Plant-Microbe Interfaces: AI-GWAPA, explainable-AI-based approaches to genome wide phytobiome association

Piet Jones[1,2]* (jonespc@ornl.gov), Benjamin Garcia[1], Manesh Shah[1], Wellington Muchero[1], Jay Chen[1], Gerald Tuskan[1], Daniel Jacobson[1,2], and Mitchel Doktycz[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN; [2]The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, TN

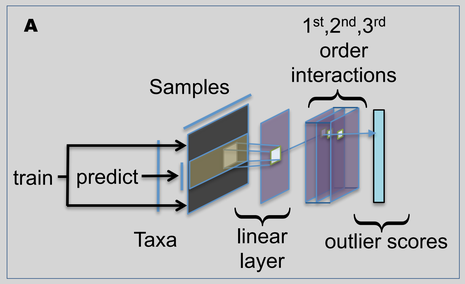
Introduction

The phytobiome consists of the plant, organismal communities and their environment. The interactions between these have significant effects on observable measurable traits that have potential economic and sustainability implications. A better systems-level understanding of the beneficial and antagonistic relationships between these components will enhance our capacity to influence these systems to produce desirable and impactful traits.

In the framework presented here: Artificial Intelligence – Genome Wide Association Phytobiome Analysis (AI-GWAPA), we utilize machine learning, deep learning and general artificial intelligence (AI) techniques), to elucidate the interactions between microbial and viral constituents of the 1000 member *Populus trichocarpa* population arrayed in common gardens in the Pacific Northwest. Metatranscriptome samples from leaf, xylem and root along with approximately 10 million SNPs called across the population allow us to associate host genetic variants to microbial/viral constituents. Using sample-specific networks, a machine learning approach, we model the contributions that a genotype has to the putative pathogenic-mutualistic relationships between taxa. Furthermore, we utilize an AI approach by training a deep learning neural network to estimate putative phytobiome-derived protein interactions among the host proteome. Together these approaches allow us to improve our fundamental understanding of the relationships between the plant and its phytobiome. (http://pmi.ornl.gov)

Factorization Machines

Taxa were identified from the leaf and xylem transcriptome, using ParaKraken, a parallel version of Kraken developed in our lab. This resulted in a phytobiome genera level classification for viruses, bacteria, archaea, fungi, nematode and aphids in a taxa-sample matrix. To improve our confidence in taxonomic assignment we processed the sparse data for putative outlier taxa. Factorization machines (FM) are an approach to approximate higher-order interactions in linear compute time, they are particularly useful for outlier detection in sparse data [1]. Here we implemented the deep learning FM-outlier approach in Pytorch [2], using a k-fold cross validation approach after initial feature engineering. Training was performed on k-1 set of taxa, with a response vector of 0, followed by prediction on the kth set of taxa. Repeating this for multiple iterations, we obtained an outlier score. A score cutoff based on the median absolute deviation from the median (MAD) was applied.



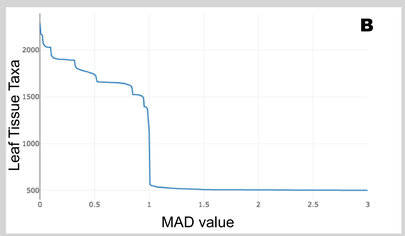


Figure 1: (A) Diagram of the FM-outlier neural network. The input matrix has taxa as rows and samples as columns. Taxa are grouped into training and test sets, respectively. After a linear layer, a combined layer of up to 3rd order interactions are estimated, the resultant mean squared error is back propagated to learn the model parameters. (B) Median absolute deviation (MAD) from the median values for the FM-outlier scores, these indicate a clear cutoff of 1. Taxa with a MAD value for their score > 1 are deemed to be outliers and discarded.

Genome Wide Association Analysis

The phytobiome taxa that remained after the FM-outiier analysis were treated as phenotypes in a genome wide association analysis. Approximately 10 million single-nucleotide polymorphisms (SNPs), were then filtered and used as the genotype information. Only SNPs with a minor allele frequency greater than 0.01 were analyzed using EMMAX [3]. Heritability was estimated from the same set of SNPs, after removing those SNPs with an ldscore > 0.5. Phenotype measurements were further masked if their MAD score > 5, and only phenotypes with non-masked observations in more than 5% of the population were analyzed. An FDR value of 0.01 was applied to correct for multiple hypotheses bias. Only SNPs that fall within a gene boundary are reported, resulting in a taxa to gene association. Results are visualized in a hive plot in Figure 2

Mutualism/antagonism

DUO was used to compare taxa abundance across the population. Taxa abundance vectors are compared pair-wise after categorizing individual measurements into High, Medium or Low, based on the quantiles of the entire dataset. The metric then evaluates how correlated the high (up) /low (down) components of the vector are. This results in 4 correlation values: UU, UD, DU and DD (U=up,D=down). The correlation is therefore directed, and has a source and sink, respectively. Mutualism is suggested by a UU or DD correlation, while antagonism is suggested by a UD or DU correlation. See Figure 3.

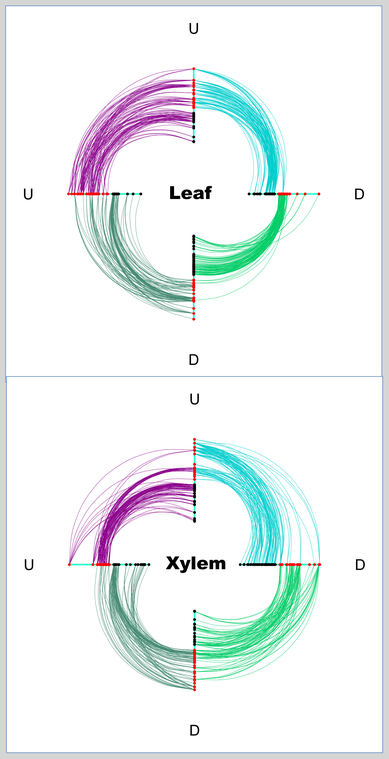
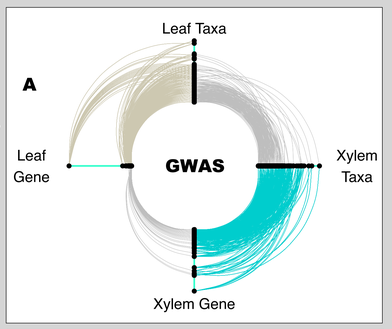
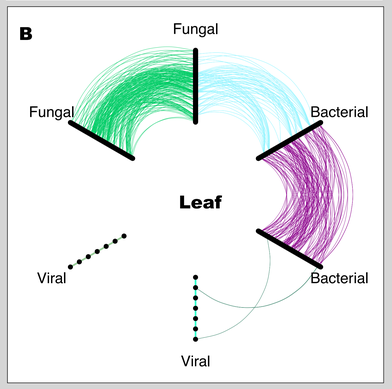


Figure 3: Hive plots of the top 100 DUO metric results of taxa in leaf and xylem samples, respectively. Nodes on the respective axes are taxa. The axes are U=Up, D= Down. The plot is read anti-clockwise, thus an edge between U and D represents a putative antagonistic relationship, as the first taxa has a higher abundance (Up) while the second has a lower abundance (Down). Red nodes are sorted based on the metric representing the sink nodes results (higher values are therefore further out from the center of the plot) . The order of the black nodes carry no meaning and merely represent the source node.





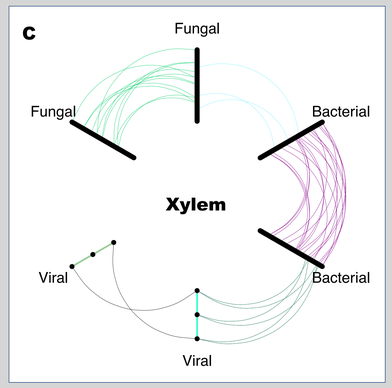


Figure 2: Hive plots of the GWAS results. A hive plot consists of a set of axis that represent categories, on these axis nodes or objects are arranged based on some measurement, for example connectivity. An arch is drawn between nodes on each axis to indicate a relationship. (A) The four axes represent either genes or taxa nodes, from the respective leaf/xylem GWAS analysis. Nodes on a gene axis are arranged based on the number of taxa associations (the outer nodes therefore have higher connectivity), similarly taxa nodes are arranged based on the number of GWAS results. Arcs between taxa and gene nodes indicate a significant GWAS result. Taxa that are both in leaf and xylem are connected, similarly genes both in leaf and xylem are connected. (B) Hive plot of the neighborhood of the gene that is associated to a particular taxa from the leaf. Nodes are arranged in a random order, consistent across axes. Arcs between taxa indicate that they share a gene, based on the significant GWAS results. We see that there is one virus that is associated with a gene which in turn associates with a bacteria. Similarly there are large fungal to fungal, bacterial to bacterial and bacterial to fungal neighborhoods. (C) Similar to B, but for the xylem GWAS instead. Here we see a viral to viral neighborhood that is absent from the leaf, a few more viral to bacterial neighborhoods.

Sample Specific Networks (SSN)

SSN values are generated by removing a genotype and recalculating the DUO metric. By doing this for all genotypes and then observing the resultant change from the original metric, we can estimate the genotype's contribution to the DUO correlations. See Figure 4.

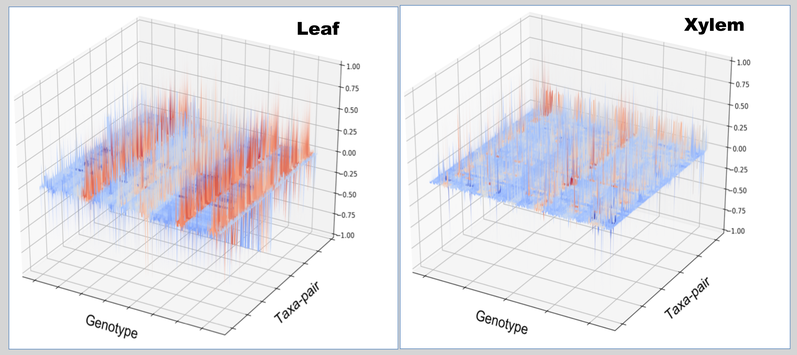
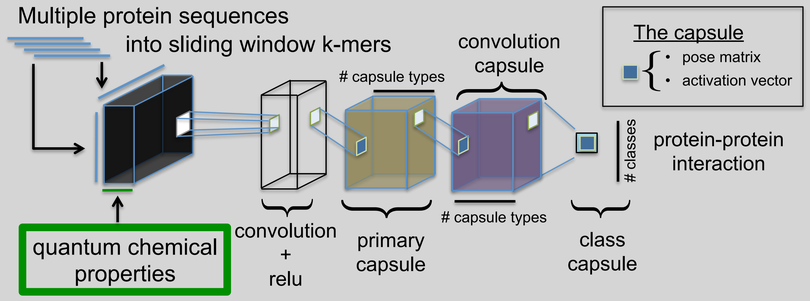


Figure 4: In the figure red indicates that the genotype has a negative effect on the metric (mutualism in this case), while blue indicates the genotype has a positive effect. (A) Surface plot of the genotype effect in leaf samples on the DUO UU value, the effect on putative mutualism. (B) Surface plot of genotype effect in xylem samples on DUO UU values.

In progress



We are currently developing a capsule network-based deep learning model that is capable of predicting protein-protein interactions from sequence data. We train the network on kmers of proteins that interact, and take into account approximately 200 quantum chemical properties of the respective amino acids. Capsule networks are capable of taking into account the localization of features. The network will therefore, through feature engineering, provide information on chemical properties and protein segments that explain the observed interaction. Training is currently underway using data from *Arabidopsis thaliana*. Transfer learning will be used to adapt the model to *Populus trichocarpa*. Protein interactions will then be predicted based on genes associated to taxa in the phytobiome.

Conclusion

Here we provide a comprehensive framework that allows for a systems biology approach in the analysis and interpretation of the complex interactions between a host and its phytobiome. From the metatranscriptome samples, we obtain a snapshot of the putative constituents of this phytobiome. The factorization machine learning approach allows us to model up to 3rd order interactions between the respective taxa, thereby retaining signal that would otherwise be missed using standard metric-based analysis. GWAS analysis helps to uncover the nature of this complex interaction. We find a few highly connected genes involved in functions such as phospholipid transportation, protein degradation, transcriptional regulation, etc. The differences in the nature of the associations when comparing leaf to xylem samples is also apparent from the above figures. With more advanced metrics, such as DUO, we see differences in the types mutualistic/antagonistic relationships when comparing leaf and xylem. With sample specific networks we can start to understand the genotypic effect on these relationships. By further using a deep learning-based protein interaction model we can work towards a protein level understanding of this dynamical system.

References:

1. Zhu, Mengxiao, et al. "Outlier Detection in Sparse Data with Factorization Machines." *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017.
2. Paszke, Adam, et al. "Automatic differentiation in PyTorch." (2017).
3. Kang, Hyun Min, et al. "Variance component model to account for sample structure in genome-wide association studies." *Nature genetics* 42.4 (2010): 348.

Acknowledgement:

This research was funded by the US DOE Office of Biological and Environmental Research, Genomic Science Program. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the US Department of Energy under Contract no. DEAC05-00OR22725. This research used resources of the Oak Ridge Leadership Computing Facility (OLCF) at the Oak Ridge National Laboratory. Additional resources used was provided by the Compute and Data Environment for Science (CADES). .