W110–W115 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* *Published online 14 May 2020*
doi: 10.1093/nar/gkaa375

Conserved unique peptide patterns (CUPP) online platform: peptide-based functional annotation of carbohydrate active enzymes

Kristian Barrett[1], Cameron J. Hunt[1], Lene Lange[2] and Anne S. Meyer[1,*]

[1]Protein Chemistry and Enzyme Technology Section, DTU Bioengineering, Department of Biotechnology and Biomedicine, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark and [2]LLa-BioEconomy, Research & Advisory, 2500 Valby, Denmark

Received March 13, 2020; Revised April 28, 2020; Editorial Decision April 29, 2020; Accepted May 12, 2020

ABSTRACT

The CUPP platform includes a web server for functional annotation and sub-grouping of carbohydrate active enzymes (CAZymes) based on a novel peptide-based similarity assessment algorithm, i.e. protein grouping according to Conserved Unique Peptide Patterns (CUPP). This online platform is open to all users and there is no login requirement. The web server allows the user to perform genome-based annotation of carbohydrate active enzymes to CAZy families, CAZy subfamilies, CUPP groups and EC numbers (function) via assessment of peptide-motifs by CUPP. The web server is intended for functional annotation assessment of the CAZy inventory of prokaryotic and eukaryotic organisms from genomic DNA (up to 30MB compressed) or directly from amino acid sequences (up to 10MB compressed). The custom query sequences are assessed using the CUPP annotation algorithm, and the outcome is displayed in interactive summary result pages of CAZymes. The results displayed allow for inspection of members of the individual CUPP groups and include information about experimentally characterized members. The web server and the other resources on the CUPP platform can be accessed from https://cupp.info.

INTRODUCTION

Large efforts have been put into alignment- and structure-based classification of carbohydrate-active enzymes (CAZymes) as well as creation of CAZyme families, and several hundred CAZyme families exist in the CAZy database today (1). A range of different tools exist for annotating enzymes to general CAZy family-level through Hidden Markov Models (2–4). Unfortunately, several of the CAZyme families harbor members that catalyze different reactions. The families thus often represent enzymes of very diverse molecular function. Sub-classification of the members within a family (or subfamilies) into groups of functionally similar enzymes is therefore highly desirable (5–7). To accomplish this goal and help promote further understanding of CAZyme catalyzed synthesis and degradation of carbohydrates reliable and robust functional annotation of CAZymes is of utmost importance. The protein databases are expanding in all directions along with more and more complex bioinformatics assessments (8). At the same time, the data numbers and the functional complexity within the individual CAZy families continuously increase based on all-versus-all BLAST (9) or automated phylogenetic tree assessments (10).

It is currently demanding to annotate complete genomes and metagenomes for identification of CAZymes, and in particular to determine if several enzymes of the same family are likely to be functionally similar or distinctly different (11,12). Also, systematic exploration and comparison of the enzymes within even a single CAZy family is an overwhelming task (8).

The detailed grouping provided by CUPP relies on an unsupervised peptide-based clustering algorithm that offers a systematical approach for exploration of CAZymes based on amino-acid sequences, but which can identify ORFs in DNA sequences and translate the mRNA sequences into proteins (13). In brief, the CUPP algorithm divides the protein sequences into smaller peptide fragments, specifically eight amino acids in length of which two are ambiguous. These fragments are then compared to a library of conserved peptide fragments, currently comprising 10,753 CUPP groups, built from assessment of more than one million proteins, covering all enzymes in the CAZy database (www.cazy.org). The purpose of the CUPP annotation web server is to provide an easy-to-use, freely accessible, robust tool to functionally annotate, sub-divide, and compare CAZymes, while simultaneously providing an overview of

*To whom correspondence should be addressed. Email: asme@dtu.dk

the functionally unexploited corners of each existing CAZy family (13).

The nuanced subdivision of the CAZy families, included in the CUPP web server, enables rapid functional annotation and identification of CAZymes. The user interface displays the nuanced subdivision of CAZyme proteins accomplished by the CUPP technology. By allowing CUPP groups not having any characterized members, the annotation furthermore enables the user to directly identify CAZymes having potential novel function via differentiation of the known and the unknown CAZymes. This service can facilitate the upload of a whole metagenome, identify all the Open Reading Frames, translate them into amino acid sequences and annotate them to CUPP groups currently spanning over 300 CAZy enzyme families http://www.cazy.org/.

In the present work and in the online platform presented, we employ a new CUPP library with all members of CAZy present ultimo 2019 included with a total of 10,753. Here, we also include a validation of the transfer of CUPP group numbers from the 2018 CUPP library version (13) onto the CUPP groups numbers of the 2020 version. The transfer of group numbers and the validation have in particular focused on ensuring that the CUPP group numbers are consistent for future versions as it is planned to update the CUPP library once a year to include all the newest research results and all CAZy database updates in the models. The newest version of the models will be available on the web server. The web server will be maintained for a minimum of 5 years with planned implementation of new features.

RESULTS AND DISCUSSION

Relevance of nuanced functional subdivision of CAZY families

The annotation of CAZymes to CUPP groups can provide biologically relevant information and background knowledge for enzyme application technology on otherwise functionally diverse CAZy families. An applied example is the enzymatic degradation of individual linkages of the abundant and complex plant cell wall carbohydrate rhamnogalacturonan II (RGII) (Figure 1).

Each of the involved CAZymes belonging to a CAZy family were found by annotation via the online CUPP annotation web server and the CUPP groups and be inspected by browsing the families in the online CUPP platform. For the two esterase activities (CE), one of them, the methyl esterase, is present in the CAZy database as 'non classified', i.e. CE0. The other esterase, which is in the RGII-associated PUL of *Bacteroides thetaiotaomicron*, is not listed on the family pages of the CAZy database, and was therefore not annotated as indicated by 'CE0*, CUPP Gr. 0*' (Figure 1) (where Gr. is Group). Additionally, the members of GH2 and GH78 were annotated by CUPP to belong to distinct groups. The individual enzymes are all in individual CUPP groups, thus functionally separating these CAZymes. For example, the four CAZymes of GH2 (Figure 1) act as a specific β-galactosidase, a β-D-galacturonidase, a β-D-glucuronidase and an α-arabinopyranosidase, respectively, according to increasing CUPP group numbering. Such a nuanced functional separation is extremely useful for elucidating and interpreting enzymatic degradation of complex carbohydrate substrates.

Basic features

The CUPP web server is an online annotation tool, accessible with any modern web browser on tablets, laptops and mobile phones. The web server supports any FASTA file containing genomic nucleotide sequences or amino acid protein sequences. The upload of gz-compressed FASTA files is supported. The query proteins can be provided by copy-and-paste, specified by a path or by drag-and-drop. The result page displays interactive tables and provides access to different graphical overview presentations of the annotations, see Figure 2.

The key points highlighted in this paper include:

- CUPP allows for molecular function annotation of CAZymes directly from genomic sequences.
- Functional annotation of CAZymes is now available through the new CUPP web server.
- Nuanced CAZyme annotations of user-queries (both DNA and protein sequences) are displayed in an intuitive and interactive way on the online platform.
- A new CUPP library is available and the group numbering is robust over time across different versions of the library.

In order to be a reliable reference today and after future updates, the subgroups of a former organization should be kept through time with a unique group name. This reliability is validated in this paper by the robustness of the clustering between the CAZymes included in the 2018 CUPP library and the current 2020 version of the CUPP library. Since the 2018 version of the CUPP library the number of enzymes has increased by ∼40% from 485,382 non-redundant domain regions to 683,873, without considering proteins in new families or CAZymes without a catalytic domain (i.e. CBMs only). This increase means that 260,319 new non-redundant domains has been added since May 2018 (246,060 of the non-redundant domains are non-fragments).

Features of the Web server - results display interface

The submission page provides basic validation of the input sequences and files as well as options for the type of submission and the option for email notification of job status. By clicking the Submit button, the user will arrive at a loading page that will automatically listen for changes to the job status and display results/errors accordingly (Figure 2A). This loading page also states the estimated time of completion and provides a link to the specific url for the current job for later access. Similar information is also sent to the user's email if this option is selected on the submission page. The link will display an interactive overview of the annotations with filtering possibilities in a Result Page (Figure 2B). The Result Page includes a dynamic table and graphical representations for summarizing the annotations

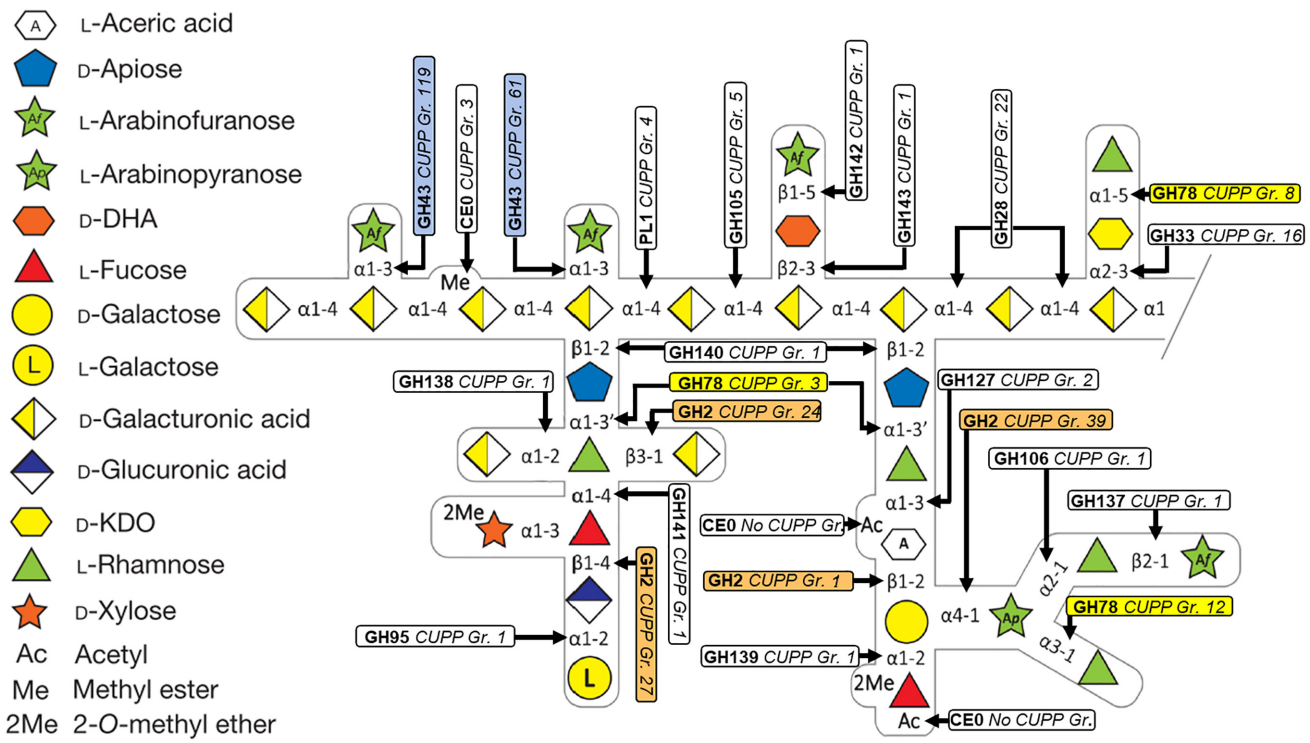

Figure 1. Enzymatic degradation of rhamnogalacturonan II (RGII). Each of the boxes represent an individual peptide signature-group of CAZymes belonging to a particular CAZy family. White boxes indicate CAZy families occurring once for the degradation of RGII, whereas colored boxes indicate different cases where different members of a CAZy family act on distinct linkages in RGII. RGII structure and layout adapted from Ndeh *et al.* (14).

in an intuitive and convenient manner. Integrated as part of the filtering, a bar chart is displayed on top of the table as the individual entries are selected in the table. Furthermore, a series of filters, searching, sorting and selecting tools allow for the specific retrieval of annotations by the user.

A typical annotation result on the web server includes the header information from the submitted sequence, linked to the predicted family, subfamily, CUPP group and molecular function (EC number). In an applied case, the genomic query protein list of *B. thetaiotaomicron* strain VPI-5482 was annotated on the web server (Figure 2B) (a subset of the annotations were previously put in the context of RGII (Figure 1)). Additional information including the score, coverage and domain of the annotated region is also provided (Figure 2B). Similarly, a bar chart and interactive table is shown on family pages (bottom of Figure 2D).

The CUPP platform also includes a browse option for the individual CAZy families, subfamilies and CUPP groups with a specific url, linking to each kind (Figure 2C). These pages display only the relevant members in an interactive and dynamic manner, similar to the annotation result pages. In this way, it is possible to get insight and biological meaningful information about the members of for example a particular CUPP group and the associated meta-data of each member. Additionally, the ability to retrieve selected members and their meta-data as a csv file, is also provided. Currently, a phylogenetic tree is included for the GH30 page embedded through iTol (15) and more families will be added in a near future (Figure 2D).

Custom export of results

The results can be downloaded as a full FASTA file or tab separated results file containing all the query proteins annotated as CAZymes. Additionally, the results of the annotations can be filtered based on families, subfamilies, CUPP groups, EC number or a search of the name of the query proteins. The current annotations displayed in the table after filtering can be downloaded as a subset of all annotations, allowing easy access to desired annotations. Additionally, the summary of the current annotations can be downloaded. These summaries are also dynamically displayed as a histogram, which can also be downloaded as displayed.

Consistency of CUPP annotations

The CAZymes added to the CAZy database since 2018 serve as a dataset independent of training sets. This dataset includes 246,060 non-redundant CAZymes. The overall family annotation performance of the CUPP analysis, using the 2018 CUPP library version, resulted in a sensitivity of 89.8%, whereas the sensitivity with the 2020 library version was improved to a sensitivity of 95.2%. The performance of the former 2018 version of the CUPP library as compared to dbCAN2 including HMM, Hotpep, and DIAMOND CAZy family annotation was reported when the CUPP technology was published in 2019 (13). It was demonstrated that CUPP runtime, *F*-score, sensitivity and precisions of family and subfamily annotations either matched or represented an improvement compared to the state-of-the-

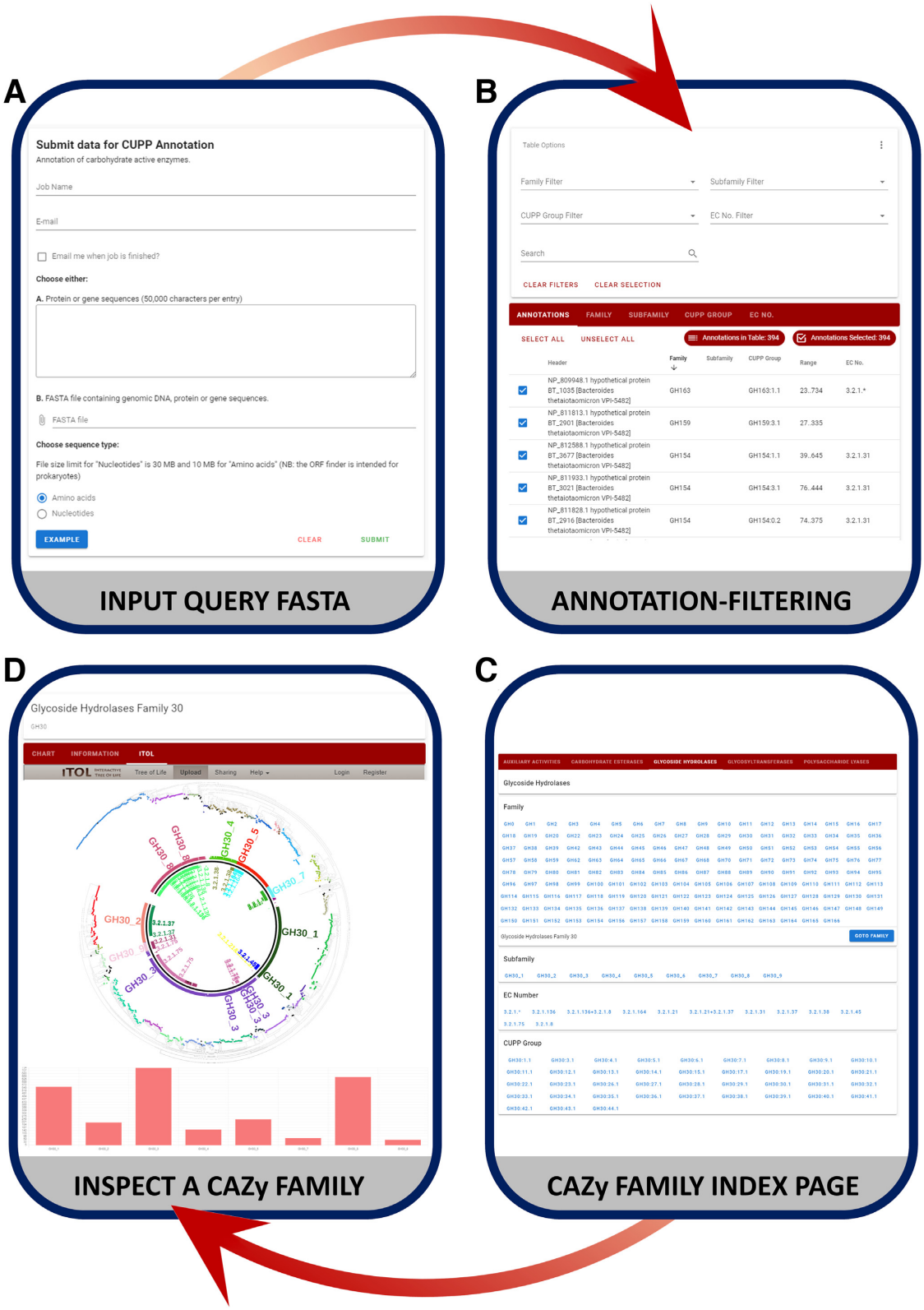

Figure 2. The user interface of the CUPP online platform and the annotation web server. (**A**) The submission page to specify the query sequences to annotate. (**B**) The dynamic table and filtering options of the table and the graphical representations of the CAZyme annotations for proteins of *B. thetaiotaomicron*. (**C**) An index page for browsing the individual CAZy families, subfamilies, and CUPP groups. (**D**) An example of the CAZy family page of GH30 with a phylogenetic tree and a bar chart representation of the family members.

Table 1. Comparison of CUPP versus eCAMI and dbCAN2 for CAZy family annotation provided as the *F*-score for selected genomes with curated CAZyme annotations. The table includes data obtained on seven annotated genomes of diverse taxonomical origin, namely: *Botrytis cinerea* B05.10, *Malassezia restricta* KCTC 27527, *Vigna angularis* Jingnong6, *Bacteroides thetaiotaomicron* VPI-5482, *Bifidobacterium bifidum* NCTC13001, *Caulobacter segnis* ATCC 21756 and *Xanthomonas campestris* ATCC 33913

			dbCAN2 tools *F*-score (+2 tools)			
Origin of genomes	*F*-score CUPP	*F*-score eCAMI	Hotpep	dbCAN	Diamond	CAZymes/proteins
B. cinerea	0.96	0.89	0.88	0.94	0.97	530/13703
M. restricta	0.95	0.91	0.88	0.91	0.98	82/4406
V. angularis	0.96	0.91	0.93	0.93	0.95	1395/37972
Eukaryotes: average	0.96	0.90	0.90	0.93	0.96	
B. thetaiotaomicron	0.97	0.88	0.95	0.83	0.96	417/4817
B. bifidum	0.98	0.90	0.88	0.91	0.94	65/1744
C. segnis	0.98	0.94	0.90	0.95	0.97	118/4103
X. campestris	0.98	0.97	0.94	0.96	0.98	156/4179
Bacteria: average	0.98	0.92	0.92	0.91	0.96	
Total average	0.97	0.91	0.91	0.92	0.96	

art tools (13). The original benchmarking result thus validated CUPP as a solid algorithm enabling peptide-based functional annotation directly from assembled genomic DNA. Furthermore, a renewed comparison of CUPP to other existing tools, including eCAMI and dbCAN2 on the six full genomes used for the original CUPP validation plus the *B. thetaiotaomicron* genome used as an example, confirms the robust performance of the CUPP algorithm (Table 1).

The *F*-score data, sensitivity, and precisions of family and subfamily annotations of the 2020 CUPP library were determined here using the same approach as used in the original CUPP publication (13), including CAZymes belonging to AA0, CE0, GH0, GT0 and PL0 in addition to all current classified CAZy families also counting fragments. Notably, as recommended by the dbCAN2 documentation (2), a CAZy family annotation was also only considered here when at least two out of three dbCAN2 tools had identified a given query protein to be a CAZyme. For the eCAMI annotations, the published library was used (16). The resulting overall performance of CUPP, including the F-score, was superior compared to eCAMI. Furthermore, when eCAMI was allowed to run on eight CPU's, the wall time of CUPP annotation (using one CPU) was more than eight times faster than the wall time of eCAMI annotation (using Intel Xeon Gold 6126 (2.60 GHz) CPU's on HDD drives). However, the eCAMI tool had a RAM usage of up to 12GB RAM for the eight CPU's, whereas the CUPP annotation used up to 16GB RAM with the 2020 version of the CUPP library. When comparing the CPU hours instead of wall-time CUPP was more than 60 times faster than eCAMI. In addition, the CUPP tool can batch annotate a folder of genomes, thus only one loading of the CUPP library is needed for processing of multiple queries.

Robustness of CUPP clustering and benchmarking

The curated proteins of the CAZy database has been obtained 13 December 2019 and is referred to as the 2020 version, whereas the 2018 version has been published (13). The robustness is determined based on a Jaccard score, where the group members of the 2018 version ideally should we organized into similar groups in the 2020 version. The equation used for the robustness was the following:

$$Robustness = \sum_{i=1}^{n} \left(\sum_{i=1}^{m} \frac{C_n \cap C_m}{C_n \cup C_m} \cdot \frac{total\ group\ members_n}{total\ family\ members} \right)$$

where m = the number of protein groups in the new version; n = the number of protein groups in the old version; C = cluster members as set of identifiers; *total group members* is the number of representative members of the protein group, whereas the *total family members* is the total members of representative sequences in the family, only including those entries present in both versions. This robustness is compared to another peptide-based clustering method (16). The robustness of CUPP clustering was 97.5% when allowing further division of existing CUPP groups whereas the similar score for eCAMI gave 85.5%. This robustness of CUPP group numbering ultimately means that continued updating of the CUPP library, to keep it aligned with new updates of the CAZy database, does not come with the high price of losing reproducibility. For this reason, we encourage researchers in the field of CAZymes to state the CUPP group number when mentioning a CAZyme in a publication as this will ease the interpretation of the result and ideally provide the characterization information to the CAZy.org database preferable with a reference (1).

Design and implementation

The CUPP online platform is primarily written in a modern JavaScript framework and hosted as a single page application on a Google Cloud app engine. The CUPP annotation server (for CAZy annotations) is hosted in the high performance clusters of Technical University of Denmark with scalable computational resources, initially fixed a eight CPU's. On average, one CPU can annotated about 100 proteins pr. second, giving more than eight million annotations pr. day pr. CPU. The CUPP online platform has been tested with major modern browsers such as Google Chrome (80+), Mozilla Firefox (61+), Microsoft Edge (42+) and Safari (12+).

DATA AVAILABILITY

The Conserved Unique Peptide Patterns online platform is freely available at the https://cupp.info. The welcome

pages provides access to the submission entry point for the CUPP webserver, which provides custom CAZy annotations for a user-defined query. Additionally, an overview of the the CAZy members of the individual CUPP groups, CAZy families and subfamilies can be browsed to or directly accessed through a url. For example, the link for family GH30 is https://cupp.info/family/GH30. The CUPP program can also be downloaded from https://bioengieering.dtu.dk/cupp for offline usage as a python script directly functional on Windows, Linux and MacOS operating systems, documentation is provided in the readme file.

ACKNOWLEDGEMENTS

We acknowledge the advice from IT Manager Mads Gleerup Christensen, DTU Bioengineering regarding capacity and data storage solutions.

FUNDING

Technical University of Denmark; H.C. Ørsted CO-FUND Postdoc Program [Marie Sklodowska-Curie grant agreement no. 713683] at the Technical University of Denmark. Funding for open access charge: Technical University of Denmark.
Conflict of interest statement. None declared.

REFERENCES

1. 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**, 490–495.
2. Zhang,H., Yohe,T., Huang,L., Entwistle,S., Wu,P., Yang,Z., Busk,P.K., Xu,Y. and Yin,Y. (2018) DbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.*, **46**, W95–W101.
3. Mistry,J., Finn,R.D., Eddy,S.R., Bateman,A. and Punta,M. (2013) Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.*, **41**, e121.
4. Li,W., Cowley,A., Uludag,M., Gur,T., Mcwilliam,H., Squizzato,S., Park,Y.M., Buso,N. and Lopez,R. (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.*, **43**, 580–584.
5. St John,F.J., Hurlbert,J.C., Rice,J.D., Preston,J.F. and Pozharski,E. (2011) Ligand bound structures of a glycosyl hydrolase family 30 glucuronoxylan xylanohydrolase. *J. Mol. Biol.*, **407**, 92–109.
6. Mewis,K., Lenfant,N., Lombard,V. and Henrissat,B. (2016) Dividing the large glycoside hydrolase family 43 into subfamilies: a motivation for detailed enzyme characterization. *Appl. Environ. Microbiol.*, **82**, 1686–1692.
7. Busk,P.K. and Lange,L. (2013) Function-based classification of carbohydrate-active enzymes by recognition of short, conserved peptide motifs. *Appl. Environ. Microbiol.*, **79**, 3380–3391.
8. 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.
9. Viborg,A.H., Terrapon,N., Lombard,V., Michel,G., Czjzek,M., Henrissat,B. and Brumer,H. (2019) A subfamily roadmap of the evolutionarily diverse glycoside hydrolase family 16 (GH16). *J. Biol. Chem.*, **294**, 15973–15986.
10. Jones,D.R., Thomas,D., Alger,N., Ghavidel,A., Douglas Inglis,G. and Wade Abbott,D. (2018) SACCHARIS: An automated pipeline to streamline discovery of carbohydrate active enzyme activities within polyspecific families and de novo sequence datasets. *Biotechnol. Biofuels*, **11**, 27.
11. Benoit,I., Culleton,H., Zhou,M., DiFalco,M., Aguilar-Osorio,G., Battaglia,E., Bouzid,O., Brouwer,C.P.J.M., El-Bushari,H.B.O., Coutinho,P.M. *et al.* (2015) Closely related fungi employ diverse enzymatic strategies to degrade plant biomass. *Biotechnol. Biofuels*, **8**, 107.
12. Gruninger,R.J., Nguyen,T.T.M., Reid,I.D., Yanke,J.L., Wang,P., Abbott,D.W., Tsang,A. and McAllister,T. (2018) Application of transcriptomics to compare the carbohydrate active enzymes that are expressed by diverse genera of anaerobic fungi to degrade plant cell wall carbohydrates. *Front. Microbiol.*, https://doi.org/10.3389/fmicb.2018.01581.
13. Barrett,K. and Lange,L. (2019) Peptide-based classification and functional annotation of carbohydrate-active enzymes by conserved unique peptide patterns (CUPP). *Biotechnol. Biofuels*, **12**, 102.
14. 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.
15. Letunic,I. and Bork,P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.*, **47**, W256–W259.
16. Xu,J., Zhang,H., Zheng,J., Dovoedo,P. and Yin,Y. (2020) eCAMI: simultaneous classification and motif identification for enzyme annotation. *Bioinformatics*, **36**, 2068–2075.