Neural Embeddings for 16S Microbiome Classification

Ryan Conrad[1], Ryan Inghilterra[1], Sean Rowan[1], Brandon Westerberg[1]

Advisors: Rob Knight[2], Daniel McDonald[2]

1 University of California San Diego, Master of Advanced Study, Data Science and Engineering

2 University of California San Diego, Departments of Pediatrics and Computer Science & Engineering

Problem Statement

The rapid proliferation of next generation sequencing technology has allowed for an ever increasing amount of gut metagenomic data to become available to researchers. While more data is available than ever before the sheer size of genomic data that can be extracted from a cohort of tens of thousands of individuals can be challenging to both interpret and model. With such large and sparse data we focus on applying state of the art embedding techniques to compress the data while retaining as much semantic information as possible about the community structure of the microbes within a sample.

Using neural embedding techniques we demonstrate how 16S sequencing data from a microbiome samples can be represented in a low dimensional vector space while still retaining their important semantic information. In our research we also show how body site classification can be accomplished using these embedding as features with accuracy rivaling the gold standard multidimensional scaling technique.

Data Science Pipeline

Acquire Data

Genomic data from the American Gut Project was queried through Qiita, an open-source microbial study management platform. Along with QIIME, an open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data, the raw sequence data was processed into distinct groups of microbes known as operational taxonomic units (OTUs).

Preprocess Data

In order to generate embeddings from the processed genomic data several intermediate preprocessing steps were required to shape the OTU data into a form that could be consumed during the embedding fitting process. Each embedding used in the study operated in fundamentally different manners and thus each one required independent preprocessing pipelines.

Fit Embedding

Once each dataset was prepared it could be transformed using its respective embedding model. To accelerate the process of fitting embeddings AWS EC2 instances with GPUs were utilized. Word2vec and Lorentz hyperbolic embeddings were chosen as the two methods to be evaluated against the control embedding technique Principal Coordinates Analysis with Unifrac distance.

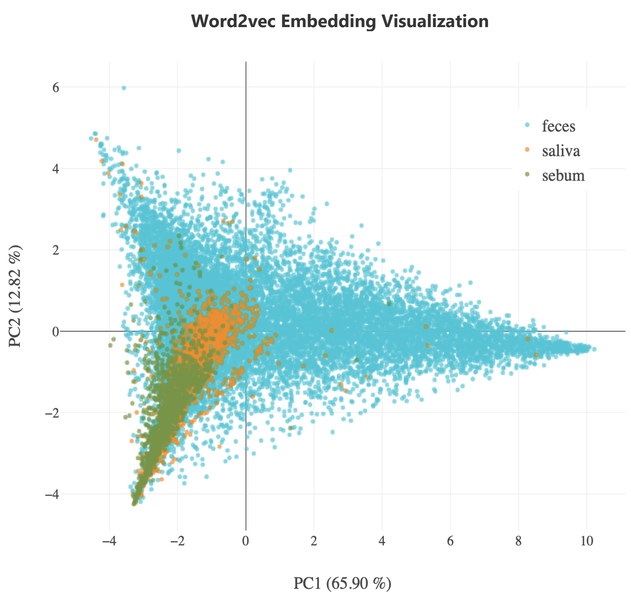
Model Data

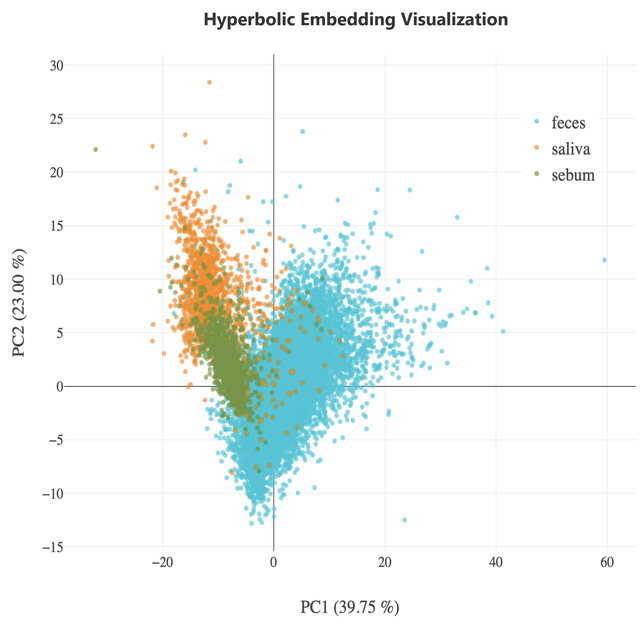
Embeddings were validated by using them as features for body site classification. The three most common sample types were chosen to evaluate the models on: feces, saliva, sebum. Embeddings were finally evaluated based on their performance on this classification task.

Key Insights

Figure 2: Embedding Visualization

Using PCA both embeddings were projected down to two dimensions to more clearly visualize the distinct body site clusters. While each embedding achieves clear separation between body sites there are distinct structural differences between each method.





Final Solution Architecture

Using Spotify's Luigi framework, an end to end ETL solution was developed to automate data extraction, preprocessing, and model training. The data pipeline allowed for modular design of each step in the process and ensured reproducible results as new data became available from the American Gut Project.

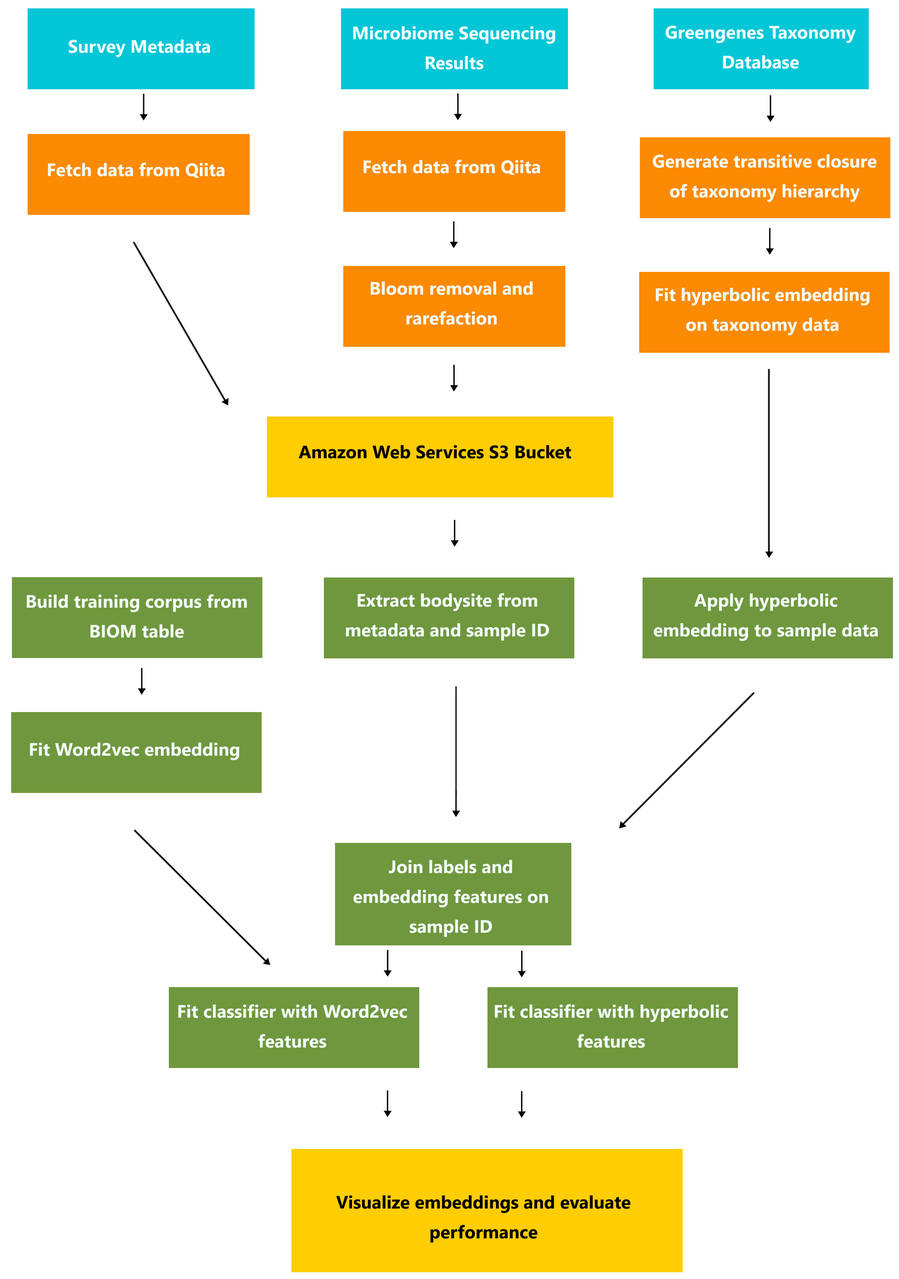


Figure 1: Dataflow diagram

Using Luigi each data processing step was broken down into a modular piece. The final pipeline described by the block diagram was used to used to generate, evaluate, and visualize embeddings given raw data as its input.

Model	PC1 Explained Variance	PC2 Explained Variance	PC3 Explained Variance	Training Accuracy	Test Accuracy	Test F1 Score
No Embedding (19105 Dimensions)	40.1 %	5.0 %	4.3 %	99.0 %	98.3 %	0.983
Unifrac Distance + PCoA (3 Dimensions)	18.4 %	19.4 %	4.4 %	95.5 %	95.8 %	0.958
Word2vec (80 Dimensions)	65.8 %	12.8 %	4.2 %	88.9 %	88.7 %	0.849
Lorentz Embedding (10 Dimensions)	39.7 %	23.0 %	9.3 %	96.6 %	96.2 %	0.961

Figure 3: Model results for body site classification task

Body site classification results were generated using different embeddings as features for a logistic regression classifier. Data was split into training and test sets containing 67% and 33% of the total data respectively. PCA was performed on embeddings to determine explained variance for the top three eigenvectors.

Acknowledgements

Thank you to our program advisor Illkay Altintas, our academic advisors Rob Knight and Daniel McDonald, and the thousands of American Gut participants who made this research possible.