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.

1 **Reassembly and co-crystallization of a family 9 processive**

2 **endoglucanase from its component parts:**

3 **Structural and functional significance of the intermodular linker**

5 **Svetlana Petkun[1], Inna Rozman Grinberg[1], Raphael Lamed[1], Sadanari Jindou[2], Tal**

6 **Burstein[1], Oren Yaniv[1], Yuval Shoham[3], Linda J.W. Shimon[4], Edward A. Bayer[5],*, and**

7 **Felix Frolow[1]**

9 *[1]Department of Molecular Microbiology and Biotechnology, The Daniella Rich Institute for*

10 *Structural Biology, Tel Aviv University, Ramat Aviv 69978 ISRAEL*

11 *[2]Department of Life Sciences, Meijo University, Nagoya 468-8502 JAPAN*

12 *[3]Department of Biotechnology and Food Engineering, Technion-Israel Institute of Technology,*

13 *Haifa 32000 ISRAEL*

14 *[4]Department of Chemical Research Support and*

15 *[5]Department of Biological Chemistry, The Weizmann Institute of Science, Rehovot 76100*

16 *ISRAEL*

18 **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

20 **Running title**: Crystal structure of reassembled Cel9I

21 **Abbreviations used**: CBM, carbohydrate-binding module; CBM3c*L,* family 3c CBM with

22 linker; CBM3c*NL,* family 3c CBM without linker; CMC, carboxymethyl cellulose; GH9, family

23 9 glycoside hydrolase; ITC, isothermal titration calorimetry; PASC, phosphoric acid-swollen

24 cellulose; PEG, polyethylene glycol; SeMet, selenium-methionine labeled derivative.

25 **Abstract.**

26   Non-cellulosomal processive endoglucanase 9I (Cel9I) from *Clostridium thermocellum* is a

27   modular protein, consisting of a family-9 glycoside hydrolase (GH9) catalytic module and two

28   family-3 carbohydrate-binding modules (CBM3c and CBM3b), separated by linker regions. GH9

29   does not show cellulase activity when expressed without CBM3c and CBM3b and the presence

30   of the CBM3c was previously shown to be essential for endoglucanase activity. Physical

31   reassociation of independently expressed GH9 and CBM3c modules (containing linker

32   sequences) restored 60-70% of the intact Cel9I endocellulase activity. However, the mechanism

33   responsible for recovery of activity remained unclear. In this work we independently expressed

34   recombinant GH9 and CBM3c with and without their interconnecting linker in *Escherichia coli*.

35   We crystallized and determined the molecular structure of the GH9/linker-CBM3c heterodimer

36   at a resolution of 1.68 Å to understand the functional and structural importance of the mutual

37   spatial orientation of the modules and the role of the interconnecting linker during their re-

38   association. Enzyme activity assays and isothermal titration calorimetry were performed to study

39   and compare the effect of the linker on the re-association. The results indicated that reassembly

40   of the modules could also occur without the linker, albeit with only very low recovery of

41   endoglucanase activity. We propose that the linker regions in the GH9/CBM3c endoglucanases

42   are important for spatial organization and fixation of the modules into functional enzymes.

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 highly efficient multi-enzyme complex, termed cellulosome (Bayer et al. 2004; Lamed et al. 1983a; Lamed et al. 1983b; Shoham et al. 1999), capable of converting a wide variety of plant-derived polysaccharides directly into soluble sugars and fermentation products (Béguin & Alzari 1998; Felix & Ljungdahl 1993; Schwarz 2001; Schwarz et al. 2004). These capabilities render *C. thermocellum* a high utility candidate for use in consolidated bioprocessing (CBP) applications [reviewed in (Akinosho et al. 2014)].

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

75   exo-acting enzymes is also reflected by the architecture of the respective class of active site,

76   whereby endoglucanases, for example, are commonly characterized by a groove or open binding

77   cleft, into which any part of the linear cellulose chain can fit. On the other hand, the

78   exoglucanases bear tunnel-like active sites, which can only accept a substrate chain via its

79   terminus (either the reducing or non-reducing end, depending on the enzyme), thereby cleaving

80   cellulose in a sequential manner. The sequential hydrolysis of a cellulose chain has earned the

81   term "processivity" (Beckham et al. 2014; Davies & Henrissat 1995; Wilson & Kostylev 2012),

82   and processive enzymes are considered to be key components which contribute to the overall

83   efficiency of a given cellulase system. Some endoglucanases, notably from GH family 9, have

84   also been shown to sequentially hydrolyze cellulose chains and are thus referred to as processive

85   endoglucanases (Gal et al. 1997; Gilad et al. 2003; Irwin et al. 1998; Jeon et al. 2012; Kuusk et

86   al. 2015; Zverlov et al. 2003). Such enzymes appear to possess extended catalytic clefts and the

87   observed processivity appears to require highly coordinated substrate-binding affinities from

88   opposite sides of the cleavage site (Bu et al. 2012; Li et al. 2010; Payne et al. 2011).

89   Cellulase 9I (Cel9I), is a non-cellulosomal family 9 processive endoglucanase from

90   *Clostridium thermocellum*, which degrades crystalline cellulose (Avicel and filter paper) as well

91   as phosphoric acid-swollen cellulose (PASC) and carboxymethyl cellulose (CMC) (Gilad et al.

92   2003). This enzyme consists of a catalytic GH9 module at its N terminus, followed by two

93   family 3 carbohydrate-binding modules (CBMs): CBM3c and CBM3b. The three modules are

94   separated by distinctive linker sequences. Such intermodular linker segments were proposed to

95   be important for the physical association of the modules in the space, and to promote

96   intermodular and/or intersubunit protein–protein interactions (Bayer et al. 1998; Bayer et al.

97   2009; Noach et al. 2008).

98   The C-terminal CBM3b module, as a classic CBM3, is responsible for targeting the Cel9I

99   enzyme to the planar surface of the crystalline cellulose substrate (Gilad et al. 2003; Su et al.

100   2012; Tormo et al. 1996). It has also been proposed to disrupt the crystalline regions of cellulose,

101   rendering it more accessible to the GH9 catalytic module (Yi et al. 2013) and to contribute to

102   enzyme processivity by preventing the desorption of the catalytic module from cellulose (Telke

103   et al. 2012). The function of the CBM3c is less straightforward. Removal of CBM3c from *C.*

104   *thermocellum* Cel9I and from *C. cellulolyticum* Cel9G. *P. Barcinonensis* Cel9B significantly

105   reduces the enzyme activity (Burstein et al. 2009; Chiriac et al. 2010; Gal et al. 1997). CBM3c

106   modules have been shown to alter the normal function of the GH9 catalytic module of

107   *Thermobifida fusca* Cel9A from the standard endo-acting mode into a processive endoglucanase

108   (Bayer et al. 1998; Irwin et al. 1998). Thus, Gilad *et* al.(Gilad et al. 2003) showed in 2003 that

109   the endoglucanase activity of Cel9I is dependent upon the presence of the CBM3c module and

110   suggested that the fused CBM3c serves an important accessory role for the catalytic domain by

111   altering its character to facilitate processive cleavage of recalcitrant cellulose substrates.

112   In addition to the Cel9 CBM3c, several other examples of CBMs that are considered to

113   modulate catalytic specificity and act cooperatively with the catalytic domain have recently been

114   discovered. These include CBM66 that directs the cognate enzyme towards highly branched

115   glucans rather than linear fructose polymers (Cuskin et al. 2012), CBM48 that contributes to

116   substrate binding at the active site of a glucan phosphatase (Meekins et al. 2014), family-43 β-

117   xylosidases where the GH43 is complemented by an additional module that confers hydrolytic

118   activity to the mature enzyme (Moraïs et al. 2012), and CBM46, that constitutes part of the

119   catalytic cleft required for the hydrolysis of β-1,3-1,4-glucans (Venditto et al. 2015). The

120   carbohydrate-binding PA14 domain is also known to affect substrate binding of the catalytic

121   domain by contributing to the formation of its active site (Gruninger et al. 2014; Zmudka et al.

122   2013).

123   We have previously shown that independently expressed GH9 and linker-containing CBM3c

124   modules of Cel9I readily re-associate *in vitro* and that this physical reassociation recovers 60-

125   70% of the intact Cel9I endoglucanase activity (Burstein et al. 2009).

126   We have examined in this work the interaction of the CBM3c with the catalytic module either

127   with or without the intermodular linker in order to better understand the function of the CBM3c

128   in the family-9 enzymes and the role of the linkers regions. The effect of the re-association of the

129   CBM3c with linker (CBM3c*L*) and the CBM3c without linker (CBM3c*NL*) on the enzymatic

130   activity of GH9 has been studied by the crystallization and structure determination of the

131   reassembled GH9-CBM3c*L* complex at a resolution of 1.68 Å. The results of this study will help

132   us to understand the contribution of ancillary modules in the action of multi-modular glycoside

133   hydrolases.

134

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′ CCATGGGCGAAGTTCCGGAGGATGAAATA 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$_2$PO$_4$, 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$_2$PO$_4$ 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

164   size-exclusion chromatography. Fast protein liquid chromatography (FPLC) was performed

165   using a Superdex 75pg column and ÄKTA Prime system (GE Healthcare, Piscataway, NJ) to

166   further purify the complex. One peak, corresponding approximately to 70 kDa, matching the

167   predicted molecular weight of the GH9-CBM3c complex, was observed in the chromatogram.

168   The 15 amino-acid linker sequence (about 1.5 kDa) did not significantly affect the elution

169   volume, compared to that of the complex without the linker, presumably due to the limited

170   resolution of the column. The relevant fractions (the purified complexed proteins) were analyzed

171   by 15% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) with

172   Coomassie brilliant blue staining. Two clear bands, of about 52 and 19.5 kDa were observed.

173   The rearranged modules were concentrated to 6 mg ml^{-1} using Centriprep YM-3 centrifugal filter

174   devices YM-3 (Amicon Bioseparation, Millipore Corporation, Bedford, USA). Protein

175   concentration was determined by measuring UV absorbance at 280 nm.

176   The full-length Cel9I was purified by affinity chromatography on Avicel as reported earlier

177   (Burstein et al. 2009; Gilad et al. 2003).

178   **Microcalorimetric analysis**

179   Isothermal titration calorimetry (ITC) experiments were carried out using a VP-ITC

180   MicroCalorimeter (MicroCal, LLC, Northampton, MA) at 298 K. About 300 µM solution of

181   CBM3c*NL* was injected into a 65 µM solution of GH9. The reaction was performed in a buffer

182   containing 50 mM Tris–HCl, pH 7.5, 150 mM NaCl, 0.05% sodium azide. Heats of dilution of

183   the titrants were subtracted from the titration data, and the corrected data were analyzed using the

184   Origin ITC analysis software package supplied by MicroCal. Thermal titration data were fit to

185   the one binding site model, and enthalpy (ΔH), entropy (ΔS), association constant (Ka) and

186   stoichiometry of binding (N) were determined. In all cases, the calculated stoichiometry (N) was

187   lower than one, most likely due to the fact that the CBM3 proteins lost their native functionality

188   with time. For the analysis, the CBM3 protein concentrations were corrected as to provide a

189   stoichiometry of one. Two titrations were performed to evaluate reproducibility.

190   **Enzyme activity assay**

191   Reactions were performed at 333 K, in 50 mM citrate buffer (pH 6.0). The soluble cellulolytic

192   substrate was carboxymethyl cellulose (CMC, Sigma Chem. Co. St. Louis, MO). The amount of

193 reducing sugars released from the substrate was determined with the 3,5-dinitrosalicylic acid

194 (DNS) reagent as described by Miller et al (Miller 1959). Activity was defined as the amount

195 (micromole) of reducing sugar released after 10 min of reaction.

Crystallization

197 Initially the protein samples containing 6 mg/ml protein solution in 1.2 mM Tris-HCl pH 7.5,

198 1.5 mM sodium chloride, 0.025% sodium azide, were screened, using the microbatch

199 crystallization method under 1:1 mixture of silicon and paraffin oil (Chayen et al. 1990), using

200 288 conditions from the Hampton Research HT screens (SaltRx, Index HT, and Crystal Screen

201 HT; Hampton Research, Aliso Viejo, CA) and 96 conditions of the Wizard Crystallization kit

202 from Emerald BioSystems (Rigaku Reagents, Bainbridge Island, WA). The dyad of GH9 and

203 CBM3c*NL* did not yield any crystals. Screening of the GH9-CBM3c*L* resulted in plate-like

204 crystals that appeared after several days under several conditions, all of which contained PEG

205 3350 and 4000. The best crystals were obtained in 30 % PEG (both 3350 and 4000), 0.2 M

206 magnesium chloride, and 0.1 M Hepes, pH 7.5. Attempts to optimize this condition using

207 microbatch, hanging-drop, and sitting drop methods were unsuccessful, as the crystals remained

208 very thin and fragile. The superfine Eyelash (Ted Pella, Inc, Redding, CA) was used to touch

209 these crystals and consequently to streak the sitting drops, composed of 5 μl of the protein

210 solution and 5 μl of the precipitating solution (24 % PEG 3350, 0.2 M magnesium chloride, 0.1

211 M Hepes, pH 8.0). After one day, crystals of different morphology, with maximum size of about

212 0.05 mm, appeared in the drop.

Data collection and crystallographic analysis

214 The crystals of the GH9-CBM3c*L* complex were harvested from the crystallization drop using

215 a nylon cryo-loop (Hampton Research, Aliso Viejo, CA). For data collection, crystals were

216 mounted on the MiTeGen stiff micro-mount (MiTeGen, Ithaca, NY) made of polyimide and

217 flash-cooled in a nitrogen stream produced by Oxford Cryostream low temperature generator

218 (Cosier & Glazer 1986) at a temperature of 100 K. Mother-liquor of the crystals served for cryo-

219 protection during the cooling in liquid nitrogen.

220 Diffraction data from the GH9-CBM3c*L* crystals were measured using the ID23-2 beam line

221 at ESRF, Grenoble, France. A MAR CCD 225 area detector and X-ray radiation of 0.873 Å

222 wavelength were used. Diffraction data of 480 images with 0.5° oscillation per image were

223 collected. Data were processed with *DENZO* and scaled with *SCALEPACK* as implemented in

224 *HKL2000* (Otwinowski & Minor 1997). The crystals diffracted to 1.68 Å resolution and belong

225 to the orthorhombic space group $P2_12_12_1$, with unit cell parameters a=70.4, b=88.5, c=106.5 Å.

226 There is one GH9-CBM3c*L* complex per asymmetric unit with a Matthews density V_M of 2.37

227 $Å^3 Da^{-1}$, corresponding to a solvent content of 48.15% (Matthews 1968). The X-ray data analysis

228 statistics are presented in Table 1 (Stout & Jensen 1968).

229   Molecular replacement was carried out with *MOLREP* (Vagin & Teplyakov 1997), using the

230 coordinates of the GH9 and CBM3c modules of endoglucanase 9G from *Clostridium*

231 *cellulolyticum* (PDB code 1G87, 66 and 51% sequence identity, respectively), as a search model.

232 The *MOLREP* calculations with the GH9 domain converged into a clear solution with 1 molecule

233 in the asymmetric unit with an R-factor of 0.533 and correlation coefficient of 0.567. This

234 solution was inserted into *MOLREP* calculations as a fixed molecule and the coordinates of

235 CBM3c module were used for the search producing a solution with an R_{cryst} of 0.505, and

236 correlation coefficient of 0.582. The resulting model with 5% of reflections forming test set

237 (Brünger 1992) was subjected to 10 cycles of restrained refinement using anisotropic B-factors,

238 yielding the R_{cryst} and R_{free} 0.329 and 0.359, respectively (*REFMAC5*) (Murshudov et al. 1997).

239 Automated model building by *ARP/wARP* (Perrakis et al. 1999) produced a complete structure

240 with R_{cryst} and R_{free} of 0.218 and 0.243 respectively. The model was manually corrected using

241 *COOT* (Emsley & Cowtan 2004) and refined using *REFMAC5* (Murshudov et al. 1997). The

242 R_{cryst} and R_{free} improved to 0.184 and 0.228, respectively. Solvent atoms were built using

243 *ARP/warp* (Perrakis et al. 1999). Refinement of TLS (rigid body translation/libration/screw

244 motions) parameters was performed (Winn et al. 2001; Winn et al. 2003). The model was

245 subjected to several additional cycles of manual rebuilding and refinement. The model

246 converged to final R_{cryst} and R_{free} factors of 0.144 and 0.176, respectively.

247   The refinement statistics of the structure are summarized in Table 2. The structure was

248 validated using *MolProbity* (Davis et al. 2007).

249 **Protein sequence analysis**

250   Sequence alignments were performed using *CLUSTALW* (Larkin et al. 2007) and the coloring

251 of residues (representing degree of conservation) using ProtSkin (Deprez et al. 2005). Sources of

252 the sequences used in this work are as follows: *Clostridium thermocellum* Cel9I GH9 module,

253 CBM3c and CBM3b (AAA20892.1); *Clostridium cellulolyticum* Cel9G GH9 module, CBM3c

254 (AAA73868.1); *Thermobifida fusca* Cel9A GH9 module and CBM3c (AAB42155.1);

255 *Cellulomonas fimi* Ce9A CBM3c (AAA23086.1); *Clostridium cellulovorans* EngH CBM3c

256 (AAC38572.2) and CbpA CBM3a (AAA23218.1); *Clostridium stercorarium* CelZ CBM3c and

257 CBM3b (CAA39010.1) and CelY CBM3b (CAA93280.1); *Clostridium thermocellum* CipA

258 CBM3a (CAA48312.1), CelQ CBM3c (BAB33148.1), Cel9V CBM3c' and CBM3b'

259 (CAK22315.1), Cel9U CBM3c' and CBM3b' (CAK22317.1) and Cbh9A CBM3b

260 (CAA56918.1); *Clostridium cellulolyticum* CipC CBM3a (AAC28899.2) and CelJ CBM3c

261 (AAG45158.1); *Acetivibrio cellulolyticus* Cel9B CBM3c' and CBM3b' (CAI94607.1) and CipV

262 (ScaA) CBM3b (AAF06064.1); *Clostridium josui* CipA (CipJ) CBM3a (BAA32429.1);

263 *Bacteroides cellulosolvens* ScaA CBM3b (AAG01230.2); *Bacillus subtilis* CelA CBM3b

264 (AAA22307.1); *Pectobacterium atrosepticum* CelVI CBM3b (X79241.2); *Bacillus*

265 *licheniformis* CelA CBM3b (CAJ70714.1).

266

267

268 **Results**

269 **Cloning, expression and purification of Cel9I and its modular components**

270 The full-length *C. thermocellum* Cel9I enzyme and its individual component parts were over-

271 expressed in *Escherichia coli,* according to Burstein et al (2009), in order to investigate the

272 contribution of the ancillary modules and their linkers to the catalytic activity of the enzyme.

273 These include the isolated GH9 module with and without a His tag, the His-tagged CBM3c

274 module together with its adjacent N-terminal linker that connects it to the GH9 module

275 (CBM3c*L*) and His-tagged CBM3c module without the N-terminal linker (CBM3c*NL*). For

276 details, see Figure 1. Following purification procedures, all recombinant proteins showed a

277 single band in SDS-PAGE of the anticipated molecular masses.

278 **Recovery of endoglucanase activity upon association of CBM3c*NL* and GH9**

279 **compared to CBM3c*L* and GH9**

280   Previous works (Burstein et al. 2009; Gilad et al. 2003) demonstrated that the Cel9I catalytic

281 module alone has no detectable activity on CMC (carboxymethyl cellulose) and that adding the

282 CBM3c*L* to form the Cel9I-CBM3c*L*-CBM3b triad serves to recover up to 70% of the lost

283 activity. To further examine the importance of the linker connecting the GH9 and the CBM3c

284 modules, we tested the ability of CBM3c*NL* to recover the CMCase activity of GH9. A fixed

285 amount of the catalytic module (70 pmol in 400 µl) was mixed with increasing amounts of

286 CBM3c*L* or CBM3c*NL*. The activity of the intact Cel9I enzyme was defined as 100%, and the

287 activity of the reconstituted complexes was measured relative to that of Cel9I. The results

288 indicated that GH9-CBM3c*NL* exhibit only about 10% of the intact Cel9I activity towards CMC,

289 whereas the reassembled GH9-CBM3c*L* provided up to 50% of the activity (Figure 2).   The fact

290 that a higher than one molar ratio was required to obtain maximum activity can be explained by

291 the fact that the CBM protein was only partly functional as was also observed in the ITC

292 experiments described below. Overall the results suggest that the linker is required for better

293 fitting of the reconstituted CBM3c which results in better recovered activity.

294 **Overall structure of the reassembled GH9-CBM3c**

295   Initially, we tried to overexpress and purify the covalently linked GH9-CBM3c, however the

296 full-length protein was unstable and proteolysis occurred during the overexpression and

297 purification stages, partially resulting in separate GH9 and CBM3c modules. Therefore the

298 obtained protein samples were not homogenous and were not suitable for crystallization trials.

299 Instead, we employed an alternative approach where we expressed the two domains separately

300 and combined them *in vitro*. Surprisingly, the combined modules crystallized and formed a

301 structure similar to those of the known GH9 cellulases.

302   The crystal structure of the reassembled *C. thermocellum* Cel9I GH9-CBM3c*L* dyad was

303 determined by molecular replacement and the coordinates are deposited in Protein Data Bank

304 with code 2XFG. Data collection and refinement statistics are given in Tables 1 and 2. The

305 catalytic GH9 and the ancillary CBM3c modules reassembled in vitro to form a dyad (Figure 3a)

306 similar in structure to the intact tandem GH9-CBM3c modules of the orthologous

307   endoglucanases: Cel9G from *C. cellulolyticum* (1G87) and Cel9A (previously termed cellulase

308   E4) from *Thermobifida fusca* (1TF4), with an RMS deviation of 0.783 Å over 468 Cα atoms

309   with Cel9G and 0.757 Å with Cel9A (Figure 3b).

Structure of the GH9 module

311   The catalytic module of the Cel9I enzyme consists of residues 1-446, comprising 15 α-

312   helices, whereby the twelve longest ones form the $(\alpha/\alpha)_6$-barrel (Figure 4A). The hydrophobic

313   core of the GH9 module is formed by 118 hydrophobic and aromatic amino acids, the vast

314   majority of which are also conserved in the GH9 modules from *C. cellulolyticum* Cel9G and *T.*

315   *fusca* Cel9A. Hydrophobic and aromatic cores have been proposed to play an important role in

316   the formation of $(\alpha/\alpha)_6$-barrels (Mandelman et al. 2003). The GH9 module of Cel9I thus shows

317   high structural similarity with the two latter GH9 structures: *C. cellulolyticum* Cel9G (0.367 Å

318   RMS deviation over 349 C-alpha atoms) and *T. fusca* Cel9A (0.532 Å RMS deviation over 359

319   C-alpha atoms).

320   The catalytic site of the GH9 module is located at the depression in the flat surface, formed by

321   the loops connecting the N termini of the barrel helices (Figure 4B). The flat face is rich in

322   charged and polar residues (Figure 4B), highly conserved also in Cel9G (1G87) and Cel9A

323   (1TF4). The GH9 modules of these cellulases (Mandelman et al. 2003; Sakon et al. 1997; Zhou

324   et al. 2004) exhibit similar flat faces and clefts, and these conserved residues (His 126, Trp 129,

325   Phe 205, Tyr 206, Trp 209, Trp 256, Asp 261, Asp 262, Trp 314, Arg 318, His 376, Arg 378, and

326   Tyr 419) have been shown to bind natural and synthetic oligosaccharides (Figure 4C). In the

327   present structure, as in the other known GH9-CBM3c bimodular structures, one end of this cleft

328   is blocked by a loop formed by residues 243-254 and the other end is fused with the flat surface

329   of the CBM3c module (Figure 4B). Details of the catalytic cleft are presented in Figure 4C.

330   One calcium ion is found near the catalytic cleft of the GH9 module of Cel9I and is

331   coordinated by a Ser 210 (OG) 2.6 Å, Gly 211 (O) 2.4 Å, Asp 261 (O) 2.4 Å, Asp 214 bifurcated

332   (OD1, OD2) 2.5 Å, and Glu 215 bifurcated (OD1, OD2) 2.5 Å (Figure 4D). Despite some minor

333   changes in the residues of coordination this ion seems to be structurally equivalent to those of *T.*

334   *fusca* Cel9A (RMS deviation 0.160 Å over 5 Cα atoms of the coordinating residues), and *C.*

335   *cellulolyticum* Cel9G (RMS deviation 0.503 Å over 4 Cα atoms). In all three cases the calcium

336   ion draws together the N-terminal ends of α-helixes 8 and 10.

Structure of the CBM3c module

338   The CBM3c module consists of 150 amino acids arranged in an eight β-stranded sandwich

339   motif homologous to other known family 3 CBM structures (Gilbert et al. 2013; Mandelman et

340   al. 2003; Petkun et al. 2010b; Sakon et al. 1997; Shimon et al. 2000b; Tormo et al. 1996; Yaniv

341   et al. 2014; Yaniv et al. 2012b; Yaniv et al. 2011). The "lower" face of the sandwich is formed

342   by β-strands 1, 2, and 7; the "upper" face is formed by β-strands 3, 3', 6, 8, and 9 (Figure 5A).

343   The structure of Cel9I CBM3c is particularly similar to the structures of the other two previously

344   described CBM3c structures (RMS deviation 0.734 Å over 116 C-alpha atoms with CBM3c from

345   *C. cellulolyticum* Cel9G; RMS deviation 0.829 Å over 113 atoms with CBM3c from *T. fusca*

346   Cel9A). Only 31% of amino acids are located in β-strands of the CBM3c module from Cel9I;

347   others are found in the loop regions.

348   One calcium ion was found in the upper β-sheet of the CBM3c molecule (Figure 5B) and is

349   coordinated by a water molecule and five residues from the upper β-sheet: Asn 500 (O), Glu 503

350   bifurcated (OE1, OE2), Asn 573 (O), Asn 576 (OD1), Asp 577 (OD1). This calcium atom is in a

351   similar location as in Cel9A and Cel9G, and probably plays a structural role for most CBM3

352   modules, as was suggested previously (Tormo et al. 1996).

353   The lower sheet forms a flat platform conserved between the CBM3c modules and the other

354   two molecular structures. This flat surface is rich in charged and polar conserved surface

355   residues: Asn 466, Glu 474, Lys 476, Ser 518, Tyr 520, Glu 559, Gln 561, and Arg 563 (Figure

356   5C). The planar region of the CBM3c modules in all three enzymes is particularly aligned in

357   continuation of the catalytic cleft of the catalytic modules, and has been proposed to bind single

358   chains of cellulose and guide them to the cleft (Mandelman et al. 2003; Sakon et al. 1997).

359   The CBM3c possesses a very interesting surface structure, formed by the β-strands on the

360   opposite side of the flat surface, called the "shallow groove" (Shimon et al. 2000b; Tormo et al.

361   1996). The "shallow groove" is lined by four aromatic rings (Phe 498, Tyr 538, Tyr 578 and Tyr

362   597), two charged or polar residues (Arg 496, and Glu 540), Leu 602, Pro 595 and Pro 608.

363   These residues are also conserved in other CBM3 modules regardless of their subgroup relation

364   (a, b, or c), their cellulose-binding ability and their effect on the activity of the catalytic module.

365   Figure 5D shows the shallow groove of the CBM3c module from the Cel9I enzyme colored

366   according to the extent of the conservation of the residues in other CBM3a, b and c modules

367   (darker blue represents more conservation). The alignment was performed over 25 CBM3

368   sequences (11 CBM3c, 12 CBM3b and CBM3b', and 4 CBM3a). Conservation of this surface

369   structure, regardless of the particular known function of the CBMs, implies that this site has

370   some kind of "generic" function. This shallow groove may serve to bind to single

371   oligosaccharide chains or to peptide chains, such as the intermodular linkers common to

372   cellulases or cellulosomal scaffoldin subunits. There is evidence that the shallow groove interacts

373   with a linker region (Petkun et al. 2010a; Shimon et al. 2000a; Yaniv et al. 2012a).

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

375   The *in vitro* reassembled GH9-CBM3c*L* complex has a large intermodular interface, the

376   contact area of which is 1108.3 Å^2, corresponding to 12.3% of the total surface-exposed area of

377   the CBM3c module and 6.2% of the exposed GH9 module (Krissinel & Henrick 2007). The GH9

378   and the CBM3c*L* modules of Cel9I are assembled into the reconstituted GH9-CBM3c complex

379   by 31 hydrogen bonds (4 main chain-main chain, 19 main chain-side chain, and 8 side chain-side

380   chain), 14 hydrophobic, 3 aromatic interactions, and 3 ionic bonds

381   (http://pic.mbu.iisc.ernet.in/index.html) (Tina et al. 2007). Sixteen residues from the GH9

382   module and seventeen residues from the CBM3c participate in these interactions (contact

383   residues are shown in Figure 6A). The vast majority of the contact residues and contacts are

384   similar to those of *C. cellulolyticum* Cel9G and of *T. fusca* Cel9A (Figure 6B).

385   Conserved residues of the linker (which spans from Gly447 to Asp462) make numerous

386   contacts with conserved residues of the GH9 module, emphasizing the importance of the linker

387   in this interaction (Figure 6). A conserved Gly447 of the linker interacts via hydrogen bonds

388   with Gly444 and Tyr440 of the GH9 module. Pro449 forms hydrophobic interactions with

389   Tyr440, and Asp450 forms hydrogen bonds with Gln11. Another linker residue, Phe453, forms

390   hydrophobic and aromatic interactions with two aromatic residues of the GH9 module, i.e.,

391   Phe15 and Tyr399. Gly455 forms hydrogen bonds with Asn33 and Arg31. Glu457 forms

392   intricate interactions with a variety of residues of the GH9 module, which include hydrogen

393   bonding with Asn33, Val397, Arg395, as well as hydrogen and ionic interactions with His396.

394   Glu461 exhibits hydrogen and ionic interactions with Arg395. Additionally, linker residues

395   Glu457 and Glu461 form hydrogen bonds and salt bridges with Trp490 and Lys487,

396   respectively. The latter belong to the CBM3c module and are part of a loop (486-498), which

397 protrudes towards and forms interactions with several amino acids of the GH9 module. There are
398 also many hydrogen bonds formed between the neighboring amino acids of the linker, thus
399 contributing to its defined conformation. Altogether the multiple, well-conserved interactions
400 between the linker, the GH9 and the CBM3c modules stabilize the spatial orientation of the
401 modules towards one another and contribute to the structural rigidity of the entire molecule,
402 resulting in an active enzyme structure.

403   As mentioned above, the mutual spatial orientation of the GH9 and CBM3c modules is very
404 similar to that in the native, intact bimodular pairs from Cel9G and Cel9A leading to the overall
405 similarity in structures. The remarkable conservation of the overall architecture in the
406 reassembled *in vitro* complex together with the striking conservation of the contact residues
407 implies its high functional importance. In all of these structures (Cel9G, Cel9A, and the
408 reassembled GH9-CBM3c*L* from Cel9I), the flat surface of the CBM3c module is aligned in
409 continuation with the catalytic cleft of the GH9 module, making an extended platform (Figure
410 4B). This platform is rich in charged and polar surface residues that are highly conserved
411 throughout the family 3 CBMc's.

412 **Microcalorimetric analysis of the GH9-CBM3c complex formation**

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

427 In the case of CBM3c*NL*, the binding process may occur faster in the absence of linker, but may

428 also lead to unspecific binding and aggregation of the modules.

429

430

431 **Discussion**

432   A striking feature of the family 9 glycoside hydrolases is their subdivision into

433 architectural themes, which are defined by their conserved modular composition (Bayer et al.

434 2006). In this context, the Theme B1 endoglucanases contain a GH9 catalytic module followed

435 by a purportedly fused family 3c CBM. Biochemical studies of some of the members of this

436 group (Arai et al. 2001; Chiriac et al. 2010; Gal et al. 1997; Irwin et al. 1998; Li et al. 2007) have

437 shown that the CBM3c acts as a modulator of the function of the catalytic module. However, the

438 exact manner in which the CBM3c functions is still unclear. It has been shown (Gal et al. 1997;

439 Gilad et al. 2003; Irwin et al. 1998) that family 3c CBMs (including the CBM3c from *C.*

440 *thermocellum* Cel9I) fail to bind insoluble cellulosic substrates, implying that they do not act as

441 targeting agents for such substrates. The targeting of the enzyme to crystalline cellulose is

442 achieved either through an additional CBM (Kostylev et al. 2012) or by attachment of the

443 enzyme to a CBM-containing scaffoldin via a cohesin-dockerin interaction (Mingardon et al.

444 2011).

445

446   The CBM3c module of Cel9A from the *T. fusca* has been proposed to loosely anchor the

447 enzyme to cellulose, to disrupt the hydrogen bonds in crystalline cellulose and to guide a single

448 cellulose strand towards the active site of the GH9 catalytic module (Bayer et al. 2006; Li et al.

449 2007). This hypothesis has been supported by molecular docking and molecular dynamics

450 simulation studies (Oliveira et al. 2009). Moreover, double point mutations indicated that high

451 coordination between the substrate affinities of the catalytic module and CBM needs to be

452 precisely controlled (Li et al. 2010). Enzyme thermostability was reported to be affected by the

453 presence of the CBM3c probably due to the formation of a compact structure (Chiriac et al.

454 2010; Su et al. 2012; Yi et al. 2013).

455

456   The previously reported structures of Cel9A from *T. fusca* (Sakon et al. 1997) and Cel9G

457   from *C. cellulolyticum* (Mandelman et al. 2003) revealed that the catalytic module and the

458   CBM3c are separated by a ~20-residue linker that forms multiple polar and hydrophobic

459   interactions mainly with the GH9 module. In an earlier report, we demonstrated that separately

460   expressed GH9 and CBM3c*L* from Cel9I of *C. thermocellum* interact with one another to form

461   an enzymatically active complex (Burstein et al. 2009). In the current article, we showed further

462   that the GH9 and CBM3c can also be reassembled without the linker, albeit at the expense of

463   catalytic activity, thus emphasizing the importance of the linker in positioning correctly the CBM

464   relative to the GH9 catalytic module.

465

466   There is evidence that linkers in multi-modular proteins may serve communication roles

467   between the modules via allosteric mechanisms and variation in their sequences affect enzyme

468   activity (Ma et al. 2011). Linker length and rigidity was shown to play a critical role in the

469   cooperative action of the catalytic module of a cellulase and a CBM (Ting et al. 2009).

470   Computational studies of *T. fusca* Cel9A suggested that thermal contributions to enzyme

471   plasticity and molecular motion at high temperatures may play a role in enhancing CBM and

472   catalytic domain synergy, and the linker may have an important role in this process (Batista et al.

473   2011). The length of the linkers may also play an important role in protein function and

474   adaptation to the environment (Sonan et al. 2007). Studies in cellulolytic fungi revealed that

475   linkers undergo modifications such as glycoslation and have also been shown to directly bind to

476   the cellulose substrate (Beckham et al. 2012; Payne et al. 2013; Sammond et al. 2012; Srisodsuk

477   et al. 1993). Point mutations in different fungal GH-CBM linkers have also been shown to

478   significantly affect the activity of the enzymes and their stability (Couturier et al. 2013; Lu et al.

479   2014).

480

481   The characteristics of the reassembled linker-containing complex are corroborated by the

482   X-ray crystallographic data. Indeed, it is quite surprising that the two separately expressed

483   entities recombined in such a way that the complex could in fact be crystallized. Moreover, the

484   resultant structure was remarkably similar to the known structures of the intact bimodular GH9-

485   CBM3c pairs from *C. cellulolyticum* Cel9G and *T. fusca* Cel9A. Accordingly, the vast majority

486   of the contact residues are similar among the three structures. Conserved residues of the linker

487   make contacts with conserved residues of the GH9 module, highlighting the importance of the

488   linker in this interaction. Multiple hydrogen, ionic and hydrophobic bonds between the linker

489   and the functional GH9 and CBM3c stabilize the spatial organization of the modules. The

490   similarity of the reassembled and native intact structures is particularly intriguing, as it suggests

491   that folding of the modular structures and emplacement of the linker during biosynthesis and

492   intermodular recognition during complex formation are governed by the same interactions,

493   which may have distinct functional consequences. In contrast to the GH9-CBM3c*L,* the re-

494   associated GH9-CBM3c*NL* complex never crystallized, suggesting that the reassembly of the

495   two modules in the absence of linker was somewhat heterogeneous in character.

496

497   Single proteins commonly fold into defined structures, wherein their N- and C-terminal

498   ends are in relatively close proximity to one another. If we view the structures of the Theme B1

499   enzymes, it is evident that their individual modules, the GH9 catalytic module and the CBM3c,

500   are consistent with this rule. The positions of the N- and C-termini of the Theme B1 catalytic

501   module are similar to those of the other GH9 thematic members, including those of Theme A,

502   which lack additional modules. Likewise, the N- and C-termini of CBM3c are essentially the

503   same as all other members of the family 3 CBMs, regardless of their source (i.e., parent

504   cellulase, scaffoldin, etc). The evolutionary significance of this observation is that, originally, the

505   functional relationship between the two modules was likely a more conventional one, whereby

506   an ancestral CBM3 played a standard targeting role to deliver the GH9 catalytic module to its

507   substrate. During the course of evolution, this relationship changed, and the precise positioning

508   and fusion of a mutated CBM3 with a GH9 catalytic module served to modulate the activity

509   characteristics of the latter. For this purpose, the flat surface of the CBM3c is aligned with the

510   flat surface of the catalytic module, and the appropriate residues that interact with the single

511   cellulose chain are thus aligned with the active site of the GH9 module. As a consequence, the

512   two closely juxtaposed modules can be considered as a single functional entity. The functional

513   positioning and fusion of the two modules, however, are at odds with the inherent locations of

514   the termini of the two modules, such that the C-terminus of the catalytic module is very distant

515   from the N-terminus of the CBM3c. Consequently, nature has provided a very distinctive type of

516   conserved linker, which both connects the two modules and helps secure their required

517   orientation for processive endoglucanase activity.

518

Conclusions

520   Cellulase 9I (Cel9I), a non-cellulosomal family 9 processive endoglucanase from *Clostridium*

521 *thermocellum*, which degrades crystalline cellulose phosphoric acid-swollen cellulose (PASC)

522 and carboxymethyl cellulose (CMC), consists of a catalytic GH9 module followed by two family

523 3 carbohydrate-binding modules (CBMs): CBM3c and CBM3b, separated by linker regions. C-

524 terminal CBM3b module, as a classic CBM3, is responsible for targeting the Cel9I enzyme to the

525 planar surface of the crystalline cellulose. The CBM3c is crucial for the GH9 enzymatic activity.

526 In this work we investigated the interaction of separately expressed catalytic module and CBM3c

527 either with or without the intermodular linker in order to better understand the function of the

528 CBM3c in the family-9 enzymes and the role of the linkers regions.

529 GH9 catalytic module and CBM3c were able to interact and reassemble both with and without

530 the linker; however the linker was essential for the endoglucanase catalytic activity.

531 Surprisingly, we were able to crystallize these two separately expressed entities, meaning that

532 their reassembly was very ordered and structurally homogeneous. The molecular structure of the

533 GH9 and CBM3c with the linker region showed that they form a complex similar in structure to

534 the intact tandem GH9-CBM3c modules of the orthologous endoglucanases Cel9G from *C.*

535 *cellulolyticum* and Cel9A from *Thermobifida fusca*. The flat, conserved surface of the CBM3c

536 module is aligned in continuation with the catalytic cleft of the GH9 module, presumably

537 forming one functional entity, which binds to the planar surface of the cellulose. Conserved

538 residues of the linker make contacts with conserved residues of the GH9 module, highlighting

539 the importance of the linker in this interaction. Overall our results demonstrate that the linker

540 regions in the GH9/CBM3c endoglucanases are necessary to achieve the right spatial

541 organization of the modules and for the fixation of the modules into functional enzymes.

542

Acknowledgements

544 This article is dedicated to the memory of Professor Felix Frolow, who passed away on 29

545 August 2014. We thankfully acknowledge the ESRF for synchrotron beam time and staff

546 scientists of the ID-29 beam line for their assistance.

548 **References**

549 Akinosho H, Yee K, Close D, and Ragauskas A. 2014. The emergence of *Clostridium*
550   *thermocellum* as a high utility candidate for consolidated bioprocessing applications.
551   *Front Chem* 2:66.

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

555 Batista PR, de Souza Costa MG, Pascutti PG, Bisch PM, and de Souza W. 2011. High
556   temperatures enhance cooperative motions between CBM and catalytic domains of a
557   thermostable cellulase: mechanism insights from essential dynamics. *Phys Chem Chem*
558   *Phys* 13:13709-13720.

559 Bayer EA, Belaich J-P, Shoham Y, and Lamed R. 2004. The cellulosomes: Multi-enzyme
560   machines for degradation of plant cell wall polysaccharides. *Annu Rev Microbiol* 58:521-
561   554.

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

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

569 Bayer EA, Shoham Y, and Lamed R. 2006. Cellulose-decomposing prokaryotes and their
570   enzyme systems. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, and
571   Stackebrandt E, eds. *The Prokaryotes, Third Edition*. New York: Springer-Verlag, 578-
572   617.

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

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

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

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

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

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

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

592 Burstein T, Shulman M, Jindou S, Petkun S, Frolow F, Shoham Y, Bayer EA, and Lamed R.
593   2009. Physical association of the catalytic and helper modules of a processive family-9
594   glycoside hydrolase is essential for activity. *FEBS Lett* 583:879-884.
595 Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, and Henrissat B. 2009. The
596   Carbohydrate-Active Enzymes database (CAZy): an expert resource for glycogenomics.
597   *Nucl Acids Res* 37:D233-238.
598 Chayen NE, Shaw Stewart PD, Maeder DL, and Blow DM. 1990. An automated system for
599   micro-batch protein crystallization and screening. *J Appl Crystallogr* 23:297-302.
600 Chiriac AI, Cadena EM, Vidal T, Torres AL, Diaz P, and Pastor FI. 2010. Engineering a family 9
601   processive endoglucanase from Paenibacillus barcinonensis displaying a novel
602   architecture. *Appl Microbiol Biotechnol* 86:1125-1134.
603 Cosier J, and Glazer AM. 1986. A nitrogen-gas-stream cryostat for general X-ray-diffraction
604   studies. *J Appl Crystalogr* 19:105-107.
605 Couturier M, Feliu J, Bozonnet S, Roussel A, and Berrin JG. 2013. Molecular engineering of
606   fungal GH5 and GH26 beta-(1,4)-mannanases toward improvement of enzyme activity.
607   *PloS ONE* 8:e79800.
608 Cuskin F, Flint JE, Gloster TM, Morland C, Basle A, Henrissat B, Coutinho PM, Strazzulli A,
609   Solovyova AS, Davies GJ, and Gilbert HJ. 2012. How nature can exploit nonspecific
610   catalytic and carbohydrate binding modules to create enzymatic specificity. *Proc Natl*
611   *Acad Sci U S A* 109:20889-20894.
612 Davies G, and Henrissat B. 1995. Structures and mechanisms of glycosyl hydrolases. *Structure*
613   3:853-859.
614 Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB,
615   3rd, Snoeyink J, Richardson JS, and Richardson DC. 2007. MolProbity: all-atom contacts
616   and structure validation for proteins and nucleic acids *Nucl Acids Res* 35(Web Server
617   issue):W375-383.
618 Demain AL, Newcomb M, and Wu JH. 2005. Cellulase, clostridia, and ethanol. *Microbiol Mol*
619   *Biol Rev* 69:124-154.
620 Deprez C, Lloubes R, Gavioli M, Marion D, Guerlesquin F, and Blanchard L. 2005. Solution
621   structure of the *E. coli* TolA C-terminal domain reveals conformational changes upon
622   binding to the phage g3p N-terminal domain *J Mol Biol* 346:1047-1057.
623 Doi RH, and Kosugi A. 2004. Cellulosomes: plant-cell-wall-degrading enzyme complexes. *Nat*
624   *Rev Microbiol* 2:541-551.
625 Emsley P, and Cowtan K. 2004. Coot: model-building tools for molecular graphics. *Acta*
626   *Crystallogr D* 60:2126-2132.
627 Felix CR, and Ljungdahl LG. 1993. The cellulosome - the exocellular organelle of Clostridium.
628   *Annu Rev Microbiol* 47:791-819.
629 Fontes CM, and Gilbert HJ. 2010. Cellulosomes: Highly efficient nanomachines designed to
630   deconstruct plant cell wall complex carbohydrates. *Annu Rev Biochem* 79:655-681.
631 Freier D, Mothershed CP, and Wiegel J. 1988. Characterization of *Clostridium thermocellum*
632   JW20. *Appl Environ Microbiol* 54:204-211.
633 Gal L, Gaudin C, Belaich A, Pagès S, Tardif C, and Belaich J-P. 1997. CelG from *Clostridium*
634   *cellulolyticum*: a multidomain endoglucanase acting efficiently on crystalline cellulose. *J*
635   *Bacteriol* 179:6595-6601.
636 Garcia-Martinez DV, Shinmyo A, Madia A, and Demain AL. 1980. Studies on cellulase
637   production by *Clostridium thermocellum*. *Eur J Appl Microbiol Biotechnol* 9:189-197.

638 Gilad R, Rabinovich L, Yaron S, Bayer EA, Lamed R, Gilbert HJ, and Shoham Y. 2003. CelI, a
639   non-cellulosomal family-9 enzyme from *Clostridium thermocellum*, is a processive
640   endoglucanase that degrades crystalline cellulose. *J Bacteriol* 185:391-398.
641 Gilbert HJ, and Hazlewood GP. 1993. Bacterial cellulases and xylanases. *J Gen Microbiol*
642   139:187-194.
643 Gilbert HJ, Knox JP, and Boraston AB. 2013. Advances in understanding the molecular basis of
644   plant cell wall polysaccharide recognition by carbohydrate-binding modules. *Curr Opin*
645   *Struct Biol* 23:669-677.
646 Gilkes NR, Henrissat B, Kilburn DG, Miller RCJ, and Warren RAJ. 1991. Domains in microbial
647   b-1,4-glycanases: sequence conservation, function, and enzyme families. *Microbiol Rev*
648   55:303-315.
649 Gruninger RJ, Gong X, Forster RJ, and McAllister TA. 2014. Biochemical and kinetic
650   characterization of the multifunctional beta-glucosidase/beta-xylosidase/alpha-
651   arabinosidase, Bgxa1. *Appl Microbiol Biot* 98:3003-3012.
652 Henrissat B. 1991. A classification of glycosyl hydrolases based on amino acid sequence
653   similarities. *Biochem J* 280:309-316.
654 Henrissat B, and Davies G. 1997. Structural and sequence-based classification of glycoside
655   hydrolases. *Curr Opin Struct Biol* 7:637-644.
656 Henrissat B, and Davies GJ. 2000. Glycoside hydrolases and glycosyltransferases. Families,
657   modules, and implications for genomics. *Plant Physiol* 124:1515-1519.
658 Henrissat B, and Romeu A. 1995. Families, superfamilies and subfamilies of glycosyl
659   hydrolases. *Biochem J* 311:350-351.
660 Irwin D, Shin D-H, Zhang S, Barr BK, Sakon J, Karplus PA, and Wilson DB. 1998. Roles of the
661   catalytic domain and two cellulose binding domains of *Thermomonospora fusca* E4 in
662   cellulose hydrolysis. *J Bacteriol* 180:1709-1714.
663 Jeon SD, Yu KO, Kim SW, and Han SO. 2012. The processive endoglucanase EngZ is active in
664   crystalline cellulose degradation as a cellulosomal subunit of Clostridium cellulovorans.
665   *N Biotechnol* 29:365-371.
666 Kostylev M, Moran-Mirabal JM, Walker LP, and Wilson DB. 2012. Determination of the
667   molecular states of the processive endocellulase Thermobifida fusca Cel9A during
668   crystalline cellulose depolymerization. *Biotechnol Bioeng* 109:295-299.
669 Krissinel E, and Henrick K. 2007. Inference of macromolecular assemblies from crystalline state.
670   *J Mol Biol* 372:774--797.
671 Kuusk S, Sorlie M, and Valjamae P. 2015. The predominant molecular state of bound enzyme
672   determines the strength and type of product inhibition in the hydrolysis of recalcitrant
673   polysaccharides by processive enzymes. *J Biol Chem.* 290: 11678-11691
674 Lamed R, Setter E, and Bayer EA. 1983a. Characterization of a cellulose-binding, cellulase-
675   containing complex in *Clostridium thermocellum*. *J Bacteriol* 156:828-836.
676 Lamed R, Setter E, Kenig R, and Bayer EA. 1983b. The cellulosome — a discrete cell surface
677   organelle of *Clostridium thermocellum* which exhibits separate antigenic, cellulose-
678   binding and various cellulolytic activities. *Biotechnol Bioeng Symp* 13:163-181.
679 Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F,
680   Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, and Higgins DG. 2007.
681   ClustalW and ClustalX version 2. *Bioinformatics* 23:2947-2948.

682   Li Y, Irwin DC, and Wilson DB. 2007. Processivity, substrate binding, and mechanism of
683   cellulose hydrolysis by *Thermobifida fusca* Cel9A. *Appl Environ Microbiol* 73:3165-
684   3172.
685   Li Y, Irwin DC, and Wilson DB. 2010. Increased crystalline cellulose activity via combinations
686   of amino acid changes in the family 9 catalytic domain and family 3c cellulose binding
687   module of *Thermobifida fusca* Cel9A. *Appl Environ Microbiol* 76:2582-2588.
688   Lu H, Luo H, Shi P, Huang H, Meng K, Yang P, and Yao B. 2014. A novel thermophilic endo-
689   beta-1,4-mannanase from Aspergillus nidulans XZ3: functional roles of carbohydrate-
690   binding module and Thr/Ser-rich linker region. *Appl Microbiol Biotechnol* 98:2155-2163.
691   Ma B, Tsai CJ, Haliloglu T, and Nussinov R. 2011. Dynamic allostery: linkers are not merely
692   flexible. *Structure (London, England : 1993)* 19:907-917.
693   Mandelman D, Belaich A, Belaich JP, Aghajari N, Driguez H, and Haser R. 2003. X-Ray crystal
694   structure of the multidomain endoglucanase Cel9G from *Clostridium cellulolyticum*
695   complexed with natural and synthetic cello-oligosaccharides. *J Bacteriol* 185:4127-4135.
696   Matthews BW. 1968. Solvent content of protein crystals. *J Mol Biol* 33:491-497.
697   Meekins DA, Raththagala M, Husodo S, White CJ, Guo HF, Kotting O, Vander Kooi CW, and
698   Gentry MS. 2014. Phosphoglucan-bound structure of starch phosphatase Starch Excess4
699   reveals the mechanism for C6 specificity. *Proc Natl Acad Sci U S A* 111:7272-7277.
700   Miller GL. 1959. Use of dinitrosalicylic acid reagent for determination of reducing sugar. *Anal*
701   *Biochem* 31:426-428.
702   Mingardon F, Bagert JD, Maisonnier C, Trudeau DL, and Arnold FH. 2011. Comparison of
703   family 9 cellulases from mesophilic and thermophilic bacteria. *Appl Environ Microbiol*
704   77:1436-1442.
705   Moraïs S, Alber O, Barak Y, Hadar Y, Wilson DB, Lamed R, Shoham Y, and Bayer EA. 2012.
706   Functional association of the catalytic and ancillary modules dictates enzymatic activity
707   in glycoside hydrolase family 43 β-xylosidase. *J Biol Chem* 287:9213-9221.
708   Murshudov GN, Vagin AA, and Dodson EJ. 1997. Refinement of macromolecular structures by
709   the maximum-likelihood method. *Acta Crystallogr D* 53:240-255.
710   Ng TK, Weimer TK, and Zeikus JG. 1977. Cellulolytic and physiological properties of
711   *Clostridium thermocellum*. *Arch Microbiol* 114:1-7.
712   Noach I, Alber O, Bayer EA, Lamed R, Levy-Assaraf M, Shimon LJW, and Frolow F. 2008.
713   Crystallization and preliminary X-ray analysis of *Acetivibrio cellulolyticus* cellulosomal
714   type II cohesin module: Two versions having different linker lengths. *Acta Crystallogr F*
715   64:58-61.
716   Oliveira OV, Freitas LCG, Straatsma TP, and Lins RD. 2009. Interaction between the CBM of
717   Cel9A from *Thermobifida fusca* and cellulose fibers. *J Mol Recognit* 22:38-45.
718   Otwinowski Z, and Minor W. 1997. Processing of X-ray diffraction data collected in oscillation
719   mode. *Meth Enzymol*, 307-326.
720   Payne CM, Bomble YJ, Taylor CB, McCabe C, Himmel ME, Crowley MF, and Beckham GT.
721   2011. Multiple functions of aromatic-carbohydrate interactions in a processive cellulase
722   examined with molecular simulation. *J Biol Chem* 286:41028-41035.
723   Payne CM, Resch MG, Chen L, Crowley MF, Himmel ME, Taylor LE, 2nd, Sandgren M,
724   Stahlberg J, Stals I, Tan Z, and Beckham GT. 2013. Glycosylated linkers in multimodular
725   lignocellulose-degrading enzymes dynamically bind to cellulose. *Proc Natl Acad Sci U S*
726   *A* 110:14646-14651.

727 Perrakis A, Morris R, and Lamzin VS. 1999. Automated protein model building combined with
728   iterative structure refinement. *Nat Struct Biol* 6:458-463.
729 Petkun S, Jindou S, Shimon LJW, Rosenheck S, Bayer EA, Lamed R, and Frolow F. 2010a.
730   Structure of a family 3b ' carbohydrate-binding module from the Cel9V glycoside
731   hydrolase from Clostridium thermocellum: structural diversity and implications for
732   carbohydrate binding. *Acta Crystallogr D* 66:33-43.
733 Petkun S, Jindou S, Shimon LJW, Rosenheck S, Bayer EA, Lamed R, and Frolow F. 2010b.
734   Structure of a family 3b' carbohydrate-binding module from the Cel9V glycoside
735   hydrolase from *Clostridium thermocellum*. Structural diversity and implications for
736   carbohydrate binding. *Acta Cryst* D66:33-43.
737 Sakon J, Irwin D, Wilson DB, and Karplus PA. 1997. Structure and mechanism of
738   endo/exocellulase E4 from *Thermomonospora fusca*. *Nat Struct Biol* 4:810-818.
739 Sammond DW, Payne CM, Brunecky R, Himmel ME, Crowley MF, and Beckham GT. 2012.
740   Cellulase linkers are optimized based on domain type and function: insights from
741   sequence analysis, biophysical measurements, and molecular simulation. *PloS one*
742   7:e48615.
743 Schwarz WH. 2001. The cellulosome and cellulose degradation by anaerobic bacteria. *Appl*
744   *Microbiol Biotechnol* 56:634-649.
745 Schwarz WH, Zverlov VV, and Bahl H. 2004. Extracellular glycosyl hydrolases from Clostridia.
746   *Advan Appl Microbiol* 56:215-261.
747 Shimon LJ, Pages S, Belaich A, Belaich JP, Bayer EA, Lamed R, Shoham Y, and Frolow F.
748   2000a. Structure of a family IIIa scaffoldin CBD from the cellulosome of Clostridium
749   cellulolyticum at 2.2 A resolution. *Acta Crystallogr D* 56:1560-1568.
750 Shimon LJW, Pagès S, Belaich A, Belaich JP, Bayer EA, Lamed R, Shoham Y, and Frolow F.
751   2000b. Structure of a family IIIa scaffoldin CBD from the cellulosome of *Clostridium*
752   *cellulolyticum* at 2.2 Å resolution. *Acta Crystallogr D* 56:1560-1568.
753 Shoham Y, Lamed R, and Bayer EA. 1999. The cellulosome concept as an efficient microbial
754   strategy for the degradation of insoluble polysaccharides. *Trends Microbiol* 7:275-281.
755 Sonan GK, Receveur-Brechot V, Duez C, Aghajari N, Czjzek M, Haser R, and Gerday C. 2007.
756   The linker region plays a key role in the adaptation to cold of the cellulase from an
757   Antarctic bacterium. *Bioch J* 407:293-302.
758 Srisodsuk M, Reinikainen T, Penttila M, and Teeri TT. 1993. Role of the interdomain linker
759   peptide of Trichoderma reesei cellobiohydrolase I in its interaction with crystalline
760   cellulose. *Journal Biol Chem* 268:20756-20761.
761 Stout GH, and Jensen LH. 1968. *X-ray structure determination. A practical guide*. London:
762   MacMillan.
763 Su X, Mackie RI, and Cann IK. 2012. Biochemical and mutational analyses of a multidomain
764   cellulase/mannanase from Caldicellulosiruptor bescii. *Appl Environ Microb* 78:2230-
765   2240.
766 Telke AA, Ghatge SS, Kang SH, Thangapandian S, Lee KW, Shin HD, Um Y, and Kim SW.
767   2012. Construction and characterization of chimeric cellulases with enhanced catalytic
768   activity towards insoluble cellulosic substrates. *Bioresource Technol* 112:10-17.
769 Tina KG, Bhadra R, and Srinivasan N. 2007. PIC: Protein interactions calculator. *Nucl Acids Res*
770   35:W473–W476.
771 Ting CL, Makarov DE, and Wang ZG. 2009. A kinetic model for the enzymatic action of
772   cellulase. *J Phys Chem B* 113:4970-4977.

773 Tormo J, Lamed R, Chirino AJ, Morag E, Bayer EA, Shoham Y, and Steitz TA. 1996. Crystal
774   structure of a bacterial family-III cellulose-binding domain: A general mechanism for
775   attachment to cellulose. *EMBO J* 15:5739-5751.
776 Vagin A, and Teplyakov A. 1997. MOLREP: an automated program for molecular replacement.
777   *J Appl Crystalogr* 30:1022-1025.
778 Venditto I, Najmudin S, Luis AS, Ferreira LM, Sakka K, Knox JP, Gilbert HJ, and Fontes CM.
779   2015. Family 46 Carbohydrate-Binding Modules contribute to the enzymatic hydrolysis
780   of xyloglucan and beta-1,3-1,4-glucans through distinct mechanisms. *J Biol Chem*
781   290:10572-10586
782 Wiegel J, Mothershed CP, and Puls J. 1985. Differences in xylan degradation by various
783   noncellulolytic thermophilic anaerobes and *Clostridium thermocellum*. *Appl Environ*
784   *Microbiol* 49:656-659.
785 Wilson DB, and Irwin DC. 1999. Genetics and properties of cellulases. *Adv Biochem Eng* 65:1-
786   21.
787 Wilson DB, and Kostylev M. 2012. Cellulase processivity. *Methods in molecular biology*
788   *(Clifton, NJ)* 908:93-99.
789 Winn MD, Isupov MN, and Murshudov GN. 2001. Use of TLS parameters to model anisotropic
790   displacements in macromolecular refinement. *Acta Crystallogr D* 57:122-133.
791 Winn MD, Murshudov GN, and Papiz MZ. 2003. Macromolecular TLS refinement in REFMAC
792   at moderate resolution. *Methods Enzymol* 374:300-321.
793 Yaniv O, Fichman G, Borovok I, Shoham Y, Bayer EA, Lamed R, Shimon LJ, and Frolow F.
794   2014. Fine-structural variance of family 3 carbohydrate-binding modules as extracellular
795   biomass-sensing components of Clostridium thermocellum anti-sigmaI factors. *Acta*
796   *Crystallogr D* 70:522-534.
797 Yaniv O, Frolow F, Levy-Assraf M, Lamed R, and Bayer EA. 2012a. Interactions between
798   family 3 carbohydrate binding modules (CBMs) and cellulosomal linker peptides.
799   *Methods Enzymol* 510:247-259.
800 Yaniv O, Petkun S, Shimon LJ, Bayer EA, Lamed R, and Frolow F. 2012b. A single mutation
801   reforms the binding activity of an adhesion-deficient family 3 carbohydrate-binding
802   module. *Acta Crystallogr D* 68:819-828.
803 Yaniv O, Shimon LJ, Bayer EA, Lamed R, and Frolow F. 2011. Scaffoldin-borne family 3b
804   carbohydrate-binding module from the cellulosome of Bacteroides cellulosolvens:
805   structural diversity and significance of calcium for carbohydrate binding. *Acta*
806   *Crystallogr D* 67:506-515.
807 Yi Z, Su X, Revindran V, Mackie RI, and Cann I. 2013. Molecular and biochemical analyses of
808   CbCel9A/Cel48A, a highly secreted multi-modular cellulase by Caldicellulosiruptor
809   bescii during growth on crystalline cellulose. *PloS One* 8:e84172.
810 Zhou W, Irwin DC, Escovar-Kousen J, and Wilson DB. 2004. Kinetic studies of Thermobifida
811   fusca Cel9A active site mutant enzymes. *Biochemistry* 43:9655-9663.
812 Zmudka MW, Thoden JB, and Holden HM. 2013. The structure of DesR from Streptomyces
813   venezuelae, a beta-glucosidase involved in macrolide activation. *Protein Sci* 22:883-892.
814 Zverlov VV, Velikodvorskaya GA, and Schwarz WH. 2003. Two new cellulosome components
815   encoded downstream of *cel*I in the genome of *Clostridium thermocellum*: the non-
816   processive endoglucanase CelN and the possibly structural protein CseP. *Microbiology*
817   149:515-524.
818

819

820 **Figure captions:**

821 **Figure 1.** Schematic diagram of the Cel9I gene product (top) and the recombinant proteins (**A-**

822 **D**) prepared for this study. The GH9 module alone (**B**) was prepared with and without an N-

823 terminal His tag (shown schematically in the figure), and the CBM3c's were prepared with C-

824 terminal His tags. Scale shows the number of amino acid residues and the boundaries of the

825 different regions of the protein.

826 **Figure 2.** Recovery of activity upon association of CBM3c (with and without linker) and GH9.

827 CMCase activity (μmol reducing sugar released in a 10-min reaction) of His-tagged GH9, mixed

828 either with CBM3c*L* (diamonds) or CBM3c*NL* (squares), was examined. A fixed amount (70

829 pmol) of the GH9 catalytic module was mixed with increasing amounts of the indicated helper

830 module, and their respective activities were compared to that of the intact Cel9I core (GH9-

831 CBM3c, set as 100%).

832 **Figure 3.** Reassembled GH9-CBM3c from Cel9I**.** C and N termini are indicated, and the break

833 between the GH9 and CBM3c modules is marked by a red ellipse. **A.** The *in vitro* reassembled

834 complex of the catalytic (GH9, wheat) and carbohydrate-binding (CBM3c, green) modules of

835 Cel9I from *C. thermocellum,* cartoon representation. Calcium atoms are shown as magenta-

836 colored spheres. **B.** Stereo-view (cross-eyed) of the superposition of the reassembled GH9-

837 CBM3c structure of *C. thermocellum* Cel9I (red) with the bimodular structures of *C.*

838 *cellulolyticum* Cel9G (blue) and *T. fusca* Cel9A (green).

839 **Figure 4**. Structural components of the reassembled *C. thermocellum* GH9-CBM3c. **A.**

840 Structure of the GH9 catalytic module*,* cartoon representation. Twelve α-helices form an (α/α)$_6$-

841 barrel fold. Pairs of helices, comprising the fold, are emphasized by red, blue, yellow, magenta,

842 cyan and green. **B.** Surface representation of the reassembled GH9-CBM3c complex. The

843 residues are shaded according to the extent of their conservation with Cel9G from *C.*

844 *cellulolyticum* and Cel9A from *T. fusca*. Darker blue indicates higher conservation. Top, birds-

845 eye view of the catalytic cleft. Bottom, lateral view, showing the flat surface (red bar). Pink

846 ellipse indicates the catalytic cleft, and green ellipse designates terminal portion of the catalytic

847 site. **C.** Close-up (same orientation as in B, top) of the catalytic cleft of the Cel9I GH9 module

848 showing functional residues. Carbohydrate-binding residue carbons are colored gray, catalytic

849 residue carbons are colored yellow. Loop 243-254 carbons are colored in light blue. **D.** Calcium-

850 binding site of the *C. thermocellum* Cel9I GH9 module. Coordinating residues are shown in stick

851 representation. The calcium ion is colored magenta, and distances to the coordinating atoms are

852 indicated.

853 **Figure 5.** Structure of the CBM3c of Cel9I from *C. thermocellum*. C and N termini are

854 indicated **A.** Cartoon representation, β-strands are numbered according to the alignment with

855 Cel9G from *C. cellulolyticum*, and Cel9A from *T. fusca*. **B.** Calcium-binding site of the CBM3c.

856 **C.** Birds-eye view of the flat surface. Residues are shaded according to their degree of

857 conservation with *C. cellulolyticum* Cel9G and *T. fusca* CEL9A. Surface-exposed conserved

858 residues are shown in stick representation. **D.** Shallow groove of the CBM3c. Conserved surface

859 residues are shown in stick representation. The residues are colored according to the degree of

860 conservation in CBM3a, CBM3b and CBM3c modules derived from the sequences listed in the

861 Methods section.

862 **Figure 6.** Contact residues of the reassembled GH9-CBM3c complex. **A.** The GH9 module is

863 colored in brown, CBM3c in green and the linker in blue. Contact residues of the GH9, CBM3c

864 and linker are colored orange, green and blue, respectively. The contact residues between the

865 linker and the domains are described in the text. **B.** Alignment of the GH9 and CBM3c modules

866 of *C. thermocellum* Cel9I, *C. cellulolyticum* Cel9G, and *T. fusca* Cel9A (E4) cellulases. Contact

867 residues are highlighted in yellow. Only the relevant regions of the alignment are shown.

868 Residues of linker sequences are shown blue font.

869 **Figure 7.** Representative ITC titration of (**A**) GH9 and CBM3c*NL* (**B**) GH9 and CBM3c*L*.   The

870 top panel shows the calorimetric titration and the bottom panel displays the integrated injection

871 heats corrected for control dilution heat. The solid line is the curve of the best fit used to derive

872 the binding parameters, and the fitted data describe an interaction of a one binding site model.

873

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.

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	P$2_1$2$_1$2$_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)

3 † $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - <I(hkl)>| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where Σ_{hkl} denotes the sum over all reflections and Σ_i the sum over all
4 equivalent and symmetry-related reflections.

Table 2 **(on next page)**

Table 2. Refinement statistics and results of *MolProbity* validation.

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

Figure 1. Schematic diagram of the Cel9I gene product (top) and the recombinant proteins (A-D) prepared for this study.

The GH9 module alone (**B**) was prepared with and without an N-terminal His tag (shown schematically in the figure), and the CBM3c's were prepared with C-terminal His tags. Scale shows the number of amino acid residues and the boundaries of the different regions of the protein.

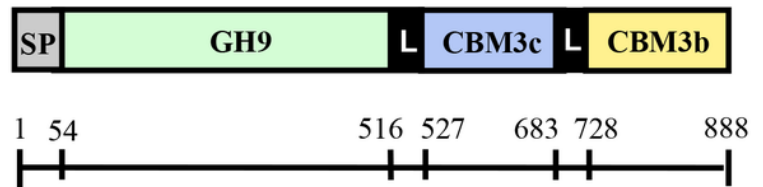
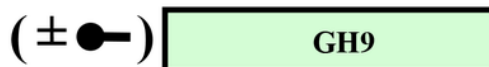
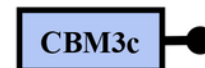



Figure 2. Recovery of activity upon association of CBM3c (with and without linker) and GH9.

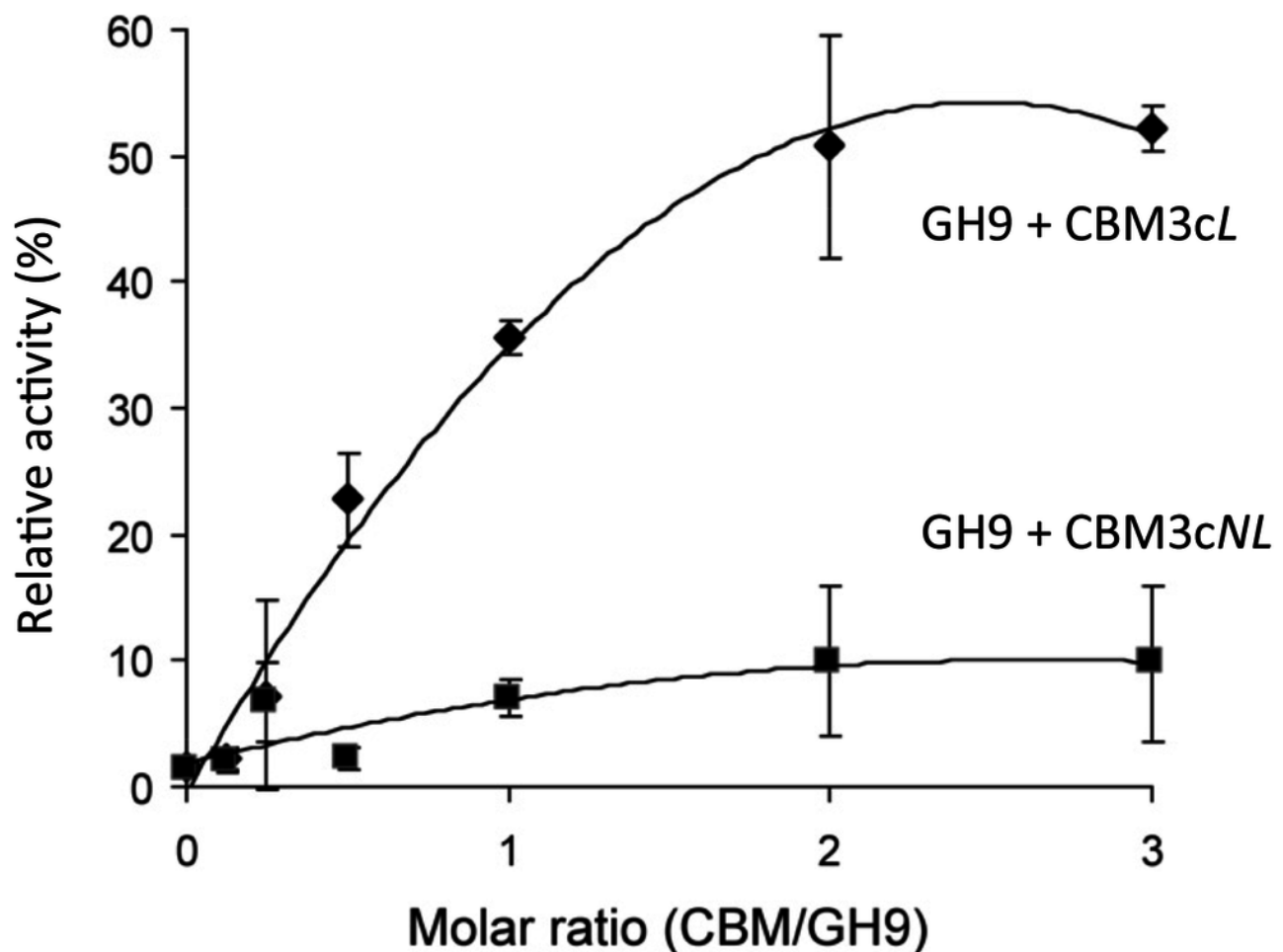
CMCase activity (μmol reducing sugar released in a 10-min reaction) of His-tagged GH9, mixed either with CBM3c*L* (diamonds) or CBM3c*NL* (squares), was examined. A fixed amount (70 pmol) of the GH9 catalytic module was mixed with increasing amounts of the indicated helper module, and their respective activities were compared to that of the intact Cel9I core (GH9-CBM3c, set as 100%).

3

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).

A

N

Linker

C

GH9

CBM3c

B

GH9

N

Linker

C

CBM3c

GH9

N

Linker

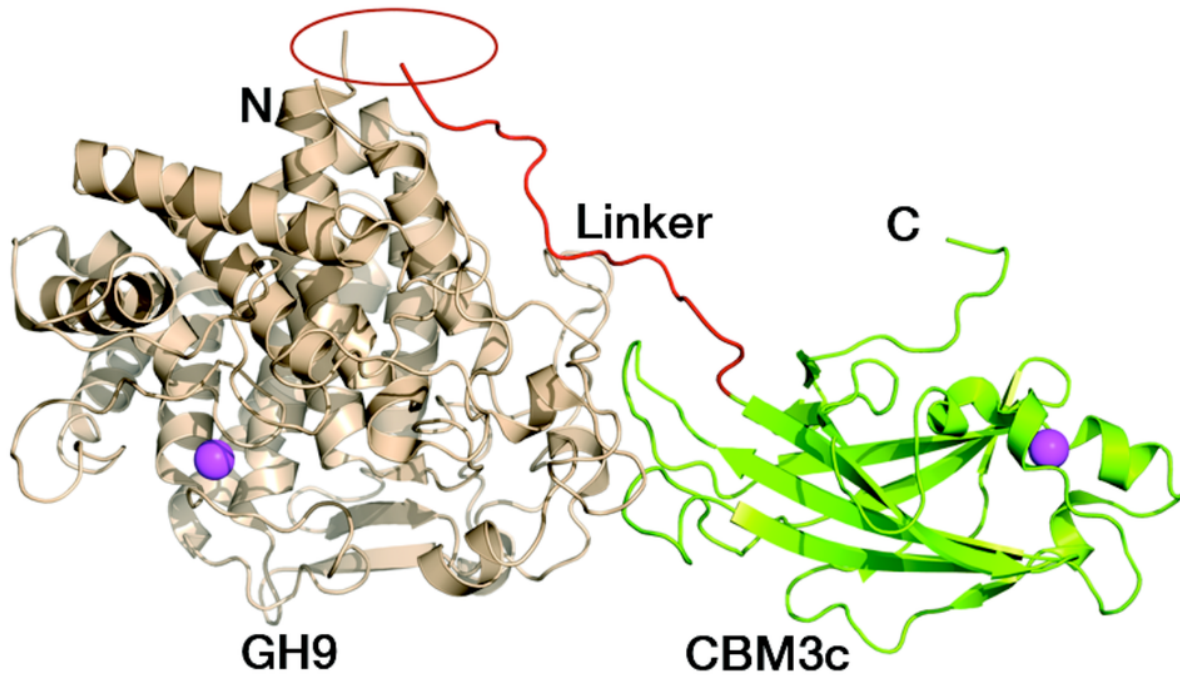
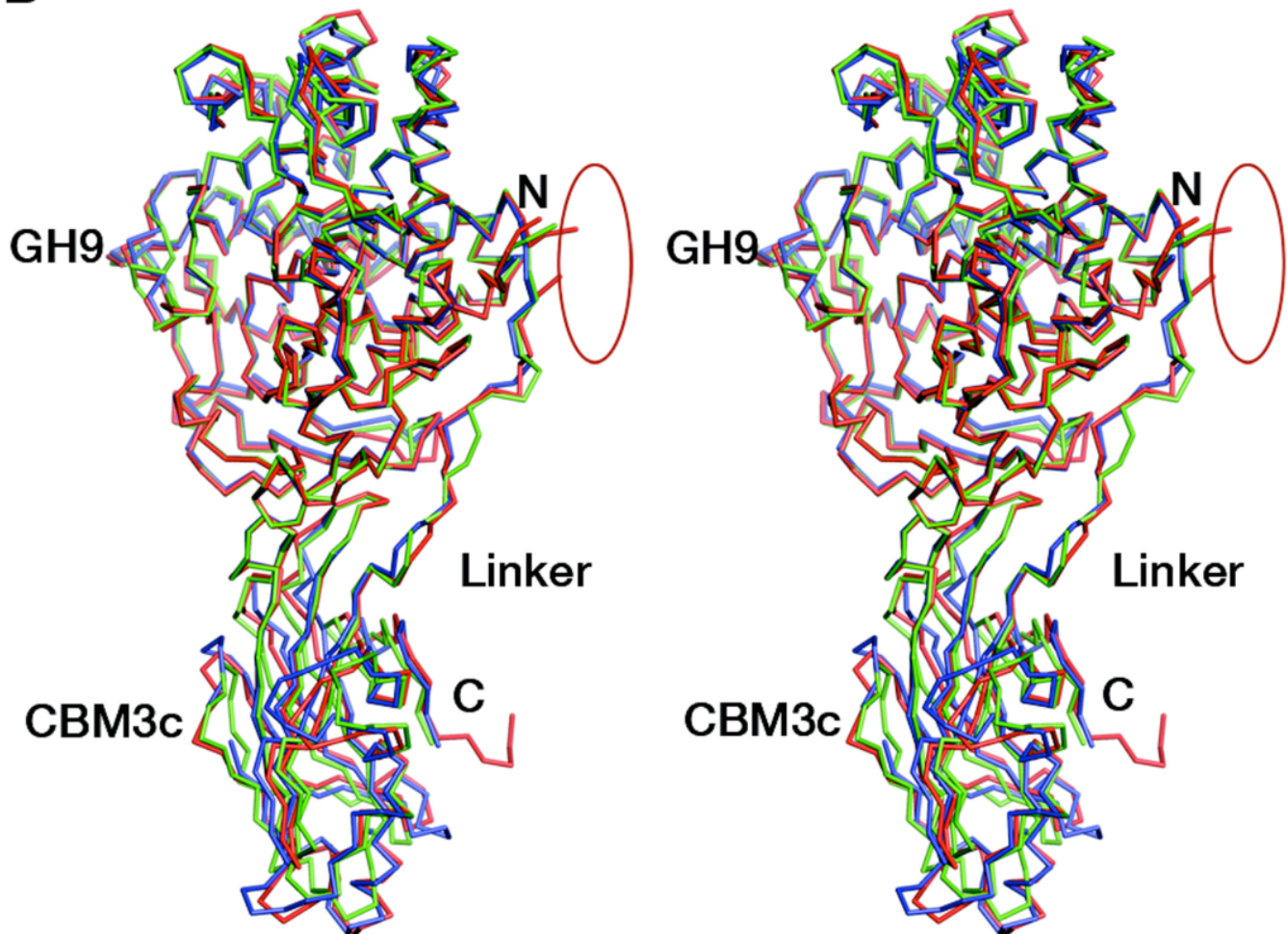
C

CBM3c

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

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

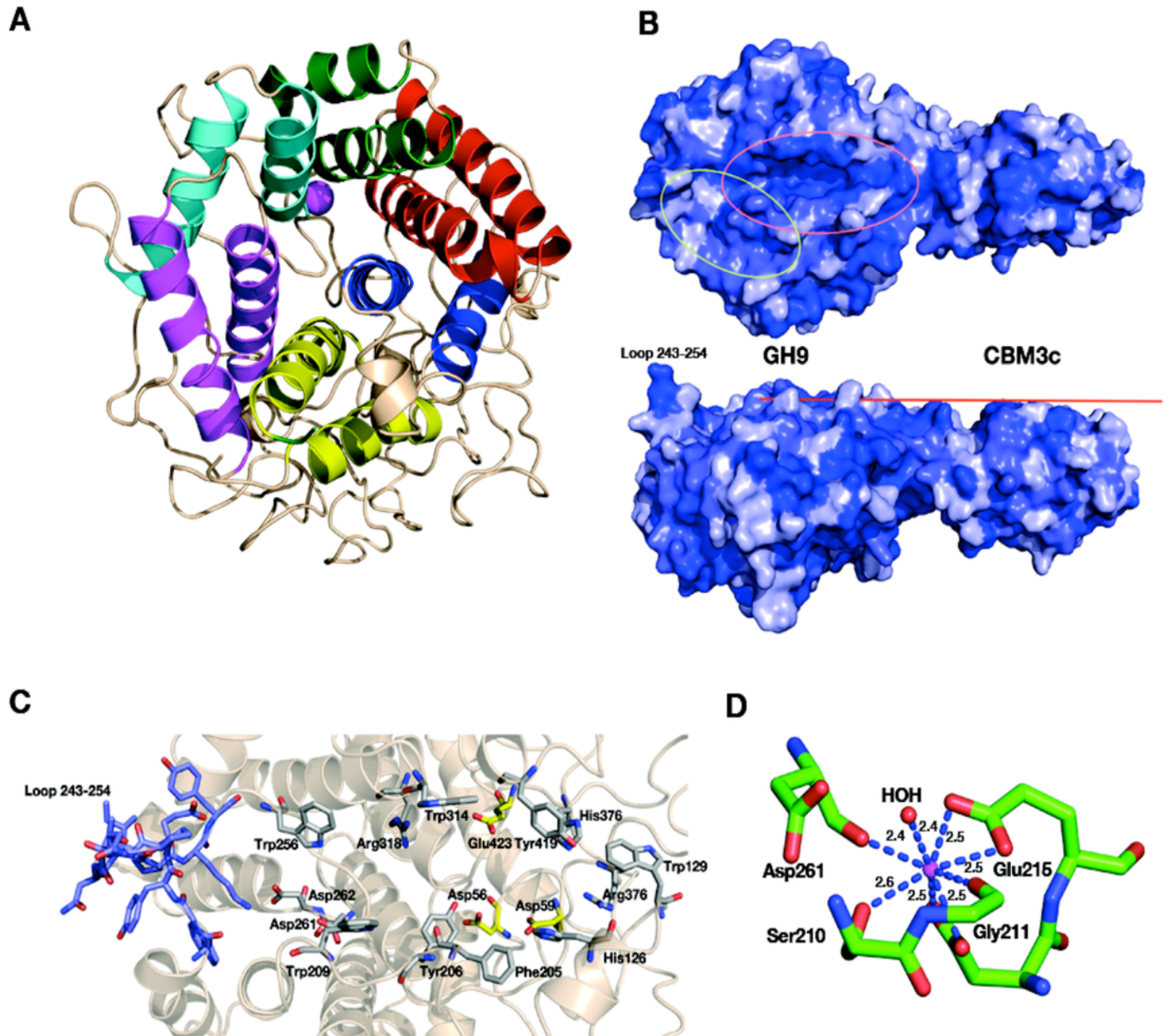



Figure 5. Structure of the CBM3c of Cel9I from *C. thermocellum.*

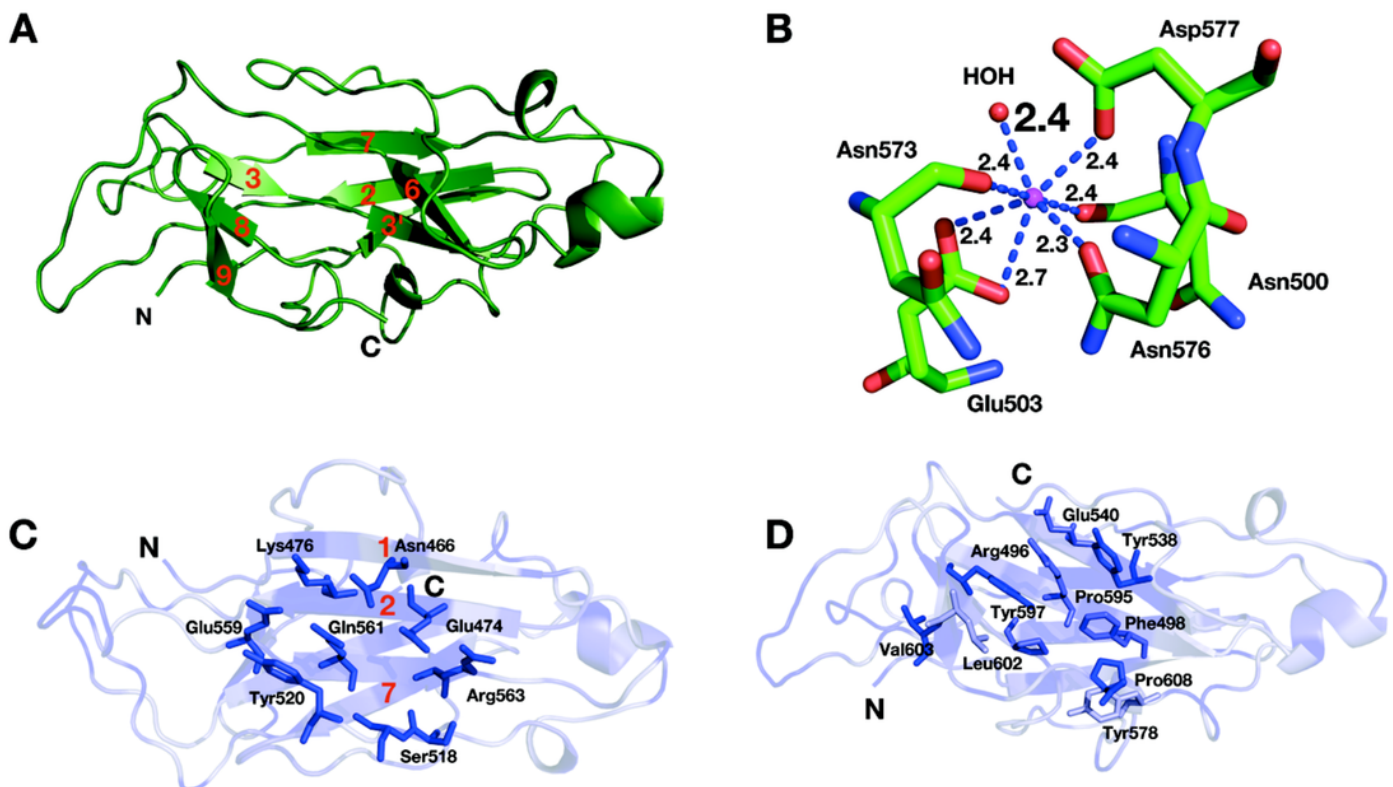
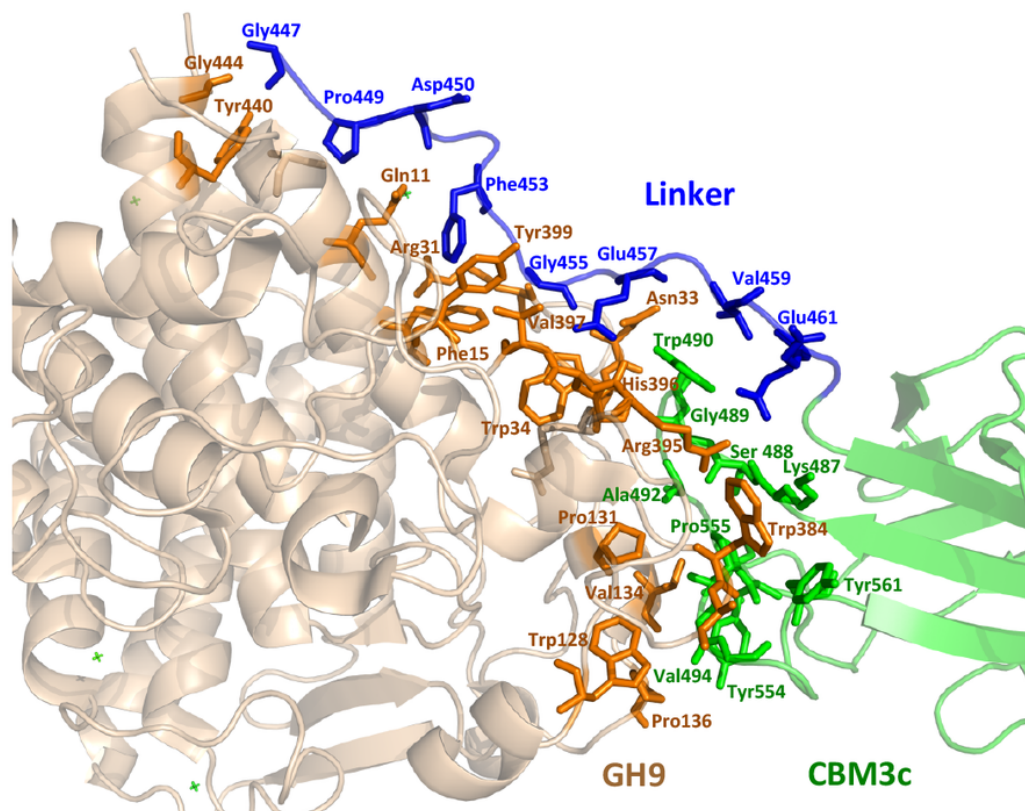
C and N termini are indicated **A.** Cartoon representation, β-strands are numbered according to the alignment with Cel9G from *C. cellulolyticum*, and Cel9A from *T. fusca*. **B.** Calcium-binding site of the CBM3c. **C.** Birds-eye view of the flat surface. Residues are shaded according to their degree 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.

6

Figure 6. Contact residues of the reassembled GH9-CBM3c complex.

A. The GH9 module is colored in brown, CBM3c in green and the linker in blue. Contact residues of the GH9, CBM3c and linker are colored orange, green and blue, respectively. The contact residues between the linker and the domains are described in the text. 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.

A

Linker — GH9 — CBM3c

B

GH9 module

```
CelI    TGAFNYGEALQKAIFFYECQRSGKLDSSTLRLNWRGDSGLDDGKDAGIDLTGGWYDAGDH    60
CelG    AGTYNYGEALQKSIMFYEFQRSGDLPAD-KRDNWRDDSGMKDGSDVGVDLTGGWYDAGDH    59
E4      EPAFNYAEALQKSMFFYEAQRSGKLPEN-NRVSWRGDSGLNDGADVGLDLTGGWYDAGDH    59

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

CelI    DGHADHAWWGPAEVMPMERPSYKVDRSSPGSTVVAETSAALAIASIIFKKVDGEYSKECL   180
CelG    DGGKDHSWWGPAEVMQMERPSFKVDASKPGSAVCASTAASLASAAVVFKSSDPTYAEKCI   179
E4      DGDADHKWWGPAEVMPMERPSFKVDPSCPGSDVAAETAAAMAASSIVFADDDPAYAATLV   179

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

CelI    G--RSFVVGFGENPPKRPHHRTAHGSWADSQMEPPEHRHVLYGALVGGPDST-DNYTDDI   416
CelG    G--RSFVVGYGVNPPQHPHHRTAHGSWTDQMTSPTYHRHTIYGALVGGPDNA-DGYTDEI   413
E4      PRNSSYVVGFGNNPPRNPHHRTAHGSWTDSIASPAENRHVLYGALVGGPGSPNDAYTDDR   417

CelI    SNYTCNEVACDYNAGFVGLLAKMYKLYGEL    446
CelG    NNYVNNEIACDYNAGFTGALAKMYKHSG —   441
E4      QDYVANEVATDYNAGFSSALAMLVEEYG —   445
```

CBM3c module

Linker
```
CelI    GSPDPKFNGIEEVPEDEIFVEAGVNASGNNFIEIKAIVNNKSGWPARVCENLSFRYFINI   506
CelG    GDPIPNFKAIEKITNDEVIIKAGLNSTGPNYTEIKAVVYNQTGWPARVTDKISFKYFMDL   501
E4      GTPLADFPPTEEPDGPEIFVEAQINTPGTTFTEIKAMIRNQSGWPARMLDKGTFRYWFTL   505

CelI    EEIVNAGKSASDLQVSSSYNQGAKLS--DV--KHYKDNIYYVEVDLSGTKIYPGGQSAYK   562
CelG    SEIVAAGIDPLSLVTSSNYSEGKNTKVSGVLPWDVSNNVYYVNVDLTGENIYPGGQSACR   561
E4      DE----GVDPADITVSSAYNQCATPE--D--VHHVSGDLYYVEIDCTGEKIFPGGQSEHR   557

CelI    KEVQFRISAPEGTV-FNPENDYSYQGLSAGTV-VKSEYIPVYDAGVLVFGREPLE       615
CelG    REVQFRIAAPQGTTYWNPKNDFSYDGLPTTSTVNTVTNIPVYDNGVKVFGNEP--       614
E4      REVQFRIAGGPG---WDPSNDWSFQGIGNE---LAPAPYIVLYDDGVPVWGTAP--       605
```

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.