Using GNNs and Microbiome Co-occurrence Networks to Predict Inflammatory Bowel Disease Diagnosis

Advait Patil, Emily Chen, Jack Michaels
Department of Computer Science
Stanford University
353 Jane Stanford Way, Stanford, CA 94305
{advaitp, emilyc02, jackfm} @ stanford.edu

1 Introduction

The human microbiome is composed of hundreds of species of bacteria, fungi, protozoa, and viruses that reside within our body. These organisms comprise dynamic, complex communities that exert significant control of host physiology through a number of mechanisms. The gut microbiome is emerging as an important manipulatable modality in the biosciences, with associations to nearly every major human body system.

The gut microbiome forms a complex interaction network, with different strains supporting or inhibiting growth of others through production of small molecules. This network of interactions between microbial strains contains important information relevant to the prognosis of many chronic diseases, including Inflammatory Bowel Disease (IBD), a chronic, immune-mediated disease that affects the gastrointestinal tract of over 3 million US adults. However, microbial interaction networks remain poorly characterized in the literature at scale, with traditional statistical methods suffering from challenges of high dimensionality and sparsity.

2 Dataset

We will use metagenomic sequencing data from the Inflammatory Bowel Disease Multi'omics Database, curated by a large collaboration spearheaded by the Huttenhower group at the Harvard Chan Center for the Microbiome in Public Health. 132 participants from five academic medical centers were studied, involving 1,785 stool samples that underwent metagenomic sequencing. These sequencing reads are aligned to >1 million clade specific marker genes with MetaPhlAn2, resulting in strain abundance measurements for approximately 17,000 reference genomes. These final data take the form of an Operational Taxonomic Unit (OTU) matrix, where columns in the matrix correspond to the sample and rows correspond to individual microbial species.

We will use SparCC (Friedman and Alm, 2012), the field standard module for computing correlations in compositional metagenomic data, to infer microbial interaction graphs for each patient from the OTU tables. An example co-abundance network can be seen in Figure 1.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

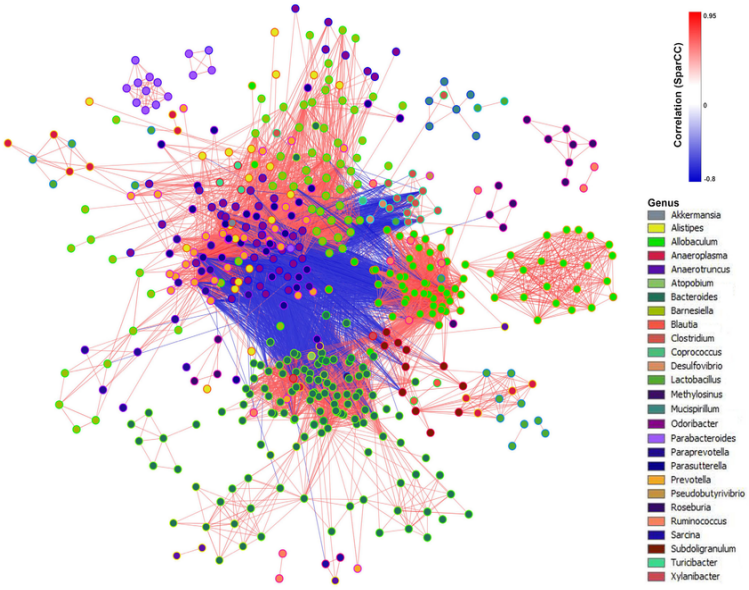

Figure 1: *An example SparCC microbial interaction graph in liver disease (Xie et al. 2016).*

Our task is to predict whether a patient has IBD or has a microbiome that is predisposed to development of IBD from microbial abundances. To our knowledge, no studies to date have attempted to use graph neural networks and microbiome co-occurrence networks to predict IBD risk.

We chose this dataset because it provides the most comprehensive description to date of host and microbial activities in inflammatory bowel diseases, permitting examination of microbial strain abundances on the largest cohort of patients to date. Clinical metadata for each sample are available as well, providing measures of IBD severity as measured by physicians during endoscopy.

3 Techniques

3.1 Baseline Model

As an initial baseline, we can naively approach the problem as an embedding task and utilize GraphSAGE. By sufficiently learning representations for microbial strains from a representative network, we can apply vector similarity metrics to said representations where a higher similarity score constitutes a more positive relation between two given microbial strains. Potential drawbacks to GraphSAGE include limited relational information between two microbial strains in regards to the specific impact a particular strain has on another and the overall magnitude of a particular relation. We foresee regularization issues arising due to the nuance GraphSAGE is required to employ to learn embeddings which capture the complex relational dynamics present in the data. Either through a lack of model sophistication or through said restrictive nuance, GraphSAGE is likely to be unable to predict novel relations between novel microbial strains, hence the necessity of a more complex model.

$$\mathbf{h}_v^{(l)} = \sigma\left(\mathbf{W}^{(l)} \cdot \mathrm{CONCAT}\left(\mathbf{h}_v^{(l-1)}, \mathrm{AGG}\left(\left\{\mathbf{h}_u^{(l-1)}, \forall u \in N(v)\right\}\right)\right)\right)$$

Figure 2: *Equation representing GraphSAGE*

3.2 Model

To capture and predict the sophisticated relations present between microbial strains, a sufficiently deep multi-layered graph neural network must be employed. To do so, we can use a graph convolutional network (GCN) as it is transductive and we have complete graph structures for training. GCN layers can be applied in succession with attention, dropout and batch normalization to advance the model

while a single graph attention network (GAT) layer can be applied before aggregation to theoretically increase nuance.

We foresee difficulty surrounding over-smoothing since multiple GCN layers may be required to increase accuracy over the baseline. Deciding how many layers of GCN to include will heavily influence the effectiveness of our model and will be an important open question which will ultimately depend on results from initial testing. Potential pre-processing involving a preliminary GraphSAGE layer to generate effective embeddings as higher dimensional input to the GCN layers may be warranted, though training time must be taken into consideration. To mitigate this and unnecessary complexity, skip connections can be applied between GCN layers. Since relations between microbial strains will be represented by some score, we can assess correct predictions via the predicted value being within some hyperparameter (epsilon) distance from the true value. Thus, we have chosen F1 and accuracy as our desired evaluation metrics.

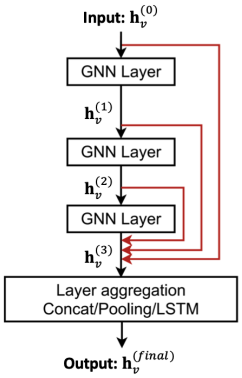

Figure 3: *Skip connections and their role in multi-layered GNNs. For our model we will be using GraphSage, GCN, and GAT layers.*

4 References

Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8(9):e1002687. doi:10.1371/journal.pcbi.1002687

Zhu Q, Jiang X, Zhu Q, Pan M, He T. Graph Embedding Deep Learning Guides Microbial Biomarkers' Identification. Front Genet. 2019 Nov 22;10:1182. doi: 10.3389/fgene.2019.01182. Erratum in: Front Genet. 2020 May 15;11:487. PMID: 31824573; PMCID: PMC6883002.

Xie G, Wang X, Liu P, Wei R, Chen W, Rajani C, Hernandez BY, Alegado R, Dong B, Li D, Jia W. Distinctly altered gut microbiota in the progression of liver disease. Oncotarget. 2016 Apr 12;7(15):19355-66. doi: 10.18632/oncotarget.8466. PMID: 27036035; PMCID: PMC4991388.

Gong H, You X, Jin M, et al. Graph neural network and multi-data heterogeneous networks for microbe-disease prediction. Front Microbiol. 2022;13:1077111. Published 2022 Dec 22. doi:10.3389/fmicb.2022.1077111

Wang, F., Huang, ZA., Chen, X. et al. LRLSHMDA: Laplacian Regularized Least Squares for Human Microbe–Disease Association prediction. Sci Rep 7, 7601 (2017). https://doi.org/10.1038/s41598-017-08127-2

Chen X, Zhu Z, Zhang W, et al. Human disease prediction from microbiome data by multiple feature fusion and deep learning. iScience. 2022;25(4):104081. Published 2022 Mar 16. doi:10.1016/j.isci.2022.104081