Integrated Predictive Genome-Scale Models to Improve the Metabolic Re-Engineering Efficiency

[Extended Abstract]

Vishwesh V. Kulkarni
University of Warwick
School of Engineering
Coventry, CV4 7AL
V.Kulkarni@warwick.ac.uk

Lina El Menjra
University of Warwick
School of Engineering
Coventry, CV4 7AL
L.el-Menjra@warwick.ac.uk

Pablo Carbonell
University of Manchester
SYNBIOCHEM
Manchester, M1 7DN
pablo.carbonell@manchester.ac.uk

Jean-Loup Faulon
University of Manchester
SYNBIOCHEM
Manchester, M1 7DN
jean-loup.faulon@manchester.ac.uk

Introduction

One of the most common applications of metabolic circuits is to produce a desired chemical in a chassis organism, such as the *Escherichia coli* (E. coli), by importing heterologous genes encoding for the enzymes that participate in the biosynthetic pathway. Recently, an automated pipeline named *RetroPath* was developed to synthesise embedded metabolic circuits [1]. These circuits are to be embedded in *E. coli* for a wide range of applications such as regulating biomass productions, sensing specific molecules, processing specific molecules, and releasing specific molecules. In RetroPath, the available circuit design space, constrained by the set of design specifications, is searched and a set of *optimal* circuits is obtained. In this process, the basic steps are as follows:

1. **Step 1:** Define the input set of metabolites $\mathbf{S}$, the output (target metabolite set) $\mathbf{T}$, and the metabolic space, i.e., the set of all possible metabolites and chemical reactions that can be generated *in vivo*, $\mathbf{M}$;
2. **Step 2:** Define the specifications of the desired circuit;
3. **Step 3:** Compute the set of enzymes involved in at least one minimal pathway that converts $\mathbf{S}$ into $\mathbf{T}$;
4. **Step 4:** Enumerate all metabolic pathways converting $\mathbf{S}$ into $\mathbf{T}$ (see [2]); and
5. **Step 5:** Solve a nonlinear optimization problem wherein the objective function comprises the expected reaction efficiencies, inhibition effects, and perturbation effects. Here, the *flux balance analysis* (FBA) is used.

The FBA (see [4]–[6]) predicts metabolic flux distributions at a steady state by solving the linear programming problem

$$\text{maximize} \;\; w^T v \quad \text{subject to} \;\; Sv = 0 \;\; \text{and} \;\; \alpha \leq v \leq \beta,$$

where $w^T v$ denotes the biomass, v denotes the vector of metabolic fluxes, S is the stoichiometric matrix, and (α, β) are the *a priori* known bounds on the fluxes. Thus, it is inherently assumed in RetroPath that the chassis organism has reached a steady state following the insertion of the heterologous gene. As shown in [4], such an assumption of optimality may not be valid for genetically engineered knockouts and bacterial strains that were not exposed to long-term evolutionary pressure. Following [4], we show that Step 5 can indeed be executed more efficiently by assuming that the metabolic fluxes undergo a minimal redistribution with respect to the flux configuration of the wild type. This *minimization of metabolic adjustment* (MOMA) is computed by solving the quadratic programming problem

$$\text{minimize} \;\; \frac{1}{2}\|v\|^2 + v^T v^* \quad \text{s. t.} \;\; Sv = 0 \;\; \text{and} \;\; \alpha \leq v \leq \beta,$$

where v^* denotes the flux distribution predicted by the FBA.

We next show that the efficiency of RetroPath can be improved if transcriptomic data is available. This is achieved by replacing the use of FBA in Step 5 first with the PROM algorithm derived in [3] and then with an extension PROM-E derived by us. PROM refines the upper bounds and the lower bounds in the metabolic flux constraints of the FBA by using Bayesian estimation on the available transcriptomic and metabolomic datasets. Here, the probabilities on the gene-TF interactions are empirically determined using available datasets; the more the number of datasets, the better is the expected performance. Effectively, PROM implements a slight modification of the FBA using linear programming and achieves an improvement in not only the capacity to generate larger genome-scale models but also in the accu-

IWBDA 2016 Newcastle-upon-Tyrne, August 2016

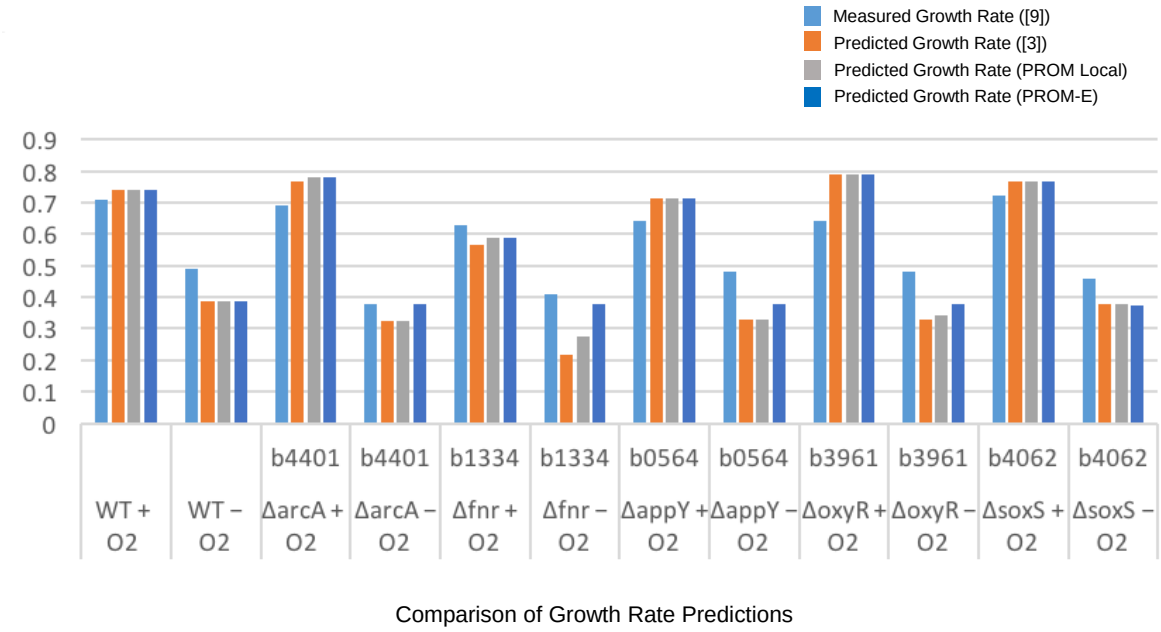

Figure 1: Our proposed algorithm PROM-E predicts the growth rates better than PROM [3] and RFBA [8] on the datasets of [8] across a variety of anaerobic conditions. For aerobic conditions, the PROM-E and PROM perform equally well and slightly better than RFBA.

racy of predicting the flux states. We then show how further improvement are obtained using our extension PROM-E.

Results

We now illustrate how our extension PROM-E of PROM estimates the flux rates more accurately than PROM on the datasets of [8] wherein growth phenotypes from *A Systematic Annotation Package* (ASAP) are predicted for community analysis of genomes database. The ASAP database has growth phenotypes of several *E. coli* gene KOs under various conditions. In [3], 15 TFs for which growth phenotypes under different 125 conditions are considered and it shown that PROM predicts the growth phenotypes more accurately than RFBA [8]. In [3], six strains with KOs of key transcriptional regulators in the oxygen response (ΔarcA, ΔappY, Δfnr, ΔoxyR, ΔsoxS, and the double KO ΔarcAΔfnr) were constructed and then the growth rates were measured in aerobic and anaerobic glucose minimal medium conditions. As Fig. 1 shows, our proposed PROM-E algorithm predicts the growth rate more accurately than PROM in anaerobic conditions and equally well in aerobic conditions. Furthermore, the standard deviation of the prediction error is significantly lower in PROM-E compared to PROM. We obtained these results in MATLAB R2015b interfaced with COBRA Toolbox 2.0 [7] and Gurobi Optimizer 6.5. As the cost of collecting omics datasets is reducing at Moore's law, we expect that our approach will soon be useful in practical contexts.

Acknowledgments

Research supported by BBSRC under grant BB/M017702/1, *Centre for Synthetic Biology of Fine and Speciality Chemicals* (SYNBIOCHEM).

1. REFERENCES

[1] Carbonell, P., Parutto, P., Baudier, C., Junot, C. and Faulon, J-L.: 'Retropath: Automated Pipeline for Embedded Metabolic Circuits', ACS Synthetic Biology, vol. 3, pp. 565–577, 2014.

[2] Carbonell, P., Fichera, D., Pandit, S. and Faulon, J-L.: 'Enumerating Metabolic Pathways for the Production of Heterologous Target Chemicals in Chassis Organisms', BMC Systems Biology, 6(10), pp. 1–18, 2012.

[3] Chandrasekaran, S. and Price, N.: 'Probabilistic Integrative Modeling of Genome-Scale Metabolic and Regulatory Networks in Escherichia Coli and Mycobacterium Tuberculosis', PNAS, 107 (41), pp. 17845–50, 2010.

[4] Segrè, D., Vitkup, D. and Church, G.: 'Analysis of Optimality in Natural and Perturbed Metabolic Networks', PNAS, 99(23), pp. 15112–17, 2002.

[5] Papoutsakis, E.: 'Equations and Calculations for Fermentations of Butyric Acid Bacteria', Biotechnology and Bioengineering, 26(2), pp. 174–187, 1984.

[6] Watson, M.: 'Metabolic Maps for the Apple II', Biochemical Society Transactions, 12(6), pp. 1093–1094, 1984.

[7] Schellenberger, J., *et al.*: 'Quantitative Prediction of Cellular Metabolism with Constraint-Based Models: The COBRA Toolbox v2.0', Nature Protocols, 6(9), pp. 1290–1307, 2011.

[8] Covert, M., Knight, E., Reed, J., Herrgard, M. and Palsson, B.: 'Integrating High-Throughput and Computational Data Elucidates Bacterial Networks', Nature, 429(6987), pp. 92–96, May 2004.