Chapter 6
The CAZy Database/the Carbohydrate-Active Enzyme (CAZy) Database: Principles and Usage Guidelines

Nicolas Terrapon, Vincent Lombard, Elodie Drula, Pedro M. Coutinho, and Bernard Henrissat

Keywords Carbohydrate-active enzymes • Protein domains • Family classification • Structural and functional annotation • Sequence • genome • and metagenome analysis • Polysaccharide Utilization Loci

Carbohydrate-Active enZymes (CAZymes) assemble, breakdown, and modify glycans and glycoconjugates using their catalytic and binding modules (functional protein domains). The CAZy database offers since 1998 an online and continuously updated classification of CAZyme modules (Lombard et al. 2014). Each module family in the CAZy classification has been created based on experimentally characterized protein modules from the literature, and the families are populated by related module sequences from public protein sequence databases. Since no universal threshold allows the systematic classification of the various CAZyme families, CAZy annotations result from an expert combination of module modeling/calibration and human curation. CAZy annotations are made publicly available for all proteins released by GenBank (Benson et al. 2012), Swiss-Prot (Boutet et al. 2016) and the Protein Data Bank (PDB; http://www.rcsb.org; (Berman et al. 2000)).

N. Terrapon • V. Lombard • P.M. Coutinho
Aix-Marseille Université, AFMB UMR 7257, 163 Avenue de Luminy, 13288 Marseille, France

Centre National de la Recherche Scientifique, AFMB UMR 7257, 163 Avenue de Luminy, 13288 Marseille, France

E. Drula
Aix-Marseille Université, AFMB UMR 7257, 163 Avenue de Luminy, 13288 Marseille, France

Institut National de la Recherche Agronomique, BBF UMR 1163, Polytech Marseille, 163 Avenue de Luminy, 13288 Marseille, France

B. Henrissat (✉)
Centre National de la Recherche Scientifique, AFMB UMR 7257, 163 Avenue de Luminy, 13288 Marseille, France

Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: Bernard.Henrissat@afmb.univ-mrs.fr

K.F. Aoki-Kinoshita (ed.), *A Practical Guide to Using Glycomics Databases*,
DOI 10.1007/978-4-431-56454-6_6

Further, functional and 3-D structural information, curated from the literature on a regular basis, constitute essential added values to the CAZy annotation. In this spirit, the display of ligand information from crystallographic complexes has been recently developed (Lombard et al. 2014). This chapter will guide the reader through the usage of CAZy to search enzyme annotations. It will also answer frequent questions such as (i) how to obtain CAZy annotations for a specific protein, a genome, or a metagenome, (ii) how to have a newly characterized family included in the CAZy classification scheme, (iii) why CAZy does not cover all protein families related to glycans/glycoconjugates, and (iv) why CAZy does not transfer functional annotation to similar sequences. Finally, we present here a recent CAZy-associated tool, namely, the Polysaccharide Utilization Loci (PUL) predictor and database in *Bacteroidetes* species (Terrapon et al. 2015).

6.1 Classes of CAZy Modules

The CAZy classification covers sequences from all taxonomic groups and provides the ground for common nomenclature for CAZymes across many glycobiologists, often specialized in some preferred taxa. Among the large diversity of proteins acting on glycoconjugates, poly- and oligosaccharides, the CAZy classification covers several enzyme classes that catalyze their assembly, breakdown, or modifications.

- Glycosyltransferases (GTs) represent the unique class in charge of glycan assembly, forming glycosidic bonds from phospho-activated sugar donors by either inverting or retaining the anomeric configuration (Campbell et al. 1997; Coutinho et al. 2003).
- Glycoside hydrolases (GHs) and polysaccharide lyases (PLs) are responsible for the cleavage of glycans (Lombard et al. 2010). GHs hydrolyze or transglycosylate glycosidic bonds, while PLs cleave the glycosidic bonds of uronic acid-containing polysaccharides by a β-elimination mechanism. Because of their widespread importance for biotechnological and biomedical applications, GHs and PLs constitute so far the best biochemically characterized set of enzymes present in the CAZy database. Interestingly, while GH-coding genes are abundant and present in the vast majority of genomes corresponding to almost half of the enzymes classified in CAZy, PLs only represent a very small proportion (Table 6.1).
- Because lignin is invariably found together with polysaccharides in the plant cell wall, CAZy recently integrated enzyme families known to be involved in lignin degradation along with lytic polysaccharide monooxygenases in a new class termed auxiliary activities (AAs) to accommodate the large range of mechanisms and substrates (Levasseur et al. 2013).
- Carbohydrate esterases (CEs) are enzymes that remove O- or N-acyl substituents on glycans (Coutinho 1999) and thereby often facilitate the action of GHs and PLs on complex polysaccharides. However, as the specificity barrier between

Table 6.1 Statistics of the CAZy website content in April 2016

CAZy class	No. of families	No. of modules in a family	No. of nonclassified	Experimentally characterized	With a 3-D structure
GH	135	289,722	4366	12,377	5065
GT	98	238,679	4449	2205	889
CBM	73	70,049	358	925	911
PL	24	7119	466	406	186
CE	16	32,843	929	311	209
AA	13	12,205	253	457	233

For each module class (first column), we indicate the number of distinct families in CAZy classification scheme (second column), the number of module occurrences classified within a specific CAZy family (distinct protein IDs, possibly having multiple module occurrences – third column), and the number of occurrences nonclassified into any family (fourth column). The numbers of corresponding proteins that have been experimentally characterized in the literature (fifth column) or have been structurally described (sixth column) are provided

carbohydrate esterases and other esterase activities is thin, it is likely that the CAZy sequence-based classification incorporates some enzymes that may act on noncarbohydrate esters (as illustrated by the high proportion of CEs falling in the "Nonclassified" category – see below).

- Carbohydrate-binding modules (CBMs) have no enzymatic activity per se but are known to potentiate the activity of many enzyme activities described above by targeting to and promoting a prolonged interaction with the substrate. If CBMs can occasionally exist in isolated or tandem forms, they are usually combined with catalytic modules within enzymes (Boraston et al. 2004). For this reason CBMs are set apart from other non-catalytic sugar-binding proteins (such as lectins and sugar transporters – see Sect. 6.6) and integrated in the CAZy classification scheme (Coutinho 1999).

The CAZy module classes are subdivided into families (see Table 6.1) based on amino acid sequence similarity, which almost invariably involves similar mechanisms. Families are designated using a simple formula including the class and a number referring to the order of family creation within the class, such as GT1 or GH130. However, the occurrence of enzymes that act on different substrates within a single family prevents the direct functional annotation of CAZymes based on family assignment. Phylogenetic analyses can frequently improve the correlation between sequence and specificity by defining subfamilies as was done for families GH5 (Aspeborg et al. 2012), GH13 (Stam et al. 2006), GH30 (St John et al. 2010), GH43 (Mewis et al. 2016), and all PL families (Lombard et al. 2010). More subfamilies are currently under development internally in CAZy and could be released when in-depth analyses confirm the stability of the subfamilies when the number of sequences increases. Finally, some of most remote homologs, for which sequence similarity is still detectable but cannot guarantee anymore any level of functional similarity nor family assignment, are also reported but without family assignment in a "Nonclassified modules" list, for each CAZy class (see

Table 6.1). The "Nonclassified modules" list is not a family per se but gathers many heterogeneous remote sequences, some that might give rise to distinct CAZy families in the future.

6.2 Browsing the CAZy Website

The homepage of the CAZy website includes a banner with several links to browse the CAZy annotation, either by CAZyme class (tabs labeled "enzyme classes" and "associated modules") or by genome (see Fig. 6.1).

6.2.1 Browsing by CAZy Class and Families

6.2.1.1 CAZy Class Webpages

The webpage dedicated to each CAZy class, illustrated in Fig. 6.2, starts with an introduction to the module function, completed by some details about the catalytic mechanisms for GHs, PLs, and GTs. Further, some statistics are given about the number of occurrence of modules in one family of this class and about the most distant homologs assigned to this class but not into a family, referred to as "Nonclassified modules." Finally, it provides the user with access to all families created in this class – links to individual webpages – in two tables: a simple ordered enumeration of existing families and a functionally oriented table that lists the different families by EC number. Please note that due to the modular nature of CAZymes, these EC numbers may not be directly associated with the family but simply borne by adjacent modules. Hence, enzymatic families with more than one known activity are repeated along this table.

6.2.1.2 CAZy Family Webpages

Each webpage dedicated to a CAZy family, illustrated in Figs. 6.3 and 6.4, contains a synthetic and updated report with all known activities (EC numbers and activity names) in the family. It should be noted that contrary to the class webpage, the activities that are listed in the header of the families correspond to the actual modules of the family and not the activity of adjacent modules. The report also

Fig. 6.1 Banner of the CAZy website where the user can choose to browse the data by CAZyme module class/family or search for a specific genome annotation

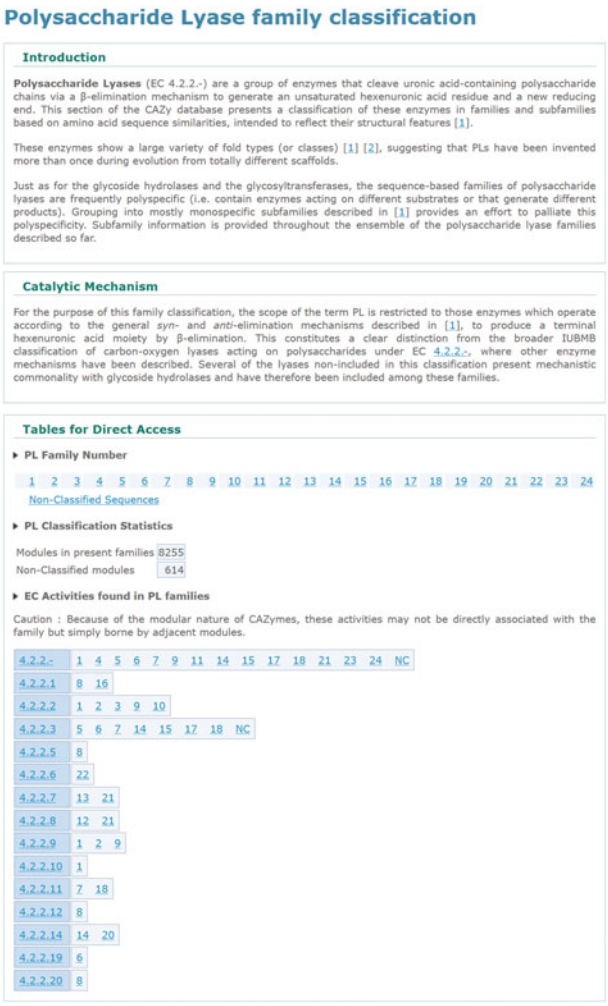


Fig. 6.2 Screenshot of the CAZy webpage that describes the PL class. Following an introductory description, statistics data and direct access to individual families are provided in different tables

specifies the mechanism (e.g., inverting or retaining), structural fold, catalytic residues, etc. where known or appropriate. More extensive encyclopedic knowledge of the biology/chemistry of some families can be obtained through links to the CAZypedia resource (see Sect. 6.8). CAZy also provides statistics about the number of known modules in each family, the number of members with a 3-D structure, and the number of functionally characterized enzymes. Finally, the complete list of modules can be browsed with a tab subdivision to see either all or restricted to a specific kingdom of life or to structurally/experimentally characterized cases. Almost all tabs present modules as lines containing the protein name, EC numbers

Glycoside Hydrolase Family 5

Known Activities	endo-β-1,4-glucanase / cellulase (EC 3.2.1.4); endo-β-1,4-xylanase (EC 3.2.1.8); β-glucosidase (EC 3.2.1.21); β-mannosidase (EC 3.2.1.25); β-glucosylceramidase (EC 3.2.1.45); glucan β-1,3-glucosidase (EC 3.2.1.58); licheninase (EC 3.2.1.73); exo-β-1,4-glucanase / cellodextrinase (EC 3.2.1.74); glucan endo-1,6-β-glucosidase (EC 3.2.1.75); mannan endo-β-1,4-mannosidase (EC 3.2.1.78); cellulose β-1,4-cellobiosidase (EC 3.2.1.91); steryl β-glucosidase (EC 3.2.1.104); endoglycoceramidase (EC 3.2.1.123); chitosanase (EC 3.2.1.132); β-primeverosidase (EC 3.2.1.149); xyloglucan-specific endo-β-1,4-glucanase (EC 3.2.1.151); endo-β-1,6-galactanase (EC 3.2.1.164); hesperidin 6-O-α-L-rhamnosyl-β-glucosidase (EC 3.2.1.168); β-1,3-mannanase (EC 3.2.1.-); arabinoxylan-specific endo-β-1,4-xylanase (EC 3.2.1.-); mannan transglycosylase (EC 2.4.1.-)
Mechanism	Retaining
Clan	GH-A
3D Structure Status	(β / α) 8
Catalytic Nucleophile/Base	Glu (experimental)
Catalytic Proton Donor	Glu (experimental)
Note	Once known as cellulase family A; many members have been assigned to subfamilies as described by Aspeborg et al. (2012) BMC Evol Biol. 12(1):186 (PMID: 22992189).
External resources	**CAZypedia**; HOMSTRAD; PROSITE;
Commercial Enzyme Provider(s)	MEGAZYME; NZYTech; PROZOMIX;
Statistics	GenBank accession (9499); Uniprot accession (1927); PDB accession (193); 3D entries (67); cryst (0)

Summary | All (8139) | Archaea (63) | Bacteria (6093) | Eukaryota (1865) | Viruses (3) | unclassified (115) | Structure (67) | Characterized (540) | Subfamilies (7294)

< | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | >

Archaea

Protein Name	EC#	Organism	GenBank	Uniprot	PDB/3D	Subf
Igag_0570		Ignisphaera aggregans DSM 17230	ADM27405.1	E0SSD3		1
Igag_0224		Ignisphaera aggregans DSM 17230	ADM27073.1	E0SQC3		1
endo-β-1,4-glucanase (EglB;TCel4;PAB0632)	3.2.1.4	Pyrococcus abyssi GE5	CAB49854.1 AEG76944.1 NP_126623.1	Q9V052		1
endo-β-1,4-glucanase (EGPh;EglB;TCel3;PH1171)	3.2.1.4	Pyrococcus horikoshii OT3	AAQ31833.1 AAQ31832.1 BAA30271.1 NP_143072.1	O58925	2ZUM[A] 2ZUN[A,B,C] 3AXX[A,B,C] 3QHM[A,B,C] 3QHN[A,B,C] 3QHO[A,B,C] 3VVG[A,B,C] 3W6L[A,B,C] 3W6M[A,B,C] 4DM1[A,B,C] 4DM2[A,B,C]	1
Py04_1787		Pyrococcus sp. ST04	AFK23355.1			1
Shell_0798		Staphylothermus hellenicus DSM 12710	ADI31911.1	D7D815		1
CHITON_2145		Thermococcus chitonophagus	CUX78924.1			1
ADU37_CDS22600		Thermococcus sp. 2319x1	ALV63957.1			1
Sequence 2 from patent US 6329187 (fragment)		unidentified archaeon AEPII1a	AAQ77850.1			1
Huta_2396		Halorhabdus utahensis DSM 12940	ACV12562.1	C7NMC3		7
Htur_4673		Haloterrigena turkmenica DSM 5511	ADB63456.1	D2S259		7
L593_05380		Salinarchaeum sp. Harcht-Bsk1	AGN01025.1			7
QT06_C0001G0664		archaeon GW2011_AR15	AJF61504.1			13
ASAC_0734		Acidilobus saccharovorans 345-15	ADL19140.1	D9Q1F2		19
SE86_04550 (fragment)		Acidilobus sp. 7A	AMD30698.1			19
SE86_03690 (fragment)		Acidilobus sp. 7A	AMD30586.1			19
SE86_03695 (fragment)		Acidilobus sp. 7A	AMD30587.1			19
Cmaq_0936		Caldivirga maquilingensis IC-167	ABW01768.1	A8MDB2		19
Igag_0312		Ignisphaera aggregans DSM 17230	ADM27158.1	E0SQT4		19
PTO1452		Picrophilus torridus DSM 9790	AAT44037.1	Q6KZ15		19
SH_2292		Sulfolobus islandicus HVE10/4	ADX83632.1			19
LD85_2652		Sulfolobus islandicus L.D.8.5	ADB88268.1	D2PGC8		19
LS215_2522		Sulfolobus islandicus L.S.2.15	ACP36477.1	C3ML75		19
SIL_2199		Sulfolobus islandicus LAL14/1	AGJ63637.1			19
M1425_2350		Sulfolobus islandicus M.14.25	ACP39078.1	C3MSD5		19
M1627_2428		Sulfolobus islandicus M.16.27	ACP56280.1	C3N1Q3		19
M164_2357		Sulfolobus islandicus M.16.4	ACR42955.1	C4KL74		19

Fig. 6.3 Screenshot of the CAZy webpage that describes the GH5 family information with the specific details on protein sequences attributed to subfamilies in the rightmost column ("Subfamilies" tab)

if any, the organism, the GenBank accessions (one reference in bold, and redundant ones below), and the UniProt and PDB identifiers (crystals not yet solved/deposited labeled as "cryst"). Further, for families with subfamily division, a tab at the very right shows the subfamily number (see Fig. 6.3). Finally, the "Structure" tab is the special case (see Fig. 6.4): it does not contain GenBank nor UniProt accessions but instead displays more detailed information from the PDB files. For each PDB file, we extract and display the resolution when the structure was solved by x-ray

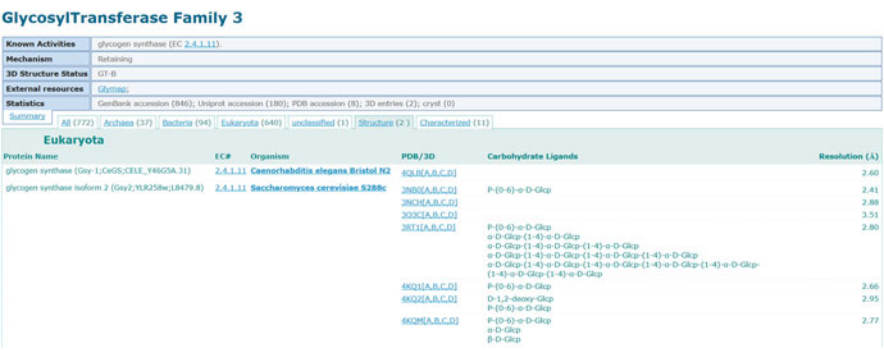


Fig. 6.4 Screenshot of the CAZy webpage corresponding to the GT3 family with the specific display (at the *bottom*) of structurally characterized proteins and related information ("Structure" tab)

crystallography (otherwise we indicate the method: powder diffraction or nuclear magnetic resonance).

6.2.1.3 Recent Addition of Carbohydrate Ligands

The PDB does not provide any option to perform a comprehensive search for carbohydrate structures found in CAZyme binding sites, and, unlike proteins or nucleic acids, the nomenclature for carbohydrate residues within PDB files is not yet standardized. Significantly, the information on how the isolated carbohydrate residues are linked to each other is not described in PDB files. For each PDB file, we thus extract the carbohydrate ligand information using PDB-care (www.glycosciences.de/tools/pdb-care/; (Lütteke and Von Der Lieth 2004)). These ligands are filtered for display as follows. N- and O-glycans covalently linked to Asn or Ser/Thr residues are discarded as they correspond to posttranslational modifications of the protein structure, and generally not directly linked to enzyme function. The remaining carbohydrate ligands are retained as they should describe functional recognition in catalytic or other binding sites in CAZymes and are displayed in the structure pages of CAZy (see Fig. 6.4) following IUPAC nomenclature (Lombard et al. 2014). Not all carbohydrate structures are susceptible to automated description by PDB-care. In a number of cases, we must manually curate and provide IUPAC descriptions for structures that are unsuitable to PDB-care such as (i) nonreducing glycans (cyclodextrins, sucrose and sucrose derivatives, trehalose, kestose, raffinose, nystose, etc.), (ii) ligands that are made of both carbohydrate and noncarbohydrate moieties such as acarbose, (iii) thio-oligosaccharides, (iv) fluorine-containing carbohydrates, and (v) oligosaccharides containing 3,6-anhydro bridges. In addition, automated scripts have been devised to handle close to 200 carbohydrate analogues that we denote <carb_like_*ligandref*> where *ligandref* corresponds to the three-letter ligand name given by the PDB. For instance, the carbohydrate-like

inhibitor 1-deoxynojirimycin appears as <carb_like_NOJ>. Significantly, nearly half (45 %) of the approx. 7500 PDB structures present in CAZy as of April 2016 bear a glycan-containing ligand or a glycan analog revealing enzyme-glycan interactions.

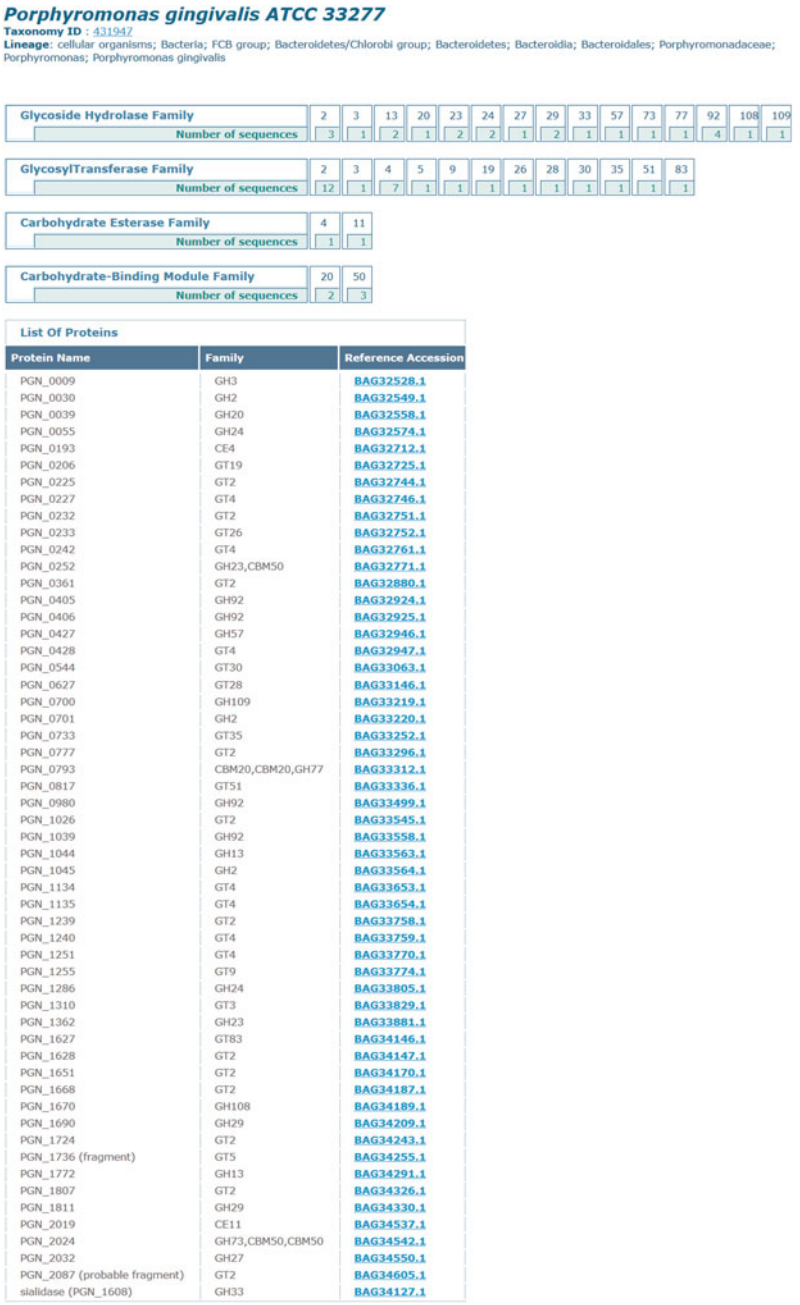
6.2.2 Browsing by Genome

The collection of carbohydrate-active enzymes encoded by the genome of an organism, hereafter referred to as "CAZome," provides an insight into the nature and extent of the metabolism of complex carbohydrates of the species. The CAZomes of free-living organisms typically correspond to 1–5 % of the predicted coding genes. Because of the massive chemical, structural, and functional variability of carbohydrates, CAZome comparisons can highlight the adaptation of the CAZymes repertoire of species to their carbohydrate environment.

The CAZy website allows to browse CAZomes by kingdom of life, where species are presented in alphabetically ordered tabs. For each organism, the complete list of CAZymes is displayed in addition to the family distribution, as illustrated in Fig. 6.5. As of April 2016, CAZy is close to 5000 public genomes, with more than 4000 bacterial genomes but less than 200 eukaryotes. This is due to the fact that the CAZomes listed in the CAZy website correspond to protein models of *finished* genomes from the daily releases of GenBank. In just a few rare cases, genomes with protein models not released as finished entries in GenBank, but publicly available, have been analyzed and are presented in CAZy. However, for these few cases, the display only shows the number of proteins in each family, but does not feature the actual list of proteins with database accessions. The taxonomical lineage of the genome is directly extracted and updated from the NCBI Taxonomy database.

6.3 Retrieving Information from the Search Form/Engine in the CAZy Website

To facilitate search of specific information, the CAZy website includes a search tool, which appears at the top right of every page. The search form is composed of a text area with a magnifying glass to enter the required query and a drop-down list to indicate the field of searched information (see Fig. 6.6, with "Site" option to search in every field). Main fields notably allow the user to search by CAZy family, organism (name, even partial, or taxonomy id), protein name or accessions (GenBank, UniProt, and PDB), ligand (indicating a sugar-like compound, a part of a chain, or the catalytic residue to which the ligand is attached, e.g., "GLU"), activity (EC number/name), etc. The result of the search either indicates the modularity of the protein or provides direct links to the relevant genome and family webpages.



Porphyromonas gingivalis ATCC 33277

Taxonomy ID : 431947

Lineage: cellular organisms; Bacteria; FCB group; Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales; Porphyromonadaceae; Porphyromonas; Porphyromonas gingivalis

Glycoside Hydrolase Family	2	3	13	20	23	24	27	29	33	57	73	77	92	108	109
Number of sequences	3	1	2	1	2	2	1	2	1	1	1	1	4	1	1

GlycosylTransferase Family	2	3	4	5	9	19	26	28	30	35	51	83
Number of sequences	12	1	7	1	1	1	1	1	1	1	1	1

Carbohydrate Esterase Family	4	11
Number of sequences	1	1

Carbohydrate-Binding Module Family	20	50
Number of sequences	2	3

List Of Proteins

Protein Name	Family	Reference Accession
PGN_0009	GH3	BAG32528.1
PGN_0030	GH2	BAG32549.1
PGN_0039	GH20	BAG32558.1
PGN_0055	GH24	BAG32574.1
PGN_0193	CE4	BAG32712.1
PGN_0206	GT19	BAG32725.1
PGN_0225	GT2	BAG32744.1
PGN_0227	GT4	BAG32746.1
PGN_0232	GT2	BAG32751.1
PGN_0233	GT26	BAG32752.1
PGN_0242	GT4	BAG32761.1
PGN_0252	GH23,CBM50	BAG32771.1
PGN_0361	GT2	BAG32880.1
PGN_0405	GH92	BAG32924.1
PGN_0406	GH92	BAG32925.1
PGN_0427	GH57	BAG32946.1
PGN_0428	GT4	BAG32947.1
PGN_0544	GT30	BAG33063.1
PGN_0627	GT28	BAG33146.1
PGN_0700	GH109	BAG33219.1
PGN_0701	GH2	BAG33220.1
PGN_0733	GT35	BAG33252.1
PGN_0777	GT2	BAG33296.1
PGN_0793	CBM20,CBM20,GH77	BAG33312.1
PGN_0817	GT51	BAG33336.1
PGN_0980	GH92	BAG33499.1
PGN_1026	GT2	BAG33545.1
PGN_1039	GH92	BAG33558.1
PGN_1044	GH13	BAG33563.1
PGN_1045	GH2	BAG33564.1
PGN_1134	GT4	BAG33653.1
PGN_1135	GT4	BAG33654.1
PGN_1239	GT2	BAG33758.1
PGN_1240	GT4	BAG33759.1
PGN_1251	GT4	BAG33770.1
PGN_1255	GT9	BAG33774.1
PGN_1286	GH24	BAG33805.1
PGN_1310	GT3	BAG33829.1
PGN_1362	GH23	BAG33881.1
PGN_1627	GT83	BAG34146.1
PGN_1628	GT2	BAG34147.1
PGN_1651	GT2	BAG34170.1
PGN_1668	GT2	BAG34187.1
PGN_1670	GH108	BAG34189.1
PGN_1690	GH29	BAG34209.1
PGN_1724	GT2	BAG34243.1
PGN_1736 (fragment)	GT5	BAG34255.1
PGN_1772	GH13	BAG34291.1
PGN_1807	GT2	BAG34326.1
PGN_1811	GH29	BAG34330.1
PGN_2019	CE11	BAG34537.1
PGN_2024	GH73,CBM50,CBM50	BAG34542.1
PGN_2032	GH27	BAG34550.1
PGN_2087 (probable fragment)	GT2	BAG34605.1
sialidase (PGN_1608)	GH33	BAG34127.1

Fig. 6.5 Screenshot of the CAZy webpage for the genome of *Porphyromonas gingivalis ATCC 33277*. The CAZy family distribution (*top*) is followed by the list of all identified CAZymes with their modularity (where relevant)

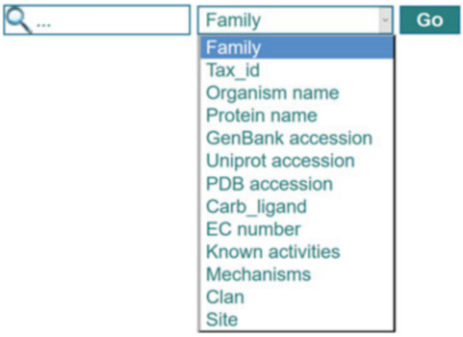


Fig. 6.6 Search tools and fields that accept queries in the CAZy database

6.4 How to Get CAZy to Annotate Your Studied Protein, Genome, or Metagenomic Sample?

The most straightforward way to obtain a CAZy annotation for a genome is to submit your sequence(s) to the NCBI with the "finished" status or by contacting us (cazy@afmb.univ-mrs.fr) to request our analysis as part of a collaborative effort. Every day, the internal CAZy tool for the semiautomatic modular assignment runs on the protein sequences from the daily release of NCBI GenBank, and our computational equipment makes it possible to perform several large-scale analyses such as the annotation of CAZyme repertoire in genomic and metagenomic investigations (the latter can be in the form of DNA or protein sequences). These putative assignments are thus manually validated (or rejected) by expert curators. Subsequent family comparisons provide insights into how similar or different might be the newly sequenced organisms compared to closely related species or how metagenomic samples differ relative to each other. Differences in the relative family size, for example, can reflect the relative diversity or complexity of the inherent biological processes and, therefore, the biology of the compared species/samples.

Automatic tools, freely available on the web, attempt to emulate the CAZy classification scheme. Our experience is that these fully automatic methods provide results that can be substantially different from actual CAZy assignments. Further, these tools sometimes include outdated module families and automatic subfamilies that are not curated. And finally (and most importantly), automatic predictors are dependent on the user's parameters for detection threshold, generally applied to e-value statistics. The issues with an e-value threshold is that (i) the e-value varies with the length of the aligned sequences for identical sequence similarity percentage, (ii) such threshold completely bypasses curation to distinguish possibly functionally related homologs from locally shared secondary structures, (iii) a unique threshold is not appropriate for families of unequal diversity, and (iv) low/significant e-values do not guarantee the completeness of modules since all detection tools (BLAST- or HMM-based) are local by nature.

6.5 What to Do If You Obtain a New Activity or the 3-D Structure for a CAZyme or If You Characterize a New CAZy Family?

We cordially invite biologists with experimental results to contact us (with appropriate material such as a peer-reviewed preprint of the work), to reduce the number of reports that was missed during our bibliography surveillance. If the subject protein has not yet been submitted to GenBank/PDB, we strongly encourage you to do so. This will allow the automatic capture and display of the annotation on the CAZy website as follows:

6.5.1 Novel Activity, 3-D Structure, or New Chemical Information for an Enzyme in an Existing CAZy Family

If the studied enzyme is already assigned to a numbered family in the CAZy website, we will complete the family records (content and description) with the new information. The new information will thus be displayed on the webpage dedicated to the family, in the corresponding sections as described in Sect. 6.2.1. The accumulation of experimental evidence will notably help in refining the classification system by the population of subfamilies based on phylogenetic analyses (see Table 6.1).

Warning If your enzyme appears in the CAZy website but in the "Nonclassified modules" listed for each CAZy class, it has to be considered as new regarding the CAZy classification (see below).

6.5.2 Novel Family in the CAZy Classification

If the newly characterized enzyme does not belong to any known CAZy family, or belongs to the "Nonclassified modules" of a CAZy class, we will create a new CAZy family, as follows. Starting from the subject sequence, we first collect the most similar homologs in GenBank by BLAST. Then, we iteratively gather more distant family members using HMMs, which capture the family diversity (flexible and constrained positions in the multiple sequence alignment and corresponding structure). The delineation of the module boundaries is guided by family conservation and is generally facilitated or refined when a 3-D structure becomes available. The creation and analysis of a new CAZy family remains private until notification by the original requestor or until publication.

6.6 Why Doesn't CAZy Extend Its Classification Scheme to Other Classes of Enzymes?

Even though the CAZy families do not always coincide with a precise substrate specificity, family assignment often gives clues on what the broad substrate category might be. And when the relatedness to a functionally characterized enzyme is high, typically at the subfamily level, then the functional predictions for CAZymes can be very good. In any case, this is substantially more informative than most other families of enzymes (kinases, proteases, esterases, etc.) whose substrates are difficult or impossible to derive from their sequence alone. Due to our limited number of expert curators and to the poorer relationship between family and function in other enzyme categories, we prefer to stay within our field of competence and do not expand the scope of CAZy beyond what it is.

6.7 Why Doesn't CAZy Propagate Experimentally Established Function to Similar Sequences?

All too often during a protein/genome study, the functional annotations automatically inferred by computational methods contain a significant amount of low-quality and even erroneous information that are then propagated to the next projects. For example, the transfer of Gene Ontology (GO) terms based on Pfam modules usually assigns excessively general terms. This can be explained by the stringent policy of module annotation that links a module solely to the GO terms common to all proteins having this module, whatever the diversity of the possible module combination and associated functions. Other widely used tools are also prone to overprediction by transferring annotation from a demonstrated example to distant homologs or by creating annotation based on hypothesis devoid of any experimental evidence, as, for example, with the BACON domain (Pfam ID PF13004) which stands for *Bacteroidetes*-Associated Carbohydrate-binding Often N-terminal based on a conjecture-only publication. This conjecture has been recently challenged in a publication showing that the BACON domain of BACOVA_02653 protein of *Bacteroides ovatus ATCC 8483* does not have any carbohydrate-binding activity (Larsbrink et al. 2014). As a consequence CAZy did not create a new CBM family for such modules. More generally, to avoid problems linked to annotation transfer, CAZy policy is to display EC numbers only for the experimentally characterized enzymes.

6.8 Links and Announcements on the CAZy Website

In addition to multiple links to essential enzymatic and glycogenomic resources, CAZy contains many cross-links to the CAZypedia resource. CAZypedia is a community-driven encyclopedic resource meant to be the logical extension of the CAZy classification. It contains extensive information about CAZy families with especial emphasis on GHs, but the other CAZy families are now being filled progressively. The CAZy website also offers an opportunity for commercial enzyme providers to present their products which follow the CAZy nomenclature and to announce scientific meetings and opened job positions related to CAZymes.

6.9 What Is the PULDB Database?

PULDB is a recent addition to CAZy that describes Polysaccharide Utilization Loci (PULs) experimentally characterized in the literature and our automated PUL predictions in *Bacteroidetes* species (Terrapon et al. 2015). A PUL is a set of physically linked genes organized around a *susCD* gene pair. Named according to the prototypic starch utilization system, *susC* is a characteristic membrane transporter, and *susD* encodes an outer membrane-binding protein (Shipman et al. 2000). PULs are prevalent in the *Bacteroidetes* phylum, with species encoding dozens of PULs, each tailored to degrade a particular glycan structure. PULs provide an evolutionary advantage to these gram-negative species by orchestrating the breakdown of complex glycans, thanks to the encoded CAZymes, and by sequestrating these nutrients away from competitors (Terrapon and Henrissat 2014). PULDB offers a query engine to search PULs by species, by (combination of) CAZy modules, and by locus tags. It also contains a JBrowse engine (Skinner et al. 2009) to visualize the genomic context of CAZymes and PULs for all the integrated genomes (source: IMG HMP project at the JGI (Markowitz et al. 2012)) as illustrated in Fig. 6.7.

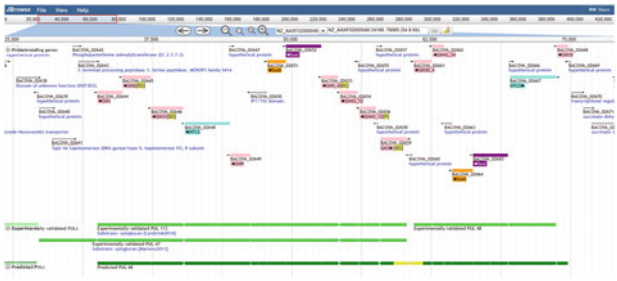


Fig. 6.7 Screenshot of the PULDB website. JBrowse visualization of a xyloglucan PUL in *Bacteroides ovatus ATCC 8483*

6.10 Conclusion

The CAZy database is based on family classification schemes that were established in the 1990s, before any genome had been completely sequenced. A key feature of the success of CAZy is the stability of its underlying classification system. The earliest GH families have survived a >500 times expansion since their creation in 1991. Other key features of CAZy are the integration of the variable modular architecture of CAZymes and its panel of expert curators to capture structural and functional data from the literature. In the near future, however, high-throughput enzymology will deliver more data in 1 year than what has accumulated during the last 50 years. Without a mechanism to capture functional information reliably, a large amount of experimental data will remain buried and underexploited.

References

Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B (2012) Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). BMC Evol Biol 12(1):186

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2012) GenBank. Nucleic Acids Res:gks1195

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242

Boraston A, Bolam D, Gilbert H, Davies G (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem J 382:769–781

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. Plant Bioinf Methods Protocols:23–54

Campbell JA, Davies GJ, Bulone V, Henrissat B (1997) A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. Biochem J 326(Pt 3):929

Coutinho P (1999) Carbohydrate-active enzymes: an integrated database approach. In Recent advances in carbohydrate bioengineering

Coutinho PM, Deleury E, Davies GJ, Henrissat B (2003) An evolving hierarchical family classification for glycosyltransferases. J Mol Biol 328(2):307–317

Larsbrink J, Rogers TE, Hemsworth GR, McKee LS, Tauzin AS, Spadiut O, Klinter S, Pudlo NA, Urs K, Koropatkin NM, Creagh AL, Haynes CA, Kelly AG, Cederholm SN, Davies GJ, Martens EC, Brumer H (2014) A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. Nature 506(7489):498–502

Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B (2013) Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol Biofuels 6(1):41

Lombard V, Bernard T, Rancurel C, Brumer H, Coutinho P, Henrissat B (2010) A hierarchical classification of polysaccharide lyases for glycogenomics. Biochem J 432:437–444

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42 (Database issue):D490–D495

Lütteke T, Von Der Lieth CW (2004) pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. BMC Bioinf 5(1):69

Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P (2012) IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 40(D1):D115–D122

Mewis K, Lenfant N, Lombard V, Henrissat B (2016) Dividing the large glycoside hydrolase family 43 into subfamilies: a motivation for detailed enzyme characterization. Appl Environ Microbiol AEM. 03453–03415

Shipman JA, Berleman JE, Salyers AA (2000) Characterization of four outer membrane proteins involved in binding starch to the cell surface of Bacteroides thetaiotaomicron. J Bacteriol 182(19):5365–5372

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH (2009) JBrowse: a next-generation genome browser. Genome Res 19(9):1630–1638

St John FJ, González JM, Pozharski E (2010) Consolidation of glycosyl hydrolase family 30: a dual domain 4/7 hydrolase family consisting of two structurally distinct groups. FEBS Lett 584(21):4435–4441

Stam MR, Danchin EG, Rancurel C, Coutinho PM, Henrissat B (2006) Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of α-amylase-related proteins. Protein Eng Des Sel 19(12):555–562

Terrapon N, Henrissat B (2014) How do gut microbes break down dietary fiber? Trends Biochem Sci 39(4):156–158

Terrapon N, Lombard V, Gilbert HJ, Henrissat B (2015) Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. Bioinformatics 31(5):647–655