

SIMS: Social Inference of Microbiome Samples

By Tomer Oron & Yuval Dotan

i.d 322549312, 318505120

The research was performed in the School of Computer Science, Faculty of Exact Sciences

Under the supervision of Mr. Omri Peleg & Yadid Algavi

October 2024

Introduction

The gut microbiome plays a critical role in shaping the health, behavior, and evolutionary dynamics of animals {}. In social species, such as baboons, the composition of gut microbiota is influenced not only by individual factors like diet, genetics, and age, but also by social interactions and environmental conditions {Tung et al.}. Social inference—understanding microbiome composition through social dynamics—has emerged as a promising approach in recent studies of social species, revealing that individuals within the same group can share microbial communities {}. The Amboseli baboon population1, which has been the subject of long-term ecological and behavioral research, presents a promising model to further explore these influences.

In this study, we aim to predict the gut microbiome composition of individual baboons by using two key sources of data: the relative abundance profiles of microbial taxa from each baboon's gut and those of their social group. Using data on the Amboseli baboon population collected between 2000 and 2013 2, we assess how well the microbiome of an individual baboon can be predicted through a combination of their own microbiome samples and the shared microbial environment within their group.

Methods

Data exploration

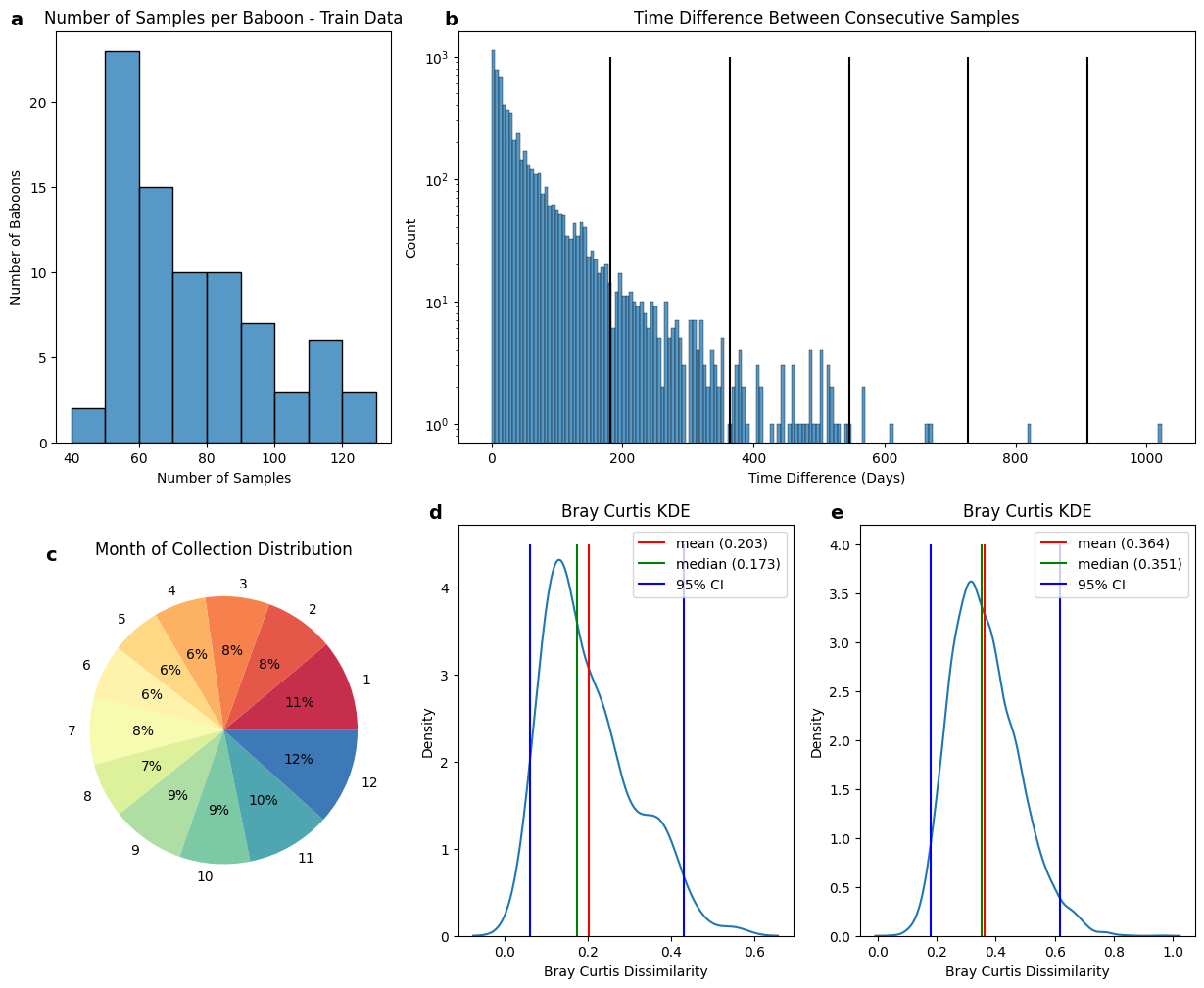
For this work, we utilized data collected from the Amboseli baboon population between 2000 and 2013. This dataset includes microbiome profiles of the gut microbiome, based on 16S rRNA gene sequencing3. The profiles were constructed using fecal samples collected opportunistically. Each sample is accompanied by metadata detailing the environmental conditions at the time of defecation and information about the baboon from which the sample was taken.

### add – general description of the training set

**The metadata** for each sample encompasses several essential elements. Temporal information includes the collection date, month, and hydrological year. At the individual level, the data specifies the baboon's unique identifier (*baboon\_id*), age at the time of collection, and sex. Social group data provides details on the baboon’s social group affiliation during sampling, the group's size, and diet composition, which is assessed using 30-day sliding windows and Principal Component Analysis (PCA)4. Environmental context is also included, covering the season (dry or wet) and the amount of rainfall during the collection month.

Analysis of the training data reveals that each baboon has an average of 76.2 samples **(Fig 1a)**. The median time difference between consecutive samples from the same baboon is 22 days, with 282 samples having a preceding sample collected more than 180 days earlier **(Fig 1b).** Although the wet season spans 42% of the year (November to May), 61% of the samples were collected during this period **(Fig 1c)**.

**For each sample**, there is a measurement of the abundance rates of the 61 most prevalent genera across all microbiome samples. Three analyses were conducted regarding samples from the same baboon, first, we compared the Bray-Curtis dissimilarity score between samples collected within the same date **(fig 1c)**, secondly, we calculated the Bray-Curtis dissimilarity score between a sample and the mean of the predeceasing samples **(fig 1d)**, and lastly, we calculated the score between every two concurrent samples **(fig 1e)**. These results suggest the predictive power of using the mean of previous samples, and using the last sample, while also providing us with the upper bound to the prediction accuracy one could expect.



**Figure 1. Analasys results of the train data**

**(a)**Histogram of number of samples collected per baboon in the train data. **(b)**Histogram of the time differences between the collection date of every 2 subsequent samples of the same baboon. **(c)**Pie chart of the distribution of samples' collection month. **(d)**Kernel Density Estimation (KDE) smoothed histogram of the Bray- Curtis dissimilarity score between every two samples collected on the same date from the same baboon. **(e)**KDE smoothed histogram of the Bray- Curtis dissimilarity score between sample and the mean of previous samples of the same baboon.

First Model

Building on our analysis, three key requirements were considered when constructing the model: 1. The model should use both the last sample and the mean of previous samples. 2. As the time difference between two consecutive samples increases, the influence of the last sample should decrease. 3. The model should present a seasonality effect.

Were is the sample to predict. The function represents the seasonal effect, and the exponent reduces the effect of the last sample as the time difference increases. and are matrices representing the effect of each genus of bacteria. These parameters are learned for each baboon individually. is a scalar, a global parameter representing desegregation constant through time effect.

Using the training set, we trained the model's parameters through iterative optimization using the L-BFGS-B{} method. For all the baboons the optimization resulted in and . Thus, we decided that our first requirement should be removed, and we suggested a new model based on the average of previous samples. Also, the results of optimizing led us to the conclusion that the effect of bacteria genus on other genera is negligible, a conclusion raised by other studies5 as well.

The Social Inference of Microbiome Samples (SIMS) Model

Based on the results of our initial model and studies showing that individuals within the same social group share similarities within their microbial composition, we refined the model to concentrate on circles of association: 1. The individual. 2. The individual’s immediate circle, its social group at the date of sampling. 3. The influence of broader circles, all the baboons which are not in its social group.

Were is the sample to predict at time t. is the average of samples taken within from , excluding those from this individual or from baboons in its social group. is the average of samples taken within from , representing baboons in the individual's social group. is the weighted average of the previous samples, where . Here, and are vectors of size 61, representing the social inference – the effect of the individual, its social group and the other baboons on its microbiome. is a scalar, a global parameter representing the decay in influence over time.

Pipeline

**The preprocessing stage** involves sorting the data by collection date, taking the mean of samples from the same baboon on the same date, and splitting the data into sub-datasets, one for each baboon.

The fitting stage includes a grid search for the initial , followed by iterative optimization of for each baboon and globally, using the L-BFGS-B optimization method. The optimization per baboon was conducted by using the mean of the Bray-Curtis dissimilarity function as a loss function, and the optimization of the parameter was conducted by minimizing the sum of this loss over all trained baboons.

Finally, the prediction is calculated, using configurable parameters. First, there is an option to choose between iterative and non-iterative modes. In the non-iterative mode, the model predicts each sample separately, whereas in the iterative mode, the model uses the baboon's previously predicted samples along with the data to predict the next one. Another configurable parameter is the threshold that specifies the minimum number of samples required to train for a new baboon. If this condition is not met, the model uses the average from the baboons already included in the model.

Validation

We validate the results using cross-validation (4-fold). The training set was split into 4 random groups of size 20. For each iteration, we trained a model containing 60 baboons and predicted the samples of the remaining 20 baboons both in iterative and non-iterative modes. The validation was conducted on three scenarios: 1. Using very-short time series (2 known samples per baboon) 2. Using short time series (10 known samples per baboon) and 3. Predicting 10 latest samples given all other samples.

Results

While fitting the model using the complete training set, the optimal was found to be . The vectors have great variance between the baboons (fig X), supporting studies that indicate the gut microbiome dynamics of baboons are highly individualized2. #### clustering? ## training score? ###

Discussion

1. Add clustering of means

Bibliography

1. Alberts, S. C. & Altmann, J. The Amboseli Baboon Research Project: 40 Years of Continuity and Change. in *Long-Term Field Studies of Primates* (eds. Kappeler, P. M. & Watts, D. P.) 261–287 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012). doi:10.1007/978-3-642-22514-7\_12.

2. Björk, J. R. *et al.* Synchrony and idiosyncrasy in the gut microbiome of wild baboons. *Nat. Ecol. Evol.* **6**, 955–964 (2022).

3. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).

4. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).

5. Roche, K. E. *et al.* Universal gut microbial relationships in the gut microbiome of wild baboons. *eLife* **12**, e83152 (2023).