TN

ML

A CRYSTAL BALL OF BACTERIAL BEHAVIOR: FROM DATA TO PREDICTION USING GENOME-SCALE MODELS

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

[1]University of California, Davis

EXECUTIVE SUMMARY

The democratization of mass sequencing and profiling technologies has resulted in a plethora of data that can reveal how organisms are organized and function on a cellular level. The goal of this project is to integrate the millions of data points collected for a model bacterium, Escherichia coli, so we can build predictive models of its behavior in novel, untested environments. We used Blue Waters for two major tasks. The first was to apply novel deep-learning algorithms for making sense of proteomics samples to find what proteins are present in given experimental settings. The second was to run massively parallel simulations with genome-scale, integrative models that predict the omics expression and bacterial behavior in novel environments. This work led to novel algorithms for omics data processing, multi-omics modeling, and the most integrative predictive model of bacterial behavior, with forward predictions validated experimentally. Parallelization of these processes on the Blue Waters supercomputer enabled a precise data-processing pipeline, search of optimal model architecture, and hyper-parameter optimization.

RESEARCH CHALLENGE

One of the grand challenges of systems and synthetic biology is the development of an accurate predictive model of any organism. If we had such a tool at our disposal, we could interrogate the cellular states and molecular abundances that exist in general under a specific selection pressure or environmental setting. In the realm of synthetic biology, it would allow the in silico testing of synthetic circuits and the genetic engineering of the chassis cells so that they exhibit optimal behavior. However, there are many obstacles to achieving this vision, including the large complexity of the cellular organization and machinery, the lack of cohesive datasets, and models that are capable of integrating them into one system that is more than the sum of its parts.

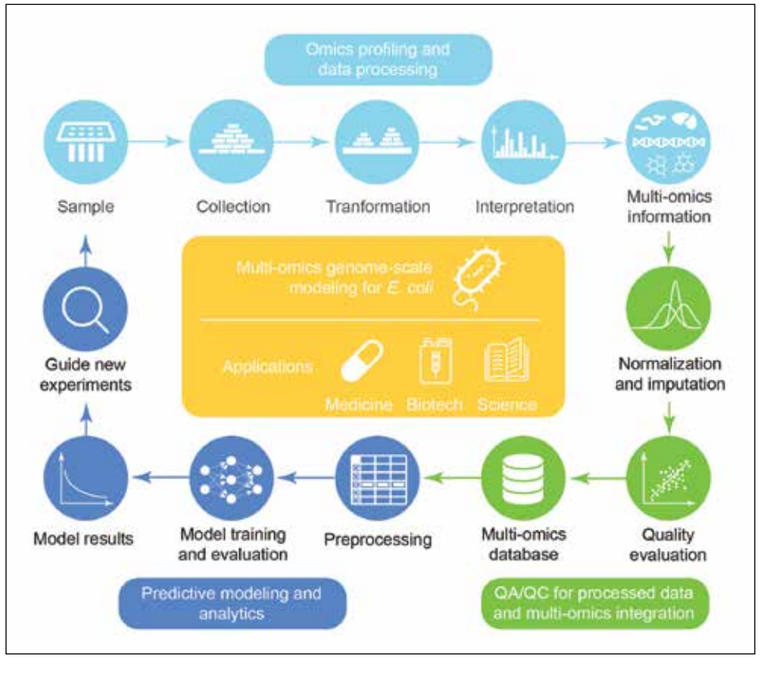

Figure 1: Overview of the processing pipeline and genome-scale modeling for *E. coli*. There are three major steps: 1) omics profiling and data processing, 2) QA/QC for processed data and multi-omics integration, and 3) predictive modeling and analysis.

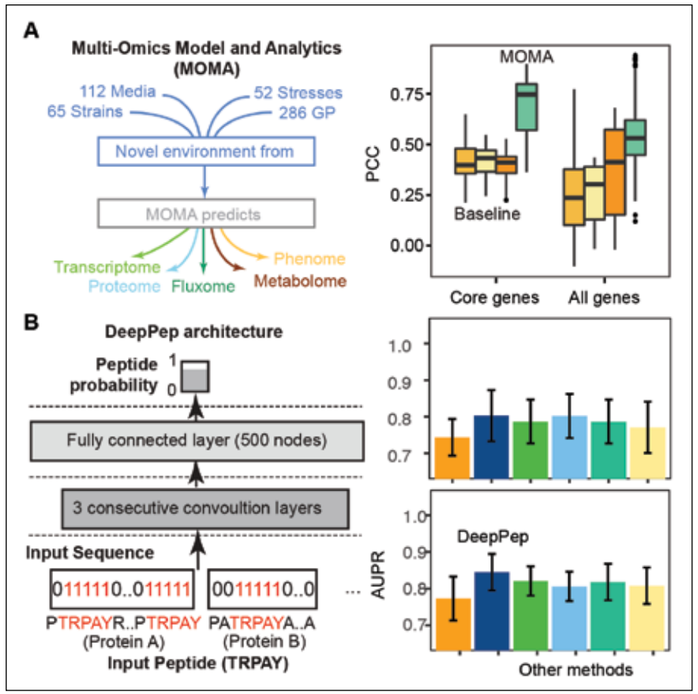

Figure 2: Performance evaluation of the proposed processing method (DeepPep) and the genome-scale model for *E. coli*. (A) DeepPep outperforms existing methods for protein inference from peptide-level proteome data. The multi-omics genome-scale model for *E. coli* shows predictive capacity in predicting molecular responses in transcriptome layer (B) and proteome layer (C) compared to multiple baseline methods.

METHODS & CODES

The proposed protein inference method (DeepPep) in the omics data-processing pipeline is based on convolutional neural networks with deep-learning techniques to efficiently train the complex model. DeepPep is written in torch/python, and the method was validated with the six external data sets (Sigma49 [1], UPS2 [2], 18Mix [3], Yeast [4], DME [5], HumanMD [6], and HumanEKC [7]). The genome-scale model employs a recurrent neural network and a constrained regression altogether to predict genome-wide responses layer by layer.

RESULTS & IMPACT

The predictive performance of the multi-scale model has shown substantial improvement over prior work as well as the ability to predict not only phenotypic information (growth, traits, etc.) but also genome-wide gene, protein, and metabolite expression. This brings us a step closer to having a crystal ball for prediction of bacterial states and behavior in novel environments. Having such a tool in our arsenal will allow the fast and inexpensive testing of experimental settings, which in turn will allow us to navigate the vast experimental space in search of suitable experimental environments of phenotypes. From biotechnology to medicine, such capability has several applications and can be transformational if applied at an industrial scale.

WHY BLUE WATERS

Our large-scale simulation for multi-scale modelling as well as the size and complexity of our datasets (30 million data points from several platforms and molecular species) necessitate the use of high-performance computing for this project. Furthermore, the effectiveness of GPU capabilities in Blue Waters for our application of deep neural network models both in multi-omics- and proteomics-specific projects allowed us to train large and complex models.

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.* (2016), DOI: 10.1038/ncomms13090.

Kim, M., A. Eetemadi, and I. Tagkopoulos, DeepPep: deep proteome inference from peptide profiling. *PLoS Computational Biology*, accepted (2017).

https://deeppep.github.io/DeepPep/
http://www.prokaryomics.com