The FEBS Journal

FEBS PRESS science publishing by scientists

The crystal structure of RsSymEG1 reveals a unique form of smaller GH7 endoglucanases alongside GH7 cellobiohydrolases in protist symbionts of termites

Topi Haataja[1], Henrik Hansson[1], Shigeharu Moriya[2], Mats Sandgren[1] and Jerry Ståhlberg[1]

1 Department of Molecular Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden
2 RIKEN Center for Advanced Photonics, Namazu-city, Japan

Keywords
cellulase enzyme; cellulose hydrolysis; glycoside hydrolase family 7; *Reticulitermes speratus*; termite symbiont

Correspondence
J. Ståhlberg, Department of Molecular Sciences, Swedish University of Agricultural Sciences, PO Box 7015, SE-750 07 Uppsala, Sweden
Tel: +46 (0)18 673182
E-mail: jerry.stahlberg@slu.se

(Received 12 July 2023, revised 31 October 2023, accepted 8 December 2023)

doi:10.1111/febs.17029

Glycoside hydrolase family 7 (GH7) cellulases are key enzymes responsible for carbon cycling on earth through their role in cellulose degradation and constitute highly important industrial enzymes as well. Although these enzymes are found in a wide variety of evolutionarily distant organisms across eukaryotes, they exhibit remarkably conserved features within two groups: exo-acting cellobiohydrolases and endoglucanases. However, recently reports have emerged of a separate clade of GH7 endoglucanases from protist symbionts of termites that are 60–80 amino acids shorter. In this work, we describe the first crystal structure of a short GH7 endoglucanase, RsSymEG1, from a symbiont of the lower termite *Reticulitermes speratus*. A more open flat surface and shorter loops around the non-reducing end of the cellulose-binding cleft indicate enhanced access to cellulose chains on the surface of cellulose microfibrils. Additionally, when comparing activities on polysaccharides to a typical fungal GH7 endoglucanase (*Trichoderma longibrachiatum* Cel7B), RsSymEG1 showed significantly faster initial hydrolytic activity. We also examine the prevalence and diversity of GH7 enzymes that the symbionts provide to the termite host, compare overall structures and substrate binding between cellobiohydrolase and long and short endoglucanase, and highlight the presence of similar short GH7s in other organisms.

Introduction

Glycoside hydrolase family 7 of carbohydrate-active enzymes contains some of the most prominent known cellulose-degrading enzymes. As one of the key enzymes fungi use for degradation of polymeric cellulose, they play a significant role in the carbon cycle in nature [1]. They are also important in many biotechnical applications and have been central in the development of cellulosic ethanol as a biofuel with perpetual improvement of the enzyme formulations used for cellulose hydrolysis, where GH7 enzymes are the major components. Significant advancements have been made to date in improving the properties of these enzymes for industrial processes, with majority of academic research having focused on fungal enzymes from ascomycetes such as *Trichoderma reesei*, and basidiomycetes such as *Phanerochaete chrysosporium* [2–4].

While a vast majority of the GH7 sequences known today belong to fungi, they are found in remarkably

Abbreviations
CBH, cellobiohydrolase; CBM, carbohydrate-binding module; DpuGH7, *Daphnia pulex* GH7 CBH; EG, endoglucanase; Endo H, endoglycosidase H; GH7, glycoside hydrolase family 7; HirCel7A, *Heterobasidion irregulare* Cel7A; LquCel7B, *Limnoria quadripunctata* Cel7B; MalCel7B, *Melanocarpus albomyces* Cel7B; MthCel7A, *Myceliophthora thermophila* Cel7A; RsSymEG1, GH7 endoglucanase 1 from symbiont of *Reticulitermes speratus*; TloCel7B, *Trichoderma longibrachiatum* Cel7B EG; TreCel7A, *Trichoderma reesei* Cel7A CBH; TreCel7B, *Trichoderma reesei* Cel7B EG.

diverse organisms across the eukaryotic tree, for example, including species of crustaceans and amoeba, suggesting either an ancient origin of the enzyme family or multiple occurrences of horizontal gene transfer [5]. While the host organisms are diverse, the sequences display strikingly conserved features with clustering into two main categories based on sequence and function, so-called endoglucanases (EG) and cellobiohydrolases (CBH) [6]. Whereas CBHs possess a long substrate binding tunnel and cleave cellulose chains through a processive mechanism with the main product being cellobiose, the active site of EGs is more open with a substrate binding groove, with their primary mode of action being endo-cleavage. Most characterized enzymes demonstrate distinct prevalence of one mode of action over the other, and recent studies have elucidated sequence and structure features important for governing the balance of exo- and endo-cleavage [1,6–8].

Studies of wood-degrading termites have shown that they produce endogenous cellulose-degrading enzymes, including glycoside hydrolases from families GH1, GH9, and GH45, but do not express GH7 enzymes. However, lower termites gain access to GH7 enzymes through intricate symbiosis with primitive unicellular eukaryotic protists possessing GH7 genes, and in fact depend on their symbionts for growth on cellulosic material [9,10]. The most prominent symbionts identified are parabasalids and oxymonads that belong to the very early diverging Excavata super-group of the eukaryotic tree [11–14]. With an increasing amount of sequencing data becoming available, it has become clear that these organisms carry genes for GH7 cellobiohydrolases, but also a previously unknown class of 60–80 amino acids shorter GH7 endoglucanase enzymes seemingly found only in symbionts of the lower termites [15,16]. Previously, biochemical characterization of a few enzymes belonging to this new class of GH7s has illustrated activity profiles diverging from those of the typical fungal GH7s, with high initial reaction rates and higher pH optimums [17–19]. However, no structural information has been published for any of these novel enzymes to date. In this work, we have strived to shed light on the novel structure architecture of the short endoglucases by solving the structure of RsSymEG1, a short GH7 endoglucanase encoded by a transcript isolated from an unknown symbiont of the termite *Reticulitermes speratus* that has previously been expressed in *Aspergillus oryzae* and characterized biochemically [19]. We also study the sequence diversity of family GH7 enzymes with a focus on the enzymes found in termite symbiont protists and possible implications to the evolution of this enzyme family.

Results

Crystal structure of RsSymEG1

Recombinant RsSymEG1 was obtained from the previously described *Aspergillus oryzae* expression strain where it is expressed as part of a fusion protein. RsSymEG1 is appended to the C terminus of *A. oryzae* alpha-amylase AmyB, linked by a Kex2 protease site that is cleaved by endogenous proteases [19]. From 3 L of culture, ~54 mg of purified RsSymEG1 was obtained, after ammonium sulfate precipitation and protein purification by ion exchange and size exclusion chromatography. The protein was further deglycosylated with Endoglycosidase H (Endo H) prior to crystallization experiments. Figure 1 shows an SDS/PAGE analysis of the purified RsSymEG1 before and after Endo H treatment.

Crystals of the novel GH7 cellulase RsSymEG1 were obtained in space group $I4_1$, with one protein molecule per asymmetric unit. An x-ray crystal structure was solved by molecular replacement, refined at a resolution of 1.85 Å, and deposited in the Protein Data Bank with PDB ID: 8POF. The structure exhibits a beta-sandwich core typical to the GH7 family, with six anti-parallel beta sheets forming a base for a substrate binding site, and several loops arranging to form a groove-like structure (Fig. 2A). At the N terminus of the protein, Gly1 is the first visible residue in the electron density, suggesting that there are no residues left from the KEX-linker region included in the construct, and confirming the absence of typical pyroglutamate (PCA) N-terminal structure. Gly1 is also the first residue following the predicted signal peptide cleavage site

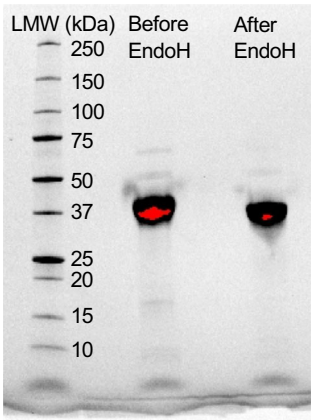


Fig. 1. SDS/page analysis of purified RsSymEG1 before and after enzymatic deglycosylation, using 1 μg Endo H per 1 mg of RsSymEG1, incubated overnight at pH 6.0 and 28 °C ($n = 1$). Red indicates oversaturated pixels.

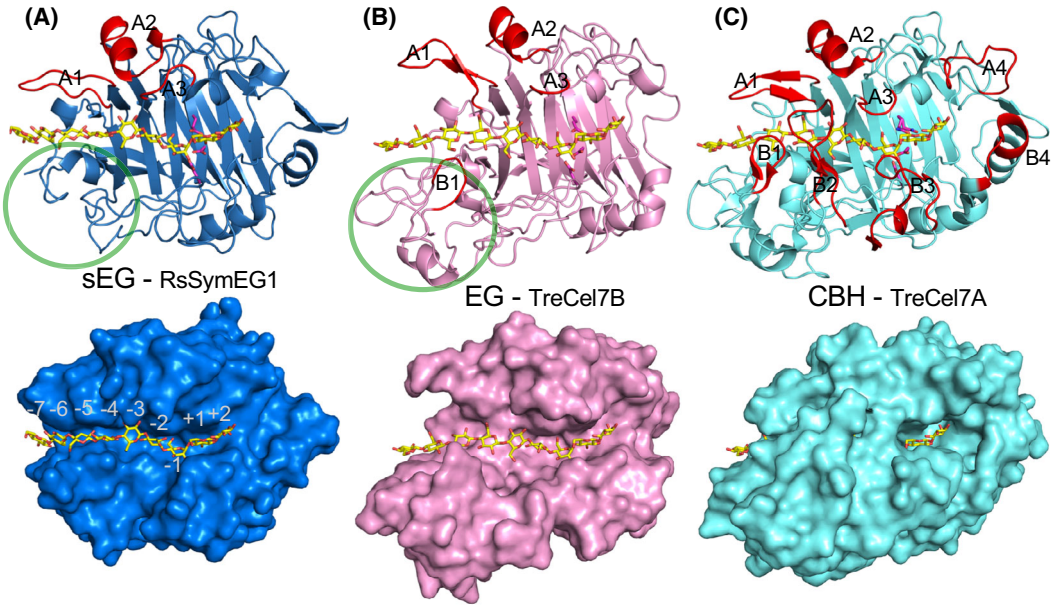


Fig. 2. Crystal structures of catalytic domains of the three types of GH7 enzymes. The GH7 catalytic domain is built on a highly conserved beta-sandwich core, with more or less extended loops (highlighted in red) flanking the active site. (A) Our new structure of RsSymEG1 presented herein is the first structure of a small GH7 EG (sEG) from termite symbionts. It lacks loops B1 and B2 and surrounding regions (green circle) as well as the B3 and B4 loops. The cellononaose molecule, with glucose-binding subsites labeled, was fitted to the enzyme using AutoDock Vina. (B) Full-length endoglucanases (EG) are represented by TreCel7B from Trichoderma reesei (PDB code 1EG1), shown with cellononaose from structure 4C4C superposed. (C) The TreCel7A complex with cellononaose (PDB code 4C4C) shows that CBHs have long loops that enclose the cellulose chain in a tunnel. Structure images were created with MACPYMOL [69].

in the native protein sequence. There is a potential N-glycosylation site at Asn38, and while the electron density suggests that there is likely at least partial glycosylation, not enough density is seen to justify modeling a N-acetylglucosamine residue. However, treatment with Endo H prior to crystallization led to decrease in apparent molecular weight on SDS/PAGE, thus confirming the N-glycosylation (Fig. 1).

The x-ray structure shows that RsSymEG1 contains four disulfide bridges, Cys100–Cys290, Cys127–Cys137, Cys154–Cys166, and Cys171-Cys245. Glu139 and Glu144 constitute the catalytic nucleophile and catalytic acid/base, respectively, while Asp141 forms the assisting residue. As an artifact from the crystallization conditions, a TRIS molecule is seen making a close contact with these three catalytic residues at the active site. Similar TRIS binding has previously been observed in PcCel7D by Ubhayasekera *et al.* [20], who also demonstrated an inhibitory effect of TRIS on that enzyme. Within the RsSymEG1 structure, one metal atom is observed, presumably Na, coordinating with the carbonyl oxygens of Gly285 and Gly288, three water molecules, and a polyethylene glycol (PEG) molecule. Overall, the structure is well defined with an average B-factor of 21.7 $Å^2$ for the peptide chain. Significant ambiguity of the electron density is seen only at the flexible Gly-Ala-Gly-Gly loop constituting residues 206–209. Relevant statistics from data collection, processing, and structure refinement are shown in Table 1.

Comparison of RsSymEG1 to existing GH7 structures shows that despite the highly similar core architecture, there are several major differences compared to other enzyme structures in the family. Side-by-side comparison with the *Trichoderma reesei* CBH TreCel7A and EG TreCel7B reveals that like TreCel7B, RsSymEG1 lacks the loops A4, B2, B3, and B4 (as denoted in Haddad Momeni *et al.* [21]) seen in TreCel7A and other GH7 cellobiohydrolases (Fig. 2). However, it also completely lacks any structure corresponding to the B1 loop in TreCel7B. This leads to RsSymEG1 having a significantly more exposed substrate binding site, especially at the early part of the binding groove corresponding to substrate binding positions −5 and −4 (as annotated for TreCel7A by Divne *et al.* [22]). The effects of missing sequence regions on the overall shape of the enzyme are seen most of all in the space corresponding to surroundings of the B1 and B2 loops in TreCel7A and TreCel7B, with RsSymEG1 missing a sizable volume in this area

Table 1. Statistics from X-ray diffraction data collection and processing, structure refinement, and final model.

A. Diffraction data	
Beamline	BioMax, MAX IV
Cell dimensions (Å)	114.6, 114.6, 50.0
Space group	I 4_1
Resolution range (Å)	45.8–1.85 (1.89–1.85)
No. of unique reflections[a]	27 804 (1666)
Completeness (%)[a]	99.8 (98.9)
Multiplicity[a]	4.2 (4.1)
I/σ(I)[a]	9.8 (2.6)
R_{merge}[a,b]	0.087 (0.50)
B. Structure refinement	
Resolution used in refinement (Å)	45.8–1.85
No. of reflections, work set	27 793
No. of reflections, test set	1405
R (work set)[c]	0.166
R_{free}[c]	0.187
No. of nonhydrogen atoms	2701
Protein atoms	2420
Water atoms	203
Other atoms	9
Average B factors (Å^2)	22.0
Protein	21.9
Water	30.6
Other atoms	28.7
RMSD bond lengths (Å)	0.004
RMSD bond angles (°)	1.23
Ramachandran plot outliers[d]	0

[a]Numbers in parentheses are for the highest resolution bin; [b]$R_{merge} = \sum_{hkl}\sum_i|I-\langle I\rangle|/\sum_{hkl}\sum_i|I|$; [c]$R = \sum||F_0|-|F_c||/\sum|F_0|$; the final *R*-factor is given; [d]wwPDB Validation Service.

(Fig. 2). This is also the region where GH7 enzymes typically contain an N-terminal PCA-residue embedded within a hydrophobic pocket in the structure, thus capping the N terminus of the peptide chain and contributing to the stability of the enzyme (Fig. 3). This whole section is missing in RsSymEG1, and it is, thus, not surprising that the N terminus of the enzyme does not contain the typical post-translational modification where an N-terminal glutamine is converted into a PCA.

Further comparison to 20 existing GH7 structures through sequence and structure alignments reveals several conserved residues within the substrate binding groove, but also highlights some features unique to the RsSymEG1 structure (Fig. 4). Starting from the non-reducing chain end of the substrate binding groove, interestingly, on the A1 loop of RsSymEG1, there are two tryptophan residues, Trp61 and Trp65, with the side chains oriented into the substrate binding groove, lining the likely substrate binding positions −6, −5, and −4 (Fig. 5A). There are no tryptophans present in corresponding positions in other published GH7

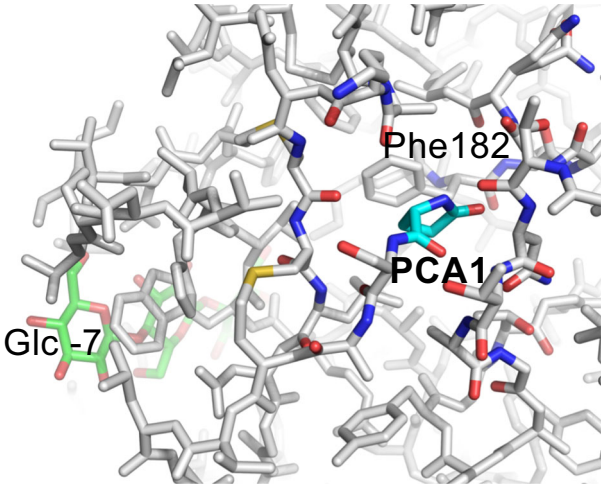


Fig. 3. N-terminal capping by pyroglutamate in GH7s. Glutamine/Q is highly conserved at the N terminus of GH7 sequences after signal peptide cleavage. All known GH7 structures to date, except for RsSymEG1, have a Gln1 residue cyclized to pyroglutamic acid (PCA). The PCA1 residue is buried in a hydrophobic pocket under loop B1 near the non-reducing end of the active site, as shown in the structure of TreCel7A with cellononaose bound (PDB code 4C4C). The image was created with MACPYMOL [69].

structures, and while structures of the CBHs LquCel7B, HirCel7A, MalCel7B, MthCel7A, and DpuGH7 (PDB entries: 4IPM, 2XSP, 2RFZ, 5W11, 4XNN) have a tyrosine at the tip of the A1 loop, it is positioned in the opposite direction, forming the beginning of the substrate binding tunnel and extending beyond the −7 site [21,23–25]. The way it is positioned in the crystal structure, Trp65 takes up significant space within the RsSymEG1 binding groove. Superposing the RsSymEG1 with other GH7 structures, including structures with bound ligands (4C4C, 4ZZU, 2RFZ, 3PL3), shows that in this region, the active site tunnel is more narrow than in other GH7 structures, and does not fit a cellulose chain in the orientation observed previously for GH7s at the −5 and −4 binding sites (Figs 2 and 5). Consequently, AutoDock Vina was run to model fitting a cellononaose chain taken from the TreCel7A structure 4C4C into the active site of RsSymEG1 [26]. The best-fit result shows glucose unit −6 aligning parallel to the Trp61 side chain, while Trp65 aligns with the −4 glucose, representing likely stacking interactions between the cellulose chain and the enzyme (Figs 2A and 5A). Whether there is enough flexibility within the A1 loop region of RsSymEG1 to allow more space for binding a cellulose chain within the −5 and −4 sites in various positions cannot be assessed solely from the crystal structure, but no ambiguity of

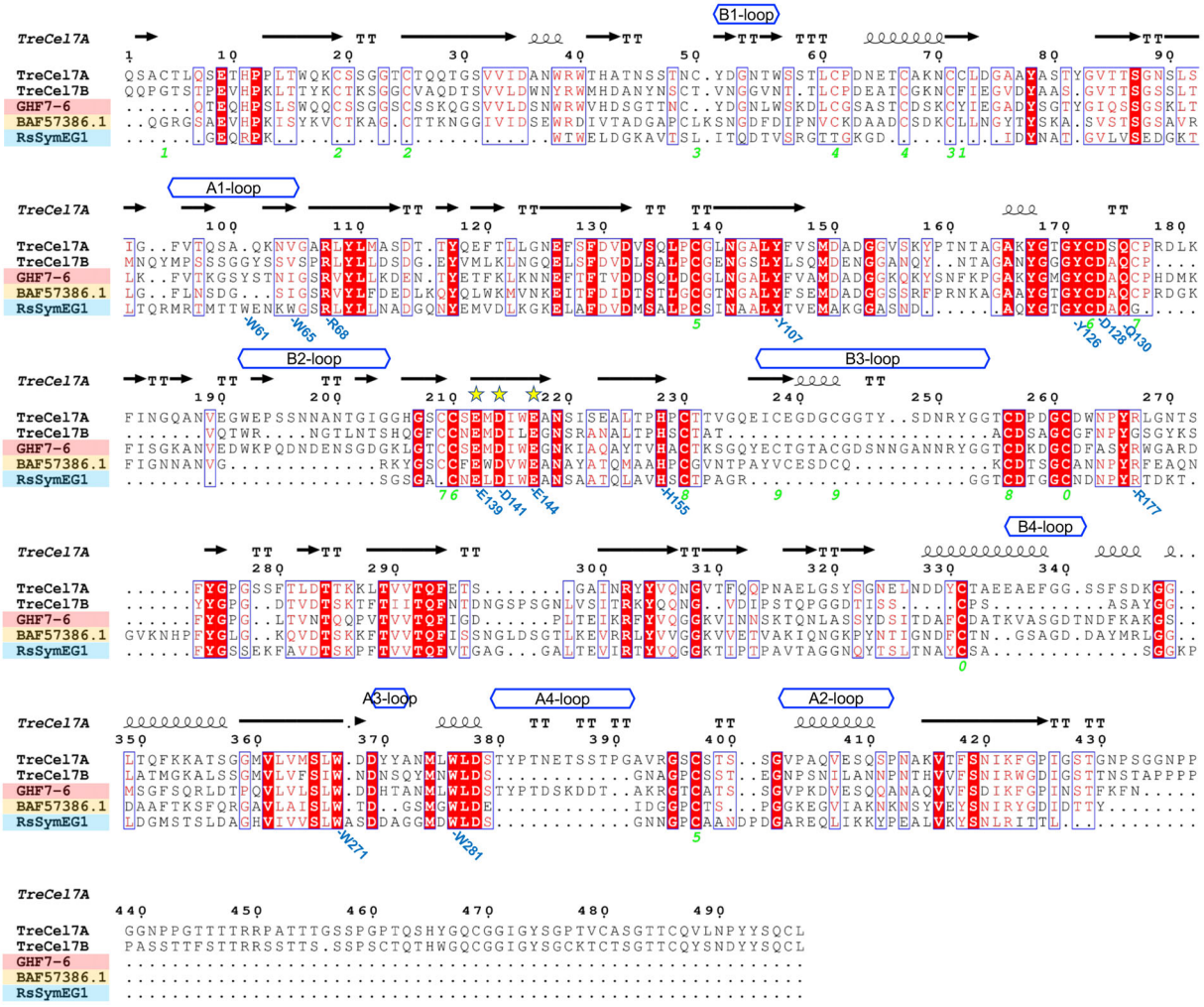


Fig. 4. Alignment of fungal cellobiohydrolases (CBH) and endoglucanases (EG), and termite symbiont GH7 amino acid sequences. TreCel7A and TreCel7B represent fungal CBHs and EGs, respectively. GHF7-6 and BAF57386.1 are the CBH and EG sequences, respectively, from termite symbionts chosen for structure prediction. RsSymEG1 represents the smaller protist endoglucanases (sEG) GH7s. The top three rows mark active site loops, secondary structure elements (helices, beta-strands as arrows, and TT as turns), and residue numbers in TreCel7A as reference. White characters on red show identical residues, red letters similar residues, and blue boxes outline conserved regions. Yellow stars mark the conserved catalytic residues, and green numbers the disulfide pairing of cysteines in TrCel7A. Residues of interest in RsSymEG1 are noted in blue at the bottom. The figure was created using ESPript [68], with further loop and residue annotations added in Microsoft PowerPoint. Figure and legend reproduced from [70] with modifications. The copyrights to the doctoral thesis [70] and the original image this has been reproduced from are held by the author, Topi Haataja. The authors and The FEBS Journal have been granted permission to reproduce the image. The images have also been licensed under the Creative Commons license CC BY NC 4.0, allowing non-commercial use.

the loop position is seen in the electron density, indicating lack of significant mobility or alternate conformations of the loop within the crystal.

Continuing toward the active site, RsSymEG1 lacks the Trp/Tyr residue at the base of the binding groove/-tunnel at the −4 subsite (Trp38 in TreCel7A, Tyr38 in TreCel7B) seen in most GH7 structures (Fig. 5). At the −3 site RsSymEG1 contains a conserved arginine, Arg68, typically interacting with a bound cellulose substrate. Trp271 and Trp281 form the typical sugar binding platforms at the substrate binding sites −2 and +1, respectively, and the −1, −2 sites are further lined by conserved tyrosine residues Tyr107 and Tyr126 (Fig. 5). In addition to the three catalytic residues, Glu139, Asp141, and Glu144, there are other conserved residues in the immediate vicinity of the active site, including Asp128, Gln130, and His155, completing the archetypal GH7 catalytic center.

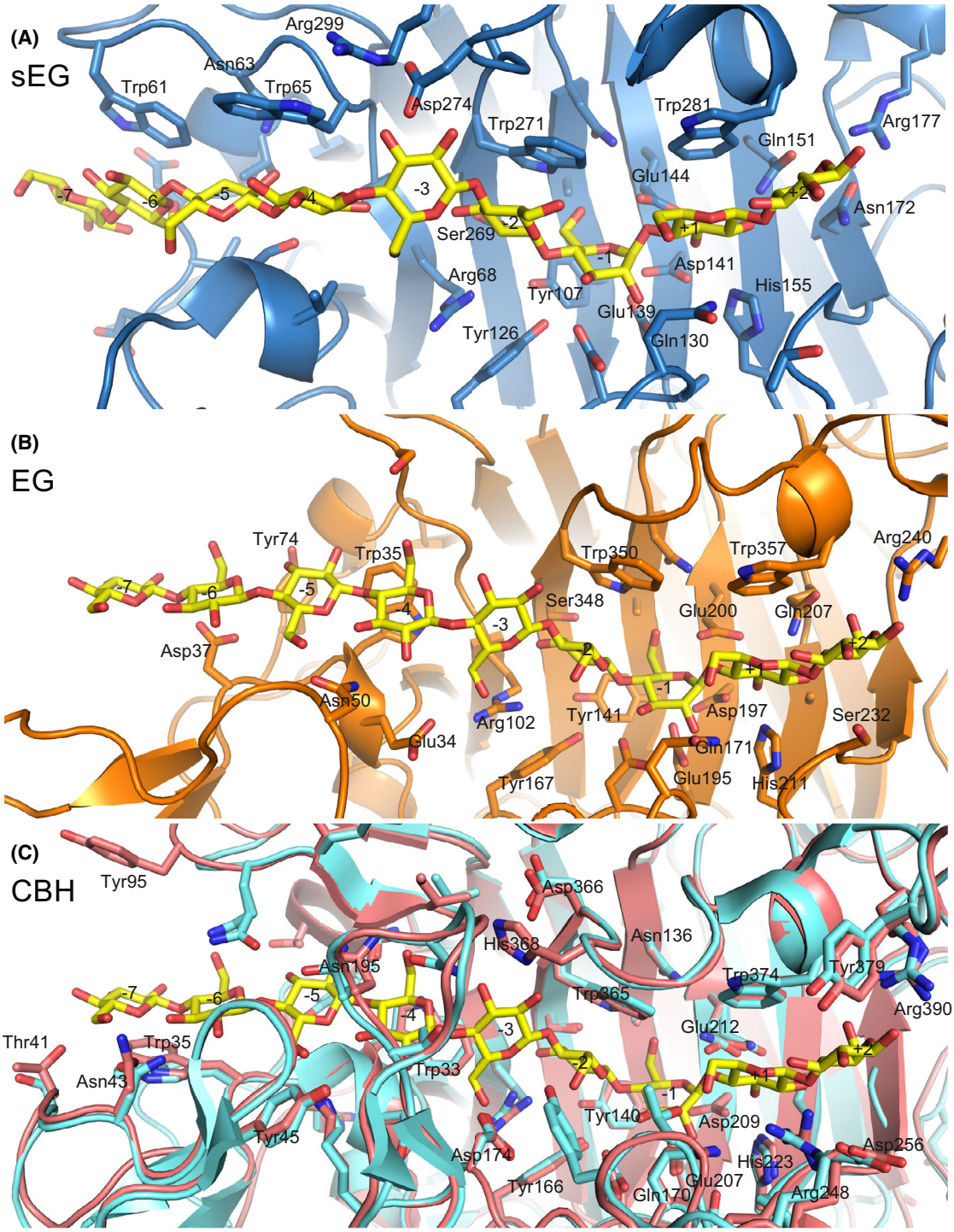


Fig. 5. Cellulose binding in GH7s from protist symbionts of termites. (A) Crystal structure of the short endoglucanase (sEG) RsSymEG1 with cellononaose fitted using AutoDock Vina. Note the two sugar-binding platforms at subsites −7 to −4 formed by Trp61 and Trp65. Glu139 is the catalytic nucleophile and Glu144 is the acid/base that cleaves the glycosidic bond between subsites −1 and +1. (B) Predicted structure from sequence BAF57386, a full-length endoglucanase (EG), with cellononaose from TreCel7A structure 4C4C. (C) Predicted structure of GHF7-6 cellobiohydrolase (CBH; red) with residue labels, superposed on the TreCel7A (cyan)/cellononaose (yellow) crystal structure (PDB code 4C4C). Structure images were created with MACPYMOL [69].

Furthermore, the residues present at the +1, +2, and +3 subsites conform to the clear endoglucanase nature of the rest of the structure with the absence of two arginines shown to promote strong cellobiose product binding in CBHs (Arg251 and Arg394 in TreCel7A), or other side chains likely to constitute strong binding of cleaved products, even if Arg177 likely contributes to substrate binding (Fig. 5) [27].

Sequences and structure models of parabasalid CBHs and EGs

In order to explore the diversity of GH7 enzymes within parabasalids in lower termites, a blast search of parabasalid sequences within the NCBI non-redundant database was conducted using the RsSymEG1 sequence as a query. Following a manual grooming of the result sequences, a sequence alignment confirms the presence of both CBH and short endoglucanase (sEG) sequences (Figs S1 and S2), examples of both of which have been previously described [17–19]. However, there are also sequences corresponding to the features of typical GH7 EGs. RoseTTAFold-server was initially used to produce structure models of example sequences of both CBHs and EGs from termite symbiont parabasalids in order to make comparisons to RsSymEG1 and other GH7 structures [28]. As the quality of the RoseTTA models regarding side-chain placement was considered insufficient, MODELLER was subsequently used to produce higher quality models based on sequence alignments with suitable GH7 templates [29,30].

GHF7-6 (GenBank id: AGT15840), a GH7 enzyme from a *Reticulitermes flavipes* symbiont, was selected as an example of a CBH for modeling as its activity and expression was previously characterized [17]. Based on the model and a sequence alignment with existing structures, GHF7-6 conforms to typical GH7 CBH features. It contains the characteristic loops A1-A4 and B1-B4 forming the substrate binding tunnel, while notably the B3 loop appears to be two residues longer than in any existing GH7 structure. Another divergence is observed at the N-terminal end of the protein, where between the N-terminal residue and the first conserved glutamic acid residue (Glu9 in TreCel7A) GHF7-6 appears to be lacking five amino acids compared to known GH7 structures (with the exception of RsSymEG1). Highlighting the probable cellulose binding residues within the substrate binding tunnel by superposing the GHF7-6 model with cellononaose from the TreCel7A-structure 4C4C shows an arrangement of typical residues (Fig. 5C). At the tunnel "entrance," similar to structures LquCel7B, HirCel7A, MalCel7B, MthCel7A, and DpuGH7 (PDB: 4IPM, 2XSP, 2RFZ, 5W11, and 4XNN), two sides of the tunnel are lined with aromatic residues, with Tyr95 at the tip of the A1-loop and Trp35 on the opposite side corresponding to binding site −7. The −4 site is formed by Trp33, corresponding to TreCel7A residue Trp38, shown to be crucial for substrate loading into the substrate binding tunnel [31,32]. Toward the catalytic site, subsites −3, −2, and −1 are lined with Tyr140, Tyr166, Trp365, Arg102, as well as His368 at the A3-loop. Beyond the catalytic residues Glu207, Asp209, and Glu212, likely product binding residues 'comprise His223, Arg 248, Trp374, and Arg390 (Fig. 5C)'.

In the absence of any omics or activity data, the sequence BAF57386 from an uncultured symbiont of *Neotermes koshunensis* was arbitrarily selected as an EG to model from three "full length" parabasalid GH7 EG sequences that were found in our queries (GenBank: BAF57386, BAF57417, and BAF57463). Despite relatively low sequence similarity to existing EG structures (< 40%), this sequence as well presents a typical array of GH7 EG substrate binding residues, including Arg240, His211, Trp357, Trp350, Arg102, and Trp35 (Fig. 4B). Interestingly, BAF57386 also contains a tyrosine not seen in other GH7 structures at position Tyr74, possibly involved in substrate binding at the early part of the binding groove. The first predicted residue after signal peptide cleavage for both GHF7-6 and BAF57386 is glutamine, consistent with a possible modification into a PCA N terminus. However, this feature does not seem to be conserved throughout the parabasalid CBHs and EGs based on the predicted signal peptide cleavage sites [33] (Fig. S2).

GH7 phylogenetics

In order to understand the potential evolutionary relationships of the different types of GH7 enzymes found in parabasalids, as well as their position in the wider context of eukaryotes, a maximum likelihood phylogenetic tree of 371 GH7 sequences was constructed. Two sequences from the distantly related GH16 family were included in order to root the tree. From this analysis, a smaller tree of 116 selected sequences across the original tree was constructed for visualization, with the clades containing termite symbiont sequences included in their entirety. The distribution of selected CBHs, EGs, and sEGs within the resulting phylogenetic tree is shown in Fig. 6. In line with a previous report by Schiano-di-Cola *et al.* [16], there appears to be two somewhat separate clades containing sEG sequences from termite symbionts. The sequence of RsSymEG1

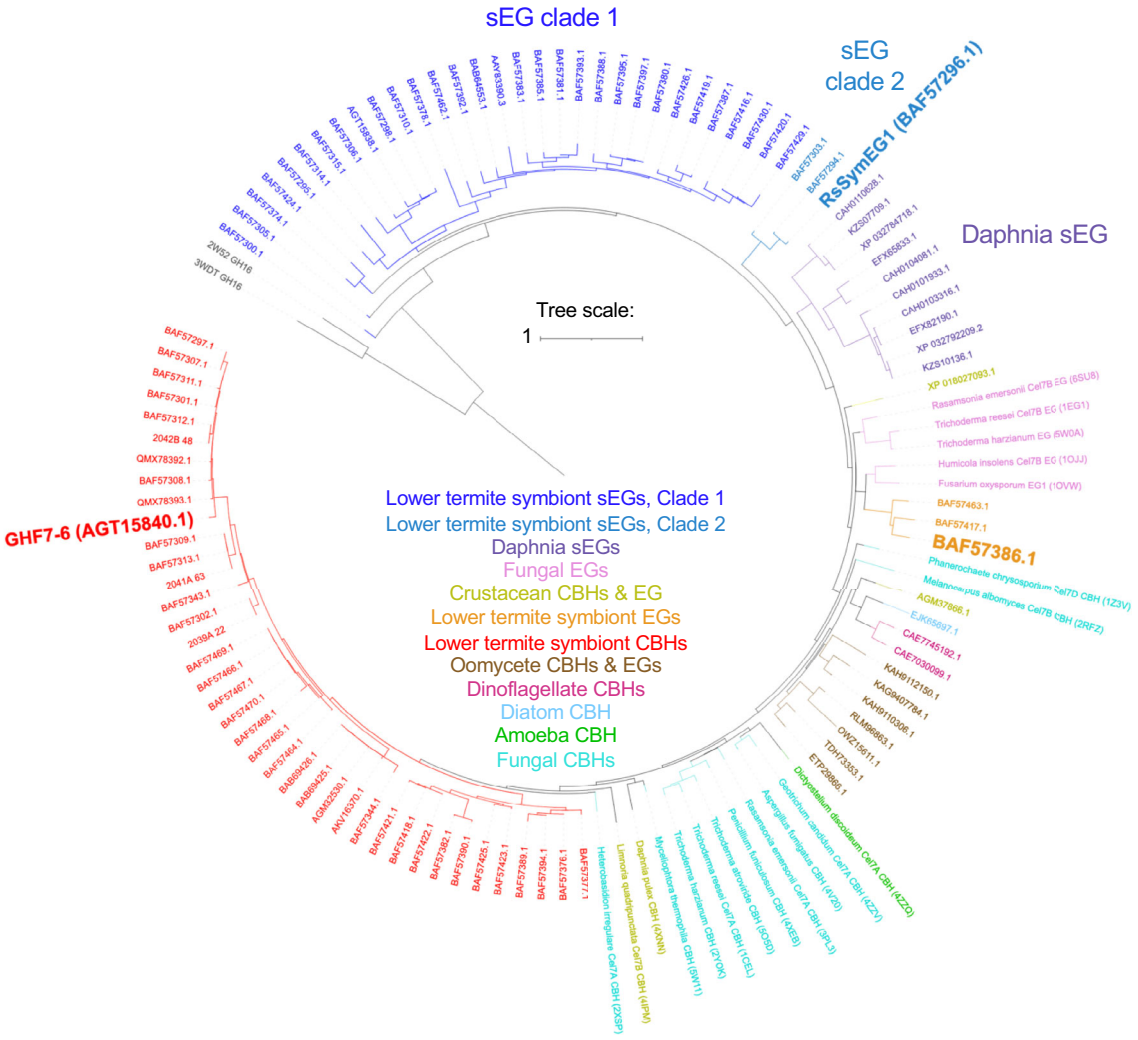


Fig. 6. Maximum likelihood phylogenetic tree of GH7 sequences. For visualization, 116 sequences were selected from the original tree of 371 GH7 sequences. Clades containing termite symbiont sequences are included in their entirety. Short endoglucanase (sEG) sequences from termite symbionts divide into two clades. RsSymEG1 clusters into the more sparsely populated clade of the two, with only two other closely related sequences. Two sequences from the distantly related family GH16 (black) were included to root the tree. The tree was built with IQ-TREE web server using the default settings (bootstrap alignments $n = 1000$), based on an alignment generated with MAFFT-L-INS-I algorithm on the MAFFT web server, and visualized using iTOL web server [64,66,67]. Figure and legend reproduced from [70] with modifications. The copyrights to the doctoral thesis [70] and the original image this has been reproduced from are held by the author, Topi Haataja. The authors and The FEBS Journal have been granted permission to reproduce the image. The images have also been licensed under the Creative Commons license CC BY NC 4.0, allowing non-commercial use.

clusters into a more sparsely populated clade of the two, with only two other closely related sequences. In contrast, the second clade consists of 27 sEG sequences. Interestingly, there is also a third clade containing sequences very similar to the termite symbiont sEGs in their features, with large deletions in similar locations. These sequences originate from different species of small water-dwelling crustaceans of the genus *Daphnia* (water fleas), with 10 sequences showing up in our screen. Structures of a GH7 CBH from *Daphnia pulex* have been deposited previously (e.g., PDB entry: 4XNN), but we are not aware of any of the sEGs from these organisms being characterized so far [34]. The full-length EGs from termite symbionts locate in a separate cluster, closer to fungal EGs. Furthermore, our analysis indicates 38 CBH sequences from lower termite symbionts, with all of these sequences clustering together as their own clade,

separately from other CBHs. Notably, none of the examined GH7 sequences from lower termite symbionts contain a carbohydrate-binding module (CBM).

Activity measurements

The enzymatic activity of RsSymEG1 was tested by analyzing release of reducing sugars from glucomannan, carboxymethyl cellulose (CMC), and barley β-glucan in order to provide a comparison to a commercially available GH7 EG, *Trichoderma longibrachiatum* Cel7B (TloCel7B). Hydrolysis experiments were conducted at two different pHs, pH 5.0 representing typical conditions for cellulase trials, and pH 6.5 which was previously determined to be optimal for RsSymEG1 for CMC hydrolysis [19]. On all three tested substrates at both pHs, RsSymEG1 demonstrated higher initial hydrolysis rates compared to TloCel7B at equimolar concentrations (Fig. 7). After the fast initial hydrolysis from 0 to 8 h, however, sugar release by RsSymEG1 plateaus more rapidly, in some cases leading to lower final released sugar concentrations at 74 h compared to TloCel7B. This effect is especially clear on glucomannan, where by 48 h the hydrolysis by TloCel7B has surpassed that of the RsSymEG1 at both pH 5.0 and 6.5. On CMC at pH 5.0, TloCel7B shows higher level of reducing sugars already at 24 h, while at pH 6.5 the performance of the *Trichoderma* enzyme is subdued and does not surpass that of RsSymEG1 even at 74 h. On barley β-glucan on the other hand, both enzymes reach very similar degrees of hydrolysis, at 74 h the final reducing sugar values being essentially equal for the two enzymes.

A qualitative high-throughput GlycoSpot assay demonstrated soluble sugar release by RsSymEG1 from dyed cellulose and barley glucan (Fig. 7D). TloCel7B was used as a control, and in contrast to RsSymEG1 showed activity on xylan, xyloglucan, and galactomannan, in addition to cellulose and barley glucan. With RsSymEG1, activity was seen also on amylose, but this stems most likely from a slight α-amylase contamination from the expression culture as RsSymEG1 was co-expressed with an *Aspergillus oryzae* amylase, as described previously [19].

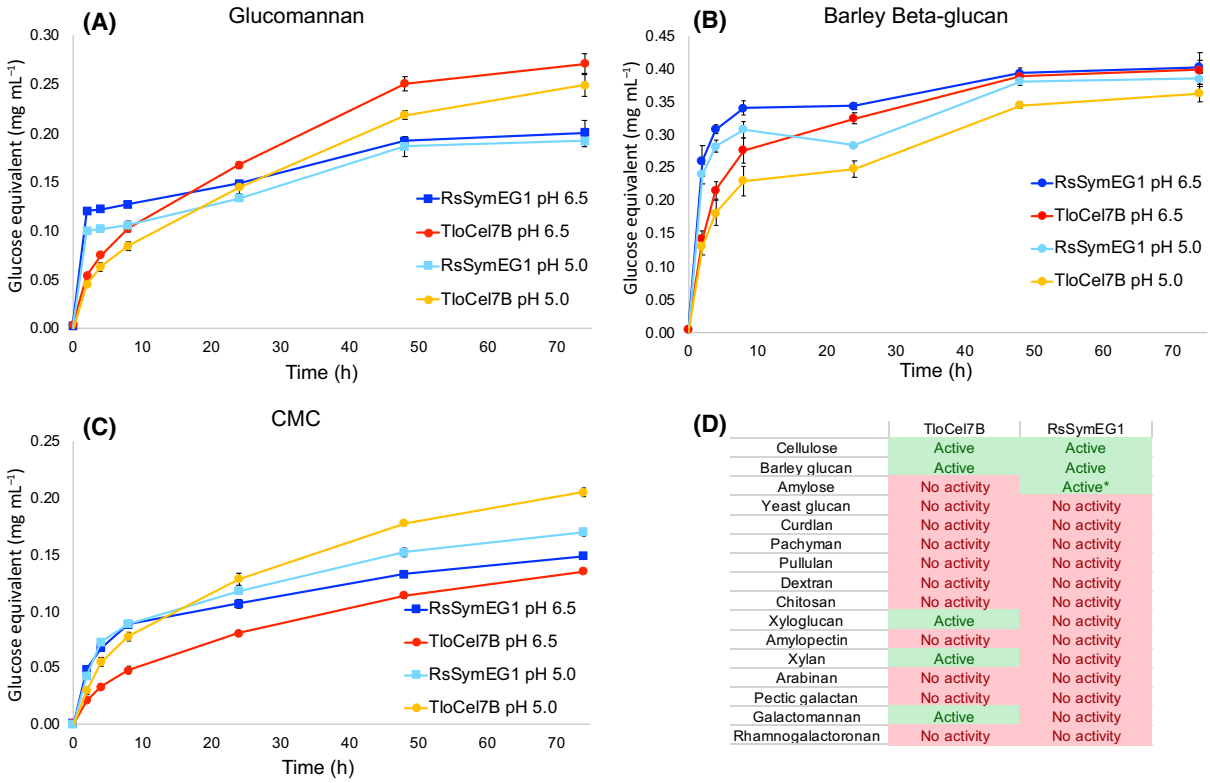
Discussion

In this work, we have successfully solved the first x-ray crystal structure of an enzyme belonging to the group of short family GH7 endoglucanases from termite symbiont protists, the RsSymEG1, and explored the diversity of GH7 enzymes in protozoa in lower termites. The short termite symbiont GH7 sequences have received surprisingly little attention in academic literature given their divergence from other GH7 sequences, and certain interesting characteristics of their sequence and activity profiles, as well as the industrial importance of this enzyme family. With increasing number of parabasalid sequencing data becoming available, more of similar short GH7 endoglucanases have consistently appeared, and now show up as distinct phylogenetic clades, highlighting that these are in fact viable, active protein sequences (Fig. 6) [16].

RsSymEG1 structure

As the first published crystal structure of this enzyme type, the RsSymEG1 structure provides important insights into the structure–function characteristics of these novel GH7 enzymes. While the core structure of the enzyme clearly conforms to the archetypal fold of GH7 enzymes, and the residues at the active site are highly conserved, the significantly shorter primary structure has major implications to the shape of the enzyme and the configuration of the substrate binding site. The complete lack of the B2 and B3 loops found in GH7 CBHs, as well as the B1 loop present in typical fungal GH7 EGs, leads to a remarkably open architecture of the substrate binding site. It is possible that this could have a major effect on accessibility of substrate at the active site. The missing B1 loop could facilitate binding of a more diverse set of carbohydrates, and the flat surface of the enzyme at the substrate binding face potentially increases accessibility to cellulose chains which might not be accessible for endo-cleavage to typical GH7 enzymes with prominent B1 and/or B2 loops, for example, when adjoining cellulose chains block binding (Fig. 8). Furthermore, looking at the shape of the binding site, it is reasonable to hypothesize that this open architecture could even accommodate productive binding of branched chain carbohydrates for cleavage. However, no indication of this was seen in the GlycoSpot assay, which only showed activity on typical GH7 EG substrates cellulose and barley β-glucan, and not, for example, on xyloglucan (Fig. 7D).

Interestingly, while the missing B1 and B2 loop structures lead to seemingly more exposed −3, −2, and −1 binding sites, the binding groove is notably narrow at subsites −6 to −4 due to the space taken up by the sidechains of Trp61 and Trp65 located on the A1 loop. Although there may be some flexibility within the A1 loop, it is difficult to see the position of Trp65 changing sufficiently to significantly open up the −4 binding site given its position adjoining a



(D)

	TloCel7B	RsSymEG1
Cellulose	Active	Active
Barley glucan	Active	Active
Amylose	No activity	Active*
Yeast glucan	No activity	No activity
Curdlan	No activity	No activity
Pachyman	No activity	No activity
Pullulan	No activity	No activity
Dextran	No activity	No activity
Chitosan	No activity	No activity
Xyloglucan	Active	No activity
Amylopectin	No activity	No activity
Xylan	Active	No activity
Arabinan	No activity	No activity
Pectic galactan	No activity	No activity
Galactomannan	Active	No activity
Rhamnogalactoronan	No activity	No activity

Fig. 7. Enzymatic activity of RsSymEG1 compared with the fungal GH7 endoglucanase TloCel7B. (A–C) Progress curves of hydrolysis of 0.5% glucomannan, barley beta-glucan, and carboxymethyl cellulose (CMC), respectively, at 20 °C, pH 5.0 and 6.5. Reducing sugar was quantified with p-hydroxybenzoic acid hydrazide (PHBAH) against glucose standards. Mean values are derived from one experiment, error bars represent SD ($n = 3$). (D) Qualitative results from Glycospot substrate specificity assay against a panel of dyed polysaccharides ($n = 1$). *The activity against amylose is probably a false-positive result caused by alpha-amylase contamination from the protein expression system.

structural β-strand. The same is true for the short α-helix (Thr24-Arg27) forming the opposite side of the binding groove at the −4 and −5 sites. Thus, while there is possibly more freedom than in other GH7s for the positioning of a cellulose chain for productive binding at the −1 to −3 sites, binding into the binding groove further away from the active site could be rather restricted regarding substrate conformation. Based on sequence alignments, the two tryptophan residues seen in the A1 loop of RsSymEG1 (Trp61 and Trp65) are not a conserved feature within the termite symbiont sEG sequences, with both of them present only in two other closely related sequences (Genbank: BAF57303.1 and BAF57294.1; Fig. S1), while one of the residues (corresponding to Trp61 in RsSymEG1) is seen in further three symbiont sEGs (Genbank: BAF57295.1, AAY83390.3, and BAB64553.1). Interestingly, also two of the *Daphnia magna* sEG sequences (Genbank: XP_032784718.1, KZS07709.1) have a tryptophan residue in this region, although likely display a significantly shorter A1 loop (Fig. S1).

Another notable feature of the RsSymEG1 structure is the architecture of the N terminus of the enzyme, where all hitherto structurally characterized GH7 enzymes contain a structurally inherent PCA residue, while no such composition is seen in the RsSymEG1 [1]. Previous studies have demonstrated the importance of this feature to the thermal stability of GH7 enzymes, with the inability of expression hosts such as *S. cerevisiae* to sufficiently perform this post-translational modification leading to reduced thermal stability [35]. Perhaps this is also an explaining factor to the relatively low thermal stability of RsSymEG1, as well as other similar sEGs in previous studies, as a consequence of the absence of stability increasing hydrophobic interactions seen in a typical PCA cleft, as well as the lack of capping structure at the peptide chain end, potentially leading it to more readily serve as an initiation point for unfolding. Interestingly, Woon *et al.* [18] showed partial recovery of activity of a similar enzyme GH1254 at lower temperatures after heat inactivation at 45 °C, suggesting spontaneous

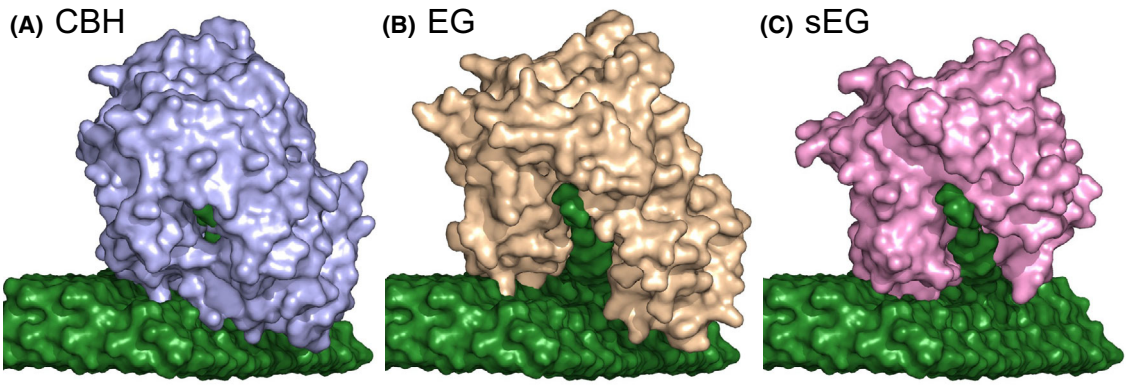


Fig. 8. Space-filling models of symbiont GH7s engaged with a cellulose chain on the surface of a cellulose microfibril. (A) Predicted structure of the cellobiohydrolase (CBH) GHF7-6. (B) Predicted structure from the endoglucanase (EG) sequence BAF57386.1. (C) Crystal structure of the short GH7 endoglucanase (sEG) RsSymEG1. The more flat surface of RsSymEG1 potentially increases its accessibility to cellulose chains at the surface of the microfibril. The models were created by superposing the structures with the catalytic domain of full-length TreCel7A (with linker-CBM) that had been docked to a cellulose microfibril by MD simulation [71]. Atom coordinates of the TreCel7A on microfibril model were kindly provided by CM Payne. The structure images were created with MACPYMOL [69].

refolding of the protein. Another factor contributing to the relatively low thermal stability could be the low number of stabilizing disulfide bridges in RsSymEG1, only four, compared to, for example, TreCel7A containing 10, and *Tre*Cel7B stability demonstrably increasing with addition of disulfide bridges through selected mutations [36]. Given that Cleveland observed early on that cellulose digesting lower termite symbiont protozoa die rapidly when temperature is raised to 36 °C, it is perhaps not terribly surprising that these enzymes have not evolved toward high thermal stability [9,10].

Structure models of a lower termite symbiont CBH and EG and significance in their hosts

Given the high degree of sequence conservation, GH7 CBHs and EGs can be modeled with relatively high confidence. In this work, we chose to model the GHF7-6 previously characterized by Sethi *et al*. [17], as an example of a termite symbiont GH7 CBH, and BAF57386 representing a more typical GH7 EG. Our modeling suggests that overall GHF7-6 presents highly typical features for GH7 CBHs, with characteristic residues within the catalytic site and substrate binding tunnel. Several studies have established the prevalence of GH7 CBH transcripts in termite symbiont protists through RNA analysis, and protein extract analysis also pointing to them as the major cellulase component in the hindguts of lower termites [15,37–39]. In the case of GHF7-6, its viability after heterologous expression was previously confirmed through enzyme activity measurements by Sethi *et al*. [17], and its expression in the native host verified through qRT-PCR experiments with protist samples isolated from *R. flavipes* gut.

When it comes to BAF57386, based on our modeling also this enzyme conforms to highly typical features seen in previous GH7 endoglucanase structures. While there is a large number of lower termite symbiont protist GH7 CBH and sEG sequences available, sequences from these organisms conforming to the typical "full length" GH7 EG features are scarce, BAF57386 being one of only three such sequences we were able to find. This could suggest that the biological significance of these GH7 EGs in these organisms is limited, with their functional role perhaps overlapping that of sEGs such as RsSymEG1. It is of course possible that these sequences originate from a contamination with a non-protist organism, for example, fungi, thus explaining the scarcity in termite symbiont protist samples. However, the closest matches in the NCBI non-redundant database share less than 45% identity to BAF57386, discounting the contamination hypothesis to some extent. While the three full-length EG sequences were identified in a study by Todaka *et al*. [38], as transcripts from enriched protist samples from lower termite hindguts, to our knowledge, their expression levels have not been assessed. However, there is nothing in the sequence or structure model of BAF57386, suggesting that it would not be a viable, active enzyme. It remains to be seen if further sequencing and omics data from lower termite symbiont protists show the full-length EGs to be as prevalent as CBHs and sEGs in these organisms, but as of now, this does not seem to be the case based on the number of sequences detected.

Phylogenetics

Given the ancient origins and conserved nature of the GH7 enzyme family, presumably due to continuous gravitation toward a few optimal structure architectures, it is likely that significant degree of evolutionary information has been lost due to repeated mutations. Therefore, any findings from a phylogenetic analysis of these sequences should be taken with appropriate caution. In our analysis, we have taken an approach where gap regions in the sequence alignments have been removed and not taken into account in the phylogenetic analysis in order to mitigate the tendency of sequences with more gaps to group together. This way we have strived to shed light on the true evolutionary relationships between CBHs, EGs, and sEGs from termite symbiont protozoa, as well as other organisms throughout the eukaryotic tree.

For the protozoan GH7s, all three types (CBHs, EGs, and sEGs) of sequences are found when non-cultured sequences from termite symbionts are included. None of these three types show a close relation to each other, suggesting that the divergence has not been a recent event within these protists. On the other hand, sEG sequences with very similar characteristics to the parabasalid sEGs are found from water flea (*Daphnia*) species, even if these two groups do not show a close relation to each other (Fig. S1; Fig. 6). The fact that this type of short GH7 endoglucanase is seen not only in lower termite symbiont protists, but in *Daphnia* species as well, could suggest that there is a functional benefit to the architecture, even if their prevalence is seemingly limited, and absence in fungi is notable. The existence of two separate clades of sEGs within the lower termite symbionts could possibly be explained by the sequences originating from two different taxa of Excavata, that is, parabasalids and oxymonads, the two main groups of symbiotic protists in lower termites [40]. While the existence of GH7 enzymes has been verified in parabasalid species in lower termite guts by single-cell transcriptomics, we are not aware of similar evidence for oxymonads [13]. However, most of the sequencing data available for sEGs is derived from aggregate samples of protists in the hindguts of lower termites, and do not, therefore, differentiate between parabasalids and oxymonads. There are no stand-out differences in the sequences of the two clades of sEGs from termite symbiont protists, suggesting that they are functionally similar.

Notably, none of the GH7 enzyme sequences from termite symbiont protists analyzed in this study contain a CBM. In fact, family 1 CBMs, the type often seen associated with GH7 enzymes in fungi, seem to be completely absent in parabasalids and oxymonads, based on available sequences. Considering that the main function of CBMs is to increase enzyme productivity by increasing their concentration on substrate surfaces, the presumably high substrate loading of fragmented cellulosic material in the confined digestive tracts of lower termites does not necessarily represent an environment where a CBM offers a significant advantage. As Kern *et al.* [23] hypothesized in the case of the marine wood borer *Limnoria quadripunctata*, this could explain why these organisms have not adopted, or have lost, this feature which is abundant in many cellulolytic organisms. On a related note, while it remains unclear to which extent the open active site architecture of the sEGs decreases the substrate binding affinity compared to their longer GH7 counterparts with prominent B1 and/or B2 loops, also this could be an adaptation to high substrate concentrations. As demonstrated by Kari *et al.* [41], optimal binding affinity for cellulases is highly dependent on substrate loading.

Activity

To date, several studies have demonstrated enzymatic activity from enzymes in this group, albeit with varying success [17–19]. In a previous study, Todaka *et al.* [19], successfully expressed RsSymEG1 in *Aspergillus oryzae*, and characterized the activity in some detail. The authors demonstrated its high initial activity rates on CMC, yet substantially lower rates on microcrystalline Avicel. A pH activity profile on CMC showed a somewhat higher pH optimum at 6.5 than most studied GH7 enzymes. Similar activity characteristics on the model substrate 4-methylumbelliferyl β-D-cellobioside were reported by Woon *et al.* [18], for another similar enzyme GH1254 from a symbiont of *Coptotermes curvignathus*, expressed in *Pichia pastoris*, although no activity on Avicel or CMC could be demonstrated. Sethi *et al.* [17] successfully utilized a baculovirus–insect cell expression system to express three GH7 enzymes from symbionts of *R. flavipes*, one sEG and two CBHs, with all three demonstrating a pH optimum of 7 on 4-nitrophenyl β-D-cellobioside [17]. In our experiments on CMC, glucomannan, and barley β-glucan, high initial hydrolysis rates by RsSymEG1 were confirmed in direct comparisons to a typical GH7 EG, the commercially available TloCel7B. While we hypothesize that the open binding site structure could facilitate substrate binding and hydrolysis in areas which are not accessible to longer GH7 EGs, it is not clear what is behind the somewhat rapid leveling off of hydrolysis rates at lower conversion levels

compared to TloCel7B, seen especially on CMC and glucomannan. As the structure and previous product profile data by Todaka *et al.* [19] do not suggest that there would be a great degree of processivity, the plateauing activity is presumably not caused by strong product inhibition as is seen in CBHs. Also, given that the assays were conducted at a moderate temperature (20 °C), lower enzyme stability is likely not an explaining factor either [19]. Overall, however, it can be said that RsSymEG1 seems to be less "persistent" when it comes to the extent of conversion of the tested substrates.

Potential for enzyme engineering

The novelty of the RsSymEG1, the ability to express it in a fungal host, and the industrial importance of GH7 family enzymes make this structure a potential template for future protein engineering through both rational design and systematic approaches. For improving stability of the enzyme, previous enzyme engineering studies of GH7 cellulases would provide ample support for targeted improvements. Interestingly, beneficial mutations in studies conducted via GH7 expression in *Saccharomyces cerevisiae* are often seen in the region surrounding the N terminus, resulting in improvement in stability of enzymes where a PCA structure is lacking, and thus perhaps providing insights for RsSymEG1 improvement as well [35,42]. The relatively small size and natively non-capped N terminus could also make RsSymEG1 and other sEGs interesting candidates for various multidomain fusion proteins. Inspired by the highly effective multidomain *Caldicellulosiruptor bescii* cellulase CelA, Brunecky *et al.*, recently constructed synthetic cellulases containing a GH7 CBH domain, a cellulose-binding module, and a GH5 EG domain, demonstrating enhanced degradation of cellulose compared to separate domains, as well as a boosting effect on lignocellulose degradation in combination with a commercial cellulase cocktail [43]. It would be highly interesting to see the performance of short GH7 endoglucanase domains in a similar setting, or in other synthetic constructs. Furthermore, it can be hypothesized that the smaller architecture with few disulfide bridges and less requirements for post-translational modifications (in terms of the missing PCA) could facilitate the use of other expression hosts that are often used for high-throughput engineering studies but have usually had limited success in expressing GH7 enzymes, for example, *Saccharomyces cerevisiae* [1,35]. Expression in *S. cerevisiae* would be especially interesting given it is commonly used in ethanol fermentations where simultaneous saccharification and fermentation (SSF) through enzyme expression in the fermenting organism can be a highly desirable feature. Indeed, Todaka *et al.* [44] reported expression of a number of sEG GH7s in *Saccharomyces*, although further studies are needed to demonstrate expression levels. Expression of short GH7s in *E. coli* has been reported as well, holding promise for simplified expression protocols for enzyme studies and production [39]. However, more investigations are required also in this regard in order to verify viability of bacterial expression in terms of expression levels and enzyme stability.

Conclusions

In this work, we have presented the structure of RsSymEG1, a novel short GH7 endoglucanase from a termite symbiont. Three GH7 architectures, CBHs, EGs, and what we have denoted here sEGs, have been identified in studies of protist symbionts of lower termites. While the prevalence of the full-length GH7 EGs in these organisms remains somewhat unclear, CBHs and sEGs have been shown to be major components of enzymatic machinery in lower termite hindguts [15,37–39]. The examples of a CBH and EG from these organisms we have examined in this study conform to typical features of these enzyme classes within GH7s. The structure of RsSymEG1, however, demonstrates a different architecture, with similar core fold, but notably open binding site groove structure. Phylogenetic analysis suggests that similar short architecture sEGs have possibly emerged separately, and are observed also in *Daphnia* species, even if the activity of these enzymes has not been demonstrated. Given the limited attention the short GH7 endoglucanases have garnered so far despite their interesting characteristics, we feel these enzymes certainly warrant further studies. The RsSymEG1 structure provides important information about the structure–function relationships within GH7s and provides a template for engineering new types of GH7 enzymes.

Materials and methods

Expression and purification of RsSymEG1

The amino acid sequence of native RsSymEG1 is found in GenBank accession no BAF57296.1. An *Aspergillus oryzae* strain expressing RsSymEG1 was a kind gift from Associate Professor Jun-ichi Maruyama and Associate Professor Manabu Arioka, Department of Biotechnology, The University of Tokyo, Japan. The expression strain and its construction has been described previously [19]. In the strain,

RsSymEG1 is expressed under a dextrin-inducible alpha-amylase promotor as a fusion protein of N-terminal *A. oryzae* alpha-amylase AmyB, linked by a Kex2 protease cleavage site containing linker peptide KRGGG to RsSymEG1 at the C terminus. The two proteins were separated in the cultures through cleavage at the linker peptide by endogenous proteases.

Three liters of 5X DPY medium in 1-L baffled flasks was inoculated with spores collected from one potato dextrose agar plate. The cultures were grown in +25 °C with 150 rpm shaking for 4 days. The cultures were harvested by filtrating consecutively through Whatman GF-B and GF-F glass filters (Cytiva, Uppsala, Sweden), and then 0.45 and 0.2 µm PES filters (VWR International, Radnor, PA, USA). RsSymEG1 was precipitated from the culture filtrate by adding ammonium sulfate to 80% saturation, at +4 °C and collected by centrifugation at 13 000 ***g*** for 20 min. The protein pellet was dissolved in 10 mM BISTRIS buffer, pH 6.0. To remove remaining ammonium sulfate and small molecular weight impurities, the protein solution was buffer exchanged on a Sephadex G-25 column (GE Healthcare, Uppsala, Sweden) into 10-mM BISTRIS, pH 6.0. The protein was subsequently purified by loading on a 20 mL SourceQ column (GE Healthcare) and eluted with a 1 M NaCl gradient, and further polished by running size exclusion chromatography on a 16/600 Superdex 75 PG column (GE Healthcare) with 10 mM BISTRIS, 0.15 M of NaCl, pH 6.0 buffer. For crystallization experiments, the enzyme was deglycosylated with Endo H by mixing 1 µg of Endo H per 1 mg of RsSymEG1 and incubating at 28 °C overnight. The protein was concentrated as needed with Vivaspin500 spin columns (Sartorius AG, Goettingen, Germany) with a 5 kDa cutoff. Protein concentrations were determined spectrophotometrically at 280 nm using extinction coefficients of 57 870 $M^{-1}\cdot cm^{-1}$ for RsSymEG1 and 75 635 $M^{-1}\cdot cm^{-1}$ for TloCel7B, calculated from the amino acid sequence using the PROTPARAM tool within Expasy portal [45].

Crystallization and structure determination

Crystals of the RsSymEG1 were obtained at 20 °C in Morpheus crystallization screen (Molecular Dimensions, Sheffield, UK) through a sitting drop vapor diffusion experiment in conditions containing 0.03 M Sodium fluoride, 0.03 M Sodium bromide, 0.03 M Sodium iodide, 12% v/v PEG 500 MME, 6% w/v PEG 20000, and 0.1 M Tris-Bicine, pH 8.5. The crystallization drop was set up with 0.6 µL of 28 $mg\cdot mL^{-1}$ RsSymEG1 in 10 mM BISTRIS pH 6.0 and 0.2 µL of the crystallization solution. Crystals were first observed 12 months after setting up the screen with approximately 50% of the mother liquor having evaporated, indicating that the crystals emerged at a very high protein concentration.

X-ray diffraction data were collected at the BioMAX beamline of the Max IV synchrotron in Lund, Sweden, using MXCUBE3 software for data collection [46,47]. The data were processed using XDSApp and Aimless in the PRESTO software package at BIOMAX computer cluster [48,49]. In Aimless, reflections with resolution higher than 1.85 Å were excluded, and 5% of reflections were assigned to an R_{free} set [49]. The structure was solved by molecular replacement using Phaser, using a homology model of RsSymEG1 as search model, which was generated using I-TASSER [50,51]. The structure was refined by several rounds of iterative refinement in Refmac and manual model building in COOT [52]. Aimless was run using the CCP4i interface, and Phaser, Refmac, and COOT were run using the CCP4i2 interface [52–55].

Coordinates for the cellononaose chain from the PDB entry 4C4C were extracted in PyMOL, and its potential binding to RsSymEG1 was modeled using the AUTODOCK VINA (version 1.2.1) extension of AMDOCK (version 1.5.2) with default settings, defining the search space as the substrate binding face of the enzyme [26,56–58].

Enzyme activity measurements

Enzyme assays with RsSymEG1 and *Trichoderma longibrachiatum* Cel7B (Megazyme, Bray, Ireland) on 0.5% w/v glucomannan (Megazyme), 0.5% w/v CMC (Sigma-Aldrich, Burlington, MA, USA), and 0.5% w/v barley beta-glucan (Megazyme) were conducted at 20 °C with 900 rpm shaking. Reactions with all substrates were conducted in two different buffers, 20 mM sodium acetate pH 5.0 and 20 mM MES hydrate pH 6.5 with enzyme concentrations of 30 nM for glucomannan and 7.5 nM for CMC and barley beta-glucan. Samples were taken at 2, 4, 8, 24, 48, and 74 h and analyzed for reducing sugars using a PHBAH assay essentially as described previously in [59]. Briefly, samples of 50 µL were mixed with 50 µL of 1 M NaOH to stop the reactions and subsequently mixed with 100 µL of PHBAH solution (0.5 M NaOH, 0.1 M PHBAH, and 0.2 M NaKTartrate). After incubating at 99 °C for 10 min, the samples were cooled down on ice for 10 min and absorbances were measured at 410 nm. Reducing sugar concentrations were calculated using glucose samples as standards. GlycoSpot enzyme assays were conducted using the GlycoSpot Discovery Kit (GlycoSpot ApS, Søborg, Denmark) according to the manufacturer's instructions. Briefly, 200 µL of 7.5 nM enzyme solutions (RsSymEG1 and TloCel7B) in 10 mM BISTRIS pH 6.0 buffer was incubated on the substrate reaction plates for 17 h at 25 °C with 130 rpm shaking, with blank samples containing only buffer used as a negative control. Subsequently, the samples were filtered through the filter of the reaction plate by vacuum, into a 96-well deep-well block. For each sample 150 µL was transferred to an optical 96-well plate and the absorbances were measured at 517 and 595 nm for the red and blue substrates, respectively. Enzyme samples showing at least twofold absorbance compared to the negative control were considered positive for activity toward the substrate.

Phylogenetic tree construction, homology models, and sequence alignments

GH7 sequences were searched with the blastp web server from the NCBI GenBank non-redundant database using the RsSymEG1 sequence as a query [60,61]. SIGNALP 6.0 was used to predict and remove signal secretion signals [33]. Sequences without a predicted secretion signal were discarded, and mature sequences were aligned with sequence of the TrCel7A catalytic domain using the MUSCLE-algorithm in UNIPRO UGENE version 37 [62,63]. All sequences were trimmed to the sequence region overlapping the TrCel7A catalytic domain, and sequences containing less than 300 amino acids within this region, or containing the word 'partial' in the sequence name were deleted. Furthermore, sequences missing amino acids corresponding to the catalytic triad (TrCel7A residues Glu212, Asp214, Glu217), or the core beta-sheets (corresponding to TreCel7A residues 83–87, 90–94, 125–131, 140–147, 223–228, 288–293, 300–305, 359–367, 416–425). Two protein sequences (Uniprot: Q874E3 and E0XN39, with respective PDB identifiers 2W52 and 3WDT) from the distantly related GH16 family were included in the sequence set in order to enable rooting of the tree. The sequences were subsequently re-aligned using MAFFT-L-INS-I algorithm on the MAFFT version 7 web server [64]. Gap regions were removed from the alignment using TRIMAL 1.3 on the Phylemon2 server [65]. IQ-TREE web server was used to build a maximum likelihood phylogenetic tree using the default settings (1000 bootstrap alignments) [66]. iTOL web server was used for visualizing selected branches of the phylogenetic tree [67]. Sequence alignments of selected termite symbiont GH7 sequences were constructed using the MUSCLE-algorithm in UNIPRO UGENE version 37 [62,63], and visualized using the ESPript 3.0 web server [68].

The structure model of AGT15840 (GHF7-6) was created with the MODELLER extension in Chimera [29,30], using a sequence alignment with sequences of the PDB structures 2RZF, 4XNN, 4IPM, 4CSI, 5W11, 4ZZQ, 5O5D, 1CEL, 2YOK, 4XEB, 4ZZV, 3PL3, 4V20, 1Z3V, and 2XSP aligned with MUSCLE algorithm in UGENE as input [62,63]. MODELLER model of BAF57386 was constructed the same way, but using a sequence alignment with sequences of the PDB structures 6SU8, 1EG1, 5W0A, 1OVW, and 1OJJ as input.

Acknowledgements

We gratefully thank Associate Professor Jun-ichi Maruyama and Associate Professor Manabu Arioka, Department of Biotechnology, The University of Tokyo, Japan, for providing the *Aspergillus oryzae* strain expressing RsSymEG1, and Associate Professor Christina M. Payne, Kentucky University, for the model of TreCel7A docked to a cellulose microfibril. Dr Igor Sabljic at our department (Molecular Sciences) is gratefully thanked for help with X-ray data collection and advice regarding the X-ray crystallography part of the study, and Professor Fabien Burki, Uppsala University, for advice regarding sequence trimming and construction of the phylogenetic tree. Funding for the research is gratefully acknowledged by the Swedish Energy Agency (Dnr 2015-009633). We acknowledge MAX IV Laboratory for time on Beamline BioMAX under Proposal 20180025. Research conducted at MAX IV, a Swedish national user facility, is supported by the Swedish Research Council under contract 2018-07152, the Swedish Governmental Agency for Innovation Systems under contract 2018-04969, and Formas under contract 2019-02496.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

TH: Investigation, Validation, Data curation, Writing – Original Draft, Writing – Review and Editing, Visualization; HH: Validation, Writing – Review and Editing; SM: Validation, Resources, Writing – Review and Editing; MS: Conceptualization, Validation, Writing – Review and Editing, Supervision, Funding acquisition; JS: Conceptualization, Writing – Original Draft, Writing – Review and Editing, Visualization, Supervision, Funding acquisition, Project administration.

Peer review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/febs.17029.

Data availability statement

GenBank: BAF57296.1 (amino acid sequence of RsSymEG1 protein). PDB ID: 8POF (crystal structure of RsSymEG1).

References

1 Payne CM, Knott BC, Mayes HB, Hansson H, Himmel ME, Sandgren M, Ståhlberg J & Beckham GT (2015) Fungal cellulases. *Chem Rev* **115**, 1308–1448.
2 Hatakka A & Hammel KE (2010) Fungal biodegradation of lignocelluloses. In Industrial Applications (Esser K & Hofrichter M, eds), pp. 319–340. Springer Berlin Heidelberg, Berlin, Heidelberg.

3 Taylor LE, Knott BC, Baker JO, Alahuhta PM, Hobdey SE, Linger JG, Lunin VV, Amore A, Subramanian V, Podkaminer K *et al.* (2018) Engineering enhanced cellobiohydrolase activity. *Nat Commun* **9**, 1–10.
4 Goedegebuur F, Dankmeyer L, Gualfetti P, Karkehabadi S, Hansson H, Jana S, Huynh V, Kelemen BR, Kruithof P, Larenas EA *et al.* (2017) Improving the thermal stability of cellobiohydrolase Cel7A from *Hypocrea jecorina* by directed evolution. *J Biol Chem* **292**, 17418–17430.
5 Hobdey SE, Knott BC, Momeni MH, Taylor LE, Borisova AS, Podkaminer KK, VanderWall TA, Himmel ME, Decker SR, Beckham GT *et al.* (2016) Biochemical and structural characterizations of two *Dictyostelium* cellobiohydrolases from the Amoebozoa kingdom reveal a high level of conservation between distant phylogenetic trees of life. *Appl Environ Microbiol* **82**, 3395–3409.
6 Gado JE, Harrison BE, Sandgren M, Ståhlberg J, Beckham GT & Payne CM (2021) Machine learning reveals sequence-function relationships in family 7 glycoside hydrolases. *J Biol Chem* **297**, 100931.
7 Schiano-di-Cola C, Røjel N, Jensen K, Kari J, Sørensen TH, Borch K & Westh P (2019) Systematic deletions in the cellobiohydrolase (CBH) Cel7A from the fungus *Trichoderma reesei* reveal flexible loops critical for CBH activity. *J Biol Chem* **294**, 1807–1815.
8 Von Ossowski I, Ståhlberg J, Koivula A, Piens K, Becker D, Boer H, Harle R, Harris M, Divne C, Mahdi S *et al.* (2003) Engineering the exo-loop of *Trichoderma reesei* cellobiohydrolase, Cel7A. A comparison with *Phanerochaete chrysosporium* Cel7D. *J Mol Biol* **333**, 817–829.
9 Cleveland LR (1923) Symbiosis between termites and their intestinal protozoa. *Proc Natl Acad Sci USA* **9**, 424–428.
10 Cleveland LR (1924) The physiological and symbiotic relationships between the intestinal protozoa of termites and their host, with special reference to *Reticulitermes flavipes* Kollar. *Biol Bull* **46**, 178–201.
11 Ohkuma M (2003) Termite symbiotic systems: efficient bio-recycling of lignocellulose. *Appl Microbiol Biotechnol* **61**, 1–9.
12 Karnkowska A, Vacek V, Zubáčová Z, Treitli SC, Petrželková R, Eme L, Novák L, Žárský V, Barlow LD, Herman EK *et al.* (2016) A eukaryote without a mitochondrial organelle. *Curr Biol* **26**, 1274–1284.
13 Nishimura Y, Otagiri M, Yuki M, Shimizu M, Inoue J-I, Moriya S & Ohkuma M (2020) Division of functional roles for termite gut protists revealed by single-cell transcriptomes. *ISME J* **14**, 2449–2460.
14 Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V *et al.* (2012) The revised classification of eukaryotes. *J Eukaryot Microbiol* **59**, 429–514.
15 Todaka N, Moriya S, Saita K, Hondo T, Kiuchi I, Takasu H, Ohkuma M, Piero C, Hayashizaki Y & Kudo T (2007) Environmental cDNA analysis of the genes involved in lignocellulose digestion in the symbiotic protist community of *Reticulitermes speratus*. *FEMS Microbiol Ecol* **59**, 592–599.
16 Schiano-di-Cola C, Koøaczkowski B, Sørensen TH, Christensen SJ, Cavaleiro AM, Windahl MS, Borch K, Morth JP & Westh P (2019) Structural and biochemical characterization of a family 7 highly thermostable endoglucanase from the fungus *Rasamsonia emersonii*. *FEBS J* **287**, 2577–2596.
17 Sethi A, Kovaleva ES, Slack JM, Brown S, Buchman GW & Scharf ME (2013) A GHF7 cellulase from the protist symbiont community of reticulitermes flavipes enables more efficient lignocellulose processing by host enzymes. *Arch Insect Biochem Physiol* **84**, 175–193.
18 Woon JSK, King PJH, Mackeen MM, Mahadi NM, Wan Seman WMK, Broughton WJ, Abdul Murad AM & Abu Bakar FD (2017) Cloning, production and characterization of a glycoside hydrolase family 7 enzyme from the gut microbiota of the termite *Coptotermes curvignathus*. *Mol Biotechnol* **59**, 271–283.
19 Todaka N, Lopez CM, Inoue T, Saita K, Maruyama JI, Arioka M, Kitamoto K, Kudo T & Moriya S (2010) Heterologous expression and characterization of an endoglucanase from a symbiotic protist of the lower termite, *Reticulitermes speratus*. *Appl Biochem Biotechnol* **160**, 1168–1178.
20 Ubhayasekera W, Muñoz IG, Vasella A, Ståhlberg J & Mowbray SL (2005) Structures of *Phanerochaete chrysosporium* Cel7D in complex with product and inhibitors. *FEBS J* **272**, 1952–1964.
21 Haddad Momeni M, Payne CM, Hansson H, Mikkelsen NE, Svedberg J, Engström A, Sandgren M, Beckham GT & Stahlberg J (2013) Structural, biochemical, and computational characterization of the glycoside hydrolase family 7 cellobiohydrolase of the tree-killing fungus *Heterobasidion irregulare*. *J Biol Chem* **288**, 5861–5872.
22 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.
23 Kern M, McGeehan JE, Streeter SD, Martin RNA, Besser K, Elias L, Eborall W, Malyon GP, Payne CM, Himmel ME *et al.* (2013) Structural characterization of a unique marine animal family 7 cellobiohydrolase suggests a mechanism of cellulase salt tolerance. *Proc Natl Acad Sci USA* **110**, 10189–10194.
24 Parkkinen T, Koivula A, Vehmaanperä J & Rouvinen J (2008) Crystal structures of *Melanocarpus albomyces*

cellobiohydrolase Cel7B in complex with cello-oligomers show high flexibility in the substrate binding. *Protein Sci* **17**, 1383–1394.
25 Kadowaki MAS, Higasi P, Godoy MO, Prade RA & Polikarpov I (2018) Biochemical and structural insights into a thermostable cellobiohydrolase from *Myceliophthora thermophila*. *FEBS J* **285**, 559–579.
26 Eberhardt J, Santos-Martins D, Tillack AF & Forli S (2021) AutoDock Vina 1.2.0: new docking methods, expanded force field, and python bindings. *J Chem Inf Model* **61**, 3891–3898.
27 Olsen JP, Kari J, Windahl MS, Borch K & Westh P (2020) Molecular recognition in the product site of cellobiohydrolase Cel7A regulates processive step length. *Biochem J* **477**, 99–110.
28 Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Dustin Schaeffer R *et al*. (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876.
29 Webb B & Sali A (2016) Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinforma* **54**, 5.6.1–5.6.37.
30 Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC & Ferrin TE (2004) UCSF chimera – a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612.
31 Nakamura A, Tsukada T, Auer S, Furuta T, Wada M, Koivula A, Igarashi K & Samejima M (2013) The tryptophan residue at the active site tunnel entrance of *Trichoderma reesei* cellobiohydrolase Cel7A is important for initiation of degradation of crystalline cellulose. *J Biol Chem* **288**, 13503–13510.
32 Nakamura A, Kanazawa T, Furuta T, Sakurai M, Saloheimo M, Samejima M, Koivula A & Igarashi K (2021) Role of tryptophan 38 in loading substrate chain into the active-site tunnel of cellobiohydrolase I from *Trichoderma reesei*. *J Appl Glycosci (1999)* **68**, 19–29.
33 Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G & Nielsen H (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol* **40**, 1023–1025.
34 Bury CS, Carmichael I & Garman EF (2017) OH cleavage from tyrosine: debunking a myth. *J Synchrotron Radiat* **24**, 7–18.
35 Dana CM, Dotson-Fagerstrom A, Roche CM, Kal SM, Chokhawala HA, Blanch HW & Clark DS (2014) The importance of pyroglutamate in cellulase Cel7A. *Biotechnol Bioeng* **111**, 842–847.
36 Zhang S, Wang Y, Song X, Hong J, Zhang Y & Yao L (2014) Improving *Trichoderma reesei* Cel7B thermostability by targeting the weak spots. *J Chem Inf Model* **54**, 2826–2833.
37 Geng A, Cheng Y, Wang Y, Zhu D, Le Y, Wu J, Xie R, Yuan JS & Sun J (2018) Transcriptome analysis of the digestive system of a wood-feeding termite (*Coptotermes formosanus*) revealed a unique mechanism for effective biomass degradation. *Biotechnol Biofuels* **11**, 1–14.
38 Todaka N, Inoue T, Saita K, Ohkuma M, Nalepa CA, Lenz M, Kudo T & Moriya S (2010) Phylogenetic analysis of cellulolytic enzyme genes from representative lineages of termites and a related cockroach. *PLoS One* **5**, e8636.
39 Watanabe H, Nakashima K, Saito H & Slaytor M (2002) New endo-β-1,4-glucanases from the parabasalian symbionts, *Pseudotrichonympha grassii* and *Holomastigotoides mirabile* of *Coptotermes* termites. *Cell Mol Life Sci* **59**, 1983–1992.
40 Yuki M, Moriya S, Inoue T & Kudo T (2008) Transcriptome analysis of the digestive organs of Hodotermopsis sjostedti, a lower termite that hosts mutualistic microorganisms in its hindgut. *Zoolog Sci* **25**, 401–406.
41 Kari J, Molina GA, Schaller KS, Schiano-di-Cola C, Christensen SJ, Badino SF, Sørensen TH, Røjel NS, Keller MB, Sørensen NR *et al*. (2021) Physical constraints and functional plasticity of cellulases. *Nat Commun* **12**, 1–10.
42 Dana CM, Saija P, Kal SM, Bryan MB, Blanch HW & Clark DS (2012) Biased clique shuffling reveals stabilizing mutations in cellulase Cel7A. *Biotechnol Bioeng* **109**, 2710–2719.
43 Brunecky R, Subramanian V, Yarbrough JM, Donohoe BS, Vinzant TB, Vanderwall TA, Knott BC, Chaudhari YB, Bomble YJ, Himmel ME *et al*. (2020) Synthetic fungal multifunctional cellulases for enhanced biomass conversion. *Green Chem* **22**, 478–489.
44 Todaka N, Nakamura R, Moriya S, Ohkuma M, Kudo T, Takahashi H & Ishida N (2011) Screening of optimal cellulases from symbiotic protists of termites through expression in the secretory pathway of *Saccharomyces cerevisiae*. *Biosci Biotechnol Biochem* **75**, 2260–2263.
45 Walker JM, Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD & Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. In The Proteomics Protocols Handbook (Walker JM, ed.), pp. 571–607. Humana Press, Totowa, NJ.
46 Mueller U, Thunnissen M, Nan J, Eguiraun M, Bolmsten F, Milàn-Otero A, Guijarro M, Oscarsson M, de Sanctis D & Leonard G (2017) MXCuBE3: a new era of MX-beamline control begins. *Synchrotron Radiat News* **30**, 22–27.
47 Ursby T, Hnberg KA, Appio R, Aurelius O, Barczyk A, Bartalesi A, Bjelcic M, Bolmsten F, Cerenius Y, Doak RB *et al*. (2020) BioMAX the first

macromolecular crystallography beamline at MAX IV laboratory. *J Synchrotron Radiat* **27**, 1415–1429.
48 Kabsch W (2010) XDS. *Acta Crystallogr D Biol Crystallogr* **66**, 125–132.
49 Evans PR & Murshudov GN (2013) How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* **69**, 1204–1214.
50 McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC & Read RJ (2007) Phaser crystallographic software. *J Appl Cryst* **40**, 658–674.
51 Yang J & Zhang Y (2015) I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res* **43**, W174–W181.
52 Emsley P, Lohkamp B, Scott WG & Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501.
53 Kovalevskiy O, Nicholls RA, Long F, Carlon A & Murshudov GN (2018) Overview of refinement procedures within REFMAC 5: utilizing data from different sources. *Acta Crystallogr D Struct Biol* **74**, 215–227.
54 Potterton L, Agirre J, Ballard C, Cowtan K, Dodson E, Evans PR, Jenkins HT, Keegan R, Krissinel E, Stevenson K *et al.* (2018) CCP 4 i 2: the new graphical user interface to the CCP 4 program suite. *Acta Crystallogr D Struct Biol* **74**, 68–84.
55 Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A *et al.* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* **67**, 235–242.
56 Valdés-Tresanco MS, Valdés-Tresanco ME, Valiente PA & Moreno E (2020) AMDock: a versatile graphical tool for assisting molecular docking with Autodock Vina and Autodock4. *Biol Direct* **15**, 1–12.
57 Trott O & Olson AJ (2009) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **31**, 455–461.
58 Schrödinger L (2010) The PyMOL Molecular Graphics System 1.5.0.4. Schrödinger, LLC, New York, NY.
59 Hori C, Igarashi K, Katayama A & Samejima M (2011) Effects of xylan and starch on secretome of the basidiomycete *Phanerochaete chrysosporium* grown on cellulose. *FEMS Microbiol Lett* **321**, 14–23.
60 Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403–410.
61 Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res* **41**, 29–33.
62 Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 1–19.
63 Okonechnikov K, Golosova O, Fursov M, Varlamov A, Vaskin Y, Efremov I, German Grehov OG, Kandrov D, Rasputin K, Syabro M *et al.* (2012) Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167.
64 Katoh K, Rozewicki J & Yamada KD (2018) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform* **20**, 1160–1166.
65 Sánchez R, Serra F, Tárraga J, Medina I, Carbonell J, Pulido L, De María A, Capella-Gutiérrez S, Huerta-Cepas J, Gabaldón T *et al.* (2011) Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic Acids Res* **39**, 470–474.
66 Trifinopoulos J, Nguyen LT, von Haeseler A & Minh BQ (2016) W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res* **44**, W232–W235.
67 Letunic I & Bork P (2021) Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293–W296.
68 Robert X & Gouet P (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res* **42**, 320–324.
69 Schrödinger L (2015) The PyMOL Molecular Graphics System, Version 1.8.6. Schrödinger, LLC, Cambridge, MA.
70 Haataja T (2023) Structure-function studies of GH7 cellulases, key enzymes in the global carbon cycle. *Acta Univ Agric Sueciae* **2023:9**. Doctoral thesis, Swedish University of Agricultural Sciences, Sweden. https://doi.org/10.54612/a.672jon1kdu
71 Payne CM, Resch MG, Chen L, Crowley MF, Himmel ME, Taylor LE, Sandgren M, Ståhlberg J, Stals I, Tan Z *et al.* (2013) Glycosylated linkers in multimodular lignocellulose-degrading enzymes dynamically bind to cellulose. *Proc Natl Acad Sci USA* **110**, 14646–14651.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Amino acid sequence alignment of selected short GH7 endoglucanases (sEG) with fungal GH7 CBH TreCel7A and EG TreCel7B.

Fig. S2. Amino acid sequence alignment of GH7 cellobiohydrolases (CBH) and endoglucanases (EG) of termite symbionts, and selected fungal CBHs and EGs.