1 On the distinct binding modes of expansin and
2 carbohydrate-binding module proteins on
3 crystalline and nanofibrous cellulose: implications
4 for cellulose degradation by designer cellulosomes
5
6 Adam Orłowski†, Lior Artzi’, Pierre-Andre Cazade†, Melissabye Gunnoo†,
7 Edward A. Bayer’, Damien Thompson*†
8
9 † Department of Physics, Bernal Institute, University of Limerick, V94 T9PX,
10 Ireland
11 ’Department of Biomolecular Sciences, The Weizmann Institute of Science,
12 Rehovot, Israel
13
14 Corresponding Author
15 * Correspondence to: Damien.Thompson@ul.ie
16
17 Keywords: Cellulose, enzyme, carbohydrate-binding module, CBM, expansin,
18 molecular dynamics simulations, binding free energy calculations
19
20 ABBREVIATIONS
21 CBM, carbohydrate-binding module; SMD, steered molecular dynamics; PMF, potential of mean
22 force; RMSD, root mean square deviation.
Abstract
Transformation of cellulose into monosaccharides can be achieved by hydrolysis of the cellulose chains, carried out by a special group of enzymes known as cellulases. The enzymatic mechanism of cellulases is well described, but the role of non-enzymatic components of the cellulose-degradation machinery is still poorly understood, and difficult to measure using experiments alone. In this study, we use a comprehensive set of atomistic molecular dynamics simulations to probe the molecular details of binding of the family-3a carbohydrate-binding module (CBM3a) and the bacterial expansin protein (EXLX1) to a range of cellulose substrates. Our results suggest that CBM3a behaves in a similar way on both crystalline and amorphous cellulose, whereas binding of the dual-domain expansin protein depends on the substrate crystallinity, and we relate our computed binding modes to the experimentally measured features of CBM and expansin action on cellulose.
**Introduction**

Cellulose is the most abundant source of raw material for production of biofuels, and much effort has therefore been devoted to its efficient catalytic degradation. In bacteria and fungi, cellulase enzymes act synergistically with a variety of other proteins to release soluble units (cellodextrins) from the insoluble and very stable cellulose substrates. Cellulases do not efficiently hydrolyse natural cellulose, which is mostly a crystalline assembly of ordered, well-packed regions of microfibrils that are inaccessible to enzymes and water. The first step is to render the substrate accessible to cellulose-degrading enzymes. Hence, efficient degradation of cellulose-containing materials requires non-catalytic disruption of the crystalline regions into a more accessible and hydrolysable substrate together with anchoring of cellulases to the cellulose. These anchoring and disruption roles are executed by groups of proteins distinct from cellulases, and multi-protein architectures have evolved that contain both the enzymatic and non-enzymatic units, such as bacterial cellulosomes that attach to cellulose and are generally thought to disturb its packed structure prior to cellulolytic hydrolysis. In recent years much effort has been spent in the fabrication of “designer cellulosomes” (DC), which are artificial constructs that combine motifs from different cellulosomes to maximize hydrolysis of cellulose. Rational design of improved DCs requires deeper knowledge of the precise role(s) performed by each docking, binding, and catalytic module.

Carbohydrate-binding modules (CBMs) are thought to enhance the adsorption of enzyme complexes to their carbohydrate substrate, to enable their alignment on fibrils, and in some cases to modify and “roughen” substrate surfaces to facilitate enzymatic hydrolysis. It was demonstrated that wild-type cellulases affect primarily the outermost surface properties of an amorphous cellulose film, while chimeric fusion proteins containing a catalytic domain and the CBM from *T. fusca* exoglucanase Cel6B improve the efficiency of the cellulase three-fold by enabling digestion within the bulk of the film. In the same study, a swelling of the cellulose film was observed together with substantial changes to its bulk properties upon incubation, which was ascribed to disruption of amorphous cellulose by the CBM.
In addition to CBMs with a potential disruptive function on crystalline cellulose \(^24,25\), other groups of proteins were proposed to perform non-catalytic disruption of recalcitrant cellulose fibres \(^{10,11,23,26,27}\). Expansins, expansin-like proteins, and swollenins are proposed to bind to cellulose and non-hydrolytically disrupt the hydrogen-bonding networks in cellulose \(^{28-30}\) and increase the efficiency of catalytic hydrolysis in a synergistic manner \(^{31-33}\). Plant expansins, which are suggested to be amongst the most effective promoters of hydrolysis, belong to two main groups: \(\alpha\)-expansins and \(\beta\)-expansins that bind to xyloglucan-enriched cellulose and to cell walls of high arabinoxylan content, respectively \(^{26,30,34,35}\). On the other hand, bacterial expansins have the ability to bind to whole (native) cell walls \(^{36-38}\) and so are especially promising for industrial applications \(^{39}\).

Bacterial expansins consist of two domains, domain D1 and D2, that form contacts with each other and are joined by a short linker \(^{40-42}\). The D1 domain lacks catalytic activity, even though it has been found to be structurally similar to family 45 glycoside hydrolases \(^{43}\), because it lacks key residues that are critical for catalytic activity. Moreover, not much is known about the D1 domain action on the substrate \(^{26,39,42}\). On the other hand, the D2 domain is a CBM with a flat aromatic-rich surface that can recognise crystalline cellulose \(^{42}\), and these interactions are well documented \(^{39,40,42}\). This domain has been assigned to the family 63 in the CAZy database \(^{39}\). Interestingly, only expansins with two intact D1 and D2 domains have been shown to exhibit disruption activity, and neither one of the domains can perform this function independently nor can they do so without being joined together \(^{42}\). Nevertheless, the CBM-like domain D2 expresses the same cellulose-binding activity as the full expansin \(^{42}\).

The role of CBMs in cellulose binding is well established (although based on indirect evidence \(^{40}\)) with proposed molecular binding mechanisms \(^{25}\). On the other hand, the role of CBMs, expansins, and other proteins in non-catalytic disruption of cellulose is less obvious and details are still elusive \(^{10,25}\), largely due to the lack of simple and efficient assays to quantify their disruptive activities. There are only a few examples of computed molecular mechanisms of non-catalytic cellulose disruption \(^{21,44}\), and from the point of view of experimental measurements, expansins may degrade pure-cellulose filter paper \(^{39}\). Expansins have been posited to act non-specifically, by enhancing cellulase action through
(i) increased cellulase stability, (ii) decreased non-productive binding to the substrate, and (iii) their action as surfactants given their amphiphilic nature.  

Nevertheless, the significance of Asp82 in expansin EXLX1 was shown in mutagenesis experiments and molecular dynamics simulations.  

In view of the above, our motivation for the present work was to elucidate the molecular details of binding of cellulose crystals and nanofibers by non-hydrolytic proteins CBM3a and EXLX1. We used multiple, long classical atomic-resolution molecular dynamics (MD) simulations to probe the molecular mechanisms of CBM and expansin binding to cellulose, and we discuss our findings relative to literature experimental and theory reports together with new control experiments in which we pre-incubated filter paper strips with CBM or expansin, transferred the strips into tubes and measured their degradation by two recombinant cellulases originated from the C. clariflavum genome.

Materials and methods

Computational models

(a) Overview of the simulated complexes

We used molecular dynamics (MD) simulations to study the structure, dynamics, and cellulose-binding energetics of a family-III carbohydrate-binding module (CBM3a) from the cellulosomal scaffoldin subunit of Clostridium thermocellum, and of expansin (EXLX1) from Bacillus subtilis. The interactions of these proteins with cellulose were studied by introducing them into two distinct cellulose environments: the crystalline form type Iβ surface consisting of three layers of cellulose, and nanocellulose fibres with two distinct non-crystalline regions. CBMs can bind to different types of cellulose, including crystalline and amorphous cellulose. However, the catalytic domains that are attached to CBM3a very often act on the loose chains of cellulose nanofibers. Therefore, to comprehensively scan the range of CBM–cellulose interactions, we performed four simulations of the protein with CBM binding to cellulose nanofibers that include both crystalline and non-crystalline regions (for visualisation see Fig. 2 C, E). Details of the cellulose models and simulation protocols are presented below.
(b) Cellulose models

A slab of cellulose cut from crystalline cellulose I\(\beta\) and stabilised by hydrogen bonds at its (100) surface terminations was modelled as three layers of ten sixteen-monomer chains (Figure 2 A, B). The terminal molecules of glucose on each chain were bonded together across the periodic walls of the simulation box. The layers extended primarily in the xy-plane, where the x-direction runs along the cellulose chain from the C1 atom to the C4 atom. The direction normal to the surface was denoted as z.

Cellulose nanofibers were created from crystalline I\(\beta\) with amorphous regions. We created one type of amorphous cellulose model (abbreviated as \(-\text{aI}\)) corresponding to a nanofiber built from layers with 3, 32, 17, 5, and 8 cellulose chains of 50, 20, 22, 24, and 26 glucose units, respectively. The chains were positioned to form an interface at which a non-crystalline amorphous region is formed between two I\(\beta\) crystalline parts. Part of the chains (three chains being 50 units long) continue from one crystalline part to the other, thereby keeping the crystalline units together, while most of the chains are broken and protrude into the non-crystalline region (Figure 2 C, D). During MD simulations these fibrils can twist and restructure but do not break apart. A second fibril model (abbreviated as \(-\text{aII}\)) had an amorphous-like coating on one of its crystalline faces (Figure 2 E, F). Here, the fibrils consisted of 59 cellulose chains forming the I\(\beta\) crystalline part, and of 6 chains that are disordered and hence form the amorphous part. All chains had the same length of 40 glucose units.

(c) Proteins

CBM3a and expansin protein atomic coordinates were obtained from the RCSB database, PDB record 1NBC for CBM at 1.75 Å resolution, and 3D30 for expansin at 1.90 Å resolution.

(d) Description of the simulated systems

The models were named after the protein (Exp for expansin and CBM for CBM) and the cellulose type present in the corresponding system (c – crystalline, aI – amorphous type I, and aII – amorphous type II), so the full simulation name consists of two parts (protein name-cellulose type) and the simulation number/repeat. All simulations and their durations are listed in Table 1, and the numbers of molecules and atoms in each simulation are given in Supplementary
Table 1. Protein starting coordinates were taken from the X-ray structures\textsuperscript{41,55}, and each of protein was introduced to the three cellulose structures to generate a total of 8 simulation types (Table 1). Simulations were performed in water at physiological salt concentrations of 150 mM NaCl and extra background counterions were introduced where necessary to neutralise the total charge of the system (Supplementary Table 1).

Table 1. List of the simulated systems with the number of simulations and run lengths (in ns). The sampling times and number of repeats were adjusted in order to identify protein binding modes and monitor cellulose dynamics within reasonable computational times.

<table>
<thead>
<tr>
<th>No</th>
<th>Name</th>
<th>Description</th>
<th>Number of simulations</th>
<th>Length (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PMF-CBM</td>
<td>Potential of mean force (CBM-crystalline cellulose)</td>
<td>1 (24 umbrella windows)</td>
<td>96 (each window)</td>
</tr>
<tr>
<td>2</td>
<td>PMF-Exp</td>
<td>Potential of mean force (expansin-crystalline cellulose)</td>
<td>1 (24 umbrella windows)</td>
<td>96 (each window)</td>
</tr>
<tr>
<td>3</td>
<td>CBM-c(1-5)</td>
<td>CBM-crystal cellulose simulation</td>
<td>5</td>
<td>4 x 1000, 863</td>
</tr>
<tr>
<td>4</td>
<td>CBM-al(1-2)</td>
<td>CBM-amorphous cellulose type I</td>
<td>2</td>
<td>2 x 500</td>
</tr>
<tr>
<td>5</td>
<td>CBM-alII(1-2)</td>
<td>CBM-amorphous cellulose type II</td>
<td>2</td>
<td>2 x 500</td>
</tr>
<tr>
<td>6</td>
<td>Exp-c1</td>
<td>Expansin-crystal cellulose</td>
<td>1</td>
<td>1000</td>
</tr>
<tr>
<td>7</td>
<td>Exp-al(1-4)</td>
<td>Expansin-amorphous cellulose type I</td>
<td>4</td>
<td>365, 668, 83, 138</td>
</tr>
<tr>
<td>8</td>
<td>Exp-alII(1-2)</td>
<td>Expansin-amorphous cellulose type II</td>
<td>2</td>
<td>412, 364</td>
</tr>
</tbody>
</table>

Simulation protocols

(a) Atomistic Molecular Dynamics Simulations

The CHARMM36 force field\textsuperscript{56–60} was used to describe all molecules in simulations of protein binding to crystalline cellulose, with water was described using the TIP3P model that is compatible with the CHARM36 parameterisation. The force field used was analogous to that employed in our previous scan of binding strengths for amino acid and small peptide adsorption on crystalline cellulose\textsuperscript{53}. The OPLS all-atom force field was used to describe all molecules\textsuperscript{61} in the simulations with amorphous cellulose, and water was modelled using the TIP3P...
variant that is compatible with the OPLS parameterization\textsuperscript{62}, similar to the
protocol used to model cellulase binding to cellulose nanofibrils in earlier work\textsuperscript{50,54,63}.

To simulate molecular dynamics, we applied boundary conditions to our
simulation cells with the usual minimum image convention employed in all three
dimensions. The length of each hydrogen atom covalent bond was preserved by
the LINCS algorithm\textsuperscript{64}, which allowed use of an integration time step of 2 fs.
Beyond this, no bias or special conditions were applied to the simulation of the
solvated interface between the proteins and cellulose. The simulations were
carried out at constant atmospheric pressure (1 bar) and room/body temperature
(310 K), controlled by the Parrinello-Rahman and velocity-rescale methods,
respectively\textsuperscript{65,66}. The temperatures of the solute and solvent were coupled
separately. For pressure, an isotropic scaling was employed. The Lennard-Jones
interactions were cut off at 1.0 nm. For the electrostatic interactions, the particle
mesh Ewald method\textsuperscript{67} was employed with a real space cut-off at 1.0 nm, beta
spline interpolation (6\textsuperscript{th} order), and a direct sum tolerance of 10\textsuperscript{-6}. All simulations
were performed using the GROMACS 5 code\textsuperscript{68}.

(b) Free Energy Calculations

The binding free energies of CBM and expansin adsorption on crystalline
cellulose were calculated using the potential of mean force (PMF) method using
umbrella sampling mixed with Hamiltonian Replica Exchange Molecular
Dynamics (H-REMD)\textsuperscript{69,70} in two steered molecular dynamics (SMD) simulations
(named PMF-CBM and PMF-Exp) with exchange attempted every 200 simulation
steps. The starting structures for SMD were obtained from the last frame of the
equilibrium simulations: CBM-c1 and Exp-c1 for the systems PMF-CBM and
PMF-Exp respectively, and the systems were prepared using the following
procedures.

Structures for the umbrella windows were obtained by desorbing the
protein (CBM or expansin) up off the crystalline cellulose surface into bulk water
by pulling along the $z$-axis in a 5 ns SMD simulation with a constant pulling force
of 1000 kJ/mol/nm$^2$ and pulling rate of 0.01 nm/ns exerted on the protein. In
umbrella sampling simulations, 24 0.05 nm-spaced windows were employed for
the two systems. The reaction coordinate was chosen as the $z$-coordinate of the
protein centre of mass. In each window, a harmonic restraining potential of 1000
3.7 kJ/mol/nm² in the z-direction was applied on the distance between the centres of mass of the protein and the cellulose substrate using the PLUMED 2 GROMACS plug-in and the protein orientation (rotation) in the xy-plane (cellulose plane) was restrained using the “angle_restraints_z” parameter in Gromacs topology to restrict rotation of the protein in the xy-plane. Finally, the free energy profile for both systems was constructed using the weighted histogram analysis method (WHAM). All other simulation parameters were kept exactly the same as described above for the equilibrium simulations.

(c) A note on sequence and structure of CcEXL1 vs. BsEXLX1

Since the three-dimensional structure of the bacterial expansin was resolved only for the *Bacillus subtilis* EXLX1⁴¹, we used this structure as the protein model in our simulations, in line with other recent computational work⁴⁴. In experiments (see Supporting Information), we also used bacterial expansin, but the one from the bacterial cellulosome which is very effective in cellulose disruption⁷³. The BsEXLX1 used in the simulations has 62% sequence identity (see Fig. S1) with the CcEXL1 used in experiments and therefore our description of the protein binding is relevant for the experimentally used protein as well. As an additional control, we used SWISS-MODEL⁷⁴–⁷⁶ to build a homology model of CcEXL1 using the BsEXLX1 structure as a template (see Fig. S1), which confirmed negligible sequence discrepancy between the computationally and experimentally studied expansin proteins. In any case, we do not consider mutations of particular residues in simulations or experiments, but rather the general effect of WT expansin and CBM binding on degradation of cellulose.

Results and Discussion

Protein structure and dynamics during the simulations

Fig. S2–S17 show the preservation of protein secondary structures throughout the simulations. Fig. S18 shows the low Root Mean Square Deviation (RMSD) of the CBM during the simulations, and provides further support of very high stability of the protein structure. For expansin, there is a clear difference between protein dynamics of the two domains of the protein (Fig. S19). Domain D1 is less ordered, as reflected in the RMSD plots (Fig. S19), which might be a result of higher content of unstructured and therefore also more flexible regions (for
secondary structure see: Fig. S11-S17). To find the regions of the protein where
the main fluctuations take place, calculated C-alpha atom Root Mean Square
Fluctuations (RMSF) are presented in Fig. S20 and S21 (for CBM and expansin,
respectively). These data support the idea that the flexible loops of domain D1 are
responsible for higher fluctuations of the protein structure.

**Binding of proteins to cellulose**

The initial position and orientation of the proteins was random in all simulations,
except for systems CBM-c1 and Exp-c1 in which proteins were placed on the
crystalline cellulose surface in binding modes suggested by previous studies and
theoretical predictions \(^{41,46,55}\). The starting configurations of systems Exp-aI3 and
Exp-aI4 were adapted from the last snapshot of simulation Exp-aI2, and the
protein was shifted in such a way that residue Asp82 is placed initially in close
contact with loose ends of cellulose chains. The rationale for this bias is based on
the results of experiments and computational studies on expansin, which indicated
that residue Asp82 plays a significant role in loosening cellulose fibers by
inducing a twist in cellulose chain orientation \(^{42,44}\). In all our simulations with a
protein randomly positioned in the water above the cellulose we observed
spontaneous binding of the proteins to both crystalline (for CBM/expansin to
cellulose distance see Fig. 1) and amorphous cellulose (movies S1-S16). The
details of the binding processes of CBM and expansin on cellulose are presented
below.
Fig. 1. The computed elevation of protein above the cellulose surface in: CBM-c1 (A), CBM-c2 (B), CBM-c3 (C), CBM-c4 (D), CBM-c5 (E), Exp-c1 – black line (domain D1 of the protein) and red line (domain D2 of the protein) (F). The data plotted represent the centre of mass (COM) distances between the protein and cellulose surface in the direction normal to the plane of the cellulose surface.

(a) Binding of CBM to crystalline cellulose

Representative snapshots from the beginning and end of each simulation of CBM binding to cellulose are shown in Fig. 2, and the entire trajectory for each simulation can be viewed in movies S1-S9. More detailed snapshots from the simulations CBM-c2 and CBM-c5 are shown in Figs. S22-S24, highlighting the residues that mediate the initial binding to the cellulose.
Fig. 2 Snapshots from the simulations of CBM binding to cellulose: CBM-c5 (A) at 0 ns, (B) after 1000 ns; CBM-aII (C) at 0 ns, (D) after 500 ns; CBM-aII1 (E) at 0 ns, (F) after 500 ns. Cellulose is depicted as a tan surface, protein is presented in “New Cartoon” representation and coloured accordingly to the secondary structure, and water and ions are not shown for clarity. All the computed MD structures and movies presented in this work were visualised using the VMD package.

For CBM binding to crystalline cellulose we calculated the distances between the centres of the mass of CBM and the cellulose slab (Fig. 1 A-E). In the simulation CBM-c1 where the protein was initially placed in the bound configuration on the cellulose, it tends to stay there for the entire simulation time (Fig. 1 A). In all other simulations of CBM with crystalline cellulose, where the protein was initially placed in water (simulations CBM-c2-c5), it eventually binds to the cellulose surface (Fig. 1 B-E, movies S2-S5). Nevertheless, only simulations CBM-c2 and CBM-c5 find the previously-calculated binding mode in which the flat side of CBM is bound to the cellulose surface (movies S2, S5). The variety of binding modes we observe in our MD simulations is most likely due to random behaviour of the protein in water solution and diffusion, which present various sides of the protein to the cellulose surface. The details of the hydrogen bond interactions and the overall number of contacts between the protein residues and cellulose is described below. We focussed our analyses and discussion on the simulations CBM-c1, CBM-c2 and CBM-c5 (flat surface bound), but data are
provided and briefly discussed for all the simulations. Computed timelines of the number of hydrogen bonds formed between the CBM and crystalline cellulose are presented in Fig. 3A-E, with the average numbers provided in Table 2. The detailed list of hydrogen bonds between CBM and cellulose is given in Tables S2-S6.

Fig. 3 Computed timelines of hydrogen bonds between CBM and cellulose in simulations: (A) CBM-c1, (B) CBM-c2, (C) CBM-c3, (D) CBM-c4, (E) CBM-c5, (F) CBM-aI1, (G) CBM-aI2, (H) CBM-aII1, (I) CBM-aII2.

Table 2. Average number of hydrogen bonds between the CBM and cellulose calculated from the simulation time given in the last column.

<table>
<thead>
<tr>
<th>Simulation Name</th>
<th>Average number of hydrogen bonds with the standard deviation</th>
<th>Time period used for analyses (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBM-c(1-5)</td>
<td>3.16 (+/- 0.01); 3.18 (+/- 0.01); 1.74 (+/- 0.01); 4.73 (+/- 0.01)</td>
<td>500-1000 ns</td>
</tr>
<tr>
<td>CBM-aI(1-2)</td>
<td>8.40 (+/- 0.02); 16.18 (+/- 0.03)</td>
<td>300-500 ns</td>
</tr>
<tr>
<td>CBM-aII(1-2)</td>
<td>3.62 (+/- 0.02); 3.88 (+/- 0.01)</td>
<td>300-500 ns</td>
</tr>
</tbody>
</table>
Based on these data we could determine that ~3 hydrogen bonds provide stable binding of the CBM to crystalline cellulose, and the main contributors are residues: Ser9, Asn10, Thr14, Asn16, Asp56, His57, Gln110, and Trp118. These hydrogen bonds are “switched on” as the CBM forms a stable encounter complex with crystalline cellulose (Fig. 3 A, B, E). Nevertheless, CBM does not lie down on the cellulose surface directly in the initial steps of the binding, but first establishes edge-on contacts with the cellulose. Specifically, the first residues to attract CBM to crystalline cellulose are Asn64 and Ser66 (through electrostatic interactions), mediated by stabilisation of the interaction with the cellulose through aromatic residue Tyr67 (Figs. S22-S24).

Since cellulose is insoluble in aqueous solution, it is expected that binding to its surface involve hydrophobic interactions, which is indeed observed and described for many CBMs\(^ {25,55}\). To retrieve information on this type of interaction in our simulations we calculated the occurrence of all inter-atomic contacts < 0.35 nm during the simulation time between CBM residues and cellulose atoms. The visual representation of the data and the exact percentage of such occurrence in simulation frames for each residue are presented in Fig. 4 and Table S18.

Fig. 4 The computed occurrence of contacts (red crosses connected by black lines) where one or more atoms of a CBM residue are within 0.35 nm of cellulose in simulations: (A) CBM-c1, (B) CBM-c2, (C) CBM-c3, (D) CBM-c4, (E) CBM-c5, (F) CBM-a1I1, (G) CBM-a1I2, (H) CBM-a1I1, (I) CBM-a1I2.
There are four CBM regions that are consistently in contact with cellulose. They span residues: Ser97-Asn19, Trp547-Tyr67, Gln110-Trp118, and Ser133-Gln134. The first three regions contain residues also establish hydrogen bonds with cellulose, as described above. The two other simulated systems showed only weakly bound CBM (CBM-c3 and CBM-c4), as is obvious from both visual inspection of the MD trajectories (movies S3-S4) and by inspection of the calculated properties. For instance, the centre of mass distance of CBM from cellulose is not stable in simulation CBM-c3 (Fig. 1 C), due to a predominantly upside-down orientation of the protein, and is slightly higher in simulation CBM-c4 (Fig. 1 D) than in simulations with the CBM bound to cellulose through its “edge”. Similarly, the number of hydrogen bonds in these two systems (Fig. 3 C, D) and inter-atomic contacts between the protein residue atoms and cellulose (Fig. 4 C, D) differ significantly from the consistent patterns observed in the three other systems (CBM-c1, CBM-c2, CBM-c5). Thus, based on these findings, we do not qualify the two systems (CBM-c3 and CBM-c4) as stable bound forms of the CBM; they constitute minor populations and/or early stage structures in the formation of bound interfaces that are not readily identifiable at experimental timescales.
(b) Binding of CBM to amorphous cellulose

Two different types of cellulose nanofibers were modelled in our simulations (CBM-aI and CBM-aII, as described in the methods section). The protein was placed randomly in the water phase above the nanofiber surface as shown in Fig. 2 C, E and the same types of MD simulations and analyses were performed as described earlier for CBM binding to crystalline cellulose.

There is no simple way to describe formation of the initial encounter complex between CBM and cellulose nanofibers, because the simulations revealed a multitude of possible sites where the protein can attach. However, visual inspection (movies S6-S9 and Fig. 4 C-F) clearly shows that the CBM sticks to nanofibrous cellulose and remains bound. The hydrogen bonds established between the CBM and cellulose are shown in Fig. 3 F-I, Table 2, and Tables S7-S10. Very different H-bond patterns are observed compared with the simulations with crystalline cellulose and the patterns vary even between the two types of amorphous region (CBM-aI and CBM-aII). The difference stems from the presence of exposed loose ends of cellulose chains that are more likely to engage in hydrogen bonding with protein residues compared to packed crystalline cellulose. This is particularly apparent in simulations CBM-aI1 and CBM-aI2 where the protein binds to loose ends of cellulose chains in the nanofiber, involving between 8 and 16 H-bonds (Fig. 3 F, G and Table 2). In simulations CBM-aII1 and CBM-aII2 the four hydrogen bonds established between the CBM and cellulose (Table 2 and Fig. 3 H, I) are close to the three H-bonds calculated for CBM binding to crystalline cellulose (Table 2 and Fig. 3 A, B, E), because model (-aII) does not contain loose ends of cellulose chains (Fig. 2 E, F) and thus shows a lower propensity to interact directly with the protein. With regard to formation of specific hydrogen bonds between the protein and cellulose nanofibers, the binding interactions (Tables S7-S10) are not sufficiently specific and repeatable to consider them with confidence. However taken as a whole the data (Fig. 4 F-I and Table S18) clearly shows that the protein region from Thr27 to Ser30 is consistently in contact with the cellulose only in simulations with non-crystalline cellulose. Two of the serine residues from this fragment Ser29 and Ser30 are involved in hydrogen bonding with cellulose (Tables S7-S10). It is striking that this result is consistent in all simulations, and we can therefore speculate that these two residues might be important for recognition of more
hydrophilic and soluble regions of cellulose fibres by CBM3a. To the best of our knowledge this is a new finding, and we propose that future site-directed mutagenesis experiments could focus on re-engineering this region to further probe the selectivity of CBM binding to crystalline vs. amorphous regions of cellulose.

(c) Binding of expansin to crystalline cellulose

The initial placement of expansin on top of crystalline cellulose in simulation Exp-c1 was based on the known orientation and interaction sites of the protein. Therefore, we did not explore the spontaneous binding of the protein to the crystalline cellulose surface as we did for the CBM. As expected, the bound form of the protein is stable throughout the full microsecond of dynamics (Fig. 1 F, Fig. 5 A-B, and movie S10). Interestingly, while domain D2 of expansin behaves like the CBM, the centre of mass distance between cellulose and domain D1 fluctuates around the average value, indicating a more dynamic interface.

Fig. 5 Snapshots from simulations of expansin binding to cellulose: Exp-c1 (A) 0 ns, (B) after 1000 ns; Exp-aI1 (C) 0 ns, (D) after 365 ns; CBM-aII1 (E) 0 ns, (F) after 412 ns. Atom representation is the same as for Fig. 2.
The computed number of hydrogen bonds between expansin and cellulose is presented in Fig. 6, and the average numbers are listed in Table 3.

**Fig. 6** Computed timelines of hydrogen bonds between expansin and cellulose (black lines – whole protein, red lines – domain D1 of the protein, green lines – domain D2 of the protein) in simulations: (A) Exp7c1, (B) Exp7aI1, (C) Exp7aI2, (D) Exp7aI3, (E) Exp7aI4, (F) Exp7aII1, (G) Exp7aII2.
Table 3. Average number of hydrogen bonds between expansin and cellulose calculated from the simulation time given in the last column.

<table>
<thead>
<tr>
<th>Simulation Name</th>
<th>Average number of hydrogen bonds with the standard deviation</th>
<th>Time used in analyses (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Whole protein</td>
<td>Domain D1</td>
</tr>
<tr>
<td>Exp-c1</td>
<td>3.30 (+/- 0.01)</td>
<td>0.40 (+/- 0.005)</td>
</tr>
<tr>
<td>Exp-al(1-4)</td>
<td>2.67 (+/- 0.02); 3.15 (+/- 0.01); 2.52 (+/- 0.02); 5.76 (+/- 0.02)</td>
<td>1.60 (+/- 0.01); 1.11 (+/- 0.06); 0.51 (+/- 0.01); 1.80 (+/- 0.01)</td>
</tr>
<tr>
<td>Exp-all(1-2)</td>
<td>5.32 (+/- 0.02); 5.74 (+/- 0.03)</td>
<td>1.83 (+/- 0.02); 2.94 (+/- 0.01)</td>
</tr>
</tbody>
</table>

The detailed hydrogen bond statistics are given in Tables S11-S18. In the simulation with crystalline cellulose the majority of hydrogen bonds (~3) are established between domain D2 and cellulose (Fig. 7 A and Table 3) and the average number of bonds is close to that calculated for CBM binding to crystalline cellulose (Fig. 4 A and Table 2). Only a minor, sporadic population of H-bonds comes from the interaction between domain D1 and cellulose. Detailed analysis of the hydrogen bonds in simulation Exp-c1 revealed that the major contributions come from residues: Glu75, Glu120, Gly121, Asp156, Tyr157, Thr163, Asn164, Gly193 and Ser195. This calculated binding mode is similar to that observed in a 2.1 Å X-ray structure (PDB: 4FER) of expansin in complex with cello-oligosaccharides in which contacts between the protein and cellulose based sugars involve Leu24, Lys119, Glu120, Gly121, Tyr157, Thr163. The occurrences of all other inter-atomic contacts < 0.35 nm between expansin residues and cellulose are presented in Fig. 7 A and the computed populations are summarised in Table S19.
Fig. 7 The computed occurrence of contacts (red crosses connected by black lines) in which one or more atoms of an expansin residue are within 0.35 nm of cellulose in simulations: (A) Exp-c1, (B) Exp-a11, (C) Exp-a12, (D) Exp-a13, (E) Exp-a14, (F) Exp-a111, (G) Exp-a112.

The data further emphasise that contacts with crystalline cellulose are concentrated mainly on domain D2. Only residues Pro74 and Glu75 from D1 are in contact with crystalline cellulose. We note that residue Glu75 was also found to establish hydrogen bonds in almost all the other simulations, including with amorphous cellulose as described below. All other significant contacts between...
expansin residues and crystalline cellulose are located on domain D2, primarily
regions: Glu120-Trp126, Met155-Tyr157, and Thr163-Ser195.

(d) Binding of expansin to amorphous cellulose

Binding of expansin to amorphous cellulose shows a significantly larger fraction
of hydrogen bonds between domain D1 and cellulose (Fig. 6 B-G and Table 3).
Moreover, in some of these simulations the overall number of hydrogen bonds is
higher (simulations Exp-all1 and Exp-all2), due to increased H-bonding for both
domains D1 and D2 (Table 3). Hydrogen bond statistics for simulations Exp-all1
and Expall2 indicate strong participation by residues: Tyr13, Thr14, Ser16, Ser19,
Gly20, Glu75, Asp96, Lys119, Glu120, Glu191, Ser192, Gly193, Ser195 in Exp-
all1; and residues: Met1, Asn15, Ala40, Asn43, Gly45, Lys48, Glu75, Gly76,
Arg78, Trp126, Glu153, Asp156, Tyr157, Glu191, Ser192, Gly193 and Thr194 in
Exp-all2. These residues are the most significant contributors to the binding of
expansin to cellulose nanofibers with non-crystalline regions via hydrogen bonds.
Analysis of all inter-atomic contacts between expansin and cellulose (Fig. 7 B-G
and Table S19) also show that domain D1 becomes involved in binding of
expansin to cellulose nanofibers. More specifically, the repeatable pattern of a
large number of contacts for the expansin region spanning residues Thr14-Leu24
(supported by hydrogen bonds analysis) indicates that residues within this region
play a significant role in binding of expansin to amorphous cellulose. It is striking
that this region shows only slight or no contact with crystalline cellulose,
supporting the hypothesis that expansin interacts differently with crystalline and
amorphous cellulose substrates.

(e) Comparison of expansin binding to crystalline and amorphous cellulose

The residues that are repeatable in at least two types of cellulose systems
(crystalline surface and amorphous nanofiber), and therefore, most important in
overall interactions with cellulose are: Glu75, Glu120, Asp156, Tyr157, Glu191,
Ser192, Gly193, Ser195. Note that residue Glu75 from domain D1 is the single
residue that establishes hydrogen bonds with cellulose in all but two of the
simulations with different cellulose types. The rest of the residues listed above are
from domain D2 which assumes the binding role. The predicted importance of
Glu75 is consistent with the experimental finding that Glu75 plays a (moderate)
role in wall creeping activity.42
(f) **Comparison of CBM and expansin binding to cellulose**

Our computed binding modes for both CBM and expansin physisorption to cellulose fit the generally accepted scheme whereby adsorption is mediated by aromatic residues \(^\text{25,40,81}\). However, some subtle details were revealed. Firstly, spontaneous binding of CBM to crystalline cellulose surface is *initiated* by polar residues. Considering the long-range nature of such electrostatic interactions, it is reasonable that they will precede short-range hydrophobic van der Waals interactions with aromatic residues. Secondly, the CBM-type role of expansin domain D2 was confirmed in our simulations to mediate stable anchoring of expansin to cellulose. By contrast, domain D1 was more flexible on the crystalline cellulose surface which suggests that it is not the preferred binding partner for D1. Indeed, we observed increase of hydrogen bond interactions of this domain with non-crystalline cellulose, which suggests that, while domain D2 will bind preferentially to crystalline cellulose, D1 will preferentially bind to non-crystalline or amorphous regions of cellulose. The main difference between CBM and expansin will therefore be related to the presence of domain D1 in expansin, which will not contribute to adhesion on the crystalline cellulose surface but might be important for interactions with more amorphous sites, typically defects and/or cellulose nanofiber surfaces.

**Role of water in the binding of the proteins to cellulose**

Given the well-known insolubility of crystalline cellulose \(^\text{82–84}\), it is tempting to investigate how soluble proteins can firmly bind to the substrate material, and to establish what is the role and behaviour of water molecules at the interface between the cellulose surface and protein. Snapshots from the simulations of both CBM and expansin show that water molecules are indeed expelled from the interface between the aromatic residues of protein and cellulose surface (Fig. 8 A), which is further supported by the quantitative profile of the water density in simulation CBM-c1 (Fig. 8 B).

**Fig. 8** (A) Heat map (from red to blue) of the density of water atoms above cellulose in simulation CBM-c1 (at 1000 ns). (B) The computed CBM-cellulose interface viewed from underneath the cellulose with the cellulose depicted as black lines, water as a cyan surface and protein residues taking part in binding to cellulose in a red licorice representation.
Free energy of CBM binding to crystalline cellulose

We estimated the magnitude of binding free energies of CBM and expansin on crystalline cellulose using Potential of Mean Force (PMF) calculations, and the computed (un)binding profiles are given in Fig. S25. We find very similar binding of both proteins, with an estimated energy of -22 kcal/mol. We did not attempt PMF calculations of binding to amorphous cellulose nanofibers, due to the large size of the simulation cells (on the order of 0.4 M atoms), and in any case a prohibitively large number of simulations would be required to obtain meaningful estimates of the experimental binding energy, given the variety of binding modes.
we observed in the simulations of protein binding to amorphous regions of cellulose. In addition, the direction of pulling would be difficult to define in these systems. Nevertheless, the PMF profiles measured on crystalline cellulose indicate that the proteins completely decouple from the substrate beyond a height of ~0.5 nm above the surface, and the computed values of both binding energy and interaction distance are almost identical for the two proteins. This is consistent with the inference from experiment that domain D2 of expansin and CBM3a are very closely related with a common role of binding to crystalline cellulose, and that domain D2 can bind effectively to cellulose, without any influence by D1. On the other hand, the similar computed $\Delta G$ values are in contrast to the experimentally measured $^{40,42}$ four times larger dissociation constant for expansin compared with CBM, which suggests more complex dissociation pathways than the simple one described here using the distance between the cellulose surface and the protein CoM as the reaction coordinate. The increased number of hydrogen bonds calculated between expansin and amorphous cellulose compared to crystalline cellulose in the equilibrium MD simulations (Figs 3 and 6) would suggest a larger affinity. The difficulty for both calculations and experimental measurement is further compounded by (a) the often unknown ratio of crystalline to amorphous regions in experiments and (b) the entropic contribution due to water molecules at the interface, and so more detailed experimental and simulation studies are needed to more precisely quantify the binding energies and identify the residues that contribute most to binding.

**Amorphogenesis of cellulose by CBM and expansin?**

Even though our simulations were extensive and to the best of our knowledge the time scales reported herein are the longest atomistic molecular dynamics simulations reported so far for CBM and expansin binding to cellulose, we did not observe any signs of cellulose disruption that would be significant for the cellulose depolymerisation (pre-incubation $^6$) process. Hence, while the simulations (and control experiments, see Supporting Information pages 6-9) indicate that the presence of CBM has little or no promoting effect on enzymatic hydrolysis, the calculated expansin-cellulose structures do not provide a straightforward explanation for the large measured improvement in cellulose degradation upon pre-incubation with expansin. Some plausible explanations are
that substrate amorphogenesis may: (i) occur at supra-microsecond time scales beyond current computational limits, (ii) involve covalent residue-glucan interactions, and/or (iii) require a more detailed model of the substrate nanostructure present in the experiments. Future work could utilise advances in HPC architecture and emerging highly-parallelised metadynamics methods to access experimentally relevant timescales to model large scale (slow) transformations in substrate topology as well as more detailed combined quantum/classical mechanics QM/MM models to include (possibly water mediated) protein-induced chemical transformations of the substrate.

Conclusions from recent contemporaneous computational studies appear to be consistent with our results presented in the current work. The authors did not observe disruption of bulk cellulose, and posited a mechanism of cellulose disruption by expansin via twisting of the cellulose monomer by Asp82. This hypothesis was based on a simulation with a single cellulose chain. Our efforts to find a similar effect of Asp82 on loose ends of non-crystalline regions of cellulose nanofiber were not successful, and we did not observe any stable contacts between loose chains of cellulose and residue Asp82 (see movie S17). Our results are also consistent with experimental inferences that expansins do not significantly change the crystal structure of cellulose, and further support the idea that the promoting effect of expansin is either not related to such changes or that it is (very) subtle, keeping in mind that the recombinant rice expansins used by Seki et al. were unlikely to be active, beyond non-specific protein effects.

**Conclusions**

Our combined simulation and experimental dataset provides atomic scale mechanisms for CBM3a and expansin binding to cellulose, highlights the key residues mediating the binding, provides estimates of the energetics of the binding process, and indicates that CBM3a and D2 of expansin act as anchors while D1 of expansin binds preferentially to non-crystalline regions. The two proteins share the same general binding mode to cellulose, in which hydrophobic interaction of aromatic residues with cellulose leads to depletion of water from the interface, creating a seal between the protein and cellulose. The binding energetics of expansin and CBM are strikingly similar and explain the experimentally observed competition of these two for binding sites on cellulose.
Binding of CBM to crystalline cellulose is initiated through creation of an initial electrostatic complex between polar residues and cellulose, after which the protein lies down flat to establish stable and firm binding mediated mainly by aromatic residues that deplete water from the protein–cellulose interface. The CBM will also bind non-specifically to non-crystalline regions of cellulose with an increased role of hydrogen bonds in interactions with amorphous cellulose. By contrast, expansin stays bound to crystalline cellulose mainly due to hydrophobic interactions of its domain D2 aromatic residues with cellulose and the binding complex strongly resembles the interface between CBM and crystalline cellulose, but the other domain D1 is mobile on the crystalline cellulose surface and does not contribute to anchoring of expansin to crystalline cellulose. Our data suggest that the flexibility and relatively low binding stability of domain D1 on crystalline cellulose compared with amorphous cellulose might lead to active searching of D1 for defects or less crystalline regions. The dual nature of expansin domains where domain D2 adheres to crystalline cellulose and the other domain D1 has a higher affinity for less crystalline regions is likely to be the perfect match for effective disruption of the cellulose substrate, which contains both crystalline and non-crystalline regions. Nevertheless, the molecular mechanism of disruption of cellulose by expansin remains unknown; in particular we observe in our simulations zero contacts between cellulose chains and residue Asp82 (shown previously to be crucial for cellulose disruption) despite scanning a variety of starting geometries. In closing, our computed D1 and D2 preferential binding to amorphous and crystalline cellulose helps explain the experimental finding that pre-incubation of cellulose with expansin improved the hydrolysis of cellulose by catalytic cellulases, which was not observed for CBM. Furthermore, our simulations indicate that the role of expansin domain D2 is almost identical to CBM, leaving domain D1 of expansin free to bind to non-crystalline regions, which may be the preliminary step in cellulose disruption.

ACKNOWLEDGMENT

We thank Science Foundation Ireland (SFI) for provision of computing time at the SFI/Higher Education Authority Irish Center for High-End Computing (ICHEC).
Funding Sources. This research has been supported by the European Framework Programme VII NMP grant 604530-2 (CellulosomePlus). DT acknowledges also support through Science Foundation Ireland (SFI) (grant no. 15/CDA/3491).

REFERENCES


