $\Delta\Delta G$ errors were less than 0.5 kcal/mol. The errors were propagated to the partition coefficients using Equation 2.27.

5.3 Results

Molecular-level Comparison of the Cel7A and Cel7B Wild Type Catalytic Domains

MD simulations were conducted to investigate molecular-level differences between the two CDs with a bound cellodextrin spanning the -7 to +2 sites. The root mean square fluctuations (RMSF) from these simulations are shown in Figure 5.4, with the appropriate tunnel loop deletions (residues 189 to 202, 234 to 252, 316 to 321, 333 to 341, and 381 to 390) inserted for Cel7B to achieve proper sequence and structural alignment [9,12]. Both systems exhibit similar RMSFs, with higher fluctuations in the loop regions surrounding the tunnel and the cleft (Figure 5.4). The tunnel loops not present in Cel7B are quite flexible in Cel7A, even though the cellodextrin remains bound in the tunnel during the simulation. The higher RMSF value in the residue 94-104 loop region for Cel7B correlates with the experimental observation that this region is more open, allowing more access to the active site than the corresponding 96-103 loop in Cel7A [9]. The Cel7B peak from residues 260-280 near the exit of the cleft corresponds to a loop structure that is not present in Cel7A. The root mean square deviations (RMSD) for each system were also found to be similar, with both structures stabilizing within 20-50 ns at less than 3 Å from the starting crystal structure (Figure 5.9).

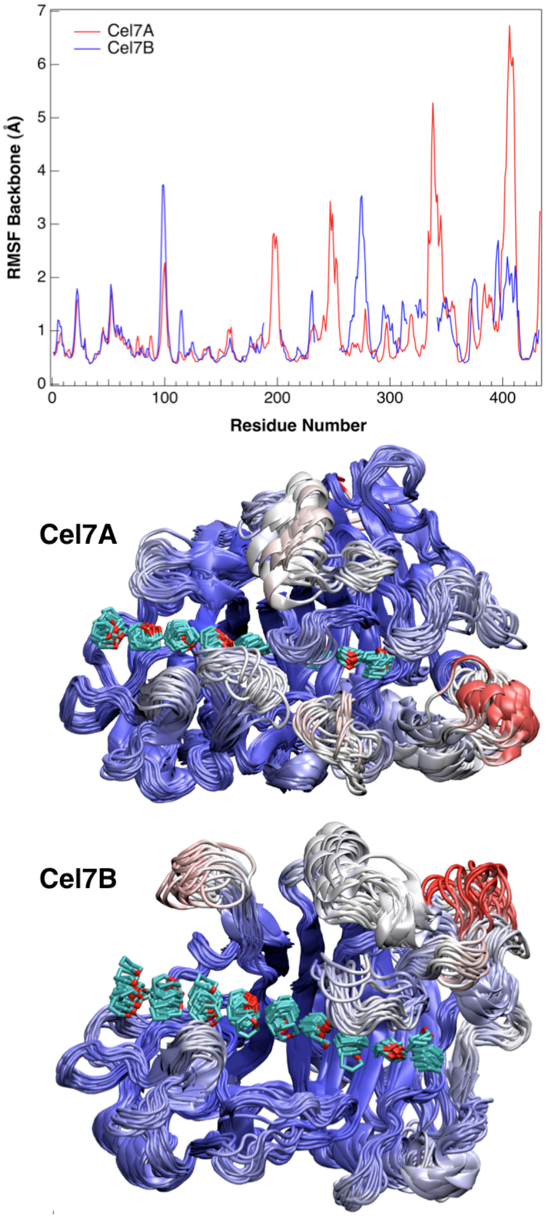


Figure 5.4: (Top) The RMSF of the protein backbones with tunnel loop residues 189 to 202, 234 to 254, 316 to 321, 333 to 341 and 381 to 390 (15) skipped (*i.e.,* designated by gaps in the blue line) in Cel7B to achieve alignment. (Middle) cluster representations of the Cel7A and (Bottom) Cel7B domains as shown from underneath, with the ligand oriented from left to right -7 to -2. The ligand ring carbon atoms are cyan and the oxygen atoms are red. The coloring of the CD from blue to white to red represents increasing RMSF, respectively, scaled to a maximum fluctuation value of 7 Å.

The interaction energy and hydrogen bonding (H-bonding) of the proteins to each cellodextrin site are shown in Figure 5.5 and confirm that the tunnel structure of Cel7A increases both hydrogen bonding to and interaction with the cellodextrin over Cel7B.

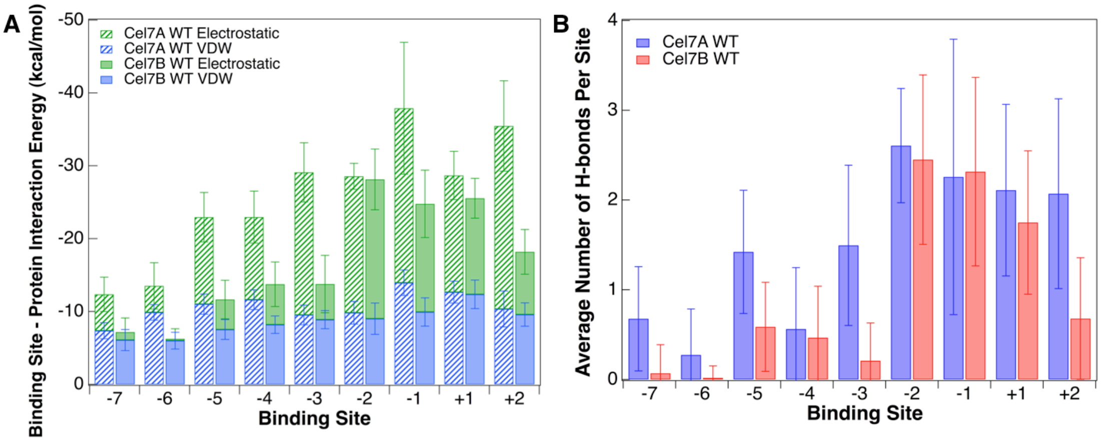


Figure 5.5: Interaction energy (A) and H-bonds (B) between the protein and each binding site. Interaction energy error bars were calculated using block averaging. Hydrogen bonds were calculated using a distance criteria of 3.0 Å and an angle of < 60° from horizontal. H-bond error bars indicate the standard deviation.

Hydrogen bonding between the protein and cellodextrin was calculated to first compare our simulations with X-ray crystallographic predictions of the number of potential protein-cellodextrin H-bonds formed in *T. reesei* Cel7A [8] and a homologous endoglucanase, Cel7B from *Melanocarpus albomyces* [15]. Our results show that the residue alignments and location of H-bonds do not deviate significantly from crystallographic studies [8,9,15], (Tables 5.1 and 5.2). Additionally, the H-bond data for the catalytically active residues are consistent with structural studies [8,15,43] and requirements for a catalytically active complex where these interactions occur to maintain the position of the glycosidic bond between +1 and -1 pointed towards the Glu-217/201 residue [8]. With the exception of electrostatic interaction at the -1 site, the interactions and H-bonding directly around the catalytic site (+1 to -2) are equal within error. The strictly conserved catalytic residues in Family 7 GHs [9] in Cel7A/B are comprised of two glutamic acid residues (Glu-217/201, pointed toward the glycosidic bond, and Glu-212/196) and one aspartic acid residue (Asp-214/198) [7,11].

Table 5.4: Count of residues involved with H-bonding to a cellodextrin predicted by structural studies [8,15] and the current study's simulation results.

	Cel7A			**Cel7B**		
Site	*Predicted*	*Average H-bond Count*	*Number Residues Involved*	*Predicted*	*Average H-bond Count*	*Number Residues Involved*
-7	2	1	2	NA	0	0
-6	3	1	2	NA	0	0
-5	3	2	4	2	1	1
-4	1	1	3	2	1	3
-3	4	2	4	4	0	2
-2	4	3	3	4	2-4	4
-1	6	2-4	5	2	2-3	4
+1	7	2-3	6	7	2	3
+2	3	2-3	5	6	1	3

Table 5.5: Hydrogen bonding patterns and occupancies in Cel7A and Cel7B wild type CD. Where applicable, active site residues are listed first in red (Glu-212/Glu-196, Asp-214/Asp-198, and Glu-201/Glu-217), followed by key aromatic tunnel residues in blue (Trp-376/Trp-329, Trp-367/Trp-320, Trp-38/Tyr-38, and Trp-40/Trp-40) for Cel7A/B. Other than the catalytic and aromatic residues, only occupancies > 5% of the total 250 ns run are displayed.

	Cel7A		Cel7B	
-7	Trp-40	<5%	Trp-40	<5%
	Gln-7	58%		
	Asn-49	10%		
-6	Trp-38	--	Tyr-38	<5%
	Trp-40	<5%	Trp-40	<5%
	Gln-101	14%		
	Asn-49	5%		
-5	Trp-38	<5%	Tyr-38	<5%
	Asn-37	32%	Asn-37	54%
	Asn-103	88%		
	Lys-181	13%		
	Gln-101	7%		
-4	Trp-38	<5%	Tyr-38	--
	Lys-181	35%	Ser-104	7%
	Asn-37	8%	Asn-37	46%
	Val-104	7%	Val-105	8%
-3	Trp-367	--	Trp-320	<5%
	Trp-38	--	Tyr-38	7%
	Arg-107	100%	Arg-108	8%
	Tyr-370	15%		
	Asp-179	12%		
	Asp-368	8%		
-2	Trp-367	<5%	Trp-320	--
	Arg-107	89%	Arg-108	104%
	Ser-365	76%	Ser-318	69%
	Tyr-145	96%	Tyr-146	60%
			Asp-322	7%
-1	Glu-212	34%	Glu-196	<5%
	Asp-214	<5%	Asp-198	<5%
	Glu-217	93%	Glu-201	21%
	Trp-367	49%	Trp-320	<5%
	Asp-173	34%	Gln-325	63%
	Gln-175	12%	Gln-174	41%
			Ser-144	79%
+1	Glu-212	<5%	Glu-196	<5%
	Asp-214	<5%	Asp-198	<5%
	Glu-217	45%	Glu-201	84%
	Trp-376	8%	Trp-329	--
	His-228	21%	His-212	16%
	Asp-259	65%	Ala-222	73%
	Thr-226	60%		
	Gln-175	8%		
+2	Trp-376	<5%	Trp-329	<5%
	Arg-394	100%	Gly-225	35%
	Arg-259	41%	Gly-223	34%
	Arg-251	7%	Ala-222	45%
	Asp-262	18%		
	Tyr-381	9%		

The primary interaction of the aromatic residues of interest in Figure 5.5 comes from van der Waals stacking over the cellodextrin ring structures. Differences in interaction and H-bonding arise in the entrance and exit regions. Figure 5.5A shows that the increase in interaction energy results from Cel7A's electrostatic interaction with the cellodextrin; the van der Waals interactions are relatively equal for the two CDs. This observation can primarily be attributed to the fact that more charged and polar residues are in contact with cellodextrin with the presence of the Cel7A loops, increasing the electrostatic interaction in Cel7A for most sites but having less impact on the van der Waals interaction at each site. Comparing the native contacts within 6.5 Å of each binding site, we find some similarities, but also that the changes in the electrostatic interaction for the -3, -1 and +2 sites between Cel7A and Cel7B can be associated with a change in native contacts, shown in Tables 5.3 and 5.4).

Table 5.6: Cel7A wild type native contact residues and fraction by binding site. VDW and electrostatic interaction energies shown with errors calculated using block averaging.

Cel7A Native Contacts and Interaction Energy

Residue		Side Chain	Binding site	NC Fraction	VDW	+/-	ELEC	+/-
Gln	**7**	Polar	-7	0.73	-0.60	0.94	-2.90	2.95
Trp	**40**	Aromatic	-7	0.99	-4.36	1.72	-0.66	1.01
Trp	**40**	Aromatic	-6	0.95	-2.89	1.11	-1.34	0.97
Gln	**101**	Polar	-6	0.84	-1.21	0.90	-2.44	3.40
Asn	**103**	Polar	-5	0.85	-0.24	0.89	-4.67	2.17
Trp	**38**	Aromatic	-4	1.00	-3.83	1.42	-2.70	1.12
Trp	**38**	Aromatic	-3	0.93	-1.63	0.78	-2.25	0.91
Arg	**107**	Charge (+)	-3	0.99	0.22	1.41	**-21.59**	8.73
Arg	**107**	Charge (+)	-2	1.00	-0.55	1.32	-7.86	3.76
Ser	**365**	Polar	-2	1.00	-0.08	0.92	-6.37	2.61
Trp	**367**	Aromatic	-2	1.00	-4.07	1.56	-1.61	1.08
Tyr	**145**	Aromatic	-1	0.98	-2.26	0.88	0.55	0.69
Asp	**173**	Charge (-)	-1	0.98	-0.64	0.97	-16.89	7.31
Glu	**212**	Charge (-)	-1	0.98	-0.26	1.11	-31.22	11.86
Asp	**214**	Charge (-)	-1	1.00	-1.50	0.60	1.89	1.08
Glu	**217**	Charge (-)	-1	1.00	-0.66	1.15	-6.46	3.02
Trp	**367**	Aromatic	-1	1.00	-1.39	1.07	-2.92	1.60
Gln	**175**	Polar	+1	0.74	-1.15	0.66	1.00	2.18
Glu	**217**	Charge (-)	+1	0.93	-0.54	0.81	-2.79	2.27
Thr	**226**	Polar	+1	1.00	-0.66	0.87	-3.93	2.91
Trp	**376**	Aromatic	+1	1.00	-3.37	1.27	-1.05	1.28
Asp	**259**	Charge (-)	+2	0.98	-1.46	1.19	-23.00	11.42
Asp	**262**	Charge (-)	+2	0.78	-0.72	0.98	-10.38	7.87
Trp	**376**	Aromatic	+2	0.84	-2.85	1.16	-0.05	0.82
Tyr	**381**	Aromatic	+2	0.93	-1.39	0.68	-0.90	1.36
Arg	**394**	Charge (+)	+2	0.78	-1.39	0.68	-0.90	1.36

Table 5.7: Cel7A wild type native contact residues and fraction by binding site. VDW and electrostatic interaction energies shown with errors calculated using block averaging.

Cel7B Native Contacts and Interaction Energy								
			Binding	**NC**				
Residue		**Side Chain**	**site**	**Fraction**	**VDW**	**+/-**	**ELEC**	**+/-**
Trp	**40**	Aromatic	-7	0.98	-4.55	1.04	-0.60	0.71
Trp	**40**	Aromatic	-6	0.69	-2.70	0.53	0.22	0.54
Tyr	**38**	Aromatic	-5	0.79	-3.82	0.62	-0.73	0.84
Tyr	**38**	Aromatic	-4	1.00	-3.23	0.69	-0.21	0.50
Ser	**106**	Polar	-3	0.99	-1.34	0.42	-0.42	0.56
Arg	**108**	Charge (+)	-3	0.99	-2.04	1.19	-3.05	4.17
Arg	**108**	Charge (+)	-2	1.00	-0.31	1.42	**-8.30**	2.00
Ser	**318**	Polar	-2	1.00	-0.42	0.90	-5.25	1.47
Trp	**320**	Aromatic	-2	1.00	-4.36	0.80	-0.47	0.47
Ser	**144**	Polar	-1	0.88	0.53	1.11	-5.62	1.74
Gln	**174**	Polar	-1	0.87	-0.67	0.87	-2.34	1.49
Asp	**198**	Charge (-)	-1	0.77	-0.89	0.32	0.19	0.35
Glu	**201**	Charge (-)	-1	0.99	-1.36	0.63	-2.49	1.91
Trp	**320**	Aromatic	-1	0.99	-2.26	0.67	-1.00	1.03
Gln	**174**	Polar	+1	0.82	-1.18	0.38	-0.07	0.87
Glu	**201**	Charge (-)	+1	0.98	-0.12	1.03	-5.64	2.02
Thr	**210**	Polar	+1	1.00	-1.26	0.50	-0.89	0.59
Trp	**329**	Aromatic	+1	1.00	-3.39	0.72	-0.78	0.64
Ala	**222**	Hydrophobic	+2	0.98	-1.32	0.86	-1.71	2.06
Trp	**329**	Aromatic	+2	0.96	-4.12	1.35	-3.46	1.07

Recent simulation studies from Lin *et al.* showed that Trp-320 dominates calculated coupling strengths in the -1 and -2 sites, but cellodextrin “clenching” weakens toward the cleft entrance in Cel7B [44]. Our results show similar behavior, in that the aforementioned open loop structure near the entrance to the cleft results in reduced to no electrostatic interaction or hydrogen bonding with the cellodextrin beginning with the entrance -7 site and extending inward to the -3 site. The cellodextrin chain twist required to “flip” the chain over in the catalytic sites is observed in both Cel7A and Cel7B, with the -2 to -4 sites key for stabilizing the cellodextrin and maintaining the twist [8,45]. Interaction energy and hydrogen bonding are both decreased in Cel7B for the -3 site, partially due to the fact that the deleted loop 189-201 in Cel7A can intermittently form H-bonds with the -3 and -4 sites, adding structure to this area (data not shown). Conserved residues Arg-107/108, Ser-365/318, and Tyr-145/146 (Cel7A/B) all form H-bonds with the -2 and -3 sites as shown by crystallography studies (Table 5.2) [8,15]; however, the Arg interactions also result in an electrostatic energy loss of -18.1 ± 7 kcal/mol for Cel7B (Table 5.3 and 5.4). While the RMSF values for Arg-107 (Cel7A) and Arg-108 (Cel7B) are nearly equal, the fluctuations of the residues directly upstream in the aforementioned 94-104 loop region are higher in Cel7B, (Figure 5.4) which likely correlates with the observed changes in the interaction energy and loss in electrostatic interaction and H-bonds for the -3 site in Cel7B. The Cel7A loop 234-254 is within 6.5 Å of the product +2 and +1 sites, allowing for increased interaction energy and formation of H-bonds with the +2 site in particular (Figure 5.5). The structure of the loops surrounding the product +1 and +2 sites in Cel7A and Cel7B are similar but differ in sequence: Cel7A forms H-bonds with three charged Arg residues (251, 267, and 394) as reported experimentally [8], while Cel7B forms H-bonds with an alanine (Ala-222) and two glycines (Gly-223 and 225) (Table 5.2). The change in sequence also results in a loss of 17 ± 8 kcal/mol in total interaction energy at the +2 and +1 sites in Cel7B (Figure 5.6), contributing to the decrease in product interaction and H-bonding at the +2 site for Cel7B observed in Figure 5.5B [15,46].

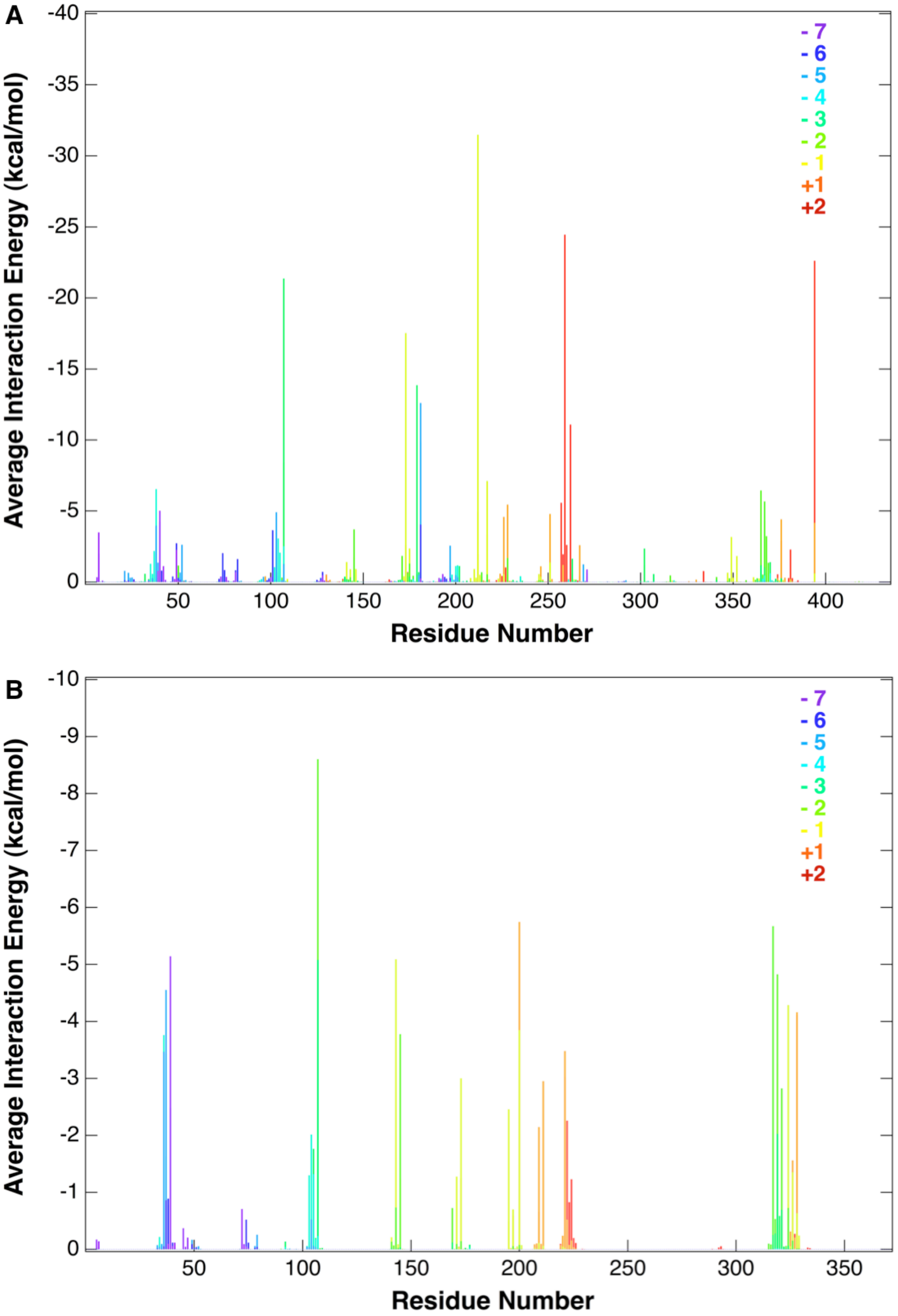


Figure 5.6: Interaction energy between each binding site and the individual protein residues for Cel7A (A) and Cel7B (B) for the first 100 ns of simulation. Binding sites are designated by color. Errors were calculated using block averaging, but are not included in the figure for simplicity and neatness. Electrostatic and van der Waals interaction energies were calculated, but not shown in the text, and all values reported in the text include the errors from the data files

Molecular-level Comparison of the Bound Cellodextrin in the Cel7A and Cel7B Wild Type

The behavior of the bound cellodextrin also changes in some sites between Cel7A and Cel7B. The RMSF and change in solvation around the cellodextrin by binding site is shown in Figure 5.7. As shown in Figure 5.7A, the fluctuations of each cellodextrin in the two wild type CDs are similar, and they correlate with the energetic and H-bond profiles. While the Cel7B loops are more open than in Cel7A from the -3 to -7 sites, demonstrated by the nearly unrestricted view of the cellodextrin in Figure 1.8, the cellodextrin RMSF values in Cel7A and Cel7B are nearly equal except at the entrance sites (Figure 5.7A), where the Cel7B 94-106 loop opens above the cellodextrin, reducing stabilizing protein-cellodextrin interactions. It is also noted that the product exit site, +2, exhibits a larger degree of error than in Cel7A, corresponding to lower interaction energy and H-bonding for Cel7B shown in Figure 5.5B. As expected, the solvation for Cel7B increases over Cel7A from the entrance to the -3 site, where the tunnel loops are removed (Figure 5.7B). Additionally, the RMSD values of the overall cellodextrin are stable in both systems, but twice as high for Cel7B (Figure 5.13).

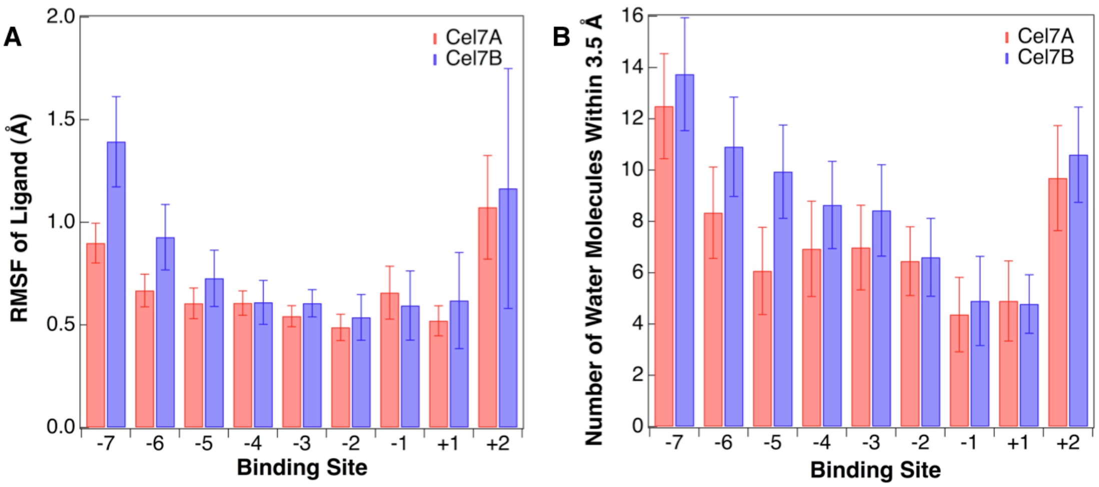


Figure 5.7: RMSF of the bound cellodextrin (A) and the solvation, or number of waters within 3.5 A, of the bound cellodextrin (B) for Cel7A and Cel7B over a 250 ns MD production run. Errors were calculated using block averaging.

We calculated the residence time of individual waters for each binding site, and generated probabilities based on the normalized histograms of the data, found in Figure 5.8. In both systems, very few waters have a probability > 1% of remaining around a binding site for longer than 0.5-1 ns, so the graphs are truncated. Additionally, most of the waters enter and leave a site in less than 0.1 ns, so to see the differences in the binding sites over a smaller range, the probabilities are truncated at 15%. Waters have a higher residence time probability near the catalytic site, -1 and +1 for Cel7A and Cel7B, respectively, and both systems have the lowest residence times at the entrance site, -7. Generally, the probability distribution is evenly distributed over each binding site in Cel7B, corresponding to similar solvation values and cellodextrin RMSF values in Figure 5.7. Conversely, Cel7A shows a slight "grouping" of probabilities from 2 to 7% for its cellodextrin sites, likely due to the tunnel conformation where waters are slightly more structured.

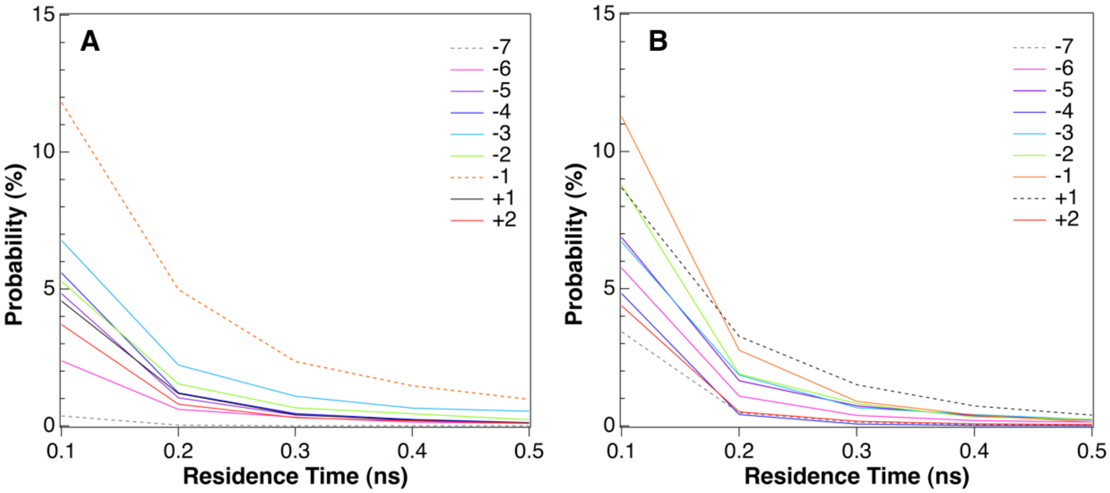


Figure 5.8: Water residence time normalized probability by binding site around Cel7A cellodextrin (A) and Cel7B cellodextrin (B) over a 250 ns MD simulation. All individual waters have < 1% probability of remaining around a site for longer than 0.5 to 1 ns, so the chart is truncated. Dotted lines represent the highest probabilities (*i.e.* longest residence times) and lowest probabilities (*i.e.* shortest residence times).

Water around the cellodextrin in the tunnel increases in the first 50-100 ns of the simulation, while the water around the cellodextrin in the cleft remains stable (data not shown). The cellodextrin data again indicates that entrance and exit binding site changes may be helpful to explaining overall processive and non-processive action in the two enzymes.

Relative Binding Free Energy from Aromatic Acid Mutation in the Cel7A and Cel7B from TI Simulation

The results of the TI simulations performed to mutate the active tunnel/cleft aromatic residues to alanine are shown in Table 5.5. Mutagenesis from aromatic to alanine is detrimental to cellodextrin binding for both Cel7A and Cel7B. We estimate partition coefficients typically calculated in experimental binding studies using Eqn 1.1:

$$\Delta\Delta G = -RT\ln\left(K_{\mathrm{Mut}} \Big/ K_{\mathrm{WT}}\right) \tag{5.1}$$

Table 5.8: Relative binding free energy changes for Cel7A and Cel7B per binding site as a result of aromatic acid mutation. The partition coefficients were estimated using Eqn 5.1 for experimental binding affinities and then inverted to show the improvement of the wild type over the mutated states.

	Mutation	Binding Site	$\Delta\Delta G$ (kcal/mol)	$K_{\mathrm{WT}}/K_{\mathrm{Mut}}$
Cel7A	W40A	-7	1.4 ± 0.4	1.0E+01
	W38A	-4	4.2 ± 0.2	1.1E+03
	W367A	-2	3.3 ± 0.3	2.5E+02
	W376A	+2/+1	2.0 ± 0.3	2.9E+01
Cel7B	W40A	-7	2.8 ± 0.2	1.1E+02
	Y38A	-4	2.7 ± 0.5	9.3E+01
	W320A	-2	4.5 ± 0.2	1.9E+03
	W329A	+2/+1	6.5 ± 0.2	5.4E+04

The results in Table 5.5 are the cumulative result of the information found in Table 5.6 and 5.7. The cumulative results are calculated from Equation 2.10.

Table 5.9: Detailed relative binding free energies and associated binding affinity (K_{WT}/K_{Mut}) calculated from TI for the *T. reesei* Cel7A system

		Bound		**Free**	
		Energy [kcal/mol]	**Error [kcal/mol]**	**Energy [kcal/mol]**	**Error [kcal/mol]**
W40A	Electrostatics	8.7	0.09	7.5	0.04
	VDW	3.6	0.40	3.4	0.10
	$\Delta\Delta G$ [kcal/mol]	**1.4 ± 0.4**			
	K_{WT}/K_{Mut}	**9.6E+00**			
W38A	Electrostatics	10.1	0.06	6.9	0.04
	VDW	2.7	0.11	1.7	0.13
	$\Delta\Delta G$ [kcal/mol]	**4.2 ± 0.2**			
	K_{WT}/K_{Mut}	**1.1E+03**			
W367A	Electrostatics	6.9	0.05	5.0	0.03
	VDW	6.0	0.17	4.6	0.17
	$\Delta\Delta G$ [kcal/mol]	**3.3 ± 0.25**			
	K_{WT}/K_{Mut}	**2.7E+02**			
W376A	Electrostatics	5.1	0.06	0.1	0.04
	VDW	6.6	0.21	5.1	
	$\Delta\Delta G$ [kcal/mol]	**2.0 ± 0.3**			
	K_{WT}/K_{Mut}	**2.7E+01**			

Table 5.10: Detailed relative binding free energies and associated binding affinity (K_{WT}/K_{Mut}) calculated from TI for the *T. reesei* Cel7B system

		Bound		**Free**	
		Energy [kcal/mol]	**Error [kcal/mol]**	**Energy [kcal/mol]**	**Error [kcal/mol]**
W40A	Electrostatics	9.1	0.03	7.4	0.03
	VDW	1.9	0.10	0.7	0.08
	$\Delta\Delta G$ [kcal/mol]	**2.8 +/- 0.2**			
	K_{WT}/K_{Mut}	**1.1E+02**			
Y38A	Electrostatics	9.4	0.03	6.1	0.04
	VDW	-1.0	0.05	-0.3	0.32
	$\Delta\Delta G$ [kcal/mol]	**2.7 +/- 0.5**			
	K_{WT}/K_{Mut}	**9.3E+01**			
W320A	Electrostatics	5.8	0.07	4.9	0.03
	VDW	6.9	0.06	3.4	0.07
	$\Delta\Delta G$ [kcal/mol]	**4.5 +/- 0.2**			
	K_{WT}/K_{Mut}	**1.9E+03**			
W329A	Electrostatics	6.7	0.01	3.8	0.02
	VDW	5.9	0.07	2.3	0.06
	$\Delta\Delta G$ [kcal/mol]	**6.5 +/- 0.2**			
	K_{WT}/K_{Mut}	**5.4E+04**			

Molecular-Level Comparison of Aromatic Acid Mutations

To investigate the molecular-level changes associated with these mutations, we ran 250 ns MD simulations to calculate the interaction energy for the protein and cellodextrin and fluctuations in the protein and cellodextrin as we did in the wild type simulations. Generally, mutations in Cel7A do not impact the structural and energetic profiles as much as mutations in Cel7B do, which is reflected in the ΔΔ*G* values in Table 5.5.

To determine if there are differences in the protein flexibility in Cel7A and Cel7B, we calculated the RMSD and RMSF values over 250 ns of MD. The WT CD in each system (shown in red for both systems) exhibits similar RMSD (Figure 5.9) and RMSF curves (Figure 5.10).

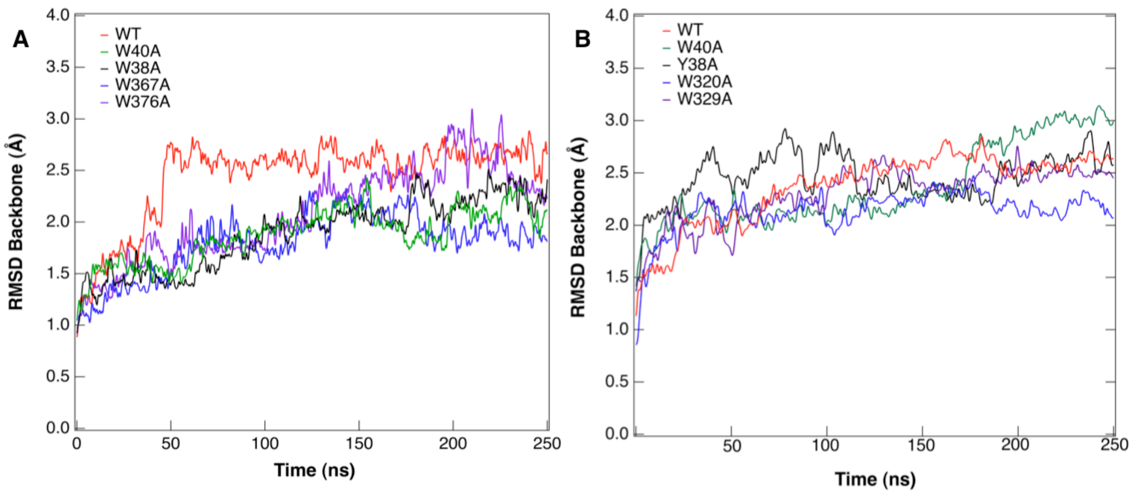


Figure 5.9: RMSD of the protein backbone over 250 ns for Cel7A wild type, WT, and mutations (A) and Cel7B WT and mutations (B).

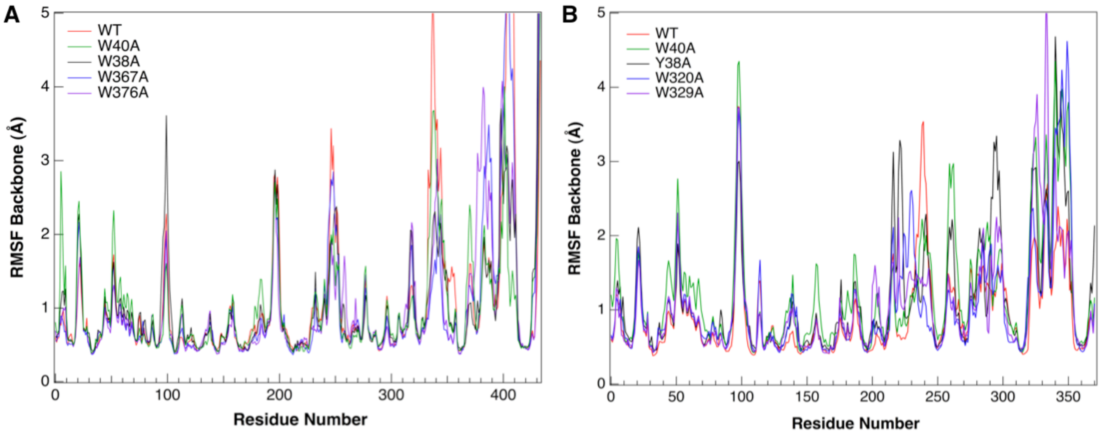


Figure 5.10: RMSF of the protein backbone over 250 ns for Cel7A wild type, WT, and mutations (A) and Cel7B WT and mutations (B).

While the protein backbone RMSD and RMSF curves do not change drastically with mutation in either enzyme (Figures 5.9 and 5.10, respectively), the interaction energy between the protein and individual binding sites and the RMSF for each binding site provide insight into the changes occurring with mutation of the tunnel aromatic residues (Figure 5.11). For Cel7A, binding is most affected at the -4 (W38A) and -2 (W367A) binding sites; Figure 6A shows that the W367A mutation decreases the average interaction energy at the +2 and -1 site by 10-20 kcal/mol (Figure 5.11A). von Ossowski *et al.* suggested that processivity relies on a balance of binding in the substrate and product sites [12]. Interestingly, the W367A and W376A mutations decrease interaction energy in the product side but only slightly increase interaction in the substrate side, which suggests that these mutations may affect processivity. Trp-367 also forms a consistent H-bond with the active site (-1) in the wild type (Table 5.2), which is not present with mutation to Ala, potentially impacting the ring conformation necessary for hydrolysis.

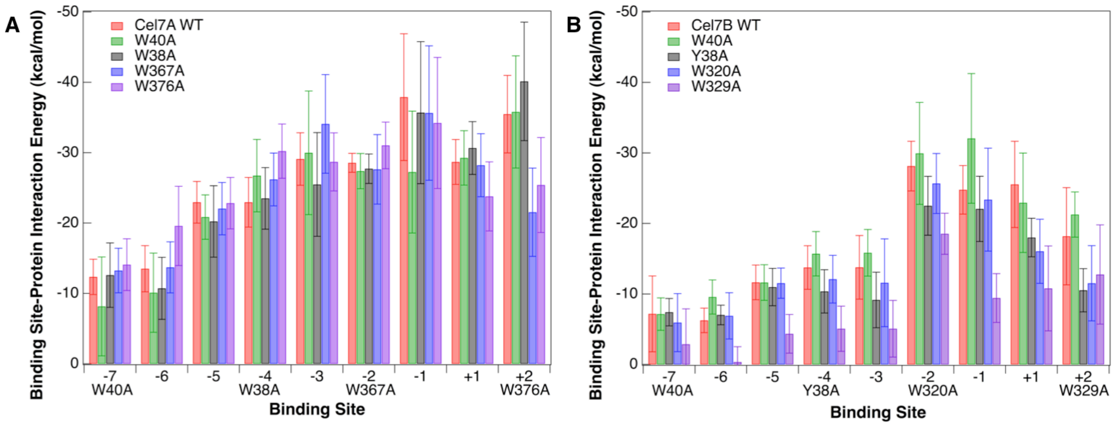


Figure 5.11: Interaction energy of the protein with each binding site in the cellodextrin for Cel7A (A) and Cel7B (B). Mutations to Ala produce a more marked response in Cel7B

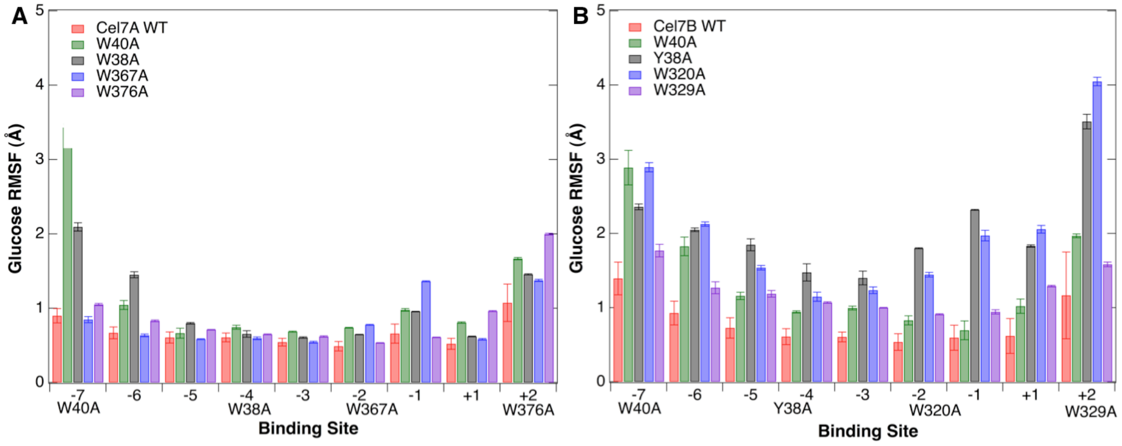


Figure 5.12: RMSF changes at each binding site resulting from aromatic mutation for Cel7A (A) and Cel7B (B)

While the W376A and W367A mutations slightly increase the fluctuation of the cellodextrin in the product and catalytic sites, the W38A and W40A mutations increase fluctuations in the entrance site, which could impact chain acquisition (Figure 5.12A), [16,19]. The relative binding affinity is most negatively impacted for W38A, despite the similarities in the protein and cellodextrin behaviors in the W40A and W38A long MD simulations. In the W38A case, the interaction energy between the protein and cellodextrin is not significantly different overall (Figure 5.11A), nor is the H-bonding of

Arg-107 and other residues to the -4 or -5 site disrupted with mutation (mutation H-bond data not shown). However in the simulations, Trp-38 was identified as the only native contact at the -4 site in Cel7A WT (Table 5.3), and experimental studies identified it as potentially important to maintaining the twist of the cellodextrin [45]. Thus, the loss of the aromatic contact at this site is a loss of an important stabilizing residue, resulting in increased cellodextrin fluctuations at the entrance, exit, and the catalytic (-1) sites (Figure 5.12A).

The tunnel in Cel6A studied by Payne *et al.* is shorter than Cel7A, and changes to the +4 site (Cel6A W272A or Cel7A W38A) substantially altered the cellodextrin fluctuations more than observed in this work for the -7 site (W40A) [16]. This increase in fluctuation at the entrance for Cel6A potentially explains the increased effect on $\Delta\Delta G$ for mutation at the entrance of Cel6A over the entrance of Cel7A. Binding affinity loss with W40A is low, but the loss of interaction energy (Figure 5.11A), dramatic increase in cellodextrin RMSF at the entrance site (Figure 5.12A), and shift in the cellodextrin RMSD (Figure 5.13A) in the simulations suggests that product acquisition could be negatively impacted in W40A, similar to the results from Igarashi *et al.* [21]. While we did not directly examine the kinetics of processivity in these simulations, our results suggest that in Cel7A, cellodextrin binding is most sensitive to mutagenesis directly upstream of the catalytically active site (W367A) and at what perhaps may be an additional substrate acquisition site (W38A).

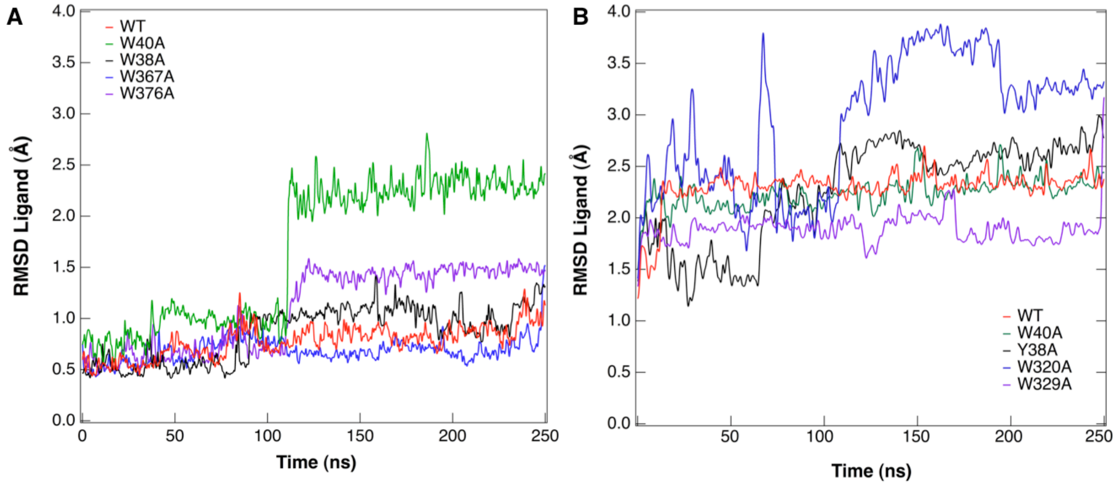


Figure 5.13: RMSD over 250 ns of the cellodextrins bound in Cel7A wild type (WT) and mutated structures (A) and Cel7B WT and mutated structures (B)

Mutation of the four aromatic residues in Cel7B produces larger energetic and structural changes than in Cel7A (Figures 5.11B and 5.12B). Cel7B binding affinity to the cellodextrin is most negatively impacted by mutation at the product site (W329A) and, like Cel7A, directly upstream of the catalytically active site (W320A). The interaction energy drops significantly at all binding sites with the W329A mutation (Figure 6B), corresponding to the $\Delta\Delta G$ results shown in Table 5.5. Examining the cross correlation maps from the MD simulation, we observe that Trp-329 positively correlates with the active site (Glu-196/Asp-198/Glu-201), so changes here may impact the actual catalysis (Figure 5.14).

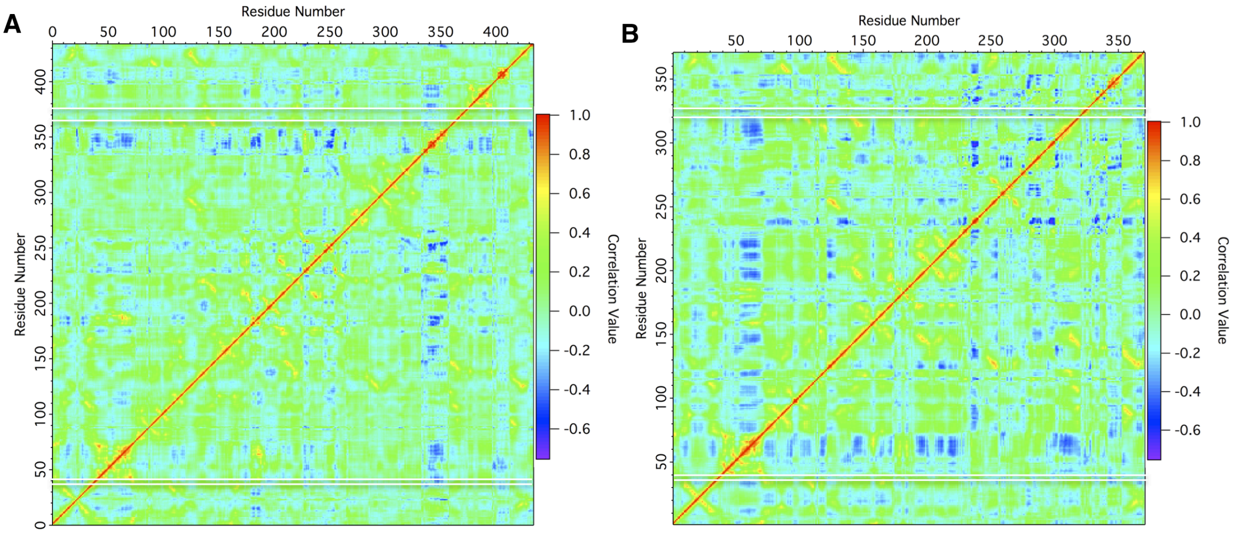


Figure 5.14: Principle component analysis of Cel7A (A) and Cel7B (B). A positive value denotes a positive correlation of motion, whereas negative values are indicative of negative correlation of motion. Zero represents a lack of correlated modes. Horizontal lines running from left to right have been placed at locations on the map corresponding to the aromatic residues of interest in Figure 2.

In viewing the Cel7B wild type and Ala-329 trajectories, we note that Ala-329 is much farther away from the product site than Trp-329, as shown in Figure 5.15A and 5.15B. This change in protein structure leads to dramatic instabilities in the cleft width, measured from the tips of the loops surrounding the +2 to -2 sites (Ala-222, a native contact in the wild type and Gln-325) (Figure 5.15C). Both the wild type and mutant exhibit some relaxation of the structure in the first 50 ns, but the wild type stabilizes whereas the W329A mutation does not**.** Since interaction energy and H-bonding are weaker outside of this region in the wild type (Figure 5.5), any disruption to the cleft here could lead to a reduction in interaction energy along the entirety of the cleft.

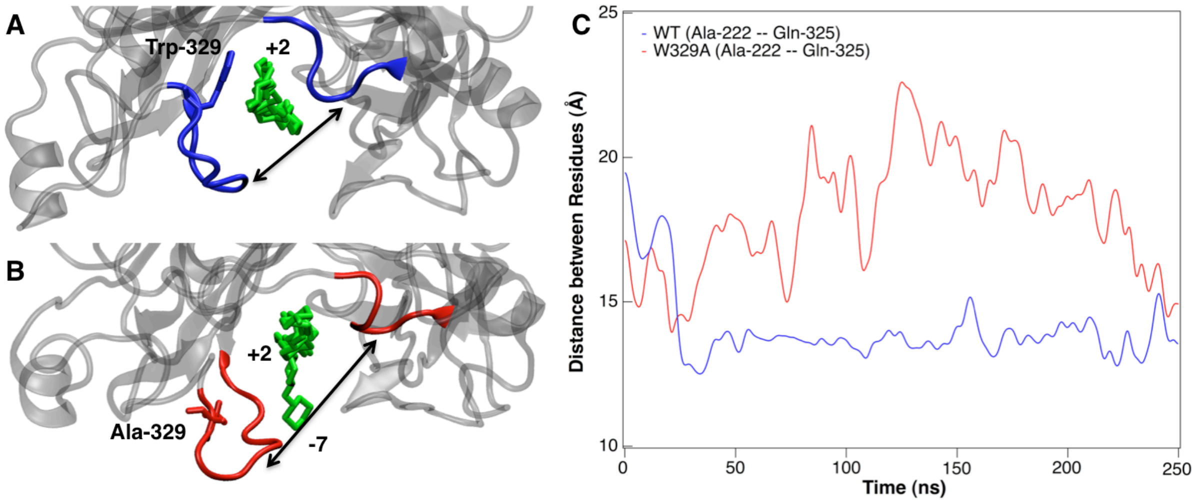


Figure 5.15: Cel7B wild type and W329A MD simulations at 250 ns. Entrance view of cellodextrin and position of Trp-329 in wild type (A) and cellodextrin and position of Ala-329 in the W329A mutation (B) showing change in position of residue 329. Trp-329 and Ala-329 are shown using the licorice rendering in VMD. The arrows correspond to the distance between the tip of the loops surrounding the +2 and +1 sites, measured by the distance between the α-carbons in Ala-222 on the left and Gln-325 on the right. The time series of this distance over the course of 250 ns is shown in (C) and illustrates the significance of W329 in structure stabilization.

W320A impacts the product sites, but not the -1/-2 sites, which may explain why its change in ΔΔ*G* is less detrimental than W329A. As previously discussed, in non-processive enzymes the cellodextrin binding free energy is likely lower than that of processive enzymes. While lower free energy may allow for faster substrate release and be important for preventing product inhibition [15,46], further reductions in interactions may cause the cellodextrin to dissociate from the protein cleft, as reflected in the poor interaction energy and high cellodextrin RMSF results for W329A. Y38A and W320A have a similar impact to protein-cellodextrin interaction energy (Figure 6B) and cellodextrin RMSF (Figure 7B), but Y38A is less detrimental than W320A and W329A to the cellodextrin binding free energy. With the reduced protein-cellodextrin interactions and H-bond formation calculated in the Cel7B wild type, it is not surprising that in general mutagenesis at Tyr-38 and Trp-40 do not impact binding as much as the entrance/active site positions in Cel7B relative to Cel7A. However, since binding is already likely lower than in Cel7A, and Tyr-38 and Trp-40 have been shown to dominate coupling in the entrance [44], mutation to alanine will still have a negative impact on overall affinity. The cross-correlation map shows that Tyr-38 correlates with the binding of Arg-108 while Trp-40 does not (Figure 5.14), with this correlation supported by the loss in H-bond duration between the aforementioned conserved residues Arg-108, Ser-318 and Tyr-146 (mutation H-bond data not shown). Despite these differences, mutagenesis at this site in Cel7B has the same negative impact as mutagenesis at Trp-40 and does not produce the same relative affinity change as in the processive Cel7A. The W40A mutation only impacts protein-cellodextrin interaction energy at the -1 site (Figure 5.11B) but increases the RMSF of the cellodextrin at each site (Figure 5.12B), which could impact binding strength (*i.e.,* a highly fluctuating cellodextrin will not bind properly). As previously discussed, studies of Family 6 GHs and endochitinases indicated that initial substrate recognition and tight binding at the entrance is not required for attachment to amorphous cellulose [3,16,19]; accordingly, our results indicate that the entrance sites in Cel7B do not impact cellodextrin binding affinity significantly, but the exit and catalytic binding sites are particularly sensitive to mutation.

5.4 Discussion and Conclusions

In this study, we examined a processive and a non-processive wild type GH7 enzyme to quantify the dynamical and structural differences in interactions with their respective cellodextrins with the overall aim to quantify key properties that are potentially important for GH processivity. First, our results suggest that the residues associated with the active sites (-2 to +1) may not be directly associated with processivity in GH7s. This hypothesis is supported by the strict sequence conservation at the active site [9] and the nearly equal protein-cellodextrin interaction energies (Figure 5.5), number of H-bonds (Table 5.2), and cellodextrin RMSF (Figure 5.7A) and solvation (Figure 5.7B) calculated at these sites for both Cel7A and Cel7B wild-type enzymes. The aromatic residues around these sites are important for stabilizing the cellodextrin in both enzymes, in that mutagenesis to alanine decreases affinity for the cellodextrin, unlike the results shown for the *T. reesei* Cel6A [16]. Experimental studies can characterize the actual changes in activity on various substrates, but based on the MD and TI results, the catalytic sites (+1 to -2) are important to both cellodextrin binding and catalysis in both processive and non-processive enzymes, but likely not directly for processivity.

In terms of GH7 processivity, the simulation results indicate that the interactions and binding affinity are different for Cel7A and Cel7B both at the entrance (-7 to -3) and exit (+2) sites. Generally, the tunnel conformation of Cel7A imparts increased protein-cellodextrin interactions and H-bonding over the non-processive Cel7B. In Cel7A, mutation at the -4 site (W38A) negatively impacts binding and increases cellodextrin fluctuations at the -7 and +2 sites, indicating that the Trp-38 residue may be critical for maintaining cellodextrin position in the tunnel, cellodextrin acquisition, and potentially processivity [3,16,19]. The weaker interactions of the Cel7B cleft residues with the cellodextrin support the notion that lack of tight binding, especially at the cleft entrance, may decrease processivity and enhance the detachment of the enzyme on insoluble substrates. Additionally, the weak interactions observed in Cel7B explain why aromatic-to-alanine mutations impact the cellodextrin binding on average more than in Cel7A.

Binding, and thus activity on even amorphous substrates could thus be impacted with single mutations in GH7 endoglucanases, such as the exit W329A mutation.

Overall, our results highlight functional differences in energetics and structure of processive and non-processive GH7s (Cel7A/B) at the molecular level. Other GH7 cellulases may exhibit similar energetic and H-bond profiles in their catalytic sites, and these structural details near the entrance and exit could potentially be used as markers in simulation studies for non-processive, endoglucanase behavior. While these results may not be directly applicable to other GH families, more detailed understanding of exo- and endoglucanases such as that elucidated here will eventually lead to a general molecular level theory of carbohydrate processivity. This will enable a clearer definition of the roles of different GH Families in enzymatic cocktails during cellulosic hydrolysis, which ultimately has valuable applicability in improving biochemical conversion of lignocellulosic biomass.

5.5 References

1. M.E. Himmel, S.Y. Ding, D.K. Johnson, W.S. Adney, M.R. Nimlos, J.W. Brady and T.D. Foust, "Biomass recalcitrance: Engineering plants and enzymes for biofuels production", *Science* **2007**, 315 (5813), 804-807.

2. L.R. Lynd, P.J. Weimer, W.H. van Zyl and I.S. Pretorius, "Microbial cellulose utilization: Fundamentals and biotechnology", *Microbiol. Mol. Biol. Rev.* **2002**, 66 (3), 506-+.

3. S.J. Horn, P. Sikorski, J.B. Cederkvist, G. Vaaje-Kolstad, M. Sorlie, B. Synstad, G. Vriend, K.M. Varum and V.G.H. Eijsink, "Costs and benefits of processivity in enzymatic degradation of recalcitrant polysaccharides", *Proc. Natl. Acad. Sci.* **2006**, 103 (48), 18089-18094.

4. B.K. Barr, Y.L. Hsieh, B. Ganem and D.B. Wilson, "Identification of two functionally different classes of exocellulases", *Biochemistry* **1996**, 35 (2), 586-592.

5. Z. Forsberg, G. Vaaje-Kolstad, B. Westereng, A.C. Bunaes, Y. Stenstrom, A. MacKenzie, M. Sorlie, S.J. Horn and V.G.H. Eijsink, "Cleavage of cellulose by a CBM33 protein", *Protein Sci.* **2011**, 20 (9), 1479-1483.

6. G. Vaaje-Kolstad, B. Westereng, S.J. Horn, Z.L. Liu, H. Zhai, M. Sorlie and V.G.H. Eijsink, "An Oxidative Enzyme Boosting the Enzymatic Conversion of Recalcitrant Polysaccharides", *Science* **2010**, 330 (6001), 219-222.

7. C. Divne, J. Stahlberg, T. Reinikainen, L. Ruohonen, G. Pettersson, J.K.C. Knowles, T.T. Teeri and T.A. Jones, "The 3-dimensional crystal-structure of the catalytic core of Cellobiohydrolase-I from *Trichodermal reesei*", *Science* **1994**, 265 (5171), 524-528.

8. C. Divne, J. Stahlberg, T.T. Teeri and T.A. Jones, "High-resolution crystal structures reveal how a cellulose chain is bound in the 50 angstrom long tunnel of cellobiohydrolase I from Trichoderma reesei", *J. Mol. Biol.* **1998**, 275 (2), 309-325.

9. G.J. Kleywegt, J.Y. Zou, C. Divne, G.J. Davies, I. Sinning, J. Stahlberg, T. Reinikainen, M. Srisodsuk, T.T. Teeri and T.A. Jones, "The crystal structure of the catalytic core domain of endoglucanase I from Trichoderma reesei at 3.6 angstrom resolution, and a comparison with related enzymes", *J. Mol. Biol.* **1997**, 272 (3), 383-397.

10. J. Stahlberg, G. Johansson and G. Pettersson, "A NEW MODEL FOR ENZYMATIC-HYDROLYSIS OF CELLULOSE BASED ON THE 2-DOMAIN STRUCTURE OF CELLOBIOHYDROLASE-I", *Bio-Technology* **1991**, 9 (3), 286-290.

11. L.F. Mackenzie, G. Sulzenbacher, C. Divne, T.A. Jones, H.F. Woldike, M. Schulein, S.G. Withers and G.J. Davies, "Crystal structure of the family 7 endoglucanase I (Cel7B) from Humicola insolens at 2.2 angstrom resolution and identification of the catalytic nucleophile by trapping of the covalent glycosyl-enzyme intermediate", *Biochem. J.* **1998**, 335, 409-416.

12. I. von Ossowski, J. Stahlberg, A. Koivula, K. Piens, D. Becker, H. Boer, R. Harle, M. Harris, C. Divne, S. Mahdi, Y.X. Zhao, H. Driguez, M. Claeyssens, M.L. Sinnott and T.T. Teeri, "Engineering the exo-loop of *Trichoderma reesei* cellobiohydrolase, Cel7A. A comparison with *Phanerochaete chrysosporium* Cel7D", *J. Mol. Biol.* **2003**, 333 (4), 817-829.

13. G.T. Beckham, J.F. Matthews, B. Peters, Y.J. Bomble, M.E. Himmel and M.F. Crowley, "Molecular-Level Origins of Biomass Recalcitrance: Decrystallization Free Energies for Four Common Cellulose Polymorphs", *Journal of Physical Chemistry B* **2011**, 115 (14), 4118-4127.

14. M. Claeyssens, H. Vantilbeurgh, J.P. Kamerling, J. Berg, M. Vrsanska and P. Biely, "STUDIES OF THE CELLULOLYTIC SYSTEM OF THE FILAMENTOUS FUNGUS TRICHODERMA-REESEI QM-9414 - SUBSTRATE-SPECIFICITY AND TRANSFER ACTIVITY OF ENDOGLUCANASE-I", *Biochem. J.* **1990**, 270 (1), 251-256.

15. T. Parkkinen, A. Koivula, J. Vehmaanpera and J. Rouvinen, "Crystal structures of Melanocarpus albomyces cellobiohydrolase Ce17B in complex with cello-oligomers show high flexibility in the substrate binding", *Protein Sci.* **2008**, 17 (8), 1383-1394.

16. C.M. Payne, Y. Bomble, C.B. Taylor, C. McCabe, M.E. Himmel, M.F. Crowley and G.T. Beckham, "Multiple Functions of Aromatic-Carbohydrate Interactions in a Processive Cellulase Examined with Molecular Simulation", *J. Biol. Chem.* **2011**, 286 (47), 41028-41035.

17. J. Rouvinen, T. Bergfors, T. Teeri, J.K.C. Knowles and T.A. Jones, "3-dimensional structure of cellobiohydrolase II from *Trichoderma reesei*", *Science* **1990**, 249 (4967), 380-386.

18. D.M.F. van Aalten, B. Synstad, M.B. Brurberg, E. Hough, B.W. Riise, V.G.H. Eijsink and R.K. Wierenga, "Structure of a two-domain chitotriosidase from *Serratia marcescens* at 1.9 A resolution", *Proc. Natl. Acad. Sci.* **2000**, 97, 5842-5847.

19. A. Koivula, T. Kinnari, V. Harjunpaa, L. Ruohonen, A. Teleman, T. Drakenberg, J. Rouvinen, T.A. Jones and T.T. Teeri, "Tryptophan 272: an essential determinant of crystalline cellulose degradation by Trichoderma reesei cellobiohydrolase Cel6A", *FEBS Lett.* **1998**, 429 (3), 341-346.

20. S.J. Williams and G.J. Davies, "Protein-carbohydrate interactions: learning lessons from nature", *Trends Biotechnol.* **2001**, 19 (9), 356-362.

21. K. Igarashi, A. Koivula, M. Wada, S. Kimura, M. Penttila and M. Samejima, "High Speed Atomic Force Microscopy Visualizes Processive Movement of Trichoderma reesei Cellobiohydrolase I on Crystalline Cellulose", *J. Biol. Chem.* **2009**, 284 (52), 36186-36190.

22. G.T. Beckham, Y.J. Bomble, E.A. Bayer, M.E. Himmel and M.F. Crowley, "Applications of computational science for understanding enzymatic deconstruction of cellulose", *Current Opinion in Biotechnology* **2011**, 22 (2), 231-238.

23. S.P.S. Chundawat, G.T. Beckham, M.E. Himmel and B.E. Dale, "Deconstruction of Lignocellulosic Biomass to Fuels and Chemicals", *Annu. Rev. Chem. Biomol. Eng.* **2011**, 2, 6.1-6.25.

24. M.E. Himmel and E.A. Bayer, "Lignocellulose conversion to biofuels: current challenges, global perspectives", *Current Opinion in Biotechnology* **2009**, 20 (3), 316-317.

25. W. Humphrey, A. Dalke and K. Schulten, "VMD: Visual molecular dynamics", *J. Mol. Graph.* **1996**, 14 (1), 33-&.

26. F.B. Sheinerman and C.L. Brooks, "Calculations on folding of segment B1 of streptococcal protein G", *J. Mol. Biol.* **1998**, 278 (2), 439-456.

27. D. Case, T.A. Darden, I. T.E. Cheatham, C.L. Simmerling, R.E.D. J. Wang, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Goetz, I. Kolossvai, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko and P.A. Kollman, "AMBER 12," University of California: San Fransisco, **2012**; pp.

28. D.A. Case, T.E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr, A. Onufriev, C. Simmerling, B. Wang and R. Woods, "The Amber biomolecular simulation programs", *J Computational Chem* **2005**, 26, 1668-1688.

29. B.R. Brooks, C.L. Brooks, A.D. Mackerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, W. Yang, D.M. York and M. Karplus, "CHARMM: The Biomolecular Simulation Program", *J. Comput. Chem.* **2009**, 30 (10), 1545-1614.

30. J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale and K. Schulten, "Scalable molecular dynamics with NAMD", *J. Comput. Chem.* **2005**, 26 (16), 1781-1802.

31. A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins", *J. Phys. Chem. B* **1998**, 102 (18), 3586-3616.

32. A.D. Mackerell, M. Feig and C.L. Brooks, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations", *J. Comput. Chem.* **2004**, 25 (11), 1400-1415.

33. O. Guvench, E. Hatcher, R.M. Venable, R.W. Pastor and A.D. MacKerell, "CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses", *Journal of Chemical Theory and Computation* **2009**, 5 (9), 2353-2370.

34. U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee and L.G. Pedersen, "A Smooth Particle Mesh Ewald Method", *J. Chem. Phys.* **1995**, 103 (19), 8577-8593.

35. J.P. Ryckaert, G. Ciccotti and H.J.C. Berendsen, "Numerical-integration of cartesian equations of motion of a system with constraints - Molecular-dynamiccs of N-alkanes", *J. Comput. Phys.* **1977**, 23 (3), 327-341.

36. N. Deshpande, M.R. Wilkins, N. Packer and H. Nevalainen, "Protein glycosylation pathways in filamentous fungi", *Glycobiology* **2008**, 18 (8), 626-637.

37. T. Eriksson, I. Stals, A. Collen, F. Tjerneld, M. Claeyssens, H. Stalbrand and H. Brumer, "Heterogeneity of homologously expressed Hypocrea jecorina (Trichoderma reesei) Cel7B catalytic module", *Eur. J. Biochem.* **2004**, 271 (7), 1266-1276.

38. D.S. Frenkel, B. , *Understanding Molecular Simulation: From Algorithms to Applications*. 2nd ed ed.; Academic Press: San Diego, CA, **2002**.

39. A. Pohorille, Jarzynski, C., & Chipot, C., "Good Practices in Free-Energy Calculations", *J. Phys. Chem. B* **2010**, (114), 10235-10253.

40. P. Kollman, "Free energy calculations: Applications to chemical and biological phenomena", *Chem. Rev.* **1993**, 93, 2395-2417.

41. C.B. Taylor, M.F. Talib, C. McCabe, L. Bu, W.S. Adney, M.E. Himmel, M.F. Crowley and G.T. Beckham, "Computational investigation of glycosylation effects on a Family 1 carbohydrate-binding module", *J. Biol. Chem.* **2012**, (287), 3147-3155

42. T. Steinbrecher, D.L. Mobley and D.A. Case, "Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations", *J. Chem. Phys.* **2007**, 127 (21).

43. G. Sulzenbacher, M. Sch√ºlein and G.J. Davies, "Structure of the Endoglucanase I from Fusarium oxysporum:‚Äâ Native, Cellobiose, and 3,4-Epoxybutyl Œ≤-d-Cellobioside-Inhibited Forms, at 2.3 √Ö Resolution‚Ä†,,Ä°", *Biochemistry* **1997**, 36 (19), 5902-5911.

44. Y. Lin, J. Silvestre-Ryan, M.E. Himmel, M.F. Crowley, G.T. Beckham and J.-W. Chu, "Protein Allostery at the Solid‚ÄìLiquid Interface: Endoglucanase Attachment to Cellulose Affects Glucan Clenching in the Binding Cleft", *Journal of the American Chemical Society* **2011**, 133 (41), 16617-16624.

45. T.T. Teeri, A. Koivula, M. Linder, G. Wohlfahrt, C. Divne and T.A. Jones, "Trichoderma reesei cellobiohydrolases: why so efficient on crystalline cellulose?", *Biochemical Society Transactions* **1998**, 26 (2), 173-178.

46. L. Bu, M.R. Nimlos, M.R. Shirts, J. St√•hlberg, M.E. Himmel, M.F. Crowley and G.T. Beckham, "Product binding varies dramatically between processive and nonprocessive cellulase enzymes", *J. Biol. Chem.* **2012**.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

The work presented provides insight into the molecular-level action of glycoside hydrolase (GH) enzymes on cellulose substrates. Knowledge of thermodynamic properties and structure-function relationships required for optimal enzymatic efficiencies is key to improving the performance of cellulases used in enzymatic hydrolysis for the conversion of lignocellulosic biomass to fermentable sugars. Rational engineering of the GH enzymatic cocktails needed in production can ultimately lead to reduction in costs needed to make biofuels more commercially viable [1-3]. Many experimental and simulation studies to date have contributed to efforts to improve the understanding of the structure mechanistic action of GHs, specifically cellulases, on various cellulose substrates [1,2,4-44]. In this thesis, we add to this knowledge base by focusing on two GH7 model systems produced in the filamentous fungi *T. reesei*, an exoglucanase (Cel7A) and its synergistic endoglucanse (Cel7B). In Chapter 3, we demonstrated the feasibility of using molecular simulation as a tool in that can be used in the conscientious design of proteins and enzymes. In particular, we demonstrated that our predictions were accurate when compared to experimental results for amino acid muations in the CBM [27,28,39]. Additionally, in Chapters 3 and 4 we highlight a novel approach to improve binding affinity, and potentially activity [39] through the addition of *O*-glycosylation to the Family 1 CBM. Finally, Chapter 5 addresses important structural components for processive and non-processive action in the Cel7A and Cel7B CDs.

We have sought to understand the impact of various *O*-glycosylation motifs from different organisms on the structure and function of carbohydrate-binding proteins. By applying TI calculations and molecular dynamics simulation [45-48], we were able to show that CBM glycosylation could be a contributor to enzyme binding affinity, which may ultimately improve activity. Using the well-defined Cel7A Family 1 CBM bound to a Iβ crystalline cellulose substrate as our model system, we systematically added simple mannan or mannan-glucan *O*-glycoforms produced in *T. reesei, A. niger, A. awamori,* and yeasts, and to our knowledge, this study is the first to test the impacts of *O*-glycosylation on a Family 1 CBM. In Cel7A, the addition of simple mannan or mannan-glucan single, di-, and trisaccharides increase the hydrogen bonding potential and interaction energy of the CBM-glycan complex with the cellulose substrate to improve binding affinity by up to 3.8 kcal/mol over a non-glycosylated wild-type CBM. The linear mannan trisaccharide produced in *T. reesei, A. niger, A. awamori,* and yeasts (Figure 4.1, Table 4.1) is found to yield the largest improvement over the non-glycosylated wild-type, and in terms of glycan chemistry, the mannan-glucan structures examined here exhibit higher affinity than their mannan counterparts (e.g., S3MG as compared to S3M2 and S3M2-16), which we attribute to the observation that a mannan-glucan structure orients more parallel to the surface than a mannan structure, increasing H-bond potential and interaction energy. Our comparison of three analogous disaccharide systems (S3M2, S3M2-16, and S3MG) indicates that linkage patterns do not impact relative binding affinity and protein structure as much as the nature of the glycan moiety.

While experimental conformation of this study is needed, our TI calculations suggest that broadening the *O*-glycan patterns available in *T. reesei* via expression of genetic modifications or via chemical synthesis could be used to tune binding affinity and overall activity of Cel7A and other cellulase enzymes. In our simulations, loss of structural integrity of the protein backbone impacted normal binding and function. Thus an important finding of these two studies is that regardless of application, in order to optimize enhancements in glycoprotein engineering care must be taken to ensure glycosylation does not change the inherent structure of the protein.

Given that glycosylation is prevalent in all kingdoms of life as a post-translational modification and plays a functional role in biochemical interactions, this work highlights the importance of the careful design of both experimental and computational studies that may be affected by changes imparted by glycosylation. Glycosylation is important in both intra- and inter-cellular recognition [49-51], and this work shows that *O*-glycosylation can impact binding affinity and protein structure in the model Family 1 CBM. Host organism and growth media selection can have a dramatic effect on glycosylation patterns [49,51,52] and failure to account for changes imparted by glycans could alter the outcomes of protein engineering efforts. Both of the CBM studies demonstrate that computational studies are a valuable tool in rational approaches to glycoprotein engineering, allowing researchers to screen various mutations in binding studies, for example, and providing molecular-level details and general visualization of glycoproteins [40,53].

This study also used molecular dynamics simulation and TI calculations to highlight differing structure-function relationships in two synergistic GH7 enzymes'

CDs, Cel7A and Cel7B. Improvements in understanding of exo- and endoglucanases can aid in efforts to define processivity on a molecular basis. Our results suggest that in GH7 CDs the residues associated with sites directly upstream and downstream of the catalytically active site are important for ligand binding and potentially activity, but may not be associated with processivity. This hypothesis is supported by strict sequence conservation in the active site [23], and the nearly equal protein-ligand energetic and structural profiles calculated in simulation at these sites for both Cel7A and Cel7B wild-type systems; differing values should highlight protein areas with processive function. Our simulation data suggests that processivity and action on crystalline (via Cel7A) versus amorpohous (via Cel7B) substrates is tied to binding and interactions at the entrance and exits of the Cel7A tunnel and Cel7B cleft where different behaviors were noted. Cel7A's tunnel conformation leads to increased protein-ligand interactions over those in the cleft of Cel7B. With these weak interactions in Cel7B, single mutations in the cleft can have a dramatic impact on binding affinity and ligand fluctuation, indicating that even single point mutations could impact activity. Actual activity for various substrates and processivity changes between systems and due to mutations could not be calculated with these simulations; however, we were able to highlight functional differences in energetics and structure of processive and non-processive GH7s (Cel7A/B) on the molecular level. Clearly defining the roles of different GH Families in enzymatic cocktails during cellulosic hydrolysis may ultimately have valuable applicability in improving biochemical conversion steps in biofuel production, and this study adds to the current knowledge base.

6.2 Future Work

Despite the progress made, there are still many aspects of cellulase structure and mechanistic action that will require further study. Below we discuss recommendations for future studies, both experimental and computational, that could help further efforts in the field of cellulase engineering for biofuel production.

Validation of Findings on Glycosylation Impacts to Binding Affinity

While our TI simulations were able to reproduce experimental data on amino acid mutation, experimental validation of the glycosylation simulations is required. Study of the impacts glycosylation can have on protein structure and function presents an interesting challenge due to microheterogeneity, in that for any glycosylation site on a protein, an organism can produce a range of glycan structure, with the added complexity that external conditions of the organism can impact these patterns [49,51,52]. The presence or absence of transferases, categorized into families and subfamilies based on gene and protein sequence identity, allow identification of an organism's possible glycosylation patterns via genomic and proteomic analysis [49,54,55]. Linkage analysis is achieved using methylation analysis, wherein a stable ether-linked methyl is attached to free hydroxyl groups in the glycan, and then the glycan glycosidic bonds are broken revealing the linkage pattern [49]. Despite the difficulties of determining heterogeneity [56], site directed mutagenesis of the genes encoding for transferases can reveal potential impacts on protein activity and function [5,14,15,54,55,57,58]. Chemical synthesis of specific linkages can be achieved through a reverse of the methylation analysis where protective groups are selectively added and removed to allow for linkage at the exposed hydroxyl groups [49]. Current efforts at the National Renewable Energy Lab (NREL) and the University of Colorado at Boulder are underway to attempt to confirm or refute the findings in Chapters III and IV using solid-state synthesis (FMOC) of the CBM glycoprotein, with their proposed method shown in Figure 6.1 [59]. Once these CBM glycoproteins are synthesized, NMR structural determination studies and experimental binding affinity studies will be performed.

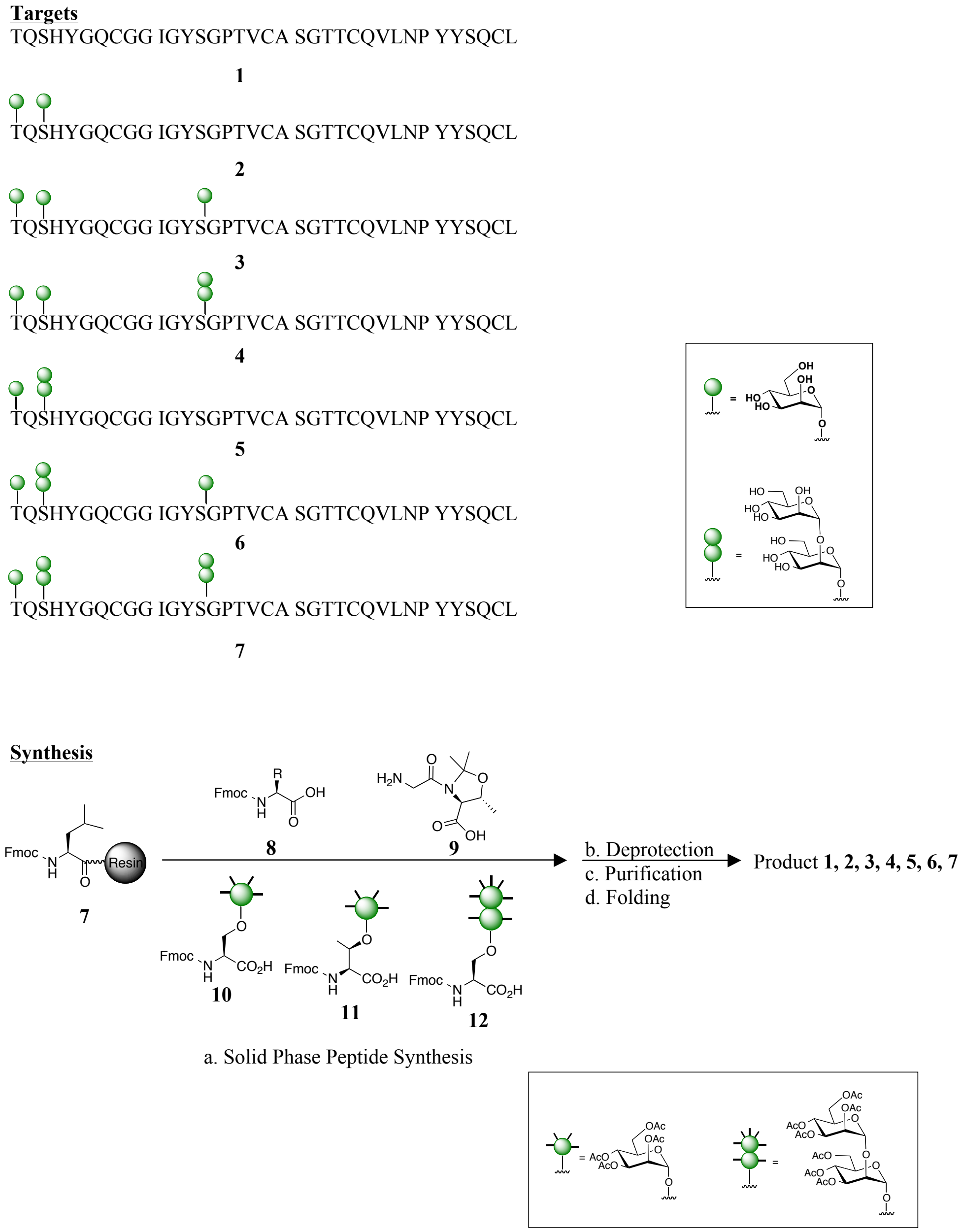

Targets
TQSHYGQCGG IGYSGPTVCA SGTTCQVLNP YYSQCL
1
TQSHYGQCGG IGYSGPTVCA SGTTCQVLNP YYSQCL
2
TQSHYGQCGG IGYSGPTVCA SGTTCQVLNP YYSQCL
3
TQSHYGQCGG IGYSGPTVCA SGTTCQVLNP YYSQCL
4
TQSHYGQCGG IGYSGPTVCA SGTTCQVLNP YYSQCL
5
TQSHYGQCGG IGYSGPTVCA SGTTCQVLNP YYSQCL
6
TQSHYGQCGG IGYSGPTVCA SGTTCQVLNP YYSQCL
7
Synthesis
Fmoc
Resin
7
8
9
10
11
12
b. Deprotection
c. Purification
d. Folding
Product 1, 2, 3, 4, 5, 6, 7
a. Solid Phase Peptide Synthesis

Figure 6.1: Solid-state Synthesis of the Cel7A Family 1 CBM with mannan patterns from Chapters III and IV on Thr-1, Ser-3, and Ser-14 shown in green. [59]

Future computational studies, as demonstrated here but perhaps on other GH Families, could also be useful tools for atomistic-level study of glycan chemistry to aid protein engineers in creating and altering glycoproteins for specific purposes [8,40,53,60,61]. As previously mentioned, glycosylation structure and function can be difficult to analyze due to the glycan complex's inherent flexibility and heterogeneity [62]. This study demonstrates that simulations can now be used effectively to measure binding affinity changes and observe and measure structural changes associated with glycosylation. Within the GH Families, there are currently 64 defined CBM Families with thousands of members, many of which demonstrate glycosylation [13,63]; the methods described in this thesis along with improving experimental tools [64-66] can be used to analyze the role and impact of glycosylation in binding in this vast catalogue of carbohydrate-active enzymes.

Processive and Non-processive Catalytic Domains in Other GH Families

Our work on two GH7 enzymes identified potential mutations and structural features that could impact binding affinity, processivity, and overall activity, but experimental binding and activity studies are needed to characterize the actual changes in activity on various substrates. Additionally, as noted in Chapter V these results may not be directly transferable to other GH Families, but the structural and energetic differences in the bound ligand for the exo- and endo-systems could be used as hallmarks for identifying processivity in other systems. The popular dogma extending from structural studies suggests that a CD with a tunnel conformation should be a processive exoglucanase while a CD with a cleft conformation should be a non-processive endoglucanase [67]; however some studies have shown examples to the contrary, with a GH6 CD (Cel6B) demonstrating endoglucanase behavior despite its tunnel formation [68], another GH6

CD (CBHA) losing exoglucanase activity with the removal of only a single tunnel loop [69], and then a switch from endo- to exoglucanase behavior with a single mutation in the GH chitinase (ChiB) cleft [21,70]. Thus, a more comprehensive goal might be to identify characteristics of processivity in GH Families based on energetic and structural profiles, as done in this study, rather than simply categorizing based on general structural forms. Development of overall molecular theories of processivity through continued computational and experimental studies could also improve understanding of how enzyme cocktails work synergistically to break down lignocellulose. Protein engineers can use findings to create new enzyme cocktails or identify efficient mixtures for use in biofuel production, lowering costs and increasing throughput.

6.3 References

1. M.E. Himmel and E.A. Bayer, "Lignocellulose conversion to biofuels: current challenges, global perspectives", *Current Opinion in Biotechnology* **2009**, 20 (3), 316-317.

2. M.E. Himmel, S.Y. Ding, D.K. Johnson, W.S. Adney, M.R. Nimlos, J.W. Brady and T.D. Foust, "Biomass recalcitrance: Engineering plants and enzymes for biofuels production", *Science* **2007**, 315 (5813), 804-807.

3. C. Schubert, "Can biofuels finally take center stage?", *Nat. Biotechnol.* **2006**, 24 (7), 777-784.

4. P.M. Abuja, M. Schmuck, I. Pilz, P. Tomme, M. Claeyssens and H. Esterbauer, "STRUCTURAL AND FUNCTIONAL DOMAINS OF CELLOBIOHYDROLASE-I FROM TRICHODERMA-REESEI - A SMALL-ANGLE X-RAY-SCATTERING STUDY OF THE INTACT ENZYME AND ITS CORE", *Eur. Biophys. J. Biophys. Lett.* **1988**, 15 (6), 339-342.

5. W.S. Adney, T. Jeoh, G.T. Beckham, Y.C. Chou, J.O. Baker, W. Michener, R. Brunecky and M.E. Himmel, "Probing the role of N-linked glycans in the stability and activity of fungal cellobiohydrolases by mutational analysis", *Cellulose* **2009**, 16 (4), 699-709.

6. B.K. Barr, Y.L. Hsieh, B. Ganem and D.B. Wilson, "Identification of two functionally different classes of exocellulases", *Biochemistry* **1996**, 35 (2), 586-592.

7. G.T. Beckham, Y.J. Bomble, E.A. Bayer, M.E. Himmel and M.F. Crowley, "Applications of computational science for understanding enzymatic deconstruction of cellulose", *Current Opinion in Biotechnology* **2011**, 22 (2), 231-238.

8. G.T. Beckham, Y.J. Bomble, J.F. Matthews, C.B. Taylor, M.G. Resch, J.M. Yarbrough, S.R. Decker, L.T. Bu, X.C. Zhao, C. McCabe, J. Wohlert, M. Bergenstrahle, J.W. Brady, W.S. Adney, M.E. Himmel and M.F. Crowley, "The O-Glycosylated Linker from the Trichoderma reesei Family 7 Cellulase Is a Flexible, Disordered Protein", *Biophys. J.* **2010**, 99 (11), 3773-3781.

9. G.T. Beckham, Z. Dai, J.F. Matthews, M. Momany, C.M. Payne, W.S. Adney, S.E. Baker and M.E. Himmel, "Harnessing glycosylation to improve cellulase activity", *Current Opinion in Biotechnology* **2012**, (0).

10. G.T. Beckham, J.F. Matthews, Y.J. Bomble, L.T. Bu, W.S. Adney, M.E. Himmel, M.R. Nimlos and M.F. Crowley, "Identification of Amino Acids Responsible for Processivity in a Family 1 Carbohydrate-Binding Module from a Fungal Cellulase", *J. Phys. Chem. B* **2009**, 114 (3), 1447-1453.

11. G.T. Beckham, J.F. Matthews, B. Peters, Y.J. Bomble, M.E. Himmel and M.F. Crowley, "Molecular-Level Origins of Biomass Recalcitrance: Decrystallization Free Energies for Four Common Cellulose Polymorphs", *Journal of Physical Chemistry B* **2011**, 115 (14), 4118-4127.

12. G.T. Beckham, C.M. Payne, D.W. Sammond, C.B. Taylor, L.T. Bu, M.R. Nimlos, C.M. McCabe, M.E. Himmel, W.S. Adney and M.F. Crowley, "Elucidating protein-carbohydrate interactions in cellulase enzymes with molecular simulation", *Abstracts of Papers of the American Chemical Society* **2011**, 242.

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

14. A.B. Boraston, L.E. Sandercock, R.A.J. Warren and D.G. Kilburn, "O-glycosylation of a recombinant carbohydrate-binding module mutant secreted by Pichia pastoris", *J. Mol. Microbiol. Biotechnol.* **2003**, 5 (1), 29-36.

15. A.B. Boraston, R.A.J. Warren and D.G. Kilburn, "Glycosylation by Pichia pastoris decreases the affinity of a family 2a carbohydrate-binding module from Cellulomonas fimi: a functional and mutational analysis", *Biochem. J.* **2001**, 358, 423-430.

16. S.P.S. Chundawat, G.T. Beckham, M.E. Himmel and B.E. Dale, "Deconstruction of Lignocellulosic Biomass to Fuels and Chemicals", *Annu. Rev. Chem. Biomol. Eng.* **2011**, 2, 6.1-6.25.

17. M. Claeyssens, H. Vantilbeurgh, J.P. Kamerling, J. Berg, M. Vrsanska and P. Biely, "STUDIES OF THE CELLULOLYTIC SYSTEM OF THE FILAMENTOUS FUNGUS TRICHODERMA-REESEI QM-9414 - SUBSTRATE-SPECIFICITY AND TRANSFER ACTIVITY OF ENDOGLUCANASE-I", *Biochem. J.* **1990**, 270 (1), 251-256.

18. C. Divne, J. Stahlberg, T. Reinikainen, L. Ruohonen, G. Pettersson, J.K.C. Knowles, T.T. Teeri and T.A. Jones, "The 3-dimensional crystal-structure of the catalytic core of Cellobiohydrolase-I from *Trichodermal reesei*", *Science* **1994**, 265 (5171), 524-528.

19. C. Divne, J. Stahlberg, T.T. Teeri and T.A. Jones, "High-resolution crystal structures reveal how a cellulose chain is bound in the 50 angstrom long tunnel of cellobiohydrolase I from Trichoderma reesei", *J. Mol. Biol.* **1998**, 275 (2), 309-325.

20. M.J. Harrison, A.S. Nouwens, D.R. Jardine, N.E. Zachara, A.A. Gooley, H. Nevalainen and N.H. Packer, "Modified glycosylation of cellobiohydrolase I from a high cellulase-producing mutant strain of Trichoderma reesei", *Eur. J. Biochem.* **1998**, 256 (1), 119-127.

21. S.J. Horn, P. Sikorski, J.B. Cederkvist, G. Vaaje-Kolstad, M. Sorlie, B. Synstad, G. Vriend, K.M. Varum and V.G.H. Eijsink, "Costs and benefits of processivity in enzymatic degradation of recalcitrant polysaccharides", *Proc. Natl. Acad. Sci.* **2006**, 103 (48), 18089-18094.

22. K. Igarashi, A. Koivula, M. Wada, S. Kimura, M. Penttila and M. Samejima, "High Speed Atomic Force Microscopy Visualizes Processive Movement of Trichoderma reesei Cellobiohydrolase I on Crystalline Cellulose", *J. Biol. Chem.* **2009**, 284 (52), 36186-36190.

23. G.J. Kleywegt, J.Y. Zou, C. Divne, G.J. Davies, I. Sinning, J. Stahlberg, T. Reinikainen, M. Srisodsuk, T.T. Teeri and T.A. Jones, "The crystal structure of the catalytic core domain of endoglucanase I from Trichoderma reesei at 3.6 angstrom resolution, and a comparison with related enzymes", *J. Mol. Biol.* **1997**, 272 (3), 383-397.

24. A. Koivula, T. Kinnari, V. Harjunpaa, L. Ruohonen, A. Teleman, T. Drakenberg, J. Rouvinen, T.A. Jones and T.T. Teeri, "Tryptophan 272: an essential determinant of crystalline cellulose degradation by Trichoderma reesei cellobiohydrolase Cel6A", *FEBS Lett.* **1998**, 429 (3), 341-346.

25. P.J. Kraulis, G.M. Clore, M. Nilges, T.A. Jones, G. Pettersson, J. Knowles and A.M. Gronenborn, "Determination of the 3-dimensional solution structure of the C-terminal domani of Cellobiohydrolase-I from Trichoderma reesei - A study using Nuclear Magnetic Resonance and hybrid distance geometry dynamical simulated annealing", *Biochemistry* **1989**, 28 (18), 7241-7257.

26. M. Kurasin and P. Valjamae, "Processivity of Cellobiohydrolases Is Limited by the Substrate", *J. Biol. Chem.* **2011**, 286 (1), 169-177.

27. M. Linder, G. Lindeberg, T. Reinikainen, T.T. Teeri and G. Pettersson, "The difference in affinity between 2 fungal cellulose-binding domains is dominated by a single amino-acid substitution", *FEBS Lett.* **1995**, 372 (1), 96-98.

28. M. Linder, M.L. Mattinen, M. Kontteli, G. Lindeberg, J. Stahlberg, T. Drakenberg, T. Reinikainen, G. Pettersson and A. Annila, "Identification of functionally important amino-acids in the cellulose-binding domain of *Trichoderma reesei* Cellobiohydrolase I", *Protein Sci.* **1995**, 4 (6), 1056-1064.

29. L.F. Mackenzie, G. Sulzenbacher, C. Divne, T.A. Jones, H.F. Woldike, M. Schulein, S.G. Withers and G.J. Davies, "Crystal structure of the family 7 endoglucanase I (Cel7B) from Humicola insolens at 2.2 angstrom resolution and identification of the catalytic nucleophile by trapping of the covalent glycosyl-enzyme intermediate", *Biochem. J.* **1998**, 335, 409-416.

30. M.L. Mattinen, M. Kontteli, J. Kerovuo, M. Linder, A. Annila, G. Lindeberg, T. Reinikainen and T. Drakenberg, "Three-dimensional structures of three engineered

cellulose-binding domains of cellobiohydrolase I from Trichoderma reesei", *Protein Sci.* **1997**, 6 (2), 294-303.

31. M.R. Nimlos, J.F. Matthews, M.F. Crowley, R.C. Walker, G. Chukkapalli, J.V. Brady, W.S. Adney, J.M. Clearyl, L.H. Zhong and M.E. Himmel, "Molecular modeling suggests induced fit of Family I carbohydrate-binding modules with a broken-chain cellulose surface", *Protein Eng. Des. Sel.* **2007**, 20 (4), 179-187.

32. C.M. Payne, Y. Bomble, C.B. Taylor, C. McCabe, M.E. Himmel, M.F. Crowley and G.T. Beckham, "Multiple Functions of Aromatic-Carbohydrate Interactions in a Processive Cellulase Examined with Molecular Simulation", *J. Biol. Chem.* **2011**, 286 (47), 41028-41035.

33. M. Penttila, P. Lehtovaara, H. Nevalainen, R. Bhikhabhai and J. Knowles, "HOMOLOGY BETWEEN CELLULASE GENES OF TRICHODERMA-REESEI - COMPLETE NUCLEOTIDE-SEQUENCE OF THE ENDOGLUCANASE-I GENE", *Gene* **1986**, 45 (3), 253-263.

34. V. Receveur, M. Czjzek, M. Schulein, P. Panine and B. Henrissat, "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), 40887-40892.

35. T. Reinikainen, M. Srisodsuk, A. Jones and T.T. Teeri, "Enzymatic hydrolysis of crystalline cellulose by *Trichoderma reesei* cellobiohydrolase I", *Protein Eng* **1993**, 6, 49-49.

36. J. Rouvinen, T. Bergfors, T. Teeri, J.K.C. Knowles and T.A. Jones, "3-dimensional structure of cellobiohydrolase II from *Trichoderma reesei*", *Science* **1990**, 249 (4967), 380-386.

37. M. Srisodsuk, J. Lehtio, M. Linder, E. MargollesClark, T. Reinikainen and T.T. Teeri, "Trichoderma reesei cellobiohydrolase I with an endoglucanase cellulose-binding domain: action on bacterial microcrystalline cellulose", *J. Biotechnol.* **1997**, 57 (1-3), 49-57.

38. J. Stahlberg, C. Divne, A. Koivula, K. Piens, M. Claeyssens, T.T. Teeri and T.A. Jones, "Activity studies and crystal structures of catalytically deficient mutants of cellobiohydrolase I from Trichoderma reesei", *J. Mol. Biol.* **1996**, 264 (2), 337-349.

39. S. Takashima, M. Ohno, M. Hidaka, A. Nakamura and H. Masaki, "Correlation between cellulose binding and activity of cellulose-binding domain mutants of Humicola grisea cellobiohydrolase 1", *FEBS Lett.* **2007**, 581 (30), 5891-5896.

40. C.B. Taylor, M.F. Talib, C. McCabe, L. Bu, W.S. Adney, M.E. Himmel, M.F. Crowley and G.T. Beckham, "Computational investigation of glycosylation effects on a Family 1 carbohydrate-binding module", *J. Biol. Chem.* **2012**, (287), 3147-3155

41. I. von Ossowski, J. Stahlberg, A. Koivula, K. Piens, D. Becker, H. Boer, R. Harle, M. Harris, C. Divne, S. Mahdi, Y.X. Zhao, H. Driguez, M. Claeyssens, M.L. Sinnott and T.T. Teeri, "Engineering the exo-loop of *Trichoderma reesei* cellobiohydrolase, Cel7A. A comparison with *Phanerochaete chrysosporium* Cel7D", *J. Mol. Biol.* **2003**, 333 (4), 817-829.

42. D.B. Wilson, "Cellulases and biofuels", *Curr. Opin. Biotech.* **2009**, 20, 295-299.

43. Y.H.P. Zhang and L.R. Lynd, "A functionally based model for hydrolysis of cellulose by fungal cellulase", *Biotechnol. Bioeng.* **2006**, 94 (5), 888-898.

44. X. Zhao, T.R. Rignall, C. McCabe, W.S. Adney and M.E. Himmel, "Molecular simulation evidence for processive motion of Trichoderma reesei Cel7A during cellulose depolymerization", *Chem. Phys. Lett.* **2008**, 460 (1-3), 284-288.

45. P. Kollman, "Free energy calculations: Applications to chemical and biological phenomena", *Chem. Rev.* **1993**, 93, 2395-2417.

46. A. Pohorille, Jarzynski, C., & Chipot, C., "Good Practices in Free-Energy Calculations", *J. Phys. Chem. B* **2010**, (114), 10235-10253.

47. D.S. Frenkel, B. , *Understanding Molecular Simulation: From Algorithms to Applications*. 2nd ed ed.; Academic Press: San Diego, CA, **2002**.

48. M. Karplus and G.A. Petsko, "MOLECULAR-DYNAMICS SIMULATIONS IN BIOLOGY", *Nature* **1990**, 347 (6294), 631-639.

49. A. Varki, *Essentials of Glycobiology*. Cold Springs Harbor Laboratory Press: Cold Springs Harbor, NY, USA, **2009**.

50. D. Spillmann and M.M. Burger, "Carbohydrate-carbohydrate interactions in adhesion", *J. Cell. Biochem.* **1996**, 61 (4), 562-568.

51. K. Drickamer and M.E. Taylor, "Evolving views of protein glycosylation", *Trends Biochem.Sci.* **1998**, 23 (9), 321-324.

52. G. Lauc and V. Zoldos, "Protein glycosylation-an evolutionary crossroad between genes and environment", *Mol. Biosyst.* **2010**, 6 (12), 2373-2379.

53. Y. Mazola, G. Chinea and A. Musacchio, "Integrating Bioinformatics Tools to Handle Glycosylation", *PLoS Comput. Biol.* **2011**, 7 (12), 7.

54. N. Deshpande, M.R. Wilkins, N. Packer and H. Nevalainen, "Protein glycosylation pathways in filamentous fungi", *Glycobiology* **2008**, 18 (8), 626-637.

55. M. Goto, "Protein O-glycosylation in fungi: Diverse structures and multiple functions", *Biosci. Biotechnol. Biochem.* **2007**, 71 (6), 1415-1427.

56. M.N. Christiansen, D. Kolarich, H. Nevalainen, N.H. Packer and P.H. Jensen, "Challenges of Determining O-Glycopeptide Heterogeneity: A Fungal Glucanase Model System", *Anal. Chem.* **2010**, 82 (9), 3500-3509.

57. T. Jeoh, W. Michener, M.E. Himmel, S.R. Decker and W.S. Adney, "Implications of cellobiohydrolase glycosylation for use in biomass conversion", *Biotechnol. Biofuels* **2008**, 1, 12.

58. K. De Pourcq, K. De Schutter and N. Callewaert, "Engineering of glycosylation in yeast and other fungi: current state and perspectives", *Applied Microbiology and Biotechnology* **2010**, 87 (5), 1617-1631.

59. G.T. Beckham and Z. Tan, CBM Synthesis.

60. G.T. Beckham, Y.J. Bomble, E.A. Bayer, M.E. Himmel and M.F. Crowley, "Harnessing glycosylation to improve cellulase activity", *Curr. Opin. Biotechnol.* **2012**, - 22 (- 2), - 238.

61. G.T. Beckham, Bomble, Y. J., Bayer, E.A., Himmel, M.E., and Crowley, M. F., "Applications of computational science for understanding enzymatic deconstruction of cellulose", *Curr. Opin. Biotechnol.* **2011**, 22, 1-8.

62. G.W. Hart and R.J. Copeland, "Glycomics Hits the Big Time", *Cell* **2010**, 143 (5), 672-676.

63. B.L. Cantarel, P.M. Coutinho, C. Rancurel, T. Bernard, V. Lombard and B. Henrissat, "The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics", *Nucleic Acids Res.* **2009**, 37, D233-D238.

64. N.J. Agard and C.R. Bertozzi, "Chemical Approaches To Perturb, Profile, and Perceive Glycans", *Accounts Chem. Res.* **2009**, 42 (6), 788-797.

65. J.E. Hudak, H.H. Yu and C.R. Bertozzi, "Protein Glycoengineering Enabled by the Versatile Synthesis of Aminooxy Glycans and the Genetically Encoded Aldehyde Tag", *Journal of the American Chemical Society* **2011**, 133 (40), 16127-16135.

66. J.A. Prescher and C.R. Bertozzi, "Chemical technologies for probing glycans", *Cell* **2006**, 126 (5), 851-854.

67. L.R. Lynd, P.J. Weimer, W.H. van Zyl and I.S. Pretorius, "Microbial cellulose utilization: Fundamentals and biotechnology", *Microbiol. Mol. Biol. Rev.* **2002**, 66 (3), 506-+.

68. G.J. Davies, A.M. Brzozowski, M. Dauter, A. Varrot and M. Schulein, "Structure and function of Humicola insolens family 6 cellulases: structure of the endoglucanase, Cel6B, at 1.6 angstrom resolution", *Biochem. J.* **2000**, 348, 201-207.

69. A. Meinke, H.G. Damude, P. Tomme, E. Kwan, D.G. Kilburn, R.C. Miller, R.A.J. Warren and N.R. Gilkes, "ENHANCEMENT OF THE ENDO-BETA-1,4-GLUCANASE ACTIVITY OF AN EXOCELLOBIOHYDROLASE BY DELETION OF A SURFACE LOOP", *J. Biol. Chem.* **1995**, 270 (9), 4383-4386.

70. V.G.H. Eijsink, G. Vaaje-Kolstad, K.M. Varum and S.J. Horn, "Towards new enzymes for biofuels: lessons from chitinase research", *Trends Biotechnol.* **2008**, 26 (5), 228-235