IFDP substrate prediction

BACKGROUND:

Part1:

Basic information

CAZymes: Enzymes(protein) that could target the glycosidic linkages to degrade, synthesize or modify carbohydrates.

PUL: polysaccharides utilization loci, the gene cluster found in bacteria including CAZymes and transport system (some functional proteins such as SusC). Each PUL in dbCAN-PUL (The database of PUL) has a characterized substrate.

CGC: Cazyme gene cluster, the computed PULs including CAZymes and TC (TF/STP).

PUL is the experimentally verified CGC.

Each PUL has a verified substrate.

Part2:

Three Substrate prediction method.

1. dbCAN\_seq update approach

For each genome file (MAG), we use faa file(protein sequences) and gff file(protein positions) to do the CAZyme and CGC annotation by dbCAN.

Genome -> CGCs (containing CAZymes and other functional protiens) -> Substrates (with 2 approaches in dbCAN\_seq update, the dbCAN-PUL approach and eCAMI subfamily approach).

1. Subfinder:

You could give some details here.

1. IFDP

Published in December 2022 (Cohen & Borenstein, 2022).

As we’ve mentioned before, CAZymes work on glycosidic linkages, and carbohydrates are linked with glycosidic linkages. As a result, they create two mapping tables:

Table 1. The relationship between Enzymes and Glycosidic linkages (Not the whole table).

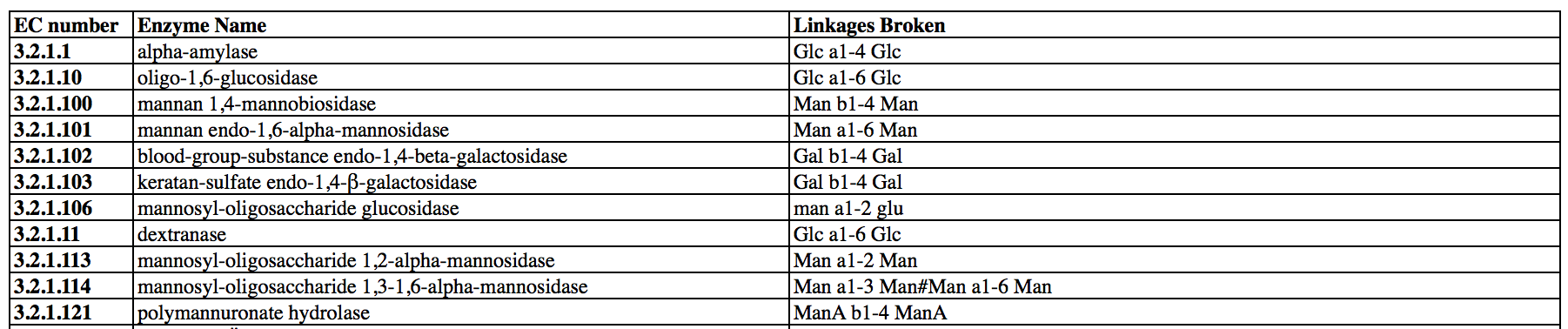
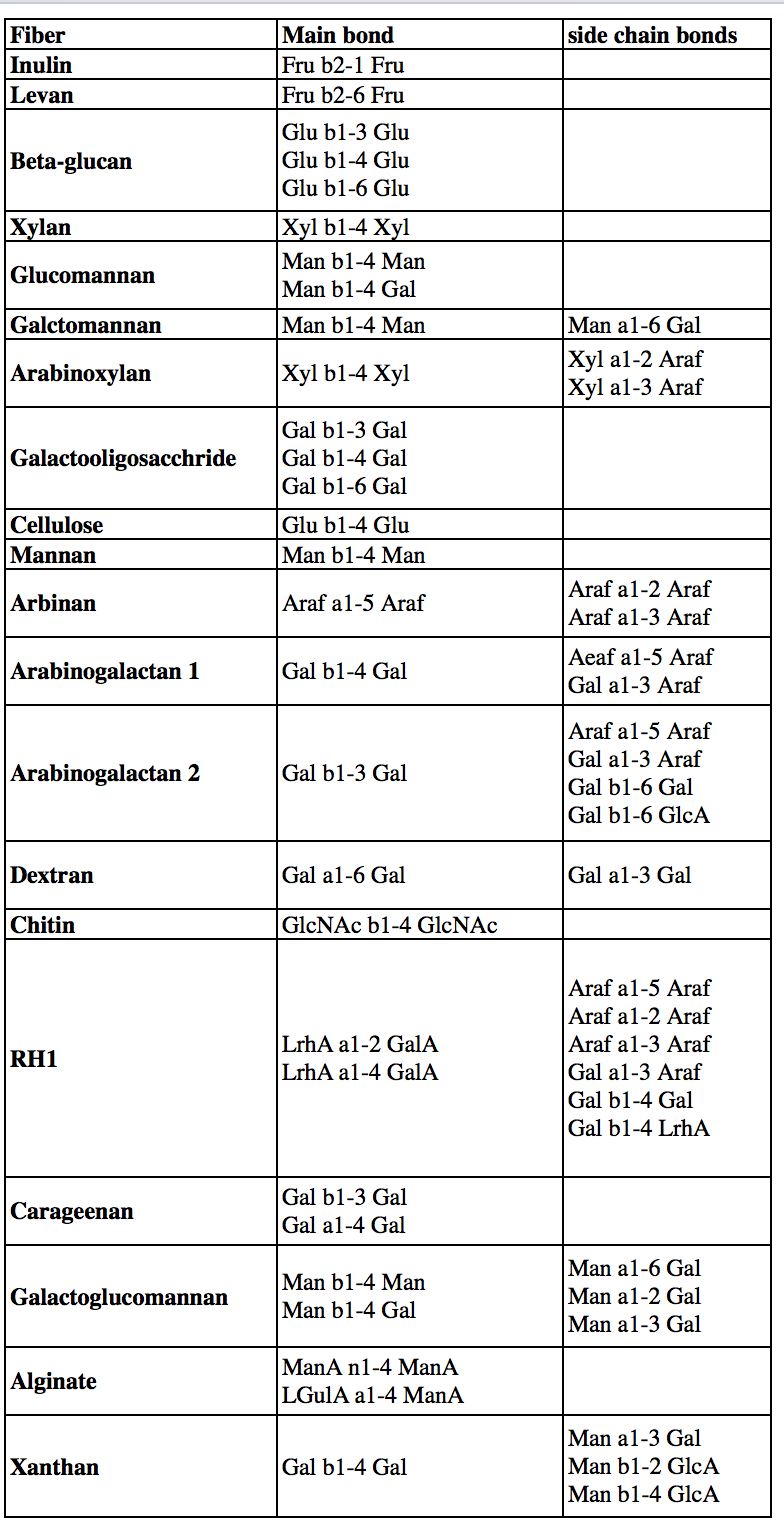


Table 2. The relationship between Glycosidic linkages and Substrates (Carbohydrates).



They only focus on 20 substrates (carbohydrates, they call dietary fiber in those paper).

After that, they use these two tables to make a matrix

Enzymes-Glycosidic linkages X Glycosidic linkages-substrates-> Enzymes-Substrate Matrix

Based on that, they use genomes files as input, and get the substrates as output.

They have 2 useful outputs for **each genome**s:

Figure1. EC number output (just a example).

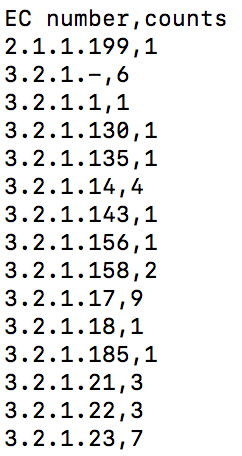
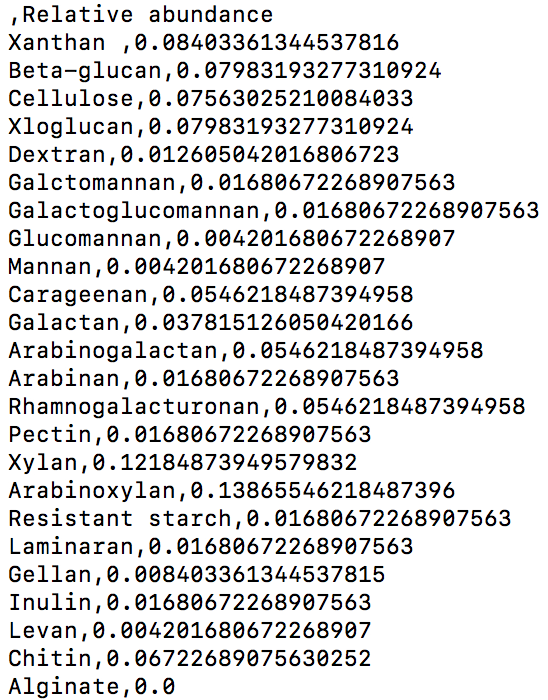


Figure 2. Substrate output (They shows all 20 substrates’ abundance).



DATA:

As we talked before, we would use 4 datasets same as dbCAN\_seq update to do the comparison.

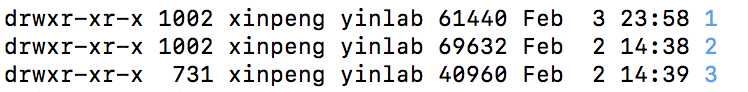
Both IFDP and dbCAN\_seq\_approach use the same input (faa, while dbcan\_seq need gff for CGCs).

Input: Genomes (faa and gff)

run\_dbCAN -> CGCs -> substrates

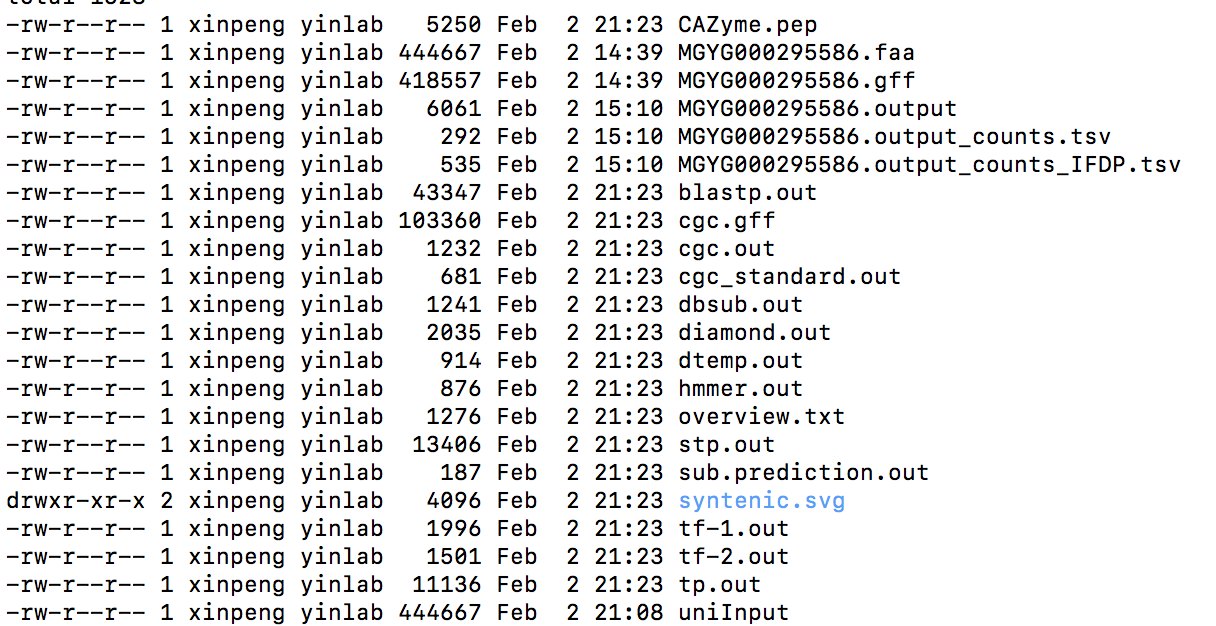
IFDP -> CAZymes (with ec number) -> substrates

Output of each folder:



Folder 1 and 2 contains 1k MAGs while folder 3 contains 729.

In MAG folder:



sub.prediction.out is the substrate-prediction result by dbCAN\_seq approaches.

MGYG\*\*\*\*\*\*\*\*.output\_counts.tsv is the result of EC in genome given by IFDP.

MGYG\*\*\*\*\*\*\*\*.output\_counts\_IFDP.tsv is the result of substrate in genome given by IFDP.

NOTICE:

MGYG000290600 and MGYG000294888 do not have CGC output.

2.5 Update

602 PULs (protein sequences) were run by IFDP and saved in the folder.