Topic modeling of soil microbiomes in droughted corn

Anastasiia Kim,[1] **Sanna Sevanto,** [2] **Nicholas Lubbers** [1]

[1] Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory
[2] Earth and Environmental Sciences Division, Los Alamos National Laboratory
akim@lanl.gov, sanna@lanl.gov, nlubbers@lanl.gov

Abstract

Plant-microbiome optimization is a potential solution to improve plant stress tolerance under water shortages. Using early data from a directed evolution experiment, our goal is to find beneficial microbiome compositions in the soil and other environmental factors affecting plant drought tolerance. In this work, we uncover the microbiome communities associated with drought using a Latent Dirichlet Allocation method with Gibbs sampling. We reveal drought-enriched microbiome at the phylum and class taxonomic levels that potentially can help iteratively guide the microbiome development process.

Introduction

The interaction of various factors such as microbiome communities, soil chemistry, plant traits, and plant chemistry plays a crucial role in plant ability to withstand limited watering conditions. Microbiome composition plays a vital role in plant functioning and development and can potentially improve the performance of biological systems (Compant S 2010; Farrar K 2014; Mendes R 2013). Several recent studies on cotton, rice, and peanut root microbiomes have revealed certain phyla and classes that are enriched in water-limited soil (Dai et al. 2019; Naylor and Coleman-Derr 2018; Ochoa-Hueso et al. 2018; Santos-Medellín et al. 2017; Ullah et al. 2019). Understanding how soil microbiomes influence plants under drought is a challenging new area of research. Nonetheless, unraveling complex interactions between plants and their microbiome could yield knowledge usable to better predict the behavior of real-world crops, and perhaps support them through many challenges in food security.

The dataset for this work comes from an ongoing greenhouse experiment applying directed evolution (Cobb, Chao, and Zhao 2013) in an attempt to iteratively optimize microbiome compositions to improve corn health in droughted conditions. In each generation, plants with desired functional traits, i.e., stomatal closure point and water use efficiency, are selected and their soil microbiome is used to grow a new generation. We hope that artificial selection on these traits will lead to better plant performance under drought and help to find stable microbiome communities that help plants to withstand water-limited conditions. At the present stage, we present results from data on 119 microbiome-plant systems from two non-directed generations (called generation 0 and 1), between which no artificial selection was applied. They were performed in order to understand the characteristics of the systems for experimental design purposes. The seed for each pot is randomly drawn from the stock, which belongs to an experimental strain of corn, USDA seed bank "B73", for which the fully sequenced genome is available. In this experiment, the corn is grown in sterile fritted clay in individual pots that were inoculated with microbiomes from the soil source using the bulk soil microbiome from prior sources. The initial microbiomes were collected from the vicinity of Los Alamos (forest soil type), NM, Fort Collins (agricultural soil type), CO, as well as from several plants grown without any inoculated microbiome. Plants from generation 1 are microbial descendants of those from generation 0, because they have been grown in soil inoculated with the microbiome from individual plants at the end of generation 0. In each generation, 64 corn plants were planted, and half of the plants were watered well (up to 65% volumetric water content 3 times a week) and the other half were not well-watered (up to 45% volumetric water content 3 times a week). After growing plants for 10 weeks, they were droughted completely. Preterminal drought measurements of the plant traits (except for drought time and stomatal closure point that were measured post-drought), soil chemical composition, and microbiome sequencing were carried out. Here we focus on the analysis of the microbiome communities, with a focus on the links between the microbiome and the watering conditions, the initial microbiome soil source, and the generation of the experiment.

Method

To address this task, we use Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), which is a probabilistic generative model developed for language modeling of a corpus – a set of documents. Each document is represented by the count of the words present in the document. The key assumption behind LDA is that documents are represented as probabilistic mixtures over latent topics, where

each topic is characterized by a distribution over words. This factorization of the data is a method to associate words with each other, and, at the same time, perform dimensionality reduction on the documents. To learn the model, we used Gibbs sampling (Griffiths 2002). Topic modeling with LDA, widely applied for text mining, has been successfully applied in a few biological studies to identify human gut, oral, and skin microbial communities (Higashi et al. 2018; Raman et al. 2019). However, less is known about how LDA performs on the soil microbiome, which is primarily analyzed using traditional statistical methods such as dimensional reduction methods and tests of significance (Jochum et al. 2019; Marasco et al. 2012; Naylor et al. 2017; Ullah et al. 2019; Zolla et al. 2013). The data itself is high-dimensional and sparse with a certain amount of unidentified taxonomic levels. We use the MALLET software (McCallum 2002) to identify topics at different taxonomic levels. This implementation of the LDA model is fast. We did not experience computational issues when running the algorithm for our data set.

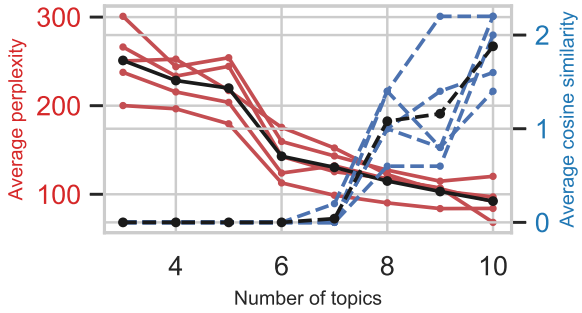

Figure 1: Perplexity score for cross-validation folds and pairwise cosine similarity between taxa in topics as functions of number of topics applied at the phylum level. Averaged curves over folds are shown in black.

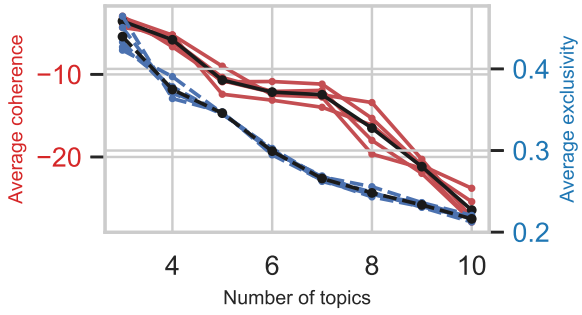

Figure 2: Coherence and exclusivity of the most abundant bacteria in the learned topics as functions of number of topics applied at the phylum level. Averaged curves over folds are shown in black.

Our goal is two-fold: one is to see if the LDA topics at different taxonomic levels have links to the treatment conditions of the pots – namely, the soil source, the watering treatment (half vs full watered), and the generation of plant.

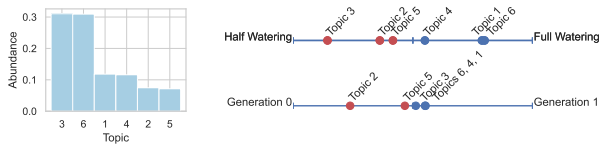

Figure 3: Left: Overall abundance of the topics across all pots. Right: Weighting of topic abundances at the phylum level in terms of watering treatment and generation.

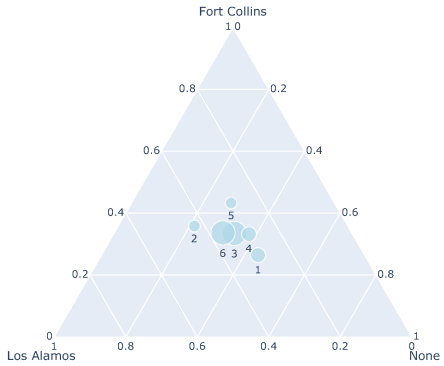

Figure 4: Topic abundance weighting at the phylum level for the different soil sources. Size indicates overall topic abundance.

Secondly, in view of these patterns, we seek to determine which species and collections of species contribute to the topics that can be associated with these treatment conditions. Thus, LDA allows a view into the most important aspects of the data, which contains more than 3000 genetic sequences from 31 phyla and 71 classes, as well as sequences for which no taxonomic classification can be assigned.

Before we start to discuss our LDA results, let us introduce the following analogy between text and microbiome analysis we used: the pot samples, bacterial species (taxa), microbiome communities are viewed as the documents, words, and topics, respectively. That is to say that individual plant microbiomes (documents) are broken down into a distribution of topics, and these topics are distributions of taxa (words). Due to the differences between text data and microbiome data, some pre-processing approaches and evaluation metrics used for text data are not appropriate, such as the removal of common words which are assumed not to be useful in describing the document (stop words). We preprocess data by aggregating basic DNA fragments from the microbiome to identified taxonomic descriptions associated with the DNA (phylum, class, ...), and handle incomplete taxonomic specification for some sequences by aggregating to the highest available taxonomic level.

To choose the appropriate number of topics, we run 5-fold cross-validation, measure several metrics such as perplexity, pairwise cosine similarity, coherence, and exclusivity to help us decide how many topics is appropriate to use, and calcu-

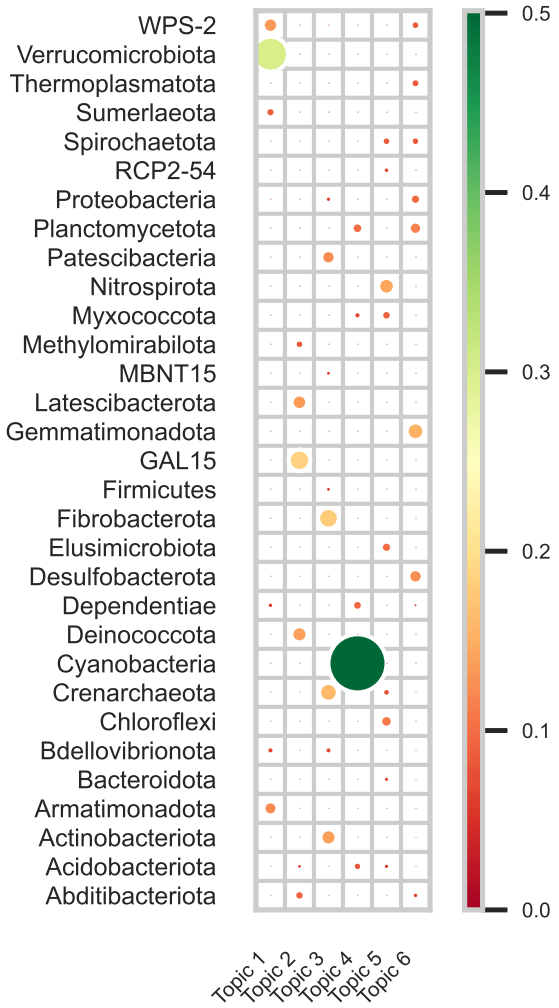

Figure 5: Relative phyla amplification for each topic, showing only the dominant phyla.

late averaged results as a function of the number of topics. A perplexity was calculated to see how well a model performs on the unseen held-out test data. A pairwise cosine similarity was calculated between words (taxa) in topics to determine how distinct the distribution of taxa is between topics. We use coherence (Stevens et al. 2012) that measures whether the most abundant taxa in a topic tend to co-occur together in other topics. The exclusivity metric (Bischof and Airoldi 2012; McCallum 2002) finds the most unique taxa for each topic. Figure 1 shows the perplexity score calculated on the test data and pairwise cosine similarity between bacteria in topics. Figure 2 shows topic coherence and exclusivity of the most abundant bacteria in the learned topics. There is no right answer when determining the number of topics, too many topics will result in indistinguishable topics, in contrast, using too few topics may hurt the explanatory power of the model. Therefore, we choose the appropriate number of topics based on described metrics behavior when they tend to be stable gradually after reaching the optimal level. Based on these metrics, we decided to use 6 topics at the phylum level, as we did not observe significant improvement for larger topic counts. After fixing the number of topics, we ran LDA once again on the full data set.

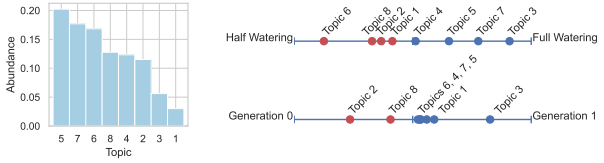

Figure 6: Left: Overall abundance of the topics across all pots. Right: Weighting of topic abundances at the class level in terms of watering treatment and generation.

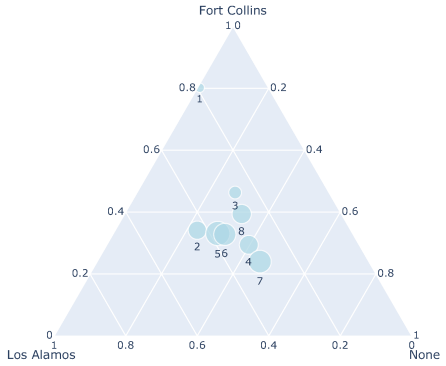

Figure 7: Topic abundance weighting at the class level for the different soil types. Size indicates overall topic abundance.

Results

Figure 3 shows the overall abundance of each topic across all pots and the weighting of each topic's abundance towards watering treatment type and generation on a 0-1 scale at phylum taxonomic level. Figure 4 shows the weighting of topic abundances for the different soil types: Los Alamos, Fort Collins, and None (no inoculated microbiome). Topic 3 is more abundant in the half-watered pots, whereas topics 1 and 6 are more abundant in the full-watered pots. We don't observe that any topic is significantly skewed towards a particular soil type at the phylum level. On average, our soil microbiome in pot predominantly consisted of the phyla *Proteobacteria* (44.15%), *Bacteroidota* (13.67 %), *Actinobacteriota* (13.08 %), *Verrucomicrobiota* (9.05 %), and *Cyanobacteria* (8.01 %). Figure 5 shows the *relative amplification* of each taxa within each topic. We define this as the probability of the phylum given the topic divided by the frequency of that phylum in the overall dataset, and then normalized for each topic across all taxa; the sum of the relative amplification for all taxa in each topic is, by definition, one. In other words, the relative amplification shows

which taxa's abundance is most amplified in that topic in comparison to their abundance in the overall dataset. We only displayed a subset of phyla whose relative amplification is greater than 0.05. We remark on the connections between fig. 3 and fig. 5. For example, topic 1 is related to full-watering treatment, and the most amplified phylum is *Verrucomicrobiota*. Similarly, topic 3 is related to half-watering, and one of the most amplified phyla in that topic is *Actinobacteriota*. Previous studies have shown these two bacteria are associated with these watering conditions in different host plants (Naylor and Coleman-Derr 2018).

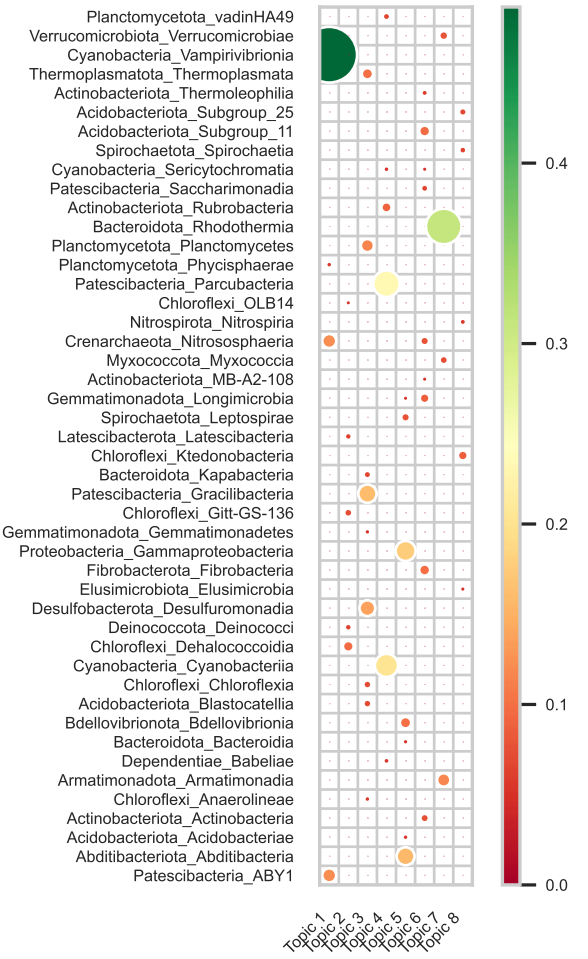

Figure 8: Relative amplification for species within each topic at the class level. The y-axis labels display the phylum name followed by the class name.

We performed a similar analysis for the class level, and 8 topics were found to balance the metrics used for topic selection. Figures 6, 7, 8 show analogous plots to those at the phylum level, but at the class level. The results show that the class-level topics are more strongly associated with the experimental conditions than the phylum-level topics. Topics 3, 5, and 7 are more abundant in the full-watered pots whereas topics 6 and 8 are more abundant in the half-watered pots. Figure 7 demonstrates that most topics are not very strongly associated with a particular soil source, with the notable exception of topic 1, which is highly abundant in the pots with Fort Collins soil microbiome inoculation, and not present in the None soil source. This is especially interesting since neither watering nor generation have a strong relationship with this topic. On average, our soil microbiome in pot predominantly consisted of the classes *Gammaproteobacteria* (25.65%), *Alphaproteobacteria* (18.73 %), *Bacteroidia* (13.52 %), *Actinobacteria* (12.41 %), *Verrucomicrobiae* (8.83 %), and *Cyanobacteriia* (7.21 %). In figure 5 we observed that *Actinobacteriota, Crenarchaeota, Fibrobacterota, Patescibacteria* phyla can be associated with water-limited soil. Figure 8 shows that all classes except *Rubrobacteria* (*Actinobacteriota*) that belong to *Actinobacteriota, Crenarchaeota, Fibrobacterota* phyla are also more amplified in the half-watered pots. However, classes from the phylum *Patescibacteria* ended up in many topics, i.e., *ABY1* and *Saccharimonadia* classes appear in the half-watered associated topics 1 or 6, whereas *Parcubacteria* and *Gracilibacteria* classes show up in the topics 4 and 3, respectively. We also note that topic 3 is strongly associated with the second generation (generation 1) of plants – where none of the phylum-level topics could be strongly ascribed to generation 1.

Conclusion

We have explored how data-driven LDA and taxonomic classification can be combined to explore the content of a large, complex data set of microbiome samples. By running LDA at different taxonomic levels we can easily detect which taxa are associated with different experimental conditions, such as water-limited soil and soil source type. Because LDA seeks a compact, unsupervised representation of the data, this is far simpler than exploring each taxon individually, and reveals associations between the taxa. Our analysis gives some results similar to previous studies of plant microbiomes, and gives rise to a host of hypotheses which might be tested in more targeted, small-scale experiments. Moving forward, we want to connect the expression of certain plant functional traits to the topic distributions for a better understanding of the plant-microbiome interaction. In this way, we may study which microbial communities may potentially improve plant performance under drought.

Acknowledgements

We thank all project team members. In particular, Abigael Nachtsheim, Christine Anderson-Cook, Michaeline Albright, Chistina Steadman, Turin Dickman, and Brent Newman for experimental design. George Perkins, Jack Heneghan, Kelsey Carter, and Dea Musa for measurements of plant traits and maintenance of the experiment. Eric Moore for analyzing microbiome. John Dunbar for oversight and guidance. We thank LDRD for funding (20200109DR). We thank USDA and US National Plant Germplasm System for providing the plant seeds for our project.

References

Bischof, J.; and Airoldi, E. M. 2012. Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 201–208.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3: 993–1022.

Cobb, R. E.; Chao, R.; and Zhao, H. 2013. Directed evolution: past, present, and future. *AIChE Journal* 59(5): 1432–1440.

Compant S, van der Heijden MG, S. A. 2010. Climate change effects on beneficial plant-microorganism interactions. *FEMS Microbiol Ecol.* 73(2). doi:doi:10.1111/j.1574-6941.2010.00900.x.

Dai, L.; Zhang, G.; Yu, Z.; Ding, H.; Xu, Y.; and Zhang, Z. 2019. Effect of drought stress and developmental stages on microbial community structure and diversity in peanut rhizosphere soil. *International journal of molecular sciences* 20(9): 2265.

Farrar K, Bryant D, C.-S. N. 2014. Understanding and engineering beneficial plant-microbe interactions: plant growth promotion in energy crops. *Plant Biotechnol J.* 12(9). doi: doi:10.1111/pbi.12279.

Griffiths, T. 2002. Gibbs sampling in the generative model of latent dirichlet allocation .

Higashi, K.; Suzuki, S.; Kurosawa, S.; Mori, H.; and Kurokawa, K. 2018. Latent environment allocation of microbial community data. *PLoS computational biology* 14(6): e1006143.

Jochum, M. D.; McWilliams, K. L.; Pierson, E. A.; and Jo, Y.-K. 2019. Host-mediated microbiome engineering (HMME) of drought tolerance in the wheat rhizosphere. *Plos one* 14(12): e0225933.

Marasco, R.; Rolli, E.; Ettoumi, B.; Vigani, G.; Mapelli, F.; Borin, S.; Abou-Hadid, A. F.; El-Behairy, U. A.; Sorlini, C.; Cherif, A.; et al. 2012. A drought resistance-promoting microbiome is selected by root system under desert farming. *PloS one* 7(10): e48479.

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit. URL http://mallet.cs.umass.edu.

Mendes R, Garbeva P, R. J. 2013. The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiol Rev.* 37(5). doi:doi:10.1111/1574-6976.12028.

Naylor, D.; and Coleman-Derr, D. 2018. Drought stress and root-associated bacterial communities. *Frontiers in plant science* 8: 2223.

Naylor, D.; DeGraaf, S.; Purdom, E.; and Coleman-Derr, D. 2017. Drought and host selection influence bacterial community dynamics in the grass root microbiome. *The ISME journal* 11(12): 2691–2704.

Ochoa-Hueso, R.; Collins, S. L.; Delgado-Baquerizo, M.; Hamonts, K.; Pockman, W. T.; Sinsabaugh, R. L.; Smith, M. D.; Knapp, A. K.; and Power, S. A. 2018. Drought consistently alters the composition of soil fungal and bacterial communities in grasslands from two continents. *Global Change Biology* 24(7): 2818–2827.

Raman, A. S.; Gehrig, J. L.; Venkatesh, S.; Chang, H.-W.; Hibberd, M. C.; Subramanian, S.; Kang, G.; Bessong, P. O.; Lima, A. A.; Kosek, M. N.; et al. 2019. A sparse covarying unit that describes healthy and impaired human gut microbiota development. *Science* 365(6449).

Santos-Medellín, C.; Edwards, J.; Liechty, Z.; Nguyen, B.; and Sundaresan, V. 2017. Drought stress results in a compartment-specific restructuring of the rice root-associated microbiomes. *MBio* 8(4): e00764–17.

Stevens, K.; Kegelmeyer, P.; Andrzejewski, D.; and Buttler, D. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 952–961.

Ullah, A.; Akbar, A.; Luo, Q.; Khan, A. H.; Manghwar, H.; Shaban, M.; and Yang, X. 2019. Microbiome diversity in cotton rhizosphere under normal and drought conditions. *Microbial ecology* 77(2): 429–439.

Zolla, G.; Badri, D. V.; Bakker, M. G.; Manter, D. K.; and Vivanco, J. M. 2013. Soil microbiomes vary in their ability to confer drought tolerance to Arabidopsis. *Applied soil ecology* 68: 1–9.