Combining metabolic modelling with machine learning accurately predicts yeast growth rate

Extended Abstract*

Christopher Culley[1], **Supreeta Vijayakumar**[2], **Guido Zampieri**[2], **Claudio Angione**[2,3]
[1]School of Electronics and Computer Science, University of Southampton, Southampton, UK
[2]Department of Computer Science and Information Systems, Teesside University, Middlesbrough, UK
[3]Healthcare Innovation Centre, Teesside University, Middlesbrough, UK
cc2u18@soton.ac.uk,{g.zampieri,s.vijayakumar,c.angione}@tees.ac.uk

ABSTRACT

New metabolic engineering techniques hold great potential for a range of bio-industrial applications. However, their practical use is hindered by the huge number of possible modifications, especially in eukaryotic organisms. To address this challenge, we present a methodology combining genome-scale metabolic modelling and machine learning to precisely predict cellular phenotypes starting from gene expression readouts. Our methodology enables the identification of candidate genetic manipulations that maximise a desired output – potentially reducing the number of *in vitro* experiments otherwise required. We apply and validate this methodology to a screen of 1,143 *Saccharomyces cerevisiae* knockout strains. Within the proposed framework, we compare different combinations of feature selection and supervised machine/deep learning approaches to identify the most effective model.

KEYWORDS

Genome-scale modelling, machine learning, deep learning, multi-omics, cellular growth, *Saccharomyces cerevisiae*.

1 INTRODUCTION

Cellular growth and gene expression are closely related in unicellular organisms, as they co-participate in mutual regulation. This relationship has yet to be fully understood, and in general predicting cellular growth following genetic manipulations is still challenging. Understanding and controlling cellular growth has important applications in biotechnology for the development of efficient cell factories, but the identification of such strains is still a complex issue [10].

We propose a novel multi-view learning framework that utilises both transcriptomics data and strain-specific metabolic fluxes to predict outputs of bio-industrial interest. To demonstrate the efficacy of this framework, we target it to predicting cellular growth of *S. cerevisiae*, one of the main eukaryotic platforms for bio-industrial production.

*Oral presentation

2 METHODS

In this work, we started from 1,143 *S. cerevisiae* gene expression (GE) profiles – our first data view – each of which are sampled from single deletion strains and are coupled with their corresponding growth rate fold change [8]. We used a genome-scale metabolic model (GSMM) of yeast metabolism [6] in conjunction with METRADE [1] – which uses gene expression to tailor reaction rate bounds – to build an equal number of strain-specific GSMMs. We next used regularised flux balance analysis (RFBA) [11] to determine reaction fluxes for the entire network by maximizing the biomass accumulation rate subject to regulatory and biochemical constraints. The solutions provide steady-state reaction rates (fluxes) for each yeast strain and every reaction in the GSMM. We used the metabolic fluxes (MF) generated in this phase as a second data view in the following prediction stage.

In the supervised learning phase, we employed the following methods: (i) support vector regression (SVR) [3]; (ii) random forest (RF) [4]; and (iii) deep neural networks (DNN). These were selected based on their suitability to build predictive models starting from high-dimensional dataset such as our transcriptomic and fluxomic profiles. We used the caret R package for SVR and RF [9], while DNN were implemented through the keras Python library [5].

Given the high dimensionality of our data, we explored whether feature selection can identify relevant genes or metabolic reactions, to build simpler and more interpretable models. We focused on three state-of-the-art techniques previously applied to omics data: (i) sparse group lasso (SGL) [12]; (ii) non-dominated sorting genetic algorithm II (NSGA-II) [7]; and (iii) iterative random forests (iRF) [2].

3 RESULTS

We developed and evaluated a computational pipeline for predicting *S. cerevisiae* growth rate from experimental and simulated omics data, which is summarised in Figure 1a. In brief, we used strain-specific GSMMs and RFBA to estimate the MF activity of 1,143 yeast mutants in log phase, starting from their GE profiles and optimising the GSMM building

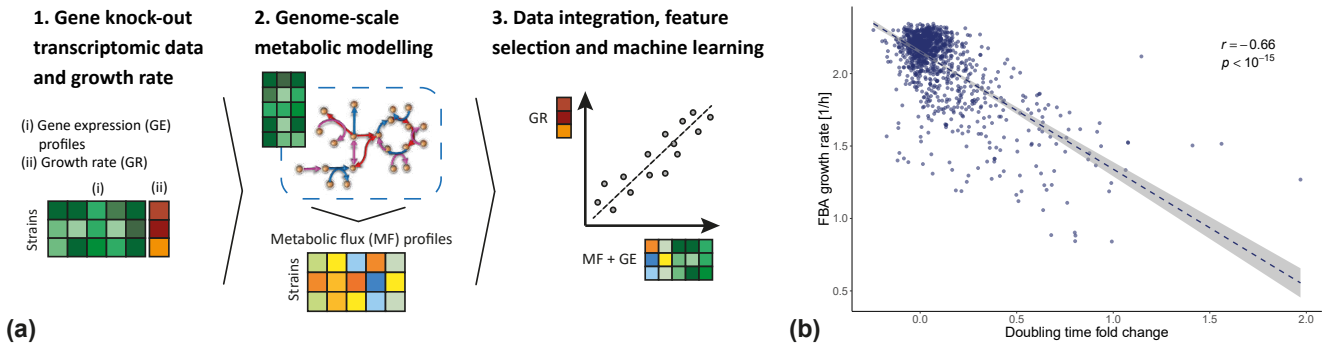

Figure 1: (a) Workflow of the proposed methodology: starting from GE profiles for different synthetic yeast strains – where the colour tones represent rates of GE (greens), target doubling time (reds) and MF (blue to orange) – we build strain-specific GSMMs from which we estimate the MF activity. Next, we build data-driven predictive models using both GE and MF information. (b) Correlation between the growth rate simulated by the strain-specific GSMMs and the relative doubling time for the same strains. This shows that the strain-specific GSMMs correctly capture the metabolic state across strains.

Table 1: Full set of accuracy scores across all dataset-method combinations tested against an unseen set of strains to determine model generalisation: mean absolute error (MAE), median absolute error (MDAE), Pearson's correlation coefficient (PCC). The final column indicates the percentage of fluxomic features (FF%) of the dataset. Here MF+GE corresponds to the full data profiles from both the gene expression and metabolic fluxes.

Dataset	Method	MAE	MDAE	PCC	FF%
MF+GE	SVR	0.080	0.054	0.845	36
MF+GE	RF	0.077	0.048	0.867	36
MF+GE	DNN	0.072	0.049	0.887	36
iRF data	SVR	**0.070**	0.048	0.886	0
iRF data	RF	0.075	0.052	0.869	0
iRF data	DNN	0.073	0.048	0.882	0
NSGA-II data	SVR	0.072	0.049	**0.889**	24
NSGA-II data	RF	0.078	**0.047**	0.843	24
NSGA-II data	DNN	0.081	0.053	0.861	24
SGL data	SVR	0.081	0.057	0.865	34
SGL data	RF	0.081	0.052	0.846	34
SGL data	DNN	0.084	0.058	0.866	34

based on the simulated growth rate (Figure 1b). Then, we built and cross-compared machine and deep learning models predicting yeast growth from integrated GE and MF information (MF+GE), with and without feature selection. In this phase, we tested SVR, RF and DNN in combination with SGL, NSGA-II and iRF. We thereby created three further datasets (SGL data, NSGA-II data and iRF data respectively) comprising the features identified by each of these approaches.

Depending on the combination of dataset and learning algorithm, we observed different trends in prediction scores. Overall, the best performing methods are SVR combined with iRF and NSGA-II, and DNN without prior feature selection. We note that in the case of SVR, feature selection can sensibly improve its prediction accuracy, while there is an opposite trend for DNN. This could suggest that effective DNN models embed non-linear relationships among genes and metabolic reactions that involve a larger set of features. Importantly, the MF variables selected allow us to mechanistically understand the factors governing cell growth and further inform potential manipulations.

4 CONCLUSIONS

Our integrative models enable the joint analysis of experimental genetic regulation patterns and knowledge-based metabolic information to predict yeast cell growth. Our results suggest that integrating multi-omics variables and metabolic modelling can improve yeast growth predictions and provide mechanistic biomarkers. Finally, our pipeline has potential applications in metabolic engineering scenarios, and can be readily extended to other hosts.

REFERENCES

[1] Claudio Angione and Pietro Lió. 2015. *Scientific reports* 5 (2015), 15147.
[2] Sumanta Basu et al. 2018. *Proceedings of the National Academy of Sciences* (2018), 201711236.
[3] Asa Ben-Hur et al. 2008. *PLoS computational biology* 4, 10 (2008), e1000173.
[4] Xi Chen and Hemant Ishwaran. 2012. *Genomics* 99, 6 (2012), 323–329.
[5] François Chollet et al. 2015. https://keras.io.
[6] Ratul Chowdhury et al. 2015. *Metabolites* 5, 4 (2015), 536–570.
[7] Kalyanmoy Deb et al. 2002. *IEEE transactions on evolutionary computation* 6, 2 (2002), 182–197.
[8] Patrick Kemmeren et al. 2014. *Cell* 157, 3 (2014), 740–752.
[9] Max Kuhn et al. 2018. https://CRAN.R-project.org/package=caret.
[10] Jiazhang Lian et al. 2018. *Metabolic Engineering* 50 (2018), 85–108.
[11] Jeffrey D Orth et al. 2010. *Nature biotechnology* 28, 3 (2010), 245.
[12] Noah Simon et al. 2013. *Journal of Computational and Graphical Statistics* 22, 2 (2013), 231–245.