Breakout Session 8: Track A

AI/ML Ready Carbohydrate Enzyme Gene Clusters in Human Gut Microbiome

Dr. Yanbin Yin (Moderator)
Professor, University of Nebraska Lincoln

AI/ML Ready Carbohydrate Enzyme Gene Clusters in Human Gut Microbiome

Yanbin Yin (UNL)

2024 NIH ODSS AI Supplement Program PI Meeting

3/28/2024

Outline

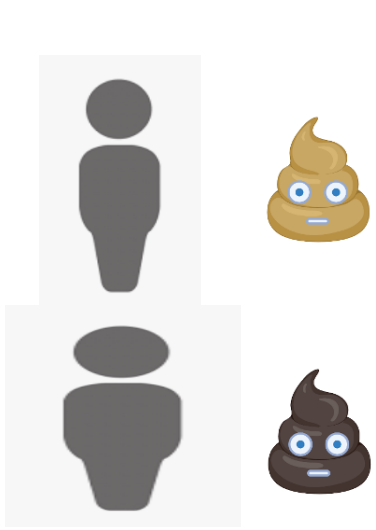
- Introduction to personalized nutrition, CAZymes, and parent R01
- dbCAN tool suite for CAZyme and CGC annotation
- AI/ML application in glycan substrate prediction for CGCs

R01 parent grant objective:

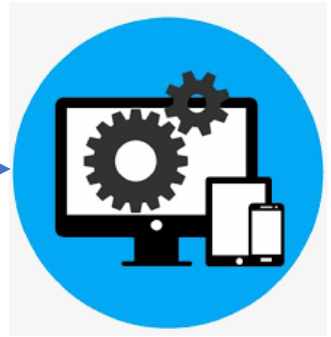
Microbiome-based personalized nutrition with bioinformatics tools

Where are CAZymes?

What fibers can you digest?



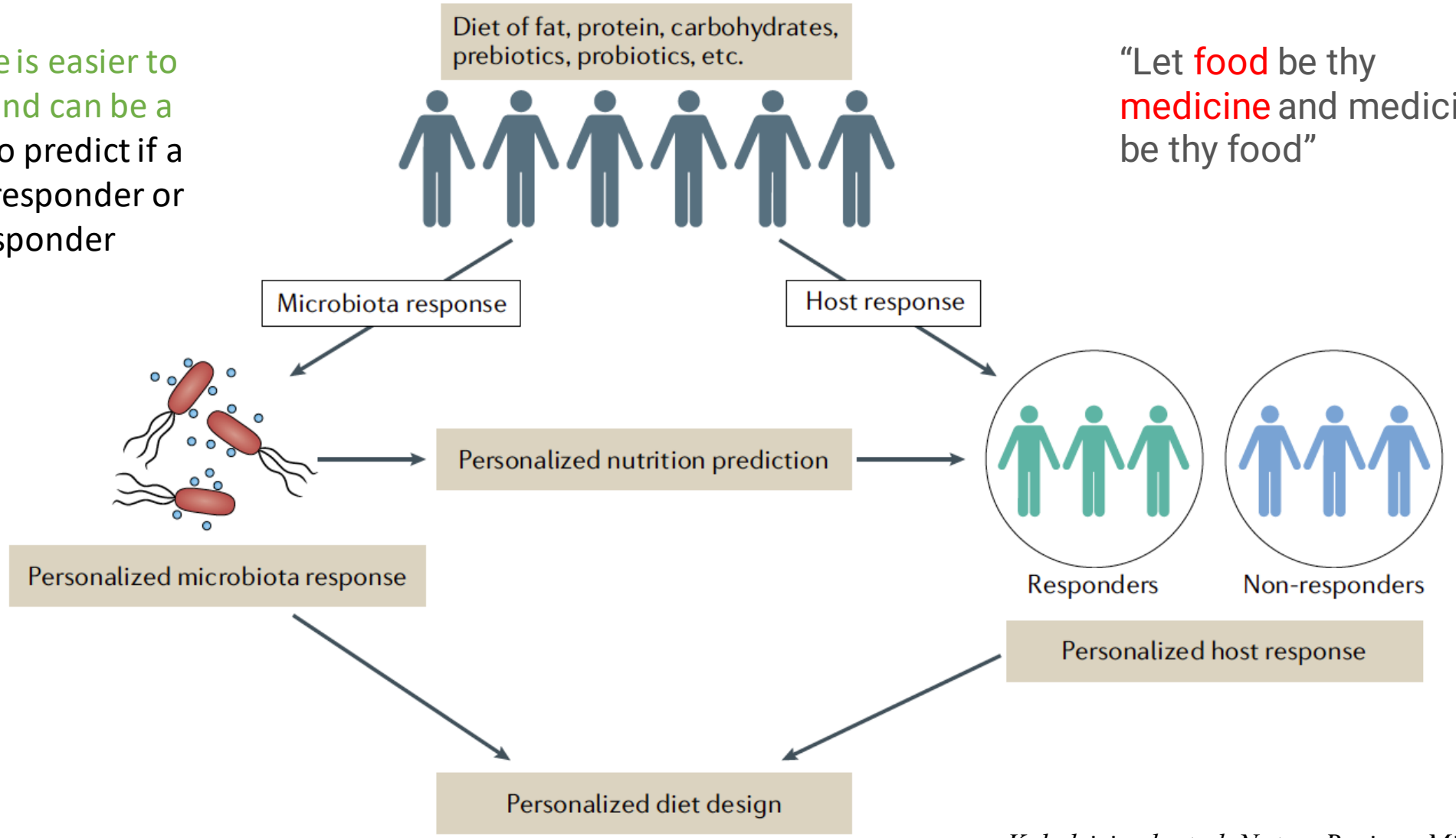
dbCAN



Personalized diet

Personalized nutrition aims to utilize inter-individual host and microbiome variations in generating data-driven personalized dietary recommendations

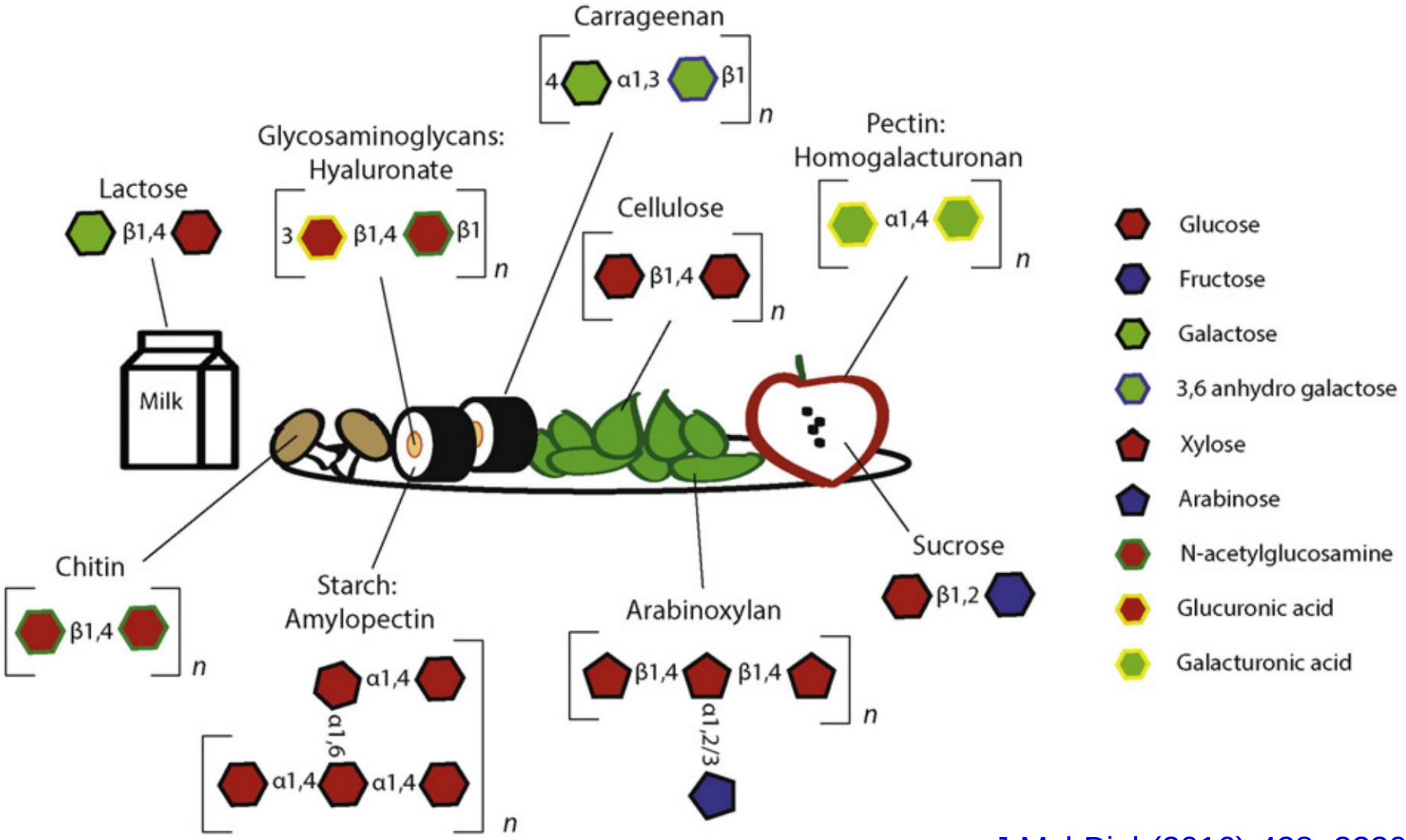
Microbiome is easier to modulate and can be a biomarker to predict if a person is a responder or non-responder



"Let food be thy medicine and medicine be thy food"

Kolodziejczyk et al, Nature Reviews Microbiology 2019

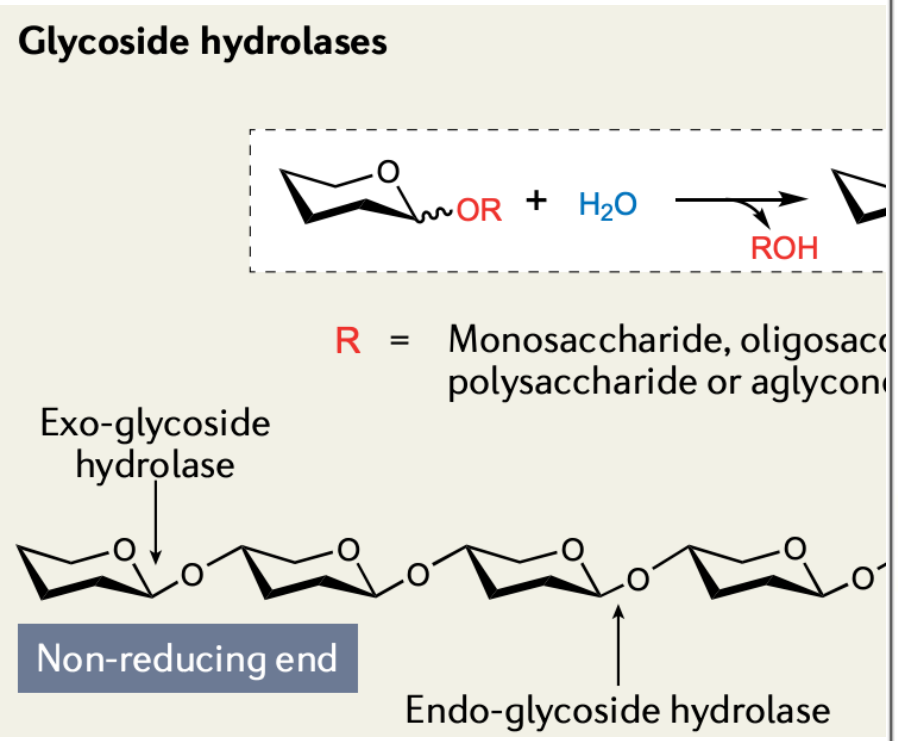
a high diversity of dietary fibers/glycans/carbohydrates

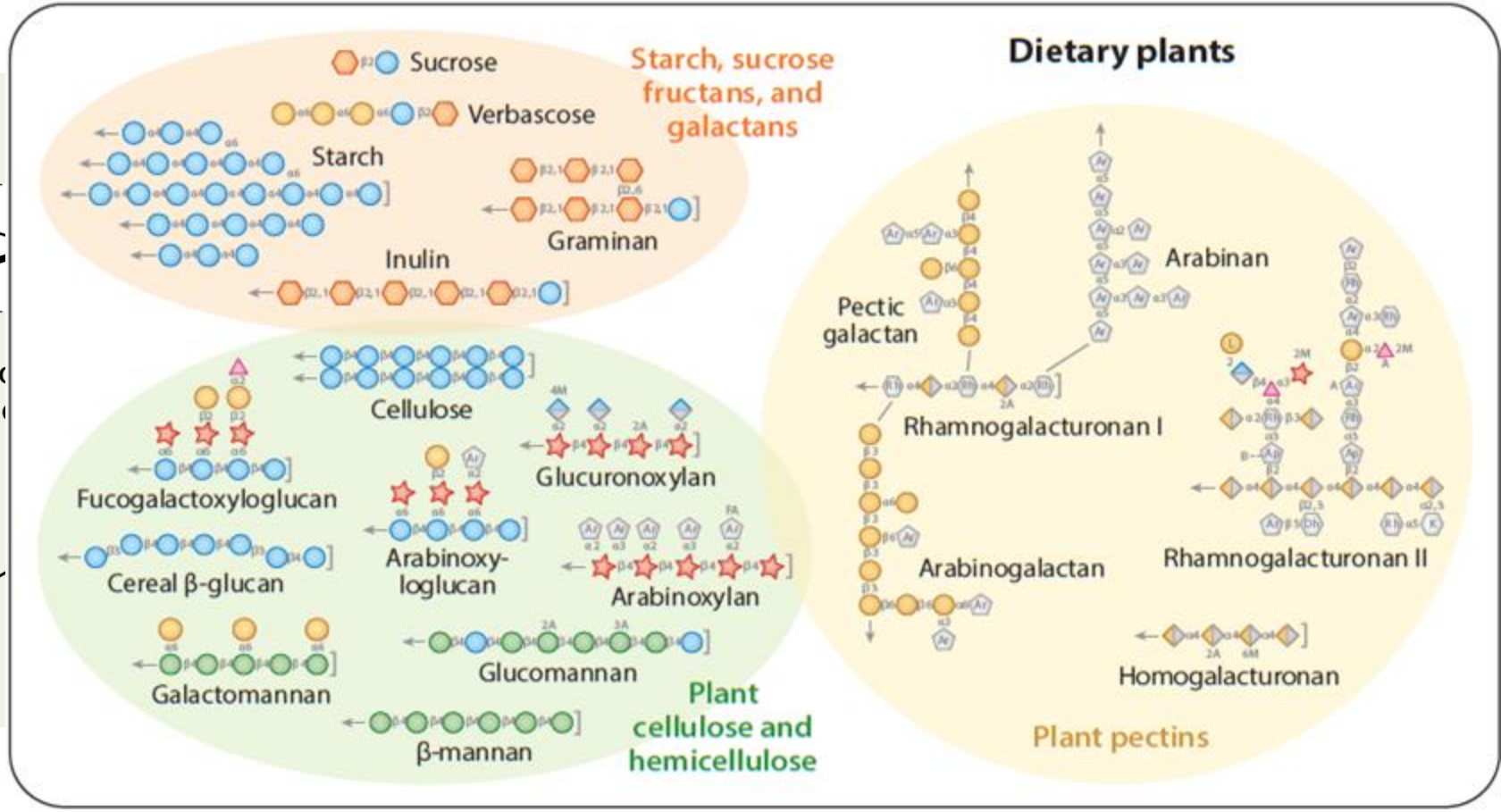


J Mol Biol (2016) 428, 3230–3252

CAZymes target glycosidic linkages in the dietary carbs

Nature Reviews Microbiology (2022)





Annu. Rev. Microbiol (2017) 71:349–69

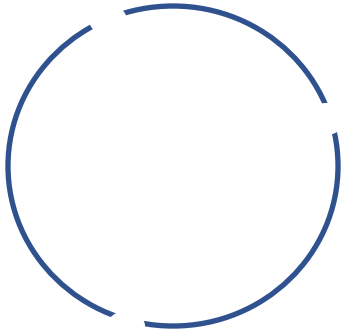
gut bacteria dedicate > 6% of their genes to CAZymes

Bacterium	Total CAZymes	GH	GT	PL	CE	Total CBMs
Bacteroides thetaiotaomicron VPI-5482	386	263	87	16	20	31
B. xylanisolvens XB1A*	349	224	81	22	22	26
B. vulgatus ATCC-8482	279	177	78	7	17	18
B. fragilis 638R	223	138	78	1	6	26
Roseburia intestinalis XB6B4*	175	115	46	0	14	11
Butyrivibrio fibrisolvens 16/4*	115	75	37	0	3	31
Ruminococcus champanellensis 18P13*	87	54	12	9	12	34
Bifidobacterium adolescentis ATCC15703	94	54	37	0	3	6

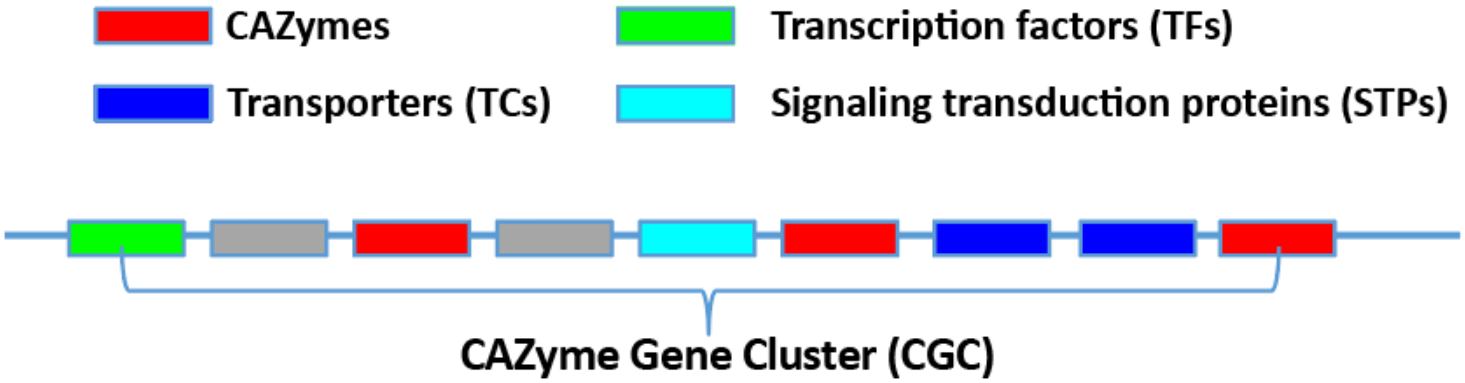
Gut Microbes 3:4, 289-306; 2012

1000 (species) x 100 (genes) = 100,000 CAZymes

dbCAN is a software for CAZyme and gene cluster prediction in bacterial genomes



predict genes
predict signature genes
call CAZyme gene clusters



Web server:

https://bcb.unl.edu/dbCAN2

300,000+ jobs in 10 years
8,000+ email addresses

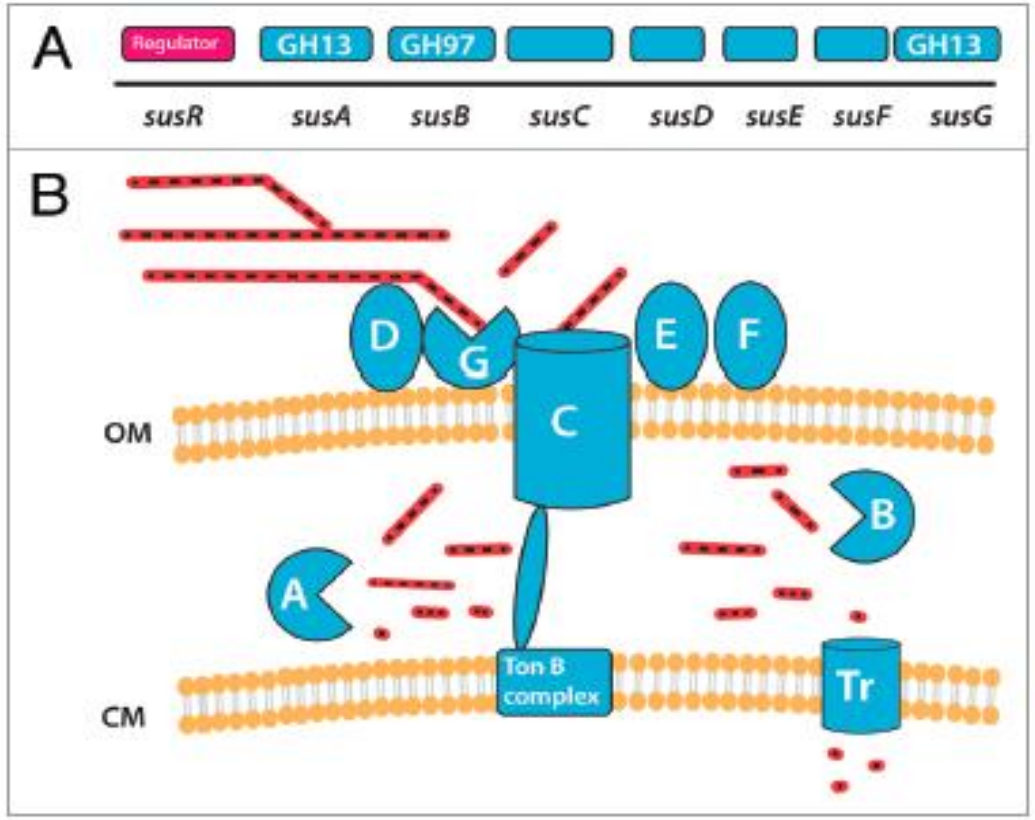
Python package:

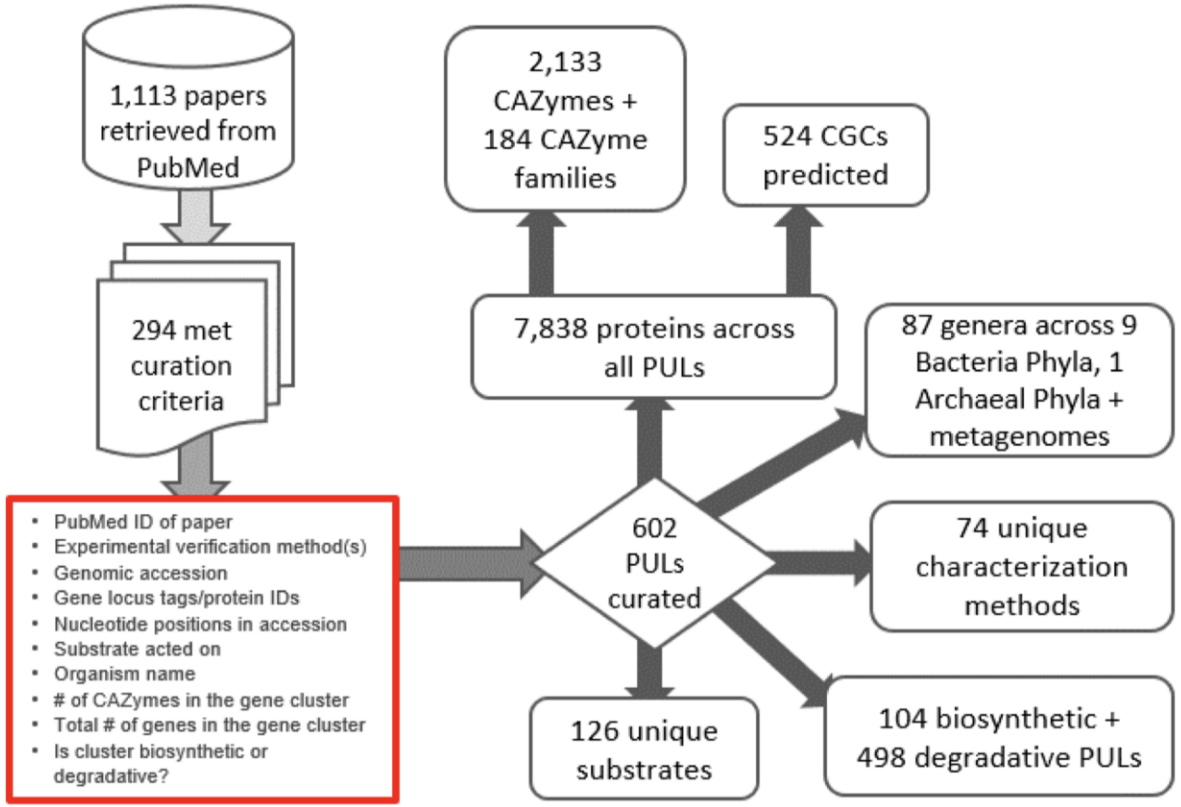
https://github.com/linnabrown/run_dbcan

dbCAN-PUL is a database with PULs/CGCs and their glycan substrates

PUL: polysaccharide utilization loci

Sus in Bacteroides thetaiotaomicron





Gut Microbes 3:4, 289-306; 2012

Ausland et al., Nucleic Acids Res 2021

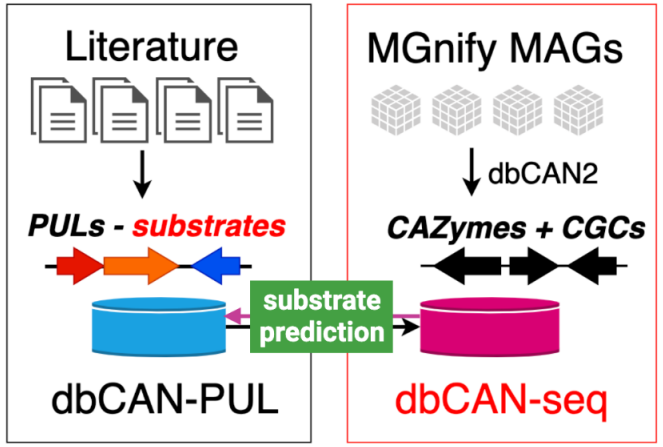
Machine learning models predict substrates for CGCs

https://bcb.unl.edu/dbCAN_seq

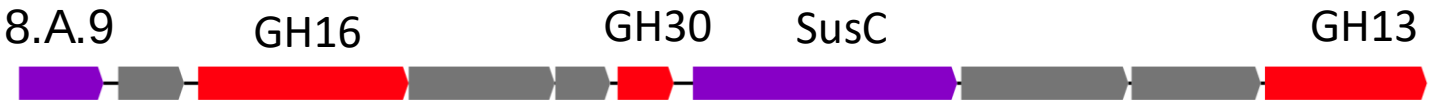
Unsupervised Data (**~250k ML/ML ready CGCs** from various microbiomes)

↓

Word2Vec Embedding



- Unsupervised ML model learns a vector representation for each family in CGCs.
- Consider the context of words in the large amount of texts

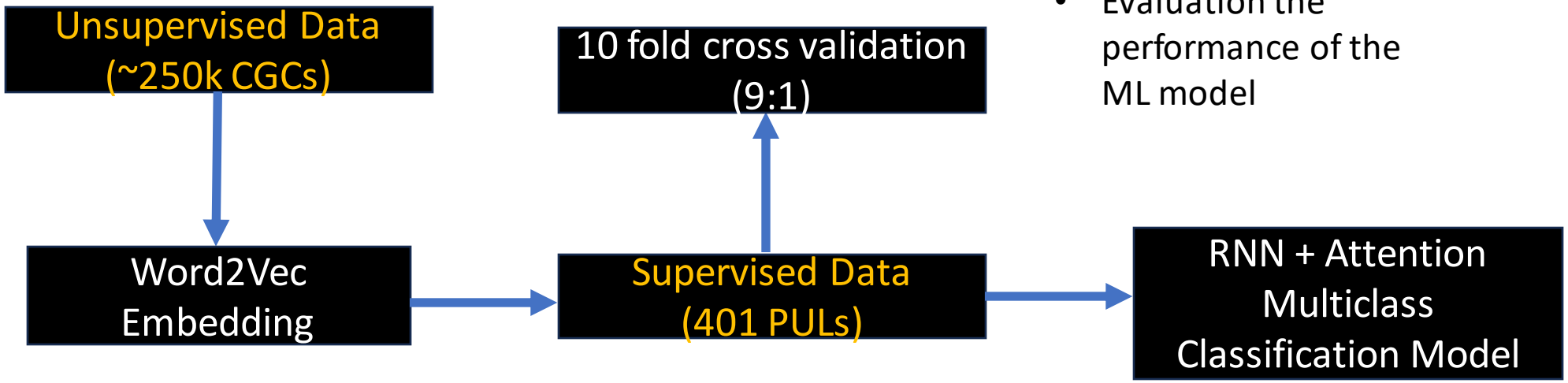


GH53,3.A.1,3.A.1,LacI,GH42
GH55,GH16_3,3.A.1,3.A.1,9.B.33
GH57,GT4,2.A.25,ACT,GH3
GH59,2.A.66,GntR,4.A.1,GH13_29
GH59,3.A.1,3.A.1,SBP_bac_1,GH30_9
GH5_1,GH9,1.A.22,Pribosyltran,2.A.40
GH5_13,GH146,HTH_AraC,1.B.14,GH146
GH5_13,GH2,3.A.1,3.A.1,GH43_32
GH5_2,2.A.38,2.A.38,Sigma70_r4,GH3
GH5_22,GH3,GH42,3.A.1,3.A.1
GH5_39,1.B.14,CE7,GerE,GH3
GH5_4,1.B.14,8.A.46,GH3,3.D.4
GH5_4,9.A.8,FeoA,FeoA,GH43_12
GH5_4,CE7,GH26,GH130,2.A.2
GH5_46,GH16_3,1.B.14,GH3,GH3
GH5_46,GH3,GH30_3,GH16_3,1.B.14
PL27,GH42,2.A.69,CBM67|GH78,2.A.66
PL37,GH154,GH88,3.A.1,3.A.1
PL38|GH88,GH2,GH3,GH30_3,1.B.14
PL42,GH105|GH154,GH43_24,2.A.37,3.A.1

Machine learning models predict substrates for CGCs

Ved

CGCs and PULs with similar family vector representations (i.e., semantic similarity) target the same glycans



- Evaluation the performance of the ML model

- unsupervised ML model learns a vector representation for each family in CGCs.
- Consider the context of words in the text

- Extract vectors for each family.
- Each PUL is a collection of families, and represented as a collection of family vectors.

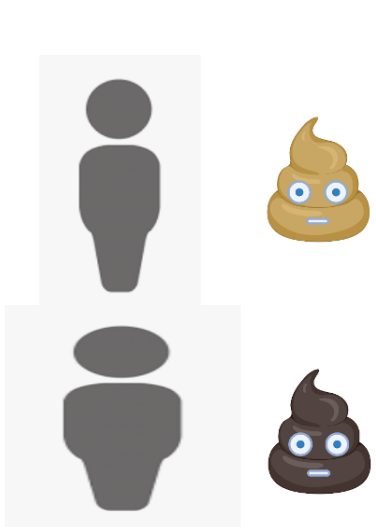
- Recurrent neural network takes the PUL vectors and predicts substrate for CGCs.
- Attention layer learns the weights for each family as importance towards the predicted category.

R01 parent grant objective:

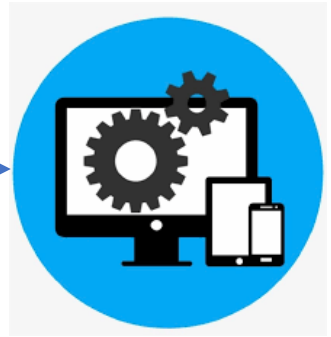
Microbiome-based personalized nutrition with bioinformatics tools

Where are CAZymes?

What fibers can you digest?



dbCAN



Personalized diet

Acknowledgements

Students:

Qiwei Ge
Yuchen Yan
Jinfang Zheng
Jerry Akresi
Xinpeng Zhang
Ved Pyrush