cazy_webscraper

For creating a local CAZy database

Emma Hobbs[1,2], Tracey Gloster[1],
Sean Chapman[2], Leighton Pritchard[3]
[1]University of St Andrews, St Andrews, UK
[2]The James Hutton Institute, Dundee, UK
[3]University of Strathclyde, Glasgow, UK

Introduction

Carbohydrate **A**ctive en**Zymes** (CAZymes) are pivotal in pathogen recognition, signalling, structure and energy metabolism. CAZy (www.cazy.org) is the most comprehensive CAZyme database [1], but it does not provide methods for automating data retrieval or submitting sequences for annotation.

cazy_webscraper retrieves user-specified datasets from CAZy, producing a local SQL database enabling thorough interrogation of the data. cazy_webscraper can also retrieve protein sequences from GenBank [2] and download structure files from RCSB PDB [3].

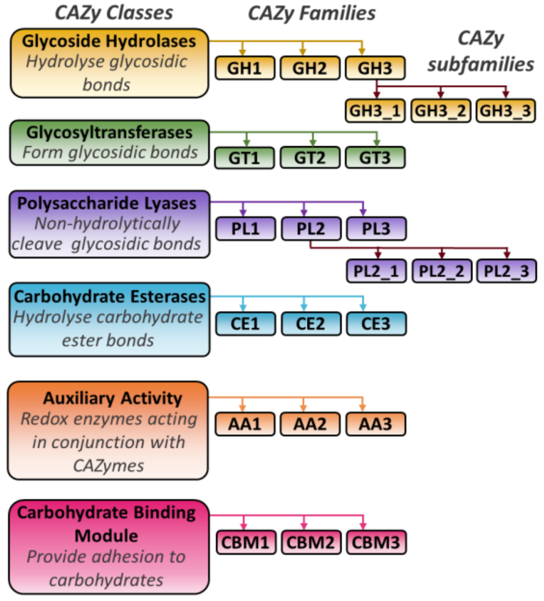

CAZy Classes
CAZy Families
CAZy subfamilies
Glycoside Hydrolases
Hydrolyse glycoslidic bonds
GH1 GH2 GH3
GH3_1 GH3_2 GH3_3
Glycosyltransferases
Form glycosidic bonds
GT1 GT2 GT3
Polysaccharide Lyases
Non-hydrolytically cleave glycosidic bonds
PL1 PL2 PL3
PL2_1 PL2_2 PL2_3
Carbohydrate Esterases
Hydrolyse carbohydrate ester bonds
CE1 CE2 CE3
Auxiliary Activity
Redox enzymes acting in conjunction with CAZymes
AA1 AA2 AA3
Carbohydrate Binding Module
Provide adhesion to carbohydrates
CBM1 CBM2 CBM3

Fig.1 CAZy database structure
CAZy catalogues proteins into classes that are divided into families, some of which are divided into subfamilies.

Method

Installation via GitHub:
https://github.com/HobnobMancer/cazy_webscraper

Scraping is invoked using the command python3 cazy_webscraper. All optional flags can be found in the GitHub repository README.

Expanding the dataset beyond CAZy is achieved using the expand module.

1. GenBank

Each unique CAZyme is identified by its **primary** GenBank accession, consolidating duplicate CAZy entries in the local database.

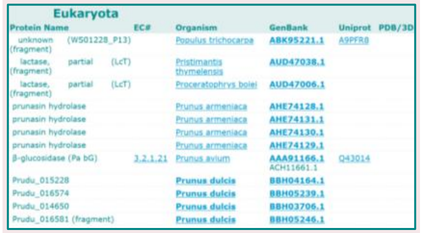
Retrieve all CAZy family annotations for a given protein by querying the local CAZyme database by its GenBank accession.

cazy_webscraper automates retrieving **protein sequences from GenBank.**

cazy_webscraper can update sequences in the local CAZyme database if a newer sequence is available in NCB, **keeping the dataset up to date.**

2. CAZy Families

cazy_webscraper automates and quickly scrapes CAZy. Scraping CAZy family GH1, containing **43,649 proteins**, takes **44 minutes,** instead of users manually reading **44 webpages**.



Eukaryota					
Protein Name	EC#	Organism	GenBank	Uniprot	PDB/3D
unknown (WS01228_P13) (fragment)		Populus trichocarpa	ABK95221.1	A9PFR8	
lactase, partial (LcT) (fragment)		Pristimantis thymelensis	AUD47038.1		
lactase, partial (LcT) (fragment)		Proceratophrys boiei	AUD47006.1		
prunasin hydrolase		Prunus armeniaca	AHE74128.1		
prunasin hydrolase		Prunus armeniaca	AHE74131.1		
prunasin hydrolase		Prunus armeniaca	AHE74130.1		
prunasin hydrolase		Prunus armeniaca	AHE74129.1		
β-glucosidase (Pa bG)	3.2.1.21	Prunus avium	AAA91166.1 ACH11661.1	Q43014	
Prudu_015228		Prunus dulcis	BBH04164.1		
Prudu_016574		Prunus dulcis	BBH05239.1		
Prudu_014650		Prunus dulcis	BBH03706.1		
Prudu_016581 (fragment)		Prunus dulcis	BBH05246.1		

Fig.2 CAZy database structure
An HTML table users had to previously parse manually to retrieve data from CAZy

Unlike previous scrapers [4], cazy_webscraper can retrieve data for **specific CAZy classes and (sub)families**, reducing waiting times from **hours to minutes**.

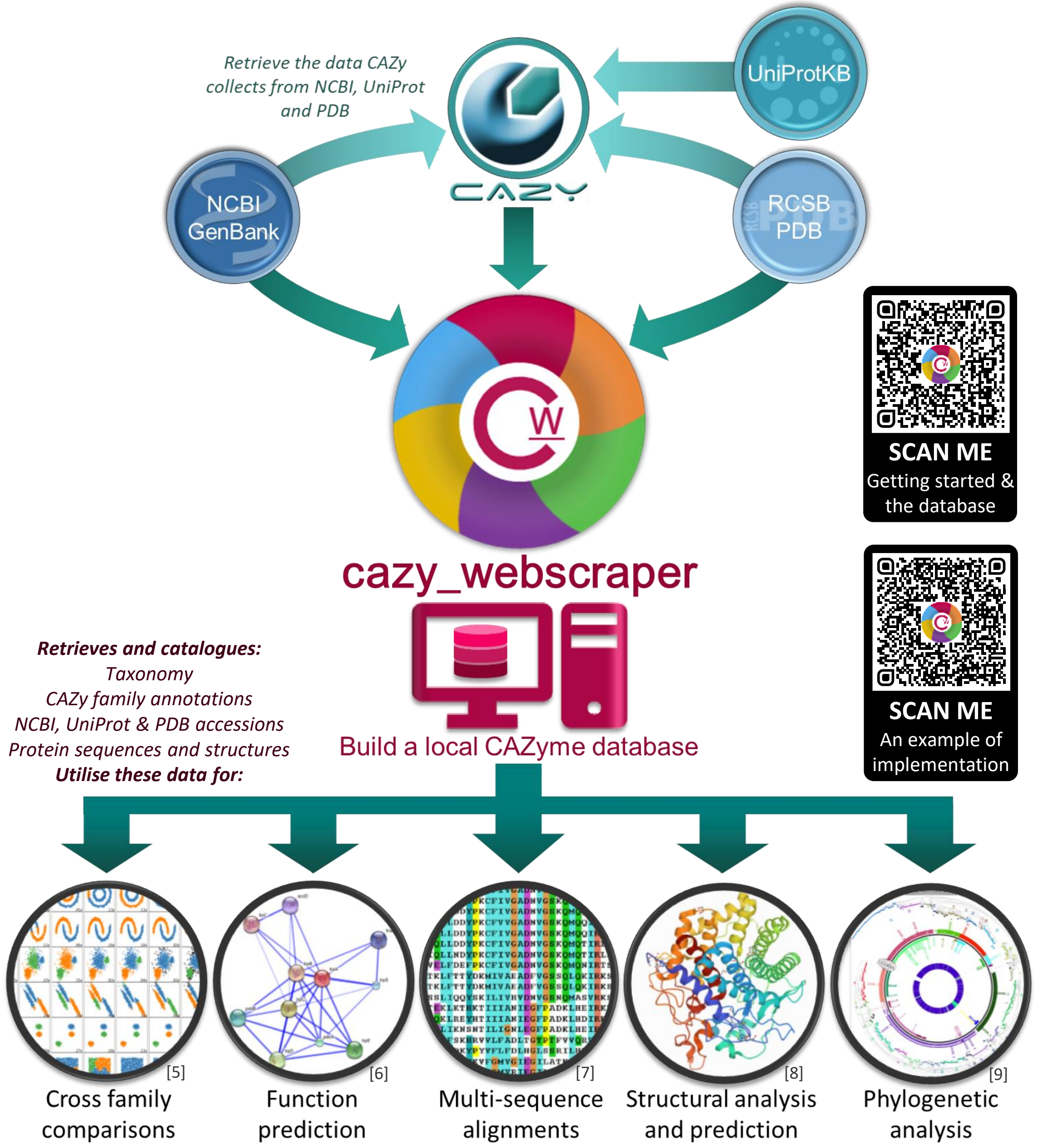
3. EC Numbers

Use cazy_webscraper to collate quickly CAZymes having similar activity by scraping by EC number or querying the local CAZyme database.

4. Taxonomy

Scrape specific taxa. Apply a combination of **kingdoms, genus, species,** and /or **strain** filters.

Use the taxonomy data to track the evolution of functions through **phylogenetic analysis**.


Retrieve the data CAZy collects from NCBI, UniProt and PDB

UniProtKB
NCBI GenBank
CAZY
RCSB PDB
cazy_webscraper
Build a local CAZyme database
SCAN ME
Getting started & the database
SCAN ME
An example of implementation
[5]
[6]
[7]
[8]
[9]

Retrieves and catalogues:
Taxonomy
CAZy family annotations
NCBI, UniProt & PDB accessions
Protein sequences and structures
Utilise these data for:

Cross family comparisons
Function prediction
Multi-sequence alignments
Structural analysis and prediction
Phylogenetic analysis

Fig. 2 Sources and application of data stored in the CAZyme database created by *cazy_webscraper*
Numbers in brackets indicate the source of the image.

5. CAZomes

Automate retrieving the CAZome (all CAZymes within a genome) of species of interest from CAZy.

Or quickly retrieve CAZomes by querying the local CAZyme database.

With one command, retrieve all protein sequences of a CAZome, ready for homolog searchers.

6. UniProt

Expand the dataset beyond CAZy by incorporating data *via* UniProt accessions. For example, retrieve CAZyme subcellular localisation data from UniProt, to **elucidate the functions** of uncharacterised CAZymes.

7. RCSB PDB

Automate rapid retrieval of **all** PDB structures for the dataset of interest in CAZy using cazy_webscraper.

Query using a combination of taxonomy, CAZy (sub)family, CAZy class and EC number filters.

8. SQL Database

Building an SQL database instead of a plaintext [5], enables thorough interrogation of the data *via* complex queries using SQL.

Perform complex queries that cannot be performed on the CAZy website.

For example, retrieve all species with at least one CAZyme in GH1 and at least one CAZyme in PL9.

Reproducibility

Use cazy_webscraper to generate **reproducible and shareable datasets**, facilitating reproduction of downstream analyses.

Optional configuration by a YAML file and generation of a log file, generates **shareable documentation** to bolster reproducibility.

Conclusions

cazy_webscraper provides new, **previously unachievable** access to the proteomic data within CAZy. This facilitates inclusion of CAZy data in functional, evolutionary, structural, genomic and metabolic studies. Thus, cazy_webscraper opens up numerous new avenues of investigation.

- **Automate** retrieving CAZy annotations, protein sequences and structure files
- **Expand** the dataset beyond that stored in CAZy
- **Thoroughly** interrogate the dataset using complex queries in SQL

References

1. Lombard, V. *et al.* (2014) 'The carbohydrate-active enzymes database (CAZy) in 2013, *Nucleic Acids Research,* 42, pp.D490–D495
2. Sayers, E. W. *et al.* (2020) 'GenBank', *Nucleic Acids Research,* 49(D1), pp.D92-96
3. Berman, Helen M. *et al.* (2000) 'The Protein Data Bank', *Nucleic Acids Research*, 28(1), pp.235-242
4. Honorato, R. V. (2016) 'CAZy-parser a way to extract information from the Carbohydrate-Active enZYmes Database', *The Journal of Open Source Software*, 1(8), 53
5. Chilamakuri, C. S. R. *et al.* (2011) 'Cross-genome comparisons of newly identified domains in Mycoplasma gallisepticum and domain architectures with other mycoplasma species', *International Journal of Genomics*, 2011, pp. 878973
6. Wikipedia (2009) 'Protein function prediction', accessed 2021.03.27
7. Andrade, M (2006) 'Multiple sequence alignment', *Wikipedia*, accessed 2021.03.27
8. Parsiegla, G. (2002) 'Crystal structure of the cellulase Cel9M enlightens structure/function relationships of the variable catalytic modules in glycoside hydrolases', *Biochemistry*, 41(37), pp.11134-11142
9. Barrett, K., Lange, L. (2019) 'Peptide-based functional annotation of carbohydrate active enzymes by conserved unique peptide patterns (CUPP)', *Biotechnology for biofuels*, 12, 102

Acknowledgements

We thank the EASTBIO Doctorial Training Partnership for funding our work.