cazy_webscraper version 2
Customising scrapes and the structure of the local CAZyme database

Emma Hobbs[1], 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]. CAZy 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 to enable thorough interrogation of the data. cazy_webscraper can also retrieve protein sequences from GenBank [2] and download structure files from RCSB PDB [3].

Methods

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 to include protein sequences from GenBank and structure files from PDB is achieved using the expand module.

Full documentation is available at:
https://cazy-webscraper.readthedocs.io/en/latest/?badge=latest

Requirements

- POSIX or Mac OS, or a Linux emulator
- Python3 version 3.8+
- Internet access while scraping CAZy
- Required Python libraries are found in the GitHub repository `requirements.txt`

1. GenBank

Each unique CAZyme is identified by its unique **primary** GenBank accession.
Primary accessions are accessions written in bold in CAZy, which CAZy defines as the 'best' model [4]. **Non-primary** accessions are accessions not written in bold.

cazy_webscraper can retrieve GenBank protein sequences for:

- Specific CAZy classes and families
- Specific kingdoms, genera, species and strains
- CAZymes without sequence records
- CAZymes whose sequence has been updated since it was last retrieved

Protein sequences are added to the local database, and can be written to FASTA files

2. CAZy Families

To scrape **CAZy subfamilies** use the --subfamilies flag.

Specific **CAZy (sub)families** can be scraped using --families flag. For example, to scrape GH1, GH3 and PL1_2 use the command:

```
python3 cazy_webscraper --families GH1,GH3,PL1_2 --subfamilies
```

To specify **CAZy classes** to scrape, use the --classes flag followed by the classes to scrape.

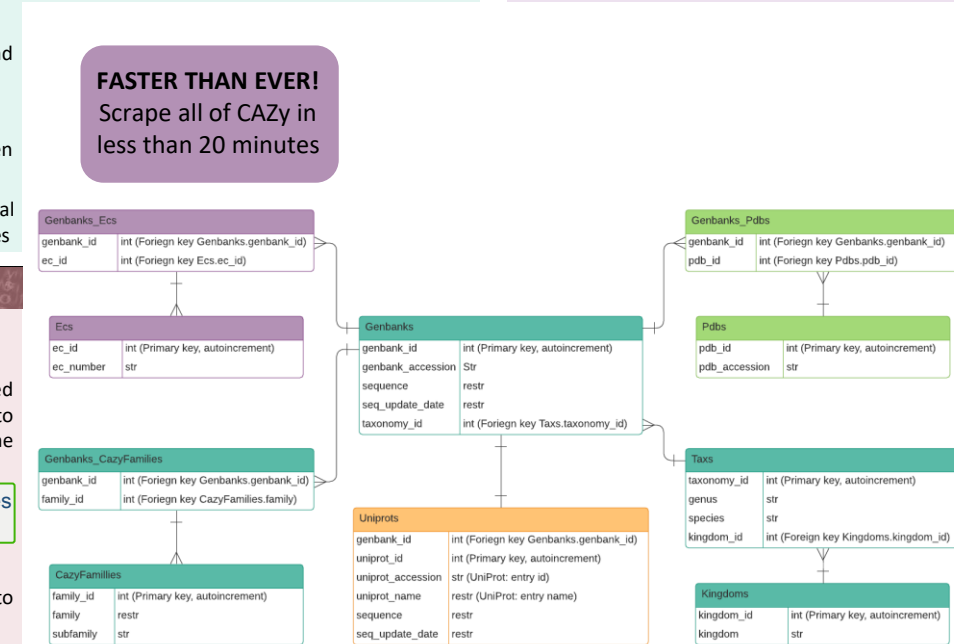Scraping specific families and classes significantly **reduces waiting times.**

Combine as many flags as you wish. For instance, to scrape bacteria from certain families and all CBMs use:

```
python3 cazy_webscraper --classes CBM --families GH5,GH7,PL8 --kingdoms Bacteria
```

7. EC Numbers

Retrieve EC numbers from UniProt by adding the --ec flag when retrieving data from UniProt.

Use EC numbers stored in the local CAZyme database to filter retrieval of data from UniProt, GenBank and PDB to CAZymes with specific EC numbers using the –ec_filters

```
cw_get_uniprot_data cazy.db --ec
```

3. Taxonomy

To scrape by taxonomy use the following flags in any combination:

Kingdoms, use the --kingdoms flag, and list Archaea, Bacteria, Eukaryota, Viruses and/or unclassified.

Genera, --genera flag, and list of genera to retrieve CAZymes from.

Species, --species flag, and list the species to scrape CAZymes from all strains of these species.

Strains, --strains flag to specify specific strains of species.

4. UniProt

CAZy no longer incorporates the accessions of associated UniProt records. UniProt [4] is a highly comprehensive protein database, incorporating data from a variety databases.

Using the cw_get_uniprot_data command retrieve UniProt accessions, protein names, EC number, PDB accessions and protein sequences from UniProt.

5. RCSB PDB

PDB is the most comprehensive protein structure database. cazy_webscraper uses the BioPython.PDB [5] module to retrieve PDB structure files for specific families, classes and taxonomies.

Use the cw_get_pdb_structures command.

6. SQL Database

Queries to the local CAZyme database are written in SQLite. To learn how to query SQL database we recommend the WiseOwl SQL tutorials https://www.wiseowl.co.uk/sql/.
The example query (left) retrieves all CAZyme records in the CAZy family GH103.
A copy of the database structure is stored in the GitHub repository to help define joins in queries.

```sql
SELECT C.cazyme_id, families.family
FROM cazymes AS C
  LEFT JOIN families ON families.family_id =
  cazymes_families.family_id
  LEFT JOIN cazymes_families ON
  cazymes_families.cazyme_id = C.cazyme_id
WHERE families.family = "GH103"
```

FASTER THAN EVER! Scrape all of CAZy in less than 20 minutes

Retrieve protein sequences, UniProt accessions, EC numbers and PDB accessions

Automatically retrieve the latest source organism from NCBI when multiple are listed or a single CAZyme in CAZy

Fig.1 Relationship-entity model of the local CAZyme database created by cazy_webscraper
A log table is also included. Logging when data is added and the commands used in each scrape

Add to an Existing Database

cazy_webscraper can scrape and add data to an existing local CAZyme database.
Use the --database flag followed by the path to the local CAZyme database.

Reproducible and cached

Every addition of data to a local CAZyme database is logged within the local database.

Raw and processed data from CAZy, UniProt and GenBank is cached, facilitating reproduction of datasets and skipping repeated retrieval of data when the retrieval was interrupted.

Conclusions

cazy_webscraper Version 2 provides new, **previously unachievable** access to the proteomic data within CAZy. This facilitates inclusion of CAZy data in many studies, including 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
2. Sayers, E. W. *et al.* (2020) 'GenBank', *Nucleic Acids Research*, 48
3. Berman, Helen M. *et al.* (2000) 'The Protein Data Bank', *Nucleic Acids Research*, 28
4. UniProt Consortium (2019) 'UniProt: a worldwide hub of protein knowledge', *Nucleic Acids Research*, 47
5. Hamelryck, T., Manderick, B. (2003) 'PDB parser and structure class implemented in Python', *Bioinformatics* 19

Acknowledgements