The laterally acquired GH5 ZgEngAGH5_4 from the marine bacterium Zobellia galactanivorans is dedicated to hemicellulose hydrolysis

Jonathan Dorival, Sophie Ruppert, Melissa Gunnoo, Adam Orlowski, Maylis Chapelais-Baron, Jérôme Dabin, Aurore Labourel, Damien Thompson, Gurvan Michel, Mirjam Czjzek, et al.

The laterally-acquired GH5 *Zg*EngA$_{GH5_4}$ from the marine bacterium *Zobellia galactanivorans* is dedicated to hemicellulose hydrolysis

Jonathan Dorival[a§], Sophie Ruppert[a§], Melissa Gunnoo[b], Adam Orłowski[b], Maylis Chapelais-Baron[a], Jérôme Dabin[a], Aurore Labourel[a], Damien Thompson[b], Gurvan Michel[a], Mirjam Czjzek[a*] and Sabine Genicot[a*]

[a] Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff (SBR), 29680 Roscoff, France

[b] Department of Physics, Bernal Institute, University of Limerick, Limerick V94 T9PX, Ireland

§These authors contributed equally to this work.

*Co-corresponding authors to whom correspondence should be addressed:

Mirjam Czjzek, Station Biologique de Roscoff (SBR), Place Georges Teissier, 29680 Roscoff, France; czjzek@sb-roscoff.fr; Phone number: + 33 2 98 29 23 75

Sabine Genicot, Station Biologique de Roscoff (SBR), Place Georges Teissier, 29680 Roscoff, France; genicot@sb-roscoff.fr; Phone number: + 33 2 98 29 23 30

Short title: Structure/ function analysis of a marine GH5_4 endoglucanase

ABBREVIATIONS: PUL: polysaccharide utilization locus; MLG, mixed-linked glucan; GH, glycoside hydrolase; GH5, family 5 of glycoside hydrolases; CAZy, Carbohydrate active Enzyme; MES , 2-(N-Morpholino)ethanesulfonic acid hydrate ; MOPS, 3-morpholino-1-propanesulfonic acid; Tris, 2-amino-2-hydroxymethyl-1.3-propanediol; CMC, Carboxymethylcellulose; G, Glucose; G2, cellobiose; G3, cellotriose; G4, cellotetraose; G5, cellopentaose; G6, cellohexaose; G3G, L2, laminaribiose; G3GG, G3A, Glucosyl-(1->3)-β-D-cellobiose; GG3G, Cellobiosyl-(1->3)-β-D-Glucose; GGG3G, Cellotriosyl-(1->3)-β-D-Glucose; GG3GG, Cellobiosyl-(1->3))-β-D-cellobiose; HPAEC, High Performance Anion Exchange Chromatography; FACE, Fluorophore assisted carbohydrate electrophoresis ; ANTS, 8-aminonaphthalene-1,3,6-trisulfonate; $NaBH_3CN$, cyanoborohydride; DMSO, dimethylsulfoxyde; SEC, size exclusion chromatography; DLS, dynamic light

scattering; SEC-MALLS, size exclusion chromatography coupled to multiple angle laser light scattering; MD, Molecular dynamics; RMSD, Root Mean Square Deviations.

KEYWORDS: Endo-β-(1.3- 1.4) glucanase; MLG; glucomannan; *Zobellia galactanivorans*; PUL; horizontal gene transfer.

SUMMARY STATEMENTS

Marine heterotrophic bacteria play a crucial role in the carbon cycle since some species have developed important enzymatic machineries to degrade algal polysaccharides. We show here that the model algae-associated bacterium *Zobellia galactanivorans* is also involved in biodegradation of hemicellulose through the activity of a horizontally-acquired endo-β-D-glucanase.

ABSTRACT

Cell walls of marine macroalgae are composed of diverse polysaccharides that provide abundant carbon sources for marine heterotrophic bacteria. Among them, *Zobellia galactanivorans* is considered as a model for studying algae-bacteria interactions. The degradation of typical algal polysaccharides, such as agars or alginate, has been intensively studied in this model bacterium, but the catabolism of plant-like polysaccharides is essentially uncharacterized. Here we identify a polysaccharide utilization locus in the genome of *Z. galactanivorans,* induced by laminarin (β-1,3-glucans), and containing a putative GH5 subfamily 4 (GH5_4) enzyme, currently annotated as a endoglucanase (*Zg*EngA$_{GH5_4}$). A phylogenetic analysis indicates that *Zg*EngA$_{GH5_4}$ was laterally acquired from an ancestral *Actinobacteria*. We performed the biochemical and structural characterization of *Zg*EngA$_{GH5_4}$ and demonstrate that this GH5 is in fact an endo-β-glucanase, most active on mixed-linked glucan (MLG). Although *Zg*EngA$_{GH5_4}$ and GH16 lichenases both hydrolyze MLG, these two types of enzymes release different series of oligosaccharides. Structural analyses of *Zg*EngA$_{GH5_4}$ reveal that all the amino acid residues involved in the catalytic triad and in the negative glucose binding subsites are conserved, when compared to the closest relative, the cellulase EngD from *Clostridium cellulovorans,* and some other GH5s. In contrast, the positive glucose binding subsites of *Zg*EngA$_{GH5_4}$ are different and this could explain the preference for MLG, with respect to cellulose or laminarin. Molecular dynamics computer simulations using different hexaoses reveal that the specificity for MLG occurs through the +1 and +2 subsites of the binding pocket that display the most important differences when compared to the structures of other GH5_4 enzymes.

INTRODUCTION

For a long time, the presence of mixed-linked glucans (β-(1,3-1,4)-glucans, MLG) in primary cell walls was considered a unique feature that has evolved in flowering plants (for review see for example [1]). This vision was first challenged by a large and systematic analysis across the plant kingdom using a glycan microarray approach, which highlighted that MLG were also present in some less commonly found, early diverging vascular plants and freshwater green algae [2, 3]. Surprisingly and more recently, β-(1,3-1,4)-glucans have been identified in the cell wall of brown macro-algae [4]. Well studied for their occurrence in the cell walls of grasses, these β-(1,3-1,4)-glucans are a major component of carbohydrate storage compounds in the endosperm of cereals, such as barley, rice, or wheat [2]. These glucans consist of linear chains of β-1,3- and β-1,4-linked glucosyl residues, and the pattern of distribution of these two linkages varies according to the plant botanical origin and growth conditions [4], in particular the distribution of β-1,3-linkages was found to be more frequent in the marine brown algae [4].

Involved in important carbon storage catabolizing processes, enzymes that efficiently hydrolyse these substrates (frequently named lichenases, mixed-linked-glucanases or termed β-(1,3-1,4)-glucanases) are found largely distributed in many kingdoms of life (*i.e.*, plants, bacteria, fungi) and their sequences are present in numerous glycoside hydrolase (GH) families, which are GH5, GH9, GH16, GH17 and GH26 according to the CAZy (Carbohydrate Active Enzymes) database (http://cazy.org; [5]). Among these different GH families, to date most characterized bacterial β-(1,3-1,4)-glucanases are found in the families GH16 [6] and GH5 [7] based on the CAZY database [5].

GH5 is one among the large families in the CAZy database, with more than 12,000 available sequences. Enzymes belonging to this family are retaining glycoside hydrolases that operate *via* the classical Koshland double-displacement mechanism [8]. The first crystallographic structure of a member of the GH5 family, solved in 1995 [9], was considered a pure β-1,4-glucanases (cellulase). It revealed a $(\beta/\alpha)_8$ barrel fold, common to several other GH families, founding the structural clan GH-A. Since then, up to 20 different activities have been reported for this large family [7], hindering assignment of enzyme specificity, although they are predicted to be involved mainly in plant cell wall

degradation. Family GH5 has recently been subdivided into 51 subfamilies to improve correspondence between specificity and sequence [7]. Several recent structure-function studies [10-12], covering various GH5 subfamilies with formerly undefined specificities, have shed new light on important residues lining the catalytic active site cleft that govern substrate specificity.

GH5 enzymes are relatively frequent in marine *Bacteroidetes*, especially in *Flavobacteriia,* which are the prevalent class of *Bacteroidetes* in marine ecosystem [13]. However *Flavobacteriia* species do not efficiently degrade crystalline cellulose [14, 15]. Regrettably, and without taking into account its polyspecificity, the GH5 family has often been used as a 'marker' for cellulose occurrence in marine environments in microbial ecology studies. Nonetheless, a study highlights that the abundance of GH5 enzymes (mainly belonging to *Gammaproteobacteria*, *Firmicutes* and *Actinobacteria*) positively correlates with chlorophyll concentration in the eastern part of the North Atlantic Ocean, and that the diversity of GH5 enzymes was greater in coastal water than in the open ocean [16].

In the present study, we have analyzed, using multiple biochemical approaches, the detailed structure-function relationship of one of the three GH5 enzymes from *Zobellia galactanivorans* Dsij^{T}, a model macroalgae-associated bacterium [13]. The gene name of this GH5 enzyme *(engA)* was given in the initial genome annotation of *Z. galactanivorans* [13] by homology to the closest characterized enzyme, the endoglucanase EngD from *Clostridium cellulovorans* [17]. The corresponding recombinant enzyme will thus be named here *Zg*EngA$_{GH5_4}$. The evolutionary trail of this enzyme leading to its presence in the genome of this marine flavobacterium is also discussed.

EXPERIMENTAL

Unless otherwise stated, all chemicals were purchased from Sigma.

Phylogenetic analysis

Homologues *Zg*EngA$_{GH5_4}$ (gene: *engA*; systematic ID: ZGAL_208) were identified using BlastP at the GenBank database. These sequences were aligned using MAFFT with the iterative refinement method and the scoring matrix Blosum62 [18]. This multiple alignment allowed calculation of model tests and maximum likelihood trees with MEGA version 6.0.6 [19]. Tree reliability was tested by bootstrap using 100 resamplings of the dataset. The trees were displayed with MEGA 6.0.6.

Cloning and site-directed mutagenesis

The *engA* gene encodes a 397 amino acids protein which includes a peptide signal (residues 1 to 20, analyzed with LipoP 1.0 [20]) and a large GH5 module (residues 21 to 397) (Figure S1). For the biochemical and structural characterizations, the precise boundaries of the catalytic module were identified using Hydrophobic Cluster Analysis (HCA) plot [21]. Genomic DNA from *Zobellia galactanivorans* was prepared as previously described [22]. The primers forward (5'-gggggg<u>agatct</u>aatatgagggagatagcccctaag-3'; BglII restriction site is underlined) and reverse (5'-ccccccc<u>aattg</u>ttacttaacaatggcctcggcaatttc-3'; MfeI restriction site is underlined), deduced from the *engA* gene of *Z. galactanivorans (*GenBank™ accession no. **<u>CAZ94281.1</u>***)*, were used to amplify the sequence encoding for the catalytic module (residues 56 to 385) (Figure S1). After digestion with the restriction enzymes BglII and MfeI, the purified PCR product was ligated using the T4 DNA ligase into the expression vector pFO4 predigested by BglII and MfeI, resulting in a recombinant protein with a N-terminal hexahistidine tag. The plasmid was then used to transform *E. coli* DH5 α strain for storage and *E. coli* BL21 (*DE3*) strain for expression as described in [23]. The sequence of the gene was checked using a genetic analyzer ABI 3130xl (Applied Biosystems) equipped with 50 cm capillaries and POP7™ polymer. Site directed mutagenesis of *Zg*EngA$_{GH5_4}$ was performed using the QuickChange II XL site-directed mutagenesis kit according to the manufacturer's instructions (Agilent Technologies). Sixteen individual mutations were produced using specific forward and reverse primers

described in Table S1. Mutated plasmids were then used to transform *E. coli* XL-10 Gold^{R} ultracompetent cells (Agilent Technologies) for storage and *E. coli* strains BL21(DE3) (Novagen^{R}) for protein expression. Mutated plasmids were sequenced to confirm the effectiveness and the position of the mutation.

Production and purification of *Zg*EngA$_{GH5_4}$ and mutant ZgEngA$_{GH5_4_E323S}$

Unless otherwise stated, experiments were performed at 20°C. Transformed *E. coli* strains BL21(DE3) (Novagen^{R}) were grown for 72 hours in 250 mL ZYP 5052 medium [24] containing 100 µg mL^{-1}ampicillin. *E. coli* BL21 (DE3) bearing pFO4 without insert was used as the negative control. Culture was stopped by centrifugation at 3,000 *g* for 20 min. at 4°C and the pellet was stored at -20°C until further use. The pellet was then suspended in 5 mL 25 mM Tris HCl (pH 7.5), 100 mM NaCl, 15 mM imidazole (Buffer A) containing 5 µL DNAse I (500 units µL^{-1}). The suspension was incubated for 20 min. at 4°C. The cells were then disrupted using a Cell disruption system (Constant Systems Ltd). After centrifugation for 1 hour at 29,000 *g* and 4°C, the cell-free supernatant was then 0.2 µm filtered before being loaded at a flow rate of 1 mL min^{-1} onto a HisPrep FF 16/10 column (1.6 x 10 cm, GE Healthcare) equilibrated in buffer A. The column was washed at a flow rate of 2 mL min^{-1} with buffer A until the absorbance at 280 nm was negligible. Elution of the protein was performed at the same flow rate using a linear gradient increasing from 15 mM to 500 mM imidazole. The final concentration of imidazole was reached after 10 column volumes. 2 mL fractions were collected during the elution step. Fractions containing the recombinant tagged enzyme were estimated by SDS-PAGE analysis and by Western blot. Transfer from SDS gel onto ready to use 0.2 µm nitrocellulose membrane (BioRad) was performed using a Trans Blot Turbo system in the conditions specified by the manufacturer (BioRad). Monoclonal anti-polyhistidine peroxidase conjugate (Sigma) was used at a final concentration of 1/10000 to specifically recognize the His-tagged fusion protein. Immuno-detection was performed by chemiluminescence using the Clarity Western ECL Substrate kit (BioRad) and visualization was achieved using the Chemi-Capt 50001 software. Fractions containing the his-tagged protein were then pooled prior being loaded at a 2 mL min^{-1} flow rate on top of an HiPrep Desalting FF 26/10 column (2.6x 10 cm, GE Healthcare) previously equilibrated in 25 mM Tris HCl

(pH 7.5), 100 mM NaCl (Buffer B). The same flow rate was used during the elution step and 1 mL fractions were collected. Purity of the desalted *Zg*EngA$_{GH5_4}$ and ZgEngA$_{GH5_4_E323S}$ was further checked by SDS PAGE analysis and dynamic light scattering (DLS).

Protein quantification

Protein amount was estimated at 280 nm using a Thermo Scientific NanoDrop One spectrophotometer. A molar extinction coefficient of 85,500 M^{-1} cm^{-1} and a molecular weight of 37.5 kDa, both deduced from the protein sequence, were used to calculate the concentration of *Zg*EngA$_{GH5_4}$ protein solutions.

Enzymatic activity assay of pure enzymes

Unless otherwise stated, assays were performed using β-D-glucan from barley (0.2 % (W/V) in 50 mM MES buffer pH 6.5) as substrate. The activity was determined using the reducing sugar assay described by Kidby and Davidson [25]. Reactions were performed at 30°C upon incubation of 180 µL of substrate with 20 µL *Zg*EngA$_{GH5_4}$ (100 nM). 20 µL of reaction mixture were withdrawn every 15 seconds and up to 105 seconds and added to 180 µL of ferricyanide reagent. The samples were then incubated for 15 minutes at 95°C and cooled down to 20°C. The absorbance was read at 420 nm using a Spark 10M microplate reader (Tecan, Switzerland). A calibration curve was performed under the same conditions, using glucose solutions at different concentrations (from 0.1 to 1.2 mM) as standard.

Substrate specificity of ZgEngA$_{GH5_4}$

To assess the enzymes specificity among glycan polysaccharides, degradation of the following substrates were assayed: β-D- glucan from barley, lichenan, glucomannan, xyloglucan, CMC, Avicel, Laminarin and curdlan. Activity was measured using the ferricyanide assay described above. Unless otherwise stated, all these substrates were used at a final concentration of 0.2% W/V in 50 mM MES pH 6.5. The enzymatic activity was expressed in min^{-1}.

To refine the characterization of substrate specificity for *Zg*EngA$_{GH5_4}$, standard commercial oligosaccharides were used as substrates and the hydrolysis products were analyzed by HPAEC coupled with pulse amperometry. Based on the major activities on polysaccharides, the following

oligosaccharide substrates were chosen: laminaribiose (G3G), and different cello oligosaccharides (G2, G3, G4, G5 and G6), and also β-(1,3-1,4)- oligosaccharides (G3GG, GG3G, GGG3G and GG3GG). All oligosaccharide substrates were purchased from Megazyme except for GG3GG (Carbosynth). Briefly, 2 mL of oligosaccharides (100 µM) were incubated with 150 µL recombinant *Zg*EngA$_{GH5_4}$ (0.5 µM). Aliquots (170 µL) of the reaction mixture were taken at different times (from 0 to 120 min.) and boiled for 15 min. to stop the reaction. Samples were then filtered through 4 mm syringe filter (Millipore) and 20 µL were injected onto a CarboPac PA1 column (4x 200 mm, Thermo Scientific) equipped with the accompanying guard column (4x 50 mm, Thermo Scientific), both thermostated at 30°C. Elution was carried out using an isocratic flow rate of 1 mL min^{-1} with 175 mM NaOH containing 50 mM NaOAc. Detection of the oligosaccharides was performed by integrated amperometry using a quadruple pulse waveform (E1 +0.1, E2, -2.0, E3 +0.6 and E4 -0.1). Integration of signal intensities was performed using the Chromeleon 6.80 software. Calibration of the different oligosaccharides was done using different concentrations of appropriate oligosaccharides from which a dose-response was determined using the Chromeleon software.

Fluorophore assisted carbohydrate electrophoresis (FACE) was performed to further analyze the specificity of *Zg*EngA$_{GH5_4}$ on oligo- and poly-saccharides. Depending on the objective, labelling of oligosaccharides was either performed prior hydrolysis with *Zg*EngA$_{GH5_4}$ or after hydrolysis. However, the applied reaction conditions were the same. Briefly, 100 µg of poly- or oligo- saccharides were labelled with 2 µL 150 mM 8-aminonaphthalene-1,3,6-trisulfonate (ANTS) and incubated for 30 min. at 37°C before the addition of 5 µL of 1M $NaBH_3CN$ in DMSO. Incubation at 37°C was further continued for about 4 to 5 hours. Samples were then dried under vacuum before being suspended. The labeled oligosaccharides were either suspended at ~2- 2.5 µg µL^{-1} in 25% glycerol (W/V) or, if used as substrate for enzymatic hydrolysis, at ~3.5 µg µL^{-1}in 50 mM MES (pH 6.5).

Conditions used for the hydrolysis of oligosaccharides were as follows: 50 µg of oligosaccharides (2 µg µL^{-1} non labeled and ~3.5 µg µL^{-1} labeled) in 50 mM MES (pH 6.5) were incubated overnight at 30°C with 4 µL of 100 nM *Zg*EngA$_{GH5_4}$. Hydrolysis of β-D- Glucan (450 µg in 50 mM MES pH 6.5) was performed overnight at 30°C using either 10 µL of 100 nM *Zg*EngA$_{GH5_4}$ or 10 µL of lichenase

(0.18 U mg^{-1}, 0.7 U mL^{-1}; Megazyme). After incubation, enzymes were inactivated for 10 minutes at 100°C. For each reaction, a blank was made under the same conditions except that the enzyme was first inactivated for 10 minutes at 100°C prior to the incubation with the poly- or oligo-saccharides. About 8-10 µg labeled oligo- and ~12.5 µg labeled β-D glucan (both in 25% glycerol (w/v) final concentration) were loaded on a chilled 27% polyacrylamide gel. The electrophoresis was performed in the dark at 125 volts (constant voltage), 4°C, using chilled migration buffer (25 mM Tris, 192 mM Glycine, pH 8.5). Visualization of the fluorescent oligosaccharides was achieved under UV using a UV Transiluminator (Thermofisher Scientific Bioblock).

Optimal pH determination of $ZgEngA_{GH5\_4}$

The Teorell and Stenhagen buffer (pH 4.2 to 8.5) [26] was used at a final concentration of 100 mM to evaluate the pH optimum. Both enzyme and MLG were diluted in this buffer prior hydrolysis reactions which were performed as described above. The enzymatic activity was estimated using the ferricyanide assay. Results are expressed as percentage of relative activity.

Optimal temperature determination of $ZgEngA_{GH5\_4}$

For this measurement both *Zg*$EngA_{GH5\_4}$ and β-D-glucan from barley were incubated at different temperatures (from 5 to 60°C), in steps of 5°C. The produced amount of reducing sugars was determined as described above. Results are expressed as the percentage of relative activity.

Thermostability analysis

The thermostability of *Zg*$EngA_{GH5\_4}$ was studied by DLS using a Zetasizer Nano instrument (Malvern). *Zg*$EngA_{GH5\_4}$ (1.15 mg mL^{-1}) was filtered through a 0.2 µm membrane filter prior to being heated from 5 to 65°C in steps of 1°C. The hydrodynamic gyration radius (Rg) was measured at each step and the denaturation temperature is defined as the temperature for which the gyration radius sharply increases.

Oligomerization state studies

The oligomerization state of *Zg*$EngA_{GH5\_4}$ was determined both by size exclusion chromatography (SEC) and by size exclusion chromatography coupled to multiple angle laser light

scattering (SEC-MALLS). For the SEC experiments, ~ 600 µg of affinity chromatography purified $ZgEngA_{GH5\_4}$ in 1 mL of buffer B were loaded on top of a Superdex 75HiLoad 16/60 column (GE Healthcare) previously equilibrated in buffer B. The elution was performed at a 0.7 mL min^{-1} flow rate. Calibration of the column was carried out in the same conditions using the appropriate calibration standards (GE Healthcare). For the experiments of size exclusion chromatography coupled with MALLS, 100 µL at 300 µg mL^{-1} of $ZgEngA_{GH5\_4}$ from the Superdex 75 chromatography were loaded onto a Superdex 200 Increase 10/300 GL column (GE Healthcare), previously equilibrated for at least 24 hours in buffer B. Elution of the protein was performed for 80 minutes at a flow rate of 0.5 mL min^{-1} and the detection was carried out using both an Optilab rEX detector (Wyatt) and Dawn Heleos light Scattering detector (Wyatt). Results were analyzed using the ASTRA V software (Wyatt Technology)

Crystallization and structure determination

Crystals for $ZgEngA_{GH5\_4}$, in solution at a concentration of 15 mg mL^{-1}, were obtained using the hanging drop vapor diffusion method by mixing 2 µL of protein solution with 1 µL of crystallization solution composed of 14% PEG 6000, 200 mM $CaCl_2$, 100 mM sodium acetate buffer at pH 5.0. Crystals were cryo-protected using the crystallization buffer supplemented with 10% glycerol and flash frozen in a N_2-stream at 100K. X-ray diffraction data were collected at 1.2 Å resolution at the European Synchrotron Radiation Facilities (ESRF, Grenoble France) on beamline ID23-1. The images were integrated using XDS [27] in the space group $P2_1$. The structure was solved by molecular replacement with MolRep [28] using EngD (PDB ID: 3NDZ) as the search model. An initial model was built automatically with the CCP4 version of ARP-wARP, [29] with several cycles of manual rebuilding in Coot [30] and refinement with Refmac5 [31].

Crystals of $ZgEngA_{GH5\_4\_E323S}$ in complex with cellotriose were obtained using the same hanging drop method, in drops containing 2 µL of $ZgEngA_{GH5\_4\_E323S}$ mutant at 7 mg mL^{-1} mixed to 1 µL of well solution consisting in 24-24.5% PEG 3350, 160 mM $MgCl_2$, 100 mM Bis- Tris pH 5.5. $ZgEngA_{GH5\_4\_E323S}$ mutant was co-crystallized with 0.04% of a mixture of oligohexa- to oligonona-saccharides obtained from limited digestion of MLG by $ZgEngA_{GH5\_4}$. Crystals were soaked in crystallization buffer supplemented with 30% glycerol before being frozen in liquid nitrogen. Data

were collected at the Soleil synchrotron on beamline Proxima1 to 2.2 Å resolution. The images were integrated using XDS [27] and the space group $P3_2$. The structure of *Zg*EngA$_{GH5_4_E323S}$ in complex with substrate was solved by molecular replacement, using the software Phaser [32] and the structure of *Zg*EngA$_{GH5_4}$ as the search model. The structure was refined using REFMAC [31] in iterative cycles with manual corrections using the graphic interface Coot [30]. All Figures representing the structures were prepared using the program PyMol (Schrödinger, LLC). The atomic coordinates and the atomic factors of both *Zg*EngA$_{GH5_4}$ and *Zg*EngA$_{GH5_4_E323S}$ have been deposited at the Protein Data Bank collection (http://wwwpdb.org/) as PDB ID: 6GL2 and PDB ID: 6GL0 respectively.

Molecular Dynamics simulations

Molecular dynamics (MD) simulations were performed to predict the cellulose recognition properties of native *Zg*EngA$_{GH5_4}$ and *Zg*EngA$_{GH5_4_E323S}$. The input starting protein structures for the calculations were generated using the experimental crystal coordinates, and the cellulose substrate was built into the binding pocket of the enzyme. Specifically, four different oligo-glucose, hexameric chains were tested in the simulations: a cellohexaose chain with β-1,4 linkages between all sugars, and three mixed-linked oligoglucans with the β-1,3 linkage at different positions (Fig. S2). The simulations are summarized in Supporting Information (Tables S2-S5) together with details of the model construction and simulation protocols. Each model was named according to the position and type of linkage present in the oligosaccharide substrate: β(1,4) for the oligosaccharide with β-1,4 linkages between all units; -1/+1 β(1,3) for that with a β-1,3 linkage between units -1 and +1; +1/+2 β(1,3) for that with a β-1,3 linkage between units +1 and +2; and +2/+3 β(1,3) for that with a β-1,3 linkage between units +2 and +3. The final coordinates for all trajectories can be accessed through the following link: https://1drv.ms/f/s!ArX4zU6cjMUQnHQ9m5ScGJkyY_Kr

RESULTS

The *engA* gene was acquired from an ancestral clostridial bacterium

The *engA* gene (systematic ID: ZGAL_208), coding a single CAZyme module, is localized within a potential polysaccharide utilization locus (PUL_4) [13] (Fig. 1). Besides *engA*, PUL_4 also includes a gene encoding a lipoprotein of unknown function (*ZGAL_209*) displaying a C-terminal carbohydrate binding module (CBM4) and two tandem susD/SusC-like pairs (*ZGAL_211/212* and *ZGAL_213/214*) (Fig. 1). PUL_4 was previously identified as strongly induced by β-1,3-glucans [33]. In the context of the development of a new screening method for carbohydrate-related proteins, the susD-like protein *ZGAL_213* was shown to specifically bind xyloglucans [34]. Altogether, these transcriptomic and biochemical results suggest that PUL_4, and thus likely the *engA*-encoded protein, which will be named *Zg*EngA$_{GH5_4}$, could be involved in the degradation of hemicelluloses.

Homology searches in the GenBank database indicate that close homologues of *Zg*EngA$_{GH5_4}$ are relatively rare in other marine flavobacteria. Unexpectedly, this protein is highly similar to numerous subfamily GH5_4 beta-glucanases from *Firmicutes* (e.g. 51% sequence identity with the cellulase EngD from *Clostridium cellulovorans* [17]). A phylogenetic analysis of the GH5_4 subfamily indicates that *Zg*EngA$_{GH5_4}$ belongs to a clade only composed of marine flavobacterial proteins. This late-diverging clade is rooted by two successive clades of GH5_4 proteins from *Firmicutes* (*Clostridia* class) (Fig. 2). Therefore, the paucity of *Zg*EngA$_{GH5_4}$ homologues in marine flavobacteria and their phylogenetic position support that marine flavobacteria have horizontally acquired these GH5_4 genes from an ancestral clostridial bacterium.

*Zg*EngA$_{GH5_4}$ is a β-glucanase with broad substrate specificity

In the genome of *Z. galactanivorans*, *engA* was annotated as encoding for an endoglucanase, referred here to as *Zg*EngA$_{GH5_4}$. To verify the prediction of this activity, the nucleotide sequence corresponding to the catalytic module was cloned into a pFO4 plasmid. The protein was produced with a yield of ~130 mg L^{-1} in *E coli* BL21(*DE3*) strain using an auto-inducible medium. The protein was purified to electrophoretic homogeneity by nickel affinity chromatography (Fig. S3). The purity of the enzyme was further confirmed by DLS (Fig. S3). Both SEC and SEC-MALLS were used to verify that

*Zg*EngA$_{GH5_4}$ is a monomer in solution (Fig. S3). SEC-MALLS additionally showed that the enzyme is characterized by a molecular weight of 36.5 kDa. This value is slightly lower than the theoretical molecular weight of 37.5 kDa deduced from the amino acid sequence of the recombinant protein using the ProtParam tool [35] (Fig. S1).

DLS was also used to study the thermostability of *Zg*EngA$_{GH5_4}$ (Fig. S4A). Indeed, in the range of temperatures from 5 to 37 °C, the protein is characterized by a hydrodynamic radius of gyration of 5.59 ± 0.12 nm. This value increases when temperature reaches 40°C and is almost doubled at a temperature of 44°C, indicating the enzymatic denaturation which then further increases with temperature.

The ferricyanide reducing sugar assay was used to screen for the hydrolytic activity of the enzyme on several soluble β-glucans such as the carboxymethyl cellulose, lichenan, β-D-glucan from barley, xyloglucan and konjac glucomannan. Activity was also screened on β-1,3-glucans such as laminarin, a β-1,3-glucan from *Euglena gracilis* and carboxy methyl curdlan (a bacterial exopolysaccharide) but revealed to be non detectable (Table 1). Although the enzyme is active on soluble cellulose derivatives, its activity towards this substrate is very low, compared to its activity towards mixed linked β-(1,4-1,3)-glucans, such as β-D-glucan from barley and lichenan from Iceland moss. *Zg*EngA$_{GH5_4}$ is also able to degrade substituted β-1,4-glycosides such as glucomannan and, to some extent, xyloglucan (Table 1). Taken together, these results classify this enzyme as a β-(1,3-1,4)-endoglucanase.

Prior to the determination of which linkage and which minimal substrate the enzyme is able to hydrolyze, an evaluation of its optimal conditions was carried out. Using β-D-glucan (MLG) from barley as substrate, the universal buffer of Teorell and Stenhagen, was used to study the pH dependence of the activity of *Zg*EngA$_{GH5_4}$. The enzyme shows activity between pH 5.0 and 8.5 but the optimum of activity is observed at pH 6.0-6.5 (Fig. S4B). At pH 5.5, the enzyme loses about 40% of its activity, as it does at pH values above 8. Similar results have been observed using biological

buffers such as MES, MOPS, phosphate and Tris, further showing that the activity of the enzyme in the MES buffer is higher than in the other ones (Fig. S4B).

β-D-glucan from barley was also used as substrate to study the influence of the temperature on the enzyme activity. In this case, the substrate was first thermostated at different temperatures, between 5 and 60°C, prior to the enzymatic reaction. As shown on Fig. S4C, the enzyme displays an optimal activity at 45°C. At 50°C, the enzyme loses almost half of its activity, which drops to only about 10 % at 60°C. In order to avoid denaturation, all the subsequent enzymatic reactions were however performed at 30°C, a compromise temperature between enzymatic activity and stability. The influence of NaCl was evaluated at different concentrations, up to 1M and seems to have no significant effect on the enzymatic activity

_Zg_EngA$_{GH5_4}$ is able to cleave both β-1,3 and β-1,4 linkages

To establish which linkages are cleaved by _Zg_EngA$_{GH5_4}$, different standard β-1,4 and β-(1,4-1,3)-oligosaccharides were used as substrates. The hydrolysis products were identified by HPAEC using a CarboPAc PA1 column, specifically dedicated to the separation of small oligosaccharides. From these experiments, it appears that, even when the reaction lasts overnight, _Zg_EngA$_{GH5_4}$ is unable to hydrolyze di- and tri-saccharides, whether they originate from cellulose or MLG polymers. A minimum of 4 glucose units (G4) is therefore essential for the activity of the enzyme.

In a first step, the nature and the concentration of different products released during hydrolysis of cello- oligosaccharides by _Zg_EngA$_{GH5_4}$ were measured as a function of time (Fig. 3). While after 60 minutes, 20 % of cellotetraose (G4) remain to be hydrolyzed (Fig. 3A), hydrolysis of cellopentaose and cellohexaose are a lot faster as they both are completely hydrolyzed within 1 and 2 minutes, respectively (Fig. 3B and C) Hydrolysis of cellopentaose (G5) is straightforward and produces only cellobiose (G2) and cellotriose (G3) (Fig. 3B). Hydrolysis of cellohexaose (G6) proceeds in two steps, since both cellotetraose (G4) and cellotriose (G3) are produced within the first two minutes, however, as hydrolysis proceeds, cellotetraose (G4) is further hydrolyzed into cellobiose (G2) (Fig. 3C).

In a next step, and to evaluate the ability of _Zg_EngA$_{GH5_4}$ to hydrolyze β-1,3 linkages, two β-(1,4-1,3)-tetrasaccharides, namely GGG3G and GG3GG, differing from each other by the position of

the β-1,3 linkage, were used as substrates. In GGG3G the β-1,3 linkage is at the reducing end, while it is flanked by a β-1,4 linkage on both sides in GG3GG. Hydrolysis of GGG3G (G4B) yielded glucose (G1), cellobiose (G2), cellotriose (G3) and laminaribiose (G3G; L2) (Fig. 3D) whilst only cellobiose (G2) was produced upon hydrolysis of GG3GG (G4C) (Fig. 3E). Altogether, these results indicate that *Zg*EngA$_{GH5_4}$ is able to accommodate both β-1,4 and β-1,3 linked glucose in the +1 binding subsite, whereas only β-1,4 linkages are accepted in the negative binding subsites (Fig. 4E and F). It also shows that the specificity of *Zg*EngA$_{GH5_4}$ is dictated by the position of the β-1,3 linkages.

Hydrolysis of oligosaccharides, as well as of β-D glucan from barley, was also followed by FACE (Fig. 4A-D). The ANTS was used as a fluorophore to label the reducing end of the oligosaccharides, which were then separated by electrophoresis. In addition to corroborating the results obtained by HPAEC about the products formed upon hydrolysis of the oligosaccharides, this technique showed in particular that cellotetraose and cellohexaose are hydrolyzed with different modes. Indeed, depending on whether labeling was performed on the substrate or on the hydrolysis products, the end products of these oligosaccharides are different: when cellotetraose (G4) is labeled before hydrolysis (Fig. 4A), the fluorescent oligosaccharides migrate as cellotriose (G3) and minor amounts of cellobiose (G2). When labeling is performed after hydrolysis (Fig. 4B), the major oligosaccharide is cellobiose (G2), although there are traces of cellotriose (G3) and even of glucose (G). Similarly, when labeling cellohexaose (G6) before hydrolysis mostly cellotriose (G3) but also some cellobiose (G2) oligosaccharides are detected (Fig. 4A), whilst when cellohexaose (G6) is first hydrolyzed and then labeled, both cellobiose (G2) and cellotriose (G3) are detected at the same intensity (Fig. 4B).

Comparison of the end products generated from the hydrolysis of β-D-glucan with the lichenase and with *Zg*EngA$_{GH5_4}$ reveals that the size of the products are similar but the cleavage sites of the enzymes are different (Fig. 4C). As expected, GG3G and GGG3G are the end products of the MLG hydrolysis by lichenase, whilst G2, G3, G4 and to some extent G5 are the end products observed after hydrolysis with *Zg*EngA$_{GH5_4}$. These oligosaccharides, as well as the complete absence of oligosaccharides with β-1,3 bonds such as GG3G, GGG3G, G3GG and GG3GG attest therefore that,

at the polymer level, the preferred cleavage site of the *Zg*EngA$_{GH5_4}$ are β-1,3 bonds that are neighbored by β-1,4 bonds, at least towards the non-reducing end (Fig. 4G).

Further hydrolysis overnight with *Zg*EngA$_{GH5_4}$ of the labeled oligosaccharides produced by the lichenase shows that GG3G is not hydrolyzed whilst the complete hydrolysis of GGG3G into laminaribiose (G3G) (Fig. 4D) is attained, again attesting thereby that, on small oligosaccharides, *Zg*EngA$_{GH5_4}$ is able to cleave the β-1,4 bond that precedes a β-1,3 bond (Fig. 4F). Altogether, these experiments allow deducing the subsites and their involvement in substrate binding, and they also show that the +1 or +2 binding subsites do not tolerate/accommodate the fluorophore (Fig. 4E and F). Further interpretation of these results is described in the discussion below.

Three-dimensional structure of *Zg*EngA$_{GH5_4}$

In order to determine the molecular basis of substrate recognition by ZgEngA$_{GH5_4}$, we solved the crystal structure of ZgEngA$_{GH5_4}$ wild-type and the E323S mutant (*Zg*EngA$_{GH5_4_E323S}$) in complex with the cellotriose (three glucose units linked by β-1,4 bonds) (Table 2). The structure of *Zg*EngA$_{GH5_4}$ was solved at 1.2 Å resolution by molecular replacement using the structure of EngD (PDB ID: 3NDZ, 51% sequence identity, Fig. 5) as a search model. There is only one molecule in the asymmetric unit. The ZgEngA$_{GH5_4}$ adopts a typical TIM-barrel $(\beta/\alpha)_8$ fold. An additional helix (α0) closes the β-barrel at its N-terminal face, consistent with other GH5 enzymes (Fig. 6A). Structural similarity searches using the DALI server [36] identified close relationship to other GH5 enzymes. The closest ones were the structure of endoglucanase E from *Ruminiclostridium thermocellum* (PDB ID: 4IM4) and of endoglucanase D from *Clostridium cellulovorans* (PDB ID: 3NDZ). Both are GH5 enzymes that exhibit broad substrate specificity, preferentially displaying high activity on β-1,4 linked glucans and xylans.

Like other GH5 enzymes, the active site is formed by a catalytic cleft, which runs across the whole protein, where specific binding subsites recognize each glucose unit. Two glutamic acid residues (E200 and E323, in ZgEngA$_{GH5_4}$) correspond to the catalytic acid-base and nucleophile respectively, and are positioned between the -1 and +1 sub-binding sites (Figs. 6B, 7A). Consistently

with all other TIM-barrel hydrolases, these residues are located at the end of β-strands β4 and β7 (Fig. 5).

The co-crystallization of *Zg*EngA$_{GH5_4_E323S}$ with a mixture of oligosaccharides (mainly hexa- to nona-saccharides, all produced by the native enzyme upon hydrolysis of MLG) resulted in the complex structure solved at 2.2 Å resolution, with three molecules in the asymmetric unit. A clear electron density corresponding to a cellotriose (G3) oligosaccharide, linked by β-1,4 bonds only, is present in the active site of each of the three monomers (Fig. 6B). The presence of this substrate molecule could either be due to a contamination of our oligosaccharide mixture by cellotriose, which are preferentially selected by *Zg*EngA$_{GH5_4_E323S}$, or additional units at the non-reducing end are completely disordered in the crystal structure. These substrate molecules (further on named cellotriose or G3) occupy the negative binding subsites from -3 to -1. The glucose unit bound to the -3 subsite establishes a stacking interaction with W89. In the -2 subsite, N77 and N358 are involved in substrate binding *via* hydrogen bonds. The glucose unit bound to the -1 subsite is the most stabilized one, stacked against W356, and hydrogen bonded to H155, H156, Y277 and E200 (Fig. 6B).

A particular feature in *Zg*EngA$_{GH5_4}$ is the loop following the β-strand β8 that is shorter by 4 residues when compared to CcEngD (PDB ID: 3NDZ) or CcCel5A (PDB ID: 1EDG). This feature creates a more open active site at the non-reducing end (negative binding subsites), which could accommodate branched substrates (Fig.7B). Indeed, in the above mentioned other GH5 enzymes, this loop binds the glucose unit occupying the -3 subsite, by forming hydrogen bonds between an Asp or Glu residue and the O6 of this glucose unit. Here, the residue E363 is located too far to interact with the substrate (Fig.7A).

Another outstanding feature is the conformation of residue Y280. Indeed, the loop between β-strand β6 α-helix α6, carrying this residue, has a completely different conformation than in other GH5 enzymes (Fig. 7B). First, the presence of T287 directed towards the short α6' helix can be noted, whereas in other GH5, this threonine is substituted by a short residue which points to the solvent. This feature forces D285 to adopt a different conformation compared to all other GH5 enzymes. To avoid a steric clash with T287 or D285, the neighboring Y280 is orientated in the opposite direction compared

to tyrosine residues at this position of other GH5 enzymes. The change of conformation of this residue is also possible by the presence of Q281, instead of an aromatic residue at this position in most of the other GH5 enzymes, which would clash with Y280 in this orientation. Overall, the presence of Y280 that changes the loop position also leads to a narrowing of the binding cleft on the positive binding subsites (Fig. S22A). Notably, when replacing Y280 by alanine (Fig. S22B) by computational mutation, the overall substrate binding cleft resembles closely that of F32EG5 (Fig. S22C).

Site directed mutagenesis of selected residues and molecular modeling to explore the catalytic active site

In order to investigate the role of a selection of residues in the active site, we undertook site directed mutagenesis experiments. Based on the 3D structure analyses, we chose to mutate residues that potentially interact with different polysaccharide substrates, outside the -1 sub-binding site, since the importance of residues surrounding the -1 sub-binding site in substrate recognition and catalytic activity has already been demonstrated [10, 37]. We also included two residues, Y82 and E363, which are not directly involved in interaction with a linear polysaccharide but that could accommodate branching in substrates, such as xyloglucan or glucomannan. However, mutation of both of these residues does not affect the activity, even on branched substrate (Table 3). On the other hand, replacement of the residues N77, H156, W210 or N358 by alanine substantially decreases or even completely abolishes the catalytic activity. These residues interact with glucose units bound to the -2, -1 and +1 subsites in the model obtained by molecular dynamics, respectively (Fig. 6B, C and D). Mutation of W210 to phenylalanine partially restores the activity (about 40% of activity when compared to *Zg*EngA$_{GH5_4WT}$), which supports the fact that this residue establishes van der Waals contacts with the glucose unit positioned in the +1 subsite. Surprisingly, mutations of Y280 and K211, which are thought to interact with glucose units bound to +2 and +3 subsites, respectively, did not decrease the activity.

Computed protein structure and molecular dynamics:

We used atomic resolution molecular dynamics computer simulations to model the binding of the full range of putative hexaose ligands (poorly resolved in the crystal structures) to wild type *Zg*EngA$_{GH5_4}$ and the mutant *Zg*EngA$_{GH5_4_E323S}$ endoglucanases. The ligand structures are described in Methods.

Both native and mutated structures show preservation of the protein secondary structure throughout the few-hundred nanosecond simulations (Figs. S5-S8 and Figs. S9-S12), even in cases where the glucan substrate leaves the binding pocket. The computed Root Mean Square Deviations (RMSD) of protein backbone non-hydrogen atoms in both mutant and native *Zg*EngA$_{GH5_4}$ (Figs. S13, S14) were within 0.2-0.3 nm, indicating a stable protein structure throughout the simulations. Calculated Root Mean Square Fluctuations (RMSF) (Figs. S15, S16) show the steric freedom of the more flexible and loose parts of the crystal structure such as turns and loops (residues 85-90, 125-126, 162-164 and 207-211).

Computed substrate dynamics – glucan in the binding site:

The -1/+1 (β-1,3) oligosaccharide with a β-1,3 linkage between units -1 and +1 (see Experimental) and the +1/+2 (β-1,3) glucan remain bound in 5 and 3 out of 8 repeats, respectively. The -1/+1 (β-1,3) and +1/+2 (β-1,3) glucans remain bound in 5 and 3 out of 8 repeats, respectively. Computed glucan RMSD values (Figs. S17-S18) show that -1/+1 (β-1,3) forms a stable binding interaction with both the native (RMSD 0.24 ± 0.04 nm) and mutated *Zg*EngA$_{GH5_4_E323S}$ (RMSD 0.26 ± 0.04 nm). The next most strongly bound ligand was +1/+2 (β-1,3), followed by +2/+3 (β-1,3). The (β-1,4) glucan either dissociates (4 out of 8 repeats) or else forms a loose complex with both the native and mutated enzyme (4 out of 8 repeats) with high glucan RMSD values of up to 0.44 ± 0.14 nm. In all other simulations, we observe either loose unstable binding of a substrate or dissociation into solution, and we did not include these dissociated structures in the analysis of binding energetics below.

Computed sugar – protein interactions:

The number of hydrogen bonds forming between the protein and substrate were monitored over time (Tables S6-S19) to identify protein residues contributing strongly to glucan binding (Figs. S19-S20). Eight hydrogen bonds (Tables S20-S21) stabilize the glucan in both native and mutated enzyme binding pockets. Namely, N77, E200, T253, H275, Y277, W356, N358, and E363.

The -1/+1 (β-1,3) glucan exhibits the most favorable affinity for both wildtype and mutated *Zg*EngA$_{GH5_4_E323S}$, as it stays strongly bound to the protein by 6 or more hydrogen bonds (Tables S20-S21) in 80% of the simulations. The computed MD structures in Fig. 6C and D show that T253 and

Y277 stabilize the sugar unit bound at the +3 subsite, W210 and E209 form H-bonds with +2, H275 and Y280 bind to +1, E200 binds to -1, and N358, W356, Y82 and N77 bind to the -2 subsite. Aromatic residues also contribute to carbohydrate recognition and orientation (Tables S22-S23 and Fig. 6C and D). Eight aromatic residues Y82, H155, W210, H275, Y277, Y280, W356 and F364 (Fig. 6C and D) interact with the substrate as it hydrogen bonds with adjacent polar and charged residues.

Computed binding energies (Table S24) show significantly stronger time-averaged substrate binding to mutated *Zg*EngA$_{GH5_4_E323S}$ than wildtype (-38.2 ± 10.6 kcal.mol^{-1} *vs*. -22.5 ± 8.5 kcal.mol^{-1}). The -1/+1 (β-1,3) glucan showed the strongest binding energy (-42.7 ± 9.0 kcal.mol^{-1}), consistent with its low RMSD (Figs. S17-S18) and extensive H-bonding (Figs. S19-S20). By contrast, (β-1,4) showed the weakest binding energies, reflecting its poor fit to the *Zg*EngA$_{GH5_4}$ active site pocket. In the most stable binding trajectories, the glucose chain is stabilized by H-bonding to approximately six polar and charged residues and makes close contacts with adjacent aromatic residues. The 'S-shaped' binding pocket better fits the natural conformation of the -1/+1(β-1,3)-linked glucan than the linear all (β-1,4) ligand.

Other insights from molecular modelling:

In the last frame of two simulations between *Zg*EngA$_{GH5_4_E323S}$ and GGGG3GG, the substrate is correctly positioned in the catalytic cleft. Then, the glucose in the +1 binding subsite interacts by stacking with W210. This interaction seems to be of high importance to position the substrate in such way to enable catalysis. The β-1,3 linkage induces a turn, which allows stacking interaction between glucose in +2 and Y280. However, this interaction seems to be more labile since it is present in only one model out of three. The glucose in +3 seems to have more degrees of liberty, and it establishes only weak contact with K211 and S252 (Fig. 6C).

In the simulations between *Zg*EngA$_{GH5_4_E323S}$ and GGG3GGG, the turn induced by the β-1,3 does not affect the stacking with W210, which is in a flexible loop. Indeed, in all simulations, it adapts its position to interact with glucose in +1. Y280 does not establish stacking contact with the glucose in +2 but it interacts with the glucose unit in +3 via a hydrogen bond (Fig. 6D)

Fixation of the whole β-1,4 substrate or GGGGG3G, seems to be weaker, as only W210 interacts with the substrate in the positive binding subsites.

DISCUSSION

The frequent classification of family GH5 enzymes as cellulases in marine *Flavobacteriia* [13, 16], despite the fact that these bacteria usually do not degrade crystalline cellulose [14, 15], is puzzling. Therefore, and in the context of recent work highlighting that GH5 enzymes belong to one of the largest, multi-specific glycoside hydrolase families [7, 10-12], covering a very large range of activities, we applied a combination of methods spanning phylogeny, enzymology, crystallography and molecular modeling to explore key enzyme-substrate interactions in *Zg*EngA$_{GH5_4}$ that define its substrate specificity. The comparison to other enzymes within GH5_4 reveals how substrate specificity is fine-tuned, even within the GH5_4 subfamily, and sheds further light on the roles of this subfamily in glucan catabolism. Taking the occurrence of this gene in a PUL that possibly is involved in the catabolism of hemicelluloses as starting point, we show that cellulose and soluble β-1,4-glucan-derivates are not the preferred substrates. Instead, we demonstrate that *Zg*EngA$_{GH5_4,}$ is a β-(1,3-1,4)-glucanase that preferably cleaves β-1,3 linkages flanked by β-1,4 linkages, but is also able to hydrolyze β-(1,4)-linkages in glucomannan, or in short oligosaccharides, depending on the linkage positions. The ability to hydrolyze β-(1,4) linkages in various substrates is common to EngD from *C. cellulovorans* (the closest structural relative of *Zg*EngA$_{GH5_4}$), which has been described as a true cellulase [38]. However, the relative activity of these two enzymes differs radically when using xyloglucan or CMC as substrates. In those cases, *Zg*EngA$_{GH5_4}$ is closer to other family GH5_4 members, such as PbGH5A from *Prevotella bryantii* and F32EG5 from *Caldicellulosiruptor* (Table 1).

Mapping the *Zg*EngA$_{GH5_4}$ active site by the combination of crystallographic structure determination together with molecular modeling and product analyses using different substrate oligosaccharides, suggests the presence of six well defined binding subsites, evenly distributed with respect to the cleavage point, three negative and three positive subsites. The crystal structure of the inactivated mutant highlights the binding subsites on the non-reducing end to which the cellotriose molecule (GGG) is bound. The molecular dynamic simulations using β-(1,3-1,4)-hexasaccharides that differ by the position of the β-1,3-linkage corroborate this biochemically observed preference, since

GGG3GGG, spanning the positions from -3 to +3 displayed the most favorable affinity for the catalytic cleft of *Zg*EngA$_{GH5_4}$. In this configuration, the β-1,3-linkage is positioned at the cleavage site, in accordance with the preferred hydrolytic activity of the enzyme on polysaccharide.

The biochemical analyses also revealed that the smallest hydrolyzed substrates are tetrasaccharides. More generally, the mode of hydrolysis of minimal substrates showed that activity is favored when oligosaccharides are spanning the cleft using the -2→+2 subsites, but hydrolysis does occur with modes spanning more subsites on the non-reducing end for GGG3G or on the reducing-end for G4. Notably, oligosaccharides containing β-1,4-linkages only, such as cellulo-oligosaccharides G4 to G6 are also hydrolyzed, however at a much slower rate than the preferred substrates. This is supported by molecular dynamics showing that cellulo-oligosaccharides were indeed much less stabilized in the active site cleft than the MLG oligosaccharides. In the case of G6, G4 is the first reaction product, meaning that for small oligosaccharides occupation of negative subsites predominates over positive ones. The hydrolysis of the mixed linkage oligosaccharides GGG3G and GG3GG also revealed the importance of the negative binding subsites in *Zg*EngA$_{GH5_4}$. GGG3G is mainly hydrolyzed according to the -3→+1 binding mode, demonstrating therefore that binding in subsite +2 is not essential for the hydrolysis of mixed linkage oligosaccharides. However, the absence of hydrolysis of GG3GG in the same mode suggests that *Zg*EngA$_{GH5_4}$ only tolerates β-1,4 bonds in the negative subsites, and that the presence of β-1,4 linkage in these positions is essential for hydrolysis of the neighboring β-1,3 linkage. In this respect, *Zg*EngA$_{GH5_4}$ is closer to cellulases [9, 37, 38]. Indeed, eight residues (Figs. 5 and 6) present in the negative binding subsites (N77, H155, H156, H275, Y277, W356 and N358) are well conserved throughout GH5_4 and typically bind to successive β-1,4-linked glucose units. This binding mode is completely different to that of family GH16 enzymes that cleave MLG (the so-called “lichenases”) that require a β-1,3 linkage in the negative subsites, a feature common to the β-glucanases *Zg*LamA$_{_GH16}$ and *Zg*LamC from *Zobellia galactanivorans* [39, 40]. Interestingly, these enzymes are also able to cleave both β-1,3 and β-1,4 linkages but, unlike *Zg*EngA$_{GH5_4}$, *Zg*LamA$_{_GH16}$ tolerate β-1,3 bonds in its negative subsites, rather than in the positive subsites.

The ability to cleave both β-1,3 and β-1,4 bonds has been previously described for GH5_4 enzymes [41] and it has been recently studied in light of 3D structures for PbGH5A [11], F32EG5 [10] and SdGluc5_26 [12]. Like *Zg*EngA$_{GH5_4}$, these enzymes have β-(1,3-1,4) glucanase activities 3 to 7 times higher than on CMC or cellulose. All of these enzymes require β-1,4-linkages between the -1 and -2 subsites and tolerate β-1,3-linkages in positive binding subsites. Nevertheless, subtle differences in accommodating the MLG substrate in the active site cleft can be noted between these enzymes. Although both *Zg*EngA$_{GH5_4}$ and F32EG5 [10] tolerate both β-1,3 and β-1,4 bonds at the +1 and +2 subsites, they differ from each other by the fact that *Zg*EngA$_{GH5_4}$ is unable to hydrolyze oligotrioses and has strict specificity for β-1,4-linkages between the -1 and -2 subsites, whereas F32EG5 only needs the -1 subsite to be occupied for activity [10]. The only structural difference between these enzymes in the negative binding sites consists in a loop that carries N358 and E363 (N362 and E370 in F32EG5 PDB ID: 4X0V; N367 and E375 in [10]) (Fig. 7A). In F32EG5, E370 interacts with O6 of the glucose-unit bound in the -1 subsite, while the different loop conformation in *Zg*EngA$_{GH5_4}$ positions this residue far too distant (more than 7 Å between E363-OE1 and O6 of the glucose unit bound in -1, making this interaction impossible (Fig. 7A). Indeed, the point mutant of E363 in our study did not have any effect on the enzymatic activity. Contrarily, this additional stabilization of a glucose unit bound to the -1 subsite in F32EG5 thus plausibly explains the major difference between these two enzymes. The need to bind several β-1,4-linked glucose units at negative subsites, in turn, is shared with PbGH5A and SdGluc5_26, although their sequence identities to *Zg*EngA$_{GH5_4}$ are lower (32% and 22% respectively) than to F32EG5 (41.5% sequence identity). Notably, both in PbGH5A and SdGluc5_26 the binding cleft displays a more open space beyond binding subsite -1 towards the non-reducing end, although the corresponding loops and residues are highly diverse in these three enzymes. Apparently less tight binding of the unit bound to -1 implies that more sites need to be occupied for substrate stabilization prior to cleavage.

Differences in loop arrangements are also present at the positive end of the active site cleft, even within the GH5_4 subfamily. In this respect, *Zg*EngA$_{GH5_4}$ has a uniquely featured loop between β6 and α6 (Fig. 7B) that influences the positioning of the substrate at the +1 and +2 binding sites. The

molecular dynamic simulations show that *Zg*EngA$_{GH5_4}$ seems to display a rather flexible binding mode in these sites, in agreement with the fact that the mutation of Y280, to our surprise, did not affect activity. Apparently the general difference of the loop structure in *Zg*EngA$_{GH5_4}$ is sufficient to shape the binding cleft such that a mixed linked chain with the β-1,3-linkage positioned at the -1→+1 cleavage site is favored. In addition, the results of the product analyses of small oligosaccharides also indicate that binding at the positive subsites +2 and +3 are not crucial for the enzymatic activity or substrate specificity. In this context, it is interesting to note that activity at the level of the MLG polysaccharide differs from that on small oligosaccharides, highlighting that although powerful and useful for dissecting subtle substrate specificities, biochemical *in vitro* product analyses of oligosaccharides artefactually show activities that might not be relevant under natural conditions. Indeed MLG polysaccharides appear to be hydrolyzed by *Zg*EngA$_{GH5_4}$ almost exclusively at the β-1,3-linkages (Fig. 4C). *Zg*EngA$_{GH5_4}$ also shows substantial activity on glucomannan as compared to CMC and no activity at all on laminarin-like substrates that only contain β-1,3-linkages. These results on polysaccharides seem to point towards the fact that the overall 3D structural conformation of the polymeric chain also plays an important role for substrate specificity, and the kinked polysaccharide chain of MLG (or a non-regular structure, as in glucomannan) is the preferred site of hydrolysis of this enzyme. Interestingly, the bent or kinked active site cleft has also been described to be an important feature of other GH5_4 members.

In summary, *in vitro Zg*EngA$_{GH5_4}$ appears to be most active on plant hemicellulose substrates, such as the polymers β-(1,3-1,4)-glucan and glucomannan, which raises the question of the functional rational behind this activity in the context of its ecologic and marine occurrence in *Z. galactanivorans*. While its evolutionary origin clearly points towards acquisition through lateral gene transfer from typical land-plant polysaccharide degrading bacteria, such as *C. cellulovorans*, the question remains whether the enzyme in the context of the physiology of *Zobellia galactanivorans* has 'specialized' for marine macro-algal cell wall components or if it remains specific of plant hemicelluloses. Arguments can be found for both scenarios: several macroalgal species of the red lineage have been reported to contain glucomannan as cell wall component [42], and mixed linkage glucans are reported in red and

brown algal species [4, 42]. Moreover, *engA* is found in a PUL like genetic context, next to hypothetical proteins that are indicative of a potential involvement in degradation of to date undescribed polysaccharide components. On the other hand, hemicellulosic polysaccharides, which strongly resemble those of land plants, can also be found in the marine environment in seagrasses, which could be the targeted natural substrate of this enzyme together with the adjacent PUL, for which the SusD-like protein was found to recognize xyloglucan. The elucidation of the biochemical activities and substrate specificities of the adjacent other components of the PUL-like structure may be the key to unravel the precise natural cell wall substrates that are targeted by these proteins.

ACCESSION NUMBERS

The atomic coordinates and the atomic factors of both *Zg*EngA$_{GH5_4}$ and *Zg*EngA$_{GH5_4_E323S}$ have been deposited at the Protein Data Bank collection (http://wwwpdb.org/) as PDB ID: 6GL2 and PDB ID: 6GL0 respectively. The modeling coordinates can be accessed through the following link: https://1drv.ms/f/s!ArX4zU6cjMUQnHQ9m5ScGJkyY_Kr

ACKNOWLEDGMENTS

We are deeply grateful to Alexandra Jeudy for technical assistance especially in the crystallization experiments of the wild-type enzyme. We thank the local contact support on the beamlines ID-14-4 and Proxima2 at the ESRF (Grenoble, France) and SOLEIL (Paris, France). JDo, SR, MG, AO, JDa and MC are grateful to the EU for its support with regards to the CellulosomePlus Program (FP7-NMP, project 604530). GM acknowledges support from the Agence Nationale de la Recherche (ANR) with regard to the "Blue Enzymes" project (Reference ANR-14-CE19-0020-01). GM and MC are also grateful to ANR for its support with regards to the investment expenditure program IDEALG (http://www.idealg.ueb.eu/, grant agreement No. ANR-10-BTBR-04). DT thanks Science Foundation Ireland (SFI) for support (Grant Number 15/CDA/3491) and for computing resources at the SFI/Higher Education Authority Irish Center for High-End Computing (ICHEC).

DECLARATION OF INTERESTS

The authors declare that there are no competing interests associated with the manuscript.

AUTHOR CONTRIBUTION STATEMENT

GM, MC and SG conceived the study; AL performed the protein expression experiments of the wild type enzyme. SR crystallized the wild type enzyme with ligand molecule. GM performed the bioinformatics and phylogenetic analysis. MCB performed some preliminary hydrolysis kinetic analysis of the wild type enzyme. SG produced and characterized the wild type enzyme. MC and JDa solved the structure of the wild type enzyme. SR produced the mutants. SR also produced the ligand she used for the crystallization of the mutant. JDo and SG performed the kinetic analysis of the mutants. SG also performed the HPAEC and FACE experiments. JDo determined the X-Ray structure of the mutants. MG, AO and DT performed the computer simulations. JDo, GM, MC and SG wrote

the manuscript with the help of AL and input from other co-authors. All authors approved the final version of the manuscript.

FIGURE LEGENDS

Figure 1: **Gene composition of the Polysaccharide Utilization Locus 4 (PUL_4) from *Zobellia galactanivorans*.** The gene encoding the *Zg*EngA$_{GH5_4}$ is colored in green; the other genes are colored in blue. Abbreviations: CBM4, family 4 of carbohydrate binding modules; TBDT, TonB-dependent transporter.

Figure 2: **Phylogenetic tree of *Zg*EngA$_{GH5_4}$ homologues.** The phylogenetic tree was derived using the maximum-likelihood approach with the program MEGA6 [19]. Numbers indicate the bootstrap values in the maximum likelihood analysis. The sequence marked by a brown diamond correspond to *Zg*EngA$_{GH5_4}$. The characterized enzymes are indicated by a black dot (biochemically characterized) or a black triangle (biochemically and structurally characterized). For these latter enzymes, the PDB code is indicated after the protein name. On the right, clades are delimited by brackets and their taxonomic affiliations are indicated. The sequences used are listed in supplementary Table S25.

Figure 3: **Substrate specificity of *Zg*EngA$_{GH5_4}$ studied by HPAEC**. Hydrolysis of cellotetraose (A), cellopentaose (B), cellohexaose (C), tetraose B (GGG3G; G4B) (D) and tetraose C (GG3GG; G4C) (E) from the mixed-linked glucan lineage with 0.5µM *Zg*EngA$_{GH5_4}$. Hydrolysis was performed as a function of time at 30°C. Aliquote of the reaction mixture were withdrawn periodically and analyzed by HPAEC-PAD on a CarboPAc-PA1 column. The oligosaccharides produced were identified and quantified via a standardization of the column performed with the different commercially available oligosaccharides used at different concentrations.

Figure 4: **Terminal products of *Zg*EngA$_{GH5_4}$ upon hydrolysis of standard oligocelluloses (A and B), β-D- Glucan (C) or its hydrolysis products (D) and schematic representation of the oligosaccharides accommodation in the active site (E and F).** In these experiments, incubations were performed overnight at 30°C using 1µL of *Zg*EngA$_{GH5_4}$ (100 nM) to hydrolyze 12.5 µg oligosaccharides (A, B and D) or 45 µg of β-D glucan from barley (C). Commercial lichenase was also used to completely hydrolyze β-D glucan from barley and to produce oligosaccharides that were then incubated for 10 minutes at 100°C prior to being labeled and further hydrolyzed overnight at

30°C with *Zg*EngA$_{GH5_4}$ (D). Commercial cello- and MLG- oligosaccharides were used as references. The ^{100} denotes oligosaccharides incubated with inactive *Zg*EngA$_{GH5_4}$ and the * indicates that the oligosaccharides were labelled before the enzymatic incubation with *Zg*EngA$_{GH5_4}$. (A and D).

(E and F): Schematic representation of cello- (E) and MLG- (F) oligosaccharides accommodation in the active site of *Zg*EngA$_{GH5_4.}$ The proposed cleavage sites deduced from HPAEC and/or FACE experiments are indicated by a grey arrow. The grey circles represent the reducing end of the oligosaccharides and the yellow circles represent the fluorophore used to label the reducing end sugar. The modes of hydrolysis observed with the FACE experiments exclusively are depicted with yellow circles and black outlines. When the mode of hydrolysis has been observed both by HPAEC and FACE, the yellow circles are outlined in grey. The arrow between the cellohexaose (G6) and the cellotetraose (G4) means that the hydrolysis product from the cellohexaose is further hydrolyzed into cellobiose (G2). Cellobiose and cellotriose are not represented as they are not hydrolyzed by *Zg*EngA$_{GH5_4.}$ (G) Hydrolysis sites of MLG by *Zg*EngA$_{GH5_4}$ deduced from FACE experiments (see above for details). The proposed cleavage sites are indicated by grey arrows. In a comparative purpose, the GH16 lichenase cleavage sites are indicated by dotted arrows.

Figure 5: Sequence alignment of *Zg*EngA$_{GH5_4}$ with structurally characterized GH5_4. The sequence alignment has been performed using MAFFT [18] and has been manually edited in Bioedit (©Tom Hall) based on the superimposition of the different crystal structures. The final figure has been created with using ESPript [43]. The sequences used in this alignment were as follows: CcEngD: the endo β-1,4-glucanase/xylanase EngD from *Clostridium cellulovorans* (GenBank accession no. **AAA23233.1**; residues 32-376; PDB ID: 3NDY); PbGH5A: the Mixed-linkage beta-Glucanase/Xyloglucanase from *Prevotella bryantii* B14 (GenBank accession no. **AAC97596.1**, residues 584-924, PDB ID: 3VDH); F32EG5: the β-(1.3-1.4) glucanase from *Caldicellulosiruptor* sp. (GenBank accession no. **AGM71677.1**, residues 38-401, PDB ID: 4XOV), BpCel5C: Cel5C from *Butyvibrio proteoclasticus* (GenBank accession no. **ADL34447.1**, residues 32-399, PDB code: 4NF7) and CcCel5A: the cellulase Cel5A from *Clostridum cellulolyticum* (GenBank accession no. **AAA23221.1**; residues 40-403, PDB ID: 1EDG). The α- and 3_{10} helices and the β-strands are represented as helices and arrows,

respectively, and β-turns are marked with TT. Dark shaded boxes enclose invariant positions, and light shaded boxes show positions with similar residues. The catalytic residues and the residues chosen for site-directed mutagenesis are marked by red triangles and blue dots, respectively.

Figure 6: Crystal structure of *Zg*EngA$_{GH5_4}$ and the relative MLG substrate locations after molecular modeling in all-atom simulations. (A) Crystal structure of *Zg*EngA$_{GH5_4}$. The central β-sheet constituting the TIM barrel is shown in yellow, the additional α helix α0 in pale blue and the loop between β-strand β6 and α-helix 6 in red. Both catalytic residues are shown in sticks. (B) View of the active site of *Zg*EngA$_{GH5_4_E323S}$. The cellotriose is shown in green and the position of the two catalytic residues are shown in grey. The experimental electron density calculated as an 2Fo-Fc map and contoured at a 2 σ level is shown as grey mesh. The stereochemistry of the substrate molecule has been validated using Privateer and the details are given in Table S26. (C) Resulting view of molecular dynamics with GGGG3GG. The active site of *Zg*EngA$_{GH5_4_E323S}$ in the last frame of the all-atom simulation and the relative position of the GGGG3GG (in orange) substrate molecule are represented. The positions of the different sub-binding sites are indicated. The arrow indicates the β-1-3 linkage. (D) Resulting view of molecular dynamics with GGG3GGG. The active site of *Zg*EngA$_{GH5_4_E323S}$ in the last frame of the all-atom simulation and the relative position of the GGG3GGG (in orange) substrate molecule are represented. The arrow indicates the β-1-3 linkage.

Figure 7: Superimposition of GH5_4 active sites. (A) Superimposition of the catalytic active sites of *Zg*EngA$_{GH5_4}$ (in yellow) with those of F32EG5 (PDB ID:4XOV in dark blue), of CcEngD in complex with cellotriose (PDB ID:3NDZ in cyan, and cellotriose in green) and that of PbGH5A (PDB ID:3VDH in light grey). The highly conserved residues surrounding the -1 binding subsite are shown, highlighting the two major features that are different in the sugar binding subsites of *Zg*EngA$_{GH5_4}$, namely E363 and Y280. (B) Superimposition of the structure of *Zg*EngA$_{GH5_4_E323S}$ (in blue) and the structure of F32EG5 from *Caldicellulosiruptor sp*. F32 (in purple) showing the conformation of the loop harboring Y280 between β-strand β6 and α-helix 6. The GGG cellotriose molecule in the crystal structure of *Zg*EngA$_{GH5_4_E323S}$ is shown in green. The residue numbers for *Zg*EngA$_{GH5_4_E323S}$ are underlined.

1

REFERENCES

1 Popper, Z. A., Michel, G., Herve, C., Domozych, D. S., Willats, W. G., Tuohy, M. G., et al. (2011) Evolution and diversity of plant cell walls: from algae to flowering plants. *Annu. Rev. Plant Biol.* **62**, 567-590

2 Sørensen, I., Pettolino, F. A., Wilson, S. M., Doblin, M. S., Johansen, B., Bacic, A., et al. (2008) Mixed-linkage (1-->3),(1-->4)-beta-D-glucan is not unique to the Poales and is an abundant component of *Equisetum arvense* cell walls. *Plant J.* **54**, 510-521

3 Eder, M., Tenhaken, R., Driouich, A. and Lutz-Meindl, U. (2008) Occurrence and Characterization of Arabinogalactan-Like Proteins and Hemicelluloses in *Micrasterias* (*Streptophyta*)(1). *J.Phycol.* **44**, 1221-1234

4 Salmean, A. A., Duffieux, D., Harholt, J., Qin, F., Michel, G., Czjzek, M., et al. (2017) Insoluble (1 --> 3), (1 --> 4)-beta-D-glucan is a component of cell walls in brown algae (*Phaeophyceae*) and is masked by alginates in tissues. *Sci. Rep.* **7**, 2880

5 Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. and Henrissat, B. (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490-495

6 Planas, A. (2000) Bacterial 1,3-1,4-beta-glucanases: structure, function and protein engineering. *Biochim. Biophys. Acta*. **1543**, 361-382

7 Aspeborg, H., Coutinho, P. M., Wang, Y., Brumer, H., 3rd and Henrissat, B. (2012) Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol. Biol.* **12**, 186

8 Davies, G. J. and Henrissat, B. (1995) Structures and mechanisms of glycosyl hydrolases. *Structure*. **3**, 853-859

9 Ducros, V., Czjzek, M., Belaich, A., Gaudin, C., Fierobe, H. P., Belaich, J. P., et al. (1995) Crystal structure of the catalytic domain of a bacterial cellulase belonging to family 5. *Structure*. **3**, 939-949

10 Meng, D. D., Liu, X., Dong, S., Wang, Y. F., Ma, X. Q., Zhou, H., et al. (2017) Structural insights into the substrate specificity of a glycoside hydrolase family 5 lichenase from *Caldicellulosiruptor* sp. F32. *Biochem. J.* **474**, 3373-3389

11 McGregor, N., Morar, M., Fenger, T. H., Stogios, P., Lenfant, N., Yin, V., et al. (2016) Structure-Function Analysis of a Mixed-linkage beta-Glucanase/Xyloglucanase from the Key Ruminal Bacteroidetes *Prevotella bryantii* B(1)4. *J. Biol. Chem.* **291**, 1175-1197

12 Lafond, M., Sulzenbacher, G., Freyd, T., Henrissat, B., Berrin, J. G. and Garron, M. L. (2016) The Quaternary Structure of a Glycoside Hydrolase Dictates Specificity toward beta-Glucans. *J. Biol. Chem.* **291**, 7183-7194

13 Barbeyron, T., Thomas, F., Barbe, V., Teeling, H., Schenowitz, C., Dossat, C., et al. (2016) Habitat and taxon as driving forces of carbohydrate catabolism in marine heterotrophic bacteria: example of the model algae-associated bacterium *Zobellia galactanivorans* DsijT. *Environ. Microbiol.* **18**, 4610-4627

14 Bernardet, J. F., Segers, P., Vancanneyt, M., Berthe, F., Kersters, K. and Vandamme, P. (1996) Cutting a Gordian knot: emended classification and description of the genus *Flavobacterium*, emended description of the family *Flavobacteriaceae*, and proposal of *Flavobacterium hydatis* nom. nov. (basonym, *Cytophaga aquatilis* Strohl and Tait 1978). *Int. J. Syst. Bacteriol.* **46**, 128-148

15 Thomas, F., Hehemann, J. H., Rebuffet, E., Czjzek, M. and Michel, G. (2011) Environmental and gut bacteroidetes: the food connection. *Front. Microbiol.* **2**, 93

16 Elifantz, H., Waidner, L. A., Michelou, V. K., Cottrell, M. T. and Kirchman, D. L. (2008) Diversity and abundance of glycosyl hydrolase family 5 in the North Atlantic Ocean. *FEMS Microbiol. Ecol.* **63**, 316-327

17 Foong, F. C. and Doi, R. H. (1992) Characterization and comparison of *Clostridium cellulovorans* endoglucanases-xylanases EngB and EngD hyperexpressed in *Escherichia coli*. *J. Bacteriol.* **174**, 1403-1409

18 Katoh, K. and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780

19 Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725-2729

20 Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H. and Krogh, A. (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* **12**, 1652-1662

21 Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J. P. (1987) Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. *FEBS Lett.* **224**, 149-155

22 Barbeyron, T., Kean, K. and Forterre, P. (1984) DNA adenine methylation of GATC sequences appeared recently in the *Escherichia coli* lineage. *J. Bacteriol.* **160**, 586-590

23 Groisillier, A., Herve, C., Jeudy, A., Rebuffet, E., Pluchon, P. F., Chevolot, Y., et al. (2010) MARINE-EXPRESS: taking advantage of high throughput cloning and expression strategies for the post-genomic analysis of marine organisms. *Microb. Cell Fact.* **9**, 45

24 Studier, F. W. (2005) Protein production by auto-induction in high density shaking cultures. *Protein. Expr. Purif.* **41**, 207-234

25 Kidby, D. K. and Davidson, D. J. (1973) A convenient ferricyanide estimation of reducing sugars in the nanomole range. *Anal. Biochem.* **55**, 321-325

26 Ostling, S. and Virtama, P. (1946) A modified preparation of the universal buffer described by Teorell and Stenhagen. *Acta Phys. Scandinav.* **11**, 289- 293

27 Kabsch, W. (2010) Xds. *Acta Crystallogr. D*. **66**, 125-132

28 Vagin, A. and Teplyakov, A. (1997) MOLREP: an Automated Program for Molecular Replacement. *J. Appl. Crystallogr.* **30**, 1022-1025

29 Perrakis, A., Sixma, T. K., Wilson, K. S. and Lamzin, V. S. (1997) wARP: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. *Acta Crystallogr. D*. **53**, 448-455

30 Emsley, P., Lohkamp, B., Scott, W. G. and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr. D*. **66**, 486-501

31 Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F., et al. (2004) REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. D*. **60**, 2184-2195

32 McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. and Read, R. J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658-674

33 Thomas, F., Bordron, P., Eveillard, D. and Michel, G. (2017) Gene Expression Analysis of *Zobellia galactanivorans* during the Degradation of Algal Polysaccharides Reveals both Substrate-Specific and Shared Transcriptome-Wide Responses. *Front. Microbiol.* **8**, 1808

34 Salmeán, A. A., Guillouzo, A., Duffieux, D., Jam, M., Matard-Mann, M., Larocque, R., et al. (2018) Double blind microarray-based polysaccharide profiling enables parallel identification of uncharacterized polysaccharides and carbohydrate-binding proteins with unknown specificities. *Sci. Rep.* **8**, 2500

35 Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., et al. (2005) Protein identification and analysis tools on the ExPASy Server. In The Proteomics Protocols Handbook (Walker, J. M., ed.). pp. 571-607, Totowa, New Jersey, USA

36 Holm, L. and Laakso, L. M. (2016) Dali server update. *Nucleic Acids Res.* **44**, W351-355

37 Bortoli-German, I., Haiech, J., Chippaux, M. and Barras, F. (1995) Informational suppression to investigate structural functional and evolutionary aspects of the *Erwinia chrysanthemi* cellulase EGZ. *J. Mol. Biol.* **246**, 82-94

38 Bianchetti, C. M., Brumm, P., Smith, R. W., Dyer, K., Hura, G. L., Rutkoski, T. J., et al. (2013) Structure, dynamics, and specificity of endoglucanase D from *Clostridium cellulovorans*. *J. Mol. Biol.* **425**, 4267-4285

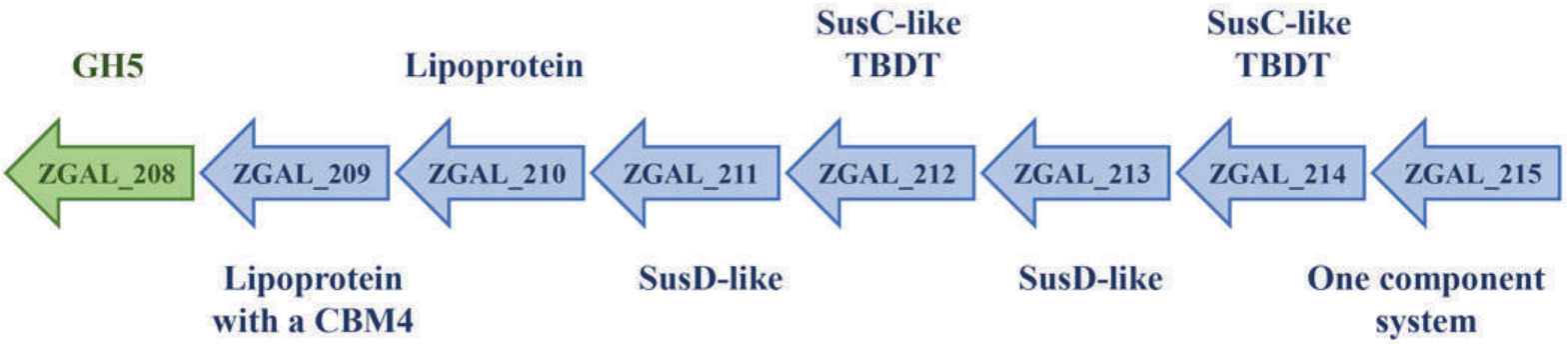
39 Labourel, A., Jam, M., Jeudy, A., Hehemann, J. H., Czjzek, M. and Michel, G. (2014) The β-glucanase *Zg*LamA from *Zobellia galactanivorans* evolved a bent active site Adapted for efficient degradation of algal laminarin. *J. Biol. Chem.* **289**, 2027–2042

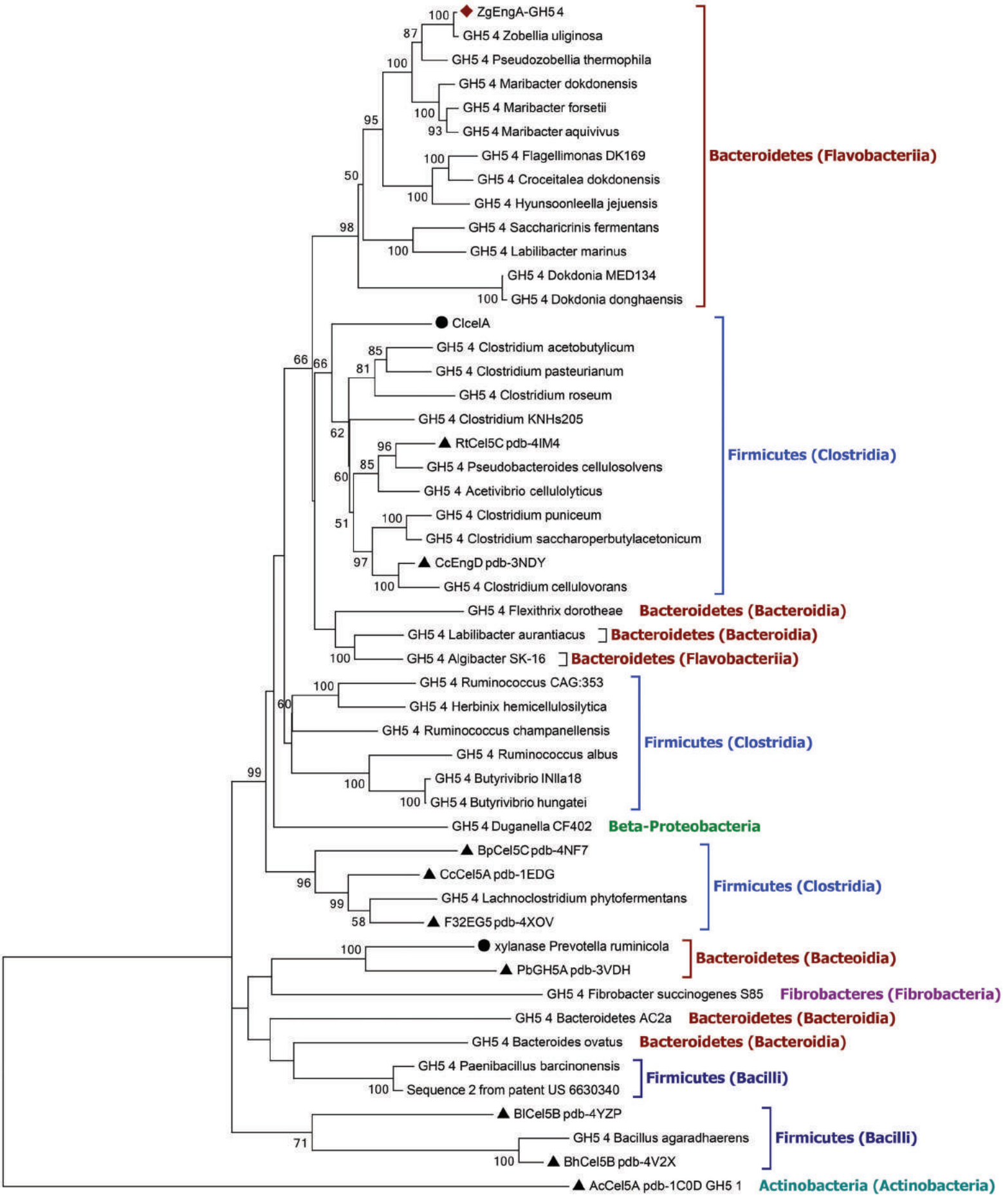
40 Labourel, A., Jam, M., Legentil, L., Sylla, B., Hehemann, J. H., Ferrieres, V., et al. (2015) Structural and biochemical characterization of the laminarinase *Zg*LamC$_{GH16}$ from *Zobellia galactanivorans* suggests preferred recognition of branched laminarin. *Acta Crystallogr. D*. **71**, 173-184

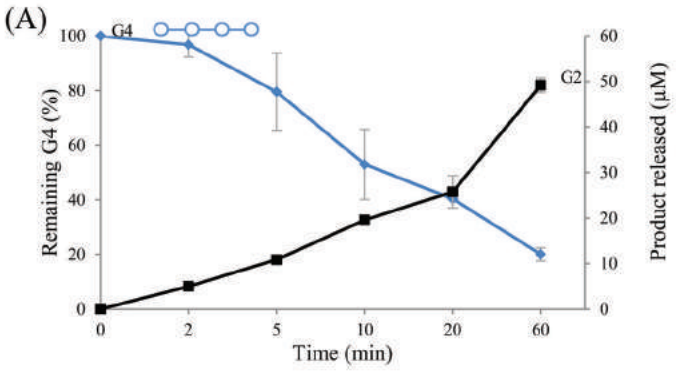
41 Iakiviak, M., Mackie, R. I. and Cann, I. K. (2011) Functional analyses of multiple lichenin-degrading enzymes from the rumen bacterium *Ruminococcus albus* 8. *Appl. Environ. Microbiol.* **77**, 7541-7550

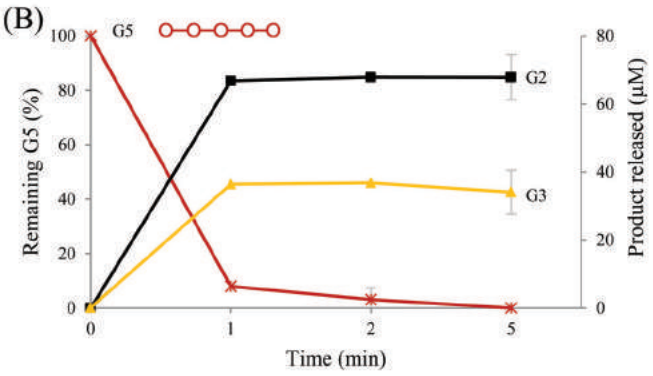
42 Lechat, H., Amat, M., Mazoyer, J., Buléon, A. and Lahaye, M. (2000) Structure and distribution of glucomannan and sulfated glucan in the cell walls of the red alga *Kappaphycus alvarezii* (*Gigartinales*, *Rhodophyta*). *J.Phycol.* **36**, 891-902

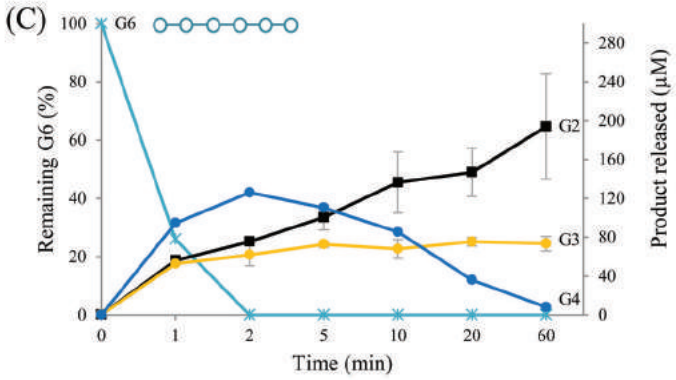
43 Robert, X. and Gouet, P. (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320-324

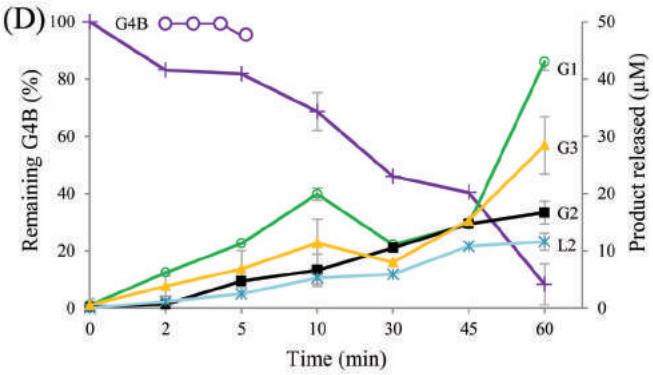

GH5
Lipoprotein
SusC-like
TBDT
SusC-like
TBDT
ZGAL_208
ZGAL_209
ZGAL_210
ZGAL_211
ZGAL_212
ZGAL_213
ZGAL_214
ZGAL_215
Lipoprotein
with a CBM4
SusD-like
SusD-like
One component
system

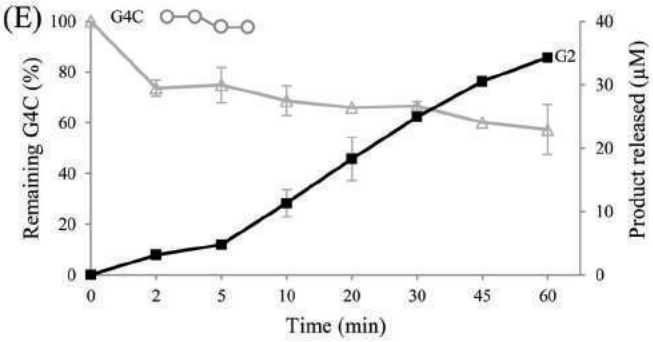

ZgEngA-GH5 4
GH5 4 Zobellia uliginosa
GH5 4 Pseudozobellia thermophila
GH5 4 Maribacter dokdonensis
GH5 4 Maribacter forsetii
GH5 4 Maribacter aquivivus
GH5 4 Flagellimonas DK169
GH5 4 Croceitalea dokdonensis
GH5 4 Hyunsoonleella jejuensis
GH5 4 Saccharicrinis fermentans
GH5 4 Labilibacter marinus
GH5 4 Dokdonia MED134
GH5 4 Dokdonia donghaensis
Bacteroidetes (Flavobacteriia)
ClCelA
GH5 4 Clostridium acetobutylicum
GH5 4 Clostridium pasteurianum
GH5 4 Clostridium roseum
GH5 4 Clostridium KNHs205
RtCel5C pdb-4IM4
GH5 4 Pseudobacteroides cellulosolvens
GH5 4 Acetivibrio cellulolyticus
GH5 4 Clostridium puniceum
GH5 4 Clostridium saccharoperbutylacetonicum
CcEngD pdb-3NDY
GH5 4 Clostridium cellulovorans
Firmicutes (Clostridia)
GH5 4 Flexithrix dorotheae
Bacteroidetes (Bacteroidia)
GH5 4 Labilibacter aurantiacus
Bacteroidetes (Bacteroidia)
GH5 4 Algibacter SK-16
Bacteroidetes (Flavobacteriia)
GH5 4 Ruminococcus CAG:353
GH5 4 Herbinix hemicellulosilytica
GH5 4 Ruminococcus champanellensis
GH5 4 Ruminococcus albus
GH5 4 Butyrivibrio INIla18
GH5 4 Butyrivibrio hungatei
Firmicutes (Clostridia)
GH5 4 Duganella CF402
Beta-Proteobacteria
BpCel5C pdb-4NF7
CcCel5A pdb-1EDG
GH5 4 Lachnoclostridium phytofermentans
F32EG5 pdb-4XOV
Firmicutes (Clostridia)
xylanase Prevotella ruminicola
PbGH5A pdb-3VDH
Bacteroidetes (Bacteoidia)
GH5 4 Fibrobacter succinogenes S85
Fibrobacteres (Fibrobacteria)
GH5 4 Bacteroidetes AC2a
Bacteroidetes (Bacteroidia)
GH5 4 Bacteroides ovatus
Bacteroidetes (Bacteroidia)
GH5 4 Paenibacillus barcinonensis
Sequence 2 from patent US 6630340
Firmicutes (Bacilli)
BlCel5B pdb-4YZP
GH5 4 Bacillus agaradhaerens
BhCel5B pdb-4V2X
Firmicutes (Bacilli)
AcCel5A pdb-1C0D GH5 1
Actinobacteria (Actinobacteria)


(A)
G4
G2
Remaining G4 (%)
Product released (μM)
Time (min)

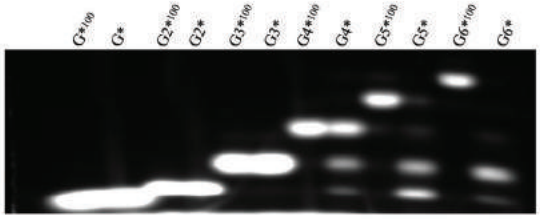

(B)
G5
G2
G3
Remaining G5 (%)
Product released (μM)
Time (min)


(C)
G6
G2
G3
G4
Remaining G6 (%)
Product released (μM)
Time (min)

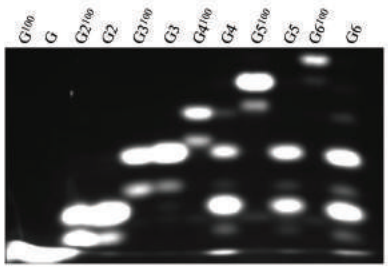

(D)
G4B
G1
G3
G2
L2
Remaining G4B (%)
Product released (μM)
Time (min)


(E)
G4C
G2
Remaining G4C (%)
Product released (μM)
Time (min)

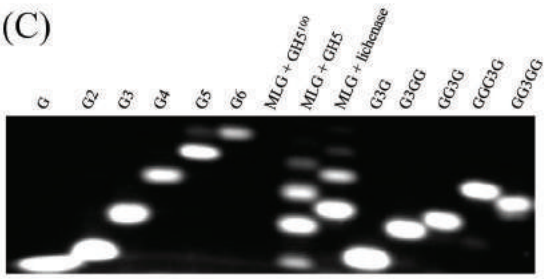
(A)


G*100
G*
G2*100
G2*
G3*100
G3*
G4*100
G4*
G5*100
G5*
G6*100
G6*

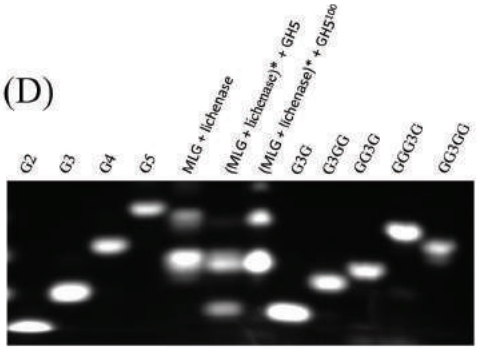
(B)


G100
G
G2100
G2
G3100
G3
G4100
G4
G5100
G5
G6100
G6

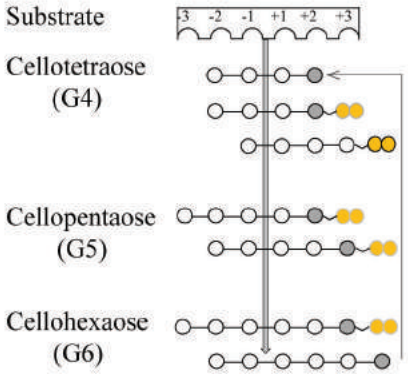
(C)


G
G2
G3
G4
G5
G6
MLG + GH5100
MLG + GH5
MLG + lichenase
G3G
G3GG
GG3G
GGG3G
GG3GG

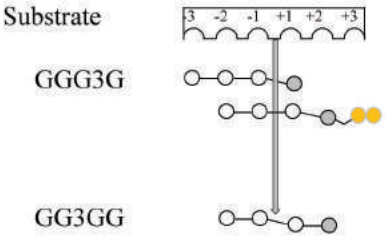
(D)


G2
G3
G4
G5
MLG + lichenase
(MLG + lichenase)* + GH5
(MLG + lichenase)* + GH5100
G3G
G3GG
GG3G
GGG3G
GG3GG

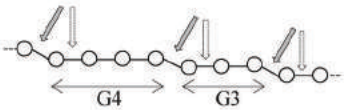
(E)

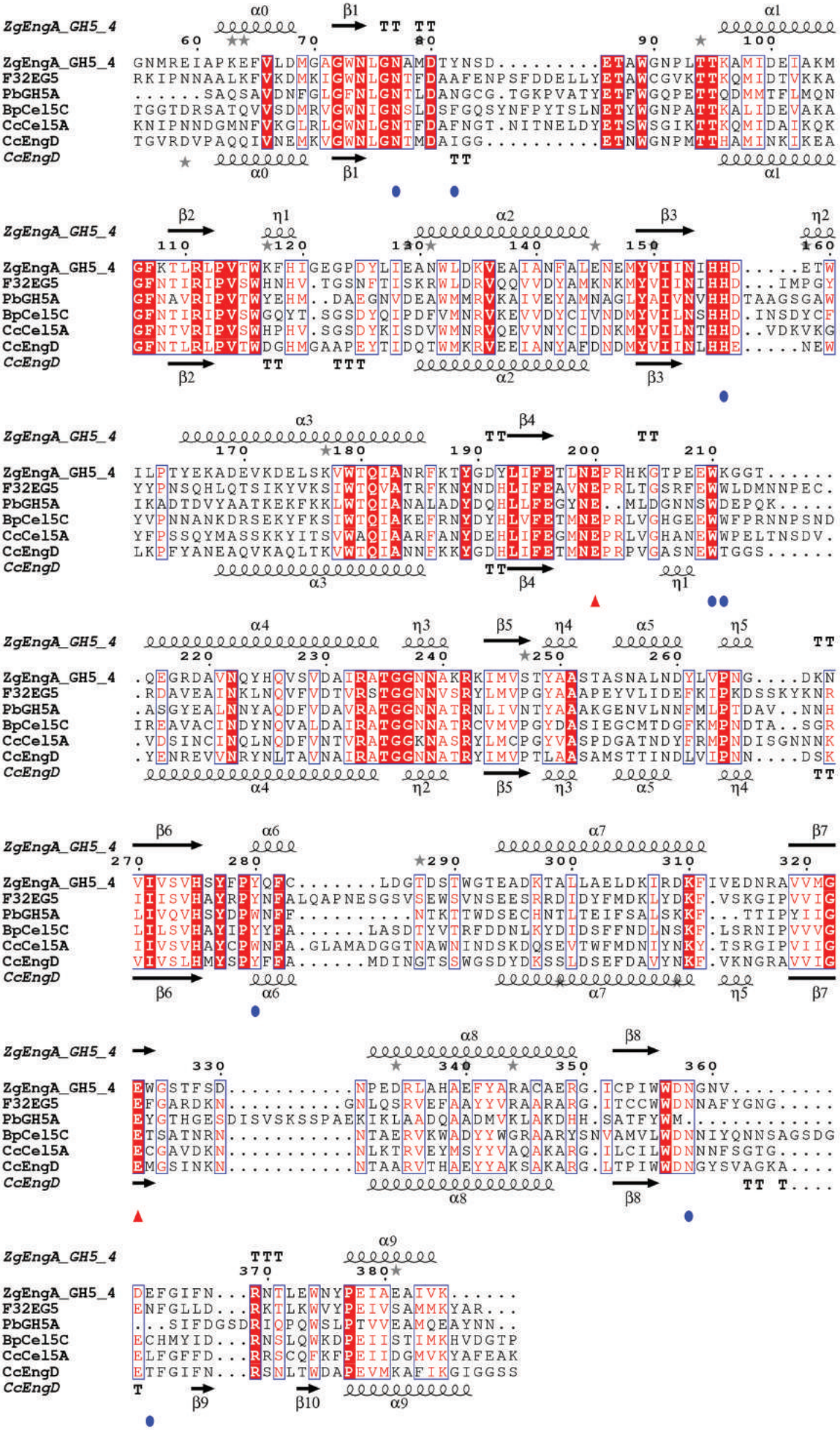

Substrate
-3
-2
-1
+1
+2
+3
Cellotetraose
(G4)
Cellopentaose
(G5)
Cellohexaose
(G6)

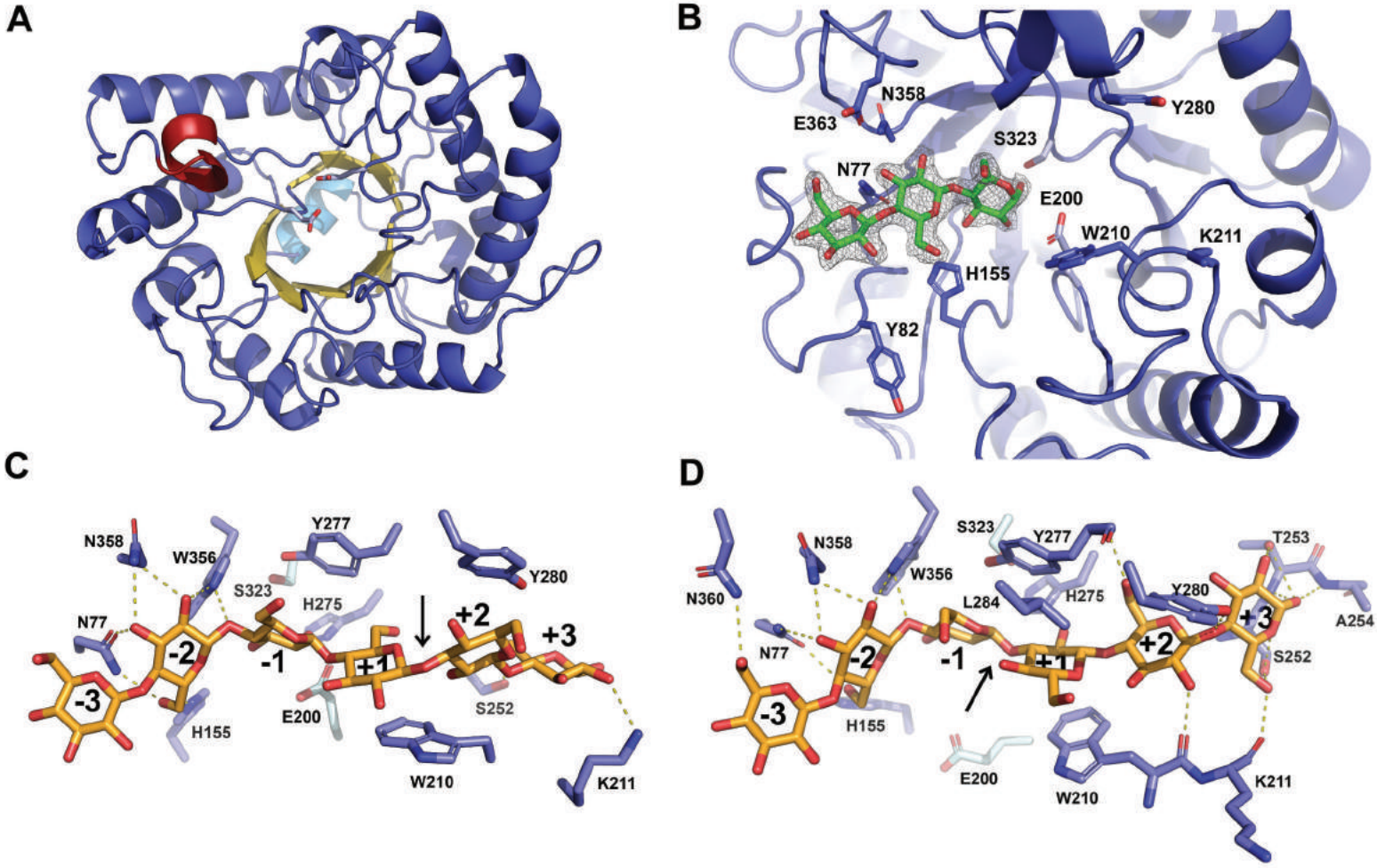
(F)

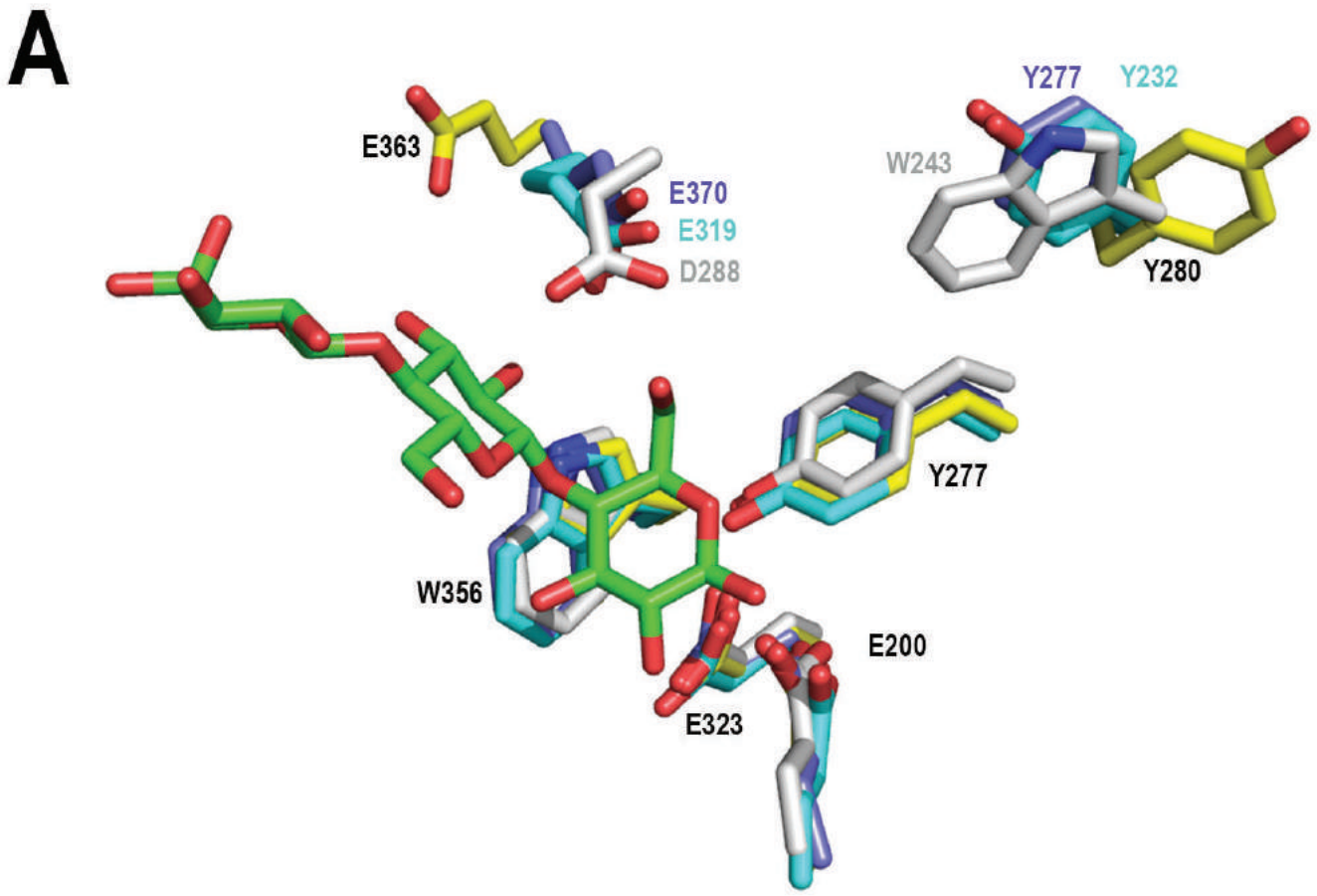

Substrate
-3
-2
-1
+1
+2
+3
GGG3G
GG3GG

(G)


MLG
G4
G3


ZgEngA_GH5_4
ZgEngA_GH5_4
F32EG5
PbGH5A
BpCel5C
CcCel5A
CcEngD
CcEngD
GNMREIAPKEFVLDMGAGWNLGNAMDTYNSD.....ETAWGNPLTTKAMIDEIAKM
RKIPNNAALKFVKDMKIGWNLGNTFDAAFENPSFDDELLYETAWCGVKTTKQMIDTVKKA
.....SAQSAVDNFGLGFNLGNTLDANGCG.TGKPVATYETFWGQPETTQDMMTFLMQN
TGGTDRSATQVVSDMRVGWNIGNSLDSFGQSYNFPYTSLNETYWGNPATTKALIDEVAKA
KNIPNNDGMNFVKGLRLGWNLGNTFDAFNGT.NITNELDYETSWSGIKTTKQMIDAIKQK
TGVRDVPAQQIVNEMKVGWNLGNTMDAIGG.....ETNWGNPMTTHAMINKIKEA
GFKTLRLPVTWKFHIGEGPDYLIEANWLDKVEAIANFALENEMYVIINIHHD.....ETW
GFNTIRIPVSWHNHV.TGSNFTISKRWLDRVQQVVDYAMKNKMYVIINIHHD...IMPGY
GFNAVRIPVTWYEHM..DAEGNVDEAWMMRVKAIVEYAMNAGLYAIVNVHHDTAAGSGAW
GFNTIRIPVSWGQYT.SGSDYQIPDFVMNRVKEVVDYCIVNDMYVILNSHHD.INSDYCF
GFNTVRIPVSWHPHV.SGSDYKISDVWMNRVQEVVNYCIDNKMYVILNTHHD..VDKVKG
GFNTLRLPVTWDGHMGAAPEYTIDQTWMKRVEEIANYAFDNDMYVIINLHHE.....NEW
ILPTYEKADEVKDELSKVWTQIANRFKTYGDYLIFETLNEPRHKGTPEEWKGGT......
YYPNSQHLQTSIKYVKSIWTQVATRFKNYNDHLIFEAVNEPRLTGSRFEWWLDMNNPEC.
IKADTDVYAATKEKFKKLWTQIANALADYDQHLLFEGYNE..MLDGNNSWDEPQK.....
YVPNNANKDRSEKYFKSIWTQIAKEFRNYDYHLVFETMNEPRLVGHGEEWWFPRNNPSND
YFPSSQYMASSKKYITSVWAQIAARFANYDEHLIFEGMNEPRLVGHANEWWPELTNSDV.
LKPFYANEAQVKAQLTKVWTQIANNFKKYGDHLIFETMNEPRPVGASNEWTGGS......
.QEGRDAVNQYHQVSVDAIRATGGNNAKRKIMVSTYAASTASNALNDYLVPNG....DKN
.RDAVEAINKLNQVFVFDTVRSTGGNNVSRYLMVPGYAAAPEYVLIDEFKIPKDSSKYKNR
.ASGYEALNNYAQDFVDAVRATGGNNATRNLIVNTYAAAKGENVLNNFMLPTDAV..NNH
IREAVACINDYNQVALDAIRATGGNNATRCVMVPGYDASIEGCMTDGFKMPNDTA..SGR
.VDSINCINQLNQDFVNTVRATGGKNASRYLMCPGYVASPDGATNDYFRMPNDISGNNNK
.YENREVVNRYNLTAVNAIRATGGNNATRYIMVPTLAAASAMSTTINDLVIPNN....DSK
VIVSVHSYFPYQFC.....LDGTDSTWGTEADKTALLAELDKIRDKFIVEDNRAVVMG
IIISVHAYRPYNFALQAPNESGSVSEWSVNSEESRRDIDYFMDKLYDKF.VSKGIPVVIG
LIVQVHSYDPWNFF.....NTKTTWDSECHNTLTEIFSALSKKF...TTIPYIIG
LILSVHAYIPYYFA.....LASDTYVTRFDDNLKYDIDSFFNDLNSKF.LSRNIPVVVG
IIVSVHAYCPWNFA.GLAMADGGTNAWNINDSKDQSEVTWFMDNIYNKY.TSRGIPVIIG
VIVSLHMYSPYFFA.....MDINGTSSWGSDYDKSSLDSEFDAVYNKF.VKNGRAVVIG
EWGSTFSD.....NPEDRLAHAEFYARACAERG.ICPIWWDNGNV.....
EFGARDKN.....GNLQSRVEFAAYYVRAARARG.ITCCWWDNNAFYGNG.....
EYGTHGESDISVSKSSPAEKIKLAADQAADMVKLAKDHH.SATFYWM.....
ETSATNRN.....NTAERVKWADYYWGRAARYSNVAMVLWDNNIYQNNSAGSDG
ECGAVDKN.....NLKTRVEYMSYYVAQAKARG.ILCILWDNNNFSGTG.....
EMGSINKN.....NTAARVTHAEYYAKSAKARG.LTPIWWDNGYSVAGKA....
DEFGIFN...RNTLEWNYPEIAEAIVK.....
ENFGLLD...RKTLKWVYPEIVSAMMKYAR...
...SIFDGSDRIQPQWSLPTVVEAMQEAYNN..
ECHMYID...RNSLQWKDPEIISTIMKHVDGTP
ELFGFFD...RRSCQFKFPEIIDGMVKYAFEAK
ETFGIFN...RSNLTWDAPEVMKAFIKGIGGSS


A
B
N358
E363
Y280
S323
N77
E200
W210
K211
H155
Y82
C
D
N358
W356
Y277
Y280
S323
H275
N77
-2
-1
+1
+2
+3
-3
H155
E200
S252
W210
K211
N360
L284
T253
A254


A
E363
E370
E319
D288
Y277
Y232
W243
Y280
Y277
W356
E200
E323

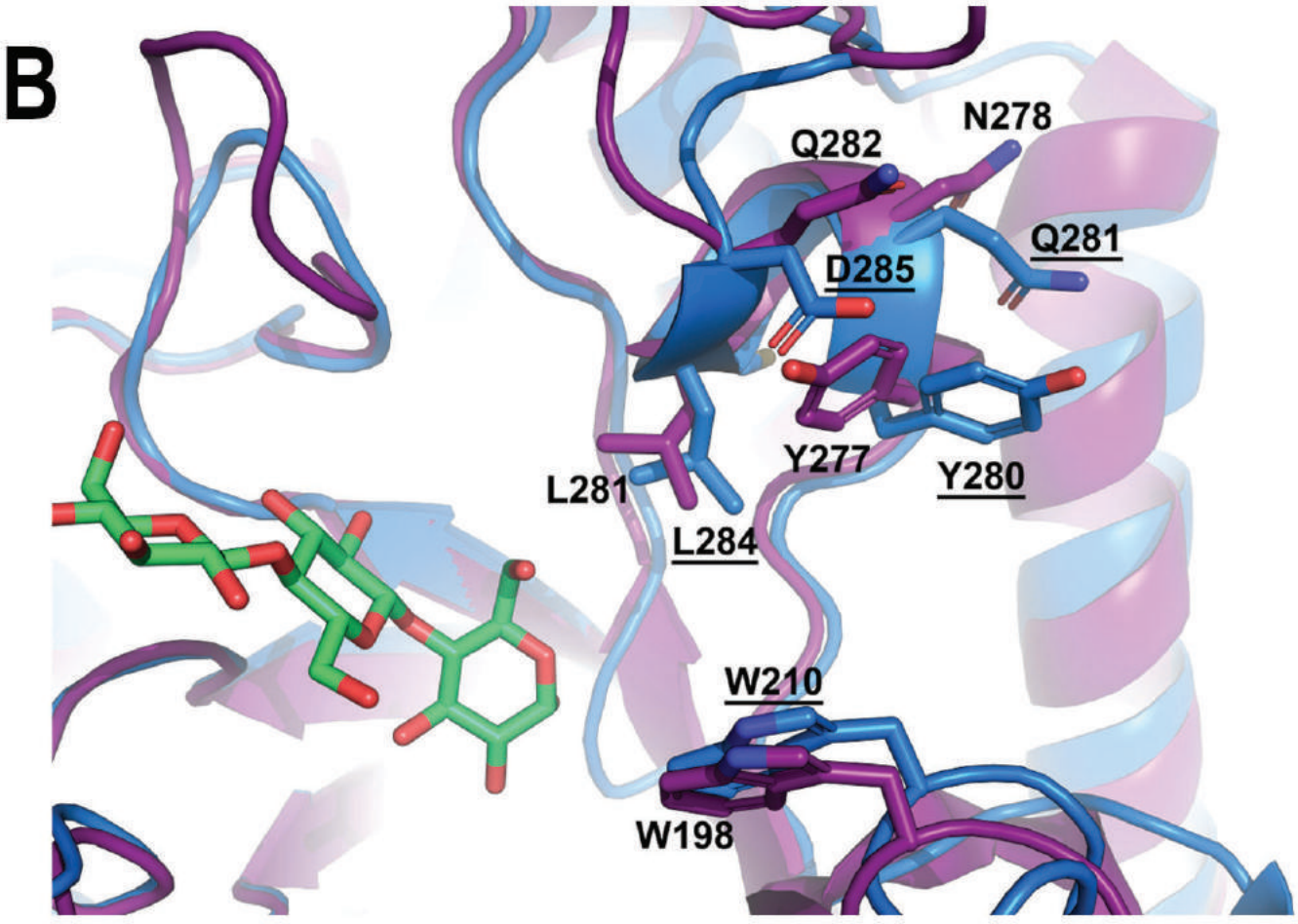

B
N278
Q282
Q281
D285
Y277
Y280
L281
L284
W210
W198

Table 1. Activity of *Zg*EngA$_{GH5_4}$ on different polysaccharides.

For comparative purpose, the relative activity of *Zg*EngA$_{GH5_4}$ is compared to the relative activity of the commercial lichenase (Megazyme) and 3 GH5_4 enzymes characterized at the 3D structure level. Results for EngD *C. cellulovorans* [38], *Caldicellulosiruptor* sp. F32 [10] and PbGH5 from *Prevotella bryantii* [11] were calculated based on published enzymatic activities.

Substrate	***Zg*EngA$_{GH5_4}$**		**Lichenase GH16**	**EngD**	**F32EG5**	**PbGH5A**
	Specific activity (μkat mg^{-1})[1]	**Relative activity (%)**	**Relative activity (%)**	**Relative activity (%)**	**Relative activity (%)**	**Relative activity (%)**
Mixed linkage glucan	5.1±0.3	100	100	100	100	100
Lichenan	1.5±0.1	29.4	37.3		52.2	
Glucomannan	3.1±0.0	60.8	n.d.	73.8		9.2
Xyloglucan	0.3±0.0	5.9	n.d.	85.7		19.5
CMC	0.2±0.0	3.9	n.d.	35.7	17.3	5.5
Avicel	n.d.	n.d.	n.d.		< 0.02	
Laminarin	n.d.	n.d.	n.d.		N.D.[3]	
Curdlan	n.d.	n.d.	n.d.			

[1] Experiments were performed in triplicate. Results are expressed as average ±S.D.
[2] n.d., not determinable, less than the limit of detection i.e. 0.01 (μkat mg^{-1})
[3] N.D., activity not detected according to mentioned reference

Table 2. Data collection and refinement statistics

	*Zg*EngA$_{GH5_4}$	*Zg*EngA$_{GH5_4_E323S}$
Data collection		
Space group	$P2_1$	$P3_2$
Cell dimensions		
a, *b*, *c* (Å)	55.49, 48.46, 59.26	84.57, 84.57, 117.66
α, β, γ (°)	90.00, 104.35, 90.00	90.00, 90.00, 120.00
Resolution (Å)	57.41-1.18 (1.24-1.18)*	45.87-2.20 (2.70-2.20)*
R_{merge}	0.07 (0.74)	0.06 (0.31)
$I / \sigma I$	10.4 (1.8)	12.9 (3.2)
CC(1/2)	0.99 (0.43)	1.00 (0.91)
Completeness (%)	98.0 (98.0)	99.3 (99.6)
Redundancy	6.3 (6.3)	2.9 (2.9)
Refinement		
Resolution (Å)	57.41-1.18	45.87-2.20
No. reflections	94103	45421
R_{work} / R_{free}	0.173 / 0.201	0.177 / 0.221
No. atoms		
Protein	2750	8066
Water	457	453
B-factors		
Protein	13.67	39.92
Water	24.15	34.79
R.m.s. deviations		
Bond lengths (Å)	0.023	0.013
Bond angles (°)	2.13	1.49
PDB ID	6GL2	6GL0

*Single crystal was used for each data set; *Values in parentheses are for highest-resolution shell.

Table 3. Comparison of the activity of *Zg*EngA$_{GH5_4}$ and its mutants using MLG and glucomannan as substrates.

Substrate	MLG (Barley)			Glucomannan		
Mutants	**Activity** [1] **x 10^3** (min^{-1})	**Specific activity** ($\mu kat\ mg^{-1}$)	**Relative activity** (% WT activity)	**Activity** [1] **x 10^3** (min^{-1})	**Specific Activity** ($\mu kat\ mg^{-1}$)	**Relative activity** (% WT activity)
Wild type (WT)	11.4 ± 0.7	5.1 ± 0.3	100	6.9 ± 0.1	3.1 ± 0.0	100
N77A	0.2 ± 0.0	0.1 ± 0.0	1.8	0.1 ± 0.0	0.04 ± 0.0	1.4
N77Q	n.d.[2]	n.d.[2]	-	n.d.[2]	n.d.[2]	-
Y82A	14.5 ± 0.5	6.5 ± 0.2	127.2	7.6 ± 0.1	3.4 ± 0.0	110.1
Y82L	15.6 ± 0.2	6.9 ± 0.1	136.8	6.7 ± 0.2	3.0 ± 0.1	97.1
H156A	2.4 ± 0.1	1.1 ± 0.1	21	1.0 ± 0.1	0.4 ± 0.0	14.5
H156I	0.4 ± 0.0	0.2 ± 0.0	3.5	0.1 ± 0.0	0.04 ± 0.0	1.4
W210A	2.3 ± 0.1	1.0 ± 0.1	20.2	0.1 ± 0.0	0.04 ± 0.0	1.4
W210F	3.6 ± 0.4	1.6 ± 0.2	31.6	3.3 ± 0.2	1.4 ± 0.1	47.8
K211A	12.7 ± 0.1	5.7 ± 0.0	111.4	8.1 ± 0.3	3.6 ± 0.1	117.4
Y280A	12.7 ± 0.1	5.6 ± 0.0	111.4	7.9 ± 0.2	3.5 ± 0.1	114.5
Y280L	10.0 ± 0.4	4.4 ± 0.2	87.7	5.7 ± 0.1	2.5 ± 0.1	82.6
E323S	n.d.[2]	n.d.[2]	-	n.d.[2]	n.d.[2]	-
N358A	1.8 ± 0.1	0.8 ± 0.0	15.7	0.6 ± 0.1	0.2 ± 0.1	8.7
N358L	n.d.[2]	n.d.[2]	-	n.d.[2]	n.d.[2]	-
E363A	12.9 ± 0.5	5.7 ± 0.2	113.1	6.9 ± 0.1	3.1 ± 0.1	100
E363S	10.1 ± 1.2	4.5 ± 0.6	88.6	7.7 ± 0.3	3.4 ± 0.1	111.6

[1] Experiments were performed in triplicate. Results are expressed as average ±S.D.
[2] n.d., not determinable, less than the limit of detection i.e. 10 (min^{-1}) or 0.01 ($\mu kat\ mg^{-1}$))

SUPPLEMENTARY INFORMATION

Description of the simulated systems

The simulated systems were named after the position and type of linkage present in the sugar chain substrate in the simulation, (β-1,4 –for β-1,4 linkages between all units, -1/+1 β-1,3 – for β-1,3 linkage between units -1 and 1, +1/+2 β-1,3 – for β-1,3 linkage between units +1 and +2, +2/+3β-1,3 – for β-1,3 linkage between units +2 and +3), so the full simulation name consists of two parts (protein name-linkage type and position) and an 'M' was added at the end in the case of the mutated structure. All the simulations and their lengths are listed in Tables S1-S2 and number of molecules and atoms in each simulation type is provided in Tables S3-S4.

Atomistic Molecular Dynamics Simulations - protocols

The CHARMM36 force field [1-3] was used to describe protein and glucan, with water simulated using the CHARMM36-compatible TIP3P model.[4] Periodic boundary conditions were employed in all three dimensions. The length of each covalent bond to hydrogen atom was preserved using the LINCS algorithm [5] which allowed a 2 fs time step. The simulations were carried out at constant pressure (1 bar) and temperature (310 K) using the Parrinello-Rahman and velocity-rescale methods, respectively [6,7]. For pressure, an isotropic scaling was employed, and the temperatures of the solute and the solvent were coupled separately. Lennard-Jones interactions were truncated at 1.0 nm and the particle mesh Ewald method [8] was used to compute all electrostatic interactions with a real space cut-off at 1.0 nm, 6^{th} order beta spline interpolation, and a direct sum tolerance of 10^{-6}. Simulations were run at physiological salt concentrations of 150 mM KCl and counter ions were introduced to neutralize the total charge of the system. The binding of each glucan chain was simulated eight times (three repeats for native and five repeats for mutated GH5). All simulations were performed using the GROMACS 5.1 simulation package [9]

Binding energies - calculations

The molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) method was used to estimate the binding energies ΔG_{bind} between glucan and protein by decomposing into contributions from gas phase energy (ΔE_{gas}), solvation energy (ΔG_{solv}), and an entropy term ($T\Delta S$) as represented in the following equation [10]:

$$\Delta G_{bind} = \Delta E_{gas} + \Delta G_{solv} - T\Delta S \qquad (1)$$

ΔE_{gas} is composed of bonded (bond, angle, torsion) and non-bonded interactions (van der Waals, electrostatic) and constitutes the MM part of MM-PBSA. The ΔG_{sol} term contains polar solvation and non-polar solvation energies and is usually computed using the solvent-accessible surface area (SASA) model, where the SASA is linearly dependent on the non-polar term.

In this work we use the gromacs tool g_mmpbsa [11] to calculate the MM-PBSA terms for the protein-glucan complex. Note the bonded contribution is by definition zero in the single-trajectory approach [11] and the entropy term is assumed negligible for similar ligands binding to the same pocket.(12) Therefore, binding energy is calculated as follows:

$$\Delta G_{bind} = \Delta E_{MM} + \Delta G_{polar} + \Delta G_{nonpolar} \qquad (2)$$

ΔE_{MM} denotes the gas-phase energy consisting of electrostatic and van der Waals interactions, ΔG_{polar} represents polar solvation energy, and $\Delta G_{nonpolar}$ is the nonpolar solvation energy. Subsequently, the energy components ΔE_{MM}, ΔG_{polar} and $\Delta G_{nonpolar}$ of each complex were calculated for 100ns of simulations when the glucan is stably bound to the protein. The vacuum and solvent dielectric constants were set at 1 and 80, respectively. The solute dielectric constant was set to 4.

Visualization

All the snapshots and movies presented in this work were prepared using the VMD package [13].

Binding energies results

The computed binding energies stem from favorable van der Waals energy, electrostatic energy, and SASA energy, which are offset by polar solvation energy which opposes binding. For β-1,4 systems (least favorable binding), the average van der Waals energy, electrostatic energy, polar solvation energy and SASA energy were -33.0, -30.0, 42.5 and -27.8 kcal/mol, respectively. The van der Waals energy contribution among the three different substrate chains that included a β-1,3 linkage varied from -34.0 to -46.8 kcal/mol, electrostatic energy varied more strongly from -42.1 to -76.3 kcal/mol with the corresponding polar solvation energy penalties varying from +44.6 to +88.9 kcal/mol. The highest magnitude values of van der Waals, electrostatic and polar solvation energy were recorded in systems with β-1,3 linkage between the -1 and +1, again supporting the hypothesis that -1/+1 β-1,3 results in effective interactions between substrate and the GH5 binding pocket. The estimated free energies of binding are relatively high compared to studies carried out on similar proteins,[14,15] which may be due to force field effects and/or choice of dielectric constant for the buried protein pocket;[16] nevertheless the rank orderings of substrate binding should be predictive unless different substrate topologies cause large-scale resculpting of the binding pocket and/or diffuse to alternative binding sites at timescales beyond the sub-microsecond sampling of the simulations.

References

1. Huang, J., and MacKerell, A. D. (2013) CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry* **34**, 2135-2145
2. Guvench, O., Greene, S. N., Kamath, G., Brady, J. W., Venable, R. M., Pastor, R. W., and Mackerell, A. D., Jr. (2008) Additive empirical force field for hexopyranose monosaccharides. *J Comput Chem* **29**, 2543-2564
3. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998) All-atom empirical potential

for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry. B* **102**, 3586-3616
4. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926-935
5. Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997) LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**, 1463-1472
6. Parrinello, M., and Rahman, A. (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52**, 7182-7190
7. Bussi, G., Donadio, D., and Parrinello, M. (2007) Canonical sampling through velocity rescaling. *The Journal of chemical physics* **126**, 014101
8. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A smooth particle mesh Ewald method. *The Journal of chemical physics* **103**, 8577-8593
9. Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., and Lindahl, E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19-25
10. Zoete, V., Irving, M. B., and Michielin, O. (2010) MM-GBSA binding free energy decomposition and T cell receptor engineering. *J Mol Recognit* **23**, 142-152
11. Kumari, R., Kumar, R., Open Source Drug Discovery, C., and Lynn, A. (2014) g_mmpbsa--a GROMACS tool for high-throughput MM-PBSA calculations. *J Chem Inf Model* **54**, 1951-1962
12. Xu, H. L., Wang, Z. J., Liang, X. M., Li, X., Shi, Z., Zhou, N., and Bao, J. K. (2014) In silico identification of novel kinase inhibitors targeting wild-type and T315I mutant ABL1 from FDA-approved drugs. *Mol Biosyst* **10**, 1524-1537
13. Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33-38
14. Knott, B., Crowley, M., E Himmel, M., Ståhlberg, J., and Beckham, G. (2014) *Carbohydrate-Protein Interactions That Drive Processive Polysaccharide Translocation in Enzymes Revealed from a Computational Study of Cellobiohydrolase Processivity*, *Journal of the American Chemical Society* **2014,** *136* (24), 8810-8819.
15. Szefler, B., Diudea, M., Putz, M., and Grudzinski, I. (2016) Molecular Dynamic Studies of the Complex Polyethylenimine and Glucose Oxidase. *International Journal of Molecular Sciences* **17**, 1796
16. Aleksandrov, A., Thompson, D., and Simonson, T. (2010) Alchemical free energy simulations for biological complexes: powerful but temperamental. *Journal of molecular recognition : JMR* **23**, 117-127

SUPPLEMENTARY FIGURE LEGENDS

Figure S1: Sequence of the gene *ZGAL_208* (A) encoding the amino acid sequence (B) of the protein *Zg*EngA$_{GH5_4}$. (A) The sequence of the forward and reverse primers used for the amplification of the catalytic module of *Zg*EngA$_{GH5_4}$ is in bold and underlined. (B) The LipoP 1.0 software [20] was used to delineate the signal peptide (amino acids 1 to 20 in red) of the protein which contain a large GH5 module (residues 21 to 397). The precise boundaries of the catalytic module (residues 56 to 385 in green and bold) used for the biochemical and structural characterizations were delineated from Hydrophobic Cluster Analysis plot [21]. The molecular weight of the recombinant protein (sequence in green) has been calculated using the ProtParam tool [35].

Figure S2: Representative snapshots of the computed oligo-glucan- *Zg*EngA$_{GH5_4_E323S}$ complexes (right) and an enlarged view of the oligosaccharide substrate alone (left): (A) Complex with cellohexaose (β-1,4 linkages between all glucose units). (B) -1/+1 β(1,3) oligoglucan. The β-1.3 linkage is marked by the blue dot in the lefthand panel and the orange sphere in the righthand panel). (C) +1/+2 β(1,3). (D) +2/+3 β(1,3). Protein (mutant *Zg*EngA$_{GH5_4_E323S}$) is shown in cartoon representation, colored according to secondary structure and overlaid with a space-filling cyan, transparent surface; glucan is shown in as sticks and the β-1,3 linkage site is marked with an orange van der Waals sphere. Water molecules are omitted for clarity.

Figure S3: **Purity (A and B) and oligomerization state (C and D) analysis of *Zg*EngA$_{GH5_4}$.** (A and B) SDS- PAGE analysis (A) and DLS (B) were performed to check the purity of *Zg*EngA$_{GH5_4}$. (A) in the SDS PAGE, the Precision Plus markers from BioRad (S) were used as standards. (B) Dynamic light scattering (DLS) was carried out to estimate the size distribution of the molecules as a function of their volume.

(C and D) Oligomerization state studies of *Zg*EngA$_{GH5_4}$ using size exclusion chromatography (C) and size exclusion chromatography coupled to multiple angle laser light scattering (MALLS)(D). Size exclusion chromatography (C) was performed using a Superdex 75 HiLoad 16/60 column previously equilibrated in 25 mM Tris HCl + 100 mM NaCl (pH 7.5) (Buffer B) and calibrated using the following standard proteins: Conalbumin (C) (MW: 75 kDa); Ovalbumin (O) (MW: 43 kDa); Carbonic anhydrase (CA) (MW: 29 kDa); Ribonuclease (R) (MW: 13.7 kDa) and Aprotinin (A) (MW: 6.5 kDa). Exclusion volume of the column was evaluated using a fresh solution of Dextran Blue 2000 (B) (MW > 2000 kDa). (D) Size exclusion chromatography coupled to MALLS was performed using a

Superdex 200 Increase 10/300GL previously equilibrated in Buffer B. Elution of *Zg*EngA$_{GH5_4}$ was performed at a flow rate of 0.5 mL min^{-1} and its calculated molar mass is shown (dotted line).

Figure S4: Thermostability of *Zg*EngA$_{GH5_4}$ (A) and influence of pH (B) and temperature (C) on its activity. All these experiments have been carried out in triplicate. The activity was measured using the ferricyanide assay and the results are expressed as percentage of relative activity. (A) The thermostability of *Zg*EngA$_{GH5_4}$ was studied by DLS in a temperature range of 5 to 65 ° C in steps of 1 ° C. The hydrodynamic gyration radius (Rg) was measured at each step and the values are the average of triplicate experiments. (B) The pH optimum was determined using the Teorell and Stenhagen buffer in a range of pH from pH 4.2 to pH 8.5. This buffer was used at a 100 mM concentration to prepare both enzyme (100 nM) and β-D-glucan (0.2%) solutions. pH optimum of the enzyme was further checked using biological buffers (MES, MOPS, Tris and Phosphate) in similar conditions. Note that these later experiments have only been performed in duplicate. (C) Optimal temperature of *Zg*EngA$_{GH5_4}$ was determined by incubating both enzyme and substrate at the appropriate temperature before performing the hydrolysis reaction.

Figure S5. Secondary structure of the protein in time for β-1,4 during three repeats. Even if no protein-glucan binding was observed for this complex, the protein shows relative structural stability in all three trajectories as seen in panels A, B and C below.

Figure S6. Secondary structure of the protein in time for model -1/+1 β-1,3 during three repeats are shown in panels A, B and C below. Out of the three trajectories, protein-glucan binding was observed in only B. However, on comparing panels A, B and C it can be seen that the relative secondary structure (mainly α-helices and β-sheets) of the protein is not affected by glucan binding/dissociation.

Figure S7. Secondary structure of the protein in time for model +1/+2 β-1,3 during three repeats are shown in panels A, B and C below. Out of the three trajectories, protein-glucan binding was observed in only C. However, on comparing panels A, B and C it can be seen that the relative secondary structure (mainly α-helices and β-sheets) of the protein is not affected by glucan binding/ dissociation.

Figure S8. Secondary structure of the protein in time for model +2/+3 β-1,3 during three repeats are shown in panels A, B and C below. Out of the three trajectories, protein-glucan binding was observed in only B. However, on comparing all three trajectories, panels A, B and C, it can be

observed that the relative secondary structure (mainly α-helices and β-sheets) of the protein, in this complex, is not affected by glucan binding/ dissociation.

Figure S9. Secondary structure of the protein in time for model β-1,4_M during five repeats are shown in panels A, B, C, D and E below. Out of the five trajectories, protein-glucan binding was observed in C and E. However, on comparing all five panels it can be seen that the relative secondary structure (mainly α-helices and β-sheets) of the protein is not affected by glucan binding/dissociation.

Figure S10. Secondary structure of the protein in time for model -1/+1 β-1,3_M during five repeats are shown in panels A, B, C, D and E below. Protein-glucan binding was observed in four (B, C, D and E) of the five trajectories and all panels below show relative secondary structure stability through the simulations.

Figure S11. Secondary structure of the protein in time for model +1/+2 β-1,3_M during five repeats are shown in panels A, B, C, D and E below. Out of the five trajectories, protein-glucan binding was observed in A and C. However, on comparing all five panels it can be seen that the relative secondary structure (mainly α-helices and β-sheets) of the protein is not affected by glucan binding/ dissociation.

Figure S12. Secondary structure of the protein in time for model +2/+3 β-1,3_M during five repeated simulations are shown in panels A, B, C, D and E below. Out of the five trajectories, protein-glucan binding was observed in A and D. However, on comparing all five panels it can be seen that the relative secondary structure (mainly α-helices and β-sheets) of the protein is not affected by glucan binding/ dissociation.

Figure S13. Root mean square deviation plots of backbone non-hydrogen atoms in mutated *Zg*EngA$_{GH5_4_E323S}$ simulations. A, B, C, D, E – -1/+1 β-1,3_M (repeats 1-5 respectively); F, G, H, I, J – +1/+2 β-1,3_M; K, L, M, N, O – +2/+3 β-1,3_M; P, Q, R, S, T – β-1,4_M are shown below. In panels I, J, P and S, simulations were not continued when the respective glucan chain was seen to dissociate from the protein binding cleft. The most stable RMSD (plateau) is observed in panels A-E, corresponding to mutated *Zg*EngA$_{GH5_4_E323S}$ complex β-1,3_M.

Figure S14. Root mean square deviation graphs of backbone non-hydrogen atoms in wildtype *Zg*EngA$_{GH5_4}$ simulations. A, B, C – -1/+1 β-1,3 (repeats 1-3 respectively); D, E, F – +1/+2 β-1,3; G, H, I – +2/+3 β-1,3; J, K, L – β-1,4 are shown below. Simulations were not continued when the glucan

chain was seen to dissociate from the protein binding cleft early on in the trajectory. These are panels A, E, G and I

Figure S15. Root mean square fluctuations of Cα atoms in the mutant *Zg*EngA$_{GH5_4_E323S}$ simulations. A, B, C, D, E – -1/+1 β-1,3_M (repeats 1-5 respectively); F, G, H, I, J – +1/+2 β-1,3_M; K, L, M, N, O – +2/+3 β-1,3_M; P, Q, R, S, T – β-1,4_M are shown below. From these, it can be observed that the more flexible and loose parts of the crystal structure such as turns and loops are along residue regions 85-90, 125-126, 162-164 and 207-211

Figure S16. Root mean square fluctuations of Cα atoms in the wildtype ZgEngA$_{GH5_4}$ simulations. A, B, C – -1/+1 β-1,3 (repeats 1-3 respectively); D, E, F – +1/+2 β-1,3; G, H, I – +2/+3 β-1,3; J, K, L – β-1,4. Similar to mutated ZgGH5, the more flexible and loose parts of the crystal structure such as turns and loops are along residue regions 85-90, 125-126, 162-164 and 207-211.

Figure S17. Root mean square deviation of glucan backbone structure atoms in mutated *Zg*EngA$_{GH5_4_E323S}$. A, B – β-1,4_M (repeats 3 and 5 respectively); C, D, E, F – -1/+1 β-1,3_M (repeats 2-5 respectively); G, H– +1/+2 β-1,3_M (repeats 1 and 3 respectively); I – +2/+3 β-1,3_M (repeat 1). Trajectories where glucan dissociation is observed are not shown in the graphs below. Panels C-F show highly stable binding of glucan -1/+1 β-1,3 with mutated *Zg*EngA$_{GH5_4_E323S}$.

Figure S18. Root mean square deviation of glucan backbone structure atoms in native ZgEngAGH5_4. A, B – β-1,4 (repeats 1 and 2 respectively); C – -1/+1 β-1,3 (repeat 2); D – +1/+2 β-1,3 (repeat 3); E – +2/+3 β-1,3 (repeat 2). Stable glucan-protein binding is observed in panels C and E.

Figure S19. Timelines of hydrogen bonds for glucan binding to mutant *Zg*EngA$_{GH5_4_E323S}$. A, B, C, D, E – -1/+1 β-1,3_M (repeats 1-5 respectively); F, G, H, I, J – +1/+2 β-1,3_M; K, L, M, N, O – +2/+3 β-1,3_M; P, Q, R, S, T – β-1,4_M are shown below. Total loss of hydrogen bonding in panels I, J, L, M, P, Q and S correspond to glucan dissociation from the protein. Most stable binding is observed in panels B-E for -1/+1 β-1,3_M complex

Figure S20. Timelines of hydrogen bonds for glucan binding to wildtype ZgEngA$_{GH5_4}$. A, B, C – -1/+1 β-1,3 (repeats 1-3 respectively); D, E, F – +1/+2 β-1,3; G, H, I – +2/+3 β-1,3; J, K, L – β-1,4. Glucan dissociates from the protein in trajectories A, E, G, I and L, hence the number of hydrogen bonds between glucan-protein goes down to zero in these cases.

Figure S21. Percentage of occurrence of interatomic contacts < 0.35 nm in the last 200 ns of the simulations of mutant *Zg*EngA$_{GH5_4_E323S}$. A – -1/+1 β-1,3_M repeat 2, B – -1/+1 β-1,3_M repeat 3, C – +1/+2 β-1,3_M repeat 1, D – +1/+2 β-1,3_M repeat 3, E – +2/+3 β-1,3 repeat 1, F – β-1,4 repeat 3 are shown below. Main residues involved in *Zg*EngA$_{GH5_4_E323S}$ and glucan binding involve H155, E200, E209, W210, T253 H275, Y277, Y280, W356, N358 and F364

Figure S22: **Conformation of the loop harboring Y280, between β-strand β6 and α-helix 6**. (A) Surface representation of the structure of the active site of *Zg*EngA$_{GH5_4_E323S}$. (B) Surface representation of the *Zg*EngA$_{GH5_4_E323S}$ Y280A model. (C) Surface representation of the structure of the active site of F32EG5. The surface of the active site of F32EG5 is quite identical to that of Y280A mutant.

Figure S1

Characteristics of the recombinant protein (sequence in green): 330 amino acids; molecular weight: 37514 Da

Figure S2

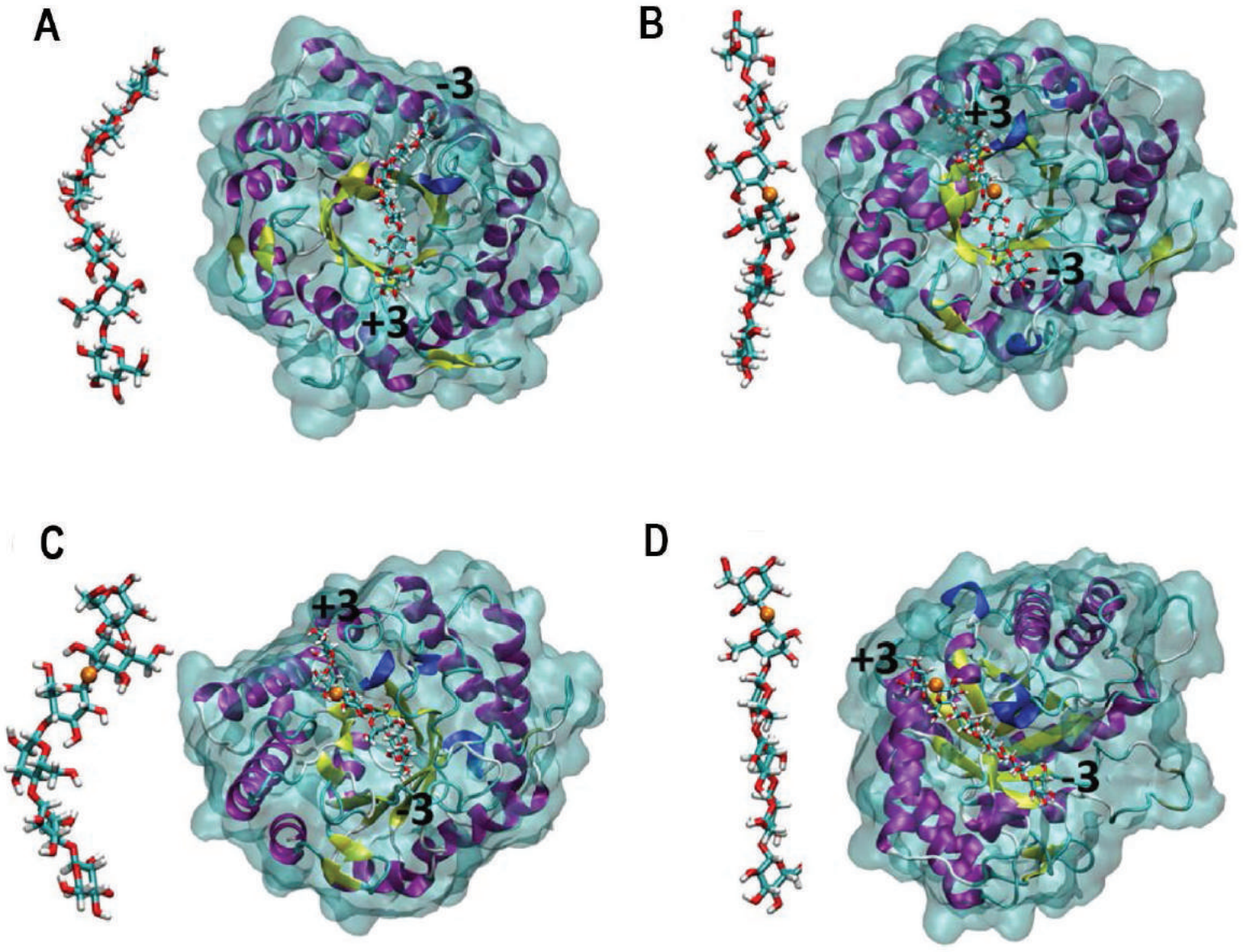


Figure S3

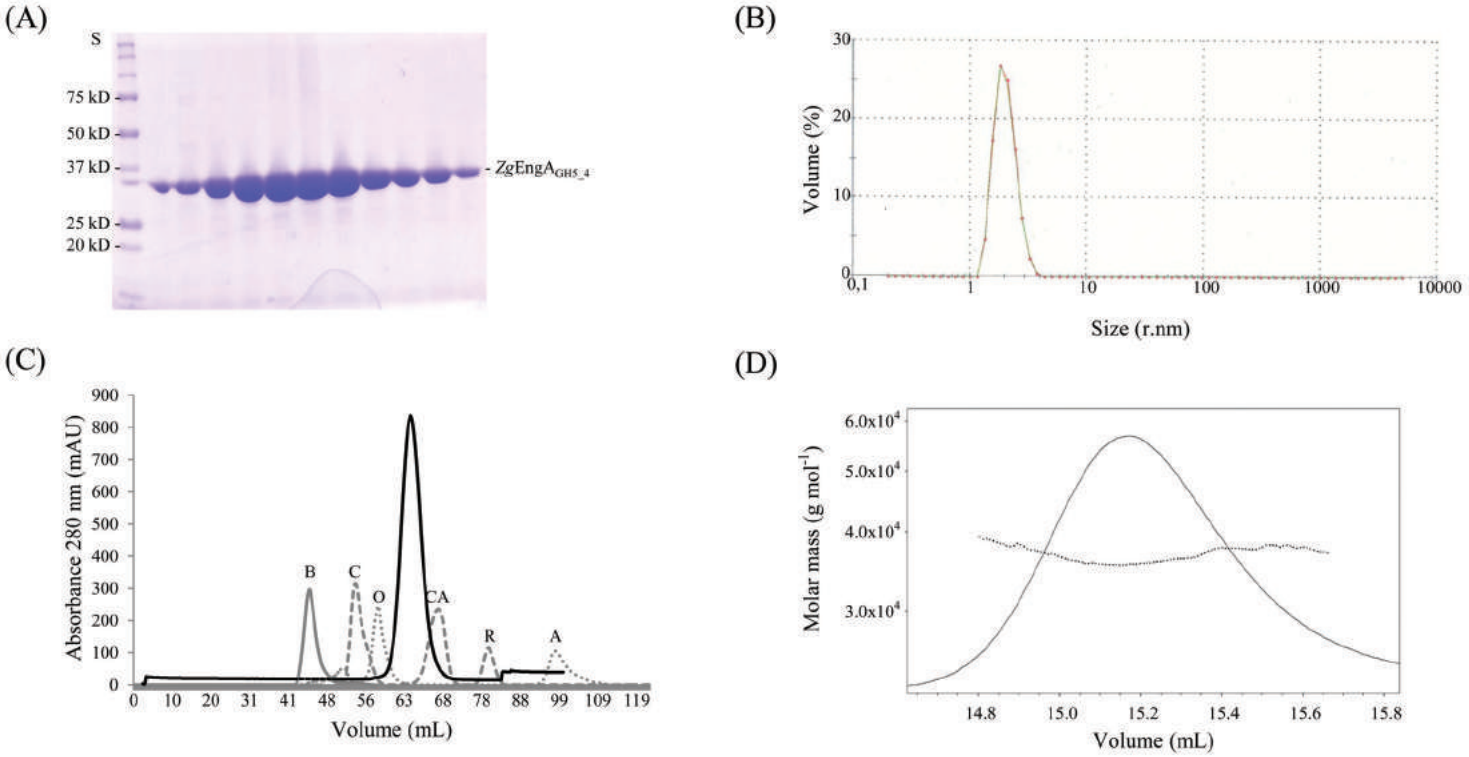

(A)
S
75 kD
50 kD
37 kD
25 kD
20 kD
ZgEngA$_{GH5_4}$
(B)
Volume (%)
Size (r.nm)
(C)
Absorbance 280 nm (mAU)
Volume (mL)
B
C
O
CA
R
A
(D)
Molar mass (g mol^{-1})
Volume (mL)

Figure S4

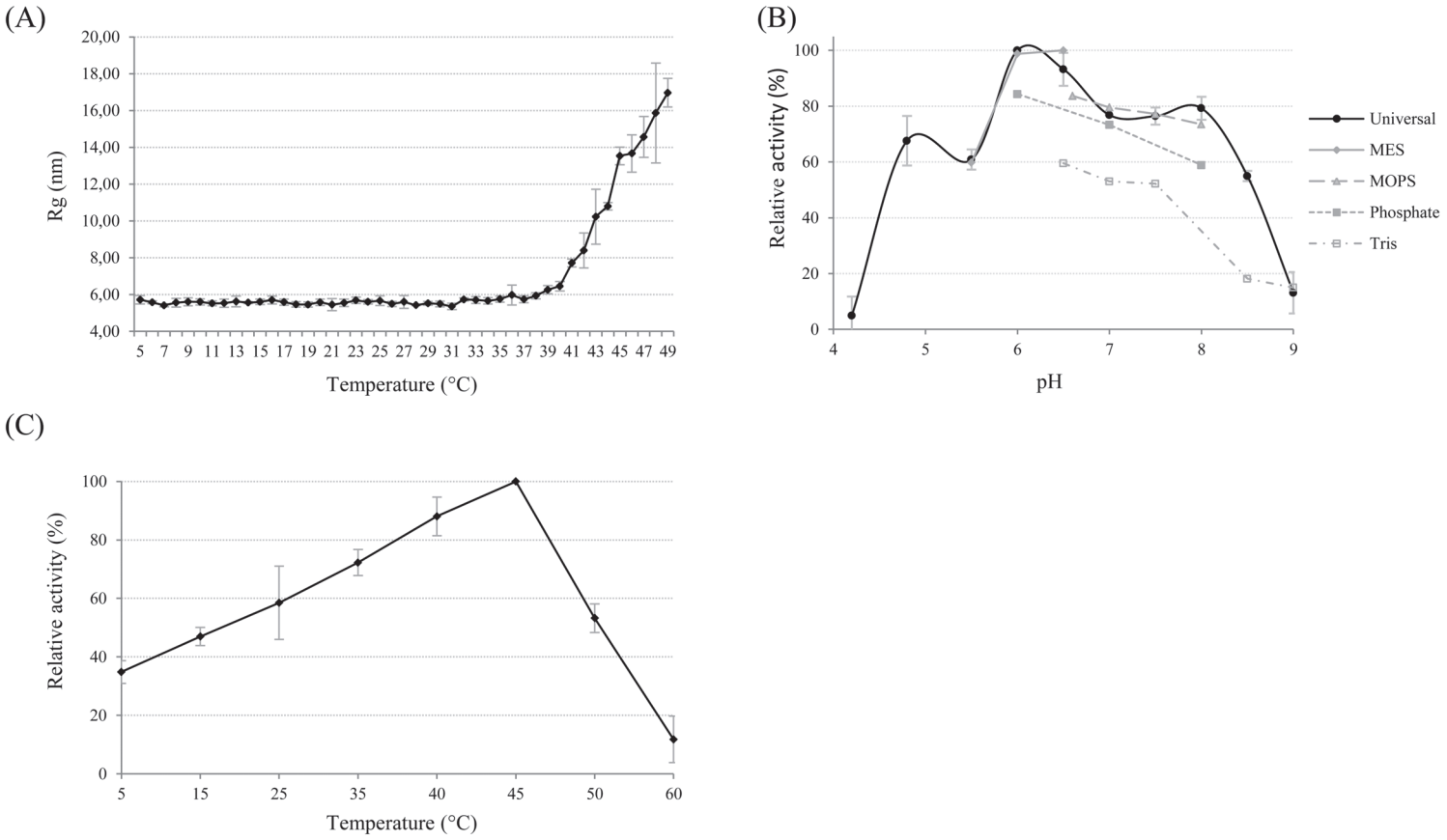

(A)
Rg (nm)
20,00
18,00
16,00
14,00
12,00
10,00
8,00
6,00
4,00
5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Temperature (°C)
(B)
Relative activity (%)
100
80
60
40
20
0
4 5 6 7 8 9
pH
Universal
MES
MOPS
Phosphate
Tris
(C)
Relative activity (%)
100
80
60
40
20
0
5 15 25 35 40 45 50 60
Temperature (°C)

Figure S5.

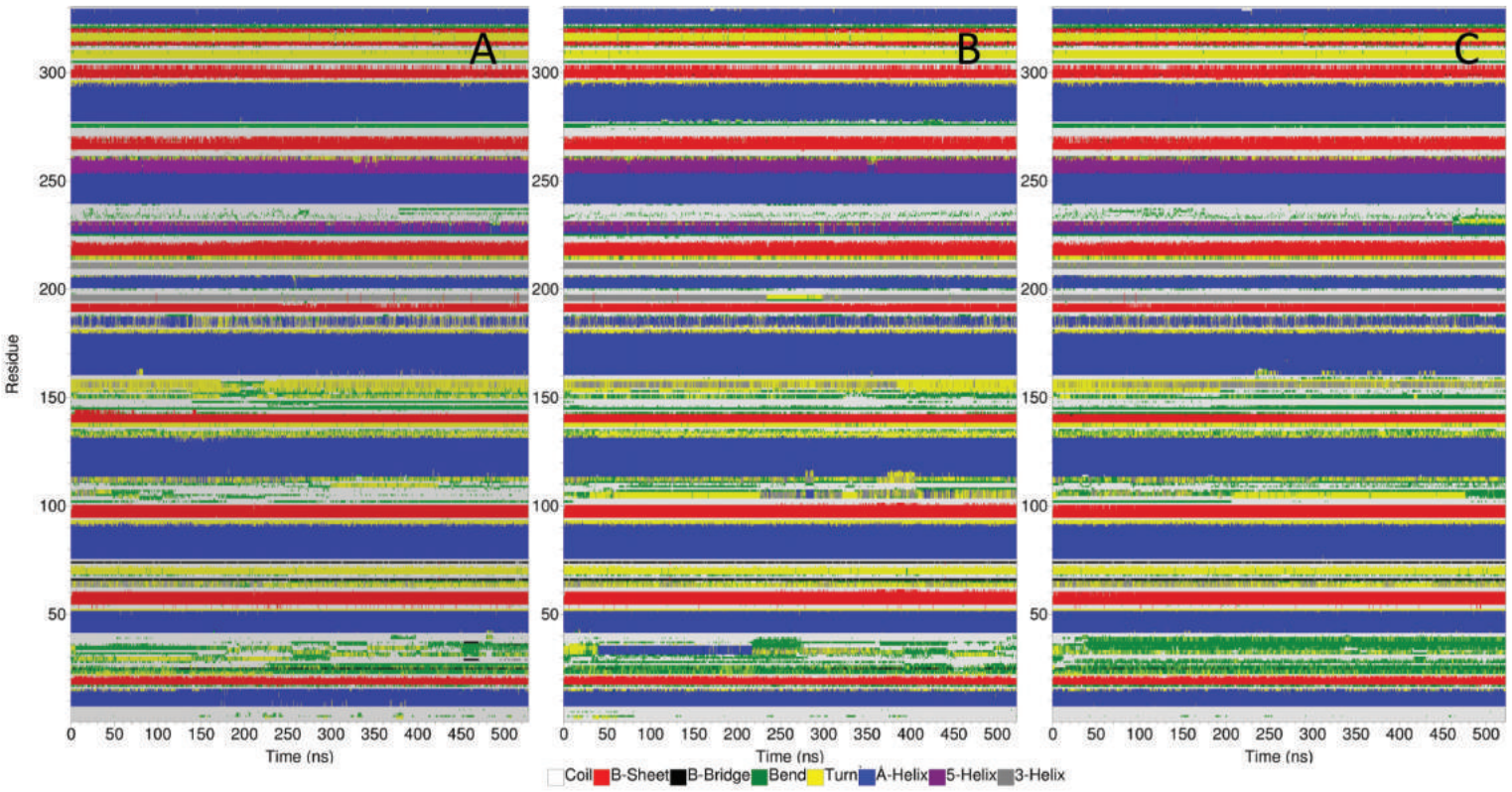

A
B
C
Residue
300
250
200
150
100
50
0 50 100 150 200 250 300 350 400 450 500
Time (ns)
Coil
B-Sheet
B-Bridge
Bend
Turn
A-Helix
5-Helix
3-Helix

Figure S6

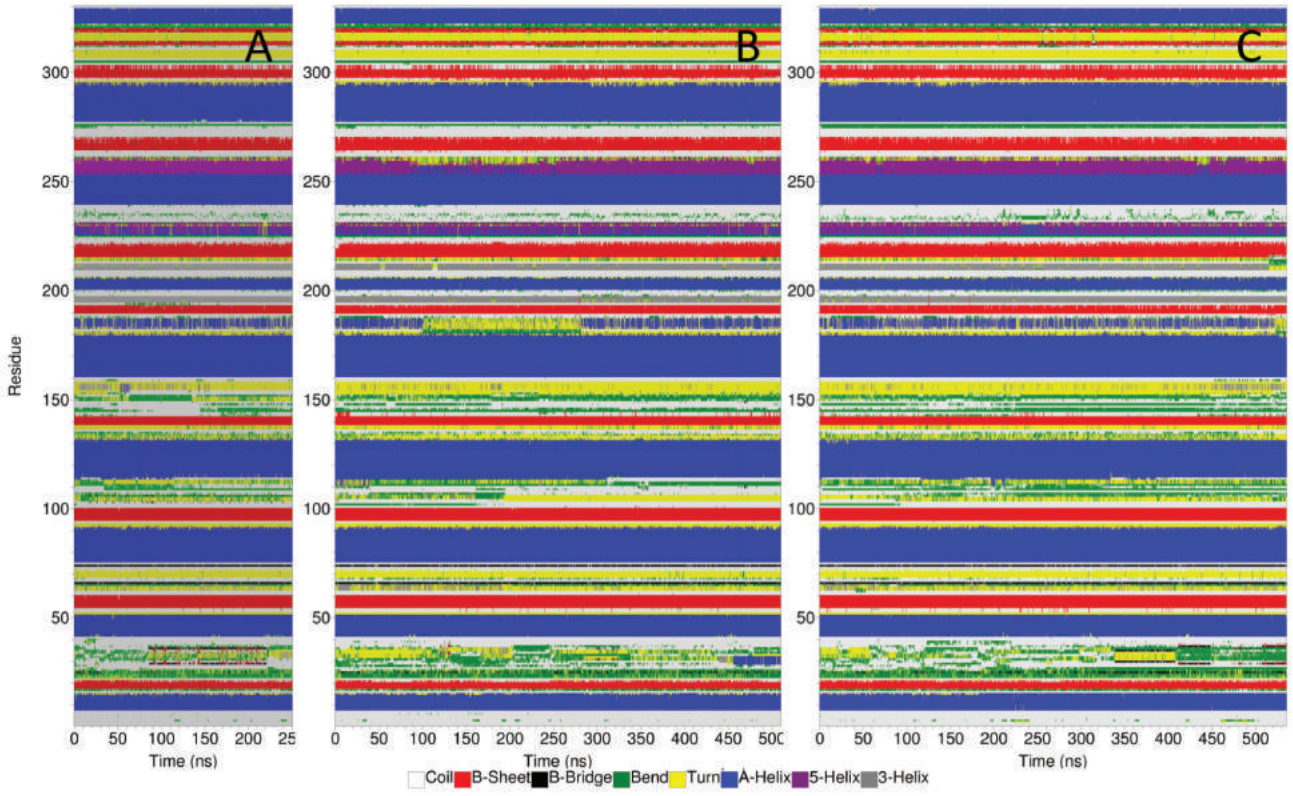

A
B
C
Residue
300
250
200
150
100
50
0 50 100 150 200 25
0 50 100 150 200 250 300 350 400 450 500
0 50 100 150 200 250 300 350 400 450 500
Time (ns)
Time (ns)
Time (ns)
Coil
B-Sheet
B-Bridge
Bend
Turn
A-Helix
5-Helix
3-Helix

Figure S7

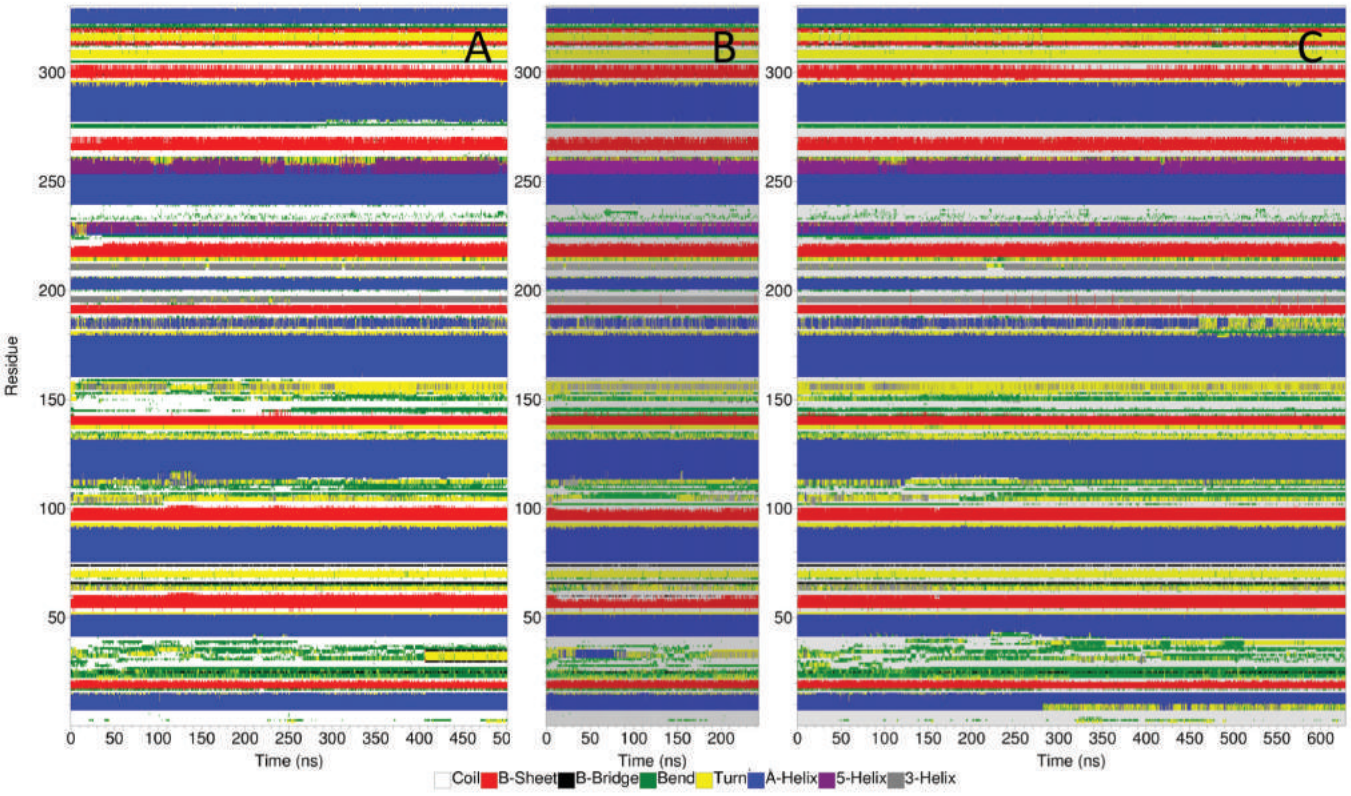

A
B
C
Residue
300
250
200
150
100
50
0 50 100 150 200 250 300 350 400 450 500
Time (ns)
0 50 100 150 200
Time (ns)
0 50 100 150 200 250 300 350 400 450 500 550 600
Time (ns)
Coil
B-Sheet
B-Bridge
Bend
Turn
A-Helix
5-Helix
3-Helix

Figure S8

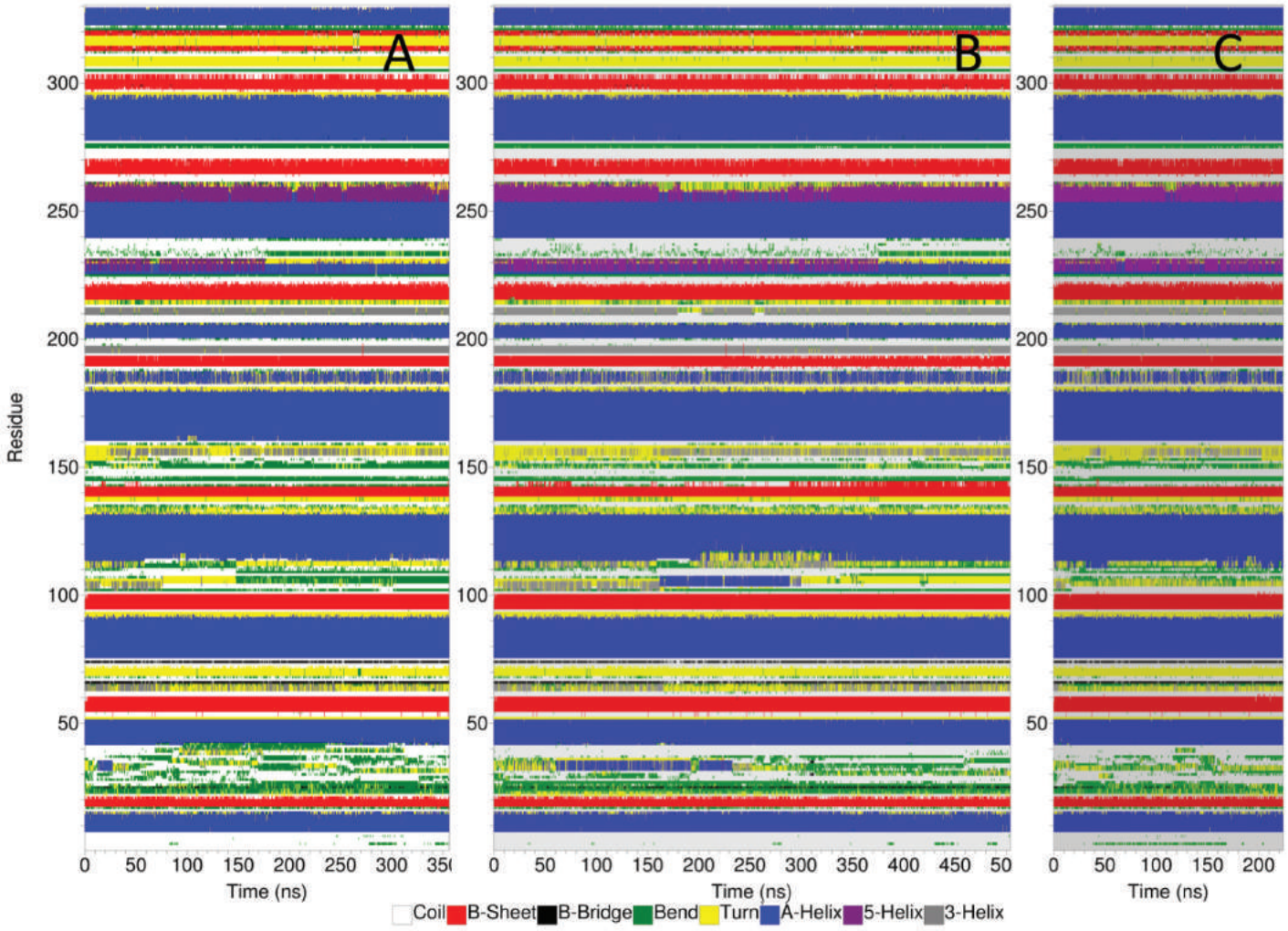

A
B
C
Residue
Time (ns)
Time (ns)
Time (ns)
Coil
B-Sheet
B-Bridge
Bend
Turn
A-Helix
5-Helix
3-Helix

Figure S9_Page1

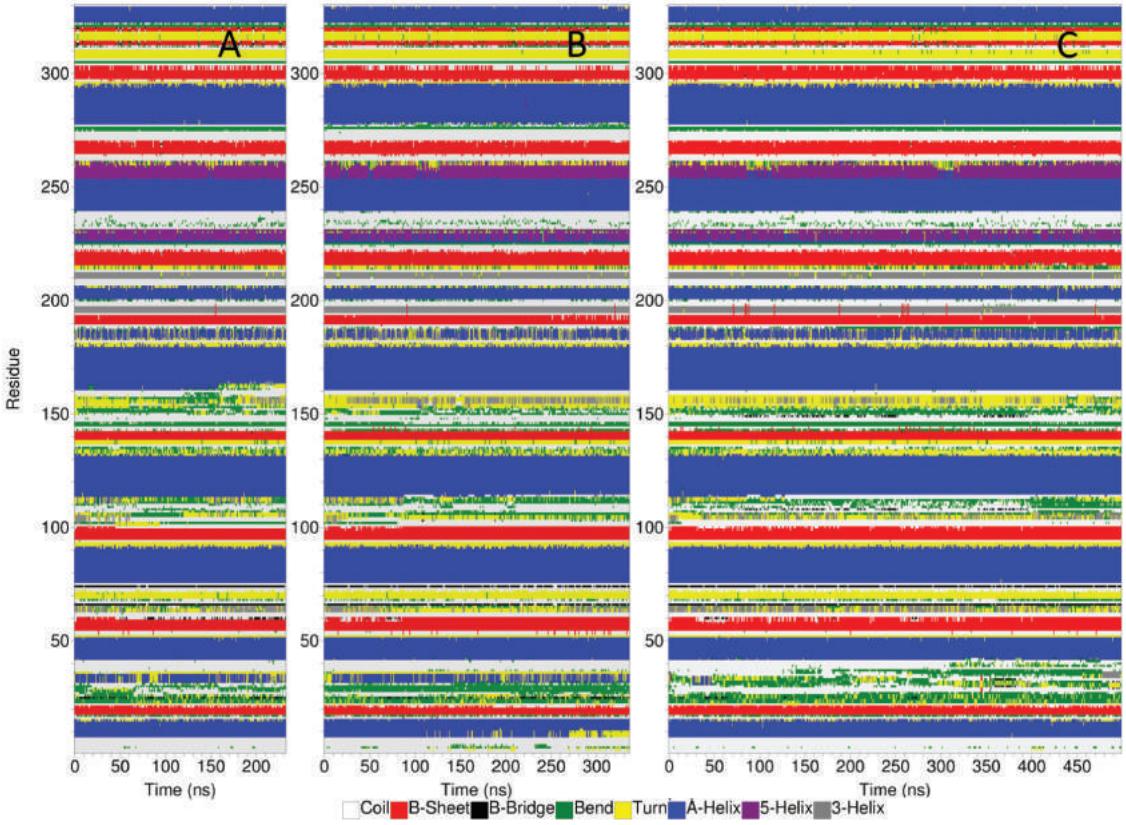


Figure S9_Page2

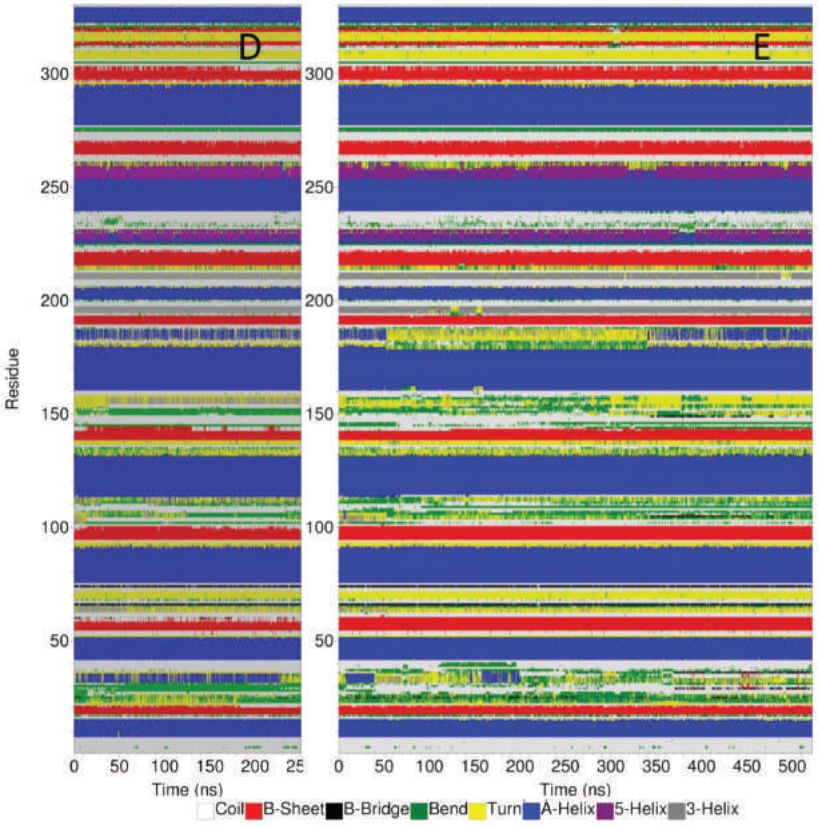


Figure S10_Page1

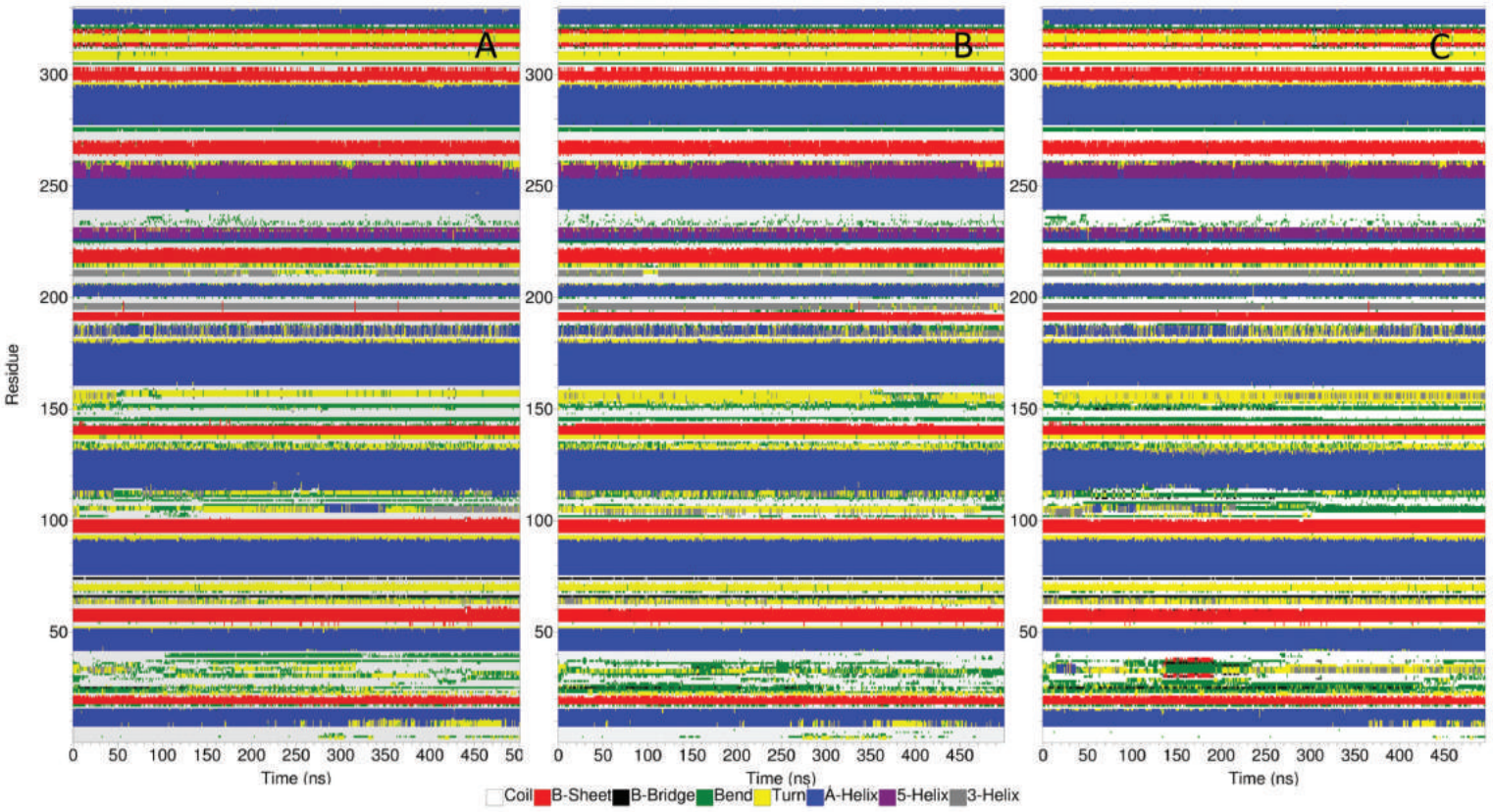


Figure S10_Page2

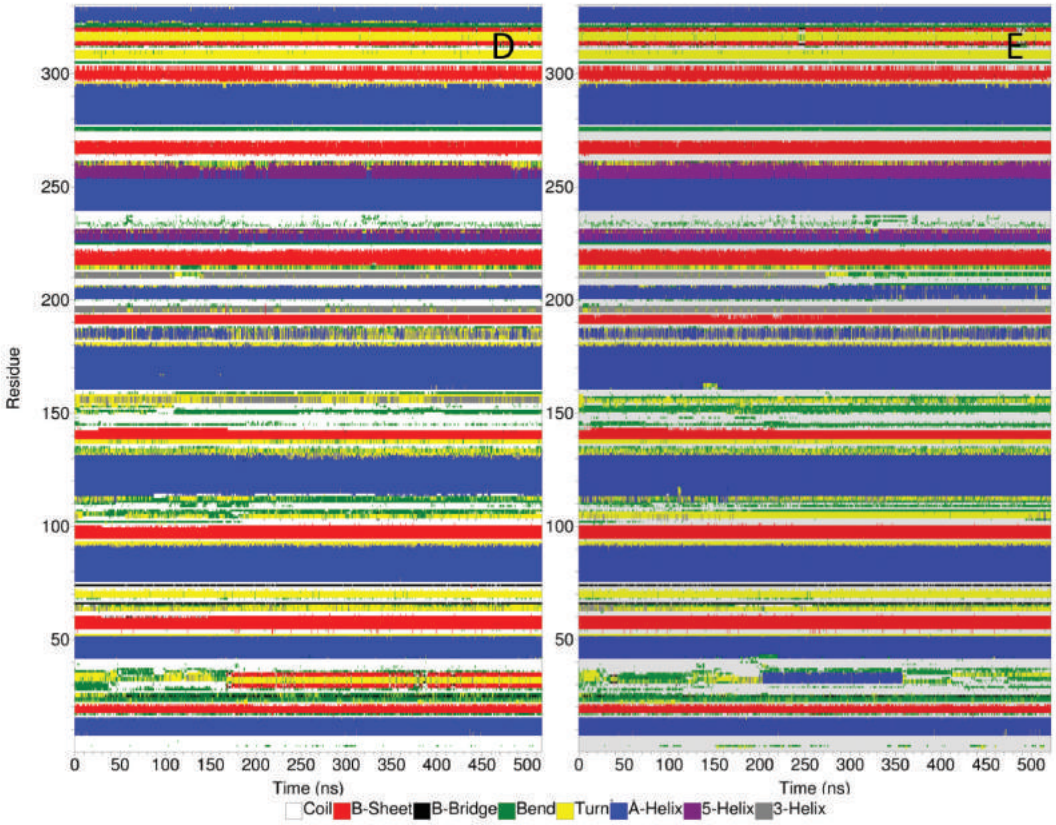


Figure S11_Page1

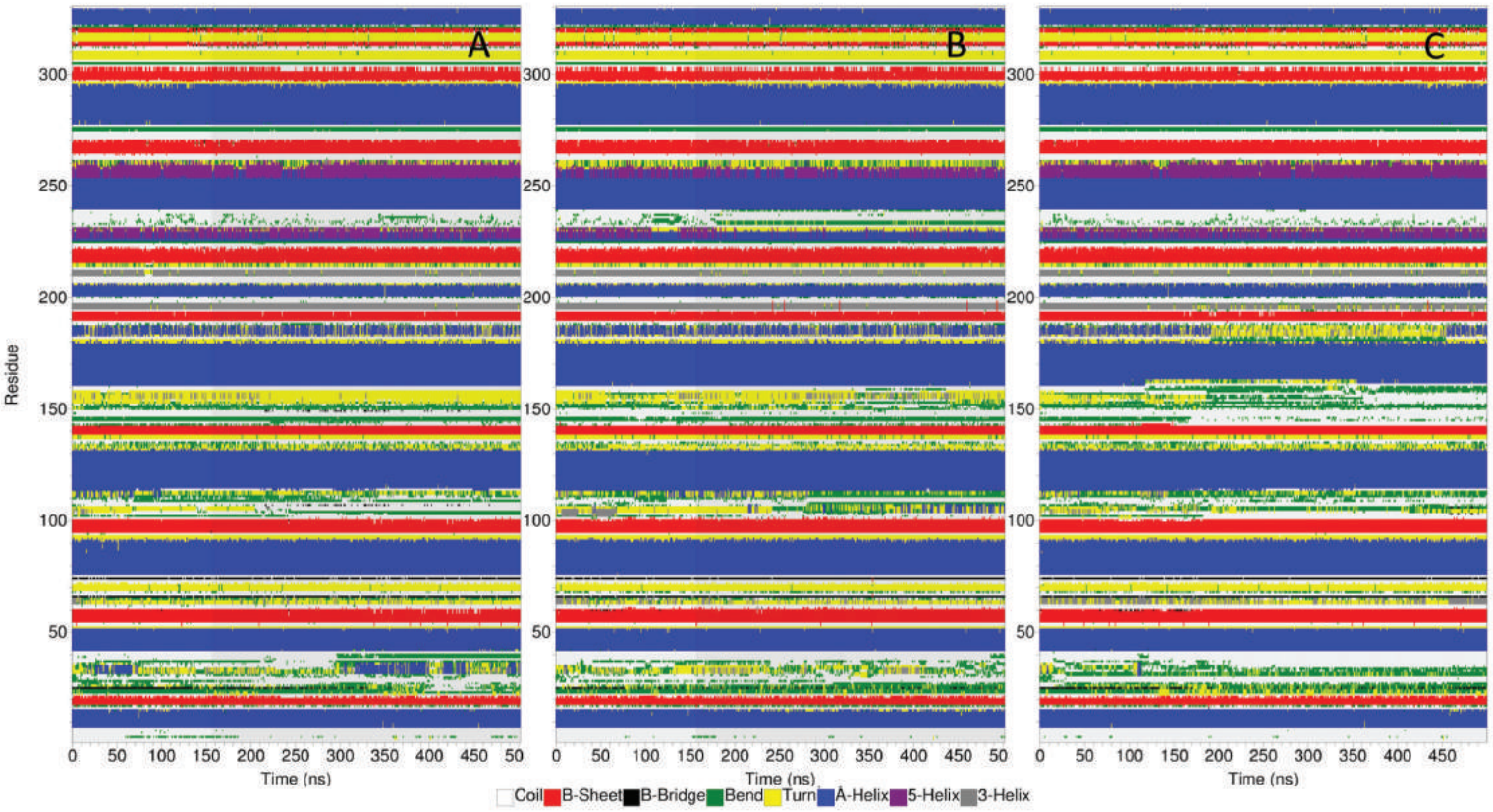


Figure S11_Page2

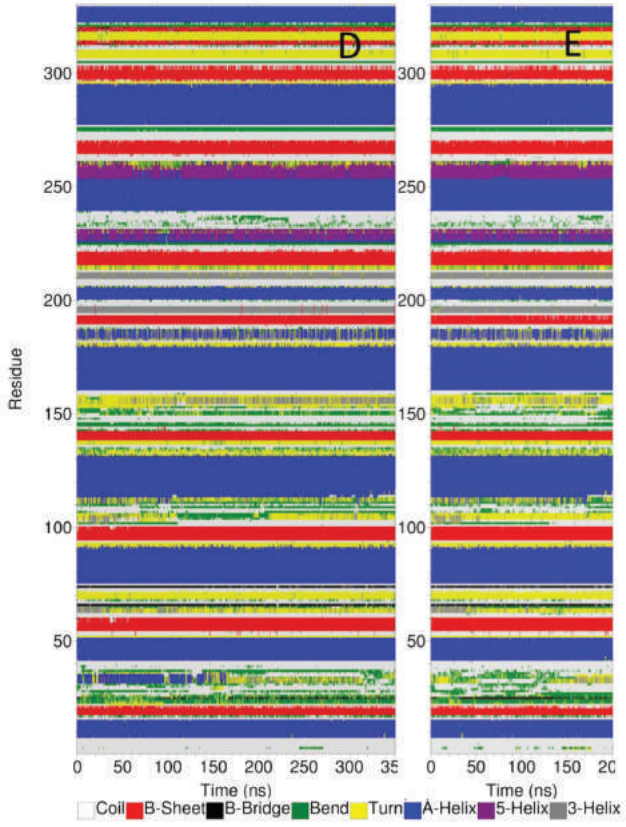


Figure S12_page1

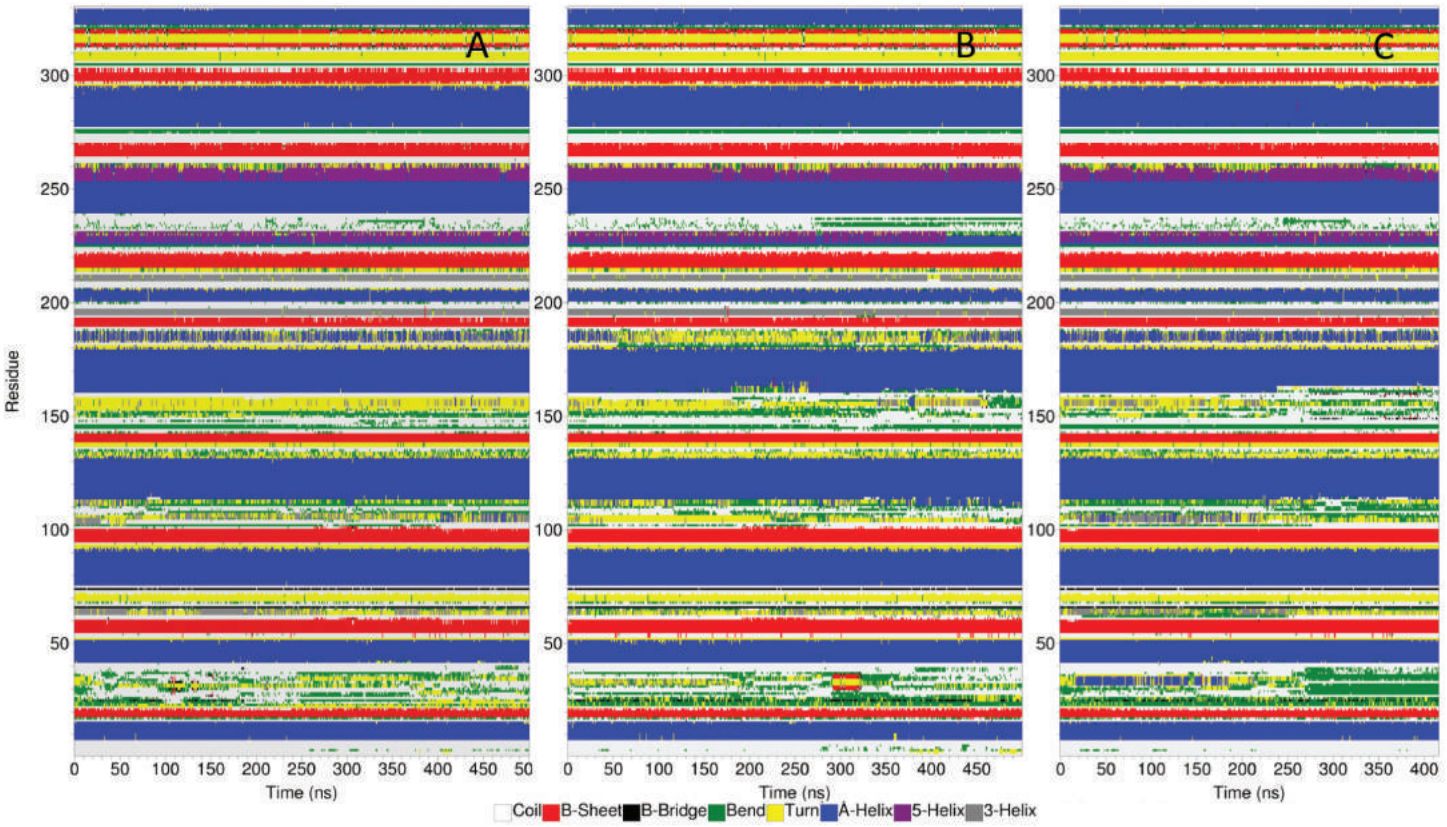


Figure S12_Page2

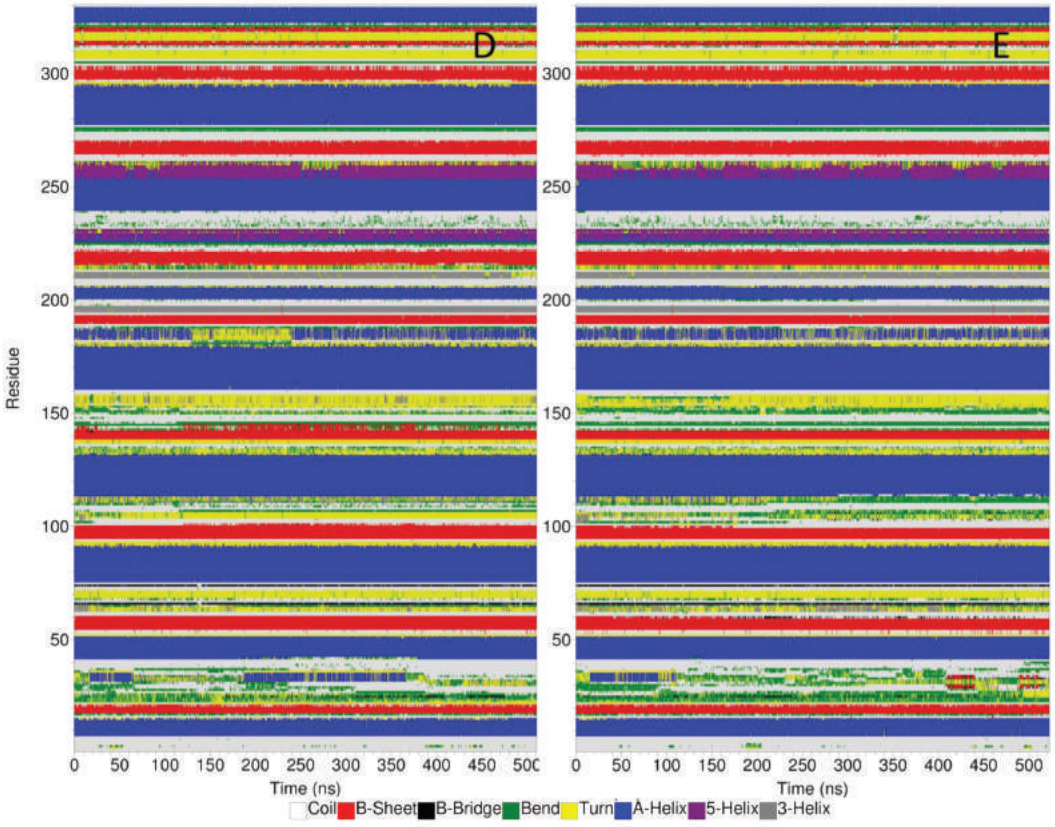


Figure S13

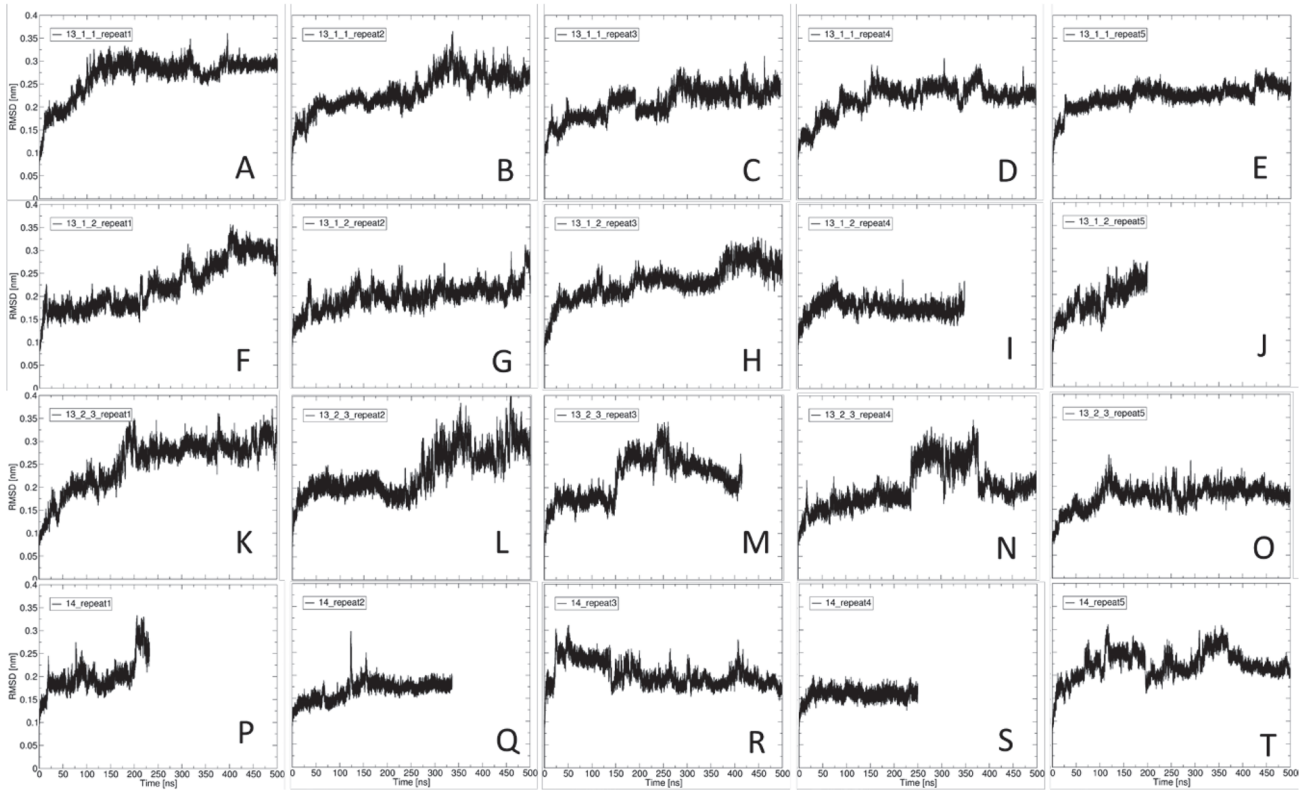

13_1_1_repeat1
13_1_1_repeat2
13_1_1_repeat3
13_1_1_repeat4
13_1_1_repeat5
13_1_2_repeat1
13_1_2_repeat2
13_1_2_repeat3
13_1_2_repeat4
13_1_2_repeat5
13_2_3_repeat1
13_2_3_repeat2
13_2_3_repeat3
13_2_3_repeat4
13_2_3_repeat5
14_repeat1
14_repeat2
14_repeat3
14_repeat4
14_repeat5
RMSD [nm]
Time [ns]
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T

Figure S14

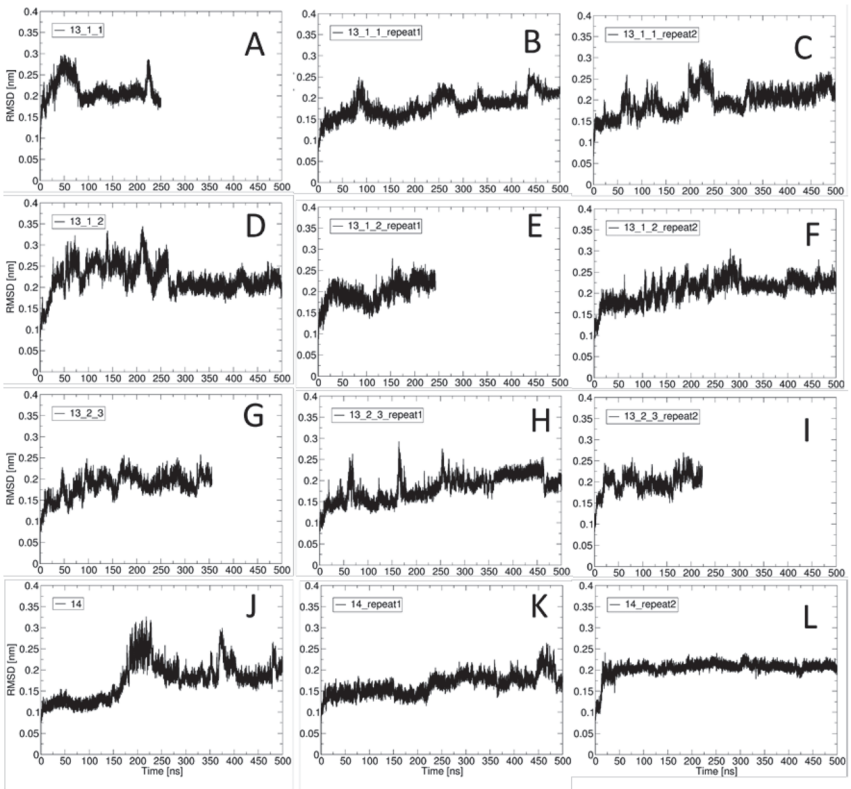

A
— 13_1_1
B
— 13_1_1_repeat1
C
— 13_1_1_repeat2
D
— 13_1_2
E
— 13_1_2_repeat1
F
— 13_1_2_repeat2
G
— 13_2_3
H
— 13_2_3_repeat1
I
— 13_2_3_repeat2
J
— 14
K
— 14_repeat1
L
— 14_repeat2
RMSD [nm]
Time [ns]

Figure S15.

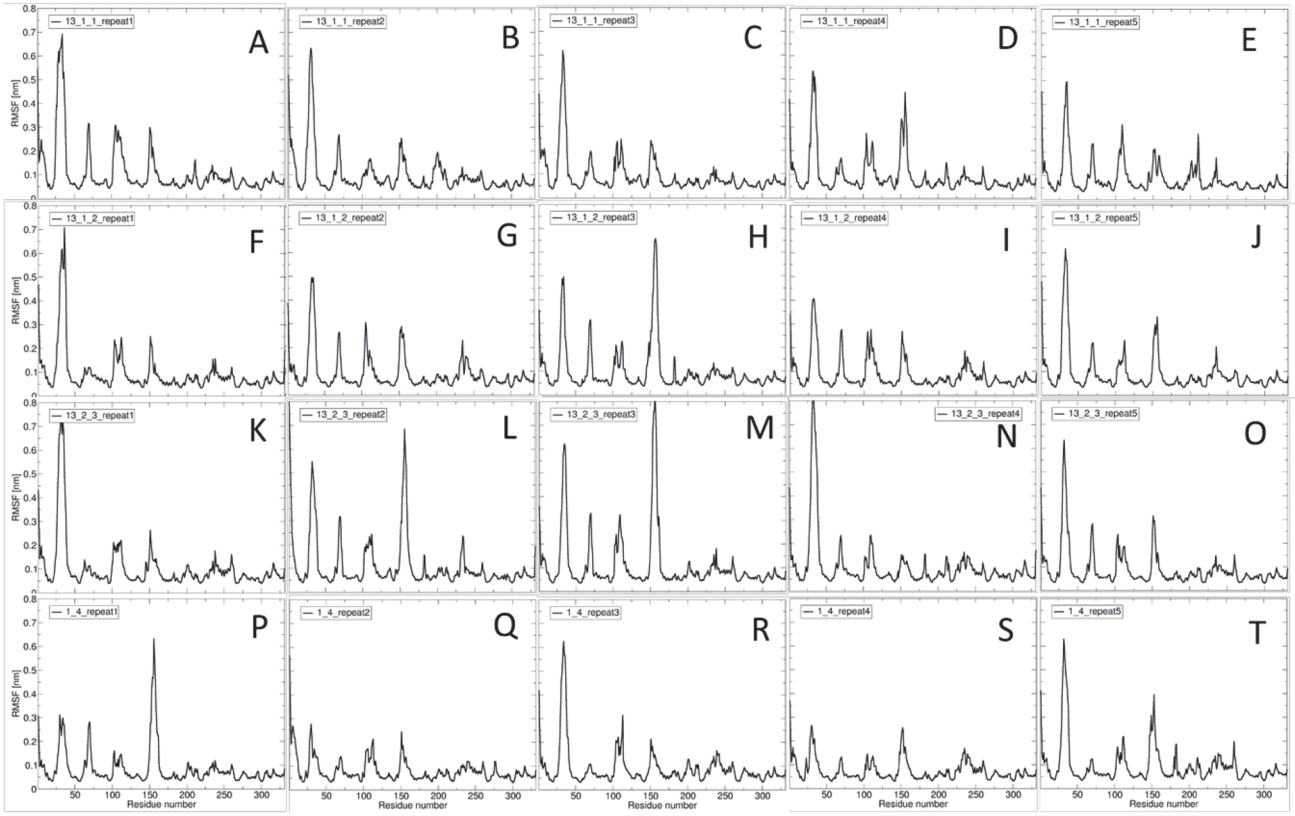

13_1_1_repeat1
A
13_1_1_repeat2
B
13_1_1_repeat3
C
13_1_1_repeat4
D
13_1_1_repeat5
E
13_1_2_repeat1
F
13_1_2_repeat2
G
13_1_2_repeat3
H
13_1_2_repeat4
I
13_1_2_repeat5
J
13_2_3_repeat1
K
13_2_3_repeat2
L
13_2_3_repeat3
M
13_2_3_repeat4
N
13_2_3_repeat5
O
1_4_repeat1
P
1_4_repeat2
Q
1_4_repeat3
R
1_4_repeat4
S
1_4_repeat5
T
RMSF [nm]
Residue number

Figure S16

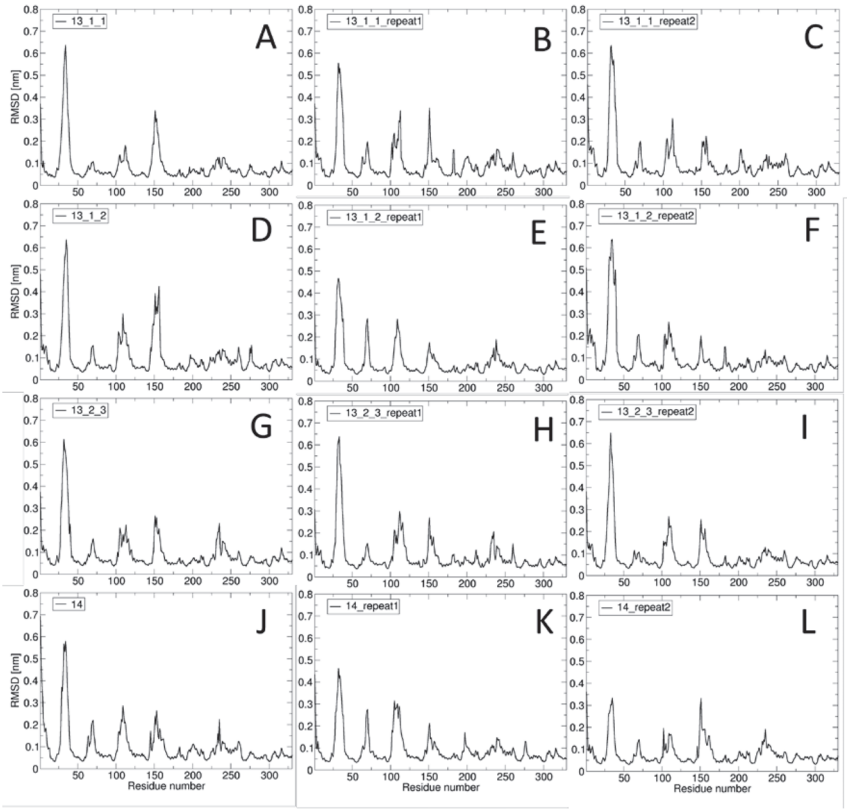

13_1_1
A
13_1_1_repeat1
B
13_1_1_repeat2
C
13_1_2
D
13_1_2_repeat1
E
13_1_2_repeat2
F
13_2_3
G
13_2_3_repeat1
H
13_2_3_repeat2
I
14
J
14_repeat1
K
14_repeat2
L
RMSD [nm]
Residue number

Figure S17

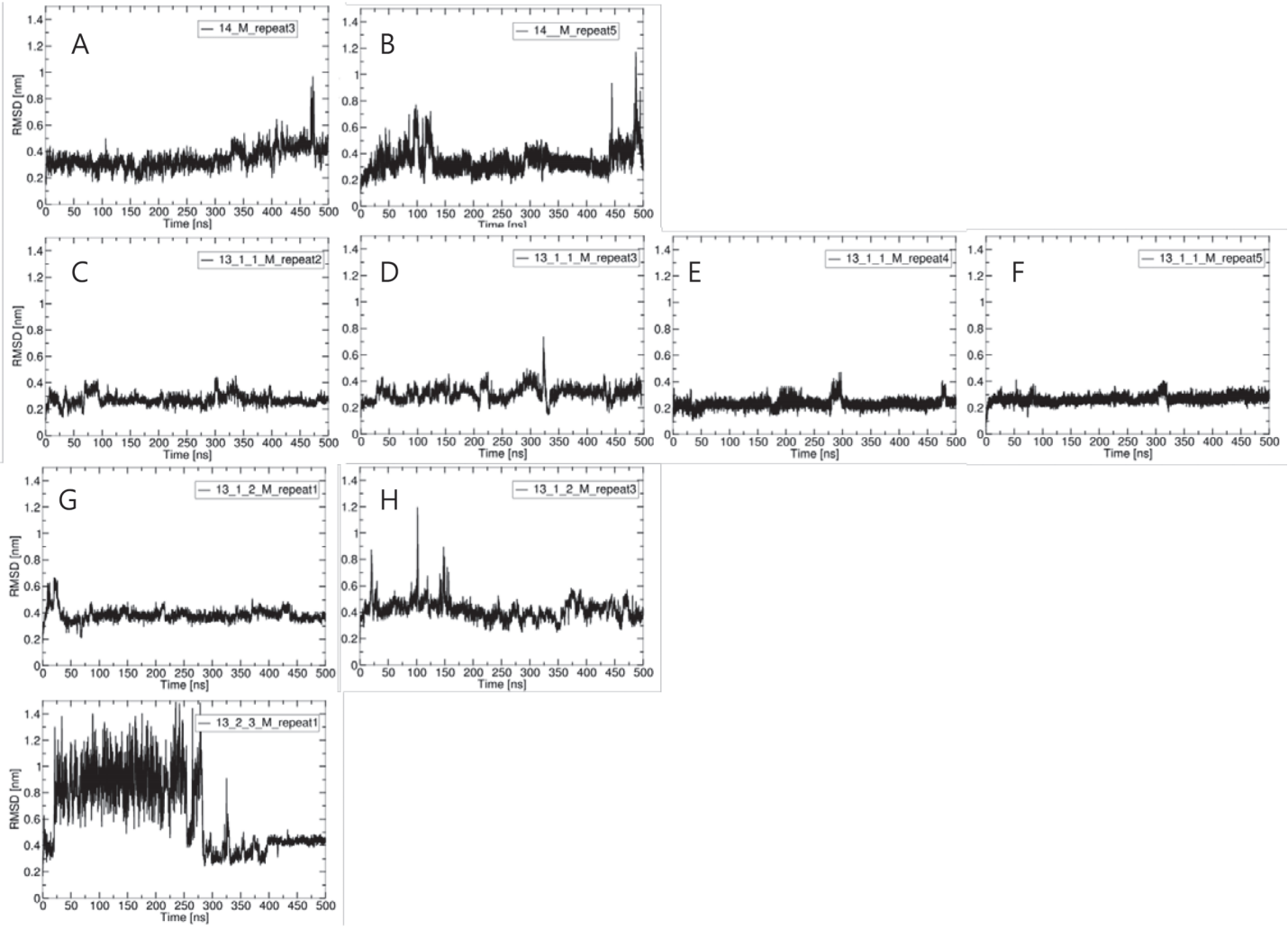

A
14_M_repeat3
B
14__M_repeat5
C
13_1_1_M_repeat2
D
13_1_1_M_repeat3
E
13_1_1_M_repeat4
F
13_1_1_M_repeat5
G
13_1_2_M_repeat1
H
13_1_2_M_repeat3
13_2_3_M_repeat1
RMSD [nm]
Time [ns]

Figure S18

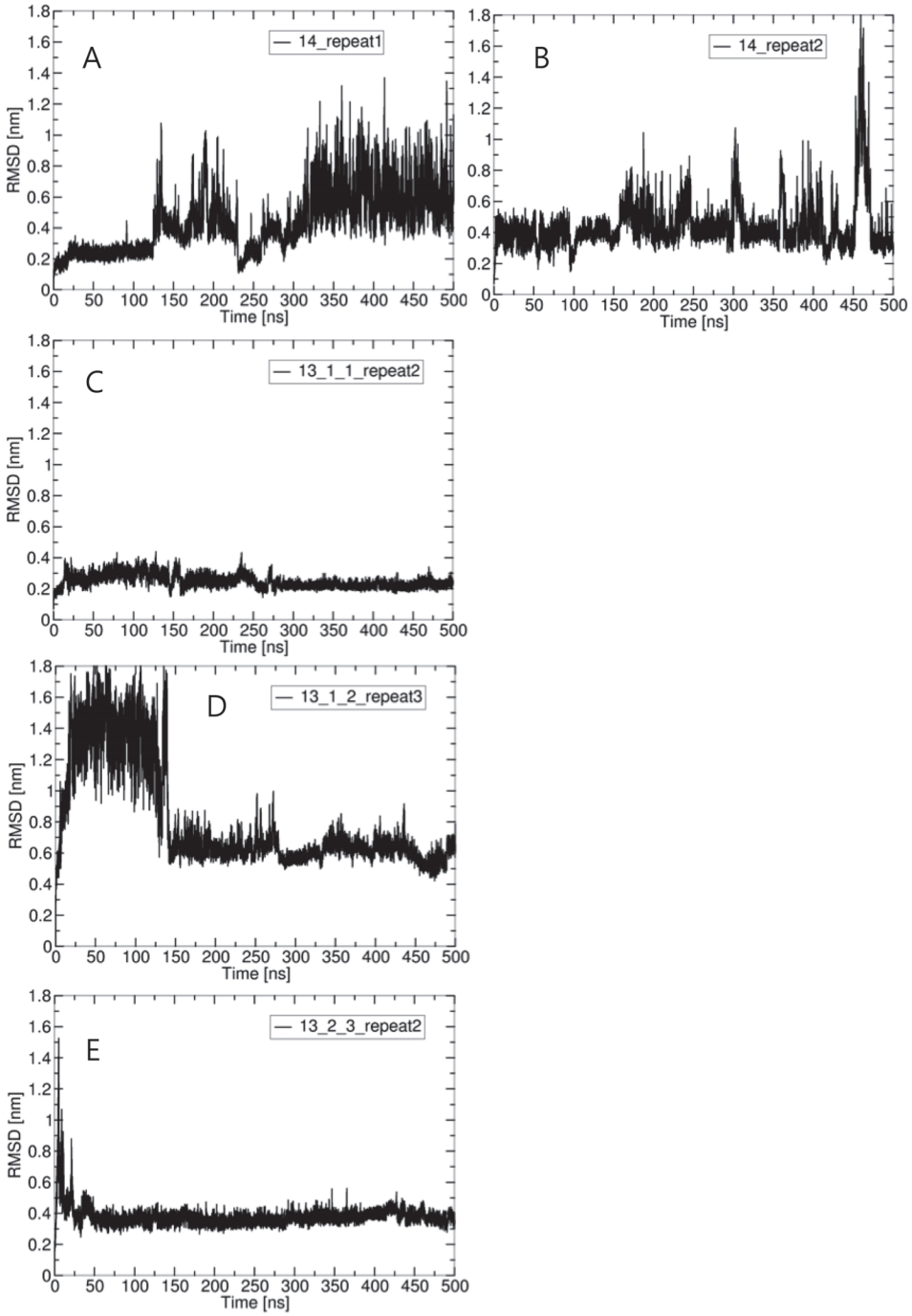

A
— 14_repeat1
B
— 14_repeat2
C
— 13_1_1_repeat2
D
— 13_1_2_repeat3
E
— 13_2_3_repeat2
RMSD [nm]
Time [ns]

Figure S19.

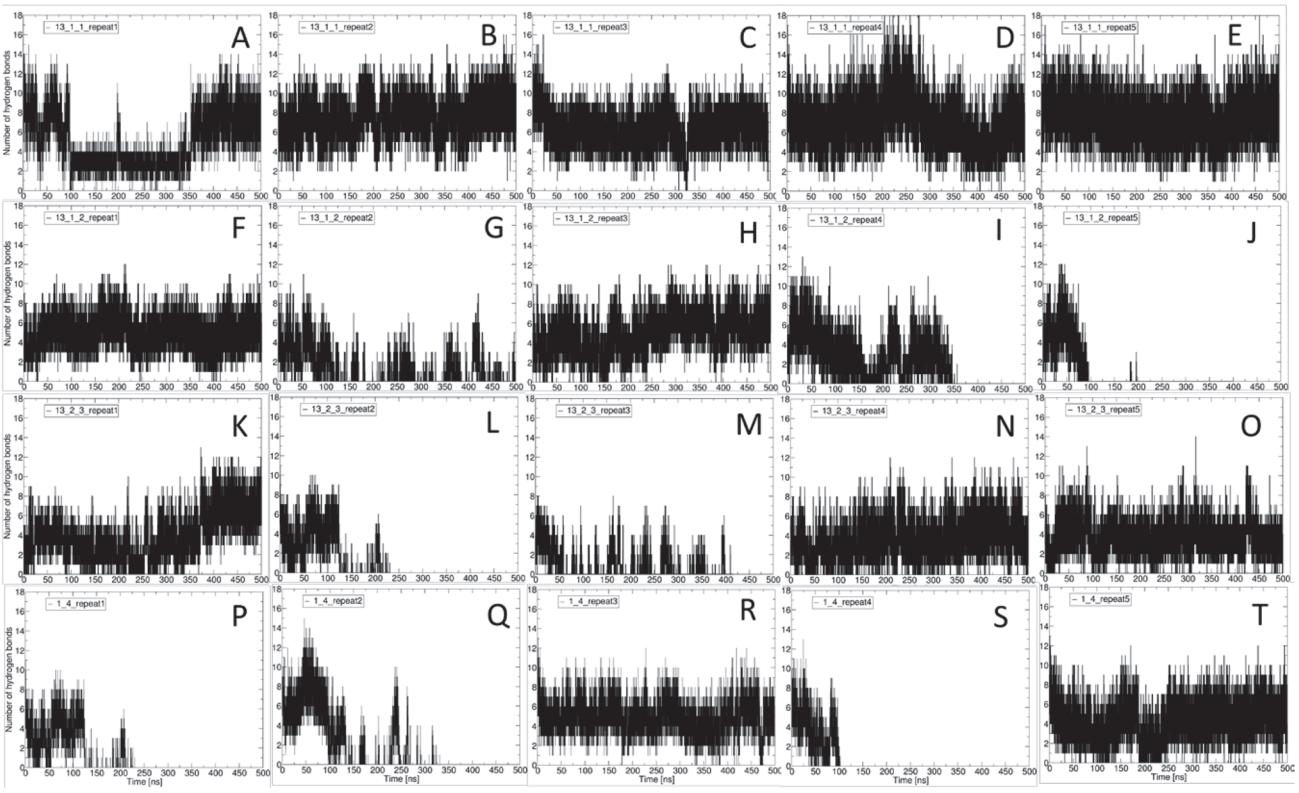

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
13_1_1_repeat1
13_1_1_repeat2
13_1_1_repeat3
13_1_1_repeat4
13_1_1_repeat5
13_1_2_repeat1
13_1_2_repeat2
13_1_2_repeat3
13_1_2_repeat4
13_1_2_repeat5
13_2_3_repeat1
13_2_3_repeat2
13_2_3_repeat3
13_2_3_repeat4
13_2_3_repeat5
1_4_repeat1
1_4_repeat2
1_4_repeat3
1_4_repeat4
1_4_repeat5
Number of hydrogen bonds
Time [ns]

Figure S20.

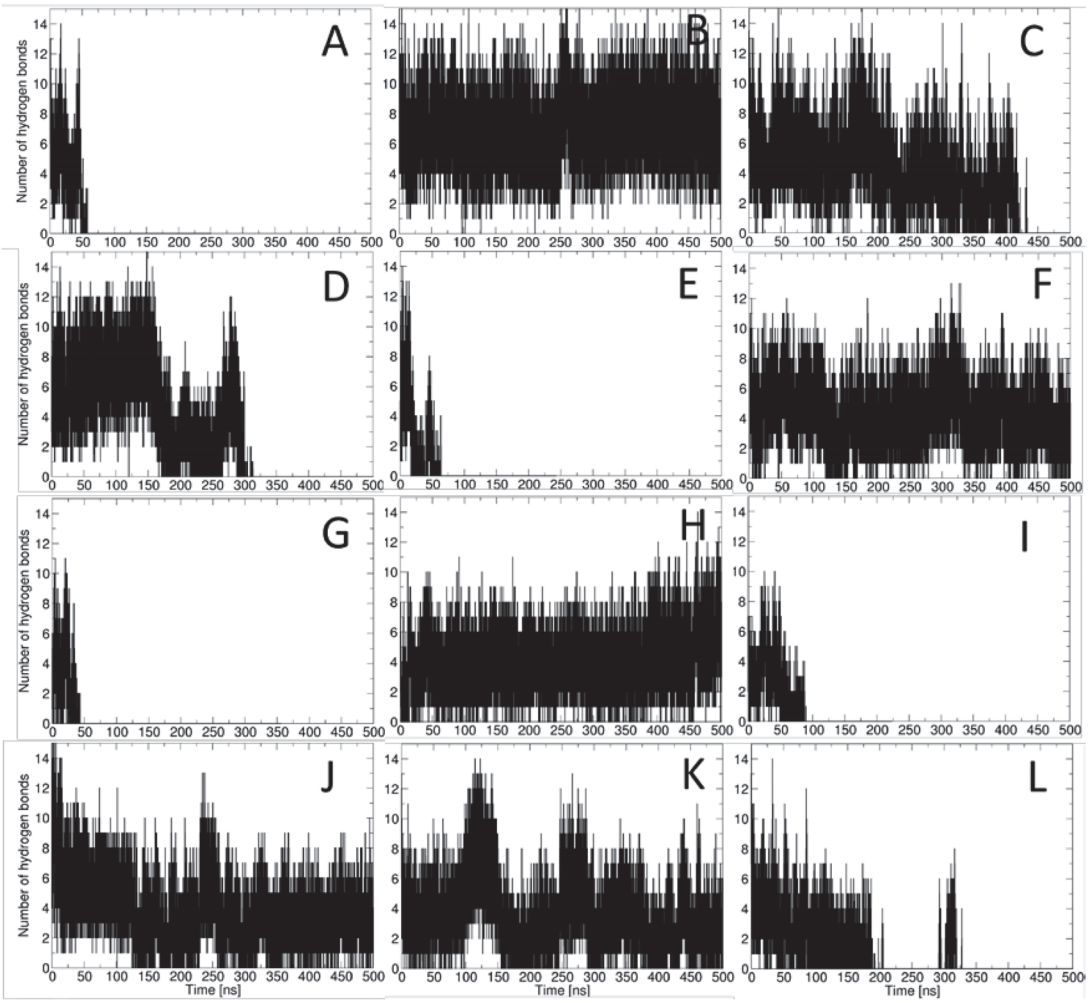


Figure S21..

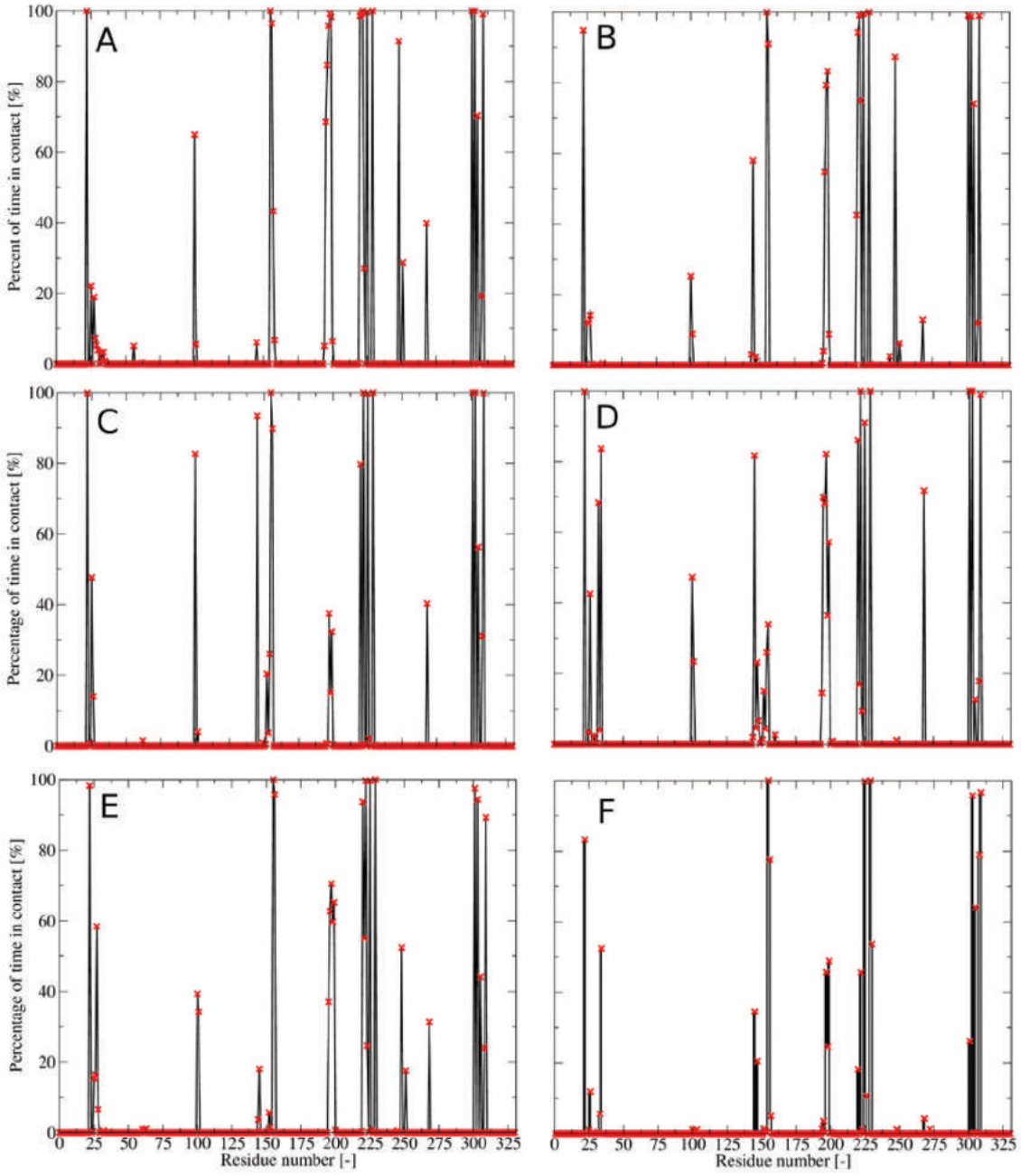

A
B
C
D
E
F
Percent of time in contact [%]
Percentage of time in contact [%]
Residue number [-]
0
20
40
60
80
100
0 25 50 75 100 125 150 175 200 225 250 275 300 325

Figure S22

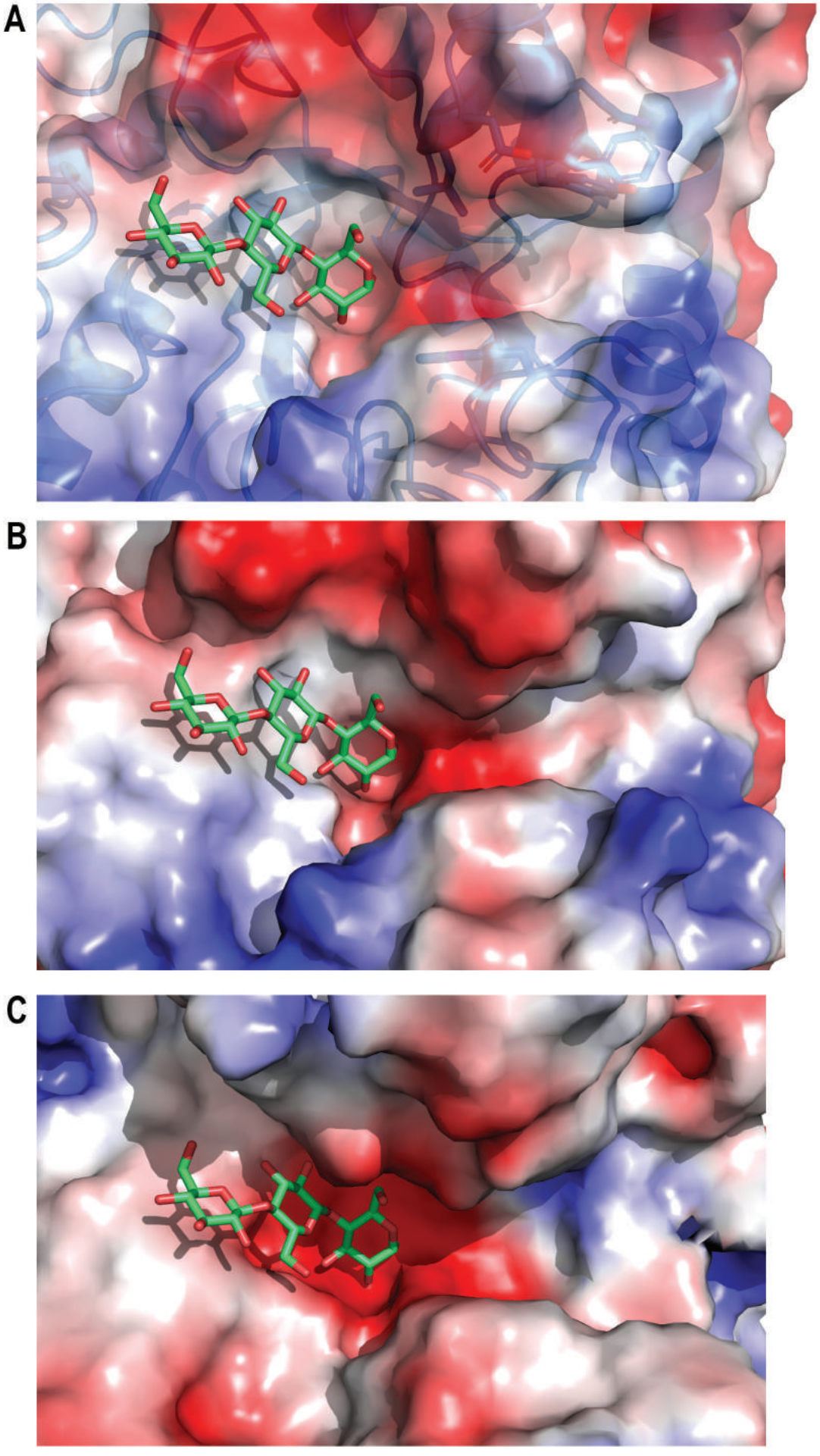


Table S1: Oligonucleotides used for the site-directed mutagenesis. The mutated codons for the punctual mutations are underlined. Mutations were performed suing the QuickChange II®XL site-directed mutagenesis kit according to the manufacturer's instructions.

Primer name	Oligonucleotide sequence (5'-> 3')
N77A_Forward	GGCCGGTTGGAACTTGGGC**GCG**GCAATGGATACCTATAACAG
N77A_Reverse	CTGTTATAGGTATCCATTGC**CGC**GCCCAAGTTCCAACCGGCC
N77Q_Forward	GCCGGTTGGAACTTGGGC**CAG**GCAATGGATACCTATAACA
N77Q_Reverse	TGTTATAGGTATCCATTGC**CTG**GCCCAAGTTCCAACCGGC
Y82A_Forward	GGCAATGCAATGGATACC**GCT**AACAGTGACGAGACGGC
Y82A_Reverse	GCCGTCTCGTCACTGTT**AGC**GGTATCCATTGCATTGCC
Y82L_Forward	GGCAATGCAATGGATACC**TTA**AACAGTGACGAGACGGCTT
Y82L_Reverse	AAGCCGTCTCGTCACTGTT**TAA**GGTATCCATTGCATTGCC
H156A_Forward	GATGTATGTGATCATAAATATACAC**GCT**GATGAAACATGGATCCTTCCTACC
H156A_Reverse	GGTAGGAAGGATCCATGTTTCATC**AGC**GTGTATATTTATGATCACATACATC
H156I_Forward	GATGTATGTGATCATAAATATACAC**ATT**GATGAAACATGGATCCTTCCTACC
H156I_Reverse	GGTAGGAAGGATCCATGTTTCATC**AAT**GTGTATATTTATGATCACATACATC
W210A_ Forward	GGGTACGCCCGAGGAA**GCG**AAAGGAGGTACACAA
W210A_Reverse	TTGTGTACCTCCTTT**CGC**TTCCTCGGGCGTACCC
W210F_Forward	AGGGTACGCCCGAGGAA**TTC**AAAGGAGGTACACAAG
W210F_Reverse	CTTGTGTACCTCCTTT**GAA**TTCCTCGGGCGTACCCT
K211A_Forward	GGTACGCCCGAGGAATGG**GCG**GGAGGTACACAAGAAGGC
K211A_Reverse	GCCTTCTTGTGTACCTCC**CGC**CCATTCCTCGGGCGTACC
Y280A_Forward	ATCGGTGCATAGCTATTTCCCT**GCG**CAGTTTTGTTTGGATGGAACG
Y280A_Reverse	CCGTTCCATCCAAACAAAACTG**CGC**AGGGAAATAGCTATGCACCGAT
Y280L_Forward	ATCGGTGCATAGCTATTTCCCT**CTG**CAGTTTTGTTTGGATGGAACGG
Y280L_Reverse	CCGTTCCATCCAAACAAAACTG**CAG**AGGGAAATAGCTATGCACCGAT
E323S_Forward	CCGTGGTCATGGGG**TCG**TGGGGCTCAACCT
E323S_Reverse	AGGTTGAGCCCCA**CGA**CCCCATGACCACGG
N358A_Forward	CATTTGTCCCATTTGGTGGGAT**GCC**GGGAATGTTGATGAG
N358A_Reverse	CTCATCAACATTCCC**GGC**ATCCCACCAAATGGGACAAATG
N358L_Forward	GGCATTTGTCCCATTTGGTGGGAT**CTA**GGGAATGTTGATGAGTTCG
N358L_Reverse	CGAACTCATCAACATTCCC**TAG**ATCCCACCAAATGGGACAAATGCC
E363A_Forward	GGTGGGATAACGGGAATGTTGAT**GCG**TTCGGTATTTTTAATAG
E363A_Reverse	CTATTAAAAATACCGAA**CGC**ATCAACATTCCCGTTATCCCACC
E363S_Forward	TGGTGGGATAACGGGAATGTTGAT**TCG**TTCGGTATTTTTAATAGAAATACC
E363S_Reverse	GGTATTTCTATTAAAAATACCGAA**CGA**ATCAACATTCCCGTTATCCCACCA

Table S2. List of the simulated systems with the number of simulations and sampling lengths (in ns) for mutated protein *Zg*EngA$_{GH5_4_E323S}$**.** Bound structures were sampled for approx. 500 ns (0.5 μs); simulations in which the substrate completely unbound were terminated early.

No	Name	Description	Number of simulations with stable bound glucan	Number of simulations	Lengths (ns)
1	**β-1,4_M**	Mutated protein *Zg*EngA$_{GH5_4_E323S}$ with 6 glucose units substrate with 1,4 linkages	2	5	232, 336, 500, 250, 521
2	**-1/+1 β-1,3_M**	Mutated protein *Zg*EngA$_{GH5_4_E323S}$ with 6 glucose units substrate with 1,3 link between -1 and +1 subsites	4	5	500, 500, 500, 515, 521
3	**+1/+2 β-1,3_M**	Mutated protein *Zg*EngA$_{GH5_4_E323S}$ with 6 glucose units substrate with 1,3 link between +1 and +2 subsites	2	5	500, 500, 500, 350, 200
4	**+2/+3 β-1,3_M**	Mutated protein *Zg*EngA$_{GH5_4_E323S}$ with 6 glucose units substrate with 1,3 link between +2 and +3 subsites	1	5	500, 500, 415, 510, 522

Table S3. List of the simulated systems with the number of simulations and lengths (in ns) for native protein *Zg*EngA$_{GH5_4}$.

No	Name	Description	Number of simulations with stable bound glucan	Number of simulations	Lengths (ns)
1	**β-1,4**	Protein with 6 glucose units substrate with 1,4 linkages	0	3	520, 520, 520
2	**-1/+1 β-1,3**	Protein with 6 glucose units substrate with 1,3 link between -1 and +1 subsites	1	3	250, 509, 535
3	**+1/+2 β-1,3**	Protein with 6 glucose units substrate with 1,3 link at position between +1 and +2 subsites	1	3	500, 242, 528
4	**+2/+3 β-1,3**	Protein with 6 glucose units substrate with 1,3 link between +2 and +3 subsites	1	3	355, 503, 223

Table S4. Composition of the simulated systems – mutated protein *Zg*EngA$_{GH5_4_E323S}$.

No	Name	Protein molecules	Substrate	K+	Cl-	Water molecules	Total number of atoms
1	**β-1,4_M**	1	1	68	48	15 660	52 391
2	**-1/+1 β-1,3_M**	1	1	68	48	15 664	52 403
3	**+1/+2 β-1,3_M**	1	1	66	46	15 189	50 974
4	**+2/+3 β-1,3_M**	1	1	66	46	15 193	50 986

Table S5. Composition of the simulated systems – native protein.

No	Name	Protein molecules	Substrate	K+	Cl-	Water molecules	Total number of atoms
1	**β-1,4**	1	1	68	48	15 660	52 420
2	**-1/+1 β-1,3**	1	1	69	48	15 666	52 414
3	**+1/+2 β-1,3**	1	1	69	48	15 669	52 423
4	**+2/+3 β-1,3**	1	1	69	48	15 193	52 438

Table S6. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation β-1,4 repeat 1. BGLC stands for the glucose unit (numbering 1 to 6 corresponds to +3 to -3 range). Values below 5% are not shown in the table.

Residue-atom	Glucan-atom	Occupancy [%]
TYR277-HN	BGLC1-O6	51.02
GLU303-OE2	BGLC1-HO1	46.84
LYS211-HZ1	BGLC1-O3	45.30
GLU303-OE1	BGLC1-HO1	31.82
SER276-OG	BGLC1-HO6	14.05
GLU303-OE2	BGLC1-HO1	13.48
TYR277-HN	BGLC2-O3	8.31
THR253-OG1	BGLC1-O6	7.48
TYR277-CD1	BGLC2-O2	5.61
TYR280-OH	BGLC1-HO6	5.14

Table S7. Percentage hydrogen bond occurrence during the last 200 ns of simulation β-1,4 repeat 2.

Residue-atom	Glucan-atom	Occupancy [%]
THR253-HN	BGLC1-O6	30.91
ASN77-HD22	BGLC5-O3	17.58
TRP356-HE1	BGLC5-O2	11.04
ALA251-O	BGLC2-HO2	10.46
ASN358-HD22	BGLC5-O3	10.24
GLU363-OE1	BGLC5-HO2	9.86
GLU363-OE2	BGLC5-HO2	9.22
GLU363-CD	BGLC5-HO3	8.95
ASP80-OD2	BGLC5-HO2	7.72
GLU200-OE1	BGLC4-HO2	7.58
TYR280-OH	BGLC1-H62	7.35
GLU363-OE2	BGLC5-HO3	7.05
GLU363-OE1	BGLC5-HO3	7.05
ASP80-OD1	BGLC5-HO3	6.96
TYR82-HN	BGLC6-O3	6.09
TRP210-O	BGLC1-H5	6.08
GLU363-CD	BGLC5-HO2	5.32

Table S8. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation -1/+1 β-1,3 repeat 2.

Residue-atom	Glucan-atom	Occupancy [%]
GLU323-OE1	BGLC4-HO2	90.97
ASN358-HD22	BGLC5-O3	50.58
ASN77-HD22	BGLC5-O3	49.49
TRP210-O	BGLC2-HO2	38.43
TRP356- HE1	BGLC5-O2	38.24
THR253- OG1	BGLC1-HO2	30.87
GLU303- OE2	BGLC1-HO2	26.30
GLU303- OE1	BGLC1-HO2	19.43
GLU303-OE2	BGLC1-HO1	18.91
GLU303- OE1	BGLC1-HO1	17.12
TYR277- HN	BGLC2-O6	16.10
ASN360- HD22	BGLC6-O6	12.17
THR253- HN	BGLC1-O3	11.79
ASN83- HD22	BGLC6-O4	5.06

Table S9. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation +1/+2 β-1,3 repeat 3.

Residue-atom	Glucan-atom	Occupancy [%]
TYR277-HN	BGLC1-O6	86.68
LYS211-HZ3	BGLC1-O3	29.77
THR253-OG1	BGLC1-HO1	22.61
GLU200-OE1	BGLC3-HO2	17.32
TRP356-HE1	BGLC4-O2	16.67
TYR82-O	BGLC6-HO3	13.27
ASN358-HD22	BGLC4-O3	11.49
GLU303-OE2	BGLC1-HO1	9.85
GLU200-OE2	BGLC3-HO2	8.76
GLU303-OE1	BGLC1-HO1	8.52
SER252-OG	BGLC1-H1	7.17
TRP210-O	BGLC1-H3	7.15
ASN77-HD22	BGLC6-O2	7.00
TRP210-CZ3	BGLC3-HO2	6.36
TYR277-HD2	BGLC2-O2	6.05
LYS211-HZ3	BGLC1-O2	5.91
THR81-OG1	BGLC6-HO2	5.84
TYR280-OH	BGLC1-H2	5.72
ASP80-OD1	BGLC5-HO6	5.36
SER276-HA	BGLC1-O6	5.01

Table S10. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation +2/+3 β-1,3 repeat 2.

Residue-atom	Glucan-atom	Occupancy [%]
ASN77-HD22	BGLC5-O3	64.45
ASN358-HD22	BGLC5-O3	53.55
TRP356-HE1	BGLC5-O2	51.59
ASN360-HD22	BGLC6-O6	19.62
GLU323-OE1	BGLC4-HO2	19.47
THR253-OG1	BGLC1-HO1	12.27
GLU200-OE1	BGLC4-HO2	11.93
THR253-HN	BGLC1-O2	10.36
GLU200-OE2	BGLC4-HO2	9.29
ASN358-HD22	BGLC5-O2	8.91
SER252-OG	BGLC1-HO2	8.30
HSD155-HE2	BGLC5-O6	6.51
TRP356-HE1	BGLC5-C2	6.40
TRP356-HE1	BGLC4-O4	6.32
THR253-HN	BGLC2-O2	5.58
ASN77-OD1	BGLC5-HO3	5.04

Table S11. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation β-1,4_M repeat 3.

Residue-atom	Glucan-atom	Occupancy [%]
GLU363-CD	BGLC5-HO3	32.13
GLU363-OE1	BGLC5-HO3	27.65
GLU363-OE2	BGLC5-HO3	29.36
GLU363-OE1	BGLC5-HO2	28.31
GLU363-OE2	BGLC5-HO2	26.19
ASP285-OD2	BGLC3-HO3	26.09
ASP285-CG	BGLC3-O2	23.03
ASP285-OD1	BGLC3-HO2	21.32
ASN77-HD22	BGLC5-O3	19.86
ASP285-OD1	BGLC3-HO3	19.16
ASP285-OD2	BGLC3-HO2	18.50
GLU363-CD	BGLC5-HO2	14.48
ASN358-HD22	BGLC5-O3	13.22
ASN360-HD22	BGLC6-O6	12.22
GLU363-OE1	BGLC6-HO6	10.86
TRP356-HE1	BGLC5-O2	9.10
GLU363-OE2	BGLC6-HO6	8.65
THR253-HN	BGLC1-O6	7.34
GLU200-OE1	BGLC4-HO2	7.04
TYR280-OH	BGLC1-HO6	6.39
GLU200-OE2	BGLC4-HO2	6.33
SER252-OG	BGLC1-HO6	6.03
GLU200-OE1	BGLC3-HO6	5.03

Table S12. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation β-1,4_M repeat 5.

Residue-atom	Glucan-atom	Occupancy [%]
TRP356-HE1	BGLC5-O2	49.76
ASN358-HD22	BGLC5-O3	40.01
GLU200-OE1	BGLC4-HO2	34.02
ASN77-OD1	BGLC5-HO3	33.72
GLU200-OE2	BGLC3-HO6	33.16
GLU200-CD	BGLC4-HO2	30.23
GLU200-OE1	BGLC3-HO6	21.25
GLU200-OE2	BGLC4-HO2	21.00
SER252-OG	BGLC2-HO2	19.32
TRP210-O	BGLC2-HO6	16.89
THR253-HN	BGLC2-O2	15.32
GLU363-OE2	BGLC5-HO2	12.69
TYR280-OH	BGLC1-HO6	10.67
GLU363-OE1	BGLC5-HO3	10.36
ASN77-HD22	BGLC5-O3	10.06
GLU200-CD	BGLC3-HO6	9.39
ASP285-CG	BGLC3-HO2	9.11
GLU363-CD	BGLC5-HO3	9.10
ASP285-OD2	BGLC3-HO2	8.83
SER252-OG	BGLC1-HO6	8.29
TRP210-O	BGLC1-HO3	8.19
ASP285-OD2	BGLC3-HO3	8.10
THR253-HN	BGLC1-O6	7.59
ASP285-OD1	BGLC3-HO2	7.42
LEU284-CD1	BGLC3-HO2	7.40
ASN360-HD22	BGLC6-O6	6.39
GLU363-OE1	BGLC5-HO2	6.29
GLU363-OE2	BGLC5-HO3	6.10
ASP285-OD1	BGLC3-HO3	6.05
GLU363-CD	BGLC5-HO2	5.83

Table S13. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation -1/+1 β-1,3_M repeat 2.

Residue-atom	Glucan-atom	Occupancy [%]
TRP356-HE1	BGLC5-O2	80.15
ASN358-HD22	BGLC5-O3	57.75
ASN77-HD22	BGLC5-O3	54.85
TRP210-O	BGLC2-HO2	49.80
TYR277-HN	BGLC2-O6	43.20
THR253-HN	BGLC1-O1	37.20
THR253-OG1	BGLC1-HO2	28.05
SER252-OG	BGLC1-HO5	18.00
GLU303-OE1	BGLC1-HO2	17.65
ASN360-HD22	BGLC6-O6	14.85
TRP356-CZ2	BGLC4-HO3	14.00
LYS211-O	BGLC1-HO6	13.75
HSD275-HD1	BGLC3-O2	12.50
GLU303-OE2	BGLC1-HO2	12.45
TRP356-HE1	BGLC5-C2	10.40
THR253-HN	BGLC1-O3	10.30
TRP356-CH2	BGLC4-HO3	8.45
ALA250-O	BGLC2-H1	7.95
HSD275-O	BGLC2-HO6	7.45
TYR280-OH	BGLC1-HO3	7.15
ASN77-OD1	BGLC5-HO3	7.10
THR253-OG1	BGLC1-HO2	7.10
ASN358-HD22	BGLC5-O2	7.10
GLU303-CD	BGLC1-HO2	6.65
ASN77-HD22	BGLC5-O6	6.65
SER252-HA	BGLC1-O5	6.40
SER323-OG	BGLC4-HO2	6.15
SER252-OG	BGLC1-HO1	5.80
SER252-OG	BGLC1-H1	5.20

Table S14. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation -1/+1 β-1,3_M repeat 3.

Residue-atom	Glucan-atom	Occupancy [%]
TYR277-HN	BGLC2-O6	73.12
TRP356-HE1	BGLC5-O2	70.77
ASN358-HD22	BGLC5-O3	59.48
ASN77-OD1	BGLC5-HO3	38.27
GLU303-OE1	BGLC1-HO2	31.12
GLU303-OE2	BGLC1-HO2	27.49
TYR280-OH	BGLC3-HO6	24.32
ASN77-HD22	BGLC5-O3	24.17
ASN360-HD22	BGLC6-O6	22.07
GLU200-OE1	BGLC4-HO2	21.61
TYR280-OH	BGLC3-HO6	21.72
GLU200-OE2	BGLC4-HO2	17.88
GLU303-CD	BGLC1-HO2	16.45
TYR277-HN	BGLC2-C6	8.38
ASN77-HD22	BGLC5-O6	7.41
SER276-HA	BGLC2-O6	6.39
TRP356-CZ2	BGLC4-HO3	5.88
TYR277-OH	BGLC4-HO2	5.67
ASN358-HD22	BGLC5-O2	5.52
TRP356-HE1	BGLC5-C2	5.01

Table S15. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation -1/+1 β-1,3_M repeat 4.

Residue-atom	Glucan-atom	Occupancy [%]
TRP356-HE1	BGLC5-O2	82.28
HSD275-HD1	BGLC3-O2	68.94
ASN77-HD22	BGLC5-O3	68.20
ASN358-HD22	BGLC5-O3	63.49
THR253-HN	BGLC1-O3	56.09
GLU303-OE2	BGLC1-HO2	29.71
GLU303-OE1	BGLC1-HO2	25.70
THR253-OG1	BGLC1-HO2	23.33
ASN360-HD22	BGLC6-O6	22.68
TRP356-CZ2	BGLC4-HO3	18.78
GLU209-OE1	BGLC2-HO2	17.47
GLU209-OE2	BGLC2-HO2	15.20
TRP356-CH2	BGLC4-HO3	12.33
SER252-OG	BGLC1-H4	10.60
TYR277-HN	BGLC2-O6	10.15
ASN77-OD1	BGLC5-HO3	8.55
GLU209-O	BGLC1-HO6	7.90
GLU303-CD	BGLC1-HO2	7.34
TRP356-HE1	BGLC5-C2	6.70
GLU209-CD	BGLC2-HO2	6.11
HSD155-CE1	BGLC5-HO6	6.03
GLU303-OE1	BGLC1-HO1	5.95
ASN77-OD1	BGLC5-H4	5.28
ASN358-HD22	BGLC5-O2	5.00

Table S16. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation -1/+1 β-1,3_M repeat 5.

Residue-atom	Glucan-atom	Occupancy [%]
THR253-HN	BGLC1-O1	80.93
ASN358-HD22	BGLC5-O3	66.32
TRP356-HE1	BGLC5-O2	56.88
THR253-OG1	BGLC1-HO2	55.36
TYR82-H	BGLC5-O6	50.22
GLU200-OE2	BGLC4-HO2	47.34
TYR277-HN	BGLC2-O6	46.11
GLU200-OE1	BGLC4-HO2	39.85
GLU200-CD	BGLC4-HO2	24.73
SER252-C1	BGLC1-HG1	22.07
GLU200-OE2	BGLC3-HO2	12.51
ASN83-ND2	BGLC6-O3	12.16
GLU200-OE1	BGLC3-HO2	11.24
ASN360-ND2	BGLC6-O6	10.72
SER252-HG1	BGLC1-O5	10.39
TRP210-O	BGLC1-HO6	9.69
TYR82-H	BGLC5-O6	8.71
ALA250-O	BGLC2-H1	8.41
TYR280-H	BGLC1-O3	7.97
TRP356-HE1	BGLC5-C2	7.38
ASN358-O	BGLC6-HO6	6.68
TYR280-OH	BGLC1-HO3	6.49
ALA250-O	BGLC1-HO6	6.33
ASN83-HD22	BGLC6-O3	6.25
ASN77-HD21	BGLC5-O3	6.14
TRP356-CZ2	BGLC4-HO3	5.91
TRP356-HE1	BGLC4-O4	5.63
HSD155-NE2	BGLC5-HO6	5.59
ALA254-HN	BGLC1-O1	5.38
TYR280-HE2	BGLC2-O6	5.37

Table S17. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation +1/+2 β-1,3_M repeat 1.

Residue-atom	Glucan-atom	Occupancy [%]
ASN358-HD2	BGLC5-O3	50.17
TRP356-HE1	BGLC5-O2	46.00
GLU200-OE1	BGLC4-HO2	38.06
GLU200-CD	BGLC4-HO2	31.07
GLU200-OE2	BGLC4-O2	29.43
ASN77-HD22	BGLC5-O3	27.94
TRP356-CZ2	BGLC4-HO3	21.49
SER252-OG	BGLC1-HO6	16.87
TRP356-CE2	BGLC4-HO3	12.16
ASP80-OD2	BGLC6-HO2	11.17
ASN358-HD22	BGLC5-O2	10.07
HSD155-HE2	BGLC5-O6	9.78
TRP356-HE1	BGLC5-C2	8.39
ASP80-OD1	BGLC6-HO2	8.29
GLU363-OE2	BGLC6-HO6	8.24
TRP356-HE1	BGLC4-O4	7.79
GLU363-CD	BGLC6-HO6	7.69
ASN77-OD1	BGLC5-HO3	7.59
ASP80-OD1	BGLC6-HO3	7.54
GLU363-OE1	BGLC6-HO6	7.39
ASP80-OD2	BGLC6-HO3	7.00
ASP80-CG	BGLC6-HO3	6.85
TRP356-CH2	BGLC4-HO3	6.15
TYR277-OH	BGLC4-HO2	5.51
ASN360-HD22	BGLC6-O6	5.31

Table S18. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation +1/+2 β-1,3_M repeat 3.

Residue-atom	Glucan-atom	Occupancy [%]
ASN77-HD22	BGLC5-O3	77.45
TRP356-HE1	BGLC5-O2	74.30
ASN358-HD22	BGLC5-O3	60.35
HSD275-HD1	BGLC3-O3	22.65
GLU200-OE1	BGLC3-HO3	22.35
HSD275-HD1	BGLC3-O6	21.55
GLU200-OE2	BGLC3-HO3	19.25
THR87-OG1	BGLC6-HO4	18.50
GLU200-CD	BGLC3-HO3	16.85
SER252-OG	BGLC1-HO6	13.20
SER252-OG	BGLC2-HO6	11.75
GLU200-CD	BGLC3-HO2	9.75
GLU200-OE2	BGLC3-HO2	9.00
THR87-OG1	BGLC6-HO4	8.30
TRP356-CZ2	BGLC4-HO3	8.15
GLU200-OE1	BGLC3-HO2	8.05
GLU200-OE1	BGLC3-HO6	6.05
GLU200-OE2	BGLC3-HO6	5.80
THR253-HN	BGLC2-O6	5.35

Table S19. Percentage of hydrogen bond occurrence during the last 200 ns of the simulation +2/+3 β-1,3_M repeat 1.

Residue-atom	Glucan-atom	Occupancy [%]
ASN77-HD22	BGLC5-O3	70.81
TRP356-HE1	BGLC5-O2	58.92
ASN358-HD22	BGLC5-O3	53.77
HSD275-HD1	BGLC3-O6	48.73
HSD275-O	BGLC2-HO2	29.99
GLU303-CD	BGLC1-HO1	17.54
THR253-OG1	BGLC1-HO2	17.19
HSD275-HD1	BGLC3-C6	14.99
GLU303-OE1	BGLC1-HO1	12.74
GLU303-OE2	BGLC1-HO1	12.24
ASN77-OD1	BGLC5-H4	9.20
SER276-HA	BGLC2-O2	8.85
TYR277-HN	BGLC2-C3	8.45
ASN360-HD22	BGLC6-O6	7.90
GLU363-OE2	BGLC6-HO6	6.45
ASN358-HD22	BGLC5-O2	6.15
TRP210-O	BGLC1-HO4	5.90
GLU363-OE1	BGLC6-HO6	5.65
TYR277-HN	BGLC2-O3	5.15
ASP80-OD2	BGLC6-HO4	5.05

Table S20. Average number of hydrogen bonds between glucose chain and protein – mutated system E323S.

Simulation Name	Average number of hydrogen bonds and standard deviation	Time used in analyses (ns)
β-1,4_M (3, 5)	**4.44** ± 1.69; **5.97** ± 2.34	300 - 500
-1/+1 β-1,3_M (2, 5)	**7.90** ± 1.89; **6.32** ± 1.64; **6.69** ± 1.99; **7.63** ± 2.05	300 - 500
+1/+2 β-1,3_M (1, 3)	**4.85** ± 1.48; **6.05** ± 1.56	300 - 500
+2/+3 β-1,3_M (1)	**5.53** ± 2.22	300 - 500

Table S21. Average number of hydrogen bonds between glucose chain and protein – native system.

Simulation Name	Average number of hydrogen bonds and standard deviation	Time used in analyses (ns)
β-1,4 (1, 2)	**3.07** ± 1.43; **2.70** ± 1.71	300 - 500
-1/+1 β-1,3 (2)	**7.31** ± 1.96	300 - 500
+1/+2 β-1,3 (3)	**4.00** ± 1.84	300 - 500
+2/+3 β-1,3 (2)	**4.26** ± 1.85	300 - 500

Table S22. Percentage of occurrence of interatomic contacts < 0.35 nm in the last 100 ns of the simulations.

Residue number	β-1,4_I	β-1,4_II	-1/+1 β-1,3_II	+1/+2 β-1,3_III	+2/+3 β-1,3_II	Average
ASN77	14	47	97	71	100	66
ASP80	0	19	3	12	8	9
THR81	11	34	14	80	37	35
TYR82	2	36	8	41	0	18
THR87	0	1	0	0	14	3
TRP89	6	1	16	0	6	6
HSD155	5	17	3	49	95	34
HSD156	26	0	0	49	27	20
GLU200	6	10	2	68	53	28
ARG202	2	5	0	9	4	4
PRO207	0	0	0	0	0	0.
GLU209	0	0	0	0	24	5
TRP210	100	100	99	100	100	100
LYS211	99	98	99	95	10	80
GLY212	0	0	1	0	0	0
TYR249	0	0	0	0	0	0
ALA250	0	9	90	0	7	21
ALA251	1	26	71	0	47	29
SER252	100	96	100	90	56	88
THR253	99	90	100	89	47	85
ALA254	99	92	100	89	49	86
HSD275	0	12	100	88	31	46
SER276	94	4	98	98	1	59
TYR277	100	74	100	100	100	95
PHE278	94	6	9	87	2	40
TYR280	100	100	100	99	99	100
LEU284	100	99	100	100	100	100
ASP285	8	7	0	0	0	3
GLU303	92	7	99	55	1	51
LYS306	12	1	41	1	5	12
SER323	1	3	99	14	74	38
TRP356	0	8	100	32	100	48
ASN358	3	35	100	39	100	55
ASN360	8	24	72	7	84	39
GLU363	11	37	24	5	17	19
PHE364	4	23	100	42	100	54

Table S23. Percentage of occurrence of interatomic contacts < 0.35 nm in the last 100 ns of the mutated *Zg*EngA$_{GH5_4_E323S}$ simulations.

Residue number	β-1,4 III_M	β-1,4 V_M	-1/+1 β-1,3 II_M	-1/+1 β-1,3 III_M	-1/+1 β-1,3 IV_M	-1/+1 β-1,3 V_M	+1/+2 β-1,3 _I_M	+1/+2 β-1,3 III_M	+2/+3 β-1,3 I_M	Average
ASN77	83	83	100	95	98	93	100	100	98	94
ASP80	1	2	22	12	8	9	48	3	15	13
THR81	12	39	0	12	28	62	14	43	16	2
TYR82	0	4	19	14	0	97	0	0	58	21
THR87	0	0	3	0	0	2	0	68	0	8
TRP89	52	0	3	0	9	0	0	84	0	16
HSD155	1	8	65	25	75	88	83	47	39	48
HSD156	1	7	6	9	37	2	0	23	34	13
GLU200	35	64	6	58	46	99	93	82	18	56
ARG202	21	3	0	2	3	0	0	23	0	6
PRO207	1	0	0	0	7	0	20	15	6	5
GLU209	0	1	0	0	45	0	26	26	0	11
TRP210	100	100	100	100	42	100	100	34	100	86
LYS211	78	93	96	91	12	36	90	0	96	66
GLY212	5	6	43	0	22	0	0	0	0	8
TYR249	1	0	5	0	0	0	0	15	0	2
ALA250	4	54	69	1	63	100	1	70	37	44
ALA251	0	44	85	4	70	100	1	68	63	48
SER252	46	83	96	55	98	100	37	82	71	74
THR253	25	84	99	79	99	100	15	36	60	66
ALA254	49	90	98	83	100	99	32	57	65	75
HSD275	18	59	99	42	97	97	80	86	94	75
SER276	0	0	99	94	96	99	0	17	55	51
TYR277	46	77	100	99	100	100	100	100	100	91
PHE278	1	2	27	75	8	28	0	9	25	19
TYR280	100	99	100	99	100	100	100	91	100	99
LEU284	100	100	100	100	100	100	100	100	100	100
ASP285	54	25	0	0	0	0	0	0	0	9
GLU303	1	2	91	87	91	95	0	1	52	47
LYS306	0	3	29	6	34	1	0	0	18	10
SER323	4	0	40	13	35	11	40	72	31	27
TRP356	26	61	100	99	100	100	100	100	98	87
ASN358	96	84	100	99	100	100	100	100	94	97
ASN360	64	48	70	74	79	32	56	13	44	53
GLU363	79	35	19	12	19	17	31	18	24	28
PHE364	97	82	99	99	99	100	100	99	89	96

Table S24. Calculated binding energies (kcal/mol) between glucan chain and protein.

System		Components[a]				
		ΔE_{vdw}	ΔE_{ele}	ΔG_{polar}	$\Delta G_{nonpolar}$	ΔG_{bind}
Native ZgEngA$_{GH5_4}$	-1/+1 β-1,3_II	-44.2 ± 3.4	-48.5 ± 11.7	72.3 ± 7.6	-6.9 ± 0.3	-27.3 ± 8.5
	+1/+2 β-1,3_III	-38.5 ± 4.2	-32.1 ± 10.5	59.7 ± 10.2	-6.5 ± 0.6	-17.4 ± 8.4
	+2/+3 β-1,3_II	-38.8 ± 3.9	-39.4 ± 11.3	61.7 ± 9.5	-6.4 ± 0.5	-22.9 ± 8.6
Mutant ZgEngA$_{GH5_4_E323S}$	β-1,4_III_M	-32.9 ± 6.8	-29.4 ± 15.6	41.8 ± 12.2	-6.7 ± 0.9	-27.3 ± 12.8
	β-1,4_V_M	-33.1 ± 6.4	-32.5 ± 10.5	43.2 ± 11.4	-5.9 ± 0.9	-28.3 ± 11.8
	-1/+1 β-1,3_II_M	-41.0 ± 4.2	-42.1 ± 14.6	44.6 ± 9.1	-3.3 ± 1.7	-41.8 ± 9.2
	-1/+1 β-1,3_III_M	-40.3 ± 4.7	-55.0 ± 13.7	57.5 ± 8.9	-6.9 ± 0.5	-44.6 ± 9.2
	-1/+1 β-1,3_IV_M	-43.4 ± 3.5	-49.5 ± 11.0	56.8 ± 7.3	-6.8 ± 0.4	-42.9 ± 8.2
	-1/+1 β-1,3_V_M	-46.8 ± 3.8	-76.3 ± 13.6	88.9 ± 8.3	-7.3 ± 0.4	-41.5 ± 9.3
	+1/+2 β-1,3_I_M	-38.7 ± 4.0	-46.5 ± 12.8	54.0 ± 8.5	-6.4 ± 0.6	-37.6 ± 9.8
	+1/+2 β-1,3_III_M	-34.0 ± 6.7	-46.1 ± 12.8	48.1 ± 11.6	-5.7 ± 0.9	-37.6 ± 12.7
	+2/+3 β-1,3_I_M	-41.2 ± 7.5	-44.8 ± 13.2	50.2 ± 12.1	-6.4 ± 0.8	-42.3 ± 12.6

Notes: [a]ΔE_{vdw} van der Waals contribution; ΔE_{ele}, electrostatic contribution; the sum of ΔE_{vdw} and ΔE_{ele} represent gas-phase energy; ΔG_{polar}, polar solvation energy; $\Delta G_{nonpolar}$, non-polar solar energy; the sum of ΔG_{polar} and $\Delta G_{nonpolar}$ is the solvation free energy; $\Delta G_{bind} = \Delta E_{ele} + \Delta E_{vdw} + \Delta G_{polar} + \Delta G_{nonpolar}$. Error values were obtained by calculating standard deviation.

Table S25: List of the sequences used in the phylogenetic analysis (Figure 2).

Label	Organism	Accession numbers	PDB code	References
ZgEngA-GH5_4	Zobellia galactanivorans Dsij	CAZ94281.1	6GL2; 6GL0	This work
GH5 4 Zobellia uliginosa	Zobellia uliginosa	SIT07898.1		
GH5 4 Pseudozobellia thermophila	Pseudozobellia thermophila	WP_072991460.1		
GH5 4 Maribacter dokdonensis	Maribacter dokdonensis	WP_074674385.1		
GH5 4 Maribacter forsetii	Maribacter forsetii	WP_051941839.1		
GH5 4 Maribacter aquivivus	Maribacter aquivivus	WP_073245446.1		
GH5 4 Hyunsoonleella jejuensis	Hyunsoonleella jejuensis	SEQ04964.1		
GH5 4 Flagellimonas DK169	Flagellimonas DK169	WP_055393410.1		
GH5 4 Croceitalea dokdonensis	Croceitalea dokdonensis	WP_054560255.1		
GH5 4 Saccharicrinis fermentans	Saccharicrinis fermentans	GAF03776.1		
GH5 4 Labilibacter marinus	Labilibacter marinus	WP_075590947.1		
GH5 4 Dokdonia MED134	Dokdonia sp. MED134	WP_021778202.1		
GH5 4 Dokdonia donghaensis	Dokdonia donghaensis	WP_052111791.1		
GH5 4 Flexithrix dorotheae	Flexithrix dorotheae	WP_020529897.1		
GH5 4 Labilibacter aurantiacus	Labilibacter aurantiacus	WP_068475339.1		
GH5 4 Algibacter SK-16	Algibacter sp. SK-16	WP_069830916.1		
GH5 4 Duganella CF402	Duganella sp. CF402	SEM71636.1		
ClcelA	Clostridium longisporum	P54937.1		Mittendorf and Thomson (1993) J. Gen. Microbiol. 139: 3233-3242
GH5 4 Clostridium KNHs205	Clostridium sp. KNHs205	WP_051685496.1		
RtCel5C_pdb-4IM4	Ruminiclostridium thermocellum	AAA23224.1	4IM4	Walker *et al* (2015) Biotechnol Biofuels 8: 220
GH5 4 Pseudobacteroides cellulosolvens	Pseudobacteroides cellulosolvens	KNY25463.1		
GH5 4 Acetivibrio cellulolyticus	Acetivibrio cellulolyticus	WP_010248927.1		
GH5 4 Clostridium pasteurianum	Clostridium pasteurianum	WP_066020423.1		
GH5 4 Clostridium acetobutylicum	Clostridium acetobutylicum	WP_010964144.1		
GH5 4 Clostridium roseum	Clostridium roseum	WP_077832505.1		
GH5 4 Clostridium cellulovorans	Clostridium cellulovorans	WP_010076241.1		
CcEngD_pdb-3NDY	Clostridium cellulovorans	AAA23233.1	3NDY	Bianchetti *et al* (2013) J. Mol. Biol. 425: 4267-4285
GH5 4 Clostridium saccharoperbutylacetonicum	Clostridium saccharoperbutylacetonicum	WP_015391601.1		
GH5 4 Clostridium puniceum	Clostridium puniceum	WP_077846784.1		
GH5 4 Ruminococcus champanellensis	Ruminococcus champanellensis	WP_054683931.1		
BpCel5C_pdb-4NF7	Butyrivibrio proteoclasticus B316	ADL34447.1	4NF7	No reference
CcCel5A_pdb-1EDG	Clostridium cellulolyticum	AAA23221.1	1EDG	Ducros *et al* (1995) Structure 3: 939-949
GH5 4 Ruminococcus CAG:353	Ruminococcus sp. CAG:353	CDE80894.1		
GH5 4 Lachnoclostridium phytofermentans	Lachnoclostridium phytofermentans	CDE80894.1		
F32EG5_pdb-4X0V	Caldicellulosiruptor sp. F32	AGM71677.1	4X0V	Meng *et al* (2017) Biochem. J. 474(20): 3373-3389
GH5 4 Herbinix hemicellulosilytica	Herbinix hemicellulosilytica	CRZ35717.1		
GH5 4 Ruminococcus albus	Ruminococcus albus KH2T6	A0A1H7KSB4,1		
GH5 4 Butyrivibrio INlla18	Butyrivibrio sp. INlla18	A0A1G5WIJ8		
GH5 4 Butyrivibrio hungatei	Butyrivibrio hungatei	WP_071175012.1		
GH5 4 Bacillus agaradhaerens	Bacillus agaradhaerens	CAD61244.1		
BhCel5B_pdb-4V2X	Bacillus halodurans	BAB04322.1	4V2X	Venditto *et al* (2015) J Biol Chem 290: 10572-86.
BlCel5B_pdb-4YZP	Bacillus licheniformis DSM 13	AAU40777.1	4YZP	Liberato *et al* (2016) Sci Rep, 6: 23473.
Xylanase Prevotella ruminicola	Prevotella ruminicola	AAC36862.1		Whitehead (1993) Curr. Microbiol. 27: 27-33
PbGH5A_pdb-3VDH	Prevotella bryantii	AAC97596.1	3VDH	McGregor *et al* (2016) J. Biol. Chem. 291: 1175-1197
GH5 4 Fibrobacter succinogenes S85	Fibrobacter succinogenes	ACX74396.1		
GH5 4 Bacteroidetes AC2a	Bacteroidetes bacterium AC2a	AIJ19564.1		
GH5 4 Bacteroides ovatus	Bacteroides ovatus	ALJ47680.1		
GH5 4 Paenibacillus barcinonensis	Paenibacillus barcinonensis	CAA73113.1		
Sequence2 patent US 6630340	n.a.	AAR65335.1		
AcCel5A_pdb-1C0D_GH5_1	Acidothermus cellulolyticus	AAA75477.1	1C0D	Baker *et al* (2005) Appl.Biochem.Biotechnol. 121-124: 129-148

Table S26. Privateer results for the validation of carbohydrate structures in the *Zg*EngA$_{GH5_4_E323S}$ mutant structure.

Name	Chain	Q[1]	Phi	Theta	Anomer	D/L[2]	Conformation	RSCC[3]	<Bfactor>	Diagnostic
GLC	A	0.578	257.993	10.875	alpha	D	$^4C_1$	0.92	28.3567	Ok
BGC	A	0.553	97.5923	5.88636	beta	D	$^4C_1$	0.92	24.6409	Ok
BGC	A	0.550	252.561	5.42196	beta	D	$^4C_1$	0.90	28.8627	Ok
GLC	B	0.541	270.446	11.5406	alpha	D	$^4C_1$	0.91	30.2825	Ok
BGC	B	0.586	71.133	3.22921	beta	D	$^4C_1$	0.91	28.07	Ok
BGC	B	0.481	141.748	14.7687	beta	D	$^4C_1$	0.87	33.9218	Ok
GLC	C	0.564	235.187	12.9257	alpha	D	$^4C_1$	0.91	31.5717	Ok
BGC	C	0.559	327.311	3.92194	beta	D	$^4C_1$	0.90	30.8973	Ok
BGC	C	0.569	299.804	3.76075	beta	D	$^4C_1$	0.87	33.0627	Ok

[1] Q is the total puckering amplitude, measured in Angstroems.
[2] Whenever N is displayed in the D/L column, it means that Privateer has been unable to determine the handedness based solely on the structure.
[3] RSCC, short for Real Space Correlation Coefficient, measures the agreement between model and positive omit density. A RSCC below 0.8 is typically considered poor.