The Board of Regents of the University of Wisconsin-System

Final Report: Systems Approach to Engineering

Period of Performance: 07/01/2012 - 06/30/2019

Contract Number: DE-SC0008103

Author: Dr. Jennifer Reed

These funds were used to support the research of 8 graduate students over six years on various projects focused on using systems approaches to improve production of biofuels and biochemicals in *Synechococcus* and other microbes. Below is a description of the research that was performed, which led to 8 publications, [3-10] 1 submitted manuscript, 1 manuscript in preparation, and 2 patents [11-12].

Engineering *Synechococcus sp.* PCC 7002 for Butanol Production

A genome-scale metabolic model was used to guide engineering of the cyanobacterium *Synechococcus sp.* PCC 7002 for the production of short- to mid-chain alcohols (e.g. 2-methyl-1-butanol) that are potential biofuels. PCC 7002 is of particular interest as a cyanobacterial production strain due to its relatively rapid growth-rate and tolerance to saline and high-light conditions. Metabolic engineering strategies for improving the production of various alcohols in PCC 7002 were investigated via the metabolic model iSyp708 [6], which was previously developed by our lab. Using this modeling approach, a potential strategy for enhanced production of 2-methyl-1-butanol in PCC 7002 was successfully identified (Figure 1). This strategy hinges on coupling the overproduction of the alcohols' metabolic precursors to nitrate assimilation by rewiring PCC 7002's native NADH-cycling pathways. Specifically, the strategy calls for replacing the native nitrate and/or nitrite reductases in PCC 7002, which are ferredoxin-dependent, with heterologous enzymes that instead use NADH to carry out their reactions. This novel engineered NADH-demand in PCC 7002 is predicted to generate a metabolic pull on several upstream reactions in the production pathway for 2-methyl-1-butanol, and potentially several other related branched-chain alcohols, thereby directing increased flux toward the desired product.

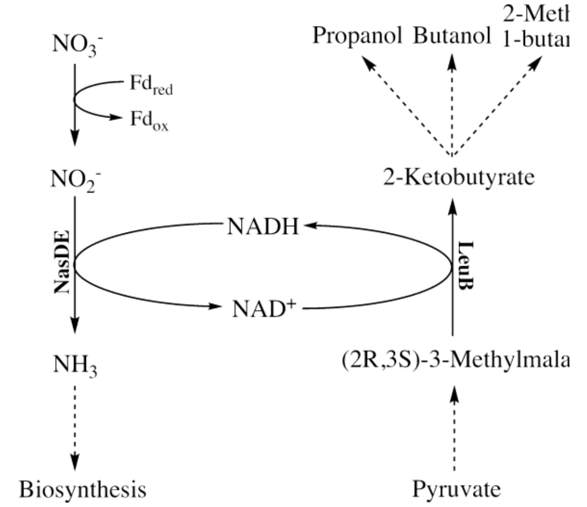

Figure 1. Engineering strategy to couple nitrate assimilation with butanol production.

To test this strategy, the native nitrite reductase gene (*nirA*) in PCC 7002 was knocked out and replaced in the same genomic locus by the nitrite reductase from *Bacillus subtilis* (*nasDE*). The nasDE enzyme has been shown to specifically use NADH as its reducing cofactor [1]. The *nirA::nasDE* strain showed a growth defect when grown with nitrate as a nitrogen source, but otherwise appeared healthy in terms of pigment content (as measured by absorption spectrum). To test whether this strain is an improved background for producing 2-methyl-1-butanol in, the branched-chain alcohol production pathway described by Shen & Liao [2] was engineered into the strain. An inducible operon containing the production pathway genes was expressed from either the chromosomal *glpK* locus or on a broad-host range RSF1010 plasmid. In both cases the engineered *nirA::nasDE* strain showed improved alcohol titers relative to a WT PCC 7002 background (Figure 2). We are currently conducting final tests to determine if expression of the alcohol pathway from other genomic loci yield improved production titers over those observed so far.

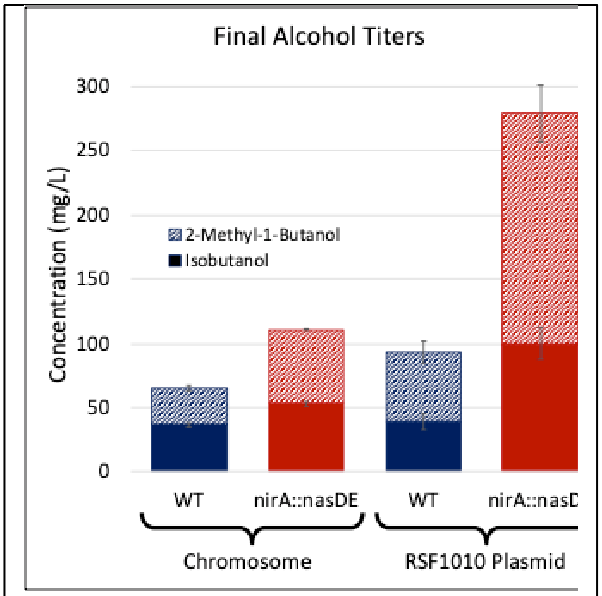

Figure 2. Alcohol production from different expression systems in wildtype (WT) and engineered PCC 7002 backgrounds.

This work demonstrates that replacing the native nitrate assimilation pathway in PCC 7002 with NADH-dependent enzymes can generate a background strain with improved production characteristics for certain classes of next-generation biofuel compounds, e.g. branched-chain alcohols. Additionally, this work further validates the utility of genome-scale metabolic models for guiding engineering studies in cyanobacterial production strains. We are completing final experiments for publication and plan to submit a manuscript describing this work later this year.

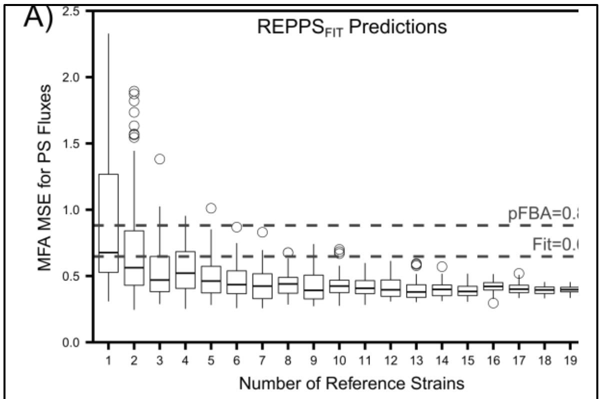

Figure 3. Sensitivity of REPPS predictions to the numbers of reference strains used. The box plots show the mean squared errors (MSE) for the parental strains estimated intracellular fluxes using either REPPS$_{FIT}$. The dashed lines indicate the MSE for the parental strains calculated directly from the pFBA and the Fit methods, which do not use any reference strain data.

Analyzing Experimental Data in the Context of Metabolic Models

Incorporating experimental data into constraint-based models can improve the quality and accuracy of the models' metabolic flux predictions. Unfortunately, routinely and easily measured experimental data such as growth rates, extracellular fluxes, transcriptomics, and even proteomics are not always sufficient to significantly improve metabolic flux predictions. We developed a new method (called REPPS) for incorporating experimental measurements of growth rates and extracellular fluxes from a set of perturbed reference strains and a parental strain to substantially improve the predicted flux distribution of the parental strain [3]. The reference strains are typically mutants derived from the parental strain. Using data from five single gene knockouts (reference strains) and the wild type strain of *Escherichia coli* (parental strain), we decreased the mean squared error of predicted central metabolic fluxes by ~47% compared to parsimonious flux balance analysis (pFBA) (Figure 3). This decrease in error further improves flux predictions for new knockout strains. Furthermore, REPPS is less sensitive to the completeness of the metabolic network than pFBA.

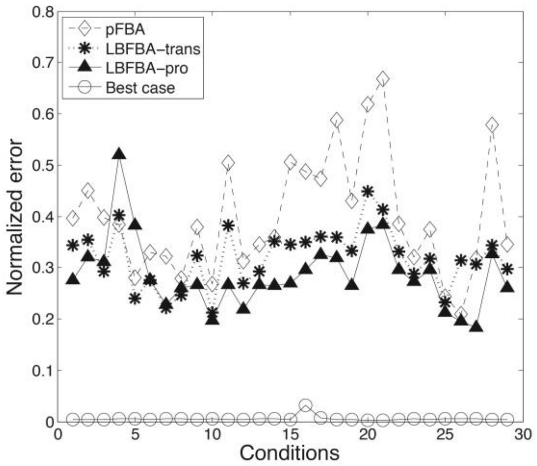

Figure 4. Simulation result for LBFBA compared with pFBA for the *E. coli* dataset. The x-axis represents the 29 conditions with measured transcriptomics, proteomics and fluxomics data. The y-axis represents the normalized flux error. For each condition, data from the 28 other conditions was used to parameterize flux bounds before LBFBA was performed on the test condition. For most cases, pFBA (hollow diamond) has a higher normalized error than LBFBA integrating transcriptomics data (star) or proteomics data (solid triangle). Integrating proteomics data was more accurate than integrating transcriptomics for most conditions. The best case (hollow circle) represents the lowest possible error achieved by fitting constraint-based models to the measured fluxes.

Transcriptomics and proteomics data have been integrated into constraint-based models to influence flux predictions. However, it has been reported recently for *E. coli* and *Saccharomyces cerevisiae*, that model predictions from parsimonious flux balance analysis (pFBA), which does not use any expression data, are as good or better than predictions from various algorithms that integrate transcriptomics or proteomics data into constraint-based models. We developed a novel constraint-based method called Linear Bound Flux Balance Analysis (LBFBA), which uses expression data (either transcriptomic or proteomic) to predict metabolic fluxes [4]. The method uses expression data to place soft constraints on individual fluxes, which can be violated. Parameters in the soft constraints are first estimated from a training expression and flux dataset before being used to predict fluxes from expression data in other conditions. We applied LBFBA to *E. coli* and *S. cerevisiae* datasets and found that LBFBA predictions were more accurate than pFBA predictions, with average normalized errors roughly half of those from pFBA (Figure 4). For the first time, we demonstrate a computational method that integrates expression data into constraint-based models and improves quantitative flux predictions over pFBA.

Analyzing Strategies to Produce Biofuels and Bioproducts

Metabolic engineering uses microorganisms to synthesize chemicals from renewable resources. Given the thousands of known metabolites, it is unclear what valuable chemicals could be produced by a microorganism and what native and heterologous reactions are needed for their synthesis. To answer these questions, a systematic computational assessment of *Escherichia coli*'s potential ability to produce different chemicals was performed using an integrated metabolic model that included native *E.coli* reactions and known heterologous reactions [5]. By adding heterologous reactions, a total of 1,777 non-native products could theoretically be produced in *E. coli* under glucose minimal medium conditions, of which 279 non-native products have commercial applications. Synthesis pathways involving native and heterologous reactions were identified from eight central metabolic precursors to the 279 non-native commercial products. These pathways were used to evaluate the dependence on, and diversity of, native and heterologous reactions to produce each non-native commercial product, as well as to identify each product's closest central metabolic precursor. Analysis of the synthesis pathways (with 5 or fewer reaction steps) to non-native commercial products revealed that isopentenyl diphosphate, pyruvate, and oxaloacetate are the closest central metabolic precursors to the most non-native commercial products. Additionally, 4-hydroxybenzoate, tyrosine, and phenylalanine were found to be common precursors to a large number of non-native commercial products. Strains capable of producing high levels of these three central metabolites could be further engineered to create strains capable of producing a variety of commercial non-native chemicals.

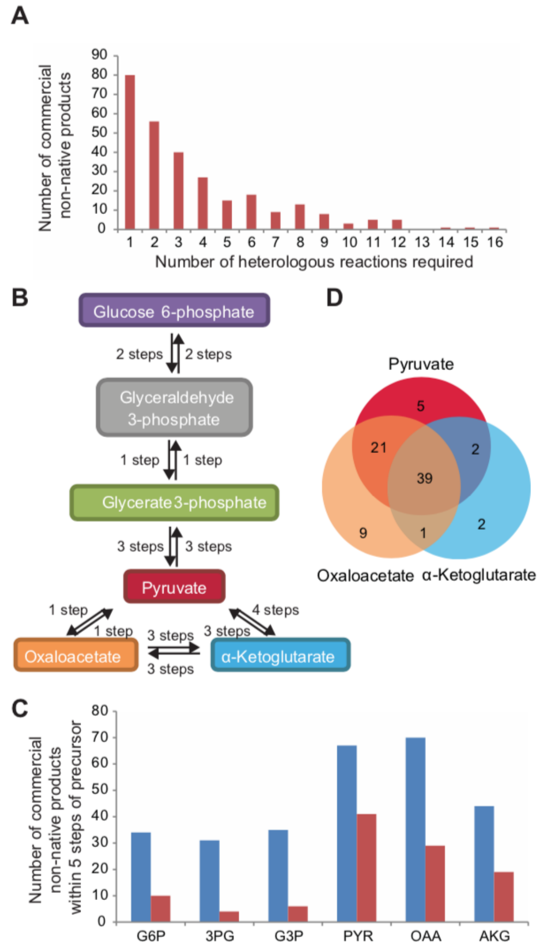

Figure 5. (A) Number of heterologous reactions needed to make the different 279 non-native products. (B) Precursors analyzed in this work. (C). The number of products derived from each of the precursors within five steps is indicated by the blue columns. The red columns indicate the number of products that are closest to each precursor. (D) Venn diagram of the non-native products that can be produced from three precursors within five reaction steps.

We have additionally extended this analysis to look at other organisms and feedstocks. We expanded our metabolic model of *Synechococcus* sp. PCC 7002 (iSYP708) [6] and used it and models of *E. coli* and *S. cerevisiae* to compare maximum theoretical yields on either acetate, methane or methanol. This work was done in collaboration with Brian Pfleger's research group. We found that the feedstock costs needed to make a fixed amount of products was lower for methane than glucose and that methanol in most scenarios was more expensive than glucose [7].

Application of Active and Machine Learning for Metabolic Engineering

Many computational and experimental approaches exist to metabolically engineer strains to produce more of a desired chemical. Computational approaches typically rely on detailed mechanistic models (e.g., kinetic/stoichiometric models of metabolism)—requiring many time- and cost-expensive experimental datasets for their parameterization—while experimental methods may require creating large mutant libraries to explore the design space, and then screening (if possible) and/or selecting (if possible) for the few mutants with desired behaviors. To address these limitations, we developed an active and machine learning approach (ActiveOpt) to intelligently guide experiments to arrive at an optimal phenotype both quickly and with minimal and easily measured datasets. ActiveOpt was applied to two separate metabolic engineering case studies that improved valine yields and neurosporene productivity in *Escherichia coli.* The first case study was one in which we used computational models to design strains of *E. coli* to produce pyruvate and then used combinations of different plasmids which express the enzymes needed to convert pyruvate into valine using different ribosome binding sites (RBSs). These same enzymes have also been used to produce various forms of butanol. Our best strain achieved an elemental carbon yield of 45% (or 54.7% of the maximum theoretical (MT) valine yield from glucose and acetate) in a defined minimal medium—the highest carbon yield reported in *E. coli*. The second case used data from a previously published study where RBS strength was varied for three genes involved in the biosynthesis of neurosporene.

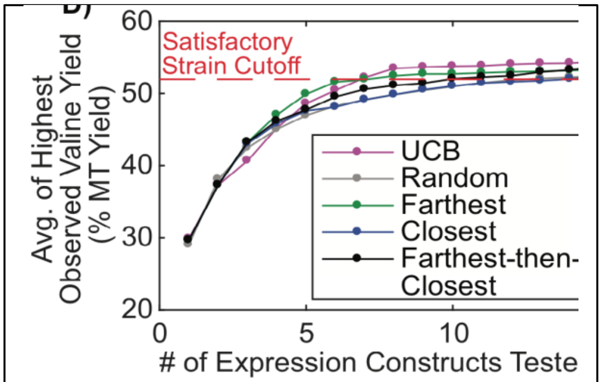

Figure 6. The average from the 89 ActiveOpt (using different objectives: randomly chosen, experiment farthest from the classifier, closest to the classifier, or alternating between farthest and closest to the classifier) or Upper Confidence Bound (UCB) runs of the highest observed % valine yield is plotted as a function of the number of total experiments performed. A total of 89 experiments were performed in the case study.

Using our valine dataset, we found that machine learning models using linear classifiers could predict whether a set of RBSs would result in high or low valine yield. From a leave-one-out cross validation analysis we found the precision and recall was 0.80 and 0.89, respectively. When the number of experiments used to train the classifier was reduced from 88 to ~11, the average precision was 0.72 and recall 0.76. We subsequently developed an active learning approach (ActiveOpt) where starting from two experimental results a classifier was built to identify another experiment predicted to have high yield, and then those experiments were used to re-train the classifier to choose another experiment (and so on). In both the cases, ActiveOpt identified the best performing strain in fewer experiments than the case studies which did not use machine learning approaches to optimize strains. This work demonstrates that machine and active learning approaches can greatly facilitate metabolic engineering efforts to rapidly achieve metabolic engineering objectives. A manuscript describing this work has been submitted.

Review Articles and Commentaries

We published several review articles and commentaries that describe the application of metabolic models for metabolic engineering [8], the integration of regulatory models with metabolic models [9], and the application of metabolic models to study and engineer microbial communities (i.e., microbiomes) [10].

Patents

This work led to two patents. The first patent describes strains of *E. coli* that were designed *in silico* to produce pyruvate [11]. We generated four strains experimentally and found they were capable of converting glucose into pyruvate with up to 95% of the maximum theoretical yield. These strains were subsequently given to a U.S. company who has been evaluating their use as a background to produce other chemicals derived from pyruvate. Our second patent [12] covers the computational method for predicting flux distributions by integrating mutant phenotyping data, which is described above.

References

1. Nakano MM, Hoffmann T, Zhu Y, and Jahn D. Nitrogen and Oxygen Regulation of *Bacillus subtilis nasDEF* Encoding NADH-Dependent Nitrite Reductase by TnrA and ResDE. *Journal of Bacteriology*, 180(20), 5344-5350 (1998).
2. Shen C and Liao J. "Photosynthetic production of 2-methyl-1-butanol from CO2 in cyanobacterium Synechococcus elongatus PCC 7942 and characterization of the native acetohydroxyacid synthase". *Energy Environ. Sci.*, 5(11), 9574-9583 (2012).
3. Long MR and JL Reed. Improving Flux Predictions by Integrating Data from Multiple Strains. *Bioinformatics*. 33(6):893-900 (2016).
4. Tian M and JL Reed. Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. *Bioinformatics*. DOI:10.1093/bioinformatics/bty445 (2018).
5. Zhang X, Tervo CJ, and JL Reed. Metabolic Assessment of *E. coli* as a Biofactory for Commercial Products. *Metabolic Engineering*. 35:64-74 (2016).
6. Vu TT, Hill EA, Kucek LA, Konopka AE, Beliaev AS, and JL Reed. Computational evaluation of *Synechococcus* sp. PCC 7002 metabolism for chemical production. *Biotechnology Journal*, 8(5):619-30 (2013).
7. Comer AD, Long MR, Reed JL, and BF Pfleger. Flux Balance Analysis Indicates that Methane Is the Lowest Cost Feedstock for Microbial Cell Factories. *Metabolic Engineering Communications.* 5:26-33. (2017).
8. Long MR, Ong WK, and JL Reed. Computational Methods in Metabolic Engineering for Strain Design. *Current Opinion in Biotechnology*. 34:135-141 (2015).
9. Kim J and JL Reed. Refining Metabolic Models and Accounting for Regulatory Effects. *Current Opinion in Biotechnology*. 29:34-38 (2014).
10. JL Reed. Genome-Scale Metabolic Modeling and Its Application to Microbial Communities, in The Chemistry of Microbiomes: Proceedings of a Seminar Series. National Academies of Sciences, Engineering, and Medicine. 2017. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/24751.
11. Zhang X and JL Reed. MICROORGANISMS AND METHODS FOR PRODUCING PYRUVATE, ETHANOL, AND OTHER COMPOUNDS. U.S. Patent (10,246,725). September 9, 2015.
12. Long MR and JL Reed. SYSTEMS AND METHODS FOR DETERMINING FLUX DISTRIBUTION. U.S. Patent Application. U.S. Patent (10,573,404). February 25, 2020.