The enzymatic degradation of plant cell walls is an important biological process of increasing environmental and industrial significance. Xylan, a major component of the plant cell wall, consists of a backbone of β-1,4-xylose (Xylp) units that are often decorated with arabinofuranose (Araf) side chains. A large penta-modular enzyme, Ctxyl5A, was shown previously to specifically target arabinoxylans. The mechanism of substrate recognition displayed by the enzyme, however, remains unclear. Here we report the crystal structure of the arabinoxylanase and the enzyme in complex with ligands. The data showed that four of the protein modules adopt a rigid structure, which stabilizes the catalytic domain. The C-terminal non-catalytic carbohydrate binding module could not be observed in the crystal structure, suggesting positional flexibility. The structure of the enzyme in complex with Xylp-β-1,4-Xylp-β-1,4-Xylp-[α-1,3-Araf]-β-1,4-Xylp showed that the Araf decoration linked O2 to the xylose in the active site is located in the pocket (−2* subsite) that abuts onto the catalytic center. The −2* subsite can also bind to Xylp and Araf, explaining why the enzyme can utilize xylose and arabinose as specificity determinants. Alanine substitution of Glu68, Tyr92, or Asn139, which interact with arabinose and xylose side chains at the −2* subsite, abrogates catalytic activity. Distal to the active site, the xylan backbone makes limited apolar contacts with the enzyme, and the hydroxyls are solvent-exposed. This explains why Ctxyl5A is capable of hydrolyzing xylans that are extensively decorated and that are recalcitrant to classic endo-xylanase attack.

The plant cell wall is an important biological substrate. This complex composite structure is depolymerized by microorganisms that occupy important highly competitive ecological niches, whereas the process makes an important contribution to the carbon cycle (1). Lignocellulosic degradation is also of continued interest to environmentally sensitive industries such as the biofuels and biorefinery sectors, where the use of sustainable or renewable substrates is of increasing importance. Given that the plant cell wall is the most abundant source of renewable organic carbon on the planet, this macromolecular substrate has substantial industrial potential (2).

An example of the chemical complexity of the plant cell wall is provided by xylan, which is the major hemicellulosic component. This polysaccharide comprises a backbone of β-1,4-xylose residues in their pyranose configuration (Xylp) that are decorated at O2 with 4-O-methyl-d-glucuronic acid (GlcA) and at O2 and/or O3 with α-1-arabinofuranose (Araf) residues, whereas the polysaccharide can also be extensively acetylated (3). In addition, the Araf side chain decorations can also be esterified to ferulic acid that, in some species, provide a chemical link between hemicellulose and lignin (3). The precise structure of xylans varies between plant species, in particular in different tissues and during cellular differentiation (4). In specialized plant tissues, such as the outer layer of cereal grains, xylans are extremely complex, and side chains may comprise a range of other sugars including l- and d-galactose and β- and α-Xylp units. Indeed, in these cereal brans, xylans have very few backbone Xylp units that are undecorated, and the side chains can contain up to six sugars (5).

Reflecting the chemical and physical complexity of the plant cell wall, microorganisms that utilize these composite structures express a large number of polysaccharide-degrading enzymes, primarily glycoside hydrolases, but also polysaccharide lyases, carbohydrate esterases, and lytic polysaccharide monooxygenases. These carbohydrate active enzymes are grouped into sequence-based families in the
Mechanism of Arabinoxylanase

FIGURE 1. Molecular architecture of GH5_34 enzymes. Modules prefaced by GH, CBM, or CE are modules in the indicated glycoside hydrolase, carbohydrate binding module, or carbohydrate esterase families, respectively. Laminin_3_G domain belongs to the concanavalin A lectin superfamily, and FN3 denotes a fibronectin type 3 domain. Segments labeled D are dockerin domains.

CAZy database (6). With respect to xylan degradation, the backbone of simple xylans is hydrolyzed by endo-acting xylanases, the majority of which are located in glycoside hydrolase (GH) families GH10 and GH11, although they are also present in GH8 (1, 7). The extensive decoration of the xylan backbone generally restricts the capacity of these enzymes to attack the polysaccharide prior to removal of the side chains by a range of α-glucuronidases, α-arabinofuranosidases, and esterases (8). Two xylanases, however, utilize the side chains as essential specificity determinants and thus target decorated forms of the hemicellulose. The GH30 glucuronoxylanases require the XyIp bound at the −2 to contain a GlcA side chain (9) (the scissile bond targeted by glycoside hydrolases is between subsites −1 and +1, and subsites that extend toward the non-reducing and reducing ends of the substrate are assigned increasing negative and positive numbers, respectively (10)). The GH5 arabinoxylanase (CtXyl5A) derived from Clostridium thermocellum displays an absolute requirement for xylans that contain Ara side chains (11). In this enzyme, the key specificity determinant is the Ara appended to O3 of the XyIp bound in the active site (−1 subsite). The reaction products generated from arabinoxylans, however, suggest that Ara can be accommodated at subsites distal to the active site.

CtXyl5A is a multimodular enzyme containing, in addition to the GH5 catalytic module (CtGH5); three non-catalytic carbohydrate binding modules (CBMs) belonging to families 6 (CtCBM6), 13 (CtCBM13), and 62 (CtCBM62); fibronectin type 3 (Fn3) domain; and a C-terminal dockerin domain Fig. 1. Previous studies of Fn3 domains have indicated that they might function as ligand-binding modules, as a compact form of peptide linkers or spacers between other domains, as cellulose-disrupting modules, or as proteins that help large enzyme complexes remain soluble (12). The dockerin domain recruits the enzyme into the cellulosome, a multienzyme plant cell wall degrading complex presented on the surface of C. thermocellum (13, 14). CtCBM6 stabilizes CtGH5 (11), and CtCBM62 binds to D-galactopyranose and 1-arabinopyranose (15). The function of the CtCBM13 and Fn3 modules remains unclear. Similarly, the mechanism of substrate recognition and its impact on specificity are key unresolved issues. This report exploits the crystal structure of mature CtXyl5A lacking its C-terminal dockerin domain (CtXyl5A_Doc), and the enzyme in complex with ligands, to explore the mechanism of substrate specificity. The data show that the plasticity in substrate recognition enables the enzyme to hydrolyze highly complex xylans that are not accessible to classical GH10 and GH11 endo-xylanases.

Results

Substrate Specificity of CtXyl5A—Previous studies showed that CtXyl5A is an arabinoxylan-specific xylanase that generates xylooligosaccharides with an arabinose linked O3 to the reducing end xylose (11). The enzyme is active against both wheat and rye arabinoxylans (abbreviated as WAX and RAX, respectively). It was proposed that arabinose decorations make productive interactions with a pocket (−2*) that is abutted onto the active site or −1 subsite. Arabinose side chains of the other backbone xylose units in the oligosaccharides generated by CtXyl5A were essentially random. These data suggest that O3* and possibly O2, on the xylose residues at subsites distal to the active site and −2* pocket are solvent-exposed, implying that the enzyme can access highly decorated xylans. To test this hypothesis, the activity of CtXyl5A against xylans from cereal brans was assessed. CtXyl5a was incubated with a range of xylans for 16 h at 60 °C, and the limit products were visualized by TLC. These xylans are highly decorated not only with Ara and GlcA units but also with 1-Gal, 2-Gal, and 2-Xyl (5). Indeed, very few xylose units in the backbone of bran xylans lack side chains. The data presented in Table 1 showed that CtXyl5A was active against corn bran xylan (CX). In contrast typical endo-xylanases from GH10 and GH11 were unable to attack CX, reflecting the lack of undecorated xylose units in the backbone (the active site of these enzymes can only bind to non-substituted xylose residues (8, 16)). The limit products generated by CtXyl5A from CX consisted of an extensive range of oligosaccharides. These data support the view that in subsites out with the active site the O2 and O3 groups of the bound xylose units are solvent-exposed and will thus tolerate decoration.

To explore whether substrate bound only at −2* and −1 in the negative subsites was hydrolyzed by CtXyl5A, the limit products of CX digested by the arabinoxylanase were subjected to size exclusion chromatography using a Bio-Gel P-2, and the smallest oligosaccharides (largest elution volume) were chosen for further study. HPAEC analysis of the smallest oligosaccharides...
TABLE 1
Kinetics of GH5_34 arabinoxylanases

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Variant</th>
<th>kcat/Km [min⁻¹ mg⁻¹ ml⁻¹]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CxXyl5A</td>
<td>CGH5-CBM6-CRM13-Fx3-CBM62</td>
<td>WAX 800 ND 460</td>
</tr>
<tr>
<td>CxXyl5A</td>
<td>CGH5-CBM6-CRM13-Fn3</td>
<td>RAX 1,232 ND 659</td>
</tr>
<tr>
<td>CxXyl5A</td>
<td>CGH5-CBM6-CRM13</td>
<td>CX 1,307 ND 620</td>
</tr>
<tr>
<td>CxXyl5A</td>
<td>CGH5-CBM6</td>
<td>NA ND 102</td>
</tr>
<tr>
<td>CxXyl5A</td>
<td>CGH5-CBM6-E68A</td>
<td>NA ND NA</td>
</tr>
<tr>
<td>CxXyl5A</td>
<td>CGH5-CBM6-Y92A</td>
<td>NA ND NA</td>
</tr>
<tr>
<td>CxXyl5A</td>
<td>CGH5-CBM6-N135A</td>
<td>NA ND ND</td>
</tr>
<tr>
<td>CxXyl5A</td>
<td>CGH5-CBM6-N139A</td>
<td>NA ND ND</td>
</tr>
<tr>
<td>AcGH5</td>
<td>Wild type</td>
<td>628 1,641 289</td>
</tr>
<tr>
<td>GpGH5</td>
<td>Wild type</td>
<td>2,600 9,986 314</td>
</tr>
<tr>
<td>V5GH5</td>
<td>Wild type</td>
<td>102 203 23</td>
</tr>
</tbody>
</table>

The small reaction products were purified by size exclusion chromatography and analyzed by HPAEC (A) and positive mode ESI-MS (B), respectively. The samples were treated with a nonspecific arabinofuranosidase (CjAbs1A) and a GH3 xylanase (XynB) that targeted β-1,3-xylidosic bonds. X, xylose; A, arabinose. The m/z = 305 species denotes a pentose disaccharide as a sodium adduct [M + Na]⁺, whereas the m/z = 587 signal corresponds to an ESI-MS dimer of the pentose disaccharide also as a sodium adduct [2M + Na]⁺.

FIGURE 2. Identification of the disaccharide reaction products generated from CX. The smallest reaction products were purified by size exclusion chromatography and analyzed by HPAEC (A) and positive mode ESI-MS (B), respectively. The samples were treated with a nonspecific arabinofuranosidase (CjAbs1A) and a GH3 xylanase (XynB) that targeted β-1,3-xylidosic bonds. X, xylose; A, arabinose. The m/z = 305 species denotes a pentose disaccharide as a sodium adduct [M + Na]⁺, whereas the m/z = 587 signal corresponds to an ESI-MS dimer of the pentose disaccharide also as a sodium adduct [2M + Na]⁺.

Mechanism of Arabinoxylanase

ion adduct, whereas a dimer of the disaccharide with a sodium adduct (m/z = 587 [2M + Na]⁺) was also evident. The monosaccharide composition of pool 4 determined by TFA hydrolysis contained xylose and arabinose in a 3:1 ratio. This suggests that the two oligosaccharides consist of two disaccharides: one consisting of two xylose residues and the other consisting of an arabinose linked to a xylose. Treatment of pool 4 with the non-specific arabinofuranosidase, CjAbs1A (17), resulted in the loss of oligosaccharide 2 and the production of both xylose and arabinose, indicative of a disaccharide of xylose and arabinose. Incubation of pool 4 with a β-1,3-xyloligosaccharide (XynB) converted oligosaccharide 1 into xylose, demonstrating that this molecule is the disaccharide β-1,3-xyloligosaccharide. This view is supported by the inability of a β-1,4-specific xyloligosaccharide to hydrolyze oligosaccharide 1 or oligosaccharide 2 (data not shown). The crucial importance of occupancy of the −2* pocket for catalytic competence is illustrated by the inability of the enzyme to hydrolyze linear β-1,4-xylooligosaccharides. The generation of Araf-Xy1p and Xyl-β-1,3-Xyl as reaction products demonstrates that occupancy of the −2 subsite is not essential for catalytic activity, which is in contrast to all endo-acting xylanases where this subsite plays a critical role in enzyme activity (18, 19). Indeed, the data demonstrate that −2* plays a more important role in productive substrate binding than the −2 subsite. Unfortunately, the inability to generate highly purified (Xyl-β-1,4)_n-[β-1,3-Xyl/Ara]-Xyl oligosaccharides from arabinoxylans prevented the precise binding energies at the negative subsites to be determined.

Crystal Structure of the Catalytic Module of CtXyl5A in Complex with Ligands—To understand the structural basis for the biochemical properties of CtXyl5A, the crystal structure of the enzyme with ligands that occupy the substrate binding cleft and the critical −2* subsite were sought. The data presented in Fig. 3A show the structure of the CtXyl5A derivative CtGH5-CtCBM6 in complex with arabinose bound in the −2* pocket. Interestingly, the bound arabinose was in the pyranose conformation rather than in its furanose form found in arabinoxylans. O1 was facing toward the active site −1 subsite, indicative of the bound arabinose being in the right orientation to be linked to the xylan backbone via an α-1,3 linkage. As discussed on below, the axial O4 of the Ara p did not interact with the −2* subsite, suggesting that the pocket might be capable of binding a xylose molecule. Indeed, soaking apo crystals with xylose showed that the pentose sugar also bound in the −2* subsite in its pyranose conformation (Fig. 3B). These crystal structures support the biochemical data presented above showing that the enzyme generated β-1,3-xyloligosaccharide from CX, which would require the disaccharide to bind at the −1 and −2* subsites. A third product complex was generated by co-crystallizing the nucleophile inactive mutant CtGH5E279SCtCBM6 with a WAX-derived oligosaccharide (Fig. 3C). The data revealed a pentasaccharide bound to the enzyme, comprising β-1,4-xylotetraose with an Ara p linked α-1,3 to the reducing end xylose. The xylotetraose was positioned in subsites −1 to −4 and the Ara p in the −2* pocket. Analysis of the three structures showed that O1, O2, O4, and the endocyclic oxygen occupied identical positions in the Ara p, Ara f, and Xy1p ligands bound in the −2* subsite and thus made identical interactions with the pocket. O1 makes a polar
Mechanism of Arabinoxylanase

FIGURE 3. Representation of the residues involved in the ligands recognition at the –2\* subsite. The protein backbone is represented as a cartoon in gray. Interacting residues are represented as stick in blue, and the catalytic residues and the mutated glutamate (into a serine) are in magenta. A, CtgHS-CBM6 in complex with an arabinopyranose. B, CtgHS-CBM6 in complex with a xylopyranose. C, CtgHS-CG2CBM6 in complex with a pentasaccharide (β1,4-xylopyranose with an L-Ara linked α1,3 to the reducing end xylose). The xylan backbone is shown transparently for more clarity. Densities shown in blue are Refmac maximum-likelihood α-weight, weighted ZF2 – F2 at 1.5 σ. The figure and all other structural figures were made with PyMOL unless otherwise stated.

contact with N\textsubscript{82} of Asn\textsuperscript{139}, O\textsubscript{2} is within hydrogen bonding distance with O\textsubscript{61} of Asn\textsuperscript{139} and the backbone N of Asn\textsuperscript{135}, and O\textsubscript{3} interacts with the N of Gly\textsuperscript{136} and O\textsubscript{2} of Glu\textsuperscript{68}. Although O\textsubscript{4} of Ara\textsubscript{p} does not make a direct interaction with the enzyme, O\textsubscript{4} and O\textsubscript{5} of Xyl\textsubscript{p} and Ara\textsubscript{p}, respectively, form hydrogen bonds with O\textsubscript{e1} of Glu\textsuperscript{68}. Finally Tyr\textsuperscript{92} makes apolar parallel interactions with the pyranose or furanose rings of the three sugars.

The importance of the interactions between the ligands and the side chains of the residues in the –2\* pocket were evaluated by alanine substitution of these amino acids. The mutants E68A, Y92A, and N139A were all inactive (Table 1), demonstrating the importance of the interactions of these residues with the substrate and reinforcing the critical role the –2\* subsite plays in the activity of the enzyme. N135A retained wild type activity because the O\textsubscript{2} of the sugars interacts with the backbone N of Asn\textsuperscript{135} and not with the side chain. Because the hydroxyls of Xyl\textsubscript{p} or Ara\textsubscript{p} in the –2\* pocket are not solvent-exposed, the active site of the arabinoxylanase can only bind to xylose residues that contain a single xylose or arabinose O\textsubscript{3} decoration. This may explain why the k\textsubscript{cat}/K\textsubscript{m} for CtXyl5A against WAX was 2-fold higher than against CX (Table 1). WAX is likely to have a higher concentration of single Ara decorations compared with CX and thus contain more substrate available to the arabinoxylanase.

In the active site of CtXyl5A the α-D-Xylp, which is in its relaxed 4C\textsubscript{1} conformation, makes the following interactions with the enzyme (Fig. 4, A–C): O\textsubscript{2} hydrogen bonds with the N\textsubscript{81} of His\textsuperscript{253} and Oe2 of Glu\textsuperscript{171} (catalytic acid-base) and makes a possible weak polar contact with the OH of Tyr\textsuperscript{255} and Oy of Ser\textsuperscript{279} (mutation of the catalytic nucleophile); O\textsubscript{3} hydrogen bonds with N\textsubscript{82} of Asn\textsuperscript{170} and OH of Tyr\textsuperscript{92}. O\textsubscript{3} (O\textsubscript{1} of the Ara at the –2\* subsite) makes a polar contact with N\textsubscript{82} of Asn\textsuperscript{139}; the endocyclic oxygen hydrogens bonds with the OH of Tyr\textsuperscript{255}. The Xylp in the active site makes strong parallel apolar interactions with Phe\textsuperscript{310}. Substrate recognition in the active site is conserved between CtXyl5A and the closest GH5 structural homolog, the endoglucanase BaCel5A (PDB code 1qi2) as noted previously (11).

The capacity of CtXyl5A to act on the highly decorated xylan CX indicates that O\textsubscript{3} and possibly O\textsubscript{2} of the backbone Xylp units are solvent-exposed. This is consistent with the interaction of the xylotetraose backbone with the enzyme distal to the active site. A surface representation of the enzyme (Fig. 4B) shows that O\textsubscript{2} and O\textsubscript{3} of xylose units at subsites –2 to –4 are solvent-exposed and are thus available for decoration. Indeed, these pyranose sugars make very weak apolar interactions with the arabinoxylanase. At –2, Xylp makes planar apolar interactions with the Ara\textsubscript{p} bound to the –2\* subsite (Fig. 4C). Xylp at subsites –2 and –3, respectively, make weak hydrophobic contact with Val\textsuperscript{118}, the –3 Xylp makes planar apolar interactions with Ala\textsuperscript{137}, whereas the xylose at –4 forms parallel apolar contacts with Trp\textsuperscript{69}. Comparison of the distal negative subsites of CtXyl5A with BaCel5A and a typical GH10 xylanase (CmXyn10B, PDB code 1uqy) highlights the paucity of interactions between the arabinoxylanase and its substrate out with the active site (Fig. 4). Thus, the cellulase contains three negative subsites and the sugars bound in the –2 and –3 subsites make a total of 9 polar interactions with the enzyme (Fig. 4, D and E). The GH10 xylanase also contains a –2 subsite that, similar to the cellulase, makes numerous interactions with the substrate (Fig. 4, F and G).

The Influence of the Modular Architecture of CtXyl5A on Catalytic Activity—CtXyl5A, in addition to its catalytic module, contains three CBMs (CtCBM6, CtCBM13, and CtCBM62) and a fibronectin domain (CfFn3). A previous study showed that although the CBM6 bound in an exo-mode to xylo- and celluooligosaccharides, the primary role of this module was to
stabilize the structure of the GH5 catalytic module (11). To explore the contribution of the other non-catalytic modules to Ctxyl5A function, the activity of a series of truncated derivatives of the arabinoxylanase were assessed. The data in Table 1 show that removal of CtxCBM62 caused a modest increase in activity against both WAX and CX, whereas deletion of the Fn3 domain had no further impact on catalytic performance. Truncation of CtxCBM13, however, caused a 4–5-fold reduction in activity against both substrates. Members of CBM13 have been shown to bind to xylans, mannose, and galactose residues in complex glycans (20–23), hinting that the function of CtxCBM13 is to increase the proximity of substrate to the catalytic module of Ctxyl5A. Binding studies, however, showed that CtxCBM13 displayed no affinity for a range of relevant glycans including WAX, CX, xylose, mannose, galactose, and birchwood xylan (BX) (data not shown). It would appear, therefore, that CtxCBM13 makes a structural contribution to the function of Ctxyl5A.

Crystal Structure of Ctxyl5A D—To explore further the role of the non-catalytic modules in Ctxyl5A the crystal structure of Ctxyl5A extending from CtxGH5 to CtxCBM62 was sought. To obtain a construct that could potentially be crystallized, the protein was generated without the C-terminal dockerin domain because it is known to be unstable and prone to cleavage. Using this construct (Ctxyl5A D), the crystal structure of the arabinoxylanase was determined by molecular replacement to a resolution of 2.64 Å with Rwork and Rfree at 23.7% and 27.8%, respectively. The structure comprises a continuous polypeptide extending from Ala36 to Trp742 displaying four modules GH5-CBM6-CBM13-Fn3. Although there was some electron density for CtxCBM62, it was not sufficient to confidently build the module (Fig. 5). Further investigation of the crystal packing revealed a large solvent channel adjacent to the area the CBM62 occupies. We postulate that the reason for the poor electron density is due to the CtxCBM62 being mobile compared with the rest of the protein. The structures of CtxGH5 and CtxCBM6 have been described previously (11, 15).

CtxCBM13 extends from Gly567 to Pro648. Typical of CBM13 proteins CtxCBM13 displays a β-trefoil fold comprising the canonical pseudo 3-fold symmetry with a 3-fold repeating unit of 40–50 amino acid residues characteristic of the Ricin superfamily. Each repeat contains two pairs of antiparallel β-strands. A Dali search revealed structural homologs from the CBM13 family with a root mean square deviation less than 2.0 Å and sequence identities of less than 20% that include the functionally relevant homologs C. thermocellum exo-β-1,3-galactanase (22) (PDB code 3vs2), Streptomyces avermitilis β-1,arabinopyranosidase (21) (PDB code 3a21), Streptomyces lividans xylanase 10A (23) (PDB code 1mc9), and Streptomyces olivaceoviridis E-86 xylanase 10A (20) (PDB code 1ve6).

The Fn3 module displays a typical β-sandwich fold with the two sheets comprising, primarily, three antiparallel strands in the order β1–β2–β5 in β-sheet 1 and β4–β3–β6 in β-sheet 2. Although β-sheet 2 presents a cleft-like topology, typical of endo-binding CBMs, the surface lacks aromatic residues that play a key role in ligand recognition, and in the context of the full-length enzyme, the cleft abuts into CtxCBM13 and thus
would not be able to accommodate an extended polysaccharide chain (see below).

In the structure of \( \text{CtXyl5A-D} \), the four modules form a three-leaf clover-like structure (Fig. 5). Between the interfaces of \( \text{CtGH5-CBM6-CBM13} \) there are a number of interactions that maintain the modules in a fixed position relative to each other. The interaction of \( \text{CtGH5 and CtCBM6} \), which buries a substantial apolar solvent-exposed surface of the two modules, has been described previously (11). The polar interactions between these two modules likely explaining why they do not fold independently compared with other glycoside hydrolases that contain CBMs (24, 25). \( \text{CtCBM13} \) acts as the central domain, which interacts with \( \text{CtGH5, CtCBM6, and CtFn3} \) via 2, 5, and 4 hydrogen bonds, respectively, burying a surface area of \( \sim 450, 350, \) and \( \sim 500 \text{ Å}^2 \), respectively, to form a compact heterotetramer. With respect to the \( \text{CtCBM6-CBM13} \) interface, the linker (SPISTGTIP) between the two modules, extending from Ser514 to Pro522, adopts a fixed conformation. Such sequences are normally extremely flexible (26); however, the two Ile residues make extensive apolar contacts within the linker and with the two CBMs, leading to conformational stabilization. The interactions between \( \text{CtGH5 and the two CBMs} \), which are mediated by the tip of the loop between \( \beta-7 \) and \( \alpha-7 \) (loop 7) of \( \text{CtGH5} \), not only stabilize the trimodular clover-like structure but also make a contribution to catalytic function. Central to the interactions between the three modules is Trp285, which is intercalated between the two CBMs. The Ne of this aromatic residue makes hydrogen bonds with the backbone carbonyl of Val415 and Gly516 in \( \text{CtCBM13} \), and the indole ring makes several apolar contacts with \( \text{CtCBM6 (Pro440, Phe489, Gly491, and Ala592)} \) (Fig. 5). Indeed, loop 7 is completely disordered in the truncated derivative of \( \text{CtXyl5A} \) comprising \( \text{CtGH5 and CtCBM6} \), demonstrating that the interactions with \( \text{CtCBM13} \) stabilize the conformation of this loop. Although the tip of loop 7 does not directly contribute to the topology of the active site, it is only \( \sim 12 \text{ Å} \) from the catalytic nucleophile Glu279. Thus, any perturbation of the loop (through the removal of \( \text{CtCBM13} \)) is likely to influence the electrostatic and apolar environment of the catalytic apparatus, which could explain the reduction in activity associated with the deletion of \( \text{CtCBM13} \).

Similar to the interactions between \( \text{CtCBM6 and CtCBM13} \), there are extensive hydrophobic interactions between \( \text{CtCBM6 and CtFn3} \), resulting in very little flexibility between these modules. As stated above, the absence of \( \text{CtCBM62} \) in the structure suggests that the module can adopt multiple positions with respect to the rest of the protein. The \( \text{CtCBM62} \), by binding to its ligands (\( \text{d-Galp and L-Arap} \)) in plant cell walls (15), may be able to recruit the enzyme onto its target substrate. Xylans are not generally thought to contain such sugars. \( \text{d-Galp} \), however, has been detected in xylans in the outer layer of cereal grains and in eucalyptus trees (5), which are substrates used by \( \text{CtXyl5A} \). Thus, \( \text{CtCBM62} \) may direct the enzyme to particularly complex xylans containing \( \text{d-Galp} \) at the non-reducing termini of the side chains, consistent with the open substrate binding cleft of the arabinoxylanase that is optimized to bind highly decorated forms of the hemicellulose. In general CBMs have little influence on enzyme activity against soluble substrates but have a significant impact on glycans within plant cell walls (27, 28). Thus, the role of CBM62 will likely only be evident against insoluble composite substrates.

**FIGURE 5.** Surface representation of the tetra-modular arabinoxylanase and zoom view on the \( \text{CtGH5} \) loop. The blue module is the \( \text{CtGH5} \) catalytic domain, the green module corresponds to the \( \text{CtCBM6} \), the yellow module is the \( \text{CtCBM13} \), and the salmon module is the fibronectin domain. Surfaces are semitransparent with the protein backbone represented as a cartoon. The \( \text{CtGH5} \) loop is stabilized between the \( \text{CtCBM6 and the CtCBM13 modules} \). The black dashes represent the hydrogen bonds. The protein backbone is represented as cartoon, and interacting residues are shown as sticks.
Exploring GH5 Subfamily 34—CtXyl5A is a member of a seven-protein subfamily of GH5, GH5_34 (29). Four of these proteins are distinct, whereas the other three members are essentially identical (derived from different strains of C. thermocellum). To investigate further the substrate specificity within this subfamily, recombinant forms of three members of GH5_34 that were distinct from CtXyl5A were generated. AcGH5 has a similar molecular architecture to CtXyl5A with the exception of an additional carbohydrate esterase family 6 module at the C terminus (Fig. 1). The GH5_34 from Verrucomicrobiae bacterium, VbGH5, contains the GH5-CBM6-CBM13 core structure, but the C-terminal Fn3-CBM62-dockerin modules, present in CtXyl5A, are replaced with a Laminin_3_G domain, which, by analogy to homologous domains in other proteins that have affinity for carbohydrates (30), may display a glycan binding function. The Verrucomicrobiae enzyme also has an N-terminal GH43 subfamily 10 (GH43_10) catalytic module. The fungal GH5_34, GpGH5, unlike the two bacterial homologs, comprises a single GH5 catalytic module lacking all of the other accessory modules (Fig. 1). GpGH5 is particularly interesting as Gonapodya prolifera is the only fungus of the several hundred fungal genomes that encodes a GH5_34 enzyme. In fact there are four potential GH5_34 sequences in the G. prolifera genome, all of which show high sequence homology to Clostridium GH5_34 sequences. G. prolifera and Clostridium occupy similar environments, suggesting that the GpGH5_34 gene was acquired from a Clostridium species, which was followed by duplication of the gene in the fungal genome. The sequence identity of the GH5_34 catalytic modules with CtXyl5A ranged from 55 to 80% (supplemental Fig. S1). All the GH5_34 enzymes were active on the arabinoxylans RAX, WAX, and CX but displayed no activity on BX (Table 1 and Fig. 6) and are thus defined as arabinoxylanases. The limit products generated by CtXyl5A, AcGH5, and GpGH5 comprised a range of oligosaccharides with some high molecular weight material. The oligosaccharides with low degrees of polymerization were absent in the VbGH5 reaction products. However, the enzyme generated a large amount of arabinose, which was not produced by the other arabinoxylanases. Given that GH43_10 is predominantly an arabinofuranosidase subfamily of GH43 (31), the arabinose generated by VbGH5 is likely mediated by the N-terminal catalytic module (see below). Kinetic analysis showed that AcGH5 displayed similar activity to CtXyl5A against both WAX and RAX and was 2-fold less active against CX. When initially measuring the activity of wild type VbGH5 against the different substrates, no clear data could be obtained, regardless of the concentration of enzyme used the reaction appeared to cease.

FIGURE 6. Products profile generated of GH5_34 enzymes. The enzymes at 1 μM were incubated with the four different xylans at 1% in 50 mM sodium phosphate buffer for 16 h at 37 °C (GpGH5, VbGH5, and AcGH5) or 60 °C. The limit products were separated by TLC. The xylooligosaccharide standards (X) are indicated by their degrees of polymerization.

Mechanism of Arabinoxylanase
after a few minutes. We hypothesized that the N-terminal GH43_10 rapidly removed single arabinose decorations from the arabinoxylans depleting the substrate available to the arabinoxylanase, explaining why this activity was short lived. To test this hypothesis, the conserved catalytic base (Asp$^{45}$) of the GH43_10 module of \textit{Vb}GH5 was substituted with alanine, which is predicted to inactivate this catalytic module. The D45A mutant did not produce arabinose consistent with the arabinoferanosidase activity displayed by the GH43_10 module in the wild type enzyme (Fig. 6). The kinetics of the GH5_34 arabinoxylanase catalytic module was now measurable, and activities were determined to be between \( \sim 1 - 10 \) fold lower than that of \textit{Ct}Xyl5A. Interestingly, the fungal arabinoxylanase displays the highest activities against WAX and RAX, \( \sim 4 - 6 \) fold higher, respectively, than \textit{Ct}Xyl5A; however, there is very little difference in the activity between the eukaryotic and prokaryotic enzymes against CX. Attempts to express individual modules of a variety of truncations of \textit{Ac}GH5 and \textit{Vb}GH5 were unsuccessful. This may indicate that the individual modules can only fold correctly when incorporated into the full-length enzyme, demonstrating the importance of intermodule interactions to maintain the structural integrity of these enzymes.

**Discussion**

A characteristic feature of enzymes that attack the plant cell wall is their complex molecular architecture (1). The CBMs in these enzymes generally play a role in substrate targeting (25, 28) and are appended to the catalytic modules through flexible linker sequences (26). \textit{Ct}Xyl5A provides a rare visualization of the structure of multiple modules within a single enzyme. The central feature of these data is the structural role played by two of the CBMs, \textit{Ct}CBM6 and \textit{Ct}CBM13, in maintaining the active conformation of the catalytic module, \textit{Ct}GH5. The crystallographic data described here are supported by biochemical data showing either that these two modules do not bind to glycans (\textit{Ct}CBM13) or that the recognition of the non-reducing end of xylan or cellulose chains (\textit{Ct}CBM6) is unlikely to be biologically significant. It should be emphasized, however, that glycan binding and substrate targeting may only be evident in the full-length enzyme acting on highly complex structures such as the plant cell wall, as observed recently by a \textit{Ct}CBM module in the eukaryotic and prokaryotic enzymes against CX. Attempts to express individual modules of a variety of truncations of \textit{Ac}GH5 and \textit{Vb}GH5 were unsuccessful. This may indicate that the individual modules can only fold correctly when incorporated into the full-length enzyme, demonstrating the importance of intermodule interactions to maintain the structural integrity of these enzymes.

**Mechanism of Arabinoxylanase**

GH5 may indicate that this arabinoxylanase is designed to target simpler arabinoxylans present in the endosperm of cereals. Although the characterization of all members of GH5_34 suggests that this subfamily is monospecific, differences in specificity are observed in other subfamilies of GHs including GH43 (31) and GH5 (29). Thus, as new members of GH5_34 are identified from genomic sequence data and subsequently characterized, the specificity of this family may require reinterpretation.

An intriguing feature of \textit{Vb}GH5 is that the limited products generated by this enzymes are much larger than those produced by the other arabinoxylanases. This suggests that although arabinose decorations contribute to enzyme specificity (\textit{Vb}GH5 is not active on xylans lacking arabinose side chains), the enzyme requires other specificity determinants that occur less frequently in arabinoxylans. This has some resonance with a recently described GH98 xylanase that also exploits specificity determinants that occur infrequently and are only evident in highly complex xylans (e.g. CX) (5).

To conclude, this study provides the molecular basis for the specificity displayed by arabinoxylanases. Substrate specificity is dominated by the pocket that binds single arabinose or xylose side chains. The open xylan binding cleft explains why the enzyme is able to attack highly decorated forms of the hemicellulose. It is also evident that appending additional catalytic modules and CBMs onto the core components of these enzymes generates bespoke arabinoxylanases with activities optimized for specific functions. The specificities of the arabinoxylanases described here are distinct from the classical endoxylanases and thus have the potential to contribute to the toolbox of biocatalysts required by industries that exploit the plant cell wall as a sustainable substrate.

**Experimental Procedures**

**Cloning, Expression, and Purification of Components of \textit{Ct}Xyl5A**—All recombinant forms of \textit{Ct}Xyl5A used in this study were expressed in the cytoplasm of \textit{Escherichia coli} because they lacked a signal peptide. DNA encoding \textit{Ct}GH5-\textit{Ct}CBM6 and \textit{Ct}Xyl5A-D (\textit{Ct}Xyl5A lacking the C-terminal dockerin domain (\textit{Ct}GH5-\textit{Ct}CBM6-\textit{Ct}CBM13-Fn3-\textit{Ct}CBM62)) were described previously (11). DNA encoding \textit{Ct}GH5-\textit{Ct}CBM6-\textit{Ct}CBM13-Fn3 and \textit{Ct}GH5-\textit{Ct}CBM6-\textit{Ct}CBM13 mature \textit{Aacetivibrio cellulosyticus} GH5 (\textit{Ac}GH5) were amplified by PCR using plasmid encoding the full-length \textit{C. thermocellum} arabinoxylanase or \textit{A. cellulosyticus} genomic DNA as the respective templates. DNA encoding the \textit{G. prolifer}a GH5 (\textit{Pp}GH5) and \textit{V. bacteria} GH5 (\textit{Vb}GH5) were initially generated by GeneArt® gene synthesis (Thermo Fisher Scientific). DNA encoding \textit{Vb}GH5 lacking the C-terminal cell surface anchoring residues was also amplified by PCR using the synthesized nucleic acid as the template. All the primers used in the PCRs required restriction sites and plasmids used are listed in supplemental Table S1. All constructs were cloned such that the encoded proteins contain a C-terminal His$^{6}$ tag. Site-directed mutagenesis was carried out using the PCR-based QuikChange method (Stratagene) deploying the primers listed in supplemental Table S1.

To express the recombinant proteins, \textit{E. coli} strain BL21(DE3), harboring appropriate recombinant plasmids, was
The pH and temperature optima were previously determined (33). When preparing the selenomethionine derivative of CtXyl5A-D for crystallography, the proteins were expressed in *E. coli* B834 (DE3), a methionine auxotroph, cultured in medium comprising 1 liter of SelenoMet Medium Base™, 50 ml of SelenoMet TM nutrient mix (Molecular Dimensions), and 4 ml of a 10 mg/ml solution of L-selenomethionine. Recombinant gene expression and protein purification were as described above except that all purification buffers were supplemented with 10 mM β-mercaptoethanol.

**Enzyme Assays—**CtXyl5A₁₇ and its derivatives were assayed for enzyme activity using the method of Miller (34) to detect the release of reducing sugar. The standard assay was carried out in 50 mM sodium phosphate buffer, pH 7.0, containing 0.1 mg/ml BSA and at substrate concentrations ranging from 1 to 6 mg/ml. The pH and temperature optima were previously determined to be 7 and 60 °C, respectively, for the CtXyl5A₁₇ and its derivatives. The optimum temperature for the other enzymes was found to be 37 °C, and pH optima of 5, 7, and 4 were determined for AcGH5, GpGH5 and VbGH5, respectively. All enzymes were assayed for activity at their individual temperature and pH optimum. A FLUOstar Omega microplate reader (BMG Labtech) was used to measure activity in 96-well plates. Over-night assays to assess end point products were carried out with 6 mg/ml substrate and 1 μM enzyme concentrations. The identification of potential reaction products was also assessed by HPAEC or TLC using methodology described previously (34).

**Oligosaccharide Analysis—**Approximately 5 g of CX or WAX were digested to completion (no further increase in reducing sugar and change in the HPAEC product profile) with 3 μM of CtXyl5A₁₇ at 60 °C for 48 h. The oligosaccharide products were purified by size exclusion chromatography using a Bio-Gel P2 column as described previously (35). The structures of the oligosaccharides were analyzed by positive ion-mode infusion/offline electrospray ionization (ESI)-MS following either dilution with 30% acetonitrile or via desalting as described previously (36).

**Crystallography—**Purified SeMet CtXyl5A₁₇ was concentrated and stored in 5 mM DTT, 2 mM CaCl₂. Crystals of selenomethionine-containing protein were obtained by hanging drop vapor diffusion in 40% (v/v) 2-methyl-2,4-pentandiol. The data were collected on Beamlines ID14-1 and ID14-4 at the European Synchrotron Radiation Facility (Grenoble, France) to a resolution of 2.64 Å. The data were processed using the programs iMOSFLM (37) and SCALA (38) from the CCP4 suite (Collaborative Computational Project, Number 4, 1994). The crystal belongs to the orthorhombic space group (P2₁,2,2) (39). The structure was solved by molecular replacement using independently solved structures of some of the modules of the CtXyl5A: C2GH5-CBM6 (PDB code 2y8k) (11), Fn3 (PDB code 3mpc) (12), and CBM13 domain was built de novo. BUCCANEER (42) and PHENIX (43) were initially used for auto building. The structure was completed by iterative cycles of manual rebuilding in COOT (44) in tandem with refinement with RefMac5 (45). The final values for Rwork and Rfree are presented in Table 2.
**Mechanism of Arabinoxylanase**

$R_{free}$ were 23.73 and 27.80%) using TLS and restraining refinement to amino acid residues 36–373 representing the CtGH5 module, 374–516 for the CtCBM6, 517–652 for CtCBM13, and 653–742 for CtFn3. Stereochemistry was assessed with COOT (44) and PDBSUM (46) (with 677 residues (96%) in preferred, 22 in allowed regions (3%), and 6 outliers (1%) in the Ramachandran plot).

To obtain structures of CtGH5-CBM6 in complex with ligand the protein was crystallized using the sitting drop vapor phase diffusion method with an equal volume (100 nl) of protein and reservoir solution (unless otherwise stated), using the robotic nanodrop dispensing systems (mosquito<sup>®</sup> LCP; TTPLabTech). Crystals of the protein (10 mg/ml) co-crystallized with arabinose (300 mM) were obtained in 1 M ammonium sulfate, 0.1 M Bis-Tris, pH 5.5, and 1% PEG 3350. Crystals with xyllose (300 mM) grew in 100 mM sodium/potassium phosphate, 100 mM MES, pH 6.5, and 2 M sodium chloride. To obtain crystals of the arabinoxylanase in complex with an oligosaccharide, the nucleophile mutant E279S was used and mixed with a range of arabinoyxooligosaccharides that was generated by digestion of WAX with CtGH5-CBM6 (see above) and thereafter by 100 nM of the Cellvibrio japonicus GH43 exo-1,4-β-xyllosidase (47). Only the inclusion of the largest purified oligosaccharide generated crystals of the arabinoxylanase. Crystals of CtGH5<sub>E279S</sub>-CBM6 were obtained by mixing an equal volume (100 nl) of the protein (11 mg/ml)/oligosaccharide (10 mM) solution and mother liquor solution consisting of 100 mM Tris-Bicine, pH 8.5, 12.5% (v/v) polyethylene glycol with an average molecular mass of 1,000 Da, 12.5% (v/v) polyethylene glycol with an average molecular mass of 3,350 Da and 12.5% (R.S)-2-methyl-2,4-pentanediol (racemic). Crystallographic data were collected on Beamlines IO2, IO4-1, and I24 at the DIAMOND Light Source (Harwell, UK). The data were processed using XDS (48) The crystal structures were solved by molecular replacement using MolRep (49) with CtGH5-CtCBM6 (PDB code 5AK1) as the search model. The refinement was done in Refmac5 (27), and COOT (26) was used for model (re)building. The final model were validated using Molprobity (32). The data collection and refinement statistics are listed in Table 2.

**Author Contributions**—A. L. obtained crystals of the GH5-CBM6 complex. L. I. C. analyzed the biochemistry of GH5 subfamilies. J. L. A. B. obtained crystals of Ctxy15A-D. A. J. analyzed the biochemistry of GH5-CBM6 and obtained crystals of GH5-CBM6. A. R. analyzed the biochemistry of GH5-CBM6 products. J. G. performed mass spectrometry. M. P. Y. provided the substrate. B. H. performed analysis of GH5 sequences. C. M. G. A. F. designed the experiments. H. J. G. designed the experiments, analyzed data, and contributed to writing the paper. S. N. solved the structure of Ctxy15A-D and contributed to writing the paper. A. B. used crystallography to solve GH5-CBM6 structures. F. C. analyzed the biochemistry of GH5-CBM6 mutants, obtained crystals of GH5-CBM6, and contributed to writing the paper.

**Acknowledgments**—We thank Diamond Light Source for access to Beamlines IO2, IO4-1, and I24 (mx1960, mx7854, and mx9948) that contributed to the results presented here.

---

**References**


Mechanism of Arabinoxylanase


