Journal of Molecular Modeling (2019) 25: 240
https://doi.org/10.1007/s00894-019-4133-1

ORIGINAL PAPER

Insights into the mechanism(s) of digestion of crystalline cellulose by plant class C GH9 endoglucanases

Siddhartha Kundu [1]

Received: 25 October 2018 / Accepted: 11 July 2019 / Published online: 23 July 2019
© The Author(s) 2019, corrected publication 2020

Abstract

Biofuels such as γ-valerolactone, bioethanol, and biodiesel are derived from potentially fermentable cellulose and vegetable oils. Plant class C GH9 endoglucanases are CBM49-encompassing hydrolases that cleave the β (1 → 4) glycosidic linkage of contiguous *D*-glucopyranose residues of crystalline cellulose. Here, I analyse 3D-homology models of characterised and putative class C enzymes to glean insights into the contribution of the GH9, linker, and CBM49 to the mechanism(s) of crystalline cellulose digestion. Crystalline cellulose may be accommodated in a surface groove which is imperfectly bounded by the GH9_CBM49, GH9_linker, and linker_CBM49 surfaces and thence digested in a solvent accessible subsurface cavity. The physical dimensions and distortions thereof, of the groove, are mediated in part by the bulky side chains of aromatic amino acids that comprise it and may also result in a strained geometry of the bound cellulose polymer. These data along with an almost complete absence of measurable cavities, along with poorly conserved, hydrophobic, and heterogeneous amino acid composition, increased atomic motion of the CBM49_linker junction, and docking experiements with ligands of lower degrees of polymerization suggests a modulatory rather than direct role for CBM49 in catalysis. Crystalline cellulose is the de facto substrate for CBM-containing plant and non-plant GH9 enzymes, a finding supported by exceptional sequence- and structural-homology. However, despite the implied similarity in general acid-base catalysis of crystalline cellulose, this study also highlights qualitative differences in substrate binding and glycosidic bond cleavage amongst class C members. Results presented may aid the development of novel plant-based GH9 endoglucanases that could extract and utilise potential fermentable carbohydrates from biomass.

Keywords Active-site geometry · Carbohydrate binding module · Class C GH9 endoglucanases · Crystalline cellulose · Glycoside hydrolase · Homology modelling · Interaction surface

Introduction

The microfibrillar structure of cellulose is constituted and strengthened by islands of hydrogen-bonded inter-glucan chains. These microcrystalline regions (I_α, I_β) render cellulose chemically inert and recalcitrant to most physical stressors, an attribute that is desirable to land plants (xylem, phloem), sporulating bacteria and fungi, and quorum sensing by microbial biofilms [1–8]. Most organisms (bacteria, fungi, protists) possess enzymes (oxidoreductases, EC 1.x.y.z; transferases, EC 2.x.y.z; hydrolases, EC 3.x.y.z) that can cleave cellulose into physiologically relevant oligo- and mono-saccharides (RXNs 1–3) [2, 9–16].

$$\{C_2, C_{2+}\} + \{e^-; 2e^-\} \rightleftharpoons \{C_2 = O, C_{2+} = O\} \quad \text{(RXN1)}$$

$$C_n + H_iPO_4 \leftrightharpoons C{-}OPO_3 + C_{n-1} \quad \text{(RXN2)}$$

$$C_n + kH_2O \leftrightharpoons + (m_1)(C_{n-i}) + kH_2O \leftrightharpoons (m_2)(C_{2-5}) + kH_2O \leftrightharpoons (m_n)(C_1) \quad \text{(RXN3)}$$

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00894-019-4133-1) contains supplementary material, which is available to authorized users.

✉ Siddhartha Kundu
siddhartha_kundu@yahoo.co.in

[1] Department of Biochemistry, Army College of Medical Sciences, Brar Square, Delhi Cantt., New Delhi 110010, India

C_n	≔	Glucan
C	≔	$D(\alpha)-$glucopyranose phosphate
i	∈	$\{1,2,3\}$
C_2	≔	Cellulose with degree of polymerization $(DP=2)$
C_{2+}	≔	Cellulose with degree of polymerization $(DP>2)$
$C_2=O$	≔	Lactone form of C_2
$C_{2+}=O$	≔	Lactone form of C_{2+}
C_{n-i}	≔	Shorter chain glucans
C_{2-5}	≔	Oligosacchrides $(DP\in\{2,3,4,5\})$ of $\beta(D)-$glucopyranose
C_1	≔	Monosaccharide of $\beta(D)-$glucopyranose
m_j	≔	Stoichiometry of short chain glucans $(m_1 < m_2 < m_3 < \ldots..m_n)$

Glycoside hydrolase 9 (GH9) endoglucanases (*EC* 3.2.1.4) hydrolytically cleave the $\beta(1\rightarrow4)$-glycoside linkage between contiguous (*D*)-glucopyranose residues and accomplish this with the aid of one or more carbohydrate binding modules (CBMs). Detailed phylogenetics analysis and molecular dating has shown that GH9 (≅480 *AA*) is very well conserved amongst taxa and has been so for ≈3000 Mya [8, 17]. The presence of active site residues in GH9 further imply that catalysis of crystalline cellulose proceeds by a relatively unchanged generic acid-base mechanism and may deploy aspartic (*D*) and/or glutamic (*E*) acids as alternating proton donors/acceptors. The arrangement of these, i.e. $\{EE, DD, DE, ED\}$, may then dictate the position of the −OH at the hemiacetal/acetal carbon (anomeric carbon; $\{C1, C2\}$) of the oligosaccharide products thereby retaining or inverting the configuration of the parent compound [18].

Carbohydrate-binding modules (CBMs) or carbohydrate-binding domains (CBDs) form distinct subsequences in eukaryotes (plants, CBM49; yeast, CBM54), protists (*Dictyostelium discoideum*, CBM8), fungi (CBM1), and bacteria (CBMs 2-4) [8, 17, 18]. Most CBMs are separated by linkers (<100 *AA*) from the GH domain(s) and vary in length (≈40 − 200 *AA*), number, position (N-, C-termini, central), substrate affinity, and contribution to catalysis [8, 17–41]. For example, GH9 endoglucanases from vascular land plants possess a unique subpopulation of CBM49-encompassing crystalline cellulose-digesting enzymes (class C) in addition to the amorphous cellulose cleaving subsets (classes A and B) [17, 18, 42–44]. The presence of one or more CBMs may also extend the range of substrates of GH9 enzymes to include complex heteropolymeric moieties (chitin, CBM5, 12, 14, 18, 33; polygalactouronic acid, CBM32; lipopolysaccharide/lipoteichoic acid, CBM39) [8, 17, 19, 35–41]. The precise mechanism(s) by which CBM-mediated catalysis proceeds is(are) debatable with several plausible explanations for the observed kinetic data [20–34]. Most CBMs possess non-contiguous aromatic amino acids (tryptophan/phenylalanine/tyrosine) interspersed with amino acids with shorter side chains. These could result in concomitant and non-uniform interactions with the glycosidic linkage(s) and consecutive cycles of stretching and relaxation. This mechanism favours the introduction of strain with consequent weakening of the glycosidic linkage [33, 34, 45–47]. Alternatively, there are reports that polar amino acids (serine/threonine/cysteine) could form complexes with calcium (CBM35, 36, 60) which, even in the absence of an overt CBM may mediate cleavage [48–50].

Extant structures of non-plant GH9 enzymes suggest that crystalline cellulose may be digested in subtle fully enclosed tunnels (processive), or in larger, open solvent accessible grooves/clefts (non-processive), although a mixed mode is likely to prevail in most enzymes [51–60]. The binding site(s) are labelled as plus (substrate, entrance) and minus (product, exit) sites with hydrolytic cleavage occurring between the +1 and −1 sites [51–57]. The length of the tunnel itself (≈50 *Ang*) is consistent amongst other GH9 enzymes and consists of about ten subsites (−7 *to* + 2), where amino acids make contact with the glucan chain [51–57]. Further insights into the mechanistic contributions of GH9, linker, and/or the CBMs may be gleaned from the X-ray structures of enzymes in complex with simple ($DP<9$; $DP=\{2,3,5\}$) or complex ($DP=10$; $-SH$) oligosaccharides [58–60]. For example, GH9 and CBM3 are distinct spatial entities (Cel9G, *Clostridium cellulolyticum*; CelE4, *Thermomonospora fusca*) with an interaction surface that comprises a network of hydrogen-bonded residues [59, 60]. However, in the absence of an active enzyme substrate (ES) complex ($DP\geq6$), the manner in which polymeric crystalline cellulose is processed by GH9 enzymes is not known [59]. Interestingly, the authors also report an inter-dependence or quasi-allostericity of the GH9 and CMBs in binding crystalline cellulose, a substrate-binding groove that is lined with polar and aromatic acid residues, and the possibility of a polyfunctional CelE4 with exo- and endoglucanase activities [59, 60]. Crystalline cellulose is the cognate substrate for GH9 endoglucanases in non-plant taxa such as bacteria, archaea, fungi, protists, and arthropods, and may predate plant GH9 enzymes by several millions of years [8]. This, when combined with the similarity between the GH9 domains, suggests that the active site architecture of plant class C enzymes and subsequent reaction chemistry may be

similar [8, 51, 52]. Whilst, the data generated vide supra is able to offer insights into the origin and evolution of plant class C enzymes, mechanistic details of the same are fundamental to comprehending the precise manner in which catalysis of crystalline cellulose may proceed. Here, I analyse homology models of putative and characterised plant class C sequences, i.e. with a single wel-defined CBM49 subsequence, to classify and infer the contribution(s) of the GH9, CBM49, and linker to the catalysis of crystalline cellulose.

Methods

Model generation, geometry optimization, equilibration, and MD of class C enzymes

A generic protocol to assess the contribution(s) of GH9, linker, and CBM49 has been outlined (Fig. 1). Laboratory-characterised full length (*FL*) and truncated (*T*) class C sequences (x) from *Oryza sativa* (*Q5NAT0*), *Gossypium hirsutum* (*Q8LJP6*), *Nicotiana tabacum* (*Q93WY9*), *Solanum lycopersicum* (*Q9ZSP9*), i.e. $x(FL) = x_{FL} = GH9 \cup L \cup CBM49$; $x(T) = x_T = GH9 \cup L$; $x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$, along with full-length putative class C sequences ($n = 92$) identified in previous work were submitted to Phyre2 (www.sbg.bio.ic.ac.uk/phyre2) [8, 18, 61]. The templates were graded in terms of the root mean squared deviation (*rmsd*) of their $C\alpha$-backbones from the predicted model, presence of an extant homologous structure (confidence), proportion of the sequence modelled (coverage), and sequence identity.

The LeAP module of AMBERTOOLS v17.0 was used to explicitly add water molecules (TIP3P) to the 3D models of characterised class C enzymes ($n = 4; x_{FL}, x_T$) and render the modelled structures electrically neutral ($\{Na^+, Cl^-\} \geq 1$) (Fig. 1) [62]. The models were optimised by minimizing their computed energies in a bi-phasic ($n_{\min 1} = n_{\min 2} = 5000$) implementation of the steepest descent algorithm with (100 Kcal $mol^{-}Ang^2$)

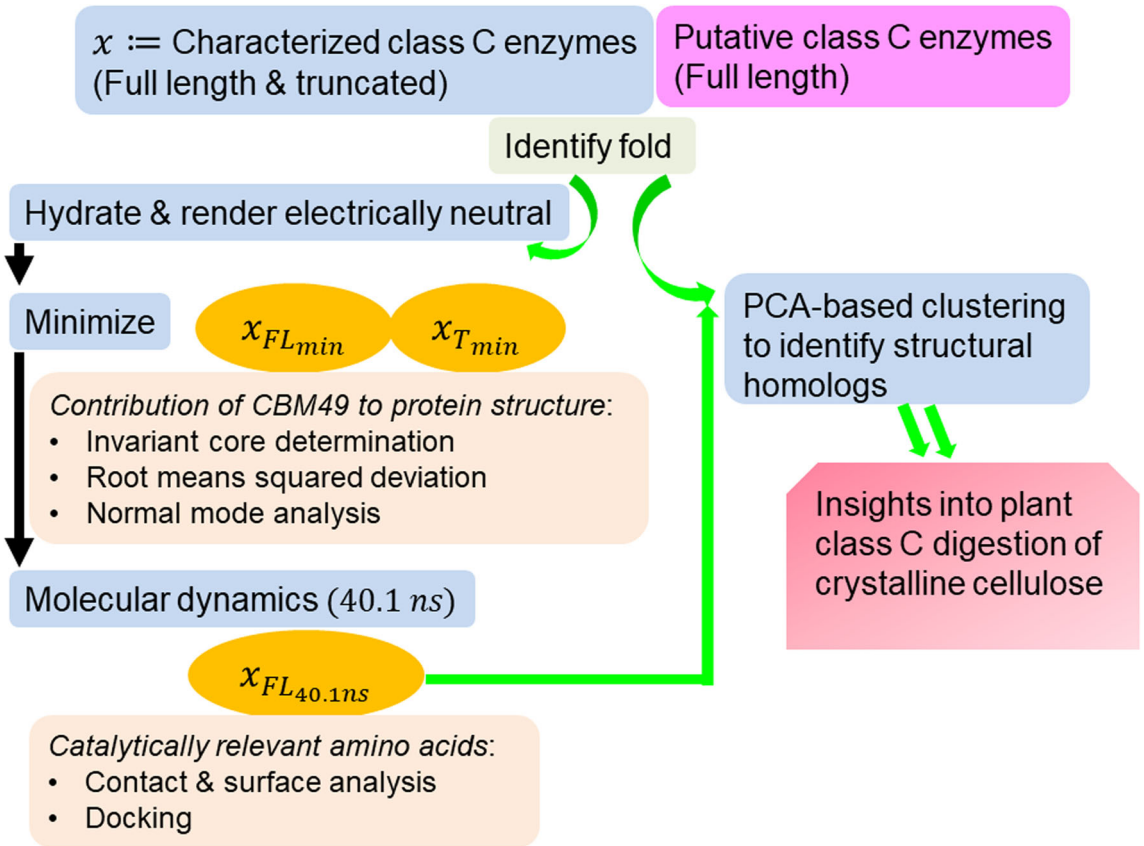

Fig. 1 Schema for biophysical characterization of class C GH9 enzymes. Generic protocol to assess contribution of GH9, CBM49, and the linker to catalysis of crystalline cellulose by plant class C enzymes. These steps consisted of fold identification, 3D protein and ligand geometry optimization, invariant core determination and normal mode analysis, surface analysis, cavity and groove delineation, and docking. Folds of characterised (full length, truncated) class C enzymes and putative class C sequences were initially identified. 3D models of class C enzymes with the top scoring templates (non-plant) were used for all further analysis; energy minimization ($E_{\min}$) of the 3D models was used to compare the effects of truncation on the structural integrity of the protein. Equilibrium structures (40.1 *ns*) were used subsequently to delineate the active site architecture of plant class C GH9 endoglucanases as well as conduct detailed docking studies with cellulose based ligands. Abbreviations—GH, glycoside hydrolase; CBM, carbohydrate binding module; Phyre2, protein homology/analogy recognition engine

and without positional restraints for the amino acids (Fig. 1, Table 1). The minimised models $(x_{FL_{\min}}, x_{T_{\min}})$ were utilised for comparative analyses to ascertain the significance and relevance of CBM49 to the structural integrity of the protein. Full length minimised structures were perturbed (Temp : 0.0K → 300.0K; constant volume; 20 *ps*) with low energy (10.0 Kcal $mol^{-1}Ang^2$) positional restraints for the amino acids, which was followed by an unrestrained (Temp = 300.0K; constant pressure; 100 *ps*) and a production grade run (40.1 *ns*) MD run with NAMD v2.13 (nanoscale molecular dynamics) and VMD v1.9.3 (visual molecular dynamics; configuration files) (Fig. 1, Table 1) [63, 64]. These models, i.e. $x_{FL_{40.1ns}}; x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$, were used to infer active site architecture, perform docking experiments, and identifying structural homologues of selected characterised class C enzymes (Fig. 1, Table 1).

Invariant core analysis of characterised and putative class C enzymes

The invariant core is a measure of inferring structural variation from the *xyz* coordinates of aligned atoms of amino acids at specific site(s) and was utilised to assess the conservation of GH9, linker, and CBM49. This was accomplished by generating multiple sequence alignments (MSA) with a standalone version of multiple sequence alignment by computing log-expectation (MUSCLE; http://drive5.com/muscle) in association with the R-package Bio3D (http://thegrantlab.org/bio3d) and with scripts developed in house (Fig. 1) [65–67]. The volume of the invariant core was then iteratively computed and is defined as the least volume ($V < 1.0\ Ang^3$) from all volumes of arbitrary ellipsoids ($V \geq 1.00\ Ang^3$). Here, an ellipsoid comprises the variance of eigenvalues along its three principle axes of the atomic *xyz* coordinates of amino acid(s) at every aligned position of the combined and ungapped MSA, whilst its volume represents the structural variation at the given position(s) [67–70]. Although Alanine is not the most hydrophobic amino acid ($kdH_{Ala} < kdH_{Met} < kdH_{Cys} < kdH_{Phe} < kdH_{Leu} < kdH_{Val} < kdH_{Ile}$; $kdH :=$ Kyte Doolittle Hydrophobicity index), its non-bulky and unbranched side chain renders it an excellent index of invariance of a given structure. Since truncating the proteins might be expected to dramatically alter the behaviour of the GH9 of the 3D models, a corrected subset (*O. sativa*, #*AA* = 456; *N. tabacum*, #*AA* = 466; *G. hirsutum*, #*AA* = 464; *S. lycopersicum*, #*AA* = 476) that comprised matched residues of full length proteins was used $(x(cFL_{\min}) = x_{cFL_{\min}})$, i.e.

$$x_{cFL_{\min}} = x_{FL_{\min}} - (x_{FL_{\min}} - x_{T_{\min}}) \tag{1}$$

for comparative analyses $(x_{cFL_{\min}}\ vs\ x_{T_{\min}})$ where $x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$. Since, the number of characterised class C enzymes was small ($n = 4$), a larger MSA, which included 3D models of putative class C enzymes ($n = 92$) was generated. The eigenvalues of the lowest invariant core ($0 < V(Ang) \leq 1.0$) were then investigated with principal component analysis (PCA), which in turn was used to cluster and identify structural homologues of characterised class C enzymes. The aligned models were thence utilised to infer plausible active-site architecture(s) of plant class C enzymes.

Structural analysis of 3D models of plant class C GH9 enzymes

Low frequency (ω) and non-trivial normal modes (NM) ($\omega(NM) > 0, NM > 6; \omega \in \mathbb{R}, NM \in \mathbb{N}$) of the superposed 3D models as well as individual protein sequences of the minimised $(NM(x_{FL_{\min}}) = NM_{x_{FL_{\min}}}, NM(x_{T_{\min}}) = NM_{x_{T_{\min}}}, NM(x_{cFL_{\min}}) = NM_{x_{cFL_{\min}}})$ and 40.1 *ns* MD trajectories $\left(NM(x_{FL_{40.1ns}}) = NM_{x_{FL_{40.1ns}}}\right)$ was done [67, 71, 72]. Each normal mode investigated was an eigenvector and was computed from the combined oscillatory motion of the *Cα*-atoms under a generic force field and possessed a characteristic eigenvalue (Fig. 1). As discussed vide supra, the corrected subset (x_{cFL}) of each protein was used for comparative analyses $(x_{cFL_{\min}}\ vs\ x_{T_{\min}})$ where $x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$. A modified rmsf-based score $(\mathrm{rmsf}(x_{cFL_{\min}}) = \mathrm{rmsf}_{x_{cFL_{\min}}}, \mathrm{rmsf}(x_{T_{\min}}) = \mathrm{rmsf}_{x_{T_{\min}}})$ was formulated as under:

$$\Delta\mathrm{rmsf}_{x_{cFL_{\min}}} = \max\left(\mathrm{rmsf}_{x_{cFL_{\min}}}\right) - \min\left(\mathrm{rmsf}_{x_{cFL_{\min}}}\right) \tag{2}$$

$$\Delta\mathrm{rmsf}_{x_{T_{\min}}} = \max\left(\mathrm{rmsf}_{x_{T_{\min}}}\right) - \min\left(\mathrm{rmsf}_{x_{T_{\min}}}\right) \tag{3}$$

These, in tandem with the standard deviation $(\sigma_{\mathrm{rmsf}}(x_{cFL_{\min}}, x_{T_{\min}}))$, were used to assess and compare the influence of atomic motion on the structural organization of characterised class C proteins. The presence of correlated displacements of residues for each full length protein after the MD run $(x_{FL_{40.1ns}})$ was also examined by the dynamic cross correlation map (DCCM), i.e. the covariance matrix of the root mean square fluctuations $\left(\mathrm{rmsf}(x_{FL_{40.1ns}}) = \mathrm{rmsf}_{x_{FL_{40.1ns}}}\right)$ of every *Cα* atom of each class C protein $(\mathrm{cov}(x_{FL_{40.1ns}}, x_{FL_{40.1ns}}))$ $\forall\, x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$ (Fig. 1). These investigations were complemented by computing the surfaces, cavities, and, grooves present in the GH9, linker, and CBM49 regions or at their interfaces using the SPDBV (Swiss protein data bank viewer) suite of programs (https://spdbv.vital-it.ch) (Fig. 1) [73]. A cylinder of minimum area and volume was used to model and thence approximate the dimensions (radius $:= r$, height $:= h$, length $:= l$; $r, h, l \in \mathbb{R}_+$) of the predicted substrate binding and cleaving groove(s) necessary to

accommodate and digest crystalline. These formulas were derived and are as under:

$$A_o \cong \varnothing + A_c = (2)(\pi)(r)(r+h) \quad (4)$$

$$V_o \cong \beta + V_c = (\pi)(r^2)(h) \quad (5)$$

Differentiating w, r, t, h and solving for r and h results in the formulae

$$r = \sqrt{A_o/(4)(\pi)} \quad (6)$$

$$h = (4)(V_o)/A_o \quad (7)$$

$$l = A_0/r \quad (8)$$

A_o := Computed area of wide groove (Ang^2)
V_o := Computed volume of wide groove (Ang^3)
r := Radius of approximating cylinder
h := Height of approximating cylinder
l := Length of groove (Ang)
A_c := Computed area of approximating cylinder (Ang^2)
V_c := Computed volume of approximating cylinder (Ang^3)
$\varnothing$:= Constant of approximation $(Area)$
β := Constant of approximation $(Volume)$

The difference data, i.e. $\varnothing = |A_o - A_c|$; $\beta = |V_o - V_c|$, was then used to quantify and characterise this approximation.

Ligand preparation and utilization

The degree of polymerization (DP) was utilised to shortlist potential candidates of cellulose oligomers ($2 \leq DP \leq 8$) and their stereoisomers, from the ZINC12 and PubChem databases (http://www.ncbi.nlm.nih.gov/pubchem;http://zinc.docking.org) [74, 75]. Briefly, for $2 \leq DP \leq 4$ ($n = 3$) and for $5 \leq DP \leq 8$ ($n = 1$) were utilised ($n = 13 = 3 * (3) + 4$) for this analysis (Fig. 1, Table 2). The ligands were downloaded in the isomeric SMILES format and built with ChemSketch installed locally. Geometry isomerization was initially performed with Chemsketch itself, followed by a further $500 - 2000$ cycles of optimization with the steepest descent and the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithms [76]. These were implemented with a local installation of Arguslab using the universal force field (UFF) parameter of the molecular mechanics component (http://www.arguslab.com/arguslab.com) [77]. Additional relevant parameters for this step were the cutoff for non-bonded interactions (8.0 Ang) and data updates after every 20 steps. The optimization converged for all the ligands tested with a net energy of < -8 Kcal mol$^{-}Ang^2$. The xyz coordinates along with other relevant information was encoded as a pdb file and uploaded to the DockingServer (https://www.dockingserver.com/web) [78]. The geometry of all the ligands ($n = 13$) uploaded were finally optimised using the semi-empirical (PM6) method of partial charge addition, the Merck molecular force field (MMFF94), with all rotatable bonds delineated and non-polar hydrogen atoms merged [79, 80].

Docking experiments of characterised plant class C GH9 endoglucanases

3D models of characterised plant class C GH9 endoglucanases $(x_{FL_{40.1ns}}; x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\})$ were uploaded to the DockingServer (https://www.dockingserver.com/web) [78]. The server with the aid of AutoDock, added the necessary hydrogens, atomic charges, and utilised a grid of $100 \times 100 \times 100$ points with a spacing of 0.375 Ang [81]. The final positions of the coordinates on this grid were modified to include the previously delineated interaction surfaces of GH9, linker, and CBM49, for all the proteins. Computation of the non-covalent bonds (van der Waals, electrostatics) was accomplished using the parameter set from AutoDock. Docking was performed using the Lamarckian genetic algorithm and a local search method after the initial position, orientation, and torsion angles of the ligand molecules were set randomly [81, 82]. Data for a single experiment was derived from 100 different runs (Δtranslation = 0.2 Ang; Δtorsion = Δquaternion = 5). These were set to terminate after a previously set limit of energy evaluations (E_evals =2500000, population =150). The contribution of these residues to the catalysis of crystalline cellulose was inferred from the free energy $\left(x(\Delta G_y) = x_{\Delta G_y}\right)$ and constant of inhibition $\left(x(Ki_y) = x_{Ki_y}\right)$ $\forall x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$; $y \in \{C21, C22, C23, C31, C32, C33, C41, C5, C6, C7, C8\}$.

Results

Data organization and arrangement

A pipeline comprising each step and the relevant data generated are presented as under the following steps:

Step 0: Parameters were defined for protocols to minimise, equilibriate, and preliminarily characterise 3D models of plant class C GH9 endoglucanases and ligands of cellulose (Fig. 1, Tables 1 and 2)
Step 1: The 3D fold of sequences of characterised (full length, truncated) and putative plant class C GH9 endoglucanases was determined (Figs. 1 and 2, Table 3; Supplementary Text 1).
Step 2: The 3D models of characterised class C enzymes were minimised and used to assess contributions of the linker and CBM49 to the structural integrity of protein (potential energy calculations, rms deviation, normal mode analysis, root mean square fluctuations) (Fig. 3, Table 4; Supplementary Texts 2–5).

Step 3: The minimised full length 3D models of characterised class C enzymes were perturb, equilibriate (300K; 120*ps*), and simulated with a molecular dynamics run (300K; 40.1 *ns*) (Fig. 4, Supplementary Text 6).

Step 4: The MD simulated characterised class C plant GH9 endoglucanases were analysed (invariant core analysis, surface contact analysis, cavity and groove delineation, normal mode analysis, docking) to garner insights into the architecture and composition of putative active sites (Figs. 5, 6, and 7, Tables 5, 6, 7, and 8; Supplementary Texts 7–9).

Step 5: Structural homologues of selected characterised and putative class C enzymes were identified with a PCA-based clustering schema and analysed to derive insignts into the mechanism(s) of digesting crystalline cellulose by plant class C GH9 endoglucanases (Figs. 8 and 9, Table 9; Supplementary Text 10).

Table 1 Parameters for minimizing, equilibrating, and simulating 3D structures of characterised class C GH9 endoglucanases

		Min_1	Min_2	Equil_1	Equil_2	40.1 ns
Algorithm		SD	SD	Shake	Shake	
imin		1	1	0	0	0
irest		0	0	0	1	–
maxcyc		5000	5000	–	–	–
ncyc		10000	10000	–	–	–
ntc		–	–	2.0	2.0	–
ntf		–	–	2.0	2.0	–
ntr		1	0	1	0	–
Force (Kcal mol^{-1}) Ang^2)		100.0	–	10.0	–	–
ntt		–	–	3.0	3.0	–
tempi (K)		–	–	0	300.0	300.0
temp0 (K)		–	–	300.0	300.0	300.0
igb		0	0	0	0	–
cut (Ang)		12.0	12.0	10.0	10.0	10.0
nstlim				10000	50000	20,050,000
dt				0.002	0.002	0.002
ntb		–	–	1	2	–
ntp		–	–	0.0	1.0	–
pres0		–	–	–	1.0	
taup		–	–	–	2.0	
ntpr		100	100	1000	10000	10000
ntwx		–	–	1000	10000	10000
ntwr		–	–	5000	10000	10000
Dielectric		–	–	–	–	1.0
Simulation space partitioning	Switching	–	–	–	–	On
	Switchdist (Ang)					9
	Pairlistdist (Ang)					12
	exclude					Scaled 1–4
	1–4scaling					1.0
firsttimestep	Timestep	–	–	–	–	1
	Stepspercycle					20
	nonbondedFreq					2
	fullElectFrequency					4
Langevin dynamics	Langevin	–	–	–	–	On
	LangevinDamping					1.0
	LangevinTemp					300
	LangevinHydrogen					No
	LangevinPiston					On
	LangevinPistonTarget					1.01325
	LangevinPistonPeriod					2000 fs
	LangevinPistonDecay					1000 fs
	LangevinPistonTemp					300
	useFlexibleCell					No
	useGroupPressure					No
	fixedAtomsForces					Off
Cell basis vectors	CellBasisVector1	–	–	–	–	x-coord,0,0
	CellBasisVector2					0,y-coord,0
	CellBasisVector3					0,0,z-coord
	wrapAll	–	–	–	–	No
	dcdUnitCell					Yes

Table 2 Ligands utilised in docking experiments

Formula	Database	ID	LID	MW	Isomeric SMILES
$C_{12}H_{22}O_{11}$	ZINC12	53683219	C21	342.297	C([C@H]1[C@@H]([C@H]([C@@H]([C@@H](O1)O[C@@H]2[C@@H](O[C@H]([C@@H]([C@H]2O)O)O)CO)O)O)O)O)O
		3978744	C22	342.297	C([C@@H]1[C@H]([C@@H]([C@H]([C@@H](O1)O[C@@H]2[C@H](O[C@@H]([C@@H]([C@H]2O)O)O)CO)O)O)O)O)O
		4097113	C23	342.297	C([C@@H]1[C@H]([C@@H]([C@H]([C@@H](O1)O[C@@H]2[C@H](O[C@H]([C@@H]([C@H]2O)O)O)CO)O)O)O)O
$C_{18}H_{32}O_{16}$	ZINC12	53683224	C31	504.438	C([C@@H]1[C@H]([C@@H]([C@H]([C@@H](O1)O[C@@H]2[C@@H](O[C@H]([C@@H]([C@H]2O)O)O[C@@H]3[C@@H](O[C@H]([C@@H]([C@H]3O)O)O)CO)CO)O)O)O)O
		8216112	C32	504.438	C([C@@H]1[C@H]([C@@H]([C@H]([C@@H](O1)O[C@@H]2[C@H](O[C@H]([C@@H]([C@H]2O)O)O[C@@H]3[C@H](O[C@@H]([C@@H]([C@H]3O)O)O)CO)CO)O)O)O)O
		8220386	C33	504.438	C([C@@H]1[C@H]([C@@H]([C@H]([C@@H](O1)O[C@@H]2[C@H](O[C@H]([C@@H]([C@H]2O)O)O[C@@H]3[C@H](O[C@H]([C@@H]([C@H]3O)O)O)CO)CO)O)O)O)O
$C_{24}H_{42}O_{22}$	ZINC12	96006066	C41	682.578	C([C@H]1[C@@H]([C@H]([C@@H]([C@@H](O1)O[C@@H]2[C@@H](O[C@H]([C@@H]([C@H]2O)O)O[C@@H]3[C@@H](O[C@H]([C@@H]([C@H]3O)O)O)O[C@@H]4[C@@H](OC([C@@H]([C@H]4O)O)(O)O)CO)CO)CO)O)O)O)O
		85603797	C42	666.579	C([C@@H]1[C@H]([C@@H]([C@H]([C@@H](O1)O[C@@H]2[C@H](O[C@H]([C@@H]([C@H]2O)O)O[C@@H]3[C@H](O[C@H]([C@@H]([C@H]3O)O)O[C@@H]4[C@H](O[C@H]([C@@H]([C@H]4O)O)O)CO)CO)CO)O)O)O)O
		87528241	C43	666.579	C([C@@H]1[C@H]([C@@H]([C@H]([C@@H](O1)O[C@@H]2[C@H](O[C@H]([C@@H]([C@H]2O)O)O[C@@H]3[C@H](O[C@H]([C@@H]([C@H]3O)O)O[C@@H]4[C@H](O[C@@H]([C@@H]([C@H]4O)O)O)CO)CO)CO)O)O)O)O
$C_{30}H_{52}O_{26}$	ZINC12	52940142	C5	828.7183	C([C@@H]1[C@H]([C@@H]([C@H]([C@@H](O1)O[C@@H]2[C@H](O[C@H]([C@@H]([C@H]2O)O)O[C@@H]3[C@H](O[C@H]([C@@H]([C@H]3O)O)O[C@@H]4[C@H](O[C@H]([C@@H]([C@H]4O)O)O[C@H]([C@@H](CO)O)[C@@H]([C@H](C=O)O)O)CO)CO)CO)O)O)O)O
$C_{36}H_{62}O_{31}$	PUBCHEM	74539963	C6	990.85888	C([C@@H]1[C@H]([C@@H]([C@H]([C@@H](O1)O[C@@H]2[C@H](O[C@H]([C@@H]([C@H]2O)O)O[C@@H]3[C@H](O[C@H]([C@@H]([C@H]3O)O)O[C@@H]4[C@H](O[C@H]([C@@H]([C@H]4O)O)O[C@@H]5[C@H](O[C@H]([C@@H]([C@H]5O)O)O[C@@H]6[C@H](OC([C@@H]([C@H]6O)O)O)CO)CO)CO)CO)O)O)O)O
$C_{42}H_{72}O_{36}$	PUBCHEM	440947	C7	1152.99948	C([C@@H]1[C@H]([C@@H]([C@H]([C@@H](O1)O[C@@H]2[C@H](O[C@H]([C@@H]([C@H]2O)O)O[C@@H]3[C@H](O[C@H]([C@@H]([C@H]3O)O)O[C@@H]4[C@H](O[C@H]([C@@H]([C@H]4O)O)O[C@@H]5[C@H](O[C@H]([C@@H]([C@H]5O)O)O[C@@H]6[C@H](O[C@H]([C@@H]([C@H]6O)O)O[C@@H]7[C@H](O[C@H]([C@@H]([C@H]7O)O)O)CO)CO)CO)CO)CO)CO)O)O)O)O
$C_{48}H_{80}O_{40}$	PUBCHEM	5287407	C8	1297.1248	C([C@@H]1[C@@H]2[C@@H]([C@H]([C@H](O1)O[C@@H]3[C@H](O[C@@H]([C@@H]([C@H]3O)O)O[C@@H]4[C@H](O[C@@H]([C@@H]([C@H]4O)O)O)O[C@@H]5[C@H](O[C@@H]([C@@H]([C@H]5O)O)O[C@@H]6[C@H](O[C@@H]([C@@H]([C@H]6O)O)O)O[C@@H]7[C@H](O[C@@H]([C@@H]([C@H]7O)O)O[C@@H]8[C@H](O[C@@H]([C@@H]([C@H]8O)O)O[C@@H]9[C@H](O[C@H](O2)[C@@H]([C@H]9O)O)CO)CO)CO)CO)CO)CO)O)O)O

ZINC, zinc is not commercial; *LID*, ligand identity; *MW*, molecular weight

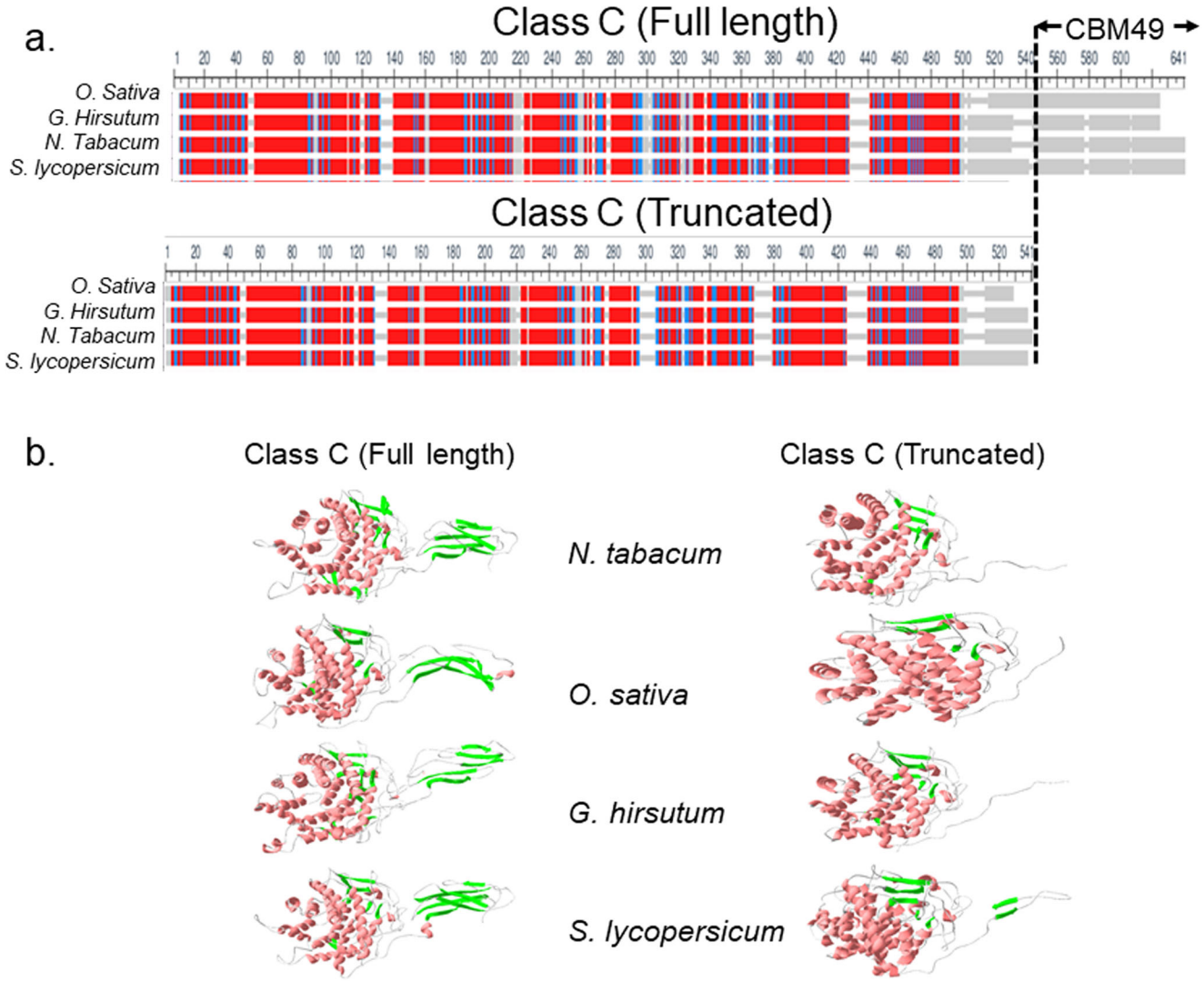

Fig. 2 3D models of full length and truncated plant class C endoglucanases. **a** Full length ($GH9 \cup L \cup CBM49$) and truncated ($GH9 \cup L$) sequences of characterised ($n = 4$; *Oryza sativa*; *Gossypium hirsutum*; *Solanum lycopersicum*; *Nicotiana tabaccum*) plant class C GH9 endonucleases along with full-length sequences of putative class C enzymes ($n = 92$) were submitted to Phyre2; **b** The 3D models that represented the best approximation to the template X-ray structures *Thermomonospora fusca* (PDB: 1JS4; *UID* : *Q8LJP6*) and *Clostridium cellulolyticum* (PDB: 1GA2; *UIDs* : *Q5NAT0*, *Q9ZSP9*, *Q93WY9*) were used for all further investigations. The parameters used to evaluate these were sequence identity, presence of an homologous structure (confidence), and the percentage of the protein that could be modelled (coverage). Abbreviations—GH9, glycoside hydrolase; L, linker sequence; CBM49, carbohydrate binding module; MUSCLE, multiple sequence comparison by log-expectation; PDB, protein data bank; Phyre2, protein homology/analogy recognition engine; *UID* : *Q5NAT0*, *O. sativa*; *UID* : *Q8LJP6*, *G. hirsutum*; *UID* : *Q93WY9*, *N. tabacum*; *UID* : *Q9ZSP9*, *S. lycopersicum*

Homology modelling and assessment of characterised class C GH9 endoglucanases

An intersequence pairwise alignment suggests that despite a high degree of identity ($\approx 75 - 83\%$) between the class C enzymes of *S. lycopersicum*, *G. hirsutum*, and *N. tabacum*, the preferred template for *G. hirsutum* was from *T. fusca* (*PDBID* : 1*JS*4). Conversely, the sequence identity for *O. sativa* was marginally lower ($\approx 62\%$ *identity*), yet shared the same top ranked template, i.e. *C. cellulolyticum* (*PDBID* : 1*GA*2), with *S. lycopersicum* and *N. tabacum* (Table 3; Supplementary Text 1). However, the average sequence identity with the templates ($\approx 32 - 40\%$) was similar for all class C enzymes investigated (Table 3). The superposed ungapped MSA of the truncated (x_T) class C proteins additionally resulted in the exclusion of the linker, i.e. $CBM49 \equiv CBM49 \cup L$, from the MSA, i.e. $x_T = GH9 - CBM49 = GH9 - (CBM49 \cup L)$; $x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$, (Fig. 2). The results (rmsd (template, x) < 2 *Ang*) suggest that the catalytic machinery for digesting crystalline may be conserved in plants and other non-plant taxa most notably bacteria (Table 3; Supplementary Text 1) [8, 17, 20–34, 59, 60]. The models

Table 3 Fold identification by homology modelling of plant GH9 endoglucanases

UID	PDBID	Organism	R(Ang)	Cv(%)	SI(%)	Co(%)	Range	Ref.
Q8LJP6	1JS4	*T. fusca*	2.0	92	37	100	26–600	60
Q5NAT0	1GA2	*C. cellulolyticum*	1.7	90	33	100	39–620	59
Q9ZSP9			1.7	89	32	100	23–621	
Q93WV9			1.7	94	33	100	31–620	

UID, uniprot identity; *PDBID*, protein data bank identity; *R*, root mean squared deviation; *Cv*, coverage; *SI*, sequence identity; *Co*, confidence; *Ref*, reference

also indicate that in addition to GH9, CBM49 and the linker (coverage = 89 − 94%) may partake in digesting crystalline cellulose (Table 3) [8, 17, 59, 60]. Since, solvent addition was explicit, minimization of energy ($E_{\min}$) was carried out exclusively by the steepest descent algorithm ($ncyc > maxcyc$) for the full length ($E_{\min}(Q5NAT0_{FL_{\min}}) \cong -4.36*10^5$ kcal mol^{-1}; $E_{\min}(Q93WY9_{FL_{\min}}) \cong -4.34*10^5$ kcal mol^{-1}; $E_{\min}(Q8LJP6_{FL_{\min}}) \cong -4.31*10^5$ kcal mol^{-1} ; $E_{\min}(Q9ZSP9_{FL_{\min}}) \cong -4.29*10^5$ kcal mol^{-1}) and truncated ($E_{\min}(Q5NAT0_{T_{\min}}) \cong -2.58*10^5$ kcal mol^{-1}; $E_{\min}(Q93WY9_{T_{\min}}) \cong -3.61*10^5$ kcal mol^{-1}; $E_{\min}(Q8LJP6_{T_{\min}}) \cong -3.49*10^5$ kcal mol^{-1}; $E_{\min}(Q9ZSP9_{T_{\min}}) \cong -3.70*10^5$ kcal mol^{-1}

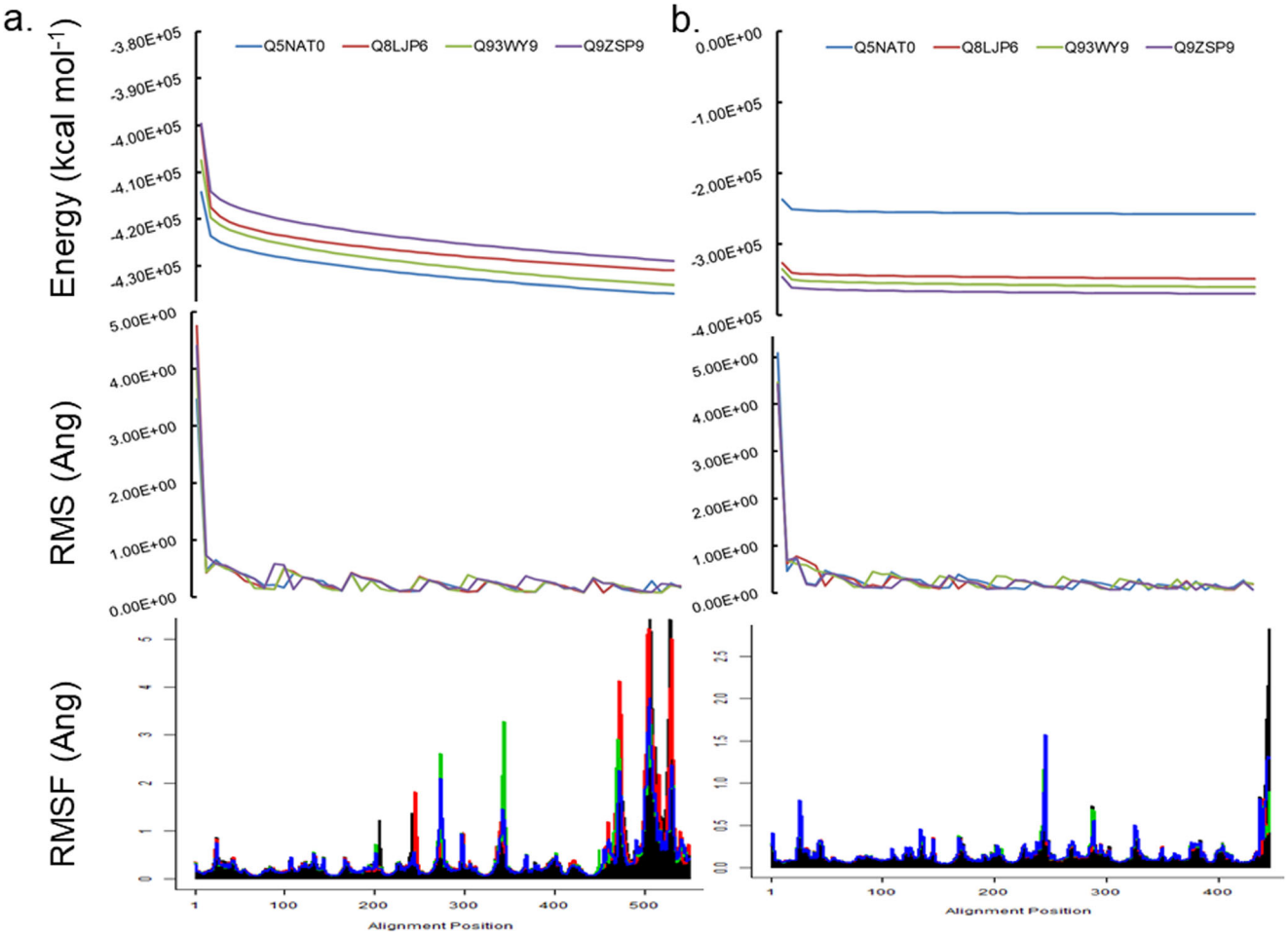

Fig. 3 Comparative analyses of full length and truncated 3D models of class C GH9 endoglucanases. **a**, **b** Energy minimization ($E_{min} < 0.0$) of 3D models of full length and truncated ($x_{FL_{min}}, x_{T_{min}}; x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$) characterised class C GH9 endoglucanases was carried out and monitored by the root mean squared deviation of the intermediate structures. The absence of significant variation of the total (ETOT) energy for the models studied suggested that these were stable and could be examined further. **c** Normal mode analysis of these minimised models suggested that the carboxy-terminal end of the linker region and the CBM49 regions experienced increased oscillatory motion, an observation which is mitigated when these were truncated. The frequencies for *O. sativa* ($Q5NAT0_{T_{\min}} \gg Q5NAT0_{FL_{\min}}$) and was more pronounced for the lower non-trivial modes as opposed to the proteins from *S. lycopersicum* and *N. tabacum* ($x_{T_{\min}} < x_{FL_{\min}}; x \in \{Q8LJP6, Q93WY9, Q9ZSP9\}$). Abbreviations—GH, glycoside hydrolase; CBM, carbohydrate binding module; FL, full length; T, truncated; *UID* : *Q5NAT0*, *O. sativa*; *UID* : *Q8LJP6*, *G. hirsutum*; *UID* : *Q93WY9*, *N. tabacum*; *UID* : *Q9ZSP9*, *S. lycopersicum*

Table 4 Frequencies of non-trivial low frequency modes of 3D models of characterised and minimised class C enzymes

x	Q5NAT0	Q8LJP6	Q93WY9	Q9ZSP9
$x_{FL\min}$				
NM	1704	1680	1728	1755
7	0.003	0.003	0.003	0.003
8	0.003	0.004	0.004	0.004
9	0.005	0.006	0.006	0.006
10	0.009	0.010	0.009	0.009
11	0.011	0.011	0.011	0.012
12	0.012	0.013	0.012	0.013
13	0.013	0.013	0.014	0.014
14	0.013	0.015	0.015	0.015
15	0.015	0.015	0.016	0.016
16	0.016	0.017	0.017	0.016
17	0.016	0.017	0.017	0.018
18	0.018	0.018	0.019	0.019
$x_{T\min}$				
NM	1368	1392	1398	1428
7	0.012	0.001	0.001	0.002
8	0.015	0.001	0.001	0.002
9	0.016	0.001	0.001	0.004
10	0.017	0.004	0.004	0.007
11	0.018	0.004	0.005	0.009
12	0.021	0.006	0.006	0.011
13	0.022	0.007	0.007	0.015
14	0.022	0.009	0.009	0.015
15	0.023	0.009	0.010	0.016
16	0.023	0.011	0.011	0.017
17	0.024	0.014	0.014	0.018
18	0.025	0.015	0.014	0.020

GH9, glycoside hydrolase 9; *M*, normal modes; $x_{FL\min}$, full length minimised 3D model of class C sequence; $x_{T\min}$, truncated minimised 3D model of class C sequence; *Q5NAT0*, *Oryza sativa*; *Q8LJP6*, *Gossypium hirsutum*; *Q93WY9*, *Nicotiana tabacum*; *Q9ZSP9*, *Solanum lycopersicum*

) models (Fig. 3, Tables 1 and 3; Supplementary Text 3). Interestingly, whilst, the data $(\text{Rank}(E_{\min}(x_{FL_{\min}})) = \text{Rank}(E_{\min}(x_{T_{\min}})) = \{2,3\}; x = \{Q93WY9, Q8LJP6\})$ were consistent for *N. tabacum* and *G. hirsutum*, there was a complete reversal of the same for *O. sativa* and *S. lycopersicum* $(\text{Rank}(E_{\min}(Q5NAT0_{FL_{\min}}, Q9ZSP9_{T_{\min}})) \propto 1/\text{Rank}(E_{\min}(Q5NAT0_{T_{\min}}, Q9ZSP9_{FL_{\min}})))$ (Fig. 3; Supplementary Text 3). These data suggest that full length class C enzymes may adopt a stable conformation earlier than their truncated counterparts. Interestingly, the rms deviations of the minimised full length class C enzymes from *O. sativa* $(E_{\min}(Q5NAT0_{FL_{\min}})/E_{\min}(Q5NAT0_{T_{\min}}) \cong 2.31)$, *G. hirsutum* $(E_{\min}(Q8LJP6_{FL_{\min}})/E_{\min}(Q8LJP6_{T_{\min}}) \cong 1.05)$, and *S. lycopersicum* $(E_{\min}(Q9ZSP9_{FL_{\min}})/E_{\min}(Q9ZSP9_{T_{\min}}) \cong 2.84)$ were higher as compared with the truncated forms while the reverse was observed for *N. tabacum* $(E_{\min}(Q93WY9_{FL_{\min}})/E_{\min}(Q93WY9_{T_{\min}}) \cong 0.89)$ (Fig. 3; Supplementary Text 3).

Assessing the contribution of CBM49 to the structural integrity of class C enzymes

The core data for the 3D models of all full length characterised plant class C enzymes suggests that while GH9 is well conserved $(\#C\alpha_{0.0<V\leq 100.0}(GH9) > 0)$, CBM49 is not $(\#C\alpha_{0.0<V\leq 100.0}(CBM49) = 0)$. The N- and C-terminal regions of the linker does, however, exhibit partial conservation $(\#C\alpha_{8.0<V\leq 100.0}(Linker) = \{1,3\})$, a trend which is unlikely to be sustained for larger datasets (Supplementary Text 2). Low frequency non-trivial modes, i.e. $NM_{x_{FL\min}} = NM_{x_{T\min}} = 7\text{–}18$, were also assessed to garner additional information about the possible role(s) of CBM49 and the linker in influencing structure of the GH9 (Table 4; Supplementary Texts 4 and 5). With the exception of *O. sativa*, the frequencies of these modes for all other full length $(\Delta\omega(NM_{x_{FL\min}}))$ class C members (*G. hirsutum*, *N. tabacum*, *S. lycopersicum*) were $\approx 2-3$ fold higher than those for their truncated forms $(\Delta\omega(NM_{x_{FL\min}}) > (\Delta\omega(NM_{x_{T\min}}))$ (Table 4; Supplementary Texts 4 and 5). The frequency of these for the truncated models $(\Delta\omega(NM_{x_{T\min}}))$ of *O. sativa* in general was $\approx 2-5$ fold higher for all modes examined or as in *S. lycopersicum* for the higher frequency modes $(\Delta\omega(NM_{x_{FL\min}}) < (\Delta\omega(NM_{x_{T\min}}))$ (Table 4; Supplementary Texts 4 and 5). The atomic fluctuation data for *G. hirsutum* $(\Delta\text{rmsf}_{x_{cFL\min}} \cong 1.74, \sigma(\text{rmsf}_{x_{cFL\min}}) \cong 0.21; \Delta\text{rmsf}_{x_{T\min}} \cong 758.00, \sigma(\text{rmsf}_{x_{T\min}}) \cong 41.12)$, *N. tabacum* $(\Delta\text{rmsf}_{x_{cFL\min}} \cong 3.28, \sigma(\text{rmsf}_{x_{cFL\min}}) \cong 0.29; \Delta\text{rmsf}_{x_{T\min}} \cong 673.61, \sigma(\text{rmsf}_{x_{T\min}}) \cong 37.56)$, and *S. lycopersicum* $(\Delta\text{rmsf}_{x_{cFL\min}} \cong 2.1, \sigma(\text{rmsf}_{x_{cFL\min}}) \cong 0.22; \Delta\text{rmsf}_{x_{T\min}} \cong 46.29, \sigma(\text{rmsf}_{x_{T\min}}) \cong 4.30)$, exhibited greater variance as compared with the full length proteins, i.e. $\sigma(\text{rmsf}_{x_{T\min}}) >> \sigma(\text{rmsf}_{x_{cFL\min}})$ (Fig. 3; Supplementary Texts 4 and 5). Interestingly, the corresponding data for *O. sativa* only differed marginally $(\Delta\text{rmsf}_{x_{cFL\min}} \cong 2.49, \sigma(\text{rmsf}_{x_{cFL\min}}) \cong 0.26; \Delta\text{rmsf}_{x_{T\min}} \cong 2.77, \sigma(\text{rmsf}_{x_{T\min}}) \cong 0.22)$ (Fig. 3; Supplementary Texts 4 and 5). The baseline rmsf values were remarkably consistent for all the proteins $(\min(\Delta\text{rmsf}_{x_{cFL\min}}) \cong \min(\Delta\text{rmsf}_{x_{T\min}}) \cong 0.07)$ examined, although for *O. sativa* there was a tangible difference, i.e. $\min(\Delta\text{rmsf}_{x_{cFL\min}}) \cong 0.07, \min(\Delta\text{rmsf}_{x_{T\min}}) \cong 0.05$ (Table 4; Supplementary Texts 4 and 5). A position-specific analysis of this data clearly demonstrates that this heightened oscillatory motion involves the residues of the linker and CBM49 (Fig. 3, Table 4; Supplementary Texts 4 and 5). These data when combined suggests that CBM49 and the linker, despite being poorly conserved even amongst class C members, may deploy corrective hypermobility to rapidly restore equilibrial status secondary to perturbation events such

as that observed for substrate binding and subsequent catalysis by enzymes.

Delineating the active site architecture of characterised plant class C enzymes

An multi-modal approach (surface contact analysis, docking, cavity and groove delineation) was adopted to ascertain the residues and their relevance to crystalline cellulose digestion by plant class C enzymes.

Analysing the DCCM to assess and characterise intra-protein residue interactions The NMA and DCCM data of mature-folded (40.1 *ns*) class C enzymes suggest that several residues that comprise the non-contiguous segments between the GH9, linker, and CBM49 exhibit positively correlated atomic displacements ($r \cong 1.00$) (Fig. 4, Supplementary Texts 6–9). These data imply that plant class C enzymes, like their bacterial counterparts may also possess well-defined interaction surface(s) $\left(IS = \left\{IS_x^{GC}, IS_x^{CL}, IS_x^{GL}\right\}\right)$ between GH9, linker, and CBM49 (Fig. 5) [59, 60]. The surface area of interacting residues was variable and ranged from $375 - 517\ Ang^2$

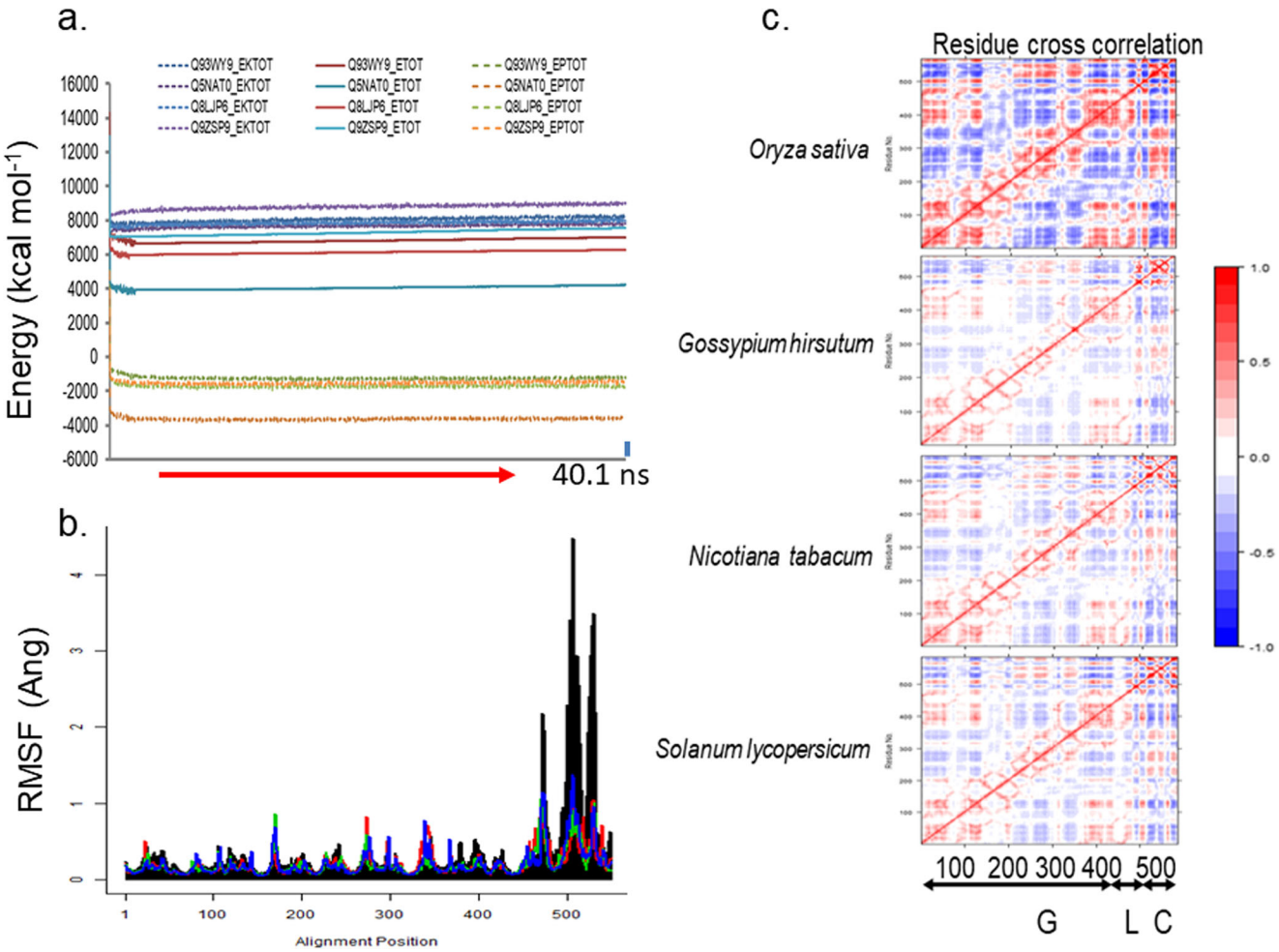

Fig. 4 Structural analyses of 3D models of class C GH9 endoglucanases. **a**, **b** 3D models of full length characterised class C GH9 endoglucanases, at equilibrium, were monitored by the root mean squared deviation of the intermediate structures, and the absence of significant variation of the kinetic (EKTOT), potential (EPTOT), and thence the total (ETOT) energies. **c** Normal mode analysis and root means square fluctuations (*rmsf*) suggested that the carboxy-terminal end of the linker region and/or CBM49 regions are flexible and may contribute to an adaptable active site geometry and **d** dynamic cross-correlation map of residues of characterised class C GH9 enzymes. The coavariance matrix of the *rmsf* values of each residue per protein was computed. The dynamic cross correlation map $\left(cov\left(rmsf_{x_{FL_{40.1ns}}}, rmsf_{x_{FL_{40.1ns}}}\right)\ \forall x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}\right)$ was examined for areas of positive correlation (red, $r \rightarrow 1.00$) across all modes of vibrational motion. The off-diagonal data suggests that like the bacterial templates, class C enzymes may also possess different non-contiguous segments (G, L, C) whose atomic displacements might be correlated. Evaluating the positive r–coefficients suggests the existence of multiple interaction surfaces between these. Abbreviations—*C*, carbohydrate binding module 49; cov, covariance matrix; *G*, glycoside hydrolase 9; *L*, linker; *r*, correlation coefficient; GH, glycoside hydrolase; CBM, carbohydrate binding module; FL, full length; *UID* : *Q5NAT0*, *O. sativa*; *UID* : *Q8LJP6*, *G. hirsutum*; *UID* : *Q93WY9*, *N. tabacum*; *UID* : *Q9ZSP9*, *S. lycopersicum*

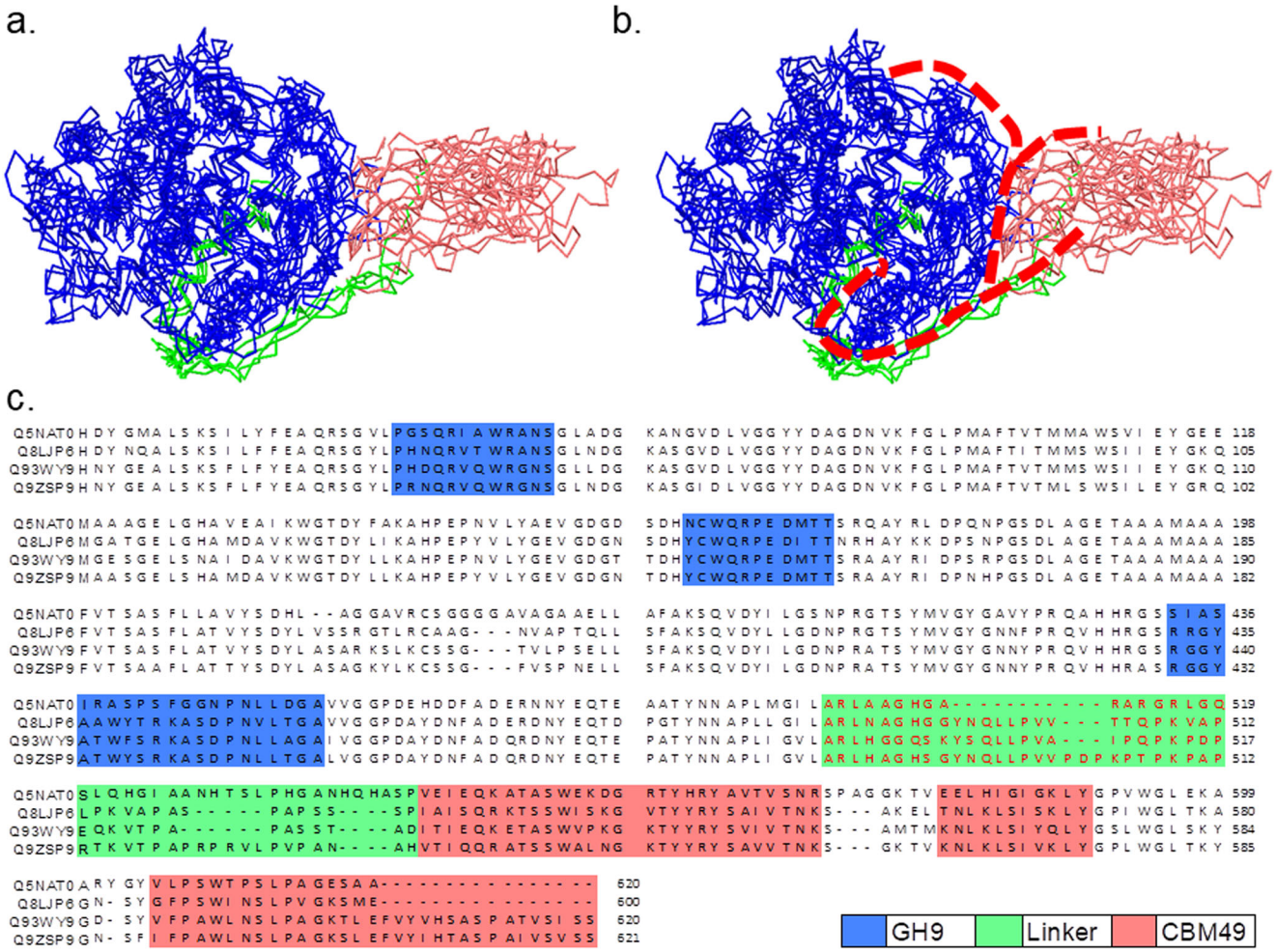

Fig. 5 Characterizing non-contiguous interacting residues of class C enzymes. The NMA- and DCCM -data of full length characterised class C enzymes $(x_{FL_{40.1ns}}; x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\})$ were analysed for the presence of interacting residues of GH9, linker, and CBM49. The MSA of these suggests that GH9 and CBM49 are comprised of three potential regions of interactions ($G1$, $G2$, $G3$; $C1$, $C2$, $C3$), by which they interact with each other as well as the intervening linker. These may be summarised as IS_x^{GC}, IS_x^{GL}, IS_x^{CL} and exceptionally IS_x^{GLC}. Here $x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$. These residues were additionally, submitted for docking with ligands of cellulose to ascertain their functional relevance. Abbreviations—CBM49≡C, carbohydrate binding module; DCCM, dynamic cross-correlation map; FL, full length; GH9≡G, glycoside hydrolase 9; IS, interaction surface; L, linker; MSA, multiple sequence alignment; NMA, normal mode analysis

(CBM49_linker≡IS_x^{CL}), 283 − 481 Ang^2 (GH9_linker≡IS_x^{GL}), and 96−208 Ang^2 (GH9_CBM49≡IS_x^{GC}) where $x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$. The surfaces themselves may be further decomposed into non-contiguous subsegments, i.e. $G = G1 \cup G2 \cup G3$ and $C = C1 \cup C2 \cup C3$. Thus, $IS_x^{GC} = G2 \cup G3 \cup C2 \cup C3$, $IS_x^{GL} = G1 \cup G2 \cup G3 \cup L$, and $IS_x^{CL} = C1 \cup C2 \cup C3 \cup L$ (Fig. 5). In general, while the contact surface formed between GH9 and CBM49 was the least, the same for CBM49 and the linker was maximal $(IS_x^{GC} < IS_x^{GL} < IS_x^{CL}, x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\})$. The only exception was for the class C enzyme from *O. sativa* $\left(IS_{Q5NAT0}^{CL} > IS_{Q5NAT0}^{GL}\right)$ which can be explained by a large interaction surface spanning GH9, CBM49, and the linker $\left(IS_{Q5NAT0}^{GLC} = G1 \cup G2 \cup G3 \cup C1 \cup L\right)$, i.e. $IS_{Q5NAT0}^{GL} \equiv IS_{Q5NAT0}^{GLC}$.

The bonds between the residues that comprised these protein-protein interaction surfaces $(AA_x^{IS}; x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\})$ were non-covalent (hydrophobic, hydrogen, van der Waals) for *N. tabacum*, *G. hirsutum*, and *S. lycopersicum*. Here, too, the contact surface for the class C enzyme from *O. sativa* was exceptional and included the possibility of a covalent and oxygen-sensitive (−*SS*−) linkage between *C*124 and *M*5/*M*132 (Fig. 5).

Docking data suggests qualitative differences between individual class C enzymes The binding energy of the ligands was lower for the higher molecular weight ligands $(x_{\Delta G_{C8}} < x_{\Delta G_{C5}} < x_{\Delta G_{C6}} \leq x_{\Delta G_{C4}} \leq x_{\Delta G_{C3}} < x_{\Delta G_{C2}} < x_{\Delta G_{C7}})$ with *C*8 possessing the lowest $(x_{\Delta G_{C8}} \cong -7.44 \text{ kcal mol}^{-1} = \min(x_{\Delta G_y})$, while interestingly, the free energy of binding for *C*7 ($x_{\Delta G_{C7}} \cong -2.67$ kcal

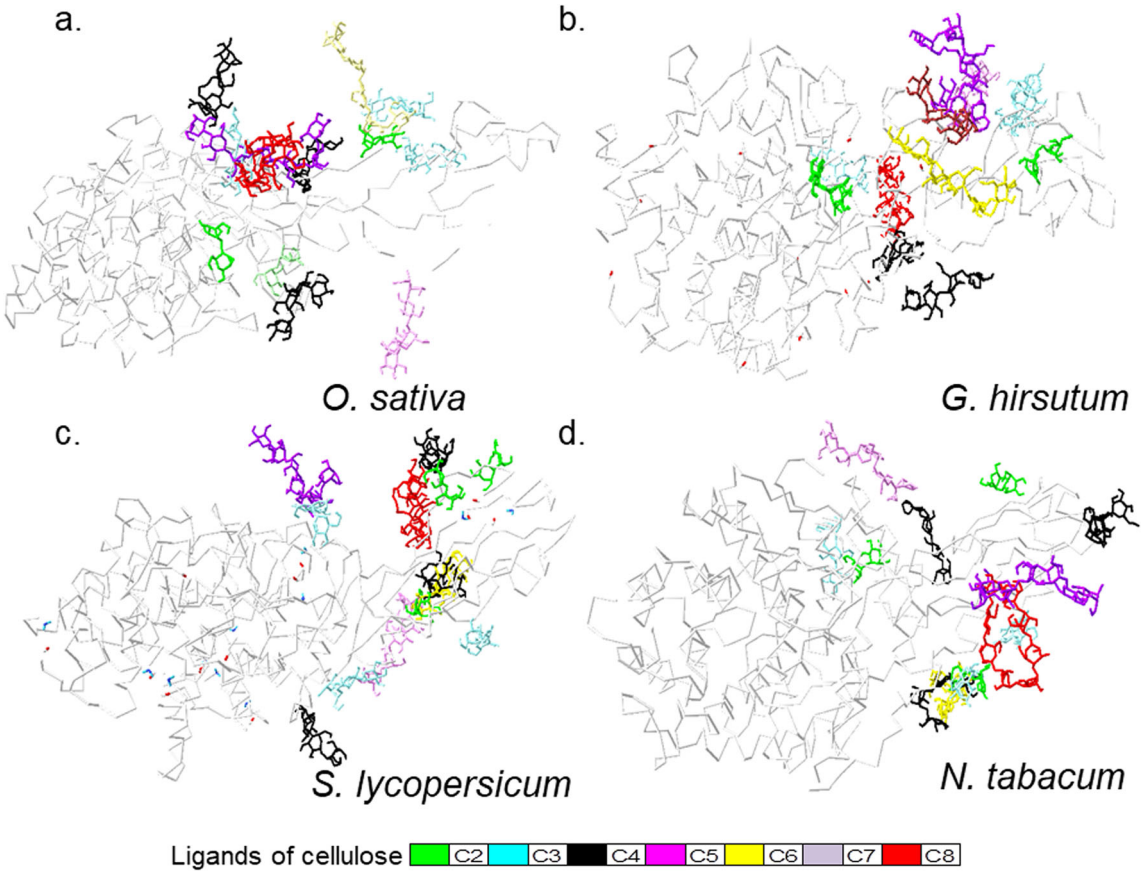

Fig. 6 Docking experiments to determine energetically favourable amino acids of characterised plant class C GH9 endoglucanases. Optimised ligands of cellulose ($2 \leq DP \leq 8$) were docked with 3D models of class C GH9 endoglucanases ($x_{FL_{40.1ns}}; x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$) using the potentially interacting surface residues of GH9, CBM49, and linker regions as potential contacts. The results were the top ranked, i.e. $\min(x_{\Delta G_y}); x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}, y \in \{C21, C22, C23, C31, C32, C33, C41, C42, C43, C5, C6, C7, C8\}$ of all considered runs. Here, despite *C*8 being the largest ligand the free energy of binding was the least $\left(\min(x_{\Delta G_y}) = \min(x_{\Delta G_{C8}}) \leq -7.36\ kcal\ mol^{-1}\right)$. In contrast, the results for *C*7 were the exact opposite $\left(\max(x_{\Delta G_y}) = \min(x_{\Delta G_{C7}}) \geq 2.97\ \text{kcal mol}^{-1}\right)$. These data while applicable to smaller ligands are unlikely to extend to the full length cellulose polymer. Here, the interactions are expected to interact uniformly with all the groove binding residues to accomplish substrate modification and catalysis. Abbreviations—CBM49, carbohydrate binding module; *DP*, degree of polymerization; GH9, glycoside hydrolase; *UID* : *Q5NAT0*, *O. sativa*; *UID* : *Q8LJP6*, *G. hirsutum*; *UID* : *Q93WY9*, *N. tabacum*; *UID* : *Q9ZSP9*, *S. lycopersicum*

$\text{mol}^{-1} = \max(\text{x}_{\Delta G_y})$) for all the class C enzymes investigated. These data were also supported by the corresponding *Ki* values, i.e. $\text{x}_{Ki_{C8}} \cong 3.56\ \mu\text{M} = \min(\text{x}_{Ki_y})$ and $\text{x}_{Ki_{C7}} \cong 7.58\ \text{mM} = \max(\text{x}_{Ki_y})$ $x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$; $y \in \{C21, C22, C23, C31, C32, C33, C41, C5, C6, C7, C8\}$) (Figs. 5 and 6, Tables 5 and 6). The distribution of specific amino acids identified by docking $\left(AA_x^{\text{Dock}} \subset AA_x^{IS}; x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}\right)$ suggests a preponderance of residues with small hydrophobic, aromatic, and basic side chains along with serine and threonine. Exceptionally, the catalytic amino acids aspartic (*D*) and glutamic (*E*) acids were almost (*D*465, *O. sativa*; *D*139, *D*451, *S. lycopersicum*) completely excluded from these calculations as were other amino acids with known proclivity to partake in catalysis, i.e. cysteine (*C*) and histidine (*H*) (Table 7).

Delineating the cavities and grooves for crystalline cellulose catalysis and modification by plant class C GH9 endoglucanases Since, solvent accessibility is a pre-requisite for hydrolytic catalysis of the glycosidic linkage by GH9 endoglucanases, the presence of amino acids identified previously by docking was examined in cavities and grooves of the 3D models of full length characterised class C enzymes. The distribution of these for *O. sativa* ($GH9 = 27, L = 0, CBM49 = 1, LC = 4, GC = 1$), *G. hirsutum* ($GH9 = 21, L = 0, CBM49 = 0, LC = 4, GC = 1$), *N. tabacum* ($GH9 = 20, L = 0, CBM49 = 0, LC = 4, GC = 0$), and *S. lycopersicum* ($GH9 = 23, L = 1, CBM49 = 2, LC = 1, GC = 0$) that CBM49/linker may function to modulate catalysis by substrate modification rather participate directly (Figs. 5 and 6, Tables 5, 6, and 7). The amino acids that comprise these were enumerated $\left(AA_x^{CvG}; x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}\right)$ and analysed

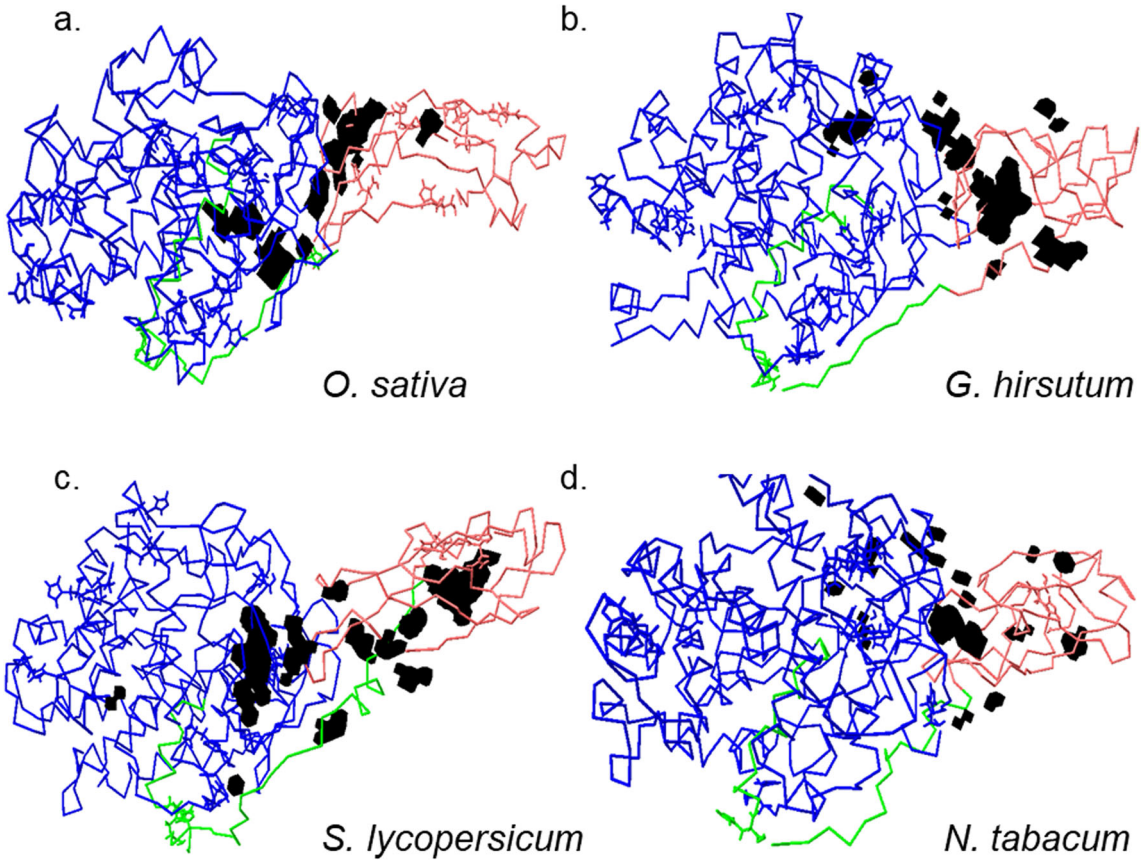

Fig. 7 Putative architecture of active site of class C GH9 endoglucanases. A combination of analytic tools were used to establish the putative active site of characterised class C enzymes. These included the amino acids that comprised the interaction surface $\left(AA_x^{IS}\right)$, resulted in energetically favourable interactions with ligands of cellulose $\left(AA_x^{Dock} \subset AA_x^{IS}\right)$, and those that formed part of numerous cavities and grooves along the surface $\left(AA_x^{CvG}\right)$ $x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$. The combined list, *i.e*, $\left(AA_x = \left(AA_x^{Dock} \cap AA_x^{CvG}\right) \subset AA_x^{IS}\right)$, was completely devoid of the aspartic (*D*), glutamic (*E*) acids, cysteine (*C*), and histidine (*H*), amino acids with known propensity for catalysis. The absence of a single continuous groove/cavity and the distribution of amino acids suggests a dual/discontinuous mode, wherein the +1 and −1 sites, are present in a subsurface cavity, while crystalline cellulose itself may interact and be modified by residues at the surface before entering the catalytic site. Despite these variations the probable length ($l \geq 100 - 200\ Ang$) of the relevant cavities and grooves suggest a well-adapted mechanism for the intact cellulose polymer. Colour codes for GH9 (blue), linker (green), and CBM49 (red), and relevant cavity and grooves (black). Abbreviations—AA, amino acids; CvG, cavities and grooves; Dock; docking experiment; *r*, *h*, *l*, radius, height, and height of groove-approximating cylinder; GH9, glycoside hydrolase 9; IS, interaction surface

(Table 7). The amino acid distribution when combined, $AA_x = \left(AA_x^{\mathrm{Dock}} \cap AA_x^{CvG}\right) \subset AA_x^{IS}; x \in \{O.sativa, G.hirsutum, N.tabacum, S.lycopersicum\}$, was utilised to compute the dimensions (length $\cong 100 - 130\ Ang$, radius $\cong 8.0 - 10.4\ Ang$, height $\cong 2.2 - 2.8\ Ang$) of a probable architecture for the active site(s) of plant class C GH9 endoglucanases (Fig. 7, Tables 7 and 8). Whilst, the volume of the approximating cylinder perfectly matched the observed data ($|V_o - V_c| = \beta \cong 0$) for all class C enzymes, the differences in the surface areas ($\varnothing \cong 250 - 510\ Ang^2$, mean $\cong 679 \pm 137.66$) could imply an intrinsic

Table 5 Dimensions of putative crystalline cellulose binding cleft of full length characterised class C enzymes after 40.1 ns MD-run

Sequence ID	Organism	A_o	V_o	r	h	l	A_c	V_c	$\Delta A = \varnothing$	$\Delta V = \beta$
Q5NAT0	*O. sativa*	905	616	8.49	2.72	106.62	597.64	616.00	307.36	0.00
Q8LJP6	*G. hirsutum*	1272	705	10.06	2.22	126.40	776.11	705.00	495.89	0.00
Q93WY9	*N. tabacum*	778	537	7.87	2.76	98.85	525.46	537.00	252.54	0.00
Q9ZSP9	*S. lycopersicum*	1352	858	10.38	2.54	130.31	841.40	858.00	510.60	0.00

MD, molecular dynamics simulations; A_o, observed area of wide groove (Ang^2); V_o, observed volume of wide groove (Ang^3); r, radius of approximating cylinder (Ang); h, height of approximating cylinder (Ang); l, length of wide groove (Ang); A_c, computed area of approximating cylinder (Ang^2); V_c, computed volume of approximating cylinder (Ang^3); ΔA, area differential $|A_o - A_c|$

Table 6 Computed data for cleaned and prepared ligands

	pKa	pKb	logP = logD	logK	TC	HAC	HDO	RBC	PSA	R(G)	molpol	MMFF94	MOPAC
C21	11.58	− 2.9	− 5.38	− 17.44	0	11	8	4	189.53	20.11	28.97	177.5862	− 469.9
C22					0.003							205.8188	− 464.852
C23					0.001							178.4911	− 468.472
C31	11.55	− 3.49	− 7.49	− 50.179	0	16	11	7	268.68	29.45	42.78	285.7057	− 663.263
C32					0.005							274.8125	− 674.849
C33					0.001							275.0043	− 674.058
C41	10.56	− 3.52	− 9.78	− 103.062	0	22	15	10	368.06	39.36	57.32	318.1659	− 979.702
C42	11.53		− 9.61	− 99.559	0.002	21	14		347.83	38.8	56.6	359.4289	− 951.688
C43					0.003							400.5905	− 930.327
C5	11.7	− 3.52	− 12.47	− 165.71	− 0.001	26	17	17	434.82	48.13	70.51	472.1967	− 1087.63
C6	11.51	− 3.52	− 13.84	− 248.049	0.002	31	20	16	506.13	57.49	84.23	567.8972	− 1298.93
C7	11.49	− 3.73	− 15.96	− 347.241	0.003	36	23	19	585.28	66.84	98.04	694.9641	− 1490.82
C8	11.95	− 3.74	− 16.93	− 444.031	0.001	40	24	8	633.2	74.77	110.52	813.1349	− 1659.38

pKa, log acid dissociation constant; *HDO*, hydrogen bond donor; *pKb*, log base dissociation constant; *RBC*, rotatable bond count; *LogP*, log partition-coefficient; *PSA*, polar surface area; *LogD*, log distribution coefficient; *R(G)*, Randic index; *LogK*, log binding constant; *MMFF94*, Merck molecular force field (kcal mol^{-1}); *HAC*, hydrogen bond acceptor; *MOPAC*, molecular orbital package (kcal mol^{-1}); *TC*, total charge

heterogeneity in the composition of amino acids viz. their side chains that comprise these grooves (Fig. 7, Tables 7 and 8).

Principal component-based clustering to identify potential class C homologues

The variance between the *xyz* coordinates of each ungapped aligned position ($n = 363$) was computed and summarised as eigenvalues ($n = 1089$). A scatter plot of the principal components ($PC1 \approx 73\ \% \equiv x$ axis; $PC3 \approx 5\ \% \equiv z$ axis) resulted in class C enzymes ($n = 96$) being clustered into 4 distinct groups ($x,\ z = \{(-,-),(-,+),(+,-),(+,+)\}$). Since most of the characterised members ($n = 3$; *O. sativa*, *S. lycopersicum*, *G. hirsutum*) belonged to a single cluster, these, and associated putative class C members ($n = 39$; *Arabidopsis* spp., *B. stricta*, *B. distachyon*, *B. rapa*, *C. rubella*, *C. sinensis*, *E. grandis*, *E. salsugineum*, *G. max*, *G. raimondii*, *L. usitatissimum*, *M. domestica*, *M. truncatula*, *M. guttatus*, *P. virgatum*, *P. trichocarpa*, *P. persica*, *S. purpurea*, *S. moellendorffii*, *S. lycopersicum*, *S. tuberosum*, *Z. mays*) (Sequence identity ≈ 3 − 49%) could be utilised to draw meaningful inferences about the generic active site and mechanism(s) deployed by plant class C enzymes to digest crystalline cellulose (Fig. 8; Supplementary Table 1 and Supplementary Text 10). Interestingly, members ($n = 22$) of the quadrant (+, −) included the bryophyte *P. patens* spp. and *O. sativa* spp. as compared with sequences ($n = 39$) present (−, −) which included the tracheophyte *S. moellendorffii* spp. (Fig. 8; Supplementary Table 1 and Supplementary Text 10). The presence of these ancestral class C members, i.e. tracheophytes, further strengthened the rationale of selecting this group since it represents organisms that may have evolved over 400 million years ago and therefore any mechanism postulated to digest crystalline cellulose would also likely have remained unchanged for that duration [8]. The quadrants (−, +) whose members ($n = 19$) included the characterised class C enzyme from *N. tabacum*, and (+, +) with $n = 13$ members possessed a similar distribution of plant members as with group 1 (−, −) (Fig. 8; Supplementary Table 1 and Supplementary Text 10).

Discussion

Contribution of the GH9, linker, and CBM49 to the architecture of the active site plant class C enzymes

Plant class C enzymes share considerable structural homology with gram-positive and -negative bacterial GH9 members (Tables 1, 2, and 3; Supplementary Texts 1 and 2). Although these results for GH9 are not entirely unexpected, data from this study also supports the involvement of the linker and CBM49 in the catalysis of crystalline cellulose by plant class C enzymes (Table 3; Supplementary Texts 1 and 2) [8, 17, 20–34, 59, 60]. The inclusion of the N- and C-terminal linker, albeit at higher volumes ($V \in (8.0,100.0]$) and the complete exclusion of CBM49 even amongst this small subset of class C enzymes suggest poor conservation of these segments (Figs. 5 and 8a; Supplementary Text 2) [8, 81]. These data raise the possibility that the linker and CBM49 may have an indirect or modulatory role in catalysing glycosidic cleavage and may

Table 7 Docking calculations to assess contribution of ligand interacting amino acids in full length class C enzyme after 40.1 ns MD-run

Ligand	Sequence	ΔG_B	Ki	NC	E	T	F	IS
C21	*N. tabacum*	−4.02	1.13 mM	−3.31	−0.18	−3.49	2%	445.487
	G. hirsutum	−4.68	3.69 uM	−4.24	−0.12	−4.36	1%	532.405
	O. sativa	−4.01	1.15 mM	−3.20	−0.07	−3.27	1%	411.143
	S. lycopersicum	−3.72	1.87 mM	−3.03	−0.09	−3.13	1%	403.271
C22	*N. tabacum*	−3.35	3.52 mM	−3.10	−0.13	−3.23	1%	371.54
	G. hirsutum	−3.27	4.04 mM	−3.44	−0.11	−3.56	1%	391.544
	O. sativa	−3.64	2.13 mM	−3.97	−0.29	−4.26	1%	561.366
	S. lycopersicum	−3.82	1.59 mM	−3.75	−0.35	−4.10	1%	471.261
C23	*N. tabacum*	−3.72	1.88 mM	−2.95	−0.13	−3.09	1%	321.025
	G. hirsutum	−3.64	2.14 mM	−3.60	−0.06	−3.65	1%	531.849
	O. sativa	−3.37	3.40 mM	−2.74	−0.01	−2.76	2%	364.286
	S. lycopersicum	−3.62	2.21 mM	−2.70	+0.02	−2.68	1%	404.028
C31	*N. tabacum*	−4.98	224.64 uM	−3.80	−0.04	−3.84	2%	515.486
	G. hirsutum	−4.72	346.9 uM	−2.77	+0.06	−2.71	1%	439.364
	O. sativa	−4.44	556.51 uM	−2.23	−0.01	−2.25	1%	353.431
	S. lycopersicum	−4.19	841.75 uM	−2.96	−0.25	−3.20	1%	431.316
C32	*N. tabacum*	−5.13	172.83 uM	−3.57	−0.09	−3.66	1%	537.325
	G. hirsutum	−5.15	166.98 uM	−2.77	−0.05	−2.82	1%	417.736
	O. sativa	−4.67	377.39 uM	−3.06	−0.01	−3.07	1%	472.404
	S. lycopersicum	−4.38	619.66 uM	−2.15	−0.02	−2.17	1%	508.931
C33	*N. tabacum*	−5.16	165.33 uM	−3.51	−0.02	−3.53	1%	454.372
	G. hirsutum	−4.81	296.47 uM	−2.80	−0.03	−2.83	1%	472.293
	O. sativa	−4.50	505.89 uM	−2.47	−0.10	−2.57	1%	422.942
	S. lycopersicum	−4.59	431.31 uM	−4.01	−0.03	−4.04	1%	565.025
C41	*N. tabacum*	−4.78	315.42 uM	−2.26	+0.02	−2.24	1%	324.88
	G. hirsutum	−5.16	165.42 uM	−2.90	−0.03	−2.93	1%	472.96
	O. sativa	−5.45	100.64 uM	−1.78	−0.10	−1.87	1%	252.991
	S. lycopersicum	−5.35	119.05 uM	−1.35	−0.02	−1.37	1%	247.16
C42	*N. tabacum*	−4.34	658.16 uM	−3.62	−0.25	−3.87	1%	569.419
	G. hirsutum	−3.90	1.37 mM	−2.13	−0.00	−2.13	1%	497.506
	O. sativa	−4.58	437.91 uM	−2.54	−0.20	−2.74	1%	402.801
	S. lycopersicum	−4.22	806.11 uM	−1.59	−0.05	−1.64	1%	358.607
C43	*N. tabacum*	−4.56	451.74 uM	−1.61	−0.06	−1.67	1%	376.196
	G. hirsutum	−5.06	200.60 uM	−3.55	−0.14	−3.69	1%	605.308
	O. sativa	−5.00	217.10 uM	−2.12	−0.01	−2.14	1%	355.299
	S. lycopersicum	−4.54	469.07 uM	−1.38	−0.01	−1.39	1%	382.023
C5	*N. tabacum*	−5.14	171.69 uM	−2.15	−0.12	−2.27	1%	529.463
	G. hirsutum	−4.56	457.13 uM	−0.37	+0.14	−0.23	1%	257.428
	O. sativa	−5.40	111.01 uM	−0.83	−0.11	−0.94	1%	221.706
	S. lycopersicum	−6.22	27.42 uM	−2.31	−0.14	−2.45	1%	439.826
C6	*N. tabacum*	−5.40	109.29 uM	−2.54	−0.02	−2.56	1%	438.28
	G. hirsutum	−4.54	470.52 uM	−2.44	−0.10	−−2.54	1%	386.471
	O. sativa	−4.19	850.37 uM	−1.57	−0.02	−1.59	1%	337.564
	S. lycopersicum	−4.74	337.11 uM	−2.08	−0.09	−2.17	1%	458.349
C7	*N. tabacum*	−2.97	6.63 mM	−0.72	−0.05	−0.77	1%	158.073
	G. hirsutum	−2.64	11.66 mM	−1.78	−0.05	−1.83	1%	437.98
	O. sativa	−2.43	16.50 mM	−1.95	−0.09	−2.04	1%	534.30
	S. lycopersicum	−2.64	11.63 mM	−2.26	+0.12	−2.14	1%	424.185
C8	*N. tabacum*	−7.60	2.70 uM	−2.01	−0.09	−2.10	1%	324.657
	G. hirsutum	−7.36	4.00 uM	−1.73	−0.05	−1.79	1%	334.515
	O. sativa	−7.40	3.79 uM	−2.34	−0.04	−2.38	1%	338.932
	S. lycopersicum	−7.40	3.75 uM	−2.56	−0.10	−2.66	1%	450.992

MD, molecular dynamics simulations; ΔG_B, estimated free energy of binding; *Ki*, inhibition constant; *NC*, non-covalent energy (van Der Waals, Hydrogen bond, desolvation); E, electrostatic energy; *T*, total energy; *F*, frequency; *IS*, interaction surface (Ang^2)

partake in substrate selection/modification rather than direct catalysis (Figs. 5 and 8a; Supplementary Text 2)).

The digestion of crystalline cellulose, in non-plant taxa may occur in a continuous groove that spans the GH9, linker,

Table 8 Distribution and composition of amino acids that may interact with cellulose-based ligands ($2 \leq DP \leq 8$)

Sequence ID	Organism	Amino acids (AA)
Q5NAT0	*O. sativa*	L60, Q64, A67, A75, K78, A79, S172, V417, Y423, R425, S442, F443, G459, P461, L463, D465, S520, L521, Q522, L532, W554, Y561, R563, Y590, V593, V605, P607, W609
Q8LJP6	*G. hirsutum*	L47, Q51, T54, A57, N62, D80, V82, F84, T157, N159, T223, V224, Q226, Y227, Y228, R414, A436, W437, Y469, Y496, Q498, L499, L500, V503, T504, L512, P513, K514, A516, I538, V542, T543, Y544, I550, N561, K563, L569, Y570, S587, W588, I589, S597, M598
Q93WY9	*N. tabacum*	L67, Y142, T149, T150, Y153, W155, T162, Y233, Y440, T442, W443, F444, P451, V508, A509, I510, P511, P513, K514, V521, T522, P525, Q536, K547, T548, Y550, M464, L567, K568, L569, Y572, K583, Y584
Q9ZSP9	*S. lycopersicum*	Y43, N47, R49, N55, L58, K62, S106, D139, N141, T142, Y145, W435, S437, D451, P502, P504, T507, K509, A511, P512, K515, P520, R521, P522, R523, V524, L525, P526, T534, L545, T549, Y550, Y551, R552, Y553, L575, P578, L579, F591, L595, N596, V607, V619

AA_x^{Dock}, amino acids that are energetically favourable (Black+Red); AA_x^{CvG}, amino acids that form part of cavities and grooves (Red); *DP*, ligands of cellulose with varying degrees of polymerization

and the associated CBMs [51–60]. Plant class C enzymes may also do so in a surface groove that is initially bounded by the GH9_linker $\left(IS_x^{GL}\right)$ at the posterior basolateral surface and continues laterally being bounded in turn by the GH9_linker $\left(IS_x^{GL}\right)$, CBM49_linker $\left(IS_x^{CL}\right)$, and GH9_CBM49 $\left(IS_x^{GC}\right)$ surfaces where $x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}$, finally terminating anteriorly in a solvent accessible cavity that might constitute the principal active site (Fig. 5). Physically, although the *IS*-bounded grooves appear discontinuous at the surface, a thorough analysis suggests the presence of several subsurface cavities that could maintain contiuity (Figs. 5, 6, and 7). Further, an almost complete absence of measurable cavities in CBM49/linker could also ensure that the substrate-facing surface through which crystalline cellulose traverses was chemically inert. The model precludes the existence of disparate active sites whilst, concomitantly asserts a preparatory/modulatory effect by CBM49/linker which may then be followed by the hydrolytic cleavage of the glycosidic bond at the active site (Fig. 7). The rmsf and DCCM data in concert with the invariant core volumes further suggests that the *IS* that bounds the linker and CBM49 $\left(IS_x^{CL}; x \in \{Q5NAT0, Q8LJP6, Q93WY9, Q9ZSP9\}\right)$ surface may exhibit heightened low frequency motion, a factor that could confer upon class C enzymes the propensity to accommodate varying lengths of crystalline cellulose (47, Tables 6, 7, and 8; Supplementary Text 2, 6–9).

Molecular dissection of a putative active site of class C enzymes

Any plausible model of the active site architecture of plant class C enzymes would have to explain as well as include extant empirical data. 3D models of full length minimised and characterised members from *O. sativa*, *G. hirsutum*, *N. tabacum*, and *S. lycopersicum* were simulated in vacuo for 40.1 *ns* and thence examined for amino acids that may contribute to substrate binding and/or catalysis. The combined list of functionally relevant amino acids, i.e. $AA_x = \left(AA_x^{\text{Dock}} \cap AA_x^{CvG}\right) \subset AA_x^{IS}; x \in \{O.sativa, G.hirsutum, N.tabacum, S.lycopersicum\}$, were enumerated and utilised for these analyses (Table 7). The paucity/absence of residues that support generic acid-base mediated cleavage of the $\beta\,(1 \rightarrow 4)$ glycosidic bond for crystalline cellulose as well as known active site amino acids $\left(\{E, C, H\} \notin AA_x^{\text{Dock}}\right)$, despite being present contiguously with those that are $(\{D, E, C, H, P, R, K, N, Q, L, I, V, A, M, W, F, Y, G, S, T\} \in AA_x^{IS} \cup AA_x^{CvG})$ suggests that catalysis might occur in a superficial cavity just below the surface of the protein (Fig. 5, Table 7). However, the preponderance of energetically favourable aromatic amino acids along the interaction surfaces and various grooves ($AAA = \{W, F, Y\} \approx 15-41\%$; $\{W, F, Y\} \in AA_x$) when taken in tandem with previously conducted mutagenesis experiments on the CBMs suggest that cellulose may physically interact with these residues on the surface prior to entering the cavity for catalysis (Figs. 5, 6, and 7, Tables 6, 7, and 8). The formation of this purported groove may be supported/strengthened by the uniform presence of proline ($P \approx 3.7-25\%$), as well as stabilizing electrostatic interactions involving arginine (*R*), lysine (*K*), asparagine (*N*), glutamine (*Q*), serine (*S*), and threonine (*T*) ($[RKNQ] \approx 12-25\%$; $[ST] \approx 10-18.5\%$), while remaining chemically inert throughout its length with several amino acids with shorter hydrophobic side chains lining the groove, i.e. leucine (*L*), isoleucine (*I*), valine (*V*), methionine (*M*), and exceptionally alanine (*A*) ($HSC \equiv [LIVAM] \approx 25-37\%$) (Table 7).

Mechanistic insights into crystalline cellulose digestion by plant class C enzymes

The aforementioned discussion notwithstanding the small sample size could preclude meaningful inference of the

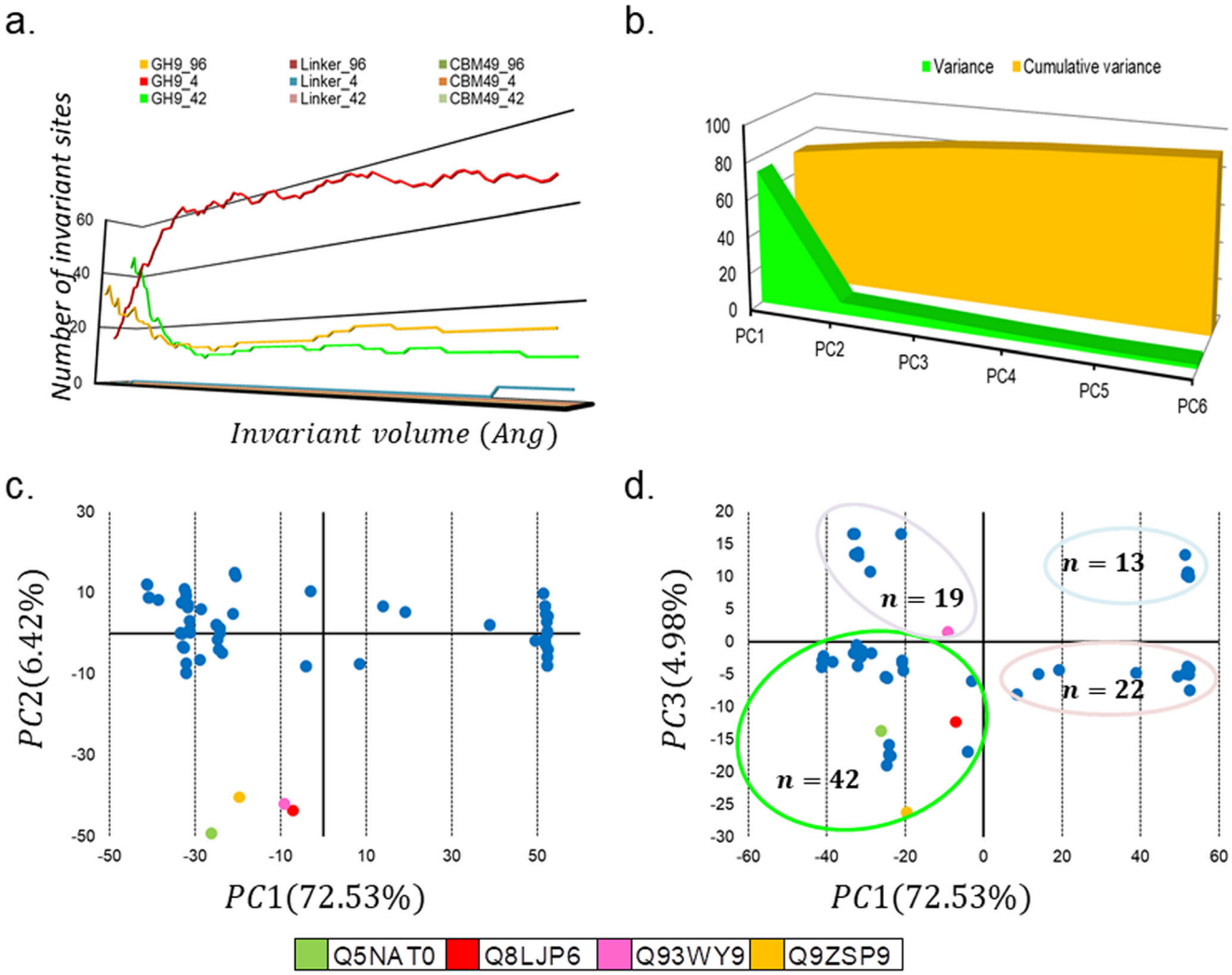

Fig. 8 PCA-based inferential clustering of plant class C enzymes. **a** The contribution of the GH9, linker, and CBM49 was assessed as a function of the volume ($0 < V(Ang^3) \leq 100$) of the invariant core computed and the number of sites that participate from each subsegment. The data suggests that while GH9 is universally conserved, CBM49 is not, even amongst class C members. **b** Principal component analysis of putative and characterised class C GH9 endoglucanases ($n = 96$) was also done to assess the variation in coordinates across the 3D models of class C enzymes Here, **c** *PC*1*vs PC*2 and **d** *PC*1*vsPC*3 were considered. Despite, the higher contribution of *PC*2 to the variance (≅6.42%) as compared with *PC*3 (≅5%), there was greater resolution of the sequences and a greater number of sequences ($n = 39$ *vs* $n = 22$) with the latter. Further, three characterised members (*O. sativa*, *G. hirsutum*, *S. lycopersicum*) also clustered into this quadrant (−, −) as opposed to only *N. tabacum* (−, +). These data suggest that the $x-$ & $z-$*axes* (*PC*1, *PC*3) might represent the principal axes of class C enzymes. Abbreviations—PC1–6, principal components 1–6; GH9, glycoside hydroxyl 9; CBM49, carbohydrate binding module; V, invariant core volume

mechanism(s) of crystalline cellulose digestion by plant class C GH9 endoglucanases. This was offset by examining 3D models of putative structural homologues of selected class C members ($n = 39$) (Fig. 8a; Supplementary Table 1 and Supplementary Text 2 and 10). Data from these suggest that the largest uninterrupted grooves that span GH9 ($l_{GH9} \cong 101 - 194$ *Ang*) and CBM49 ($l_{CBM49} \cong 71 - 183$ *Ang*) are disjoint and distinct, the only exceptions being the sequences from *L. usitatissimum* spp. and *B. distachyon* spp. (Fig. 9, Table 9). Further support for the mechanism(s) purported for digesting crystalline cellulose plant class C enzymes may be gleaned by examining the 3D models for *IS*-bounded surface grooves $\left(IS_x^{GL}, IS_x^{CL}, IS_x^{GC}\right)$ in *A. coerulea*, *C. sinensis*, *C. rubella*, *S. purpurea*, *P. persica* spp., *M. domestica*, *P. virgatum* spp., *A. lyrata*, *B. rapa* spp., *Z. mays* spp., and *B. distachyon* spp. (Fig. 9, Table 9). Interestingly, the groove located at the interaction surface and bounded by GH9, linker, and CBM49 concomitantly $\left(IS_x^{GLC}\right)$ as in *L usitatissimum* spp., *P. trichocarpa*, *M. truncatula* spp., *G. max*, and *B. distachyon* spp. may exert significant influence on crystalline cellulose in comparison with the distally located and smaller CBM49-bounded grooves ($l_{CBM49} \leq 100$ *Ang*) (Fig. 9). This data further complements the hypothesis that plant class C GH9 endoglucanases may possess a dual mode

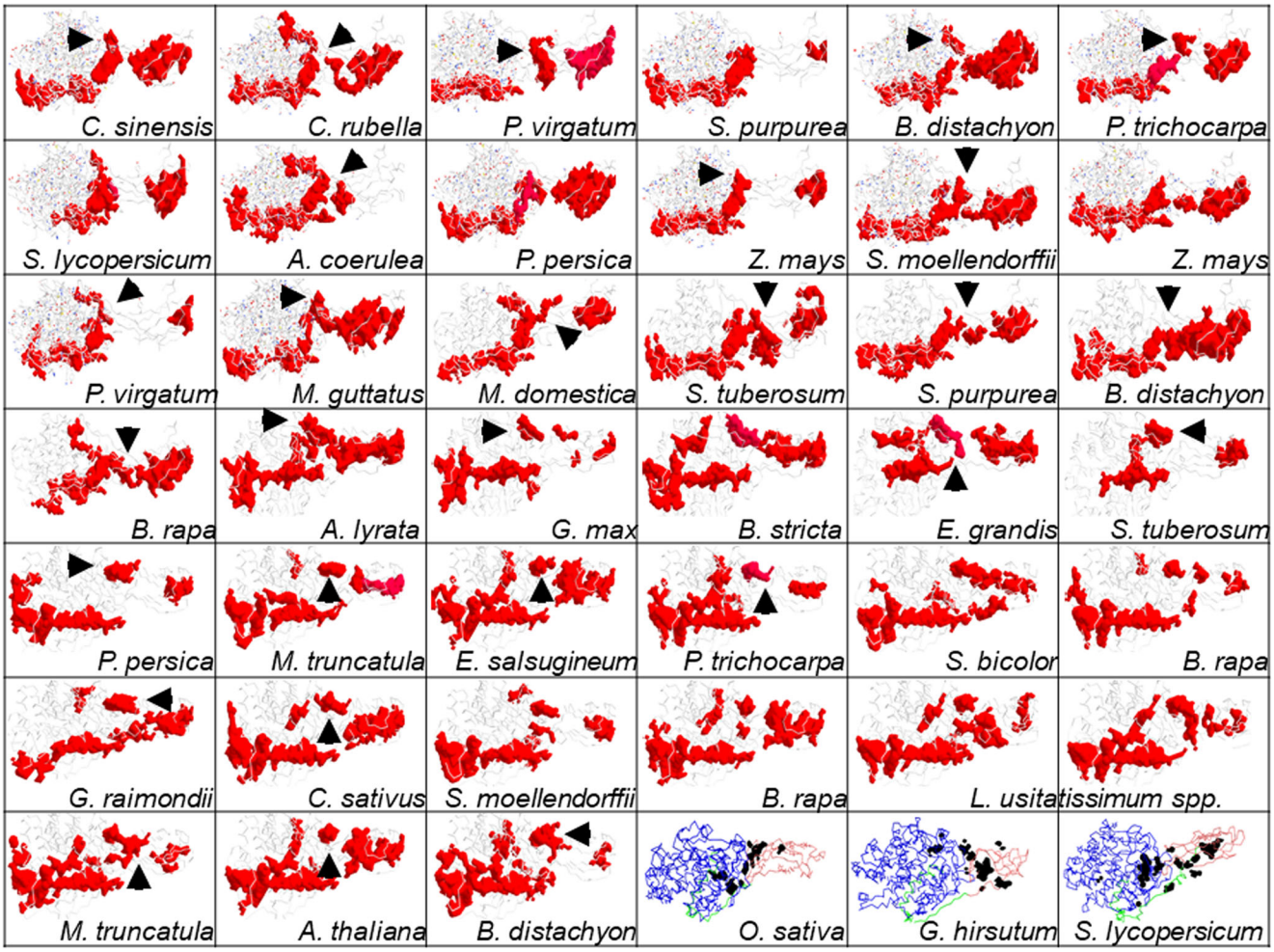

Fig. 9 Mechanistic insights into digestion of crystalline cellulose by plant class C endoglucanases. The data presented suggest that CBM49 along with the linker is poorly conserved and exhibits considerable heterogeneity, even amongst plant class C enzymes. Since, the effects of similar CBMs on catalysis are well characterised at least in non-plant taxa, any model would have to consider modulation by CBM49 of the catalytic residues which are present on GH9. This would imply that while catalysis may occur in a solvent accessible subsurface cavity, the surface groove(s) leading to it must involve CBM49 and the linker. The multimodal approach adopted here (interactions surface definition and amino acid enumeration, docking, cavity and surface analysis) suggests that the extended side chains of aromatic amino acid effect could interact and thereby render crystalline cellulose amenable to subsequent cleavage. Residues such as proline and stabilizing electrostatic interations involving arginine, lysine, asparagine, glutamine, serine, and threonine, along with several smaller hydrophobic residues along the interaction surfaces of the linker and CBM49, heightened oscillatory motion, could result in physical alteration of the groove itself, whilst concomitantly influencing the reactions that cleave crystalline cellulose. Additionally, the selection of substrates/polymer may also be determined by these residues. In support of these analyses 3D models of several homologues were analysed. Clearly, a large and extended groove formed by GH9, linker, and CBM49 and could lead to the catalytic site is observed with *C. sinensis*, *C. rubella*, *A. coerulea*, *P. persica*, and *M. domestica*. Although, segment spanning and overlapping grooves (*B. distachyon*, *M. truncatula*) are also present, it is unlikely that these may contribute to catalysis. However, the clear presence of large disjoint grooves along the interaction surfaces, along with the complete absence of catalytically competent residues corroborates a dual mode of interaction/modification and catalysis by plant class C enzymes. Abbreviations—CBM49, carbohydrate binding module; GH9, glycoside hydrolase

(processive, non-processive) of action wherein crystalline cellulose is initially acted upon and thereby modified by the indenting side chains of aromatic amino in a quasi-continuous surface groove at the interface(s) of GH9, linker, and CBM49, which is inert and stable. Once modified (induced strain on the glycosidic linkage), crystalline cellulose is driven towards a solvent accessible subsurface cavity. Here, the GH9 conserved catalytic residues of aspartic (*D*) and/or glutamic (*E*) acids utilise an acid-base catalytic mechanism to cleave the β $(1 \rightarrow 4)$ linkage between glucopyranose units. These may then be acted upon by exoglucanases to release oligosaccharides ($C2-C4$). This mechanism not only corroborates extant kinetic data such as CBM-mediated modulatory catalysis, but also offers a molecular explanation for substrate promiscuity observed for this group of enzymes, whilst conforming to available structural data from non-plant taxa (Figs. 4, 5, 6, 7, 8, and 9, Tables 4, 5, 6, 7, 8, and 9; Supplementary Tables 1 and Supplementary Texts 2–10) [51–60].

Table 9 Major groove dimensions of putative plant class C enzymes ($n = 39$)

Sequence ID	Organism	r_{GH9}	h_{GH9}	l_{GH9}	r_{CBM49}	h_{CBM49}	l_{CBM49}
orange1.1g043219m	*Citrus sinensis*	13.60	3.50	170.78	11.29	4.40	141.76
ppa022524m	*Prunus persica*	11.65	3.71	146.38	13.29	4.06	166.87
SapurV1A.0237s0330.1.p	*Salix purpurea*	13.68	3.74	171.84	6.09	3.73	76.50
Carubv10003874m	*Capsella rubella*	13.50	3.28	169.59	10.11	3.90	126.99
Pavir.Eb00189.1.p	*Panicum virgatum*	13.55	3.53	170.15	8.76	4.06	110.04
Bradi5g026010.1.p	*Bradipodium distachyon*	9.79	3.64	123.02	14.46	4.08	181.58
Brara.E01714.1.p	*Brassica rapa*	8.06	4.00	101.18	10.23	4.28	128.52
Pavir.Ea00142.1.p	*Panicum virgatum*	12.12	3.66	152.23	6.31	2.70	79.25
GRMZM2G143747_P01	*Zea mays*	15.45	3.32	194.11	9.71	4.61	121.95
SapurV1A.0035s0560.1.p	*Salix purpurea*	14.64	3.47	183.88	8.48	3.78	106.56
Bradi2g07150.1.p	*Bradipodium distachyon*	13.45	3.94	168.96	–	–	–
PGSC0003DMP400021750	*Solanum tuberosum*	14.02	3.50	176.10	9.77	2.94	122.77
Migut.D01909.1.p	*Mimulus guttatus*	10.15	3.26	127.54	11.60	4.19	145.69
234652	*Selaginella moellendorffii*	9.30	4.07	116.79	10.93	4.09	137.30
MDP0000131267	*Malus domestica*	12.20	3.81	153.26	8.13	4.22	102.16
Potri.001G092200.1	*Populus trichocarpa*	12.44	3.79	156.30	9.47	3.20	118.92
GRMZM2G453565_P01	*Zea mays*	12.86	3.71	161.55	7.27	4.68	91.25
Solyc02g014220.2.1	*Solanum lycopersicum*	8.73	3.05	109.64	9.66	4.43	121.38
Aquca_037_00141.1	*Aquilegia coerulea*	12.83	3.96	161.16	6.80	2.67	85.35
Glyma.05G216400.1.p	*Glycine max*	13.34	3.68	167.55	5.91	3.24	74.26
489943	*Arabidopsis lyrata*	12.62	3.74	158.53	14.59	3.44	183.19
Bostr.25463s0223.1.p	*Boechera stricta*	13.46	3.67	169.11	12.65	3.65	158.85
Eucgr.J00862.1	*Eucalyptus grandis*	9.11	4.41	114.46	12.83	3.54	161.13
ppa002939m	*Prunus persica*	14.83	3.79	186.22	8.16	3.85	102.47
PGSC0003DMP400034548	*Solanum tuberosum*	8.54	3.40	107.32	8.66	3.21	108.72
Thhalv10028514m	*Eutrema salsugineum*	13.92	3.71	174.81	13.35	3.97	167.66
Medtr4g074960.1	*Medicago truncatula*	9.77	3.75	122.77	8.91	4.48	111.90
Gorai.005G210200.1	*Gossypium raimondii*	11.43	3.57	143.52	10.27	3.78	128.96
Potri.003G139600.1	*Populus trichocarpa*	13.60	3.89	170.81	7.71	3.66	96.86
Sobic.003G015700.1.p	*Sorghum bicolor*	13.31	3.57	167.13	10.85	3.60	136.29
Brara.I01325.1.p	*Brassica rapa*	13.90	3.84	174.63	6.05	3.10	75.93
Cucsa.107370.1	*Cucumis sativus*	14.22	4.11	178.58	12.41	4.04	155.86
99802	*Selaginella moellendorffii*	9.71	4.13	122.00	8.00	4.70	100.49
Brara.C02656.1.p	*Brassica rapa*	13.86	3.68	174.09	12.82	3.95	160.97
Lus10032377	*Linum usitatissimum*	12.70	3.99	159.48	7.76	3.98	97.44
Lus10003888	*Linum usitatissimum*	12.45	4.17	156.38	6.27	2.83	78.77
Medtr8g099410.1	*Medicago truncatula*	13.71	3.81	172.20	7.92	3.04	99.49
AT4G11050.1	*Arabidopsis thaliana*	13.53	3.84	169.89	12.86	4.03	161.55
Bradi2g32270.1.p	*Bradipodium distachyon*	14.21	3.61	178.47	5.64	2.68	70.88

r, radius of approximating cylinder (*Ang*); *h*, height of approximating cylinder (*Ang*); *l*, length of wide groove (*Ang*); *GH9*, glycoside hydrolase 9; *CBM49*, carbohydrate binding module

Evolutionary significance for CBM49-mediated digestion of crystalline cellulose

The ability to cleave crystalline cellulose by plant class C members is dependent on the presence of CBM49 and may have evolved directly from non-plant taxa (≈500 Mya) [8, 17, 20–34, 59, 60]. An additional premise explored previously was that plant class C enzymes may not just predate but, could potentially diverge into classes A and B after CBM49 was excised during processing of the mature mRNA transcript [8, 18, 46, 83–85]. A mechanistic understanding of these processes is clearly desirable with much of the aforementioned

generated data involving kinetic parameters, mRNA expression levels, and sequence information. The present study highlights variations in the CBM49/linker even amongst class C enzymes, provides insights into the architecture, position, plasticity, and composition of the *IS*-enclosed surface grooves, delineates the position and composition of a contiguous subsurface cavity for catalytic cleavage of the glycosidic linkage, enumerates functionally relevant amino acids that participate in substrate selection/modification, and offers a mechanistic explanation of CBM49-mediated reaction chemistry (Figs. 1, 2, 3, 4, 5, 6, 7, 8, and 9; Tables 1, 2, 3, 4, 5, 6, 7, 8, and 9; Supplementary Table 1 and Supplementary Texts 1–10). Additionally, a definitive body of literature indicates that hyperflexible regions may be intrinsically disordered and therefore have short $t_{1/2}$ [63, 86, 87]. This would imply that proteins with the CBM49_linker may be evolutionarily at a disadvantage than those without. Alternatively, these might be encoded by nucleotides with a tendency to form higher order substructures in mRNA such as stem loops, bulges, and bends. These in turn could delay or irreversibly interrupt the ribosomal apparatus and prevent effective translation of the mRNA, and thereby contribute to decreased expression of class C enzymes. Since CBM49 is central to the ability of plant class C enzymes to digest crystalline cellulose, it would follow this loss could lead to a decrease in class C enzymes or conversely an increase in classes A and B [8].

Conclusions

A detailed biophysical analysis of homology models of characterised and putative class C endoglucanases was carried out to assess the contribution(s) of the GH9, linker, and CBM49 to catalysis/modification of crystalline cellulose. The work presented in this manuscript corroborates the notion that the linker and CBM49 may complement generic acid-base catalysis by aspartic/glutamic residues of GH9, and may do so in a multitude of ways. These include an influence on the structural organization of the protein, participation in critical intra-protein interactions, facilitate formation of inert and structurally plastic surface grooves, and render crystalline cellulose amenable to hydrolytic cleavage. Despite being entirely computational, the findings presented here offer profound insights into not just the active site geometry of plant class C GH9 endoglucanases, but also offer valuable clues into their evolutionary divergence. Whilst, most these findings await experimental valiation the analyses conducted suggests that plant-based conversion of biomass is feasible and may constitute a viable alternative to bacterial-, fungal-, and algal-based protocols.

Acknowledgments SK wishes to formally thank Dr. Rita Sharma for her suggestions and unflinching moral support.

Author contribution SK collated the data, conducted the analysis, developed the scoring indices, wrote all the necessary code, and the manuscript.

Abbreviations *AA,* Amino acids; *CBM,* Carbohydrate binding module; *DP,* Degree of polymerization; *EC,* Enzyme commission; *FL,* Full length; *GH,* Glycoside hydrolase; *IS,* Interaction surface; *L,* Linker; *NMA,* Normal mode analysis; *T,* Truncated

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Klemm D, Heublein B, Fink HP, Bohn A (2005) Cellulose: fascinating biopolymer and sustainable raw material. Angew Chem Int Ed Eng 44(22):3358–3393
2. Augimeri RV, Varley AJ, Strap JL (2015) Establishing a role for bacterial cellulose in environmental interactions: lessons learned from diverse biofilm-producing proteobacteria. Front Microbiol 6: 1282
3. Reardon-Robinson ME, Wu C, Mishra A, Chang C, Bier N, Das A, Ton-That H (2014) Pilus hijacking by a bacterial coaggregation factor critical for oral biofilm development. Proc Natl Acad Sci U S A 111(10):3835–3840
4. Updegraff DM (1969) Semimicro determination of cellulose in biological materials. Anal Biochem 32(3):420–424
5. Yoshida Y, Palmer RJ, Yang J, Kolenbrander PE, Cisar JO (2006) Streptococcal receptor polysaccharides: recognition molecules for oral biofilm formation. BMC Oral Health 6(Suppl 1):S12
6. Agarwal V, Dauenhauer PJ, Huber GW, Auerbach SM (2012) Ab initio dynamics of cellulose pyrolysis: nascent decomposition pathways at 327 and 600 degrees C. J Am Chem Soc 134(36):14958–14972
7. Paulsen AD, Hough BR, Williams CL, Teixeira AR, Schwartz DT, Pfaendtner J, Dauenhauer PJ (2014) Fast pyrolysis of wood for biofuels: spatiotemporally resolved diffuse reflectance in situ spectroscopy of particles. ChemSusChem
8. Kundu S, Sharma R (2018) Origin, evolution, and divergence of plant class C GH9 endoglucanases. BMC Evol Biol 18:79
9. del Campillo E, Gaddam S, Mettle-Amuah D, Heneks J (2012) A tale of two tissues: AtGH9C1 is an endo-beta-1,4-glucanase involved in root hair and endosperm development in Arabidopsis. PLoS One 7(11):e49363
10. Kundu S (2015) Co-operative intermolecular kinetics of 2-oxoglutarate dependent dioxygenases may be essential for system-level regulation of plant cell physiology. Front Plant Sci 6:489
11. Tan TC, Kracher D, Gandini R, Sygmund C, Kittl R, Haltrich D, Hallberg BM, Ludwig R, Divne C (2015) Structural basis for cellobiose dehydrogenase action during oxidative cellulose degradation. Nat Commun 6:7542

12. Westermark U, Eriksson K-E, Daasvatn K, Liaaen-Jensen S, Enzell CR, Mannervik B (1974) Cellobiose:quinone oxidoreductase, a new wood-degrading enzyme from white-rot fungi. Acta Chem Scand 28b:209–214
13. Schimz KL, Broll B, John B (1983) Cellobiose phosphorylase (EC 2.4.1.20) of cellulomonas: occurrence, induction, and its role in cellobiose metabolism. Arch Microbiol 135(4):241–249
14. Sheth K, Alexander JK (1969) Purification and properties of beta-1, 4-oligoglucan:orthophosphate glucosyltransferase from Clostridium thermocellum. J Biol Chem 244(2):457–464
15. Ye X, Zhu Z, Zhang C, Zhang YH (2011) Fusion of a family 9 cellulose-binding module improves catalytic potential of Clostridium thermocellum cellodextrin phosphorylase on insoluble cellulose. Appl Microbiol Biotechnol 92(3):551–560
16. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42(Database issue):D490–D495
17. Davison A, Blaxter M (2005) Ancient origin of glycosyl hydrolase family 9 cellulase genes. Mol Biol Evol 22:1273–1284
18. Kundu S, Sharma R (2016) In silico identification and taxonomic distribution of plant class C GH9 endoglucanases. Front Plant Sci 7: 1185
19. Ficko-Blean E, Boraston AB (2006) The interaction of a carbohydrate-binding module from a Clostridium perfringens N-acetyl-beta-hexosaminidase with its carbohydrate receptor. J Biol Chem 281(49):37748–37757
20. Duan CJ, Feng YL, Cao QL, Huang MY, Feng JX (2016) Identification of a novel family of carbohydrate-binding modules with broad ligand specificity. Sci Rep 6:19392
21. Prates ET, Stankovic I, Silveira RL, Liberato MV, Henrique-Silva F, Pereira Jr N, Polikarpov I, Skaf MS (2013) 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(3):e59069
22. Boraston AB, Nurizzo D, Notenboom V, Ducros V, Rose DR, Kilburn DG, Davies GJ (2002) Differential oligosaccharide recognition by evolutionarily-related beta-1,4 and beta-1,3 glucan-binding modules. J Mol Biol 319(5):1143–1156
23. Charnock SJ, Bolam DN, Nurizzo D, Szabo L, McKie VA, Gilbert HJ, Davies GJ (2002) Promiscuity in ligand-binding: the three-dimensional structure of a Piromyces carbohydrate-binding module, CBM29-2, in complex with cello- and mannohexaose. Proc Natl Acad Sci U S A 99(22):14077–14082
24. Crennell SJ, Cook D, Minns A, Svergun D, Andersen RL, Nordberg Karlsson E (2006) Dimerisation and an increase in active site aromatic groups as adaptations to high temperatures: X-ray solution scattering and substrate-bound crystal structures of Rhodothermus marinus endoglucanase Cel12A. J Mol Biol 356(1):57–71
25. Kim SJ, Kim SH, Shin SK, Hyeon JE, Han SO (2016) Mutation of a conserved tryptophan residue in the CBM3c of a GH9 endoglucanase inhibits activity. Int J Biol Macromol 92:159–166
26. Mattinen ML, Kontteli M, Kerovuo J, Linder M, Annila A, Lindeberg G, Reinikainen T, Drakenberg T (1997) Three-dimensional structures of three engineered cellulose-binding domains of cellobiohydrolase I from Trichoderma reesei. Protein Sci 6(2):294–303
27. Morrill J, Kulcinskaja E, Sulewska AM, Lahtinen S, Stalbrand H, Svensson B, Abou Hachem M (2015) The GH5 1,4-beta-mannanase from Bifidobacterium animalis subsp. lactis Bl-04 possesses a low-affinity mannan-binding module and highlights the diversity of mannanolytic enzymes. BMC Biochem 16:26
28. Nishijima H, Nozaki K, Mizuno M, Arai T, Amano Y (2015) Extra tyrosine in the carbohydrate-binding module of Irpex lacteus Xyn10B enhances its cellulose-binding ability. Biosci Biotechnol Biochem 79(5):738–746
29. Parsiegla G, Reverbel-Leroy C, Tardif C, Belaich JP, Driguez H, Haser R (2000) Crystal structures of the cellulase Cel48F in complex with inhibitors and substrates give insights into its processive action. Biochemistry 39(37):11238–11246
30. Simpson HD, Barras F (1999) Functional analysis of the carbohydrate-binding domains of Erwinia chrysanthemi Cel5 (endoglucanase Z) and an Escherichia coli putative chitinase. J Bacteriol 181(15):4611–4616
31. Simpson PJ, Xie H, Bolam DN, Gilbert HJ, Williamson MP (2000) The structural basis for the ligand specificity of family 2 carbohydrate-binding modules. J Biol Chem 275(52):41137–41142
32. Strobel KL, Pfeiffer KA, Blanch HW, Clark DS (2015) Structural insights into the affinity of Cel7A carbohydrate-binding module for lignin. J Biol Chem 290(37):22818–22826
33. Taylor CB, Talib MF, McCabe C, Bu L, Adney WS, Himmel ME, Crowley MF, Beckham GT (2012) Computational investigation of glycosylation effects on a family 1 carbohydrate-binding module. J Biol Chem 287(5):3147–3155
34. Yaniv O, Petkun S, Shimon LJ, Bayer EA, Lamed R, Frolow F (2012) A single mutation reforms the binding activity of an adhesion-deficient family 3 carbohydrate-binding module. Acta Crystallogr D Biol Crystallogr 68(Pt 7):819–828
35. Abbott DW, Hrynuik S, Boraston AB (2007) Identification and characterization of a novel periplasmic polygalacturonic acid binding protein from Yersinia enterolitica. J Mol Biol 367(4):1023–1033
36. Abramyan J, Stajich JE (2012) Species-specific chitin-binding module 18 expansion in the amphibian pathogen Batrachochytrium dendrobatidis. MBio 3(3):e00150–e00112
37. Bachman ES, McClay DR (1996) Molecular cloning of the first metazoan beta-1,3 glucanase from eggs of the sea urchin Strongylocentrotus purpuratus. Proc Natl Acad Sci U S A 93(13): 6808–6813
38. Janecek S, Svensson B, MacGregor EA (2011) Structural and evolutionary aspects of two families of non-catalytic domains present in starch and glycogen binding proteins from microbes, plants and animals. Enzym Microb Technol 49(5):429–440
39. Li S, Yang X, Bao M, Wu Y, Yu W, Han F (2015) Family 13 carbohydrate-binding module of alginate lyase from Agarivorans sp. L11 enhances its catalytic efficiency and thermostability, and alters its substrate preference and product distribution. FEMS Microbiol Lett:362(10)
40. Newstead SL, Watson JN, Bennet AJ, Taylor G (2005) Galactose recognition by the carbohydrate-binding module of a bacterial sialidase. Acta Crystallogr D Biol Crystallogr 61(Pt 11):1483–1491
41. Palomo M, Kralj S, van der Maarel MJ, Dijkhuizen L (2009) The unique branching patterns of Deinococcus glycogen branching enzymes are determined by their N-terminal domains. Appl Environ Microbiol 75(5):1355–1362
42. Libertini E, Li Y, McQueen-Mason SJ (2004) Phylogenetic analysis of the plant endo-beta-1,4-glucanase gene family. J Mol Evol 58(5):506–515
43. Molhoj M, Pagant S, Hofte H (2002) Towards understanding the role of membrane-bound endo-beta-1,4-glucanases in cellulose biosynthesis. Plant Cell Physiol 43(12):1399–1406
44. Urbanowicz BR, Bennett AB, Del Campillo E, Catala C, Hayashi T, Henrissat B, Hofte H, McQueen-Mason SJ, Patterson SE, Shoseyov O et al (2007) Structural organization and a standardized nomenclature for plant endo-1,4-beta-glucanases (cellulases) of glycosyl hydrolase family 9. Plant Physiol 144(4):1693–1696
45. Flint J, Bolam DN, Nurizzo D, Taylor EJ, Williamson MP, Walters C, Davies GJ, Gilbert HJ (2005) Probing the mechanism of ligand recognition in family 29 carbohydrate-binding modules. J Biol Chem 280(25):23718–23726

46. Montanier C, Flint JE, Bolam DN, Xie H, Liu Z, Rogowski A, Weiner DP, Ratnaparkhe S, Nurizzo D, Roberts SM et al (2010) Circular permutation provides an evolutionary link between two families of calcium-dependent carbohydrate binding modules. J Biol Chem 285(41):31742–31754
47. Roske Y, Sunna A, Pfeil W, Heinemann U (2004) High-resolution crystal structures of Caldicellulosiruptor strain Rt8B.4 carbohydrate-binding module CBM27-1 and its complex with mannohexaose. J Mol Biol 340(3):543–554
48. Zhang C, Zhang W, Lu X (2015) Expression and characteristics of a Ca(2)(+)-dependent endoglucanase from Cytophaga hutchinsonii. Appl Microbiol Biotechnol 99(22):9617–9623
49. Tunnicliffe RB, Bolam DN, Pell G, Gilbert HJ, Williamson MP (2005) Structure of a mannan-specific family 35 carbohydrate-binding module: evidence for significant conformational changes upon ligand binding. J Mol Biol 347:287–296
50. Uni F, Lee S, Yatsunami R, Fukui T, Nakamura S (2009) Role of exposed aromatic residues in substrate-binding of CBM family 5 chitin-binding domain of alkaline chitinase. Nucleic Acids Symp Ser (Oxf) 53:311–312
51. Divne C, Ståhlberg J, Reinikainen T, Ruohonen L, Pettersson G, Knowles JKC, Teeri TT, Jones TA (1994) The 3dimensional crystal-structure of the catalytic core of cellobiohydrolase-I from Trichoderma reesei. Science 265:524–528
52. Divne C, Ståhlberg J, Teeri TT, Jones TA (1998) High resolution crystal structures reveal how a cellulose chain is bound in the 50 Å long tunnel of cellobiohydrolase I from Trichoderma reesei. J Mol Biol 275:309–325
53. Kleywegt GJ, Zou JY, Divne C, Davies GJ, Sinning I, Ståhlberg J, Reinikainen T, Srisodsuk M, Teeri TT, Jones TA (1997) The crystal structure of the catalytic core domain of endoglucanase I from Trichoderma reesei at 3.6 Å resolution, and a comparison with related enzymes. J Mol Biol 272:383–397
54. Mackenzie LF, Sulzenbacher G, Divne C, Jones TA, Woldike HF, Schulein M, Withers SG, Davies GJ (1998) Crystal structure of the family 7 dndoglucanase I (Cel7B) from Humicola insolens at 2.2 Å resolution and identification of the catalytic nucleophile by trapping of the covalent glycosyl-enzyme intermediate. Biochem J 335:409–416
55. Ståhlberg J, Johansson G, Pettersson G, New A (1991) Model for enzymatic-hydrolysis of cellulose based on the 2-domain structure of cellobiohydrolase-I. Biotechnol Biofuels 9:286–290
56. Payne CM, Baban J, Horn SJ, Backe PH, Arvai AS, Dalhus B, Bjørås M, Eijsink VGH, Sørlie M, Beckham GT et al (2012) Hallmarks of processivity in glycoside hydrolases from crystallographic and computational studies of the Serratia marcescens Chitinases. J Biol Chem 287:36322–36330
57. Taylor CB, Payne CM, Himmel ME, Crowley MF, McCabe C, Beckham GT (2013) Binding site dynamics and aromatic-carbohydrate interactions in processive and non-processive family 7 glycoside hydrolases. J Phys Chem B 117:4924–4933
58. Payne CM, Resch MG, Chen L, Crowley MF, Himmel ME, Taylor 2nd LE, Sandgren M, Ståhlberg J, Stals I, Tan Z, Beckham GT (2013) Glycosylated linkers in multimodular lignocellulose-degrading enzymes dynamically bind to cellulose. Proc Natl Acad Sci U S A 110:14646–14651
59. Mandelman D, Belaich A, Belaich JP, Aghajari N, Driguez H, Haser R (2003) X-ray crystal structure of the multidomain endoglucanase Cel9G from Clostridium cellulolyticum complexed with natural and synthetic cello-oligosaccharides. J Bacteriol 185(14):4127–4135
60. Sakon J, Irwin D, Wilson DB, Karplus PA (1997) Structure and mechanism of endo/exocellulase E4 from Thermomonospora fusca. Nat Struct Biol 4(10):810–818
61. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845–858
62. Case DA, Cerutti DS, Cheatham III TE, Darden TA, Duke RE, Giese TJ, Gohlke H, Goetz AW, Greene D, Homeyer N, Izadi S, Kovalenko A, Lee TS, LeGrand S, Li P, Lin C, Liu J, Luchko T, Luo R, Mermelstein D, Merz KM, Monard G, Nguyen H, Omelyan I, Onufriev A, Pan F, Qi R, Roe DR, Roitberg A, Sagui C, Simmerling CL, Botello-Smith WM, Swails J, Walker RC, Wang J, Wolf RM, Wu X, Xiao L, York DM, Kollman PA (2017) AMBER 2017. University of California, San Francisco
63. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26(16):1781–1802
64. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14(1):33–38 27-8
65. Edgar RC (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113
66. Edgar RC (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797
67. Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS (2006) Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22:2695–2696
68. Altman RB, Gerstein M (1994) Finding an average core structure: application to the globins. Proc Int Conf Intell Syst Mol Biol 2:19–27
69. Gerstein M, Altman RB (1995) Average core structures and variability measures for protein families: application to the immunoglobulins. J Mol Biol 251(1):161–175
70. Gerstein M, Chothia C (1991) Analysis of protein loop closure. Two types of hinges produce one motion in lactate dehydrogenase. J Mol Biol 220(1):133–149
71. Durand P, Trinquier G, Sanejouand Y-H (1994) A new approach for determining low-frequency normal modes in macromolecules. Biopolymers 34(6):759–771
72. Hinsen K, Petrescu A-J, Dellerue S, Bellissent-Funel M-C, Kneller GR (2000) Harmonicity in slow protein dynamics. Chem Phys 261(1–2):25–37
73. Kaplan W, Littlejohn TG (2001) Swiss-PDB viewer (deep view). Brief Bioinform 2:195–197
74. Irwin JJ, Shoichet BK (2005) ZINC–a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182
75. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC: a free tool to discover chemistry for biology. J Chem Inf Model:52, 1757–1768
76. Das B, Meirovitch H, Navon IM (2003) Performance of hybrid methods for large-scale unconstrained optimization as applied to models of proteins. J Comput Chem 24(10):1222–1231
77. Rappe AK, Casewit CJ, Colwell KS, Goddard WA, Skiff WM (1992) UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. J Am Chem Soc 114(25):10024–10035
78. Bikadi Z, Hazai E (2009) Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock. Aust J Chem 1:15
79. Stewart JJ (2009) Application of the PM6 method to modeling proteins. J Mol Model 15:765–805
80. Halgren TA (1996) Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem 17:490–519
81. Morris GM, Goodsell DS (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19(14):1639–1662

82. Solis FJ, Wets RJB (1981) Minimization by random search techniques. Math Oper Res 6(1):19–30
83. Urbanowicz BR, Catala C, Irwin D, Wilson DB, Ripoll DR, Rose JK (2007) A tomato endo-beta-1,4-glucanase, SlCel9C1, represents a distinct subclass with a new family of carbohydrate binding modules (CBM49). J Biol Chem 282(16):12066–12074
84. Buchanan M, Burton RA, Dhugga KS, Rafalski AJ, Tingey SV, Shirley NJ, Fincher GB (2012) Endo-(1,4)-beta-glucanase gene families in the grasses: temporal and spatial co-transcription of orthologous genes. BMC Plant Biol 12:235
85. Xie G, Yang B, Xu Z, Li F, Guo K, Zhang M, Wang L, Zou W, Wang Y, Peng L (2013) Global identification of multiple OsGH9 family members and their involvement in cellulose crystallinity modification in rice. PLoS One 8:e50171
86. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT et al (2014) Classification of intrinsically disordered regions and proteins. Chem Rev 114(13):6589–6631
87. van der Lee R, Lang B, Kruse K, Gsponer J, Sanchez de Groot N, Huynen MA, Matouschek A, Fuxreiter M, Babu MM (2014) Intrinsically disordered segments affect protein half-life in the cell and during evolution. Cell Rep 8(6):1832–1844

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.