Computational Microbiome Analysis Workshop Report

Nir Borger (ID: 313580920)

Naama Tsiben (ID: 319013926)

Introduction

Predicting the composition of gut microbiomes over time is a critical challenge in computational biology, with profound implications for understanding host health, ecology, and evolution. This project focuses on developing predictive models for the microbiome composition of wild baboons in the Amboseli ecosystem, leveraging extensive longitudinal data from the Amboseli Baboon Research Project. The primary goal is to extrapolate microbial relative abundances at future time points, thereby enhancing our ability to anticipate changes in microbiome structure and function.

The dataset encompasses microbial relative abundance profiles and detailed metadata for 60 baboons, including variables such as age, sex, social group, diet, rainfall, and seasonality. Building upon the findings of Björk et al. (2022), which revealed largely idiosyncratic gut microbiome dynamics with weak synchrony even among baboons sharing the same environment, this project addresses the inherent challenges posed by individual variability in microbiome prediction.

Our current computational pipeline employs Random Forest regression to interpolate missing data and Vector Autoregression (VAR) for temporal modeling. Model performance is evaluated using the Bray-Curtis distance to quantify dissimilarities between predicted and observed microbiome compositions. Despite integrating advanced statistical methods and accounting for both temporal and non-temporal predictors, the pipeline has yielded suboptimal results. The high dissimilarity scores indicate that the models struggle to capture the complex, individualized dynamics of the baboon gut microbiome.

In light of these challenges, the project aims to refine the predictive approach by exploring alternative modeling strategies that can better account for individual-specific factors and the idiosyncratic nature of microbiome dynamics. By improving predictive accuracy, we hope to contribute to a deeper understanding of microbial ecology in wild populations and inform future research on the interplay between microbiomes and host health.

Methods

The pipeline predicts gut microbiome composition by integrating machine learning and temporal modeling. It involves data preprocessing, interpolation, temporal modeling, and validation.

Data Preprocessing: The dataset includes microbial relative abundances and metadata with variables such as age, social group, diet, and seasonality. Missing metadata values were imputed using forward and backward filling for categorical variables and linear regression for continuous ones. Synthetic samples for unobserved dates were generated using metadata imputation and date-filling procedures. Features were standardized for compatibility with machine learning models.

Data Interpolation: Missing microbiome compositions were interpolated using a Random Forest Regressor trained on metadata and available microbiome data. Predicted values were normalized to ensure they summed to one, maintaining the constraints of relative abundance data.

Temporal Modeling: Temporal dependencies were modeled using Vector Autoregression (VAR) to iteratively forecast future microbiome compositions based on interpolated data.

Model Validation: A five-fold cross-validation scheme stratified by baboon ID was used. For each fold, Random Forest interpolation and VAR modeling were applied to the training set, and predictions for the validation set were compared to actual values using Bray-Curtis dissimilarity. This metric quantified prediction accuracy and was averaged across samples in each fold.

Testing Phase: The trained models were applied to test data, including short and sliced time series. Predicted compositions were generated for missing samples and saved for evaluation.

Software and Tools: The pipeline was implemented in Python using pandas, NumPy, scikit-learn, statsmodels, and scipy. Code was version-controlled on GitHub for reproducibility.

Results

The pipeline's performance was evaluated using Bray-Curtis dissimilarity scores, comparing predicted microbiome compositions to observed values. These results were assessed against both a naïve prediction baseline and across specific temporal patterns.

The average Bray-Curtis dissimilarity for the model predictions was 0.4, compared to 0.38 for naïve predictions, indicating that the model performed similarly to the baseline overall. This reflects challenges in capturing the inherent variability of microbiome compositions.

When stratified by the type of time series, for regular time series, the model showed a mean decrease in performance (-0.022), as compared to the naïve baseline.
For long time series, the model outperformed the naïve baseline with a mean improvement of 0.041. This suggests that the pipeline's temporal modeling component is better suited to extrapolate patterns over extended periods.

Bray-Curtis dissimilarities were also analyzed per baboon ID. Variability in performance indicated that the pipeline was sensitive to individual-specific microbiome dynamics, aligning with prior findings of strong idiosyncratic patterns in baboon gut microbiomes.

A mixed linear model revealed no significant relationship between prediction error (Bray-Curtis dissimilarity) and the temporal difference in days between samples ($p = 0.287$). The intercept value of 0.383 aligns with the average baseline Bray-Curtis scores.

Discussion

This study underscores the complexities of predicting microbiome compositions, particularly in dynamic and idiosyncratic systems like those of wild baboons. One notable observation was the frequent alignment of model predictions with the overall average microbiome composition, particularly for short time series or sparse data. This could reflect a stabilizing effect, where microbiome compositions are distributed around a central tendency. This may partially explain why predictions close to the average sometimes yielded competitive Bray-Curtis scores, even when individual variability was not fully captured.

There was a noticeable difference in Bray-Curtis scores when the validation set was interpolated during cross-validation compared to using the original, non-interpolated set. This suggests that interpolating the validation data may contribute to overfitting. However, when interpolation was removed entirely, the results remained unchanged, indicating that the current interpolation approach does not effectively enhance the dataset as intended and requires improvement.

Performance varied based on data density. For periods with multiple samples, averaging across them produced better results, whereas for sparse time points, predictions based on individual baboons were more accurate. This highlights the importance of adapting prediction strategies to the data's temporal and individual-specific characteristics.

Finally, the application of Vector Autoregression (VAR) may not be ideal for modeling microbiome dynamics. The weak temporal relationships observed suggest that the nonlinear and complex nature of microbiome data requires alternative models better suited to capture these dynamics.

Future Work

Future research should focus on improving interpolation techniques to better represent temporal and individual-specific patterns without overfitting. Alternative temporal models, such as neural networks or ensemble approaches, could offer more flexibility for capturing nonlinear relationships. These refinements are necessary to address the current limitations and advance microbiome predictive modeling.

Appendix

GitHub repository with the code and predictions file here.