THERMAL DENATURATION AND REFOLDING OF CARBOHYDRATE BINDING MODULES TO IMPROVE ENZYME RECYCLING IN A LIGNOCELLULOSIC BIOREFINERY

by

Jose A. Sanchez Hernandez

B.A. (Hood College) 2012

THESIS

Submitted in partial satisfaction of the requirements

for the degree of

MASTER OF SCIENCE

in

BIOMEDICAL SCIENCE

in the

GRADUATE SCHOOL

of

HOOD COLLEGE

July 2019

Accepted:

_____	_____
Ricky R. Hirschhorn, Ph.D. Committee Member	Ann L. Boyd, Ph.D. Director, Biomedical Science Program
_____	
Dominic Esposito, Ph.D. Committee Member	
_____	_____
Craig S. Laufer, Ph.D. Thesis Adviser	April Boulton, Ph.D. Dean of the Graduate School

STATEMENT OF USE AND COPYRIGHT WAIVER

I, Jose A. Sanchez Hernandez, do authorize Hood College to lend this thesis, or reproductions of it, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

DEDICATION

I dedicate this work to my family and friends for their unconditional love and support. To my parents, Jose Sanchez Miramontes and Rocio Sanchez, I want to thank you for giving me life in a country filled with opportunity. I want to thank you for raising me in a home where the values of faith, family, and the love of progress were always enforced. To my siblings Cristina Zuniga and Zaidel Sanchez, I want to thank you for playing a part in guiding me through the most difficult moments of my adulthood, especially in recent years. I navigate each day with confidence knowing that if I ever need anything, I can always count on you both to be at my side and to give me strength. To my brothers, David Bartee, James Watts, and Justin Nguyen, since high school I have been thankful for your friendship. I am grateful for the memories we have made while driving across state lines, procrastinating college assignments, making runs to Wawa after countless hours of gaming, and am especially grateful for all your advice when I needed it most. To my friends and family, thank you for loving me and for pushing me to be better— I would not be here today without your combined efforts.

ACKNOWLEDGEMENTS

I would like to start by expressing my sincerest gratitude to my advisor Craig Laufer for the continuous motivation and feedback in the last few years. Since starting in the biomedical science graduate program at Hood College, I have counted on his unwavering enthusiasm regarding my project and my growth as a scientist. I am lucky to have been a student of Dr. Laufer for many reasons. Firstly, some of the best ideas I have had for my thesis project were inspired by papers that were assigned in his classes. Dr. Laufer handpicked incredibly interesting papers, which naturally made the discussion of meaningful concepts about protein engineering and biochemistry enjoyable—he truly makes it easy to be excited about anything he lectures about in class. Secondly, I have appreciated the opportunity to be creative in lab; if I ever wanted to try a new experiment, or design a new assay, I could always count on Dr. Laufer's support. Most importantly, the many scientific discussions we had in his office, in the hallways of Hodson between classes, during class breaks, or even during lunch were invaluable for my experience as a student of the Biomedical Science program. It has been an honor to work on such an interesting and meaningful project under his guidance.

Next, I would like to thank my committee member Ricky Hirschhorn for being my mentor since my time as an undergraduate at Hood College. Dr. Hirschhorn played a key role in helping me become the scientist I am today starting back when she hired me as her Summer Institute Research Scholar when I was an undergraduate. Working for Dr. Hirschhorn was what first inspired me to pursue a career in biomedical science; she was the first person to help me appreciate how elegantly chemistry is always at work in "the cell," and over the years has treated me like family with her support in both my

professional and personal life. I have enjoyed being her student and ultimately owe my appreciation of biochemistry to her.

Additionally, I would like to thank my committee member Dominic Esposito. I was recently hired by the protein expression laboratory at the Frederick National Laboratory for Cancer Research (FNLCR) and have thoroughly enjoyed learning from his leadership and his knowledge of protein biochemistry and assay development. Despite his many responsibilities, Dr. Esposito always managed to carve out time in his day to meet with me regarding my thesis proposal, my poster presentations, and my career at Leidos. I am grateful to have had his feedback and attention to detail during the completion of this project.

I would also like to thank Robert Kozak of Atlantic Biomass for the collaborative efforts on this project and for the many late nights conversations we have had in the Hodson Science and Technology Center. Further, I would like to thank the many faculty members at Hood who were always willing to share their expertise and knowledge with me during my graduate course of study. When in a bind, I always relied on Dana Lawrence in the chemistry and physics department, Ann Boyd, Oney Smith, and Susan Carney in the biology department. I have lost count of how often I would come to them with questions about homework, seeking tips on experiments, or simply to borrow their equipment for long periods of time. I would like to especially thank Georgette Jones in the biology department for allowing me to shadow her introductory biology lectures and lab sections. Seeing Dr. Jones in action as an instructor prepared me to successfully teach my own classes and set the tone for a great experience as a graduate research assistant during my time working with the Hood College biology department.

Special thanks to my friends and colleagues Troy Taylor, Nitya Ramakrishnan, Matt Drew, Mukul Sherekar, Shelley Perkins, Gulcin Gulten, and Vanessa Wall at the FNLCR for their help with cloning, expression, purification, and optimization of assays—their help was key to the completion of my thesis project. A huge thanks to Anastazia Jablunovsky and Codi West, members of the Laufer lab, for their help with the cloning of CtCBM-GSF constructs. This work has been funded by the USDA.

TABLE OF CONTENTS

ABSTRACT

Second generation bioethanol using lignocellulosic waste is a promising source of renewable energy. Cellulose, lignin, and other biopolymers make up lignocellulose. Saccharification of lignocellulose requires cellulases, which have a catalytic domain (CD) and a carbohydrate binding module (CBM) that binds the enzyme to substrates. Cellulase activity is known to decrease during saccharification when CBMs irreversibly bind lignin that the CD cannot hydrolyze for release and further reactions. While costly, cellulases must be supplemented after each cycle of saccharification. Here we demonstrate that following denaturation when heated to 5°C above the melting temperature (T_m), CBMs 11 and 44 (CAZy families) from *Hungateiclostridium thermocellum* (CtCBM11 and CtCBM44) can be released from a bound substrate. Once cooled to a temperature below the T_m, CtCBM11 and CtCBM44, Type B CBMs with a β-sandwich fold, spontaneously refold and regain binding function. Using temperature tunable CBMs could drastically reduce saccharification costs by improving enzyme recycling strategies.

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF ABBREVIATIONS

ADA	N-(2-Acetamido) iminodiacetic acid
ABSCO	Absolute contact order
AGE	Affinity gel electrophoresis
RsgI-N	Anti-sigma factor N-terminus
CO	Contact order
CAZy	Carbohydrate-Active enzymes
CBM	Carbohydrate binding module
CE	Carbohydrate esterase
CO_2	Carbon dioxide
CMC	Carboxymethyl cellulose
CD	Catalytic domain
CtCBM	*Clostridium thermocellum* CBM
CAPS	N-cyclohexyl-3-aminopropanesulfonic acid
DNS	3,5-Dinitrosalicylic acid
FPA	Filter paper assay
GDT	Global distance test
GH	Glycoside hydrolase
GT	Glycosyl transferases
GFPS11	Green fluorescent protein strand 11
GFPS1-S10	Green fluorescent protein strands 1-10
IMAC	Immobilized metal affinity chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
KmR	Kanamycin selection marker
LacI	Lac repressor gene

LB	Luria-Bertani
T_m	Melting temperature
MEA	Microextraction automated instrument
NEB	New England Biolabs®
MES	2-(N-Morpholino) ethanesulfonic acid monohydrate
MOPS	3-(N-morpholino) propanesulfonic acid
CelD-N	N-terminal Ig-like domain of cellulase
pDNA	Plasmid DNA
PCR	Polymerase chain reaction
PL	Polysaccharide lyase
PDB	Protein data base
RFU	Relative fluorescence units
SLH	S-layer homology domain
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
PT7	T7 promoter
PTet	Tet promoter
TetR	Tetracycline repressor
UNFAO	United Nations Food and Agriculture Organization
uGDT	Un-normalized GDT
Xylan BW	Xylan from beechwood
Xylan OS	Xylan from oat spelts

INTRODUCTION

The production of carbon neutral energy sources has become one of the biggest challenges of our time. Combustion of fossil fuels has become one of the major contributors to global warming through the release of human-generated greenhouse gases which is largely dominated by the emission of carbon dioxide (CO_2) as it accounts for an estimated 77% of greenhouse gases and hugely impacts the environment (Rahman et al. 2017). Decades of research have shown that biofuels are a viable source of renewable energy, which contrast from the limited nature and detrimental global impact of fossil fuel energy. A key advantage of biofuels over fossil fuels is that upon complete combustion, the fully oxidized carbon released into the above-ground carbon cycle is taken up by fuel crops and prevents a net increase in CO_2 emissions. Biofuels are high-energy chemicals that are produced through biological processes or are derived from chemical conversions from the biomass of prior living organisms. There are two types of biofuels: primary and secondary biofuels. Primary biofuels are obtained by the direct combustion of organic materials in an unprocessed form, such as the burning of fuelwood, wood chips and pellets. Secondary biofuels involve the indirect production of bioethanol or other fuels from plant or animal materials and are further classified into three generations. First generation biofuel is ethanol produced directly from food crops that are rich in starch, or oils amenable to biodiesel production. While rich in energy, first generation biofuel manufacturing has important limitations. Most notably, relying on food crops as a fuel source will naturally contribute to food shortages and rises in food prices that could devastate the developing world. For instance, between 2001 and 2007 first generation ethanol production tripled from 4.9 billion gallons to almost 15.9 billion

gallons, according to C. Ford Runge, a professor of agricultural economics at the University of Minnesota. However, in December 2007, the United Nations Food and Agriculture Organization (UNFAO) calculated that world food prices rose 40% in 12 months prior, and the price hikes affected all major biofuel feedstocks, including sugarcane, corn, rapeseed oil, palm oil, and soybeans (Tenenbaum 2008). More recent models show that world food prices could further rise by 32 percent by 2022 with half of this increase stemming from the use of food crops for biofuel production (Chakravorty et al. 2017).

For these reasons, most research is directed towards the production of second and third generation biofuels. Second generation biofuel is bioethanol derived from non-food, lignocellulosic biomass. The third generation of biofuel is obtained from cyanobacteria, microalgae, or other photosynthetic microbes. Third generation biofuels have been found to be especially attractive; based on species and cultivation methods alone, biohydrogen, biomethanol, bioethanol, and biodiesel can be produced (Poudyal et al. 2016). Efforts in third generation biofuel production have been directed towards optimizing the dark fermentation of bacteria where carbohydrates are converted to biohydrogen or other biofuels; upscaling the photobiological methods of biofuel production of microalgae; metabolic and genetic engineering of cyanobacteria or microalgea to enhance biohydrogen production; genetic engineering of yeast to increase ethanol tolerance to increase alcohol production; and fermentation of plant cell wall carbohydrates by the co-culture of microorganisms to produce biofuels.

Despite the potential of third generation biofuel production, not much has been done to increase the low yields and low efficiencies of these processes. According to a

recent and helpful review by Rodionava et al. (2017), the production of biohydrogen is currently not competitive enough to replace the hydrogen production from fossil fuels. The many limiting factors for third generation biofuels include the different efficiencies of light utilization by phototrophs at varying levels of sunlight intensity, the impairment of hydrogen production enzymes and pathways in cells by abundant atmospheric oxygen, and the rate of carbon dioxide assimilation during photosynthesis necessary for efficient biomass accumulation and its further conversion into biofuel is low. While certainly a key player in finding a solution to the global energy crisis, currently, third generation biofuel production alone cannot remedy this problem.

Fortunately, alternative sources exist for production of secondary biofuel, particularly in the form of bioalcohols. Ethanol is the most common bioalcohol as it accounts for more than 90% of total biofuel usage, while biopropanol and biobutanol are less common. To date, the largest source for bioalcohol is lignocellulosic biomass that is rich in complex polysaccharides in the form of non-food plant materials like switchgrass and agricultural waste, which can be readily obtained from such sources as corn stover, citrus peel, or beet pulp. As shown in Figure 1, such biomass consists of mainly cellulose, hemicellulose, and lignin, along with minor amounts of pectin, ash, protein, and extractives (Jørgensen and Pinelo 2017).

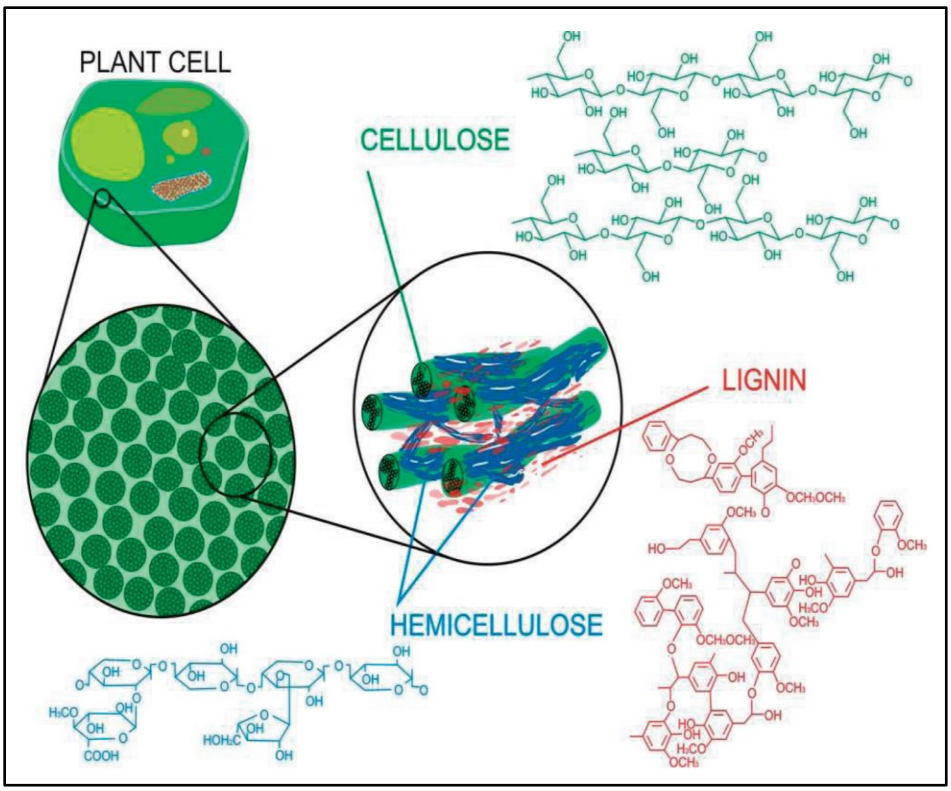


Figure 1: Chemical structures for three of the main components of lignocellulosic biomass most important to the saccharification reactions in plant based biofuel production (modified from Bamdad et al. 2018).

Releasing simple monosaccharides from these structural polysaccharide components gives rise to a sugar platform that, upon biochemical processing, yields bioethanol among other products. In recent years, the launch of the first biorefineries based on biochemical conversion of biomass into ethanol finally have made renewable energy a commercial reality. Ørsted A/S plant in Denmark (formerly DONG Energy) started up in 2009 (Larsen et al. 2012), Beta Renewables started in Italy during 2013, and three full-scale plants inaugurated in the USA in 2015: Abengoa (closed operations after a short time in operation) (2015 Survey of Non-Starch Ethanol and Renewable Hydrocarbon Biofuels Producers 2016), POET-DSM, and Dupont. Despite current low

oil prices, production of higher value chemicals and materials, along with ethanol, has attracted much attention and could make the costs of lignocellulosic biomass biorefining competitive with fossil fuel processes or first generation biofuel production from food crops (Henning and Manuel 2017).

The ingredients for lignocellulosic biorefining are complex polysaccharides in the form of biomass, catabolic enzyme cocktails of glycoside hydrolases (GH- cellulases, hemicellulases, and other auxiliary enzymes) that can release monosaccharides by hydrolysis in saccharification reactions, and microbes that can use these simple sugars as substrates for alcoholic fermentation followed by distillation to bring the alcohol concentrations to suitable levels to employ as fuels. The key for such an enterprise is the combination of cost-efficient biomass pre-treatment along with a low-cost cellulolytic process. Cellulases have a tremendous plasticity, which allows them to recognize a wide range of β-1,4-glucosidic bonds in a variety of polysaccharides (e.g. cellulose, xyloglucan, glucomannan, and mixed-linked β -1,4- β -1,3-glucans). Unfortunately, the enzymatic degradation of insoluble polysaccharides can often be inefficient as lignocellulose is insoluble and is present as hydrogen-bonded crystalline fibers, coated with hemicellulose chains and pectin all "glued" into an intricate 3D network (Viegas et al. 2013). As a result, target substrates are often inaccessible to the active site of the appropriate enzymes. To overcome these problems many glycoside hydrolases are composed of two domains: in addition to a catalytic domain (CD), we find that catabolic enzymes also contain a carbohydrate binding module (CBM) linked to the CD via a linker peptide (Figure 2).

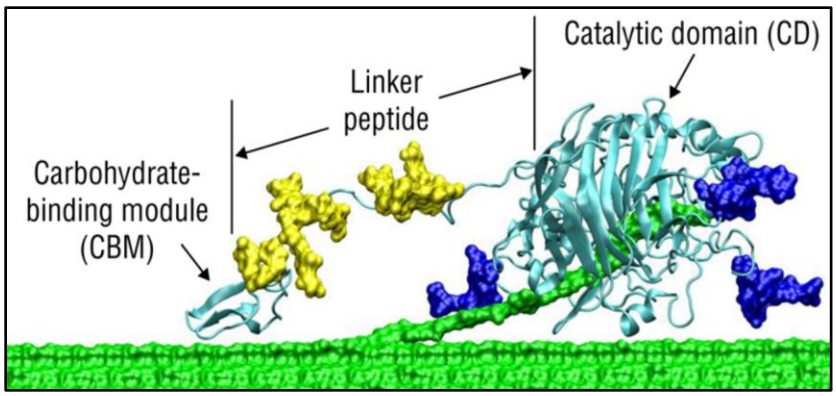


Figure 2: The Family 7 processive cellulase enzyme from biomass-degrading fungus *Trichoderma reesei*. The cellulose crystal that the enzyme is digesting is shown in green. The enzyme consists of a CBM, a flexible linker with attached glycosylation (in yellow), and a large catalytic domain that threads single cellulose chains into a long tunnel and hydrolyzes the chains into soluble sugar products. Additional glycosylation (in blue) is shown on the catalytic domain. This enzyme moves down a single chain of cellulose and is the primary enzyme of interest in fungal enzyme cocktails for biofuel production (modified from Taylor et al. 2012).

A CBM is defined as a contiguous amino acid sequence within a carbohydrate-active enzyme with a discrete fold having carbohydrate-binding activity. CBMs have three functions with respect to their cognate CD: (i) targeting function, (ii) proximity effect, and a (iii) substrate disruptive function. Through their binding activity, CBMs concentrate enzymes on the polysaccharide substrates and are thought to increase the rate of glycoside hydrolysis (Boraston et al. 2004). Currently, CBMs are grouped into 84 different families based on amino acid similarity on the Carbohydrate-Active enZYmes (CAZy) database (CAZy web site: http://www.cazy.org/Carbohydrate-Binding-Modules.hTml). These groupings are intended to aid in the identification of CBMs, in identifying functional residues, help reveal evolutionary relationships, and can also help predict polypeptide folds. Since structural folds of proteins are better conserved than amino acid sequences, CBMs are further classified into fold families, of which there are

seven, as shown below in Table 1: β-sandwich, β-trefoil, cysteine knot, unique, OB fold, hevein fold and "hevein-like" fold.

Table 1: CBM families further grouped based on their structural fold (Boraston et al. 2004).

Fold family	Fold	CBM families
1	β-Sandwich	2, 3, 4, 6, 9, 15, 17, 22, 27, 28, 29, 32, 34, 36
2	β-Trefoil	13
3	Cysteine knot	1
4	Unique	5, 12
5	OB fold	10
6	Hevein fold	18
7	Unique; contains hevein-like fold	14

As suggested by Figure 3 below, by far the dominant fold among CBMs is the β-sandwich (fold family 1). CBMs belonging to this family fold have a β-jelly roll with two β-sheets, each consisting of three to six antiparallel β-strands. In most cases, β-sandwich CBMs have bounded metal ions (usually calcium) which are believed to play a structural role. With few known exceptions, the binding site in these CBMs is localized in the concave side of the β-barrel lined with solvent exposed hydrophobic residues (namely tryptophan and tyrosine). The β-trefoil fold family (fold family 2) is generally associated with the ricin toxin β-chain. CBMs belonging to this fold contain twelve β-sheet strands that form six hairpin turns. Six of the β-strands form a β-barrel structure attendant with three hairpin turns. The other three hairpins form a triangular cap on one end of the β-barrel denominated "hairpin triplet" resulting in a pseudo 3-fold symmetry. The three functional binding sites are an advantage as they lead to significantly enhanced affinities. (Hashimoto 2006).

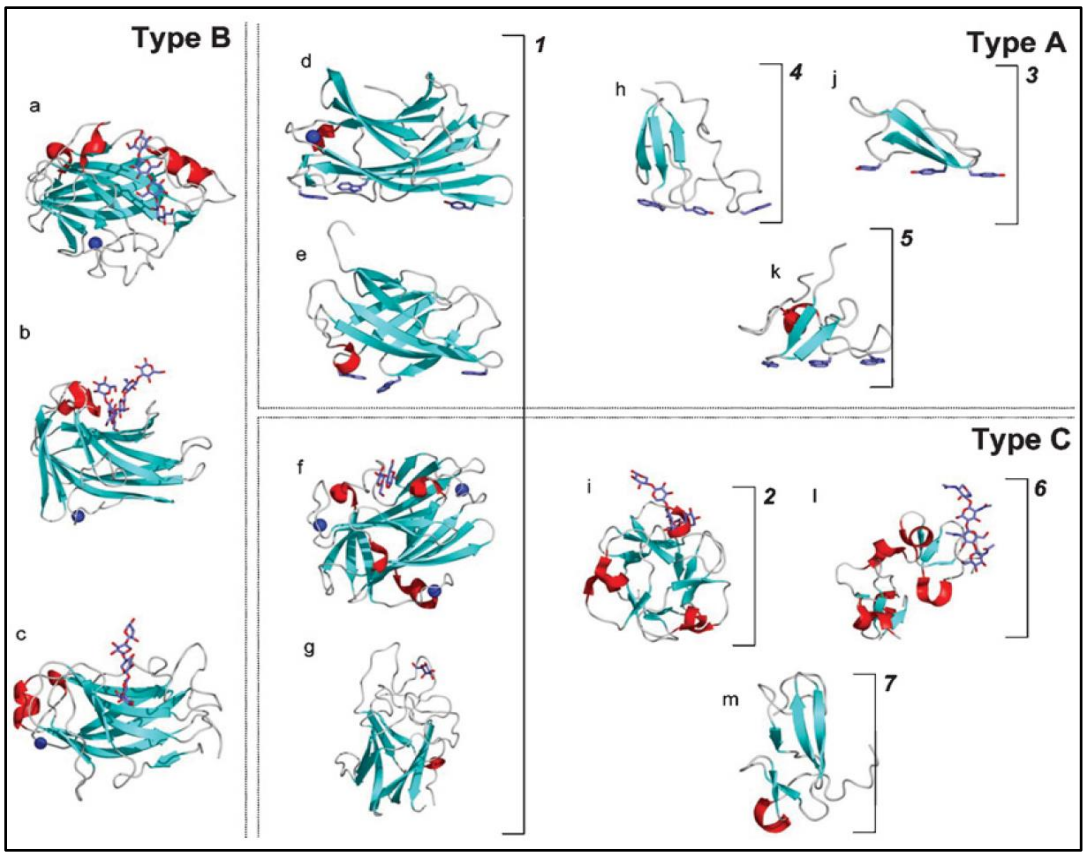


Figure 3: Classification of CBMs, shown as ribbon structures, based on their fold families. Brackets with numbers indicate examples of CBMs belonging to fold families 1–7. Bound ligands are shown as 'liquorice' representations, while bound metal ions are shown as blue spheres. Dotted boxes surround examples of CBMs belonging to the functional Types A, B, and C, which is based on the CBMs substrate affinity (modified from (Boraston et al. 2004)). The CBMs included in this figure are as follows: (a) family 17 CBM from *Hungateiclostridium cellulovorans* in complex with cellotetraose (PDB code 1J84 (Notenboom, Alisdair B Boraston, et al. 2001)); (b) family 4 CBM from *Thermotoga maritima* in complex with laminariohexaose (PDB code 1GUI (Boraston et al. 2002)); (c) family 15 CBM from *Cellvibrio japonicus* in complex with xylopentaose (PDB code 1GNY (Pires et al. 2004)); (d) family 3 CBM from *Hungateiclostridium thermocellum* (PDB code 1NBC (Tormo et al. 1996)); (e) family 2 CBM from *Cellulomonas fimi* (PDB code 1EXG (Xu et al. 1995)); (f) family 9 CBM from *T. maritima* in complex with cellobiose (PDB code 1I82 (Notenboom, Alisdair B. Boraston, et al. 2001)); (g) family 32 CBM from *Micromonospora viridifaciens* in complex with galactose (PDB code 1EUU (Gaskell et al. 1995)); (h) family 5 CBM from *Erwinia chrysanthemi* (PDB code 1AIW (Brun et al. 1997)); (i) family 13 CBM from *Streptomyces lividans* in complex with xylopentaose (PDB code 1MC9 (Notenboom et al. 2002)); (j) family 1 CBM from *Trichoderma reesi* (PDB code 1CBH (Kraulis et al. 1989)); (k) family 10 CBM from *Cellvibrio japonicus* (PDB code 1E8R (Raghothama et al. 2000)); (l) family 18 CBM from *Urtica dioca* in complex with chitotriose (PDB code 1EN2 (Saul et al. 2000)); (m) family 14 CBM from *Tachypleus tridentatus* (PDB code 1DQC (Suetake et al. 2000)).

CBMs from fold families 3 to 5 are small amino acid polypeptides (30-60 amino acids) that contain only a β-sheet and coil. They appear to be specialized in binding cellulose and/or chitin. The majority of these CBMs have planar surfaces, which contain hydrophobic residues arranged in a planar orientation complementary to the flat surface of the crystalline polysaccharides they bind. Fold families 6 and 7 contain small CBMs with approximately 40 amino acids, originally identified in plants as chitin-binding proteins. This fold is dominated by coil with two small β-sheets and an α-helix. The minimal hevein fold in fold family 6 is found in family 18 CBMs. The family 14 CBMs belong to fold family 7 as their fold also incorporates a hevein fold but is also fused with a small β-sheet structure, which changes the overall topology of the family fold.

Although CBM families can be grouped into fold families based on the evolutionary conservation of a protein fold, unfortunately such groupings are not consistently predictive of their function. Enough diversity exists among fold family members that functional elements like amino acids or binding-site topographies are not conserved. A more useful approach to predicting function is based on a type classification: Type A, surface-binding CBMs; Type B, glycan-chain binding CBMs; and Type C, small-sugar binding CBMs (see Figure 3 above and Table 2 below). Surface-binding, Type A, CBMs are known to bind insoluble, highly crystalline cellulose and/or chitin using flat, or platform-like, aromatic residues in the binding site that are complementary to the flat surfaces presented by cellulose or chitin crystals. These so-called surface binding CBMs have little to no affinity for soluble carbohydrates.

Table 2: CBM folds and families further classified by surface binding (Type A), glycan chain-binding (Type B), or small-sugar binding (Type C) (Boraston et al. 2004).

Type	Fold family	CBM families
A	1, 3, 4, 5	1, 2a, 3, 5, 10
B	1	2b, 4, 6, 15, 17, 20, 22, 27, 28, 29, 34, 36
C	1, 2, 6, 7	9, 13, 14, 18, 32

Next, structural studies of glycan-chain binding Type B CBMs have revealed carbohydrate-binding sites often described as grooves or clefts, which comprise several subsites that are able to accommodate individual sugar units of polymeric ligands. Much like Type A CBMs, aromatic residues play a key role in ligand binding. In contrast to Type A CBMs, the orientation of these amino acids is not planar— amino acids are instead orientated into the groove so that hydrogen bonds between key tyrosine residues and polysaccharide substrates also help define ligand specificity. CBMs that belong to this family have clearly evolved binding site topographies that are equipped to interact with individual glycan chains due to their negligible affinities for oligosaccharides with a degree of polymerization of three or less, or for highly flat and crystalline surfaces.

Lastly, small-sugar binding, Type C, CBMs have a high affinity for mono-, di-, or tri-saccharides and are unable to bind larger polymers due to lacking the extended binding-site grooves of Type B CBMs. Type C CBMs are mostly found in xylanases and specialize in binding only the reducing end sugars of xylan or cellulose. Compared to Type A and Type B CBMs, identification and characterization of Type C CBMs lags behind as their presence is limited in glycoside hydrolases and mostly found in toxins and enzymes that attack eukaryotic cell surfaces. Notably, the distinction between Type B

and Type C CBMs can be subtle. For example, the Type B CBM6 from *Hungateiclostridium stercorarium* xylanase has a very similar fold to the Type C CBM32 family but can bind to longer polymer chains and thus accounts for its Type B classification. Additionally, CBM6 from *Cellvibrio mixtus* contains two discrete binding sites that display characteristics of Type B and Type C CBMs (Pires et al. 2004; Henshaw et al. 2006: 6). However, it is apparent that in the Type C binding site a stronger hydrogen-bonding network is apparent when binding a small oligosaccharide ligand, which matches what would be expected between interactions of smaller oligosaccharides compared to largely hydrophobic interactions that drive the binding interactions of amino acids in Type B CBMs to longer polymers of saccharides. The interactions of Type A, B, and C CBMs with their corresponding types of saccharide substrates are summarized in Figure 4 below.

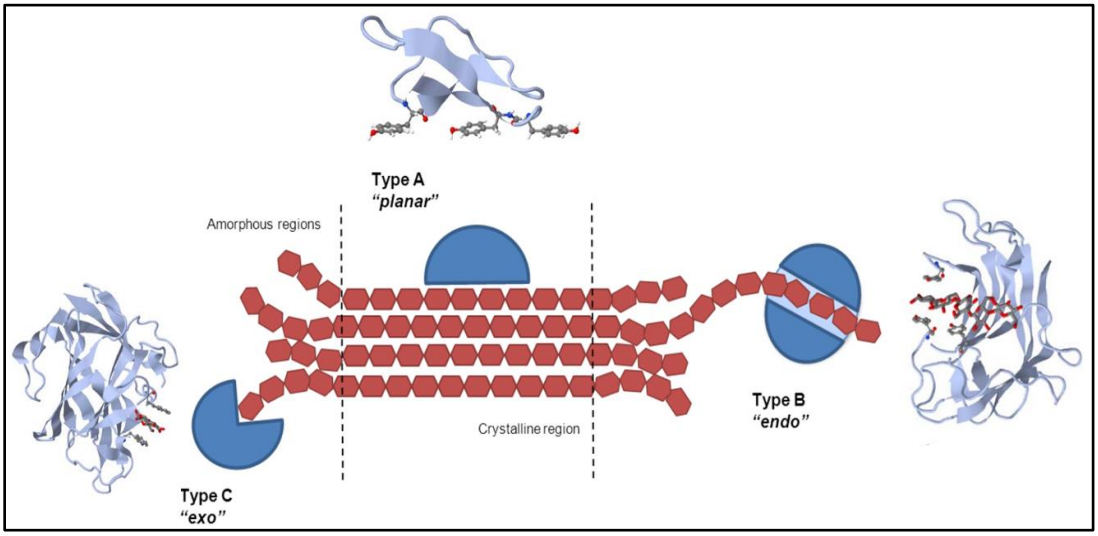


Figure 4: CBM structures from different sources interacting with their respective substrates. Type A CBM from *Trichoderma reesei* cellobiohydrolase I (PDB code 1CBH), a Type B CBM from *Cellulomonas cimi* endo-1,4-glucanase C (PDB code 1GU3), and a Type C CBM from *Thermotoga maritime* xylanase 10A (PDB code 1I82) (modified from (Nakamura et al. 2008), and (Guillén et al. 2010)). Secondary structural elements are all shown in grey. Functional amino acids at binding sites are shown in a ball-and-stick representation. Polysaccharide substrates are shown in gray and red.

Despite the added benefit of CBMs being present in most glycoside hydrolases, due to the complex chemical nature of the structural polysaccharides that make up biomass, conversion of these polysaccharides to sugars useable for fermentation requires multiple catabolic enzymes, often at high concentrations. Additionally, at industrially relevant scales with high biomass loadings, saccharification reaction efficiency is often limited by poor enzyme stability and detrimental interactions between enzymes and lignin, a complex organic polymer that forms important structural elements in the support tissues of plants and algae. These limitations manifest as measurable decreases in enzyme activity and sugar production yields after multiple rounds of saccharification. As shown in Figure 5, time course experiments of cellulase activity done in the presence of lignocellulose from beet pulp show that the activity of saccharification enzymes can drop by as much as 50 percent compared to the activity of cellulases incubated in the same conditions, but devoid of substrate. The greater loss of cellulase activity at these time points suggests that the adsorption of cellulases onto lignin present in the roller bottles lowers activity over time. Interestingly, the opposite trend was observed for lyase enzymes. In the presence of lignocellulose, substrate or product stabilization appeared to increase the stability of lyases compared to the negative control devoid of substrate and adsorption to the substrate is not an issue as these enzymes lack CBMs.

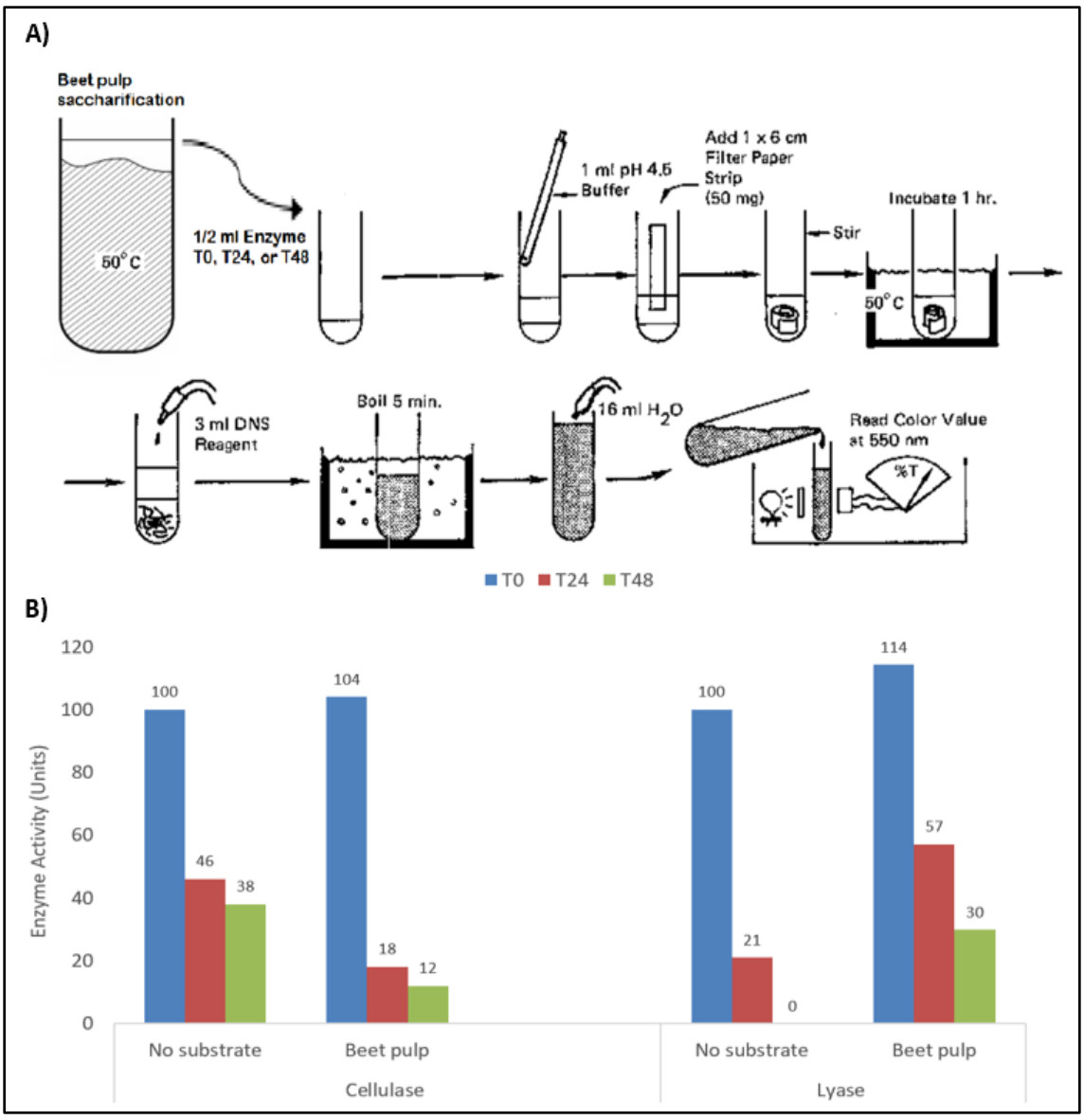


Figure 5: A) Schematic for the filter paper assay (FPA), which measures the glucose liberated from a cellulose substrate after incubation with saccharification enzymes. As part of this time course experiment, cellulases containing both a CD and CBM, and lyases composed of only of a CD, were recovered from a roller bottle containing lignocellulose in the form of beet pulp. Negative controls in this assay were devoid of substrate. Liquid aliquots were taken at three time points: 0 hours, 24 hours, and 48 hours. The assay consisted of using Whatman No. 1 filter paper cut into 1 × 6 cm strips (50 mg), buffer = 50mM sodium citrate pH= 4.8, glucose standards in buffer, and dinitrosalicyclic acid (DNS) to measure reducing sugars. B) Cellulase and lyase activity taken from roller bottles with and without lignocellulose substrate at time points zero (blue), 24 hours post incubation (red), and 48 hours post incubation (green).

The detrimental interactions between lignin and cellulases cannot be overstated and has been well-documented in the literature. As stated by Ooshima et al. (Ooshima et al. 1990), “the adsorption of enzyme on the lignacious residue as well as cellulose must be taken into account in the development of the hydrolysis kinetics.” Thus, the decrease in enzyme activity after every saccharification cycle likely results from cellulases preferentially binding non-digestible and lignin-rich materials in spent biomass, rather than being released to the soluble and recyclable fraction for continued use. This further means that after each saccharification cycle, a smaller population of soluble cellulases and auxiliary enzymes are available for deconstruction of any newly added biomass substrate. This ultimately creates a bottleneck that limits how cheaply bioethanol can be made from non-food lignocellulosic substrates.

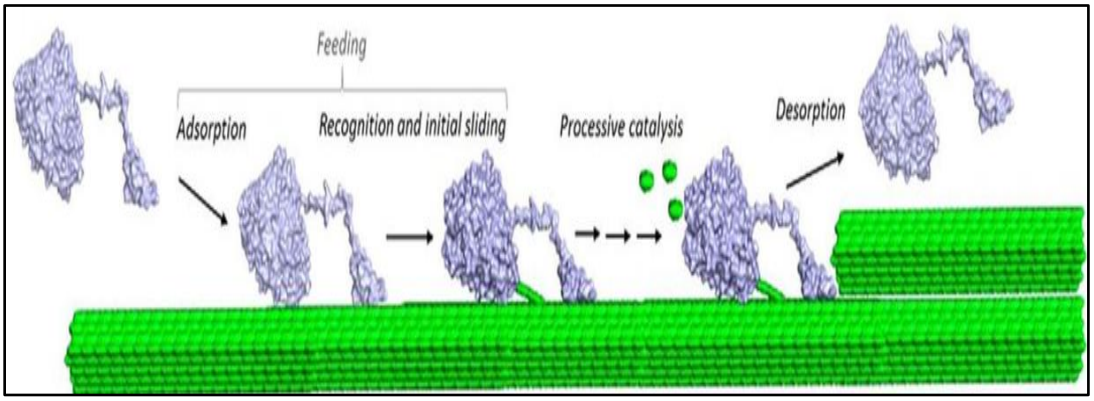


Figure 6: Schematic representation of the catalytic domain, linker, and CBM that make up TrCel7A, a glycoside hydrolase, outlining the molecular steps involved in the hydrolysis of cellulose. Cellulose strands are represented as green sticks. Steps leading from the free enzyme in the solution to the enzyme with the reducing end of the cellulose chain in the binding site are collectively referred to as the feeding step. Processive catalysis includes the formation of a Michaelis enzyme-substrate complex, hydrolysis of the glycosidic bond, and expulsion of smaller oligosaccharides (green ellipses). Processive catalysis is repeated until the enzyme meets an obstacle (depicted here as upper cellulose fibril), happens to dissociate, or runs out of substrate (Kont et al. 2016)

Fortunately, more recent studies have improved our understanding of how to mitigate enzyme-lignin interactions. As outlined in Figure 5, ideally the CBM should land the enzyme on a substrate, the CD should hydrolyze the substrate, and the entire enzyme should then release itself from the biomass and diffuse to another substrate for further saccharification reactions. Studies by Strobel et al. have uncovered some of the biochemical mechanisms for cellulase-lignin interactions that prevent this ideal process from occurring. While it is well-known that hydrophobic and electrostatic interactions along with hydrogen bonding are responsible for enzyme-lignin interactions (Strobel et al. 2016), this study further showed that the affinity of a cellulase to lignin could be altered by simple mutagenesis of the Trichoderma Cel7A CBM. In fact, beneficial mutations of CBMs were combined to generate a mutant cellulase with 2.5-fold less lignin affinity while fully retaining cellulose affinity and its CD activity. It is likely that the current limitations of saccharification reactions mostly arise from CBMs irreversibly binding the lignin-rich, insoluble matter in a reaction vat that the CD cannot hydrolyze. Considering this new information, an enzyme recycling strategy could be designed focused on rescuing lignin-trapped cellulases and auxiliary enzymes by reducing or eliminating CBM-lignin affinity after each saccharification cycle.

When developing methods that rescue enzymes that have become irreversibly bound to insoluble lignin through a CBM-lignin interaction, it is important that the proposed enzyme-saving process not create waste that affects downstream processing of monosaccharides and be detrimental to the stability of the saved enzyme once released from the lignin, nor should it be costlier than the purchase cost of fresh enzymes. Many strategies have been attempted to improve enzyme recycling, which include recycling of

free enzymes using changes in pH and the addition of surfactants, reabsorption to fresh material, recycling of solids, membrane filtration, enzyme immobilization, and protein engineering of cellulases devoid of a CBM entirely (Jørgensen and Pinelo 2017). Regrettably, most of these strategies do little to rescue enzyme activity or to reduce the overall costs of the biorefining process. For example, changes in pH trigger conformational changes have been reported to recover greater than 90% of cellulase protein. However, this was only possible after changing the pH of the reaction vessel to greater than 11.5 using calcium hydroxide treatments— these harsh treatments resulted in less than 10% of enzyme activity and in some cases no greater than 50% of the initial activity (Otter et al. 1984; Rodrigues et al. 2012) likely resulting from the denaturation of the CD in the cellulase.

Better results for enzyme recycling have been obtained in the deconstruction of softwood feedstock. Data analysis showed the optimized conditions for releasing lignin-trapped enzymes were temperatures of 44.4 ^{o}C, pH 5.3 and 0.5% Tween 80. In fact, these conditions showed great promise since most added cellulose substrate was converted after just a few rounds of saccharification. The additives were likely efficient at lowering non-specific CBM-lignin interactions after each round of saccharification and unlike previous methods, did so without affecting the structure and function of the CD in the rescued enzyme. Despite these encouraging results, one cannot ignore that the efficient recovery of enzymes in these conditions comes at the expense of introducing high concentrations of detergent substances that are costly and likely deleterious to the downstream processing of the generated monosaccharides, which could later interfere with conversion into ethanol by yeasts (Tu et al. 2009).

Other attempts have focused on the issue of the CBM impeding the rescue of lignin-trapped enzymes by engineering enzymes variants that contain a CD but are devoid of their natural CBM. The rationale for such an idea makes sense: removing the protein domain responsible for protein-lignin interactions could be beneficial and improve enzyme recycling strategies. Such endeavors have only reduced enzyme activity (likely because the CBM plays a key role in enzyme-substrate binding), requires greater biomass pre-treatment, higher enzyme concentrations, and increases production costs (Mes-Hartree et al. 1987). These results suggest that the CBM is ultimately needed for efficient saccharification reactions. In fact, a study published by Walker (Walker et al. 2015) shows that fusing a broad affinity CBM to a single multifunctional CD can actually increase rates of saccharification with different pure polysaccharides and with pretreated biomass. As shown in this study, fusing CBMs from families that have a broad substrate affinity to CDs can form new cellulases that are more efficient than those found in nature. Combined with lowering the costs required for saccharification through better enzyme recycling strategies, this type of protein engineering could create an entirely new avenue for improving plant biomass saccharification processes.

Currently cost-analyses have reported that enzyme costs can form up to 28% of the total ethanol selling price depending upon whether the enzymes are produced on or off-site (Jørgensen Henning and Pinelo Manuel 2017). Other cost-analyses report that cellulases are the second most expensive element in the overall process. Together with the pre-treatment of the lignocellulosic raw material, these two processes make up a significant part of the final bioethanol cost (Aden and Foust 2009). Such costs pose a major obstacle that threatens the economic viability of the lignocellulosic biorefining

enterprise. Thus, the goal of this project is to lower the cost associated with enzyme saccharification of lignocellulosic biomass through novel engineering of CBMs and generation of cellulases currently not present in nature that can be released from spent plant biomass for use in multiple rounds of saccharification.

We believe that saccharification costs can be reduced dramatically by implementing an enzyme recycling strategy that involves reversibly denaturing and refolding CBMs to release them from lignin without compromising the CD. At a temperature even just a few degrees above the T_m of the CBM, the folding equilibrium lies essentially 100% with the unfolded state where binding function is lost. Conversely, a few degrees below the T_m of the CBM, puts the folding equilibrium essentially 100% towards the folded state where binding function is possible. Thus, a slight increase in temperature above the melting temperature (T_m) of the CBM will denature the CBM releasing the whole enzyme from the remaining spent biomass and lignin-rich material. Following the desorption of the CBM from a spent batch of biomass, simple phase partitioning could recover the soluble enzymes. Addition of new substrate with a return to the operating temperature below the T_m of the CBM would allow the CBM to refold and regain binding function. A change in temperature as small as 5 °C could result in a temperature tunable transition of entire cellulases and other auxiliary enzymes that contain a CBM between saccharification cycles, which would allow for the same batch of enzymes to be used in multiple rounds of saccharification reactions.

In this study we have shown that CBMs from CAZy families 11 and 44, from thermophile *H. thermocellum* (formerly *Clostridium thermocellum,* CtCBM11 and CtCBM44), a thermostable organism that produces many well-defined cellulases that

have been reported in the literature (Walker et al. 2015; Hirano et al. 2016), are capable of refolding spontaneously after being thermally denatured. Using functional binding assays against many cellulose substrates (microcrystalline cellulose, cellulose membranes, and substrate retardation assays with dissolved carboxymethyl cellulose and xylan), we show that CtCBM11 and CtCBM44, Type B CBMs that display a β-sandwich fold, regain binding ability upon refolding by cooling from a thermally perturbed state. These results suggest that saccharification costs could be dramatically reduced by engineering temperature tunable CBMs and fusing them to thermostable and high activity CDs. This type of strategy could lead to the generation of novel cellulose degrading enzymes that could more effectively degrade plant biomass and lead to more effective enzyme recycling strategies than those currently being employed by lignocellulosic biorefineries.

MATERIALS AND METHODS

Polymerase chain reaction for CtCBM-GFPS11 and CtCBM-GSF constructs

Recombinant DNA sequences of CtCBMs were made by PCR from the genomic DNA of *H. thermocellum* ATCC 27405 with a concentration of approximately 5 ug/mL, which was then diluted to 1:10 and 1:100 DNA using molecular biology grade water. All PCR primers were prepared to a concentration of 50mM using the appropriate amount of molecular biology water. The master mix for these reactions contained 1/100 the volume of forward and reverse primers each—the remaining volume contained molecular biology grade water. The PCR reactions were carried out and purified as outlined in the illustra PuReTaq Ready-To-Go PCR Beads (GE Healthcare UK Limited) and analyzed using 1.4% agarose gel electrophoresis. The primers and corresponding amino acid sequences for proteins CtCBM11, CtCBM30, and CtCBM44 are shown below in Table 3.

Table 3: Primers and amino acid sequences used for the cloning of CtCBM-GFPS11 and CtCBM-GSF constructs. All primers contain Nde1 or BamH1 restriction sites (in bold) needed for digestion of the CBM DNA for insertion into the GFPS11 or GSF vector.

CtCBM-GFPS11	**Forward primer**	**Reverse primer**
CtCBM11	5'-GATATA **CATATG** GCT GTC GGT GAA AAA ATG-3'	5'-CTATAT **GGATCC** AGC ACC AAT CAG CTT GAT-3'
CtCBM30	5'-GCTATA **CATATG** AGT GCC GAA ACA GTT GC-3'	5'-CTATAT **GGATCC** CTT GAT TGC AGG AGC GGA C-3'
CtCBM44	5'-GATATA **CATATG** TTT ACA GCT ACC ATA AAA GTA ACC-3'	5'-CTATAT **GGATCC** CCA GTC AAT AGC ATC TAC-3'
CtCBM-GSF	**Forward primer**	**Reverse primer**
CtCBM11	5'-GATATA **CATATG** GCT GTC GGT GAA AAA ATG-3'	5'-CTATAT **CATATG** AGC ACC AAT CAG CTT GAT-3'

CtCBM30	5’-GATATA **CATATG** AGT GCC GAA ACA GTT GC-3’	5’-CTATAT **CATATG** CTT GAT TGC AGG AGC GGA C-3’
CtCBM44	5’-GATATA **CATATG** TTT ACA GCT ACC ATA AAA GTA ACC-3’	5’-CTATAT **CATATG** CCA GTC AAT AGC ATC TAC-3’
CtCBM	**Amino acid sequence**	
CtCBM11	**AVGEKMLDDFEGVLNWGSYSGEGAKVSTKIVSGKTGNGMEVSYTGTTDGYWGTV YSLPDGDWSKWLKISFDIKSVDGSANEIRFMIAEKSINGVGDGEHWVYSITPDSSWK TIEIPFSSFRRRLDYQPPGQDMSGTLDLDNIDSIHFMYANNKSGKFVVDNIKLIGA**	
CtCBM30	**SASAETVAPEGYRKLLDVQIFKDSPVVGWSGSGMGELETIGDTLPVDTTVTYNGLP TLRLNVQTTVQSGWWISLLTLRGWNTHDLSQYVENGYLEFDIKGKEGGEDFVIGF RDKVYERVYGLEIDVTTVISNYVTVTTDWQHVKIPLRDLMKINNGFDPSSVTCLVFS KRYADPFTVWFSDIKITSEDNEKSAPAIK**	
CtCBM44	**KFNFEDGTLGGFTTSGTNATGVVVNTTEKAFKGERGLKWTVTSEGEGTAELKLDGGTIVV PGTTMTFRIWIPSGAPIAAIQPYIMPHTPDWSEVLWNSTWKGYTMVKTDDWNEITLTLPED VDPTWPQQMGIQVQTIDEGEFTIYVDAIDW**	

Engineering of CtCBM-GFPS11 and CtCBM-GSF constructs

Purified PCR CBM sequences were cloned into a green fluorescent protein strand 11 (GFPS11) vector (SandiaBiotech) to form constructs that encode a CBM fused to the N-terminus of a split GFPS11 fragment. Additionally, CBM sequences were cloned into a SuperFolder GFP (GSF) vector to form constructs that encode a CBM fused to the N-terminus of GSF (SandiaBiotech). Vector information for GFPS11 and GSF can be found in Figure 5 and 6. The digestion reaction consisted of 2/5 reaction volume of CtCBM DNA, the appropriate 3.1 10X buffer (NEB), 1ul BamH1 (NEB), 1ul Nde1 (NEB), and the remaining volume consisted of high-quality molecular biology grade water. In a separate reaction tube, the GFPS11 or GSF vector digestion was carried out in the same way. Digestion reactions were carried out at 37 °C for one hour and stopped by purification. Digestion products were purified using the QIAGEN PCR Purification Kit. Ligation reactions were carried out using 1 ul T4 DNA Ligase (Invitrogen), 1ul of digested pDNA, 10ul of digested CtCBM DNA, and the appropriate amount of 5X buffer

and high-quality molecular biology grade water. Ligation reactions were carried out at 16 °C overnight and stopped by freezing until transformation.

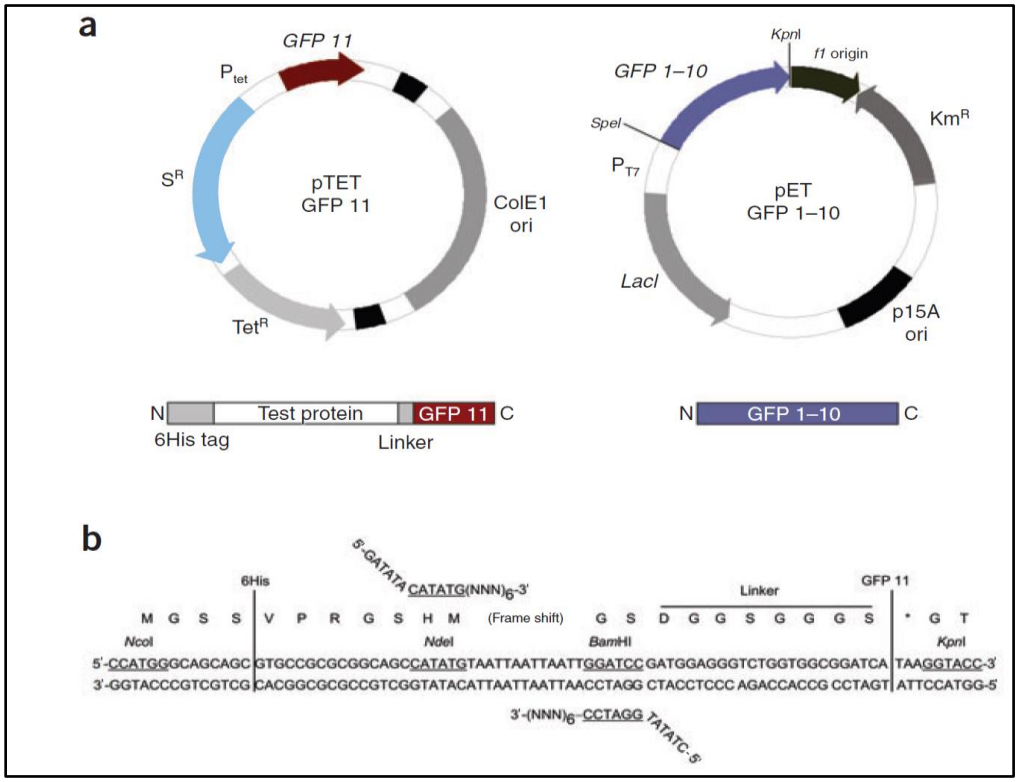


Figure 7: Split-GFPS11 and GFPS1-S10 vectors. (a) The CBMs are expressed, as N-terminal fusion with the GFPS11 tag, under the control of the tet promoter (Ptet) from a pTET plasmid, which contains the spectinomycin resistance marker (SR—targets the class of aminoglycoside antibiotics), a ColE1 origin of replication and the gene encoding tetracycline repressor (TetR). GFP S1–10 is expressed under the control of the T7 promoter (PT7) from a pET plasmid that contains a p15 origin of replication compatible for coexpression with the GFPS11 pTET plasmid, a kanamycin selection marker (KmR) and the Lac repressor gene (LacI). (b) Cloning sites in the GFPS11 vector cassette. The insert protein is cloned using NdeI and BamHI restrictions sites. A frame-shift stuffer with three translational stops (one in each frame) prevents false positives from a self-ligated plasmid. A 6His tag followed by a thrombin cleavage site is located at the N terminus. An 8-amino-acid linker provides a spacer between the test protein and the C-terminal GFPS11 tag. Unique restriction sites in the plasmids are indicated above the 5′ sequence of the cassette (Cabantous and Waldo 2006).

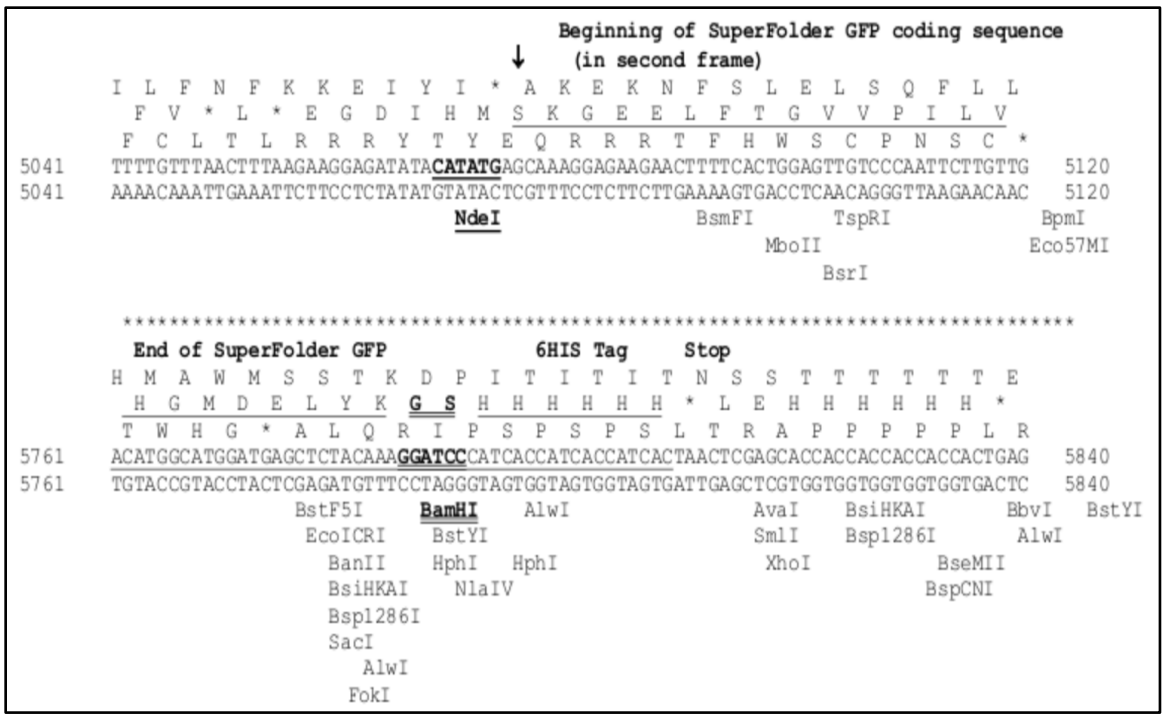


Figure 8: GSF vector: CBMs are expressed fused to the N-terminal end of the GSF. The insert protein is cloned using NdeI restrictions sites on both the forward and reverse primer. More information on the GSF vector can be found in the supplementary information of (Pédelacq et al. 2006).

Transformation of CtCBM-GFPS11 and CtCBM-GSF constructs

CtCBM-GFPS11 pDNA constructs (concentrations of pDNA was not determined) were first transformed into chemically competent Ultracompetent XL10 Gold cells (Agilent Technologies). CtCBM-GSPS11 pDNA and CtCBM-GSF pDNA were isolated using a QIAGEN mini-prep purification and transformed into the expression strain BL21 (DE3) pLysS (Invitrogen) using 1ul of pDNA in 50ul of chemically competent cells. The pDNA was incubated with cells on ice for 20 minutes, heat shocked at 42 °C for 45 seconds, and placed back on ice for one minute. Following heat shock, 80ul of Luria-Bertani (LB) was used as a recovery media and incubated with transformed cells at 37 °C under vigorous shaking for one hour. The entire volume of transformed cells was plated

onto LB plates containing a working concentration of Kanamycin (50ug/mL) and incubated overnight at 37 °C for colony selection the next day.

Expression and purification of CtCBM-GFPS11 and CtCBM-GSF constructs

CtCBM-GFPS11 and CtCBM-GSF constructs were expressed using the so-called "Dynamite media," which has been shown to maximize soluble heterologous protein expression in *E. coli* with cell densities as high as 25 optical density (OD) units (Taylor et al. 2017). After induction with a final concentration of 0.5mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) at an optical density (OD) range between 6 and 8, expression was carried out at 16 °C overnight for a total of 24 hours of growth. *E. coli* cells expressing CtCBM-GFPS11 and CtCBM-GSF expressions are then harvested and resuspended in TNG buffer (50mM Tris pH 7.4, 100mM NaCl, 10% glycerol) (Cabantous, 2006) supplemented with 5mM $CaCl_2$ to aid in stabilization of soluble CtCBM proteins, some of which are known to bind calcium ions. Lysis is done in a ratio of 1mL of lysis buffer per every 100 OD units, where an OD unit is defined as the product of the final OD and the volume of the culture in ml.

Lysis was done by mechanical methods using the LV-1 microfluidizer from microfluidics at 10,000 PSI and two passes. Lysates were clarified at 8,000 x *g* for 90 minutes at 4 ^{o}C. Purification of CtCBM-GFPS11 constructs was done using a microextraction automated instrument (MEA) by immobilized metal affinity chromatography (IMAC) purification tips. CtCBM-GFPS11 constructs were eluted using a gradient imidazole concentration ranging from 125mM imidazole to 500mM imidazole. IMAC purified samples were pooled and frozen by liquid nitrogen in small volumetric

aliquots of 50ul and thawed by water bath for use. CtCBM-GSF constructs were frozen by liquid nitrogen in small volumetric aliquots of 250ul following high speed clarification (as described above) and thawed by water bath for use.

Melting temperature determination for CtCBM-GFPS11 constructs

The T_m of IMAC purified CtCBM-GFPS11 proteins were determined by the SYPRO Orange assay (Crowther et al. 2010). To account for expected substrate-stabilization, CtCBM-GFPS11 constructs were assayed in the presence of 5mM $CaCl_2$. In some trials, xylan from oat spelt (0.01% from 0.4% w/v stock) and carboxymethyl cellulose (0.1% w/v from 5% w/v stock) were also included in the solution to try to account for substrate-stabilization of the measured T_m. Buffers ranging in pH from 4 to 10 were used in a screen to obtain a clean melting curve. The buffers included in this screen were sodium acetate trihydrate pH= 4.3; sodium citrate tribasic trihydrate pH=5.6; 2-(N-Morpholino) ethanesulfonic acid (MES) monohydrate pH= 6.2; Bis Tris propane pH=6.4; N-(2-Acetamido) iminodiacetic acid (ADA) pH=6.4; 3-(N-morpholino) propanesulfonic acid (MOPS) pH= 7.1; TNG buffer pH=7.4; Bicine pH= 8.4; and N-cyclohexyl-3-aminopropanesulfonic acid (CAPS) pH= 10. Reactions were carried out in 50ul volumes using a 96 well-plate.

All buffers used were from a 10X stock (1M), final working buffer concentration in all cases was 100mM. Assuming a protein concentration of approximately 4 mg/ml, protein samples were diluted to a working concentration of 0.25mg/ml. SYPRO orange: S-1234, sold as 5000X stock, was diluted to a 50X stock in 1:100 dilution using water, and used with final working concentration of 5X. All reactions were thus prepared with

33.75 ul of water, 5 ul of 10X buffer, 6.25 ul of IMAC purified CtCBM-GFPS11 constructs, and 5ul of 50X SYPRO orange dye (this was added last to prevent destabilization of protein from the 1% DMSO in the SYPRO dye). The 96-well plate was incubated on ice and was then centrifuged at 2,000 x *g* for 2 minutes, 4°C. The SYPRO assay was conducted on a Biorad CFX96 Real Time System qPCR machine using the thermal melting protocol. First, the temperature is held at 25 °C for ten minutes to ensure uniform temperature of samples. The temperature then ramps up from 25 °C to 95 °C in 0.2-degree increments. Optimal excitation occurs at 480nm and optimal emission occurs at 568nm for the SYPRO orange dye.

CtCBM-GFPS11 and GFPS1-S10 split GFP complementation assay

IMAC purified CtCBM-GFPS11 samples were prepared by diluting in a 50/50 solution of TNG buffer, which also contained 5mM $CaCl_2$, xylan from oat spelt (0.01% from 0.4% w/v stock) and carboxymethyl cellulose (0.1% w/v from 5% w/v stock). These samples were heated to 10°C above the established T_m for five minutes. Heat treated samples were then allowed to cool on ice for 5 minutes and were then centrifuged at 13,000 x *g* for 5 minutes at room temperature using a tabletop centrifuge. Following sample preparation and while still in the presence of dissolved substrates, these samples were mixed with an excess of GFPS1-10 to monitor the complementation kinetics between GFPS1-10 and thermally denatured and refolded CtCBM-GFPS11, and non-heated controls, for 17 hours according to the S1-10 and S11 split GFP Fold-N-Glow complementation assay (SandiaBiotech) by Cabantous (Cabantous et al. 2005; Cabantous and Waldo 2006).

Affinity gel electrophoresis of temperature tunable CtCBM-GFPS11 constructs

Affinity gel electrophoresis (AGE, a native retardation assay) of CtCBM11-GFPS11 and CtCBM44-GFPS11 was done in the presence of 0.1% carboxymethyl cellulose (CMC), xylan from beechwood (Xylan BW), and xylan from oat spelts (Xylan OS) to ensure functionality when in a native state. This protocol was modified from (Foumani et al. 2015). AGE gels were prepared with the following components: 7.5% w/v polyacrylamide/bis-acrylamide made from a 4X stock in 25mM Tris-base/250mM glycine buffer pH=8.3 (made from a 5X running buffer stock), and with 0.1% w/v test polysaccharide made from the appropriate stock as described previously. A small volume of 2.5ul of protein lysate was added per well with 5mg/mL BSA used as a non-binding control. The gel was held at a constant 90V for 2 hours at room temperature. Binding was compared to migration in native gels devoid of substrates and to non-binding controls. The stock solutions were prepared as follows: 5X running buffer stock was made with 15.1g Tris-base, 94g glycine, in 1 liter of deionized water. The polyacrylamide/bis-acrylamide was made from a dry powder blend to form a 30% w/v stock, which was made by dissolving 60g polyacrylamide/bis-acrylamide (19:1 ratio) in 40ml of 5X running buffer and 160ml of deionized water. The separating gel was prepared as follows: 5ml of 4X acrylamide/bis-acrylamide, 4ml of 5X Tris/glycine running buffer, 0.1%w/v test polysaccharide from the appropriate stock solution, and deionized water was used to bring the total volume to 20ml. This volume of solution was enough to make two separating gels each containing 10ml, to which 20ul TEMED and 70ul (10% w/v) degassed ammonium persulfate was added prior to pouring and casting.

Functional evaluation of thermally denatured and refolded temperature tunable CtCBM-GSF constructs against microcrystalline cellulose (Avicel)

Temperature tunable CBMs that passed our split GFP complementation assay after being thermally denatured, CtCBM11 and CtCBM44, were expressed as a fusion to the N-terminus of GSF from Pédelacq et al. (Pédelacq et al. 2006). The ability of thermally denatured and refolded CtCBM11-GSF and CtCBM44-GSF to regain function was evaluated using crystalline cellulose (Avicel). Samples containing 250ul of frozen CtCBM-GSF samples were prepared by thawing in a room temperature water bath and were heated to 60 °C for ten minutes and then centrifuged in a tabletop centrifuge at 13,000 x *g* for ten minutes at room temperature to precipitate *E. coli* proteins. The remaining lysate was diluted four-fold in TNG buffer containing 5mM $CaCl_2$ and stored on ice. The binding assays were done by adding 100 ul of prepared CtCBM-GSF sample to 100ul of 100mM MES pH= 6.2 in an Eppendorf tube and 20mg of Avicel. These samples were gently mixed by tapping and incubated at room temperature for 30 minutes.

In a separate incubation vessel, an identically prepared sample of CBM-GSF sample was heated to 5°C above the established T_m in the absence of substrate for 5 minutes, incubated on ice for 5 minutes, and centrifuged in a tabletop centrifuge at max speed for 5 minutes at room temperature. This sample corresponds to thermally denatured and refolded CtCBM-GSF constructs (denoted Δ), of which 100 ul was added to 100ul of 100mM MES pH=6.2 and incubated with 20mg of Avicel as previously described. All CtCBM-GSF samples were then incubated at 50 °C and at 10 °C above the established T_m. After centrifugation at 13,000 x *g* on a tabletop centrifuge for 5 minutes, 20 ul of sample was taken from the incubation vessel and loaded into a plate reader for excitation

at 485 nm and emission at 520 nm. Loss of fluorescence intensity in the supernatants were taken to demonstrate binding of fusion protein to pelleted Avicel.

Functional evaluation of thermally denatured and refolded CtCBM-GSF constructs against cellulose membranes

The ability of thermally denatured and refolded CtCBM11-GSF and CtCBM44-GSF to regain function was evaluated using cellulose acetate/cellulose nitrate 0.45um membranes from Millipore. Samples of temperature tunable CtCBM11-GSF and CtCBM44-GSF constructs were prepared as previously described and normalized to have equal fluorescence units/ul compared to the control GSF (devoid of a CBM subunit) using TNG buffer containing 5mM $CaCl_2$. In a separate Eppendorf tube, normalized CtCBM-GSF samples were heated to 5°C above the established T_m in the absence of substrate for 5 minutes, incubated on ice for 5 minutes, and centrifuged in a tabletop centrifuge at max speed for 5 minutes at room temperature. These samples correspond to thermally denatured and refolded CtCBM-GSF constructs (denoted Δ).

Normalized GSF control samples were also heated to the same conditions to serve as a non-binding control and to ensure stability of the reporter fluorophore. Cellulose membranes (HA 0.45 µm, Millipore) were cut into small strips and blocked with Odyssey Blocking Buffer (TBS) from LI-COR. Cellulose membranes were then blotted with 10 ul containing equal fluorescence units from CtCBM-GSF, CtCBM-GSFΔ, GSF, or GSFΔ samples. Blotted membranes containing samples were immediately imaged under UV light and under inverted contrast. After a brief incubation at room temperature, the cellulose membranes were placed in Eppendorf tubes containing 1ml of 100mM MES

pH=6.2 and were incubated for 10 minutes at 50°C, 5°C below the T_m of the CtCBM, or 5 to 10°C above the T_m of the CtCBM. At the end of the incubation, gentle mixing was done by repeated inversion of tubes to ensure non-specific binding was removed from the cellulose membranes. Membranes were imaged once again under UV light for analysis of CtCBM binding function.

Contact order calculations for CtCBM constructs and structural alignments

The CAZy database (CAZy web site: http://www.cazy.org/Carbohydrate-Binding-Modules.hTml) was used to screen CtCBMs. Using the live link in CAZy to the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), 3-D coordinates of CtCBMs that have been cloned independently from their natural enzymes were downloaded to be used in contact order (CO) calculations. However, due to the limited number of CtCBMs structures that exist in the PDB without being part of an entire enzymatic complex, most files would compromise absolute CO (ABSCO) values and would not be true representations of just the CBMs in question. To resolve this issue, amino acid sequences corresponding to many CtCBMs were obtained from (Walker et al. 2015). These sequences were submitted to RaptorX for 3D structure prediction (Källberg et al. 2012). Predicted 3D structures from CtCBMs sequences were downloaded as PDB files from RaptorX. Using RaptorX, a few of these predicted structures were then compared by structural alignment to the actual 3D structures of CtCBMs in the PDB to verify the accuracy of the predicted models (Wang et al. 2013). Once it was verified that 3D structures predicted by RaptorX contained acceptable RMSD when overlaid the 3D structure found for a CtCBM in the PDB, the PDB files for predicted structures were submitted to the University of Washington Contact Order

(UWCO) calculator (https://depts.washington.edu/bakerpg/contact_order/) (Plaxco et al. 1998). The obtained ABSCO values were plotted against amino acid length for comparisons between the topologies of different CtCBM folds.

RESULTS

Three CBMs from *H. thermocellum* ATCC 27405 were successfully cloned into a GFPS11 plasmid to form CtCBM-GFPS11 constructs, as shown below in Figure 9: CtCBM11-GFPS11, CtCBM30-GFPS11, and CtCBM44-GFPS11. An attempt to clone and express CtCBM35-GFPS11 failed—the plasmid backbone appears to differ from other samples and is the likely reason for failed expression and purification of CtCBM35-GFPS11.

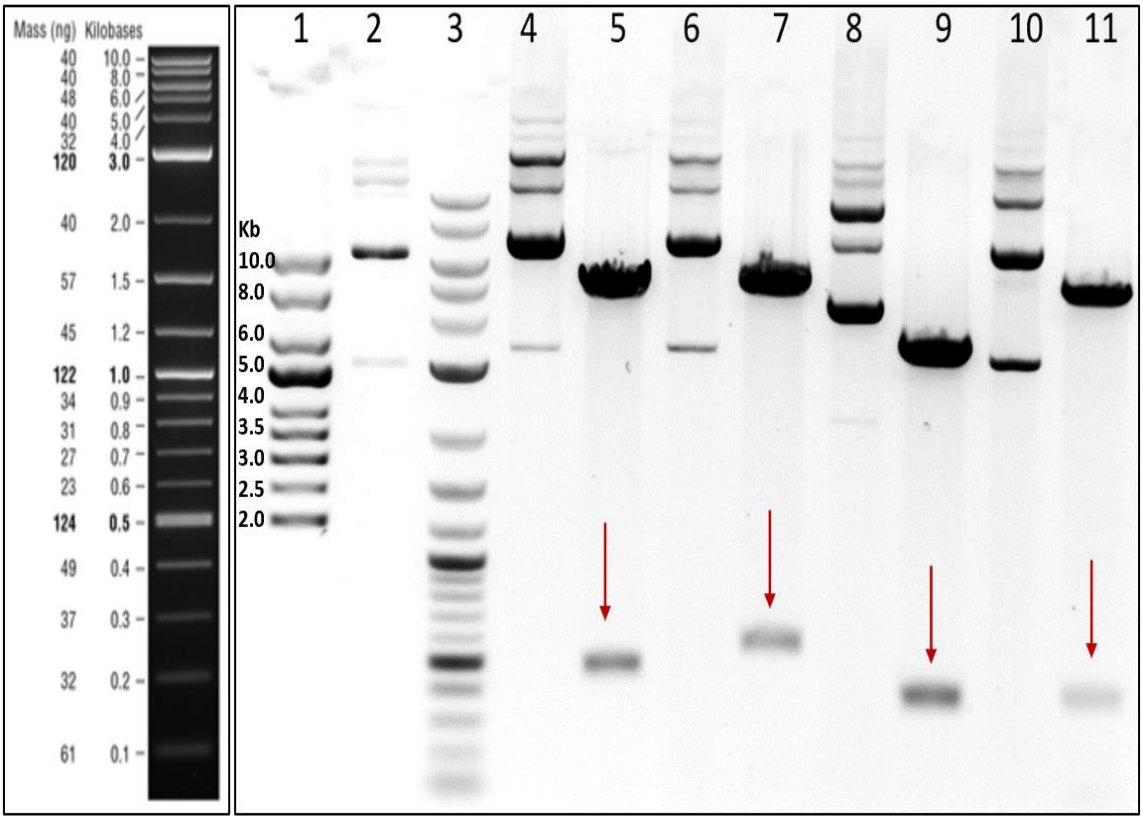


Figure 9: Molecular cloning of CtCBM-GFPS11 constructs. PCR, digestion, and ligation of CtCBM11, CtCBM30, CtCBM35, and CtCBM44 into the GFPS11 plasmid was done to create complete CtCBM-GFPS11 constructs. Lane 1 contains a ladder of supercoiled pDNA, lane 2 contains undigested GFPS11 pDNA. Lane 3 contains a DNA ladder (annotated in left panel), while the remaining lanes 4-11 contain undigested and digested CtCBM11-GFPS11 (lanes 4 and 5), CtCBM30-GFPS11 (lanes 6 and 7), CtCBM35-GFPS11 (lanes 8 and 9), and CtCBM44-GFPS11 (lanes 10 and 11), respectively. Each CtCBM insert appears around 500 bp as outlined by the red arrows.

IMAC was used to purify CtCBM-GFPS11 constructs. Elution samples were pooled and analyzed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) as shown below in Figure 10. IMAC purified samples were used for thermal melting analysis, which is shown in Figure 11. Although CtCBM30-GFPS11 immediately produced a clean melting curve in 50mM TNG buffer devoid of glycerol at pH= 7.4, CtCBM11-GFPS11 and CtCBM44-GFPS11 required a thorough buffer screen to obtain a clean melting curve result. The T_m for CtCBM11 and CtCBM44 was not found to differ greatly between buffers with a pH of at least 6 or greater, thus the shown data was selected based on sharpness of the melting curve. Our results suggest (data not shown) that pH levels below 6 are destabilizing to our constructs as the T_m of CBM11-GFPS11 decreased by 6-8 °C compared to the established T_m of 67°C and the T_m of CtCBM44-GFPS11 decreased by 3-9 °C compared to the established T_m of 75°C. The cleanest melting curves for both CtCBM11-GFPS11 and CtCBM44-GFPS11 were obtained in 100mM MES buffer at pH=6.2, which was used as the buffer for binding assays discussed later in this section.

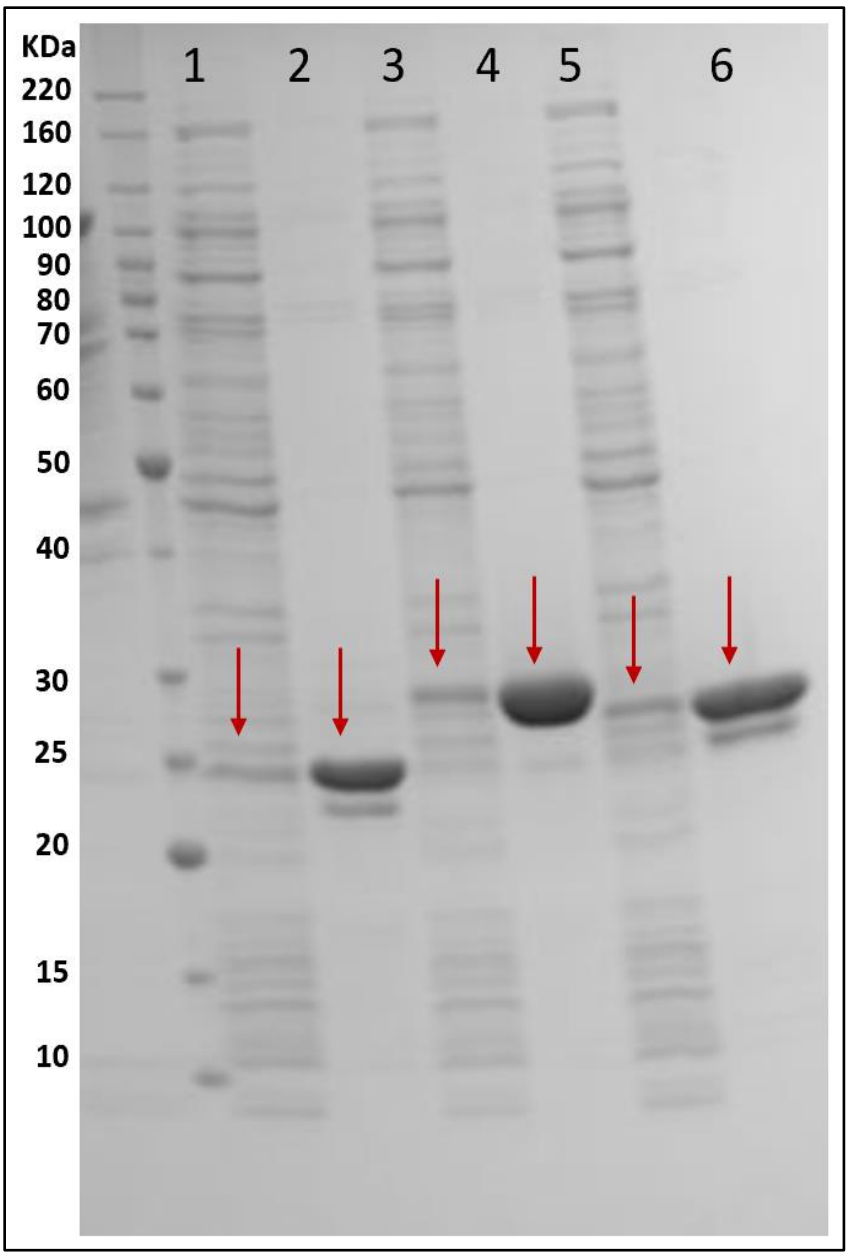


Figure 10: SDS-PAGE of expression and purification by IMAC of CtCBM11-GFPS11 (lanes 1-2, 25kDa), CtCBM30-GFPS11 (lanes 3-4, 29kDa), and CtCBM44-GFPS11 (lanes 5-6, 26kDa), respectively. Red arrows point to the key bands present for each CtCBM-GFPS11 construct.

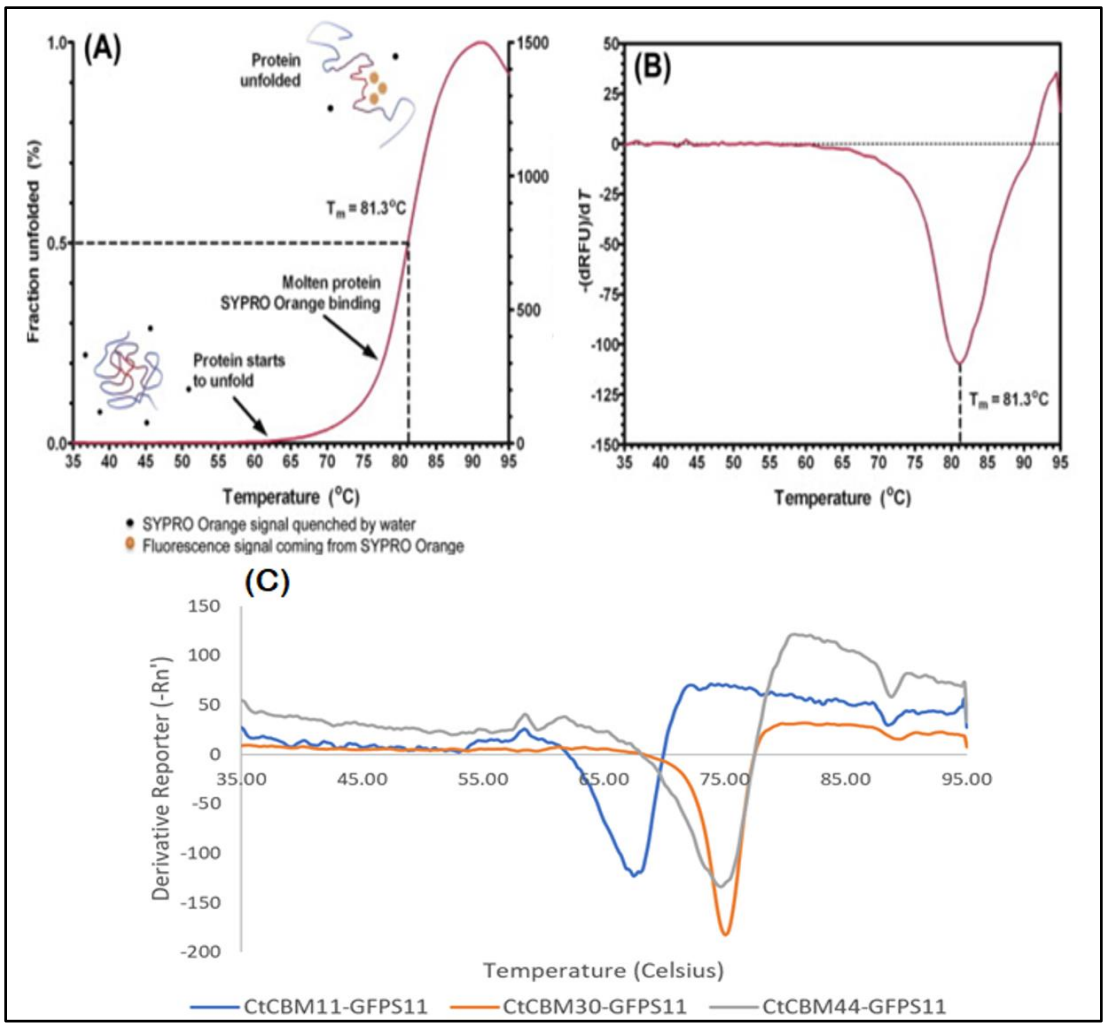


Figure 11: T_m determination of CtCBM-GFPS11 constructs by SYPRO Orange**.** (A) The SYPRO assay works by detecting changes in fluorescence signals of fluorescent dyes when they interact with the hydrophobic residues of thermally denatured proteins, which are normally buried in the folded state (Modified from ThermoFluor®). (B) The temperature at which this change in fluorescence is detected corresponds to the T_m of the protein of interest (Modified from ThermoFluor®). (C) CtCBM11-GFPS11(blue) was determined to have a T_m of 67°C in 100mM MES buffer pH=6.2; CtCBM30-GFPS11 (orange) was determined to have a T_m of 75°C in 50mM TNG buffer devoid of glycerol pH= 7.4; and CtCBM44-GFPS11 (gray) was determined to have a T_m of 75°C in 100mM MES buffer pH=6.2.

Once the melting temperatures of each CtCBM-GFPS11 construct was established, the split GFP complementation assays using the Fold-N-Glow kit from SandiaBiotech was employed to assess whether CtCBMs could refold spontaneously

following thermal denaturation. As shown in Figure 12, the split GFP Fold-N-Glow assay relies on GFPS11, a soluble, self-associating fragment of GFP that can be used to tag proteins without changing the solubility of the fused protein. A protein of interest is fused to a small GFP fragment (beta strand 11, residues 215-230) via a flexible linker of glycine and serine. Complementary GFP fragment 1-10 (beta strands 1-10, residues 1-214) is expressed separately. Neither fragment alone is fluorescent, but when mixed the small and large GFP fragments spontaneously associate, resulting in the formation of a complete GFP fluorophore. Misfolding or aggregation of the fusion protein makes the fluorescent protein tag inaccessible and prevents complementation, thus preventing fluorescence. Therefore, misfolded or aggregated proteins are not included in the quantification of the protein of interest.

CtCBM-GFPS11 constructs that can refold after thermal denaturation and complement GFPS1-S10 without a change in fluorescence gain or folding kinetics compared to the non-heat-treated controls were further evaluated in functional binding tests. Figure 12 shows that the thermally denatured and refolded samples of CtCBM11-GFPS11 and CtCBM44-GFPS11 match the gain in fluorescence of the non-heated controls. CtCBM30-GFPS11 was not moved forward in our tests as the gain in fluorescence for the thermally denatured and cooled samples only reached 48% of the maximum fluorescence gain compared to the non-heated controls.

The encouraging results obtained for CtCBM11-GFPS11 and CtCBM44-GFPS11 using the split GFP complementation Fold-N-Glow assay suggest that when heated to a temperature a few degrees above the T_m, these CBMs can refold spontaneously by cooling. As a result, these CBMs were moved forward on to functional binding tests to

ensure the functionality of each CtCBM before and after denaturation and refolding. To start, we tested the binding affinities of CtCBM11 and CtCBM44 by AGE retardation assays, which are shown in Figure 13. In the absence of substrates, CtCBM11-GFPS11 migrates to the same position as the prominent BSA band, while CtCBM44-GFPS11 appears slightly higher in the gel matrix compared to the prominent band in BSA. As evident by the retardation of prominent CtCBM11-GFPS11 bands in gels containing binding substrates, this CBM appears to strongly interact with both CMC and Xylan OS.

Similarly, although less clear than the binding of CtCBM11-GFPS11, CtCBM44-GFPS11 also appears to bind Xylan OS. If one focuses on the location of prominent bands in lanes that have been heat-treated, a band is observed that does not appear in other gels, which means there could be high affinity interactions between CtCBM44-GFPS11 and Xylan OS. Notably, in this experiment samples were boiled instead of matching the conditions used in the Fold-N-Glow complementation assay. This heat treatment was done for convenience as these binding tests were done prior to the determination of T_m for CtCBM11 and CtCBM44 and to obtain preliminary data showing that these CtCBMs were functional before thermal perturbation and could recover at least some function following their cooling and refolding. Without the presence of SDS to help break apart aggregates, most of the BSA proteins that were boiled precipitated out of solution and were not successfully loaded into to the gel.

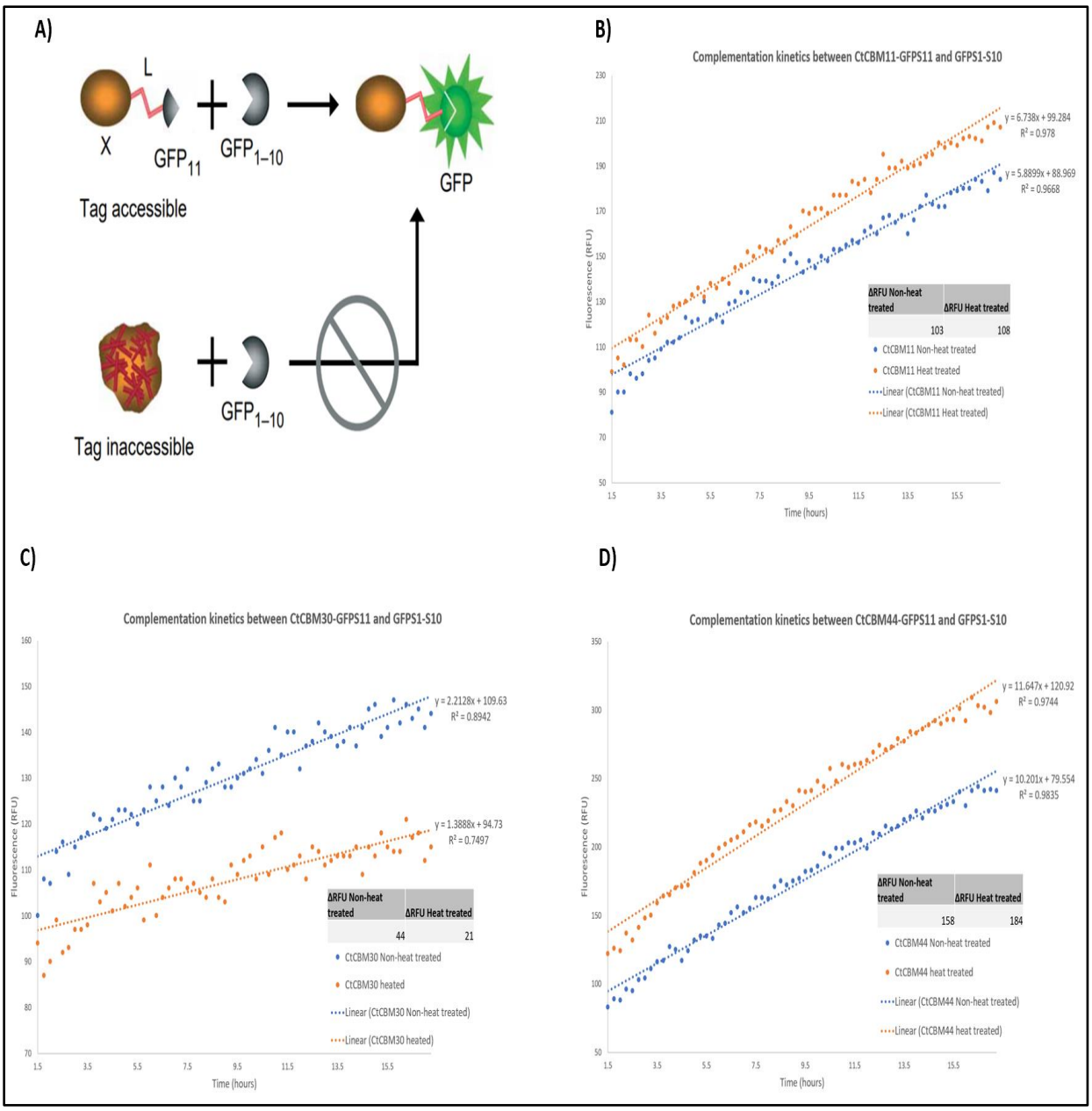


Figure 12: CtCBM-GFPS11 and GFPS1-S10 split GFP complementation assays using the Fold-N-Glow kit from SandiaBiotech. (A) Fold-N-Glow complementation assays of cargo protein fused to GFPS11 and GFPS1-S10 (Sandia Biotech). (B) CtCBM11-GFPS11 (N=2), (C) CtCBM30-GFPS11 (N=2), and (D) CtCBM44-GFPS11 with GFP S1-10 (N=2) complementation kinetics are shown along with the rate of complementation and total gain in blank corrected relative fluorescence units (RFU). Non-heat-treated controls (blue) and thermally denatured and refolded samples (orange) are included on each plot. The rate of increase in fluorescent signal due to complementation as well as total gain in fluorescence are indicated. Following thermal denaturation and cooling, heat treated CtCBM11-GFPS11 and CtCBM44-GFPS11 samples were found to refold as the gain in fluorescence matched the non-heated controls. Heat treated CtCBM30-GFPS11 did not fully refold following thermal denaturation and cooling as only 48% of the fluorescence gain was obtained compared to the non-heated control.

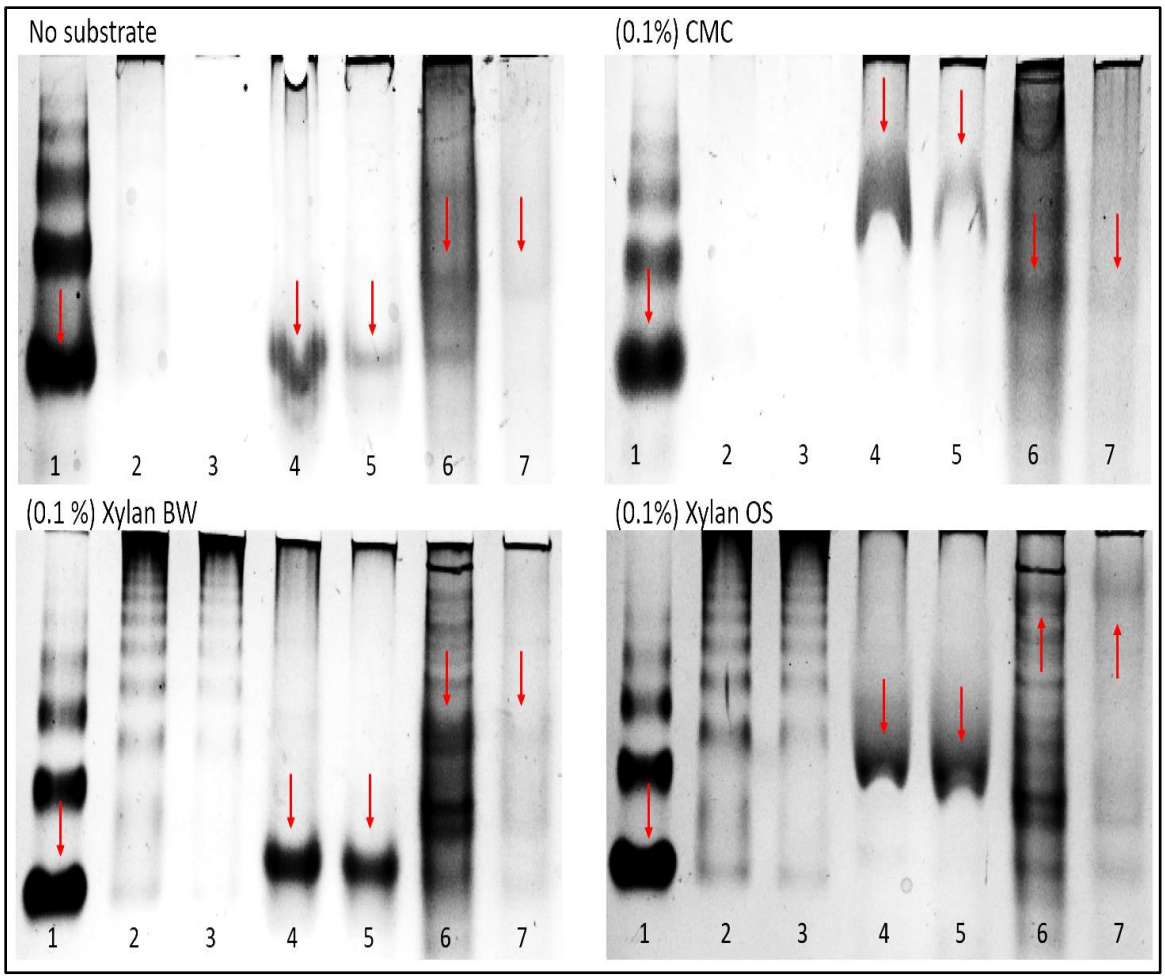


Figure 13: Functional evaluation using AGE of CtCBM11-GFPS11 and CtCBM44-GFPS11 in the absence of substrate, in the presence of 0.1% CMC, Xylan BW, and Xylan OS. Prominent protein bands of interest are highlighted by a red arrow in each gel. Lanes 1 in each gel contain BSA (5mg/mL) and were used as non-binding controls. Lanes 2 and 3 contain BSA samples that were boiled for 10 minutes and then allowed to cool at room temperature for at least 10 minutes. Lanes 4 and 5 contain CtCBM11-GFPS11—lanes 4 were loaded with non-heat-treated sample and lanes 5 contain sample that was boiled as previously described. Lanes 6 and 7 contain CtCBM44-GFPS11— lanes 6 were loaded with non-heat-treated sample and lanes 7 contain sample that was boiled as previously described. The retardation of CtCBM11-GFPS11 in gels containing substrate when compared to the migration of these samples in gels devoid of substrate suggests that CtCBM11-GFPS11 is functional and capable of binding a substrate when expressed before and after being thermally perturbed at boiling temperatures. The above results for CtCBM44-GFPS11 are less clear compared to CtCBM11-GFPS11, but some interactions with Xylan OS are observed in lanes 6 and 7.

Upon verification of function from the native CtCBM11 and CtCBM44 samples, we decided to create new constructs with fusions to the N-terminus of GSF. GSF is reportedly resistant to high temperatures and denaturant conditions (95 °C, 9M Urea) (Pédelacq et al. 2006) and fluoresces regardless of whether the fused CtCBM is folded or denatured, which makes it an excellent reporter to show if a CBM is folded and bound to a substrate, or is denatured and found in the supernatant. SDS-PAGE analysis was used to confirm the expression of soluble GSF and CtCBM-GSF constructs for analysis (Figure 14). Notably, while CtCBM30-GSF was cloned and successfully expressed, it was not used in further experiments as it failed preliminary refolding experiments following thermal denaturation and cooling shown in Figure 12. Figure 15 shows the results of a refolding functional evaluation of CtCBM11-GSF and CtCBM44-GSF constructs in the presence of an Avicel substrate where changes in fluorescence are assumed to be due to binding of a CtCBM to pelleted Avicel. At every temperature tested, fluorescence signals taken from incubation of Avicel with GSF remained consistent indicating that GSF does not bind to Avicel. Figure 15 also suggests that upon refolding from a thermally denatured state, both CtCBM11-GSF and CtCBM44-GSF regain a functional structure that is capable of binding Avicel with equal efficiency as non-heated controls. What is interesting, is that upon incubation at 10°C above the established $T_{m,}$ CtCBM11-GSF fluorescence signals return to almost 100% of the load sample, which means that CtCBM11-GSF is released from the bound Avicel and is found in solution. However, we did not expect to find that even at temperatures as high as 85°C, CtCBM44-GSF remained bound to Avicel as evident by the constant fluorescence signals at both 50°C and 85°C.

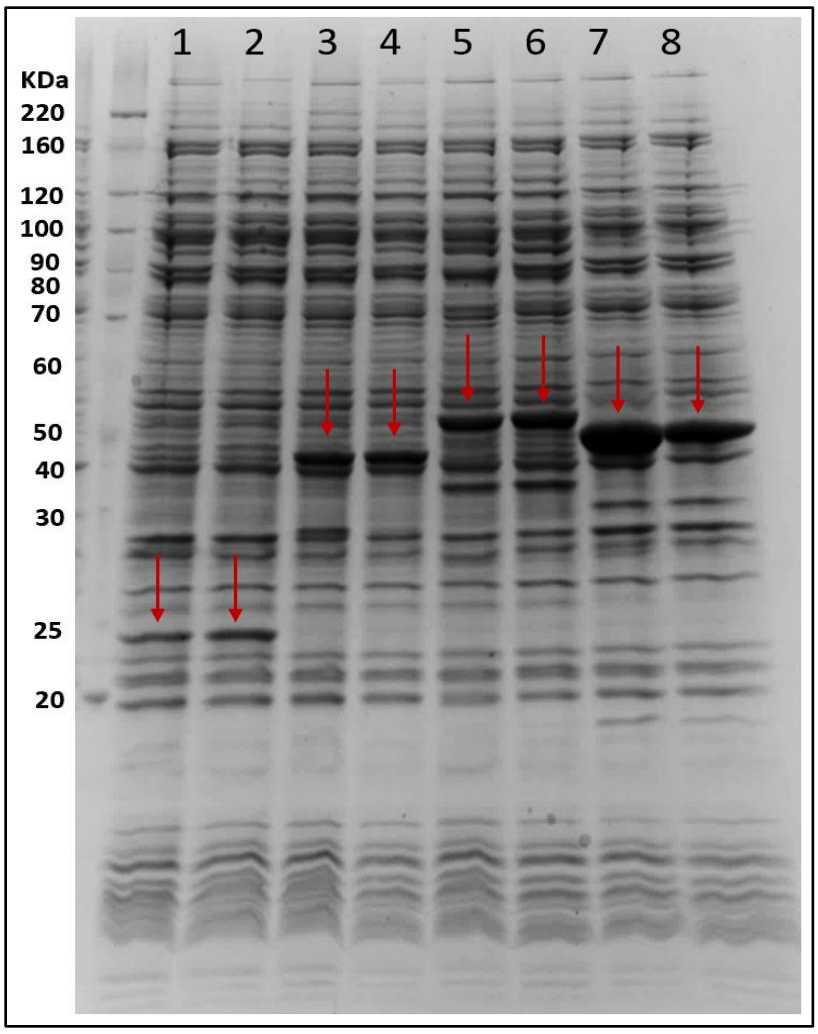


Figure 14: SDS-PAGE analysis of total and soluble fractions for expression of GSF (lane 1 and 2, 26kDa), CtCBM11-GSF (lanes 3 and 4, 45kDa), CtCBM30-GSF (lanes 5 and 6, 49kDa) and CtCBM44-GSF (7 and 8, 46kDa). Red arrows point to the key bands present for each expression.

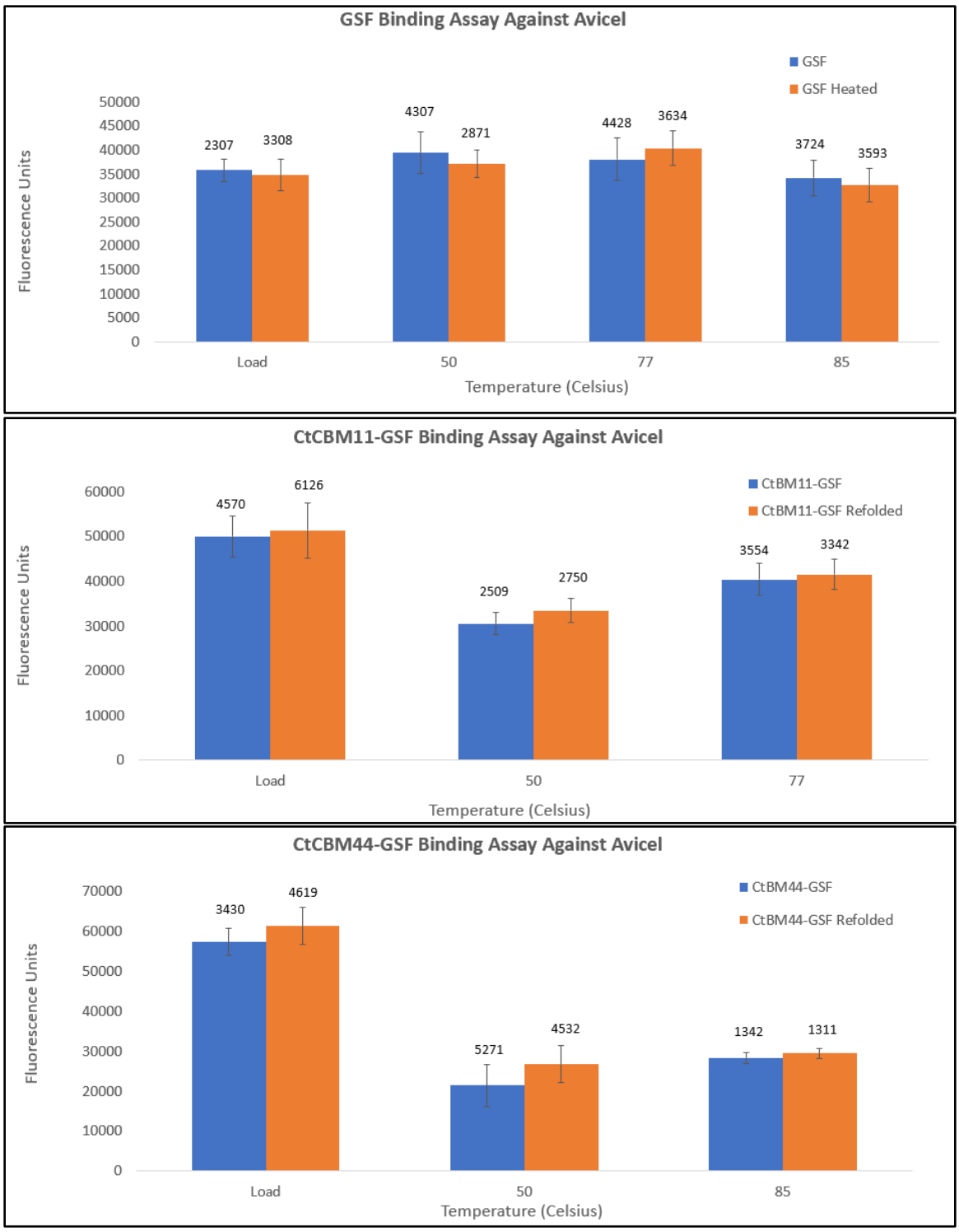


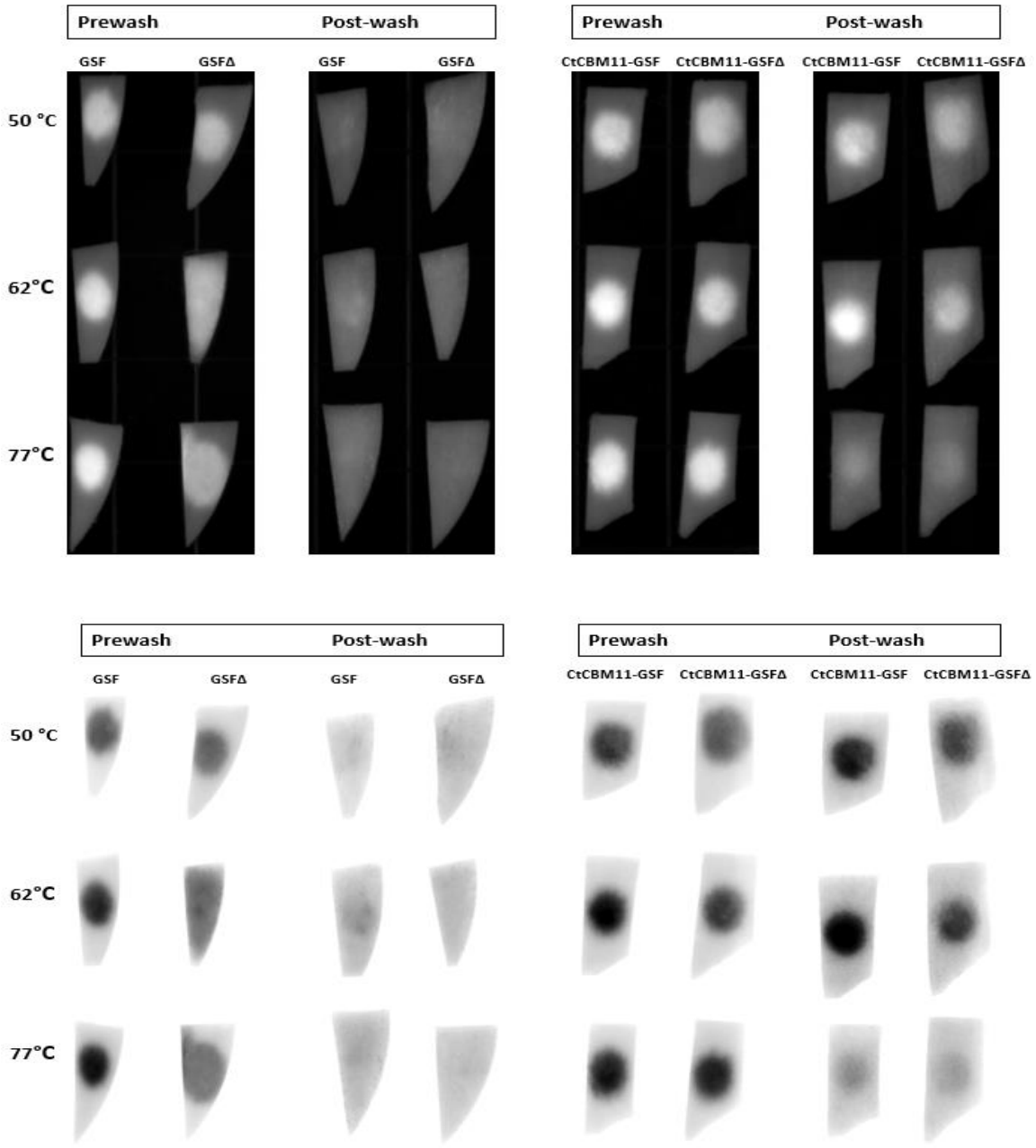
Figure 15: Functional evaluation of thermally denatured and refolded temperature tunable CtCBM11-GSF and CtCBM44-GSF constructs against microcrystalline cellulose (Avicel) at 50 °C and at 10 °C above the established T_m. Non-heat-treated samples are shown in blue, heat treated samples are shown in orange. The fluorescence signals at each temperature correspond to the available CtCBM-GSF or GSF

in the supernatant after incubation with the available Avicel substrate. The fluorescence units provided for the load corresponds to the amount of CtCBM-GSF available in the supernatant prior to binding to Avicel at each temperature condition. The results provided were taken from three different trials with freshly prepared samples. GSF without a fused CtCBM is not observed to bind to Avicel at any temperature. Upon cooling from a thermally perturbed state at temperatures 5°C above the established T_m, both CtCBM11-GSF and CtCBM44-GSF become fully active with virtually 100% efficiency compared to the non-heated controls. CtCBM44-GSF appears to bind Avicel with much greater affinity than CtCBM11-GSF and appears to still bind Avicel at temperatures well past its T_m, while CtCBM11-GSF appears to mostly be released into the supernatant once incubated at a temperature above its established T_m.

A complimentary set of data is found in Figure 16 where refolding functional evaluations are done on cellulose membranes as opposed to pelleted Avicel. In contrast to the previous experiment where fluorescence signal is taken from the amount of CtCBM-GSF found in solution, in this experiment fluorescence signals are imaged using cellulose membranes to see how much of CtCBM is bound to the substrate. Much like the data shown in Figure 15, GSF does not appear to have an affinity for a cellulose substrate. Data provided in Figure 16 further provides evidence that both CtCBM11-GSF and CtCBM44-GSF refold and regain a functional conformation upon cooling with virtually 100% efficiency compared to non-heated controls. In this experiment, we are additionally able to show efficient temperature tunability where binding ability can be essentially turned on or off with a small change in temperature of 5°C - 10°C. The CtCBM-GSF constructs are functional at all temperatures below the T_m as evident by the fluorescent cellulose membranes. Much like the data shown in Figure 15, Figure 16 shows that following denaturation once the temperature is raised to 10°C greater than the T_m, fluorescence almost entirely disappears from the cellulose membrane. In contrast to the data shown in Figure 15 for CtCBM44-GSF, we see that in this experiment once the temperature is raised to 10°C greater than the T_m, fluorescence almost entirely disappears

from the cellulose membrane as CtCBM44-GSF loses its affinity for the cellulose acetate/cellulose membrane—our rationale for these different results are outlined in greater detail in the Discussion section of this paper.

(A)



(B)

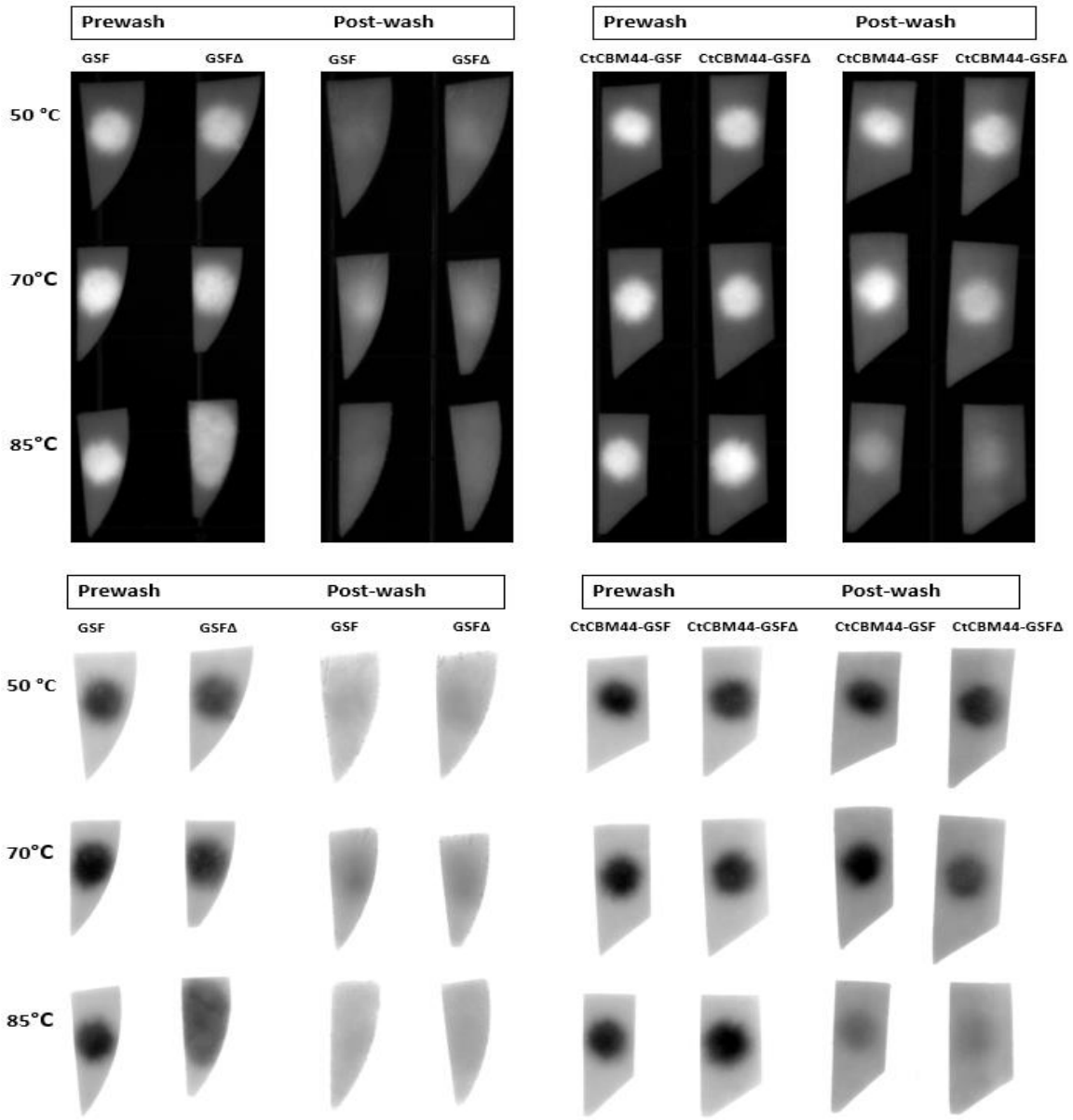


Figure 16: Functional evaluation of thermally denatured and refolded CtCBM11-GSF and CtCBM44-GSF using cut out pieces of cellulose acetate/ cellulose nitrate 0.45um membranes from Millipore. All samples that were heat treated are labeled Δ. Figure 16A shows CtCBM11-GSF and CtCBM11-GSFΔ. As a control, GSF without a fused CtCBM was incubated in identical conditions as the temperature tunable CtCMB-GSF constructs. GSFΔ was also heated to conditions that matched the heat treatment of CtCBM11-GSFΔ. Figure 16B shows CtCBM44-GSF and CtCBM44-GSFΔ. GSF and GSFΔ without a fused CtCBM were once again used as controls. Each membrane has been blotted with 10ul of fluorescence normalized samples that contain equal fluorescence units per ul. Each membrane was imaged under UV light before and after

washing in buffer at 50°C, 5°C below the established T_m where each CtCBM-GSF construct should still be functional, and denaturing and non-functional conditions 10°C above the established T_m. In both figures 16A and 16B, images have been shown in normal UV light and with inverted contrast as a multiple representation for the same results acquired from the refolding binding assay. The images shown are representative images of three different trials done with freshly prepared samples. GSF is not observed to bind to cellulose membranes at any temperature. Upon cooling from a thermally perturbed state at temperatures 5°C above the established T_m, both CtCBM11-GSF and CtCBM44-GSF become fully active with virtually 100% efficiency compared to the non-heated controls but lose their affinity for the cellulose membrane once heated to 10°C above the T_m.

The results presented thus far strongly support that CtCBM11 and CtCBM44 are both temperature tunable CBMs capable of spontaneously refolding and regaining binding function when cooling from a thermally denatured state. Our next efforts in this study focused on trying to explain the properties of CtCBM11 and CtCBM44 that allow for this unique temperature tunability, which is both inherently interesting and could also help improve future screening efforts for temperature tunable CBMs. From the revolutionary work published by Anfinsen (1973), we have learned that all the information required for a protein to form native contacts and to fold into its native state is present in the amino acid sequence. Since ABSCO of proteins measures how closely residues that contact one another in the native state are in the primary sequence of the protein, and due to the known correlation between the ABSCO of proteins and their folding kinetics, we chose to compare the ABSCO of CtCBMs investigated in this study as well as CtCBMs in families we have not yet studied. We used comparisons between the ABSCO of CtCBMs to understand why CtCBM11 and CtCBM44 reversibly denature and refold to their functional native state, and why CtCBM30 does not. Table 4 and Figure 17 show the results of comparisons made between CO calculations of CtCBMs that have been grouped into different families. As mentioned in the Materials and

Methods section of this paper, the sequences we have used for this part of our study were largely taken from (Walker et al. 2015) due to the limited number of CtCBMs structures that exist in the PDB without being part of an entire enzymatic complex. We opined we could use amino acid sequences for protein structure prediction done by RaptorX for 3D (Källberg et al. 2012) as a convenient and reliable way to get around this issue. While we accept the unavoidable limitations that exist in protein structure prediction, we are hopeful that enough CtCBM families of different fold families have been uploaded to the PDB to obtain reliable structure predictions. Additionally, when superimposing the RaptorX models of CBMs with actual structures of CBMs (without a fused CD) from the PDB, the predicted models and actual CBM structures were in complete agreement with a RSMD of less than 2 and high-quality P-value. This positive quality control result increased our confidence that the predicted models by RaptorX were reliable structural models for CtCBMs.

Predicted 3D structures from CtCBMs sequences were downloaded as PDB files from RaptorX. Given an amino acid sequence, RaptorX predicts its secondary and tertiary structures, contacts, solvent accessibility, disordered regions, and binding sites. RaptorX also assigns some confidence scores to indicate the quality of a predicted 3D model: P-value for the relative global quality, GDT (global distance test) and uGDT (un-normalized GDT) for the absolute global quality, and modeling error at each residue. The P-value evaluates the relative quality of a model compared to randomly generated models for the query. The smaller the P-value, the higher quality the model. For mainly alpha proteins, P-value less than 10^{-3} is a good indicator. For mainly beta proteins, P-value less than 10^{-4} is a good indicator. For a protein with more than 100 residues, as in the case of

most CBMs, a uGDT greater than 50 is a good indicator of a reliable model. All the used structures had p-values indicative of a high-quality model. More on the way which RaptorX threading works can be found in detail at (http://raptorx.uchicago.edu/).

Table 4: Table of CBMs used for structure prediction from sequence along with their gene locus, gene structure, amino acid chain length, and ABSCO calculation. This table was taken from supplementary information found in (Walker et al. 2015). Abbreviation Key: RsgI-N - Anti-sigma factor N-terminus, CelD-N - N-terminal Ig-like domain of cellulase, CE - carbohydrate esterase, GH- glycoside hydrolase, GT - GlycosylTransferases, PL - polysaccharide lyase, SLH - S-layer homology domain. * Indicates that these CBMs are from *Thermoanaerobacterium* sp. MYST/2012-07.

CBM family construct	**Gene Locus**	**Gene Domains (CAZY / NCBI / Pfam)**	**Chain Length (N)**	**ABSCO**
3-1	Cthe_0059	RsgI1_RsgI-N,CBM3	83	8.8295
3-2	Cthe_0271	CBM3	83	8.5615
3-3	Cthe_0433	GH9,CBM3,Dockerin	78	8.0661
3-4	Cthe_0578	CelR_GH9,CBM3,Dockerin	86	8.7499
3-5	Cthe_0625	CelQ_GH9,CBM3, Dockerin	87	8.4112
3-6	Cthe_2360	CelU_GH9,CBM3,CBM3,Dockerin	92	8.9167
3-7	Cthe_2360	CelU_GH9,CBM3,CBM3,Dockerin	83	8.5242
3-8	Cthe_2423	CBM3	85	8.5529
3a	Cthe_3077	CipA Scaffoldin_2XCohesin,CBM3,7XCohesin,Dockerin	159	30.973
4-1	Cthe_0412	CelK_CBM4,CelD-N,GH9,Dockerin	139	17.631

4-2	Cthe_0413	CbhA_CBM4,CelD-N,GH9,CBM3,Dockerin	139	17.782
4-3	Cthe_1257	CBM3,CBM4	130	17.557
4-4	Cthe_2809	LicA_SLH,CBM54,GH16,CBM4,CBM4,CBM4,CBM4	123	18.905
4-5	Cthe_2809	LicA_SLH,CBM54,GH16,CBM4,CBM4,CBM4,CBM4	133	16.643
4-6	Cthe_2809	LicA_SLH,CBM54,GH16,CBM4,CBM4,CBM4,CBM4	131	17.386
4-7	Cthe_2809	LicA_SLH,CBM54,GH16,CBM4,CBM4,CBM4,CBM4	124	16.437
6	Cthe_2972	XynA/U_GH11,CBM6,Dockerin,CE4	121	23.992
9-1*	XynX	XynX_CBM22,GH10,CBM9,CBM9,SLH	185	28.3112
9-2*	XynX	XynX_CBM22,GH10,CBM9,CBM9,SLH	169	25.759
11	Cthe_1472	CelH_GH26,GH5,CBM11,Dockerin	167	31.187
13	Cthe_0661	GH43,CBM13,Dockerin	134	19.32
16	Cthe_3095	CBM16,GT39	103	22.722
22-1*	XynX	XynX_CBM22,GH10,CBM9,CBM9,SLH	138	17.942
22-2	Cthe_0912	XynY_CBM22,GH10,CBM22,Dockerin,CE1	110	17.973
22-3	Cthe_0912	XynY_CBM22,GH10,CBM22,Dockerin,CE1	136	17.849
22-4	Cthe_1838	XynC_CBM22,GH10,Dockerin	132	17.535
22-5	Cthe_2590	XynD_CBM22,GH10,Dockerin	133	17.137
25	Cthe_1080	CBM25	81	13.647
30	Cthe_0624	CelJ_CBM30,GH9,GH44,Dockerin,CBM44	197	27.856

32	Cthe_0821	Man5A_GH5,CBM32,Dockerin	116	22.473
34	Cthe_0795	CBM34,GH13	107	15.564
35-1	Cthe_2137	GH39,CBM35,CBM35,Dockerin	131	25.338
35-2	Cthe_2137	GH39,CBM35,CBM35,Dockerin	137	24.842
35-3	Cthe_0246	Dockerin,CBM35,PL11	124	24.547
42	Cthe_0015	CBM42,Dockerin,GH43	130	14.443
44	Cthe_0624	CelJ_CBM30,GH9,GH44,Dockerin,CBM44	151	28.091
48	Cthe_2191	CBM48,CBM48,GH13	105	13.756
50	Cthe_1800	CBM50,CBM50,GH18	45	7.60347
54	Cthe_2809	LicA_SLH,CBM54,GH16,CBM4,CBM4,CBM4,CBM4	236	11.669

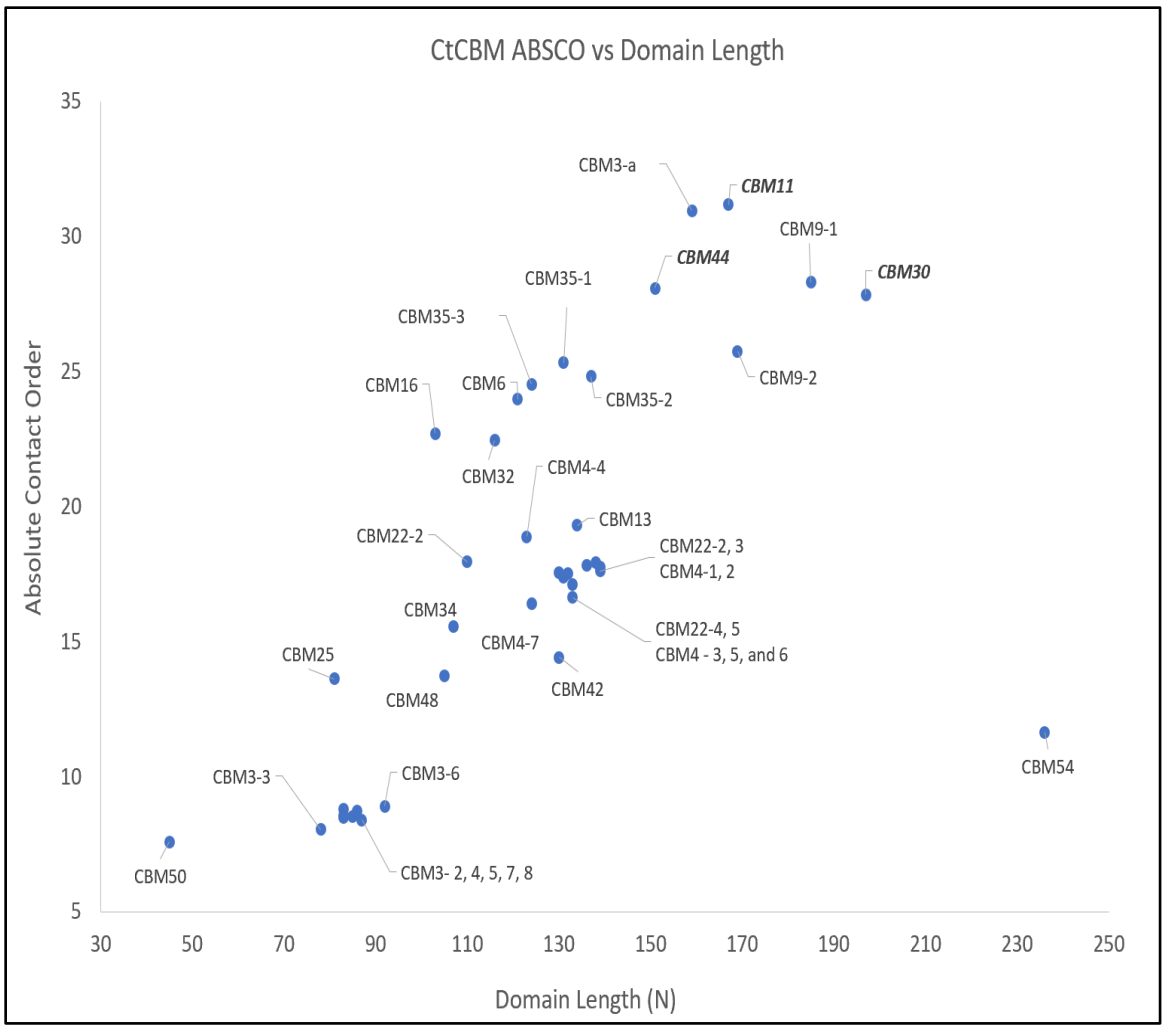


Figure 17: ASCO is plotted as a function of domain length. Each data point corresponds to a specific CBM family. Nomenclature of CBMs matches the nomenclature of CBM constructs in Table 4 above. All CO calculations were done using a predicted 3D model generated by the software RaptorX. By plotting Absolute CO as a function of domain length, useful comparisons between the topology of different CBM families and domain length are possible. Temperature tunable CtCBM 11 and 44, and CtCBM30 which was also investigated in this study, are bolded and italicized and were found to have some of the highest ABSCO values.

We found that CtCBM11, CtCBM30, and CtCBM44 all have some of the highest CO values of all CtCBM families: ABSCO= 31.18, ABSCO= 27.85, and ABSCO=28.05, respectively. In fact, the lowest CO values were found in CBM families we have not studied: one member from the CtCBM50 family (ABSCO=7.6, Figure 23) and many

members that belong to the CtCBM3 family (Average ABSCO=8.58). These results suggest that ABSCO cannot be used to explain the observed denaturation and refolding behavior of CtCBM11 and CtCBM44. Considering these findings, we looked further into structural biology to help elucidate a meaningful explanation for our results. To do this, we used RaptorX to compare CtCBM11, CtCBM30, and CtCBM44 by structural alignments. We hypothesized that there would be clearly observable structural fold differences between CtCBM11 and CtCBM30, and CtCBM44 and CtCBM30, but some clear similarities between CtCBM11 and CtCBM44. Several scores were generated by the structural alignment generated by RaptorX, which are shown in Figure 18: Lali, RMSD, uGDT(GDT), and the TMscore. Of most importance are RMSD and the TMscore. RMSD is the mean-square deviation for which scores < 2 angstroms represents a high-quality match. For multiple alignments, RMSD is calculated only on the core residues. TMscore is between 0 to 1 and is an approximate but quantitative criterion for protein topology classification. If TMscore >0.6, it is very likely (90% of chance) that two proteins share a similar fold. When TMscore <0.4, it is very likely (90% of chance) that two proteins have different folds. More on the Lali score and on uGDT(GDT), the unnormalized GDT score, and supplementary information on the TMscore can be found at http://bioinformatics.oxfordjournals.org/content/26/7/889, Wang et al. 2011, and Wang et al. 2013.

What we have found is that despite the RMSD being a little higher than ideal, the TMscore reveals that CtCBM11, CtCBM30, and CtCBM44 are clearly all members of the β-sandwich fold family and are “very likely” (90% of chance) to share a similar fold. These results suggest that the topology of the native fold of each protein alone does not

help account for the reversible denaturation and refolding of CtCBM11 and CtCBM44 compared to CtCBM30. Despite falling short of finding a predictive indicator of which proteins will reversibly refold following denaturation up until this point, Figures 19-22 finally shed some light on what may account for the reversible denaturation in CtCBM11 and CtCBM44. As shown in Figure 19, as is common in most CBMs that share the β-sandwich fold, both CtCBM11 and CtCBM44 bind calcium ions. In fact, CtCBM11 binds two calcium ions—surprisingly, CtCBM30 does not. As explored further in the Discussion section of this paper, we believe that calcium ion binding could contribute to changes in folding equilibrium when CtCBM11 and CtCBM44 are heated to 5-10°C higher than the reported T_m. Further, as shown in Figures 20-22 results from RaptorX structure models of CtCBM11, CtCBM30, and CtCBM44 show that CtCBM11 and CtCBM44 have similar β-sheet character and only small regions of disorder on the N-terminal ends (3% and 2%, respectively). In stark contrast to CtCBM11 and CtCBM44, an astonishing 12% of the CtCBM30 structure contains regions of high intrinsic disorder on both the N-terminal and C-terminal end. As discussed in more detail in the next section, we believed this region of intrinsic disorder could play a detrimental role in the refolding pathway of CtCBM30 when cooling from a thermally perturbed state. Figure 24, which was taken from the PDB, further supports this hypothesis as the crystal structure of CtCBM30 (PDB code: 2C24) shows that the N-terminal and C-terminal regions (1-14 and 186-197, respectively) of the protein fail to form any secondary structure that could help stabilize the native fold of CtCBM30 during refolding.

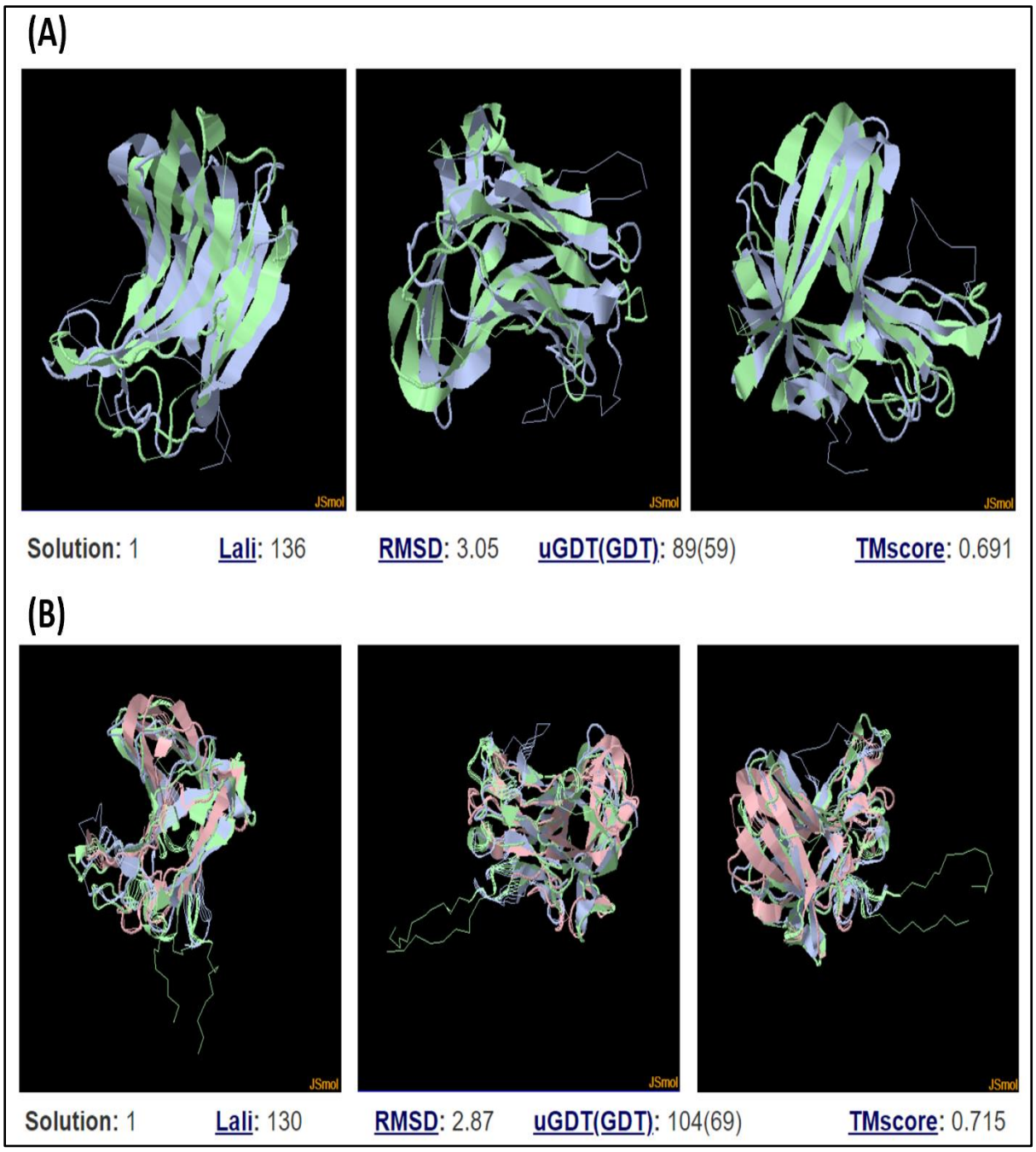


Figure 18: (A) RaptorX structural alignment of CtCBM11 (blue) and CtCBM44 (green). (B) RaptorX structural alignment of CtCBM11 (blue), CtCBM30 (green), and CtCBM44 (pink). The TMscore from these structural alignments suggest that CtCBM11, 30, and 44 all have the same β-sandwich fold.

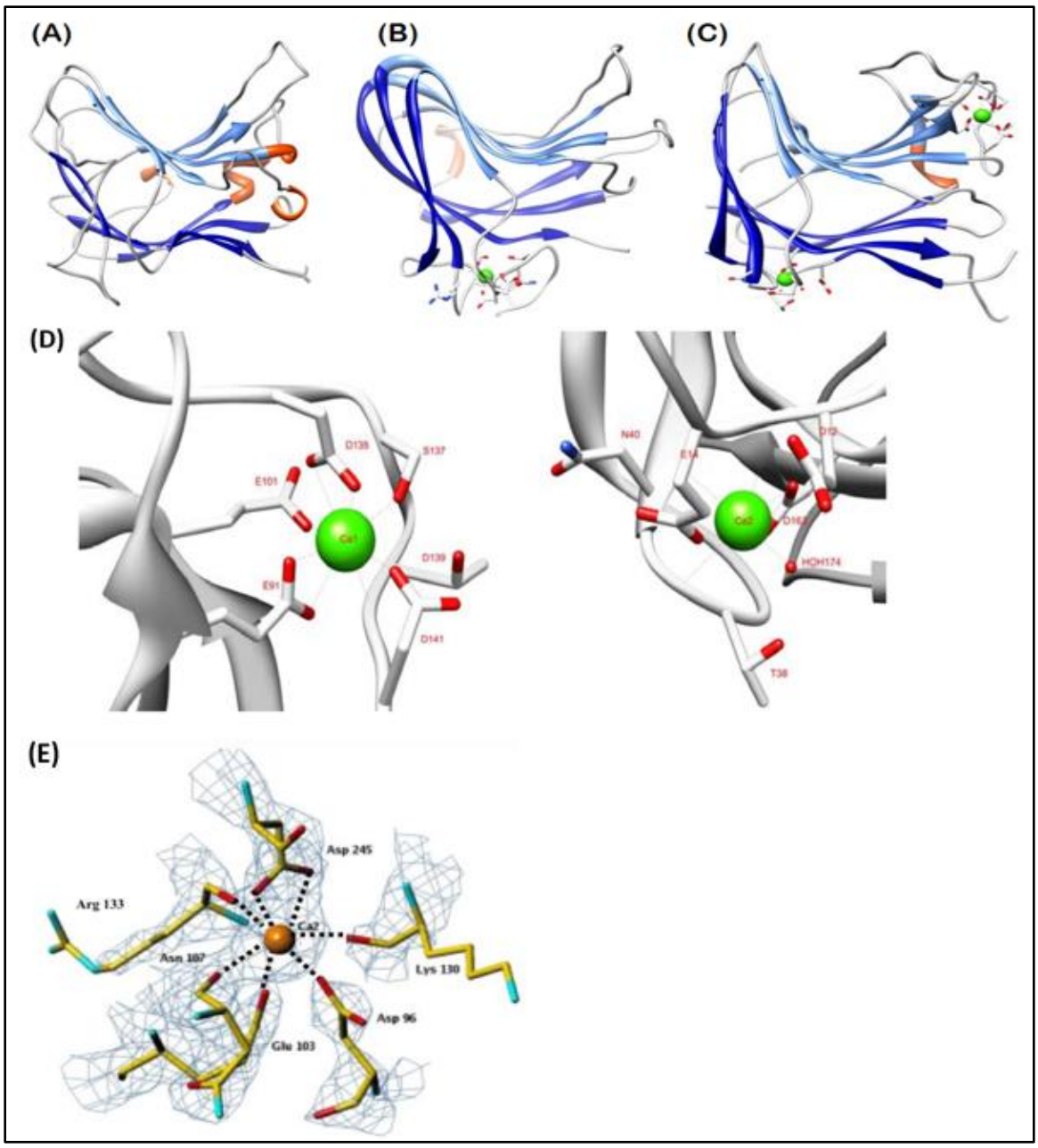


Figure 19: 3D structure of CtCBM30 (A), CtCBM44 (B), and CtCBM11 (C) obtained by X-ray crystallography (Viegas 2012). All structures reveal a classical distorted β-jelly roll fold consisting of two six-strand anti-parallel β-sheets, which forms a convex side (light blue, binding pocket) and a concave side (dark blue). As shown in CtCBM44, one calcium ion is depicted as a green sphere. In the case of CtCBM11, two calcium ions are shown (the residues that bind calcium are depicted as sticks). The α-helical regions are depicted in orange. CtCBM30 does not have a calcium ion binding site. (D) Both calcium ions in CtCBM11 show an octahedral coordination and are bound to main chain and side chain oxygens. (E) Calcium binding site (Ca2) in the CBM44 domain and their corresponding coordinating amino acid residues. (Najmudin et al. 2006). The Ca1 site in a neighboring domain not shown.

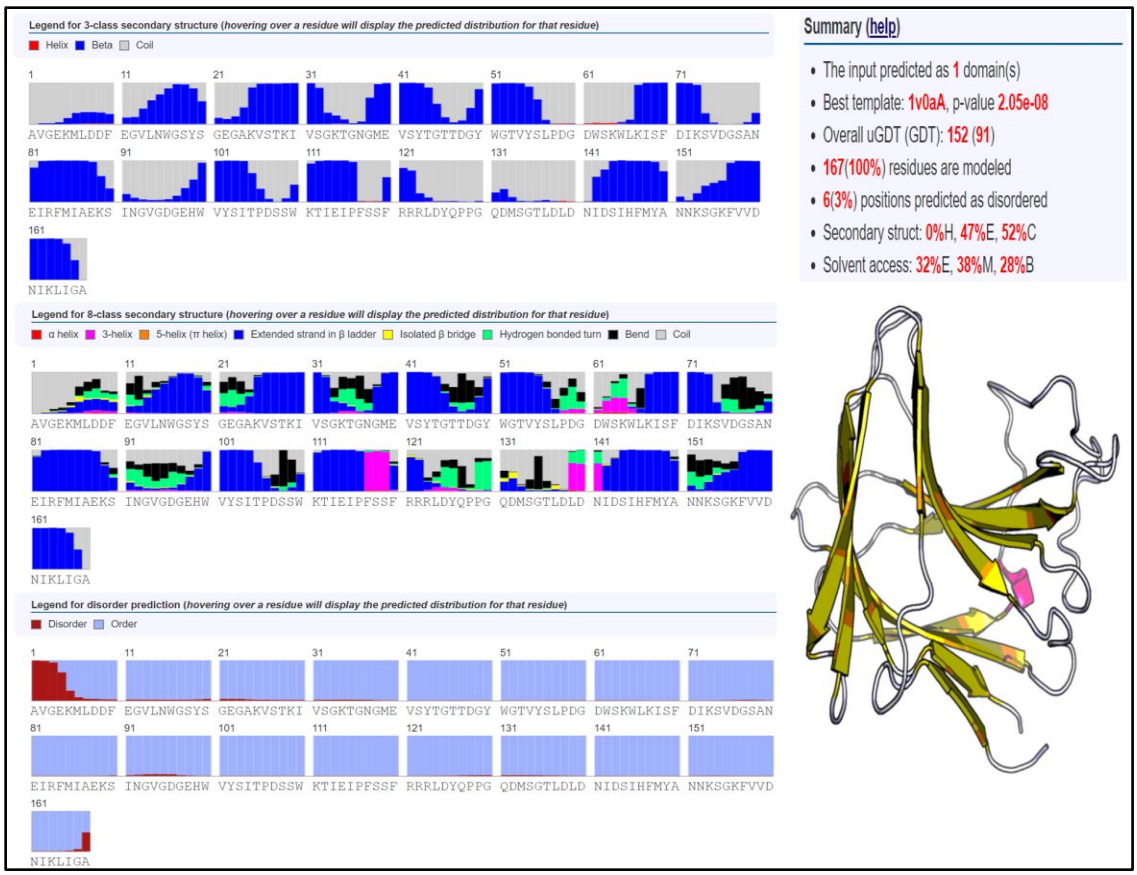


Figure 20: CtCBM11 RaptorX structure prediction result, which uses the known structure for CtCBM11 in the PDB (1v0aA) as a template for threading. All 167 (100%) of CtCBM11 residues are modeled in the shown structure for which a p-value of 2.05×10^{-8} was obtained. Only 3% of the amino acids are found to be disordered in the 3D structure, 0% of the structure contains an alpha-helix, 47% contains beta-sheets, and 52% of the structure is made up of loops. Solvent accessibility is divided into three states: *buried, medium, and exposed*. A *buried* protein contains less than 10% of the protein exposed, an *exposed* protein contains a score larger than 42%, and a *medium* protein contains between 10% and 42% of the protein exposed. CtCBM11 contains 32% of the protein solvent exposed, 28% of it buried, and 38% medium, which corresponds to a protein of medium solvent accessibility.

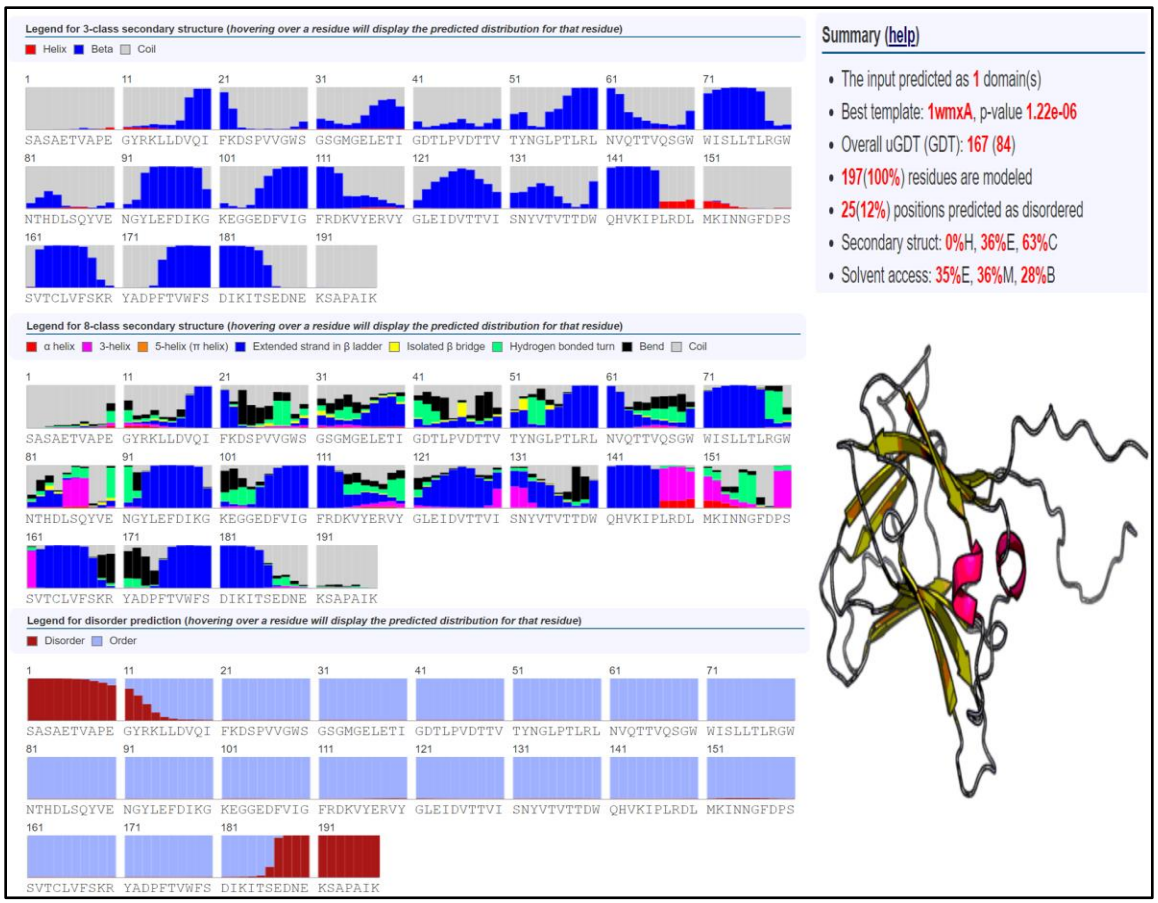


Figure 21: CtCBM30 RaptorX structure prediction result, which uses the known structure for CtCBM30 in the PDB (1wmxA) as a template for threading. All 197 (100%) of CtCBM30 residues are modeled in the shown structure for which a p-value of $1.22x10^{-6}$ was obtained. A total of 12% of the amino acids are found to be disordered in the 3D structure, 0% of the structure contains an alpha-helix, 36% contains beta-sheets, and 62% of the structure is made up of loops. Solvent accessibility is divided into three states: *buried, medium, and exposed* as described previously. CtCBM30 contains 35% of the protein solvent exposed, 28% of it buried, and 36% medium, which corresponds to a protein of medium solvent accessibility.

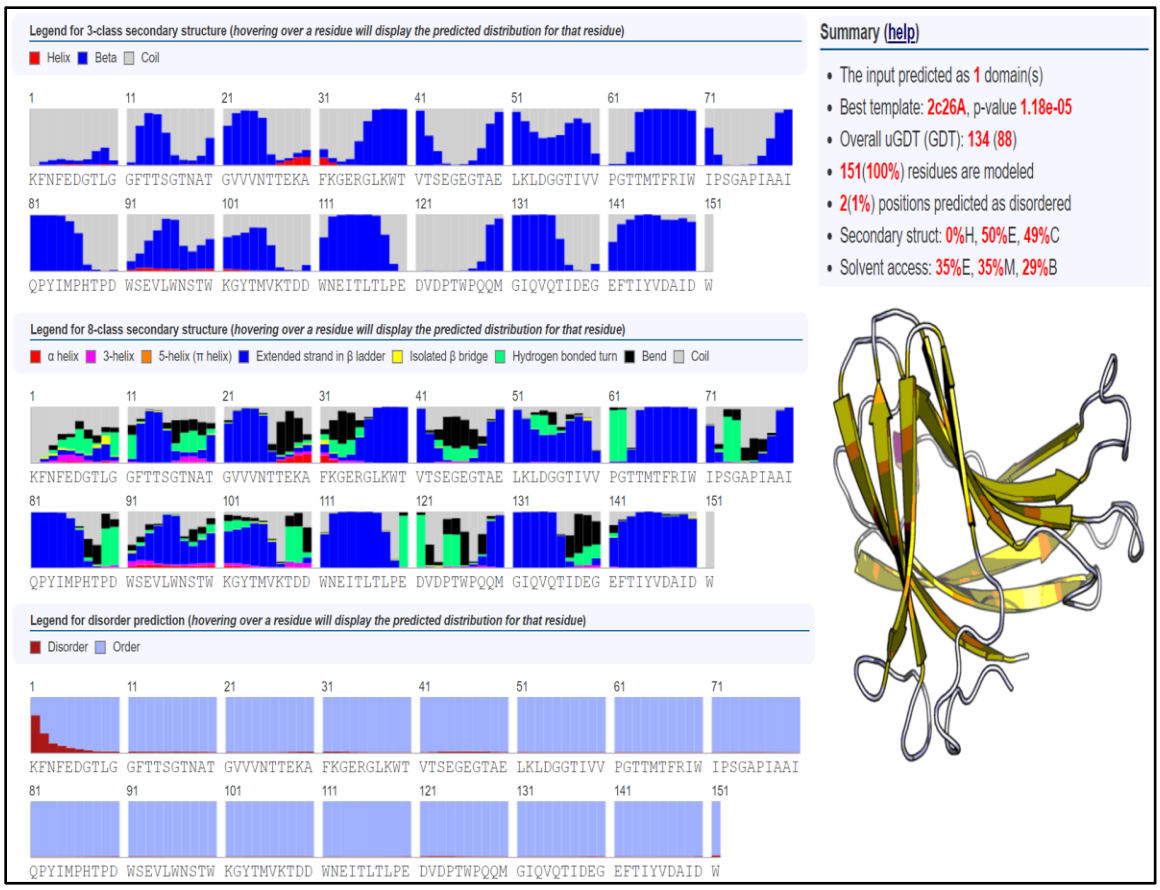


Figure 22: CtCBM44 RaptorX structure prediction result, which uses the known structure for CtCBM44 in the PDB (2c26A) as a template for threading. All 151 (100%) of CtCBM44 residues are modeled in the shown structure for which a p-value of 1.18×10^{-5} was obtained. Only 2% of the amino acids are found to be disordered in the 3D structure, 0% of the structure contains an alpha-helix, 50% contains beta-sheets, and 49% of the structure is made up of loops. CtCBM44 contains 35% of the protein solvent exposed, 29% of it buried, and 35% medium, which corresponds to a protein of medium solvent accessibility.

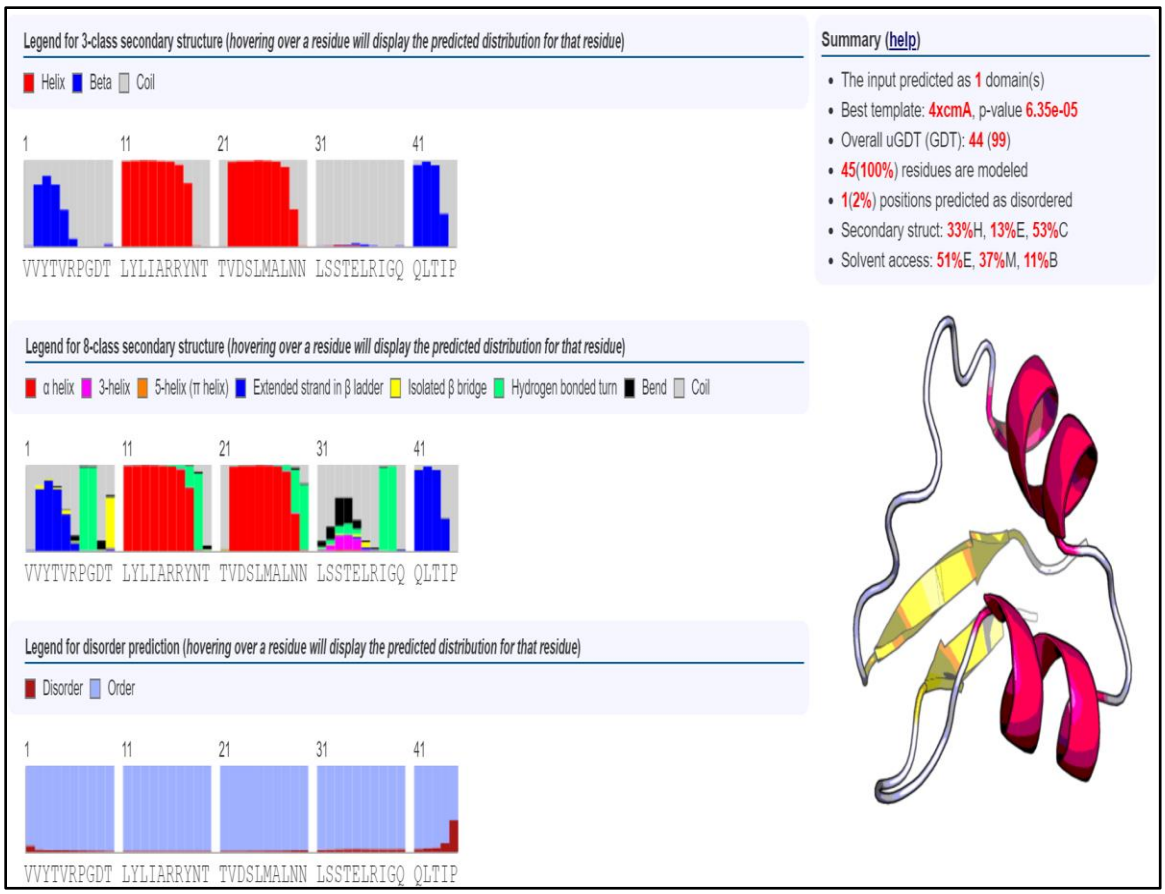


Figure 23: CtCBM50 RaptorX structure prediction result, which uses the PDB file 4xcmA template for threading. This construct resulted in the lowest ABSCO: 7.60347 and an amino acid chain length of only 45 amino acids, which is intriguing as it has the lowest ABSCO value of any CtCBM present in the literature. The fold closely resembles the hevein fold of family 6. All 45 (100%) of CtCBM50 residues are modeled in the shown structure for which a p-value of $6.3x10^{-5}$ was obtained. Only 2% of the amino acids are found to be disordered in the 3D structure, 33% of the structure contains an alpha-helix, 13% contains beta-sheets, and 53% of the structure is made up of loops. CtCBM11 contains 51% of the protein solvent exposed, 11% of it buried, and 37% medium, which corresponds to an exposed protein with great solvent accessibility.

Figure 24: Secondary structures present in the native crystal structure of CtCBM30 (PDB: 2C24). Secondary structures are only present between amino acids 15-185, which are composed of 3 helices (10 residues, 4%) and 15 β-strands (77 residues, 37%).

DISCUSSION

The present study shows that CtCBMs of the families 11 and 44 are capable of regaining binding function following thermal denaturation. We also show that CtCBM30, which is a CBM belonging to the bifunctional modular cellulase CtCel9D-Cel44A (CtCBM44 also belongs to this cellulase) does not efficiently refold following thermal denaturation. These CtCBMs were selected using three screening parameters: (1) thermostable CBMs from a cellulolytic thermophile as tuning down the melting temperature (T_m) of a CBM to be lower than the T_m of a fused CD will be much easier than stabilizing a mesophilic CBM domain up to a higher T_m; (2) CBMs that are known to bind to substrates found in lignocellulose as opposed to searching for putative CBMs without confirmed function; and (3) CBMs that exist in the PDB with a resolved 3-D structure, which will facilitate cloning, expression, and rationale engineering efforts. The main tool used to facilitate these screening efforts was the CAZy database, which lists thousands of characterized and putative CBMs organized into 84 families based on sequence, structural fold, and substrate affinity. Through our screening, we came across a study described by Hirano et al., (2016) where 40 components from the cellulosome of *H. thermocellum* ATCC 27405 (formerly *C. thermocellum* ATCC 27405), which can grow at temperatures between 50 and 68°C (Akinosho, et al 2014), were synthesized in-vitro using a wheat germ cell-free protein synthesis system. These 40 components represented exoglucanases and endoglucanases (cellulases), hemicellulases, pectinases, xylanases, xylan esterases, mannanases, and other enzymes shown to be important for synergistic cellulose degradation—most of which contain a carbohydrate binding module. A search in the CAZy database confirmed that *H. thermocellum* has cellulases and auxiliary

enzymes that have many CBMs classified in 18 different families. A study by Walker (2015) showed that some members of the 18 CBM families found in *H. thermocellum* could be fused to the CD of CelE (which can hydrolyze cellulose, mannan and xylan), and result in the formation of novel enzymes of higher activity than those found in nature. Due to the tremendous body of work published on CBMs from *H. thermocellum,* we decided to focus our investigation on CtCBMs. In fact, the three CtCBMs described in this study (CtCBM11, CtCBM30, and CtCBM44) were chosen because they were shown to have high affinity to hydroxyethyl cellulose by AGE (Walker et al. 2015) and have also had their 3-D structure resolved and uploaded to the PDB database. These characteristics made CtCBM11, CtCBM30, and CtCBM44 promising candidates to assess for temperature tunability in hopes of finding a CBM that can be used to release and recycle lignin-trapped cellulases following saccharification reactions.

Figure 11 shows that CtCBM11, CtCBM30, and CtCBM44 are (as expected) thermostable CBMs— CtCBM11 has a T_m of 67°C and both CtCBM44 and CtCBM30 have a T_m of 75°C. Determining the melting temperatures of these CBMs was important to establish in order to optimize conditions for thermal denaturation and refolding by cooling instead of simply boiling the samples. This is because high-temperature denaturation of proteins results from changes in the entropy of the solvent water, diminishing the hydrophobic effect, and the increased entropy of the protein chain itself at higher temperatures. Some proteins, but not others, spontaneously refold to their native structure when returned to temperatures below their T_m. There are currently no bioinformatics tools that readily predict which proteins spontaneously refold to the active, native sate following thermal denaturation and which do not.

To evaluate the refolding of CtCBM11, 30, and 44 following thermal denaturation we used the Fold-N-Glow complementation assay from SandiaBiotech between CtCBM-GFPS11 constructs and GFPS1-S10 split GFP (Figure 12). CtCBM11-GFPS11 without heat treatment obtained a gain in fluorescence of 103 RFU, while CtCBM1-GFPS11 that had been thermally denatured and refolded by cooling obtained a gain in 108 RFU. As shown by the slopes of these graphs, the rate at which CtCBM-GFPS11 and GFPS1-S10 complemented were nearly identical, which suggests that CtCBM-GFPS11 adopts the native compact, stable fold after heating above its T_m and then cooling. Further, these results indicate that this conformation can be reached without aggregation or precipitation that would otherwise prevent the formation of the complete GFPS1-S11 fluorophore. A similar result is also observed with CtCBM44-GFPS11, although the gain in fluorescence in this case went from 158 RFU in the non-heated controls to 184 RFU following thermal denaturation and refolding by cooling. Though not a tremendous difference in RFU, this gain in fluorescence can be explained if one considers that perhaps the starting materials contained small populations of CtCBM44-GFPS11 that were soluble aggregates or monomeric non-native constructs that prevented complementation with GFPS1-S10. However, once denatured and allowed to cool, the equilibrium conditions of the assay could have favored the folded state compared to the original misfolded conformation. Any misfolded protein originally present in the sample could have thus been given another chance at properly folding, which could explain the 16% gain in RFU compared to non-heat-treated controls.

Following thermal denaturation, less than 50% of the population of CtCBM30-GFPS11 could complement GFPS1-S10, suggesting that during the refolding process

while the solution cooled, most of the CtCBM30-GFPS11 in solution misfolded and/or precipitated. As a follow up experiment, CtCBM30-GFPS11 was heated to 60^{o}C (much lower than the reported T_m), which resulted in fluorescence gains that matched the non-heated control as the heat-treated sample did not denature. This supports that upon heating to 5^{o}C above the T_m and denaturing CtCBM30, the complementation between CtCBM30-GFPS11 and GFPS1-S10 does not occur due to a significant portion of the CtCBM30-GFPS11 misfolding during the cooling process. Notably, while the total gain in fluorescence was much lower compared to the complementation assays between CtCBM11-GFPS11 and GFPS1-S10, and CtCBM44-GFPS11 and GFPS1-S10, this is likely just a result of lower protein concentrations obtained by the IMAC purified samples. Future experiments could easily fix this issue by loading more CtCBM30-GFPS11 into the assay.

While these differences in refolding following cooling from a thermally perturbed state are interesting, justifying the differences in our results is far from trivial. In his pioneering work on refolding ribonuclease, Anfinsen (1973) concluded that the 3D structure of a protein is determined by the amino-acid sequence and that a protein regains its native state when the denaturant condition is removed. Thanks to this work and advances in protein biophysics, we now understand that as a protein folds it must form contacts that stabilize the native state over all other non-native (misfolded) conformations. However, non-native contacts that stabilize misfolded conformations are in most cases unavoidable and result in local minima—this phenomenon is referred to as "energetic frustration." How high the energy barriers are between unfolded states and the folded native state, or between local minima and the folded state, depends on many

factors including the entropy of the main chain, steric constraints experienced by amino acids during folding, the formation of secondary elements, and the topological requirements of secondary elements (among other parameters like changes in solvent environments as water is unbound following hydrophobic collapse). This understanding along with a growing body of evidence from protein folding experiments has strongly suggested that most proteins fail to refold correctly from the unfolded state and instead precipitate, aggregate, or reach more stable and non-native kinetic traps. Fortunately, there are certain biophysical qualities of proteins that can favor spontaneous refolding following denaturation and bias populations of unfolded proteins towards a native state. In particular, proteins that have small molecular weights (Bhattacharjee and Das 2000), secondary elements of low contact order (Plaxco et al. 1998), and metal ion binding sites (Anfinsen 1973), have been shown to be good candidates to refold following denaturation due to having "downhill" protein folding pathways with fewer non-native kinetic traps.

Knowing little about the actual structures of CtCBM11, CtCBM30, and CtCBM44 at the time, we hypothesized that members of CBM families belonging to fold families with the lowest ABSCO would be ideal candidates for reversible denaturation and refolding. ABSCO is a measure of how close residues that contact one another in the tertiary structure of a protein are in the primary sequence of the protein. Contacts made by amino acids near each other in the native fold of a protein are not necessarily near each other in the primary sequence. Thus, native contacts between sequentially close residues in proteins with a small ABSCO are expected to form more frequently than native contacts between sequentially distant residues in proteins with a large ABSCO. Studies have shown that proteins with a small ABSCO are known to correlate with "downhill" protein

folding that avoids non-native, kinetic traps in the folding process due to the formation of local contacts that stabilize the native fold much more than other non-native contacts (Plaxco et al. 1998).

We explored whether ABSCO could be used to explain reversible denaturing and refolding of CtCBMs 11 and 44, and in hopes of potentially finding a clear way to screen for temperature tunable CBMs in the future. To do this we used RaptorX to predict the structure from sequence of many CBMs. These predicted structures were then used for ABSCO calculations and plotted against amino acid chain length, as shown in Figure 17. Ideally, CBMs with resolved structures could be used directly, but unfortunately our database screening on CAZy showed that most CBM structures exist as part of a cellulase, or another neighboring domain from the parent enzyme, which would not give a representative ABSCO value. While we expected that perhaps a trend could be observed between different CBM family members, we were surprised to find that CtCBM11 and CtCBM44 had ABSCO larger than most other CBM families. When comparing the three tested CBMs in this study, the ABSCO of CtCBM44 (ABSCO=28.05, RELCO=0.186) is slightly greater than the ABSCO of CtCBM30 (ABSCO= 27.85, RELCO= 0.141), and the ABSCO of CtCBM11 (ABS CO= 31.18, RELCO=0.187), is also higher than the ABSCO of CtCBM30.

Further, the 3-dimensional structures from Figure 18 and Figures 20-22 help account for the large ABSCO values of CtCBM11, CtCBM30, and CtCBM44. These large values appear to be due to the high β-sheet character present in CtCBM11, CtCBM30, and CtCBM44, which are driven mainly by non-local interactions between amino acids in the native structure of the protein that are far apart in the amino acid chain. Additionally, we

were further surprised to find that CtCBM11, CtCBM30, and CtBM44 had similar ABSCO values to each other. In fact, structural alignments of CtCBM11, CtCBM30, and CtCBM44 shown in Figure 18 strongly suggest these CBMs have the same fold (i.e., the β-sandwich family), as shown by the TMScore >0.60. This score suggests they are 90% likely to share the same fold, which would explain why the obtained ABSCO values for CtCBM11, 30, and 44 are so similar. What is more surprising is that the addition of CtCBM30 to the structural alignment of CtCBM11 and CtCBM44 improved TMScore and uGDT scores. This can be reasoned by considering the gene structure of CtCBM30 and CtCBM44 as they are both part of the gene that lead to the expression of the same parent enzyme (Table 4). This suggests that CtCBM30 and CtCBM44 likely arose from gene duplication and drifted apart in sequence over time, but the folds and substrate affinity were partially conserved. This would explain why both CtCBM30 and CtCBM44 can be grouped in the same fold family and in the same glycan-chain binding, Type B family.

While our original hypothesis was not correct, the concept of ABSCO is still intriguing and was worth further exploring. To further assess differences in ABSCO with actual topologies of CBMs, we compared the topology of the CBM with the lowest ABSCO (as well as other properties reported by RaptorX) to the structures of CtCBM11, CtCBM30, and CtCBM44. The structure for the CBM with lowest ABSCO is shown in Figure 23—CtCBM50. Interestingly, in contrast to the β-sandwich fold of Type B CBMs, the fold of CtCBM50 suggests that it could be grouped in the hevein fold of the fold family 6, which are small (approx. 40 amino acids) chitin-binding proteins (Costa et al. 2016) with a fold that comprises predominantly coil but does have two small β-sheets and a small region of

helix (Boraston et al. 2004). The fold of CBM50 is largely solvent exposed and has a much greater character of alpha-helix in the native structure— by nature, alpha-helices have more local contacts than those found in β-sheets, which is consistent with what a protein of low ABSCO should look like (small molecular weight and largely dominated by alpha-helical secondary structures). However, much like CtCBM11, CtCBM30, and CtCBM44 (and most CtCBMs discovered to date), CtCBM3 has been grouped in the β-sandwich fold family. We remain perplexed by the tremendous differences in ABSCO between CtCBM families 3, 11, 30, and 44 despite similarities in the protein fold. In future work it might be interesting to test the refolding properties of CBM50 and CBMs from family 3 which have the lowest ABSCO of any CtCBM family.

Thus far, a reliable way to predict temperature tunability remained elusive. The complementation assay we obtained continued to raise an important question: following thermal denaturation, why do CtCBM44 and CtCBM11 refold, but CtCBM30 does not? As shown in Figure 19, CtCBM44 is known to bind a calcium ion in solution and CtCBM11 is known to have two calcium ion binding sites. Surprisingly, despite having the same fold and being classified as Type B CBMs due to substrate binding affinities, CtCBM30 does not have a calcium ion binding site. This is especially important since all samples when prepared and assayed contained 5mM $CaCl_2$, which is strongly believed to play a structural role in Type B CBMs like CtCBM11 and CtCBM44. Careful analysis of these structures in Figure 19 shows that these calcium ion binding sites occur in loops where the oxygens of both main chain and side chain (often small acidic residues) can coordinate the bound ion. Perhaps when denatured, these loop regions form quickly following hydrophobic collapse, which limits the protein entropy and allows calcium ion

binding to shift equilibrium during protein folding far to the folded state. Alternatively, perhaps only heating to 5-10°C above the T_m leaves the protein as a molten globule where loops remain intact or once again form first in the presence of excess calcium ions (as opposed to a complete loss of structure). A similar result was reported in an interesting study by Bhattacharjee and Das (2000), where even at 85-90 °C, beta-lactoglobulin does not completely lose its folded structure. The idea of residual structure remaining at temperatures only 5-10°C higher than the T_m is especially appealing when we consider in experiments where CtCBM11-GFPS11 and CtCBM44-GFPS11 were thermally perturbed to boiling temperatures for five minutes and allowed to cool to refold: we found that the complementation assays with GFPS1-S10 did not result in 100% fluorescence gain when compared to non-heated controls (data not shown). Perhaps the loss of the hydrophobic effect at such high temperatures takes CtCBM11 and CtCBM44 from a molten globule where residual structure remains and calcium ion binding can still occur to a completely unfolded state from which refolding is just not favorable compared to the enthalpy gained from protein-protein interactions of misfolded CtCBMs. Without the ability to bind calcium ions, it makes sense that CtCBM30 may not have similar refolding pathways to CtCBM11, CtCBM44, and other calcium ion binding CBMs that favor the native state.

Alternatively, a comparison of the secondary structure elements of CtCBM11, CtCBM30, and CtCBM44 shows another possibility (Figures 20-22) for why despite having the same fold, CtCBM11 and CtCBM44 can refold while CtCBM30 does not when cooling from a thermally perturbed state. The RaptorX model for CtCBM11 shows that only 3% of the amino acids are found to be disordered in the 3D structure, 0% of the

structure contains an alpha-helix, 47% contains beta-sheets, and 52% of the structure is made up of loops. In CtCBM44 only 2% of the amino acids are found to be disordered in the 3D structure, 0% of the structure contains an alpha-helix, 50% contains beta-sheets, and 49% of the structure is made up of loops. And finally, in CtCBM30 12% of the amino acids are found to be disordered in the 3D structure, 0% of the structure contains an alpha-helix, 36% contains beta-sheets, and 62% of the structure is made up of loops. All CBMs in question are classified as having medium solvent accessibility. Closer analysis reveals that the N and C termini of CtCBM30 lack any secondary structure and are highly disordered. This lack of structure in the RaptorX model is supported by experimental data that is shown in Figure 24: the crystal structure for CtCBM30 (PDB: 2C24) also lacks any detail on the N-terminal and C-terminal end as structure is only present between amino acids 15-185. While it is possible that these regions in the protein were conserved for spatial/steric requirements involved in feeding the substrate into the parent enzyme, in applications of temperature tunability where refolding is important following denaturation, this much greater degree of disorder in CtCBM30 increases entropy of the entire protein chain. We can thus reason that the large protein disorder on the N and C termini of CtCBM30 are detrimental to the refolding pathway as the increased protein entropy of CtCBM30 likely increases the possibility of non-native contacts forming in the refolding pathway. This must result in deep local minima (kinetic traps) of misfolds, or disordered aggregates, that are much more stable than the native state of CtCBM30. Alternatively, these areas could also prevent residual structure remaining when thermally perturbed to 5^{o}C above the established T_m.

The above results strongly suggest that CtCBM11 and CtCBM44 could be used as temperature tunable domains that could be fused to a cellulose degrading CD to form cellulases that can be efficiently recycled for use in multiple rounds of saccharification through the reversible thermal denaturation and refolding of the fused CBM. However, as a prerequisite to functional experiments of that caliber, we first needed to ensure that the cloned CBMs were functional in addition to being properly folded. The function of CtCBM11 and CtCBM44 was confirmed through retardation assays shown in Figure 13. While previous papers have reported that CtCBM11 and CtCBM44 are functional and can bind cellulose-based substrates, it was helpful and encouraging to personally verify these results. Our results are in agreements with reports in the literature and suggest that both CtCBM11 and CtCBM44 are functional and capable of binding a cellulose-based substrate when expressed. These results also suggest that both CtCBM11 and CtCBM44 can refold and regain binding function following thermal denaturation. Notably, heat treated samples were simply boiled for convenience as this assay was done prior to establishing the T_m of each construct. As expected, and previously discussed, the boiling conditions did result in the misfolding (and potential precipitation) of CtCBM11 and CtCBM44 as evident by decreases in band intensity between non-heat-treated samples and boiled samples. This supports the notion that successfully refolding from a thermally perturbed state is contingent upon the entropic contributions of water, which can be greatly diminished at higher (near-boiling) temperatures.

To better test binding of CtCBM11 and CtCBM44 following thermal denaturation, different CtCBM constructs were engineered. In this case, each CBM was fused to the N-terminus of GSF, which serves as a thermostable reporter to help visualize and measure

fluorescence regardless of whether the fused CBM is folded or in a misfolded and non-functional state. This allows us to use fluorescence as a direct measure of how much binding occurs in the presence of a cellulose substrate. Figure 16 and 17 show functional binding assays of both CtCBM11-GSF and CtCBM44-GSF constructs in the presence of Avicel or a cellulose membrane. Figure 16 shows that GSF alone does not interact with Avicel through non-specific interactions as equal amounts of GSF are present in the supernatant of each condition, which is evident by the matching fluorescence signals of both samples incubated with Avicel and "load" samples (negative controls devoid of Avicel substrate). This suggests that changes in fluorescence shown in binding experiments between Avicel and CtCBM11-GSF or CtCBM44-GSF must be due to the direct binding of the CBMs to glycan-chains present in the Avicel substrate. Additionally, GSF samples used in this study were heated to varying temperatures to ensure stability of the fluorophore as reported in the literature. We have found that incubation in temperatures that exceed 86-87°C compromise the fluorophore and leads to noticeable decreases in fluorescence signals (data not shown). As a result, our binding experiments were limited to temperatures no higher than 85°C as decreases in fluorescence cannot be discerned as resulting from binding or destruction of the GSF fluorophore.

In all cases binding assays were done at 50°C and at 10°C above the established T_m where all CtCBMs should be denatured and no longer capable of binding the Avicel substrate. The fluorescence units provided for the load correspond to equal volumes of sample for each temperature condition, but devoid of Avicel substrate and are used as a reference to determine how much of the loaded CtCBM-GSF binds to Avicel at each temperature condition. The observed decrease in fluorescence for both CtCBM11 and

CtCBM44 at 50°C following thermal denaturation and cooling suggests that these CBMs reach a functional state once a temperature below the established T_m is reached. What is encouraging for CtCBM11 is that upon reaching 77°C, which is 10°C above the established T_m, the fluorescence mostly returns within margins of error that make it equivalent to the fluorescence of the load. This result suggests that CtCBM11 is an ideal candidate for temperature-tunability. In contrast, an interesting result is observed for CtCBM44; while in the absence of substrate 3-5°C above the T_m can be enough to denature virtually 100% of the protein population in solution, in the presence of a substrate the T_m of a protein can be much greater than samples devoid of substrate. Even after heating to a temperature of 85°C, the fluorescence signals for CtCBM44-GSF are virtually identical to those at 50°C, which suggests that CtCBM44 interacts with Avicel and is stabilized well past its T_m established in the absence of substrate. Alternatively, it is possible that during centrifugation at 13,000 x *g* for five minutes, the small volumes of solution used for the binding assay may reach temperatures below the T_m of CtCBM44 and therefore could allow for some of the CtCBMs in solution to refold or return to their native and functional state. This could perhaps explain why even CtCBM11 fluorescence signals, on average, did not completely return to the level of signals obtained from the load samples devoid of Avicel substrate. While perhaps heating an additional 5°C could resolve the observed result with CtCBM44, unfortunately, as previously mentioned the poor stability of the fluorophore at higher temperatures that exceed approximately 86-87°C does not allow for this experiment to be done. Alternatively, a shorter processing time post-heating or use of a heated centrifuge might resolve the situation.

To resolve this issue, we took a different approach using a cellulose membrane that could perhaps lead to lower substrate-stabilization by nature of having less surface area available for binding. Here, the fluorescent signal is due to bound protein as opposed to measuring unbound protein in the Avicel binding assay. In this experiment equal fluorescence units of GSF, CtCBM11-GSF, and CtCBM44-GSF were loaded onto a cellulose membrane made up cellulose acetate/cellulose nitrate polymers. Fluorescence that can be visualized on the membranes is assumed to be due to binding and can be represented with a normal fluorescence image and with a more sensitive, but inverted contrast image. We hypothesized that by using cellulose membranes, we could image the membranes to show whether samples of CtCBMs that have been thermally denatured and refolded by cooling are still bound after incubation in buffer at 50°C, at 5°C below the T_m, and at 5-10°C above the established T_m. As expected, virtually no interaction occurs between the GSF and cellulose membrane as incubation at all temperatures and gentle mixing by inversion removed any fluorescence from the membranes. An encouraging result was observed with CtCBM11; once again CtCBM11 showed binding at all temperatures below the established T_m of 67°C but could be almost entirely removed after heating to 10°C higher than the established T_m. This result matches the results from Figure 16 perfectly; i.e., the absence of fluorescence on the membrane must be due to the loss of structure and function of CtCBM11. As expected, a nearly identical result can be observed for CtCBM44. Notably, unlike the Avicel experiment shown in Figure 16 where substrate-stabilization was observed at 85°C and binding was still observed, in this case almost entirely all the fluorescence is removed from the cellulose membrane. This suggests that CtCBM44 denatures and loses its ability to bind a substrate at 85°C, likely

due to a lower degree of substrate stabilization on the cellulase acetate membrane compared to the high surface area of purified crystalline cellulose (Avicel).

The data shown in this study strongly suggest that CtCBM11 and CtCBM44 are excellent candidates to be used as temperature tunable domains in protein engineering endeavors. Broad substrate, Type B CBMs of the β-sandwich fold family with bound metal ions may be excellent CBMs to fuse to CDs often employed in saccharification reactions of lignocellulose substrates. This is important because as it stands, current low oil prices and the low efficiency of second-generation biofuel production strategies have prevented lignocellulosic biorefineries from thriving as an enterprise over fossil fuels. The two largest obstacles that must be overcome are the major costs of biomass pretreatment and the production costs of saccharification enzymes. The temperature tunable system we have implemented in this study removes the need for additional saccharification enzymes to be added to fresh substrate at the beginning of each cycle. After a round of saccharification, any cellulase that is bound to lignin could be released from a trapped substrate and reused simply by allowing cooling from a thermally perturbed state of the fused temperature tunable CBM. This will allow the same batch of engineered cellulases to be used for multiple rounds of saccharification. Future work will focus on the mutagenesis of CtCBM11 and CtCBM44 to lower the T_m compared to the T_m of a fused thermostable cellulolytic CD. Next steps would also include engineering and expression of a novel cellulase composed of a thermostable and high activity CD fused to a temperature tunable CBM. The activity of our temperature tunable cellulases would be first tested against purified cellulose substrate (Avicel) before and after thermal denaturation of the CBM to ensure recovery of binding function. Any adverse effects on

the CBM would become immediately obvious by changes in activity of the temperature tunable cellulase since the CBM is integral to the formation of the E-S complex. Finally, the ultimate test will be to test the activity of a temperature tunable cellulase in a lignocellulosic biorefinery for multiple rounds of saccharification. It will be important to demonstrate that rescuing lignin-trapped saccharification enzymes in spent biomass by thermal induction can lead to higher enzyme activity and sugar production than conventional enzyme recycling methods, which could drop the cost of saccharification and make the lignocellulose biofuel enterprise more competitive with fossil fuel energy in the future.

REFERENCES

Aden A, Foust T. 2009. Technoeconomic analysis of the dilute sulfuric acid and enzymatic hydrolysis process for the conversion of corn stover to ethanol. Cellulose. 16: 535-545. 10.1007/s10570-009-9327-8.

Anfinsen CB. 1973. Principles that govern the folding of protein chains. Science. 181(4096):223–230. doi:10.1126/science.181.4096.223.

Bamdad H, Hawboldt K, MacQuarrie S. 2018. A review on common adsorbents for acid gases removal: Focus on biochar. Renewable and Sustainable Energy Reviews. 81:1705–1720. doi:10.1016/j.rser.2017.05.261.

Bhattacharjee C, Das KP. 2000. Thermal unfolding and refolding of β-lactoglobulin. European Journal of Biochemistry. 267(13):3957–3964. doi:10.1046/j.1432-1327.2000.01409.x.

Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. 2004. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem J. 382(Pt 3):769–781. doi:10.1042/BJ20040892.

Boraston AB, Nurizzo D, Notenboom V, Ducros V, Rose DR, Kilburn DG, Davies GJ. 2002. Differential oligosaccharide recognition by evolutionarily-related β-1,4 and β-1,3 glucan-binding modules. Journal of Molecular Biology. 319(5):1143–1156. doi:10.1016/S0022-2836(02)00374-1.

Brun E, Moriaud F, Gans P, Blackledge MJ, Barras F, Marion D. 1997. Solution structure of the cellulose-binding domain of the endoglucanase Z secreted by *Erwinia chrysanthemi*. Biochemistry. 36(51):16074–16086. doi:10.1021/bi9718494.

Cabantous S, Terwilliger TC, Waldo GS. 2005. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. Nat Biotechnol. 23(1):102–107. doi:10.1038/nbt1044.

Cabantous S, Waldo GS. 2006. In vivo and in vitro protein solubility assays using split GFP. Nat Methods. 3(10):845–854. doi:10.1038/nmeth932.

Chakravorty Ujjayant, Hubert Marie-Hélène, Moreaux Michel, Nøstbakken Linda. 2017. Long-run impact of biofuels on food prices. The Scandinavian Journal of Economics. 119(3):733–767. doi:10.1111/sjoe.12177.

Crowther GJ, He P, Rodenbough PP, Thomas AP, Kovzun KV, Leibly DJ, Bhandari J, Castaneda LJ, Hol WGJ, Gelb MH, et al. 2010. Use of thermal melt curves to assess the quality of enzyme preparations. Anal Biochem. 399(2):268–275. doi:10.1016/j.ab.2009.12.018.

Foumani M, Vuong TV, MacCormick B, Master ER. 2015. Enhanced polysaccharide binding and activity on linear β-glucans through addition of carbohydrate-binding modules to either terminus of a glucooligosaccharide oxidase. PLOS ONE. 10(5):e0125398. doi:10.1371/journal.pone.0125398.

Gaskell A, Crennell S, Taylor G. 1995. The three domains of a bacterial sialidase: a β-propeller, an immunoglobulin module and a galactose-binding jelly-roll. Structure. 3(11):1197–1205. doi:10.1016/S0969-2126(01)00255-6.

Guillén D, Sánchez S, Rodríguez-Sanoja R. 2010. Carbohydrate-binding domains: multiplicity of biological roles. Appl Microbiol Biotechnol. 85(5):1241–1249. doi:10.1007/s00253-009-2331-y.

Hashimoto H. 2006. Recent structural studies of carbohydrate-binding modules. Cell Mol Life Sci. 63(24):2954–2967. doi:10.1007/s00018-006-6195-3.

Henshaw J, Horne-Bitschy A, van Bueren AL, Money VA, Bolam DN, Czjzek M, Ekborg NA, Weiner RM, Hutcheson SW, Davies GJ, et al. 2006. Family 6 carbohydrate binding modules in beta-agarases display exquisite selectivity for the non-reducing termini of agarose chains. J Biol Chem. 281(25):17099–17107. doi:10.1074/jbc.M600702200.

Hirano K, Kurosaki M, Nihei S, Hasegawa H, Shinoda S, Haruki M, Hirano N. 2016. Enzymatic diversity of the *Clostridium thermocellum* cellulosome is crucial for the degradation of crystalline cellulose and plant biomass. Scientific Reports. 6:35709. doi:10.1038/srep35709.

Jørgensen Henning, Pinelo Manuel. 2017. Enzyme recycling in lignocellulosic biorefineries. Biofuels, Bioproducts and Biorefining. 11(1):150–167. doi:10.1002/bbb.1724.

Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J. 2012. Template-based protein structure modeling using the RaptorX web server. Nat Protoc. 7(8):1511–1522. doi:10.1038/nprot.2012.085.

Kont R, Kari J, Borch K, Westh P, Väljamäe P. 2016. Inter-domain synergism is required for efficient feeding of cellulose chain into active site of cellobiohydrolase Cel7A. J Biol Chem. 291(50):26013–26023. doi:10.1074/jbc.M116.756007.

Kraulis J, Clore GM, Nilges M, Jones TA, Pettersson G, Knowles J, Gronenborn AM. 1989. Determination of the three-dimensional solution structure of the C-terminal domain of cellobiohydrolase I from Trichoderma reesei. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. Biochemistry. 28(18):7241–7257.

Larsen J, Haven MØ, Thirup L. 2012. Inbicon makes lignocellulosic ethanol a commercial reality. Biomass and Bioenergy. 46:36–45. doi:10.1016/j.biombioe.2012.03.033.

Mes-Hartree M, Hogan CM, Saddler JN. 1987. Recycle of enzymes and substrate following enzymatic hydrolysis of steam-pretreated aspenwood. Biotechnol Bioeng. 30(4):558–564. doi:10.1002/bit.260300413.

Najmudin S, Guerreiro CIPD, Carvalho AL, Prates JAM, Correia MAS, Alves VD, Ferreira LMA, Romão MJ, Gilbert HJ, Bolam DN, et al. 2006. Xyloglucan is recognized by carbohydrate-binding modules that interact with beta-glucan chains. J Biol Chem. 281(13):8815–8828. doi:10.1074/jbc.M510559200.

Nakamura T, Mine S, Hagihara Y, Ishikawa K, Ikegami T, Uegaki K. 2008. Tertiary structure and carbohydrate recognition by the chitin-binding domain of a hyperthermophilic chitinase from *Pyrococcus furiosus*. Journal of Molecular Biology. 381(3):670–680. doi:10.1016/j.jmb.2008.06.006.

Notenboom V, Boraston Alisdair B, Chiu P, Freelove ACJ, Kilburn DG, Rose DR. 2001. Recognition of cello-oligosaccharides by a family 17 carbohydrate-binding module: an X-ray crystallographic, thermodynamic and mutagenic study11Edited by R. Huber. Journal of Molecular Biology. 314(4):797–806. doi:10.1006/jmbi.2001.5153.

Notenboom V, Boraston Alisdair B., Kilburn DG, Rose DR. 2001. Crystal structures of the family 9 carbohydrate-binding module from *Thermotoga maritima* xylanase 10A in native and ligand-bound forms,. Biochemistry. 40(21):6248–6256. doi:10.1021/bi0101704.

Notenboom V, Boraston AB, Williams SJ, Kilburn DG, Rose DR. 2002. High-resolution crystal structures of the lectin-like xylan binding domain from *Streptomyces lividans* xylanase 10A with bound substrates reveal a novel mode of xylan binding,. Biochemistry. 41(13):4246–4254. doi:10.1021/bi015865j.

Ooshima H, Burns DS, Converse AO. 1990. Adsorption of cellulase from *Trichoderma reesei* on cellulose and lignacious residue in wood pretreated by dilute sulfuric acid with explosive decompression. Biotechnol Bioeng. 36(5):446–452. doi:10.1002/bit.260360503.

Otter DE, Munro PA, Scott GK, Geddes R. 1984. Elution of *Trichoderma reesei* cellulase from cellulose by pH adjustment with sodium hydroxide. Biotechnol Lett. 6(6):369–374. doi:10.1007/BF00138007.

Pédelacq J-D, Cabantous S, Tran T, Terwilliger TC, Waldo GS. 2006. Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol. 24(1):79–88. doi:10.1038/nbt1172.

Pires VMR, Henshaw JL, Prates JAM, Bolam DN, Ferreira LMA, Fontes CMGA, Henrissat B, Planas A, Gilbert HJ, Czjzek M. 2004. The crystal structure of the family 6 carbohydrate binding module from *Cellvibrio mixtus* endoglucanase 5A in complex with oligosaccharides reveals two distinct binding sites with different ligand specificities. Journal of Biological Chemistry. 279(20):21560–21568. doi:10.1074/jbc.M401599200.

Plaxco KW, Simons KT, Baker D. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. Journal of Molecular Biology. 277(4):985–994. doi:10.1006/jmbi.1998.1645.

Poudyal Roshan Sharma, Tiwari Indira, Najafpour Mohammad Mahdi, Los Dmitry A., Carpentier Robert, Shen Jian-Ren, Allakhverdiev Suleyman I. 2016 Apr 29. Current insights to enhance hydrogen production by photosynthetic organisms. Hydrogen Science and Engineering: Materials, Processes, Systems and Technology. doi:10.1002/9783527674268.ch20. [accessed 2018 Mar 27]. https://onlinelibrary.wiley.com/doi/10.1002/9783527674268.ch20.

Raghothama S, Simpson PJ, Szabo L, Nagy T, Gilbert HJ, Williamson MP. 2000. Solution structure of the CBM10 cellulose binding module from Pseudomonas xylanase A. Biochemistry. doi:10.1021/bi992163+. [accessed 2019 Apr 18]. https://eprint.ncl.ac.uk/63675.

Rahman FA, Aziz MMA, Saidur R, Bakar WAWA, Hainin MR, Putrajaya R, Hassan NA. 2017. Pollution to solution: capture and sequestration of carbon dioxide (CO_2) and its utilization as a renewable energy source for a sustainable future. Renewable and Sustainable Energy Reviews. 71:112–126. doi:10.1016/j.rser.2017.01.011.

Rodionova MV, Poudyal RS, Tiwari I, Voloshin RA, Zharmukhamedov SK, Nam HG, Zayadan BK, Bruce BD, Hou HJM, Allakhverdiev SI. 2017. Biofuel production: challenges and opportunities. International Journal of Hydrogen Energy. 42(12):8450–8461. doi:10.1016/j.ijhydene.2016.11.125.

Rodrigues AC, Leitão AF, Moreira S, Felby C, Gama M. 2012. Recycling of cellulases in lignocellulosic hydrolysates using alkaline elution. Bioresour Technol. 110:526–533. doi:10.1016/j.biortech.2012.01.140.

Saul FA, Rovira P, Boulot G, Van Damme EJ, Peumans WJ, Truffa-Bachi P, Bentley GA. 2000. Crystal structure of Urtica dioica agglutinin, a superantigen presented by MHC molecules of class I and class II. Structure. 8(6):593–603. doi:10.1016/S0969-2126(00)00142-8.

Strobel KL, Pfeiffer KA, Blanch HW, Clark DS. 2016. Engineering Cel7A carbohydrate binding module and linker for reduced lignin inhibition. Biotechnol Bioeng. 113(6):1369–1374. doi:10.1002/bit.25889.

Suetake T, Tsuda S, Kawabata S, Miura K, Iwanaga S, Hikichi K, Nitta K, Kawano K. 2000. Chitin-binding proteins in invertebrates and plants comprise a common chitin-binding structural motif. Journal of Biological Chemistry. 275(24):17929–17932. doi:10.1074/jbc.C000184200.

Taylor CB, Talib MF, McCabe C, Bu L, Adney WS, Himmel ME, Crowley MF, Beckham GT. 2012. Computational investigation of glycosylation effects on a family 1 carbohydrate-binding module. J Biol Chem. 287(5):3147–3155. doi:10.1074/jbc.M111.270389.

Taylor T, Denson J-P, Esposito D. 2017. Optimizing expression and solubility of proteins in *E. coli* using modified media and induction parameters. Methods Mol Biol. 1586:65–82. doi:10.1007/978-1-4939-6887-9_5.

Tenenbaum DJ. 2008. Food vs. Fuel: Diversion of crops could cause more hunger. Environ Health Perspect. 116(6):A254–A257.

Tormo J, Lamed R, Chirino AJ, Morag E, Bayer EA, Shoham Y, Steitz TA. 1996. Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. EMBO J. 15(21):5739–5751.

Tu M, Zhang X, Paice M, MacFarlane P, Saddler JN. 2009. The potential of enzyme recycling during the hydrolysis of a mixed softwood feedstock. Bioresour Technol. 100(24):6407–6415. doi:10.1016/j.biortech.2009.06.108.

Viegas A, Sardinha J, Freire F, Duarte DF, Carvalho AL, Fontes CMGA, Romão MJ, Macedo AL, Cabrita EJ. 2013. Solution structure, dynamics and binding studies of a family 11 carbohydrate-binding module from *Clostridium thermocellum* (CtCBM11). Biochem J. 451(2):289–300. doi:10.1042/BJ20120627.

Viegas AJM. 2012. Molecular determinants of ligand specificity in carbohydrate-binding modules: an NMR and X-ray crystallography integrated study. The FEBS Journal. doi.org/10.1111/j.1742-4658.2008.06401.x.

Walker JA, Takasuka TE, Deng K, Bianchetti CM, Udell HS, Prom BM, Kim H, Adams PD, Northen TR, Fox BG. 2015. Multifunctional cellulase catalysis targeted by fusion to different

carbohydrate-binding modules. Biotechnology for Biofuels. 8:220. doi:10.1186/s13068-015-0402-0.

Wang S, Ma J, Peng J, Xu J. 2013. Protein structure alignment beyond spatial proximity. Scientific Reports. 3:1448. doi:10.1038/srep01448.

Xu GY, Ong E, Gilkes NR, Kilburn DG, Muhandiram DR, Harris-Brandts M, Carver JP, Kay LE, Harvey TS. 1995. Solution structure of a cellulose-binding domain from *Cellulomonas fimi* by nuclear magnetic resonance spectroscopy. Biochemistry. 34(21):6993–7009.