

# SCIENTIFIC REPORTS



OPEN

## Molecular characterization of a family 5 glycoside hydrolase suggests an induced-fit enzymatic mechanism

Received: 21 October 2015

Accepted: 25 February 2016

Published: 01 April 2016

Marcelo V. Liberato<sup>1,\*</sup>, Rodrigo L. Silveira<sup>2,\*</sup>, Érica T. Prates<sup>2</sup>, Evandro A. de Araujo<sup>1</sup>, Vanessa O. A. Pellegrini<sup>1</sup>, Cesar M. Camilo<sup>1</sup>, Marco A. Kadowaki<sup>1</sup>, Mario de O. Neto<sup>3</sup>, Alexander Popov<sup>4</sup>, Munir S. Ska<sup>2</sup> & Igor Polikarpov<sup>1</sup>

Glycoside hydrolases (GHs) play fundamental roles in the decomposition of lignocellulosic biomaterials. Here, we report the full-length structure of a cellulase from *Bacillus licheniformis* (BLcel5B), a member of the GH5 subfamily 4 that is entirely dependent on its two ancillary modules (Ig-like module and CBM46) for catalytic activity. Using X-ray crystallography, small-angle X-ray scattering and molecular dynamics simulations, we propose that the C-terminal CBM46 caps the distal N-terminal catalytic domain (CD) to establish a fully functional active site via a combination of large-scale multidomain conformational selection and induced-fit mechanisms. The Ig-like module is pivoting the packing and unpacking motions of CBM46 relative to CD in the assembly of the binding subsite. This is the first example of a multidomain GH relying on large amplitude motions of the CBM46 for assembly of the catalytically competent form of the enzyme.

The production of biofuels from renewable sources is an important element of the global strategy for generating sustainable energy with reduced environmental impact. Current technologies for obtaining liquid biofuels and green chemicals rely on the enzymatic digestion of lignocellulosic biomass from a variety of feedstocks<sup>1</sup>. Plant biomass—the most abundant source of carbohydrates on Earth—is primarily composed of cellulose microfibrils surrounded by a hydrated heteropolymeric matrix of hemicellulose and lignin. Plant biomass may be subjected to thermo-chemical pretreatments and enzymatic reactions to produce soluble fermentable sugars.

The canonical model of hydrolytic degradation of cellulose requires at least three classes of enzymes<sup>1</sup>. Cellulohydrolases (CBHs) processively cleave the glycosidic bonds at the reducing and non-reducing ends of cellulose chains in crystalline regions to produce cellobiose. Endoglucanases (EGs) introduce random cuts in the amorphous regions of cellulose and create new chain extremities for CBH attack; thus, these enzymes act synergistically. The released cellobiose molecules are then enzymatically converted into glucose by  $\beta$ -glucosidases.

The molecular architecture of glycoside hydrolases (GHs) frequently consists of a catalytic domain (CD), where hydrolysis occurs, and one or more ancillary modules (AMs), which are usually connected by less structured linkers. The most common type of AMs are carbohydrate-binding modules (CBMs), which are able to recognize and bind specific carbohydrate chains<sup>2</sup>. Generally distinct and independent structural domains, the CBMs facilitate carbohydrate hydrolysis by increasing the local concentration of enzymes at the surface of insoluble substrates, thereby targeting the CD component to its cognate ligands. CBMs might also disrupt the crystalline structure of cellulose microfibrils, although the underlying mechanism remains poorly understood. Thus, CBMs enhance the accessibility of CDs to carbohydrate chains to improve enzymatic activity, making them important candidates for the development of effective biomass-degrading enzymes in industrial settings.

Although there are examples of active GHs that lack AMs<sup>3,4</sup>, the majority of the enzymes depend on AMs for activity<sup>2</sup>. In several cases, CBMs were shown to extend and complement the CD substrate-binding site in

<sup>1</sup>São Carlos Institute of Physics, University of São Paulo, São Carlos 13566-590 São Paulo, Brazil. <sup>2</sup>Institute of Chemistry, University of Campinas, Campinas 13084-862, São Paulo, Brazil. <sup>3</sup>Institute of Bioscience, State University of São Paulo, Botucatu 18618-970, São Paulo, Brazil. <sup>4</sup>European Synchrotron Radiation Facility, Grenoble CS40220, France. <sup>\*</sup>These authors contributed equally to this work. Correspondence and requests for materials should be addressed to M.S.S. (email: skaf@iqm.unicamp.br) or I.P. (email: ipolikarpov@ifsc.usp.br)

multimodular carbohydrate-active enzymes, such as endo/exocellulase E4 from *Thermobifida fusca*<sup>5</sup>, chitinase B from *Serratia marcescens*<sup>6</sup>, a starch phosphatase from *Arabidopsis thaliana*<sup>7</sup> and a GH5 subfamily 4 (GH5\_4) endoglucanase from *Bacillus halodurans* (*Bh*Cel5B)<sup>8</sup>. A pioneer work of Sakon *et al.*<sup>5</sup> revealed that rigid structural extension of the GH9 CD by a type C CBM3 imprints a processive mode of action to this endoglucanase. Further publications showed that CBM-based structural extensions of the active site are important for substrate engagement and recognition<sup>7–9</sup>.

Recently, Venditto *et al.*<sup>8</sup> reported the X-ray structure of the tri-modular GH5\_4 endoglucanase from *Bacillus halodurans* (31% sequence identity to *Bi*Cel5B), with the CBM46 extension of the active site appended to the CD via an immunoglobulin (Ig)-like module. Removal of the CBM46 caused a ~60-fold reduction of the activity of the enzyme against  $\beta$ -glucans, but showed little or no effect against xyloglucan hydrolysis. Moreover, the CBM46 mediated a significant increase in the *Bh*Cel5B activity in plant cell wall settings. Modeling of cellobiose in the negative subsites of the active site of *Bh*Cel5B demonstrated the structural conservation of the -1 position, but provided little information about direct interactions between CBM46 and the substrate. It was speculated that  $\beta$ -1,3 kink of the  $\beta$ -glucan might allow the ligand to reach for the CBM46, whereas pure  $\beta$ -1,4 linkages in the backbone of xyloglucan chains would restrict binding to the CD, thus explaining the lack of influence of the CBM46 on the enzymatic activity of *Bh*Cel5B against xyloglucans in solution<sup>8</sup>. It was also argued that the CBM46 could potentialize the activity by driving *Bh*Cel5B towards xyloglucan-rich regions in the context of the plant cell walls, but no large-scale conformational adjustments of the AMs have been shown to occur or suggested to take part in the enzymatic activity<sup>8</sup>.

The mechanisms of ligand binding mediated by large-scale conformational changes in proteins following the induced-fit<sup>10</sup> or conformational selection<sup>11</sup> models have recently attracted considerable attention<sup>12–14</sup>. Although initially introduced as contradictory theories, these two limiting cases can be unified considering the flux description concept<sup>13</sup> or the extended conformational selection model<sup>14</sup>. While local ligand-induced conformational adjustments have been reported for carbohydrate-active enzymes<sup>5,15–17</sup>, cognate ligands recognition and hydrolysis mediated by a large-scale conformational mobility of distinct domains in multidomain settings is uncommon for endoglucanases.

Here, we report the crystal structure of a full-length GH5\_4 enzyme from *Bacillus licheniformis* (*Bi*Cel5B) that exhibits two AMs (Ig-like module and CBM46) appended to the CD. We structurally and functionally characterize the enzyme using a combination of protein crystallography, small-angle X-ray scattering (SAXS), molecular dynamics computer simulations and site-directed mutagenesis, and show that the AMs and their conformational mobility are essential for the enzymatic activity of *Bi*Cel5B. We find that the large-scale conformational adjustments of the distal CBM46 mediated by the Ig-like hinge domain are crucial in active-site assembly for optimal substrate binding and hydrolysis. We propose that the *Bi*Cel5B conformational selection/induced-fit mechanism of hydrolysis represents a novel paradigm that applies to several GH5\_4 members and, possibly, to a number of other multidomain GHs.

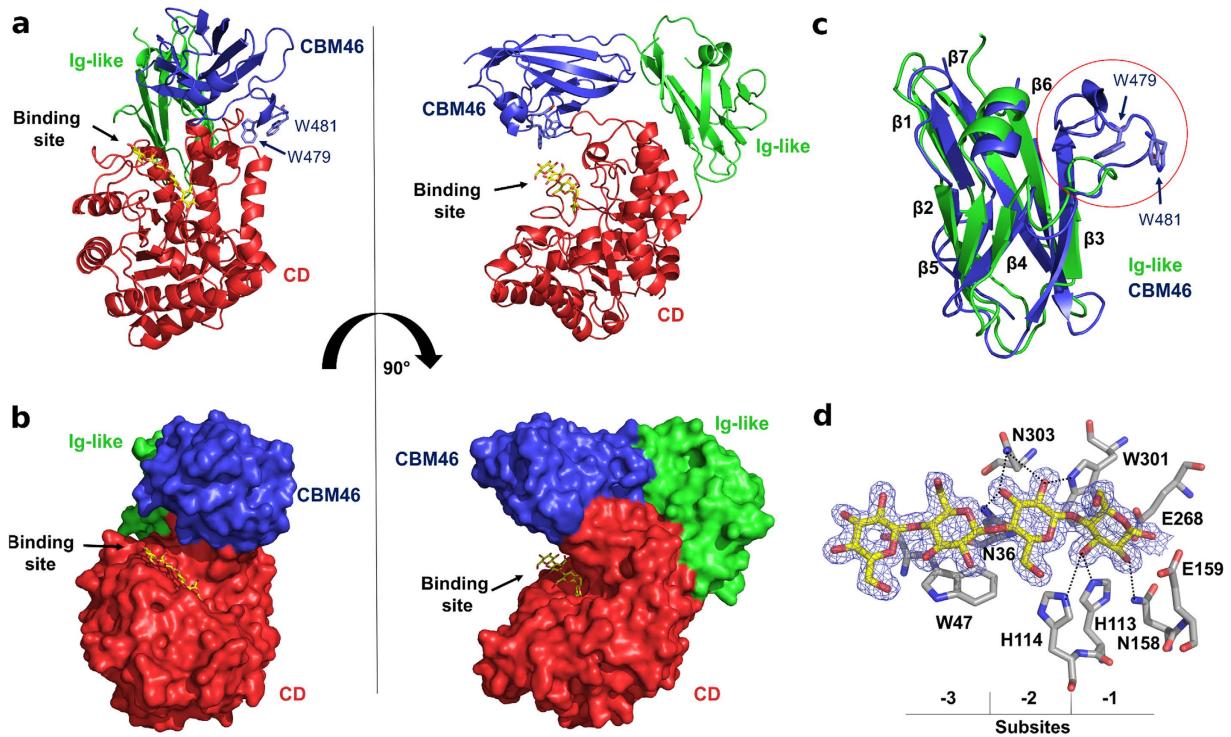
## Results

***Bi*Cel5B Crystal Structure.** *Bi*Cel5B crystals in the substrate-free form and complexed with cellobiose (C5) were obtained and diffracted to 1.7 Å and 1.75 Å resolutions, respectively (Supplementary Table 1). The substrate-free and complexed structures exhibited no substantial conformational differences (with the exception of the substrate). Because of minor variations in the loops located distal to the substrate-binding site, a root mean squared deviation (rmsd) of 0.33 Å between the complexed and substrate-free structures was observed. A single protein chain occupies the asymmetric unit, and most of the residues were built, with the exception of the first 17 residues and those in the loop between L398 and P405 due to weak electron density.

The *Bi*Cel5B structure comprises three distinct domains: an N-terminal CD (residues 18 to 330), an Ig-like module (residues 335 to 428) and a family 46 CBM (residues 432 to 533) (Fig. 1A,B). Similarly to other members of the GH5 family, the CD of *Bi*Cel5B has a typical TIM barrel fold with eight inner  $\beta$ -strands and eight outer  $\alpha$  helices that are interconnected by loops and three short  $\alpha$  helices. Very short linkers, D429-D430-P431 and V331-P332-N333-A334, connect the CBM46 to the Ig-like module and the Ig-like module to the CD, respectively. Both Ig-like module and CBM46 have a  $\beta$ -sandwich fold composed of two  $\beta$ -sheets of four and three antiparallel  $\beta$ -strands interconnected by loops and a short  $\alpha$  helix between strands  $\beta$ 3 and  $\beta$ 4 (Fig. 1C). A structural comparison between the Ig-like module and the CBM46 using the Dali server<sup>18</sup> yielded an rmsd of 2.3 Å and a Z-score of 10.2. However, despite their structural resemblance, these modules share only 17% sequence identity. A structure-based search performed using the same server showed that the Ig-like module is similar to the Ig-like module from a recently solved crystal structure of a tri-modular GH5\_4 enzyme from *Bacillus halodurans*<sup>8</sup>, *Bh*Cel5B, with rmsd = 1.3 Å and Z-score = 15.3. The CBM46 from *Bh*Cel5B is the most structurally similar to *Bi*Cel5B CBM46, with rmsd = 1.6 Å and Z-score = 12.4. The sequence identity relative to *Bh*Cel5B, however, is low (28% for Ig-like and 25% for CBM46).

The Ig-like module, adjacent to the CD, contains only one tyrosine (Y367) exposed to solvent and no tryptophan residues. Because aromatic residues play a major role in glucose recognition, this observation suggests that substrate binding may not be the primary function of Ig-like module. In contrast, the CBM46 has three tryptophan residues, two of which face the CD substrate binding site (Fig. 1A), indicating that it may be actively engaged in the carbohydrate binding.

Electron density maps clearly reveal the presence of a cellobiose (C4) and not a soaked cellobiose (C5) in the CD negative substrate-binding subsites (Fig. 1D), indicating that *Bi*Cel5B is catalytically active in the crystal state and able to cleave a C5 molecule. The lack of electron density verifies the absence of the fifth glucose moiety from the soaked C5, and a closer inspection of the structure confirmed that the presence of a fifth glucose unit would be sterically hindered by the catalytic residues on the reducing end and by residue R234 of a symmetry-related enzyme molecule on the non-reducing end. The ability of *Bi*Cel5B to cleave C5 into



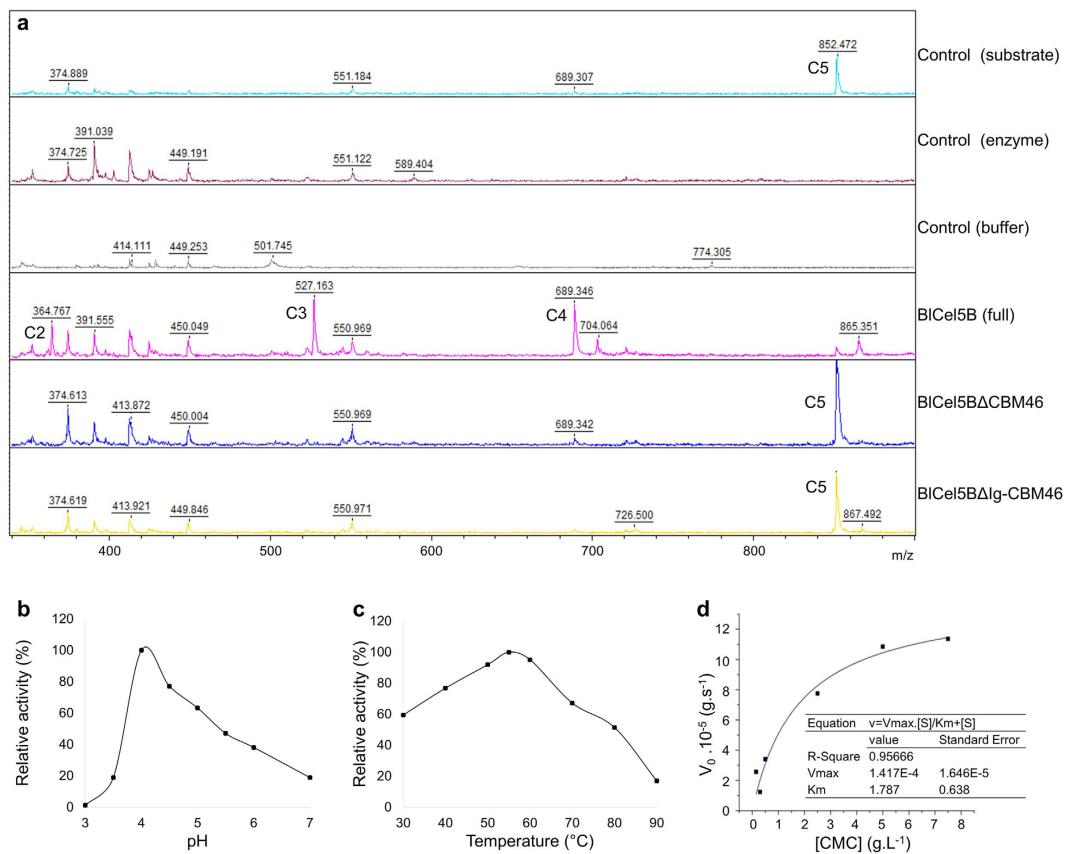
**Figure 1. Crystal models of *Bl*Cel5B.** Complete structure is shown as a cartoon illustration in (a) and a van der Waals surface in (b). The CD module (red) has a typical TIM-barrel fold, and its substrate-binding site is adjacent to CBM46 (blue). Despite the proximity of the binding site in the crystallographic model, the CBM46 residues W479 and W481 are distant from the substrate cellotetraose (yellow). The Ig-like domain (green) has a lateral position, serving as a connector between the CD and CBM46. (c) A superposition of the Ig-like domain and CBM46 illustrates their structural similarity, with most of the structural differences present in the loop highlighted by a red circle. (d) Cellotetraose occupies subsites -1 to -3 and is primarily coordinated by the residues represented in gray.

glucose and C4 molecules in solution was demonstrated by enzymatic product profile mass spectrometry analysis (Fig. 2A). The C4 oligomer in the *Bl*Cel5B binding site is coordinated by hydrogen bonds to residues N36, H113, H114, N158, W301, and N303 and by a CH- $\pi$  interaction with residue W47 (Fig. 1D). These residues belong to the CD and are conserved in the GH5 family.

***Bl*Cel5B enzymatic activity.** *Bl*Cel5B exhibits optimum activity toward carboxymethylcellulose (CMC; 8.7 U/mg) at a pH of 4.0 and 55 °C and retains approximately half of its maximum activity at 80 °C, demonstrating considerable thermal stability (Fig. 2B,C). *Bl*Cel5B is also active on  $\beta$ -glucan (34 U/mg), lichenan (17.8 U/mg) and xyloglucan (15.7 U/mg) substrates (Table 1), whereas no activity was detected on galactomannan, rye arabinoxylan, 1,4- $\beta$ -mannan or the insoluble substrate Azo-Avicel. Kinetic parameters were calculated assuming Michaelis-Menten behavior with CMC as substrate:  $K_m = 1.78 \text{ g L}^{-1}$  and  $V_{max} = 1.41 \times 10^{-4} \text{ g s}^{-1} \text{ mg protein}^{-1}$  (Fig. 2D). Although *Bl*Cel5B is not a highly active enzyme against one specific substrate as compared to others GH5\_4<sup>19</sup>, it has the advantage of being active against different substrates with  $\beta$ -1,3 and/or  $\beta$ -1,4 glycosidic linkages.

To understand the importance of the ancillary modules for *Bl*Cel5B activity, enzymatic assays were carried out using four enzyme mutants: a CBM46 deletion ( $\Delta$ CBM46) and an Ig-like + CBM46 deletion ( $\Delta$ Ig-CBM46) as well as point mutations of the CBM46 inner surface residues W479A and W481A. These mutants were expressed and purified as described for the wild-type enzyme. Strikingly, neither of the deletion variants exhibited detectable activity toward any of the substrates tested using full-length *Bl*Cel5B (Table 1), demonstrating that the Ig-like module and the CBM46 are essential for *Bl*Cel5B activity. Thermal shift assays were conducted to confirm structural stability of the mutants (Supplementary Fig. 1). All of the constructs showed similar melting temperatures: 62 °C for *Bl*Cel5B, 58 °C for *Bl*Cel5B $\Delta$ CBM46, 56 °C for *Bl*Cel5B $\Delta$ Ig-CBM46, 65 °C for *Bl*Cel5B W479A and 59 °C for *Bl*Cel5B W481A, thus confirming their proper overall fold.

We also examined the function of the CBM46 inner surface residues W479 and W481 (Fig. 1A) in *Bl*Cel5B activity by performing enzymatic assays with W479A and W481A mutants. Both mutations reduced enzymatic activity toward all tested substrates (Table 1), with W481A having a stronger effect than W479A (~64% vs. 79% activity relative to wt *Bl*Cel5B using  $\beta$ -glucan and ~10% vs. 50% using CMC). This indicates that CBM46 must interact with the substrate via residues W479 and W481. However, since the *Bl*Cel5B crystal structure exhibits no close contact between these residues and the substrate, these results suggest the existence of large-amplitude interdomain motions that may enable direct interactions between CBM46 and the carbohydrate.

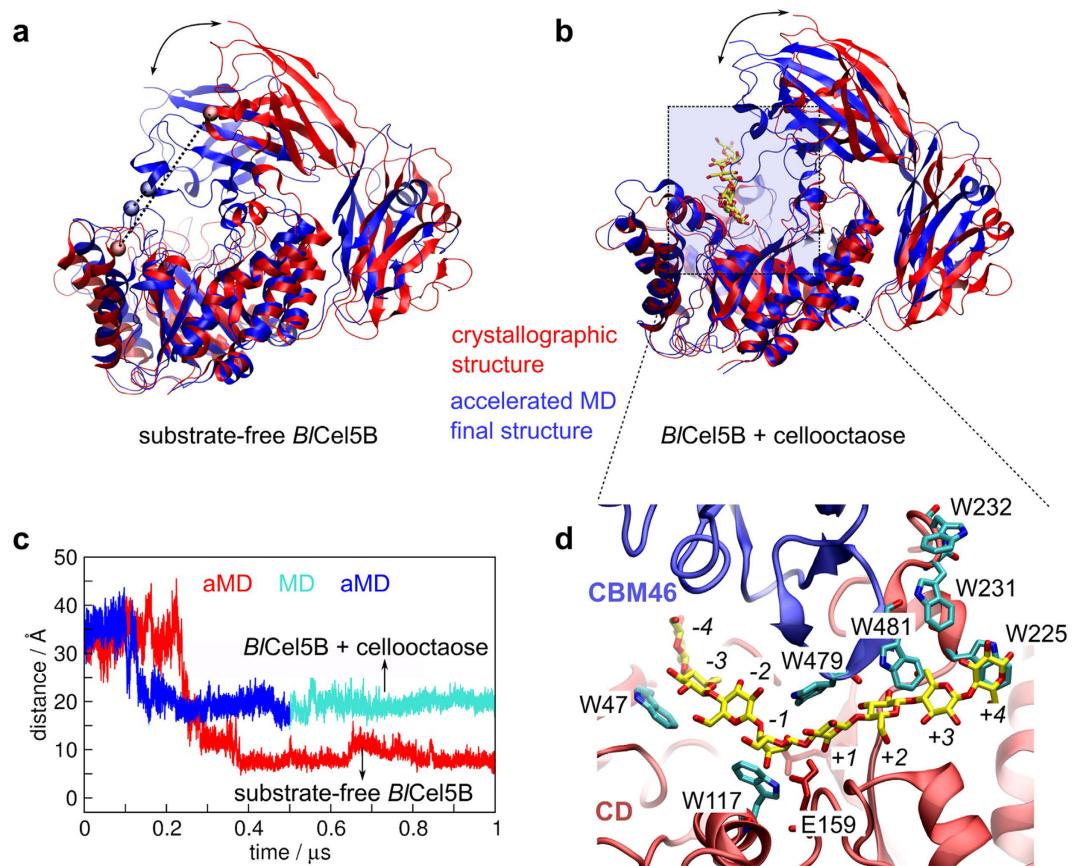


**Figure 2. *BICel5B* enzymatic activity characterization.** (a) MALDI/TOF-MS spectra of the products released after incubation of *BICel5B* and its two deletion constructs ( $\Delta$ CBM46 and  $\Delta$ Ig-CBM46) with the substrate cellopentaose (C5). The first three spectra show the substrate, enzyme and buffer controls. The forth spectrum reveals that full length *BICel5B* is capable of enzymatic hydrolysis of C5 into smaller oligosaccharides such as C4, C3 and C2. The last two spectra show that the C-terminal deletions eliminate the enzyme activity. *BICel5B* activities on CMC as functions of pH and temperature are shown in (b) and (c), respectively. The enzyme exhibits optimal pH of 4.0 and optimal temperature of 55 °C, retaining about 50% of its activity at 80 °C. (d) Michaelis-Menten curve using CMC as a substrate.

Substrate (1%)	Relative Activity (%)				
	WT*	W479A	W481A	$\Delta$ CBM46	$\Delta$ Ig-CBM46
$\beta$ -glucan	100	79.1	63.6	nd	nd
CMC	25.5	12.2	2.4	nd	nd
Lichenan	52.4	41	28.6	nd	nd
Xyloglucan	45.2	41.2	30.8	nd	nd
Azo-Avicel	nd**	nd	nd	nd	nd
Arabinoxylan	nd	nd	nd	nd	nd
Galactomannan	nd	nd	nd	nd	nd
1,4- $\beta$ -mannan	nd	nd	nd	nd	nd

**Table 1. Activity of *BICel5B* constructs against tested substrates.** \*WT = wild type. \*\*nd = not detected.

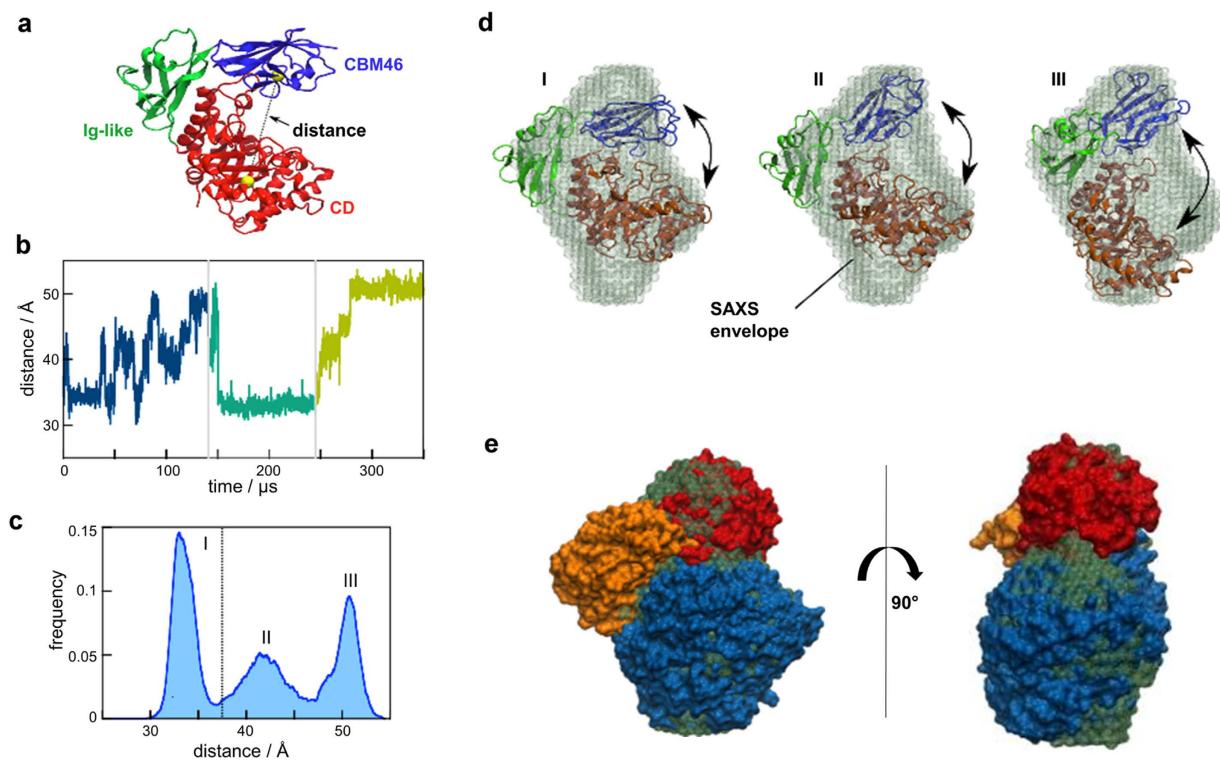
***BICel5B* dynamics and binding-site architecture.** Molecular dynamics (MD) simulations were performed to investigate the conformational mobility of *BICel5B*. In the simulations of the crystal structure for *BICel5B* bound to C4, the substrate dissociates from the protein within the first 100 ns of the simulation time (Supplementary Fig. 2A). This observation suggests that cellobetaose does not exhibit detectable affinity for this specific *BICel5B* conformation in solution, as one might otherwise expect for a reaction product. No changes beyond local fluctuations were observed in any of the three *BICel5B* domains within the time scale of these runs (400 ns; Supplementary Fig. 2B). However, the CBM46 and Ig-like domains did exhibit rigid body-like motions relative to the CD, with rmsd values around 2.3 Å and 1.8 Å, respectively, suggesting that *BICel5B* may execute large-amplitude interdomain motions over longer time scales (Supplementary Fig. 2B,C).



**Figure 3. Open-close transitions of B/Cel5B.** (a) B/Cel5B in the absence of substrate and (b) in the presence of cellobiose, as observed in our aMD simulations. The distance between the  $\alpha$  carbon of residues I120 (CD) and E477 (CBM46), illustrated as spheres in (a), is plotted in (c), revealing a transition by the decrease in the distance from 40  $\text{\AA}$  to 7  $\text{\AA}$  (substrate-free) or 20  $\text{\AA}$  (in presence of cellobiose). For the substrate-free enzyme, the red line refers to a 1  $\mu\text{s}$ -long aMD; for the B/Cel5B-cellobiose complex, the first 500 ns refers to aMD (in blue) and the second 500 ns to conventional MD (in turquoise). (d) A snapshot of the B/Cel5B-cellobiose complex, highlighting the tryptophan residues that interact with the glucan chain in subsites  $-4$  to  $+4$ . Residues W479 and W481 belong to CBM46 and only become available for substrate interactions in the closed configuration of B/Cel5B.

Accordingly, simulations were then performed using accelerated molecular dynamics (aMD) techniques to probe B/Cel5B interdomain motions. aMD enhances conformational sampling by raising the basins of the dihedral potential energy surface without affecting the general form of the atomistic potential, thereby increasing transition rates between different local minima. aMD trajectories corresponding to more than 1.0  $\mu\text{s}$  of conventional MD runs were generated<sup>20</sup>. During these simulations, we observed occlusive conformations between CBM46 and CD that resulted in a rearrangement of the enzyme's architecture around the active site (Video S1). Figure 3A shows B/Cel5B in the crystallographic conformation (red) and in a selected configuration obtained with aMD (blue) in the absence of the substrate. Interdomain motions were gauged by the time evolution of the distance between the  $\alpha$  carbons of residues I120 and E477 (represented as spheres in Fig. 3A), belonging to the CD and CBM46, respectively. Figure 3C shows that the I120-E477 distance (red curve) gradually decreases from  $\sim$ 35  $\text{\AA}$  to  $\sim$ 7  $\text{\AA}$  within the first half of the 1.0  $\mu\text{s}$  aMD trajectory, indicating a transition between the semi-open (crystallographic) and occluded (aMD sampled) configurations. During the second half of the aMD simulation, the full-length enzyme remained in the closed conformation, with the CBM46 covering the carbohydrate-binding site. These results suggest that B/Cel5B undergoes large-scale interdomain movements that enable interactions between CBM46 and the substrate bound to the CD.

To study the interactions of B/Cel5B with a non-hydrolyzed glucan chain, we built a model structure with a cellobiose (C8) chain spanning the entire positive (+1 to +4) and negative ( $-4$  to  $-1$ ) subsites of the enzyme. Starting from the crystallographic B/Cel5B conformation, the C8 molecule deviated significantly from the active site and assumed a non-productive binding mode (Supplementary Fig. 2D). This observation suggests that the open conformation of B/Cel5B is not able to hold the substrate in a position suitable for hydrolysis (Supplementary Fig. 2E). However, after subjecting the B/Cel5B-C8 complex to a 0.5  $\mu\text{s}$  aMD simulation with harmonic restraints on the C8 chain to prevent it from deviating from the productive binding mode, the CBM46 readily closed over the CD and trapped the C8 chain in position for hydrolysis (Fig. 3B). In the presence of



**Figure 4. Large-scale movements of *Bl*Cel5B modules and superposition of their representative conformations with the SAXS envelope.** (a) *Bl*Cel5B structure showing the distance between the backbone beads of residues I120 and E477, which are centrally located in CD and CBM46, respectively, as a metric for the relative disposition between the two domains. (b) Time history of the I120-E477 distance computed using CG-MD simulations. Different colors separated by vertical lines correspond to independent simulations of approximately 120  $\mu$ s. (c) The distance distribution indicates three major peaks: closed or occluded CBM46/CD conformations (I); semi-open (II), which is similar to the crystallographic structure; and extended conformers (III). (d) Superimposition of the three representative molecular conformations of *Bl*Cel5B with the SAXS model. (e) Average structures obtained from the simulation segments corresponding to population groups I-III, which are individually superposed on the SAXS envelope.

the substrate, CBM46 adopts a final conformation intermediate between the crystallographic structure and that observed in the substrate-free *Bl*Cel5B aMD simulations; this is illustrated by the I120-E477 distance, which stabilizes near 20 Å in the closed configuration that traps the C8 molecule (in contrast to ~7 Å for substrate-free *Bl*Cel5B) (Fig. 3C). This *Bl*Cel5B-C8 configuration remains stable over an additional 500 ns of conventional MD simulation with no restraints (Fig. 3C cyan line, Supplementary Fig. 2E,F).

A closer inspection of the productive binding mode obtained from these extensive simulations reveals that the CBM46 tryptophan residues W479 and W481 (along with CD tryptophan residues) play important roles in carbohydrate recognition and orientation by creating a tunnel-like topology along the *Bl*Cel5B binding cleft, as depicted in Fig. 3D. Together, these results indicate that CBM46 is a key component of the catalytic active complex, providing an explanation as to why CBM46 is essential for the enzymatic activity of *Bl*Cel5B.

To enable substantially longer time scales compared to atomistic simulations, we further explored the dynamics of *Bl*Cel5B using coarse-grained MD (CG-MD) simulations. We performed three independent ~120  $\mu$ s CG-MD simulations, for a total of approximately 360  $\mu$ s of sampling. The distance between the  $\alpha$  carbons of two residues centrally positioned in the CD and CBM46 (Fig. 4A) was monitored, and the results shown in Fig. 4B indicate that the wide-amplitude events described above frequently appear in this time scale. The computed distance distribution depicted in Fig. 4C indicates three main conformational states ranging from (I) closed conformations similar to those encountered in the substrate-free aMD simulations, in which CBM46 interacts with the CD to shape the substrate binding site, to (II) semi-open conformations similar to the crystallographic structure, and (III) extended *Bl*Cel5B conformations in which the CD and CBM46 are even further apart than in the crystal structure.

***Bl*Cel5B conformers fit the SAXS envelope.** SAXS experiments were conducted to assess *Bl*Cel5B conformational states in solution, and the results revealed the enzyme in its monomeric form, with average values of  $R_g = 27.17$  Å and  $D_{\max} = 87.59$  Å (Supplementary Table 2). The *ab initio* dummy atom model (DAM) demonstrated that the SAXS-derived *Bl*Cel5B molecular envelope could not be single-handedly filled by any of the main conformational states encountered in the simulations (Fig. 4D).

It is known that a Kratky plot exhibits a peak with an elevated baseline at high  $q$  for a monodisperse system composed of multi-domain particles with flexible extensions<sup>21,22</sup>. Indeed, an elevation of the baseline toward a hyperbolic-like curve was observed for *B/Cel5B*, indicating a considerable degree of molecular mobility in solution (Supplementary Fig. 3). Thus, the conformational heterogeneity of the enzyme can be decomposed in structural terms as a combination of conformational states identified in our crystallographic and MD studies. We found that the SAXS envelope can be well represented by considering the superimposition of three different representative molecular conformations of *B/Cel5B* (Fig. 4E): a closed or CBM46/CD-occluded conformation extracted from the simulations with a relative weight of 26%, a semi-open conformation represented by the crystal structure corresponding to 40%, and an extended conformation based on simulations that is responsible for 34% of the SAXS envelope. The resulting average scattering curve from this model fits the experimental protein scattering intensity, with  $\chi = 1.89$  (Supplementary Fig. 3).

**GH5\_4 phylogenetic analysis.** To date, there are 427 sequences classified as subfamily 4 members in the CAZy database<sup>23</sup>. After the exclusion of partial sequences and the suppression of highly identical members (higher than 90% identity), 144 sequences containing between 277 and 400 residues were aligned and used to construct a phylogenetic tree (Supplementary Fig. 4A). According to PFAM database<sup>24</sup> conserved domain classification, 128 GH5 enzymes have an architecture consisting of an N-terminal catalytic module, a CBM\_X2 module and an unknown module of approximately 100 residues at the C-terminus (Supplementary Fig. 4B). Of these, 12 enzymes have an additional CBM1, and 5 have a CBM2 at the N-terminal region. Based on this PFAM architecture and CAZy subfamily classification, all the 144 enzymes (including *B/Cel5B*) belong to the GH5\_4 subfamily and group together in the same branch of the phylogenetic tree, evidencing a common ancestor. These results support the hypothesis that the enzymes may employ the same mechanism by which ligand binding is mediated by an extensive conformational breathing of the enzyme that involves the large-scale movement of CBM46 around the Ig-like module (CBM\_X2) as a structural hinge.

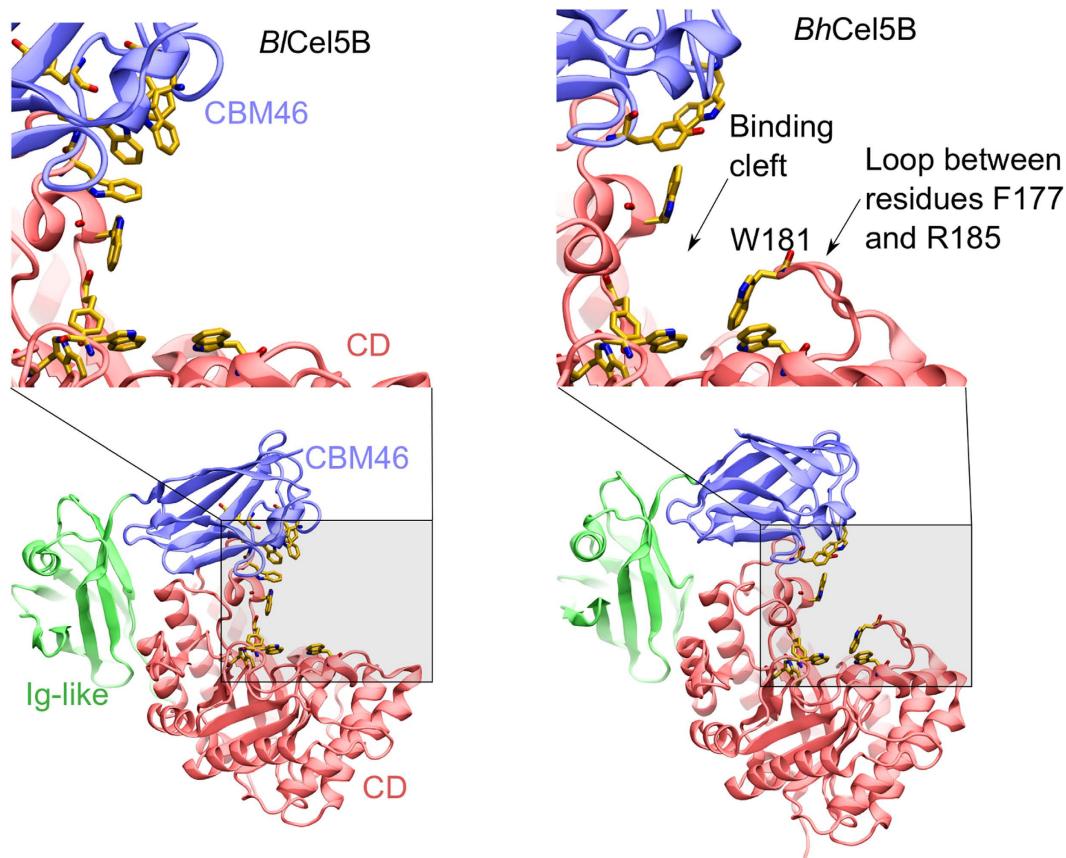
## Discussion

Growing interest in biotechnological applications of enzymes exhibiting activity toward lignocellulosic biomass has sparked efforts in the discovery and development of novel enzymes, as well as the search for a deeper understanding of their mechanisms of action. Here, we elucidate the trimodular molecular architecture of the full-length *B/Cel5B*, a member of the GH5\_4 subfamily, for which large-scale conformational dynamics appears to play a central role in its enzymatic activity. Full-length *B/Cel5B* is active on both cellulosic and hemicellulosic substrates and auxiliary modules are crucial for its activity.

Most carbohydrate-active enzymes are modular and consist of a catalytic domain appended to one or more separate AMs. AMs, such as CBMs, typically recognize carbohydrates and target their cognate catalytic domains toward the substrate. Because the structural analysis of the protein is challenging if the linkers connecting the structural subunits of the enzyme are long and flexible, the standard approach is to study the domains separately. In this work, a combination of protein crystallography, computational molecular dynamics, and SAXS analyses enabled the identification of a new conformational selection-based molecular mechanism that involves GH5 catalytic domain and two AMs in full-length *B/Cel5B*. We observed that the *B/Cel5B* distal CBM46 is directly involved in shaping the local architecture of the substrate-binding site. Although the CD alone appears unable to bind the substrate for catalysis, the AMs exhibit open-close motions that allow the substrate to be captured in a suitable position for hydrolysis. Here, we advocate that large-amplitude motions of AMs are crucial for assembling the enzyme into its active conformation, highlighting a new function of CBMs. This mechanism of substrate binding closely resembles the extended conformational selection model<sup>13,14</sup>, with the induced-fit mechanism of reaction<sup>10</sup> as its limiting case. To the best of our knowledge, this enzymatic mechanism has not been proposed previously for any GH.

The CD binding site of *B/Cel5B* is open and relatively flat and is thus barely able to properly hold the substrate in position for catalysis without assistance from the CBM46. In contrast, other GH5s belonging to subfamily 4 listed in the Protein Data Bank<sup>19,25–28</sup> exhibit a deep binding cleft or tunnel that can effectively entrap the substrate for catalysis (Fig. 5). Due to the marked interdomain conformational rearrangement observed in our simulations, the CBM46 generates a confined binding site in *B/Cel5B* that resembles the binding site architecture of the other GH5 enzymes that lack AMs. Thus, *B/Cel5B* appears to have adopted a strategy of CBM46-mediated interactions for proper functioning. Although the homologous *BhCel5B* has the same domain architecture of *B/Cel5B* and belongs to the same subfamily (a comparison of the sequence and structure of *B/Cel5B* and *BhCel5B* is presented in Supplementary Fig. 5), its binding site exhibits important differences that may impact the catalytic mechanism. The *BhCel5B* binding site is V-shaped and deeper than the *B/Cel5B* binding site (Figs 5 and 6). This is due to the loop between residues F177 and R185 from *BhCel5B* (absent in the *B/Cel5B*), which contains residue W181 that forms part of the binding cleft (Fig. 6). Consistently, although *BhCel5B* CBM46 is important for  $\beta$ -1,3-1,4-glucan hydrolysis (*BhCel5B* is about 60-fold less active without CBM46), the truncated enzyme is completely active against xyloglucan<sup>8</sup>, suggesting that the CBM46, in this case, is necessary for the binding to specific substrates. A closer inspection of results of the phylogenetic analysis, more specifically of the clade composed by GH5\_4 enzymes with trimodular architecture (Supplementary Fig. 4C), reveals subclades whose main characteristic is the varying length of the loop located between residues 161 and 163 (*B/Cel5B* residue numbering). Therefore, our results show that *B/Cel5B* represents a smaller group of enzymes that are completely dependent on its AMs for hydrolysis of plant cell wall polysaccharides, and that the underlying mechanism may rely on large-scale interdomain motions.

The amino acid sequence of the *B/Cel5B* Ig-like module is recognized by BLASTP as belonging to CBM\_X2, a poorly described group that has been compared with CBM-like accessory modules without a defined function<sup>29</sup>. Despite the similarity of *B/Cel5B* Ig-like module to CBMs, it lacks an identifiable aromatic residue-rich



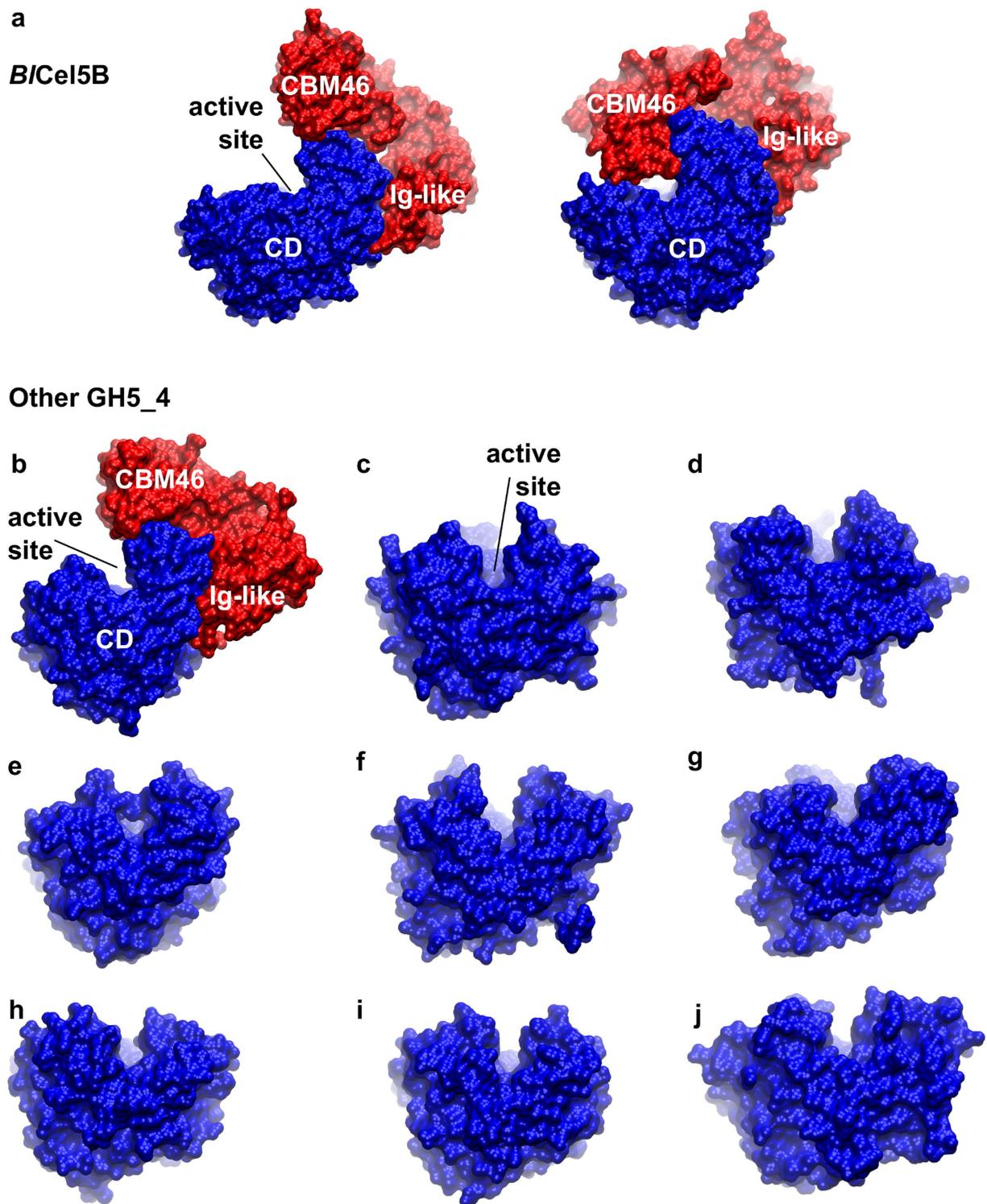
**Figure 5. Comparison of the binding site shape of GH5\_4 enzymes available on the Protein Data Bank.** (a) *B/Cel5B* in the crystallographic and closed configuration; (b) *Bacillus halodurans Cel5B* (*BhCel5B*) (PDB id: 4V2X); (c) *Piromyces rhizinflata* GH5 endoglucanase (PDB id: 3AYR); (d) *Clostridium cellulolyticum* GH5 endoglucanase (PDB id: 1EDG); (e) *Clostridium cellulovorans* GH5 endoglucanase (PDB id: 3NDY); (f) *Bacteroides ovatus* GH5 xyloglucanase (PDB id: 3ZMR); (g) *Paenibacillus pabuli* GH5 xyloglucanase (PDB id: 2JEP); (h) *Prevotella bryantii* GH5 endoglucanase (PDB id: 3VDH); (i) *Ruminiclostridium thermocellum* multifunctional GH5 cellulase, xylanase and mannase (PDB id: 4IM4); (j) *Bacteroides bacterium* AC2a endocellulase (PDB id: 4YHE).

carbohydrate-binding site. Nonetheless, according to our results, the Ig-like module seems to play an important function as a structural hinge, dynamically holding the CBM46 and CD in positions that are appropriate for enzymatic activity.

Based on the results of our crystallographic, computer simulation, and SAXS structural analyses, as well as site-directed mutagenesis and activity assays, we propose a molecular mechanism for *B/Cel5B* substrate binding, which might apply to other GH5\_4 subfamily enzymes that share this tri-modular architecture. *B/Cel5B* can be found in several different conformational states ranging from CBM46/CD closed (or occluded) to extended conformations (Fig. 7). In extended configurations, the substrate may dock at the shallow substrate binding site of CD in one of the semi-closed conformations of the enzyme; however, its binding is properly stabilized for hydrolysis only with the aid of induced-fit repositioning mediated by CBM46. After cleavage, the intrinsic dynamics of *B/Cel5B* would eventually allow the opening of the active site for product release. The proposed mechanism is consistent with our mutagenesis and enzymatic activity assays, which show that the Ig-like module and CBM46 are indispensable for *B/Cel5B* catalytic activity and, together with the CD, form the unique catalytic domain of the enzyme. These experiments reveal a novel function for CBMs in which they are intimately involved in the assembly of the active site and catalytic process. Computer simulations suggest that large-scale motions of the CBM46 and Ig-like domains mediate conformational selection and final induced-fit adjustments to trap the substrate at the active site and promote hydrolysis. SAXS data support the modeling results, providing compelling evidence for highly mobile domains in solution.

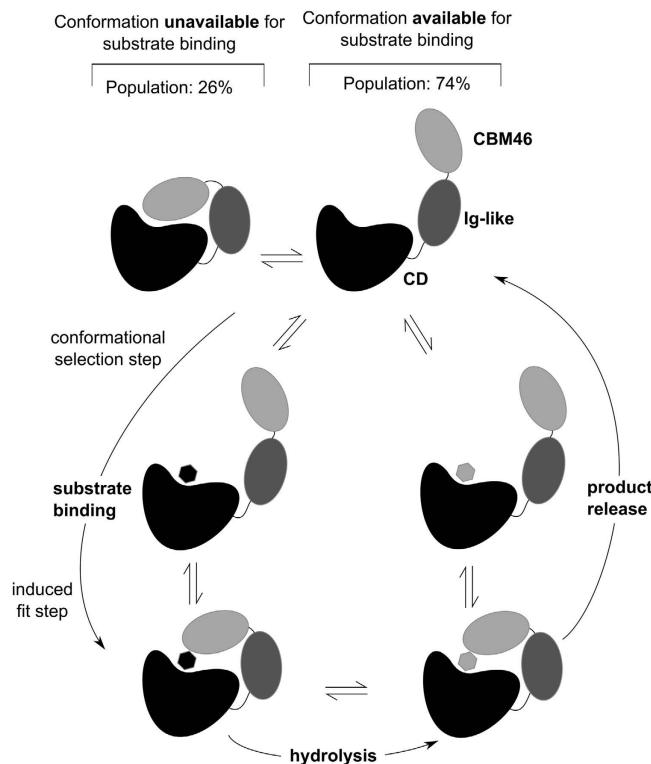
## Methods

**Cloning, Expression and Purification.** The gene encoding *B/Cel5B* (GenBank: AAU23417.1) was amplified from *Bacillus licheniformis* genomic DNA (ATCC 14580) without the predicted signal peptide sequence (nucleotides 1 to 81) using the primers *B/Cel5B\_Fw* and *B/Cel5B\_Rv* (Supplementary Table 3). The fragment was cloned into the expression vector pETTRXA-1a/LIC by ligation-independent cloning (LIC), as described elsewhere<sup>30</sup>.



**Figure 6. Comparison of the binding cleft of the *B/Cel5B* and *BhCel5B*.** The main difference between *B/Cel5B* and *BhCel5B* is that the latter exhibits a deeper cleft due to the presence of residue W181 in the loop between F177 and R185. We conjecture that this difference in the binding site architecture relates to the importance that the CBM46 plays in the *B/Cel5B* enzymatic mechanism.

The same method was used for construction of domain deletions. For Ig-like + CBM46 deletion,  $\Delta$ (Ig-CBM46), the fragment encoding the CD (nucleotides 82 to 1086) was amplified using the primers *B/Cel5B\_Fw* and *B/Cel5BΔ1087-1683\_Rv*. For CBM46 deletion,  $\Delta$ CBM46, the fragment encoding the CD + Ig-like (nucleotides 82 to 1377) was amplified using the primers *B/Cel5B\_Fw* and *B/Cel5BΔ1378-1683\_Rv* (Supplementary Table 3). Both fragments were cloned into pETTRXA-1a/LIC.



**Figure 7. Proposed molecular mechanism of *Bl*Cel5B conformational selection.** As suggested by the simulations and SAXS data, *Bl*Cel5B spans multiple conformations ranging from closed to extended CBM46/CD states. In a given open state, the substrate may reach the active site and become entrapped by the capping of CBM46 onto CD and induced-fit conformational adjustments. After hydrolysis, the reaction product is released to yield apo-*Bl*Cel5B, which becomes ready for a new cycle.

The wt protein *Bl*Cel5B, mutated proteins and AM deletions were expressed in *E. coli* Rosetta2 (DE3) strain. The cells were grown at 37 °C and 150 RPM in Luria Bertani Broth medium supplemented with 50 µg/mL kanamycin to an  $A_{600}$  of 1.5–2.0, after which the temperature was reduced to 20 °C and protein expression was induced with 1 mM IPTG for 6 h.

The extract was then loaded onto a NiNTA resin (Qiagen) equilibrated with a washing buffer (5 mM imidazole, 100 mM NaCl, 50 mM Tris-HCl, pH 8.0). Non-absorbed material was washed with ten times column volume with washing buffer and the purified protein was eluted with 200 mM imidazole, 100 mM NaCl, 50 mM Tris-HCl at pH = 7.0. His6 tag was removed by overnight digestion with TEV (Tobacco Etch Virus) at 4 °C, and untagged protein was purified by gel filtration through a HiLoad 16/60 Superdex 200 column in buffer containing 50 mM NaCl, 25 mM Tris-HCl at pH 7.0.

**Site-directed Mutagenesis.** The *Bl*Cel5B point mutations W479A and W481A were obtained by the inverse PCR method of site-directed mutagenesis<sup>31</sup>. Phusion® “High-Fidelity” DNA polymerase (NEB, USA) was used for amplifications with the plasmid pETTRXA-1a/LIC-*Blcel5B* as a template. Mutagenic primers *Blcel5BW479A\_Fw/Rv* and *Blcel5BW481A\_Fw/Rv* (Supplementary Table 3) were generated by HTP-OligoDesigner tool (<http://www.ifsc.usp.br/htpoligo/>).

**Activity Assays.** Enzymatic activity assays were performed by a colorimetric method using the 3,5-dinitrosalicylic acid (DNS)<sup>32</sup>, with glucose being a standard for the calibration curves. Assays of optimal temperature and pH were performed in triplicate with 1% medium-viscosity CMC as the substrate. For optimal temperature, the reaction mixture containing 10 µL of enzyme at 0.1 mg/mL, 50 µL of 1% (w/v) CMC and 40 µL of 50 mM sodium citrate buffer (pH 5.0) was incubated at 30 to 80 °C for 15 min and stopped by adding 100 µL of DNS solution. After this, the mixture was incubated again for 5 min at 100 °C and the absorbance was measured at 540 nm with a spectrophotometer. For optimal pH determination, the same amount of enzyme and substrate were diluted in 40 mM acetate/borate/phosphate buffer (ABF) with different pH values ranging from 2.0 to 10.0. The reactions were carried out under the predetermined optimal temperature.

The substrate specificity of the enzyme was determined using rye arabinan, xyloglucan, β-glucan, galactomannan, lichenan, β-mannan, Azo-Avicel and CMC as substrates. The substrates were diluted in water to 1% (w/v), and the reaction mixture was composed of 10 mL of purified enzyme at a concentration of 0.1 mg/mL, 0.4 mL of 50 mM sodium citrate buffer at pH 5.0, and 0.5 mL of 1% (w/v) substrate aqueous solution. The reaction was incubated at 50 °C for 15 min, followed by treatment with DNS as mentioned above. Enzyme unit was defined as the amount of enzyme that produces 1.0 µM of glucose in one minute for each substrate.

The kinetic parameters were determined by increasing concentrations of CMC. Reactions were performed in 50 mM sodium citrate buffer (pH = 4.0) at 50 °C, and measured by DNS method as well. Kinetic constants were determined by non-linear regression using OriginPro 8.0.

**Thermal Shift Assays.** The thermal denaturation assays were performed using a Real Time PCR Machine (Stratagene Mx3005P) as described by Dupeux and co-workers<sup>33</sup>. Briefly, the enzymes were diluted to 10 μM in 50 mM sodium citrate buffer (pH = 4.0) containing 1x SYPRO orange dye (Thermo Fisher Scientific). The fluorescence emission of the probe was monitored (excitation and emission at 492 and 516 nm, respectively) varying the temperature between 25 and 75 °C with the rate of 1 °C/min.

**Cellopentaose Cleavage Experiment.** The full-length *B/Cel5B* and AM deletion constructs were tested for product formation from cellopentaose. Cellopentaose (1.0 mM) was incubated with 25 μg of purified enzyme in 10 mM ammonium bicarbonate buffer (pH 7.0) in a total volume of 50 μL. The reaction was incubated for 90 min at 50 °C and then stopped by treatment at 100 °C for 5 min. After centrifugation for 10 min at 16,000 g the samples were subjected to MALDI/TOF-MS. Samples were supplemented with NaCl to a final concentration of 20 mM and 1 μL of the supernatant was co-crystallized with 1 μL 2,5-dihydroxybenzoic acid (10 mg/mL) in acetonitrile 30% and spotted on the target plate. The products were analyzed on Microflex LT MALDI-TOF (Bruker Daltonics) operating in positive ion mode. A single spectrum was obtained by averaging four independent spectra generated by 300 laser shots at 60% potency.

**Crystallization, Data Collection, and Structure Determination.** After purification, *B/Cel5B* was concentrated to 10 mg/mL for crystallization trials. Crystallization screens were set up using the sitting-drop vapor-diffusion method on a Cartesian PixSys 4200 (Genomic Solutions, United Kingdom) in a 96-well plate with drops formed by 100 nL protein solution plus 100 nL reservoir solution. The commercial kits Crystal Screen and Index (Hampton) were used as initial conditions. Crystals were grown at 18 °C between 3 and 7 days, and screened for diffraction.

Crystals were supplemented with cryoprotection solution, flash cooled in liquid nitrogen and diffraction data were collected at 100 K, at beamline ID23-1 (wavelength of 0.97 Å) from the European Synchrotron Radiation Facility (Grenoble, France). A crystal grown in condition containing 22.5% PEG 4000, 14% isopropanol and 0.1 M sodium citrate, pH 6.0, was selected to collect diffraction data to 1.7 Å resolution. The complex of the enzyme with substrate was obtained by crystal soaking with five times molar excess of cellopentaose for 24 hours. Diffraction data for the complexed enzyme were collected at 1.75 Å resolution.

Data were integrated with iMosflm<sup>34</sup> and scaled with Aimless<sup>35</sup>. The structure was solved by molecular replacement with Phaser<sup>36</sup> using an endoglucanase from *Clostridium cellulovoran* (PDB code: 3NDY) as the search model. Coot<sup>37</sup> was used for density fitting, and refinement was performed with PHENIX<sup>38</sup>.

**Atomistic simulations.** We took the *B/Cel5B* structure complexed with cellobiose as the starting configuration for the MD simulations. The missing residues were taken from the apo *B/Cel5B* structure after structural alignment using the LovoAlign server<sup>39</sup>. Hydrogen atoms were then added according to the protonation states determined at the optimum pH of 4.0 using the H + + server<sup>40</sup>. The following residues were considered protonated: H55, H77, D89, E96, E103, H114, E129, E159, E197, D198, E202, H205, E208, D211, H220, E245, E248, E260, H278, H292, D306, E312, E371, E375, E476, H416, E477, E489, D497, and E524. The remaining protonatable residues were considered in the standard protonation state. The *B/Cel5B*-cellobiose complex was then immersed in a rectangular simulation box of dimensions such that a solvent layer at least 16 Å thick surrounded the protein. The simulation box, built with Packmol<sup>41</sup>, also contained 0.10 M NaCl aqueous solution with excess counter ions to keep the system electrically neutral. The final system comprised approximately 85500 atoms.

The simulations were performed using NAMD<sup>42</sup> with the CHARMM force field and the TIP3P water model<sup>43–45</sup>. Periodic boundary conditions were employed, using particle mesh Ewald<sup>46</sup> to handle electrostatics and a 12-Å cutoff radius for truncating short-range potentials. Bonds involving hydrogen atoms were constrained at their equilibrium lengths and a time step of 2 fs was used to integrate the equations of motion. The simulations were carried out under constant pressure and temperature of 1 atm and 310 K, respectively, employing the Langevin barostat and thermostat<sup>42</sup>.

**Accelerated Molecular Dynamics.** In accelerated molecular dynamics<sup>20,47</sup>, the trajectory is propagated on a modified potential aimed to enhance conformational sampling. Whenever the potential energy drops below a given threshold *E*, a boost  $\Delta V(r)$  is applied, so that the escaping rates of local minima increase. When the potential energy gets over the threshold *E*, the system evolves on the original energy surface. This method has the advantage of conserving the general shape of the potential energy surface and of requiring no prior definition of reaction coordinates, so the system is allowed to explore freely its conformational space. Here, we restricted the energy boost only to the dihedral potential energy, as changes in torsion angles are the main source of conformational changes in proteins. The energy boost assumes the form of equation (1) that depends on the energy threshold *E* and on the parameter  $\alpha$  – which modulates the shape of the potential energy surface where the boost is applied. We set the parameters *E* and  $\alpha$  according to previous studies<sup>48</sup>, which recommend that *E* equals the average dihedral energy obtained from a conventional MD simulation plus 4 kcal/mol times the number of residues, and  $\alpha$  equals 0.8 kcal/mol times the number of residues. The average dihedral energy was 2275.5 kcal/mol and the *B/Cel5B* has 516 residues, so we set  $E = 2275.5 + 4 \times 516 = 4339.5$  kcal/mol and  $\alpha = 0.8 \times 516 = 418.8$  kcal/mol.

$$\Delta V(r) = \frac{(E - V(r))^2}{a + (E - V(r))} \quad (1)$$

**Simulation procedures.** *B/Cel5B-cellotetraose* – Having built the system, we carried out the following steps for equilibration: (i) 1000 steps of energy minimization followed by 100 ps of MD simulation with all non-solvent heavy atoms fixed; (ii) same as (i), but with only the  $\alpha$  carbons fixed; (iii) 5 ns of MD with all atoms free. After these preliminary steps, a trajectory lasting 400 ns was generated using conventional MD and then the aMD dihedral boost was applied for additional 1.0  $\mu$ s. After 100 ns of conventional MD, the cellotetraose dissociated and the simulation began to represent the dynamics of unbound state of *B/Cel5B*.

***B/Cel5B-cellooctaose*.** To build the cellooctaose chain, we extended the original cellotetraose chain in the crystal structure with 4 additional glucose residues spanning regions around the *B/Cel5B* positive subsites. Then, we submitted the system to the following procedure: (i) 1000 steps of energy minimization followed by 1 ns of MD keeping all the non-solvent heavy atoms fixed, except the 4 modeled glucose residues of the cellooctaose chain; (ii) same as step (i), but with only the  $\alpha$  carbons fixed; (iii) 1 ns of MD with only the non-modeled glucose residues fixed. We then performed a 200-ns-long MD with three harmonic potentials involving cellooctaose chain: first, between C3 atom (CHARMM atom names) of the second glucose residue from the cellooctaose non-reducing end and the CD2 atom of the W47 tryptophan residue; second, between the OH3 atom of the forth glucose residue from the cellooctaose non-reducing end and HE2 atom of the H113 histidine residue; and third, between the HE2 atom of the catalytic residue E159 and O4 glycosidic oxygen between the fourth and fifth glucose unit of the cellooctaose chain. After these preliminary relaxation steps, the harmonic potentials were removed and the trajectory was propagated by 400 ns using MD. To get a model of the *B/Cel5B-cellooctaose* complex in the closed conformation, we took the configuration after 80 ns of the restrained 200-ns MD simulation as the starting point for a 500-ns-long restrained aMD simulation, in which the CBM46 moved towards the CD in the presence of the harmonically-restrained cellooctaose chain. After this procedure, we released the restraints and propagated the closed *B/Cel5B-cellooctaose* complex for additional 500 ns of conventional, restraint-free MD simulation.

**Coarse-grained MD simulations.** The coarse-grained model was constructed from the minimized all-atom protein. We have used the domELNEDIN CG model for the protein. In this representation, an elastic network is used within each domain as a structural scaffold in order to maintain the overall shape of the protein, and a slightly modified version of MARTINI CG model describes the interactions involving beads not connected by harmonic springs<sup>49–51</sup>.

The delimitation of each domain was quite clear considering the short linkers connecting them and the recognition of their structural patterns in databases. We assumed CD, Ig-like module, and CBM46 as consisted of residues 18–331, 332–430, and 431–533, respectively. Therefore, there were elastic network bonds only within these domains (domELNEDIN CG model in Supplementary Fig. 6A).

The protonation state of each residue bead in the protein was the same adopted in the atomistic simulations. The system was then solvated by 10000 standard MARTINI CG water beads, including 10% of antifreeze particles. Also, 58 chloride and 48 sodium ions were added for charge neutrality. The size of final system was 109  $\text{\AA}$   $\times$  109  $\text{\AA}$   $\times$  109  $\text{\AA}$ .

Preliminary simulations were performed to test the elastic network (EN) parameters. We have tested six different ENs in 100 ns of simulations, using combinations of cut-off distance ( $R_c$ ) of 8  $\text{\AA}$  and 9  $\text{\AA}$  with spring force constant ( $k_s$ ) of 500, 800 and 1000  $\text{kJ mol}^{-1} \text{nm}^{-2}$ . The time evolution of root mean square deviation relative to the crystal structure as well as the mobility profile of the protein in these simulations were compared to the correspondent data from a 100 ns atomistic simulation. From this procedure, the parameters  $R_c = 9 \text{\AA}$  and  $k_s = 500 \text{kJ mol}^{-1} \text{nm}^{-2}$  resulted in the best match between atomistic and coarse-grained simulations (Supplementary Fig. 6B).

The coarse-grained simulations were carried out using GROMACS<sup>52</sup>. Periodic boundary conditions were employed. Van der Waals interactions were shifted to zero in the range 0.9–1.2 nm, and the electrostatic interactions, in the range 0.0–1.2 nm. The simulations were performed in the isothermal-isobaric ensemble (NpT), employing the Berendsen thermostat and barostat for temperature and pressure control, respectively, with time constants  $\tau_T = 0.5 \text{ ps}$  and  $\tau_p = 1.2 \text{ ps}$ .

The CG simulations were carried out using the following protocol: the system was first minimized for 1000 steps using the steepest descent method. Then, it was submitted to a relaxation procedure comprising gradual increasing in time step or temperature. In the first stage of relaxation, the protein beads were restrained with a 1000  $\text{kJ mol}^{-1} \text{nm}^{-2}$  force constant and a 50 ps simulation was carried out at 50 K, using the short time step of 1 fs. In the second stage, the time step was increased up to 5 ps lasting 1000 ps of simulation time. In the last stage of relaxation, all the system is released to move and it underwent a gradual increase in temperature, consisting on five segments of 100 ps at 50, 100, 150, 200 and 310 K. After achieving the desired temperature of 310 K, we performed three production simulations using 20-fs timestep. We have used a random number generator for assigning velocities to generate three independent simulations.

In general, smoothing of the energy surface in CG model makes the time scales faster. A speed up factor of 4 is typically employed to rescale the time scale of MARTINI CG systems<sup>51</sup>. Therefore, all CG simulations times described here and in the main text are effective times, i.e.,  $\times 4$  simulation time.

**Small Angle X-ray Scattering.** SAXS data were collected at the SAXS2 beamline of the Brazilian Synchrotron Light Laboratory-LNLS (Campinas, Brazil) on a bi-dimensional position sensitive CCD detector (MarResearch, USA) using the radiation wavelength 1.54  $\text{\AA}$ . The sample-detector distance of 1000 mm allowed covering the momentum transfer range  $0.01 \text{\AA}^{-1} < q < 0.35 \text{\AA}^{-1}$  ( $q = 4\pi \sin \theta / \lambda$ , where  $2\theta$  is the scattering angle).

The protein samples were prepared in McIlvaine's buffer at 50 mM, pH 5 and 20 °C<sup>53</sup>. In each measurement, two successive frames of 300 s were recorded for each sample at 1 and 2 mg/mL to monitor radiation damage. The patterns were integrated using the FIT2D program<sup>54</sup>. The comparative analysis for each scattering curve at 1 and 2 mg/mL of *B/Cel5B* (data not shown), as well as the radius of gyration values ( $R_g$ ), indicated that

concentration and aggregation effects did not exist. The linearity of the Guinier plot indicated that the preparation was monodisperse.

The radius of gyration of the molecules ( $R_g$ ) was estimated by two methods, using the Guinier equation  $I(q) = I(0) \exp[-(q^2 R_g^2)/3]$ ,  $q \cdot R_g < 1.3$  and also with the inverse Fourier transform in GNOM<sup>55</sup>. The same program was used to obtain the distance distribution function  $P(r)$  and the maximum diameter  $D_{\max}$ . Ten independent dummy atom models (DAMs) were restored by the *ab initio* proceeding implemented in DAMMIN package<sup>56</sup>. The best model, selected using normalized spatial discrepancy parameter computed by DAMAVER program, was superimposed on the crystallographic model with the SUPCOMB.

Then, based on the enzyme conformations reported by MD and protein crystallography, the computed X-ray scattering profile was fitted to a given experimental SAXS data by minimizing the  $\chi$  function in the FOXS program<sup>57</sup>.

To assess the inter-domain information, the contribution of individual conformer and the flexibility of *B/Cel5B*, we proceed in two approaches. First, the theoretical profiles and experimental data comparison was performed to infer the best-fit conformation of the ensemble-based analysis by the ensemble optimization method – EOM<sup>21</sup>, which assumes coexistence of a range of conformations in solution for which an average scattering intensity fits the experimental SAXS data; all models were generated with the three individual domains (Ig-like, CBM46, and CD) free to randomly move in order to cover the entire conformational space. The second approach was based on a fractional volume calculation from three conformation members extracted from the MD simulations, each with a distinct scatter curve. OLIGOMER<sup>58</sup> provided solution of a system of linear equations between the experimental and generated conformations by MD.

The simulated scattering curves from the MD and crystallographic models were obtained using the CRYSTAL<sup>59</sup>.

**Phylogenetic assignment.** Sequences for all GH5 members, in which only the catalytic domain were considered, were downloaded from PFAM database<sup>24</sup> and their classification into subfamilies were obtained within the CAZy database<sup>23</sup>. The sequences belonging to subfamily 4 were selected and those that had over 90% identity or represented partial coverage were rejected. Based on their multiple sequence alignment, the phylogenetic tree was constructed using the maximum likelihood method implemented in the MEGA program version 6.0<sup>60</sup>. One hundred Bootstrap replications were performed to examine the reliability of the phylogenetic tree.

## References

- Chundawat, S. P. S., Beckham, G. T., Himmel, M. E. & Dale, B. E. Deconstruction of lignocellulosic biomass to fuels and chemicals. *Annu. Rev. Chem. Biomol. Eng.* **2**, 121–45 (2011).
- Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J.* **382**, 769–81 (2004).
- Kern, M. *et al.* Structural characterization of a unique marine animal family 7 cellobiohydrolase suggests a mechanism of cellulase salt tolerance. *Proc. Natl. Acad. Sci. USA* **110**, 10189–94 (2013).
- Prates, É. T. *et al.* X-ray structure and molecular dynamics simulations of endoglucanase 3 from *Trichoderma harzianum*: structural organization and substrate recognition by endoglucanases that lack cellulose binding module. *PLoS One* **8**, e59069 (2013).
- Sakon, J., Irwin, D., Wilson, D. B. & Karplus, P. A. Structure and mechanism of endo/exocellulase E4 from *Thermomonospora fusca*. *Nat. Struct. Biol.* **4**, 810–8 (1997).
- van Aalten, D. M. *et al.* Structural insights into the catalytic mechanism of a family 18 exo-chitinase. *Proc. Natl. Acad. Sci. USA* **98**, 8979–84 (2001).
- Meekins, D. A. *et al.* Phosphoglucan-bound structure of starch phosphatase Starch Excess4 reveals the mechanism for C6 specificity. *Proc. Natl. Acad. Sci. USA* **111**, 7272–7 (2014).
- Venditto, I. *et al.* Family 46 Carbohydrate-binding Modules Contribute to the Enzymatic Hydrolysis of Xyloglucan and  $\beta$ -1,3-1,4-Glucans through Distinct Mechanisms. *J. Biol. Chem.* **290**, 10572–86 (2015).
- Lammerts van Bueren, A. *et al.* The conformation and function of a multimodular glycogen-degrading pneumococcal virulence factor. *Structure* **19**, 640–51 (2011).
- Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. USA* **44**, 98–104 (1958).
- Bosshard, H. R. Molecular recognition by induced fit: how fit is the concept? *News Physiol. Sci.* **16**, 171–3 (2001).
- Sullivan, S. M. & Holyoak, T. Enzymes with lid-gated active sites must operate by an induced fit mechanism instead of conformational selection. *Proc. Natl. Acad. Sci. USA* **105**, 13829–34 (2008).
- Hammes, G. G., Chang, Y.-C. & Oas, T. G. Conformational selection or induced fit: a flux description of reaction mechanism. *Proc. Natl. Acad. Sci. USA* **106**, 13737–41 (2009).
- Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35**, 539–46 (2010).
- Domínguez, R., Souchon, H., Lascombe, M. & Alzari, P. M. The crystal structure of a family 5 endoglucanase mutant in complexed and uncomplexed forms reveals an induced fit activation mechanism. *J. Mol. Biol.* **257**, 1042–51 (1996).
- Varrot, A., Schülein, M. & Davies, G. J. Insights into ligand-induced conformational change in Cel5A from *Bacillus agaradhaerens* revealed by a catalytically active crystal form. *J. Mol. Biol.* **297**, 819–28 (2000).
- Ohtaki, A., Mizuno, M., Tonozuka, T., Sakano, Y. & Kamitori, S. Complex structures of *Thermoactinomyces vulgaris* R-47 alpha-amylase 2 with acarbose and cyclodextrins demonstrate the multiple substrate recognition mechanism. *J. Biol. Chem.* **279**, 31033–40 (2004).
- Holm, L. & Rosenström, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, W545–9 (2010).
- Naas, A. E., MacKenzie, A. K., Dalhus, B., Eijsink, V. G. H. & Pope, P. B. Structural Features of a Bacteroidetes-Affiliated Cellulase Linked with a Polysaccharide Utilization Locus. *Sci. Rep.* **5**, 11666 (2015).
- Hamelberg, D., Mongan, J. & McCammon, J. A. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120**, 11919–29 (2004).
- Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* **129**, 5656–64 (2007).
- Rambo, R. P. & Tainer, J. A. Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers* **95**, 559–71 (2011).
- Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–5 (2014).
- Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–30 (2014).
- Ducros, V. *et al.* Crystal structure of the catalytic domain of a bacterial cellulase belonging to family 5. *Structure* **3**, 939–49 (1995).

26. Larsbrink, J. *et al.* A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature* **506**, 498–502 (2014).
27. Gloster, T. M. *et al.* Characterization and three-dimensional structures of two distinct bacterial xyloglucanases from families GH5 and GH12. *J. Biol. Chem.* **282**, 19177–89 (2007).
28. Tseng, C. W. *et al.* Substrate binding of a GH5 endoglucanase from the ruminal fungus *Piromyces rhizinflata*. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.* **67**, 1189–94 (2011).
29. Mosbath, A. *et al.* Solution structure of the module X2 1 of unknown function of the cellulosomal scaffolding protein CipC of *Clostridium cellulolyticum*. *J. Mol. Biol.* **304**, 201–17 (2000).
30. Camilo, C. M. & Polikarpov, I. High-throughput cloning, expression and purification of glycoside hydrolases using Ligation-Independent Cloning (LIC). *Protein Expr. Purif.* **99**, 35–42 (2014).
31. Rabhi, I. *et al.* A novel simple and rapid PCR-based site-directed mutagenesis method. *Mol. Biotechnol.* **26**, 27–34 (2004).
32. Miller, G. L. Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar. *Anal. Chem.* **31**, 426–428 (1959).
33. Dupeux, F., Röwer, M., Seroul, G., Blot, D. & Márquez, J. A. A thermal stability assay can help to estimate the crystallization likelihood of biological samples. *Acta Crystallogr. D. Biol. Crystallogr.* **67**, 915–9 (2011).
34. Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. W. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D. Biol. Crystallogr.* **67**, 271–81 (2011).
35. Evans, P. R. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr. D. Biol. Crystallogr.* **67**, 282–92 (2011).
36. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
37. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D. Biol. Crystallogr.* **66**, 486–501 (2010).
38. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D. Biol. Crystallogr.* **66**, 213–21 (2010).
39. Martínez, L., Andreani, R. & Martínez, J. M. Convergent algorithms for protein structural alignment. *BMC Bioinformatics* **8**, 306 (2007).
40. Gordon, J. C. *et al.* H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* **33**, W368–71 (2005).
41. Martínez, L., Andrade, R., Birgin, E. G. & Martínez, J. M. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **30**, 2157–64 (2009).
42. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–802 (2005).
43. MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–616 (1998).
44. Guvench, O., Hatcher, E. R., Venable, R. M., Pastor, R. W. & MacKerell, A. D. CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *J. Chem. Theory Comput.* **5**, 2353–2370 (2009).
45. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926 (1983).
46. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089 (1993).
47. Wang, Y., Harrison, C. B., Schulten, K. & McCammon, J. A. Implementation of Accelerated Molecular Dynamics in NAMD. *Comput. Sci. Discov.* **4**, 015002 (2011).
48. Pierce, L. C. T., Salomon-Ferrer, R., Augusto F de Oliveira, C., McCammon, J. A. & Walker, R. C. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* **8**, 2997–3002 (2012).
49. Monticelli, L. *et al.* The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **4**, 819–834 (2008).
50. Periole, X., Cavalli, M., Marrink, S.-J. & Ceruso, M. A. Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *J. Chem. Theory Comput.* **5**, 2531–2543 (2009).
51. Siuda, I. & Thøgersen, L. Conformational flexibility of the leucine binding protein examined by protein domain coarse-grained molecular dynamics. *J. Mol. Model.* **19**, 4931–45 (2013).
52. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **4**, 435–447 (2008).
53. McIlvaine, T. A buffer solution for colorimetric comparison. *J. Biol. Chem.* **49**, 183–186 (1921).
54. Hammersley, A. FIT2D: An Introduction and Overview. *ESRF Internal Report* ESRF97HA02T (1997).
55. Svergun, D. I. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.* **25**, 495–503 (1992).
56. Svergun, D. I. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* **76**, 2879–86 (1999).
57. Schneidman-Duhovny, D., Hammel, M. & Sali, A. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* **38**, W540–4 (2010).
58. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J. & Svergun, D. I. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *J. Appl. Crystallogr.* **36**, 1277–1282 (2003).
59. Svergun, D., Barberato, C. & Koch, M. H. J. CRYSTOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Crystallogr.* **28**, 768–773 (1995).
60. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–9 (2013).

## Acknowledgements

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grants 2008/56255-9, 2009/52840-7, 2010/18773-8, 2013/08293-7, 2013/15582-5 and 2014/10448-1, and by CNPq (grants #490022/2009-0, 301981/2011-6, 500091/2014-5 and 310177/2011-1). The authors acknowledge LNLS and ESRF for support with synchrotron data collection.

## Author Contributions

M.V.L. and I.P. designed the experiments; C.M.C. performed the gene mutations and cloning; M.V.L. and V.O.A.P. expressed and purified the enzymes and characterized the enzymatic activities; M.V.L. crystallized and determined the crystal structures, with data collection supervised by A.P.; M.A.K. performed mass spectrometry experiment; E.A.A. and M.O.N. collected and treated SAXS data. R.L.S., E.T.P. and M.S.S. designed the computer simulations; R.L.S. performed Accelerated Molecular Dynamics; E.T.P. performed Coarse-grained MD Simulations; M.V.L., R.L.S., M.S.S. and I.P. wrote the manuscript with the input from all the other authors; M.S.S. and I.P. supervised the project.

### Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Liberato, M. V. *et al.* Molecular characterization of a family 5 glycoside hydrolase suggests an induced-fit enzymatic mechanism. *Sci. Rep.* **6**, 23473; doi: 10.1038/srep23473 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

## Supplementary Information

### Molecular characterization of a family 5 glycoside hydrolase suggests an induced-fit enzymatic mechanism

**Marcelo V. Liberato<sup>a,#</sup>, Rodrigo L. Silveira<sup>b,#</sup>, Érica T. Prates<sup>b</sup>, Evandro A. de Araujo<sup>a</sup>,  
Vanessa O. A. Pellegrini<sup>a</sup>, Cesar M. Camilo<sup>a</sup>, Marco A. Kadowaki<sup>a</sup>, Mario de O. Neto<sup>c</sup>,  
Alexander Popov<sup>d</sup>, Munir S. Skaf<sup>b,\*</sup> and Igor Polikarpov<sup>a,\*</sup>**

<sup>a</sup> São Carlos Institute of Physics, University of São Paulo, São Carlos 13566-590 São Paulo, Brazil

<sup>b</sup> Institute of Chemistry, University of Campinas, Campinas 13083-970, São Paulo, Brazil

<sup>c</sup> Institute of Bioscience, University of São Paulo State, Botucatu 18618-970, São Paulo, Brazil

<sup>d</sup> European Synchrotron Radiation Facility, Grenoble CS40220, France

<sup>#</sup> These authors contributed equally.

\* Corresponding authors:

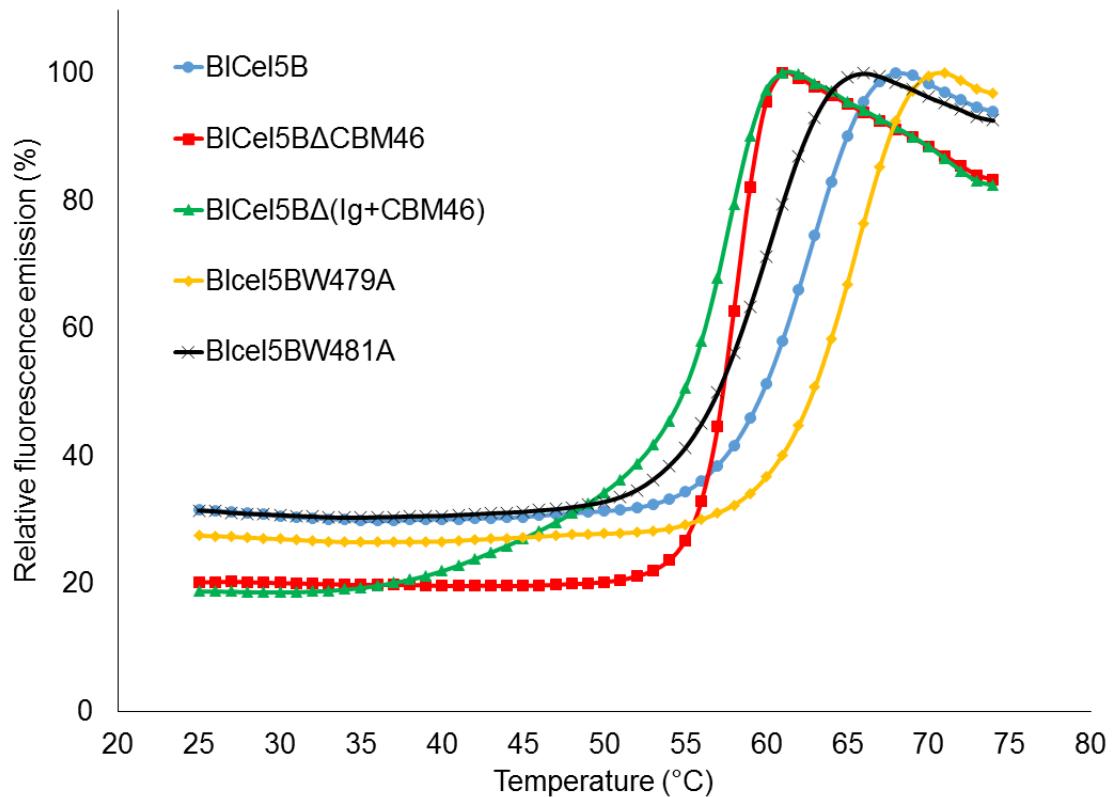
Munir S. Skaf, Institute of Chemistry, University of Campinas, Cx. P. 6154 Campinas, SP, 13084-862, Brazil

Phone: +55(19)3521-3093; E-mail: [skaf@iqm.unicamp.br](mailto:skaf@iqm.unicamp.br) and

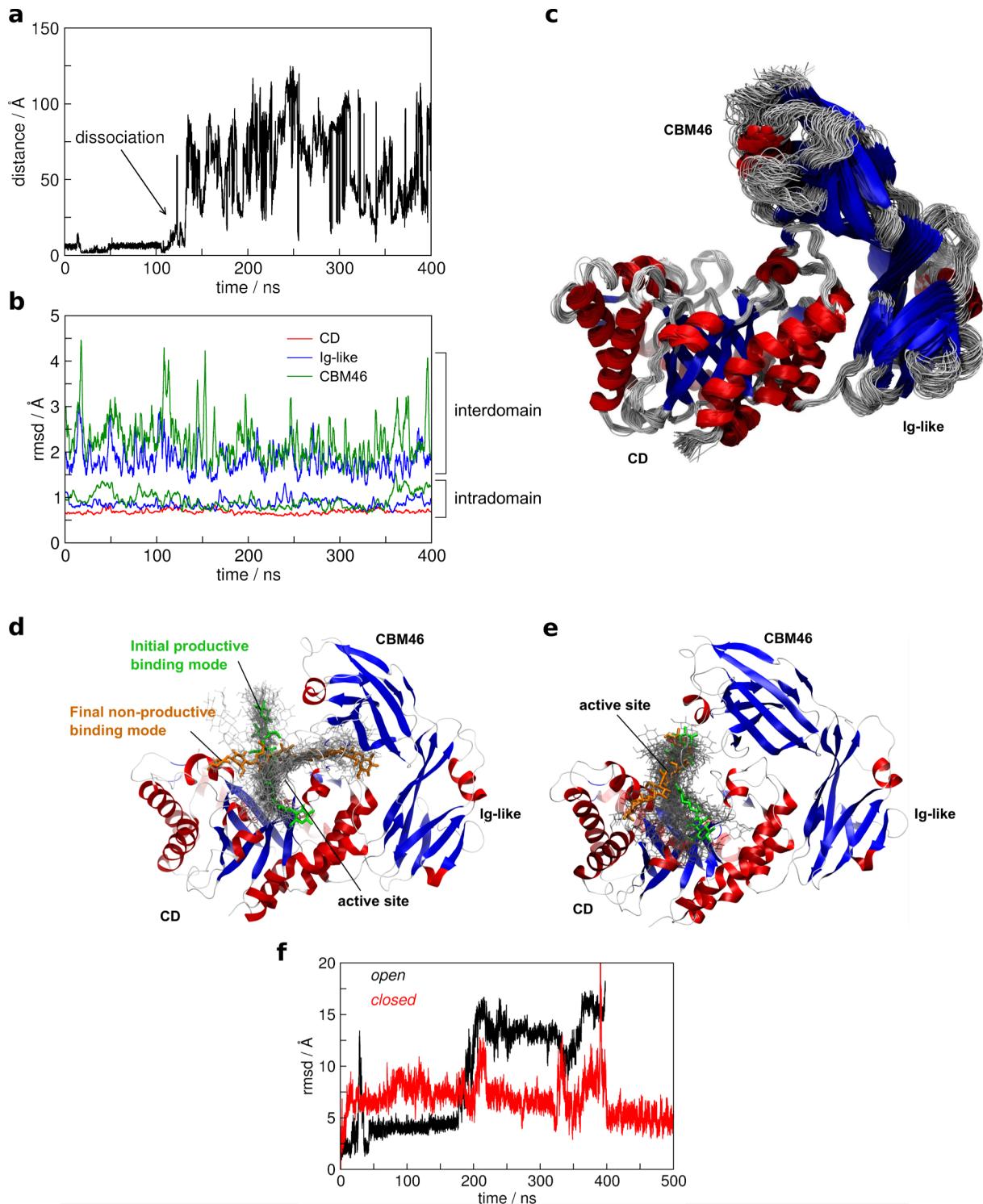
Igor Polikarpov, São Carlos Institute of Physics, University of São Paulo, Av. Trabalhador São-Carlense, 400 Pq. Arnold Schimidt, São Carlos, SP, 13566-590, Brazil

Phone: +55(16) 3373-8088; E-mail: [ipolikarpov@ifsc.usp.br](mailto:ipolikarpov@ifsc.usp.br)

## SUPPLEMENTARY FIGURES

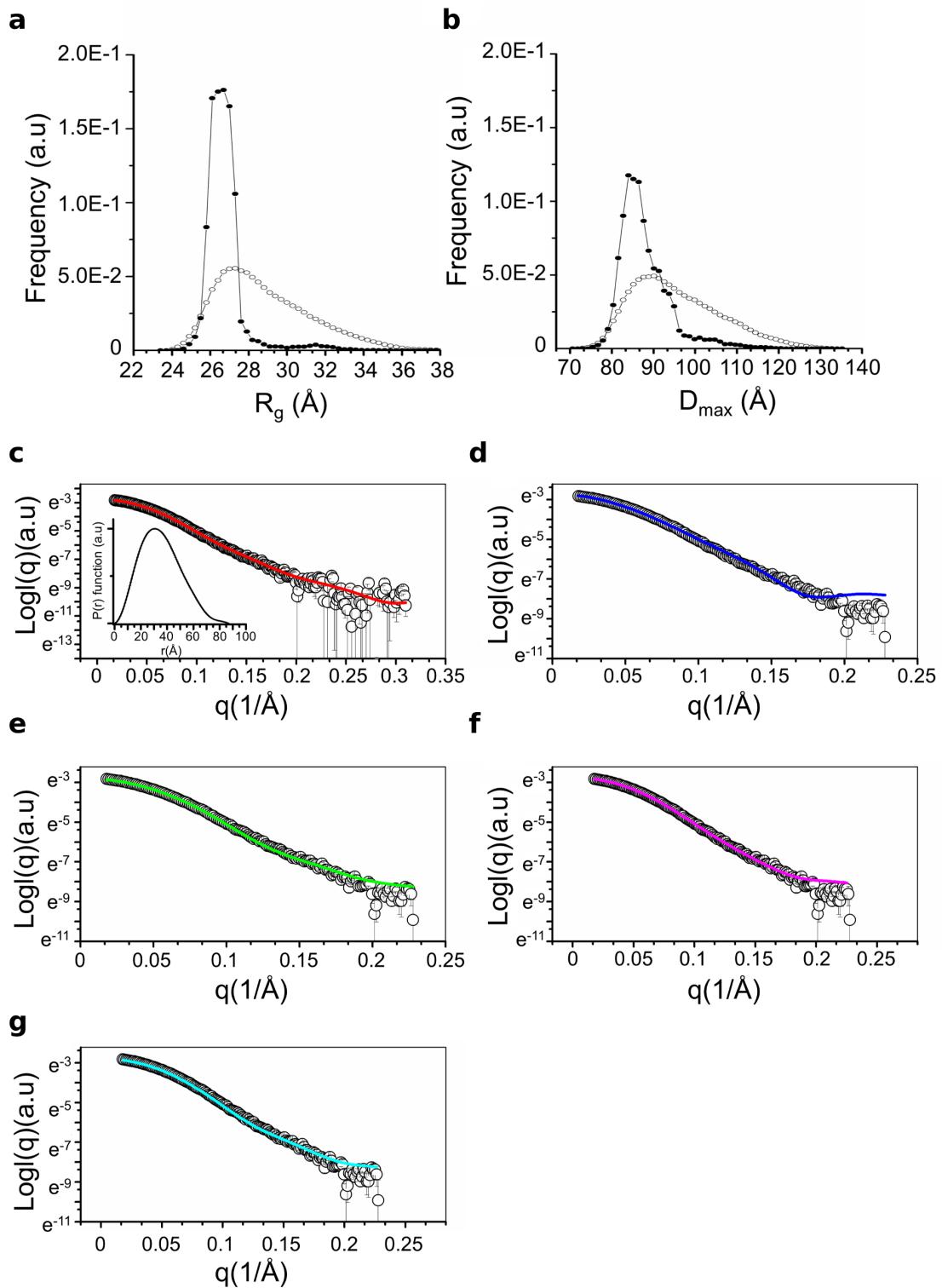


**Supplementary Fig. 1 | Effect of deletions and point mutations on *B/Cel5B* thermostability.** Thermal Shift Assay was applied to evaluate possible enzyme destabilization caused by deletions and site-directed point mutations. All of the constructs have melting temperature higher than 55 °C, which is the optimum temperature for *B/Cel5B* activity.



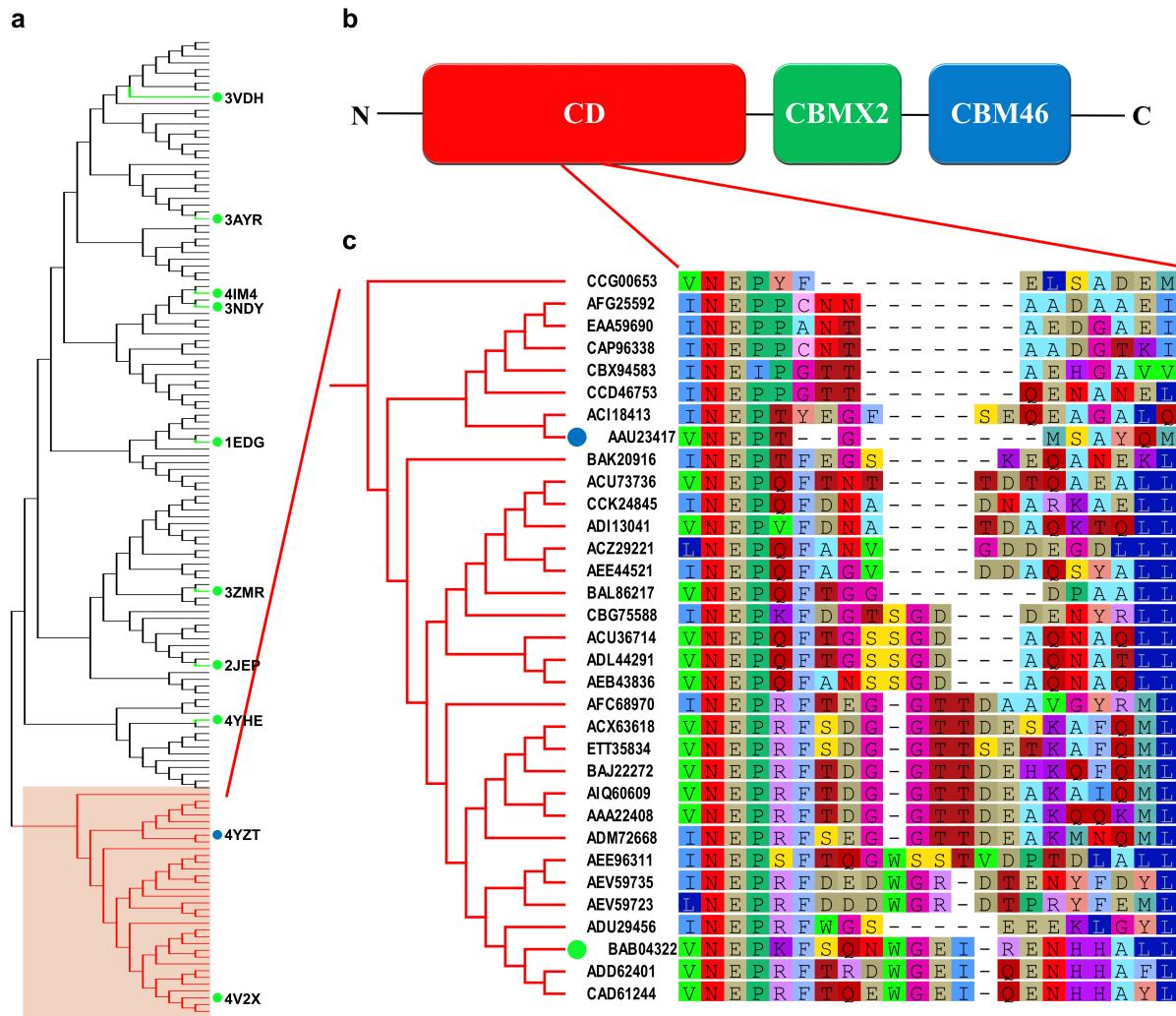
**Supplementary Fig. 2 | Molecular dynamics of *B/Cel5B*.** **(a)** Rmsd of the cellotetraose chain during a conventional molecular dynamics simulations strarting from its configuration in the *B/Cel5B* crystal structure. The rmsd – computed after alignment of the *B/Cel5B* CD – fluctuates around low and constant values until  $\sim 100$  ns, indicating that the cellotetraose remains bound to the enzyme during this time. After 100 ns, the rmsd increases, indicating that the substrate left the enzyme and started exploring the bulk. **(b)**  $\alpha$  carbon rmsd relative to the average structure of the domains CD, Ig-like module and CBM46. The intradomain

rmsd's were computed after structural alignment of the same domain for which the rmsd was computed. As these intradomain rmsd's fluctuates around 1 Å, the *B/Cel5B* domains do not exhibit conformational changes. The interdomain rmsd's were computed for each AM after structural alignment of the CD  $\alpha$  carbons. The higher values of interdomain rmsd's compared to the intradomain rmsd indicates that the *B/Cel5B* is more prone to exhibit large-scale interdomains than internal conformational changes. Running averages were performed on time windows of 2 ns for smoothing. **(c)** Dynamic picture of the *B/Cel5B*, showing superposed configurations of the enzyme taken every 5 ns from a 400-ns MD simulation. Before the superposition, the CD was aligned so that the interdomain motions between the AMs became evident. As the substrate dissociated in 100 ns, this picture is representative of the substrate-free *B/Cel5B*. **(d)** Dynamic picture of a cellobiose chain (in gray sticks) in the open, crystallographic-like configuration of *B/Cel5B*, showing that the substrate, initially in a productive binding mode (in green), assumes a non-productive binding mode (in orange) by the end of a 400-ns MD simulation, which is nearly perpendicular to the initial configuration and away from the active site. **(e)** Same as **(d)**, but after the *B/Cel5B* transited to the closed configuration through AMD. In the closed configuration, the substrate is confined in the space between the flat binding surface of the CD and CBM46 lying over it. **(f)** Rmsd of the cellobiose chain from its initial configuration in the simulations, showing that the closed *B/Cel5B* configuration reduce the substrate freedom to move away from the initial, productive binding mode.



**Supplementary Fig. 3 | Conformational distribution and *B/Cel5B* SAXS data.** Given the significant conformational flexibility of *B/Cel5B*, we used EOM to analyze the distribution of different conformers of the protein. The  $R_g$  (a) and  $D_{max}$  (b) distributions reflect the overall size and shape properties of the molecules in solution. An initial ensemble (10,000 randomized models) of *B/Cel5B* (open circles) with different conformations were generated and the scattering profiles of these ensemble were computed using EOM. The genetic

algorithm employed in EOM was used to select more representative structures (black circles) from the initial pool such that the average calculated scattering from the selected set of more representative conformations fits best the SAXS data. The  $R_g$  and  $D_{max}$  distributions reflect the relative populations of inter-domain conformational distributions (from extended to compact configuration) of *B/Cel5B* in solution. **(c)** Experimental SAXS intensity is given as black open circles with errors bars and simulated SAXS scattering computed from DAM model is shown as red line. Inset, the distance distribution functions  $p(r)$  for *B/Cel5B*. *B/Cel5B* experimental SAXS data fitted with the theoretical scattering profile of the atomic structure from coarse-grained MD models: **(d)** open model scattering is given as a blue line; **(e)** closed model scattering is shown as a green line; **(f)** intermediate model scattering is presented as a pink line; **(g)** average model computed with OLIGOMER is shown as a cyan line.

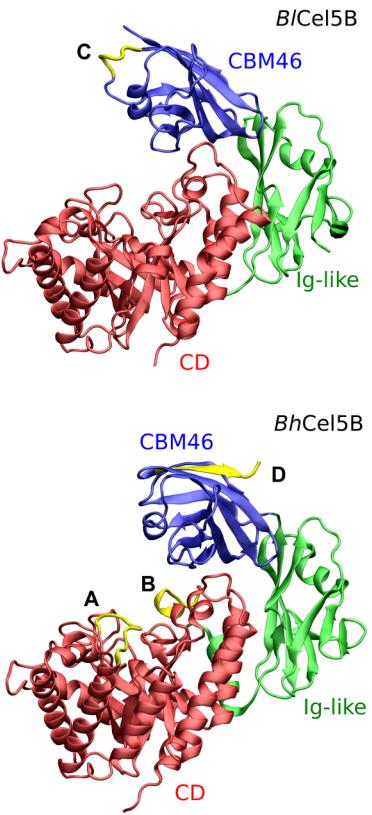


**Supplementary Fig. 4 | Phylogeny of GH5 subfamily 4 (GH5\_4).** (a) The phylogenetic tree was constructed using all GH5 sequences assigned in CAZY to the subfamily 4, excluding the partial sequences and those that share identity greater than 90%. The members with the known crystallographic structure are highlighted with a green sphere followed by the PDB id code. The blue sphere represents *B/Cel5B*. (b) Architecture representation of several GH5\_4 enzymes, including *B/Cel5B*. (c) Highlighted sub-tree composed by all GH5\_4 members with the tri-modular architecture described in (b). Each member is represented by its Genbank number. Alignment of the sequences evidencing the length of the loop between the residues P160 and M163 (with *B/Cel5B* as a reference) is given at the right-hand side. First sub-group includes *B/Cel5B* (blue sphere) and have a small loop, the second sub-group have an intermediary loop, and the third subgroup includes *BhCel5B* (green sphere) and have a long insertion loop.

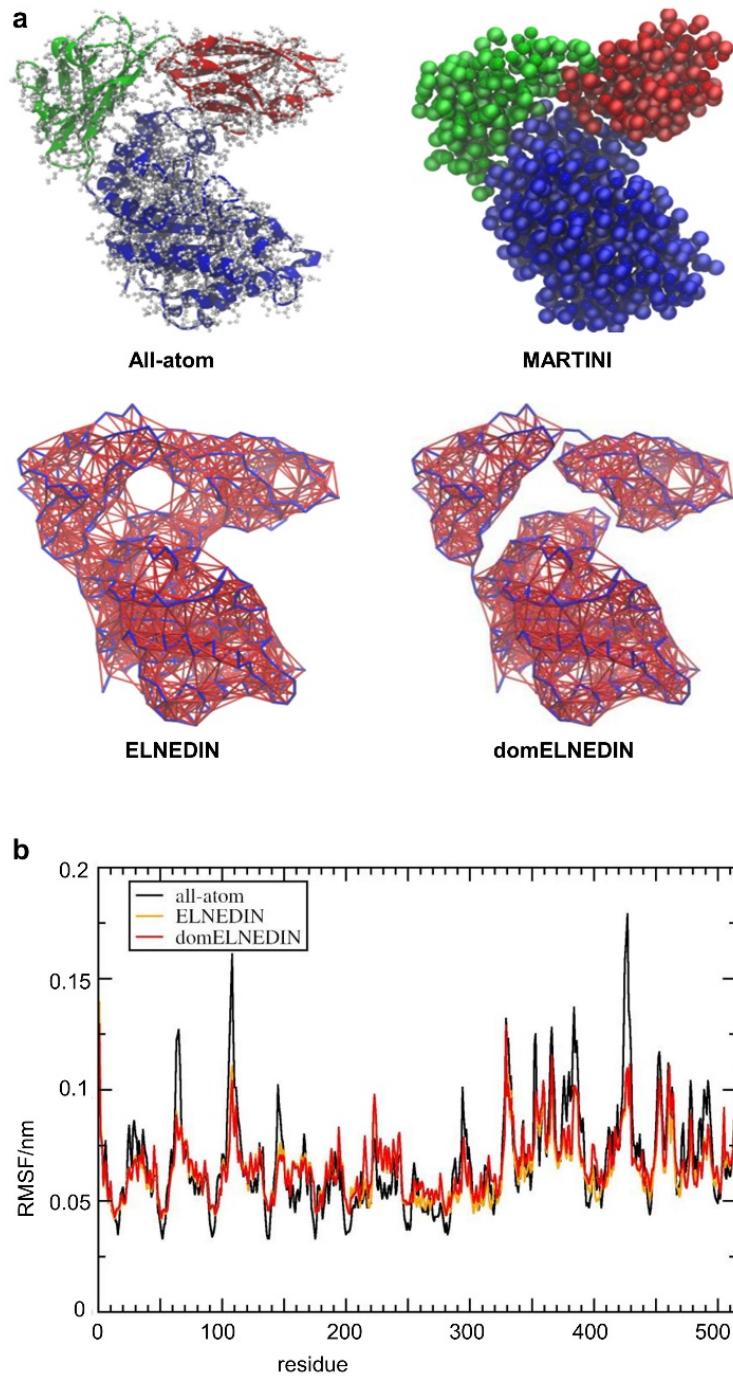
**a**

<i>B/Cel5B</i>	18	-IITPADLLENMSPGWNLGNTLDAVP-TEGSWNPPVREHTFDDIRDAG	CD
<i>BhCel5B</i>	31	KTLDIQ-SYVRDMQPGWNLGNTFDAVGQDETAWGNPRVTRELIEQIADEG	
	65	FKSVRIPVTWDSHIGSAPEYPIDDTTDWMRNRVEEVTDWALEREFYVVLNI	
	80	YKSIRIPVTWENRIGGAPDPI-DP-QFLNRVDQVQWALEEDLYVMNL	
	113	HHDSWLWISRMGSQQETLDDKLKGKVWQIAERFKNKSERLLFEIIVNEP	
	128	HHDSWLWIVYMEHNYNGVMV-KYRSLWEQLSNHFKDYPPTKLMFES-VNEP	
	161	<b>A</b> TGM-----SAYQQMNLNREMLNIIRSTGGNGQRLVIVGGLEDNKD	
	176	KFSQNWGEIRENHHALLDDLNTVFEIIRQSGGQNDIRPLVLPTMETATS	
	202	ELLHSF--EP--PDDDRIVLTYHYYSPWDYVSNWWGRTTWGSAAEISSE	
	226	QPFLNNLYQTIDKLDDPNLIIATVHYYGWPFSVNIAGYTRFEE-NSK-QE	
	246	MEEDIKPVYKEFVREGYPIIGEYGTG-----ANEKHSKWLYHDTFVR	
	274	IIEAFDRVHHTFVARGIPVVLGEFGLLGFDKHTGVIQQGEKLKFFELIH	
	290	LAHKYQMVPMMWDNGNDQFDRAERQWRDPVVKEEIVIQAQGRGVPNAAIKP	Ig-like
	324	HLNERDITHMLWDNGQ-HFNRHTYEWDQELFDM-MRASWE-GRSSVAES	
	339	ADLFIKKGQSISDQTVDIQLNGNVLTGIYQKSEPLKEGSDYTVDAGKTVS	
	371	NFIYLKQGDRRIADATVSLQHGNELTGLRANGQLTPGQDYELNG-ERLT	
	390	IKASCLAKLLNGAGQPGVKAQLTFTFHKGASQVMDDIIYDPPKLEKSEFT	CBM46
	420	VKAHVLSAIASS-GTGLTNGMVTAEFNRGADWHFRVNTYRTPVL-QSTQG	
	440	<b>C</b> ISQSAISGDLKIPASLNGTKLATVKGVVWDSTGRPVLEEVWSWTPYLNYDE	
	468	H-----V-SNFSIPASFNGMSLATMEAVYY-DGGNAG--PQDWTSFKEFGY	
	490	DFYEKDG--DLYLKERVLKYLKSD-STFTFELWPKGVEAVVKVKITP---	
	510	AFSPSYDANEMLKTEAFFREVVRDGEVRLTFHFWSGETVNYTIIKNGNQVT	
	533		D
	560	<b>D</b> GIAAQ	

Sequence identity: 31%

**b**

**Supplementary Fig. 5 | Comparison between *B/Cel5B* and *BhCel5B*.** **(a)** Sequence alignment, with insertions/deletions highlighted and labeled as A, B, C and D. **(b)** Structure of *B/Cel5B* and *BhCel5B*, with the insertions highlighted in yellow. The insertion A is a loop that shapes the *BhCel5B* binding cleft. The insertion B is located below the CBM46 in *BhCel5B*, and might be important for the relative mobility of CD and CBM46 in *BhCel5B* structure. The insertion C is located in the *B/Cel5B* CBM46 and does not face the binding cleft. The insertion D is a C-terminal prolongation of the *BhCel5B* CBM46 and does not face the binding cleft. All other regions of *B/Cel5B* and *BhCel5B* from the structural viewpoint are nearly identical.



**Supplementary Fig. 6 | Tested coarse-grained protein models.** **(a)** Different representations of *Bl*Cel5B: atomistic, MARTINI coarse-grained, elastic networks (ELNEDIN) and intradomain elastic networks (domELNEDIN). In this study, we have performed simulations using all-atom (atomistic) and domELNEDIN coarse-grained models. **(b)** Comparison between the mobility profiles of *Bl*Cel5B in 100-ns all-atom MD simulation and the MARTINI-elastic networks simulations (ELNEDIN and domELNEDIN).

## SUPPLEMENTARY TABLES

**Supplementary Table 1 | Statistics for crystallographic data collection and refinement.**

	<i>B</i> /Cel5B+cellotetraose	<i>apo</i> <i>B</i> /Cel5B
<b>Data collection</b>		
Space group	<i>P</i> 4 <sub>3</sub> 2 <sub>1</sub> 2	<i>P</i> 4 <sub>3</sub> 2 <sub>1</sub> 2
Cell dimensions <i>a</i> , <i>b</i> , <i>c</i> (Å)	91.3, 91.3, 124.7	90.7, 90.7, 120.6
Resolution (Å)	73.66-1.66 (1.69-1.66) *	64.13-1.7 (1.73-1.7)
<i>R</i> <sub>merge</sub>	0.146 (0.802)	0.078 (0.552)
<i>I</i> / <i>σI</i>	6.9 (2.2)	7.1 (1.7)
Completeness (%)	99.8 (99.9)	94.1 (95.9)
Redundancy	5.0 (5.2)	2.5 (2.5)
<b>Refinement</b>		
Resolution (Å)	57.33-1.67	64.13-1.7
# of reflections	62078	52066
<i>R</i> <sub>work</sub> / <i>R</i> <sub>free</sub> ,	0.175 / 0.212	0.166 / 0.198
# of atoms		
Protein	4108	4040
Ligand/ion	53	0
Water	738	593
<i>B</i> -factors		
Protein	19.3	21.8
Ligand/ion	22.2	
Water	32.8	35.2
R.m.s. deviations		
Bond lengths (Å)	0.007	0.007
Bond angles (°)	1.09	1.06
Ramachandran		
Favored	97.0	97.0
Allowed	2.81	2.8
Outliers	0.19	0.2

\* Highest resolution shell is shown in parentheses. Both structures are based on single crystals.

**Supplementary Table 2 | SAXS data and analysis.**

Data-collection parameters		
Wavelength	1.54 Å	
q range	0.015 to 0.338 Å <sup>-1</sup>	
Exposure time	300 s	
Concentration	2 mg/mL	
Temperature	20 °C	
Structural parameters		
SAXS		
R <sub>g</sub> (Å) from Guinier plot	27.40 ± 0.06	None
R <sub>g</sub> (Å) from P(r)	27.10 ± 0.03	24.99
D <sub>max</sub> (Å)	87.00	78.00
Fitting SAXS profile by minimizing $\chi$ function using experimental data		
$\chi_{\text{Exp/DAM}}$	1.40	
Fitting SAXS profile by minimizing $\chi$ function between experimental data and MD models		
Intermediate		
Open		Average
$\chi \chi$	2.62	3.27
		4.45
		1.89

**Supplementary Table 3 | Sequences of the primers used in *BlCel5B* constructs cloning.**

Primer name	Sequence (5'-3')
<i>Blcel5B_Fw</i>	CAGGGCGCCATGGTTCCGAAAGCTTCGGG
<i>Blcel5B_Rv</i>	GACCCGACGCGGTTAAGGCCTGATTTCACCTTCACG
<i>Blcel5BΔ1006-1602_Rv</i>	GACCCGACGCGGTTAGATCGCATTGGGAACGC
<i>Blcel5BΔ1297-1602_Rv</i>	GACCCGACGCGGTTACTTCGGGTACATCAAAGAACG
<i>Blcel5BW479A_Fw</i>	TTGAAGAAGTAGCGTCATGGACG
<i>Blcel5BW479A_Rv</i>	GAACCGGCCTTCCTGTAGAATC
<i>Blcel5BW481A_Fw</i>	AGTATGGTCAGCGACGCCCTAC
<i>Blcel5BW481A_Rv</i>	TCTTCAAGAACCGGCCTTCC

**SUPPLEMENTARY VIDEO S1 |** Video obtained from the molecular trajectories generated by MD simulations of the substrate-free *BlCel5B* showing open-close transition of the CBM46 over the catalytic domain (CD).