BIG DATA ON SMALL ORGANISMS: GENOME-SCALE MODELING AND PHENOTYPIC PREDICTION OF ESCHERICHIA COLI

Allocation: NSF PRAC/300 Knh
PI: Ilias Tagkopoulos[1]

[1]University of California-Davis

EXECUTIVE SUMMARY

We developed semi-supervised normalization pipelines and performed experimental characterization (growth, transcriptional, proteome) to create a consistent, quality-controlled multiomics compendium for *Escherichia coli* with cohesive metadata information. We then used this resource to train on Blue Waters a multi-scale model that integrates four omics layers to predict genome-wide concentrations and growth dynamics. Large scale simulations and subsequent validation led to several interesting results. First, the genetic and environmental ontology reconstructed from the omics data was found to be substantially different and complementary to the genetic and chemical ontologies. Second, the integration of all layers led to the predictor with the highest performance, although the increase in accuracy was incremental. We have validated 16 new predictions by genome-scale transcriptional profiling. This work constitutes the largest omics-based simulation and demonstrates how large high-performance computing infrastructure, big data and novel computational methods can lead to an integrative framework to guide biological discovery.

INTRODUCTION

Predicting microbial behaviors through data-driven computational tools is key to a) understand how microbial behaviors emerge and b) engineer the organisms in a fast, robust and accurate manner for biomedical and biotechnological applications. Realistic models of bacterial cells can lead to new paradigms related to bacterial physiology as well as the rational redesign of genomes with desired behavior. For this task, multi-scale modeling and visualization tools are of paramount importance as they provide insight of the systems under study.

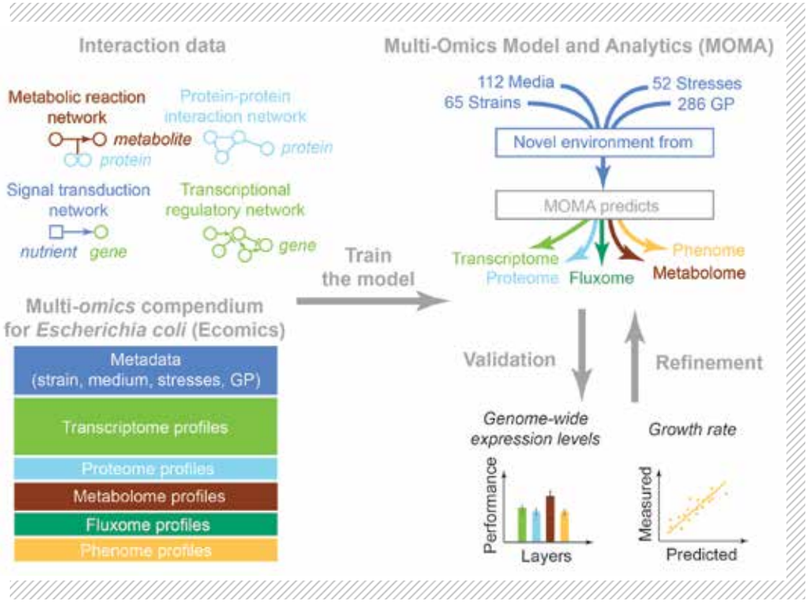

FIGURE 1: Dataset and Methodology

METHODS & RESULTS

This work had two parts, the generation of a consistent omics compendium and the development of the integrative model that uses it as a training set. As such, we first constructed Ecomics, a normalized, well-annotated, multi-omics database for *E. coli*, developed to provide high-quality data and associated meta-data for performing predictive analysis and training data-driven algorithms. This compendium houses 4,389 normalized expression profiles across 649 different conditions. We then proceeded to create the Multi-Omics Model and Analytics (MOMA) platform, an integrated model that learns from the Ecomics and other available network data to predict genome-wide expression and growth.

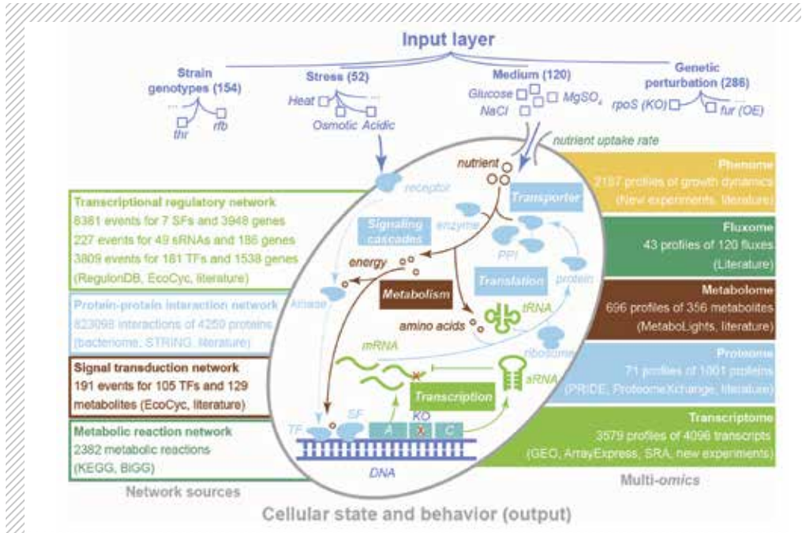

FIGURE 2: Layers of biological organization captured and information used to train the model.

WHY BLUE WATERS

Due to the complexity and size of the datasets (18 million points from various platforms and molecular species), the computational complexity of the algorithms that range from constrained regression to artificial neural networks, **Blue Waters was critical for the completion of the large-scale simulations** we performed.

NEXT GENERATION WORK

Our first publication appeared in *Nature Communications*. We have developed a deep learning method for proteome reconstruction, which has been tested already in Blue Waters and will be integrated into the model. Aside from algorithmic improvements, we aspire to move to population simulations of this model to investigate and predict population-level phenomena, similarly what we have done before, with simpler models [1,2]

PUBLICATIONS AND DATA SETS

Kim, M., N. Rai, V. Zorraquino, and I. Tagkopoulos, Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli, *Nat. Commun.* 7:13090 EP (2016), doi:10.1038/ncomms13090.