Modeling Microbiota: Using LSTM to Study Vaginal Dynamics

Anonymous submission

Abstract

The complex microbe-microbe interactions within the human body are crucial in determining human health. However, current models struggle to capture community microbial interactions. Existing "bottom-up" models struggle due to the presence of both pairwise and higher order interactions. Additionally, generating a model based on external observation is difficult because of the large number of unknowns. However, natural language processing techniques such as the Long Short Term Memory (LSTM), have shown promise for learning complex context and time dependent functions. Furthermore, utilizing AI Explainability methods such as SHapley Additive exPlanations (SHAP) values, insight can be gained into complex community microbial interactions. Thus, in this paper we combine the power of a LSTM and SHAP to create a tool for generating insight into complex community microbial interactions.

Introduction

Microbes have a wide ranging number of essential roles in nearly all environments, including the human body, strongly correlated with general health and disease (de Vos and de Vos 2012). The role each microbe plays in an environment is shaped by numerous factors that can vary based on context and time. However, gaining strong insight has proven difficult due to the complexity of the potential interactions within a microbial community. For example, past work has found considering only pairwise interactions neglects highly informative higher order interactions (Sanchez-Gorostiaga et al. 2019).

However, deep learning has demonstrated the ability to learn complex functions in order to solve problems. Similar to microbial interactions, natural language processing is a problem where the the function of specific words can change based on context and order. Thus, natural language processing techniques are a natural avenue to investigate microbial community interactions. Natural language processing techniques such as the LSTM have already demonstrated significant success for modeling interactions within a microbial community (Baranwal et al. 2022). Also, using Explainability AI methods such as SHAP value calculator algorithms (Lundberg and Lee 2017) and attention can provide valuable insight into the relationship between the inputs and the outputs. Thus, this work delivers a tool to gain helpful, interpretable insight into microbial interactions from any time longitudinal dataset.

Case Study

In order to test the tool, an LSTM was fit onto a set of time longitudinal measures of relative abundances of vaginal microbiota. It was obtained from a clinical trial (Song et al. 2020) that studied microbiota fluctuations in relation to the hormonal cycle. We parsed the data by first removing any irregularities (duplications or long blanks in time), and then averaging in any timesteps where the preceding and following timestep were not missing. After that, we selected training data by choosing all contiguous sequences of appropriate length. Before input into the model, the data is then rescaled from 0 to 1 with the MinMaxScaler from scikit-learn. To maximize the amount of training data, only a single test trajectory was completely held out from training to evalu-

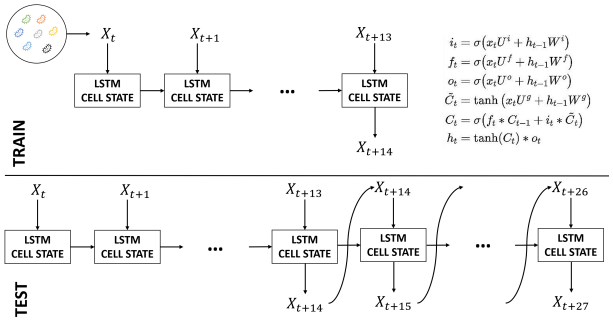

Figure 1: The procedure utilized to train and test the LSTM (above and below the line, respectively). Different time steps of normalized microbial relative abundances are utilized as input for training to produce a single output predicting the relative abundances at time step $t + 14$. The testing scheme utilizes a similar initial setup with normalized inputs for time steps t through $t + 13$, however, after outputting time step $t+14$; time step $t+14$ becomes an input, and the inputs $t+1$ through $t + 14$ are utilized to predict time step $t + 15$, and then $t+2$ through $t+15$ are utilized to predict the next time step. This continues until the final time step $t+27$, resulting in 14 time steps of predicted relative abundances ($t + 14$ to $t + 27$).

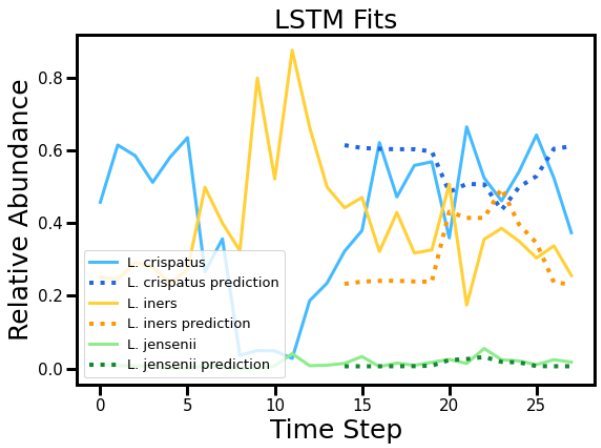

Figure 2: The LSTM's test trajectory performance. The solid lines are the true data, while the dotted lines are the LSTM's predictions. As explained, the LSTM starts outputting predicted values from the 15th to the 28th time steps with the first 14 time steps as the initial input data. This plot uses normalized values and only demonstrates 3 of the 15 species.

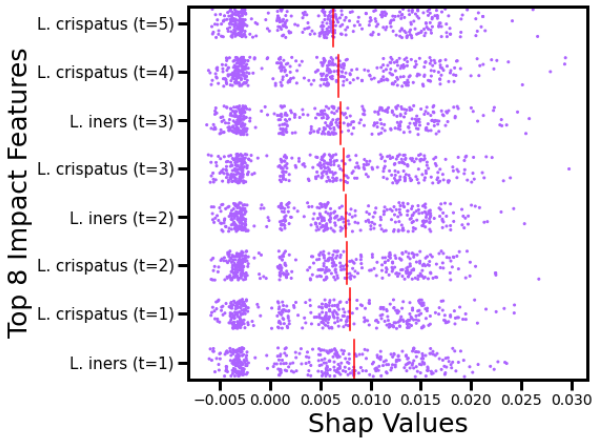

Figure 3: Plot of the explanation for the model's prediction of *L. crispatus*. This plot demonstrates the top 8 highest distinct input features, meaning distinct time step and species, by mean absolute value impacting the predictions of the LSTM for *L. crispatus* across the entire training dataset. The scatter plot has the SHAP values on the x-axis demonstrating the impact of the particular feature while the feature is plotted on the y-axis with jitter to help visualize point density. Furthermore, each feature has a vertical red line symbolizing the mean absolute SHAP value.

ate the model resulting in 413 total training instances. The model was built using PyTorch, and consisted of a 3-layer unidirectional LSTM feeding into a linear layer, and then a sigmoid activation resulting in 57,084 trainable parameters. The LSTM's hyperparameters were fitted using random hyperparameter search and model selection was based on performance on the hold out test case. The testing procedure consisted of feeding 14 timesteps as inputs to the LSTM and having it output the 15th timestep in that sequence (as shown in Figure 1). For the testing scheme (illustrated in Figure 1), the LSTM would initially begin with 14 time steps as input and continually utilize the most recent 14 time steps to predict the next time step one at a time until 14 time steps were output. Despite the relatively low volume of data compared to the number of trainable parameters, Figure 2 demonstrates the ability of the LSTM to learn complex relations.

Results

In order to gain insight into microbial community interactions, we built a tool to address this need and tested it on a time longitudinal vaginal microbial dataset. In order to examine the learned relations between the inputs and outputs from the LSTM, the DeepSHAP algorithm was utilized to generate the SHAP values (Lundberg and Lee 2017). Figure 3 demonstrates some of the potential insights that can be gained utilizing this method. Figure 3 demonstrates the importance of the early time steps for the final prediction along with the importance of 2 major species in determining the final value of *L. crispatus*, *L. crispatus* and *L. iners*. This result could signify a type of direct interaction between the two species. A similar process can be carried out to observe the influential inputs on each output feature (i.e. each species in our microbial system). This tool can direct future researchers to uncover important microbial interactions in fields such as healthcare improving health outcomes.

References

Baranwal, M.; Clark, R. L.; Thompson, J.; Sun, Z.; Hero, A. O.; and Venturelli, O. S. 2022. Recurrent neural networks enable design of multifunctional synthetic human gut microbiome dynamics. *eLife*, 11: e73870.

de Vos, W. M.; and de Vos, E. A. 2012. Role of the intestinal microbiome in health and disease: from correlation to causation. *Nutrition reviews*, 70(suppl_1): S45–S56.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 4765–4774. Curran Associates, Inc.

Sanchez-Gorostiaga, A.; Bajić, D.; Osborne, M. L.; Poyatos, J. F.; and Sanchez, A. 2019. High-order interactions distort the functional landscape of microbial consortia. *PLOS Biology*, 17(12): 1–34.

Song, S. D.; Acharya, K. D.; Zhu, J. E.; Deveney, C. M.; Walther-Antonio, M. R. S.; Tetel, M. J.; and Chia, N. 2020. Daily Vaginal Microbiota Fluctuations Associated with Natural Hormonal Cycle, Contraceptives, Diet, and Exercise. *mSphere*, 5(4): e00593–20.

Acknowledgments

YL is thankful to Kia Khezeli for teaching and mentorship.