Downloaded from orbit.dtu.dk on: Apr 24, 2025

DTU Library

Conserved unique peptide patterns (CUPP) online platform 2.0: implementation of +1000 JGI fungal genomes

Barrett, Kristian; Hunt, Cameron J.; Lange, Lene; Grigoriev, Igor V.; Meyer, Anne S.

Published in:
Nucleic Acids Research

Link to article, DOI:
10.1093/nar/gkad385

Publication date:
2023

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

Link back to DTU Orbit

Citation (APA):
Barrett, K., Hunt, C. J., Lange, L., Grigoriev, I. V., & Meyer, A. S. (2023). Conserved unique peptide patterns (CUPP) online platform 2.0: implementation of +1000 JGI fungal genomes. *Nucleic Acids Research*, *51*(W1), W108-W114. Article gkad385. https://doi.org/10.1093/nar/gkad385

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

W108–W114 *Nucleic Acids Research, 2023, Vol. 51, Web Server issue* *Published online 22 May 2023*
https://doi.org/10.1093/nar/gkad385

Conserved unique peptide patterns (CUPP) online platform 2.0: implementation of +1000 JGI fungal genomes

Kristian Barrett [1,2], Cameron J. Hunt[1], Lene Lange[4], Igor V. Grigoriev [2,3] and Anne S. Meyer [1,*]

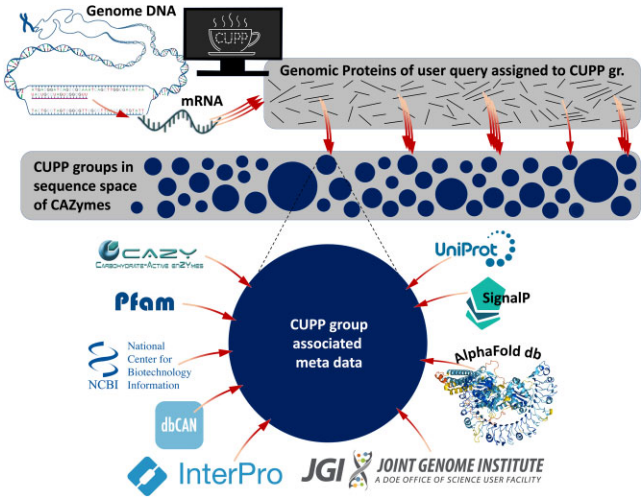
[1]Department of Biotechnology and Biomedicine, Technical University of Denmark, Søltofts Plads, DK-2800 Kgs. Lyngby, Denmark, [2]Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [3]Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 94720, USA and [4]BioEconomy, Research & Advisory, Valby DK-2500, Denmark

Received March 26, 2023; Revised April 27, 2023; Editorial Decision April 28, 2023; Accepted April 29, 2023

ABSTRACT

Carbohydrate-processing enzymes, CAZymes, are classified into families based on sequence and three-dimensional fold. Because many CAZyme families contain members of diverse molecular function (different EC-numbers), sophisticated tools are required to further delineate these enzymes. Such delineation is provided by the peptide-based clustering method CUPP, Conserved Unique Peptide Patterns. CUPP operates synergistically with the CAZy family/subfamily categorizations to allow systematic exploration of CAZymes by defining small protein groups with shared sequence motifs. The updated CUPP library contains 21,930 of such motif groups including 3,842,628 proteins. The new implementation of the CUPP-webserver, https://cupp.info/, now includes all published fungal and algal genomes from the Joint Genome Institute (JGI), genome resources MycoCosm and PhycoCosm, dynamically subdivided into motif groups of CAZymes. This allows users to browse the JGI portals for specific predicted functions or specific protein families from genome sequences. Thus, a genome can be searched for proteins having specific characteristics. All JGI proteins have a hyperlink to a summary page which links to the predicted gene splicing including which regions have RNA support. The new CUPP implementation also includes an update of the annotation algorithm that uses only a fourth of the RAM while enabling multi-threading, providing an annotation speed below 1 ms/protein.

GRAPHICAL ABSTRACT



INTRODUCTION

Enzymes that catalyze modification of carbohydrates, i.e. CAZymes, are generally highly specific due to the huge stereochemical diversity of their substrates. Based on sequence and three-dimensional fold, CAZymes are classified into families covering 5 types of catalytic reactions (glycoside hydrolases, glycosyl transferases, polysaccharide lyases, carbohydrate esterases, and 'auxiliary activities' which are mainly redox enzymes) (1). So far, about 400 CAZy families have been created, curated, and kept up-to-date for decades by the dedicated work of the CAZy group at Aix Marseille University, France (1–3). The CAZy group provides robust family and subfamily delineations while keeping track of relevant, characterized enzymes. As more genomes become sequenced, more CAZyme members are added into each family, and in cases of new enzyme activity descriptions, i.e.

*To whom correspondence should be addressed. Tel: +45 4525 2600; Email: asme@dtu.dk

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article/51/W1/W108/7175339 by guest on 03 October 2024

when new molecular function information becomes available, new families with potentially novel structure-function relations are created (1).

Several CAZy families comprise members with distinct molecular functions, meaning that they have unique specificities as described by an EC number approved by the International Union of Biochemistry and Molecular Biology Enzyme Commission. However, even though members of the same CAZy family have the same protein fold, the fold similarity does not always mean that their molecular functions are the same, which is why several CAZy families cover different enzyme functions as specified by EC numbers. Thus, an approach for automated and dynamic subdivision for capturing differences is a desirable supplement to the robust family and subfamily delineations provided in the CAZy database. A range of CAZy family annotation services based on the HMM's of Pfam (4), InterProScan (5) or dbCAN (6) already exist. Yet, although some efforts have been successful using SACCHARIS (7) and eggNOG (8,9), the establishment of phylogenetic trees for creation of branches of distinct molecular function via genome annotation is not trivial; it is thus a major effort even for a trained bioinformatician to manually divide the enzymes (sequences) of large families into relevant areas (7). The CUPP clustering and annotation tool was first launched as a stand-alone algorithm (10), but subsequently the webserver and database (https://cupp.info/) were published (11).

Here, we present an updated version of the CUPP-webserver (https://cupp.info/), which features an improved overall user interface, and not least inclusion of all the published JGI fungal and algal genomes into the CAZy family architecture to ease genome comparison amongst these genomes. Annotation of CAZymes from fungal and algal sequences is considered a new frontier exploration element for novel enzyme discoveries. In the updated version of the CUPP-webserver, the new features include options for direct genome comparison from a user query to the CUPP groups in the pre-annotated database which, in addition to 44,544 other strains, now includes 1418 published fungal/algal genomes (see Supplementary material for references to each JGI genome) from the JGI MycoCosm (12) and JGI PhycoCosm (13). This inclusion thus enables users to browse the JGI genomes with the user-friendly interface for the advanced querying, searching, filtering and retrieval of the CUPP annotated CAZy database. In this way, the updated CUPP-webserver gives access to visualizations of protein structure, domains, sequence alignments and summary charts for CUPP groups and queries on the database. The CUPP-webserver will be maintained for minimum 5 years with the newest version of the models available.

MATERIALS AND METHODS

Expansion of CAZy families

The protein accessions were obtained from the CAZy.org database (1) on November 2022 and sequences of the CAZy accessions were downloaded from NCBI nr db version 63 (14). All proteins which have a known molecular function or crystal structure listed in CAZy.org were treated as seeds along with a single member of each unknown group of the previous CUPP database (11). Each family was processed individually on our DTU High Performance Cluster in parallel in the following steps: 1) The seed proteins were truncated individually to their catalytic domains by dbCAN (6) using HMMER3 (15) and each domain retrieved up to 5000 proteins from the NCBI nr database using Diamond BLAST (16) with default setting. Additionally, the expansion was also done on the combined list of protein of the JGI genomes, i.e. all proteins from all published genomes imported from JGI MycoCosm and PhycoCosm resources, were added. Secondly, additional CAZymes were predicted in the published JGI genomes using the former CUPP library to highlight additional CAZymes. These additional CAZymes are in https://cupp.info/ marked as 'MycoCosm+' or 'PhycoCosm+', if they indeed become a member of the CUPP group after the all-vs-all clustering. This expansion included the retrieval of more than ten million possible CAZymes of which about 3,842,628 made it into one of the 21,930 CUPP groups.

For the annotation benchmark analysis, the family and EC annotations for CUPP were performed using the new CUPP-webserver including only hits with a significance above the default significance score of 5. The eggNOG 5.0 annotations were performed on the online webserver with default settings (9). The dbCAN annotations were performed using the online webserver (6) using default settings. The sensitivity is defined as the fraction of the true positives (CAZy families or EC numbers) found by each program. The precision is calculated per protein as the number of true positives divided by the total positives (i.e. the sum of true and false positives) for the particular protein. The sensitivity and precision results are presented as the average for the query proteins assessed.

The catalytic domains of each protein were identified using dbCAN with a less strict e-value (e-value < 0.001) if they originated from www.CAZy.org whereas sequences retrieved by BLAST from NCBI required a more strict e-value, namely an e-value $<10^{-15}$ to be accepted. The collection of catalytic domains was reduced by CDHIT (17) (setting the clustering threshold to 90% identity) to remove nearly identical proteins. Redundant sequences retrieved by BLAST from NCBI, were not included into the CUPP-webserver (https://cupp.info/). The collection of representative domains was subjected to all-vs-all CUPP clustering (10) to identify sequence-motif within subbranches of each family, hence placing the JGI genome proteins in distinct subbranches of the family. Motif groups without any official CAZy family member (according to www.CAZy.org) were moved from the library as the expansion was performed to capture the diversity within the groups, not to expand the families beyond the outer boundaries of the families. The motif groups with all their associated annotations including Signalp 6.0 (18), Pfam domains (4), Uniprot links (19), MycoCosm links (12), PhycoCosm links (13), dbCAN domains (6) and more were uploaded to the https://cupp.info/ webserver for user interaction.

RESULTS

Systematic genome comparison including JGI genomes

The new CUPP.info webserver allows any user to submit a genome or any list of proteins up to 32MB in a file for free.

Downloaded from https://academic.oup.com/nar/article/51/W1/W108/7175339 by guest on 03 October 2024

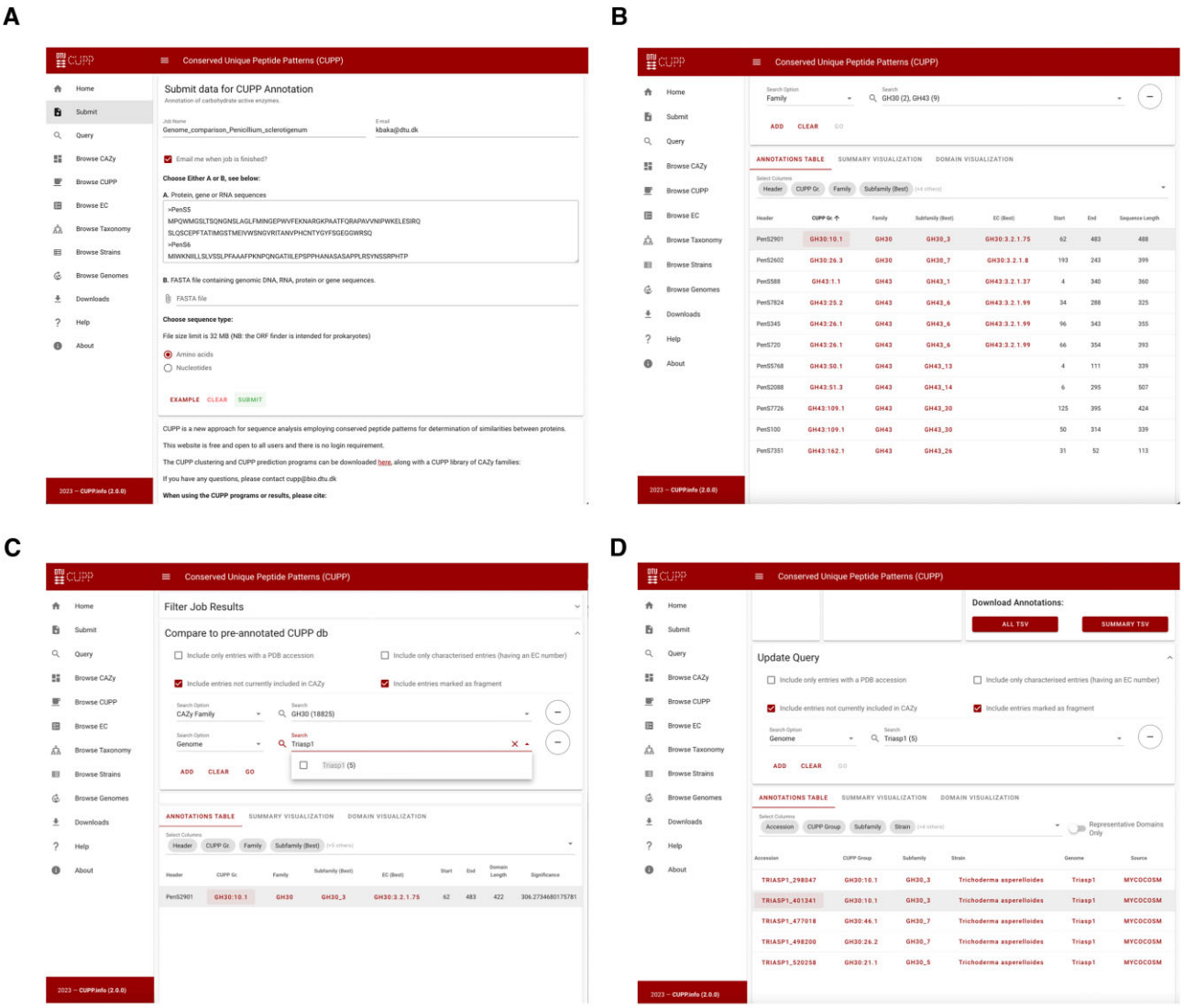


Figure 1. A tour through the webserver and the associated pre-annotations. (**A**) Submission of user sequences. (**B**) Filtering of user sequences for example by CAZy family GH30 and GH43. (**C**) Comparison of the user defined protein to the CUPP groups shared with the proteins of the JGI sequenced *Trichoderma asperelloides*. (**D**) The overview of the five GH30 hits found in the JGI sequenced *Trichoderma* with links on the accession to the JGI website for a much more elaborate documentation of the protein.

Once a query has been submitted, here exemplified by the genome of *Penicillium sclerotigenum* (20) (Figure 1A), the delineation and annotation will commence. After about 9 s of annotation (e.g. for a genome containing about 9000 proteins), a summary page will appear in which the results can be filtered, in this example limited to CAZy family GH30 and GH43 (Figure 1B).

In case the user wants to compare the current GH30 annotations of e.g. the *Trichoderma asperelloides* JGI genome (21) this can be seamlessly done by selecting the portal name 'Triasp1' using the 'Compare to pre-annotated CUPP db' filters which will show shared CUPP groups between the genomes, in this case GH30:10.1 (Figure 1C). Alternatively, all JGI MycoCosm proteins combined (12) or a specific taxonomic class can be selected within the webserver interface. In this example, the user can also opt to use the 'Browse Genomes'-tab to go directly to the JGI genome of *T. asperelloides* from the left control panel. The *T. asperelloides* genome has five GH30 hits, and one of the genes 'TRIASP1_401341' belonging to group GH30:10.1 could, for example, be selected for experimental characterization as the JGI predicted genes moreover have transcriptomics support (Figure 1D). To inspect the transcriptomics result of individual genes, proteins originating from JGI have a hyperlink to a summary page which links to a 'Genome browser' page that displays the predicted gene splicing, including which regions have RNA support (Figure 1D).

Hence, in the protein specific site in the JGI website under 'To Genome Browser', the current protein (GeneCatalog) can be seen together with several other alternative predictions of the gene splicing, which is essential to have correct, for the protein to function naturally (Figure 2).

As the RNA coverage supports the exon/intron splicing, it is possible to infer whether a particular gene, in this case

Downloaded from https://academic.oup.com/nar/article/51/W1/W108/7175339 by guest on 03 October 2024

Table 1. Comparison between the dbCAN webserver, the eggNOG webserver for both family and functional annotation of CAZymes and the updated CUPP-webserver using the recommended significance cut-off at 5. The column 'CAZy - All characterized' encompasses all 10784 characterized proteins in the CAZy database used for the training, whereas the 'CAZy - Newly characterized' designate 199 characterized CAZymes that were added after the training ended

	CAZy – all characterized			CAZy – newly characterized		
	CUPP	eggNOG	dbCAN	CUPP	eggNOG	dbCAN
Family sensitivity	99.9%	50.7%	98.1%	100.0%	46.0%	97.4%
Family precision	99.6%	94.8%	99.9%	99.7%	94.1%	99.7%
EC sensitivity	84.0%	59.7%	93.0%	54.6%	40.9%	58.7%
EC precision	95.1%	88.9%	76.3%	93.3%	89.2%	79.2%

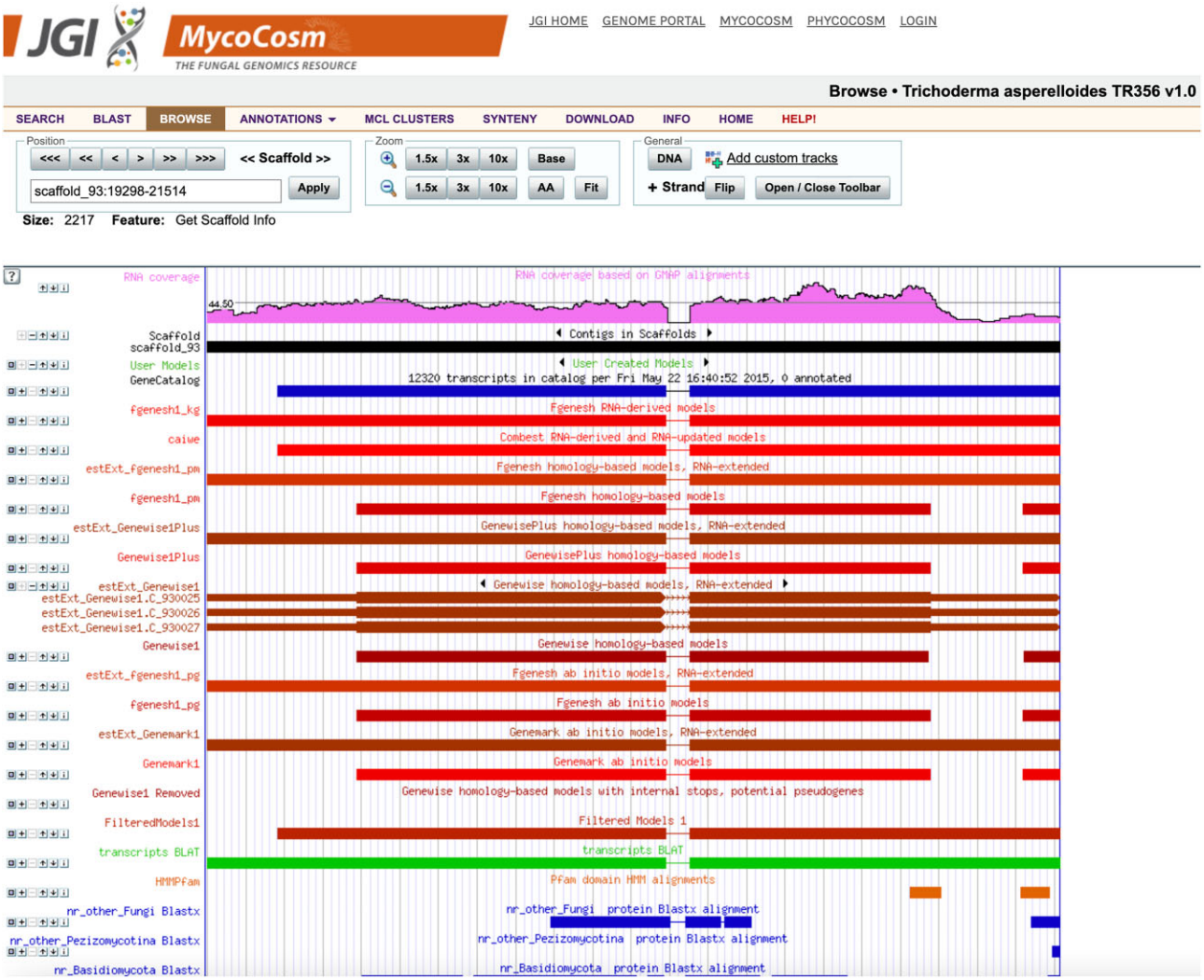


Figure 2. An example of the many descriptive pages for each JGI protein, here it is the 'Genome Browse' subpage. The blue bar induces the two exons of the final transcript from which the protein is based. In the pink RNA coverage graph, a drop to zero signal can be observed in the intron region between the two exons. The bars belong show predicted transcripts based on alternative gene prediction tools showing that a region toward to N-terminus is sometimes not considered part of the protein.

a gene such as the one selected in Figure 2, is more likely to work after heterologous gene expression.

To further improve the enzyme selection, all NCBI Genbank accessions were mapped to Uniprot ID to link to the specific Uniprot accession page including the predicted AlphaFold structures, Go annotations and InterPro annotations and more (19).

Pre-annotations of JGI genomes and browse options

The proteins in the CUPP database can be displayed in various ways including a 'Summary visualization' as a bar-plot which could compare GH30 occurrence in the 21 genomes in MycoCosm of *Trichoderma* spp. (Figure 3A).

Downloaded from https://academic.oup.com/nar/article/51/W1/W108/7175339 by guest on 03 October 2024

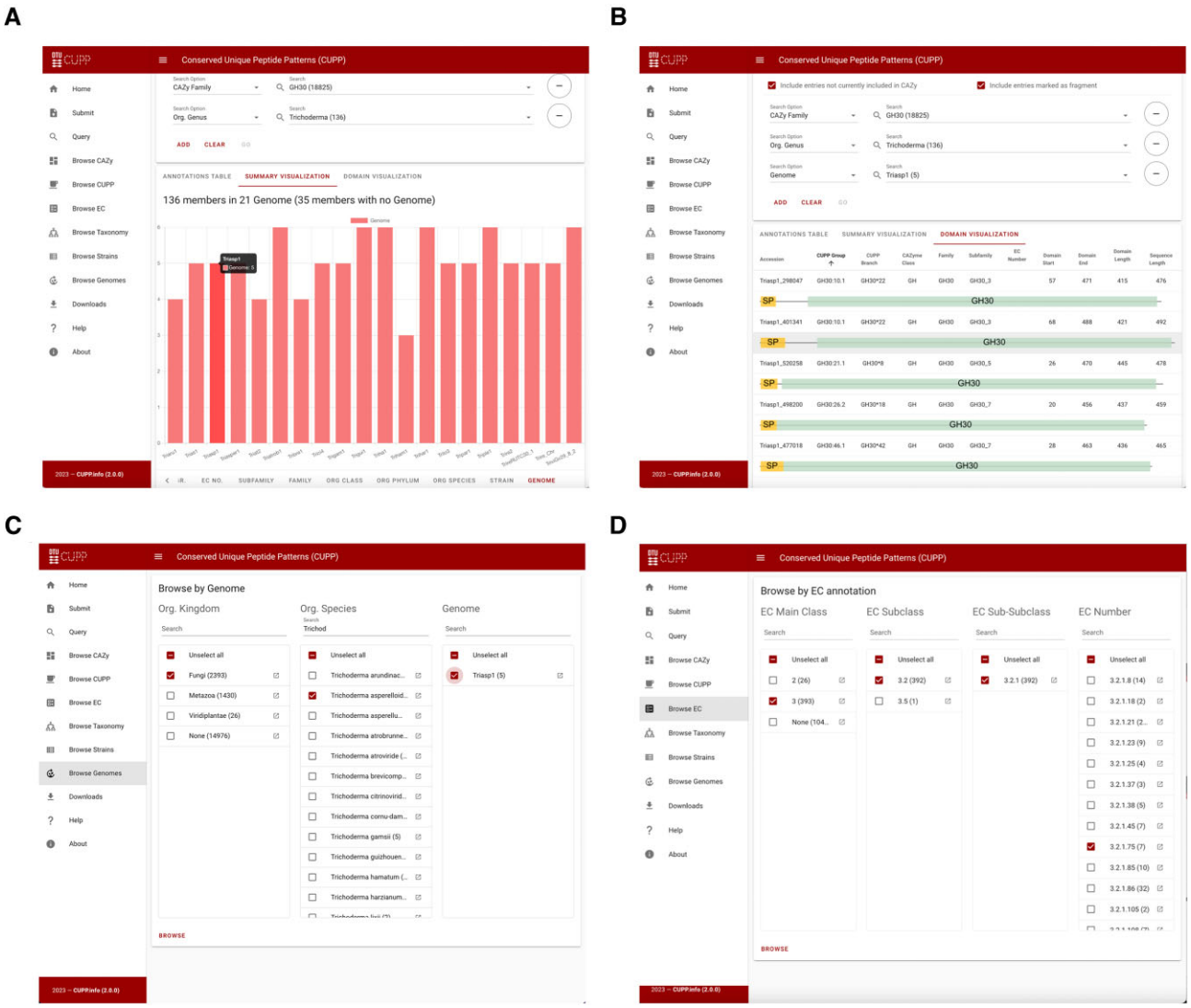


Figure 3. Examples of visualization options provided by the new CUPP-webserver interface. (**A**) The bar-plot of the 21 JGI genomes belonging to *Trichoderma* with predicted GH30 enzymes in the pre-annotated CUPP db shown by clicking on 'Summary visualization'. (**B**) The five GH30 of *Trichoderma asperelloides* displayed with their domain modularity including predicted signal peptide, shown by clicking 'Domain visualization'. (**C**) The short-cut to inspect a JGI genome, named 'Browse genomes' in the left control panel. (**D**) A short-cut to inspect a particular EC number across all families.

By clicking on the 'Domain visualization' tab, the domain of each protein can be inspected to see possible secondary domains and signal peptides (SP), here shown for GH30 protein in *T. asperelloides* (Figure 3B). To ease the accessibility to the genomes, a new browse panel has been implemented which allows quick inspection of particular genomes under the 'Browse genomes' tab in the left control panel (Figure 3C); an alternative option is to 'Browse by EC numbers' (Figure 3D).

Stand-alone improvements

The former Python implementation of the CUPP annotation algorithm did not allow the multi-threading required for optimal maneuvering and speedy data processing. This was problematic as the library files occupied 16 GB RAM during any run. With the new implementation, the RAM usage is reduced to a fourth while allowing efficient multi-threading. The annotation speed for an average genomic protein is now <1 ms using only one core.

DISCUSSION

Annotation comparison and benchmarking

The overall family annotation performance of the CUPP algorithm is considered highly satisfactory with nearly maximum sensitivity and precision using CUPP for full collection of both the characterized proteins included and those not included in the training (called newly characterized) (Table 1). The family annotation by dbCAN is also high, only missing a few percent (Table 1). For the EC numbers, the annotation by CUPP shows a lower sensitivity than dbCAN, but - more importantly - a better precision. The online webserver for EC annotation by eggNOG performed with lower sensitivity and precision for both the protein col-

Downloaded from https://academic.oup.com/nar/article/51/W1/W108/7175339 by guest on 03 October 2024

Table 2. Comparison of CUPP CAZy family annotation versus the CAZy family annotations of dbCAN and eggNOG. The true CAZy family annotations and the genomic proteins were obtained from MycoCosm. The selected genomes include: *Aaosphaeria arxii* CBS 175.79 belonging to *Ascomycota*, class *Dothideomycetes* (Aaoar1) (22), *Acremonium* sp. TS7 belonging to *Ascomycota*, class *Sordariomycetes* (AcreTS7_1) (23), *Abortiporus biennis* CIRM-BRFM 1778 belonging to phylum *Basidiomycota* (Abobie1) (24), *Anaeromyces* sp. S4 belonging to phylum *Chytridiomycota* (Anasp1) (25,26), and*Absidia repens* NRRL1336 belonging to phylum *Mucoromycota* (Absrep1) (26)

MycoCosm	Genomic	True	CAZy family sensitivity			CAZy family precision		
Genus	Proteins	CAZy	CUPP	eggNOG	dbCAN	CUPP	eggNOG	dbCAN
Aaosphaeria	14,203	585	98.9%	30.7%	95.7%	99.7%	99.9%	99.5%
Acremonium	9964	429	97.9%	37.9%	93.8%	99.7%	99.8%	99.5%
Abortiporus	11,767	372	97.7%	36.6%	94.1%	99.8%	99.9%	99.7%
Anaeromyces	12,832	503	94.9%	25.7%	91.6%	98.3%	99.8%	99.8%
Absidia	14,919	297	97.5%	41.2%	91.9%	99.3%	99.9%	99.7%
Total	63,685	2186	97.4%	33.7%	93.6%	99.4%	99.9%	99.6%

lection for which CUPP was trained and for the collection of newly added characterized CAZy enzymes (Table 1).

When comparing the annotation performance of CUPP on a set of representative genomes using genomic proteins from MycoCosm (Table 2), the sensitivity of the CUPP outperforms dbCAN and eggNOG (Table 2). The precision of eggNOG was only slightly below that of CUPP, however, the sensitivity of eggNOG was far below that of CUPP, whereas dbCAN was far better than eggNOG, but still below CUPP (Table 2). The high granularity of the CUPP groupings thus ensures that only a very limited number of incorrect EC assignments occur, with a minor negative effect on sensitivity (Table 1).

Sensitivity and precision for CAZy family annotation is likely better when the query sequences are the founding members or central members of the sequence space of the CAZy families, as evident from the 98–99.9% performance results of CUPP and dbCAN (Table 1). However, when CAZy family annotation is carried out on a full genome, some of the query sequences are likely near the outermost boundary of the CAZy family sequence space, thus causing the sensitivity to be lower (Table 2) than the CAZy family annotation of the characterized enzymes (Table 1).

DATA AVAILABILITY

The data underlying this article are available in the article and in its online supplementary material. The webserver is freely available: https://cupp.info. The entry page provides easy access to the annotation of existing genomes as well as a submission page for user-defined queries. For offline usage, the new implementation of the CUPP program can be downloaded from https://cupp.info/downloads as a Python script directly functional on Windows, Linux and MacOS operating systems with documentation provided in the readme file.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Many thanks to the CAZy group at Aix Marseille University, Marseille, France, and Professor Bernard Henrissat for providing the up-to-date CAZy.org database which lays the foundation for CAZymes and now underpins all research with carbohydrate-active enzymes.

Author contributions: Kristian Barrett: Conceptualization, Data curation, Methodology, Validation, Writing – original draft. Cameron J. Hunt: Investigation, Software, Methodology, Validation, Visualization. Igor Grigoriev: Data curation, Resources. Lene Lange: Conceptualization. Anne S. Meyer: Project administration, Validation, Writing – review & editing.

FUNDING

Novo Nordisk Foundation [NNF21OC0066330 and NNF22OC0072911 to A.S.M.]; Technical University of Denmark, DTU Bioengineering; U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy [DE-AC02-05CH11231]. Funding for open access charge: Novo Nordisk Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Drula,E., Garron,M., Dogan,S., Lombard,V., Henrissat,B. and Terrapon,N. (2021) The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.*, **50**, D571–D577.
2. 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.
3. Henrissat,B. (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.*, **280**, 309–316.
4. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
5. Paysan-Lafosse,T., Blum,M., Chuguransky,S., Grego,T., Pinto,B.L., Salazar,G.A., Bileschi,M.L., Bork,P., Bridge,A., Colwell,L. *et al.* (2023) InterPro in 2022. *Nucleic Acids Res.*, **51**, D418–D427.
6. 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.
7. 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.
8. Hernández-Plaza,A., Szklarczyk,D., Botas,J., Cantalapiedra,C.P., Giner-Lamia,J., Mende,D.R., Kirsch,R., Rattei,T., Letunic,I., Jensen,L.J. *et al.* (2023) eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res.*, **51**, D389–D394.

Downloaded from https://academic.oup.com/nar/article/51/W1/W108/7175339 by guest on 03 October 2024

9. Huerta-Cepas,J., Szklarczyk,D., Heller,D., Hernández-Plaza,A., Forslund,S.K., Cook,H., Mende,D.R., Letunic,I., Rattei,T., Jensen,L.J. *et al.* (2019) EggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.
10. 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.
11. Barrett,K., Hunt,C.J., Lange,L. and Meyer,A.S. (2020) Conserved unique peptide patterns (CUPP) online platform: peptide-based functional annotation of carbohydrate active enzymes. *Nucleic Acids Res.*, **48**, W110–W115.
12. Grigoriev,I.V., Nikitin,R., Haridas,S., Kuo,A., Ohm,R., Otillar,R., Riley,R., Salamov,A., Zhao,X., Korzeniewski,F. *et al.* (2014) MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.*, **42**, D699–D704.
13. Grigoriev,I.V., Hayes,R.D., Calhoun,S., Kamel,B., Wang,A., Ahrendt,S., Dusheyko,S., Nikitin,R., Mondo,S.J., Salamov,A. *et al.* (2021) PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res.*, **49**, D1004–D1011.
14. Sayers,E.W., Bolton,E.E., Brister,J.R., Canese,K., Chan,J., Comeau,D.C., Farrell,C.M., Feldgarden,M., Fine,A.M., Funk,K. *et al.* (2023) Database resources of the national center for biotechnology information in 2023. *Nucleic Acids Res.*, **51**, D29–D38.
15. 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.
16. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
17. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
18. Teufel,F., Almagro Armenteros,J.J., Johansen,A.R., Gíslason,M.H., Pihl,S.I., Tsirigos,K.D., Winther,O., Brunak,S., von Heijne,G. and Nielsen,H. (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.*, **40**, 1023–1025.
19. The UniProt Consortium (2023) UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
20. Barrett,K., Zhao,H., Hao,P., Bacic,A., Lange,L., Holck,J. and Meyer,A.S. (2022) Discovery of novel secretome CAZymes from *Penicillium sclerotigenum* by bioinformatics and explorative proteomics analyses during sweet potato pectin digestion. *Front. Bioeng. Biotechnol.*, **10**, 950259.
21. Kubicek,C.P., Steindorff,A.S., Chenthamara,K., Manganiello,G., Henrissat,B., Zhang,J., Cai,F., Kopchinskiy,A.G., Kubicek,E.M., Kuo,A. *et al.* (2019) Evolution and comparative genomics of the most common *Trichoderma* species. *Bmc Genomics [Electronic Resource]*, **20**, 485.
22. Haridas,S., Albert,R., Binder,M., Bloem,J., LaButti,K., Salamov,A., Andreopoulos,B., Baker,S.E., Barry,K., Bills,G. *et al.* (2020) 101 *Dothideomycetes* genomes: a test case for predicting lifestyles and emergence of pathogens. *Stud. Mycol.*, **96**, 141–153.
23. Hagestad,O.C., Hou,L., Andersen,J.H., Hansen,E.H., Alternark,B., Li,C., Kuhnert,E., Cox,R.J., Crous,P.W., Spatafora,J.W. *et al.* (2021) Genomic characterization of three marine fungi, including *Emericellopsis atlantica* sp. nov. with signatures of a generalist lifestyle and marine biomass degradation. *IMA Fungus*, **12**, 21.
24. Hage,H., Miyauchi,S., Viragh,M., Drula,E., Min,B., Chaduli,D., Navarro,D., Favel,A., Norest,M., Lesage-Meesen,L. *et al.* (2021) Gene family expansions and transcriptome signatures uncover fungal adaptations to wood decay. *Environ. Microbiol.*, **23**, 5716–5732.
25. Haitjema,C.H., Gilmore,S.P., Henske,J.K., Solomon,K.V., de Groot,R., Kuo,A., Mondo,S.J., Salamov,A.A., LaButti,K., Zhao,Z. *et al.* (2017) A parts list for fungal cellulosomes revealed by comparative genomics. *Nat. Microbiol.*, **2**, 17087.
26. Mondo,S.J., Dannebaum,R.O., Kuo,R.C., Louie,K.B., Bewick,A.J., LaButti,K., Haridas,S., Kuo,A., Salamov,A., Ahrendt,S.R. *et al.* (2017) Widespread adenine N6-methylation of active genes in fungi. *Nat. Genet.*, **49**, 964–968.

Downloaded from https://academic.oup.com/nar/article/51/W1/W108/7175339 by guest on 03 October 2024

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com