The Raymond and
Beverly Sackler Faculty
of Exact Sciences
Tel Aviv University

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 & Mr. Yadid Algavi

November 2024

Introduction

The gut microbiome plays a major role in animal health and behavior[1–3], influenced by individual factors like diet and genetics, as well as social interactions and environmental conditions[4]. Recent studies on social species have shown that individuals in the same social group share similar microbial communities[5–7]. In this project, we use data from the Amboseli baboon population[8] to predict individual baboon microbiome composition by combining their microbial profiles with those of their social group, based on data collected between 2000 and 2013[9].

Methods

Data Exploration

In this project, we used data from 6096 fecal samples of 80 baboons for training and validation. For each sample, there is a measurement of the abundance rates of the 61 most prevalent genera across all microbiome samples based on 16S rRNA gene sequencing[10], accompanied by metadata regarding the environment and the baboon from whom the sample was collected.

The metadata includes the collection date, month, and hydrological year. At the individual level, data includes the baboon's ID, age at collection, and sex. Social group information covers the baboon's affiliation, the size of the social group, and the diet composition of the group (assessed with a 30-day sliding window and PCA[11]). Environmental data includes season (dry or wet) and monthly rainfall. Analysis showed each baboon has an average of 76.2 samples, with a median gap of 22 days between consecutive samples; 282 samples were collected over 180 days apart. Although the wet season (November–May) spans 42% of the year, 61% of samples were collected during this period.

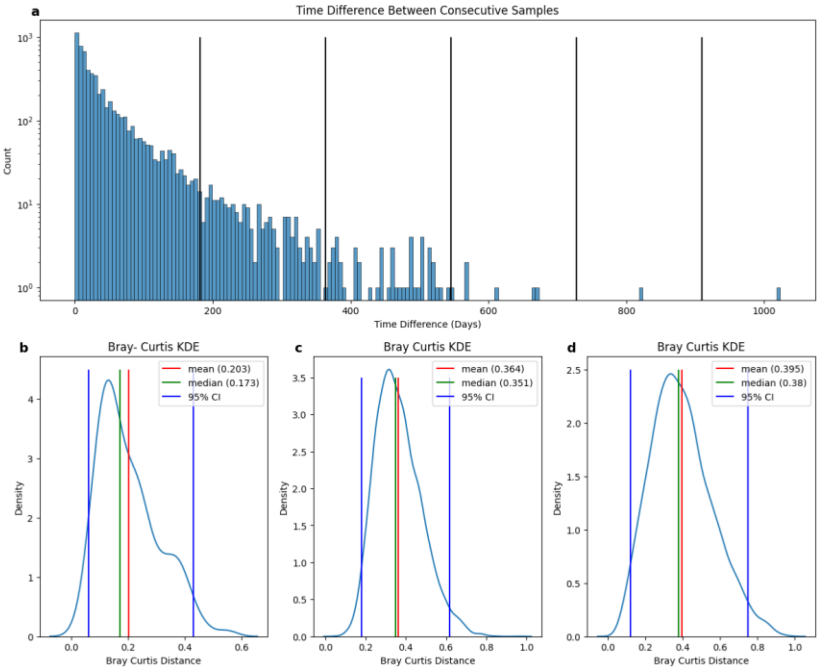

Figure1 . Analysis results of the train data

(a) Histogram of the time differences between the collection date of every two subsequent samples of the same baboon. **(b)** 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. **(c)** KDE smoothed histogram of the Bray- Curtis dissimilarity score between a sample and the mean of previous samples of the same baboon. **(d)** KDE smoothed histogram of the Bray- Curtis dissimilarity score between a sample and the previous sample of the same baboon.

We conducted three analyses regarding the predictive power of samples from the same baboon. First, we compared the Bray-Curtis dissimilarity score between samples collected on the same date (fig 1a). Secondly, we calculated the Bray-Curtis dissimilarity score between a sample and the mean of the predeceasing samples (fig 1b). Lastly, we calculated the score between every two concurrent samples (fig 1c). These results suggest the predictive power of using the mean of previous and last samples while providing us with the upper bound for the prediction accuracy one could expect.

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.

$$d_i = e^{-\lambda\Delta t}\cos\frac{2\pi\Delta t}{365}\ \alpha D_{i-1} + \left(1 - e^{-\lambda\Delta t}\cos\frac{2\pi\Delta t}{365}\right)\beta\overline{\{D_1,\dots,D_{\imath-2}\}}$$

Were d_i is the i^{th} sample to predict. The cosine function represents the seasonal effect, and the exponent reduces the effect of the last sample as the time difference increases. α and β are 61×61 matrices representing the effect of each genus of bacteria. These parameters are learned for each baboon individually. λ 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[12] method. For all the baboons the optimization resulted in $\alpha = 0_{61\times 61}$ and $\beta = I$. 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 β led us to the conclusion that the effect of bacteria genus on other genera is negligible, a conclusion raised by other studies[13] 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.

$$d_t = \alpha D_{O_{\Delta t_1}} + (1 - \alpha - \beta) D_{S_{\Delta t_2}} + \beta D_I$$

Were d_t is the sample to predict at time t. $D_{O_{\Delta t_1}}$ is the average of samples taken within Δt_1 from d_t, excluding those from this individual or baboons in its social group. $D_{S_{\Delta t_2}}$ is the average of samples taken within Δt_2 from d_t, representing baboons in the individual's social group. D_I is the weighted average of the previous samples, where $w_j = \frac{1}{(\Delta t_j)^{\gamma} + 10^{-10}}\left(0.5 + 0.5\cos\left(\frac{2\pi\Delta t_j}{365}\right)\right)$. Here, α and β are vectors of size 61, representing the social inference – the effect of the individual, its social group, and the other baboons on its microbiome. γ is a scalar, a global parameter representing the decay in influence over time.

Pipeline

In the preprocessing stage, the data is sorted by collection date, and the mean of samples from the same baboon on the same date is calculated. Subsequently, the data is split into sub-datasets, each corresponding to a specific baboon.

During the fitting stage, an initial grid search is conducted, followed by iterative optimization of β for each baboon and γ globally, utilizing the L-BFGS-B optimization method. The optimization process for each baboon involves using the mean of the Bray-Curtis dissimilarity function as a loss function. At the same time, the parameter γ is optimized by minimizing the sum of this loss across all trained baboons.

The prediction is calculated using configurable parameters. First, there is an option to choose between iterative and non-iterative prediction. In non-iterative prediction, the model predicts only based on the known samples, whereas in the iterative mode, the model also uses previously predicted samples. Another configurable parameter is the threshold, which specifies the minimum number of samples required to train beta 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 validated the results using cross-validation (4-fold). The training set was split into four random groups of size 20. We trained a model containing 60 baboons for each iteration and predicted the samples of the remaining 20 baboons 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 ten latest samples given all other samples.

Code

All code and analysis were written in Python, with NumPy, Matplotlib, SciPy, and Pandas. The source code is available at https://github.com/yuvaldotan/workshop_microbiome.

Results

While fitting the model using the complete training set, the optimal γ was found to be 0.957. The β vectors have great variance between the baboons (fig 2), supporting studies that indicate the gut microbiome dynamics of baboons are highly individualized[9].

In scenario 1, the mean BC using an iterative mode was noticeably better than the non-iterative mode, whereas in scenario 3, the non-iterative prediction was preferable (fig 3). For scenario 2, we also checked the usage of β trained per baboon in comparison to using the average β of the other 60 baboons. The results showed for both cases that the iterative mode is slightly better for prediction and that using ten samples to train β is not enough.

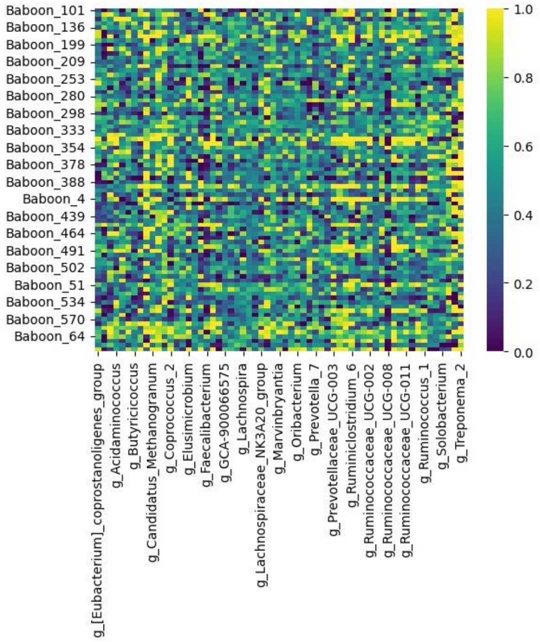

Figure 2. Heatmap of the learned β vectors.

Each row represents a baboon (y-axis), and each column corresponds to a microbial genus (x-axis). The color represents the value given to each entry, where blue means a higher weight given to the social data and yellow indicates a higher weight to previous samples of the same baboon.

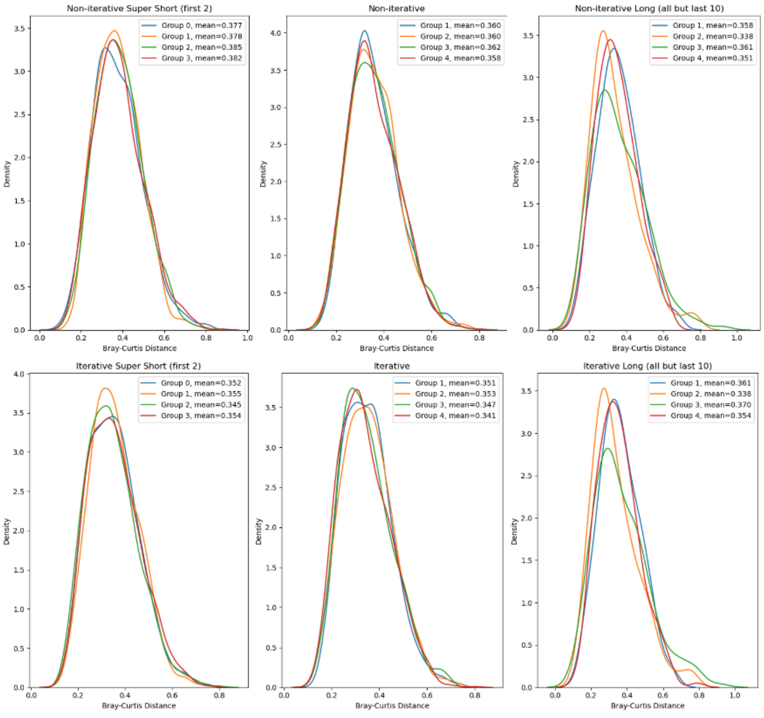

Figure 3. Cross-validation results

Kernel Density Estimation (KDE) plots of Bray-Curtis dissimilarity scores across different scenarios (short and long) and modes (non-iterative and iterative). Each line represents a test set of 20 baboons, with mean distances shown in the legend. The left column shows results for prediction based on the first 2 samples, the middle for prediction based on the first 10 samples, and the right for prediction of the last 10 samples based on all others.

Discussion

During the development of our model, we questioned how to handle the temporal aspect of the data, specifically the seasonal effect and distance time gaps between sampling. Throughout the model construction, we began with a simple model assigning equal weights to samples, then added seasonal effect, and finally considered the time-decay factors. Using $e^{-\gamma\Delta_t}$ made the weights too small (effectively zero) due to the time differences between samples. Therefore, we decided to replace it with a function presenting slower decay, $\Delta_t^{-\gamma}$. In every model update, we have seen an increase in the prediction's accuracy. Future work could explore other functions for seasonality and time decay to capture temporal trends better.

Another question we faced was whether using Newton-Raphson optimization algorithms was the correct method to find the best parameter. The Bray—Curtis dissimilarity function is defined using the L1 norm and, therefore, is not differentiable at every point. Throughout our work, we considered other optimization algorithms that do not require the function to be differentiable. Still, they did not present better parameters than the ones obtained through the L-BFGS-B algorithm and took more time to run.

While analyzing the data, we applied the K-means algorithm to cluster baboons based on their average microbiome samples. This resulted in five distinct clusters (Fig. 4) with an inertia of 0.1997. We observed no correlation between these clusters and the baboons' social groups, suggesting that these clusters may instead reflect other forms of interaction among baboons, which may improve the model. To do so, more metadata is required, such as family relations between baboons and more detailed data about each baboon's participation in social activities, such as grooming. Another method of finding relations between baboons is by clustering the trained betas of the baboons. The betas may also be used to uncover insights about correlations and interactions between bacteria.

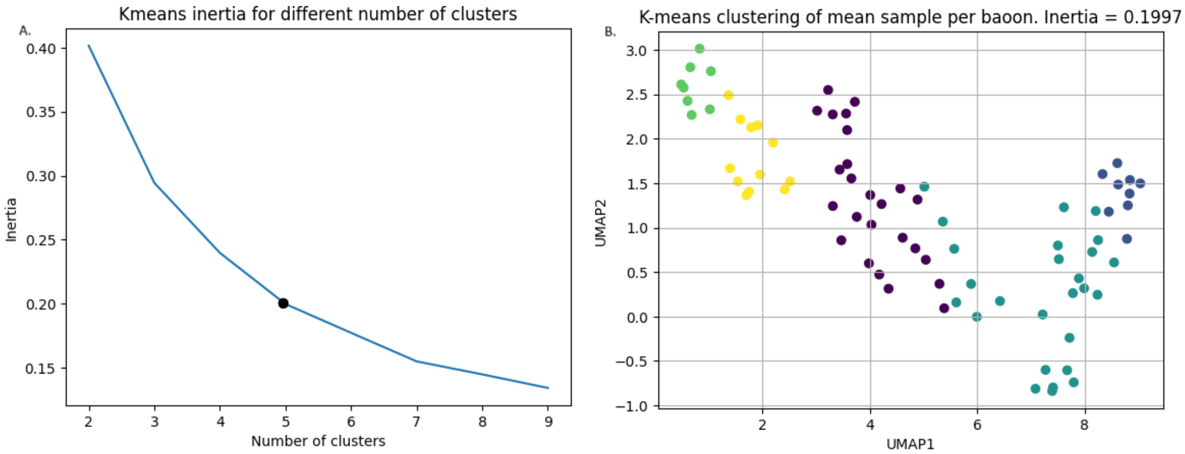

Figure 4. Results of clustering the average sample per baboon using K-means.

(A) Presents elbow method analysis for finding the best number of clusters. (B) UMAP projection of clusters found for K=5.

Bibliography

1. Iacob, S., Iacob, D. G. & Luminos, L. M. Intestinal Microbiota as a Host Defense Mechanism to Infectious Threats. *Front. Microbiol.* **9**, 3328 (2019).
2. Appleton, J. The Gut-Brain Axis: Influence of Microbiota on Mood and Mental Health. *Integr. Med. Clin. J.* **17**, 28–32 (2018).
3. Davenport, E. R. *et al.* The human microbiome in evolution. *BMC Biol.* **15**, 127 (2017).
4. Debray, R., Tung, J. & Archie, E. A. Ecology and Evolution of the Social Microbiome. *Annu. Rev. Ecol. Evol. Syst.* (2024) doi:10.1146/annurev-ecolsys-102622-030749.
5. Tung, J. *et al.* Social networks predict gut microbiome composition in wild baboons. *eLife* **4**, e05224.
6. Raulo, A. *et al.* Social networks strongly predict the gut microbiota of wild mice. *ISME J.* **15**, 2601–2613 (2021).
7. Wikberg, E. C., Christie, D., Sicotte, P. & Ting, N. Interactions between social groups of colobus monkeys (*Colobus vellerosus*) explain similarities in their gut microbiomes. *Anim. Behav.* **163**, 17–31 (2020).
8. 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.
9. Björk, J. R. *et al.* Synchrony and idiosyncrasy in the gut microbiome of wild baboons. *Nat. Ecol. Evol.* **6**, 955–964 (2022).
10. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
11. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).
12. Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization.

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