Comprehensive Evaluation of CAZyme Prediction Tools in Fungal and Bacterial Species

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 is the most comprehensive CAZyme database, cataloguing CAZymes into sequence-based CAZy families [1]. The CAZyme prediction tools **dbCAN** [2], **CUPP** [3] and **eCAMI** [4] annotate CAZymes with CAZy families. However, these tools have not been independently evaluated on a common high-quality dataset. Additionally, previous evaluations did not evaluate the **binary classification** of CAZymes/non-CAZymes, and the **multilabel classification** of CAZymes to multiple CAZy families.

Method

The bioinformatic pipeline **pyrewton** was developed for this independent evaluation (Fig.1).
GitHub: https://github.com/HobnobMancer/pyrewton
The ground truths were retrieved using **cazy_webscraper**.
GitHub: https://github.com/HobnobMancer/cazy_webscraper
70 genomic assemblies:
40 Bacteria
16 Fungal
11 Yeast
3 Eukaryotes

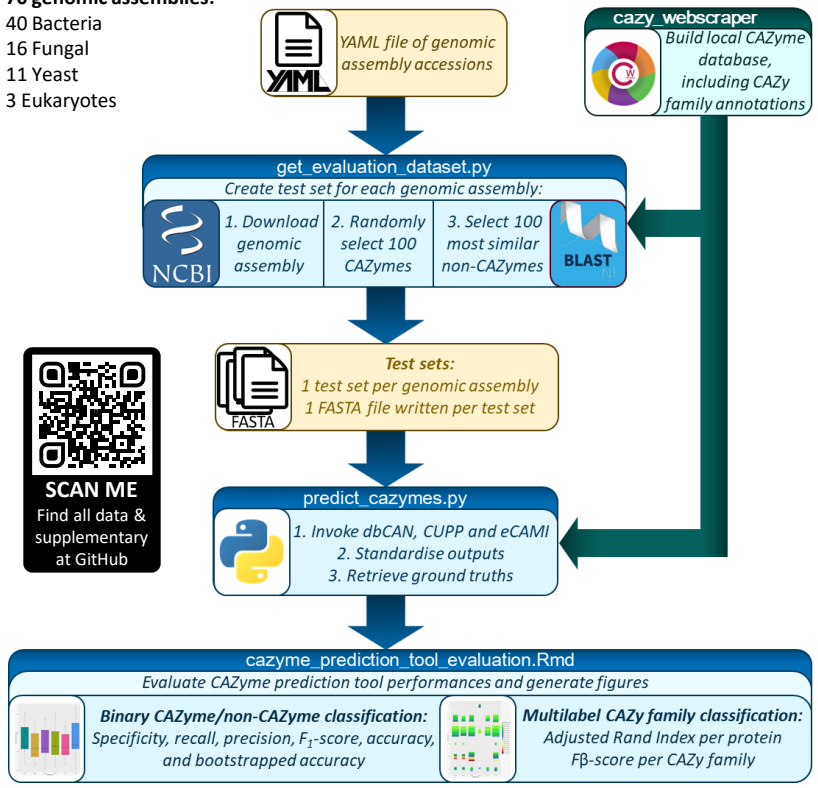

Fig.1 Schematic of the bioinformatic pipeline **pyrewton** *for evaluating CAZyme prediction tools*

Results

Binary CAZyme/non-CAZymes classification evaluation

dbCAN invokes the function prediction tools HMMER, Hotpep and DIAMOND. All predictiontools showed a low probability of misidentifying non-CAZymes as CAZymes, but also showed a tendency to miss identify a small proportion of CAZymes as non-CAZymes (Fig.2).

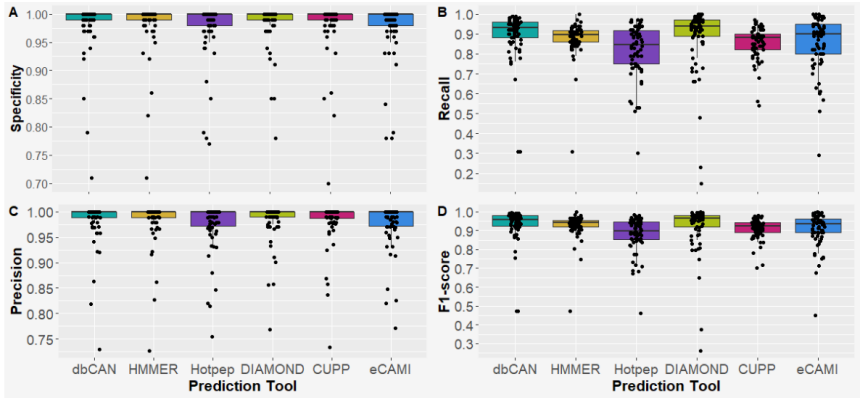

Fig.2 Evaluation of CAZyme/non-CAZyme differentiation performance.
One-dimensional scatterplots overlaying boxplots for [A] specificity, [B] recall, [C] precision and [D] F1-score.

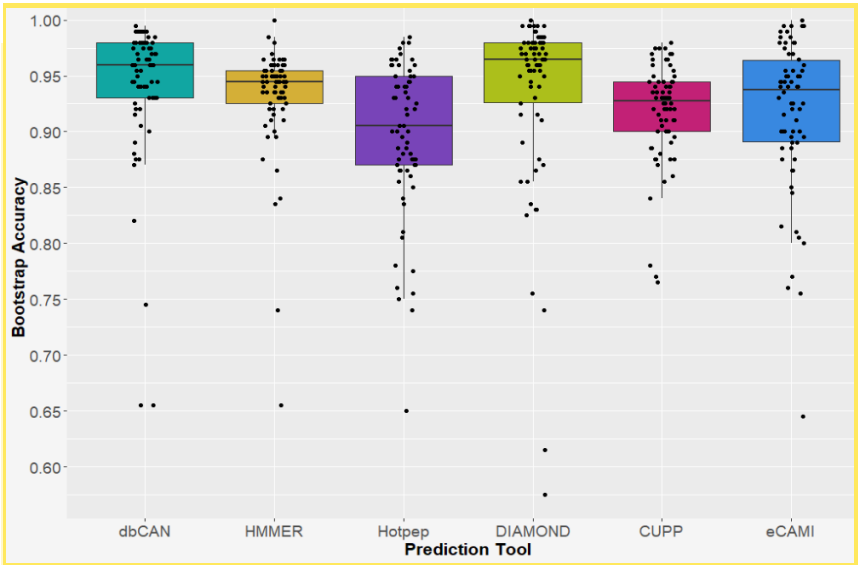

Fig.3 Expected range of performance of CAZyme prediction tools dbCAN, CUPP and eCAMI
Bootstrapping the CAZyme/non-CAZyme prediction accuracy was performed 10,000 times per test set. The median bootstrapped accuracy of each test set was plotted.

Multilabel CAZy family classification evaluation

Multilabel classification arises from the ability of a CAZyme to be assigned multiple CAZy families. The Adjusted Rand Index (ARI) was calculated per protein (Fig.4[A]) and the Fβ-score (β=1) calculated for each CAZy family, true negative non-CAZyme predictions were excluded (Fig.4[B])

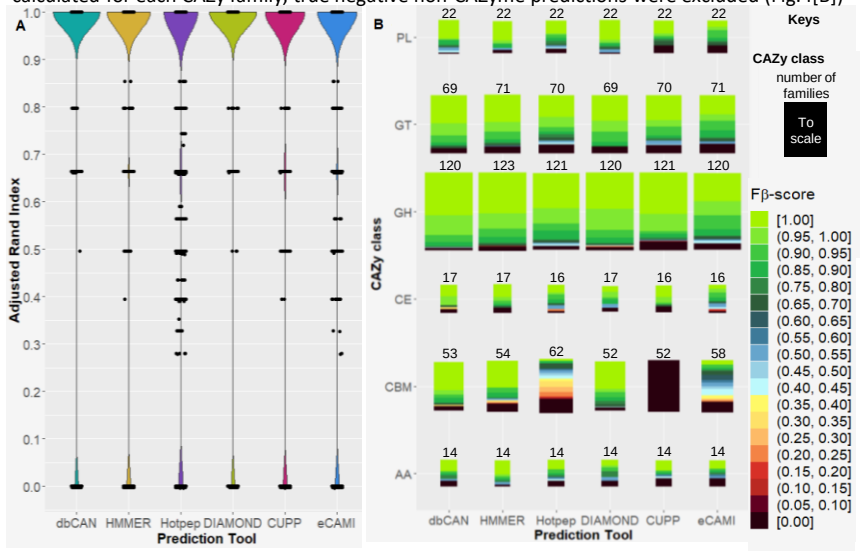

Fig.4 Evaluation of CAZy family multilabel classification
[A] Adjusted Rand Index per protein sequence. [B] Proportional area plot of CAZy classes sized by the number of families analysed, and coloured by the proportion of CAZy family Fβ-scores within each range of the scale, (β=1).

Conclusions

- Created a bioinformatic pipeline for the reproducible evaluation of CAZyme predictions tools, and benchmarked dbCAN, CUPP and eCAMI against a high quality test set
- Evaluated the binary and multilabel classification of CAZymes for the first time
- Statistically evaluated the expected range of performance for the first time
- dbCAN was best overall but the weakest was Hotpep, which is incorporated into dbCAN
- Best performance may be achieved by replacing Hotpep with CUPP and/or eCAMI
- Next steps are to expand the dataset and evaluate substituting Hotpep with CUPP and/or eCAMI

References

1. Lombard, V. *et al.* (2014) 'The carbohydrate-active enzymes database (CAZy) in 2013, *Nucleic Acids Research*, 42, pp.D490–D495
2. Zhange *et al.* (2018) 'dbCAN2: a meta server for automated carbohydrate-active enzyme annotation', *Nucleic Acids Research*, 46, W1, pp. W95-W101
3. Barrett, K., Lange, L. (2019) 'Peptide-based functional annotation of carbohydrate active enzymes by conserved unique peptide patterns (CUPP)', *Biotechnology for biofuels*, 12, 102
4. *Xu et al.* (2020) 'eCAMI: simultaneous classification and motif identification for enzyme annotation', *Bioinformatics*, 36, 7, pp.2068-2075

Acknowledgements

We would like to thank the EASTBIO Doctoral Training Partnership (BBSCR) for funding our research.