The carbohydrate-active enzyme database: functions and literature

Elodie Drula, Marie-Line Garron, Suzan Dogan, Vincent Lombard, Bernard Henrissat, Nicolas Terrapon

The carbohydrate-active enzyme database: functions and literature

Drula, Elodie; Garron, Marie Line; Dogan, Suzan; Lombard, Vincent; Henrissat, Bernard; Terrapon, Nicolas

Published in:
Nucleic Acids Research

Link to article, DOI:
10.1093/nar/gkab1045

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Published online 29 November 2021 *Nucleic Acids Research, 2022, Vol. 50, Database issue* ***D571–D577***
https://doi.org/10.1093/nar/gkab1045

The carbohydrate-active enzyme database: functions and literature

Elodie Drula[1,2], Marie-Line Garron[1,2], Suzan Dogan[1,2], Vincent Lombard[1,2], Bernard Henrissat[1,2,3,4] and Nicolas Terrapon[1,2,*]

[1]Aix Marseille Univ, CNRS, UMR7257 AFMB, Marseille, France, [2]INRAE, USC1408 AFMB, Marseille, France, [3]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia and [4]Technical University of Denmark, DTU Bioengineering, Kgs Lyngby, Denmark

Received September 14, 2021; Revised October 13, 2021; Editorial Decision October 14, 2021; Accepted October 14, 2021

ABSTRACT

Thirty years have elapsed since the emergence of the classification of carbohydrate-active enzymes in sequence-based families that became the CAZy database over 20 years ago, freely available for browsing and download at www.cazy.org. In the era of large scale sequencing and high-throughput Biology, it is important to examine the position of this specialist database that is deeply rooted in human curation. The three primary tasks of the CAZy curators are (i) to maintain and update the family classification of this class of enzymes, (ii) to classify sequences newly released by GenBank and the Protein Data Bank and (iii) to capture and present functional information for each family. The CAZy website is updated once a month. Here we briefly summarize the increase in novel families and the annotations conducted during the last 8 years. We present several important changes that facilitate taxonomic navigation, and allow to download the entirety of the annotations. Most importantly we highlight the considerable amount of work that accompanies the analysis and report of biochemical data from the literature.

INTRODUCTION

The family classification of carbohydrate-active enzymes (CAZymes) and the annotation of the constitutive modules from the primary sequences is now primarily used to analyze, understand, and compare the ability of an organism or a community thereof to assemble and breakdown complex carbohydrates. To carry out detailed and meaningful analyses and comparisons, bioinformatics algorithms must be completed by a deep knowledge of the peculiarities of carbohydrates and their enzyme families.

Since its inception, literature-based functional data have underpinned our database. Each family is built around at least one biochemically characterized member. There is no family of unknown function in CAZy and thus we avoid the transmission of unverified information. This is also for this reason that our families are not named but simply numbered: indeed, because sequence-based families can group together enzymes of different specificity, the function of the first characterized family member is rarely representative of the functional diversity of the family. Once a family is created, it is populated by homologous sequences and literature surveillance begins to identify newly characterized family members which may or not share the same activity (specificity) as the founding member. Enzymes that perform catalysis at the anomeric carbon of carbohydrates such as glycoside hydrolases and glycosyltransferases have catalytic mechanisms that either retain or invert the stereochemistry at the reaction site. Although substrate specificity is variable in the families, it has been found that the catalytic mechanism is extremely well conserved and thus predictable once established for a family member. Depending on the families and their biotechnological interest, the number of experimentally characterized members can vary from one (the minimum, for 60 of the 340 families) to several hundred (>800 in family GH13). The inventory of the particular activities (specificities) found in a given family is yet an essential prerequisite for functional prediction via sequence similarity. Families that have been insufficiently sampled in terms of specificity are thus less adapted to functional predictions based on best Blast hits. In the case of multifunctional families, the time-consuming task of searching the literature for functionally characterized members can be harnessed to divide families in subfamilies of narrower specificity, and therefore of increased functional prediction power. In spite of our efforts, we are aware that we miss several and perhaps many biochemical characterizations of CAZymes. The reasons are multiple: insufficient number of trained staff, dispersion of biochemical data on CAZymes over a forever increasing number of journals, data buried in sometimes very long supplementary materials and, most importantly, lack of a simple tabular format that would

*To whom correspondence should be addressed. Tel: +33 491 825 587; Fax: +33 491 266 720; Email: nicolas.terrapon@univ-amu.fr

allow authors to report their biochemical data and link them to a particular amino acid sequence accession to a database. Yet in the era of high-throughput biochemical assays, the capture of unformatted biochemical information is a growing concern and translates into pain-staking efforts with little reward and no funding. Here we describe our efforts to facilitate this process.

Maintenance of the modular classification and daily updates

For now over 20 years, the daily releases of GenBank protein sequences (www.ncbi.nlm.nih.gov/genbank) (1) are analyzed and the sequence modules corresponding to existing families of glycoside hydrolases (GH), glycosyltransferases (GT), polysaccharide lyases (PL), carbohydrate esterases (CE), auxiliary activities (AA including lytic polysaccharide monooxygenases; LPMO) and carbohydrate-binding modules (CBM) are systematically annotated and assigned to their respective families. The assignment is based on Blastp hits of these query sequences against target sequences already classified. Extremely confident results (e-value less than 1e–50 and 100% coverage) allow the automatic assignment to the same family(ies) as the target. In all other cases, a human curator cross-examines Blastp (https://blast.ncbi.nlm.nih.gov/Blast.cgi) (2) and HMMER3 (http://hmmer.janelia.org/) (3) results based on search against various libraries of isolated modules (e.g. modules derived from functionally characterized CAZymes; family multiple alignment/consensus).

Although the accumulated annotations for over 20 years constitute a strong basis to facilitate the task of the curators, there is still a considerable amount of curation to perform. This manual curation has essentially been performed by a single individual until 2015, but the workload is now shared among the members of the team.

Since the previous description of our database eight years ago (4), the number of CAZymes listed on our site has been multiplied by 6.7 (Table 1). The growth of the number of LPMO sequences appears to be less (×3.2) but this is due to the fact that LPMOs are mostly found in eukaryotes and that eukaryotic genomes are rarely brought to the finished status, and thus the encoded proteins do not appear in the daily releases of GenBank. On the other hand, the number of sequences comprising a carbohydrate-binding module (CBM) has increased by ×8.6, a performance most likely due to the fact that CBMs can occur alone or appended to proteins that are not necessarily acting on carbohydrates (e.g. esterases, peptidases, etc.) to which they confer carbohydrate-binding ability.

The increase in database size is naturally augmented by the discovery of novel families (Table 1). For instance, the number of PLs has increased eight times, i.e. more than the average, due in part to the addition of 19 PL families during the last 8 years (+83%). However, the increase in the number of sequences to analyze and assign to families is mostly due to the ever-growing number of sequenced genomes. In our 2013 paper (published in 2014, (4)), the CAZy database reported the list of CAZymes for the genomes of 2351 Bacteria, 158 Archaea, and 73 Eukaryota. As of July 2020, these numbers have grown to 20 954 Bacteria (×8.9), 427 Archaea (×2.7) and 346 Eukaryota (×4.7). Importantly, since 2020, we have systematically renamed genomes according to changes in the NCBI taxonomy. The constant increase in the number and redundancy of sequenced genomes has led us to make strategic choices. In 2015, the UniProt database decreased its size from almost 110 million to 60 million sequences by removing highly similar proteomes (5). In a similar vein, we have decided to no longer systematically analyze genomes for which at least 60 strains have already been analyzed in CAZy. However, we occasionally perform a custom analysis of such genomes at the demand of a colleague. Similarly, genomes released as complete but originating from metagenomics data assembly are not systematically analyzed, nor are those from the myriads of clinical isolates of strains already present in CAZy. We made these choices to focus our limited resources on the exploration of the novel and more distant genome/sequence space, which requires more human curation given the distance to already annotated references.

At present, human curation remains indispensable, in order to identify in the deluge of sequences those that correspond to poorly explored regions of the sequence space, hence where functional exploration remains necessary. These distant relatives require to be manually analyzed with an adjustment of the inclusion thresholds, inspection of the conservation of the catalytic machinery and critical assessment of the literature. For instance, the non-LPMO families of auxiliary activities (AA1 to AA8, and AA12) of fungi have been shown to cooperate with CAZymes in the deconstruction of plant cell walls, but such cooperation has not been established so far for bacteria (6); we thus do not include bacterial sequences in these families, except if they carry a CBM. Similarly, the criteria for inclusion of distant relatives of carbohydrate esterases have been tightened since we observed that the distinction of esterases active on carbohydrates vs. esterases that act on non-carbohydrates, or even non-esterases, is particularly difficult. For example, several carbohydrate esterase families belong to the GDSL hydrolase superfamily (Pfam clan CL0264) which displays activities such as thio-/arylesterase, protease and lysophospholipase activities (7). Such observations are possible thanks to human curation, which remains essential to detect and adjust criteria that escape automated methods. An example is the necessity to modify the Hidden Markov models to take into account specific features such as increased family sequence diversity or the precise domain boundaries that appear only after a first 3D structure is described in a family. The human curation thus ensures continuously improved annotations, which in turn improve further genomic or metagenomics analyses.

Functional annotations

The display of functional information in the CAZy database is essential to inform the user community of progress made by others. In order to display this information, three steps are necessary: (i) literature survey, (ii) assessment of the pertinence of the data and (iii) storage/integration into the database. Literature surveys are presently performed in an empirical manner by performing searches on PubMed in order to detect publications that report one or several biochemical characterizations. While the title of the papers

Table 1. Growth of the CAZy database during the past 8 years. For the distinct protein classes (rows), columns indicate: the total number of modules annotated in CAZy ('Modules'); the number of modules which have been biochemically characterized ('Characterized'); the number of modules with at least one 3D-structure in the PDB ('With Structure'); and the number of created families in each class ('Families')

	Modules		Characterized		With Structure		Families	
Protein class	Dec-2013	Sept-2021	Dec-2013	Sept-2021	Dec-2013	Sept-2021	Dec-2013	Sept-2021
GH	162 550	995 295	6 094	7 248	790	1 567	133	171
GT	122 853	849 449	1 507	1 862	137	298	94	114
PL	4 114	31 710	246	357	52	99	23	42
CE	16 467	97 226	198	216	61	99	16	19
CBM	33 793	277 412	-	-	269	408	68	88
AA	4 921	18 935	134	255	56	118	11	17

rarely discloses such characterizations, the abstract and/or a complete reading of the articles (and their ever-growing supplementary materials) allow to identify and verify functional claims. This task is becoming really challenging with the advances of experimental enzymology, and the massive increase in the number of publications and of journals (of unequal quality). To avoid unverified claims, the CAZy database has decided to not integrate functional data from preprint servers and to focus exclusively on published literature (mostly publications, but sometimes patents).

Members of the scientific community involved in CAZyme research sometimes spontaneously notify us of novel functional data, although this is generally to obtain a family number for the publication reporting a novel family. Although useful to describe new families, this does not capture the bulk of the biochemical characterizations that appear in the literature. In order to improve functional data capture, we have implemented a simple online form (http://www.cazy.org/Functional-Data.html; accessible from any CAZy webpage in the top menu; Figure 1, box 1) that enables an author to report one or several functional characterization(s), using a minimal number of fields, namely a sequence database accession (GenBank (1), PDB www.rcsb.org (8), or UniProt www.uniprot.org (5)), a textual description of the activity or the EC number (www.enzyme-database.org) (9), the PubMed ID (pubmed.ncbi.nlm.nih.gov) (10) or DOI of the publication reporting the function (only published peer-reviewed data are taken into account) and an email address for occasional clarification. Although the authors of a paper are generally best placed to fill this form, anyone can contribute and in particular for older data that have escaped our attention. The submitted data are then assessed by a curator before deciding on whether the evidence is sufficient to be posted on our website. Importantly, it is possible to submit several proteins or functions (separated by semicolons) in the case of multiple characterizations within a given paper.

We thus cordially invite the community to submit functional information using this form. In return, after verification by a curator, a link from the characterized enzyme to the suitable PubMed ID or DOI is provided on the CAZy website for the enzyme in question (Figure 1, box 6; discussed in more detail in the 'New display of information' section hereafter). We hope that the display of links to the publications on CAZy website will provide an increased visibility for the biochemists who performed the work and will facilitate bibliography survey prior to future analysis and discoveries.

Other sources of functional information

Every week all new structures released by the PDB are analyzed for the presence of CAZymes using the same criteria as for the GenBank sequences. Because functional information is often associated with a PDB entry, we systematically check the PubMed ID associated with each structure. Unfortunately, associated publications frequently appear lately *after* release of the coordinates, limiting a direct capture of functional information from the PDB. To improve this procedure, we regularly reanalyze entire flat files from the PDB for updated publications.

Swiss-Prot, the manually curated part of UniProt, is another widely recognized source of functional information for proteins (565k proteins in Swiss-Prot; 219M in uncurated TrEMBL (11,12)). We systematically searched the correspondence between functionally characterized CAZymes and UniProt entries (Swiss-Prot or TrEMBL). We observe that currently two thirds of the ∼10k characterized CAZymes only have a reference in TrEMBL, and that Swiss-Prot records a literature reference for only 55% of the cases. This illustrates the relevance of our efforts for literature display. Inversely, we identified proteins in Swiss-Prot with a supporting literature reference that are missing in CAZy. These instances (currently ∼140) are being reviewed and, if deemed sufficient, will be integrated (see below).

Once a pertinent reference paper is detected, functional annotation can begin. This can be a tedious effort as high-throughput Biology allied to the emergence of ever-growing supplementary materials can deliver over a dozen enzyme characterizations in a single paper (13–15), or several dozen (16) or even hundreds (17). The functional annotation requires to identify in the published paper a sequence database accession (GenBank, PDB, UniProt) for each characterized protein in order to link the activity to a particular entry. Then the functional information reported in the paper must be decrypted, and this is the most difficult (yet important!) part as there is no general format to report these activities, and annotation requires a suitably precise description of the activity based on pertinent experiments/results. A desirable parameter for the functional annotation is the purification and assay of the protein (preferably recombinant to ensure coincidence between the enzyme activity and the sequence), which is a lot more reliable than indirect observations by knockout or phenotype observations whose results may be the consequence of a chain of events. Finally, several substrates might undergo the enzymatic reaction, with similar or large differences in

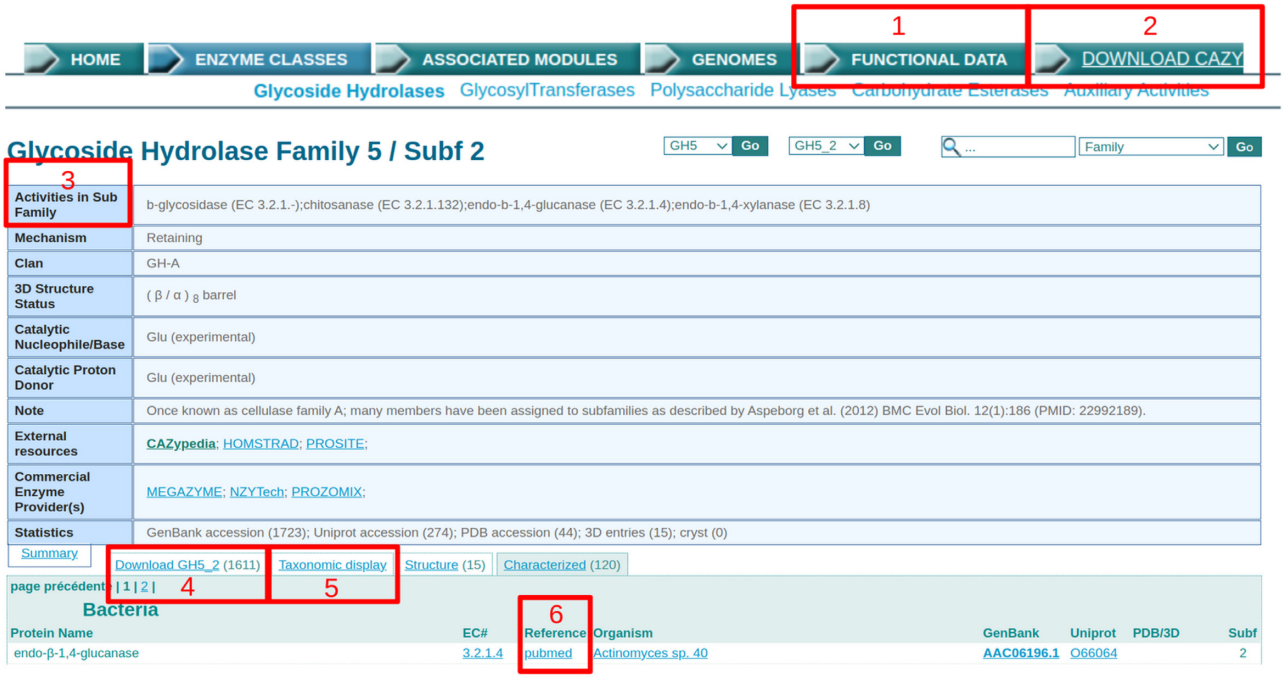

Figure 1. The new headers of CAZy (sub)family webpages. Boxes highlight the novel features: (1) direct access to the form to report functional characterization(s); (2) download link to the complete list of protein accessions and their CAZy modules; (3) for subfamilies with characterized members, only the subfamily-specific functions are now listed; (4) taxonomic tabs have been removed and a 'Download' tab gives access to a text file corresponding to previous 'ALL' tab with the complete list of protein accessions belonging to the (sub)family; (5) a 'Taxonomic display' tab links to a Krona visualization of the family members as illustrated in Figure 2; (6) links to the publications that describe the functions of the enzymes are now given (preferentially PubMed, otherwise DOI or occasionally URL).

magnitude, which should be summarized. The reference system currently and widely utilized to describe enzymatic reactions is the EC system developed by the Enzyme Nomenclature committee of the IUBMB (9). This classification was developed to avoid the proliferation of names that had little to do with the catalyzed reaction (subtilisin, papain, invertase, etc.). Because they were defined for another purpose in the pre-genomic (and even the pre-sequencing) era, EC numbers can cause problems for genomic functional annotation. For instance, EC 3.2.1.4 (cellulase) does not specify whether the enzyme operates with a retaining or an inverting mechanism. To date 214 EC numbers have been issued by the Enzyme Nomenclature committee to describe the substrates and products of GHs. Unfortunately, this already large choice does not cover the extent of glycan diversity, even though new EC numbers are introduced each year. In fact, current trends in research and research funding strongly favor the quest for novelty, e.g., CAZymes reactions not described before. This precludes the utilization of EC numbers in these papers since new EC numbers are defined only *after* a paper has reported the novel activity. To manage this constant lag between the report of CAZyme reactions and the assignment of a novel EC number by the IUBMB, the CAZy database creates 'open' EC numbers such as 3.2.1.- where the dash character represents a precise activity. For example, since 2018 several β-carrageenases have been characterized but no corresponding EC number was created yet (18,19). As a consequence, we have to follow updates of the EC system on a regular basis to identify new EC numbers that correspond to those 'open' EC numbers. At the time of writing ∼100 such activities in ∼400 proteins are waiting for a formal EC number.

New display of information

We have made several changes to our web pages in order to improve access to information. The first major change results from the above efforts on functional data capture. Starting with the GHs and the PLs, we now display a link (to PubMed preferentially, or a DOI) to the paper(s) that describe the function of the enzyme (Figure 1; box 6). We hope that this initiative will also encourage colleagues to inform us of their future CAZyme characterization results as well as the old ones we have missed. Linking each activity to the relevant CAZy entry is a colossal work as each and every entry must be reviewed, and several are removed during this process, i.e. when the evidence for function is insufficient or missing. Indeed, some functional characterizations were integrated in CAZy at a time where the paucity of biochemical data made us consider more indirect proofs (e.g. personal communications prior to publication and convincing annotations from GenBank/Swiss-Prot). We began this work a year ago and as of September 2020, we now provide a literature reference to families for 96% of GH and PL families (76% of the functionally characterized enzymes in these classes). Entries with an activity not yet curated, and therefore possibly subject to changes, carry an EC number but do not carry a bibliographic link. We will complete the work

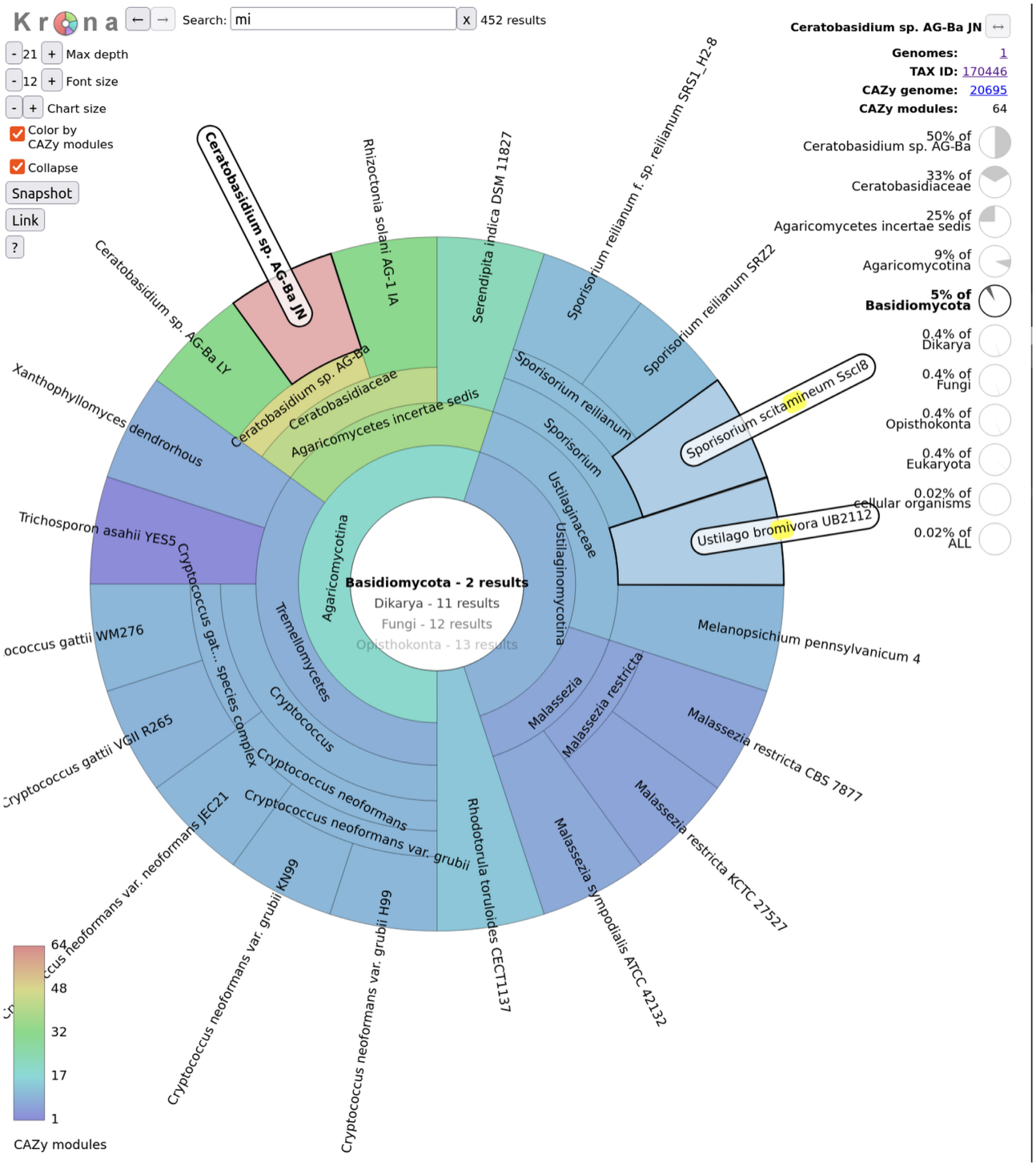

Figure 2. Krona chart browsing into family GH5. After navigation through taxonomic levels, this figure shows the display of the Basidiomycota, in bold in the center. The central part also recalls the results of the text search performed from the top form with the string 'mi'. Results are highlighted both within the Basidiomycota, two matching genomes at right have their sector highlighted and the 'mi' string in yellow background, as well as in the levels above as the center shows for example 11 results in Dikarya. All sectors are color-coded according to the number of GH5 modules in the complete genome, and one sector was selected, the genome of *Ceratobasidium* sp. AG-Ba JN. Once a sector is selected, various links appear at the top right, to CAZy or NCBI, and multiple pie charts illustrate the representativeness of this genome in its taxonomic lineage.

on GHs and PLs before extending the same principle to other CAZyme categories. Moreover, CAZy homepage now displays a table that highlights the most recently integrated functions/publications at each update.

The second main change applies to the family webpages (and similarly for subfamilies which are built on the same model). Previously, browsing through family members was difficult as protein accessions were split into taxonomic tabs, itself divided into multiple table sheets. Therefore, we have simplified the display and a family page now only displays members that have been characterized either structurally or functionally. Another reason for this simplification is the multiplication of programs and users regularly scanning the whole CAZy website to copy its data, slowing down its usage. The complete lists of protein accessions are now available as downloadable text files. The links to these files are shown in Figure 1, for each family (box 4) as well as for all families simultaneously (box 2). In a simplification process, we have changed the display of PDB codes, by merging chains with an identical underlying sequence in a single reference *ID[comma-separated chains]*. Finally, the subfamily pages were modified to only display the functional information of the subfamily and not that of the whole family (Figure 1; box 3).

The last change in the new version of our website concerns the access to genome information. In addition to the current access by alphabetical listings for Bacteria, Archaea, Eukaryota and Viruses, we now provide access via Krona charts (Figure 2). These circular multilayer diagrams allow a dynamic navigation across the different taxonomic levels along with a user-friendly text search to facilitate the navigation among matching strings. All taxonomic levels (chart sectors) are color-coded according to the average number of total CAZy modules, and HTML links are provided to both the NCBI taxonomy and to a flat file that lists all genomes in this taxonomic level. When finally clicking over a genome (last taxonomic level), an additional link appears to redirect towards the CAZy genome page (for the detailed CAZy annotation). We also integrated Krona charts specific to each family (on each CAZy family webpage; Figure 1, box 5) allowing the visualization of the family member distribution across the taxonomic levels, outlined by the color-scale (Figure 2). Krona charts will thus enhance taxonomy-guided comparative analysis.

CONCLUDING REMARKS

We hope that with the novel display of information and the recent added features presented here, the CAZy database will continue being a useful resource to the community in the years to come. The maintenance of this >20-year old database does not rely (and never did) on dedicated funding as data capture, curation and organization does not appear to be a funding priority for French scientific decision makers. Yet many external bots download the entirety of our site almost daily, and this frequently slows down its operation. While this systematic download of our resource is a sign of usefulness in an open science context, it may jeopardize our long term efforts by creating competing and funded resources based on our hard curation efforts. Fortunately, the continuation of the CAZy database has benefitted from multiple scientific collaborations for omics data analyses and from the amazing acceptance and support of the (glyco)biological and biotechnological communities interested in CAZymes, which we acknowledge here with gratitude.

FUNDING

No external funding.
Conflict of interest statement. None declared.

REFERENCES

1. Sayers,E.W., Cavanaugh,M., Clark,K., Pruitt,K.D., Schoch,C.L., Sherry,S.T. and Karsch-Mizrachi,I. (2021) GenBank. *Nucleic Acids Res.*, **49**, D92–D96.
2. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
3. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
4. Lombard,V., Golaconda Ramulu,H., Drula,E., Coutinho,P.M. and Henrissat,B. (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.*, **42**, D490–D495.
5. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
6. Brown,M.E. and Chang,M.C.Y. (2014) Exploring bacterial lignin degradation. *Curr. Opin. Chem. Biol.*, **19**, 1–7.
7. Akoh,C.C., Lee,G.-C., Liaw,Y.-C., Huang,T.-H. and Shaw,J.-F. (2004) GDSL family of serine esterases/lipases. *Prog. Lipid Res.*, **43**, 534–552.
8. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Christie,C.H., Dalenberg,K., Di Costanzo,L., Duarte,J.M. *et al.* (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.
9. Barrett,A.J. (1997) Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB). enzyme nomenclature. recommendations 1992. slement 4: corrections and additions (1997). *Eur. J. Biochem.*, **250**, 1–6.
10. White,J. (2020) PubMed 2.0. *Med. Ref. Serv. Q.*, **39**, 382–387.
11. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M. and Bairoch,A. (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol. Clifton NJ*, **406**, 89–112.
12. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
13. Ndeh,D., Rogowski,A., Cartmell,A., Luis,A.S., Baslé,A., Gray,J., Venditto,I., Briggs,J., Zhang,X., Labourel,A. *et al.* (2017) Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nature*, **544**, 65–70.
14. Luis,A.S., Briggs,J., Zhang,X., Farnell,B., Ndeh,D., Labourel,A., Baslé,A., Cartmell,A., Terrapon,N., Stott,K. *et al.* (2018) Dietary pectic glycans are degraded by coordinated enzyme pathways in human colonic Bacteroides. *Nat. Microbiol.*, **3**, 210–219.
15. Cuskin,F., Lowe,E.C., Temple,M.J., Zhu,Y., Cameron,E., Pudlo,N.A., Porter,N.T., Urs,K., Thompson,A.J., Cartmell,A. *et al.* (2015) Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism. *Nature*, **517**, 165–169.
16. Helbert,W., Poulet,L., Drouillard,S., Mathieu,S., Loiodice,M., Couturier,M., Lombard,V., Terrapon,N., Turchetto,J., Vincentelli,R. *et al.* (2019) Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 6063–6068.
17. Glasgow,E.M., Kemna,E.I., Bingman,C.A., Ing,N., Deng,K., Bianchetti,C.M., Takasuka,T.E., Northen,T.R. and Fox,B.G. (2020) A structural and kinetic survey of GH5_4 endoglucanases reveals

determinants of broad substrate specificity and opportunities for biomass hydrolysis. *J. Biol. Chem.*, **295**, 17752–17769.
18. Schultz-Johansen,M., Bech,P.K., Hennessy,R.C., Glaring,M.A., Barbeyron,T., Czjzek,M. and Stougaard,P. (2018) A novel enzyme portfolio for red algal polysaccharide degradation in the marine bacterium paraglaciecola hydrolytica S66T encoded in a sizeable polysaccharide utilization locus. *Front. Microbiol.*, **9**, 839.
19. Hettle,A.G., Hobbs,J.K., Pluvinage,B., Vickers,C., Abe,K.T., Salama-Alber,O., McGuire,B.E., Hehemann,J.-H., Hui,J.P.M., Berrue,F. *et al.* (2019) Insights into the κ/ι-carrageenan metabolism pathway of some marine *Pseudoalteromonas* species. *Commun. Biol.*, **2**, 474.