SCAN ME ARTICLE

Predicting *Clostridioides difficile* infections from microbiome data using Sourcepredict

Maxime Borry

Max Planck Institute for the Science of Human History
borry@shh.mpg.de

Max-Planck-Institut für Menschheitsgeschichte
Max Planck Institute for the Science of Human History

INTRODUCTION

- Predicting the category, or source, of a sequenced microbiome sample from its composition is a problem known as **sourcetracking.**
- Using dimension reduction algorithms, machine learning models, and labelled reference datasets, I developed Sourcepredict[1] to predict the sources of microbiome samples from their taxonomic composition.
- Here I showcase Sourcepredict on a *Clostridioides difficile* infection (CDI) 16s targeted sequencing dataset.
- *Clostridioides difficile* is a Gram positive opportunistic pathogenic bacteria that can proliferate in the human gut, usually after an antibiotic treatment disrupting the gut microbiome. CDI is one of the main causes of hospital associated infections in the US.

DATA

- **194** 16s healthy human gut microbiome samples[2]
- **146** 16s CDI human gut microbiome samples[2]

METHODS

- 16s OTU clustering and Taxonomic assignation with Dada 2[3] and Nextflow: **github.com/maxibor/dada2-nf**

DADA2 Amplicon Sequencing. Exactly. Version 1.12

- GMPR[4] normalization to account for variable sequencing depth.
- Sourcepredict performs t-SNE dimension reduction followed by K-Nearest Neighbors (KNN) classification.

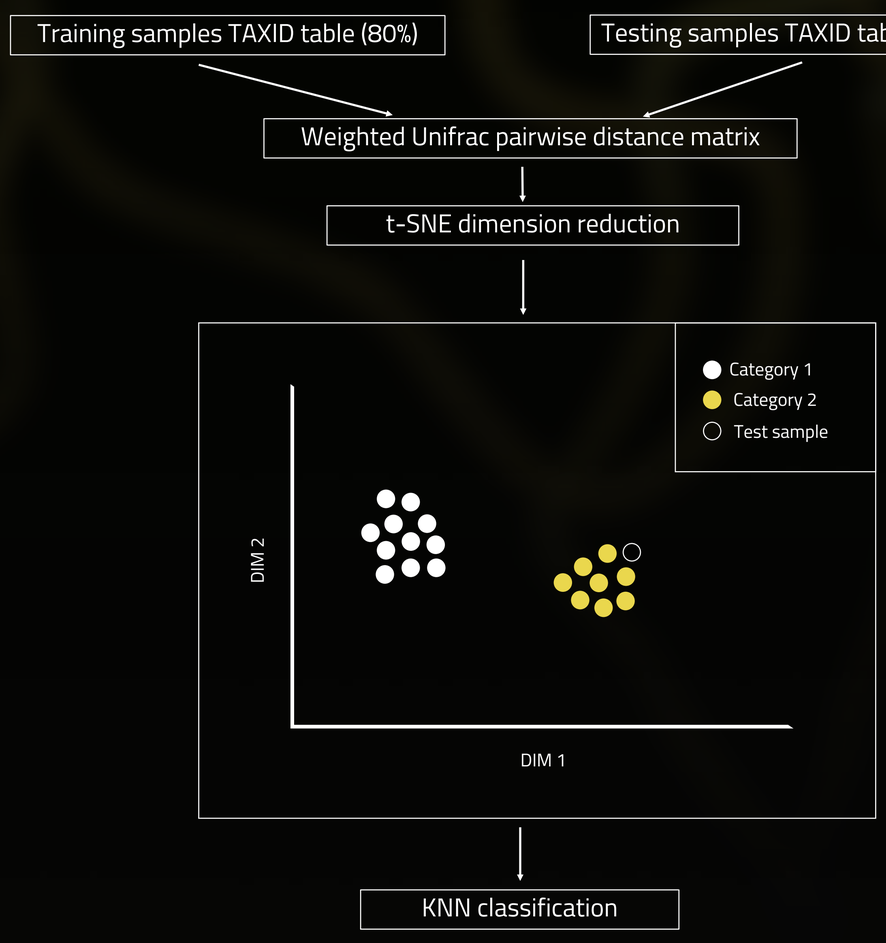

Figure 1 – Simplified overview of Sourcepredict.

RESULTS

1

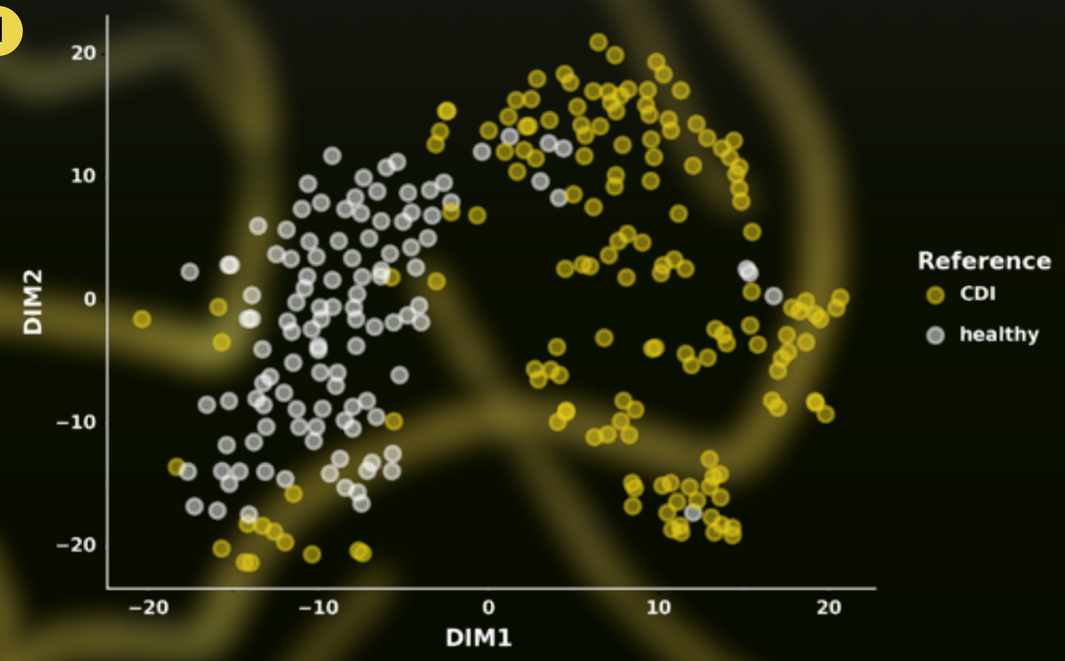

Figure 2– t-SNE embedding of the 272 training samples by Sourcepredict.

2

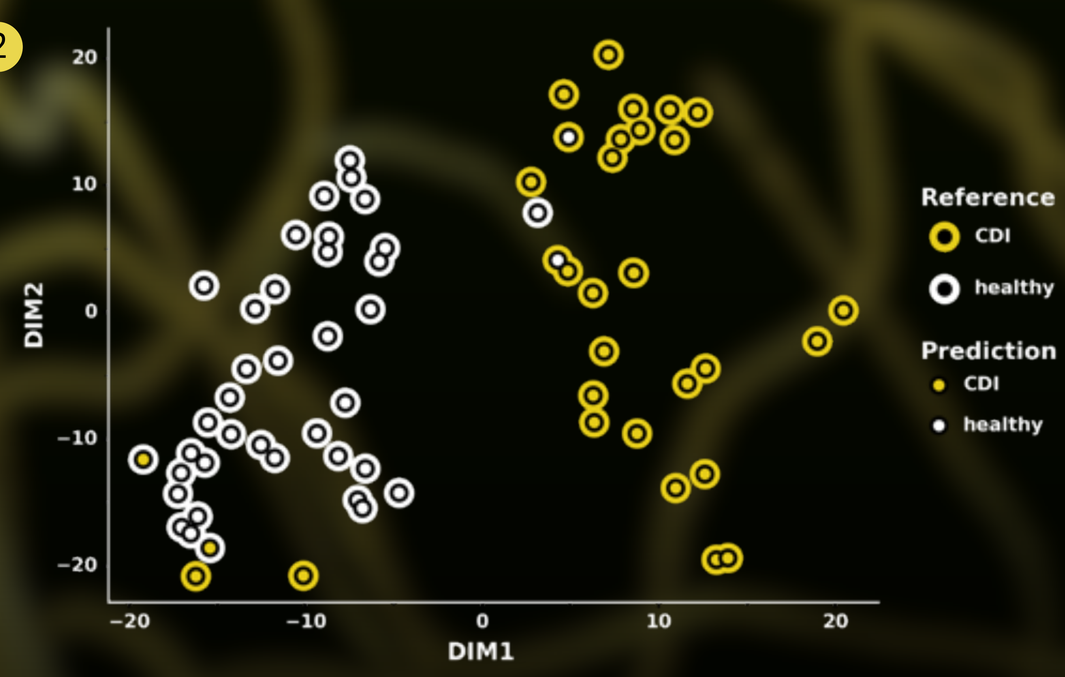

Figure 3– t-SNE embedding of the 68 test samples and their class as predicted by Sourcepredict.

3

	Sourcepredict	Sourcetracker2*[5]
Accuracy	0.94	0.8
Wall time	45s	2h 31m 18s

Table 1 – Comparison of Sourcepredict (v0.35) and Sourcetracker2 (v2.0.1-dev) performances.
*Sourcetracker2 didn't run to completion and classified only 46 samples out of 68.

CONCLUSION

1. Embedding the Unifrac distance matrix in two dimensions with t-SNE shows a good separation between classes.
2. With KNN classification, Sourcepredict can accurately classify most of the test samples.
3. Sourcepredict performs a better source prediction than Sourcetracker2 on this CDI dataset, both in accuracy and runtime.

References

[1] Borry, Maxime "Sourcepredict: Prediction of metagenomic sample sources using dimension reduction followed by machine learning classification". Journal of Open Source Software, 4.41, (2019):1540
[2] Seekatz, Anna Maria, et al. "Dynamics of the fecal microbiome in patients with recurrent and nonrecurrent Clostridium difficile infection." *Genome medicine* 8.1 (2016): 47.
[3] Callahan, Benjamin J., et al. "DADA2: high-resolution sample inference from Illumina amplicon data." *Nature methods* 13.7 (2016): 581.
[4] Chen, Li, et al. "GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data." *PeerJ* 6 (2018): e4600.
[5] Knights, Dan, et al. "Bayesian community-wide culture-independent microbial source tracking." *Nature methods* 8.9 (2011): 761.

Acknowledgements

- Dr.Christina Warinner, Dr. Alexander Herbig, Dr.AB Rohrlach, and Alexander Hübner for proofreading and their valuable comments the Sourceprecict manuscript.
- The JOSS reviewers for their valuable reviews and comments which improved Sourcepredict manuscript and software.
- This work was funded by the Max Planck Society and the Deutsche Forschungsgemeinschaft, project code: EXC 2051 #390713860.
- Background image: Kateryna Kon/Shutterstock.com