Reassembly and co-crystallization of a family 9 processive endoglucanase from its component parts: Structural and functional significance of the intermodular linker

Svetlana Petkun, Inna Rozman Grinberg, Raphael Lamed, Sadanari Jindou, Tal Burstein, Oren Yaniv, Yuval Shoham, Linda J.W. Shimon, Edward A Bayer, Felix Frolow

Non-cellulosomal processive endoglucanase 9I (Cel9I) from *Clostridium thermocellum* is a modular protein, consisting of a family-9 glycoside hydrolase (GH9) catalytic module and two family-3 carbohydrate-binding modules (CBM3c and CBM3b), separated by linker regions. GH9 does not show cellulase activity when expressed without CBM3c and CBM3b and the presence of the CBM3c was previously shown to be essential for endoglucanase activity. Physical reassociation of independently expressed GH9 and CBM3c modules (containing linker sequences) restored 60-70% of the intact Cel9I endocellulase activity. However, the mechanism responsible for recovery of activity remained unclear. In this work we independently expressed recombinant GH9 and CBM3c with and without their interconnecting linker in *Escherichia coli*. We crystallized and determined the molecular structure of the GH9/linker-CBM3c heterodimer at a resolution of 1.68 Å to understand the functional and structural importance of the mutual spatial orientation of the modules and the role of the interconnecting linker during their re-association. Enzyme activity assays and isothermal titration calorimetry were performed to study and compare the effect of the linker on the re-association. The results indicated that reassembly of the modules could also occur without the linker, albeit with only very low recovery of endoglucanase activity. We propose that the linker regions in the GH9/CBM3c endoglucanases are important for spatial organization and fixation of the modules into functional enzymes.

Reassembly and co-crystallization of a family 9 processive endoglucanase from its component parts: Structural and functional significance of the intermodular linker

Svetlana Petkun[1], Inna Rozman Grinberg[1], Raphael Lamed[1], Sadanari Jindou[2], Tal Burstein[1], Oren Yaniv[1], Yuval Shoham[3], Linda J.W. Shimon[4], Edward A. Bayer[5,*], and Felix Frolow[1]

[1]Department of Molecular Microbiology and Biotechnology, The Daniella Rich Institute for Structural Biology, Tel Aviv University, Ramat Aviv 69978 ISRAEL
[2]Department of Life Sciences, Meijo University, Nagoya 468-8502 JAPAN
[3]Department of Biotechnology and Food Engineering, Technion-Israel Institute of Technology, Haifa 32000 ISRAEL
[4]Department of Chemical Research Support and
[5]Department of Biological Chemistry, The Weizmann Institute of Science, Rehovot 76100 ISRAEL

Corresponding author:

Edward A. Bayer
Department of Biological Chemistry
The Weizmann Institute of Science
Rehovot 76100 ISRAEL
Tel: 972-8-934-2373
Fax: 972-8-946-8256
Email: ed.bayer@weizmann.ac.il

Running title: Crystal structure of reassembled Cel9I

Abbreviations used: CBM, carbohydrate-binding module; CBM3c*L,* family 3c CBM with linker; CBM3c*NL,* family 3c CBM without linker; CMC, carboxymethyl cellulose; GH9, family 9 glycoside hydrolase; ITC, isothermal titration calorimetry; PASC, phosphoric acid-swollen cellulose; SeMet, selenium-methionine labeled derivative.

Abstract.

Non-cellulosomal processive endoglucanase 9I (Cel9I) from *Clostridium thermocellum* is a modular protein, consisting of a family-9 glycoside hydrolase (GH9) catalytic module and two family-3 carbohydrate-binding modules (CBM3c and CBM3b), separated by linker regions. GH9 does not show cellulase activity when expressed without CBM3c and CBM3b and the presence of the CBM3c was previously shown to be essential for endoglucanase activity. Physical reassociation of independently expressed GH9 and CBM3c modules (containing linker sequences) restored 60-70% of the intact Cel9I endocellulase activity. However, the mechanism responsible for recovery of activity remained unclear. In this work we independently expressed recombinant GH9 and CBM3c with and without their interconnecting linker in *Escherichia coli*. We crystallized and determined the molecular structure of the GH9/linker-CBM3c heterodimer at a resolution of 1.68 Å to understand the functional and structural importance of the mutual spatial orientation of the modules and the role of the interconnecting linker during their re-association. Enzyme activity assays and isothermal titration calorimetry were performed to study and compare the effect of the linker on the re-association. The results indicated that reassembly of the modules could also occur without the linker, albeit with only very low recovery of endoglucanase activity. We propose that the linker regions in the GH9/CBM3c endoglucanases are important for spatial organization and fixation of the modules into functional enzymes.

Key words: *Clostridium thermocellum*; family-9 glycoside hydrolase; carbohydrate-binding module (CBM); X-ray structure; protein-protein interaction

Introduction

Cellulose is a major component of the plant cell wall, lending structural stability and resilience to an otherwise flaccid material. The propensity of cellulose to form ordered, tightly packed, para-crystalline fibrils hinders its enzymatic degradation. Indeed, the recalcitrant properties of cellulose are such that numerous enzymes are required to act synergistically in achieving its efficient degradation. Many types of bacteria and fungi are capable of degrading cellulose and other plant cell wall polysaccharides in an effective manner, producing a variety of various cellulases and related enzymes, either existing in the free state, or associated with a multi-enzyme complex known as the cellulosome (Bayer et al. 2004; Bayer et al. 2008; Demain et al. 2005; Doi & Kosugi 2004; Fontes & Gilbert 2010). *Clostridium thermocellum* is an anaerobic thermophilic bacterium, known for its efficient degradation of cellulose and other plant cell wall polysaccharides (Béguin et al. 1992; Freier et al. 1988; Garcia-Martinez et al. 1980; Ng et al. 1977; Wiegel et al. 1985). The cellulase system of this bacterium includes a remarkable variety of enzymes, some existing in the free state but most associated with a cellulosome (Béguin & Alzari 1998; Felix & Ljungdahl 1993; Schwarz 2001; Schwarz et al. 2004).

Cellulases are a class of modular enzymes with a catalytic glycoside hydrolase (GH) module that hydrolyzes the β-1,4-glucosidic bond of the cellulose chain (Cantarel et al. 2009; Davies & Henrissat 1995; Gilbert & Hazlewood 1993; Henrissat 1991; Henrissat & Davies 1997; Wilson & Irwin 1999). The catalytic module is usually associated with various numbers of accessory modules that serve to modulate the enzyme activity, and the enzymes have been categorized into families according to the amino-acid sequence of the GH domain (Cantarel et al. 2009; Gilkes et al. 1991; Henrissat & Davies 1997; Henrissat & Davies 2000; Henrissat & Romeu 1995). Cellulases have been broadly divided into two types: endoglucanases that can hydrolyze bonds internally in cellulose chain, and exoglucanases that act preferentially on chain ends, progressively cleaving off cellobiose as the main product. The distinction between endo- and exo-acting enzymes is also reflected by the architecture of the respective class of active site, whereby endoglucanases, for example, are commonly characterized by a groove or open binding cleft, into which any part of the linear cellulose chain can fit. On the other hand, the exoglucanases

bear tunnel-like active sites, which can only accept a substrate chain via its terminus (either the reducing or non-reducing end, depending on the enzyme), thereby cleaving cellulose in a sequential manner. The sequential hydrolysis of a cellulose chain has earned the term "processivity" (Beckham et al. 2014; Davies & Henrissat 1995; Wilson & Kostylev 2012), and processive enzymes are considered to be key components which contribute to the overall efficiency of a given cellulase system. Some endoglucanases, notably from GH family 9, have also been shown to sequentially hydrolyze cellulose chains and are thus referred to as processive endoglucanases (Gal et al. 1997; Gilad et al. 2003; Irwin et al. 1998; Jeon et al. 2012; Kuusk et al. 2015; Zverlov et al. 2003). Such enzymes appear to possess extended catalytic clefts and the observed processivity appears to require highly coordinated substrate-binding affinities from opposite sides of the cleavage site (Bu et al. 2012; Li et al. 2010; Payne et al. 2011).

Cellulase 9I (Cel9I), is a non-cellulosomal family 9 processive endoglucanase from *Clostridium thermocellum*, which degrades crystalline cellulose (Avicel and filter paper) as well as phosphoric acid-swollen cellulose (PASC) and carboxymethyl cellulose (CMC) (Gilad et al. 2003). This enzyme consists of a catalytic GH9 module at its N terminus, followed by two family 3 carbohydrate-binding modules (CBMs): CBM3c and CBM3b. The three modules are separated by distinctive linker sequences. Such intermodular linker segments were proposed to be important for the physical association of the modules in the space, and to promote intermodular and/or intersubunit protein–protein interactions (Bayer et al. 1998; Bayer et al. 2009; Noach et al. 2008).

The C-terminal CBM3b module, as a classic CBM3, is responsible for targeting the Cel9I enzyme to the planar surface of the crystalline cellulose substrate (Gilad et al. 2003; Su et al. 2012; Tormo et al. 1996). It has also been proposed to disrupt the crystalline regions of cellulose, rendering it more accessible to the GH9 catalytic module (Yi et al. 2013) and to contribute to enzyme processivity by preventing the desorption of the catalytic module from cellulose (Telke et al. 2012). The function of the CBM3c is less straightforward. Removal of CBM3c from *C. thermocellum* Cel9I and from *C. cellulolyticum* Cel9G *P. Barcinonensis* Cel9B significantly reduces the enzyme activity (Burstein et al. 2009; Chiriac et al. 2010; Gal et al. 1997). CBM3c modules have been shown to alter the normal function of the GH9 catalytic module of

Thermobifida fusca Cel9A from the standard endo-acting mode into a processive endoglucanase (Bayer et al. 1998; Irwin et al. 1998). Thus, Gilad *et* al.(Gilad et al. 2003) showed in 2003 that the endoglucanase activity of Cel9I is dependent upon the presence of the CBM3c module and suggested that the fused CBM3c serves an important accessory role for the catalytic domain by altering its character to facilitate processive cleavage of recalcitrant cellulose substrates.

In addition to the Cel9 CBM3c, several other examples of CBMs that are considered to modulate catalytic specificity and act cooperatively with the catalytic domain have recently been discovered. These include CBM66 that directs the cognate enzyme towards highly branched glucans rather than linear fructose polymers (Cuskin et al. 2012), CBM48 that contributes to substrate binding at the active site of a glucan phosphatase (Meekins et al. 2014), family-43 β-xylosidases where the GH43 is complemented by an additional module that confers hydrolytic activity to the mature enzyme (Moraïs et al. 2012), and CBM46, that constitutes part of the catalytic cleft required for the hydrolysis of β-1,3-1,4-glucans (Venditto et al. 2015). The carbohydrate-binding PA14 domain is also known to affect substrate binding of the catalytic domain by contributing to the formation of its active site (Gruninger et al. 2014; Zmudka et al. 2013).

We have previously shown that independently expressed GH9 and linker-containing CBM3c modules of Cel9I readily re-associate *in vitro* and that this physical reassociation recovers 60-70% of the intact Cel9I endoglucanase activity (Burstein et al. 2009).

We have examined in this work the interaction of the CBM3c with the catalytic module either with or without the intermodular linker in order to better understand the function of the CBM3c in the family-9 enzymes and the role of the linkers regions. The effect of the re-association of the CBM3c with linker (CBM3c*L*) and the CBM3c without linker (CBM3c*NL*) on the enzymatic activity of GH9 has been studied by the crystallization and structure determination of the reassembled GH9-CBM3c*L* complex at a resolution of 1.68 Å. The results of this study will help us to understand the contribution of ancillary modules in the action of multi-modular glycoside hydrolases.

Materials and methods

Cloning of the GH9, CBM3c*L* and CBM3c*NL* proteins

Cloning of the DNA fragments encoding the C-terminally His-tagged CBM3c with the linker and the untagged GH9 module from Cel9I of *C. thermocellum* (GenBank accession code L04735) was described earlier (Burstein et al. 2009; Gilad et al. 2003). C-terminally His-tagged CBM3c without the linker connecting it to the GH9 was amplified using the same procedure and the following primers: F′-5′CCATGGCGAAGTTCCGGAGGATGAAATA and R′-5′CTCGAGCGGTTCCCTTCCAAATACCAG. The PCR products were purified and cleaved with restriction enzymes *Nco*I and *Xho*I and inserted into the pET-28a(+) expression vector (Novagen, Madison, WI, USA).

Expression and purification of recombinant proteins

The GH9 and CBM3c modules both with (GH9L, CBM3c*L*) and without (GH9*NL* and CBM3c*NL*) the linker regions were expressed independently by the identical expression procedure. *Escherichia coli* strain BL21(DE3)RIL harboring the plasmids was aerated at 310 K in 3-liters Terrific Broth supplemented with 25 mg ml^{-1} kanamycin. After 3 h, the culture reached an A_{600} of 0.6; 0.1 mM isopropyl-β-D-1-thiogalactopyranoside was added to induce gene expression, and cultivation was continued at 310 K for an additional 12 h. Cells were harvested by centrifugation (5,000 × g for 15 min) at 277 K and were subsequently re-suspended in 50 mM NaH_2PO_4, pH 8.0, containing 300 mM NaCl at a ratio of 1 g wet pellet to 4 ml buffer solution. A few micrograms of DNase powder were added prior to the sonication procedure. The suspension was kept on ice during sonication, after which it was centrifuged (20,000 × g at 277 K for 20 min), and the supernatant was collected.

The soluble expressed His-tagged CBM3c modules with or without the linker, according to the type of the experiment, were applied batchwise to Ni-IDA resin during 1-h incubation with gentle stirring at 4 °C. Non-specifically bound proteins were washed with a buffer containing 50 mM NaH_2PO_4 pH 6, 300 mM NaCl, 10% glycerol and 10 mM imidazole. Crude extract supernatant fluids, containing the expressed GH9 module, were added to the CBM3c-bound Ni-IDA resin, and the mixture was incubated overnight with gentle stirring at 4 °C. The adsorbed protein

complexes were eluted with 300 mM imidazole and subjected to further purification by size-exclusion chromatography. Fast protein liquid chromatography (FPLC) was performed using a Superdex 75pg column and ÄKTA Prime system (GE Healthcare, Piscataway, NJ) to further purify the complex. One peak, corresponding approximately to 70 kDa, matching the predicted molecular weight of the GH9-CBM3c complex, was observed in the chromatogram. The 15 amino-acid linker sequence (about 1.5 kDa) did not significantly affect the elution volume, compared to that of the complex without the linker, presumably due to the limited resolution of the column. The relevant fractions (the purified complexed proteins) were analyzed by 15% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) with Coomassie brilliant blue staining. Two clear bands, of about 52 and 19.5 kDa were observed. The rearranged modules were concentrated to 6 mg ml^{-1} using Centriprep YM-3 centrifugal filter devices YM-3 (Amicon Bioseparation, Millipore Corporation, Bedford, USA). Protein concentration was determined by measuring UV absorbance at 280 nm.

The full-length Cel9I was purified by affinity chromatography on Avicel as reported earlier (Burstein et al. 2009; Gilad et al. 2003).

Microcalorimetric analysis

Isothermal titration calorimetry (ITC) experiments were carried out using a VP-ITC MicroCalorimeter (MicroCal, LLC, Northampton, MA) at 298 K. About 300 μM solution of CBM3c*NL* was injected into a 65 μM solution of GH9. The reaction was performed in a buffer containing 50 mM Tris–HCl, pH 7.5, 150 mM NaCl, 0.05% sodium azide. Heats of dilution of the titrants were subtracted from the titration data, and the corrected data were analyzed using the Origin ITC analysis software package supplied by MicroCal. Thermal titration data were fit to the one binding site model, and enthalpy (ΔH), entropy (ΔS), association constant (Ka) and stoichiometry of binding (N) were determined. In all cases, the calculated stoichiometry (N) was lower than one, most likely due to the fact that the CBM3 proteins lost their native functionality with time. For the analysis, the CBM3 protein concentrations were corrected as to provide a stoichiometry of one. Two titrations were performed to evaluate reproducibility.

Enzyme activity assay

Reactions were performed at 333 K, in 50 mM citrate buffer (pH 6.0). The soluble cellulolytic substrate was carboxymethyl cellulose (CMC, Sigma Chem. Co. St. Louis, MO). The amount of reducing sugars released from the substrate was determined with the 3,5-dinitrosalicylic acid (DNS) reagent as described by Miller et al (Miller 1959). Activity was defined as the amount (micromole) of reducing sugar released after 10 min of reaction.

Crystallization

Initially the protein samples containing 6 mg/ml protein solution in 1.2 mM Tris-HCl pH 7.5, 1.5 mM sodium chloride, 0.025% sodium azide, were screened, using the microbatch crystallization method under 1:1 mixture of silicon and paraffin oil (Chayen et al. 1990), using 288 conditions from the Hampton Research HT screens (SaltRx, Index HT, and Crystal Screen HT; Hampton Research, Aliso Viejo, CA) and 96 conditions of the Wizard Crystallization kit from Emerald BioSystems (Rigaku Reagents, Bainbridge Island, WA). The dyad of GH9 and CBM3c*NL* did not yield any crystals. Screening of the GH9-CBM3c*L* resulted in plate-like crystals that appeared after several days under several conditions, all of which contained PEG 3350 and 4000. The best crystals were obtained in 30 % PEG, 0.2 M magnesium chloride, and 0.1 M Hepes, pH 7.5. Attempts to optimize this condition using microbatch, hanging-drop, and sitting drop methods were unsuccessful, as the crystals remained very thin and fragile. The superfine Eyelash (Ted Pella, Inc, Redding, CA) was used to touch these crystals and consequently to streak the sitting drops, composed of 5 μl of the protein solution and 5 μl of the precipitating solution (24 % PEG 3350, 0.2 M magnesium chloride, 0.1 M Hepes, pH 8.0). After one day, crystals of different morphology, with maximum size of about 0.05 mm, appeared in the drop.

Data collection and crystallographic analysis

The crystals of the GH9-CBM3c*L* complex were harvested from the crystallization drop using a nylon cryo-loop (Hampton Research, Aliso Viejo, CA). For data collection, crystals were mounted on the MiTeGen stiff micro-mount (MiTeGen, Ithaca, NY) made of polyimide and flash-cooled in a nitrogen stream produced by Oxford Cryostream low temperature generator (Cosier & Glazer 1986) at a

temperature of 100 K. Mother-liquor of the crystals served for cryo-protection during the cooling in liquid nitrogen.

Diffraction data from the GH9-CBM3c*L* crystals were measured using the ID23-2 beam line at ESRF, Grenoble, France. A MAR CCD 225 area detector and X-ray radiation of 0.873 Å wavelength were used. Diffraction data of 480 images with 0.5° oscillation per image were collected. Data were processed with *DENZO* and scaled with *SCALEPACK* as implemented in *HKL2000* (Otwinowski & Minor 1997). The crystals diffracted to 1.68 Å resolution and belong to the orthorhombic space group $P2_12_12_1$, with unit cell parameters a=70.4, b=88.5, c=106.5 Å. There is one GH9-CBM3c*L* complex per asymmetric unit with a Matthews density V_M of 2.37 $Å^3$ Da^{-1}, corresponding to a solvent content of 48.15% (Matthews 1968). The X-ray data analysis statistics are presented in Table 1 (Stout & Jensen 1968).

Molecular replacement was carried out with *MOLREP* (Vagin & Teplyakov 1997), using the coordinates of the GH9 and CBM3c modules of endoglucanase 9G from *Clostridium cellulolyticum* (PDB code 1G87, 66 and 51% sequence identity, respectively), as a search model. The *MOLREP* calculations with the GH9 domain converged into a clear solution with 1 molecule in the asymmetric unit with an R-factor of 0.533 and correlation coefficient of 0.567. This solution was inserted into *MOLREP* calculations as a fixed molecule and the coordinates of CBM3c module were used for the search producing a solution with an R_{cryst} of 0.505, and correlation coefficient of 0.582. The resulting model with 5% of reflections forming test set (Brünger 1992) was subjected to 10 cycles of restrained refinement using anisotropic B-factors, yielding the R_{cryst} and R_{free} 0.329 and 0.359, respectively (*REFMAC5*) (Murshudov et al. 1997). Automated model building by *ARP/wARP* (Perrakis et al. 1999) produced a complete structure with R_{cryst} and R_{free} of 0.218 and 0.243 respectively. The model was manually corrected using *COOT* (Emsley & Cowtan 2004) and refined using *REFMAC5* (Murshudov et al. 1997). The R_{cryst} and R_{free} improved to 0.184 and 0.228, respectively. Solvent atoms were built using *ARP/warp* (Perrakis et al. 1999). Refinement of TLS (rigid body translation/libration/screw motions) parameters was performed (Winn et al. 2001; Winn et al. 2003). The model was subjected to several additional cycles of manual rebuilding and refinement. The model converged to final R_{cryst} and R_{free} factors of 0.144 and 0.176, respectively.

The refinement statistics of the structure are summarized in Table 2. The structure was validated using *MolProbity* (Davis et al. 2007).

Protein sequence analysis

Sequence alignments were performed using *CLUSTALW* (Larkin et al. 2007) and the coloring of residues (representing degree of conservation) using ProtSkin (Deprez et al. 2005). Sources of the sequences used in this work are as follows: *Clostridium thermocellum* Cel9I GH9 module, CBM3c and CBM3b (AAA20892.1); *Clostridium cellulolyticum* Cel9G GH9 module, CBM3c (AAA73868.1); *Thermobifida fusca* Cel9A GH9 module and CBM3c (AAB42155.1); *Cellulomonas fimi* Ce9A CBM3c (AAA23086.1); *Clostridium cellulovorans* EngH CBM3c (AAC38572.2) and CbpA CBM3a (AAA23218.1); *Clostridium stercorarium* CelZ CBM3c and CBM3b (CAA39010.1) and CelY CBM3b (CAA93280.1); *Clostridium thermocellum* CipA CBM3a (CAA48312.1), CelQ CBM3c (BAB33148.1), Cel9V CBM3c' and CBM3b' (CAK22315.1), Cel9U CBM3c' and CBM3b' (CAK22317.1) and Cbh9A CBM3b (CAA56918.1); *Clostridium cellulolyticum* CipC CBM3a (AAC28899.2) and CelJ CBM3c (AAG45158.1); *Acetivibrio cellulolyticus* Cel9B CBM3c' and CBM3b' (CAI94607.1) and CipV (ScaA) CBM3b (AAF06064.1); *Clostridium josui* CipA (CipJ) CBM3a (BAA32429.1); *Bacteroides cellulosolvens* ScaA CBM3b (AAG01230.2); *Bacillus subtilis* CelA CBM3b (AAA22307.1); *Pectobacterium atrosepticum* CelVI CBM3b (X79241.2); *Bacillus licheniformis* CelA CBM3b (CAJ70714.1).

Results

Cloning, expression and purification of Cel9I and its modular components

The full-length *C. thermocellum* Cel9I enzyme and its individual component parts were over-expressed in *Escherichia coli,* according to Burstein et al (2009), in order to investigate the contribution of the ancillary modules and their linkers to the catalytic activity of the enzyme. These include the isolated GH9 module with and without a His tag, the His-tagged CBM3c module together with its adjacent N-

terminal linker that connects it to the GH9 module (CBM3c*L*) and His-tagged CBM3c module without the N-terminal linker (CBM3c*NL*). For details, see Figure 1. Following purification procedures, all recombinant proteins showed a single band in SDS-PAGE of the anticipated molecular masses.

Recovery of endoglucanase activity upon association of CBM3c*NL* and GH9 compared to CBM3c*L* and GH9

Previous works (Burstein et al. 2009; Gilad et al. 2003) demonstrated that the Cel9I catalytic module alone has no detectable activity on CMC (carboxymethyl cellulose) and that adding the CBM3c*L* to form the Cel9I-CBM3c*L*-CBM3b triad serves to recover up to 70% of the lost activity. To further examine the importance of the linker connecting the GH9 and the CBM3c modules, we tested the ability of CBM3c*NL* to recover the CMCase activity of GH9. A fixed amount of the catalytic module (70 pmol in 400 μl) was mixed with increasing amounts of CBM3c*L* or CBM3c*NL*. The activity of the intact Cel9I enzyme was defined as 100%, and the activity of the reconstituted complexes was measured relative to that of Cel9I. The results indicated that GH9-CBM3c*NL* exhibit only about 10% of the intact Cel9I activity towards CMC, whereas the reassembled GH9-CBM3c*L* provided up to 50% of the activity (Figure 2). The fact that a higher than one molar ratio was required to obtain maximum activity can be explained by the fact that the CBM protein was only partly functional as was also observed in the ITC experiments described below. Overall the results suggest that the linker is required for better fitting of the reconstituted CBM3c which results in better recovered activity.

Overall structure of the reassembled GH9-CBM3c

The crystal structure of the reassembled *C. thermocellum* Cel9I GH9-CBM3c*L* dyad was determined by molecular replacement and the coordinates are deposited in Protein Data Bank with code 2XFG. Data collection and refinement statistics are given in Tables 1 and 2. The catalytic GH9 and the ancillary CBM3c modules reassembled in vitro to form a dyad (Figure 3a) similar in structure to the intact tandem GH9-CBM3c modules of the orthologous endoglucanases: Cel9G from *C. cellulolyticum* (1G87) and Cel9A (previously termed cellulase E4) from *Thermobifida fusca* (1TF4), with an RMS deviation of 0.783 Å over 468 Cα atoms with Cel9G and 0.757 Å with Cel9A (Figure 3b).

Structure of the GH9 module

The catalytic module of the Cel9I enzyme consists of residues 1-446, comprising 15 α-helices, whereby the twelve longest ones form the $(\alpha/\alpha)_6$-barrel (Figure 4A). The hydrophobic core of the GH9 module is formed by 118 hydrophobic and aromatic amino acids, the vast majority of which are also conserved in the GH9 modules from *C. cellulolyticum* Cel9G and *T. fusca* Cel9A. Hydrophobic and aromatic cores have been proposed to play an important role in the formation of $(\alpha/\alpha)_6$-barrels (Mandelman et al. 2003). The GH9 module of Cel9I thus shows high structural similarity with the two latter GH9 structures: *C. cellulolyticum* Cel9G (0.367 Å RMS deviation over 349 C-alpha atoms) and *T. fusca* Cel9A (0.532 Å RMS deviation over 359 C-alpha atoms).

The catalytic site of the GH9 module is located at the depression in the flat surface, formed by the loops connecting the N termini of the barrel helices (Figure 4B). The flat face is rich in charged and polar residues (Figure 4B), highly conserved also in Cel9G (1G87) and Cel9A (1TF4). The GH9 modules of these cellulases (Mandelman et al. 2003; Sakon et al. 1997; Zhou et al. 2004) exhibit similar flat faces and clefts, and these conserved residues (His 126, Trp 129, Phe 205, Tyr 206, Trp 209, Trp 256, Asp 261, Asp 262, Trp 314, Arg 318, His 376, Arg 378, and Tyr 419) have been shown to bind natural and synthetic oligosaccharides (Figure 4C). In the present structure, as in the other known GH9-CBM3c bimodular structures, one end of this cleft is blocked by a loop formed by residues 243-254 and the other end is fused with the flat surface of the CBM3c module (Figure 4B). Details of the catalytic cleft are presented in Figure 4C.

One calcium ion is found near the catalytic cleft of the GH9 module of Cel9I and is coordinated by a Ser 210 (OG) 2.6 Å, Gly 211 (O) 2.4 Å, Asp 261 (O) 2.4 Å, Asp 214 bifurcated (OD1, OD2) 2.5 Å, and Glu 215 bifurcated (OD1, OD2) 2.5 Å (Figure 4D). Despite some minor changes in the residues of coordination this ion seems to be structurally equivalent to those of *T. fusca* Cel9A (RMS deviation 0.160 Å over 5 Cα atoms of the coordinating residues), and *C. cellulolyticum* Cel9G (RMS deviation 0.503 Å over 4 Cα atoms). In all three cases the calcium ion draws together the N-terminal ends of α-helixes 8 and 10.

Structure of the CBM3c module

The CBM3c module consists of 150 amino acids arranged in an eight β-stranded sandwich motif homologous to other known family 3 CBM structures (Gilbert et al. 2013; Mandelman et al. 2003; Petkun et al. 2010b; Sakon et al. 1997; Shimon et al. 2000b; Tormo et al. 1996; Yaniv et al. 2014; Yaniv et al. 2012b; Yaniv et al. 2011). The "lower" face of the sandwich is formed by β-strands 1, 2, and 7; the "upper" face is formed by β-strands 3, 3', 6, 8, and 9 (Figure 5A). The structure of Cel9I CBM3c is particularly similar to the structures of the other two previously described CBM3c structures (RMS deviation 0.734 Å over 116 C-alpha atoms with CBM3c from *C. cellulolyticum* Cel9G; RMS deviation 0.829 Å over 113 atoms with CBM3c from *T. fusca* Cel9A). Only 31% of amino acids are located in β-strands of the CBM3c module from Cel9I; others are found in the loop regions.

One calcium ion was found in the upper β-sheet of the CBM3c molecule (Figure 5B) and is coordinated by a water molecule and five residues from the upper β-sheet: Asn 500 (O), Glu 503 bifurcated (OE1, OE2), Asn 573 (O), Asn 576 (OD1), Asp 577 (OD1). This calcium atom is in a similar location as in Cel9A and Cel9G, and probably plays a structural role for most CBM3 modules, as was suggested previously (Tormo et al. 1996).

The lower sheet forms a flat platform conserved between the CBM3c modules and the other two molecular structures. This flat surface is rich in charged and polar conserved surface residues: Asn 466, Glu 474, Lys 476, Ser 518, Tyr 520, Glu 559, Gln 561, and Arg 563 (Figure 5C). The planar region of the CBM3c modules in all three enzymes is particularly aligned in continuation of the catalytic cleft of the catalytic modules, and has been proposed to bind single chains of cellulose and guide them to the cleft (Mandelman et al. 2003; Sakon et al. 1997).

The CBM3c possesses a very interesting surface structure, formed by the β-strands on the opposite side of the flat surface, called the "shallow groove" (Shimon et al. 2000b; Tormo et al. 1996). The "shallow groove" is lined by four aromatic rings (Phe 498, Tyr 538, Tyr 578 and Tyr 597), two charged or polar residues (Arg 496, and Glu 540), Leu 602, Pro 595 and Pro 608. These residues are also conserved in other CBM3 modules regardless of their subgroup relation (a, b, or c), their cellulose-binding ability and their effect on the activity of the catalytic module. Figure 5D shows the shallow groove of the CBM3c module from the Cel9I enzyme colored

401 according to the extent of the conservation of the residues in other CBM3a, b and c
402 modules (darker blue represents more conservation). The alignment was performed
403 over 25 CBM3 sequences (11 CBM3c, 12 CBM3b and CBM3b', and 4 CBM3a).
404 Conservation of this surface structure, regardless of the particular known function of
405 the CBMs, implies that this site has some kind of "generic" function. This shallow
406 groove may serve to bind to single oligosaccharide chains or to peptide chains, such
407 as the intermodular linkers common to cellulases or cellulosomal scaffoldin subunits.
408 There is evidence that the shallow groove interacts with a linker region (Petkun et al.
409 2010a; Shimon et al. 2000a; Yaniv et al. 2012a).

410 **Contact residues between the GH9, linker and CBM3c**

411 The *in vitro* reassembled GH9-CBM3c*L* complex has a large intermodular
412 interface, the contact area of which is 1108.3 $Å^2$, corresponding to 12.3% of the total
413 surface-exposed area of the CBM3c module and 6.2% of the exposed GH9 module
414 (Krissinel & Henrick 2007). The GH9 and the CBM3c*L* modules of Cel9I are
415 assembled into the reconstituted GH9-CBM3c complex by 31 hydrogen bonds (4
416 main chain-main chain, 19 main chain-side chain, and 8 side chain-side chain), 14
417 hydrophobic, 3 aromatic interactions, and 3 ionic bonds
418 (http://pic.mbu.iisc.ernet.in/index.html) (Tina et al. 2007). Sixteen residues from the
419 GH9 module and seventeen residues from the CBM3c participate in these interactions
420 (contact residues are shown in Figure 6A). The vast majority of the contact residues
421 and contacts are similar to those of *C. cellulolyticum* Cel9G and of *T. fusca* Cel9A
422 (Figure 6B). Conserved residues of the linker make contacts with conserved residues
423 of the GH9 module, emphasizing the importance of the linker in this interaction.
424 As mentioned above, the mutual spatial orientation of the GH9 and CBM3c
425 modules is very similar to that in the native, intact bimodular pairs from Cel9G and
426 Cel9A leading to the overall similarity in structures. The remarkable conservation of
427 the overall architecture in the reassembled *in vitro* complex together with the striking
428 conservation of the contact residues implies its high functional importance. In all of
429 these structures (Cel9G, Cel9A, and the reassembled GH9-CBM3c*L* from Cel9I), the
430 flat surface of the CBM3c module is aligned in continuation with the catalytic cleft of
431 the GH9 module, making an extended platform (Figure 4B). This platform is rich in
432 charged and polar surface residues that are highly conserved throughout the family 3
433 CBMc's.

Microcalorimetric analysis of the GH9-CBM3c complex formation

The binding constants of GH9 and the CBM3c were obtained by performing isothermal titration calorimetry (ITC) experiments in which a solution of GH9 was titrated with a solution of CBM3c with or without the linker (Figure 7). Control experiments for each of the components alone were conducted and subtracted from the titration data. In both cases the titration curve could be fitted to a one-site binding model although the calculated stoichiometry was less than one. The low stoichiometry is probably a result of the fact that the soluble CBM module lost its functionality with time and its true active concentration was less than the measured protein concentration. To estimate the binding constants for the two CBM3c forms the CBM3c concentrations were corrected to provide a stoichiometry of one. In all cases the binding reactions were enthalpy driven with a negative entropy contribution. CBM3cL provided binding constants (K_d) between 1.3-2.0 x 10^{-6} M, whereas CBM3cNL exhibited stronger binding constants, K_d between 2.9-4.3 x 10^{-7} M. Thus, the linker may serve as a mitigating factor for the binding process, ensuring specific binding orientation. This is consistent with the structural data and the activity assays, which emphasizes the important role of the linker in enzyme functioning. In the case of CBM3c*NL*, the binding process may occur faster in the absence of linker, but may also lead to unspecific binding and aggregation of the modules.

Discussion

A striking feature of the family 9 glycoside hydrolases is their subdivision into architectural themes, which are defined by their conserved modular composition (Bayer et al. 2006). In this context, the Theme B1 endoglucanases contain a GH9 catalytic module followed by a purportedly fused family 3c CBM. Biochemical studies of some of the members of this group (Arai et al. 2001; Chiriac et al. 2010; Gal et al. 1997; Irwin et al. 1998; Li et al. 2007) have shown that the CBM3c acts as a modulator of the function of the catalytic module. However, the exact manner in which the CBM3c functions is still unclear. It has been shown (Gal et al. 1997; Gilad et al. 2003; Irwin et al. 1998) that family 3c CBMs (including the CBM3c from *C. thermocellum* Cel9I) fail to bind insoluble cellulosic substrates, implying that they do

not act as targeting agents for such substrates. The targeting of the enzyme to crystalline cellulose is achieved either through an additional CBM (Kostylev et al. 2012) or by attachment of the enzyme to a CBM-containing scaffoldin via a cohesin-dockerin interaction (Mingardon et al. 2011).

The CBM3c module of Cel9A from the *T. fusca* has been proposed to loosely anchor the enzyme to cellulose, to disrupt the hydrogen bonds in crystalline cellulose and to guide a single cellulose strand towards the active site of the GH9 catalytic module (Bayer et al. 2006; Li et al. 2007). This hypothesis has been supported by molecular docking and molecular dynamics simulation studies (Oliveira et al. 2009). Moreover, double point mutations indicated that high coordination between the substrate affinities of the catalytic module and CBM needs to be precisely controlled (Li et al. 2010). Enzyme thermostability was reported to be affected by the presence of the CBM3c probably due to the formation of a compact structure (Chiriac et al. 2010; Su et al. 2012; Yi et al. 2013).

The previously reported structures of Cel9A from *T. fusca* (Sakon et al. 1997) and Cel9G from *C. cellulolyticum* (Mandelman et al. 2003) revealed that the catalytic module and the CBM3c are separated by a ~20-residue linker that forms multiple polar and hydrophobic interactions mainly with the GH9 module. In an earlier report, we demonstrated that separately expressed GH9 and CBM3c*L* from Cel9I of *C. thermocellum* interact with one another to form an enzymatically active complex (Burstein et al. 2009). In the current article, we showed further that the GH9 and CBM3c can also be reassembled without the linker, albeit at the expense of catalytic activity, thus emphasizing the importance of the linker in positioning correctly the CBM relative to the GH9 catalytic module.

There is evidence that linkers in multi-modular proteins may serve communication roles between the modules via allosteric mechanisms and variation in their sequences affect enzyme activity (Ma et al. 2011). Linker length and rigidity was shown to play a critical role in the cooperative action of the catalytic module of a cellulase and a CBM (Ting et al. 2009). Computational studies of *T. fusca* Cel9A suggested that thermal contributions to enzyme plasticity and molecular motion at

499 high temperatures may play a role in enhancing CBM and catalytic domain synergy,
500 and the linker may have an important role in this process (Batista et al. 2011). The
501 length of the linkers may also play an important role in protein function and
502 adaptation to the environment (Sonan et al. 2007). Studies in cellulolytic fungi
503 revealed that linkers undergo modifications such as glycoslation and have also been
504 shown to directly bind to the cellulose substrate (Beckham et al. 2012; Payne et al.
505 2013; Sammond et al. 2012; Srisodsuk et al. 1993). Point mutations in different
506 fungal GH-CBM linkers have also been shown to significantly affect the activity of
507 the enzymes and their stability (Couturier et al. 2013; Lu et al. 2014).
508
509 The characteristics of the reassembled linker-containing complex are
510 corroborated by the X-ray crystallographic data. Indeed, it is quite surprising that the
511 two separately expressed entities recombined in such a way that the complex could in
512 fact be crystallized. Moreover, the resultant structure was remarkably similar to the
513 known structures of the intact bimodular GH9-CBM3c pairs from *C. cellulolyticum*
514 Cel9G and *T. fusca* Cel9A. Accordingly, the vast majority of the contact residues are
515 similar among the three structures. Conserved residues of the linker make contacts
516 with conserved residues of the GH9 module, highlighting the importance of the linker
517 in this interaction. The similarity of the reassembled and native intact structures is
518 particularly intriguing, as it suggests that folding of the modular structures and
519 emplacement of the linker during biosynthesis and intermodular recognition during
520 complex formation are governed by the same interactions, which may have distinct
521 functional consequences. In contrast to the GH9-CBM3c*L,* the re-associated GH9-
522 CBM3c*NL* complex never crystallized, suggesting that the reassembly of the two
523 modules in the absence of linker was somewhat heterogeneous in character.
524
525 Single proteins commonly fold into defined structures, wherein their N- and C-
526 terminal ends are in relatively close proximity to one another. If we view the
527 structures of the Theme B1 enzymes, it is evident that their individual modules, the
528 GH9 catalytic module and the CBM3c, are consistent with this rule. The positions of
529 the N- and C-termini of the Theme B1 catalytic module are similar to those of the
530 other GH9 thematic members, including those of Theme A, which lack additional
531 modules. Likewise, the N- and C-termini of CBM3c are essentially the same as all

other members of the family 3 CBMs, regardless of their source (i.e., parent cellulase, scaffoldin, etc). The evolutionary significance of this observation is that, originally, the functional relationship between the two modules was likely a more conventional one, whereby an ancestral CBM3 played a standard targeting role to deliver the GH9 catalytic module to its substrate. During the course of evolution, this relationship changed, and the precise positioning and fusion of a mutated CBM3 with a GH9 catalytic module served to modulate the activity characteristics of the latter. For this purpose, the flat surface of the CBM3c is aligned with the flat surface of the catalytic module, and the appropriate residues that interact with the single cellulose chain are thus aligned with the active site of the GH9 module. As a consequence, the two closely juxtaposed modules can be considered as a single functional entity. The functional positioning and fusion of the two modules, however, are at odds with the inherent locations of the termini of the two modules, such that the C-terminus of the catalytic module is very distant from the N-terminus of the CBM3c. Consequently, nature has provided a very distinctive type of conserved linker, which both connects the two modules and helps secure their required orientation for processive endoglucanase activity.

Conclusions

Cellulase 9I (Cel9I), a non-cellulosomal family 9 processive endoglucanase from *Clostridium thermocellum*, which degrades crystalline cellulose phosphoric acid-swollen cellulose (PASC) and carboxymethyl cellulose (CMC), consists of a catalytic GH9 module followed by two family 3 carbohydrate-binding modules (CBMs): CBM3c and CBM3b, separated by linker regions. C-terminal CBM3b module, as a classic CBM3, is responsible for targeting the Cel9I enzyme to the planar surface of the crystalline cellulose. The CBM3c is crucial for the GH9 enzymatic activity. In this work we investigated the interaction of separately expressed catalytic module and CBM3c either with or without the intermodular linker in order to better understand the function of the CBM3c in the family-9 enzymes and the role of the linkers regions.

GH9 catalytic module and CBM3c were able to interact and reassemble both with and without the linker; however the linker was essential for the endoglucanase catalytic activity. Surprisingly, we were able to crystallize these two separately expressed

565 entities, meaning that their reassembly was very ordered and structurally
566 homogeneous. The molecular structure of the GH9 and CBM3c with the linker region
567 showed that they form a complex similar in structure to the intact tandem GH9-
568 CBM3c modules of the orthologous endoglucanases Cel9G from *C. cellulolyticum*
569 and Cel9A from *Thermobifida fusca*. The flat, conserved surface of the CBM3c
570 module is aligned in continuation with the catalytic cleft of the GH9 module,
571 presumably forming one functional entity, which binds to the planar surface of the
572 cellulose. Conserved residues of the linker make contacts with conserved residues of
573 the GH9 module, highlighting the importance of the linker in this interaction. Overall
574 our results demonstrate that the linker regions in the GH9/CBM3c endoglucanases are
575 necessary to achieve the right spatial organization of the modules and for the fixation
576 of the modules into functional enzymes.
577

578 ## Acknowledgements

579 This article is dedicated to the memory of Professor Felix Frolow, who passed away
580 on 29 August 2014. We thankfully acknowledge the ESRF for synchrotron beam time
581 and staff scientists of the ID-29 beam line for their assistance.
582

References

Arai T, Ohara H, Karita S, Kimura T, Sakka K, and Ohmiya K. 2001. Sequence of celQ and properties of celQ, a component of the Clostridium thermocellum cellulosome. *Appl Microbiol Biotechnol* 57:660-666.

Batista PR, de Souza Costa MG, Pascutti PG, Bisch PM, and de Souza W. 2011. High temperatures enhance cooperative motions between CBM and catalytic domains of a thermostable cellulase: mechanism insights from essential dynamics. *Physical Chemistry Chemical Physics* 13:13709-13720.

Bayer EA, Belaich J-P, Shoham Y, and Lamed R. 2004. The cellulosomes: Multi-enzyme machines for degradation of plant cell wall polysaccharides. *Annu Rev Microbiol* 58:521-554.

Bayer EA, Lamed R, White BA, and Flint HJ. 2008. From cellulosomes to cellulosomics. *Chem Rec* 8:364-377.

Bayer EA, Morag E, Lamed R, Yaron S, and Shoham Y. 1998. Cellulosome structure: four-pronged attack using biochemistry, molecular biology, crystallography and bioinformatics. In: Claeyssens M, Nerinckx W, and Piens K, eds. *Carbohydrases from Trichoderma reesei and other microorganisms*. London: The Royal Society of Chemistry, 39-65.

Bayer EA, Shoham Y, and Lamed R. 2006. Cellulose-decomposing prokaryotes and their enzyme systems. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, and Stackebrandt E, eds. *The Prokaryotes, Third Edition*. New York: Springer-Verlag, 578-617.

Bayer EA, Smith SP, Noach I, Alber O, Adams JJ, Lamed R, Shimon LJW, and Frolow F. 2009. Can we crystallize a cellulosome? In: Sakka K, Karita S, Kimura T, Sakka M, Matsui H, Miyake H, and Tanaka A, eds. *Biotechnology of lignocellulose degradation and biomass utilization*: Ito Print Publishing Division, 183-205.

Beckham GT, Dai Z, Matthews JF, Momany M, Payne CM, Adney WS, Baker SE, and Himmel ME. 2012. Harnessing glycosylation to improve cellulase activity. *Curr Opinion Biotechnol* 23:338-345.

Beckham GT, Stahlberg J, Knott BC, Himmel ME, Crowley MF, Sandgren M, Sorlie M, and Payne CM. 2014. Towards a molecular-level theory of carbohydrate processivity in glycoside hydrolases. *Curr Opinion Biotechnol* 27:96-106.

Béguin P, and Alzari PM. 1998. The cellulosome of *Clostridium thermocellum*. *Biochem Soc Trans* 26:178-185.

Béguin P, Millet J, and Aubert J-P. 1992. Cellulose degradation by *Clostridium thermocellum:* From manure to molecular biology. *FEMS Microbiol Lett* 100:523-528.

Brünger TA. 1992. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355:472-475.

Bu L, Nimlos MR, Shirts MR, Stahlberg J, Himmel ME, Crowley MF, and Beckham GT. 2012. Product binding varies dramatically between processive and nonprocessive cellulase enzymes. *J Biol Chem* 287:24807-24813.

Burstein T, Shulman M, Jindou S, Petkun S, Frolow F, Shoham Y, Bayer EA, and Lamed R. 2009. Physical association of the catalytic and helper modules of a processive family-9 glycoside hydrolase is essential for activity. *FEBS Lett* 583:879-884.

630 Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, and Henrissat B.
631 2009. The Carbohydrate-Active Enzymes database (CAZy): an expert
632 resource for glycogenomics. *Nucl Acids Res* 37:D233-238.
633 Chayen NE, Shaw Stewart PD, Maeder DL, and Blow DM. 1990. An automated
634 system for micro-batch protein crystallization and screening. *J Appl*
635 *Crystallogr* 23:297-302.
636 Chiriac AI, Cadena EM, Vidal T, Torres AL, Diaz P, and Pastor FI. 2010.
637 Engineering a family 9 processive endoglucanase from Paenibacillus
638 barcinonensis displaying a novel architecture. *Applied microbiology and*
639 *biotechnology* 86:1125-1134.
640 Cosier J, and Glazer AM. 1986. A nitrogen-gas-stream cryostat for general X-ray-
641 diffraction studies. *J Appl Crystalogr* 19:105-107.
642 Couturier M, Feliu J, Bozonnet S, Roussel A, and Berrin JG. 2013. Molecular
643 engineering of fungal GH5 and GH26 beta-(1,4)-mannanases toward
644 improvement of enzyme activity. *PloS one* 8:e79800.
645 Cuskin F, Flint JE, Gloster TM, Morland C, Basle A, Henrissat B, Coutinho PM,
646 Strazzulli A, Solovyova AS, Davies GJ, and Gilbert HJ. 2012. How nature can
647 exploit nonspecific catalytic and carbohydrate binding modules to create
648 enzymatic specificity. *Proceedings of the National Academy of Sciences of the*
649 *United States of America* 109:20889-20894.
650 Davies G, and Henrissat B. 1995. Structures and mechanisms of glycosyl hydrolases.
651 *Structure* 3:853-859.
652 Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW,
653 Arendall WB, 3rd, Snoeyink J, Richardson JS, and Richardson DC. 2007.
654 MolProbity: all-atom contacts and structure validation for proteins and nucleic
655 acids *Nucl Acids Res* 35(Web Server issue):W375-383.
656 Demain AL, Newcomb M, and Wu JH. 2005. Cellulase, clostridia, and ethanol.
657 *Microbiol Mol Biol Rev* 69:124-154.
658 Deprez C, Lloubes R, Gavioli M, Marion D, Guerlesquin F, and Blanchard L. 2005.
659 Solution structure of the *E. coli* TolA C-terminal domain reveals
660 conformational changes upon binding to the phage g3p N-terminal domain *J*
661 *Mol Biol* 346:1047-1057.
662 Doi RH, and Kosugi A. 2004. Cellulosomes: plant-cell-wall-degrading enzyme
663 complexes. *Nat Rev Microbiol* 2:541-551.
664 Emsley P, and Cowtan K. 2004. Coot: model-building tools for molecular graphics.
665 *Acta Crystallographica Section D-Biological Crystallography* 60:2126-2132.
666 Felix CR, and Ljungdahl LG. 1993. The cellulosome - the exocellular organelle of
667 Clostridium. *Annu Rev Microbiol* 47:791-819.
668 Fontes CM, and Gilbert HJ. 2010. Cellulosomes: Highly efficient nanomachines
669 designed to deconstruct plant cell wall complex carbohydrates. *Annu Rev*
670 *Biochem* 79:655-681.
671 Freier D, Mothershed CP, and Wiegel J. 1988. Characterization of *Clostridium*
672 *thermocellum* JW20. *Appl Environ Microbiol* 54:204-211.
673 Gal L, Gaudin C, Belaich A, Pagès S, Tardif C, and Belaich J-P. 1997. CelG from
674 *Clostridium cellulolyticum*: a multidomain endoglucanase acting efficiently
675 on crystalline cellulose. *J Bacteriol* 179:6595-6601.
676 Garcia-Martinez DV, Shinmyo A, Madia A, and Demain AL. 1980. Studies on
677 cellulase production by *Clostridium thermocellum*. *Eur J Appl Microbiol*
678 *Biotechnol* 9:189-197.

679 Gilad R, Rabinovich L, Yaron S, Bayer EA, Lamed R, Gilbert HJ, and Shoham Y.
680 2003. CelI, a non-cellulosomal family-9 enzyme from *Clostridium*
681 *thermocellum*, is a processive endoglucanase that degrades crystalline
682 cellulose. *J Bacteriol* 185:391-398.
683 Gilbert HJ, and Hazlewood GP. 1993. Bacterial cellulases and xylanases. *J Gen*
684 *Microbiol* 139:187-194.
685 Gilbert HJ, Knox JP, and Boraston AB. 2013. Advances in understanding the
686 molecular basis of plant cell wall polysaccharide recognition by carbohydrate-
687 binding modules. *Current opinion in structural biology* 23:669-677.
688 Gilkes NR, Henrissat B, Kilburn DG, Miller RCJ, and Warren RAJ. 1991. Domains
689 in microbial b-1,4-glycanases: sequence conservation, function, and enzyme
690 families. *Microbiol Rev* 55:303-315.
691 Gruninger RJ, Gong X, Forster RJ, and McAllister TA. 2014. Biochemical and kinetic
692 characterization of the multifunctional beta-glucosidase/beta-xylosidase/alpha-
693 arabinosidase, Bgxa1. *Applied microbiology and biotechnology* 98:3003-3012.
694 Henrissat B. 1991. A classification of glycosyl hydrolases based on amino acid
695 sequence similarities. *Biochem J* 280:309-316.
696 Henrissat B, and Davies G. 1997. Structural and sequence-based classification of
697 glycoside hydrolases. *Curr Opin Struct Biol* 7:637-644.
698 Henrissat B, and Davies GJ. 2000. Glycoside hydrolases and glycosyltransferases.
699 Families, modules, and implications for genomics. *Plant Physiol* 124:1515-
700 1519.
701 Henrissat B, and Romeu A. 1995. Families, superfamilies and subfamilies of glycosyl
702 hydrolases. *Biochem J* 311:350-351.
703 Irwin D, Shin D-H, Zhang S, Barr BK, Sakon J, Karplus PA, and Wilson DB. 1998.
704 Roles of the catalytic domain and two cellulose binding domains of
705 *Thermomonospora fusca* E4 in cellulose hydrolysis. *J Bacteriol* 180:1709-
706 1714.
707 Jeon SD, Yu KO, Kim SW, and Han SO. 2012. The processive endoglucanase EngZ
708 is active in crystalline cellulose degradation as a cellulosomal subunit of
709 Clostridium cellulovorans. *New biotechnology* 29:365-371.
710 Kostylev M, Moran-Mirabal JM, Walker LP, and Wilson DB. 2012. Determination of
711 the molecular states of the processive endocellulase Thermobifida fusca
712 Cel9A during crystalline cellulose depolymerization. *Biotechnology and*
713 *bioengineering* 109:295-299.
714 Krissinel E, and Henrick K. 2007. Inference of macromolecular assemblies from
715 crystalline state. *J Mol Biol* 372:774--797.
716 Kuusk S, Sorlie M, and Valjamae P. 2015. The predominant molecular state of bound
717 enzyme determines the strength and type of product inhibition in the
718 hydrolysis of recalcitrant polysaccharides by processive enzymes. *The Journal*
719 *of biological chemistry*.
720 Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H,
721 Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, and
722 Higgins DG. 2007. ClustalW and ClustalX version 2. *Bioinformatics* 23:2947-
723 2948.
724 Li Y, Irwin DC, and Wilson DB. 2007. Processivity, substrate binding, and
725 mechanism of cellulose hydrolysis by *Thermobifida fusca* Cel9A. *Appl*
726 *Environ Microbiol* 73:3165-3172.
727 Li Y, Irwin DC, and Wilson DB. 2010. Increased crystalline cellulose activity via
728 combinations of amino acid changes in the family 9 catalytic domain and

family 3c cellulose binding module of *Thermobifida fusca* Cel9A. *Appl Environ Microbiol* 76:2582-2588.

Lu H, Luo H, Shi P, Huang H, Meng K, Yang P, and Yao B. 2014. A novel thermophilic endo-beta-1,4-mannanase from Aspergillus nidulans XZ3: functional roles of carbohydrate-binding module and Thr/Ser-rich linker region. *Applied microbiology and biotechnology* 98:2155-2163.

Ma B, Tsai CJ, Haliloglu T, and Nussinov R. 2011. Dynamic allostery: linkers are not merely flexible. *Structure (London, England : 1993)* 19:907-917.

Mandelman D, Belaich A, Belaich JP, Aghajari N, Driguez H, and Haser R. 2003. X-Ray crystal structure of the multidomain endoglucanase Cel9G from *Clostridium cellulolyticum* complexed with natural and synthetic cello-oligosaccharides. *J Bacteriol* 185:4127-4135.

Matthews BW. 1968. Solvent content of protein crystals. *J Mol Biol* 33:491-497.

Meekins DA, Raththagala M, Husodo S, White CJ, Guo HF, Kotting O, Vander Kooi CW, and Gentry MS. 2014. Phosphoglucan-bound structure of starch phosphatase Starch Excess4 reveals the mechanism for C6 specificity. *Proceedings of the National Academy of Sciences of the United States of America* 111:7272-7277.

Miller GL. 1959. Use of dinitrosalicylic acid reagent for determination of reducing sugar. *Anal Biochem* 31:426-428.

Mingardon F, Bagert JD, Maisonnier C, Trudeau DL, and Arnold FH. 2011. Comparison of family 9 cellulases from mesophilic and thermophilic bacteria. *Applied and environmental microbiology* 77:1436-1442.

Moraïs S, Alber O, Barak Y, Hadar Y, Wilson DB, Lamed R, Shoham Y, and Bayer EA. 2012. Functional association of the catalytic and ancillary modules dictates enzymatic activity in glycoside hydrolase family 43 β-xylosidase. *J Biol Chem* 287:9213-9221.

Murshudov GN, Vagin AA, and Dodson EJ. 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D* 53:240-255.

Ng TK, Weimer TK, and Zeikus JG. 1977. Cellulolytic and physiological properties of *Clostridium thermocellum*. *Arch Microbiol* 114:1-7.

Noach I, Alber O, Bayer EA, Lamed R, Levy-Assaraf M, Shimon LJW, and Frolow F. 2008. Crystallization and preliminary X-ray analysis of *Acetivibrio cellulolyticus* cellulosomal type II cohesin module: Two versions having different linker lengths. *Acta Cryst* F64:58-61.

Oliveira OV, Freitas LCG, Straatsma TP, and Lins RD. 2009. Interaction between the CBM of Cel9A from *Thermobifida fusca* and cellulose fibers. *Journal of Molecular Recognition* 22:38-45.

Otwinowski Z, and Minor W. 1997. Processing of X-ray diffraction data collected in oscillation mode. *Meth Enzymol*, 307-326.

Payne CM, Bomble YJ, Taylor CB, McCabe C, Himmel ME, Crowley MF, and Beckham GT. 2011. Multiple functions of aromatic-carbohydrate interactions in a processive cellulase examined with molecular simulation. *The Journal of biological chemistry* 286:41028-41035.

Payne CM, Resch MG, Chen L, Crowley MF, Himmel ME, Taylor LE, 2nd, Sandgren M, Stahlberg J, Stals I, Tan Z, and Beckham GT. 2013. Glycosylated linkers in multimodular lignocellulose-degrading enzymes dynamically bind to cellulose. *Proceedings of the National Academy of Sciences of the United States of America* 110:14646-14651.

779 Perrakis A, Morris R, and Lamzin VS. 1999. Automated protein model building
780 combined with iterative structure refinement. *Nature Struct Biol* 6:458-463.
781 Petkun S, Jindou S, Shimon LJW, Rosenheck S, Bayer EA, Lamed R, and Frolow F.
782 2010a. Structure of a family 3b ' carbohydrate-binding module from the Cel9V
783 glycoside hydrolase from Clostridium thermocellum: structural diversity and
784 implications for carbohydrate binding. *Acta Crystallographica Section D-*
785 *Biological Crystallography* 66:33-43.
786 Petkun S, Jindou S, Shimon LJW, Rosenheck S, Bayer EA, Lamed R, and Frolow F.
787 2010b. Structure of a family 3b' carbohydrate-binding module from the Cel9V
788 glycoside hydrolase from *Clostridium thermocellum*. Structural diversity and
789 implications for carbohydrate binding. *Acta Cryst* D66:33-43.
790 Sakon J, Irwin D, Wilson DB, and Karplus PA. 1997. Structure and mechanism of
791 endo/exocellulase E4 from *Thermomonospora fusca*. *Nature Struct Biol*
792 4:810-818.
793 Sammond DW, Payne CM, Brunecky R, Himmel ME, Crowley MF, and Beckham
794 GT. 2012. Cellulase linkers are optimized based on domain type and function:
795 insights from sequence analysis, biophysical measurements, and molecular
796 simulation. *PloS one* 7:e48615.
797 Schwarz WH. 2001. The cellulosome and cellulose degradation by anaerobic bacteria.
798 *Appl Microbiol Biotechnol* 56:634-649.
799 Schwarz WH, Zverlov VV, and Bahl H. 2004. Extracellular glycosyl hydrolases from
800 Clostridia. *Advan Appl Microbiol* 56:215-261.
801 Shimon LJ, Pages S, Belaich A, Belaich JP, Bayer EA, Lamed R, Shoham Y, and
802 Frolow F. 2000a. Structure of a family IIIa scaffoldin CBD from the
803 cellulosome of Clostridium cellulolyticum at 2.2 A resolution. *Acta*
804 *Crystallogr D Biol Crystallogr* 56:1560-1568.
805 Shimon LJW, Pagès S, Belaich A, Belaich JP, Bayer EA, Lamed R, Shoham Y, and
806 Frolow F. 2000b. Structure of a family IIIa scaffoldin CBD from the
807 cellulosome of *Clostridium cellulolyticum* at 2.2 Å resolution. *Acta*
808 *Crystallogr D Biol Crystallogr* 56:1560-1568.
809 Sonan GK, Receveur-Brechot V, Duez C, Aghajari N, Czjzek M, Haser R, and
810 Gerday C. 2007. The linker region plays a key role in the adaptation to cold of
811 the cellulase from an Antarctic bacterium. *The Biochemical journal* 407:293-
812 302.
813 Srisodsuk M, Reinikainen T, Penttila M, and Teeri TT. 1993. Role of the interdomain
814 linker peptide of Trichoderma reesei cellobiohydrolase I in its interaction with
815 crystalline cellulose. *The Journal of biological chemistry* 268:20756-20761.
816 Stout GH, and Jensen LH. 1968. *X-ray structure determination. A practical guide*.
817 London: MacMillan.
818 Su X, Mackie RI, and Cann IK. 2012. Biochemical and mutational analyses of a
819 multidomain cellulase/mannanase from Caldicellulosiruptor bescii. *Applied*
820 *and environmental microbiology* 78:2230-2240.
821 Telke AA, Ghatge SS, Kang SH, Thangapandian S, Lee KW, Shin HD, Um Y, and
822 Kim SW. 2012. Construction and characterization of chimeric cellulases with
823 enhanced catalytic activity towards insoluble cellulosic substrates.
824 *Bioresource technology* 112:10-17.
825 Tina KG, Bhadra R, and Srinivasan N. 2007. PIC: Protein interactions calculator.
826 *Nucl Acids Res* 35:W473–W476.
827 Ting CL, Makarov DE, and Wang ZG. 2009. A kinetic model for the enzymatic
828 action of cellulase. *The journal of physical chemistry B* 113:4970-4977.

829 Tormo J, Lamed R, Chirino AJ, Morag E, Bayer EA, Shoham Y, and Steitz TA. 1996.
830 Crystal structure of a bacterial family-III cellulose-binding domain: A general
831 mechanism for attachment to cellulose. *EMBO J* 15:5739-5751.
832 Vagin A, and Teplyakov A. 1997. MOLREP: an automated program for molecular
833 replacement. *J Appl Crystalogr* 30:1022-1025.
834 Venditto I, Najmudin S, Luis AS, Ferreira LM, Sakka K, Knox JP, Gilbert HJ, and
835 Fontes CM. 2015. Family 46 Carbohydrate-Binding Modules contribute to the
836 enzymatic hydrolysis of xyloglucan and beta-1,3-1,4-glucans through distinct
837 mechanisms. *The Journal of biological chemistry*.
838 Wiegel J, Mothershed CP, and Puls J. 1985. Differences in xylan degradation by
839 various noncellulolytic thermophilic anaerobes and *Clostridium thermocellum*.
840 *Appl Environ Microbiol* 49:656-659.
841 Wilson DB, and Irwin DC. 1999. Genetics and properties of cellulases. *Adv Biochem*
842 *Eng* 65:1-21.
843 Wilson DB, and Kostylev M. 2012. Cellulase processivity. *Methods in molecular*
844 *biology (Clifton, NJ)* 908:93-99.
845 Winn MD, Isupov MN, and Murshudov GN. 2001. Use of TLS parameters to model
846 anisotropic displacements in macromolecular refinement. *Acta Crystallogr D*
847 57:122-133.
848 Winn MD, Murshudov GN, and Papiz MZ. 2003. Macromolecular TLS refinement in
849 REFMAC at moderate resolution. *Methods Enzymol* 374:300-321.
850 Yaniv O, Fichman G, Borovok I, Shoham Y, Bayer EA, Lamed R, Shimon LJ, and
851 Frolow F. 2014. Fine-structural variance of family 3 carbohydrate-binding
852 modules as extracellular biomass-sensing components of Clostridium
853 thermocellum anti-sigmaI factors. *Acta crystallographica Section D,*
854 *Biological crystallography* 70:522-534.
855 Yaniv O, Frolow F, Levy-Assraf M, Lamed R, and Bayer EA. 2012a. Interactions
856 between family 3 carbohydrate binding modules (CBMs) and cellulosomal
857 linker peptides. *Methods in enzymology* 510:247-259.
858 Yaniv O, Petkun S, Shimon LJ, Bayer EA, Lamed R, and Frolow F. 2012b. A single
859 mutation reforms the binding activity of an adhesion-deficient family 3
860 carbohydrate-binding module. *Acta crystallographica Section D, Biological*
861 *crystallography* 68:819-828.
862 Yaniv O, Shimon LJ, Bayer EA, Lamed R, and Frolow F. 2011. Scaffoldin-borne
863 family 3b carbohydrate-binding module from the cellulosome of Bacteroides
864 cellulosolvens: structural diversity and significance of calcium for
865 carbohydrate binding. *Acta crystallographica Section D, Biological*
866 *crystallography* 67:506-515.
867 Yi Z, Su X, Revindran V, Mackie RI, and Cann I. 2013. Molecular and biochemical
868 analyses of CbCel9A/Cel48A, a highly secreted multi-modular cellulase by
869 Caldicellulosiruptor bescii during growth on crystalline cellulose. *PloS one*
870 8:e84172.
871 Zhou W, Irwin DC, Escovar-Kousen J, and Wilson DB. 2004. Kinetic studies of
872 Thermobifida fusca Cel9A active site mutant enzymes. *Biochemistry* 43:9655-
873 9663.
874 Zmudka MW, Thoden JB, and Holden HM. 2013. The structure of DesR from
875 Streptomyces venezuelae, a beta-glucosidase involved in macrolide activation.
876 *Protein science : a publication of the Protein Society* 22:883-892.
877 Zverlov VV, Velikodvorskaya GA, and Schwarz WH. 2003. Two new cellulosome
878 components encoded downstream of *cel*I in the genome of *Clostridium*

879 *thermocellum*: the non-processive endoglucanase CelN and the possibly
880 structural protein CseP. *Microbiology* 149:515-524.
881
882
883

884

Table 1 Diffraction data of the GH9-CBM3c *in vitro* reassembled complex from Cel9I from *C. thermocellum*. Values shown in parentheses are for the highest resolution cell.

GH9-CBM3c	ESRF
Space group	$P2_12_12_1$
Number of crystals	1
Total rotation range (°)	240
a (Å)	70.39
b (Å)	88.54
c (Å)	106.49
V (Å^3)	663743.40
Resolution range (Å)	30-1.68 (1.71-1.68)
Total number of reflections	676571
Unique reflections	76727
Mosaicity range (°)	0.18-0.46
Average redundancy	9.0
Completeness, overall (%)	97.9 (74.8)
Mean I/σ(I)	34.72 (2.08)
R_{merge}†(%)	7.4 (49.8)

885

886 † $R_{\text{merge}} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl) \rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where Σ_{hkl} denotes the sum over all reflections and Σ_i the sum
887 over all equivalent and symmetry-related reflections. (Stout & Jensen 1968)
888

889

890

891

892

893

894

Table 2 Refinement statistics and results of *MolProbity* validation

†Clash score is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.

Protein	Reassembled GH9 and CBM3c (Cel9I)
Space group	$P2_12_12_1$
Resolution range	30-1.68
No. of reflections in working set	71559
No. of reflections in test set	3580
No. of protein atoms	5071
No. of solvent atoms	835
No. of Cl ion atoms	3
No. of Ca ion atoms	2
Overall B factor from Wilson plot ($Å^2$)	16.06
Averaged B factor ($Å^2$)	21.12
R_{cryst}	0.1441
R_{free}	0.1759
Geometry	
RMS bonds (Å)	0.014
RMS bond angles (°)	1.371
MolProbity validation	
Ramachandran favored (%) (goal >98%)	96.7
Ramachandran outliers (%) (goal <0.2%)	0.5
C_β deviations >0.25Å (goal 0)	1
† Clash score (all atoms)	2.88
Rotamer outliers (%) (goal<1%)	0.8
Residues with bad bonds (%) (goal <1%)	0.00
Residues with bad angles (%) (goal <0.5)	0.33

895
896
897

898 **Figure captions:**

899 **Figure 1.** Schematic diagram of the Cel9I gene product (top) and the recombinant
900 proteins (A-D) prepared for this study. The GH9 module alone (B) was prepared with
901 and without an N-terminal His tag (shown schematically in the figure), and the
902 CBM3c's were prepared with C-terminal His tags. Scale shows the number of amino
903 acid residues and the boundaries of the different regions of the protein.

904 **Figure 2.** Recovery of activity upon association of CBM3c (with and without linker)
905 and GH9. CMCase activity (μmol reducing sugar released in a 10-min reaction) of
906 His-tagged GH9, mixed either with CBM3c*L* (diamonds) or CBM3c*NL* (squares),
907 was examined. A fixed amount (70 pmol) of the GH9 catalytic module was mixed
908 with increasing amounts of the indicated helper module, and their respective activities
909 were compared to that of the intact Cel9I core (GH9-CBM3c, set as 100%).

910 **Figure 3.** Reassembled GH9-CBM3c from Cel9I**.** C and N termini are indicated,
911 and the break between the GH9 and CBM3c modules is marked by a red ellipse. A.
912 The *in vitro* reassembled complex of the catalytic (GH9, wheat) and carbohydrate-
913 binding (CBM3c, green) modules of Cel9I from *C. thermocellum,* cartoon
914 representation. Calcium atoms are shown as magenta-colored spheres. B. Stereo-view
915 (cross-eyed) of the superposition of the reassembled GH9-CBM3c structure of *C.*
916 *thermocellum* Cel9I (red) with the bimodular structures of *C. cellulolyticum* Cel9G
917 (blue) and *T. fusca* Cel9A (green).

918 **Figure 4**. Structural components of the reassembled *C. thermocellum* GH9-CBM3c.
919 A. Structure of the GH9 catalytic module*,* cartoon representation. Twelve α-helices
920 form an $(\alpha/\alpha)_6$-barrel fold. Pairs of helices, comprising the fold, are emphasized by
921 red, blue, yellow, magenta, cyan and green. B. Surface representation of the
922 reassembled GH9-CBM3c complex. The residues are shaded according to the extent
923 of their conservation with Cel9G from *C. cellulolyticum* and Cel9A from *T. fusca*.
924 Darker blue indicates higher conservation. Top, birds-eye view of the catalytic cleft.
925 Bottom, lateral view, showing the flat surface (red bar). Pink ellipse indicates the
926 catalytic cleft, and green ellipse designates terminal portion of the catalytic site. C.
927 Close-up (same orientation as in B, top) of the catalytic cleft of the Cel9I GH9
928 module showing functional residues. Carbohydrate-binding residue carbons are
929 colored gray, catalytic residue carbons are colored yellow. Loop 243-254 carbons are

colored in light blue. D. Calcium-binding site of the *C. thermocellum* Cel9I GH9 module. Coordinating residues are shown in stick representation. The calcium ion is colored magenta, and distances to the coordinating atoms are indicated.

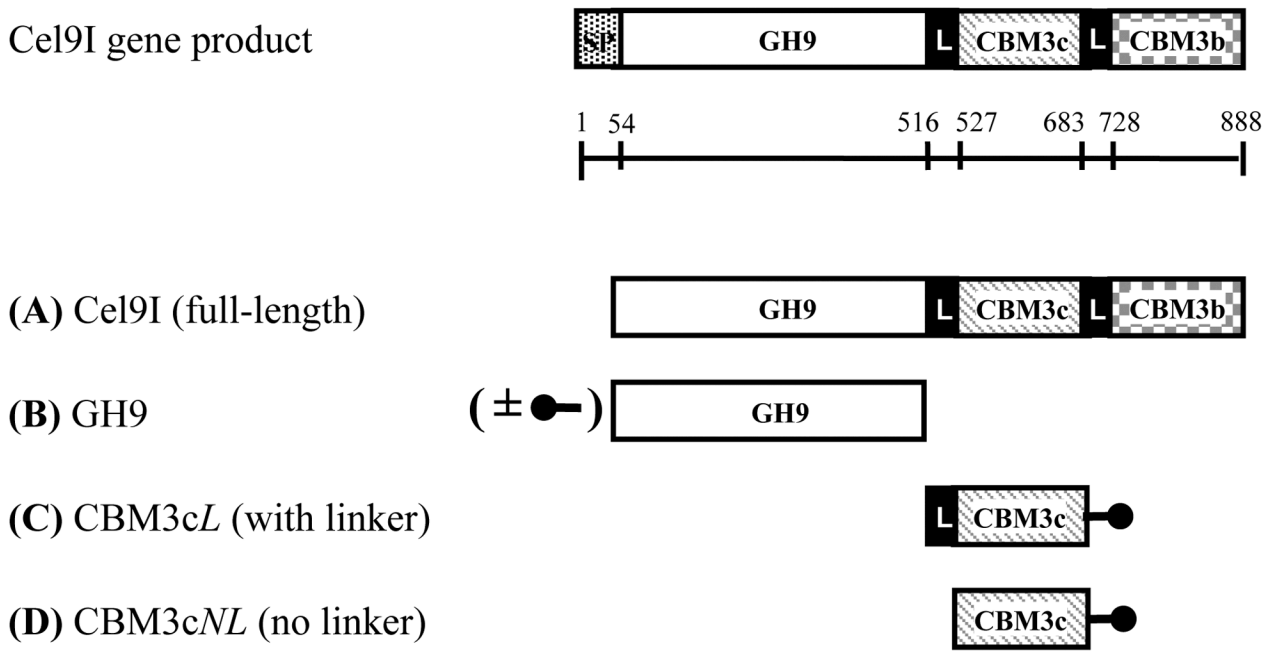
Figure 5. Structure of the CBM3c of Cel9I from *C. thermocellum.* C and N termini are indicated A. Cartoon representation, β-strands are numbered according to the alignment with Cel9G from *C. cellulolyticum*, and Cel9A from *T. fusca*. B. Calcium-binding site of the CBM3c. C. Birds-eye view of the flat surface. Residues are shaded according to their degree of conservation with *C. cellulolyticum* Cel9G and *T. fusca* CEL9A. Surface-exposed conserved residues are shown in stick representation. D. Shallow groove of the CBM3c. Conserved surface residues are shown in stick representation. The residues are colored according to the degree of conservation in CBM3a, CBM3b and CBM3c modules derived from the sequences listed in the Methods section.

Figure 6. Contact residues of the reassembled GH9-CBM3c complex. A. Contact residues of the GH9 module are colored orange, of the CBM3c module green. Contact residues between the linker and the catalytic domain are indicated (in green: CBM3c residues, in brown: GH9 residues). B. Alignment of the GH9 and CBM3c modules of *C. thermocellum* Cel9I, *C. cellulolyticum* Cel9G, and *T. fusca* Cel9A (E4) cellulases. Contact residues are highlighted in yellow. Only the relevant regions of the alignment are shown. Residues of linker sequences are shown blue font.

Figure 7. Representative ITC titration of (A) GH9 and CBM3c*NL* (B) GH9 and CBM3c*L*. The top panel shows the calorimetric titration and the bottom panel displays the integrated injection heats corrected for control dilution heat. The solid line is the curve of the best fit used to derive the binding parameters, and the fitted data describe an interaction of a one binding site model.

956 **Figure 1. Schematic diagram of the Cel9I gene product and the recombinant**
957 **proteins prepared for this study**

958

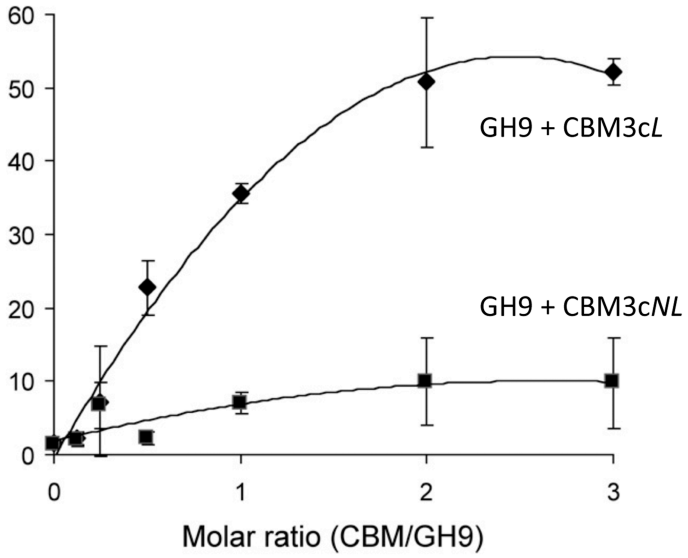


959

960
961
962
963
964

965 **Figure 1.** Schematic diagram of the Cel9I gene product (top) and the recombinant
966 proteins (A-D) prepared for this study. The GH9 module alone (B) was prepared with
967 and without an N-terminal His tag (shown schematically in the figure), and the
968 CBM3c's were prepared with C-terminal His tags. Scale shows the number of amino
969 acid residues and the boundaries of the different regions of the protein.
970

971 **Figure 2. Recovery of activity upon association of CBM3c (with and without linker) and**
972 **GH9.**
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989



990 Figure 2. Recovery of activity upon association of CBM3c (with and without linker)
991 and GH9. CMCase activity (μmol reducing sugar released in a 10-min reaction) of
992 His-tagged GH9, mixed either with CBM3c*L* (diamonds) or CBM3c*NL* (squares),
993 was examined. A fixed amount (70 pmol) of the GH9 catalytic module was mixed
994 with increasing amounts of the indicated helper module, and their respective activities
995 were compared to that of the intact Cel9I (set as 100%).
996

Figure 3. Reassembled GH9-CBM3c from Cel9I.

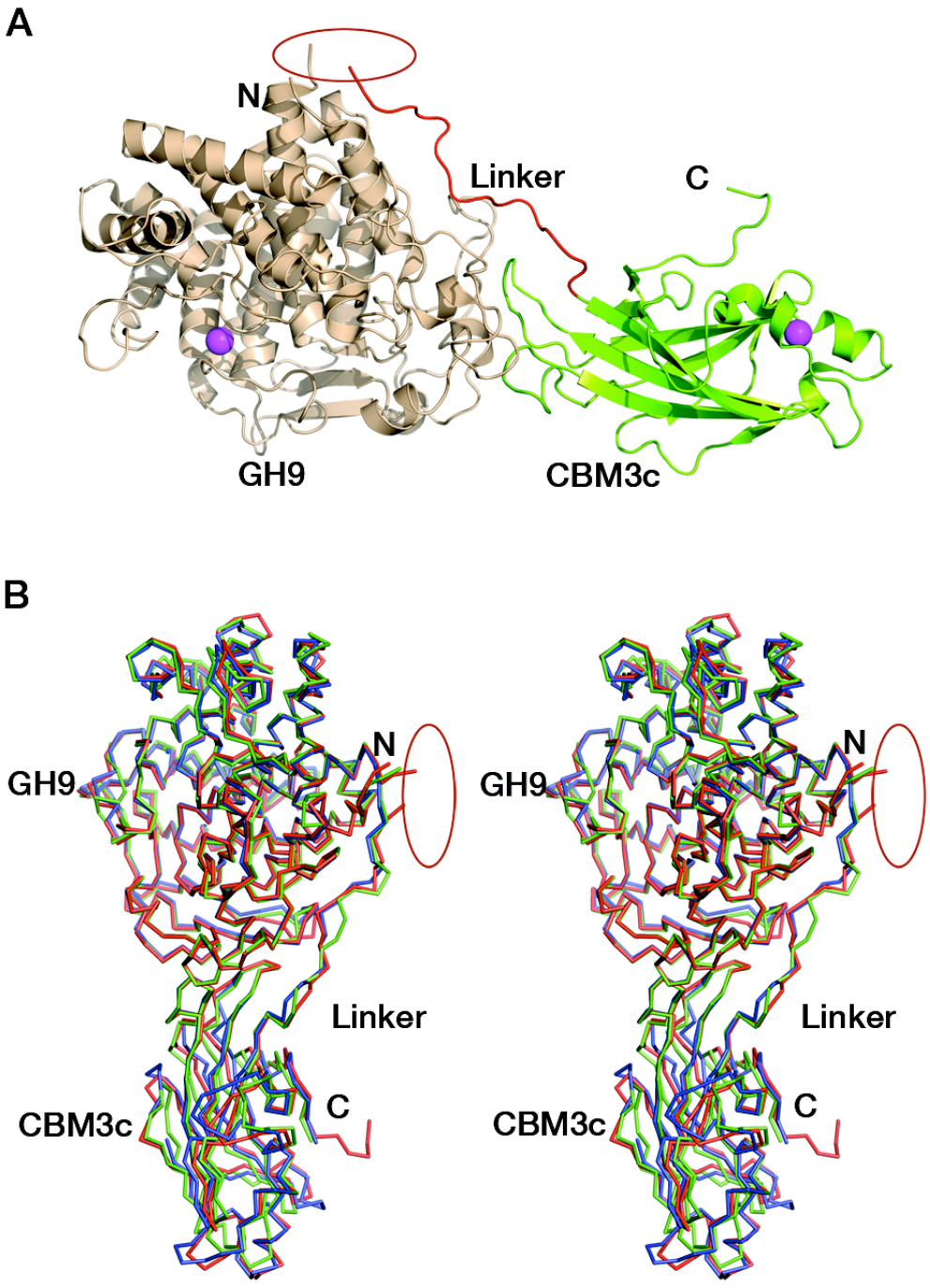
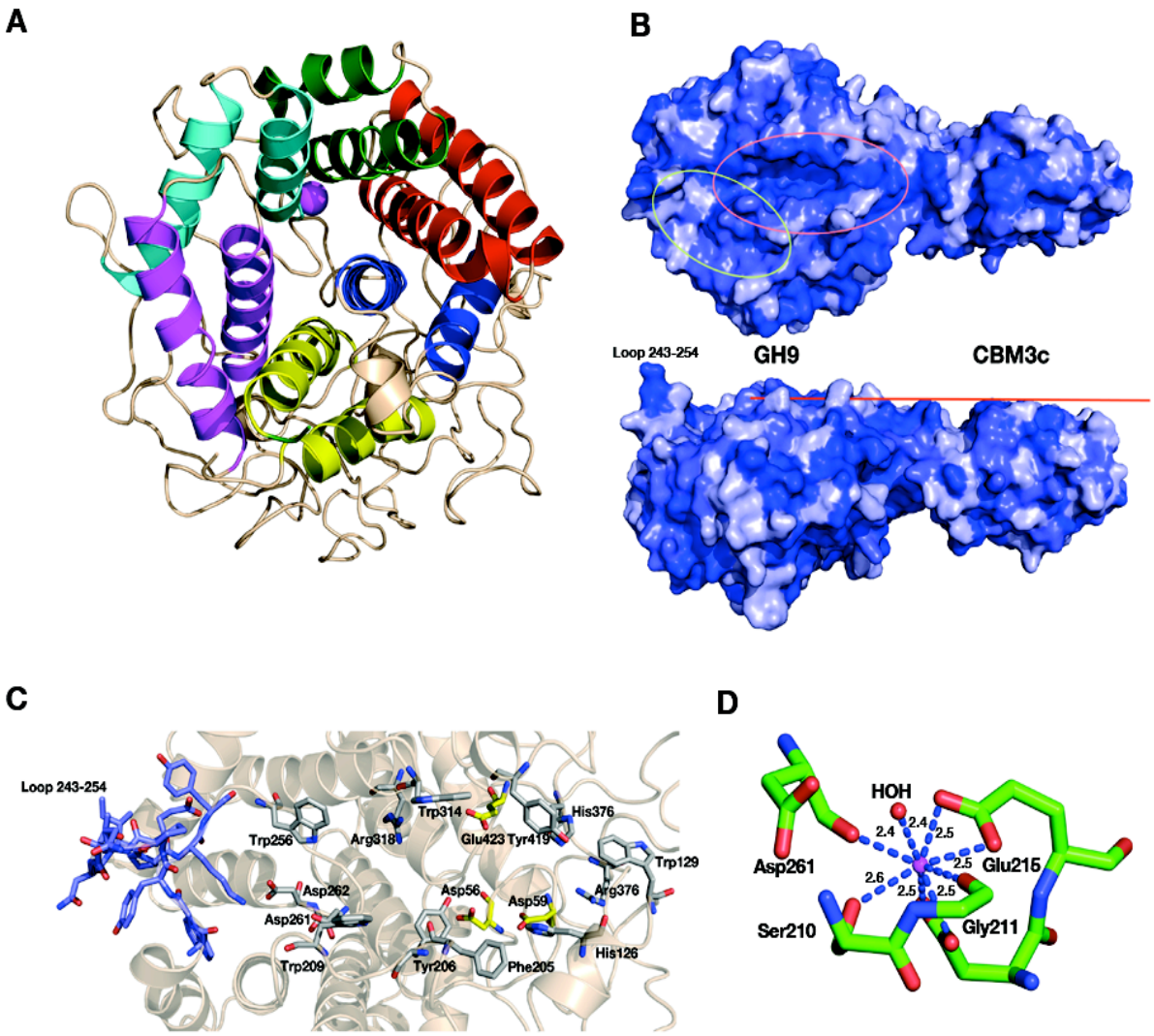


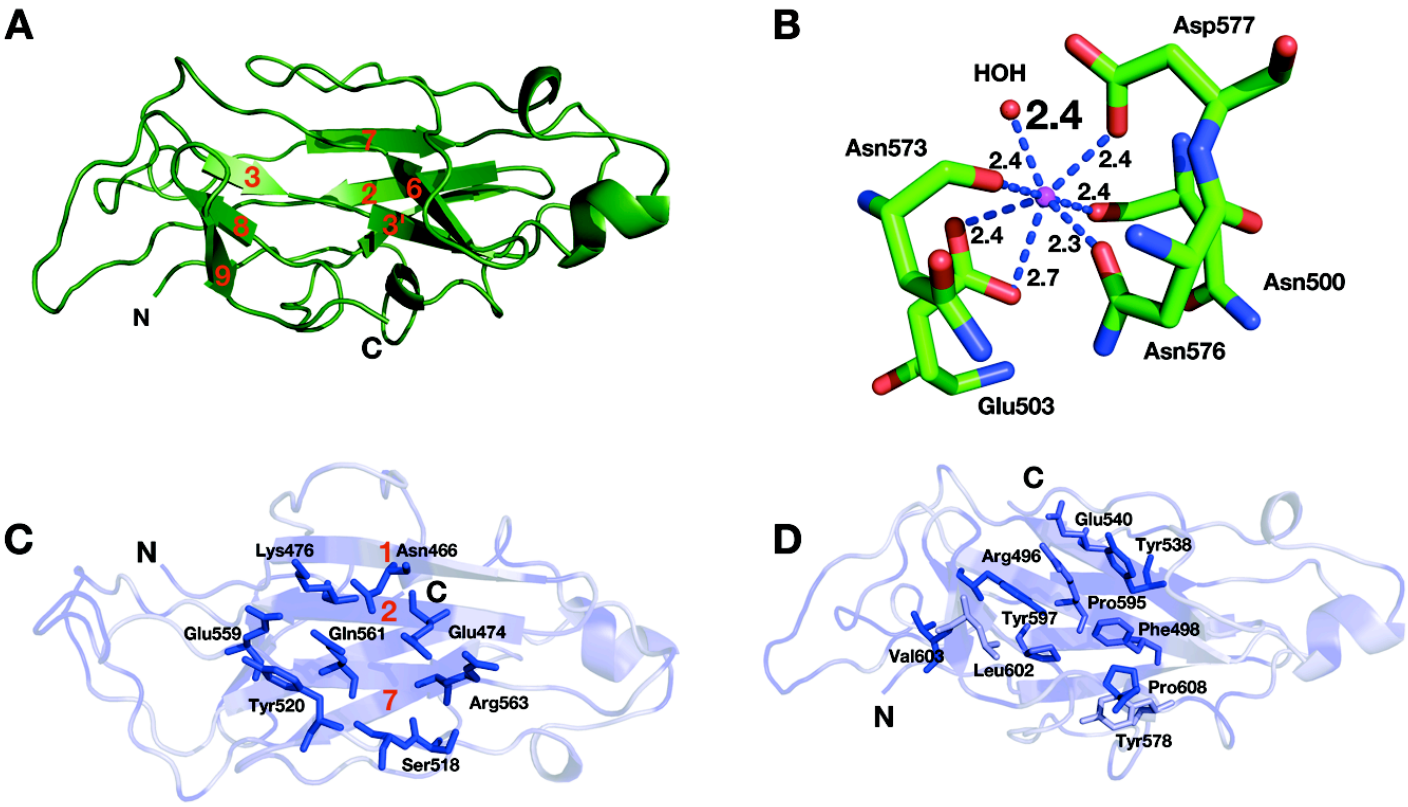
Figure 3. Reassembled GH9-CBM3c from Cel9I. C and N termini are indicated, and the break between the GH9 and CBM3c modules is marked by a red ellipse. A. The *in vitro* reassembled complex of the catalytic (GH9, wheat) and carbohydrate-binding (CBM3c, green) modules of Cel9I from *C. thermocellum,* cartoon representation. Calcium atoms are shown as magenta-colored spheres. B. Stereo-view (cross-eyed) of the superposition of the reassembled GH9-CBM3c structure of *C. thermocellum* Cel9I (red) with the bimodular structures of *C. cellulolyticum* Cel9G (blue) and *T. fusca* Cel9A (green).

1035 **Figure 4. Structural components of the reassembled GH9-CBM3c.**



1036
1037
1038
1039 **Figure 4.** Structural components of the reassembled GH9-CBM3c. A. Structure of the
1040 GH9 catalytic module, cartoon representation. Twelve α-helices form an $(\alpha/\alpha)_6$-barrel
1041 fold. Pairs of helices, comprising the fold, are emphasized by red, blue, yellow,
1042 magenta, cyan and green. B. Surface representation of the reassembled GH9-CBM3c
1043 complex. The residues are shaded according to the extent of their conservation with
1044 Cel9G from *C. cellulolyticum* and Cel9A from *T. fusca*. Darker blue indicates higher
1045 conservation. Top, birds-eye view of the catalytic cleft. Bottom, lateral view, showing
1046 the flat surface (red bar). Pink ellipse indicates the catalytic cleft, and green ellipse
1047 designates terminal portion of the catalytic site. C. Close-up (same orientation as in B,
1048 top) of the catalytic cleft of the Cel9I GH9 module showing functional residues.
1049 Carbohydrate-binding residue carbons are colored gray, catalytic residue carbons are
1050 colored yellow. Loop 243-254 carbons are colored in light blue. D. Calcium-binding
1051 site of the *C. thermocellum* Cel9I GH9 module. Coordinating residues are shown in
1052 stick representation. The calcium ion is colored magenta, and distances to the
1053 coordinating atoms are indicated.
1054

1055 **Figure 5. Structure of the CBM3c of Cel9I from *C. thermocellum.***



1056

1057

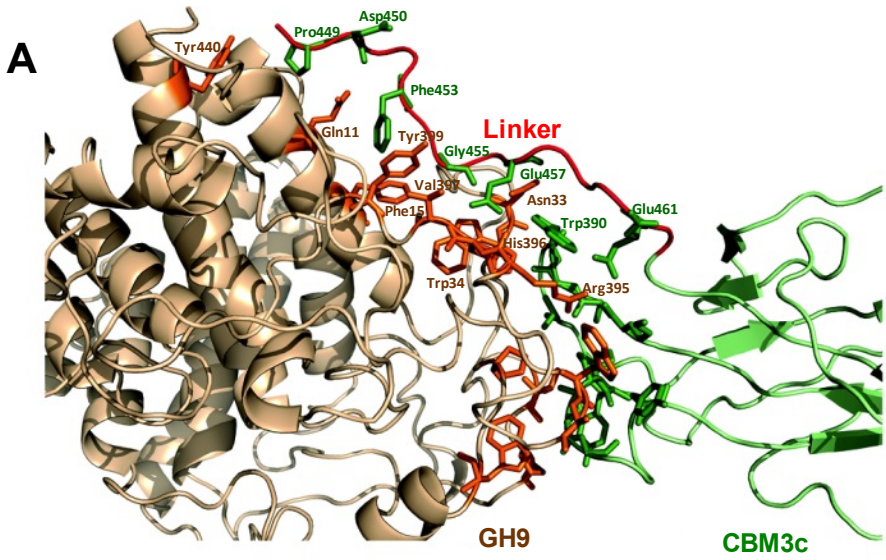
1058

1059

1060 **Figure 5.** Structure of the CBM3c of Cel9I from *C. thermocellum.* C and N termini
1061 are indicated A. Cartoon representation, β-strands are numbered according to the
1062 alignment with Cel9G from *C. cellulolyticum*, and Cel9A from *T. fusca*. B. Calcium-
1063 binding site of the CBM3c. C. Birds-eye view of the flat surface. Residues are shaded
1064 according to their degree of conservation with *C. cellulolyticum* Cel9G and *T. fusca*
1065 CEL9A. Surface-exposed conserved residues are shown in stick representation. D.
1066 Shallow groove of the CBM3c. Conserved surface residues are shown in stick
1067 representation. The residues are colored according to the degree of conservation in
1068 CBM3a, CBM3b and CBM3c modules derived from the sequences listed in the
1069 Methods section.

1070

1071 **Figure 6. Contact residues of the reassembled GH9-CBM3c complex.**
1072
1073



B **GH9 Modules**

```
CelI gAFNYGEALQKAIFFYECQRSGKLDPsTLRLNWRGDSGLDDGKDAGIDLTGGWYDAGDHV   60
CelG .TYNYGEALQKSIMFYEFQRSGDLPA.DKRDNWRDDSGMKDGSDVGVDLTGGWYDAGDHV   58
E4   PAFNYAEALQKSMFFYEAQRSGKLP.ENNRVSWRGDSGLNDGADVGLDLTGGWYDAGDHV   60

     ------------------------------------------------------------

CelI GHADHAWWGPAEVMPMERPSYKVDRSSPGSTVVAETSAALAIASIIFKKVDGEYSKECLK  180
CelG GGKDHSWWGPAEVMQMERPSFKVDASKPGSAVCASTAASLASAAVVFKSSDPTYAEKCIS  178
E4   GDADHKWWGPAEVMPMERPSFKVDPSCPGSDVAAETAAAMAASSIVFADDDPAYAATLVQ  180

     ------------------------------------------------------------

CelI GRSFVVGFGENPPKRPHHRTAHGSWADSQMEPPEHRHVLYGALVGGP.DSTDNYTDDISN  420
CelG GRSFVVGYGVNPPQHPHHRTAHGSWTDQMTSPTYHRHTIYGALVGGP.DNADGYTDEINN  416
E4   NSSYVVGFGNNPPRNPHHRTAHGSWTDSIASPAENRHVLYGALVGGPgSPNDAYTDDRQD  422

     ------------------------------------------------------------

CelI YTCNEVACDYNAGFVGLLAKMYKLYGE  444
CelG YVNNEIACDYNAGFTGALAKMYKHSGG  440
E4   YVANEVATDYNAGFSSALAMLVEEYG.  445
```

CBM3c Modules

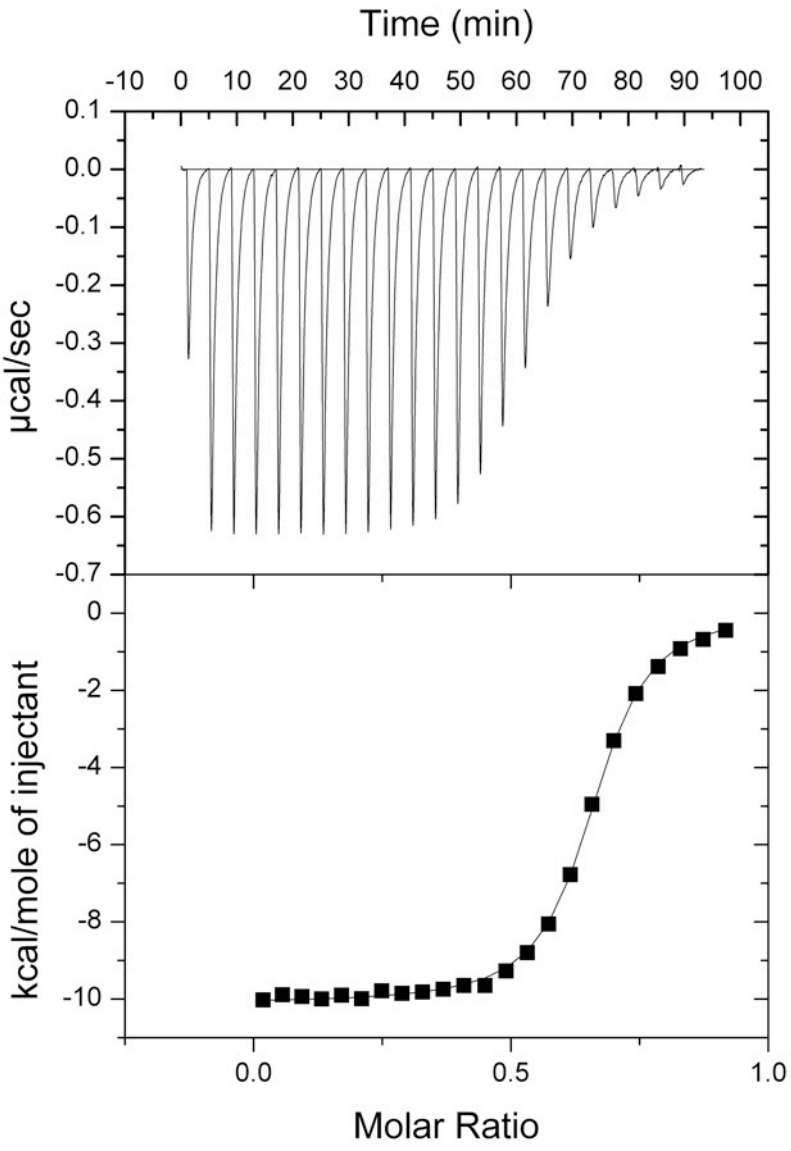
```
CelI SPDPKFNGIEEVPEDEIFVEAGVNASGNNFIEIKAIVNNKSGWPARVCENLSFRYFINIE   60
CelG DPIPNFKAIEKITNDEVIIKAGLNSTGPNYTEIKAVVYNQTGWPARVTDKISFKYFMDLS   60
E4   TPLADFPPTEEPDGPEIFVEAQINTPGTTFTEIKAMIRNQSGWPARMLDKGTFRYWFTLD   61

CelI EIVNAGKSASDLQVSSsYNQG..AKLS.DVKHYK..DNIYYVEVDLSGTKIYPGGQSAYKK 116
CelG EIVAAGIDPLSLVTSS.YSEGknTKVS.GVLPWDvsNNVYYVNVDLTGENIYPGGQSACRR 119
E4   EGV....DPADITVSSAYNQ...CATPeDVHHVS..GDLYYVEIDCTGEKIFPGGQSEHRR 111
```

1074
1075

1076 **Figure 6.** Contact residues of the reassembled GH9-CBM3c complex. A. Contact
1077 residues of the GH9 module are colored orange, of the CBM3c module green. Contact
1078 residues between the linker and the catalytic domain are indicated (in green: CBM3c
1079 residues, in brown: GH9 residues). B. Alignment of the GH9 and CBM3c modules of
1080 *C. thermocellum* Cel9I, *C. cellulolyticum* Cel9G, and *T. fusca* Cel9A (E4) cellulases.
1081 Contact residues are highlighted in yellow. Only the relevant regions of the alignment
1082 are shown. Residues of linker sequences are shown blue font.

1083 **Figure 7. Representative ITC titration of GH9 and CBM3c*NL*.**



1084

1085
1086
1087
1088
1089
1090

1091 Figure 7. Representative ITC titration of GH9 and CBM3c*NL*. The top panel shows
1092 the calorimetric titration and the bottom panel displays the integrated injection heats
1093 corrected for control dilution heat. The solid line is the curve of the best fit used to
1094 derive the binding parameters, and the fitted data describe an interaction of a one
1095 binding site model.
1096

1097

1098