Application of machine learning and active learning to enhance chemical yields in microbes

Prashant Kumar[1]* (pkumar29@wisc.edu), Paul A. Adamczyk[1,2], Xiaolin Zhang[1], Parameswaran Ramanathan[3], and **Jennifer L. Reed**[1,2]
[1]Department of Chemical and Biological Engineering, UW-Madison, Madison, WI 53706
[2]DOE Great Lakes Bioenergy Research Center, Madison, WI 53706
[3]Department of Electrical and Computer Engineering, UW-Madison, Madison, WI 53706

Project Goals: Cyanobacteria offer a promising route for directly converting solar energy and CO_2 into biofuels. The objectives of this research are to integrate modeling and experimental approaches to guide development of a butanol producing cyanobacterium, *Synechococcus* sp. PCC 7002. New computational approaches will be developed to facilitate these efforts which will (1) design experiments and analyze their results, and (2) identify genetic engineering strategies for improving butanol production in *Synechococcus* sp. PCC 7002. Experiments will subsequently be performed to construct and analyze *Synechococcus* sp. PCC 7002 strains engineered for butanol production. The developed approaches will be systematically applied to suggest genetic engineering strategies for improving production of a variety of biofuels in five other microorganisms. This research will support the U.S. Department of Energy's mission for developing renewable ways of producing advanced biofuels.

Renewable sources of transportation fuels are needed to reduce the amount of oil used to satisfy transportation energy needs in the U.S. and to alleviate our dependence on foreign sources of oil. Microbes can be used to produce a wide variety of liquid biofuels including: ethanol, butanol, isobutanol, isoprene, hydrogen, and alkanes. Cyanobacteria offer an alternative route for converting solar energy and CO_2 into biofuels, without the need for using lignocellulosic biomass as an intermediate. The biofuel production capabilities of microbes can be improved through metabolic engineering, where metabolic and regulatory processes are adjusted using targeted genetic manipulations. Traditionally, metabolic engineering strategies are found through manual inspection of metabolic pathways, where enzymes involved in biosynthesis are overexpressed or added, competing pathways are eliminated, and the performance of resulting strains are evaluated. However, such manual approaches cannot predict the effects that these changes will have on metabolism and the enzyme levels needed to optimize flux through a metabolic pathway.

Efforts have been made to develop computational methods to study metabolic and regulatory networks of microbes and identify the genetic interventions needed to produce desired high-value metabolite(s) from low-cost substrates. Genome-scale constraint-based metabolic models rely on information on reaction stoichiometry to identify reaction or gene deletion and addition strategies to enhance product yield. However, these models cannot predict how changes to enzyme levels will impact metabolic fluxes. In contrast, smaller kinetic models can be used to suggest additional strategies based on changes in enzyme levels and/or kinetic properties, but these kinetic models need a lot of experimental data (e.g., proteomic, metabolomic, and fluxomic data) to parameterize them. Hence, there is a need for computational methods that can predict expression levels needed to achieve metabolic engineering goals with limited amounts of experimental data and no kinetic details about the system.

We developed an active learning framework called **ActiveOpt** to design expression constructs for a metabolic pathway of interest. ActiveOpt does not need a detailed kinetic model and instead uses a linear Support Vector Machine (SVM) classifier to predict product yields or productivities (either high or low) from ribosome binding site strengths estimated by the RBS Calculator [1]. ActiveOpt initially trains a SVM classifier from a few experiments, where RBSs in gene expression constructs are varied and product yields are measured and labeled (as high/low yield), and then proposes subsequent experiments to be conducted. ActiveOpt, with relatively little experimental data and no mechanistic or kinetic details of the pathway, can be used to design experiments to achieve high biochemical yields in a small number of experiments.

ActiveOpt was tested on two separate datasets: (1) a newly generated valine yield dataset and (2) a published neurosporene productivity dataset [2]. The valine dataset included 91 experiments, in which two plasmids, that express nine valine biosynthesis and exporter genes (*ilvBNIHCDE* and *ygaZH*) with varying RBS strengths, were transformed into *Escherichia coli* and valine yields were measured in glucose+acetate minimal medium. A leave-one-out cross validation showed that SVM classifiers built from this dataset have high precision (75%) and recall (87%). Starting with just a few of the 91 possible experiments, ActiveOpt could identify expression constructs resulting in at least 95% of the highest measured valine yield (across all conducted 91 experiments) in a small number of experiments (typically <7) and identify the genes whose RBS strengths significantly affect valine yield. Further, ActiveOpt was used to propose four new experiments (beyond the original 91 valine experiments) that were predicted to have high yields. Valine yields in those four new experiments were found to be high, with one strain having 53.38% of the maximum theoretical yield, which is close to the best yield found in the 91 previously conducted experiments (54.70%). ActiveOpt was also tested on a previously published neurosporene dataset, and the algorithm could again identify the expression constructs with high productivity in less than 10 experiments as compared to the 101 experiments conducted in the original study.

These results show that ActiveOpt can efficiently design gene expression constructs that lead to high chemical yield in organisms in very small numbers of experiments. It can also identify the genes whose expression (as predicted by RBS strengths) significantly influence biochemical production. Our next step is to use ActiveOpt to design experiments to identify strains with high butanol yield in the cyanobacterium *Synechococcus* sp. PCC 7002.

References

1. Salis, H.M., E.A. Mirsky, and C.A. Voigt, *Automated design of synthetic ribosome binding sites to control protein expression.* Nat Biotechnol, 2009. **27**(10): p. 946-50.
2. Farasat, I., et al., *Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria.* Mol Syst Biol, 2014. **10**: p. 731.

Funding statement. This work was supported by the Office of Science (BER), U.S. Department of Energy (DE-SC0008103), the W.M. Keck Foundation, and the College of Engineering at the University of Wisconsin-Madison.