Supporting Information

Glycosylation is vital for industrial performance of hyper-active cellulases

Daehwan Chung, Nicholas S. Sarai, Brandon C. Knott, Neal Hengge, Jordan F. Russell, John M. Yarbrough, Roman Brunecky, Jenna Young, Nitin Supekar, Todd VanderWall, Deanne W. Sammond, Michael F. Crowley, Christine M. Szymanski, Lance Wells, Parastoo Azadi, Janet Westpheling, Michael E. Himmel, and Yannick J. Bomble

Biosciences Center, National Renewable Energy Laboratory, 15013 Denver West Parkway,
Golden CO, 80401, USA

Department of Genetics, University of Georgia, Athens GA, 30602, USA

Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602

Supplementary Information

Additional Materials and Methods

Strains, Media, and Growth Conditions

Caldicellulosiruptor bescii and *Escherichia coli* strains used in this study are listed in Table S1. All *C. bescii* strains were grown anaerobically at 65°C on solid or in liquid low osmolarity defined (LOD) medium (*1*), as previously described, with .5 % w/v D(+)-Cellobiose (Acros Organics, NJ, U.S.A., Code: 108465000, Lot: A0384025) as the sole carbon source, final pH 6.8, for routine growth and transformation experiments (*2*). For growth of uracil auxotrophic strains based off of JWCB029 (*ΔpyrFA ldh::*IS*Cbe4 Δcbe1 ΔcelA*), the LOD medium contained 40 µM uracil. This concentration of uracil does not support growth of *C. bescii* as sole carbon source. *E. coli* strain DH5α was used for plasmid DNA construction and preparation using standard techniques as described (*3*). *E. coli* cells were cultured in LB broth supplemented with apramycin (50 µg/mL) and plasmid DNA was isolated using a Qiagen Mini-prep kit (Qiagen, Hilden, Germany). Chromosomal DNA from *Caldicellulosiruptor* strains was extracted using the Quick-gDNA™ MiniPrep (Zymo Research, Irvine, CA, U.S.A.) as previously described (*4*).

Construction and Transformation of CelA and CelA Derivative Expression Vectors

Plasmids in this study were generated using Q5 High-Fidelity DNA polymerase (New England BioLabs, Ipswich, MA, U.S.A.), restriction enzymes (New England BioLabs, Ipswich, MA, U.S.A.), and Fast-link^{TM} DNA Ligase (Epicentre Technologies, Madison, WI, U.S.A.) according to the manufacturer's instructions. For the construction of pJYW008 (Figure S1, Table S1), a 10.64 kb DNA fragment was synthesized with the reverse primer JY017 and forward primer JY018 using pDCW173 (*4*) as a template. After amplification, the PCR product was ligated via blunt-end ligation. The 9.96 kb (amplified using DCB151 and DCB152) and 10.62 kb (amplified using DCB152 and DCB153) PCR amplified DNA fragments, using pDCW173 as a template, were synthesized for the construction of pDCYB037 and pDCYB038 (Table S1), respectively. These two linear DNA fragments were digested with SphI and ligated to construct pDCYB037 and pDCYB038, respectively. The *C. bescii* CelA gene sequence (Cbes_1867; GenBank accession number Z86105) was codon optimized for expression in *E. coli* and cloned into a pET28b(+) vector using NcoI and XhoI sites (GenScript, Piscataway, NJ, U.S.A.). The sequence for a 6x histidine tag was placed at the *C*-terminus to facilitate protein purification. It was referred as pDCYB017 (Table S1). For the construction of pDCYB018 (Table S1), a 7.81 kb DNA fragment was synthesized with the reverse primer DCB068 and forward primer DCB069 using pDCYB017 as a template. After amplification, the PCR product was ligated via blunt-end ligation. The 7.78 kb (amplified using DCB155 and DCB157) and 7.11 kb (amplified using DCB157 and DCB256) PCR amplified DNA fragments, using pDCYB017 as a template, were synthesized for the construction of pDCYB075 and pDCYB076 (Table S1), respectively. These two linear DNA fragments were digested with SphI and ligated to construct pDCYB075 and pDCYB076, respectively.

To construct JWCB056, YBCB008, and YBCB009, plasmids pJYW024, pDCYB037, and pDCYB038 were electro-transformed into JWCB029 (*ΔpyrFAΔldh::ISCbe4 Δcbe1ΔcelA*) cells as previously described (*2*). Cultures, electro-pulsed with plasmid DNA (0.5 ~ 1.0 µg), were recovered in low osmolarity complex (LOC) growth medium (*1*) at 75°C. Recovery cultures were transferred to liquid LOD medium without uracil to allow selection of uracil prototrophs. Cultures were plated on solid LOD media to obtain isolated colonies, and total DNA was isolated from transformants. PCR amplification using primers (DCB008 and DCB010) outside the gene cassette on the plasmid was used to confirm the presence of the plasmid with the gene of interest intact. Primers used for plasmid constructions and confirmation are listed in Table S1. *E. coli* strain DH5a cells were transformed by electroporation in a 2-mm-gap cuvette at 2.5 V and transformants were selected for apramycin resistance. The sequences of all plasmids were confirmed by automatic sequencing (Genewiz, NJ, U.S.A.). All plasmids are available on request.

Protein Expression in *E. coli*

E. Coli BL21(DE3) was used for all experiments. Recombinant strains were grown in LB broth supplemented with apramycin (50 µg/mL). Cultures were induced at 15°C with 0.25 mM IPTG when OD_{600}=0.4. Cultures were centrifuged at 5,000xg for 10 min when OD_{600}≥1.2. To recover intracellular protein, centrifuged cells were enzymatically lysed in Buffer A (50 mM Tris-HCl, 40mM NaCl, 10mM Imidazole, pH 8.0) supplemented with lysozyme, protease inhibitors, and nuclease (Pierce, Waltham, MA, U.S.A.) at 4°C. Enzymatically lysed cells were subjected to 1 minute of sonication in a water bath at 10 second intervals punctuated by 30 s on ice. Lysed cells were centrifuged at 10,000xg for 30 min. The lysate was recovered.

Fermentation of *C. bescii*

Each relevant strain of *C. bescii* was inoculated from frozen stocks into serum vials containing 20 mL of low osmolarity defined growth media (LOD) (*1*). Cultures were grown at 60°C for approximately 24 h. A 1% subculture into 200 mL of LOD from the 20 mL culture was then carried out. These cultures were grown between 18-24 h with occasional agitation if the cell growth was slow. All 200 mL of seed culture was then inoculated into 10-L fermentation vessels containing LOD media with 0.5% (w/v) cellobiose. Cultures were agitated at 50 RPM and sparged continuously with N_2 at a flow rate of 0.5 liters per min while maintaining a pH of 6.8. A constant temperature of 65°C was maintained.

The fermentation broth was separated from the cells by passing it through a polycap glass fiber filter (GE Healthcare Bio-Sciences, Marlborough, MA) before being concentrated using hollow fiber concentrators containing a 10 kDa cutoff membrane (GE Healthcare, Piscataway, NJ, U.S.A.). The concentrator was also used to buffer exchange the broth into Buffer A.

Molecular simulations (Additional information)

Each system was solvated in a box of explicit water molecules. For linkers in solution, approximate dimensions of the simulation boxes were 100 x 100 x 100 $Å^3$; the total system size was approximately 100,000 atoms. For the three domains on cellulose, approximate dimensions of the simulation boxes were 230 x 125 x 90 $Å^3$; the total system size was approximately 260,000 atoms.

For all systems, the CHARMM force field with CMAP was used to describe the proteins,(*46*) the carbohydrates were described with the C35 force field,(*47, 48*) and water was described with the TIP3P model.(*49*) All MD simulations utilized explicit solvent. The systems were built and minimized in CHARMM,(*50*) performing a stepwise protocol of minimization in which restraints are gradually released on protein, glycans, and cellulose surface. For the surface systems, all minimizations and simulations were performed with a harmonic restraint (force constant of 5 kcal•mole^{-1}•Å^{-2}) on the glucopyranose ring atoms of the bottom cellulose layer as well as the first and last glucose ring of each cellulose chain (i.e. glucose residues 1 and 40). Following minimization, the system was heated from 100 K to 348 K over the course of 1 ns in the NVT ensemble. Systems were then density equilibrated for 1 ns in the NPT ensemble at a constant pressure of 1 atmosphere and constant temperature of 348 K (Nosé-Hoover barostat and thermostat); subsequent production runs were performed with constant volume and temperature (348 K) in NAMD.(*51*) All simulations were performed at a temperature of 348 K. This is higher than molecular simulations of enzymes are typically performed, but we target this temperature because this is the temperature that the assays for binding and hydrolytic activity of thermophilic CelA were performed. The SHAKE algorithm was utilized to fix all bonded hydrogen distances.(*52*) The timestep was 2 fs. Nonbonded cutoff distance of 10 Å was utilized, with a switching distance of 9 Å, and a nonbonded pair list distance of 13 Å. The Particle Mesh Ewald method was used to describe the long-range electrostatics(*53*) with a sixth order b-spline, a Gaussian distribution with a width of 0.312 Å, and 1 Å grid spacing. The velocity Verlet multiple timestepping integration scheme was used evaluating the full nonbonded interactions every timestep, with full electrostatics interactions every 3 timesteps, and 6 timesteps between atom reassignments.

Table S1: ***Caldicellulosiruptor bescii*** **strains and plasmids used in this study**

Strains	**Description**	**Source**
JWCB029	*ΔpyrFAΔldh::ISCbe4 Δcbe1 ΔcelA* (ura$^-$/5-FOA^R)	(*4*)
JWCB047	JWCB029 harboring pDCYB173	(*4*)
JWCB056	JWCB029 harboring pJYW008	This study
YBCB008	JWCB029 harboring pDCYB037	This study
YBCB009	JWCB029 harboring pDCYB038	This study
Plasmids		
pDCW173	*C. bescii* expression vector for full-length CelA (Apramycin^R)	(*4*)
pJYW008	*C. bescii* expression vector for C-terminus of CelA (CBM3-GH48) (Apramycin^R)	This study
pDCYB037	*C. bescii* expression vector for N-terminus of CelA (GH9-CBM3) (Apramycin^R)	This study
pDCYB038	*C. bescii* expression vector for N-terminus of CelA (GH9-CBM3-CBM3) (Apramycin^R)	This study
pDCYB017	*E. coli* expression vector for full-length CelA (Apramycin^R)	This study
pDCYB018	*E. coli* expression vector for C-terminus of CelA (CBM3-GH48) (Apramycin^R)	This study
pDCYB075	*E. coli* expression vector for N-terminus of CelA (GH9-CBM3-CBM3) (Apramycin^R)	This study
pDCYB076	*E. coli* expression vector for N-terminus of CelA (GH9-CBM3) (Apramycin^R)	This study

Table S2: List of primers used in this study. The underlined sequences indicate the recognition sites of the corresponding restriction enzymes.

Name	Sequence (5’ → 3’)	Restriction enzyme	Description

DCB008	AGAGTAGAGCGTGATGACATAGA	-	To confirm transformants
DCB010	ATCATCCCCTTTTGCTGATGGA	-	To confirm transformants
JY017	AAACGAACCAGCCCTAACCTCTTGC	-	To construct pJYW008
JY018	GTA GCAGGCGGGCAGATAAAG	-	To construct pJYW008
DCB151	GCATGCGAAAACTTGTATTTCCAGGGCCAT	SphI	To construct pDCYB037
DCB152	GCATGCTGTCGGTGTTGCTCCAGAAG	SphI	To construct pDCYB037/038
DCB153	GCATGCTGTTGGTGTCGCTCCACTC	SphI	To construct pDCYB038
DCB068	GAAGGAGCCCATGGTATATCTC		To construct pDCYB018
DCB069	GTTGCGGGTGGCCAAATCAAAG		To construct pDCYB018
DCB155	GGATCCACCGCTCGGTTCCTGGCC	BamHI	To construct pDCYB075
DCB157	GGATCCGAAAACTTGTATTTCCAGGGCCTCG	BamHI	To construct pDCYB075/076
DCB206	GGATCCACCGCTGGTACCCGGTTCTTC	BamHI	To construct pDCYB076

Table S3: Actual protein concentrations with percent difference between the control and the unbound fraction found in Figure 6.

	CelA expressed in *C. bescii*	CelA expressed in *E. coli*
Control	0.123 mg/ml	0.130 mg/ml
Unbound to Avicel	0.033 mg/ml	0.013 mg/ml
% Protein unbound to Avicel	26%	10%
Control	0.103 mg/ml	0.142 mg/ml
Unbound to Lignin	0.089 mg/ml	0.065 mg/ml
% protein unbound to lignin	86%	45%

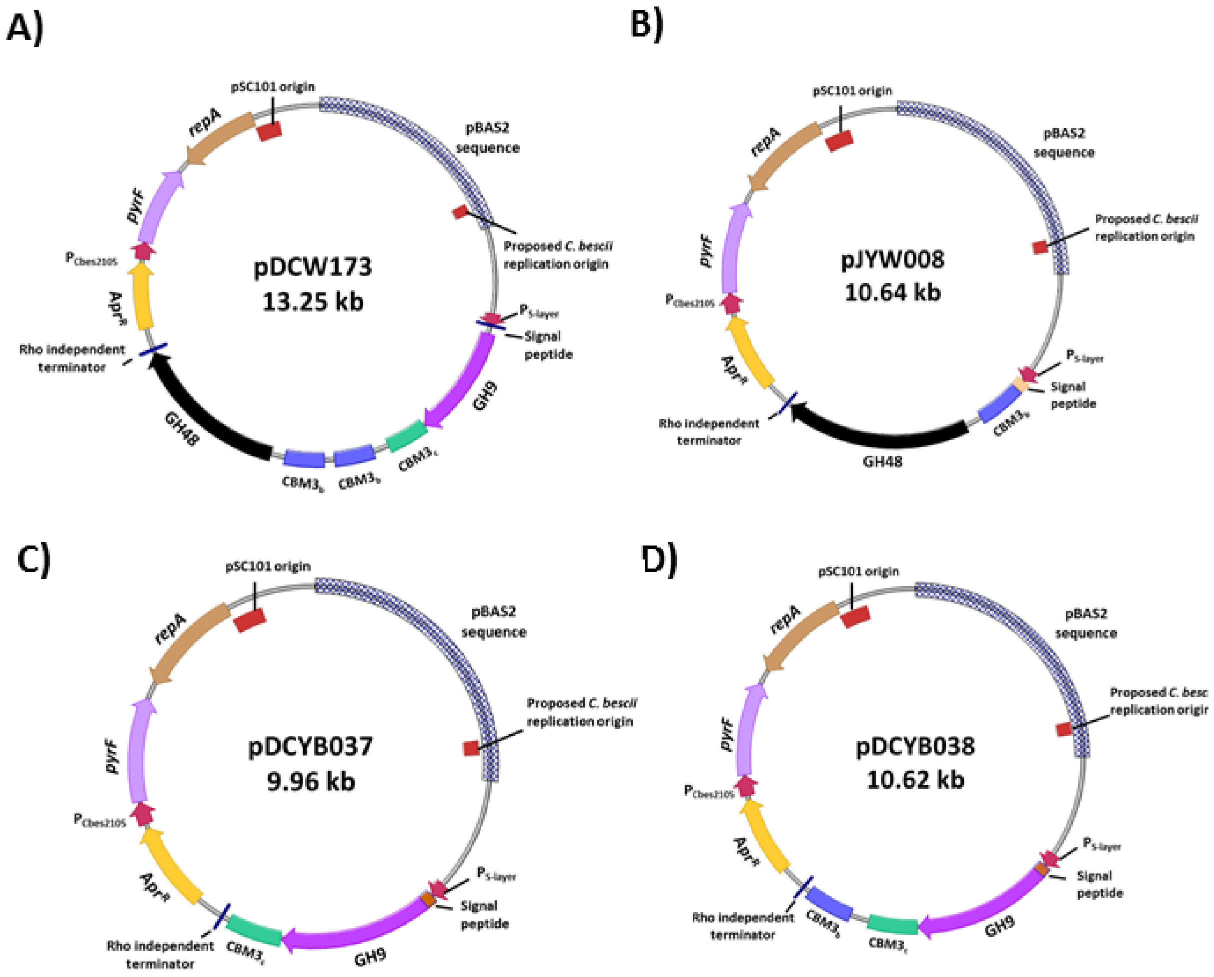


Figure S1: **Plasmid maps of *Caldicellulosiruptor bescii* expression vectors for full-length (FL; GH9-CBM3$_c$-CBM3$_b$-CBM3$_b$-GH48) (A), C-terminal part of CelA (CT; CBM3$_b$-GH48) (B), N-terminal part of CelA (NT1; GH9-CBM3$_c$-CBM3$_b$) (C), and N-terminal part of CelA (NT2; GH9-CBM3$_c$) (D) expression.** Various CelA derivatives were expressed under the control of the regulatory region of the *C. bescii* S-layer protein. The expression vectors contain a signal peptide sequence derived from CelA, a C-terminal 6X His-tag, a Rho independent terminator, the *pyrF* (from *C. thermocellum*) cassette for selection, and pBAS2 sequences for replication in *C. bescii.* The apramycin resistant gene cassette (*Apr^R*), pSC101 low copy replication origin in *E. coli*, and *repA*, a plasmid-encoded gene required for pSC101 replication are indicated.

A) Full-length (FL; GH9-CBM3_c-CBM3_b-CBM3_b-GH48)

MKRYRRIIAMVVTFIFILGVVYGVKPWQEVRAGSFNYGEALQKAIMFYEFQMSGKLPNWVRNNWRGDSALKDGQDNGLDLTGGWFDAGDHVKFNLPMSYTGTMLSWAVYEYK
DAFVKSGQLEHILNQIEWVNDYFVKCHPSKYVYYYQVGDGSKDHAWWGPAEVMQMERPSFKVTQSSPGSTVVAETAASLAAASIVLKDRNPTKAATYLQHAKELYEFAEVTKSDA
GYTAANGYYNSWSGFYDELSWAAVWLYLATNDSTYLTKAESYVQNWPKISGSNTIDYKWAHCWDDVHNGAALLLAKITGKDIYKQIIESHLDYWTTGYNGERIKYTPKGLAWLDQ
WGSLRYATTTAFLAFVYSDWVGCPSTKKEIYRKFGESQIDYALGSAGRSFVVGFGTNPPKRPHHRTAHSSWADSQSIPSYHRHTLYGALVGGPGSDDSYTDDISNYVNNEVACDYN
AGFVGALAKMYQLYGGNPIPDFKAIETPTNDEFFVEAGINASGTNFIEIKAIVNNQSGWPARATDKLKFRYFVDLSELIKAGYSPNQLTLSTNYNQGAKVSGPYVWDASKNIYYILVDF
TGTLIYPGGQDKYKKEVQFRIAAPQNVQWDNSNDYSFQDIKGVSSGSVVKTKYIPLYDGDVKVWGEEPGTSGATPTPTATATPTPTPTVTPTPTPTPTSTATPTPTPTPTVTPTPTPT
PTATPTATPTPTSTPSSTPVAGGQIKVLYANKETNSTTNTIRPWLKVVNTGSSSIDLSRVTIRYWYTVDGDKAQSAISDWAQIGASNVTFKFVKLSSSVSGADYYLEIGFKSGAGQLQA
GKDTGEIQIRFNKSDWSNYNQGNDWSWMQSMTNYGENVKVTAYIDGVLVWGQEPSGATPTPTATPAPTVTPTPTPTPTSTPTATPTATPTPTPTPSSTPVAGGQIKVLYANKETN
STTNTIRPWLKVVNTGSSSIDLSRVTIRYWYTVDGDKAQSAISDWAQIGASNVTFKFVKLSSSVSGADYYLEIGFKSGAGQLQAGKDTGEIQIRFNKSDWSNYNQGNDWSWMQSM
TNYGENVKVTAYIDGVLVWGQEPSGATPTPTATPAPTVTPTPTPAPTPTPTPTPTATPTPTPTPTPTATPTVTATPTPTPSSTPSVLGEYGQRFMWLWNKIHDPANGYFNQDGIPYH
SVETLICEAPDYGHLTTSEAFSYYVWLEAVYGKLTGDWSKFKTAWDTLEKYMIPSAEDQPMRSYDPNKPATYAGEWETPDKYPSPLEFNVPVGKDPLHNELVSTYGSTLMYGMHW
LMDVDNWYGYGKRGDGVSRASFINTFQRGPEESVWETVPHPSWEEFKWGGPNGFLDLFIKDQNYSKQWRYTDAPDADARAIQATYWAKVWAKEQGKFNEISSYVAKAAKMG
DYLRYAMFDKYFKPLGCQDKNAAGGTGYDSAHYLLSWYYAWGGALDGAWSWKIGSSHVHFGYQNPMAAWALANDSDMKPKSPNGASDWAKSLKRQIEFYRWLQSAEGAIA
GGATNSWNGRYEKYPAGTATFYGMAYEPNPVYHDPGSNTWFGFQAWSMQRVAEYYYVTGDKDAGALLEKWVSWVKSVVKLNSDGTFAIPSTLDWSGQPDTWNGAYTGNSN
LHVKVVDYGTDLGITASLANALLYYSAGTKKYGVFDEGAKNLAKELLDRMWKLYRDEKGLSAPEKRADYKRFFEQEVYIPAGWIGKMPNGDVIKSGVKFIDIRSKYKQDPDWPKLEA
AYKSGQAPEFRYHRFWAQCDIAIANATYEILFGNQ

B) C-terminal truncation mutant (CT; CBM3_b-GH48)

MKRYRRIIAMVVTFIFILGVVYGVKPWQEVRAGSFVAGGQIKVLYANKETNSTTNTIRPWLKVVNTGSSSIDLSRVTIRYWYTVDGDKAQSAISDWAQIGASNVTFKFVKLSSSVSGA
DYYLEIGFKSGAGQLQAGKDTGEIQIRFNKSDWSNYNQGNDWSWMQSMTNYGENVKVTAYIDGVLVWGQEPSGATPTPTATPAPTVTPTPTPAPTPTPTPTPTATPTPTPTPTPT
ATPTVTATPTPTPSSTPSVLGEYGQRFMWLWNKIHDPANGYFNQDGIPYHSVETLICEAPDYGHLTTSEAFSYYVWLEAVYGKLTGDWSKFKTAWDTLEKYMIPSAEDQPMRSYDP
NKPATYAGEWETPDKYPSPLEFNVPVGKDPLHNELVSTYGSTLMYGMHWLMDVDNWYGYGKRGDGVSRASFINTFQRGPEESVWETVPHPSWEEFKWGGPNGFLDLFIKDQNY
SKQWRYTDAPDADARAIQATYWAKVWAKEQGKFNEISSYVAKAAKMGDYLRYAMFDKYFKPLGCQDKNAAGGTGYDSAHYLLSWYYAWGGALDGAWSWKIGSSHVHFGYQ
NPMAAWALANDSDMKPKSPNGASDWAKSLKRQIEFYRWLQSAEGAIAGGATNSWNGRYEKYPAGTATFYGMAYEPNPVYHDPGSNTWFGFQAWSMQRVAEYYYVTGDKDA
GALLEKWVSWVKSVVKLNSDGTFAIPSTLDWSGQPDTWNGAYTGNSNLHVKVVDYGTDLGITASLANALLYYSAGTKKYGVFDEGAKNLAKELLDRMWKLYRDEKGLSAPEKRAD
YKRFFEQEVYIPAGWIGKMPNGDVIKSGVKFIDIRSKYKQDPDWPKLEAAYKSGQAPEFRYHRFWAQCDIAIANATYEILFGNQ

C) N-terminal truncation mutant (NT1; GH9-CBM3_c-CBM3_b)

MKRYRRIIAMVVTFIFILGVVYGVKPWQEVRAGSFNYGEALQKAIMFYEFQMSGKLPNWVRNNWRGDSALKDGQDNGLDLTGGWFDAGDHVKFNLPMSYTGTMLSWAVYEYK
DAFVKSGQLEHILNQIEWVNDYFVKCHPSKYVYYYQVGDGSKDHAWWGPAEVMQMERPSFKVTQSSPGSTVVAETAASLAAASIVLKDRNPTKAATYLQHAKELYEFAEVTKSDA
GYTAANGYYNSWSGFYDELSWAAVWLYLATNDSTYLTKAESYVQNWPKISGSNTIDYKWAHCWDDVHNGAALLLAKITGKDIYKQIIESHLDYWTTGYNGERIKYTPKGLAWLDQ
WGSLRYATTTAFLAFVYSDWVGCPSTKKEIYRKFGESQIDYALGSAGRSFVVGFGTNPPKRPHHRTAHSSWADSQSIPSYHRHTLYGALVGGPGSDDSYTDDISNYVNNEVACDYNA
GFVGALAKMYQLYGGNPIPDFKAIETPTNDEFFVEAGINASGTNFIEIKAIVNNQSGWPARATDKLKFRYFVDLSELIKAGYSPNQLTLSTNYNQGAKVSGPYVWDASKNIYYILVDFT
GTLIYPGGQDKYKKEVQFRIAAPQNVQWDNSNDYSFQDIKGVSSGSVVKTKYIPLYDGDVKVWGEEPGTSGATPTPTATATPTPTPTVTPTPTPTPTSTATPTPTPTPTVTPTPTPTPT
ATPTATPTPTSTPSSTPVAGGQIKVLYANKETNSTTNTIRPWLKVVNTGSSSIDLSRVTIRYWYTVDGDKAQSAISDWAQIGASNVTFKFVKLSSSVSGADYYLEIGFKSGAGQLQAGK
DTGEIQIRFNKSDWSNYNQGNDWSWMQSMTNYGENVKVTAYIDGVLVWGQEPSGA

D) N-terminal truncation mutant (NT2; GH9-CBM3_c)

MKRYRRIIAMVVTFIFILGVVYGVKPWQEVRAGSFNYGEALQKAIMFYEFQMSGKLPNWVRNNWRGDSALKDGQDNGLDLTGGWFDAGDHVKFNLPMSYTGTMLSWAVYEYK
DAFVKSGQLEHILNQIEWVNDYFVKCHPSKYVYYYQVGDGSKDHAWWGPAEVMQMERPSFKVTQSSPGSTVVAETAASLAAASIVLKDRNPTKAATYLQHAKELYEFAEVTKSDA
GYTAANGYYNSWSGFYDELSWAAVWLYLATNDSTYLTKAESYVQNWPKISGSNTIDYKWAHCWDDVHNGAALLLAKITGKDIYKQIIESHLDYWTTGYNGERIKYTPKGLAWLDQ
WGSLRYATTTAFLAFVYSDWVGCPSTKKEIYRKFGESQIDYALGSAGRSFVVGFGTNPPKRPHHRTAHSSWADSQSIPSYHRHTLYGALVGGPGSDDSYTDDISNYVNNEVACDYNA
GFVGALAKMYQLYGGNPIPDFKAIETPTNDEFFVEAGINASGTNFIEIKAIVNNQSGWPARATDKLKFRYFVDLSELIKAGYSPNQLTLSTNYNQGAKVSGPYVWDASKNIYYILVDFT
GTLIYPGGQDKYKKEVQFRIAAPQNVQWDNSNDYSFQDIKGVSSGSVVKTKYIPLYDGDVKVWGEEPGTSGA

Figure S2: Amino acid sequences of CelA and its truncation mutants; (A) Full-length (FL), (B) C-terminal part of CelA (CT), (C) N-terminal part of CelA (NT1), and (D) N-terminal part of CelA (NT2). The color-coded letters denote domain organization; letters in brown (CelA signal peptide), purple (GH9 domain), red (linker sequences), green (CBM3c domain), blue (CBM3b domain), black (GH48 domain).

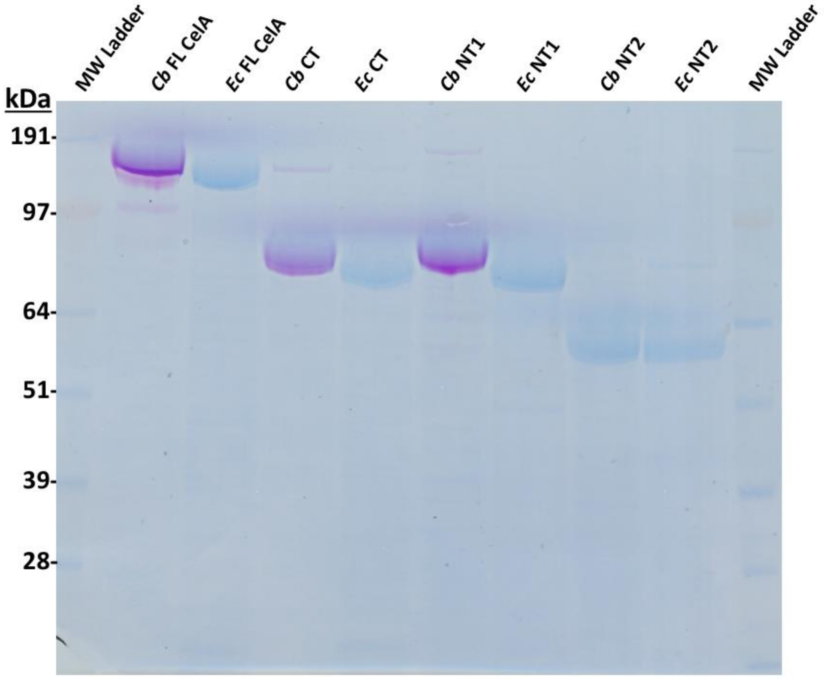


Figure S3: SDS-PAGE and glycoprotein staining demonstrates heavy glycosylation of the linker regions within *C. bescii* CelA. Following SDS-PAGE, a periodic acid Schiff glycoprotein stain was applied that preferentially stained the glycosylated proteins magenta. Following glycoprotein staining, a standard colloidal blue stain revealed non-glycosylated proteins. *C. bescii* expressed CelA, CT, and NT1 are all glycosylated as seen by the magenta color and higher MW compared to *E. coli* expressed proteins. *C. bescii* and *E. coli* expressed NT2 ran at the same electrophoretic mobility and neither was specifically stained (See Figure 2 in the main text).

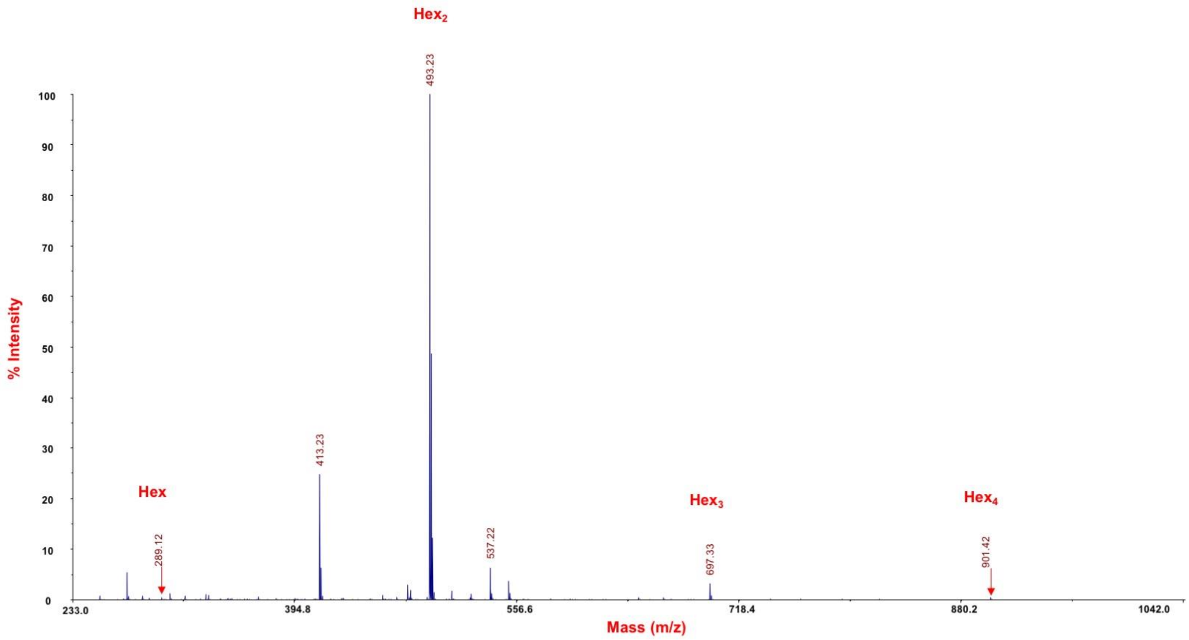


Figure S4: MALDI-TOF-MS spectrum of glycomics analysis of CelA (GH9-CBM3$_c$-CBM3$_b$-CBM3$_b$-GH48) sample by ß-elimination. The presence of Hex$_1$, Hex$_2$, Hex$_{3,\ \text{and}}$ Hex$_4$ is shown.

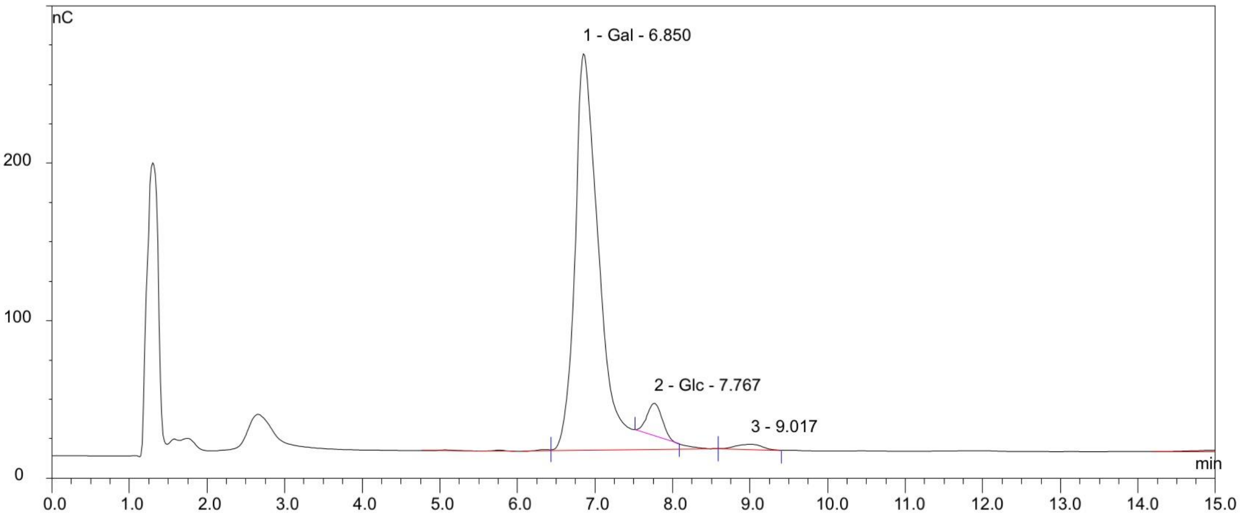


Figure S5: O-linked glycans in CelA (GH9-CBM3$_c$-CBM3$_b$-CBM3$_b$-GH48) HPAEC chromatogram of monosaccharides from CelA (GH9-CBM3$_c$-CBM3$_b$-CBM3$_b$-GH48) O-linked glycans separated by CarboPacPA 20 column shows presence of, galactose (Gal), glucose (Glc), and mannose (Man). Although Glc was detected, we could not determine if this residue is part of the O-glycans structures or simply a contaminant as often as the case.

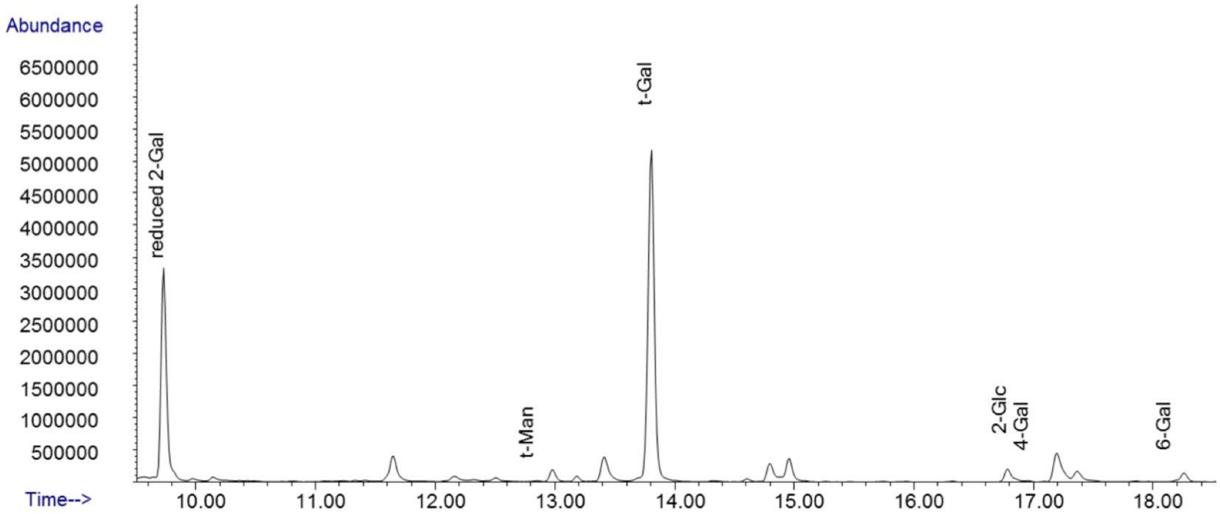


Figure S6: Linkage analysis by GC-MS of partially methylated alditol acetates of O-glycans from CelA (GH9-CBM3$_c$-CBM3$_b$-CBM3$_b$-GH48)

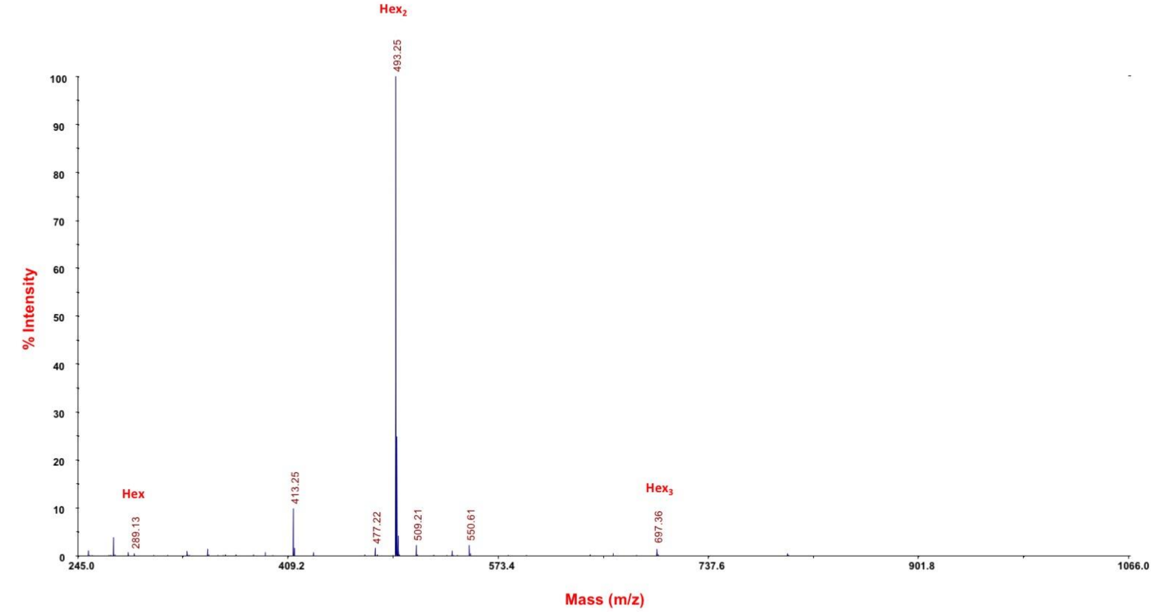


Figure S7: MALDI-TOF-MS spectrum of glycomics analysis of NT1 (GH9-$CBM3_c$-$CBM3_b$) sample by ß-elimination. The presence of Hex_1, Hex_2, and Hex_3 is shown.

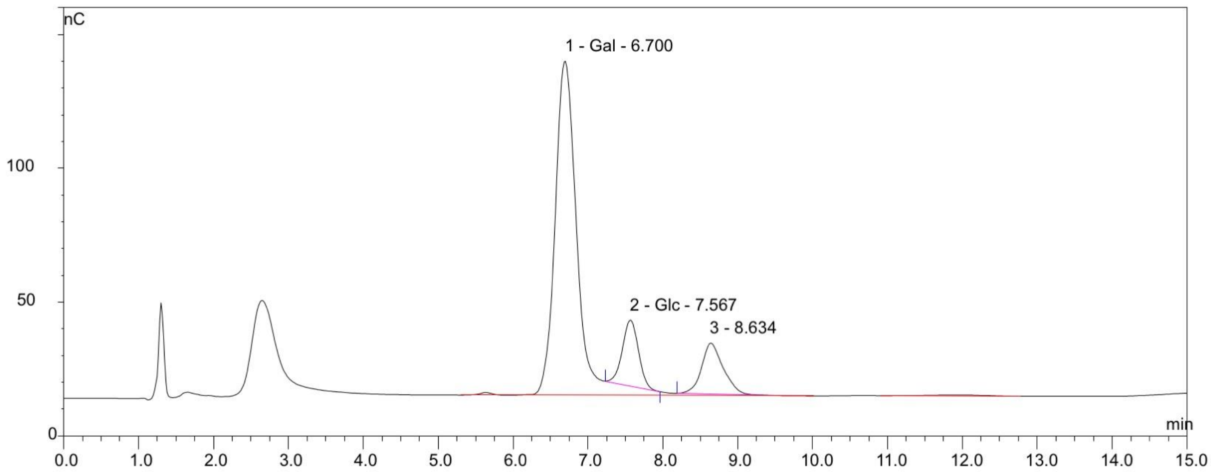


Figure S8: O-linked glycans in NT1 (GH9-$CBM3_c$-$CBM3_b$) HPAEC chromatogram of monosaccharides from NT1 (GH9-$CBM3_c$-$CBM3_b$) O-linked glycans separated by CarboPacPA 20 column shows presence of; galactose (Gal), glucose (Glc), and mannose (Man). Although Glc was detected, we could not determine if this residue is part of the O-glycans structures or simply a contaminant as often is the case.

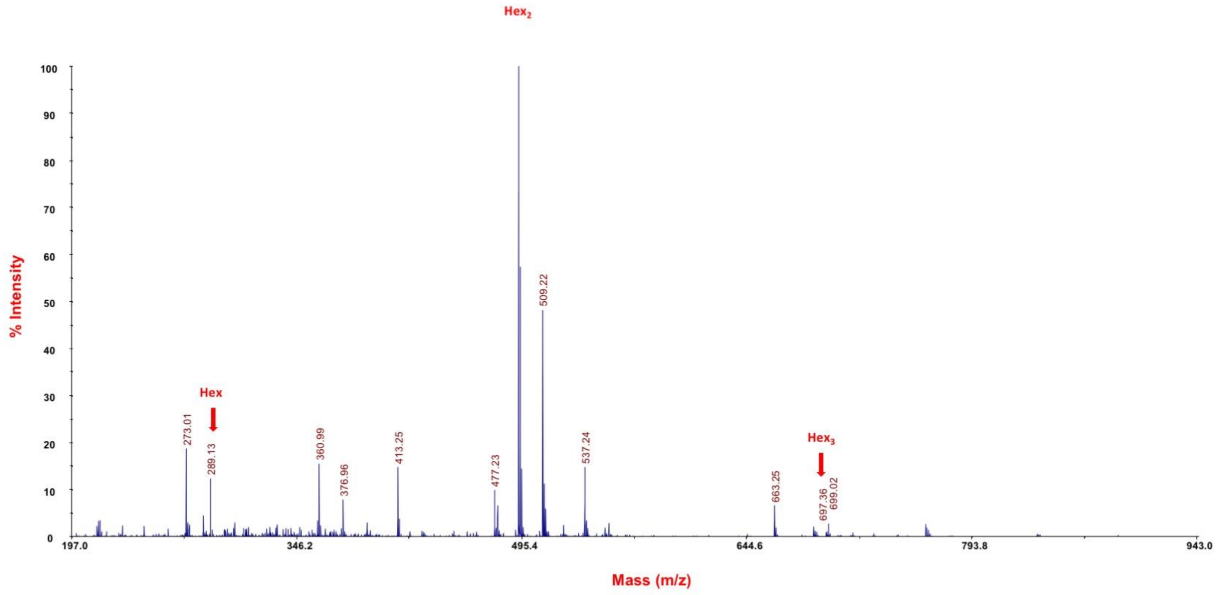


Figure S9: MALDI-TOF-MS Spectrum of glycomics analysis of CT (CBM3$_b$-GH48) sample by ß-elimination. The presence of Hex$_1$, Hex$_2$, and Hex$_3$ is shown.

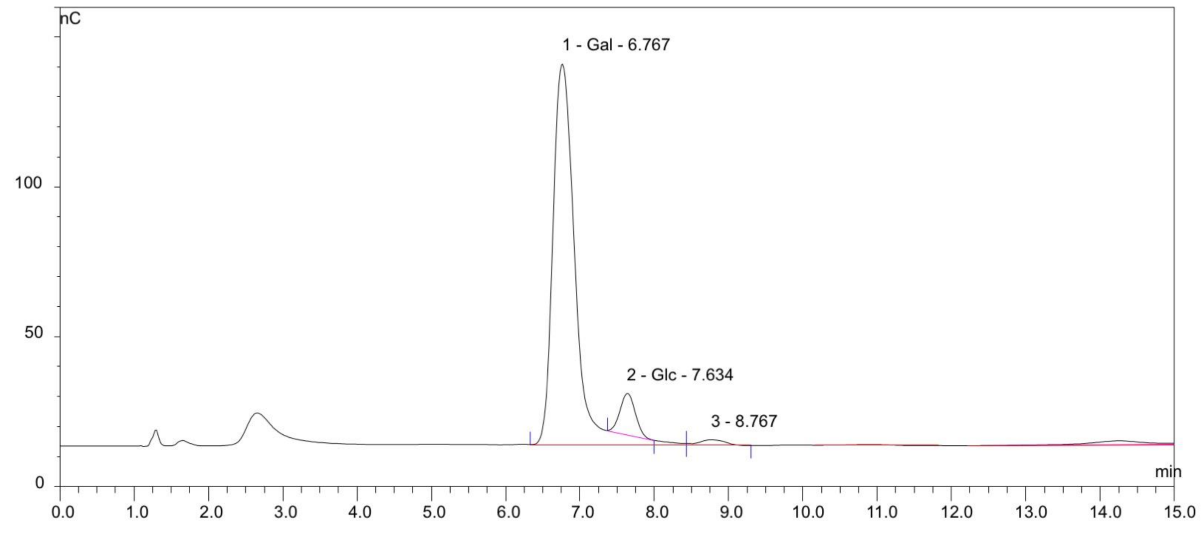
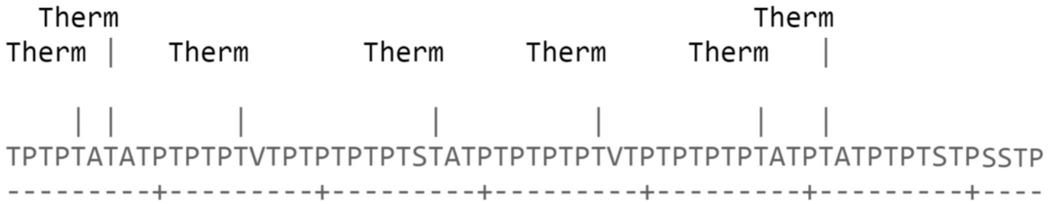


Figure S10: O-linked glycans in CT (CBM3$_b$-GH48) HPAEC chromatogram of monosaccharides from CT (CBM3$_b$-GH48) O-linked glycans separated by CarboPacPA 20 column shows presence of, galactose (Gal), glucose (Glc), and mannose (Man). Although Glc was detected, we could not determine if this residue is part of the O-glycans structures or simply a contaminant as often as the case.



```
  Therm                                       Therm
Therm |   Therm       Therm     Therm     Therm   |

    | |       |           |         |         |   |
TPTPTATATPTPTPTVTPTPTPTPTSTATPTPTPTPTVTPTPTPTPTATPTATPTPTSTPSSTP
---------+---------+---------+---------+---------+---------+----
```

Figure S11: Predicted thermolysin cleavage sites in linker between CBM3$_c$ and CBM3$_b$. The ExPASY tool PeptideCutter predicts sites for CelA linker cleavage based on the canonical recognition sequence of thermolysin. Within the CelA linker sequences, threonine-alanine and threonine-valine motifs are common; thermolysin preferentially cleaves the scissile bond N-terminal to a valine or alanine.

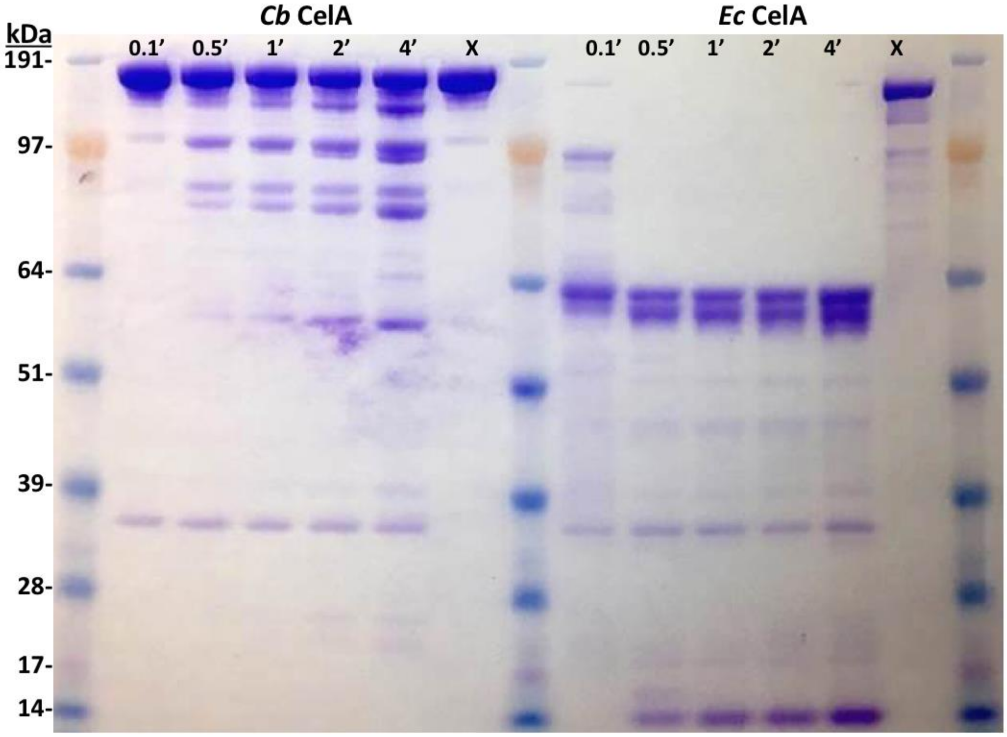


Figure S12: Cleavage with a promiscuous, thermostable protease highlights protective role of glycosylation. CelA expressed in *C. bescii* and *E. coli* were digested at 70°C by thermolysin from *Bacillus thermoproteolyticus* for four hours. Thermolysin scissile bonds are present throughout the amino acid sequence of CelA, including in the linker peptides (Figure S11). *C. bescii* CelA is robust to thermolysin degradation. Even after four hours, most of the protein remains intact. However, *E. coli* CelA is fully degraded in only thirty minutes. Glycosylation distributed throughout the linker peptides provides a strong protective barrier against thermolysin cleavage. (X) Protein subjected to heat only, no thermolysin, for 4 hours.

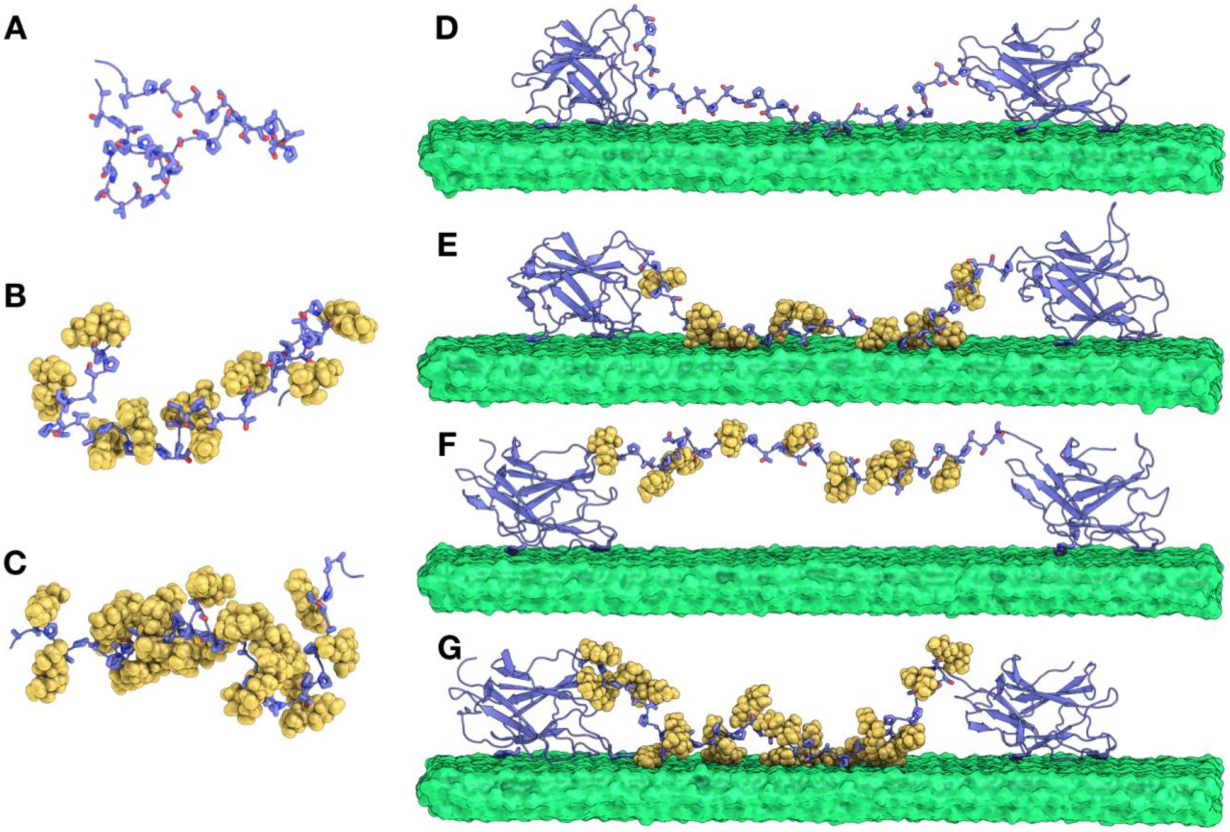


Figure S13: Setup configurations for molecular dynamics (MD) simulations. Linker 3 (see Figure 2A in main text) in solution: A) non-glycosylated, B) with "base case" glycosylation, and C) "doubly glycosylated." Linker 3 with adjacent CBM3$_b$ domains on the hydrophobic surface of cellulose: D) non-glycosylated, E) with "base case" glycosylation, F) with "base case" glycosylation started with linker unbound, and G) "doubly glycosylated."

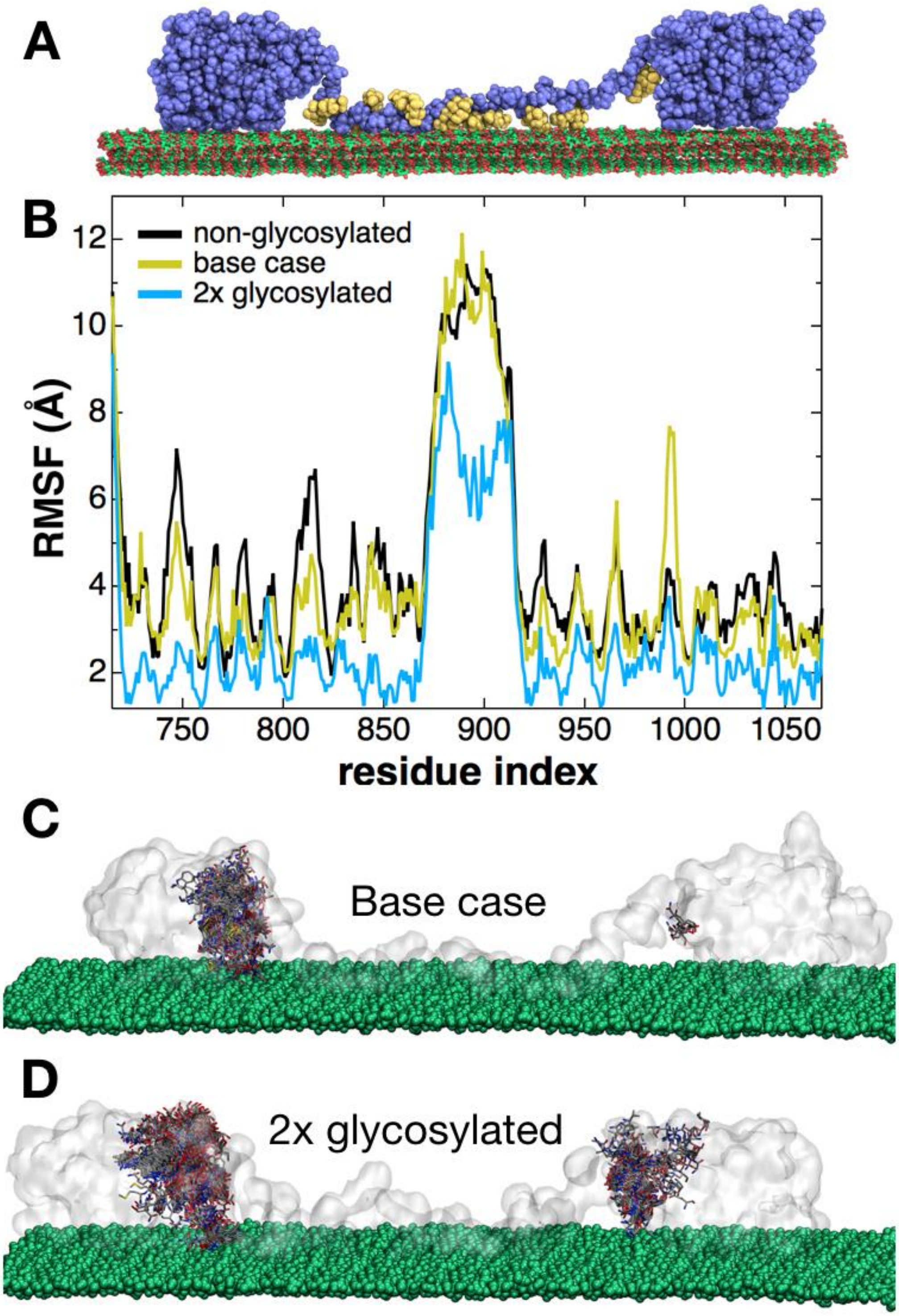


Figure S14: Effect of glycosylation on spatial fluctuations from MD simulations. A) System set-up with “base case” glycosylation; B) Root mean squared fluctuations (RMSF) for the three levels of glycosylation studied. C) The residues on each CBM3$_b$ domain within 4 Å of any linker glycan are shown in gray “sticks” every 2 ns over the course of the 340 ns simulation for C) base case glycosylation and D) doubly glycosylated system. The top level of the cellulose surface is shown as green spheres, the CBM3$_b$ and linker domains are shown as a transparent surface.

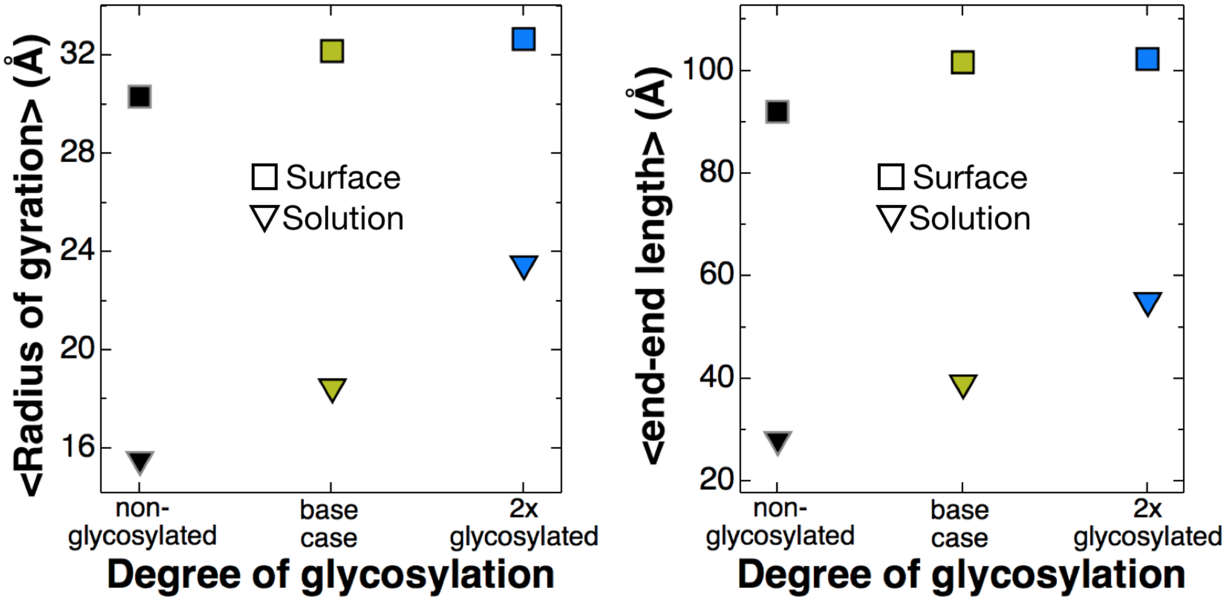


Figure S15: Effect of glycosylation on linker extension from MD simulations. As measured by both radius of gyration (left) and end-to-end length (right), addition of linker glycans extends the linker domain, both with isolated linker in solution and with adjacent $CBM3_b$ domains on the surface of cellulose

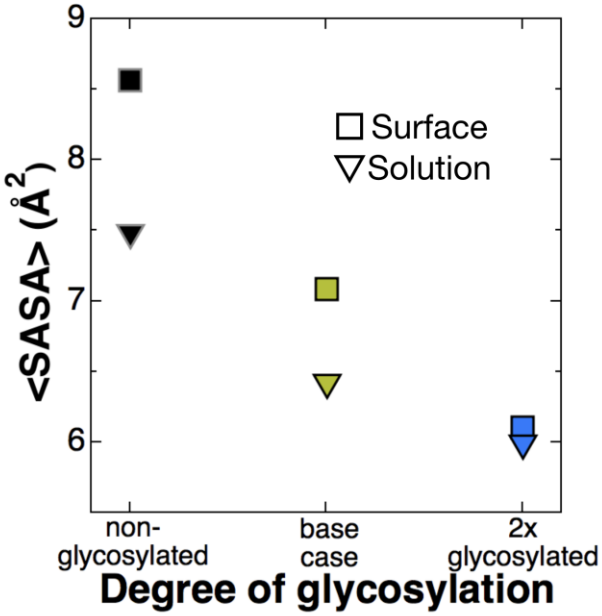


Figure S16: Effect of glycosylation on solvent accessible surface area (SASA) from MD simulations. Both with isolated linkers in solution and for linkers attached to $CBM3_b$ domains on the surface of cellulose, SASA is reduced by the addition of linkers. This is despite the extending effect of the linker glycans (Figure S15). This is calculated with a 1.4 Å probe size, as appropriate for a water molecule.

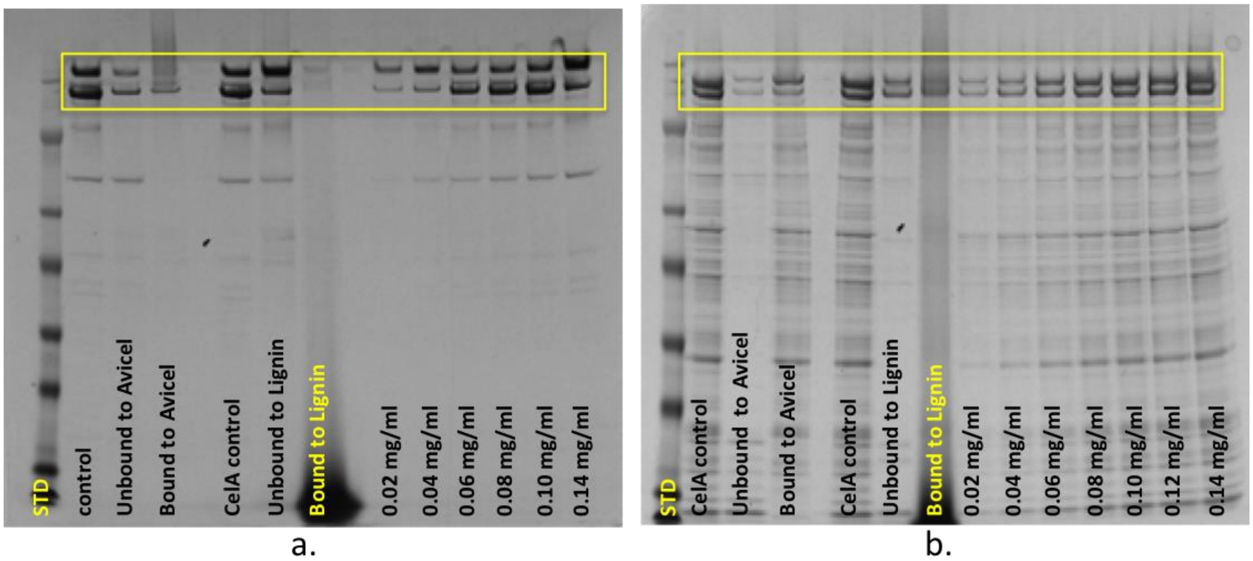


Figure S17: SDS gels of *C. bescii* (a) and *E. coli* (b) CelA after incubation with Avicel and lignin.

At first glance, there appears to be a difference in the affinity of CelA toward Avicel and lignin when comparing expression in *C. bescii* and *E. coli*. In Figure S16a, there is a CelA protein band in the lane of the unbound supernatant for Avicel as well as in the unbound supernatant to lignin, whereas in the case with CelA expressed in *E. coli* the CelA protein band appears to be primarily in the bound fraction of Avicel but it less obvious in the case of lignin. Therefore, densitometry was performed on these gels to measure and quantify the amount of protein lost to both Avicel and lignin utilizing ImageJ gel analysis (developed at the National Institutes of Health (http://rsb.info.nih.gov/ij/docs/menues/analyze.html#gels)). Densitometry was used in conjunction with ImageJ to quantify the amount of protein and the percent difference between the control and unbound fractions (*5, 6*).

A calibration curve between pixel density and total protein concentration was developed (Figure S18) for CelA expressed in both *C. bescii* and *E. coli* using the purified enzymes at varying concentration from 0.02 mg/ml to 0.14 mg/ml. From this standard curve, protein concentrations were calculated for the control and unbound fraction. Protein quantification was not performed on the bound fraction due to the increased background within the lane caused by the Avicel and lignin residues.

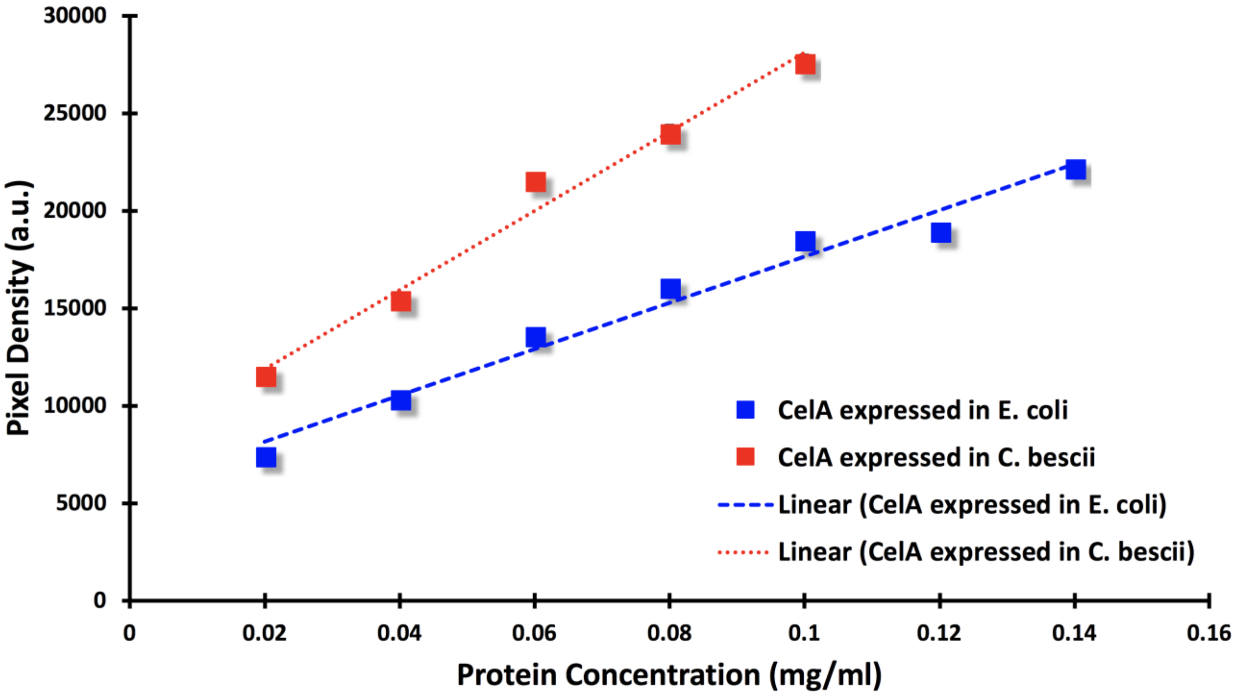


Figure S18: Calibration curves for CelA expressed in *C. bescii* (red) and CelA expressed in *E. coli* (blue) for quantification of protein concentration. These curves were developed using gel analysis software in Image J and used to measure the amount of protein in a given band (http://rsb.info.nih.gov/ij/docs/menues/analyze.html#gels).

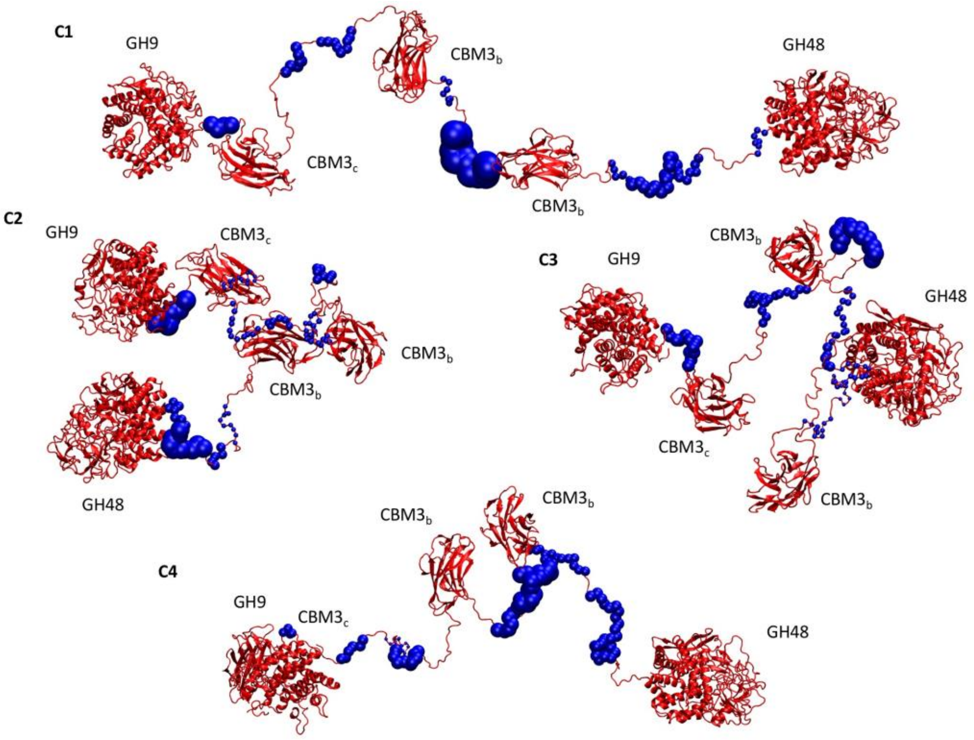


Figure S19: Location of hydrophobic patches in CelA linkers. Hydrophobic patches for CelA linkers calculated using the structures from four different CelA conformations are shown in blue as VDW spheres scaled by scores.

	Linker 1	Linker 2	Linker 3	Linker 4	Total linkers
C1	5.76	7.84	40.96	7.84	62.4
C2	16	5.76	5.76	16	43.52
C3	12.96	19.36	7.84	4	44.16
C4	4	10.24	36	27.04	77.28

Figure S20: Hydrophobic patch score of CelA linkers. The highest hydrophobic patch score for each linker was calculated using the structures from four different CelA conformations. The total score is the sum of the highest scores for each linker. In a previous study, we reported that proteins with scores above 35 were susceptible to non productive binding to lignin.

1. J. Farkas *et al.*, Improved growth media and culture techniques for genetic analysis and assessment of biomass utilization by Caldicellulosiruptor bescii. *J Ind Microbiol Biotechnol* **40**, 41-49 (2013).
2. D. Chung, M. Cha, J. Farkas, J. Westpheling, Construction of a stable replicating shuttle vector for Caldicellulosiruptor species: use for extending genetic methodologies to other members of this genus. *PLoS One* **8**, e62881 (2013).
3. J. Sambrook, D. Russell, *Molecular Cloning: A Laboratory Manual*. (Cold Spring Harbor Laboratory Press, 2001).
4. D. Chung *et al.*, Homologous expression of the Caldicellulosiruptor bescii CelA reveals that the extracellular protein is glycosylated. *PLoS One* **10**, e0119508 (2015).
5. V. Boissonneault, I. Plante, S. Rivest, P. Provost, MicroRNA-298 and microRNA-328 regulate expression of mouse β-amyloid precursor protein-converting enzyme 1. *Journal of Biological Chemistry* **284**, 1971-1981 (2009).
6. J. G. Walsh *et al.*, Executioner caspase-3 and caspase-7 are functionally distinct proteases. *Proceedings of the National Academy of Sciences* **105**, 12815-12819 (2008).