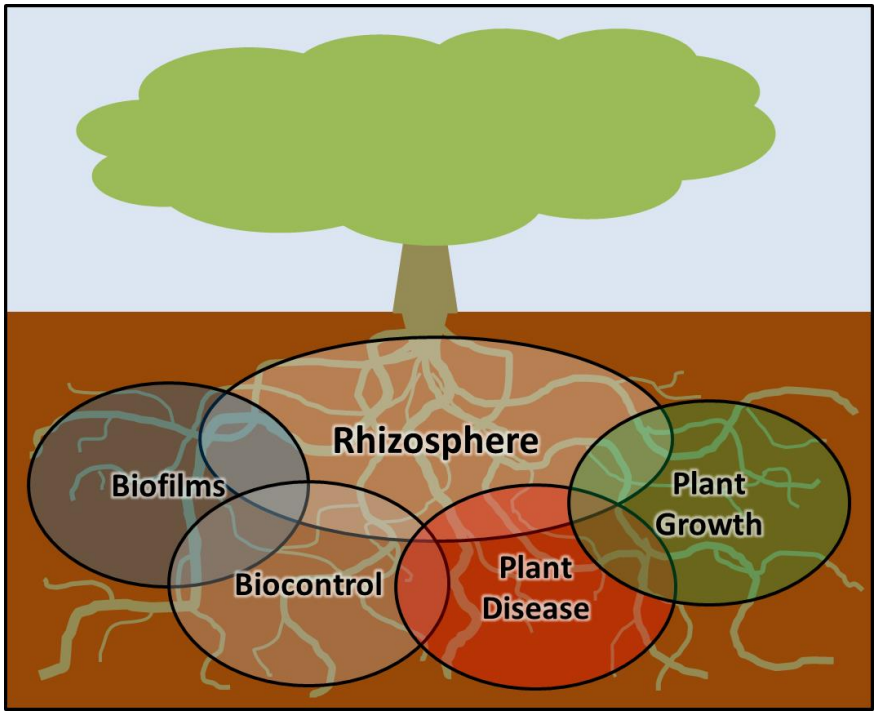
Predicting Pseudomonads' ecological roles in the rhizosphere using machine learning and 'omics' computational modeling

Peter E. Larsen[1,2,*] (plarsen@anl.gov), Yang Dai[2], Mark W. Silby[3], Frank R. Collart[1], **Philippe Noirot**[1]

[1] Biosciences Division, Argonne National Laboratory, Lemont IL; [2] Department of Bioengineering, University of Illinois at Chicago, Chicago IL; [3] Department of Biology, University of Massachusetts Dartmouth, N. Dartmouth, MA

The ability to obtain complete genome sequences from bacteria in environmental samples, such as soil samples from the rhizosphere, has highlighted the microbial diversity and complexity of environmental communities. However, new algorithms to analyze genome sequence information in the context of community structure are needed to enhance our understanding of the specific ecological roles of these organisms in soil environments. We present a machine learning approach using sequenced Pseudomonad genomes coupled with outputs of metabolic and transportomic computational models for identifying the most predictive molecular mechanisms indicative of a Pseudomonad's ecological role in the rhizosphere: a biofilm, biocontrol agent, promoter of plant growth, or plant pathogen. The capacity to form biofilms are indicative of Pseudomonad's capacity to form sessile colonies of plant roots. Biocontrol capacity is Pseudomonads' ability to defend plant roots against plant bacterial, fungal, or animal pathogens. Plant pathogenicity is the ability of certain Pseudomonads to cause disease in plants. Plant growth promotion is the capacity of Pseudomonads to increase plant biomass through providing access to nutrients, remediating abiotic stress, or biosynthesis of plant hormone analogues that influence plant growth.

Application of this model requires the input of sequence, annotated Pseudomonads that can confidently be ascribed to one of the selected rhizosphere ecological niche types. Genomes are re-annotated using a custom database of over 754,000 enzymes and 164,000 transporter annotated protein sequences to insure uniformity of annotations across all genomes. From re-annotated genomes, enzyme function profiles, metabolomic models, and transportomic models are generated. For this analysis, a novel modeling approach to quantify a bacteria's realative capacity to transport specific ligands across the membrane, Predicted Relative Trans-membrane Transport (PRTT) has been developed. These datatypes are used to train Support Vector Machines (SVMs) that can determine membership to rhizosphere niche type. Most predictive



features identified by SVM provide valuable insights into the specific molecular meahcnisms by which Pseudomonads are adapted to the rhizosphere environment and their interactions with plant roots.

Computational predictions of ecological niche were highly accurate overall with models trained on transportomic model output being the most accurate (Leave One Out Validation F-scores between 0.82 and 0.89). The strongest predictive molecular mechanism features for rhizosphere ecological niche overlap with many previously reported analyses of Pseudomonad interactions in the rhizosphere, suggesting that this approach successfully informs a system-scale level understanding of how Pseudomonads sense and interact with their environments. Specific metabolic and transportomic functions are identified that are important for Pseudomonad adaptations to rhizosphere ecological niche types. Transport activities that are identified as predictive for inhabiting the rhizosphere involve carbohydrate transporters (e.g. 2-O-alpha-manosyl-D-glycerate) suggestive for osmoregulation in soils and 3-hydroxyphenylpropionic, one of many lignin breakdown products, which are ubiquitous in soils. Biocontrol is most predicted by its transportome, specifically by transport of cobamide coenzyme, and monosaccharides. The most predictive metabolic activities for biofilm formation are for anti-biofilm compounds protoporphyrin and methyglyoxal. Fatty acid biosynthesis pathways were identified as features predictive for plant pathogenicity in Pseudomonads. Metabolomic input type predicts that synthesis of a number of plant signaling compounds is predictive of plant growth promotion by Pseudomonads including indole and flavones eriodictyol, neringenin. C4-dicarboxylate, calcium, and glutathione transport are also predictive of plant growth promotion The ability to transport of a number of simple sugars (i.e. malonate, mannose, sucrose, galactose, and hexose) is found to be predictive of plant growth promotion by Pseudomonads and is suggestive of an ecological niche that is able to take advantage of exuded photosynthetic sugars present in the rhizosphere. The observation that an organism's transportome is highly predictive of its ecological niche is a novel discovery and may have implications in our understanding microbial ecology. The framework developed here can be generalized to the analysis of any bacteria across a wide range of environments and ecological niches important to carbon cycling and plant-rhizosphere community interactions making this approach a powerful tool for providing insights into functional predictions from bacterial genomic data for DOE mission applications.