Predicting rice plant age from soil microbiome data

Aleksa Krsmanović[1,2], Sanja Brdar[1], Jelena Slivka[2]

Background

A plant, during its life cycle, affects the chemical and microbial qualities of the soil it grows in, which in return, affect further growth and development of the plant[1,2]. Even though the importance of microbial communities in soil has been known since as early as the mid-twentieth century[3], the exact relationships between different microbes and plant species still remain unknown[4,5]. Such knowledge would enable the creation of perfect microbial conditions, which, in return, could cause preferable plant growth, an increase of yield[6] and enhancement of drought resistance[7]. Interpretable machine learning models could provide valuable insight into these correlations. However, performance of different machine learning models has not been compared on these problems yet, nor have the best practices been defined. We believe that interpretation of the best performing machine learning model provides the most accurate information about these correlations.

Methods

In this study, we analysed bacterial and archeal sequences from the rhizosphere root compartment of rice plants (*Oryza sativa*), whose life cycle lasts for 150 days on average. We evaluated different machine learning pipelines for predicting plant age from the relative abundances of microbes.

We used the work of *Edwards et al.*[7] who collected 495 soil samples across three consecutive seasons in two geographical locations from six different rice cultivars. They performed 16S pair-ended sequencing (V4 region), resulting in reads with a length of 2 x 250 bp, using Illumina MiSeq. This data was made publicly available in .fastq format.

In our study, we analyzed the sequences from the mentioned research[7] using QIIME2 software[8]. Firstly we performed the trimming of the sequences to retain a Phred score above, or equal to 30. We used DADA2 package to perform denoising and joining of the reads [9]. From the generated ASV table, we removed the amplicon sequence variants appearing less than 20 times across samples and samples with a total count of sequences less than 1500. We then performed closed-reference OTU picking using Silva database[10] and QIIME2's Naive Bayes taxonomic classifier with a similarity threshold of 99%. Finally, we filtered out mitochondria and chloroplasts, as well as any OTU that has not been classified to at least a phylum level. The resulted OTU table was observed at five taxonomic levels (phylum, class, order, family, genus) to determine which level predicts plant age the best.

Instead of rarefying, microbe counts were transformed into compositional data. To map the data from the simplex into the Euclidean space, we used Aitchison's centralized log-ratio transformation with the addition of pseudo-values[11]. To predict rice plant age we applied 11 machine learning pipelines on all five taxonomic levels[12,13,14,15]:

1. Random Forest with recursive feature elimination based on RF feature importance (**RF+RF IMP**)
2. Random Forest with univariate feature selection based on F-value (**RF+F-VAL**)

1 Biosense Institute
2 Faculty of Technical Sciences, University of Novi Sad

3. Random Forest with univariate feature selection based on Spearman's correlation (**RF+SPRMN**)
4. Random Forest with Principle Component Analysis (**RF+PCA**)
5. Linear Support Vector Machine regression with recursive feature elimination based on the coefficient value (**SVM+SVM IMP**)
6. Support Vector Machine with univariate feature selection based on F-value (**SVM+F-VAL**)
7. Support Vector Machine with univariate feature selection based on Spearman's correlation (**SVM+SPRMN**)
8. Support Vector Machine with Principle Component Analysis (**SVM+PCA**)
9. Ridge regression with recursive feature elimination based on the coefficient value (**RIDGE+RFE**)
10. Lasso regression with recursive feature elimination based on the coefficient value (**LASSO+RFE**)
11. Elastic Net regression with recursive feature elimination based on the coefficient value (**ENET+RFE**)

To evaluate the models, we split the available data into approximately 80% for training and 20% for testing. When splitting the data, we left the whole season of 2015 in the test set, as usually done in data science projects in agriculture. We also made sure to split the dataset in a stratified fashion: we performed stratification by rice age group (young, middle-aged, old), season, geographical location and cultivar. To find the optimal hyperparameters and dimension counts, we performed a stratified five-fold-cross validation on the training set.

Results

In total, we evaluated 55 different prediction (11 machine learning pipelines using five different taxonomic levels as inputs). As a benchmark, we used a model that always predicts the mean of the training data.

All 11 pipelines on all taxonomic levels performed significantly better than the benchmark model (*Supplementary material - Figure 3*). The best result was achieved using Elastic Net with recursive feature elimination on the family level (*Figure 1*). The minimum achieved mean squared error (MSE) was 85.82 (9.43 days of error). In comparison, in the original research *Edwards et al.* [7] applied a Random Forest model resulting in a MSE of 166.75 (12.89 days of error). It should be noted that the test/train split performed by *Edwards et al.* differed from our setting: they did not perform a stratified split, although, among other data in the test set, they did perform testing on the whole 2015 season.

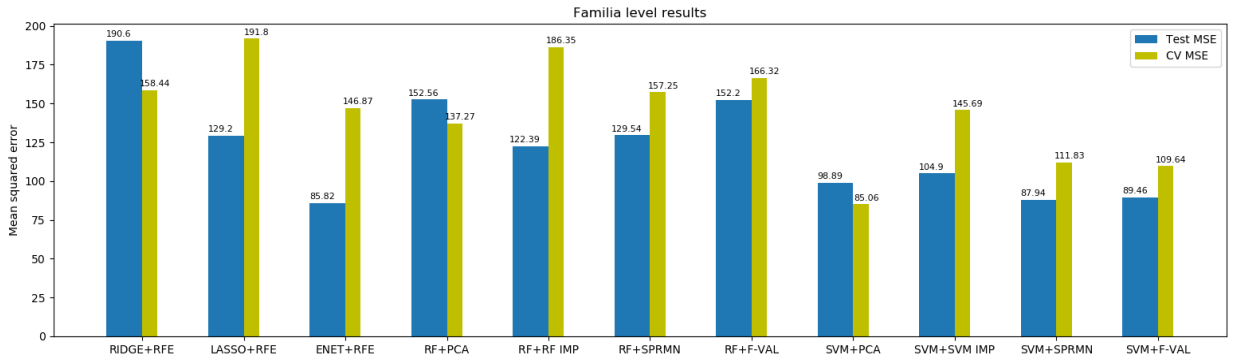

Figure 1 – method comparison on family taxonomic level without the benchmark regressor

We interpreted the best model to find out which microbes are the most correlated to plant age. This is done by examining the coefficient values of the generated regression function (*Figure 2*).

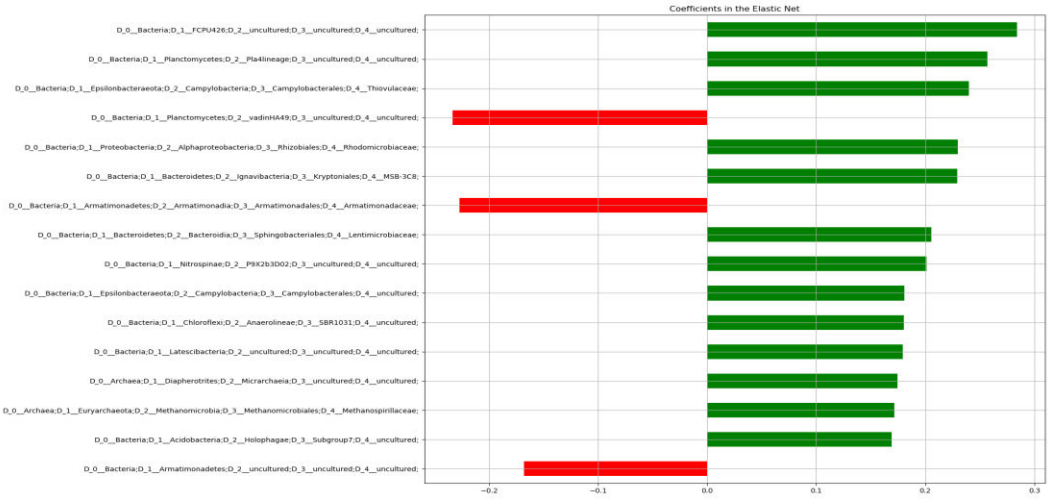

Figure 2 – microbial model-wise significance sorted by absolute value. Positive and negative correlation is denoted by green and red colours, respectively.

As seen in *Figure 2*, some microbes have not been classified to a family level. Unlike the human or mice gut microbiota, microbial communities in soil still remain heavily unexplored[16], thus, it is possible that some still uncultured microbes may be significantly correlated to plant age.

Table 1 lists the top 10 results obtained by applying different pipeline/taxonomic level input combinations. All linear shrinkage methods, including Elastic Net, demonstrated a significantly shorter execution time. Relative microbe abundances at a family taxonomic level proved to be the best predictor of plants age. Family level was followed by order, genus, class and phylum levels, respectively. In general, excluding Elastic Net, SVM outperforms Random Forest and linear shrinkage methods. This high performance of SVM comes at the cost of not being able to determine the model-wise importance of each microbe, when not using the linear kernel, and long execution time (*Table 1*). Relatively low performance of RF models can probably be explained by poor performance of decision tree regressors, that RF consits of, on linear problems.

Method	**Taxonomy level**	**Test MSE**	**CV MSE**	**Execution time** [minutes]	**Optimal dimension count**
Elastic Net + RFE	Family	85.82	146.87	65.34	347/1024
RBF SVM + Spearman's rank	Family	87.94	111.83	249.78	699/1024
RFB SVM + F-Value	Family	89.46	109.64	266.13	689/1024
RBF SVM + PCA	Family	98.89	85.06	238.75	20/1024
RFB SVM + F-Value	Order	101.49	115.66	177.4	507/661
RBF SVM + Spearman's rank	Order	101.49	117.64	178.61	507/661
SVM + SVM's Importance	Family	104.9	145.69	1346	499/1024
RBF SVM + PCA	Order	105.17	100.96	155.8	10/661
RBF SVM + F-Value	Genus	108.41	121.78	346.09	482/1305
RBF SVM + PCA	Genus	112.56	96.74	396.23	20/1305

Table 1 – top ten regressors of all taxonomic levels

Discussion

Even though the pipelines were tested on soil data, the same 11 methods can be applied to any other microbiome prediction problem, and in order to ascertain a state-of-the art predictor, more datasets need to be experimented on. Neural network architectures, such as autoencoders, hold potential for solving these kind of problems and should be explored as well.

However, the effect of different sequence analysis pipelines on the final prediction remains poorly explored. Therefore, in our future research we plan on comparing and evaluating the following:

1. Different normalization techniques
2. Usage of ASV tables instead of the OTU based ones
3. Usage of open-reference OTU picking compared to closed-reference
4. Different sequence filtering techniques, taxonomic classifiers, reference databases and similarity thresholds.

References

[1] Wang, Minggang, et al. "Removal of soil biota alters soil feedback effects on plant growth and defense chemistry." New Phytologist 221.3 (2019): 1478-1491.

[2] Hartmann, Anton, et al. "Plant-driven selection of microbes." Plant and Soil 321.1-2 (2009): 235-257.

[3] Hata, Kosay. "Studies on Plant Growth Accelerating Substances: Part I. The Isolation Method of Soil Microbes which Produce Plant Growth Accelerating Substances." Agricultural and Biological Chemistry 26.5 (1962): 278-287.

[4] Peiffer, Jason A., et al. "Diversity and heritability of the maize rhizosphere microbiome under field conditions." Proceedings of the National Academy of Sciences 110.16 (2013): 6548-6553.

[5] Shakya, Migun, et al. "A multifactor analysis of fungal and bacterial community structure in the root microbiome of mature Populus deltoides trees." PloS one 8.10 (2013): e76382.

[6] Nezarat, S., and A. Gholami. "Screening plant growth promoting rhizobacteria for improving seed germination, seedling growth and yield of maize." Pakistan journal of biological sciences 12.1 (2009): 26.

[7] Edwards, Joseph A., et al. "Compositional shifts in root-associated bacterial and archaeal microbiota track the plant life cycle in field-grown rice." PLoS biology 16.2 (2018): e2003862.

[8] Bolyen, Evan, et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. No. e27295v1. PeerJ Preprints, 2018.

[9] Callahan, Benjamin J., et al. "DADA2: high-resolution sample inference from Illumina amplicon data." Nature methods 13.7 (2016): 581.

[10] Quast, Christian, et al. "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools." Nucleic acids research 41.D1 (2012): D590-D596.

[11] Aitchison, John. "The statistical analysis of compositional data." Journal of the Royal Statistical Society: Series B (Methodological) 44.2 (1982): 139-160.

[12] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.

[13] Drucker, Harris, et al. "Support vector regression machines." Advances in neural information processing systems. 1997.

[14] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58.1 (1996): 267-288.

[15] Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." Journal of the royal statistical society: series B (statistical methodology) 67.2 (2005): 301-320.

[16] Blum, Winfried EH, Sophie Zechmeister-Boltenstern, and Katharina M. Keiblinger. "Does Soil Contribute to the Human Gut Microbiome?." Microorganisms 7.9 (2019): 287.

Supplementary material

All data, including microbe importance lists, Python scripts and figures used in the research are available at: github.com/ciganche/RiceAgePrediction.

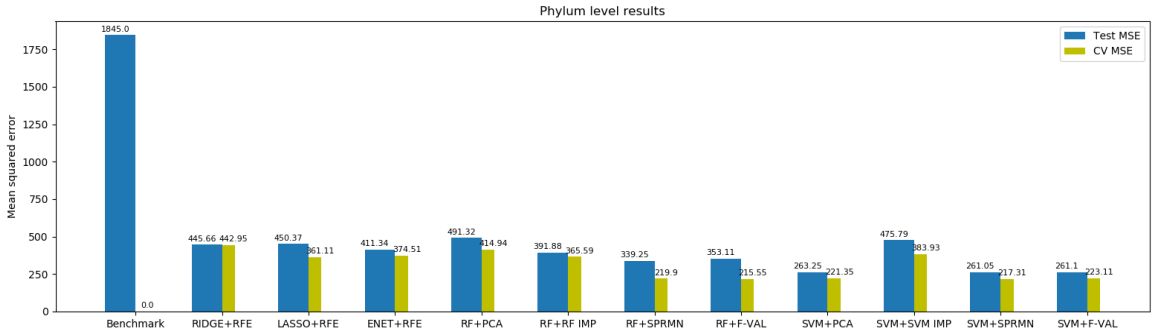

Figure 3 – pipeline comparison on the phylum taxonomic level with the benchmark regressor

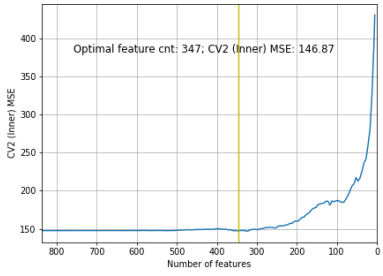

Figure 4 – optimal feature count for the best model

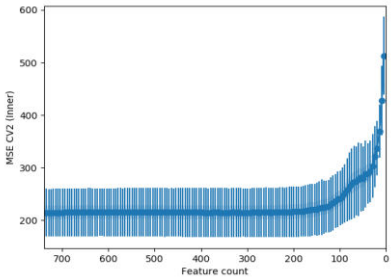

Figure 5 – mean and standard deviation of CV MSE for Elastic Net models with different sets of hyperparameters during feature selection