IMPACT OF A CONSERVED TYROSINE RESIDUE ON BINDING OF FAMILY 1 CARBOHYDRATE BINDING MODULES TO CELLULOSE ALLOMORPHS

By

AKASH DAGIA

A thesis submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfilment of the requirements

For the degree of

Master of Science

Graduate Program in Chemical and Biochemical Engineering

Written under the direction of

Dr. Shishir P. S. Chundawat

And approved by

_____

_____

_____

New Brunswick, New Jersey

[October, 2018]

ABSTRACT OF THE THESIS

Impact of a conserved tyrosine residue on binding of Family 1 carbohydrate binding modules to cellulose allomorphs

By Akash Dagia

Thesis Director:

Dr. Shishir P. S. Chundawat

The recalcitrance of cellulose, coupled with non-productive binding of cellulases, is considered to be a major bottleneck in the deconstruction of biomass into biofuels (Jeoh, T., Cardona, M. J., Karuna, N., Mudinoor, A. R. & Nill, J. Mechanistic kinetic models of enzymatic cellulose hydrolysis—A review. Biotechnol. Bioeng. 114, 1369–1385 (2017). doi:10.1002/bit.26277). Past research to address the recalcitrance of cellulose, has led the development of pre-treatment technologies like the Extractive Ammonia process (Sousa, L. et al. Next-generation ammonia pretreatment enhances cellulosic biofuel production. Energy Environ. Sci. 9, 1215–1223 (2016). doi:10.1039/c5ee03051j), which can modify the ultrastructure of native crystalline cellulose-I to cellulose-III allomorph (Chundawat, S. P. S. et al. Restructuring the crystalline cellulose hydrogen bond network enhances its depolymerization rate. *J. Am. Chem. Soc*. 133, 11163–11174 (2011). doi:10.1021/ja2011115). Surprisingly, it was found previsouly that some full-length cellulases bind with lower apparent affinity to crystalline cellulose-III, while the enzymatic hydrolysis rate for this modified cellulose-III allomorph was between two to five-folds higher by fungal cellulase enzyme cocktails (Gao, D. et al. Increased enzyme binding to substrate is not necessary for more efficient cellulose hydrolysis. *Proc. Natl. Acad. Sci*. 110, 10922–10927 (2013). doi

10.1073/pnas.1213426110). Our results attest to better understanding the role of carbohydrate-binding modules (CBMs) on the reduced binding affinity of full-length fungal cellulases seen towards cellulose-III. Here, we closely explore the role of key amino acid residues of a Family 1 CBM that are likely to impact protein binding interactions with the surface of cellulose-III. Single-site saturation mutagenesis libraries were generated at such key positions to better understand impact on CBM-1 adsorption to both cellulose allomorphs. We report results here relating to the: (i) Expression and purification of green fluorescent protein (GFP) tagged wild-type CBM-1 protein construct and its single-site saturation mutagenesis protein library for subsequent binding/structural characterization, (ii) Effect of pH and salt concentration on the apparent binding affinity or partition coefficient of wild-type CBM-1 protein and its saturation mutagenesis mutants library towards cellulose allomorphs. We also report regression correlations between the experimentally measured binding parameters and various *in silico* sequence/structural modeling derived Rosetta software estimated parameters that are indicative of protein function. We also discuss a roadmap for future studies that include analysis of full scale binding isotherm of wild type and some of the key mutants to native and pre-treated cellulose, cloning and testing of overall cellulase activity with mutant CBM1-cellulases to understand the correlation between CBM-mediated overall binding affinity and cellulolytic activity on different cellulose allomorphs. This work has important implications for creation of more efficient cellulase enzymes, which can pave the way towards sustainable production of biofuels from ammonia-pretreated lignocellulosic biomass.

ACKNOWLEDGEMENTS

I would like to extend my gratitude towards my thesis advisor Dr. Shishir Chundawat (Assistant Professor, Department of Chemical and Biochemical Engineering, Rutgers University) for giving me the opportunity to carry out this research project. His constant guidance, periodic meetings, and updates have helped me keep the project on track. I would also like to thank our collaborators Dr. Brian Fox (University of Wisconsin – Madison) for providing the original pEC-GFP-CBM plasmids and Dr. Leonardo Sousa (Michigan State University) for providing cellulose-III used in this work. I am extremely grateful towards the National Science Foundation (NSF) for partially supporting this research work via the following two NSF awards (#1604421 and #1236120).

I would also like to thank the Chemical and Biochemical Engineering department and the whole staff at Rutgers University for their support. I cannot thank enough my fellow laboratory members for bearing with me, especially Bhargava Nemmaru and Vibha Narayanan for supporting me through thick and thin. They both have been my constant source of support for nearly everything. I would like to thank Patrick as well for running the Rosetta simulations for me and Nicholas for carrying out some additional supplemental experiments.

Last but not the least, I would like to thank my parents who believed in me and have done and sacrificed so much to make sure that I come to the United States of America to pursue my Master's Degree, I will be forever grateful towards them.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Plant biomass is a renewable feedstock to produce biofuels

In the 21st century, there has been increased emphasis worldwide to explore viable renewable alternatives to meet our growing energy demand. Cellulosic biomass is recognised as one of the promising feedstocks to produce biofuels due to its renewable source and plentiful abundance. Plant cell walls are primarily comprised of sugar polymers like cellulose, hemicellulose, and aromatic polymers like lignin. Here, β-1-4-glucose polymers or cellulose makeup the most dominant components of plant cell walls. Cellulose polymers can be hydrolysed using hydrolytic enzymes or acidic catalysts to produce glucose monomers or dimers (like cellobiose) that can be then fermented by yeast or bacteria to produce biofuels like ethanol. The overall process schematic using cellulase enzymes to convert cellulosic biomass like corn stover or switchgrass to fuels/chemicals is outlined in Figure 1 below.

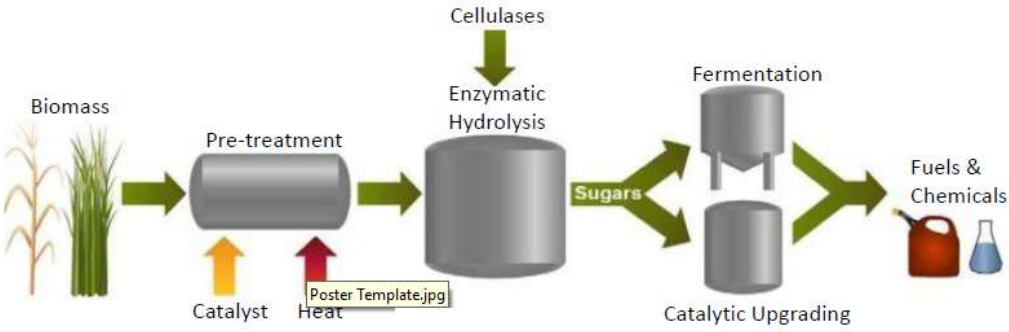


Figure 1: Outline of biochemical processing of cellulosic feed stock into fermentable sugars that are upgraded into fuels or chemicals. Figure taken from Payne, C. et al [23].

One way of hydrolysing the plant-based sugar polymers is by using concentrated acid based catalytic treatments to decrystallize cellulose and simultaneously hydrolyze the glycosidic bonds to produce soluble sugars like glucose. However, the soluble sugar yields achieved are low due to the extensive degradation of polymer via side-reactions like

dehydration of glucose instead of selective hydrolysis of glycosidic bonds in cellulose. Naturally occurring cellulolytic microbes that secrete various enzymes like cellulases can give also near theoretical yields during hydrolysis of cellulosic biomass. However, cellulosic microcrystalline fibrils are not easily accessible to these enzymes which results in slow hydrolysis rates. Therefore, with the help of mild thermochemical pre-treatments of biomass, the subsequent enzymatic hydrolysis rate for pre-treated biomass can be improved significantly [8]. Various thermochemical pre-treatment methods have been developed over the years such as steam explosion, Ammonia Fiber Expansion (AFEX), dilute acid, and ionic liquid based pre-treatments [32]. Most aqueous based thermochemical pretreatments can increase the overall accessibility of embedded sugar polymers in plant cell walls, but do not significantly decrystallize native cellulose microfibrils [7].

1.2 Cellulose ultrastructure

Cellulose is the most abundant organic polymer found in nature. It is a polysaccharide consisting of hundreds of β-D glucose monomers linked together via a β-1-4 glycosidic bonds that result in the formation of a single cellulose polymer chain. These individual chains self-associate via strong inter/intra-chain non-covalent bonds to form elementary semi-crystalline micro fibrils component of 24 or 36 individual chains in plant cell walls [29]. The hydrolysis of the covalent bonds are a minor hurdle to bioconversion, but the first major hurdle for the enzymatic hydrolysis process is the disruption of the non-covalent hydrogen bonding and stacking interactions between the polymer chains within each cellulose micro fibril that limits overall enzyme accessibility/activity [8, 11].

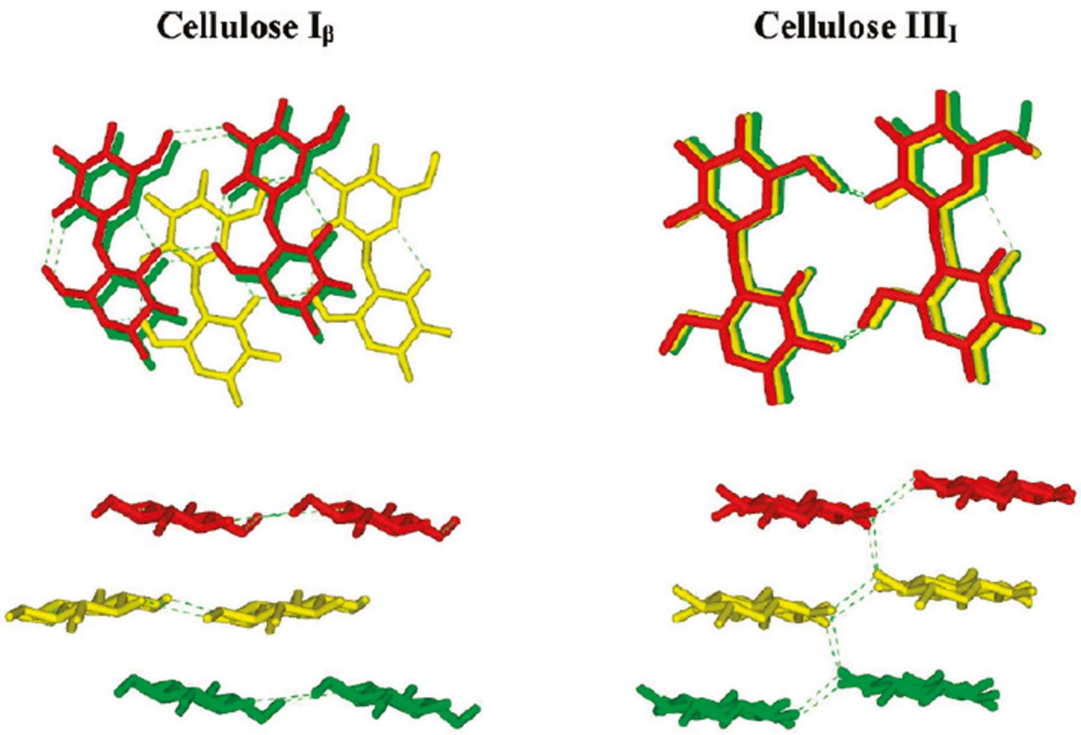


Figure 2: Side and top view of unit cells from published crystal structures of cellulose allomorph viz. cellulose-Iβ and cellulose-III. Dotted lines depict the intra-molecular and inter-molecular cellulose polymer chains hydrogen bonding interactions. Figure taken from Parthasarathi, R. et al [22]

There are two naturally occurring allomorphs of cellulose viz. cellulose-Iβ or cellulose-Iα with distinct crystal structures depending on the original source [24]. Hydrogen bonding pattern and interlayer stacking interactions of the chains are distinctly different for the two natural cellulose allomorphs. There are no major inter-sheet hydrogen bonding in either case, but a highly coupled dynamic network of hydrogen-bonds are observed between and within individual cellulose chains. Natural cellulose allomorphs are produced in plant, algal and bacterial cell walls via cellulose synthase enzyme complexes but other unnatural cellulose allomorphs can be made in laboratories using various thermochemical pre-treatment methods. Cellulose-III, is one such unnatural allomorph of cellulose that can be formed by treating native cellulose-I crystals with anhydrous liquid ammonia [8]. Cellulose-III crystal

has a staggered layer surface of cellulose chains with hydrogen bonding within the sheet as well as between the sheets, as shown in Figure 2. Initially, Ammonia Fiber EXpansion process (AFEX) with low water loading [8] was used to convert cellulose-I to cellulose-III. The Extractive Ammonia (EA) pre-treatment process is a recently developed efficient pre-treatment process that can also partially extract lignin while producing cellulose-III allomorph during pretreatment with anhydrous liquid ammonia based solvent mixtures [30]. EA pretreatment uses low moisture levels, unlike the conventional AFEX process, that facilitates interconversion of cellulose-I to cellulose-III allomorph. The pretreatment process with ammonia increases the hydrogen bonding network between the layers/sheets of individual chains and decreases the hydrogen bonding network within the layer/sheets of chains, thus increasing the accessibility of cellulose chains on the surface of cellulose III towards some specific families of cellulases [8]. A higher enzymatic activity was also observed on cellulose after pretreatment and thus it was expected to have higher binding for these cellulases towards pre-treated cellulose-III versus native cellulose-I. However, while the enzymatic hydrolysis rate was significantly higher for pretreated cellulose using fungal cellulase cocktails, the overall binding (as characterized by partition coefficient measurements for full-length Trichoderma cellulases) of all key cellulase enzymes towards cellulose-III was always lower than that of cellulose-I [11]. According to a molecular dynamics (MD) simulation study conducted in collaboration with Dr. Gnanakaran's group at Los Alamos National Laboratory [8] and recent experimental work reported in Ms. Vibha Narayanan's Masters thesis [21], the reduced binding observed for Trichoderma cellulases is likely caused due to lower affinity of enzyme to the relatively more hydrophilic surface of cellulose-III as compared to cellulose-I. Additional unpublished MD simulations of CBM1 bound to the surface of cellulose-III has further revealed that the Y5 position of CBM1 domain of the Trichoderma cellulase is likely encountering steric clashes with the stepped

surface of cellulose-III that ultimately is the underlying cause of reduced cellulase binding (unpublished results from Dr. Gnankaran/Dr. Chundawat). These observations encouraged us to more closely investigate the binding mechanism of CBM1 wild type protein domains and the role of Y5 position on binding interactions with cellulose-I and cellulose-III through this research project.

1.3 *Trichoderma reesei* and its extracellularly secreted TrCel7A exo-cellulase

Enzymatic hydrolysis of cellulosic biomass using enzymes has been extensively studied for the past several decades and many microorganisms have been found to evolve complex enzymatic machinery to break down plant biomass into a carbon or energy source. One such well-studied filamentous fungus, *Trichoderma reesei,* was accidently discovered by the Natick Army Research Lab in the 1940s. This fungus was found to secrete a highly synergistic cocktail of Carbohydrate Active enZymes (CAZymes) that deconstruct lignocellulose biomass into fermentable sugars like glucose and xylose. Enzymes that degrade cellulose specifically are called cellulases and are predominantly Glycosyl Hydrolases (GH) family of enzymes. GH can be classified into two groups as endoglucanases and exoglucanases (or cellobiohydrolases). Endoglucanase are non-processive enzymes that cleave the cellulose polymer mid-chain where it is more easily accessible, while exoglucanases are processive enzymes (like CBH1) that specifically bind to either the reducing or non-reducing chain-ends of cellulose before processively depolymerizing it into soluble sugars like cellobiose or glucose [5]. CBH1 (recently renamed as TrCel7A) exoglucanase is a major cellulase enzyme secreted by *T. reesei* that consists of bi-functional domain organization viz. a Carbohydrate Binding Module (CBM) and a large Catalytic Domain (CD) both of them linked by a short O-glycosylated peptide linker (Figure 3).

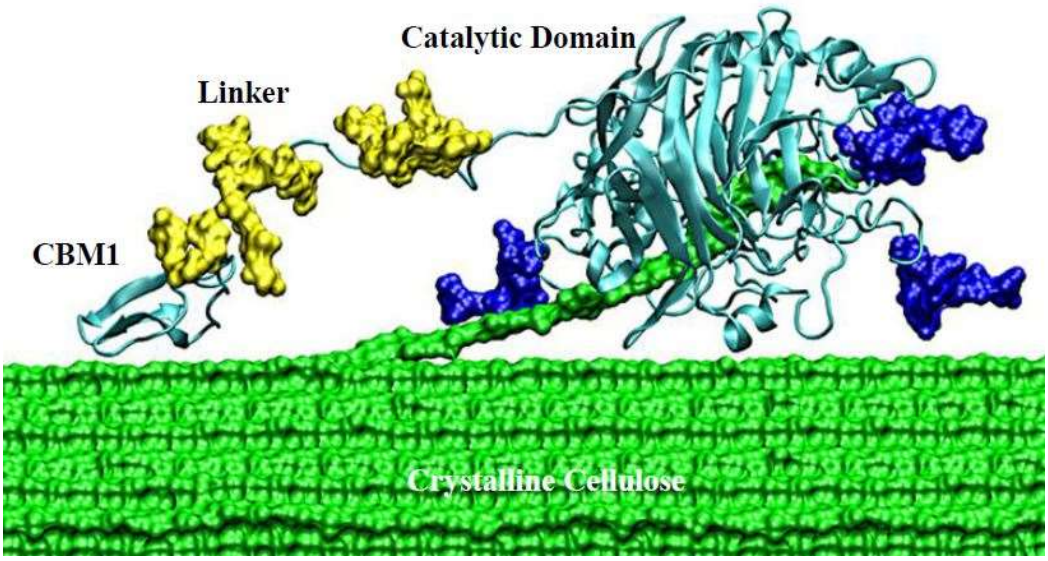


Figure 3: TrCel7A acting on crystalline cellulose surface (side view) with protein domains highlighted. A single cellulose chain is extracted from the cellulose microfibril surface and fed into the catalytic tunnel active site. TrCel7A has a Catalytic Domain (CD), a small family 1 Carbohydrate Binding Module (CBM1) linked via O-glycosylated linker domains (in yellow). N-linked protein glycosylation in blue is shown here as well. Figure taken from Chundawat, S. P. S. et al [7].

CBMs are currently classified into 84 distinct families currently based on structural homology and binding function (www.cazy.org). These were renamed from Cellulose Binding Domain (CBDs) to CBMs in the early 2000's [22], since initially it was thought that these domains bind to cellulose alone. However, later discoveries showed that other families of similarly folded protein domains could bind to carbohydrates other than cellulose as well. CBMs are a contiguous amino acid sequence of a larger protein sequence that can fold into distinct modules. CBMs are linked to the other part of the protein sequence, namely a Catalytic Domain (CD), via a peptide linker domain. Typically, CBMs have preferential binding affinity to certain soluble/insoluble carbohydrates or carbohydrate-polymers towards which the CD has evolved to have preferential catalytic activity. The main role of the CBM is mostly to bind to the carbohydrate and direct the CD machinery towards the substrate but the CBM by itself is devoid of any hydrolytic activity. Upon binding to the substrate, these

CBMs are thought to not undergo any major structural changes but due to the unavailability of the crystal structures of true enzyme-ligand complexes (for typically insoluble substrates like cellulose), it is very difficult to fully validate this hypothesis. The topography of CBM binding site is generally found to complement the shape of the target carbohydrate substrate via presence of specific amino acid residues that facilitate binding. CBMs are thus classified into 3 main superstructure fold types based on their 3D structure and functional similarities viz. Type A, Type B and Type C. Type A CBMs binds with geometrically planar surface of sugars or polymer surfaces like the hydrophobic crystalline face of cellulose-I which forms a flat surface. Type B CBMs bind to highly amorphous glycan chains that have 4 or more monosaccharides binding within a cleft or sandwich type CBM binding motif. Type C CBMs bind to the termini of short carbohydrates since binding sites have small pocket type motifs that can accommodate only short chains of sugar (www.cazypedia.org).

T. reesei has a family 1 Carbohydrate Binding Module (CBM1) classified as Type A CBM appended to most secreted cellulases like TrCel7A. The CBM is appended to the CD via a linker peptide, which may extensively glycosylated as well [6]. The CD and CBM can both bind to the cellulose substrate individually but the binding of the CD alone is significantly lesser than the CBM-appended full-length enzyme. Though, the catalytic activity of the full-length enzyme (CBM-CD) and CD alone are on par with each other on soluble or highly amorphous substrates, the activity of CD alone is highly reduced on insoluble or highly crystalline substrates [26] like cellulose which further confirms that CBMs have a critical functional role to play as well.

1.4 Structure-function relationship of *T. reesei* Cel7A CBM1

TrCel7A has a family 1 CBM domain that consists of 36 amino acids (PDB: 1cbh). Its wedge shape structure is possible due to three irregularly shaped β-sheets (Figure 5). This structure is further stabilized with the help of four cysteine residues lying within the β-sheets that form two disulfide bonds (not shown in figure 5). There are two faces of the CBM1 structure, one is a rough/irregular face and the other surface is relatively planar/flat. The flat face contains multiple polar and aromatic residues that are thought to be extremely important in the functionality of the CBM1 for binding to cellulose. It contains five critical residues that are involved in binding to cellulose; namely Y5, N29, Y31, Y32 and Q34 (Driscoll, Gronenborn, Beress, & Clore, 1989). In Figure 4, except Y5 and Y31 residues, which are the flanking residues at either edge of the planar surface, all other planar face residues are well conserved. But, both Y5 and Y31 positions are conserved for their aromaticity within the GH7 family of CAZymes (see Figure 4 below).

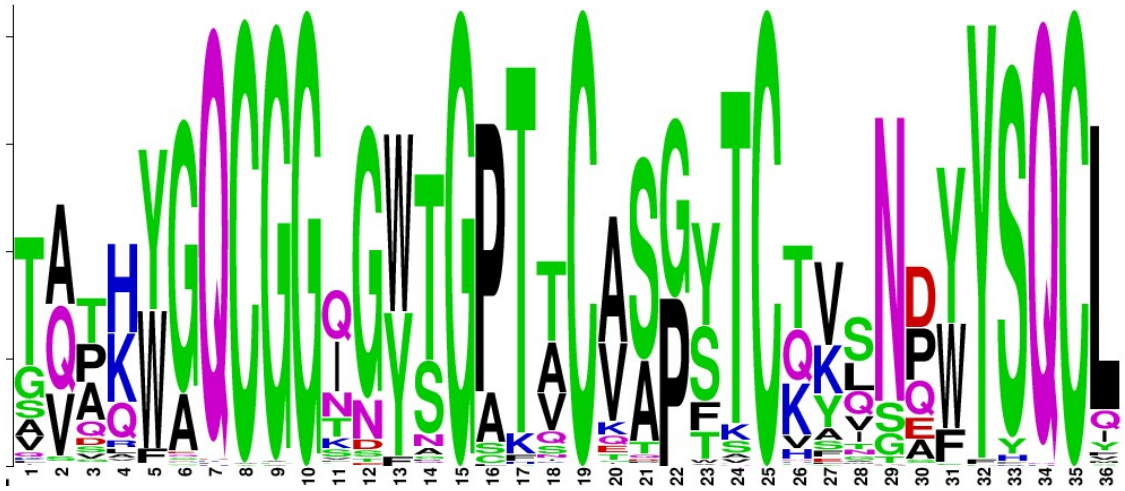


Figure 4: Multiple Sequence Alignment (MSA) analysis of CBM1 domains of all the GH7 cellulases containing CBM1 available in Pfam database. All GH7 CBM1 sequences were extracted from Pfam (https://pfam.xfam.org/), Aligned in Geneious 11.0.5 bioinformatics software, and the figure generated using WebLogo server (weblogo.berkeley.edu). Details provided later in the thesis.

From Figure 5 below, we see that the three aromatic tyrosine residues are aligned along the planar surface of the protein. The complementary structure of protein binding residues and the cellulose surface suggests that the aromatic rings are frequently involved in binding interaction due to stacking of interactions of the aromatic rings with the sugar rings of a single cellulose chain. These interactions are further assisted by hydrogen bonding interactions and hydrophobic interactions. The periodicity of glucose rings in cellulose and that of the aromatic rings of the protein are remarkably complementary. Both, the polar residues and aromatic residues are expected to form strong hydrogen bonds with the cellulose chain. The van der Waals interactions between CBM aromatic residues and glucopyranose rings within cellulose chains are also important for the binding of CBM to cellulose, as shown by another MD simulation study [2]. The sliding processivity of CBM1 is attributed to the electrostatic interactions observed after binding of CBM to the cellulose chain. The overall interaction of CBM1 with cellulose structure is thought to be driven primarily by positive entropy change which is a critical thermodynamic force thought to drive binding of CBMs to insoluble glycan ligands like cellulose [3].

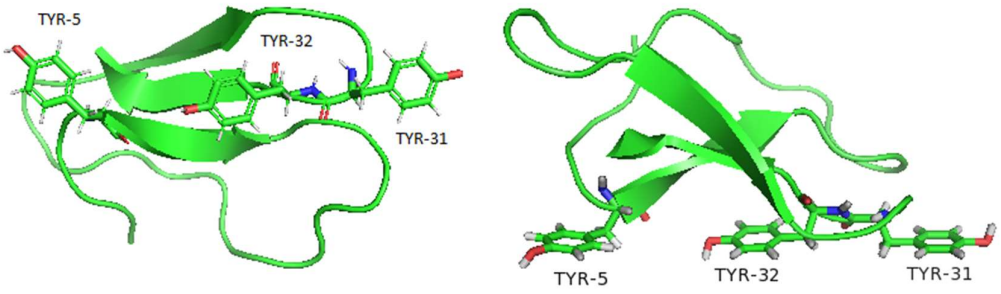


Figure 5: Bottom and side view of CBM1 from TrCel7A (pdb: 1cbh) in a cartoon representation with Tyr-5, Tyr-31 and Tyr-32 shown separately. Figure was generated using PyMol software.

The tyrosine residues at positions 5, 31 and 32 that are a part of the flat face of the protein are considered to be functionally important for the binding of CBM1 to cellulose [16].

The highly conserved asparagine N29 and glutamine Q34 residues are commonly involved in interactions of proteins with carbohydrates. This could be because of the high hydrogen bonding capacity of the polar residues. Site-directed mutagenesis has been performed on the flat face of the protein to understand the role of these residues on the fold of the protein and the binding interactions with cellulose substrate [20]. Alanine (A) was substituted at each of the flat face residues i.e. Y5, N29, Y31, Y32 and Q34 and structural characterization was carried out for those mutants. The mutation on N29 and Q34 did not show any significant impact except slight conformational changes for the mutant protein. Y31 mutation to alanine also showed only minor changes in the structure of flat face as compared to wild type. For Y32 mutation to alanine, the distance between N29 and Q34 was reduced and the tyrosine rings had gone out of the planar geometry [20] thus affecting both planarity and periodicity of the ring structures. For Y5 position, there was overall loss of compactness of the structure that also disturbed the planarity of the flat face [16, 20]. The Y5 position was thus be hypothesized to be important for structural integrity of the N terminus of the peptide. Also, the Y5 position of the peptide is in a type II turn and an unfavorable conformation (usually occupied by a glycine), but stabilized by positive histidine (His-4) residue besides the Tyr-5 position [19]. This in particular makes the N terminus sensitive to mutations. Y31 and Y32 residues are not as important to structural integrity of the backbone protein but they are functionally important in binding of CBM1 to cellulose. Y5 residue is also considered to interact with the adjacent chains of cellulose on the cellulose microfibril surface that could explain the reduced CBM1 binding to unnatural cellulose allomorphs like cellulose-III due to its stepped-like surface that likely causes steric hindrance during binding with the Y5 position [11].

1.5 Rosetta software based protein structure modeling and prediction

Predicting protein three-dimensional (3D) structures directly from amino acid sequences or homologous protein 3D models has been a critical to be able to successfully map out structure-to-function relationships for protein engineers. In a strictly physics-based *ab initio* or *de novo* approach to develop such models, interactions between atoms should be based on quantum mechanics. However, due to the lack of computational resources required for such calculations, a more practical starting point for *ab initio* protein modelling is to use compromised force fields for selected atom types. Examples of software that use such an all-atom physics-based force fields, often coupled with molecular dynamics (MD) simulations, to predict protein structures include AMBER and CHARMM. But physics-based models coupled with MD simulations are not always successful in accurate and fast structure prediction. In the last decade, faster conformational space sampling methods (such as Monte Carlo simulations) combined with physics-based potentials using softwares like Rosetta/FoldIt have yielded better results in structure prediction and [15].

The Rosetta software suite for macromolecular modeling has been used extensively by structural biologists in the last two decades to develop accurate *ab initio* structural models of biomacromolecules like proteins [15, 27]. It has been shown that 3D structural predictions for small proteins (<125 amino acids) using this software regularly has backbone root-mean-square (RMS) deviations smaller than 5.0 Å compared to experimentally determined structures using NMR or XRD. More importantly, there are several examples where the software can predict *ab initio* structures with atomic level accuracy better than conventional experimental methods (2.5 Å). Along with *ab initio* structure prediction, Rosetta has been also used for performing molecular docking, homology modeling, determining protein structures from experimental data, and protein engineering/design. Like other structure prediction algorithms, like iTASSER, AMBER etc, Rosetta performs two tasks *in silico* [15].

First, the software efficiently samples the conformational space, and in case of designing mutants, it also samples the sequence space. Next, the software accurately ranks or quantitatively evaluates the computed free energy of the resulting structural models. We have briefly described below the general outline of steps taken during 3D structural modeling using Rosetta to predict protein structures using sequences or homology models. Rosetta software implements a knowledge-guided Metropolis Monte Carlo (MMC) sampling approaches coupled with knowledge-based energy functions to solve protein structures. Knowledge-based energy functions assume that most molecular properties of folded protein domains can be derived from the Protein Data Bank (PDB). Simulations can be conducted using Rosetta with; (1) amino acid side chains represented by super atoms or centroids in a low-resolution mode, or (2) at fully atomistic detail in the high-resolution simulation mode. Both modes come with tailored energy functions that are reviewed elsewhere [28].

During the first low-resolution knowledge-based energy function simulation amino acid side chains are treated as simple centroids. The energy function models solvation, electrostatics, hydrogen bonding between β strands, and steric clashes. Solvation effects are modeled as the probability of having a particular amino acid with a given number of α-carbons within an amino acid-dependent cutoff distance. Electrostatic interactions are modeled as the probability of observing a given distance between the centroids of amino acids. Hydrogen bonding between β strands is evaluated on the basis of the relative geometric arrangement of strand fragments. Backbone atom and side chain centroid overlap is penalized and thus provides the repulsive component of van der Waals forces. The radius of gyration term is used to model the effect of van der Waals attraction. All probability profiles are derived using Bayesian statistics based on crystal structures available from the PDB. The lower resolution of centroid-based energy functions smoothes the energy landscape at the expense of accuracy. These energy landscapes allow prediction of reasonable starting

structures close to the true global minima to maintain a low energy state. Whereas, a fully atomistic energy function would penalize the structural defects seen in such low-resolution structures.

During the next fully atomistic high-resolution energy function simulations the software combines the 6-12 Lennard-Jones (LJ) potential for modeling van der Waals forces, a solvation approximation, an orientation-dependent hydrogen bonding potential, a knowledge-based electrostatics term, and a knowledge-based conformation-dependent amino acid internal free energy term. One critical detail in the construction of this potential was that all energy terms are pairwise decomposable which thus limits the total number of energy contributions to $N * (N - 1)/2$, where N is the total number of atoms within the system. This allows storage of several energy contributions in the computer memory to facilitate rapid execution of the MMC sampling strategies employed by the software during protein design and structure prediction.

Chundawat and co-workers have used Rosetta software in recent years to computationally screen protein surface residues that were ideal for making multiple mutations without deleteriously impacting protein fold stability [13, 33]. Since Rosetta uses a biased Monte Carlo search strategy with quasi-Newtonian minimization using an empirically derived scoring function, this facilitates rapidly ranking homologous protein variants with respect to a starting 3D template model sharing reasonable homology (>30%). This approach allows one to readily screen protein libraries *in silico* to facilitate prediction of overall stability and potentially binding properties of proteins to desired ligands. Ranking of mutations are based on a holistic Rosetta energy function, which is correlated to the change in Gibbs free energy vs. wild-type protein (ΔΔG). This Rosetta energy function can be also further deconvoluted into its various component free energy terms that facilitate gaining detailed quantitatively modeled protein structural insights with atomistic level resolution.

Rosetta also provides several optional user interfaces for interacting with the software and stored deconvoluted energy functions in the library. In addition to the standard command line interface, pyROSETTA (http://pyrosetta.org) has been also developed in recent years that contains a set of Python bindings to the Rosetta libraries that integrates many aspects of the software into Python scripts. A simple XML-based scripting language is available that allows users without significant programming experience to be able to quickly generate custom protocols to run simulations and extract relevant structural models/data upon running simulations

Objectives

The main objectives of this thesis project are as follows:

1. To attest the role of family 1 Carbohydrate-Binding Module (CBM) in reduced binding of full-length cellulases to cellulose-III.
2. Understanding the physico-chemical parameters that drive binding of CBM1 to both cellulose allomorphs for efficient designing of catalytically active enzymes.

To fulfill these objectives, the Y5 position of CBM1 appended at the N-terminus to a marker protein GFP was mutated to all other amino acids to create a single site saturation mutagenesis library of GFP-CBM1 proteins to cover the whole range of polar, non-polar and charged residues which might exhibit their own interesting effects on binding of CBM1 to cellulose allomorphs. These proteins were then expressed in Rosettagami expression strain of *Escherichia Coli* and purified via a two-step in-house optimized purification protocol using an AKTA FPLC system. For binding characterization, partition coefficient assay experiments were carried out which gave limited insight into the mechanism of CBM1 adsorption to cellulose. Subsequently, detailed pH dependent assays at single protein loading with and without the presence of suitable salts was conducted to compare binding of GFP-CBM1 mutant library to cellulose-I and cellulose-III. Lastly, detailed full-scale binding assays were conducted to estimate bulk binding affinity and total available binding sites for a select set of GFP-CBM1 mutants to native cellulose-I. Experimentally determined binding parameters were then correlated to *in silico* modeled physicochemical properties of the CBM1 mutant library using Rosetta software to infer structure-function relationships using a structure-guided first-principles approach. Specifically, here we use Rosetta software to model our protein mutants of interest and extract useful structural information from these models to

draw mechanistic insights about protein function and also develop regression based model that can predict protein binding to cellulose.

Chapter 2: Expression/purification of GFP-CBM1 and its Y5 position site-saturation mutagenesis library

2.1 Materials and Methods

2.1.1 Cloning of CBM1 gene from *T. reesei* Cel7A and its Y5 site-saturation library

The CBM1 gene from *T. reesei*, synthesized by Genscript, was fused at the C terminus of eGFP gene with 8x Histidine tagged on N-terminus using the pEC-GFP-CBM plasmid [18]. The CBM1 DNA gene was inserted between AflII and BamHI restriction sites using the relevant restriction enzymes followed by ligation using a T4 DNA ligase (NEB labs). The ligation products were then transformed in *E. cloni* 10G chemically competent cells (Lucigen) and plated on Kanamycin resistant LB agar plates. The colonies were screened by colony PCR and DNA sequencing using upstream NcoI and T7 terminator primers (Figure 6). Extracted plasmid DNA was transformed into Rossettagami expression strains of *E. coli* and stored as 20% glycerol stock solutions at -80^{o}C. Site-saturation mutagenesis library at the Y5 position of CBM1 was generated by Life Technologies (Invitrogen & Applied Biosystems) and sub-cloned into the pEC-GFP-CBM1 plasmid. These plasmids were also transformed into the Rossettagami expression strains and stored at -80^{o}C for future use.

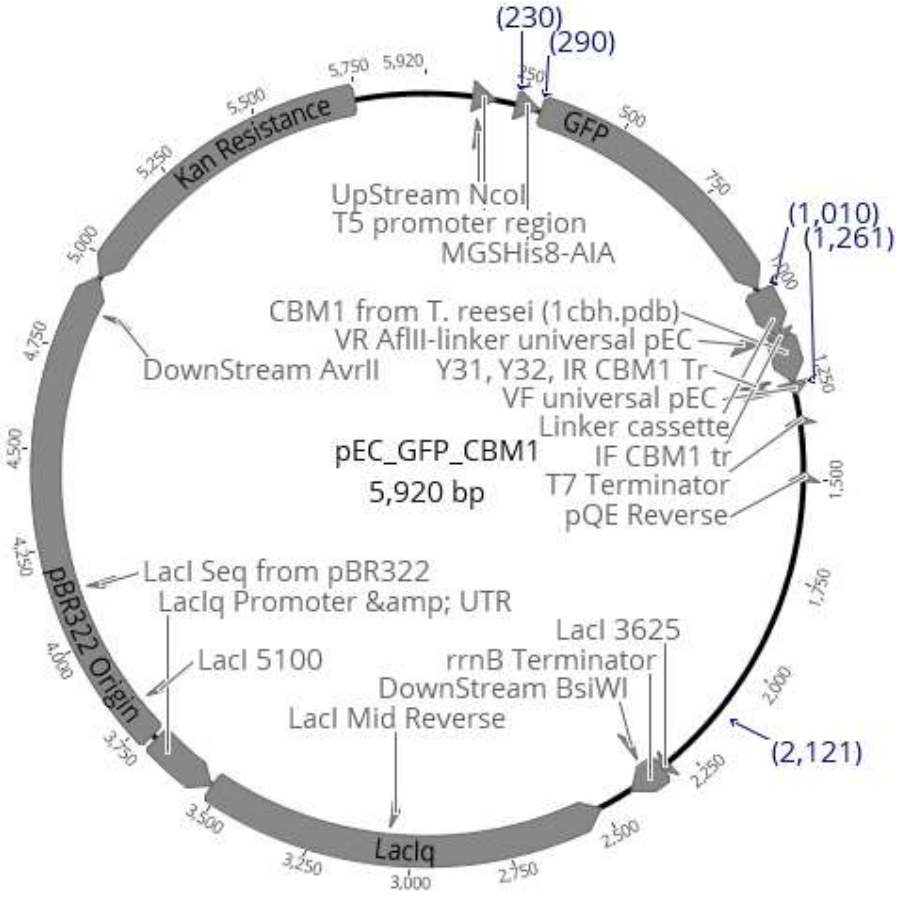


Figure 6. pEC-GFP-CBM1 plasmid DNA map generated using Geneious 11.0.5 bioinformatics software.

2.1.2 Small-scale protein expression studies

All mutant plasmids transformed cell lines were grown using the Rossettagami stocks using sterile LB media at 37°C for 16 hours in a shaker incubator. Total culture volume of 2 mL was used as starter culture for each mutant that was composed of 1% Tryptone, 1% Yeast Extract, and 1% NaCl. Kanamycin (final concentration 50 μg/mL) was added to the culture to avoid contamination. After 16 hours of growth, 50 μL of the overnight starter culture was transferred to TB+G media (5% v/v of TB+G media) that was composed of 1.2% tryptone, 2.4% yeast extract, 2.3% KH_2PO_4, 12.5% K_2HPO_4, 0.375% aspartate, 2 mM $MgSO_4$, 0.8% glycerol, 0.015% glucose, and 0.5% α-lactose, which is an auto-inducing media [18] in a 2

mL deep well plate. Total culture volume of 1 mL was used as production culture for each mutant. Kanamycin (50 μg/mL) was included as well and the production cultures were grown in a shaker incubator at 37°C till the cells reached an Optical Density (OD) corresponding to the exponential phase of growth curve. This was typically achieved between 4-6 hours of growth at 37°C. The incubator temperature was then reduced to 25°C and the cells were allowed to then grow and express the targeted proteins for an additional 24 hours. 50 uL of the culture cells was transferred to a 365 well flat bottom white plate and the OD at 600nm was measured post 24 hours of the growth phase. 100 uL of the culture cells were transferred to 96 well flat bottom black opaque microplate and the fluorescence (ext.480nm emm.512nm cutoff495nm) of the cells was measured post 24 hours of the growth phase. The culture plate was then centrifuged down (Eppendorf centrifuge) at 3600 RPM for 10 minutes to allow cells to settle down and the liquid media was carefully pipetted out and discarded. Next, 200 μL of B-PER II (ThermoFisher Scientific) bacterial cell lysis reagent was added to the wet cell pellets and gently mixed to ensure that the cells were lysed. The plate was then centrifuged again and 100 μL supernatant was removed and transferred to a 96 well black opaqe microplate and was analyzed for GFP fluorescence to determine total protein expression. All the OD600 values were measured in a 365 well plate and were converted to cuvette equivalent values by multiplying it by its conversion factor (Data in appendix). Then, the average of culture fluorescence values for all the mutant proteins and the average of supernatant fluorescence values for all the mutant proteins were normalized with the new OD600 values of the respective mutant protein. Finally, the culture and supernatant fluorescence values for various mutants were normalized with respect to that of the wild type protein.

2.1.3 Large-scale cell culture and protein expression

GFP-CBM1 wild type and all mutants were grown using similar protocol as described in Section 2.1.2 using a 25 mL of starter culture volume in LB/Kan media for 16 hours at 37^{o}C. This starter culture volume was then transferred to a larger batch of 1L TB+G media (5% v/v of TB+G/Kan media) and grown till the cell OD reached exponential phase prior to turning down the temperature to 25^{o}C. Finally, after 24 hours growth, the cell cultures were centrifuged in large bottles at 7500 RPM for 15 minutes and the wet cell pellets were harvested by separating the supernatant. The wet cell pellets were weighed in a falcon tube and stored in -80^{o}C for future use.

2.1.4 Large-scale cell pellet lysis and protein purification

Three grams of wet cell pellets stored at -80^{o}C were thawed and lysed using 15 mL cell lysis buffer. In the cell lysis buffer (composed of 20 mM Sodium Phosphate, 500 mM NaCl, 20% v/v glycerol at pH 7.4), 200 µL of Protease Inhibitor Cocktail (Composition: 1 µM E-64 from Sigma #E3132, 0.5 mM Benzamidine from Calbiochem #199001, and 1mM EDTA tetrasodium dihydride) and 15 µL Lysozyme (Sigma Aldrich, USA) were added as well to promote cell lysis. The mixture was well mixed and kept on ice for 10 minutes to make minimize protease activity. The lysis mixture was then mechanically lysed using a sonicator (Misonix Sonicator 3000) on ice with an output level of 4.5, 10 seconds ‘on time’, 30 seconds ‘off time’ and total process time of 5 minutes. The temperature never exceeded 7^{o}C - 8^{o}C during cell lysis. The cell lysate was then centrifuged at 20,000 RPM for 1 hr at 4^{o}C and the supernatant with protein of interest was collected and filtered using 0.22 µm syringe filter (Fisherbrand) prior to FPLC based protein purification.

The library of mutants along with wild type GFP-CBM1 were all purified using a two-step protein purification protocol using the AKTA FPLC (GE Healthcare) system. First step included a Immobilized Metal Affinity Chromatography (IMAC) purification stage where a 5 mL HisTrap FF Crude column (GE Healthcare) was cleaned with 5 Column Volumes (CVs) of elution buffer IMAC B (Composition: 100 mM MOPS, 500 mM Imidazole, 500 mM NaCl; pH 7.4) to make sure there were no residual proteins bound to the column. Next the column was equilibrated with 5 CVs of running/binding buffer IMAC A (Composition: 100 mM MOPS, 10 mM Imidazole, 500 mM NaCl; pH 7.4) before loading the filtered cell lysate supernatant containing the His-tagged protein of interest to the column. All background proteins through the column to waste, except the His-tagged proteins of interest that bind strongly to the Nickel-IMAC media on the HisTrap FF Crude column. The protein of interest were then eluted using IMAC B and the eluted fractions were found to contain the partially proteolyzed his-tagged GFP domain and the targeted full-length his-tagged GFP-CBM1 construct.

As optimized previously by Ms. Vibha Narayanan [21], the collected IMAC B eluents were further purified using conventional Hydrophobic Interaction Chromatography (HIC). This method allows separation of the intact GFP-CBM1 from the proteolyzed GFP domains due to differences in hydrophobicity. This method is similar to the cellulose affinity purification protocol, but is less laborious and gives marginally higher protein yields upon completion of purification [31]. Briefly, the eluent from IMAC purification were first pooled together and desalted into the HIC loading buffer composed of 1 M Ammonium Sulfate and 50 mM Sodium Phosphate at pH 7 using a HiPrep 26/10 Desalting column. The desalted eluent was then loaded onto a 5 mL HiTrap Butyl Sepharose column (GE Healthcare) using a low flowrate of 0.5 ml/min to make sure the protein binds to the column properly. The protein elution was done using a linear gradient over a period of 3 hours, at low flowrates to

ensure efficient separation, using an elution buffer composed of 50 mM Sodium Phosphate at pH 7. As the concentration of loading buffer decreases with the linear gradient, the less hydrophobic GFP only domains elute off the column first followed by more hydrophobic intact GFP-CBM1 constructs. Multiple smaller volume fractions of eluent were collected to ensure that we obtain highly pure GFP-CBM1 containing eluents for SDS-PAGE binding characterization prior to pooling fractions.

A small scale binding assay was done prior to SDS-PAGE analysis as a preliminary step in identifying and confirming that the intact GFP-CBM1 protein was eluted off the column in the collected fractions. Briefly, 100 μL of 100 mg/mL cellulose-I slurry was added along with 100 μL of the collected protein solution from the respective fraction number to the respective number of PCR tubes to maintain a 1:1 volume ratio of cellulose slurry to the added eluent protein fractions. The tubes were gently mixed for 10 minutes and then centrifuged in a PCR tube centrifuge. Then 100 μL of the supernatant which contained the unbound protein was carefully pipetted out ensuring no solid particles of cellulose-I and was transferred to a 96 well flat bottom black opaque microplate. Next, 100 μL of the protein solution from respective FPLC fraction number was also transferred to the plate as a substrate blank control for total protein. The plate well fluorescences were measured (ext.480 nm, em.512 nm, cut off495 nm) to estimate the free and total protein concentrations. The subtraction of the GFP fluorescence readings for free protein wells from the total protein added substrate blank control wells allowed us to estimate the amount of protein bound to the added cellulose. Since, some of the initial eluting fractions have proteolyzed GFP fragments and the latter fractions have mostly pure/intact GFP-CBM1 proteins, it is expected that the binding should increase for the latter collected fractions as observed in the reported results (Table 2). This assay helps validate, along with SDS-PAGE,

which eluted FPLC fractions contain fully-functional intact GFP-CBM1 protein for further analysis.

SDS-PAGE analysis was completed for the collected eluents along with the standard ladder which helped identify the bands with correct Molecular Weight of GFP-CBM1. All eluent fractions containing the right size proteins were pooled together and concentrated using the Amicon 10 kDa centrifugal concentrators at 6000 RPM for 40 minutes. The concentrated proteins were then pipetted out and then desalted into a storage buffer composed of 50 mM MES at pH 6.5 using PD-10 gravity desalting columns (GE Healthcare). The final desalted protein were then aliquoted and flash frozen using liquid N_2 and stored at -80^{o}C. Protein concentration for aliquots was estimated using previously reported extinction coefficients and reading A280 on Spectramax M5e (Molecular Devices).

2.2 Results

The small scale expression studies were done in a 2 mL deep well plate with the exact same conditions also used for large scale expression work. The cells containing the GFP-CBM1 mutants/wild-type were expressed followed by cell lysis to estimate total GFP fluorescence in the supernatant. The results from the small-scale expression studies are shown in Table 1 below. We can clearly see that the wild type (Y5) always expresses the best and the expression of the mutant proteins mostly resulted in lower yield compared to the wild type (based on GFP fluorescence directly). Y5N mutant gave the highest expression yields amongst all mutants that was nearly equivalent to the wild-type protein based on RFU (data in appendix). The only other mutant constructs that gave higher RFU based protein yields were Y5E and Y5I which are not statistically different than that of wild type protein (Data in appendix). Overall these results are not entirely surprising considering that mutations made

at the Y5 position to alanine are reported to slightly destabilize the disulfide bond linked β-sheet protein structure that results in lower final expressed yield [16, 20, 25].

CBM1 Y5 Mutant Type	RFU/RFUwt in Cells	Standard Deviation in RFU/RFUwt in Cells	RFU/RFUwt in Supernatant	Standard Deviation in RFU/RFUwt in Supernatant
Wild type	1.000	0.193	1.000	0.188
N*	0.857	0.140	0.850	0.143
E*	0.781	0.158	0.802	0.164
I*	0.774	0.182	0.899	0.210
C	0.623	0.102	0.675	0.098
R	0.612	0.091	0.472	0.076
L	0.604	0.089	0.575	0.084
H	0.585	0.087	0.480	0.070
A	0.581	0.164	0.704	0.199
D	0.430	0.071	0.486	0.068
G	0.408	0.063	0.387	0.059
W	0.383	0.060	0.326	0.053
F	0.322	0.071	0.306	0.071
T	0.298	0.070	0.235	0.058
V	0.290	0.044	0.339	0.066
Q	0.216	0.034	0.224	0.033
K	0.194	0.034	0.146	0.023
S	0.170	0.024	0.156	0.023
P	0.160	0.027	0.141	0.024
M**	N/A	N/A	N/A	N/A

Table 1: Average of 3 replicates of Relative Fluorescence Units (RFU) with the Standard Error obtained was normalised with respect to OD600 of the respective mutant protein and also with respect to RFU measured for the wild type protein. *Values are not significantly ($P<0.05$) different than that of reported wild type protein values (data in appendix). **Values for Y5M mutant could not be obtained from small-scale expression studies since this construct did not grow in 1st stage of growth in LB media itself. N/A stands for not available.

The pEC-GFP-CBM1 plasmid transformed *E. coli* Rossettagami strains were grown in a non-inducing LB media till the OD value reached 0.5 to 0.7 and then the cells were transferred to a larger auto-induction TB+G media and grown for an additional 24 hours.

The cells were harvested after 24 hours and the total amount of cell pellet obtained for all the mutant proteins were noted down to calculate final purified protein yields.

CBM1 Y5 Mutant Type	Total grams of cell pellet obtained per liter of TB+G culture volume	Final purified protein yield (in mg) per gram of starting wet cell pellet
A*	7.89	0.74
C	7.21	0.78
D	7.50	0.72
E	8.93	0.50
F*	9.63	0.26
G	7.25	0.98
H	7.00	0.77
I	8.18	0.35
K	9.43	1.00
L	5.83	0.28
M	6.39	0.66
N*	7.96	0.51
P	5.36	0.91
Q	9.63	0.93
R	7.40	0.70
S	8.46	0.33
T	6.64	0.66
V	7.37	0.73
W*	10.17	0.24
Y (wild-type)	9.00	0.37

Table 2: Final purified protein yield for all Y5 mutant proteins calculated per gram of wet cell pellet obtained from culture volume and grams of cell pellet obtained per liter of TB+G culture volume. *Values taken from Narayanan, V. et al [21].

Next, large scale protein expression and purification work was conducted to purify the 8x-His tagged GFP-CBM1. The 5 mL HisTrap FF crude has a total binding capacity of ~200 mg of His-tagged proteins, though after visually inspecting the purification run it was observed that the flow through started to turn green after 40 mL of the cell lysate was loaded on to the column. All the runs were thus done loading less than 40 mL of filtered cell lysate to prevent over loading the column. AKTA FPLC (Amsheram Biosciences) runs on UNICORN 5.0 software where the user interface shows the live chromatogram that can be saved for future record keeping and analysis. A typical chromatogram for IMAC purification of the his-tagged GFP-CBMs is shown in Figure 7. It can be seen that during elution, the His tagged protein comes off as a single peak of UV absorbance at 280 nm. The high UV background during loading step represents all background cell lysate proteins flowing through the column in to the waste. The UV trace in the chromatogram also helps track the elution of protein during the desalting step which was done using a HiPrep 26/10 column (GE healthcare). This column has a maximum pore volume of 15 mL. The protein comes off the desalting first followed by the IMAC B eluents. IMAC B alone has significant absorbance at 280 nm due to presence of imidazole, and hence we see a tiny spike in the UV trace even after the protein has eluted off the column. Fractions were collected to avoid any traces of IMAC B contamination with the desalted protein.

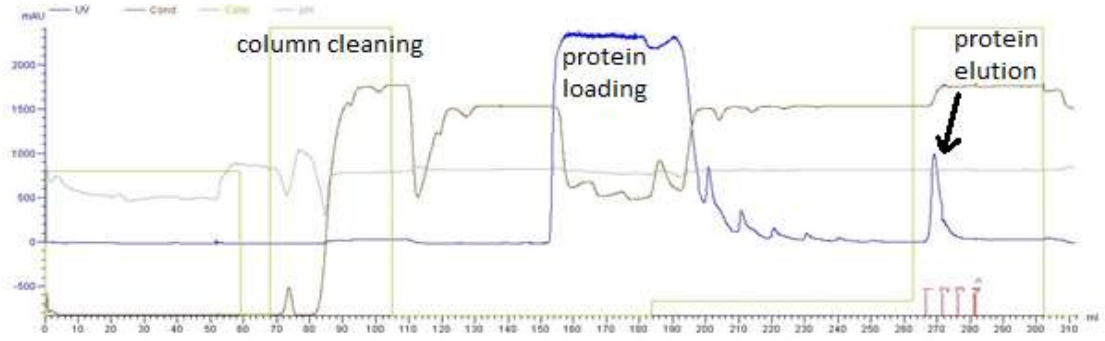


Figure 7: Chromatograph from immobilized metal affinity chromatography (IMAC) purification of GFP-CBM1 Y5K showing loading and elution steps. Figure generated using GE Healthcare Unicorn 5.1 software.

The second stage of protein purification was done using Hydrophobic Interaction Chromatography (HIC). The start buffer was chosen from the Hoffmeisters series of salts. Using high concentration of a suitable “salting out” salt, the interaction between the ligand in the column and the protein is increased to facilitate the binding of the protein to the column. The elution buffer had less “salting out” effect thus promoting decreased hydrophobic interaction which causes elution the protein off the column. The elution is done in a gradient fashion for 3 hours at 0.5 ml/min. The low flowrate along with gradient elution step facilitates in better resolution between the eluting protein peaks and also avoids co-elution of intact proteins with proteolyzed protein domains. The HIC chromatogram shown in Figure 8 shows that the proteolyzed GFP only domain comes out first and the intact GFP-CBM1 elutes later since GFP alone has a lower hydrophobicity and GFP-CBM1 has higher hydrophobicity. A small-scale binding assay to cellulose-I was carried out immediately after the HIC fractions enriched in the intact GFP-CBM1 proteins are collected. Here, a 1:1 volume ratio of cellulose slurry to the eluted protein fraction was added together in a PCR tube and mixed for 10 mins. The cellulose was pelleted down pulling down bound GFP-CBM1 proteins and supernatant fluorescence was analyzed to confirm residual unbound GFP-CBM concentration to estimate % drop in fluorescence intensity for the solid depletion assay. The earlier eluting fractions, which contain mostly proteolyzed GFP were found to show a significantly lesser drop in fluorescence as compared to the later eluting fractions enriched in full length GFP-CBM1 (Table 2). This preliminary binding assay along with SDS-PAGE allowed us to confirm if the HIC separation step was conducted properly and identify the eluent fractions that need to be pooled together.

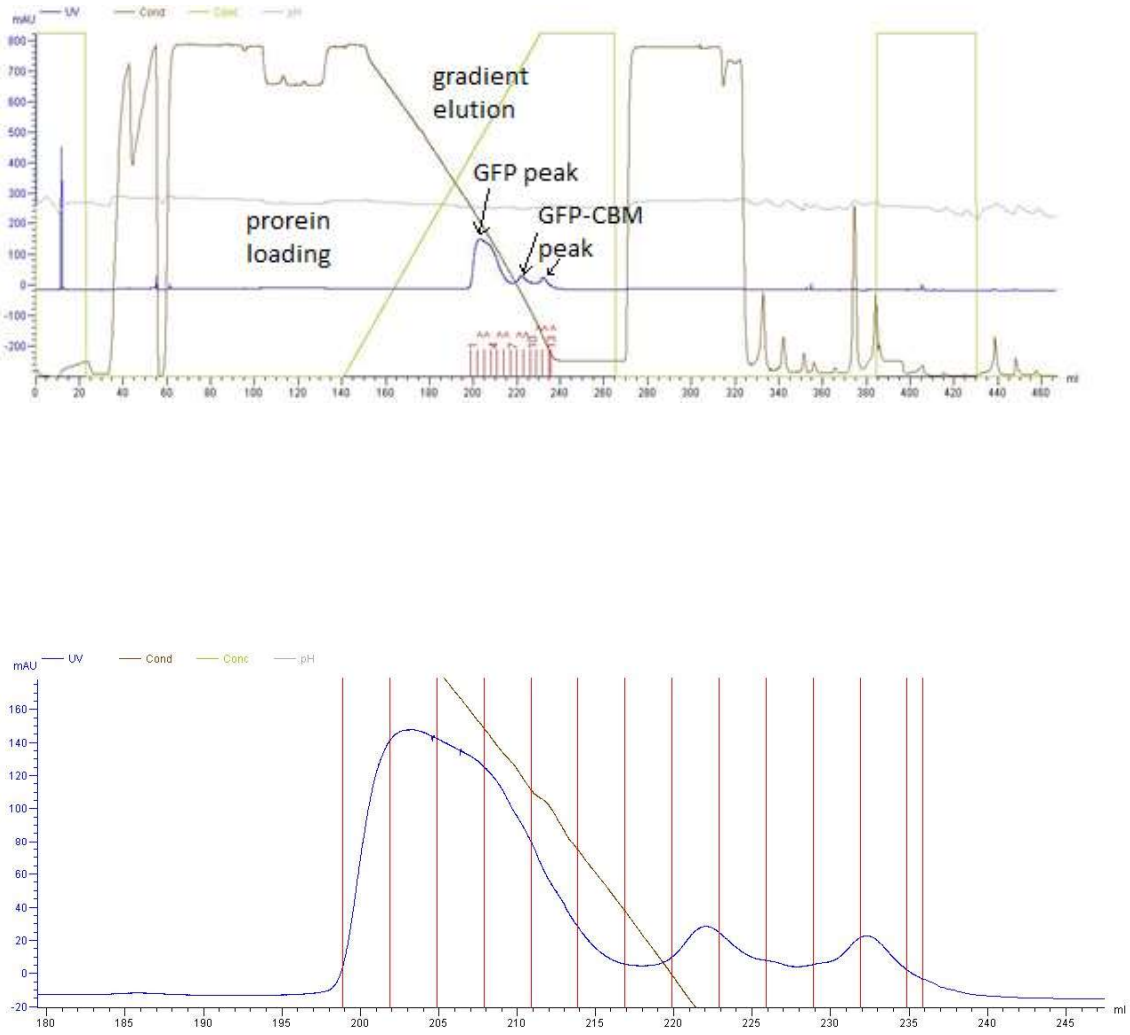


Figure 8: Top image is typical chromatogram showing 2nd step of purification using hydrophobic interaction chromatography (HIC) for GFP-CBM1 Y5K and bottom image represents the magnified image of the chromatogram depicting the eluting peaks/fractions enriched in proteolyzed GFP fragments and intact full-length GFP-CBM1 (Y5K mutant) protein. Fractions 1-13 were collected here and are shown by the vertical red lines. Figure generated using GE Healthcare Unicorn 5.1 software.

Fraction number	1	2	3	4	5	6	7
Free protein	23010.11	55122.15	54116.73	50570.05	30748.31	15281.1	6783.76
Total protein	40912.93	83842.13	82252.57	75460.53	56844.43	29366.27	15597.89
Bound protein	17902.82	28719.98	28135.84	24890.48	26096.12	14085.16	8814.13
% bound	43.76	34.25	34.21	32.98	45.91	47.96	56.51

Fraction number	8	9	10	11	12	13
Free protein	10248.43	10837.3	5855.15	5470.53	7701.57	4291.07
Total protein	21257.28	24260.95	14501.19	14852.15	23296.26	13768.22
Bound protein	11008.85	13423.65	8646.05	9381.62	15594.69	9477.15
% bound	51.79	55.33	59.62	63.17	66.94	68.83

Table 3: Small scale binding assay data for GFP-CBM1 Y5K eluted HIC column fractions are shown here. The fraction numbers correspond to the same fractions indicated in Figure 8 and Figure 9. The values indicate fluorescence values of free protein in the supernatant in presence of cellulose-I and total protein which was added in absence of cellulose (substrate blank controls). The bound protein was calculated by subtracting free protein values from total protein values.

The eluting protein fractions are further validated by running SDS-PAGE. In Figure 9, we see that there are multiple protein bands at around the 27 kDa mark in the earlier eluting fractions while increased number of bands around 37 kDa are clearly seen in the latter eluting HIC protein fractions. The lanes with bands at 37 kDa likely correspond to the full length GFP-CBM1. All fractions with bands only at 37 kDa (and higher cellulose binding affinity, Table 2) were pooled together and then concentrated using 10 kDa spin filters. The concentrated proteins were then desalted into a storage buffer (10mM MES pH 6.5) using a gravity flow desalting column (GE Healthcare) and the protein concentrations were estimated using Spectradrop at 280 nm.

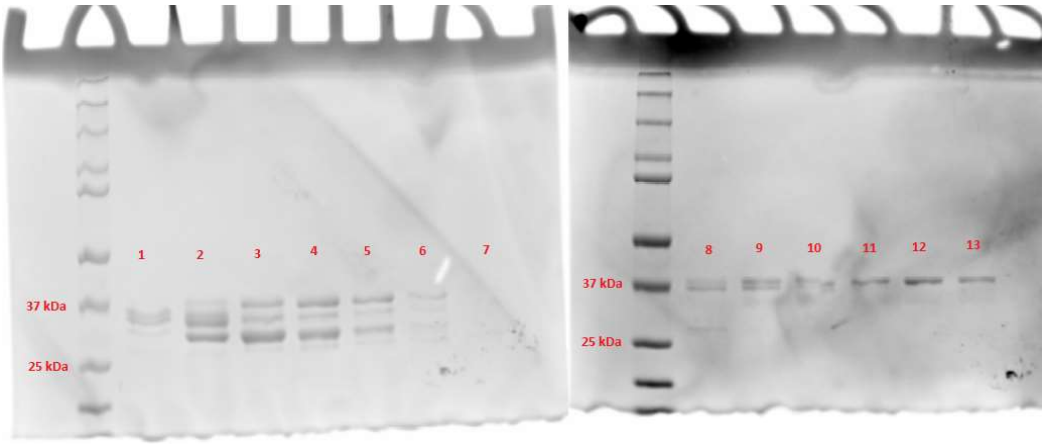


Figure 9: SDS-PAGE coomassie stained gel image of eluted protein fractions corresponding to the chromatogram shown in Figure 8. The 1st lane is the protein standard ladder. Rest of the lanes are labelled according to the fraction numbers of the eluting HIC column fractions.

2.3 Discussion

The three tyrosine residues on the flat face of the CBM1 protein play a critical role in cellulose binding affinity. Among them, the Y5 position is likely the most important due to its structural and functional role in CBM binding affinity towards cellulose. Site-directed saturation mutagenesis library of 19 mutants for Y5 position were generated to carefully study the structure-function relation of the mutants and the wild type proteins. All the mutants were characterized for their various properties and parameters.

The GFP fused to CBM1 facilitates rapid detection of the binding proteins using fluorescence measurements. Green Fluorescence Protein, as the name suggests, is a highly fluorescent protein (excitation 488 nm and Emission 512 nm) that makes it easy to complete the protein purification process as the proteins can be visually tracked. The chromatograms obtained from UNICORN software used to run AKTA FPLC further validates the visual inspection of the protein during the purification process. The His-tagged GFP fused with

CBM1 was chosen as it can be easily purified using Immobilized Metal Affinity Chromatography (IMAC). The 8x Histidine tag on GFP interacts with the metal (Ni^{2+} NTA) immobilized on the resin and binds to it. The bound proteins are then eluted using high concentrations of Imidazole [4] which has preferential binding to Ni^{2+} NTA. IMAC purification gives mostly two protein fractions, proteolysed GFP domain and the intact full length GFP-CBM1. These proteins were then separated using Hydrophobic Interaction Chromatography (HIC) based on the extent of hydrophobicity of individual proteins. The purification process was optimized to use Butyl Sepharose column which has a medium affinity interactions as compared to Phenyl Sepharose column which had very high irreversible interactions with the protein. It can be seen in Figure 8 that there are 3 peaks, 1st peak corresponding to GFP only and the last two of intact-GFPCBM1 (confirmed by small scale assay and SDS-PAGE), which can be because the higher amounts of GFP might drag the GFP-CBM1 down in the column to some extent along with itself while eluting and when the GFP-CBM1 starts eluting, this dragged protein comes off before the non-displaced GFP-CBM1. SDS-PAGE analysis of the collected fractions helped identify the eluting fractions that had intact GFP-CBM1 proteins alone.

The yields of all the final purified intact proteins were not equal (Table 2). Interestingly, here the wild type protein along with several mutants yield the highest amounts of final purified intact proteins (mg protein/L of culture medium). These large scale results were similar to the the small scale expression level experiments where the wild type GFP-CBM1 gave the highest normalized RFU yield.

Chapter 3: *In silico* and *in vitro* characterization of GFP-CBM1 Y5 mutant library

All 19 mutants and the wild type GFP-CBM1 proteins were successfully expressed and purified using the 2-step purification method outlined in chapter 2. Next we performed several adsorption studies on cellulose-I and cellulose-III to characterize the binding properties of the mutant library. However, single protein loading experiments in the linear binding concentration range can also allow us to estimate an apparent partition coefficient value that is fairly close to the actual partition coefficient. Since we had limited amount of protein available, we used this modified apparent partition coefficient estimation method for all the mutants at various pH and salt concentration. We also analyzed the *in silico* properties of all the mutants using Rosetta software and ran regression analysis to develop an empirical model capable of predicting the apparent binding affinity of CBM1 to cellulose based on the underlying structure-based modeled parameters of CBM1. All these results were then closely evaluated to decide which Y5 mutants should be cloned into a full-length cellulase construct (e.g., CelE-CBM1) for future enzymatic activity studies.

3.1 Materials and Methods

3.1.1 Multiple sequence alignment of all available GH7 CBM1s

All CBM1 sequences reported exclusively for Glycosyl Hydrolase Family 7 were extracted from PFAM (https://pfam.xfam.org/) and imported in Geneious 11.0.5 bioinformatics software. Next, the regions of GH7 and CBM1 were annotated based on uniprot entry (also provided on the website along with the sequence). The DNA regions coding for CBM1 alone were extracted and saved in a separate file. All data files containing the sequence for CBM1s alone were then aligned using the Geneious Alignment tool with alignment type "global alignment with free end gaps" and the cost matrix of "Blossom 62". The gap open penalty chosen was 12, gap extension penalty chosen was 3, and the refinement iterations was set at 2. The new file which had all the sequences aligned was exported as a fasta file and saved locally. This fasta file was then uploaded on the WebLogo server (weblogo.berkeley.edu) to generate an image file that also shows the level of conservity of all amino acid residues at every position in the CBM1 sequence.

3.1.2 pH/Salt dependent apparent partition coefficient estimation cellulose binding assays

All full-scale partition coefficient assays were initially conducted at pH 6.5 since GFP shows optimal fluorescence at close to neutral pH (data shown in appendix). Details regarding the full-scale parition coefficient assay are highlighted in another paper published by Dr. Chundawat elsewhere [18]. However, due to the limited amount of purified protein available it was not possible to run full scale partition coefficient assays for all mutants at various buffer compositions to explore the effect of pH and salt concentration on CBM binding to cellulose. We suspected that both pH and salt type/concentrations would have a

significant impact on the binding of some GFP-CBM1 mutants due to the complex role of electrostatic and hydrophobic interactions known to drive CBM binding to insoluble cellulose. Preliminary assays also showed that both pH and Salt have a drastic effect on the binding behaviour of wild-type GFP-CBM1. To explore these effects in more detail, all protein constructs including the wild type were screened for binding to cellulose at 4 different pHs (pH 5, 5.5, 6, 6.5), both in the presence and absence of a suitable salt. Avicel PH-101 (Sigma Aldrich) based microcrystalline cellulose-I derived from plant biomass and cellulose-III were used for these screening assays. Cellulose-III was produced by pre-treating cellulose-I using anhydrous liquid ammonia (Ammonia pretreatment conditions: 90^{o}C, 30 minutes, ~1000 psi, mass ratio of cellulose to ammonia loading was 1:6). Cellulose III was prepared and kindly gifted by Dr. Leonardo Sousa at Michigan State University. The screening assay was set up for only one protein loading fixed at either 2.5 mg/g cellulose-I or 0.625 mg/g cellulose-III. These protein loadings were picked based on earlier assays with wild-type GFP-CBM1 to make sure that the protein loading lies in the mid-linear range used to estimate the partition coefficient (data shown in appendix). Binding assays were performed in a 300 μL volume based 96 well clear round bottom propylene microplate (USA scientific). Each assay condition were run in triplicates for both cellulose-I and cellulose-III with equal number of controls (i.e., no substrate added). First, 25 μL of well-mixed 100 mg/mL (w/v) cellulose-I slurry and 100 μL of 100 mg/mL cellulose-III slurry was added to the empty wells. Then, 50 μL of BSA-buffer mixture composed of MES at desired pH was added to get an effective final concentration of 50 mM MES in the well. For the plate which had salt, certain amount of water was replaced with 1M NaCl to get 100 mM effective final salt concentration in the well. BSA was added to prevent non-specific interaction of proteins with the tube/well walls. The final amount of targeted protein solution and DI water added to each well was adjusted to get a final reaction volume of 200 μL, to

maintain a 2.5 mg protein loading/g cellulose-I or 0.625 mg protein loading/g cellulose-III. The plates were then covered with a plate mat and then mixed in an end-over-end fashion using a VWR Hybridization for 1 hour at room temperature. The microplates were then centrifuged at 2000 rpm for 2 minutes using an Eppendorf centrifuge. Then, 100 μL of supernatant was carefully pipetted out, ensuring no solids were picked up, and transferred in to a 96 well flat bottom black opaque microplate. The residual GFP fluorescence (480 nm excitation, 512 nm emission, 495 nm cutoff) was then read for each well in the plate using a Molecular Devices Spectramax M5e to estimate protein concentration. Control plates containing known concentrations of targeted proteins, without cellulose, were used to create a protein calibration curve to estimate the concentration of unknown samples.

3.1.3 GFP-CBM1 library *in silico* Rosetta parameters modeling

The library of GFP-CBM1 mutants were scored for 17 different protein properties using the Rosettascript scripting interface [10, 15]. The ROSETTA software interface was used to score all protein mutants (script used are provided in appendix). Here, Rosetta utilizes the score function “talaris2014” to score the folded protein energy score of the given protein and then the script relaxes the structure to obtain the minimum energy structure. Typically, around 150 iterations (in steps of 10 iterations per run) are needed to be run to get to the minimum energy structure where the difference between two subsequent minimized structures is at least 0.1 Rosetta Energy Units (REU). Once this minimized energy structure is obtained, then the relevant structural properties and modeled energetic terms are extracted from the Rosetta library using a standard pyRosetta script. Once, the properties of wild type protein are noted/stored, a desired mutation is then introduced at the Y5 position. The Rosetta

script is run again to obtain a minimized energy structure and the structural properties and energy terms are again noted/stored for that respective mutant.

3.1.4 Regression analysis of Rosetta model parameters with experimental binding data

A linear regression model was built to predict the apparent binding affinity values using various empirically chosen variables available from *in silico* modeling (and some experimental values such as RFU). The input variables to the regression model include pI of the whole CBM1 protein at pH 5 determined using the empirical formula based on number of charged residues and operating pH, pKaAA (pKa of individual amino acid at Y5 position of CBM1 domain), H-patch score at pH 5 determined using Rosetta, RFU (relative fluorescence units) from small scale protein expression experimental studies, and REU (Rosetta Energy Units) determined using Rosetta (refer section 3.1.3). The software used was multivariate regression analysis was RStudio v3.5.0.

The output apparent binding values and all input variable values were first imported to R using “read.csv” command. Then, “aov” and “lm” commands were used to perform ANOVA and fit a linear regression model. Finally, “step” function was used to refine the model and include only significantly important (>95% significance) parameters at all levels of variable interaction terms. The regression model coefficients were then exported in a .csv file for detailed analysis. The values of significant parameters at all levels of interaction terms were then fitted into the model using the coefficient terms. Next, the predicted apparent partition coefficients were back calculated to check the accuracy of the model compared to the fitted data.

3.2 Results

The output image file from WebLogo analysis for all available CBM1 sequences for GH7 enzymes in the genbank database is shown in Figure 4. We can see that residues from 5 to 9, from 24 to 28 and from 33 to 36 that represent the β sheets of the CBM1 protein are highly conserved along with the 4 Cysteine residues at positions 8, 19, 25 and 35 that helps maintain the structural stability of the protein. The flanking residues on the flat face of the protein i.e. Y5 and Y31 are not as highly converved as residues at N29, Y32 and Q34 and among the flanking residues, Y5 is the least conserved.

Here we also conducted all pH based protein adsorption studies on both cellulose-I and cellulose-III either in the presence or absence of 100 mM NaCl salt for pHs ranging between 5 to 6.5. These pH dependent assays are a good starting point to gain mechanistic insights into structure-function relationships driving protein stability and ligand binding for the Y5 mutant library of GFP-tagged CBM1 proteins. From Figure 10, we can see that for cellulose-I, as we increase pH from 5 to 5.5 without addition of any excess salt, the average apparent partition coefficient at pH 5 (i.e., 1.654 ± 0.1 mg GFP-CBM1/g cellulose-I) for all the mutants reduces by around 1.5-fold except for wild-type CBM1. Interestingly, at pH 5 itself, the average apparent partition coefficient reduces by 5-fold upon addition of 100 mM NaCl salt (data shown in the appendix). Furthermore, increasing the pH from 5 to 5.5 in presence of 100 mM salt, also reduces the average apparent partition coefficient (i.e., 0.581 ± 0.143 mg GFP-CBM1/g cellulose-I) at pH 5 by 1.2-fold. Overall, increasing the pH from 5 to 6.5 (data shown in appendix), the apparent partition coefficient is seen to progressively reduce for all mutants. Interestingly, there was no significant loss in binding observed for the wild type protein at increasing pH's, which was also observed by Linder, M. et al (1999) [17]. While, the addition of salt resulted in further decreasing the apparent partition coefficient for all Y5 mutants, when pH was increased (data shown in appendix), the wild-

type CBM1 affinity towards native cellulose-I was not significantly affected by the addition of salt for all pHs. For cellulose-III, similar trends were also observed, where increasing the pH caused a reduction in the average apparent partition coefficient and inclusion of higher salts concentrations further reduced the affinity for all mutants. Even for cellulose-III, the wild-type CBM1 affinity was not affected either by either increasing pH or addition of salt to the system. Since CBM1 is appended to the catalytic cellulase domain for native TrCel7A that gives the highest hydrolytic activity at pH close to 5, we originally suspected that the wild-type CBM1 domain might have co-evolved along with the catalytic domain to give maximum binding at close to pH 5. However, surprisingly the native CBM1 domain was found to give comparable apparent partition coefficient to cellulose at pH's ranging from pH 5 to 6.5 at a fixed low protein loading. It is possible that there are differences observed in the binding of the native CBM1 to cellulose at higher concentrations with saturation of available binding sites due to the increased likelihood of protein-protein interactions that could impact apparent binding affinity measurements. Nevertheless, it was interesting to find that most Y5 CBM1 mutants were highly sensitive to both pH and salt concentration on their measured low protein loading based partition coefficient measurement. Furthermore, there were clear differences in the binding patterns seen for the various variants depending on the type of Y5 mutation. This led us to more closely examine the underlying molecular-origins for the differences observed in partition coefficients between mutants at pH 5, since this value is closest to the physiologically relevant pH for CBM1 and its full-length TrCel7A construct.

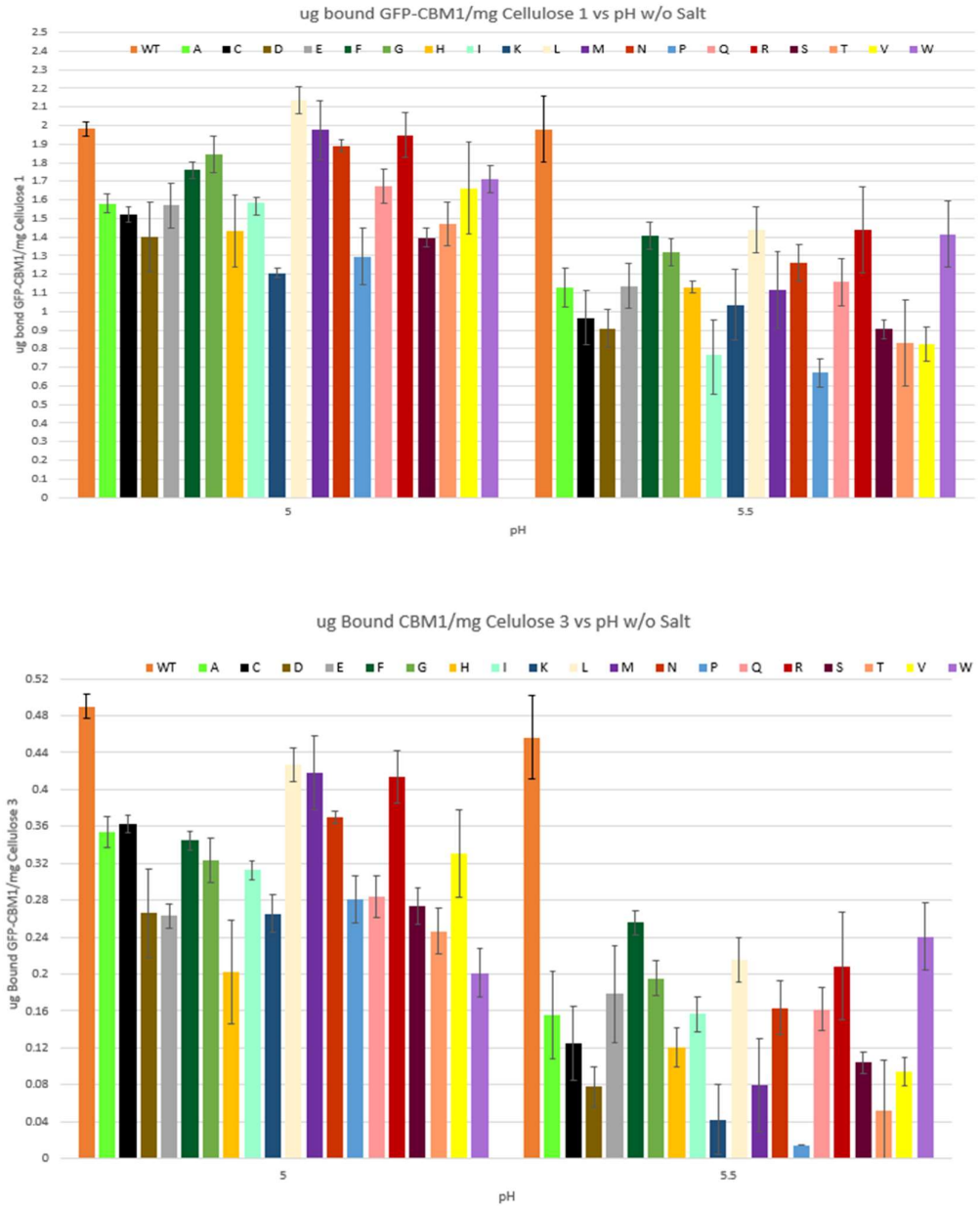


Figure 10: Apparent partition coefficients estimated for the library of CBM1 Y5 site-saturation mutants appended to GFP reporter fluorescent proteins. The top figure is for cellulose-I and the bottom figure is for cellulose-III for pH 5 and pH 5.5. The color legend (from left to right) provides details on the wild-type (WT) protein and the Y5 mutant library with single-letter amino acid notations used here (e.g., A represents Y5A CBM1 mutant).

Cellulose-I			Cellulose-III		
Mutant	Apparent Partition Coefficient (mg GFP-CBM1/g cellulose)	Standard Deviation	Mutant	Apparent Partition Coefficient (mg GFP-CBM1/g cellulose)	Standard Deviation
L	2.135	0.08	Wild type	0.489	0.01
Wild type	1.983	0.04	L	0.427	0.02
M	1.975	0.16	M	0.418	0.04
R	1.95	0.12	R	0.414	0.03
N	1.891	0.03	N	0.369	0.01
G	1.847	0.10	C	0.362	0.01
F	1.759	0.04	A	0.354	0.02
W	1.71	0.07	F	0.345	0.01
Q	1.673	0.09	V	0.331	0.05
V	1.663	0.25	G	0.323	0.02
I	1.584	0.06	I	0.312	0.01
A	1.58	0.05	Q	0.283	0.02
E	1.57	0.12	P	0.281	0.03
C	1.522	0.04	S	0.274	0.02
T	1.471	0.12	D	0.266	0.05
H	1.434	0.19	K	0.265	0.02
D	1.402	0.19	E	0.263	0.01
S	1.398	0.05	T	0.246	0.03
P	1.295	0.15	H	0.202	0.06
K	1.207	0.03	W	0.201	0.03

Table 4: Rank ordering of the library of CBM1 Y5 mutants based on experimentally determined apparent partition coefficient (from highest to lowest) maintaining 3 replicates for each construct estimated at pH 5 for both cellulose-I and cellulose-III.

At pH 5, the mutants have an average apparent partition coefficient of 1.654 ± 0.1 mg GFP-CBM1/g cellulose-I as compared to the highest partition coefficient of 2.1 seen for the Y5L mutant for cellulose-I. This is about 25-30% drop in binding observed on average for the mutants compared to the highest binding mutant. However, there are several mutants (e.g., Y5L, Y5R, Y5M, Y5N, Y5G) which have apparent partition coefficient value within ±10% of the wild-type for cellulose-I. However, there are certain mutants have binding much lower than the binding observed for the wild-type CBM1. To more systematically understand the behaviour of the mutants, we used Rstudio statistical software package, to

develop a model upon fitting the apparent partition coefficient data for both cellulose-I and cellulose-III at pH 5 with relevant input variables mostly from Rosetta analysis (e.g., pI of the protein, pKa of individual amino acid at Y5 position etc). Table 5 shows the linear regression model coefficients for cellulose-I.

Co-efficient	Value
(Intercept)*	-3.80E+03
pI*	3.87E+03
pKaAA*	1.16E+04
RFU*	1.07E+04
REU*	-9.94E+04
Hpatch*	9.82E+03
pI:pKaAA*	-1.18E+04
pI:RFU*	-1.09E+04
pKaAA:RFU*	-1.75E+04
pI:REU*	1.01E+05
pKaAA:REU***	3.32E+02
RFU:REU***	1.38E+03
pI:Hpatch*	-9.97E+03
pKaAA:Hpatch	-4.91E+01
RFU:Hpatch***	-1.33E+02
REU:Hpatch*	-5.42E+01
pI:pKaAA:RFU*	1.79E+04
pKaAA:RFU:REU***	-2.54E+03
pKaAA:RFU:Hpatch***	2.51E+02

Significance legend for the parameters listed in the table:

*P < 0.05

**P < 0.01

***P < 0.001

Table 5: Regression model coefficients for fitted apparent binding affinity values for the library of GFP-CBM1 Y5 mutants to cellulose-I. The final R-Squared fitted value for the model was 0.9504.

The model is refined by using the "step" function in Rstudio to include only those parameters which are significant (refer to significance legend in Table 5) in contributing to explanining the variance in the binding seen for CBM1 towards cellulose-I. It should be noted that the 1st level interaction parameters pI and REU are the 'global' property terms for CBM1 protein while pKaAA & Hpatch scores are 'local' property terms that are thought to significantly contribute to the net binding of CBM1 to cellulose. Global property terms can

be used to explain binding interactions assuming the folded protein behaves as a charged colloidal sphere interacting with a charged surface, while the local property term can be used to explain higher resolution structural features of the protein not readily captured in global property terms. Figure 11 shown below is a scatter plot of relationship between the binding values versus the 1st level interaction parameters obtained from Rscript to have a better understanding of the trends various independent parameter chosen follow with the apparent partition coefficient data recorded. We can see in the plots below, that the binding increases with increases in RFU (which relates to the stability of the protein) and Hpatch score and the binding decreases with increase in REU.

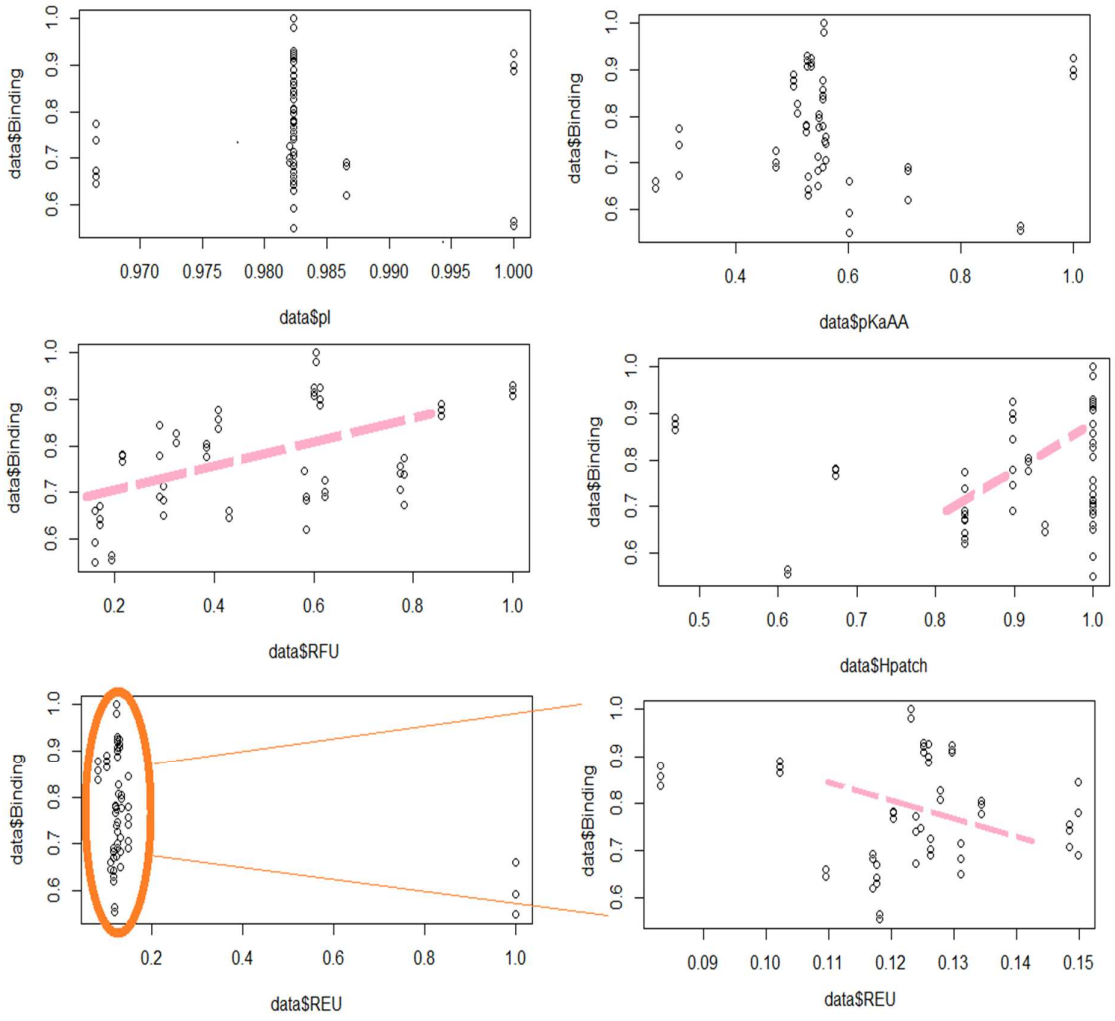


Figure 11: Scatter plots of apparent partition coefficients versus 1st level interaction parameters for cellulose-I included in the model. Inset for REU dataset (orange oval) is zoomed into on the x-axis scale and shown here as well. Dotted pink trend lines (with positive or negative slopes) have been plotted to aid the eye.

Further, a similar regression model was developed to fit the apparent partition coefficients of GFP-CBM1 wild type and the library of Y5 CBM1 mutants to cellulose-III with the same input variables. Table 6 below shows the model coefficients for the fit of apparent partition coefficient values of the CBM1 library to cellulose-III.

Co-efficient	Values
(Intercept)***	-5.48E+03
pI***	6.26E+03
pKaAA**	-5.23E+03
RFU***	-4.75E+03
REU***	7.41E+04
Hpatch**	-6.50E+02
pI:pKaAA**	4.08E+03
pI:RFU***	4.74E+03
pKaAA:RFU*	4.92E+03
pI:REU***	-8.11E+04
pKaAA:REU**	1.02E+04
RFU:REU*	1.39E+03
pKaAA:Hpatch*	1.19E+03
RFU:Hpatch	-9.35E+01
REU:Hpatch**	5.47E+03
pI:pKaAA:RFU*	-4.85E+03
pKaAA:RFU:REU**	-2.57E+03
pKaAA:RFU:Hpatch	2.01E+02
pKaAA:REU:Hpatch**	-1.00E+04

Significance legend for the parameters listed in the table:

*P < 0.05

**P < 0.01

***P < 0.001

Table 6: Regression model coefficients for fitted apparent binding affinity values for the library of GFP-CBM1 Y5 mutants to cellulose-III. The final R-Squared fitted value for the model was 0.9302.

The model is refined by using the "step" function in Rstudio to include only those parameters which are significant (refer to significance legend in Table 6) in contributing to explanining the variance in the binding seen for CBM1 towards cellulose-III. Similarly, figure 12 below is a scatter plot of relationship between the binding values versus the 1st level interaction parameters obtained from Rscript. Again, here we can see that the binding increases with RFU but decreases with Hpatch and REU. It can be attributed to the fact that cellulose-III is partially more hydrophilic than cellulose-I which is why the binding decreases for the proteins which have higher Hpatch score.

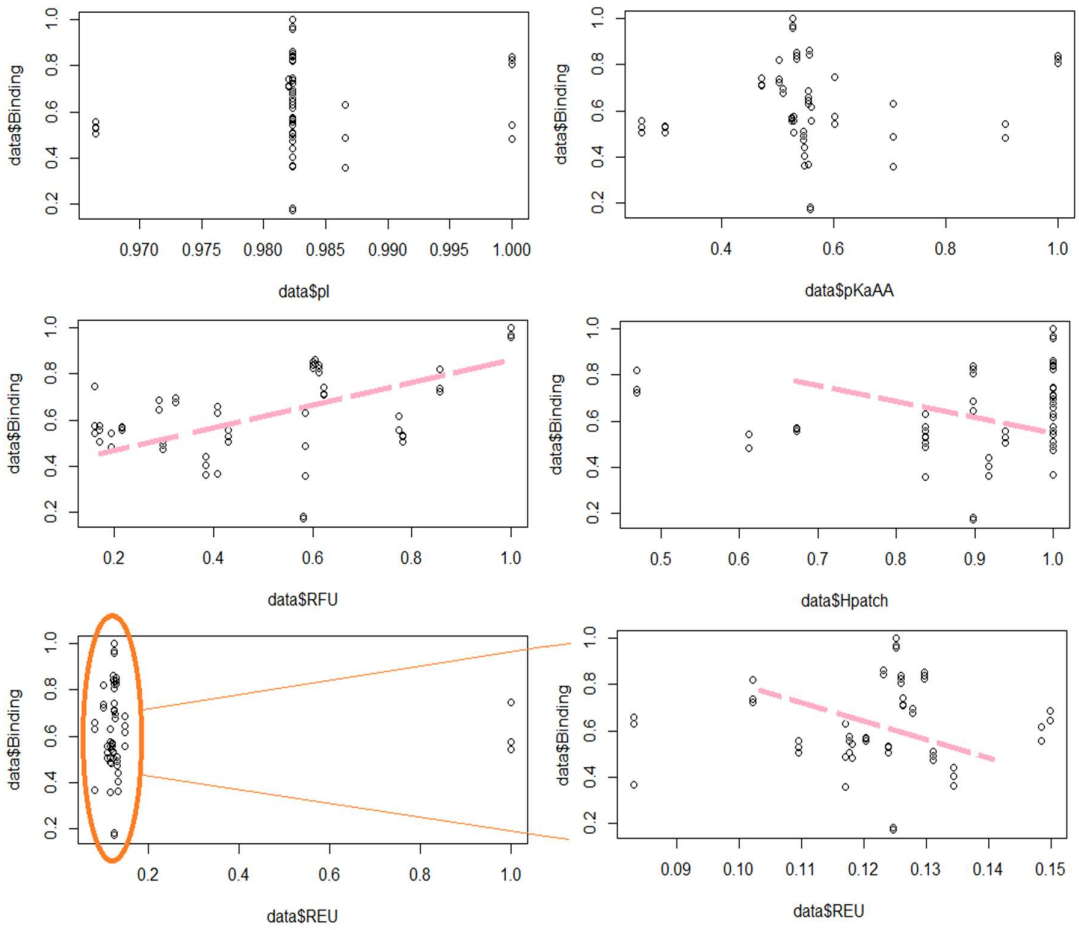


Figure 12: Scatter plots of binding partition coefficients versus 1st level interaction parameters for cellulose-III included in the model. Inset for REU dataset (orange oval) is zoomed into on the x-axis scale and shown here as well. Dotted pink trend lines (with positive or negative slopes) have been plotted to aid the eye.

Since, the Rosetta Energy Units (REU) score from *in silico* simulations is a summation of a various intrinsic structural properties of the folded mutant protein that could provide some structural insights in to how individual amino acid mutations could impact binding of CBM1 to cellulose. Therefore, the individual intrinsic parameters that were computed to estimate the overall REU score were deconvoluted and further explored to detailed structural insight into understanding how CBM structure impacts function and also drawing qualitative correlations with experimentally measured binding properties of the mutants towards both cellulose-I and cellulose-III. Table 5 listed below summarizes some of the intrinsic energetic terms and/or free energy relevant structural modeling parameters of all

the mutants of the library of GFP-CBM1 mutants, extracted from the *in silico* Rosetta based simulation studies.

Mutation	Total score	fa_elec	fa_rep	fa_sol	rama	H patch
A	-45.775	-0.006	-1.25	-2.788	0.186	7.04
C	-45.597	-0.185	-0.435	-1.085	0.063	7.84
D	-47.641	-1.626	-0.772	1.43	-0.16	7.36
E	-45.885	-0.259	-0.993	-1.098	0.155	6.56
F	-45.406	-0.413	-0.392	-0.054	-0.06	7.84
G	-50.849	-0.017	-2.32	-1.777	-0.751	7.04
H	-46.721	-0.828	-0.565	0.312	-0.129	6.56
I	-42.891	-0.447	-0.595	-1.151	0.768	7.84
K	-46.578	-0.036	-1.159	-1.634	-0.099	4.8
L	-45.983	-0.161	-0.722	-1.071	0.221	7.84
M	-45.172	-0.214	-0.881	-1.532	-0.025	7.84
N	-48.535	-1.129	-0.593	0.51	-0.412	3.68
P	60.995	-0.613	80.614	-2.189	2.103	7.84
Q	-46.314	-0.756	-1.018	-0.884	0.071	5.28
R	-45.638	-1.443	0.091	1.102	0.026	7.04
S	-46.633	-0.429	-1.231	-1.436	0.171	6.56
T	-45.011	-0.983	-1.267	-0.715	0.541	7.84
V	-42.712	-0.563	-1.013	-1.69	0.933	7.04
W	-44.594	-0.363	-0.098	1.346	0.132	7.2
Y	-45.719	-11.274	13.488	84.055	-2.578	7.84

Table 7: List of estimated *in silico* calculated structural and/or energetic term parameters extracted from Rosetta based structural modelling of CBM1 and its mutants.

The terms fa_sol, fa_elec, fa_rep and rama listed in the above table are used to gain more detailed structural insights into the variable binding behaviour of seen for the high vs. low binding mutant proteins seen in rank order list in Table 4. Since only single-point mutations were created on the protein, we expect that most of the variations seen in these computed Rosetta parameters largely arise due to the single amino acid substitution at the Y5 position (unless the mutation significantly destabilizes the overall protein fold, as

predicted in some cases). The term fa_sol represents Lazaridis-Karplus solvation energy which is computed using an isotropic model that describes it as the energy required to desolvate (remove contacting water) an atom when it is approached by a neighbouring atom [1]. For example, higher fa_sol value corresponds to lower solvated residue which indicates lower amounts of solvent present around the residue and similarly a lower fa_sol value corresponds to highly solvated residue, which means higher amounts of solvent would be present around the residue. This parameter is critical since it is known that CBM1-cellulose interaction is largely driven by the hydrophobic effect and thus less solvated residues will pave way for stronger hydrophobic effect driven interactions between the residue (and hence protein) and the ligand or cellulose chain. The term fa_elec represents Columbic electrostatic potential is a distance-dependent dielectric which indicates the energy of interaction between two non-bonded charged atoms separated by certain distance [1]. This can thus cause further destabilization of the amino acid residue at the Y5 position. The terms rama and fa_rep, represent the Ramachandran angle preferences and Lennard Jones repulsive energy (energy between two atoms on different residue separated by certain distance) are both energetically unfavourable.

3.3 Discussion

'Structure-determines-function' is the central tenet of structural biology and therefore availability of the structure of any protein/mutant is critical for developing a first-principles based understanding of its function. Generally, experimental protein structure determination is achieved using techniques such as X-ray crystallography/diffraction (XRD), nuclear magnetic resonance (NMR), and cryoelectron microscopy (Cryo-EM) that requires significant investments of both time and effort. Rosetta protein structural modeling software can more readily and inexpensively provide detailed structural information for proteins by predicting the high probability structures for any given amino acid sequence with a high degree of accuracy [15]. Here, we used structural predictions from Rosetta software for the single-site saturation library of CBM1 Y5 mutants and characterized the complex relationships of *in silico* structure modeling guided protein property predictions with actual experimental cellulose binding data.

Table 5 summarizes the regression coefficient values of the linear model which fits the output measured apparent partition coefficient values to various input computed variables like pI, pKaAA, RFU, REU and Hpatch for cellulose-I (Note: RFU was the only experimentally measured value used as input parameter). We can see from Table 5 that the regression coefficient value for the 1st interaction term of Rosetta Energy Units (REU) is negative. This makes sense since higher the overall energy score of the protein, lower its likely fold stability and hence lower expected binding to cellulose due to the destabilized protein structure. This is also confirmed with the fact that GFP-CBM1-Y5P is amongst the lowest binding mutant proteins to cellulose-I (Table 4). However, an overall REU score does not provide any detailed insights into the conformation of the amino acid at the planar binding motif and its interaction with other residues/solvent that could also drive increased/reduced binding interactions with cellulose. Also, the results for GFP-CBM1-Y5P

could be clearly interpreted in terms of the Ramachandran preferences (rama) and also the Lennard-Jones repulsive term (fa_rep) (Figure 14 & Figure 15). Both these terms are energetically unfavourable because of the ring structure of Proline side chain, possibly leading to an unstable conformation of overall CBM1. On the other hand, the regression coefficient value for the 1st interaction term of Hpatch is positive which means higher the hydrophobic patch score of the protein, higher the binding. This result is also justifiable since it is known that the binding of Type-A CBMs to cellulose-I is largely driven by the hydrophobic effect [9]. We see that higher Hpatch scored mutants, like GFP-CBM1 wild type/Y5L/Y5M, all show higher binding towards cellulose (Table 4). Similarly, the regression coefficient value for the 1st interaction term of pI (overall isoelectric point of the protein) is also positive, this means that higher the pI of the mutant CBM1, higher it's expected binding interactions with cellulose. From Figure 11, it can be seen that the plot of binding partition coefficients shown on the y-axis versus RFU plotted on the x-axis seems increase linearly that explains the positive coefficient trend. RFU corresponds to Relative Fluorescence Units normalized with respect to wild type expression results, which means that the level of overall expression of the protein inside the cell was normalized with respect to the wild-type protein. It is well known that the more stable a protein structurally, the more likely it will be well expressed in the cell and higher its expected yield. Since RFU is showing a positive trend in the regression results, it means with higher RFU values for the mutants greater binding is seen. This makes sense again since higher RFU likely correlates to greater protein stability which could explain the overall improved binding seen towards cellulose-I. For example, proteins like GFP-CBM1-Y5P which are energetically less stable lie in the bottom of the table of RFU (Table 1). This mutant also shows up in the bottom of the table of rank order of the mutants for binding to cellulose-I (Table 4). While, proteins

like GFP-CBM1 wild type/N/R which are among the top ones in Table 1 with high RFU score, also show higher binding to cellulose as well.

We hypothesized also that the energy contribution terms from Rosetta modeling might yield additional insights into the binding behaviour of GFP-CBM1 mutants which do not follow the trend based on the empirical linear regression model that lacks a clear mechanistic basis. This is because hydrophobic effect driven interactions between the protein and ligand are a combination of van der Waals interactions coupled with solvation effects caused by the exclusion of solvent from interface. Rosetta energy terms such as fa_sol (for solvation) can be used to help better understand such interactions. For instance, a low fa_sol score can be interpreted as reduced solvation of the mutant surface residues. GFP-CBM1-Y5R and GFP-CBM1-Y5N are two mutants that have a positive solvation energy (fa_sol), indicating that these mutant proteins are most likely not highly solvated. This indicates the possibility of solvent exclusion near the binding interface although more information may be required to definitively make a claim about the same. In the future, we intend to obtain more information from Rosetta for understanding how specific residues that may be interacting with amino acid position 5 using PyRosetta which provides improved functionalities as compared to Rosettascripts to extract structural information from the *in silico* simulations. However, in the case of GFP-CBM1-Y5L and GFP-CBM1-Y5M, the solvation energy terms are negative, and we currently cannot point to a single energetic term which might explain or be responsible for increased binding seen towards the ligand.

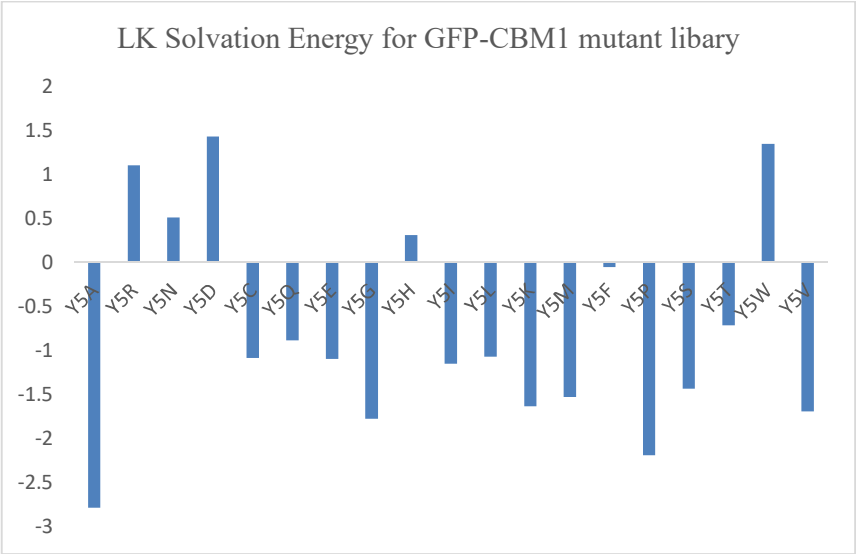


Figure 13: Lazaridis-Karplus solvation energy (fa_sol) calculated using Rosetta software for all the Y5 mutants after subtraction of wild type energy score.

In the case of low cellulose binding mutants, GFP-CBM1-Y5K has the least columbic electrostatic potential of all mutants, possibly due to the resulting electrostatic repulsion with H4 (Histidine residue at position 4) which might be causing the resulting structure to be unstable hence leading to reduced binding to cellulose. It is important to note that this unfavourable columbic electrostatic potential is not observed in the case of Y5R (which could partly explain why these mutants give better binding to cellulose).

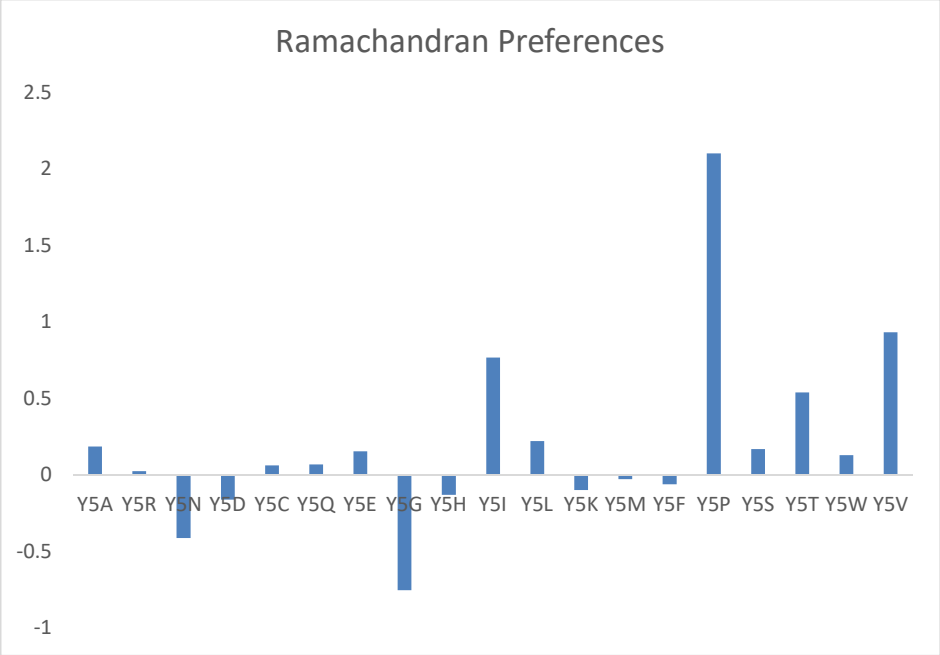


Figure 14: Ramachandram preference (rama) calculated using Rosetta software for all the Y5 mutants after subtraction of wild type energy score.

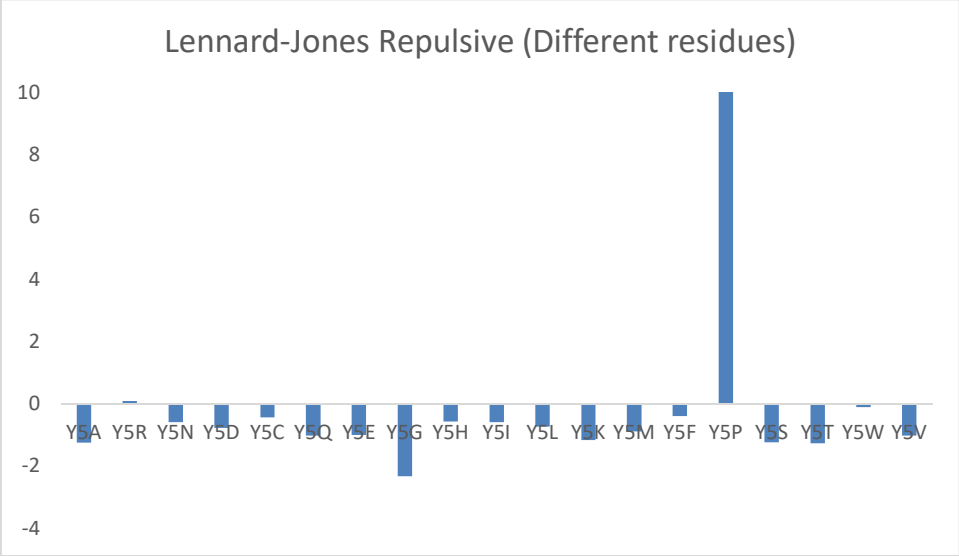


Figure 15: Lennard-Jones repulsive between atoms in different residues (fa_rep) of all the Y5 mutants subtracted from wild type score.

In the case of cellulose-III, from Table 6, we see that the sign of Hpatch has turned negative. It implies that lower the hpatch score of the protein, higher the expected binding towards cellulose. This is likely considering that cellulose-III has a slightly hydrophilic

surface and has increased propensity to form H-bonds with bulk water or solutes as shown in a recent MD simulation report [8]. This could explain why increased H-bonding capacity for some CBM1 mutants could show increased binding towards cellulose-III. This suggests that the binding interactions between type-A CBMs and cellulose-III is likely not driven as strongly by the hydrophobic effect as seen for cellulose-I. On-going MD simulations between CBM1-cellulose-III vs. CBM1-cellulose-I are shedding further light into the complex role of enthalpic versus hydrophobic interactions driving binding interactions in both cases. From figure 12, we can also see that even for cellulose-III, increase in RFU increases the binding. Higher RFU indicates higher expression stability and with higher stability, we could expect the protein to possibly showed better binding to the substrate (Note: RFU is a much more complex variable that depends on not only overall protein stability but also toxicity of the protein to *E. coli* that hence impacts overall expression levels). In Table 4, proteins like GFP-CBM1 wild type, Y5N/R have higher RFU values and also showed higher binding whereas mutant proteins like GFP-CBM1 Y5K/S/P which have lower RFU end up lower in the table of rank order for binding to cellulose-III. In the case binding of GFP-CBM1 Y5W to cellulose-III, its rank order ended up at the bottom of the table. We believe this likely happens because Tryptophan (W) being a double ringed amino acid, and its bulkier side chains cause massive steric clashes with the stepped-like surface of cellulose-III. This could explain this anomalous result as to why this mutant gave lower binding to wards cellulose-III and also colloborates with ongoing CBM1-cellulose-III MD simulation studies (unpublished data).

It is critical to note that here the protein structure model was used to compute its various biophysical parameters which were then correlated to explain the apparent binding affinity seen towards cellulose-I and cellulose-III. However, correlation is not causation and

in no way are these Rosetta modeling derived protein structural properties fully causative of the experimentally observed functional binding measurements reported here.

Chapter 4: Future studies

Three decades of research on Family 1 CBMs have unearthed information about the structure of the CBM, the functionality of the CBM in the process of cellulose degradation, and the mechanism of its adsorption to carbohydrates. Studies have been performed detailing the effect of single point mutations (typically to alanine alone) on the structure of CBM and its affinity to the surface of cellulose. The effect of physical parameters affecting binding of the CBM to cellulose-I have allowed to develop conditions of optimal affinity and catalytic activity of the appended catalytic domain. Structural changes to cellulose-I such as the ammonia pre-treatment process that results in the formation of an unnatural allomorph called cellulose-III is another alternative substrate currently being explored as a substrate to produce cellulosic biofuels. However, while the change in the hydrogen bonding interactions in cellulose-III allow for better accessibility for some endocellulase enzymes to the surface of the polysaccharide, this does not guarantee an increase in binding of the enzyme to the substrate surface. During the course of this thesis, a lower binding of CBM1 (and the Y5 mutant library) to cellulose-III has also been clearly observed. This provides direct evidence for the reduced binding of Cel7A to cellulose-III that has been reported in the literature.

Furthermore, we have studied in detail the effect of single point site saturation mutations at Y5 position of GFP-CBM1 on adsorption to both cellulose-I and cellulose-III and those effects were compared with that of the wild type protein. The Y5 position was chosen based on previous studies where it was shown that this is an important residue on the flat binding face of CBM1 which contributes to the binding activity of overall protein. It was also hypothesized that the interaction of Y5 with H4 gives the 5^{th} amino acid residue its stability. Future work should include exploring the effects of how the mutations on the rest

of the flat face residues viz. Y31 and Y32 can drive the binding of CBM1 to cellulose-I and cellulose-III.

The purification strategies, developed originally by Ms. Vibha Narayanan in her MS thesis work [21], used a 2-step purification with Immobilized Metal Affinity Chromatography (IMAC) and Hydrophobic Interaction Chromatography (HIC) to get a maximum yield of CBM1 proteins. Parameters such as resin media, flow rate, concentration of the buffers and gradient length can be further tweaked to further fine-tune the purification method. A full scale protein loading adsorption studies need to be done to get detailed insights into the maximum number of binding sites available on the surface of ligand as well as to determine the dissociation, therefore affinity constant [12]. Studies should also be conducted to confirm the reversibility of binding of CBM1 Y5 mutants to cellulose allomorphs.

With regards to the binding vs. hydrolysis activity relationship for full-length cellulases, work has already been started to generate a library of CelE-CBM1 mutants where the GFP is replaced with catalytically active CelE domain. Half the library has already been successfully cloned. Once the library is fully generated, a small scale activity assay should be performed to check the relative activity of the proteins and select a few mutants to be grown, purified, and characterized on a large scale to determine clearly the relationship between binding and activity with the mutation on Y5 position of CBM1.

Appendix

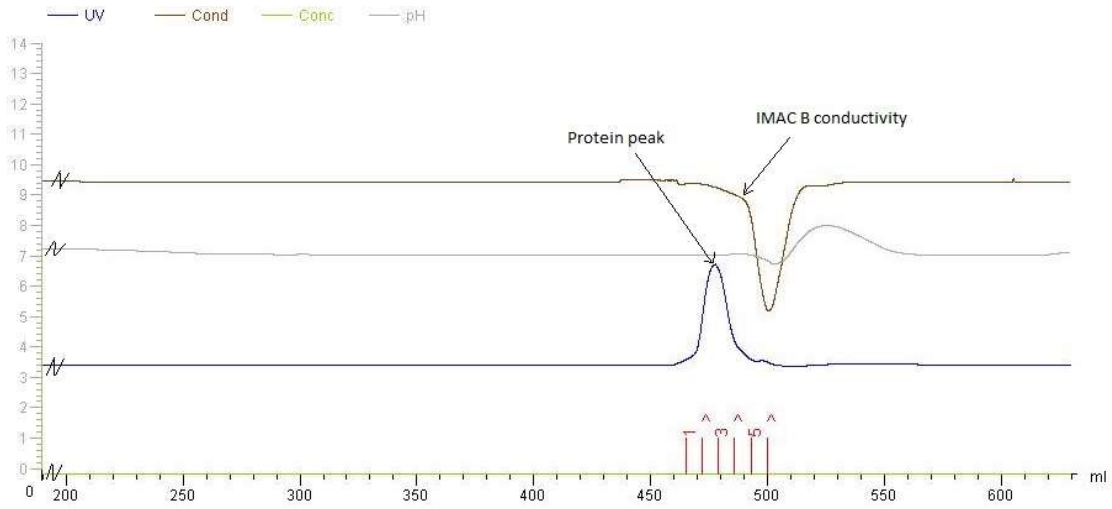


Figure A1: Example chromatogram showing desalting of GFP-CBM1 Y5K into desired buffer after IMAC purification. The protein peak represents GFP-CBM1 Y5K containing fractions eluting off the column followed by IMAC B as pointed in the figure. Only those fractions were pooled where there were no traces of IMAC B.

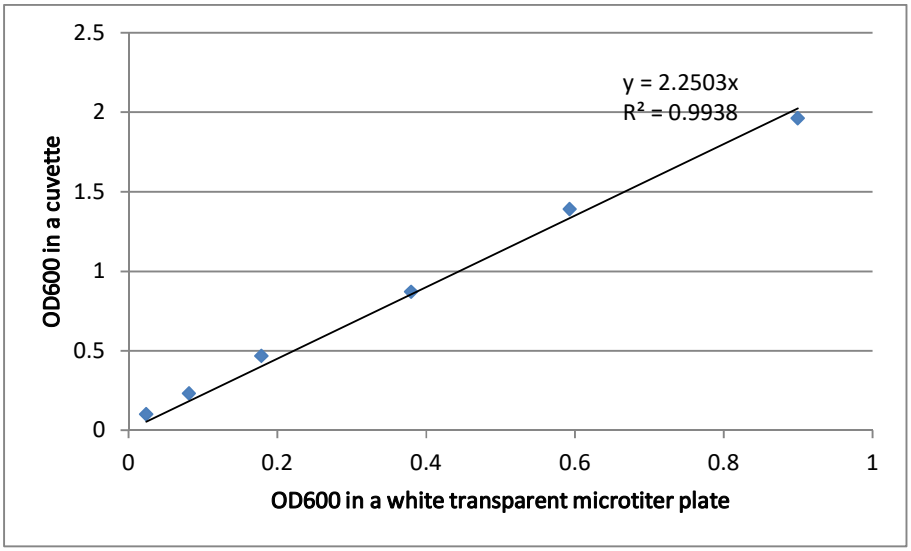


Figure A2: Calibration curve used to convert optical density of a culture medium (at 600 nm) measured in a white transparent microtiter well plate to standard cuvettes used to estimate cell density. A total of 6 samples were used with 3 replicates each with varying cell concentrations and a blank control to generate the above curve. The numbers here denote the actual OD values measured.

	Wild type	Y5N		Wild type	Y5I
Mean	43945.28	37425.63	Mean	43945.28	34443.94
Variance	34529432	10852140	Variance	34529432	43136192
Observations	3	3	Observations	3	3
Hypothesized Mean Difference	0		Hypothesized Mean Difference	0	
df	3		df	4	
t Stat	1.676276		t Stat	1.867374	
P(T<=t) one-tail	0.096139		P(T<=t) one-tail	0.067621	
t Critical one-tail	2.353363		t Critical one-tail	2.131847	
P(T<=t) two-tail	**0.192278**		P(T<=t) two-tail	**0.135241**	
t Critical two-tail	3.182446		t Critical two-tail	2.776445	

	Wild tye	Y5E		Wild type	Y5C
Mean	43945.28	34477.96	Mean	43945.28	27136.62
Variance	34529432	29360608	Variance	34529432	9407654
Observations	3	3	Observations	3	3
Hypothesized Mean Difference	0		Hypothesized Mean Difference	0	
df	4		df	3	
t Stat	2.051499		t Stat	4.392161	
P(T<=t) one-tail	0.054752		P(T<=t) one-tail	0.010934	
t Critical one-tail	2.131847		t Critical one-tail	2.353363	
P(T<=t) two-tail	**0.109504**		P(T<=t) two-tail	**0.021867**	
t Critical two-tail	2.776445		t Critical two-tail	3.182446	

Table A1: Student's t-test performed on Culture Fluorescence Unit normalized with OD600 for Wild type with Y5N, Y5I, Y5E and Y5C. The mean value indicates CFU/OD600 for the respective wild type or mutant-protein. The t-test was performed assuming unequal variance and alpha as 0.05.

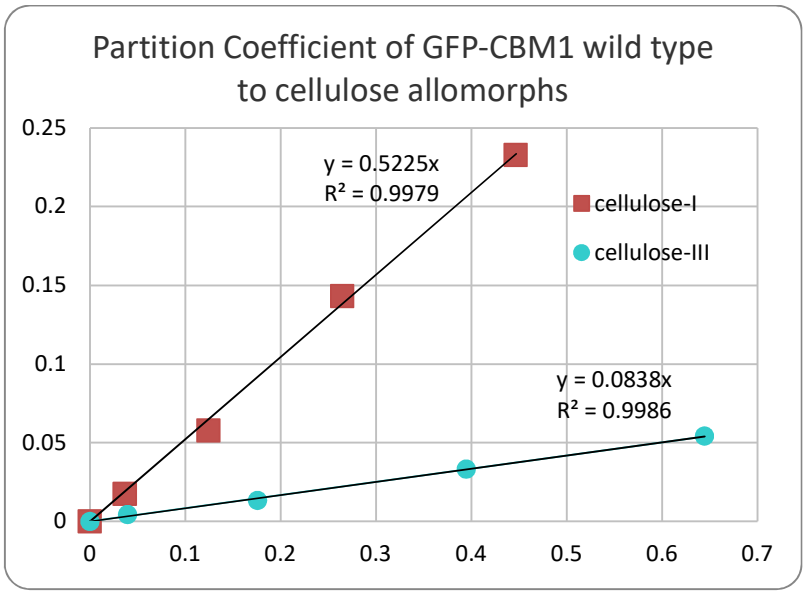


Figure A3: Full scale partition coefficient assay for GFP-CBM1 wild type to cellulose-I and cellulose-III. The slope of the respective trendlines represent the partition coefficient value in liters/gram for respective cellulose. The assay was done at pH 6.5 with effective concentration of 10 mM MES in the well. Protein loading for cellulose-I ranges from 0 to 11 mg of protein per gram cellulose-I and from 0 to 2.75 mg of protein per gram cellulose-III.

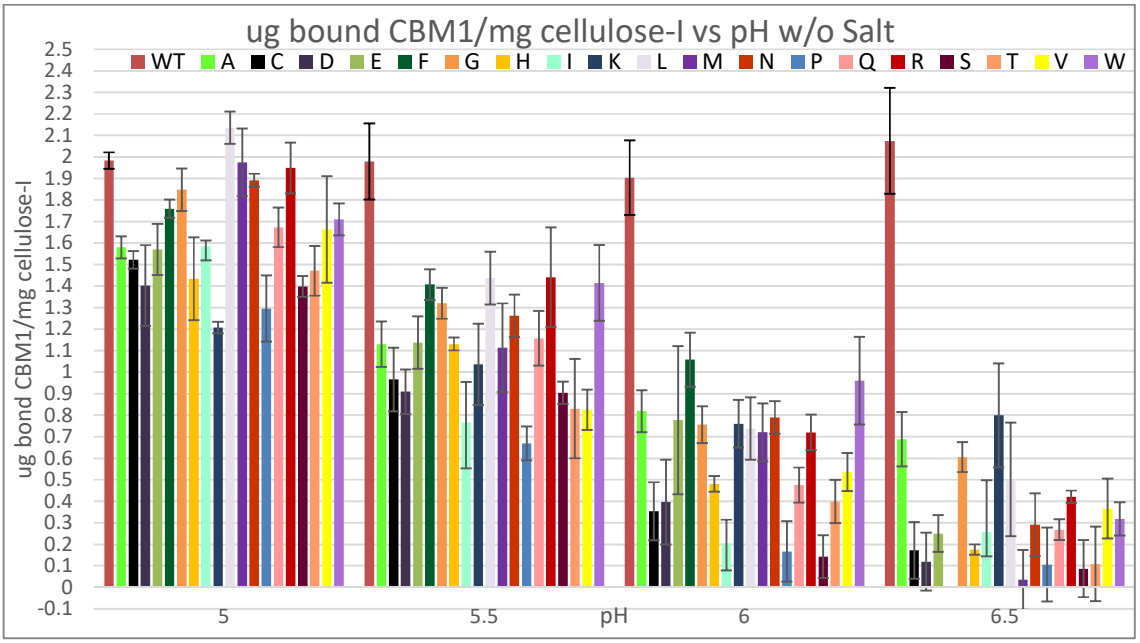


Figure A4: Plot of apparent partition coefficient for the library of GFP-CBM1 mutants vs various pHs for cellulose-I without the presence salt.

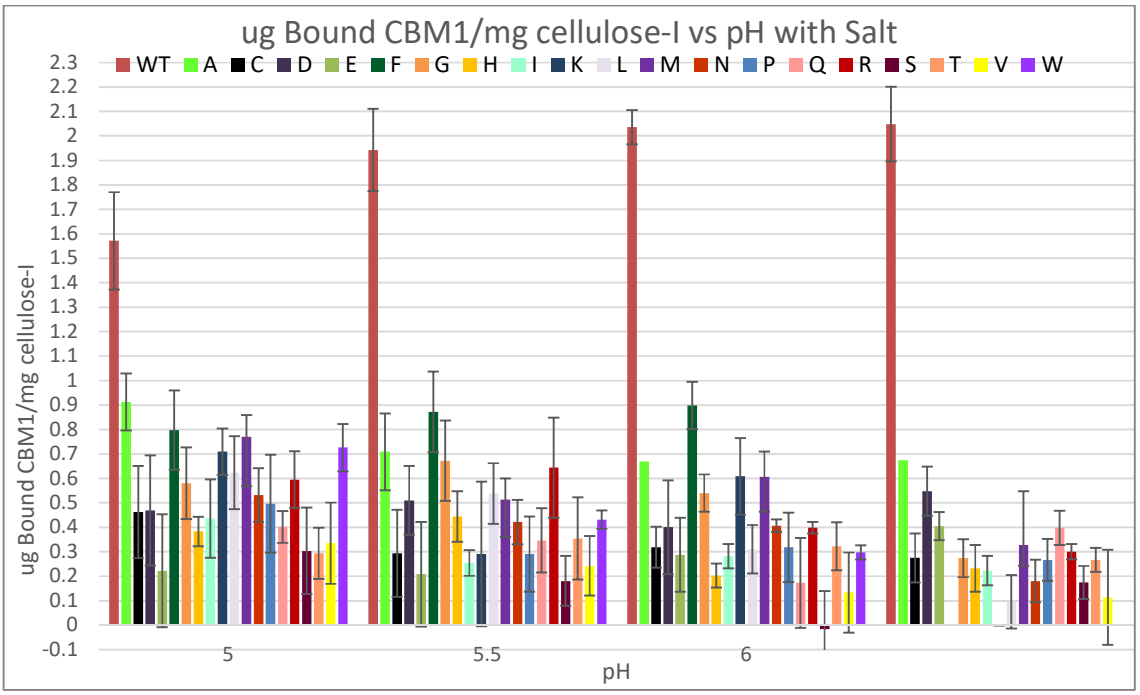


Figure A5: Plot of apparent partition coefficient for the library of GFP-CBM1 mutants vs various pHs for cellulose-I in the presence of 100mM NaCl salt.

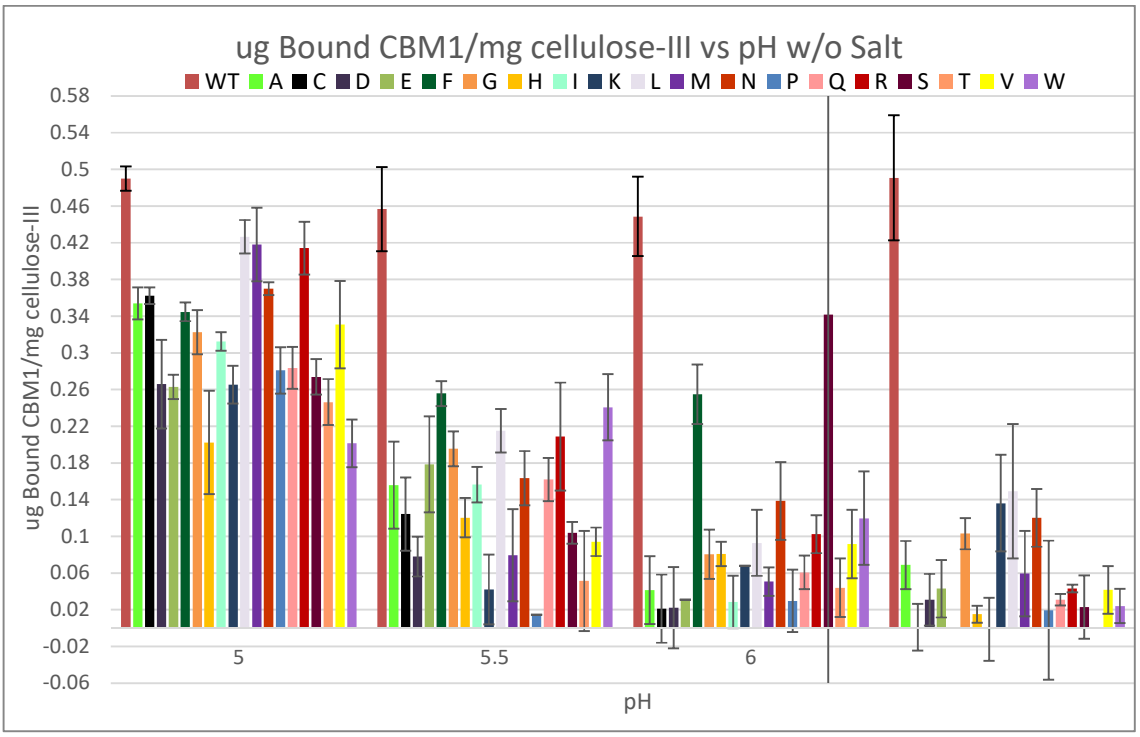


Figure A6: Plot of apparent partition coefficient for the library of GFP-CBM1 mutants vs various pHs for cellulose-III without the presence salt.

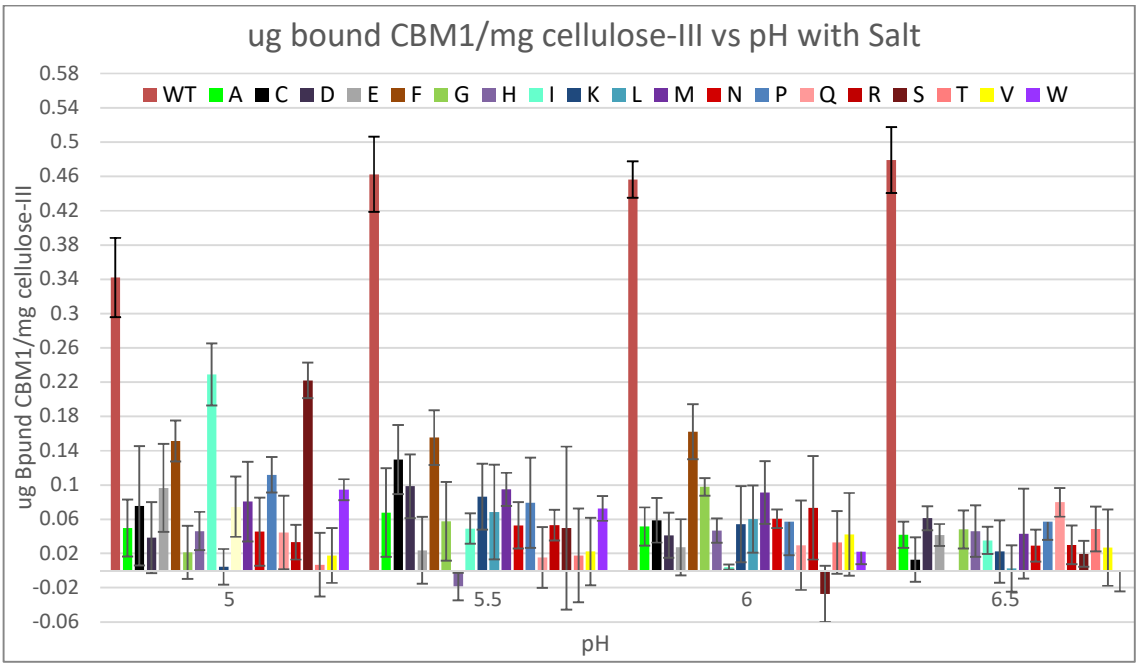


Figure A7: Plot of apparent partition coefficient for the library of GFP-CBM1 mutants vs various pHs for cellulose-III in the presence of 100mM NaCl salt.

Commands for Rosettascript:

#!/bash/org

/home/path/Rosetta/main/source/bin/rosetta_scripts.static.linuxgccrelease -s /home/path/GFP-CBM1.pdb
database /home/path/Rosetta/main/database -parser:protocol /home/path/GEL_fastrelax.xml
@/home/path/flagsfile2

Flag Flies for Rosettascript:

-ex1
-ex2
-auto_setup_metals
-overwrite
-extrachi_cutoff 1
-linmem_ig 10
-use_input_sc

RosettaScript:

<ROSETTASCRIPTS>
<SCOREFXNS>
<ScoreFunction name="myscore" weights="talaris2014_cst.wts"/>
</SCOREFXNS>

```
<RESIDUE_SELECTORS>
</RESIDUE_SELECTORS>
<TASKOPERATIONS>
        <RestrictToRepacking name="restrict"/>
        <InitializeFromCommandline name="init"/>
        <IncludeCurrent name="curr"/>
</TASKOPERATIONS>
<FILTERS>
</FILTERS>
<MOVERS>
        <AtomCoordinateCstMover name="coord_cst" coord_dev="0.5" bounded="true"
        bound_width="0.2" sidechain="false" native="false" task_operations="restrict" />
        <FastRelax name="fastrelax" repeats="10" scorefxn="myscore"
        task_operations="restrict,init,curr" />
</MOVERS>
<PROTOCOLS>
        <Add mover_name="coord_cst" />
        <Add mover_name="fastrelax"/>
</PROTOCOLS>
</ROSETTASCRIPTS>
```

Mutation	A	C	D	E	F	G
total_score	-45.78	-45.60	-47.64	-45.89	-45.41	-50.85
dslf_fa13	-2.47	-2.49	-2.51	-2.48	-2.49	-2.48
fa_atr	-131.49	-133.40	-134.77	-133.96	-137.92	-132.71
fa_dun	27.06	28.01	29.88	30.50	31.13	27.42
fa_elec	-11.28	-11.46	-12.90	-11.53	-11.69	-11.29
fa_intra_rep	0.27	0.27	0.27	0.27	0.30	0.27
fa_rep	12.24	13.05	12.72	12.50	13.10	11.17
fa_sol	81.27	82.97	85.49	82.96	84.00	82.28
hbond_bb_sc	-2.59	-2.61	-1.90	-2.51	-2.51	-2.86
hbond_lr_bb	-9.26	-9.51	-9.39	-9.24	-9.44	-9.69
hbond_sc	0.00	0.00	-1.32	0.00	0.00	0.00
hbond_sr_bb	-4.87	-4.92	-5.06	-4.86	-4.75	-5.10
omega	2.63	2.43	2.72	2.82	2.52	2.34
p_aa_pp	-8.50	-8.69	-9.28	-8.79	-8.46	-9.87
pro_close	0.10	0.09	0.02	0.10	0.09	0.09
rama	-2.39	-2.52	-2.74	-2.42	-2.64	-3.33
ref	3.51	3.18	1.10	0.77	3.35	2.91
yhh_planarity	0.00	0.01	0.03	0.00	0.01	0.01

Mutation	H	I	K	L	M	N
total_score	-46.72	-42.89	-46.58	-45.98	-45.17	-48.54
dslf_fa13	-2.53	-2.50	-2.49	-2.49	-2.49	-2.46
fa_atr	-137.06	-133.77	-133.97	-134.88	-134.05	-134.38
fa_dun	30.54	29.56	29.12	28.31	29.72	29.08
fa_elec	-12.10	-11.72	-11.31	-11.44	-11.49	-12.40
fa_intra_rep	0.27	0.39	0.27	0.27	0.27	0.27
fa_rep	12.92	12.89	12.33	12.77	12.61	12.90
fa_sol	84.37	82.90	82.42	82.98	82.52	84.57
hbond_bb_sc	-1.86	-1.92	-2.46	-2.48	-2.54	-1.90
hbond_lr_bb	-9.21	-9.55	-9.23	-9.54	-9.25	-9.34
hbond_sc	-1.26	-1.16	0.00	0.00	0.00	-1.37
hbond_sr_bb	-4.69	-5.26	-4.83	-4.75	-4.88	-5.04
omega	2.59	2.84	2.82	2.46	2.66	2.62
p_aa_pp	-9.24	-7.64	-9.04	-8.43	-8.74	-9.67
pro_close	0.10	0.02	0.10	0.09	0.09	0.02
rama	-2.71	-1.81	-2.68	-2.36	-2.60	-2.99
ref	3.12	3.81	2.37	3.49	2.98	1.54
yhh_planarity	0.02	0.03	0.00	0.01	0.00	0.03

Mutation	P	Q	R	S	T	V
total_score	61.00	-46.31	-45.64	-46.63	-45.01	-42.71
dslf_fa13	-2.37	-2.50	-2.51	-2.47	-2.49	-2.48
fa_atr	-133.90	-134.05	-136.34	-131.91	-131.28	-131.99
fa_dun	40.18	30.59	30.95	27.21	27.88	28.31
fa_elec	-11.89	-12.03	-12.72	-11.70	-12.26	-11.84
fa_intra_rep	0.29	0.27	0.28	0.27	0.37	0.38
fa_rep	94.10	12.47	13.58	12.26	12.22	12.48
fa_sol	81.87	83.17	85.16	82.62	83.34	82.37
hbond_bb_sc	-1.90	-2.46	-1.88	-2.54	-1.98	-1.92
hbond_lr_bb	-8.69	-9.23	-9.68	-9.25	-9.75	-9.55
hbond_sc	-1.07	-0.49	-1.30	-0.66	-1.80	-1.17
hbond_sr_bb	-4.86	-4.84	-4.73	-4.84	-5.16	-4.65
omega	6.06	3.05	2.11	2.60	2.75	2.73
p_aa_pp	-7.32	-9.06	-8.51	-8.81	-7.81	-7.57
pro_close	8.46	0.10	0.08	0.10	0.02	0.10
rama	-0.48	-2.51	-2.55	-2.41	-2.04	-1.65
ref	2.48	1.22	2.41	2.90	2.93	3.71
yhh_planarity	0.03	0.00	0.01	0.00	0.03	0.02

Mutation	W	Wild Type
total_score	-44.59	-45.72
dslf_fa13	-2.49	-2.48
fa_atr	-140.08	-138.41
fa_dun	31.84	31.54
fa_elec	-11.64	-11.27
fa_intra_rep	0.29	0.30
fa_rep	13.39	13.49
fa_sol	85.40	84.06
hbond_bb_sc	-2.62	-2.67
hbond_lr_bb	-9.47	-9.78
hbond_sc	0.00	0.00
hbond_sr_bb	-4.84	-5.01
omega	2.42	2.61
p_aa_pp	-8.42	-8.51
pro_close	0.09	0.11
rama	-2.45	-2.58
ref	3.97	2.90
yhh_planarity	0.01	0.01

Table A2: List of all the estimated parameters extracted from Rosetta based structural modelling of CBM1 and its mutants.

dslf_fa13	Disulfide geometry potential
fa_atr	Lennard-Jones attractive between atoms in different residues
fa_dun	Internal energy of sidechain rotamers as derived from Dunbrack's statistics
fa_elec	Coulombic electrostatic potential with a distance-dependent dielectric
fa_intra_rep	Lennard-Jones repulsive between atoms in the same residue
fa_rep	Lennard-Jones repulsive between atoms in different residues
fa_sol	Lazaridis-Karplus solvation energy
hbond_bb_sc	Sidechain-backbone hydrogen bond energy
hbond_lr_bb	Backbone-backbone hbonds distant in primary sequence
hbond_sc	Sidechain-sidechain hydrogen bond energy
hbond_sr_bb	Backbone-backbone hbonds close in primary sequence
omega	Omega dihedral in the backbone. A Harmonic constraint on planarity with standard deviation of ~6 deg.
p_aa_pp	Probability of amino acid, given torsion values for phi and psi
pro_close	Proline ring closure energy and energy of psi angle of preceding residue
rama	Ramachandran preferences (with separate lookup tables for pre-proline positions and other positions)
ref	Reference energy for each amino acid. Balances internal energy of amino acid terms. Plays role in design.
yhh_planarity	A special torsional potential to keep the tyrosine hydroxyl in the plane of the aromatic ring

Table A3: Description of all the Rosetta parameters mentioned in Table A2.

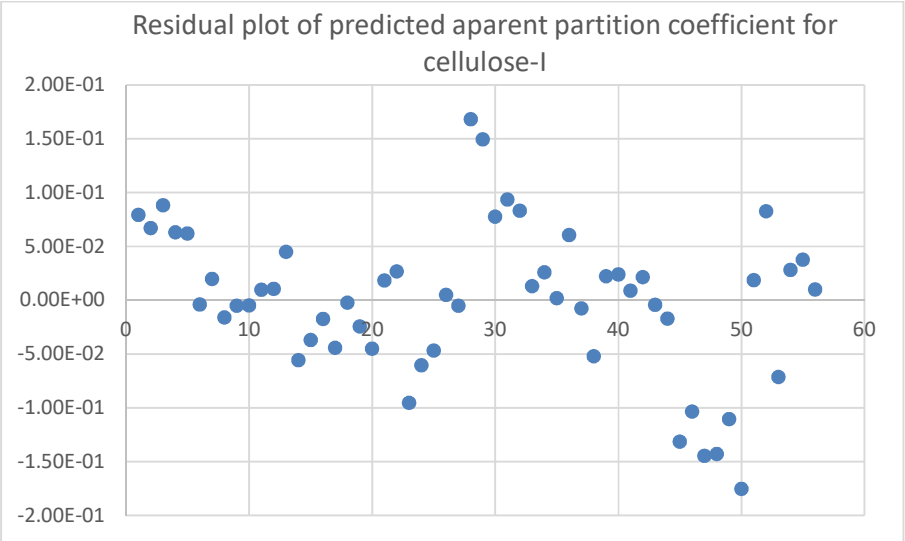


Figure A8: Plot of residuals for predicted versus actual apparent partition coefficient determined for cellulose-I dataset.

From Figure A8, it can be seen that majority of the residual values are below +/-0.1, and that the R-squared value is 0.9504, thus we can conclude that the model is accurate.

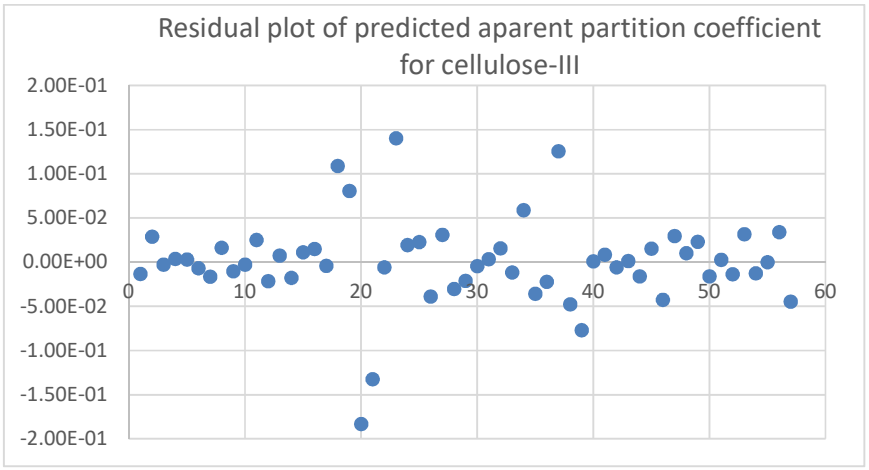


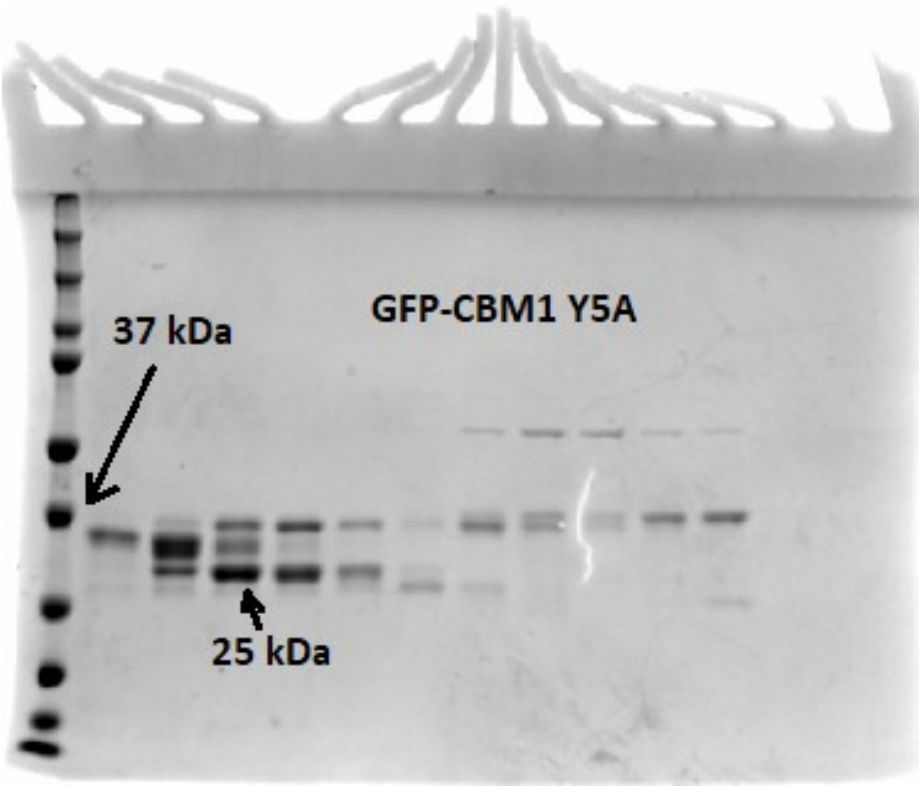
Figure A9: Plot of residuals for predicted versus actual apparent partition coefficient determined for cellulose-III dataset.

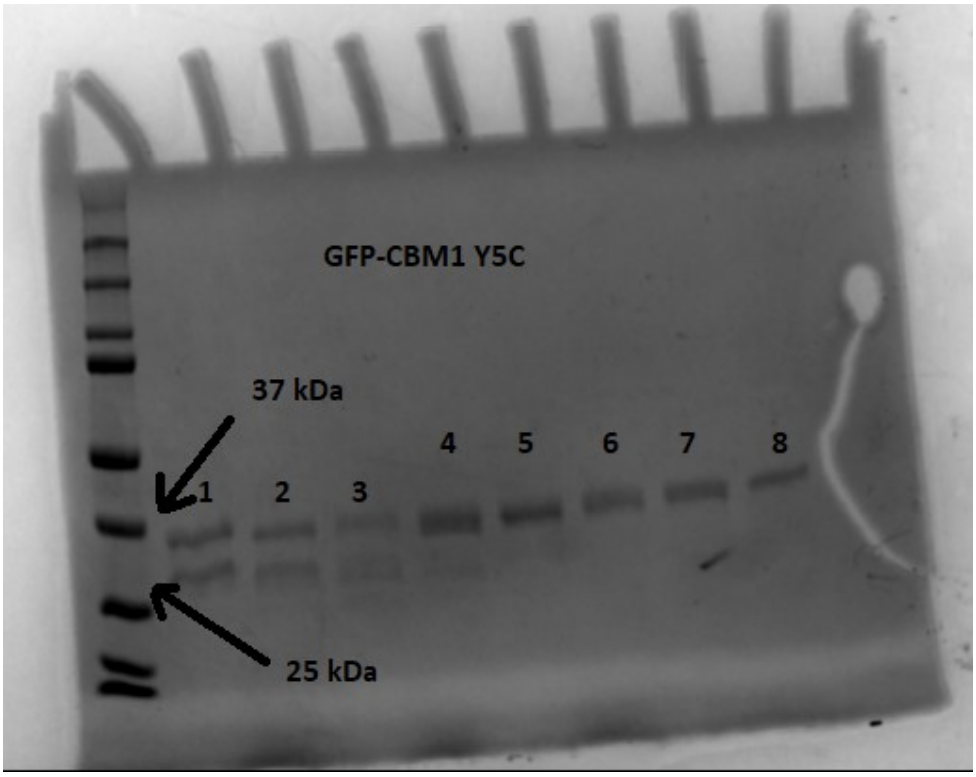
We can again see, from Figure A9, that the residual of very few sample set is above +/-0.1 and the R-squared being 0.9302.

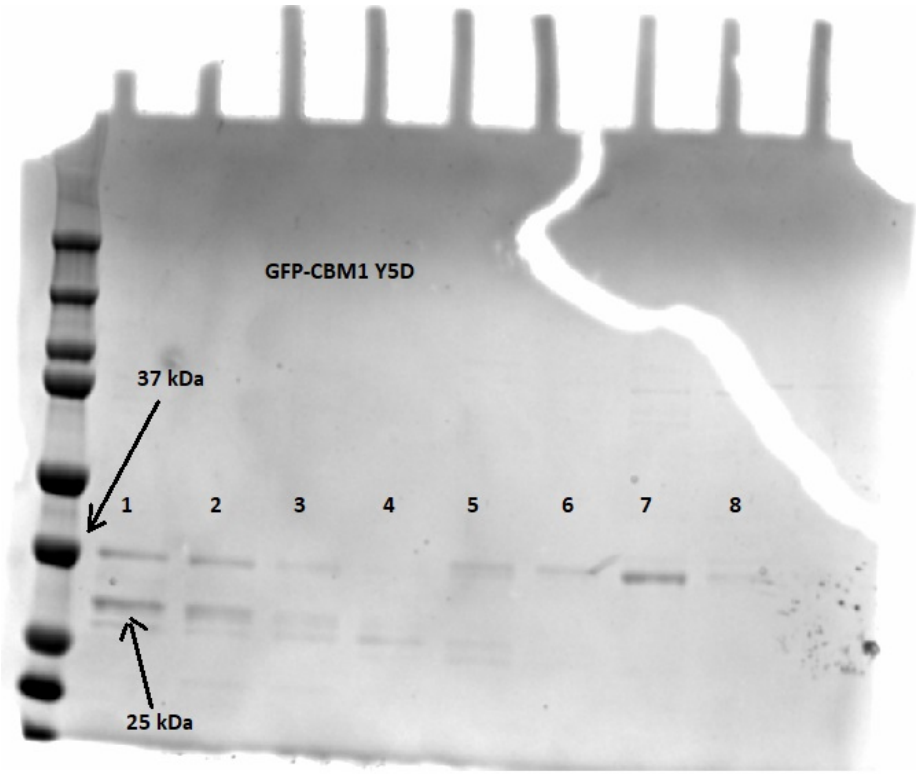
Wildtype	Number of binding sites N (µmoles/g cellulose)	N Error	Dissociation constant K_d (µM)	K_d Error	Adjusted R-square for dataset fit to model
Cellulose-I	0.95	0.10	0.22	0	0.979
Cellulose-III	0.226	0.05	0.448	0.188	0.978

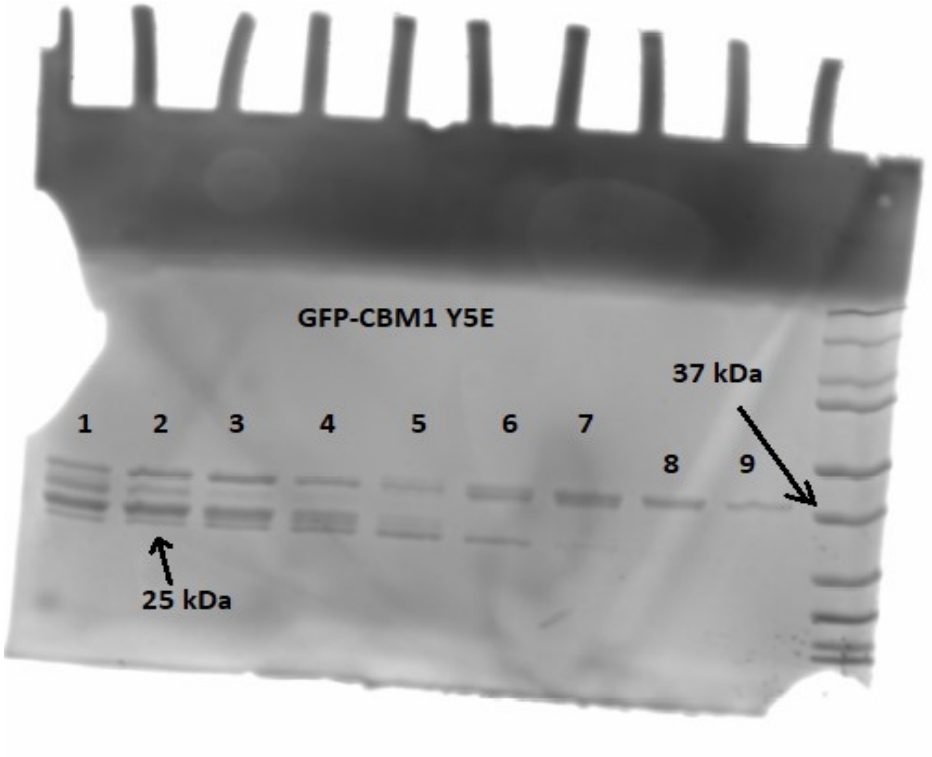
Table A4: Langmuir single-site model based adsorption parameters determined from solid-depletion based full-scale protein loading binding assays of select GFP-CBM1 Y5 wild type with cellulose-I and cellulose-III.

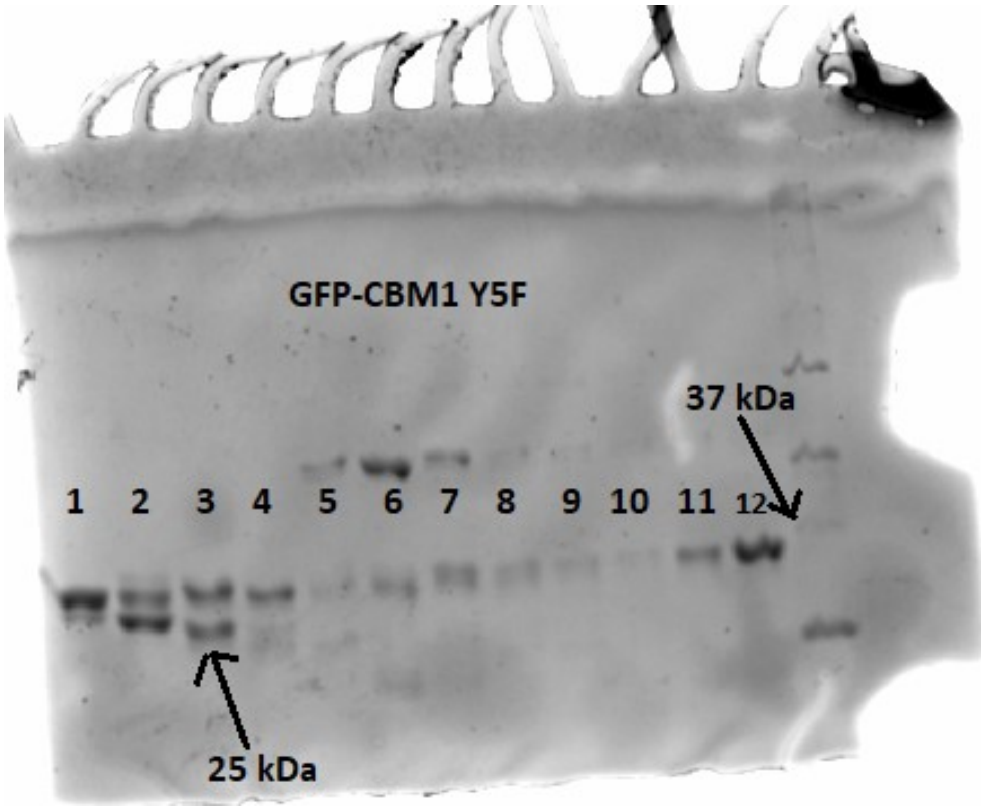
Figure A10 Below are the SDS_PAGE image of the library of purified GFP-CBM1 mutants. Ladder represents Biorad protein ladder. The numbers on top represent the fraction numbers. GFP-CBM1s typically shows up as a thick band at 37kDa. Mutants GFP-CBM1 Y5N and GFP-CBM1 Y5W were purified by Ms Vibha Narayanan (Narayanan, V. et al [21])

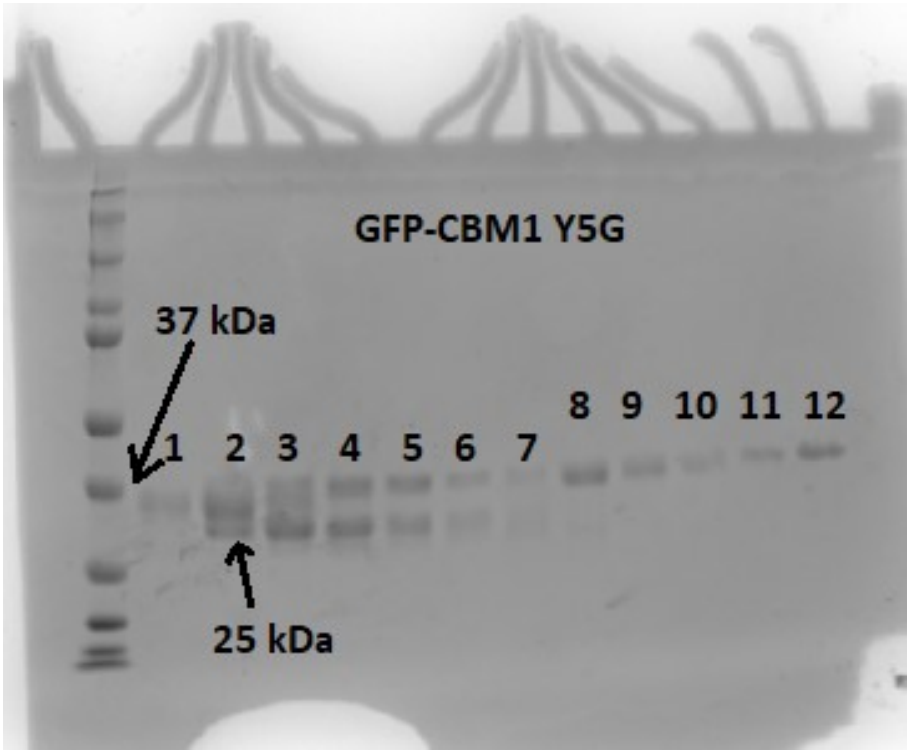


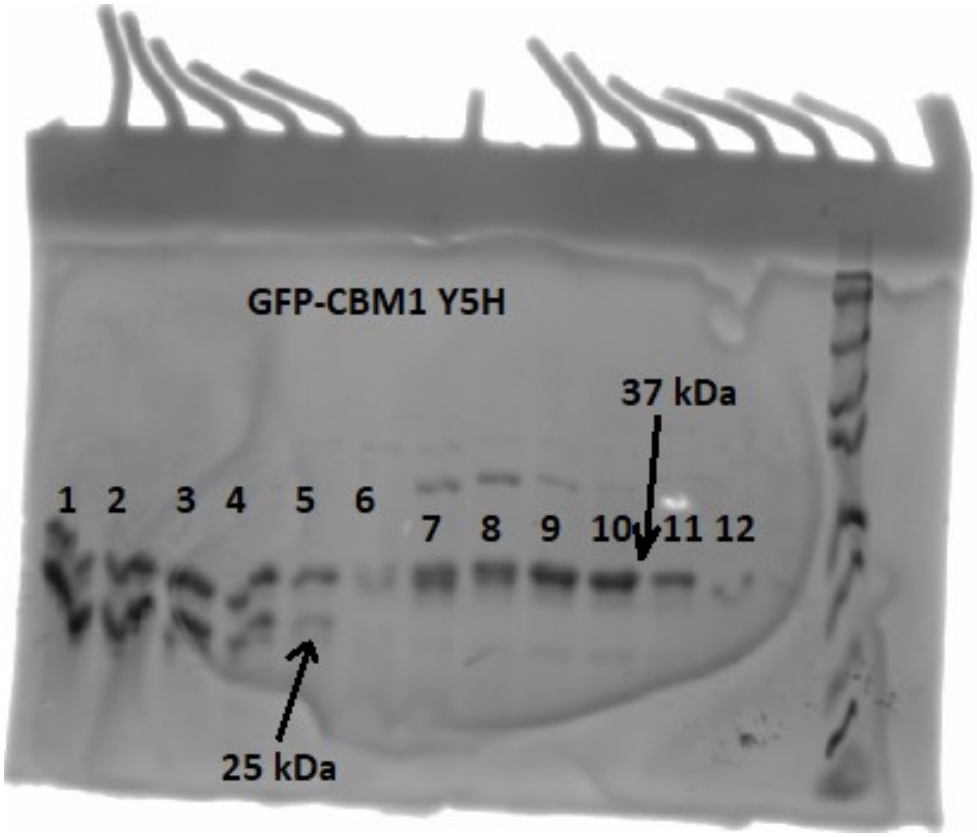

GFP-CBM1 Y5C
37 kDa
1
2
3
4
5
6
7
8
25 kDa

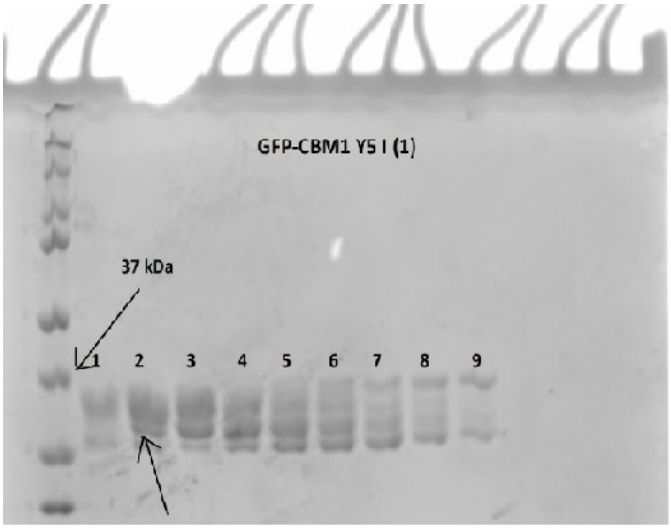

GFP-CBM1 Y5D
37 kDa
1
2
3
4
5
6
7
8
25 kDa

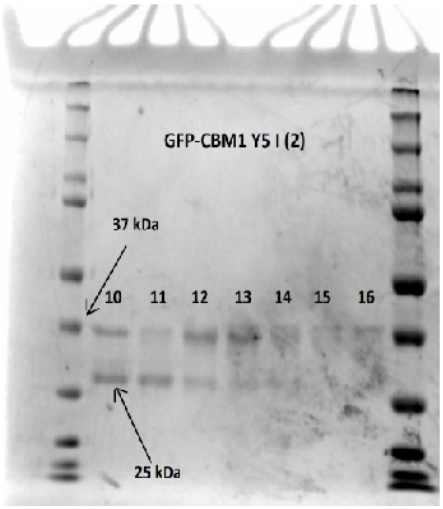

GFP-CBM1 Y5E
37 kDa
1 2 3 4 5 6 7 8 9
25 kDa

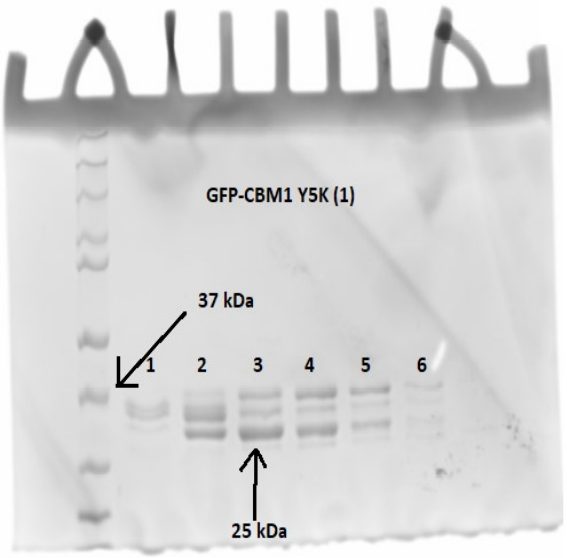

GFP-CBM1 Y5F
37 kDa
1 2 3 4 5 6 7 8 9 10 11 12
25 kDa

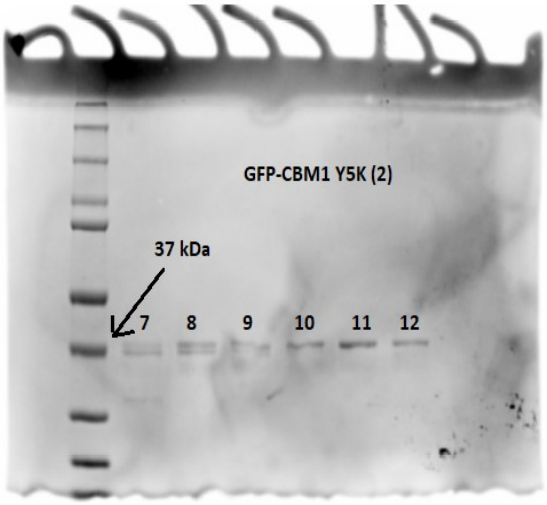

GFP-CBM1 Y5G
37 kDa
1 2 3 4 5 6 7 8 9 10 11 12
25 kDa

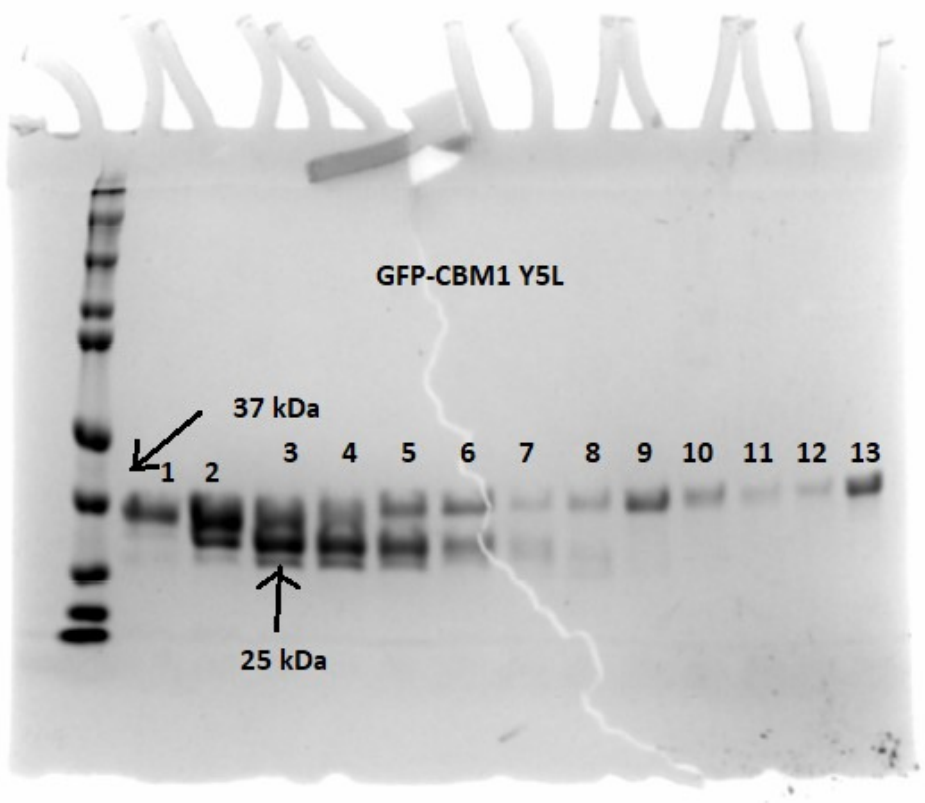

GFP-CBM1 Y5H
37 kDa
1 2 3 4 5 6 7 8 9 10 11 12
25 kDa

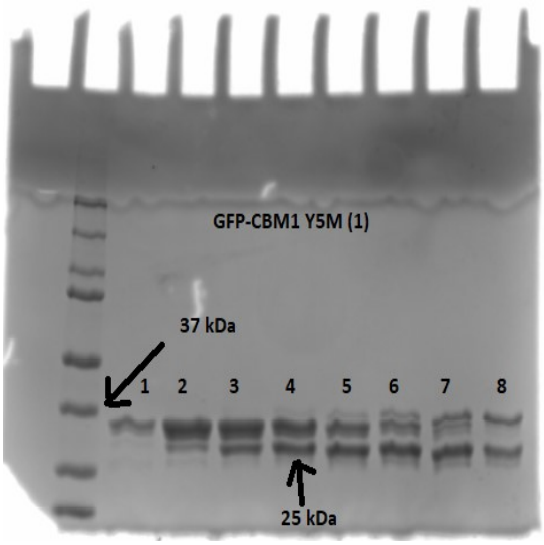

GFP-CBM1 Y5 I (1)
37 kDa
1 2 3 4 5 6 7 8 9

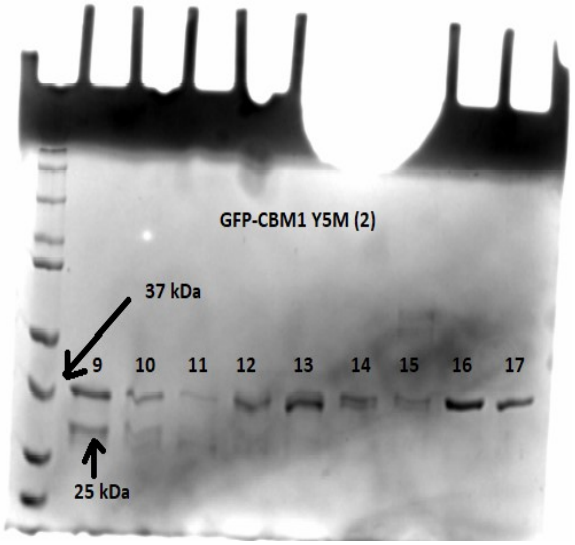

GFP-CBM1 Y5 I (2)
37 kDa
10 11 12 13 14 15 16
25 kDa

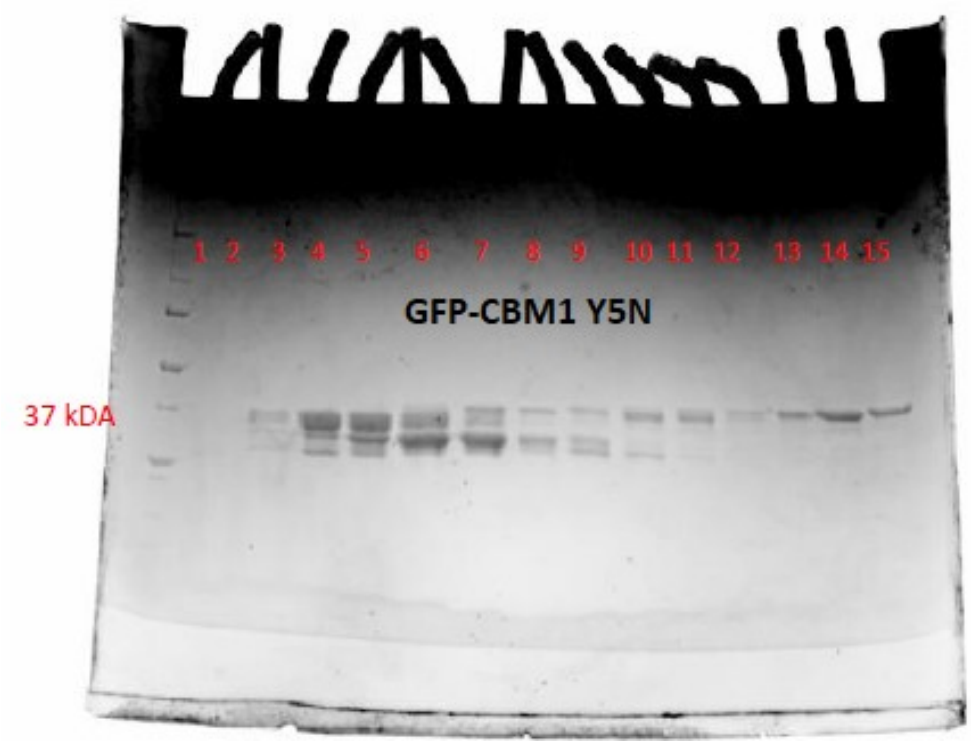

GFP-CBM1 Y5K (1)
37 kDa
1 2 3 4 5 6
25 kDa

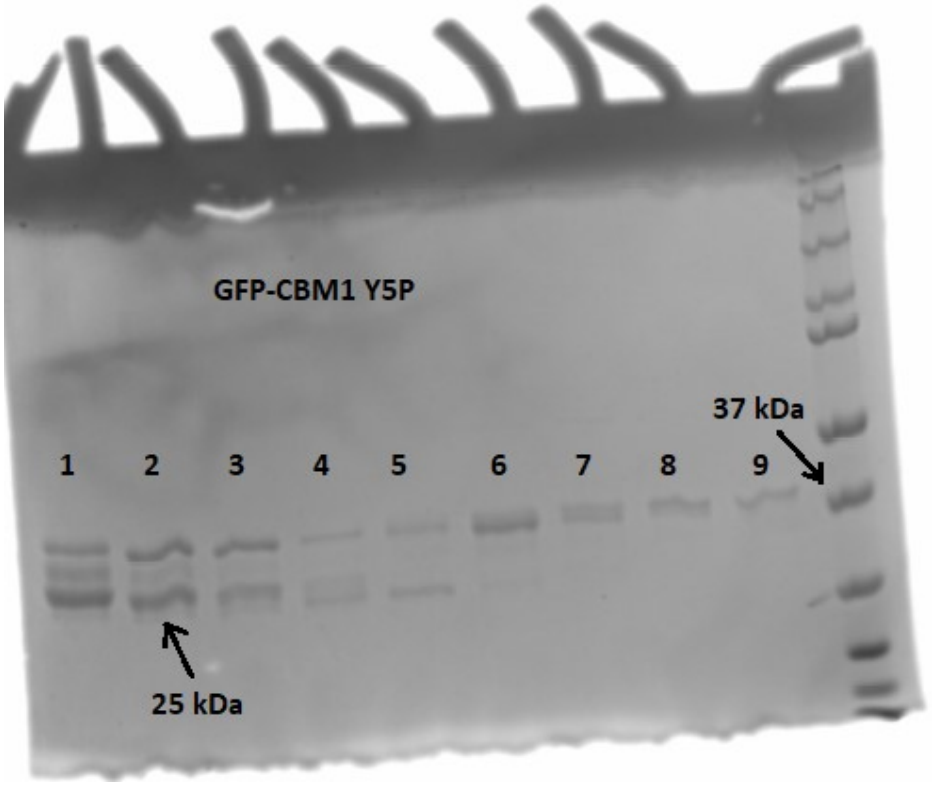

GFP-CBM1 Y5K (2)
37 kDa
7 8 9 10 11 12


GFP-CBM1 Y5L
37 kDa
1
2
3
4
5
6
7
8
9
10
11
12
13
25 kDa

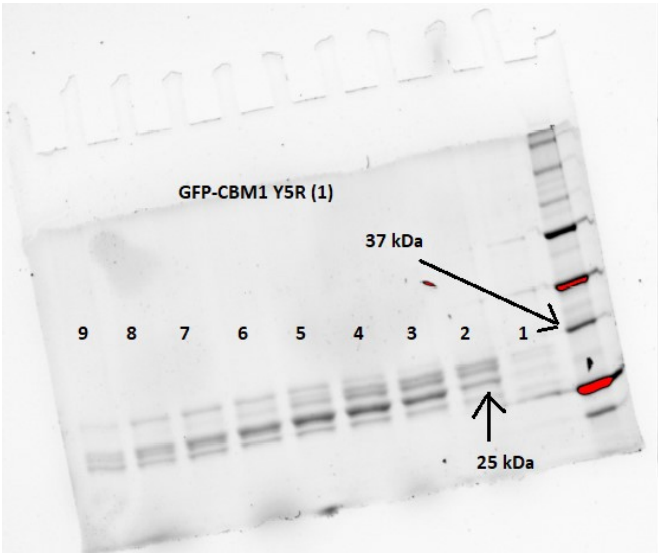

GFP-CBM1 Y5M (1)
37 kDa
1
2
3
4
5
6
7
8
25 kDa

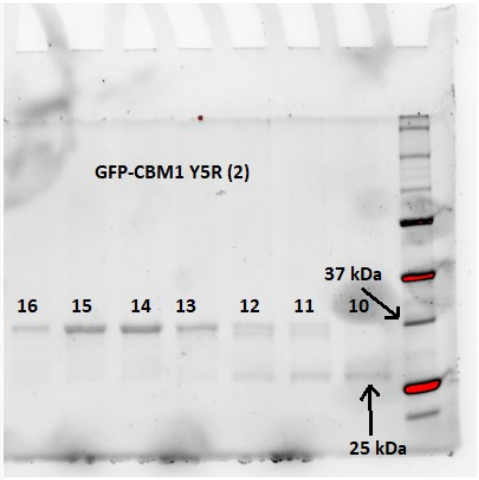

GFP-CBM1 Y5M (2)
37 kDa
9
10
11
12
13
14
15
16
17
25 kDa

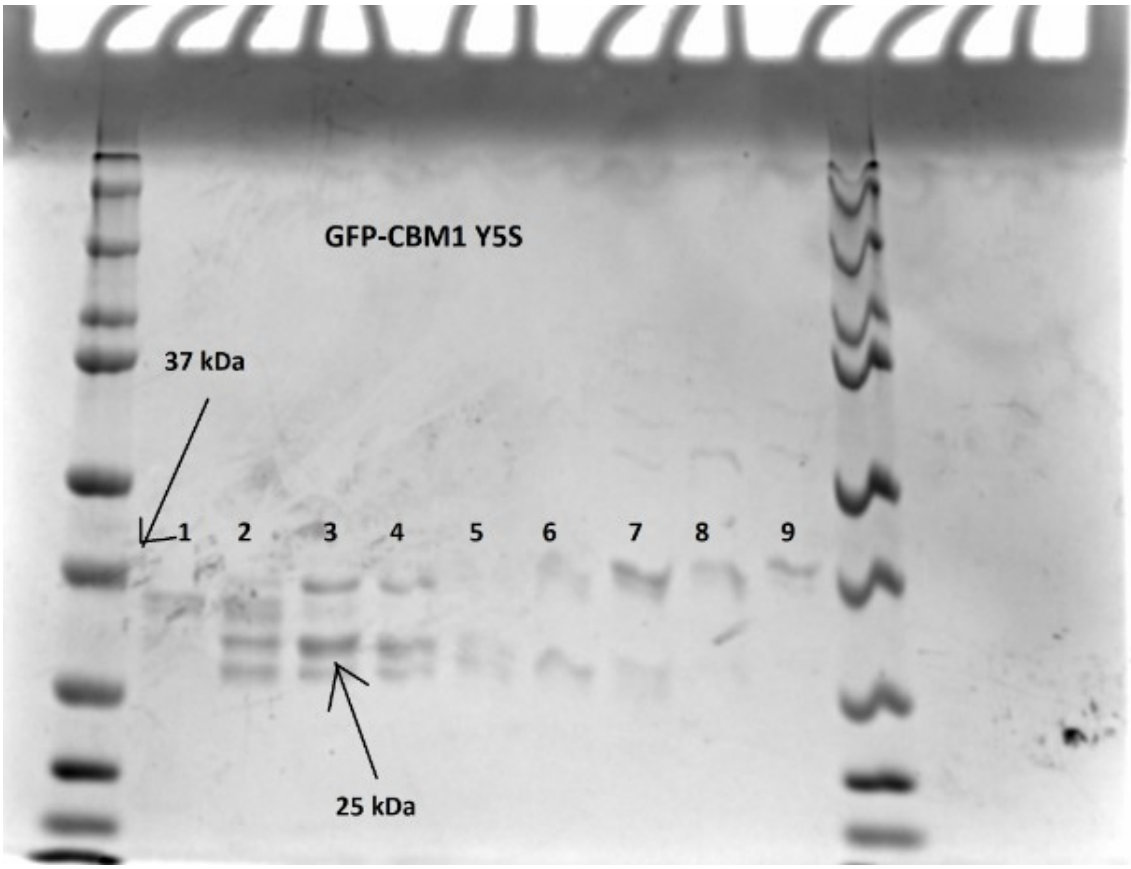

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
GFP-CBM1 Y5N
37 kDA

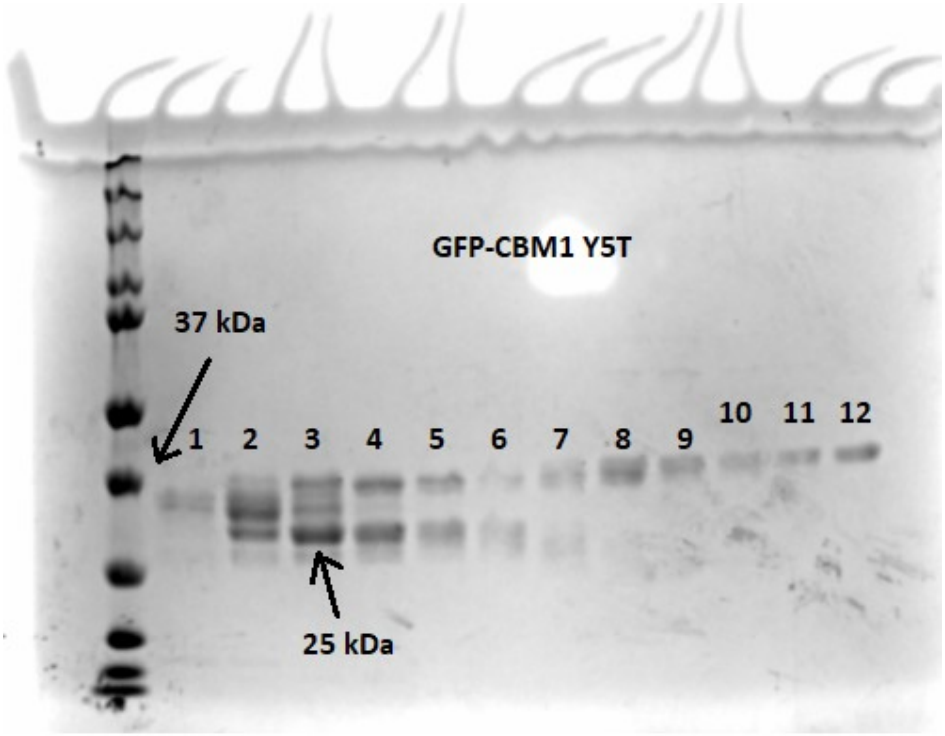

GFP-CBM1 Y5P
37 kDa
1 2 3 4 5 6 7 8 9
25 kDa

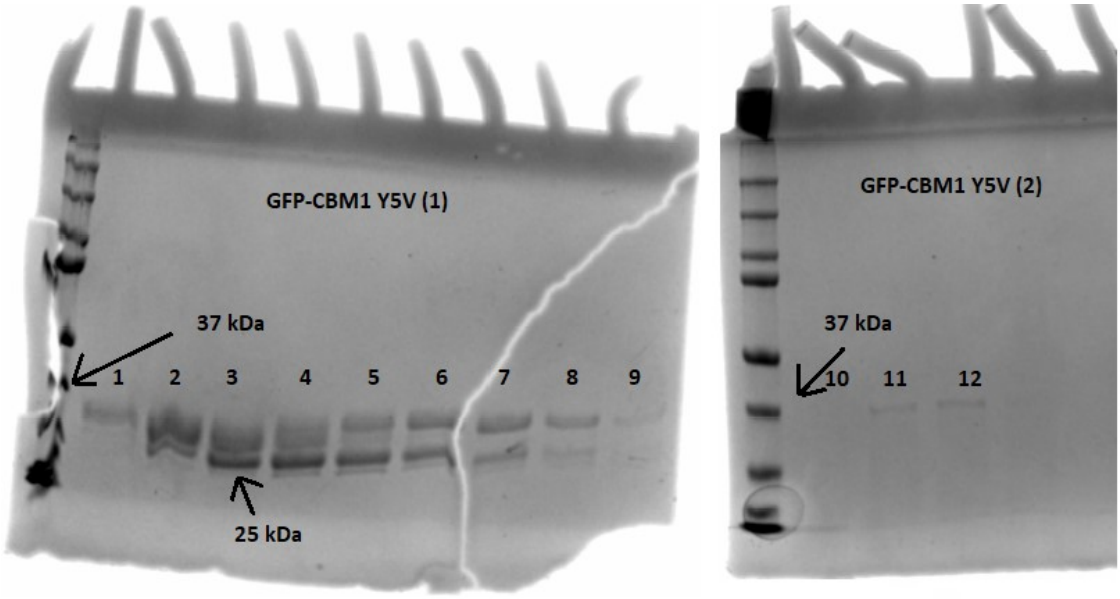

GFP-CBM1 Y5Q
37 kDa
1 2 3 4 5 6 7 8 9 10 11
25 kDa

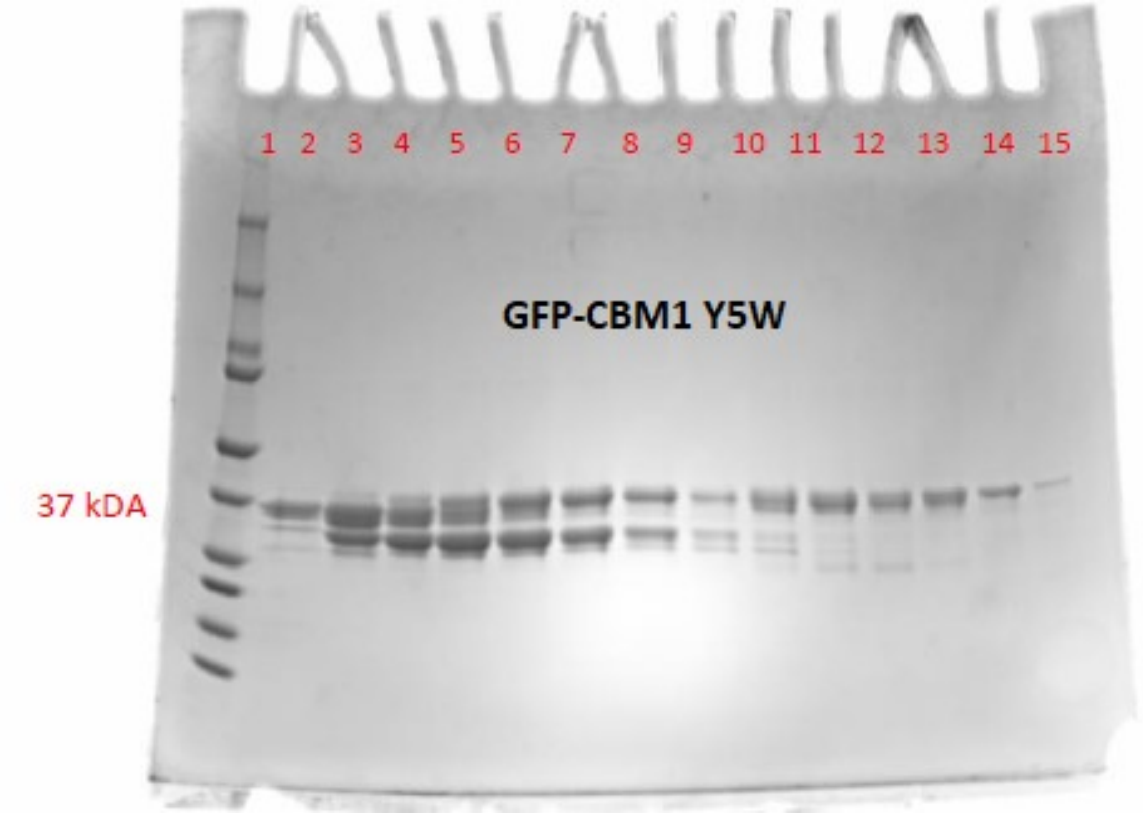

GFP-CBM1 Y5R (1)
37 kDa
9 8 7 6 5 4 3 2 1
25 kDa

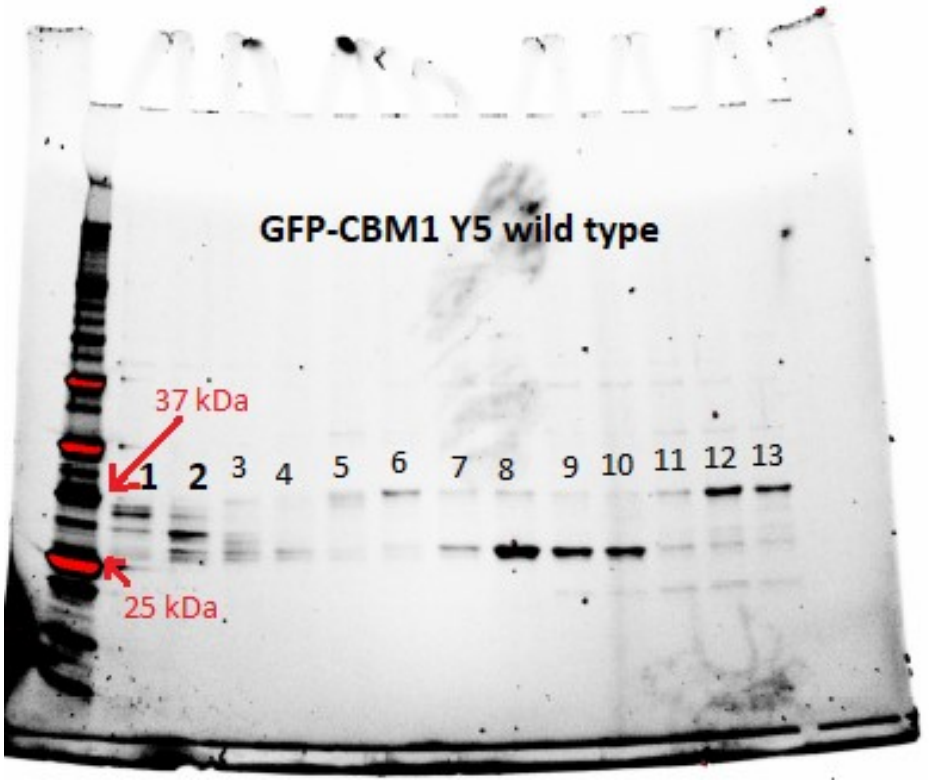

GFP-CBM1 Y5R (2)
37 kDa
16 15 14 13 12 11 10
25 kDa


GFP-CBM1 Y5S
37 kDa
1 2 3 4 5 6 7 8 9
25 kDa


GFP-CBM1 Y5T
37 kDa
1 2 3 4 5 6 7 8 9 10 11 12
25 kDa


GFP-CBM1 Y5V (1)
37 kDa
1 2 3 4 5 6 7 8 9
25 kDa
GFP-CBM1 Y5V (2)
37 kDa
10 11 12


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
GFP-CBM1 Y5W
37 kDA


GFP-CBM1 Y5 wild type
37 kDa
25 kDa
1 2 3 4 5 6 7 8 9 10 11 12 13

References

1. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13,** 3031–3048 (2017). doi:10.1021/acs.jctc.7b00125

2. Beckham, G. T. *et al.* Identification of amino acids responsible for processivity in a family 1 carbohydrate-binding module from a fungal cellulase. *J. Phys. Chem. B* **114,** 1447–1453 (2010). doi:10.1021/jp908810a

3. Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J.* **382,** 769–781 (2004). doi:10.1042/BJ20040892

4. Bornhorst, J. & Falke J. Purification of Proteins Using Polyhistidine Affinity Tags. *Methods in Enzymology*, *2000, 326*, 245–254.

5. Brady, S. K., Sreelatha, S., Feng, Y., Chundawat, S. P. S. & Lang, M. J. Cellobiohydrolase 1 from Trichoderma reesei degrades cellulose in single cellobiose steps. *Nat. Commun.* **6,** 1–9 (2015). doi: 10.1038/ncomms10149

6. Chen, L. *et al.* Specificity of O-glycosylation in enhancing the stability and cellulose binding affinity of Family 1 carbohydrate-binding modules. *Proc. Natl. Acad. Sci.* **111,** 7612–7617 (2014). doi: 10.1073/pnas.1402518111

7. Chundawat, S. P. S., Beckham, G. T., Himmel, M. E. & Dale, B. E. Deconstruction of Lignocellulosic Biomass to Fuels and Chemicals. *Annu. Rev. Chem. Biomol. Eng.* **2,** 121–145 (2011). doi:10.1146/annurev-chembioeng-061010-114205

8. Chundawat, S. P. S. et al. Restructuring the crystalline cellulose hydrogen bond network enhances its depolymerization rate. *J. Am. Chem. Soc*. 133, 11163–11174 (2011). doi:10.1021/ja2011115

9. Creagh, A. L., Ong, E., Jervis, E., Kilburn, D. G. & Haynes, C. A. Binding of the cellulose-binding domain of exoglucanase Cex from *Cellulomonas fimi* to insoluble microcrystalline cellulose is entropically driven. *Proc. Natl. Acad. Sci.* **93,** 12229–12234 (1996). doi:10.1073/pnas.93.22.12229

10. Fleishman, S. J. *et al.* Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. *PLoS One* **6,** 1–10 (2011). doi:10.1371/journal.pone.0020161

11. Gao, D. et al. Increased enzyme binding to substrate is not necessary for more efficient cellulose hydrolysis. *Proc. Natl. Acad. Sci*. 110, 10922–10927 (2013). doi 10.1073/pnas.1213426110

12. Guo, J. & Catchmark, J. M. Binding specificity and thermodynamics of cellulose-binding modules from trichoderma reesei Cel7A and Cel6A. Biomacromolecules 14, 1268–1277 (2013). doi: 10.1021/bm300810t

13. Haarmeyer, C. N., Smith, M. D., Chundawat, S. P. S., Sammond, D. & Whitehead, T. A. Insights into cellulase-lignin non-specific binding revealed by computational redesign of the surface of green fluorescent protein. *Biotechnol. Bioeng.* **114,** 740–750 (2017). doi:10.1002/bit.26201

14. Jeoh, T., Cardona, M. J., Karuna, N., Mudinoor, A. R. & Nill, J. Mechanistic kinetic models of enzymatic cellulose hydrolysis—A review. Biotechnol. Bioeng. 114, 1369–1385 (2017). doi:10.1002/bit.26277

15. Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H. & Meiler, J. Practically useful: What the R osetta protein modeling suite can do for you. *Biochemistry* **49,** 2987–2998 (2010). doi: 10.1021/bi902153g

16. Linder, M. et al. Identification of functionally important amino acids in the cellulose-binding domain of Trichoderma reesei cellobiohydrolase I. Protein Sci. 4, 1056–1064 (1995). doi:10.1002/pro.5560040604

17. Linder, M., Nevanen, T. & Teeri, T. T. Design of a pH-dependent cellulose-binding domain. FEBS Lett. 447, 13–6 (1999). doi:10.1016/S0014-5793(99)00253-7

18. Lim, S., Chundawat, S. P. S. & Fox, B. G. Expression, purification and characterization of a functional carbohydrate-binding module from Streptomyces sp. SirexAA-E. Protein Expr. Purif. 98, 1–9 (2014). doi: 10.1016/j.pep.2014.02.013

19. Loewenthal, R., Sancho, J. & Fersht, A. R. Histidine-aromatic interactions in barnase. J. Mol. Biol. 224, 759–770 (1992). doi: 10.1016/0022-2836(92)90560-7

20. Mattinen, M. et al. Three-dimensional structures of three engineered cellulose-binding domains of cellobiohydrolase I from Trichoderma reesei. Protein Sci. 6, 294–303 (1997). doi:10.1002/pro.5560060204

21. Narayanan, V. Binding interactions of family 1 carbohydrate binding modules with cellulose allomorphs. (Rutgers, The State University of New Jersey, 2017). doi:10.7282/t3zg6wck

22. Parthasarathi, R. *et al.* Insights into hydrogen bonding and stacking interactions in cellulose. *J. Phys. Chem. A* **115,** 14191–14202 (2011). doi: 10.1021/jp203620x

23. Payne, C. M. et al. Fungal Cellulases. Chem. Rev. 115, 1308–1148 (2015). doi:10.1021/cr500351c

24. Pérez, S. & Samain, D. *Structure and Engineering of Celluloses*. *Advances in Carbohydrate Chemistry and Biochemistry* **64,** (2010). doi: 10.1016/S0065-2318(10)64003-6

25. Reinikainen, T. *et al.* Investigation of the function of mutated cellulose-binding domains of Trichoderma reesei cellobiohydrolase I. *Proteins Struct. Funct. Bioinforma.* **14,** 475–482 (1992). doi: 10.1002/prot.340140408

26. Reyes-ortiz, V. *et al.* Addition of a carbohydrate-binding module enhances cellulase penetration into cellulose substrates. Biotechnol. Biofuels 6, 1 (2013). doi:10.1186/1754-6834-6-93

27. Rigden, D. J. *From Protein Structure to Function with Bioinformatics*. (2009). doi:10.1007/978-1-4020-9058-5

28. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein Structure Prediction Using Rosetta. *Methods Enzymol.* **383,** 66–93 (2004).

29. Sethaphong, L. et al. Tertiary model of a plant cellulose synthase. Proc. Natl. Acad. Sci. 110, 7512–7517 (2013). doi: 10.1073/pnas.1301027110

30. Sousa, L. et al. Next-generation ammonia pretreatment enhances cellulosic biofuel production. Energy Environ. Sci. 9, 1215–1223 (2016). doi:10.1039/c5ee03051j

31. Sugimoto, N., Igarashi, K. & Samejima, M. Cellulose affinity purification of fusion proteins tagged with fungal family 1 cellulose-binding domain. *Protein Expr. Purif.* **82,** 290–296 (2012). doi:10.1016/j.pep.2012.01.007

32. Wi, S. G. *et al.* Biotechnology for Biofuels Lignocellulose conversion for biofuel : a new pretreatment greatly improves downstream biocatalytic hydrolysis of various lignocellulosic materials. *Biotechnol. Biofuels* 1–11 (2015). doi:10.1186/s13068-015-0419-4

33. Whitehead, T. A., Bandi, C. K., Berger, M., Park, J. & Chundawat, S. P. S. Negatively Supercharging Cellulases Render Them Lignin-Resistant. *ACS Sustain. Chem. Eng.* **5,** 6247–6252 (2017). doi: 10.1021/acssuschemeng.7b01202