Machine Learning Aided Advances in Synthetic Biology

A 90-minute workshop for IWBDA 2021

Organizer: Dr Vishwesh Kulkarni

Organizer's Affiliation: University of Warwick Organizer's Email: V.Kulkarni@warwick.ac.uk

Synthetic biology aims to understand, refine, control, re-engineer, and evolve nature. Data is a key element in this process. Rapid progress in synthetic biology is making increasingly large, increasingly complex, and novel datasets available to us. These exciting developments have direct parallels in machine learning which is facilitating methods and paradigms that have increasingly superior computational speed as well as increasingly superior ability to transform data into useful information. What are the possibilities if synthetic biology and machine learning come together with great synergy? In this workshop, we present four talks that answer some specifics for this exciting open-ended question.

Talk 1

Title: Tapestry Pooling – A Compressed Sensing Approach to Pooled RT-PCR Testing for COVID-19 Detection

Presenter: Sabyasachi Ghosh and Dr Manoj Gopalkrishnan (IIT Bombay)

Abstract: We present 'Tapestry', a single-round pooled testing method with application to COVID-19 testing using quantitative *Reverse Transcription Polymerase Chain Reaction* (RT-PCR) that can result in shorter testing time and conservation of reagents and testing kits, at clinically acceptable false positive or false negative rates. This innovative approach towards deconvolution of pooled tests is realised using concepts and techniques from compressed sensing and combinatorial group testing. Unlike Boolean group testing algorithms, the input is a quantitative readout from each test and the output is a list of viral loads for each sample relative to the pool with the highest viral load. For guaranteed recovery of k infected samples out of $n \gg k$ being tested, Tapestry needs only $O(k \log(n))$ tests with high probability. We then show that deterministic binary pooling matrices can ensure acceptable balance between good reconstruction properties and matrix sparsity. This enables large savings using Tapestry at low prevalence rates while maintaining viability at prevalence rates as high as 9.5%. Empirically, single-round Tapestry pooling improves over two-round Dorfman pooling by almost a factor of 2 in the number of tests required. We evaluate Tapestry in simulations with synthetic data obtained using a novel noise model for RT-PCR

and validate it in wet lab experiments with oligomers in quantitative RT-PCR assays. Lastly, we describe use-case scenarios for deployment.

Talk 2

Title: CAMP (Co-culture/Community Analyses for Metabolite Production) – Constraint-based Optimization to Maximize the Target Metabolite Production

Presenter: Maziya Ibrahim and Dr Karthik Raman (IIT Madras)

Abstract: Understanding compatibility and interactions based on growth between the members of a microbial community is imperative to exploit these communities for biotechnological applications. We introduce a computational analysis framework "CAMP" (Co-culture/Community Analyses for Metabolite Production) that evaluates all possible two-species communities generated from a given set of microbial species on single or multiple substrates to achieve optimal production of a target metabolite. We present results on how it can be used on the *genome-scale metabolic models* (GSMMs) belonging to *Lactobacillus*, *Leuconostoc*, and *Pediococcus* species from the Virtual Metabolic Human (https://vmh.life/) resource and well-curated GSMMs of *L. plantarum* WCSF1 and *L. reuteri* JCM 1112. We analyse 1,176 two-species communities using a constraint-based modelling method for steady-state flux-balance analysis of communities and determine the maximum lactate flux in the communities. We determine which substrates, when used separately or in combination, result in parasitism, amensalism, or mutualism in the bacterial communities. Metabolic engineering strategies, such as the reaction knockouts that can improve product flux while retaining the community's viability, are identified using *in silico* optimisation methods. Reaction knockouts that enhance lactate production are determined on the same lines. Our approach can guide in the selection of the most promising communities for experimental testing and validation to produce valuable bio-based chemicals.

Talk 3

Title: Large scale active-learning-guided exploration for in vitro protein production optimization

Presenter: Dr Jean-Loup Faulon (INRA)

Abstract: Today, lysate-based cell-free systems constitute a major platform to study gene expression but suffer from poor predictive accuracy concerning the protein production. A reason for this limitation is the batch-to-batch variation. We show how this limitation can

be overcome using advanced machine learning techniques such as an active learning approach. Our solution is implemented using an acoustic liquid handling robot (Echo 550, Labcyte, USA) and a plate reader (Infinite MF500, Tecan, USA) to measure cell-free reactions, inclusive of controls and triplicates. This generates the data required to train our machine learning models. Our active learning algorithm explores a combinatorial space of nearly 4 million cell-free buffer compositions. It not only maximizes protein production but also identifies which parameters are critical parameters in cell-free productivity. We also facilitate a one-step-method to achieve high quality predictions for protein production. Our approach gives precious information about the extent to which the protein production can be improved, how a home-made cell-free system can be made more efficient, and the efficiency of the transcription/translation processes can be improved.

Talk 4

Title: An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals

Presenter: Dr Jonathan Tellechea and Dr Pablo Carbonell (Polytechnic University of Valencia)

Abstract: The microbial production of fine chemicals is emerging as a bio-sustainable manufacturing paradigm. It has been implemented successfully to realize small scale production of natural products and high-value chemicals. But many daunting challenges must be overcome before it can be adopted for the large-scale industrial productions. One such challenge stems from excessive resource requirements. To overcome it, we present an automated Design–Build-Test–Learn (DBTL) pipeline using which a rigorous discovery and optimization of biosynthetic pathways can be implemented. We first show how our DBTL pipeline can be applied to produce flavonoid (2S)-pinocembrin in *Escherichia coli* and demonstrate rapid iterative DBTL cycling with automation at each stage. In this case, application of two DBTL cycles successfully established a production pathway improved by 500-fold, with competitive titers up to 88 mg L^{-1}. We also show how DBTL can facilitate a rapid optimization of microbial strains for production of any chemical compound of interest.