Identifying Data-Driven Gene Targets to Control Bacterial Fitness

Shara Balakrishnan[1]* (sbalakrishnan@ucsb.edu), Adam Deustchbauer[2], Enoch Yeung[3], and **Rob Egbert**[4]

Institutions: [1]Department of Electrical and Computer Engineering, University of California Santa Barbara, Santa Barbara, California; [2]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Lab, Berkeley, California; [3]Department of Mechanical Engineering, University of California Santa Barbara, Santa Barbara, California; [4]Earth & Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington

Website: https://genomicscience.energy.gov/research/sfas/pnnlbiosystemsdesign.shtml

Project Goals: The Pacific Northwest National Laboratory Persistence Control Scientific Focus Area aims to gain a fundamental understanding of factors governing the persistence of engineered microbial functions in rhizosphere environments. From this understanding, we will establish design principles to control the environmental niche of native rhizosphere microbes. In our first funding period, we are examining the efficacy of genome reduction and metabolic addiction to plant root exudates in environmental isolates as persistence control strategies using the bioenergy crop sorghum and defined microbial communities as a model ecosystem. Effective persistence control will lead to secure plant–microbe biosystems that promote stress-tolerant and highly productive biomass crops.

Abstract: An important factor in plant growth is the interdependent relationship between the plant and the soil microbiome. By genetically modifying the microbes, we can modulate the soil composition to promote robust plant growth in low-resource environments. A microbe with engineered functions, when introduced into a new environment, interacts with the native microbiome, and subjects itself to competitors and predators [1]. A first step towards introducing engineered microbial functions to promote plant growth is to control the environmental persistence of the microbe of interest in the target environment. Our objective is to develop growth harness actuators, genetic devices that can control gene expression to regulate the microbial growth. In this poster, we present how to identify "fitness genes" for a bacterium in a target environment using time-series RNAseq data along with time series growth data. We propose a novel data-driven sensor fusion technique that combines dynamical systems theory and machine learning to discover dynamic genotype-to-phenotype models that can simulate single as well as combinatorial gene knockouts to identify optimal sets of genes corresponding to the phenotype. We further validate these genes using a novel time-series transposon knockout assay.

Current approaches to identify genes for a phenotype of interest hinge on differential RNA expression of genes either across time or across media conditions; the genes that exhibit maximal differential mRNA expression across conditions are the important genes for that condition. These approaches only use the mRNA expression data and check for individual phenotypes. We represent the genotype-to-phenotype dynamics as a dynamical state-space model by assuming gene expression to represent the state of the system and phenotype to represent the output of the system. The state-space model contains the state equation which captures the gene dynamics and the output model captures the mapping of instantaneous gene activity to the growth, the phenotype of interest. This predictive modeling framework can be extended to fuse other types of data like metabolomics, proteomics, fluorescence, and microscopic data either as a

state or output depending on the availability of data and the problem being solved. We developed an algorithm called output constrained deep dynamic mode decomposition (OC-DeepDMD) algorithm which uses multilayer feedforward neural networks to identify a high dimensional linear Koopman model of a relatively lower-dimensional nonlinear system such that the state-space model becomes linear in both the state and output equation [2]. Using this model, we simulate the effect of single-gene knockouts on growth output.

To validate the model predictions, we do Random Barcoded Transposon Sequencing (RB-TnSeq) experiments which is an alternate method to identify genes that relate to fitness in the media conditions used in the experiment [3]. RB-TnSeq employs a pool of single-gene knockout mutants of a single strain via transposon insertions with unique genetic barcodes and captures the fitness of the individual strains in various media conditions; the more negative the strain fitness value in a certain media condition, the more important the gene for that condition. Typically, RB-TnSeq experiments are done by considering an initial and a final time point. We extend the experiment to include multiple intermediate time points and compute the fitness curves as a function of time for each mutant and establish a benchmark for the comparison of the fitness predictions obtained for the single-gene knockout mutants from the Koopman models.

In this work, we consider the growth of the soil bacterium *Pseudomonas putida* in R2A media with varying concentrations of two nutrients - glucose as a carbon source and casein hydrolysate as a source for amino acids. By observing the growth curves of *P. putida* under varying concentrations of the two nutrients, we selected the condition under which a maximum growth rate (MAX condition) is observed and the negative control (NC) condition in which the nutrients are absent. We performed time-series RNA sequencing and RB-TnSeq experiments for the MAX and NC conditions while obtaining optical density (OD) growth measurements. We used the RNAseq and OD data in the OC-DeepDMD algorithm to learn a Koopman operator representation of the state-space model. We simulated fitness with single-gene knockouts and validated our model predictions using the RB-TnSeq data.

In summary, we propose to learn dynamic genotype-to-phenotype models using the OC-DeepDMD algorithm, predict the fitness of single-gene knockouts and experimentally validate them using time-series RB-TnSeq experiments. In the future, we plan to simulate combinatorial gene knockout and identify optimal gene targets to control microbial fitness.

References/Publications

[1] Wang, F. and Zhang, W., 2019. Synthetic biology: recent progress, biosafety and biosecurity concerns, and possible solutions. *Journal of Biosafety and Biosecurity*, pp.22-30

[2] Balakrishnan, S., Hasnain, A., Egbert, R. and Yeung, E., 2021. The Effect of Sensor Fusion on Data-Driven Learning of Koopman Operators. *arXiv preprint arXiv:2106.15091*.

[3] Wetmore, Kelly M., et al. "Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons." *MBio* 6.3 (2015): e00306-15.

Funding statement: *This research was supported by the U.S. Department of Energy (DOE), Office of Biological and Environmental Research (BER), as part of BER's Genomic Science Program (GSP), and is a contribution of the Pacific Northwest National Laboratory (PNNL) Secure Biosystems Design Science Focus Area "Persistence Control of Engineered Functions in Complex Soil Microbiomes" through a subcontract to the University of California Santa Barbara. PNNL is a multi-program national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RL01830.*