MACHINE-LEARNING BASED PREDICTION OF GLYCOSYLTRANSFERASE SUBSTRATES

Ditte Hededam Welner, Technical University of Denmark, Denmark
diwel@biosustain.dtu.dk
David Teze, Technical University of Denmark, Denmark
Christian Degnbol Madsen, University of Melbourne, Australia
Tiia Kittila, Technical University of Denmark, Denmark
Mads Langhorn, Technical University of Denmark, Denmark
Hani Gharabli, Technical University of Denmark, Denmark
Mandy Hobusch, Technical University of Denmark, Denmark
Onur Kírtel, Technical University of Denmark, Denmark
Felipe Mejia Otalvaro, Technical University of Denmark, Denmark
Evelyn Travnik, Technical University of Denmark, Denmark

machine learning, glycosyltransferase, substrate prediction, random forest, high-throughput screening

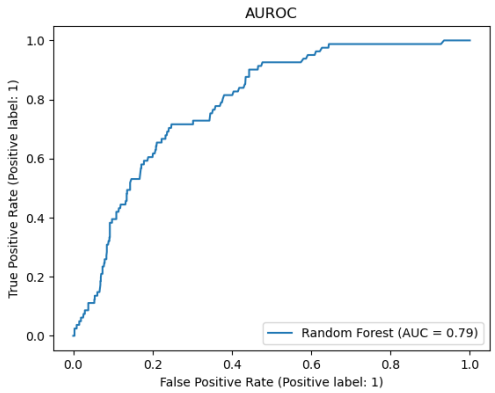

Figure 1 – Predictor performance on independent data

Functional prediction from enzyme sequence remains a major challenge in biocatalysis and enzyme engineering. For some enzyme families, sequence-function relationships are particularly elusive, and seem to be governed by complex patterns that escape our elucidation. Machine learning is emerging as a powerful tool in enzymology, due to its strength in recognizing patterns in complex data. Here, we apply a random-forest predictor to glycosyltransferase family 1 (GT1) enzymes[1]. This enzyme family is promiscuous and notorious for escaping elucidation of robust structure-function relationships[2]. It is also an enzyme family with large industrial potential, due to its capability of regioselective and stereoselective glycosylation of a vast array of industrially relevant molecules, including pharmaceuticals, nutraceuticals, and cosmetics. Efforts to apply machine learning to predict substrates for GT1 enzymes seem promising[3,4], although a pan-specific predictor that would truly enable efficient GT1 enzyme discovery is still missing. To provide this, we have constructed a predictor capable of parsing all known GT1 sequences, as well as all chemicals, the latter through a pipeline for automated generation of 153 chemical features for any given molecule (freely available at https://github.com/degnbol/GT). We have validated this predictor on an independent dataset from our lab consisting of 1001 datapoints, from 88 chemically diverse acceptors and 24 plant GT1 enzymes. We obtain an AUROC of 0.79 and a balanced accuracy of 73%. We then demonstrate predictor performance on two use cases: firstly, to find GT1 enzymes for glycosylation of DIBOA, a plant autotoxin. Secondly, to find GT1 enzymes for glycosylation of niclosamide, an essential medicine that can be used to treat COVID-19 and other severe diseases. The latter is a new-to-nature glycoside for which random screening did not yield any active enzyme.

1. Coutinho PM, Deleury E, Davies GJ, Henrissat B (2003) An evolving hierarchical family classification for glycosyltransferases. J. Mol. Biol. 328:307-317
2. Zhang P, Zhang Z, Zhang L, Wang J, Wu C (2020) Glycosyltransferase GT1 family: Phylogenetic distribution, substrates coverage, and representative structural features. Comp. Struct. Biotech. J. 18: 1383-1390
3. Yang M, Fehl C, Lees KV, Lim EK, Offen WA, Davies GJ, Bowles DJ, Davidson MG, Roberts SJ, Davis BG. (2018) Functional and informatics analysis enables glycosyltransferase activity prediction. Nat. Chem. Biol. 14:1109-1117
4. Goldman S, Das R, Yang KK, Coley CW (2022) Machine learning modeling of family wide enzyme-substrate specificity screens. PLoS Comput Biol 18(2): e1009853