**SubFinder: machine learning predicts glycan utilization in human gut microbiome**

**Deep learning predicts glycan utilization in gut microbiome**

**Introduction**

1. Importance of carbohydrate metabolism in different environments (global carbon recycling on earth, climate change, human health, soil health, crop production)

A screenshot of a computer

Description automatically generated with medium confidence

1. Glycan utilization in human and animal gut and personalized nutrition
2. State of the art of CAZyme bioinformatics

Diagram

Description automatically generatedGraphical user interface

Description automatically generated with medium confidence

1. Research gap: lack of computer tools for glycan substrate prediction

***Our hypothesis:*** *PULs targeting the same or similar substrates will have similar gene composition (i.e., protein family/domain combinations)*

A picture containing chart

Description automatically generatedText

Description automatically generated

**Results**

1. Collection of polysaccharide utilization loci (PULs) with experimentally characterized glycan substrates: training data

**Figure 1**: Stats of PULs from literature (example: figure 1 of <https://www.biorxiv.org/content/10.1101/2022.01.13.476228v1.full>, Catie had some plots that need redone)

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Different substrate groups:

Cellulose

Xylan

Beta-mannan

Beta-glucan

Pectin

Alpha-glucan

Chitin

Host glycan

Inulin

Peptidoglycan

Mixed-linkage galactan

Ulvan

Alginate

Fucoidan

Alpha-galactan

Arabinogalactan

1. Sequence features (protein family signatures) in PULs that can distinguish different glycan substrates: what protein families or family combinations determine substrate specificity of PULs?

**Figure 2**: Pfam/dbCAN/TCDB protein families over-represented in PULs targeting different glycans (Catie had some plots that need redone)

Chart, calendar

Description automatically generatedChart, bar chart

Description automatically generated

1. Develop a machine learning tool (subFinder) to learn sequence features and classify PULs targeting different glycans: what is the methodological innovation and what is the performance of this new tool?

**Figure 3**: Stats of training and testing data, graphical presentation of subFinder algorithm, performance of subFinder (example in <https://academic.oup.com/nar/article/47/18/e110/5545735> and <https://www.nature.com/articles/s41589-019-0400-9>)

We have finished the entire subFinder pipeline composed of five different steps. Step 1 converts the raw sequence of CGCs/PULs into formats that the machine learning models can consume. Step 2 uses the unsupervised CGC sequences to train three types of embedding models Word2vec, Doc2Vec, and FastText. These embedding models engineer features for the CGC sequences by taking the sequential order of the genes into account. Step 3 uses the embedding models from the Step 2 to convert the supervised PUL sequences into feature vectors and then create shallow and deep learning models to predict the high-level substrates. Finally, Step 4 uses the trained, supervised models from the Step 3 to predict the substrates for various unsupervised CGC sequences. In addition to these four steps, we have also developed scripts to generate the p-values for each subFinder prediction and provide signature genes that most impact the substrate prediction category. We are also developing a novel statistical methodology that combines the predictions from the various models we have developed. This would be beneficial because the statistical model combination method can further help boost the accuracy by combining these models.

We collected 411 PULs from dbCAN-PUL as the testing data to compare the performance of IFDP and subFinder.

IFDP predicted 105 (out of 411) PULs with single substrate labels. 64 PULs are from the top 10 substrate groups with more PULs (minimum 6).

These 64 PULs are used as the testing data for both IFDP and subFinder. The average IFDP accuracy is calculated to be 35.93%. For subFinder, we retrained our model by excluding the 105 PULs. So with 411-105 PULs for training, we had a new subFinder model and applied it to 64 PULs. The average accuracy is 64.06%.

1. Genome mining 200,000 human gut MAGs (metagenome assembled genomes) for CGCs (CAZyme gene clusters)

**Figure 4**: Stats of CGCs from gut MAGs (example in <https://www.biorxiv.org/content/10.1101/2021.02.25.432841v1.full> and <https://journals.asm.org/doi/full/10.1128/aem.01851-21>)

1. Predict glycan substrates for gut MAGs using subFinder

**Figure 5**: Stats of predicted PULs or pPULs (glycans in CGCs from gut MAGs, example in <https://www.biorxiv.org/content/10.1101/2021.02.25.432841v1.full>, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4164201/>)

1. Prevalence and abundance of pPULs in different gut bacterial taxonomic and geographical groups

**Figure 6**: Distribution of predicted PULs with respect to taxonomy and geography (and <https://journals.asm.org/doi/full/10.1128/aem.01851-21>, https://pubmed.ncbi.nlm.nih.gov/25036635/)

1. Clustering of pPULs into pPUL families

**Figure 7**: Distribution of predicted PULs with respect to taxonomy and geography (<https://academic.oup.com/nar/article/49/D1/D490/5917658?login=true> )

1. GitHub and web server of subFinder and utility