Plant-Microbe Interfaces: Application of machine-learned protein-metabolite binding prediction models to plant-microbe interfaces

Omar N. Demerdash[1]* (demerdashon@ornl.gov), Amy L. Schaefer,[2] Caroline S. Harwood,[2] Dale A. Pelletier,[1] and **Mitchel J. Doktycz**[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN; [2]The University of Washington, Seattle, WA

http://pmiweb.ornl.gov/

Project Goals: The goal of the PMI SFA is to characterize and interpret the physical, molecular, and chemical interfaces between plants and microbes and determine their functional roles in biological and environmental systems. *Populus* and its associated microbial community serve as the experimental system for understanding the dynamic exchange of energy, information, and materials across this interface and its expression as functional properties at diverse spatial and temporal scales. To achieve this goal, we focus on 1) defining the bidirectional progression of molecular and cellular events involved in selecting and maintaining specific, mutualistic *Populus*-microbe interfaces, 2) defining the chemical environment and molecular signals that influence community structure and function, and 3) understanding the dynamic relationship and extrinsic stressors that shape microbiome composition and affect host performance.

Computational prediction of the binding of small organic metabolites and other ligands to biological macromolecules has far-reaching implications for a range of problems, particularly metabolomics. Small metabolites are implicated in a host of roles, including symbiotic relationships between plant and microbe. Nonetheless, critical tasks such as predicting the bound structure of a protein-ligand complex along with its affinity have proven to be very difficult, owing largely to the inherent approximations in generating physically reasonable bound conformations of the ligand and an accurate free energy or proxy thereof. In recent years, machine learning-based methods have proven to be more robust than the standard linear sum of energetic terms, suggesting a complex, potentially non-linear interaction among terms. However, these methods are often trained on a small set of features, with a single functional form for any given energetic or physical effect, and often with little mention of the rationale behind choosing one functional form over another. Moreover, a systematic investigation of the effect of machine learning method is not undertaken, with a single method being favored for reasons that are often obscure. Here we undertake a comprehensive effort towards developing high-accuracy, machine-learned scoring functions, systematically investigating the effects of machine learning method and choice of features, and, providing insights into the relevant physics using methods that assess feature importance. Here, we show synergism among disparate features, yielding Pearson correlations (R^2) with experimental binding affinities of up to 0.865 and enrichment for native bound structures of up to 0.913 in an independent test set consisting of the well-known CASF-2013 benchmark. We deploy these models to predict the relative activity of metabolites in two systems of importance in plant-microbe symbiosis, one plant-bacterial (the LuxI enzyme and its potential substrates), and the other an enzyme that synthesizes plant defense hormones. We show the ability to discriminate

low- from high-activity substrates and describe further how these methods shall be deployed on a larger scale to screen larger sets of molecules.

The Plant Microbe Interfaces Scientific Focus Area is sponsored by the Genomic Science Program, U.S Department of Energy, Office of Science, Biological and Environmental Research.