Computational and Structural Biotechnology Journal 19 (2021) 5931–5942

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Expanded analyses of the functional correlations within structural classifications of glycoside hydrolases

Dan-dan Li [a], Jin-lan Wang [b], Ya Liu [a], Yue-zhong Li [a,1,*], Zheng Zhang [a,c,2,*]

[a] State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao 266237, China
[b] National Administration of Health Data, Jinan 250002, China
[c] Suzhou Research Institute, Shandong University, Suzhou 215123, China

ARTICLE INFO

Article history:
Received 1 July 2021
Received in revised form 30 October 2021
Accepted 30 October 2021
Available online 2 November 2021

Keywords:
Glycoside hydrolase
Superfamily
Fold
Sequence–structure–function
Evolution

ABSTRACT

Glycoside hydrolases (GHs) are greatly diverse in sequences and functions, but systematic studies of GH relationships based on structural information are lacking. Here, we report that GHs have multiple evolutionary origins and are structurally derived from 27 homologous superfamilies and 16 folds, but GHs are highly biased to distribute in a few superfamilies and folds. Six of these superfamilies are widely encoded by archaea, bacteria, and eukaryotes, indicating that they may be the most ancient in origin. Most superfamilies vary in enzyme function, and some, such as the superfamilies of $(\beta/\alpha)_8$-barrel and $(\alpha/\alpha)_6$-barrel structures, exhibit extreme functional diversity; this is highly positively correlated with sequence diversity. More than one-third of glycosidase activities show a phenomenon of convergent evolution, especially the degradation functions of GHs on polysaccharides. The GHs of most superfamilies have relatively narrow environmental distributions, normally with the highest abundance in host-associated environments and a distribution preference for moderate low-temperature and acidic environments. Overall, our expanded analysis facilitates an understanding of complex GH sequence–structure–function relationships and may guide our screening and engineering of GHs.

1. Introduction

Proteins are the main actors that perform a myriad of exquisite functions in life, and this diversity has been achieved through the diversification of protein sequences and structures [1]. However, compared with the unlimited number of protein sequences, the number of basic shapes of protein structures is finite, with probably only thousands of folds in existence [2,3]. Point mutations and insertions/deletions drive the diversification of protein sequences and structures [4–7], while natural selection tends to preserve the folds of core structures of proteins and alter their sequences to generate diverse functions [8,9]. Exploring the relationships among the protein sequence space, structure space, and function space is a fundamental scientific focus. Characterizing these relationships may carry practical implications for protein function prediction [3,10].

Glycoside hydrolases (GHs) are a vast class of enzymes that catalyze the hydrolysis of glycosidic linkages and are mainly responsible for carbohydrate degradation in nature [11]. Carbohydrates, as the most abundant photosynthesis-fixed carbon form on Earth, possess unrivaled structural and chemical diversities that enable their various functions, such as structural maintenance (cellulose, alginate, and chitin) and energy storage (glycogen and starch) [12]. With the importance of substrates and products, GHs play crucial roles in many biological processes and applications. For instance, in the carbon cycle, microorganisms in the soil decompose plant cells to produce CO_2 and various fermentation products with the assistance of GHs [13–15]. Cellulases, xylanases, and other glucosidases have been used to produce sugars from pretreated biomass substrates, and the sugars are then fermented to produce ethanol or butanol as renewable substitutes for gasoline [16–19]. Therefore, detailed knowledge of GH function is invaluable for understanding their ecological effects and industrial applications.

The known GH protein sequences have exceeded 1,000,000, and GHs are classified into more than 160 families based on their sequence similarities [11]. However, sequence classification of these enzymes is complicated to determine the overall situation

* Corresponding authors at: State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao 266237, China.

E-mail addresses: lilab@sdu.edu.cn (Y.-z. Li), zhangzheng@sdu.edu.cn (Z. Zhang).

[1] ORCID: 0000-0001-8336-6638.

[2] ORCID: 0000-0001-9971-6006.

https://doi.org/10.1016/j.csbj.2021.10.039

of these enzymes. Compared with sequences, protein structures are more conserved during evolution; GH families descending from a common ancestor are beyond the limit of detection based on their sequence similarities, but they might still have the same folds in structure [12]. Thus, it is possible to reveal the distant evolutionary relationships of GHs based on their structural information. Proteins of known three-dimensional structures have already been categorized according to their structural and evolutionary relationships [20,21]. For instance, the CATH database classifies domains into four levels: class, architecture, topology/fold, and homologous superfamily. Members of homologous superfamilies share a conserved structural core and a common ancestor [21]. The carbohydrate-active enzymes (CAZy) database adopts the classification of GH enzymes fundamentally based on amino acid sequences. Although structural information supports the CAZy classification, the classification of folds is still poor. In this study, through the analysis of all GH family members deposited in the CAZy database, protein structure classification information released by the CATH database, and further prediction of glycoside hydrolase structures with AlphaFold [22], we systematically investigated the complex sequence–structure–function relationships of GHs.

2. Materials and methods

2.1. Glycoside hydrolase identification

CAZy is a specialized database dedicated to the display and analysis of genomic, structural and biochemical information on carbohydrate-active enzymes [11,23]. GHs (EC 3.2.1.x) are a widespread group of enzymes that hydrolyze the glycosidic bond between two or more carbohydrates or between carbohydrate and noncarbohydrate moieties [24]. Information pertaining to the classification and taxonomic sources of GHs was acquired from the latest release (current until September 2021) of the CAZy database, whereby 163 families of GHs were classified based on amino acid sequence similarities (Table S1).

Protein sequences originating from all complete genomes could be assigned to GH families. The complete genomes are those released by the NCBI as regular entries in the daily releases of GenBank [25]. GH gene information from 21,244 bacterial, 424 archaeal, 456 viral, and 352 eukaryotic genomes was derived from the CAZy database. Taxonomic classification information was obtained from the NCBI Taxonomy database [26].

2.2. Identification of superfamilies and folds associated with glycoside hydrolases

CATH (class, architecture, topology/fold, homologous superfamilies) is a hierarchical protein domain classification database [9,21]. At the class level, domains are assigned according to their secondary structural contents; at the architecture level, information on the secondary structure arrangement in a three-dimensional space is used for assignment; at the topology/fold level, information on how the secondary structural elements are connected and arranged is used; and assignments are made to the homologous superfamily level if there is good evidence that the domains are related by evolution, i.e., they are homologous.

Based on the GH crystal structure information, the CATH classifications were searched and matched with GH family classifications. Since many GHs are multidomains, we only analyzed the catalytic domains. For GH families whose crystal structures have not yet been resolved, we used the newly developed AlphaFold algorithm to model the three-dimensional structures of their representative members (Table S2). AlphaFold is an AI system that predicts a protein's 3D structure from its amino acid sequence and regularly achieves accuracy competitive with experiment [22,27]. The three-dimensional structures were displayed in a cartoon model by PyMol (Schrödinger, LLC). The GH sequence diversity and functional diversity in each homologous superfamily were calculated based on the Shannon index [28].

2.3. Corresponding relationship between the EMP OTUs and prokaryotic genomes

The Earth Microbiome Project (EMP) was founded in 2010 to sample the Earth's microbial communities to advance our understanding of the organizing biogeographic principles that govern microbial community structures on Earth [29]. A total of 262,011 operational taxonomic units (OTUs) were obtained from a set of 10,000 EMP samples using Deblur software [30]. Chimera filtering relied on the EMP project.

The 21,244 bacterial and 424 archaeal genomes annotated by the CAZy database were compared with the sequence data of 10,000 EMP samples. Alignments between the EMP OTUs and prokaryotic genomes were performed using BLASTn [31,32], and the corresponding relationships were determined with a 16S rRNA (V4 region) identity greater than 97% as the standard. Number and sequence abundance proportions of identified OTUs of no<5% relative to the total OTUs in one sample were used as the criterion to screen samples for subsequent analysis.

2.4. Environmental distribution of homologous superfamilies

The environmental distribution and abundance of prokaryotic GH genes were analyzed based on the corresponding relationships between prokaryotic genomes and OTUs. For OTUs identified in the EMP samples, combined with the copy number of GH genes, the copy number of 16S rRNA genes, total number of genes, and cell abundance, the GH gene abundance (GH/1000) in each sample was calculated. The GH/1000 value represents the number of GH genes per thousand prokaryotic genes.

The environmental distributions of GH genes were analyzed based on the Earth Microbiome Project Ontology (EMPO), which classified 17 microbial environments (level 3) as free-living or host-associated (level 1) and saline or nonsaline (if free-living) or animal or plant (if host-associated) (level 2) [29].

On the basis of 2381 EMP samples with recorded temperature information and 1183 samples with recorded pH values, the composition and abundance of GH systems in prokaryotic communities were calculated under different temperature and pH conditions. According to environmental temperatures, the samples were classified into 5 groups: low temperature (≤10 °C), moderate low temperature (greater than10 °C and ≤ 20 °C), medium temperature (greater than20 °C and < 30 °C), moderate high temperature (≥30 °C and < 45 °C), and high temperature (≥45 °C). Each group contained no less than 87 samples. The samples were also classified into 5 groups according to the pH value of the environment: acidic (≤5), slightly acidic (≤5 and ≤6.5), neutral (>6.5 and <7.5), slightly alkaline (≥7.5 and <9) and alkaline (≥9). There were at least 115 samples in each pH group.

3. Results

3.1. Multiple evolutionary origins and highly uneven distribution of glycoside hydrolases

The CAZy database contains 171 GH families, 8 of which have been deleted. GHs are currently divided into 163 families according to sequence similarities, with 141 families including members with known crystal structures. We analyzed the CATH hierarchical classification of these GH families based on crystal structure information. Since many GHs contain multiple domains, we only analyzed the catalytic domains. The results showed that among the 141 GH families, the catalytic domains of 136 GH families each corresponded to specific homologous superfamilies (Table S1). Other families were excluded: members of families GH46, GH80 and GH133 all have two domains directly involved in catalysis, while NAD-dependent GHs of families GH4 and GH109 contain cofactor binding domains for the catalytic process.

Notably, there are 22 GH families whose crystal structures have not yet been resolved. In this study, we used the newly developed AI system [22] to model the three-dimensional structures of their representative members (Fig. 1 **and** Table S2). The three-dimensional structures showed that the catalytic domains of families GH71, GH96, GH111, GH147, GH148, GH151, GH157, GH168 and GH169 all had the classic $(\beta/\alpha)_8$ TIM-barrel fold, and the GH119 family adopted the $(\beta/\alpha)_7$-barrel fold (pseudo-TIM-barrel). The catalytic domains of families GH139, GH154 and GH161 all consisted of the $(\alpha/\alpha)_6$-barrel fold, while families GH159 and GH165 displayed the five bladed β-propeller fold, the GH122 family adopted the seven bladed β-propeller, the GH118 and GH160 families had parallel β-helix architectures, and the GH132 family showed a β-sandwich fold. Additionally, the three-dimensional structures of families GH75, GH150 and GH163 were not determinable in the CATH structural classification.

Based on information from the crystal structures and AlphaFold-modeled structures, the CATH hierarchical classifications of 155 GH families were determined, and further, the distant evolutionary correlations of these GH families were analyzed

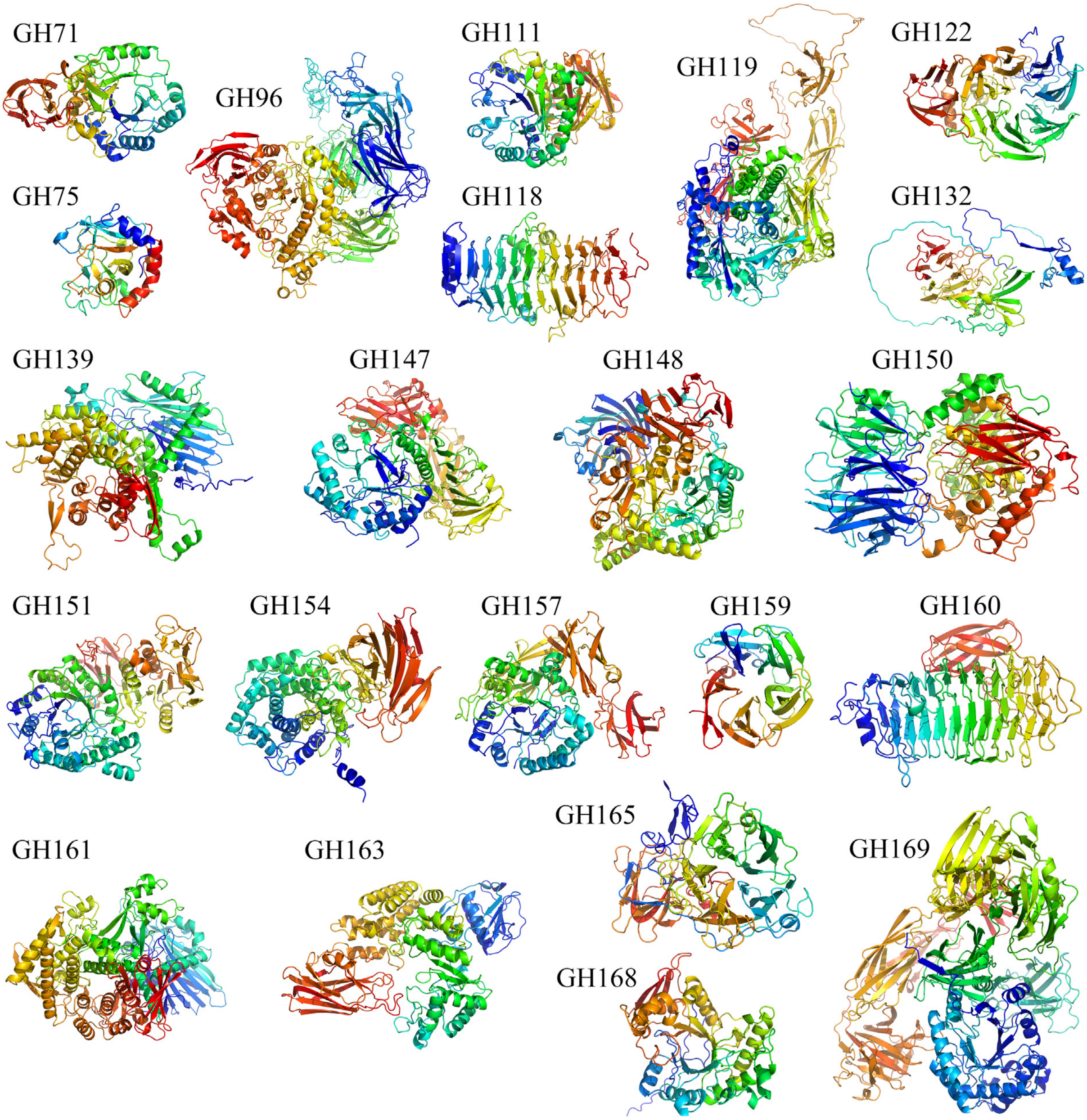

Fig. 1. Schematic diagram of representative structures of GH families. Twenty-two GH families without crystal structures are displayed in the figure. The three-dimensional structures of whole proteins were obtained by AI system modeling and are shown as cartoon models, with the coloring highlighting of N-terminals in blue and C-terminals in red. Specific information on the GHs is shown in Table S2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(Fig. 2). More than 1,000,000 known GH homologues belonged to only 27 homologous superfamilies. Remarkably, more than 65% of GH families (n = 105) were distributed in 4 homologous superfamilies: the superfamily CATH 3.20.20.80 contained up to 62 families (38.0% of the total GH families), CATH 1.50.10.10 contained 27 families, and CATH 2.160.20.10 and CATH 2.115.10.20 included 12 and 10 families, respectively (Fig. 3**A**). In addition, 9 of these superfamilies contained 2–7 GH families, and the other 14 superfamilies included only one GH family. In terms of the number of GH genes, more than 75% of the sequences were concentrated in 4 homologous superfamilies: CATH 3.20.20.80 (38.3%), CATH 1.10.530.10 (16.8%), CATH 2.120.10.10 (12.8%), and CATH 1.50.10.10 (8.3%) (Fig. 3**B**). Therefore, the distribution of GHs is extremely uneven and concentrated within only a few homologous superfamilies.

At the class level in the CATH structural hierarchy, the core structures of 40 GH families and 28.1% GH genes were mainly formed by α-helices (mainly α), 41 GH families and 22.8% GH genes were primarily composed of β-sheets (mainly β), and 74 GH families and 46.6% GH genes were constructed by a mixture of α-helices and β-sheets (mixed α-β) (Fig. S1). Folds can reflect how the secondary structure elements are connected and arranged [21]. The 27 homologous superfamilies belonged to 16 different folds (Fig. 2), with up to 69 GH families and 44.3% GH genes exhibiting TIM barrels (CATH 3.20.20, $(\beta/\alpha)_8$) as core structures. More than 85% of GH families and over 90% of GH genes were concentrated in 7 types of folds (Fig. 3**C and** Fig. 3**D**). In conclusion, most GHs belong to only a few folds, and the distribution of enzymes among these folds is highly uneven.

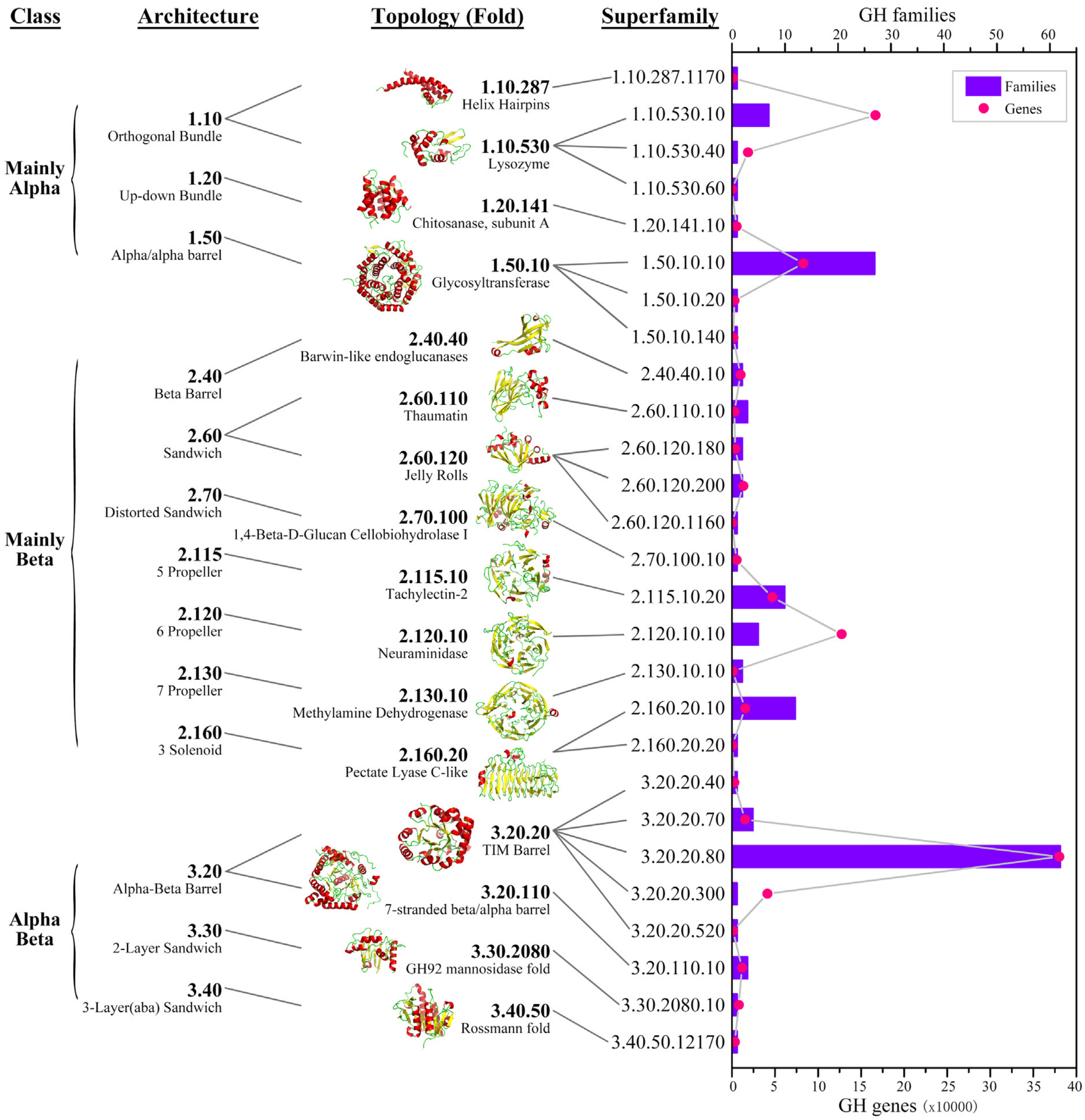

Fig. 2. Homologous superfamily associated with glycoside hydrolases in the CATH hierarchy. The four main levels of CATH hierarchical classification are the protein class, architecture, topology/fold, and homologous superfamily. Histograms and line charts show the numbers of GH families and genes in each superfamily, respectively. The typical structure of each fold is displayed as a cartoon model.

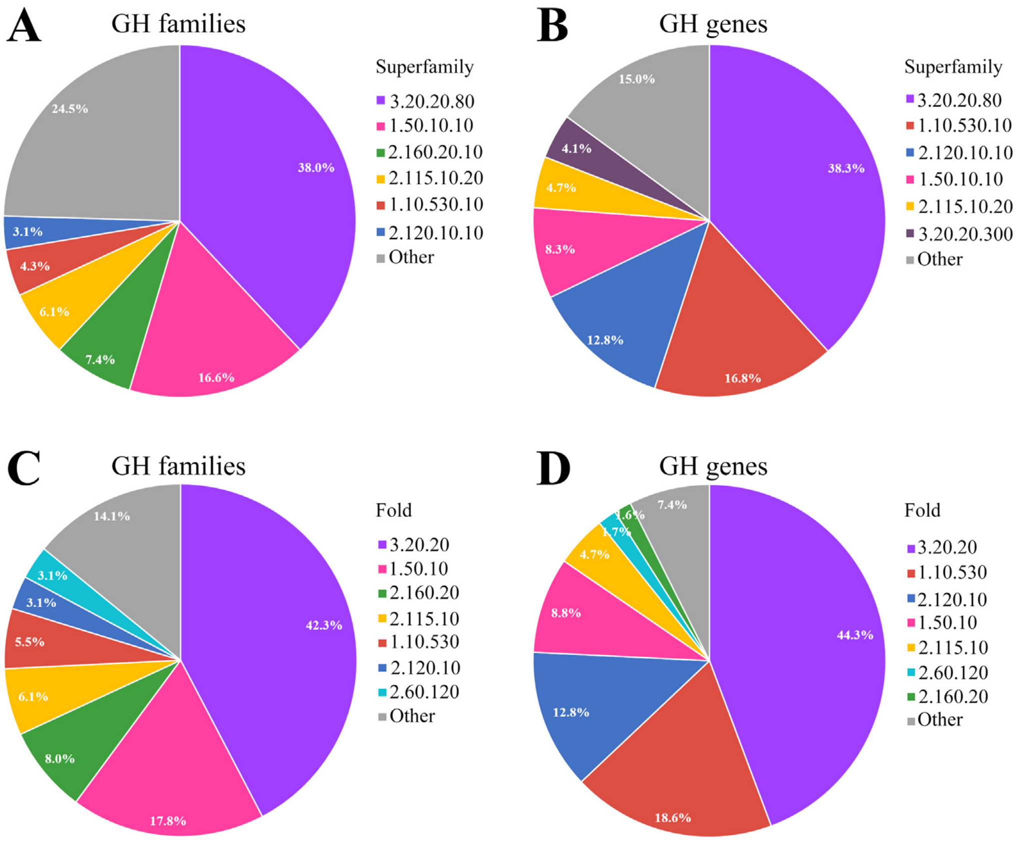

Fig. 3. Distributions of known glycoside hydrolases in homologous superfamilies and folds. The distribution of GH families (A) and GH genes (B) in homologous superfamilies are shown. The distribution of GH families (C) and GH genes (D) in folds are also shown.

It should be noted that in the CAZy database, 69 GH families were combined into 18 clans at a higher hierarchical level. Enzymes of the same clan have a common evolutionary origin of their genes and share the most important functional characteristics, such as the composition of the active center, anomeric configuration of cleaved glycosidic bonds, and molecular mechanism of the catalyzed reaction (inverting or retaining) [12,33]. We found that the 18 clans were clustered into 8 CATH superfamilies, of which the superfamily CATH 1.50.10.10 contained 6 clans, CATH 3.20.20.80 encompassed 5 clans, CATH 2.115.10.20 included 2 clans, and the other 5 superfamilies each contained a single clan (Table S1). Overall, there were 87 GH families that did not belong to any clan, but fell into CATH homologous superfamilies. In addition, more than 20,000 GH sequences have not been classified (grouped in GH0), among which 11 proteins have known crystal structures belonging to the superfamilies mentioned above.

3.2. Most GH superfamilies are multifunctional, and a few are extremely versatile

<1% of GHs have been functionally identified, but more than 150 types of glycosidase activities (EC 3.2.1.x) have been discovered to date [11]. Over two-thirds of the superfamilies (n = 20) had at least 2 glycosidase activities, among which 6 superfamilies were associated with at least 10 glycosidase activities (Fig. 4**A**). In particular, the superfamily CATH 3.20.20.80 was related to as many as 107 types of glycosidase activities, occupying 70.9% of the total. Phosphorylases are a group of carbohydrate-active enzymes to cleave glycosidic bonds using phosphate as a nucleophile. GH families contain diverse phosphorylases. We found that phosphorylases related to GH families were mainly located in three superfamilies of CATH 3.20.20.80, CATH 1.50.10.10 and CATH 2.115.10.20 (**Fig. S2**), and these three superfamilies also included the most types of glycosidase activities. Therefore, most superfamilies are multifunctional in terms of glycosidase activities and phosphorylase activities, and a few are extremely versatile. However, it is important to note that because an increasing number of novel enzymatic activities are being revealed, we probably still underestimate the functional diversity of GHs.

Subsequent analysis showed a positive correlation between GH sequence diversity and functional diversity in homologous superfamilies, with a higher sequence diversity yielding a higher functional diversity ($r = 0.745$, $p < 0.001$) (Fig. 4**B**). The three CATH structural classes (mainly α, mainly β, and mixed α-β) all had the potential to evolve high GH sequence diversity and functional diversity. Moreover, compared with other superfamilies, the superfamilies CATH 3.20.20.80 with the $(\beta/\alpha)_8$-barrel fold and CATH 1.50.10.10 with the $(\alpha/\alpha)_6$-barrel fold exhibited particularly outstanding GH sequence diversity and functional diversity.

Remarkably, the hydrolysis of glycosidic bonds is catalyzed by two amino acid residues in the enzyme: a general acid (proton donor) and a nucleophile/base in most cases [34]. Depending on the spatial positions of these catalytic residues, hydrolysis occurs via overall retention or overall inversion of the anomeric configuration. In this study, we found that there were at least 8 superfamilies that could perform both inverting and retaining activities (Table S1). Therefore, diversity in catalytic mechanisms (inverting and retaining) is common in homologous superfamilies, and the same is true for the three classes.

In addition, 55 glycosidase activities existed in two or more superfamilies, accounting for 36.4% of the total (Fig. 5**A and** Table S3). The results suggested that many GHs convergently

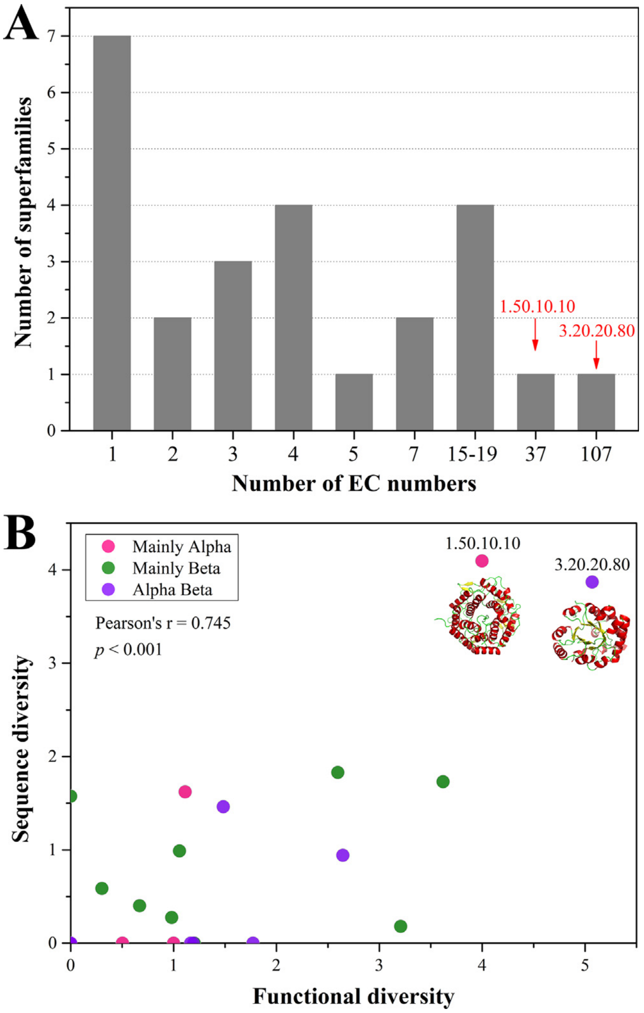

Fig. 4. Diversification of glycoside activities in homologous superfamilies. (A) Distribution of the number of different enzymatic functions associated with each superfamily, as indicated by the number of different EC numbers (EC 3.2.1.x). (B) Positive correlation between GH sequence diversity and functional diversity in homologous superfamilies. The GH sequence diversity and functional diversity were calculated based on the Shannon index. The color of each point reflects the first-level CATH class as follows: pink (mainly α), green (mainly β), and purple (mixed α-β). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

evolved toward similar enzymatic functions, even though they descended from multiple evolutionary origins. In particular, EC 3.2.1.4 (cellulases) and EC 3.2.1.8 (xylanases) both existed in as many as 8 different superfamilies (Fig. 5**B**). More than 95% of xylanases were distributed in the superfamilies CATH 3.20.20.80 and CATH 2.60.120.180, while they were rare in the other 6 superfamilies. However, the proportions of cellulases from the superfamilies CATH 3.20.20.80, CATH 2.40.40.10, CATH 2.60.120.180, and CATH 1.50.10.10 all exceeded 5%, suggesting that glycosidase activity could be generated by the convergent evolution of three distinct structural classes (mainly α, mainly β, and mixed α-β). Notably, in addition to cellulases and xylanases, other GHs associated with multiple evolutionary origins also mainly target polysaccharides such as xyloglucan, chitosan, and lichenan. These findings further suggested that, compared with other glycosidase activities, degradation activities for those particularly abundant and recalcitrant substrates are more likely to evolve independently in different superfamilies. However, unlike glycosidase activities, almost each phosphorylase activity occurs only in a single superfamily.

3.3. Glycoside hydrolases in a few superfamilies can be widely distributed in life domains

Based on the genomes annotated by the CAZy database, the proportions of members in each GH family among different life domains were analyzed (**Fig. S3**). The results showed that 156, 106, and 74 GH families existed in the bacterial, eukaryotic, and archaeal genomes, respectively, whereas only 27 families appeared in viral genomes. Notably, no single GH family accounted for more than 10% of the genomes in each of the four life domains at the same time. Only 9 families (GH1, GH2, GH3, GH5, GH13, GH15, GH31, GH36, and GH38) coexisted in the genomes of bacteria, eukaryotes and archaea, while the GH18 family was present in bacteria, eukaryotes and viruses. In addition, 42.1% of GH genes in bacteria, 29.7% in archaea, 80.6% in eukaryotes, and 2.2% in viruses have not yet been assigned to a GH family (GH0).

For the 27 homologous superfamilies containing GHs, members of 25, 23, 18, and 11 superfamilies were present in the genomes of bacteria, eukaryotes, archaea, and viruses, respectively (Fig. 6**A and** Table S4). Only the GHs of CATH 3.20.20.80, the superfamily with the most diverse glycosidase activities, had an occurrence frequency of more than 30% of the genomes in all four life domains. The TIM barrel (CATH 3.20.20, $(\beta/\alpha)_8$) is regarded as one of the oldest folds and is traceable to the last universal common ancestor (LUCA) [35,36]. In addition, GHs of 5 superfamilies were codistributed in bacterial, eukaryotic, and archaeal genomes (at least 10%), among which core structures of 3 superfamilies were α/β barrels (CATH 3.20), while the other 2 superfamilies were α/α barrels (CATH 1.50) and β propellers (CATH 2.115). Furthermore, in bacteria, GHs of 23, 24, 21, 18, and 24 superfamilies were found in the FCB group, Proteobacteria, PVC group, Spirochaetes, and Terrabacteria group, respectively, among which 7 superfamilies each accounted for more than 10% of the five phyla (Fig. 6**B and** Table S4). In eukaryotes, we found that 23, 15, and 14 GH superfamilies appeared in the genomes of fungi, animals, and plants, respectively. Among these superfamilies, 10 superfamilies each accounted for more than 10% of the three taxa.

Among the 21,244 bacterial genomes annotated by the CAZy database, an average of 37.2 GH genes were encoded in each genome (Fig. 7**A**). Correspondingly, the 352 eukaryotic genomes encoded up to 101.3 GH genes on average. Comparably, the average numbers of GH genes encoded by the 424 archaeal genomes and 456 viral genomes were only 12.4 and 2.0, respectively. With respect to the number of families, the average number of GH families encoded by single eukaryotic, bacterial, archaeal, and viral genomes was 25.4, 16.7, 6.9, and 1.5, respectively (Fig. 7**B**). In terms of the number of superfamilies, the average values were 9.4, 6.7, 3.7, and 1.2 in the corresponding life domains (Fig. 7**C**). In short, the numbers of GH genes, families, and superfamilies encoded by a single genome were the largest in eukaryotes, followed by bacteria, and both were significantly higher than those of archaea and viruses. Meanwhile, in view of the number of GH genes, families, and homologous superfamilies encoded by a single genome, the disparity between eukaryotes and prokaryotes trended to be smaller. Furthermore, in bacteria, the FCB group encoded the most GH genes, families and superfamilies, while Spirochaetes encoded the fewest. In eukaryotes, a plant genome encoded nearly 400 GH genes on average, while that in animals or fungi were only approximately 100. However, the average number of fungal-encoded superfamilies was higher than that of plant-encoded superfamilies. Overall, the difference in GH gene numbers between eukaryotic and prokaryotic genomes was probably mainly due to various copy numbers of paralogous genes in the superfamilies rather than the number of superfamilies.

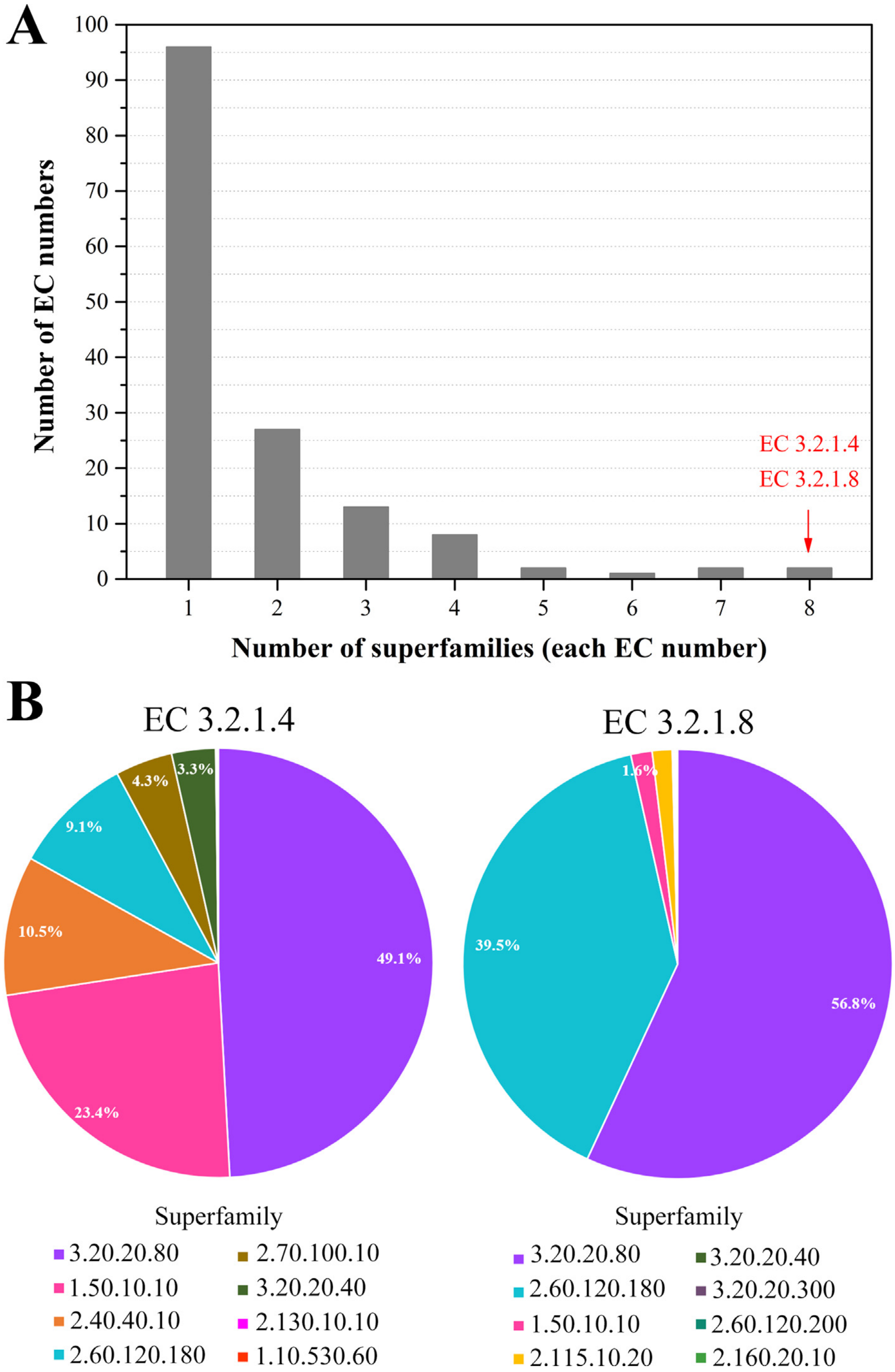

Fig. 5. Convergent evolution of glycoside activities. (A) Distribution of the number of different enzymatic structures associated with each EC number, as indicated by the number of different superfamilies. (B) Distributions of cellulases (EC 3.2.1.4) and xylanases (EC 3.2.1.8) in diverse homologous superfamilies, shown in the form of pie charts.

3.4. Glycoside hydrolases of most superfamilies have specific environmental distributions

Prokaryotes are the main source of GHs on Earth. Among over 1,000,000 known GH genes, 80.2% are from bacteria and archaea. The EMP samples the Earth's microbial communities at an unprecedented scale to evaluate the distribution of prokaryotes in global environments [29]. Our previous work showed that the sequenced proportion of the global prokaryotic genomes reached a high level, i.e. the median proportions of genome-sequenced cells and taxa (at 100% identities in the 16S-V4 region) in different biomes reached 38.1% (16.4–86.3%) and 18.8% (9.1–52.6%), respectively [37]. The EMPO classifies microbial environments as free-living or host-associated, with further subdivision into 17 environmental types [29]. Here, by large-scale sequence alignments between the 10,000 samples released by the EMP and approximately 20,000 fully sequenced prokaryotic genomes, we analyzed the environmental distributions of GHs from diverse superfamilies (Fig. 8**A**).

Our results showed that a total of 23 homologous superfamilies were found in EMP prokaryotic communities, among which the superfamily CATH 3.20.20.80 had the highest GH gene abundances

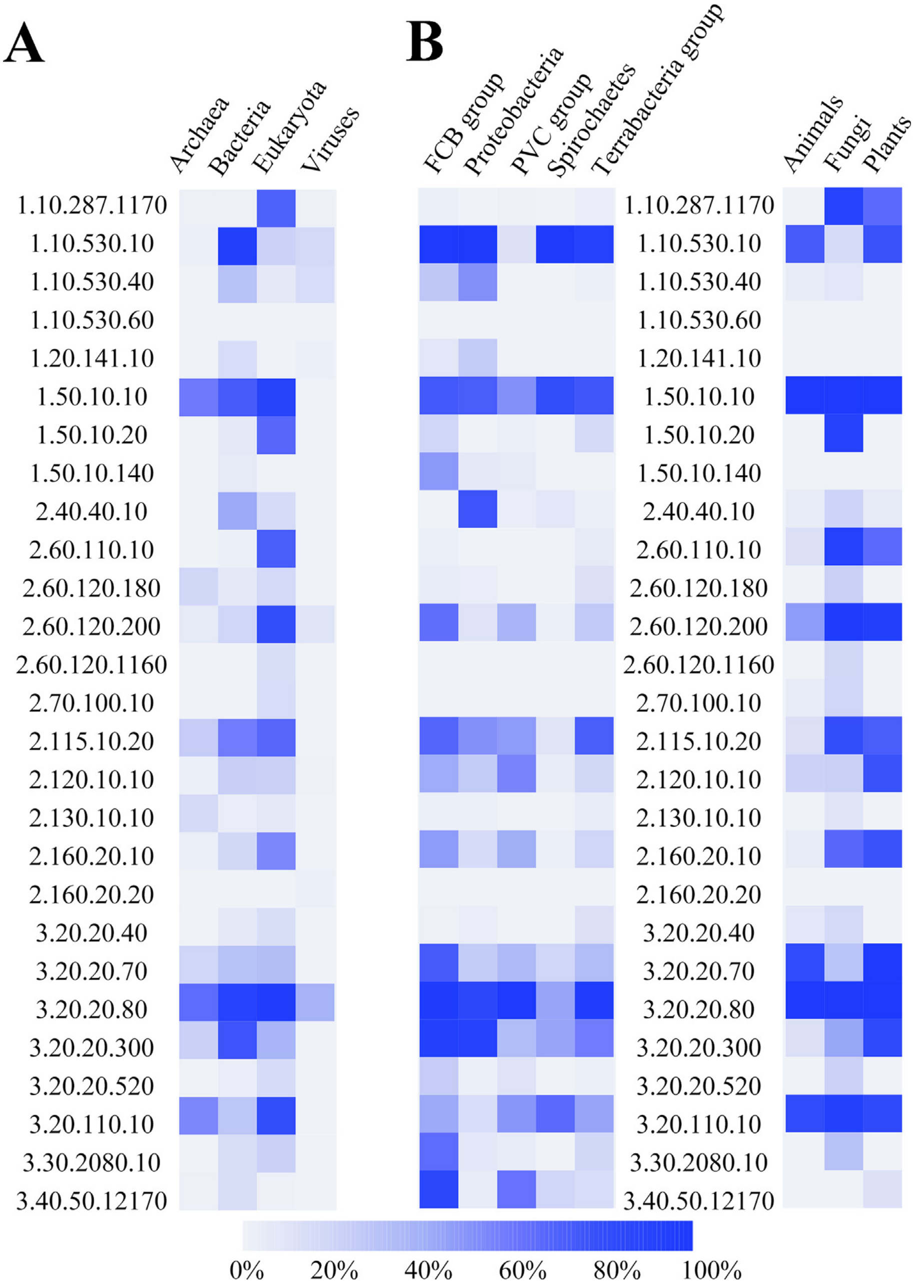

Fig. 6. Distributions of glycoside hydrolases from different homologous superfamilies in life domains. (A) Distributions of GHs from homologous superfamilies in different life domains. Colors ranging from white to blue represent occurrence frequency values from low to high. (B) Distributions of GH superfamilies in the genomes of three taxa in eukaryotes and five phyla in bacteria. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in all 17 environments, especially in animal proximal gut and animal distal gut samples, with the GH/1000 medians of 8.81 and 7.96, respectively. The GH gene abundances of the superfamily CATH 3.20.20.80 in free-living environments, particularly saline environments, were lower than those in host-associated environments. The GH/1000 medians in the sediment (saline) and water (saline) samples were 2.66 and 2.52, respectively, and the lowest was 2.01 in the hypersaline (saline) samples. The second most abundant superfamily was CATH 1.10.530.10, whose GH/1000 medians exceeded 1 in all 17 environments, with minor differences ranging from 2.23 in animal corpus samples to 1.25 in plant surface samples. The CATH 1.50.10.10 and CATH 2.115.10.20 superfamilies were also abundant in environments, while the GH/1000 medians of all the other 21 superfamilies were below 1 in all 17 environments. In general, superfamilies that are widely distributed in prokaryotic genomes are also extensively present in different prokaryotic communities.

Of the 23 superfamilies that were revealed in the EMP prokaryotic communities, the highest GH gene abundances of 14 superfamilies appeared in animal host-associated environments: 7 in the animal proximal gut environment, 4 in the animal distal gut environment, and 3 in the animal corpus environment. Similarly, 9 superfamilies exhibited the highest abundances in plant host-associated environments: 5 in the plant rhizosphere environment, 3 in the plant surface environment, and 1 in the plant corpus environment (Fig. 8A). Notably, no superfamily had the highest abundance in free-living environments. In particular, in saline-, water-, and surface-environments, no more than 4 superfamilies reached 50% of the highest abundances in other environments. From a superfamily perspective, more than 65% of superfamilies

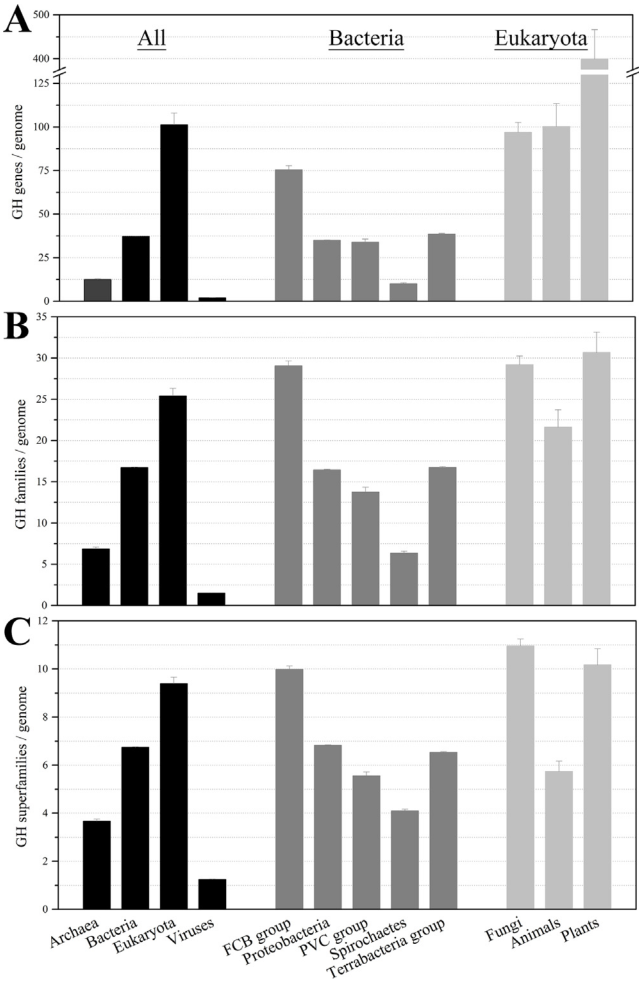

Fig. 7. Statistics of the numbers of glycoside hydrolase genes (A), families (B), and homologous superfamilies (C) in a single genome. Statistical analyses were performed using the average and standard deviation in the following order: four life domains (archaea, bacteria, eukaryotes, and viruses), bacteria (FCB group, Proteobacteria, PVC group, Spirochaetes, and Terrabacteria group), and eukaryotes (fungi, animals, and plants).

(n = 15) maintained 50% of the highest abundance in fewer than 5 environments. The widely distributed superfamilies were mainly concentrated in 3 different folds (CATH 1.10.530, CATH 3.20.20, and CATH 1.50.10), of which only CATH 1.10.530.10 maintained more than 50% of the highest abundance in all 17 environments. Therefore, most superfamilies exhibit narrow environmental distributions, and their highest abundances are concentrated in host-associated environments.

Furthermore, we analyzed the effects of environmental temperatures and pH values on the distributions of GHs from different superfamilies (Fig. 8**B**). A total of 22 superfamilies were detected based on 2381 EMP samples with temperature information and 1183 samples with pH values. According to environmental temperatures, samples were categorized into 5 groups: low temperature (≤10 °C), moderate low temperature (>10 °C and ≤20 °C), medium temperature (>20 °C and <30 °C), moderate high temperature (≥30 °C and <45 °C), and high temperature (≥45 °C). The results showed that 21 superfamilies exhibited the highest abundances in low- or moderate low-temperature environments, whereas only the highest abundance of superfamily CATH 3.20.110.10 appeared in the high-temperature environment. Interestingly, only 4 superfamilies that contain the α/β barrels (CATH 3.20) core structures maintained at least 80% of their highest abundances in the high-temperature environment. Among GHs with mainly α structures, only 2 superfamilies retained more than 40% of their highest abundances in the high-temperature environment. However, no superfamily with mainly β structures maintained more than 15% of its highest abundance in the high-temperature environment. Hence, GHs of the mixed α-β structures maintain the best stability in the high-temperature environment, while GHs with mainly β structures are the weakest.

Based on environmental pH values, the samples were also classified into 5 groups for analysis: acidic (≤5), slightly acidic (greater than5 and ≤ 6.5), neutral (greater than6.5 and < 7.5), slightly alkaline (≥7.5 and < 9) and alkaline (≥9). The highest abundances of 15 superfamilies were observed in the acidic environment, while the superfamily CATH 3.20.110.10 had the highest abundance in the alkaline environment (Fig. 8**C**). Furthermore, a total of 10 superfamilies maintained at least 50% of their highest abundances in the alkaline environment, of which 4 had a mixed α-β structure as the core structure, 4 possessed a mainly α structure, and 2 had a mainly β structure. Therefore, the abundances of superfamilies associated with GHs were less sensitive to environmental pH values than temperatures. Interestingly, the superfamily CATH 3.20.110.10 was the single group with a distribution preference for high-temperature and alkaline environments. The fold of CATH 3.20.110.10 is 7-stranded β/α barrels, including the families GH38, GH57, and GH119, along with diverse glycosidase functions.

4. Discussion

GHs play multiple key roles in nature and have many applications in health, nutrition, and biotechnology [18,38–40]. In this study, we analyzed the classification of GHs at the homologous superfamily and fold levels based on protein structure information. Our work not only analyzed the crystal structures of GHs but also modeled the structures of 22 GH families without crystal structures. Structural modeling was performed using the newly developed AlphaFold algorithm, which achieves an accuracy competitive with that of experimental results [22,27]. This allowed our analysis to cover all 163 GH families, at a level different from and complementary to the 18 existing clans that cover 69 GH families in the CAZy database.

GHs include a huge sequence space, with more than 1,000,000 known members that have been divided into more than 160 families based on sequence similarities [11]. Due to the large variety of possible linkages, the diversity of monosaccharides, the modification of carbohydrates by acetylation or sulfation, and the form and length adopted by polysaccharides, glycans provide the widest diversity of all biomolecules [12]. Therefore, glycosidase activities are also extremely diverse, with at least 151 classified EC numbers (EC 3.2.1.x). However, structurally, we found that known GHs were from at least 27 homologous superfamilies and 16 folds, among which the distributions of GH sequences and enzymatic activities were highly uneven; a few homologous superfamilies and folds were extremely abundant, but most were rare.

A few superfamilies and folds exhibited extremely high versatility of glycosidase activities, mainly the α-β barrel (CATH 3.20), α/α barrel (CATH 1.50), and β propeller (CATH 2.115, CATH 2.120, and CATH 2.130). Although these structures were unrelated and quite distinct from one another, certain general structural features that allowed them to accommodate diverse glycosidase catalytic activities could be discerned (Fig. 2). The α/β-barrel structure includes $(\beta/\alpha)_8$- and $(\beta/\alpha)_7$- barrel folds, in which β-strands and α-helices alternate along the protein sequence, with β-strands forming the inner barrel and α-helices flanking as the exterior. The α/α-barrel structure consists of 6 parallel α-helices forming the inner barrel, which are flanked by 6 external α-helices. The β-propeller structure is composed completely of β-sheets and is characterized by

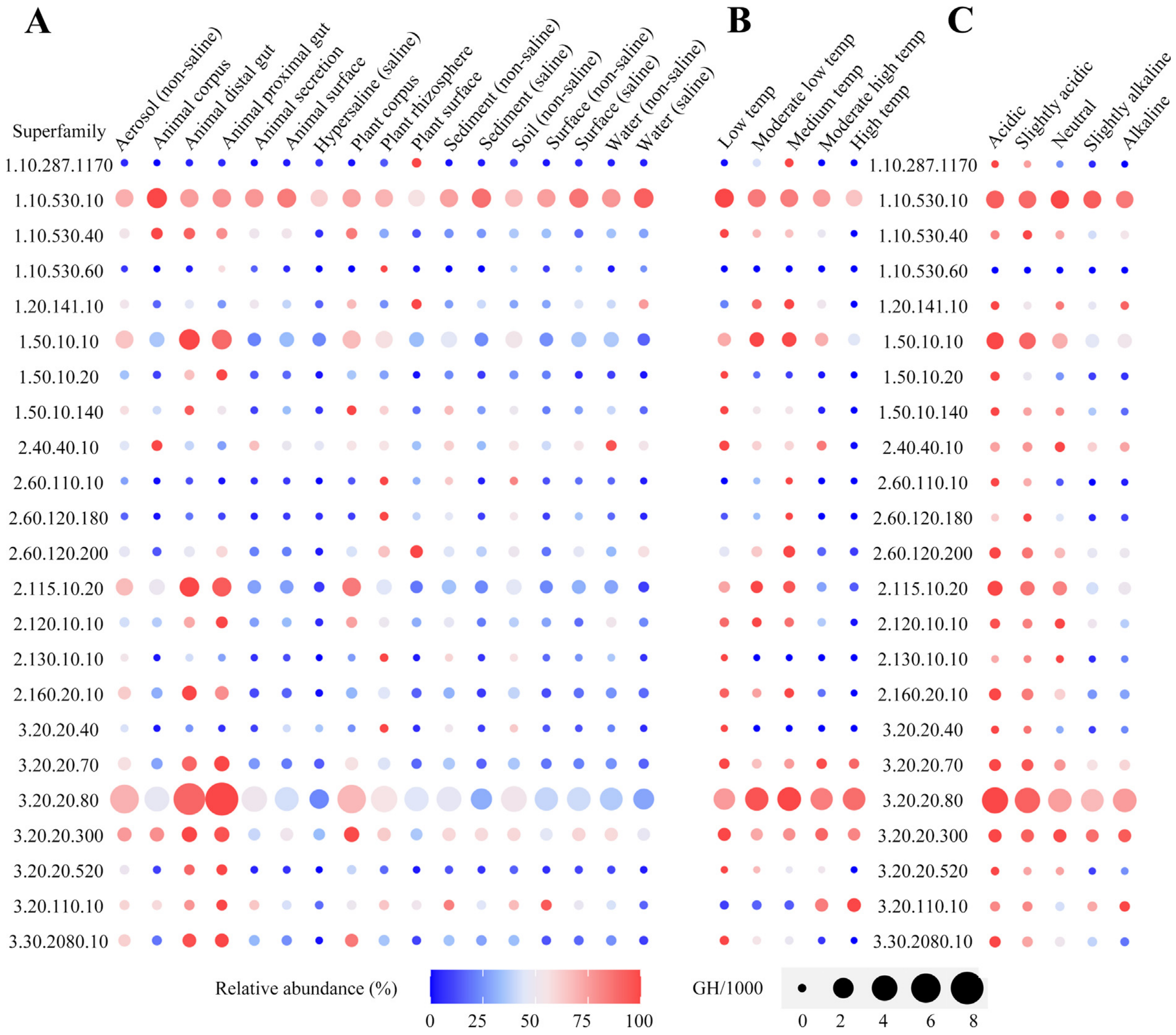

Fig. 8. Environmental distributions of glycoside hydrolases from homologous superfamilies. (A) Distributions of GHs from homologous superfamilies in 17 EMPO environments. (B) Environmental temperatures affect the distributions of GHs from homologous superfamilies. According to environmental temperatures, samples were classified into 5 groups: low temperature (≤10 °C), moderate low temperature (>10 °C and ≤ 20 °C), medium temperature (>20 °C and < 30 °C), moderate high temperature (≥30 °C and < 45 °C), and high temperature (≥45 °C). Each group contained no<87 samples. (C) Influence of environmental pH on the distributions of GHs from homologous superfamilies. Samples were also classified into 5 groups according to the pH value of the environments for analysis: acidic (≤5), slightly acidic (>5 and ≤ 6.5), neutral (>6.5 and < 7.5), slightly alkaline (≥7.5 and < 9) and alkaline (≥9). There were at least 115 samples in each group. The circle size represents the GH gene abundance, which is displayed as the GH/1000 value. Colors from blue to red represent the relative abundance of GHs in each superfamily from low to high, with the highest GH/1000 value in the environment defined as 100%. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5–7 sets of blade-shaped sheets arranged symmetrically around the central axis. These structures all contain a central pocket that can bind to their substrates, with approximately cyclic symmetry. The inherent pockets in these structures can accommodate various substrate molecules through low-specificity interactions. Furthermore, the intrinsic symmetry of the central pockets offers the potential for different catalytic residues to appear on the surface of the substrate-binding site, enabling the evolution of multiple activities.

Previous observations have hinted that functional characterization by structure determination is not straightforward, and the mechanisms generating functional diversity during evolution are complex [41,42]. For example, the active site residues of TIM-barrel enzymes are distributed at the eight βα motifs [43]. The N/C-terminal and loop regions on TIM barrel proteins are capable of hosting structural inserts ranging from simple secondary structural motifs to complete domains [44]. We found that the GH sequence diversity was positively correlated with its functional diversity in homologous superfamilies, indicating no contradiction between the robustness of protein structures that tolerate sequence variations and the innovation that drives the acquisition of new functions. The $(\beta/\alpha)_8$-barrel and $(\alpha/\alpha)_6$-barrel structures exhibit particularly high diversity in sequence, and the core structures of these proteins are large, containing many secondary structural elements with many interactions and high contact density. Thus, they are highly robust to mutations. Moreover, GHs with these structures are also highly resistant to high temperatures. Most superfamilies of $(\beta/\alpha)_8$-barrel GHs could maintain at least 80% of their highest gene abundances in the high-temperature environment (≥45 °C), and superfamilies with the $(\alpha/\alpha)_6$-barrel structure preserved at least 60%, both of which were much higher than those of other folds. Thus, the combination of rigid core structures and flexible central pockets on the surfaces of the $(\beta/\alpha)_8$-barrel and $(\alpha/\alpha)_6$-barrel structures enables enzymatic activities. Robustness to point mutations and insertions/deletions results in high sequence diversity and a large number of variations, especially in functional regions, which provides new catalytic mechanisms and promotes the production of high functional diversity.

Interestingly, GHs also showed apparently convergent evolution. More than one-third of GH activities existed in several homologous superfamilies, indicating multiple evolutionary origins. In addition, we found that the degradation functions of polysaccharides, such as cellulose and xylan, were more likely than other functions to be obtained through convergent evolution. For instance, EC 3.2.1.4 (cellulases) and EC 3.2.1.8 (xylanases) both occurred in at least 8 different homologous superfamilies. Cellulose, as the most abundant polysaccharide on Earth, and hemicellulose with xylan as the main component together constitute plant cell wall material [45,46]. Therefore, we speculate that the large biomass and high degradation difficulty might be the reasons for the widespread multiorigin convergent evolution of the degradation activities of polysaccharides.

Altogether, the diversity of carbohydrate substrates promotes the evolution of diverse GH activities. Nevertheless, the evolutionary mechanisms are complex, including divergent evolution from a versatile structure to acquire new specificities and convergent evolution of different structures toward similar catalytic mechanisms. These findings not only increase our understanding of the sequence–structure–function relationships and evolution of GHs but also may be useful in designs and modifications in future protein engineering of GHs.

CRediT authorship contribution statement

Dan-dan Li: Conceptualization, Methodology, Investigation, Visualization, Writing – original draft, Validation. **Jin-lan Wang:** Resources, Validation. **Ya Liu:** Data curation, Validation. **Yue-zhong Li:** Conceptualization, Writing – review & editing, Project administration, Funding acquisition. **Zheng Zhang:** Conceptualization, Supervision, Writing – review & editing, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Gregory Poole and Qianqian Ma for useful discussion and suggestions. The scientific calculations in this paper were performed on the HPC Cloud Platform at Shandong University. This work was financially supported by the National Natural Science Foundation of China (32070030), the Special Investigation on Scientific and Technological Basic Resources (2017FY100300), the National Key Research and Development Program (2018YFA0900400 and 2018YFA0901704), the Key Research and Development Program of Shandong Province (2018GSF121015) (YZL), the Natural Science Foundation of Jiangsu Province (BK20190199), and the Fundamental Research Funds of Shandong University (2020GN113) (ZZ).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.10.039.

References

[1] Romero-Romero S, Kordes S, Michel F, Höcker B. Evolution, folding, and design of TIM barrels and related proteins. Curr Opin Struct Biol 2021;68:94–104.

[2] Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature 2002;420(6912):218–23.

[3] Kolodny R, Pereyaslavets L, Samson AO, Levitt M. On the universe of protein folds. Annu Rev Biophys 2013;42(1):559–82.

[4] Zhang Z, Wang Y, Wang L, Gao P, Kolokotronis S-O. The combined effects of amino acid substitutions and indels on the evolution of structure within protein families. PLoS ONE 2010;5(12):e14316.

[5] Zhang Z, Huang J, Wang Z, Wang L, Gao P. Impact of indels on the flanking regions in structural domains. Molecular biology and evolution 2011;28:291-301.

[6] Zhang Z, Wang J, Gong Y, Li Y. Contributions of substitutions and indels to the structural variations in ancient protein superfamilies. BMC Genomics 2018;19:771.

[7] Zhang Z, Xing C, Wang L, Gong B, Liu H. IndelFR: a database of indels in protein structures and their flanking regions. Nucleic acids research 2012;40:D512-8.

[8] Anantharaman V, Aravind L, Koonin EV. Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. Curr Opin Chem Biol 2003;7(1):12–20.

[9] Sillitoe I, Dawson N, Thornton J, Orengo C. The history of the CATH structural classification of protein domains. Biochimie 2015;119:209–17.

[10] Osadchy M, Kolodny R. Maps of protein structure space reveal a fundamental relationship between protein structure and function. PNAS 2011;108(30):12301–6.

[11] Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 2014;42(D1):D490–5.

[12] Garron M-L, Henrissat B. The continuing expansion of CAZymes and their families. Curr Opin Chem Biol 2019;53:82–7.

[13] Himmel ME, Ding S-Y, Johnson DK, Adney WS, Nimlos MR, Brady JW, et al. Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science 2007;315(5813):804–7.

[14] Bomble YJ, Lin C-Y, Amore A, Wei H, Holwerda EK, Ciesielski PN, et al. Lignocellulose deconstruction in the biosphere. Curr Opin Chem Biol 2017;41:61–70.

[15] Bardgett RD, Freeman C, Ostle NJ. Microbial contributions to climate change through carbon cycle feedbacks. The ISME journal 2008;2(8):805–14.

[16] Ragauskas AJ, Williams CK, Davison BH, Britovsek G, Cairney J, Eckert CA, et al. The path forward for biofuels and biomaterials. Science 2006;311(5760):484–9.

[17] Schubert C. Can biofuels finally take center stage? Nat Biotechnol 2006;24(7):777–84.

[18] López-Mondéjar R, Algora C, Baldrian P. Lignocellulolytic systems of soil bacteria: A vast and diverse toolbox for biotechnological conversion processes. Biotechnol Adv 2019;37(6):107374. https://doi.org/10.1016/j.biotechadv.2019.03.013.

[19] Wilson DB. Cellulases and biofuels. Curr Opin Biotechnol 2009;20(3):295–9.

[20] Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic acids research 2020;48:D376-D82.

[21] Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic acids research 2021;49:D266-D73.

[22] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9.

[23] Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 2009;37:D233–8.

[24] Cornish-Bowden A. Current IUBMB recommendations on enzyme nomenclature and kinetics. Perspect Sci 2014;1(1-6):74–87.

[25] Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, et al. GenBank. Nucleic acids research 2021;49:D92-D6.

[26] Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database : the journal of biological databases and curation 2020;2020.

[27] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature 2021;596(7873):590–6.

[28] Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal 1948;27:379-423.

[29] Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 2017;551(7681):457–63.

[30] Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems 2017;2.

[31] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinf 2009;10(1). https://doi.org/10.1186/1471-2105-10-421.

[32] Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, et al. BLAST: a more efficient report with usability improvements. Nucleic acids research 2013;41:W29-33.

[33] Naumoff DG. Hierarchical classification of glycoside hydrolases. Biochemistry Biokhimiia 2011;76(6):622–35.

[34] Davies G, Henrissat B. Structures and mechanisms of glycosyl hydrolases. Structure 1995;3(9):853–9.

[35] Copley RR, Bork P. Homology among $(\beta/\alpha)_8$ barrels: implications for the evolution of metabolic pathways. J Mol Biol 2000;303:627–41.

[36] Sterner R, Hocker B. Catalytic versatility, stability, and evolution of the $(\beta/\alpha)_8$-barrel enzyme fold. Chem Rev 2005;105:4038–55.

[37] Zhang Z, Wang J, Wang J, Wang J, Li Y. Estimate of the sequenced proportion of the global prokaryotic genome. Microbiome 2020;8(1). https://doi.org/10.1186/s40168-020-00903-z10.21203/rs.3.rs-106686/v1.

[38] Elleuche S, Schäfers C, Blank S, Schröder C, Antranikian G. Exploration of extremophiles for high temperature biotechnological processes. Curr Opin Microbiol 2015;25:113–9.

[39] Lopez-Mondejar R, Zuhlke D, Vetrovsky T, Becher D, Riedel K, Baldrian P. Decoding the complete arsenal for cellulose and hemicellulose deconstruction in the highly efficient cellulose decomposer Paenibacillus O199. Biotechnol Biofuels 2016;9:104.

[40] Tiwari R, Nain L, Labrou NE, Shukla P. Bioprospecting of functional cellulases from metagenome for second generation biofuel production: a review. Crit Rev Microbiol 2018;44(2):244–57.

[41] Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001;307:1113–43.

[42] Scossa F, Fernie AR. The evolution of metabolism: How to test evolutionary hypotheses at the genomic level. Comput Struct Biotechnol J 2020;18:482–500.

[43] Nagano N, Orengo CA, Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 2002;321(5):741–65.

[44] Nagarajan D, Deka G, Rao M. Design of symmetric TIM barrel proteins from first principles. BMC Biochem 2015;16:18.

[45] Gao Y, Lipton AS, Wittmer Y, Murray DT, Mortimer JC. A grass-specific cellulose-xylan interaction dominates in sorghum secondary cell walls. Nat Commun 2020;11:6081.

[46] Scheller HV, Ulvskov P. Hemicelluloses. Annu Rev Plant Biol 2010;61(1):263–89.