Function, distribution, and annotation of characterized cellulases, xylanases, and chitinases from CAZy

Stanley T. C. Nguyen[1], **Hannah L. Freund**[1], **Joshua Kasanjian**[1], and **Renaud Berlemont**[1,*]

[1]Department of Biological Sciences, California State University - Long Beach, Long Beach, USA

Abstract

The enzymatic deconstruction of structural polysaccharides, which relies on the production of specific glycoside hydrolases (GHs), is an essential process across environments. Over the past few decades, researchers studying the diversity and evolution of these enzymes have isolated and biochemically characterized thousands of these proteins. The carbohydrate active enzymes database (CAZy) lists these proteins and provides some metadata. Here, the sequences and metadata of characterized sequences derived from GH families associated with the deconstruction of cellulose, xylan, and chitin were collected and discussed. First, although few polyspecific enzymes are identified, characterized GH families are mostly monospecfic. Next, the taxonomic distribution of characterized GH mirrors the distribution of identified sequences in sequenced genomes. This provides a rationale for connecting the identification of GH sequences to specific reactions or lineages. Finally, we tested the annotation of the characterized GHs using HMM scan and the protein families database (Pfam). The vast majority of GHs targeting cellulose, xylan and chitin can be identified using this publicly accessible approach.

Keywords

Cellulase; Chitinase; Xylanase; CAZy; Glycoside Hydrolase

Introduction

Cellulose and xylan, from the plant cell wall, represent the largest pool of organic carbon in land ecosystems whereas chitin, although abundant in terrestrial ecosystems, dominates in marine ecosystems. The deconstruction of these polysaccharides, outside the cell, by specific enzymes releases short oligo/di-saccharides which then can be translocated into the cell and further processed to release energy (e.g., glycolysis, fermentation). Glycoside hydrolases (GHs) are essential enzymes required for the breakdown of these polysaccharides. These proteins together with other carbohydrate active enzymes (e.g., carbohydrate esterases (CE), polysaccharide lyases (PL), and lytic polysaccharide monooxigenases (LPMO))(Lombard et al. 2014) support key processes across ecosystems. However, although abundant across

environments (Berlemont and Martiny 2016), GHs are not randomly distributed (Berlemont and Martiny 2013). Most identified GHs are from microbes [e.g., (Medie et al. 2012; Berlemont and Martiny 2013; Lombard et al. 2014; Berlemont and Martiny 2015; Berlemont 2017)] and invertebrates [see (Guo et al. 2008; Rahman et al. 2014)].

Large-scale comparisons of sequenced bacterial genomes reveal that not all the lineages have genes for GHs potentially involved in polysaccharide deconstruction, whereas most microbes have genes for the processing of polysaccharide deconstruction products (i.e., oligosaccharides) [e.g.,(Berlemont and Martiny 2013)]. Microorganisms targeting short deconstruction products, sometime referred to as the opportunists (Berlemont et al. 2014), contribute indirectly to the process of polysaccharide deconstruction by keeping local oligosaccharide concentrations low and, thus, prevent the product from inhibiting the enzyme [see (Rignall et al. 2002; Gefen et al. 2012; Xu et al. 2013; Bailey et al. 2013)]. In contrast, organisms targeting larger substrates have evolved several strategies to degrade polysaccharides (Wilson 2011; Talamantes et al. 2016). Briefly, microbes produce many single domain enzymes consisting of one unique GH-catalytic domain sometime associated with accessory non-catalytic domain(s) such as carbohydrate binding modules (CBM). These CBMs direct their associated catalytic domains to specific substrates, increase the local concentration of enzymes, reduce the enzyme diffusion, and help relax the crystalline structure of the substrate. In so doing, they improve the overall catalytic efficiency of the hydrolytic systems (Din et al. 1991; Hervé et al. 2010). Interestingly, still other degraders produce proteins with multiple catalytic domains (Gibbs et al. 1992; Brunecky et al. 2013; Talamantes et al. 2016). Finally, some bacteria and fungi evolved non-covalent modular multi-protein complexes consisting of several GHs called cellulosomes (Artzi et al. 2016; Haitjema et al. 2017).

Large-scale comparisons of sequenced genomes highlight the phylogenetic conservatism of enzymes involved in polysaccharide deconstruction, and variability of the domain organization in GHs from closely related strains [e.g.,(Talamantes et al. 2016; Berlemont 2017)]. In bacteria, most members of the same genus share similar abilities for cellulose, xylan, and chitin deconstruction. Identified degrader lineages frequently display redundant enzymes from the same GH family and assumed to target the same substrate (Berlemont and Martiny 2015; Berlemont 2017). However, the extensive biochemical characterization of the "CAZome" of isolated microbes highlighted subtle variations in the substrate specificity, enzymology, and regulation of these apparently redundant enzymes [e.g., (Ravachol et al. 2014)].

Detailed information about characterized GHs are centralized on the carbohydrate-active enzymes database (CAZy, http://www.cazy.org)(Lombard et al. 2014). The CAZy database is an essential resource for scientists studying the processing of carbohydrates. Amongst others, CAZy provides a framework for the sequence-based classification of GHs (and other carbohydrate-active enzymes; CAZymes), a listing of the characterized enzymes, and some taxonomic information. CAZy also lists and classifies the many identified yet uncharacterized GHs from sequenced genomes. Basically, the classification of CAZymes reflects their structural and/or sequence similarity (Henrissat and Davies 1997). According to CAZy most GHs families are polyspecific and target various substrates (Henrissat and

Davies 1997; Aspeborg et al. 2012). For example, at least 20 distinct enzymatic activities are listed for members of the GH family 5, whereas monospecific families display narrow substrate specificity. Specificity can also be investigated at the protein level. The vast majority of characterized GHs are monospecific whereas few proteins or catalytic domains, able to accommodate multiple substrates, are polyspecific [e.g., (Berger et al. 1989)]. Sometime, multidomain proteins [e.g., (Brunecky et al. 2013)] artificially inflate the polyspecificity of identified GH families on CAZy. For example, the unique xylanase in GH family 62 is a multidomain xylanase/arabinofuranosidase from *Streptomyces chattanoogensis* UAH23, whose xylanolytic activity likely results from a GH10 domain (acc. num. AAD32559.2)(Hernández et al. 2001).

Although listing GHs, CAZy provides no analysis tools to annotate or analyze new sequences, the removal of the links to the Pfam and InterPro databases (in 2008) resulted in the development of alternative annotation systems for GHs and related enzymes (Park et al. 2010; Yin et al. 2012; Berlemont and Martiny 2013; Talamantes et al. 2016). In this context, three important questions remain. First, can the polyspecificity of GH families associated with the deconstruction of abundant polysaccharides be estimated? Knowing the targeted substrates and the activities in each family will provide an easy way to connect specific sequences to environmental process (e.g., carbon cycling)(Treseder and Lennon 2015). Next, what are the taxonomic distribution and the substrate specificity of characterized GHs? Identifying clustered distribution of enzymes with particular substrate specificity could highlight their recent evolution and provide a comprehensive framework to isolate new enzymes with specific activities (Aspeborg et al. 2012). Finally, as the classification of CAZymes was first intended to be "more friendly to the needs of bioinformatics" (Henrissat and Davies 1997), we asked the question: can the annotation of GHs for cellulose, xylan, and chitin be achieved using publicly accessible tools? The rapid and reliable annotation of GHs in the growing number of sequenced genomes and microbiomes is essential because GHs enzymes support key functions in cells and ecosystems (Knight et al. 2012). In order to answer these questions, we reviewed the functional and taxonomic distribution of characterized enzymes listed on the CAZy database, as of summer 2017. Sequences from characterized proteins from the GH families of interest were analyzed using Pfam-scan against the entire PFam A database (Eddy 2011; Punta et al. 2012; Finn et al. 2014).

Cellulases

The enzymatic deconstruction of cellulose requires the interaction between some endo-acting glycoside hydrolases (i.e., endocellulase EC.3.2.1.4) and some exo-acting glycoside hydrolases (i.e., exocellulase, EC.3.2.1.91/176)(Wilson 2011). These activities release cellooligosaccharides and cellobiose that are further degraded to glucose by β-glucosidases (EC3.2.1.21, not discussed here). Beside these enzymes, 4-β-D-glucan glucohydrolases (EC. 3.2.1.74) are exocellulases active on cellulose and releasing glucose directly. Both endo- and exocellulases are listed in the GH families 5, 6, 7, 9, and 48. In addition, according to CAZy, endocellulases have also been identified in GH families 8, 10, 12, 26, 44, 45, 51, 74, and 124 (table 1, S1).

Most characterized cellulolytic enzymes are from the polyspecific GH family 5 (n=547 listed proteins)(Aspeborg et al. 2012). Respectively, 61% and 1% of the characterized GH5s act as endocellulases (EC.3.2.1.4) or exocellulases (EC.3.2.1.74/91). Non-cellulolytic GH5s target other plant cell wall polysaccharides such as 109 endo-β-1,4-mannosidases (EC. 3.2.1.78), few licheninases (EC.3.2.1.73, 2%), and some xylanases (EC.3.2.1.8, 1.6%), and some xyloglucanases (EC.3.2.1.151, 1%). Beside theses well characterized enzymes, 46 of the listed GH5s have unspecified substrates (i.e., EC.3.2.1.-). Interestingly, 19 cellulolytic GH5s are polyspecific enzymes. Some of these proteins are multi-activity GHs (Talamantes et al. 2016), such as a GH5-CBM3-CBM3-GH44 protein from *Caldicellulosiruptor saccharolyticus* (acc. num. AAA71887.1) (Gibbs et al. 1992) and the GH5-GH26 protein from an uncultured bacterium (acc. num. ABB46200.1) (Palackal et al. 2007). Conversely, some enzymes with only one catalytic domain, such as the GH5-CBM2 protein identified in *Butyrivibrio fibrisolvens* H17c (acc. num. CAA35574.1)(Berger et al. 1989), also target multiple substrates, sometime marginally, and are thus considered polyspecific enzymes. Similarly, although most GH5 endo-β-1,4-mannosidases target mannans only, few enzymes such as cel5B mannanase from *Thermotoga maritima* MSB8 target multiple substrates (acc. num. AAD36817.1) (Nelson et al. 1999). The GH family 8 (n=74 characterized enzymes) is also polyspecific (table 1). Endocellulases (n=36), chitosanases (EC.3.2.1.132, n=23), and few xylanases (EC.3.2.1.8, n=12) are listed. Few proteins from *Bacillus* (n=1), *Paenibacillus* (n=3), *Lysobacter* (n=1) are polyspecific and target lichenan and cellulose or chitosan [e.g., (Ogura et al. 2006)]. Licheninase activity is never observed alone, however, suggesting that lichenan is not the primary target of characterized GH8s. The GH family 12 (n=68), also polyspecific, lists enzymes targeting cellulose or xyloglucan (EC.3.2.151) (table 1). TrCel12A from *Trichoderma reesei* is the only polyspecific GH12 listed (EC.3.2.14/151, acc. num. AAE59774.1) (US patent #6187732).

Proteins in GH families 9, 44, and 45 are mostly monospecific endocellulases (>95%) with only few characterized enzymes from each family targeting a different substrate (table 1). Non-cellulolytic GH9s target other plant polysaccharides (e.g., Cel9X xyloglucanase from *Clostridium cellulolyticum* H10, acc. num. ACL76949.1 (Ravachol et al. 2014)) and oligosaccharides (e.g., exo-β-D-glucosaminidase from *Photobacterium profundum* SS9, acc. num. CAG18943.1 (Honda et al. 2011)). Similarly, GH families 6, 7, and 48 are mostly monospecific exocellulases (table 1). Finally, few cellulolytic enzymes are listed in GH families 10, 26, 51, and 74. Among others, 3 GH5 endocellulases were associated with catalytic domains from GH family 10 (xylanase) in *Caldicellulosiruptor* (Talamantes et al. 2016) and an uncultured bacterium (Saul et al. 1989). Conversely, in GH26, among the 72 characterized proteins listed, all 4 cellulolytic enzymes are polyspecific, suggesting that cellulose is not the primary target of GH26 (von Freiesleben et al. 2016). Finally, in GH families 51 and 74, only 6 and 2 proteins are cellulases, out of 78 and 27 characterized proteins, respectively. Most characterized GH51 are α-L-arabinofuranosidases (EC.3.2.1.55) whereas most GH74 are xyloglucanases (EC.3.21.151).

Few characterized GHs targeting cellulose have been identified in animals, including cellulases from termites (e.g., *Coptotermes*), crustaceans (e.g., *Limnoria*), and mollusks (e.g., *Ampularia* and *Aplysia*) [e.g., (Byrne et al. 1999; Guo et al. 2008; King et al. 2010)]. Most of these enzymes are likely involved in digestive functions (Watanabe and Tokuda

2010) whereas GH5 and GH9 cellulases from plants (e.g., *Arabidopsis, Nicotiana*) are likely involved in plant cell wall synthesis and remodeling (Vain et al. 2014). Few hydrolytic cellulases from archaea (e.g., GH5 and 12 from *Crenarchaeota*) (Huang et al. 2005; Graham et al. 2011) and algae (e.g., GH9 from *Eisenia fetida,* acc. num. BAM14716.1) have been characterized. However, beside these enzymes, the overwhelming majority of characterized cellulases are from bacteria or fungi. Characterized GH5, 6, and 12 from fungi account for 28%, 34%, and 36% of the characterized proteins, whereas bacterial enzymes account for 53%, 62%, and 55%, respectively (excluding the enzymes of unknown origin). Next, 86% of characterized GH7s derived from fungi (mostly from *Ascomycota*) whereas some are from termites (e.g., *Coptotermes*) and their symbiotic protozoa *Holomastigotoides.* Similarly, in GH family 45, besides two bacterial enzymes from *Cellvibrio* (acc. num. ACE82688.1) and *Fibrobacter* (acc. num. ACX75523.1), 49% of characterized enzymes are derived from fungi (mostly from *Ascomycota* and *Mucoromycota* and few from *Basidiomycota* and *Neocallimastigomycota*) and 47% from animals (mostly arthropods and mollusks). The systematic investigation of sequenced bacterial and fungal genomes support the skewed distribution of characterized GH7 and 45 in fungi (Berlemont and Martiny 2013; Berlemont 2017). Thus, due to their abundance in sequenced genomes, identifying sequences for GH7 can be used to estimate the contribution of fungi to plant cell wall deconstruction in the environment (Berlemont et al. 2014; Treseder and Lennon 2015). Conversely, no characterized enzymes from GH families 8, 9, and 44 are from fungi. More precisely, 49% and 35% of the characterized GH8s are from *Firmicutes* and *Proteobacteria*, respectively but few are from *Fibrobacters* and *Actinobacteria.* In GH9 family, 19, 20, and 48% of characterized enzymes derive from plants, animals, and bacteria (mostly *Firmicutes* and *Proteobacteria*), respectively. Similarly, most characterized GH44s and GH48s are from bacteria (mostly *Firmicutes*) whereas one GH44 is from the mollusk *Bankia gouldi* (acc. num. CAH68691.1) and one GH48 is from the insect *Gastrophysa atrocyanea* (acc. num. BAE94320.1). However, despite a skewed distribution in characterized enzymes, identified sequences for GH9 and 44 are abundant in bacteria and fungi (Berlemont 2017). Thus, amongst the cellulase families, only GH8 and GH48 can be predominantly associated with bacterial lineages. However, unlike GH7, many GH8s are not hydrolytic cellulases. Indeed many GH8s are chitosanases in *Firmicutes* or non-cellulolytic cellulases associated with the bacterial cellulose synthesis operon in *Proteobacteria* (Berlemont and Martiny 2013; Römling and Galperin 2015) whereas GH48s are relatively rare in sequenced genomes.

Xylanases

Xylan, abundant in hemicellulose from plant cell wall, consists of a linear backbone made of β-1,3/4 linked β-D-xylose "decorated" with various side groups such as acetyl-groups in O-2 and O-3 positions. Larger groups can substitute the xylan backbone (e.g., arabinofuranosyl and 4-O-methyl glucuronyl) (Grantham et al. 2017). The enzymatic deconstruction of xylan requires first the removal of the side chains and then the deconstruction of the xylan backbone (Dodd and Cann 2009). Carbohydrate esterases (CE) are carbohydrate-active enzymes involved in the removal of the side chains from the substituted xylose units (Dodd et al. 2009) whereas specfic glycoside hydrolases, called xylanases, are involved in the xylan backbone hydrolysis (Kulkarni et al. 1999; Dodd and

Cann 2009). The removal of cumbersome side chains by CEs expose the xylan backbone to xylanases and improve the overall deconstruction process (Vardakou et al. 2008). Amongst others, acetyl xylan esterases (EC.3.1.1.72) are found in CE families 1, 2, 3, 4, 5, 6, 7, 12, and 16 whereas feruloyl esterases (EC 3.1.1.73) are found in CE family 1 (Lombard et al. 2014). CE and endo-1,4-β-xylanase domains are frequently identified in multi-domain proteins [e.g., Xyn10D-Fae1A, acc. num. ACN78954.1 (Dodd et al. 2009)].

Xylanases/endo-β-1,4-xylanases (EC.3.2.1.8) are endo-acting GHs targeting the backbone of xylan from plants and seaweed whereas endo-β-1,3-xylanases (EC.3.2.1.32) target β-1,3-linked xylose in seaweed (Konishi et al. 2012). Xylanases release xylobiose that is further degraded by β-xylosidases (EC.3.2.1.37, not discussed here). According to CAZy, endo-β-1,4-xylanases are found in GH families 3, 5, 8, 9, 10, 11, 12, 16, 26, 30, 43, 44, 51, 62, 98, and 141 whereas endo-β-1,3-xylanases are found in GH families 11 and 26 only. The vast majority of characterized xylanases are from GH families 10 and 11 (Lombard et al., 2014). In GH10 (n= 350 listed proteins) and GH11 (n= 271), 96.8% and 99.6% of the characterized enzymes are monospecific endo-1,4-β-xylanases, respectively whereas few characterized endo-1,3-β-xylanases are found in GH11 (0.01%). In GH10, the few non-xylanolytic enzymes are 2 endocellulases (EC.3.2.1.4) and 8 enzymes with unspecified substrate (EC. 3.2.1.-). Most characterized xylanases have one catalytic domain such as Xyn10A (acc. num. AGA16736.1)(Bai et al. 2012), however some proteins consist of multi-domain and multi-activity enzymes such as a Xyn10D-Fae1A (i.e., GH10-CE1, acc. num. ACN78954.1) from *Prevotella ruminicola* 23 (Dodd et al. 2009).

Although polyspecific, the GH family 30 (n=38 listed proteins) is dominated by endo-1,4-β-xylanases (n=17). In this family, other activities include glucuronoarabinoxylan endo-1,4-β-xylanases (EC.3.2.1.136, n=8) and β-xylosidases (EC.3.2.1.37, n=5) also involved in xylan deconstruction. GH30 also contains a few non-xylanolytic activities such as β-glucosidase (EC.3.2.1.21, n=2), β-1,6-glucanase (EC.3.2.1.75, n=7), endo-β-1,6-galactanase (EC. 3.2.1.164, n=3), and some unspecified activity (EC 3.2.1.-) (n=3).

Beside GH10, 11 and 30, a few xylanolytic enzymes were identified in cellulolytic GH families (see cellulase section) and in some polyspecific families including GH26, 43, 51, 62, and 98, according to CAZy (Lombard et al. 2014). More precisely, only 6 characterized GH26 are bacterial β-1,3-xylanases (EC.3.2.1.32) whereas most GH26s target mannans [e.g., 74% β-mannanases (EC.3.2.1.78), 8% exo-β-1,4-mannobiohydrolase (EC.3.2.1.100)]. Similarly, although most of the 151 listed GH43s target arabinans [e.g., 37% α-L-arabinofuranosidases (EC.3.2.1.55), 22% arabinanase (EC.3.2.1.99)] or xylobiose (EC. 3.2.1.37, 34%), 6 proteins with domain from GH family 43 are xylanolytic. Most of these xylanases are bifunctional multidomain proteins with two catalytic domains (e.g., β-1,4-xylanase/α-L-arabinosidase (XynA) from *Caldicellulosiruptor sp.* Tok7B.1, acc. num. AAD30363.1)(Gibbs et al. 2000). Similarly the unique "xylanase" from GH family 62 is a multidomain xylanase/arabinofuranosidase from *Streptomyces chattanoogensis* UAH23 (acc. num. AAD32559.2)(Hernández et al. 2001). Finally, the only characterized GH51, out of 78 listed proteins, endowed with xylanolytic activity is a distantly related single polyspecific catalytic domain targeting cellulose and xylan from *Alicyclobacillus*

acidocaldarius subsp. acidocaldarius DSM 446 (acc. num. ACV57112.1)(Mavromatis et al. 2010).

Investigating the taxonomic origin of characterized proteins from GH family 10 and 11 supports their broad distribution in sequenced bacterial and fungal genomes (Berlemont and Martiny 2015; Berlemont 2017). In addition, 2 GH10 xylanases are from plants (e.g., maize xylanase) (Wu et al. 2002), and one GH11 is from the insect *Phaedon cochleariae* (acc. num. AGK45632.1)(Pauchet and Heckel 2013). Finally, in GH family 30, all the xylanases, except one from the nematode *Meloidogyne incognita* (acc. num. AAF37276.1) (Mitreva-Dautova et al. 2006) and one exo-xylanase from *Trichoderma reesei* RutC30 (acc. num. AAP64786.1) (US patent #6555335), are from bacteria. However, not all the bacterial GH30s are xylanases, and most non-xylanolytic GH30s target hemicellulose [e.g., glucuronoarabinoxylan-specific endo-β-1,4-xylanases (EC.3.2.1.136) and endo-β-1,6-galactanase (EC.3.2.1.164)]. Most characterized eukaryotic GH30s are fungal enzymes also targeting hemicellulose (e.g., endo-β-1,6-galactanase (EC.3.2.1.164) from *Trichoderma viride* (acc. num. BAC84995.1) (Kotake et al. 2004)].

Chitinases

Chitin is a linear polysaccharide made of β-1,4 linked N-acetylglucosamine produced by fungi and arthropods. The enzymatic deconstruction of chitin requires chitinases. These GHs release the disaccharide chitobiose that is further processed by β-N-acetylhexosaminidase (EC.3.2.1.52). Chitin can also be deacetylated by specific chitin deacetylases (EC.3.5.1.41) to produce chitosan (not discussed here). Chitinolytic enzymes are endochitinases (EC. 3.2.1.14) (Lombard et al. 2014) or exochitinases acting on either the reducing (EC. 3.2.1.201) or the non-reducing end of chitin (EC.3.2.1.200) [e.g., (Wang et al. 2001)]. Although there is no mention of exo-chitinase on CAZy, many chitinolytic enzymes are listed in the mostly monospecific GH families 18 and 19. In addition few chitinases are listed in GH families 23 and 48.

More precisely in GH18, amongst 477 listed proteins, 91% display endochitinase activity whereas 3 proteins are lysozymes targeting bacterial peptidoglycan (EC.3.2.1.17), 21 are endo-β-N-acetylglucosaminidases (EC.3.2.1.96) possibly involved in the deconstruction of chito-oligosaccharides (Gooday 1990; Jhaveri et al. 2015) and 25 target unspecified substrates (EC.3.2.1.-). Although most enzymes are monospecific, 3 polyspecific chitinases-lysozymes from *Bacillus cf. pumilus* SG2 (acc. num. ABI15082.1)*, Hevea brasiliensis* (acc. num. CAA07608.1), and *Nicotinia tabacum* (acc. num. CAA55128.1) are identified [e.g., (Bokma et al. 2000)]. Next, 172 out of 176 proteins from GH family 19 are monospecific endochitinases, 3 are lysozymes, and one targets an unspecified substrate (acc. num. BAE86996.1). Finally GH families 23 and 48, described mostly as lysozyme (type G) (Wohlkönig et al. 2010) and exocellulase families respectively, contain one chitinolytic enzyme each. Interestingly, the GH48 chitinase from the beetle *Gastrophysa atrocyanea* is the only characterized member of the GH family 48 derived from an eukaryote (acc. num. BAE94320.1) (Fujita et al. 2006).

When excluding the enzymes of unknown origin, 51% of characterized GH18 are derived from eukaryotes. More precisely, 20.4, 11.9, 8.4, and 4.5% are from *Ascomycota*, plants, arthropods, and chordates. Conversely, 46% of characterized GH18 are bacterial enzymes: 20.7, 16.2, 7.4 and 1.7% derive from *Proteobaceria*, *Actinobacteria*, *Firmicutes*, and *Bacteroidetes* respectively. Finally 2.4% derive from *Euryarchaeota*. Most characterized GH19 derived from plants (88.8%) whereas few originate from bacteria (mostly *Acinobacteria* and some *Proteobacteria*), and none from fungi. The taxonomic origin of characterized chitinases from GH18 and GH19 reflects the broad distribution of identified chitinases in sequenced genomes (Suzuki et al. 2001; Kawase et al. 2006; Bussink et al. 2007; Berlemont 2017). The skewed distribution of GH19 provides some rationale to link the identification of GH19 in sequenced microbiomes to the environmental deconstruction of chitin by bacteria.

Glycoside Hydrolases identification

The identification of GHs is a prerequisite for understanding the polysaccharide deconstruction by isolated (Youssef et al. 2013) and complex communities of microorganisms (Hess et al. 2011) and for the identification of new enzymes for biotechnological applications (Brunecky et al. 2013). As the CAZy database provides no tool for sequence annotation and is not designed to ease the extraction and the analysis of information at large, we created a custom bioinformatic program to link (i) the functional annotation as listed on CAZy, (ii) the sequences retrieved from the NCBI database, and (iii)– if possible- the taxonomic information for all characterized proteins from GH families involved in the deconstruction of cellulose, xylan, and chitin (supplementary data). Next, sequences were analyzed using HMMscan (Eddy 2011) against the Pfam A database (Finn et al. 2014) as described by Talamantes, and referred to as the Pfam-based annotation here after (Talamantes et al. 2016).

We focused on GH families predominantly involved in cellulose, xylan, and chitin deconstruction. The Pfam-based annotation of characterized proteins (table 1) correctly identified 2,326 proteins out of 2,409 tested sequences (96.6%). More precisely, HMMscan identified 514 PF00150-cellulase domains out 537 sequences retrieved from the GH family 5 (96%). Similarly, more than 90% of the domains from analyzed GH families were categorized as expected with the exception of proteins from GH family 30 and 45. Noteworthy, a detailed analysis of sequences from GH family 30 highlighted its polyspecificity and its structural complexity (St John et al. 2010). Eventually this resulted in the creation of up to 8 subfamilies and all the xylanolytic GH30s fall into the subfamily GH30_8. In GH family 45, 9 of the 55 characterized proteins were miss-annotated. However, these cellulases are derived from invertebrates (e.g., *Aplysia*, acc. num. BAP19116.1) and are known to be distantly related to the other GH45s (Rahman et al. 2014). Finally, the identification of GH families marginally associated with cellulose, xylan, and chitin deconstruction was also supported by Pfam-based annotation (table S1).

Future directions

The vast majority of sequences from the GH families involved in cellulose, xylan, and chitin deconstruction can be identified using Pfam-based annotation, a documented and publicly accessible system (Eddy 2011; Finn et al. 2014). Although similar annotation systems exist [e.g., (Yin et al. 2012)], using the Pfam A database allows for the identification of accessory domains not listed in the CAZy database (Talamantes et al. 2016). However, some recently created GH families containing a reduced number of sequences (e.g., 6 identified and 1 characterized sequences in GH124, acc. num. ABN51673.1) cannot yet be identified using this approach.

The conserved substrate specificity in monospecific GH families provides a way to link the occurrence of specific GH domains to environmental processes (Stursová et al. 2012; Berlemont and Martiny 2016). Moreover, based on the skewed taxomonic distribution of domains from GH families the studied processes can be attributed to specific lineages (Berlemont et al. 2014; Treseder and Lennon 2015; Lladó et al. 2017). In polyspecific families (GH5, 8, and 30) the predictive power is reduced. However, characterized GH5s target mostly cellulose or hemicellulosic substrates, GH8 target mostly cellulose or chitosan (i.e., deacetylate chitin), and GH30 target mostly xylan or hemicellulosic substrates. In these polyspecific families, the substrate specificity cannot be identified using the HMM-based annotation. Detailed characterization of these families eventually led to the creation of monospecific GH subfamilies (St John et al. 2010; Aspeborg et al. 2012). In the future the growing number of sequences in poorly represented GH families (e.g., GH124) and in each subfamilies (e.g., GH30_8) will allow the creation of specific HMM profile to further identify proteins from these groups. In the meantime, the compiled sequences dataset (supplementary data) will be more friendly to the needs of bioinformatics. This dataset can be used for the creation of a custom database to be used with the Basic Local Alignment Search Tool (BLAST) in order to perform detailed sequence comparison, identify close relatives, and help predict the mode of action or the substrate specificity of proteins from polyspecific GH families.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Artzi L, Bayer EA, Moraïs S. Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. Nat Rev Microbiol. 2016; 15:83–95. DOI: 10.1038/nrmicro.2016.164 [PubMed: 27941816]

Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). BMC Evol Biol. 2012; 12:186.doi: 10.1186/1471-2148-12-186 [PubMed: 22992189]

Bai W, Xue Y, Zhou C, Ma Y. Cloning, expression and characterization of a novel salt-tolerant xylanase from *Bacillus sp.* SN5. Biotechnol Lett. 2012; 34:2093–2099. DOI: 10.1007/s10529-012-1011-7 [PubMed: 22864505]

Bailey VL, Fansler SJ, Stegen JC, McCue LA. Linking microbial community structure to β-glucosidic function in soil aggregates. ISME J. 2013; 7:2044–53. DOI: 10.1038/ismej.2013.87 [PubMed: 23719152]

Berger E, Jones WA, Jones DT, Woods DR. Cloning and sequencing of an endoglucanase (end1) gene from *Butyrivibrio fibrisolvens* H17c. Mol Gen Genet. 1989; 219:193–8. [PubMed: 2615759]

Berlemont R. Distribution and diversity of enzymes for polysaccharide degradation in fungi. Sci Rep. 2017; 7:222.doi: 10.1038/s41598-017-00258-w [PubMed: 28302998]

Berlemont R, Allison SD, Weihe C, Lu Y, Brodie EL, Martiny JBH, Martiny AC. Cellulolytic potential under environmental changes in microbial communities from grassland litter. Front Microbiol. 2014; 5:639.doi: 10.3389/fmicb.2014.00639 [PubMed: 25505459]

Berlemont R, Martiny AC. Phylogenetic distribution of potential cellulases in bacteria. Appl Environ Microbiol. 2013; 79:1545–54. DOI: 10.1128/AEM.03305-12 [PubMed: 23263967]

Berlemont R, Martiny AC. Genomic potential for polysaccharides deconstruction in bacteria. Appl Environ Microbiol. 2015; 81:1513–19. DOI: 10.1128/AEM.03718-14 [PubMed: 25527556]

Berlemont R, Martiny AC. Glycoside Hydrolases across Environmental Microbial Communities. PLOS Comput Biol. 2016; 12:e1005300.doi: 10.1371/journal.pcbi.1005300 [PubMed: 27992426]

Bokma E, Barends T, Terwissch van Scheltingab AC, Dijkstr BW, Beintema JJ. Enzyme kinetics of hevamine, a chitinase from the rubber tree *Hevea brasiliensis*. FEBS Lett. 2000; 478:119–22. [PubMed: 10922481]

Brunecky R, Alahuhta M, Xu Q, Donohoe BS, Crowley MF, Kataeva IA, Yang S-J, Resch MG, Adams MWW, Lunin VV, Himmel ME, Bomble YJ. Revealing nature's cellulase diversity: the digestion mechanism of *Caldicellulosiruptor bescii* CelA. Science. 2013; 342:1513–6. DOI: 10.1126/science.1244273 [PubMed: 24357319]

Bussink AP, Speijer D, Aerts JMFG, Boot RG. Evolution of mammalian chitinase(-like) members of family 18 glycosyl hydrolases. Genetics. 2007; 177:959–970. DOI: 10.1534/genetics.107.075846 [PubMed: 17720922]

Byrne KA, Lehnert SA, Johnson SE, Moore SS. Isolation of a cDNA encoding a putative cellulase in the red claw crayfish *Cherax quadricarinatus*. Gene. 1999; 239:317–24. [PubMed: 10548733]

Din N, Gilkes NR, Tekant B, Miller RC, Warren RAJ, Kilburn DG. Non–Hydrolytic Disruption of Cellulose Fibres by the Binding Domain of a Bacterial Cellulase. Bio/Technology. 1991; 9:1096–1099. DOI: 10.1038/nbt1191-1096

Dodd D, Cann IKO. Enzymatic deconstruction of xylan for biofuel production. Glob Change Biol Bioenergy. 2009; 1:2–17. DOI: 10.1111/j.1757-1707.2009.01004.x [PubMed: 20431716]

Dodd D, Kocherginskaya SA, Spies MA, Beery KE, Abbas CA, Mackie RI, Cann IKO. Biochemical analysis of a β-D-xylosidase and a bifunctional xylanase-ferulic acid esterase from a xylanolytic gene cluster in *Prevotella ruminicola* 23. J Bacteriol. 2009; 191:3328–3338. DOI: 10.1128/JB.01628-08 [PubMed: 19304844]

Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011; 7:e1002195.doi: 10.1371/journal.pcbi.1002195 [PubMed: 22039361]

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. Pfam: the protein families database. Nucleic Acids Res. 2014; 42:D222–30. DOI: 10.1093/nar/gkt1223 [PubMed: 24288371]

Fujita K, Shimomura K, Yamamoto K, Yamashita T, Suzuki K. A chitinase structurally related to the glycoside hydrolase family 48 is indispensable for the hormonally induced diapause termination in a beetle. Biochem Biophys Res Commun. 2006; 345:502–507. DOI: 10.1016/j.bbrc.2006.04.126 [PubMed: 16684504]

Gefen G, Anbar M, Morag E, Lamed R, Bayer EA. Enhanced cellulose degradation by targeted integration of a cohesin-fused β-glucosidase into the *Clostridium thermocellum* cellulosome. Proc Natl Acad Sci U S A. 2012; 109:10298–303. DOI: 10.1073/pnas.1202747109 [PubMed: 22689961]

Gibbs MD, Reeves RA, Farrington GK, Anderson P, Williams DP, Bergquist PL. Multidomain and multifunctional glycosyl hydrolases from the extreme thermophile *Caldicellulosiruptor* isolate Tok7B.1. Curr Microbiol. 2000; 40:333–40. [PubMed: 10706665]

Gibbs MD, Saul DJ, Lüthi E, Bergquist PL. The beta-mannanase from *Caldocellum saccharolyticum*, is part of a multidomain enzyme. Appl Environ Microbiol. 1992; 58:3864–7. [PubMed: 1476429]

Gooday GW. Physiology of microbial degradation of chitin and chitosan. Biodegradation. 1990; 1:177–190. DOI: 10.1007/BF00058835

Graham JE, Clark ME, Nadler DC, Huffer S, Chokhawala HA, Rowland SE, Blanch HW, Clark DS, Robb FT. Identification and characterization of a multidomain hyperthermophilic cellulase from an archaeal enrichment. Nat Commun. 2011; 2:375.doi: 10.1038/ncomms1373 [PubMed: 21730956]

Grantham NJ, Wurman-Rodrich J, Terrett OM, Lyczakowski JJ, Stott K, Iuga D, Simmons TJ, Durand-Tardif M, Brown SP, Dupree R, Busse-Wicher M, Dupree P. An even pattern of xylan substitution is critical for interaction with cellulose in plant cell walls. Nat Plants. 2017; 3:859–865. DOI: 10.1038/s41477-017-0030-8 [PubMed: 28993612]

Guo R, Ding M, Zhang S-L, Xu G, Zhao F. Molecular cloning and characterization of two novel cellulase genes from the mollusc *Ampullaria crossean*. J Comp Physiol B. 2008; 178:209–215. DOI: 10.1007/s00360-007-0214-z [PubMed: 17952442]

Haitjema CH, Gilmore SP, Henske JK, Solomon KV, de Groot R, Kuo A, Mondo SJ, Salamov AA, LaButti K, Zhao Z, Chiniquy J, Barry K, Brewer HM, Purvine SO, Wright AT, Hainaut M, Boxma B, van Alen T, Hackstein JHP, Henrissat B, Baker SE, Grigoriev IV, O'Malley MA. A parts list for fungal cellulosomes revealed by comparative genomics. Nat Microbiol. 2017; 2:17087.doi: 10.1038/nmicrobiol.2017.87 [PubMed: 28555641]

Henrissat B, Davies G. Structural and sequence-based classification of glycoside hydrolases. Curr Opin Struct Biol. 1997; 7:637–644. [PubMed: 9345621]

Hernández A, Copa-Patiño JL, Soliveri J. xln23 from *Streptomyces chattanoogensis* UAH23 encodes a putative enzyme with separate xylanase and arabinofuranosidase catalytic domains. DNA Seq. 2001; 12:167–77. [PubMed: 11762192]

Hervé C, Rogowski A, Blake AW, Marcus SE, Gilbert HJ, Knox JP, Herve C, Rogowski A, Blake AW, Marcus SE, Gilbert HJ, Knox JP. Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. Proc Natl Acad Sci U S A. 2010; 107:15293–8. DOI: 10.1073/pnas.1005732107 [PubMed: 20696902]

Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, Mackie RI, Pennacchio La, Tringe SG, Visel A, Woyke T, Wang Z, Rubin EM. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science. 2011; 331:463–7. DOI: 10.1126/science.1200387 [PubMed: 21273488]

Honda Y, Shimaya N, Ishisaki K, Ebihara M, Taniguchi H. Elucidation of exo-β-d-glucosaminidase activity of a family 9 glycoside hydrolase (PBPRA0520) from *Photobacterium profundum* SS9. Glycobiology. 2011; 21:503–511. DOI: 10.1093/glycob/cwq191 [PubMed: 21098515]

Huang Y, Krauss G, Cottaz S, Driguez H, Lipps G. A highly acid-stable and thermostable endo-β-glucanase from the thermoacidophilic archaeon *Sulfolobus solfataricus*. Biochem J. 2005; 385:581–588. DOI: 10.1042/BJ20041388 [PubMed: 15456402]

Jhaveri P, Papastamatiou YP, German DP. Digestive enzyme activities in the guts of bonnethead sharks (*Sphyrna tiburo*) provide insight into their digestive strategy and evidence for microbial digestion in their hindguts. Comp Biochem Physiol -Part A Mol Integr Physiol. 2015; 189:76–83. DOI: 10.1016/j.cbpa.2015.07.013

Kawase T, Yokokwa S, Saito A, Fujii T, Nikaidou N, Miyashita K, Watanabe T. Comparison of Enzymatic and Antifungal Properties between Family 18 and 19 Chitinases from *S. coelicolor* A3(2). Biosci Biotechnol Biochem. 2006; 70:988–998. DOI: 10.1271/bbb.70.988 [PubMed: 16636468]

King AJ, Cragg SM, Li Y, Dymond J, Guille MJ, Bowles DJ, Bruce NC, Graham IA, McQueen-Mason SJ. Molecular insight into lignocellulose digestion by a marine isopod in the absence of gut microbes. Proc Natl Acad Sci. 2010; 107:5345–5350. DOI: 10.1073/pnas.0914228107 [PubMed: 20212162]

Knight R, Jansson J, Field D, Fierer N, Desai N, Fuhrman JA, Hugenholtz P, van der Lelie D, Meyer F, Stevens R, Bailey MJ, Gordon JI, Kowalchuk GA, Gilbert JA. Unlocking the potential of metagenomics through replicated experimental design. Nat Biotechnol. 2012; 30:513–20. DOI: 10.1038/nbt.2235 [PubMed: 22678395]

Konishi T, Nakata I, Miyagi Y, Tako M. Extraction of β-1,3 Xylan from Green Seaweed, *Caulerpa lentillifera*. J Appl Glycosci. 2012; 59:161–163. DOI: 10.5458/jag.jag.JAG-2011_025

Kotake T, Kaneko S, Kubomoto A, Haque MA, Kobayashi H, Tsumuraya Y. Molecular cloning and expression in Escherichia coli of a *Trichoderma viride* endo-beta-(1-6)-galactanase gene. Biochem J. 2004; 377:749–55. DOI: 10.1042/BJ20031145 [PubMed: 14565843]

Kulkarni N, Shendye A, Rao M. Molecular and biotechnological aspects of xylanases. FEMS Microbiol Rev. 1999; 23:411–456. [PubMed: 10422261]

Lladó S, López-Mondéjar R, Baldrian P. Forest Soil Bacteria: Diversity, Involvement in Ecosystem Processes, and Response to Global Change. Microbiol Mol Biol Rev. 2017; 81:e00063–16. DOI: 10.1128/MMBR.00063-16 [PubMed: 28404790]

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014; 42:D490–5. DOI: 10.1093/nar/gkt1178 [PubMed: 24270786]

Mavromatis K, Sikorski J, Lapidus A, Glavina Del Rio T, Copeland A, Tice H, Cheng J-F, Lucas S, Chen F, Nolan M, Bruce D, Goodwin L, Pitluck S, Ivanova N, Ovchinnikova G, Pati A, Chen A, Palaniappan K, Land M, Hauser L, Chang Y-J, Jeffries CD, Chain P, Meincke L, Sims D, Chertkov O, Han C, Brettin T, Detter JC, Wahrenburg C, Rohde M, Pukall R, Göker M, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Klenk H-P, Kyrpides NC. Complete genome sequence of *Alicyclobacillus acidocaldarius* type strain (104-IAT). Stand Genomic Sci. 2010; 2:9–18. DOI: 10.4056/sigs.591104 [PubMed: 21304673]

Medie FM, Davies GJ, Drancourt M, Henrissat B. Genome analyses highlight the different biological roles of cellulases. Nat Rev Microbiol. 2012; 10:227–234. [PubMed: 22266780]

Mitreva-Dautova M, Roze E, Overmars H, de Graaff L, Schots A, Helder J, Goverse A, Bakker J, Smant G. A Symbiont-Independent Endo-1,4-β-Xylanase from the Plant-Parasitic Nematode *Meloidogyne incognita*. Mol Plant-Microbe Interact. 2006; 19:521–529. DOI: 10.1094/MPMI-19-0521 [PubMed: 16673939]

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature. 1999; 399:323–329. DOI: 10.1038/20601 [PubMed: 10360571]

Ogura J, Toyoda A, Kurosawa T, Chong AL, Chohnan S, Masaki T. Purification, Characterization, and Gene Analysis of Cellulase (Cel8A) from *Lysobacter sp.* IB-9374. Biosci Biotechnol Biochem. 2006; 70:2420–2428. DOI: 10.1271/bbb.60157 [PubMed: 17031054]

Palackal N, Lyon CS, Zaidi S, Luginbühl P, Dupree P, Goubet F, Macomber JL, Short JM, Hazlewood GP, Robertson DE, Steer BA. A multifunctional hybrid glycosyl hydrolase discovered in an uncultured microbial consortium from ruminant gut. Appl Microbiol Biotechnol. 2007; 74:113–124. DOI: 10.1007/s00253-006-0645-6 [PubMed: 17103163]

Park BH, Karpinets TV, Syed MH, Leuze MR, Uberbacher EC. CAZymes Analysis Toolkit (CAT): Web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. Glycobiology. 2010; 20:1574–1584. DOI: 10.1093/glycob/cwq106 [PubMed: 20696711]

Pauchet Y, Heckel DG. The genome of the mustard leaf beetle encodes two active xylanases originally acquired from bacteria through horizontal gene transfer. Proc R Soc B Biol Sci. 2013; 280:20131021–20131021. DOI: 10.1098/rspb.2013.1021

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database. Nucleic Acids Res. 2012; 40:D290–301. DOI: 10.1093/nar/gkr1065 [PubMed: 22127870]

Rahman MM, Inoue A, Ojima T. Characterization of a GHF45 cellulase, AkEG21, from the common sea hare *Aplysia kurodai*. Front Chem. 2014; 2:60.doi: 10.3389/fchem.2014.00060 [PubMed: 25147784]

Ravachol J, Borne R, Tardif C, de Philip P, Fierobe H-P. Characterization of all family-9 glycoside hydrolases synthesized by the cellulosome-producing bacterium *Clostridium cellulolyticum*. J Biol Chem. 2014; 289:7335–48. DOI: 10.1074/jbc.M113.545046 [PubMed: 24451379]

Rignall TR, Baker JO, McCarter SL, Adney WS, Vinzant TB, Decker SR, Himmel ME. Effect of Single Active-Site Cleft Mutation on Product Specificity in a Thermostable Bacterial Cellulase. Appl Biochem Biotechnol. 2002; 98–100:383–394. DOI: 10.1385/ABAB:98-100:1-9:383

Römling U, Galperin MY. Bacterial cellulose biosynthesis: diversity of operons, subunits, products, and functions. Trends Microbiol. 2015; 23:545–557. DOI: 10.1016/j.tim.2015.05.005 [PubMed: 26077867]

Saul DJ, Williams LC, Love DR, Chamley LW, Bergquist PL. Nucleotide sequence of a gene from *Caldocellum saccharolyticum* encoding for exocellulase and endocellulase activity. Nucleic Acids Res. 1989; 17:439. [PubMed: 2789517]

St John FJ, González JM, Pozharski E. Consolidation of glycosyl hydrolase family 30: a dual domain 4/7 hydrolase family consisting of two structurally distinct groups. FEBS Lett. 2010; 584:4435–41. DOI: 10.1016/j.febslet.2010.09.051 [PubMed: 20932833]

Stursová M, Zif áková L, Leigh MB, Burgess R, Baldrian P. Cellulose utilization in forest litter and soil: identification of bacterial and fungal decomposers. FEMS Microbiol Ecol. 2012; 80:735–46. DOI: 10.1111/j.1574-6941.2012.01343.x [PubMed: 22379979]

Suzuki M, Morimatsu M, Yamashita T, Iwanaga T, Syuto B. A novel serum chitinase that is expressed in bovine liver. FEBS Lett. 2001; 506:127–130. DOI: 10.1016/S0014-5793(01)02893-9 [PubMed: 11591385]

Talamantes D, Biabini N, Dang H, Abdoun K, Berlemont R. Natural diversity of cellulases, xylanases, and chitinases in bacteria. Biotechnol Biofuels. 2016; 9:133.doi: 10.1186/s13068-016-0538-6 [PubMed: 27366206]

Treseder KK, Lennon JT. Fungal traits that drive ecosystem dynamics on Land. Microbiol Mol Biol Rev. 2015; 79:243–262. DOI: 10.1128/MMBR.00001-15 [PubMed: 25971588]

Vain T, Crowell EF, Timpano H, Biot E, Desprez T, Mansoori N, Trindade LM, Pagant S, Robert S, Hofte H, Gonneau M, Vernhettes S. The Cellulase KORRIGAN Is Part of the Cellulose Synthase Complex. PLANT Physiol. 2014; 165:1521–1532. DOI: 10.1104/pp.114.241216 [PubMed: 24948829]

Vardakou M, Dumon C, Murray JW, Christakopoulos P, Weiner DP, Juge N, Lewis RJ, Gilbert HJ, Flint JE. Understanding the Structural Basis for Substrate and Inhibitor Recognition in Eukaryotic GH11 Xylanases. 2008; :1293–1305. DOI: 10.1016/j.jmb.2007.11.007

von Freiesleben P, Spodsberg N, Blicher TH, Anderson L, Jørgensen H, Stålbrand H, Meyer AS, Krogh KBRM. An *Aspergillus nidulans* GH26 endo-β-mannanase with a novel degradation pattern on highly substituted galactomannans. Enzyme Microb Technol. 2016; 83:68–77. DOI: 10.1016/j.enzmictec.2015.10.011 [PubMed: 26777252]

Wang S-Y, Moyne A-L, Thottappilly G, Wu S-J, Locy RD, Singh NK. Purification and characterization of a *Bacillus cereus* exochitinase. Enzyme Microb Technol. 2001; 28:492–498. [PubMed: 11267643]

Watanabe H, Tokuda G. Cellulolytic Systems in Insects. Annu Rev Entomol. 2010; 55:609–632. DOI: 10.1146/annurev-ento-112408-085319 [PubMed: 19754245]

Wilson DB. Microbial diversity of cellulose hydrolysis. Curr Opin Microbiol. 2011; 14:259–63. DOI: 10.1016/j.mib.2011.04.004 [PubMed: 21531609]

Wohlkönig A, Huet J, Looze Y, Wintjens R. Structural Relationships in the Lysozyme Superfamily: Significant Evidence for Glycoside Hydrolase Signature Motifs. PLoS One. 2010; 5:e15388.doi: 10.1371/journal.pone.0015388 [PubMed: 21085702]

Wu SSH, Suen DF, Chang HC, Huang AHC. Maize tapetum xylanase is synthesized as a precursor, processed and activated by a serine protease, and deposited on the pollen. J Biol Chem. 2002; 277:49055–64. DOI: 10.1074/jbc.M208804200 [PubMed: 12368281]

Xu Q, Ding S-Y, Brunecky R, Bomble YJ, Himmel ME, Baker JO. Improving activity of minicellulosomes by integration of intra- and intermolecular synergies. Biotechnol Biofuels. 2013; 6:126.doi: 10.1186/1754-6834-6-126 [PubMed: 23987588]

Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2012; 40:W445–W451. DOI: 10.1093/nar/gks479 [PubMed: 22645317]

Youssef NH, Couger MB, Struchtemeyer CG, Liggenstoffer AS, Prade RA, Najar FZ, Atiyeh HK, Wilkins MR, Elshahed MS. The Genome of the Anaerobic Fungus *Orpinomyces sp*. Strain C1A Reveals the Unique Evolutionary History of a Remarkable Plant Biomass Degrader. Appl Environ Microbiol. 2013; 79:4620.doi: 10.1128/aem.00821-13 [PubMed: 23709508]

Table 1

Distribution of substrate specificity and activities of characterized enzymes from GH families associated with the deconstruction of cellulose, xylan, and chitin, according to the CAZy database (in August 2017).

GH family (listed proteins) (# unique sequences retrieved from CAZYdb)	Polyspecificity activity (%)	PFam annotation
GH5 (537)	**Cellulose** EC.3.2.1.4 (61%); Mannan EC.3.2.1.78 (20%); Lichenan EC.3.2.1.73 (2%); **Xylan** EC.3.2.1.8 (1.6%); Xyloglucan EC.3.2.1.151 (1%); *Unspecified EC.3.2.1.-(8%)*	PF00150 (98.7%)
GH6 (65)	**Cellulose** EC.3.2.1.91 (64%); **Cellulose** EC.3.2.1.4 (37%); **Cellulose** EC.2.4.1.12 (1.5%); *Unspecified EC.3.2.1-(3%)*	PF01341 (100%)
GH7 (84)	**Cellulose** EC.3.2.1.176 (69%); **Cellulose** EC.3.2.1.4 (32%)	PF00840 (98.7%)
GH8 (74)	**Cellulose** EC.3.2.1.4 (49%); Chitosan EC.3.2.1.132 (31%); **Xylan** EC.3.2.1.8 (11%); **Xylan** EC.3.2.1.156 (5%); *Unspecified EC.3.2.1.-(5%)*	PF01270 (100%)
GH9 (170)	**Cellulose** EC.3.2.1.4 (86%); Lichenan EC.3.2.1.73 (3%); **Cellulose** EC.3.2.191 (1.7%); *Unspecified EC.3.2.1.-(7%)*	PF00759 (100%)
GH10 (350)	**Xylan** EC.3.2.1.8 (96%); **Xylan** EC.2.4.2-(0.5%); *Unspecified EC.3.2.1.-(3.4%)*	PF00331 (96.7%)
GH11 (271)	**Xylan** EC.3.2.1.8 (100%); **Xylan** EC.3.2.1.32 (0.3%)	PF00457 (95.8%)
GH12 (68)	**Cellulose** EC.3.2.1.4 (76%); Xyloglucan EC.3.2.1.151 (19%); Lichenan EC.3.2.1.73 (3%); **Xylan** EC.2.4.1.2 (1.5)	PF01670 (100%)
GH18 (477)	**Chitin** EC.3.2.1.14 (91%); **Chitin** EC.3.2.1.96 (4%); *Unspecified EC.3.2.1.-(5%)*	PF00704 (93%)
GH19 (176)	**Chitin** EC.3.2.1.14 (98%); Peptidoglycan EC.3.2.1.17 (1.7%)	PF00182 (99%)
GH30 (38)	**Xylan** EC.3.2.1.8 (37%); β-1,6-glucan EC.3.2.1.75 (18%); Xylan EC.3.2.1.37 (11%); …	PF17189 (79%)
GH44 (16)	**Cellulose** EC.3.2.1.4 (69%); **Cellulose** EC.3.2.1.151 (37%); *Unspecified EC.3.2.1.-(6%)*	PF12891 (100%)
GH45 (53)	**Cellulose** EC.3.2.1.4 (100%)	PF02015 (83.02%)
GH48 (19)	**Cellulose** EC.3.2.1.176 (58%); **Cellulose** EC.3.2.1.4 (16%); **Chitin** EC.3.2.1.14 (5%); *Unspecified EC.3.2.1.-(26%)*	PF02011 (100%)
GH85 (11)	**Chitin** EC.3.2.1.96 (100%)	PF03644 (100%)