Evolutionary-Informed Contrastive Learning for the Microbiome

Linnea M. Wolniewicz [1], Yannik Glaser [1], Peter Sadowski [1]

Department of Information and Computer Sciences, University of Hawai'i at Mānoa [1]

Introduction

Advances in metagenomic sequencing technology have enabled scientists to quantify the composition of microbial communities from environmental samples [1]. These sequencing methods enable scientists to quantify microbe abundance in samples, given 16S amplicon sequence variant (ASV) data. From quantified abundances, biologists may attempt to understand the dependencies within microbe communities, their stability, and their interactions with their host.

The small sample size and high dimensionality of metagenomic data has made it difficult to find patterns. The goal of this work is to address this issue by using self-supervised learning techniques to embed sample data to a useful, lower-dimensional latent space. These embedded samples can then be used for the classification of various metagenomic metadata, such as sample provenance. Specifically, this work utilizes sample data taken from the Earth Microbiome Project (EMP) [3] and the contrastive learning method Barlow Twins [2].

Dataset

Earth Microbiome Project (EMP): The EMP contains 27,151 unique samples taken from 97 individual studies [3]. Each sample contains the ASV abundances from a set of 309,457 possible ASVs, all sequenced from the 16S gene. Each sample is associated with various metadata, including the classification tasks listed in Table 1 below.

We reduce the dimensions of the raw microbe abundance data before training by removing ASVs with counts ≤ 500, resulting in a set of 41,854 possible observed ASV sequences in each sample. Our model embeds this ASV abundance data to a final dimension of 32, evaluated on the downstream classification tasks listed below.

Task	Description	N Classes
Habitat	Global ecological context of a sample	13
Biome	Broad ecological context of a sample	43
Material	General material displaced by the sample	45
Feature	Local ecological context of a sample	97
Sample Type	Specific material displaced by the sample	120

Table 1:Microbe provenance classification tasks

Self-Supervised Learning

Contrastive learning is a self-supervised learning technique developed for high-dimensional images, as an alternative to the autoregressive self-supervised learning used to train model such as ChatGPT. Specifically, the contrastive Barlow Twins framework introduced by [2] feeds two identical networks distorted versions of the sample, and minimizes the cross-correlation matrix between the outputs of both networks.

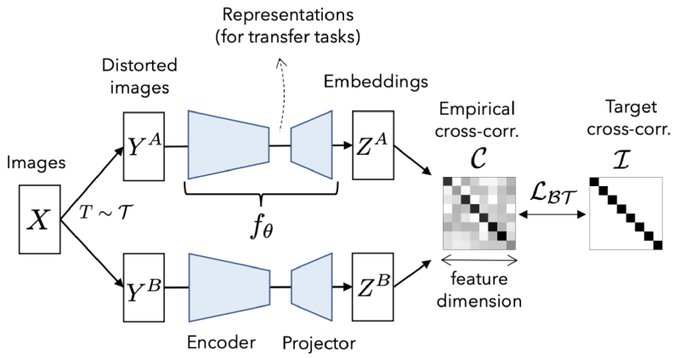

Figure 1:The Barlow Twins objective function [2]

In our application of the Barlow Twins framework we apply the following data augmentations:

- Bootstrap observed ASVs
- Set random observed ASV counts to a normally sampled value
- Add noise from a normal distribution to all observed ASV counts
- Take the log of all ASV counts

Evolution-Informed Neural Networks

Our encoder architecture takes advantage of the sparsity of ASV abundance data, as well as the hierarchical clustering of correlated features found in phylogenetic trees, by using the evolutionary-informed neural networks (NNs) of (shown below) [1].

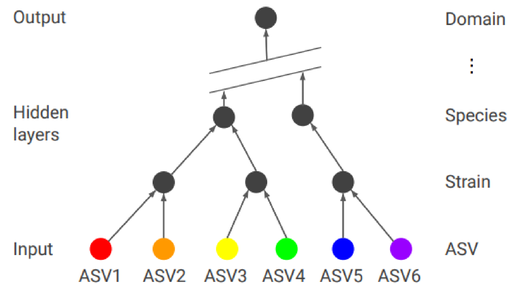

Figure 2:The evolutionary-informed neural network structure from [1]

Results

Task	Raw Data AUROC	Raw Data ACC	Evolutionary-Informed AUROC	Evolutionary-Informed ACC	Fully Connected AUROC	Fully Connected ACC
Habitat	0.990	0.982	0.991	0.983	0.990	0.982
Biome	0.912	0.814	0.508	0.038	0.775	0.559
Material	0.920	0.828	0.527	0.076	0.813	0.634
Feature	0.903	0.800	0.515	0.040	0.773	0.543
Sample Type	0.885	0.759	0.500	0.006	0.769	0.534

Table 2:Accuracy and AUROC of a linear classifier trained on 41,584 dimension raw data compared to a classifier trained on 32 dimension encoder embeddings. The evolutionary-informed neural network was trained for 50 epochs, and the fully-connected neural network was trained for 1,000 epochs

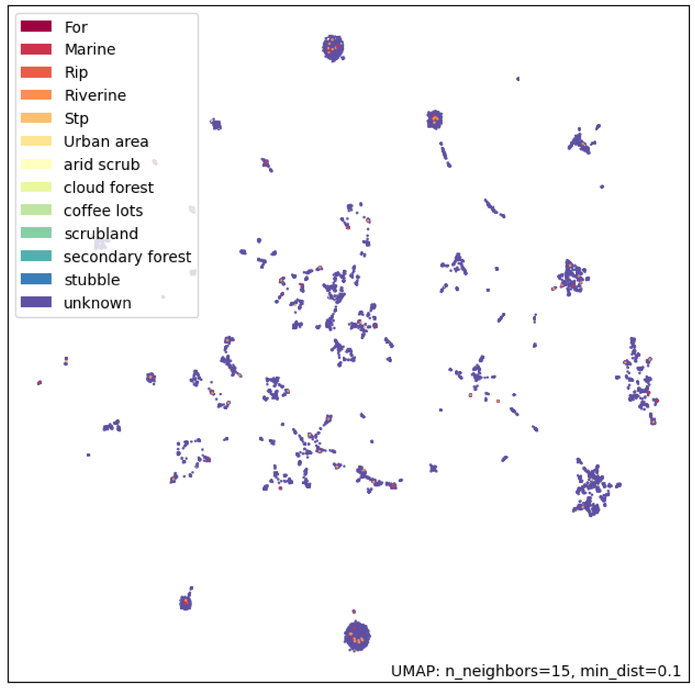

Figure 3:Visualization of the 32 dimensional embedding from the highest performing evolutionary-informed neural network encoder, colored by Habitat

Contributions

- Learned low-dimensional embedding of microbe abundance data
- Useful metagenomic sample data embeddings, which are applicable to many downstream shallow machine learning tasks
- Demonstration of the usefulness of scientifically-informed model architectures, specifically evolutionary-informed neural networks

References and Acknowledgments

This work was funded by NSF Award 2124922.

[1] Michael Ito, Yannik Glaser, and Peter Sadowski. "Evolution-Informed Neural Networks for Microbiome Data Analysis." 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2021): 3386-3391.

[2] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction." arXiv preprint 2103.03230.