**Optimal Experimental Design (OED) of Biological Systems**

**Francis J. Alexander,[1,*] (falexander@bnl.gov),** Ian Blaby[2], Edward Dougherty[3], Maria Soto[2], Xiaoning Qian[3], and Byung-Jun Yoon[1,3]

[1]Brookhaven National Laboratory, Upton, NY; [2]Lawrence Berkeley National Laboratory, Berkeley, CA; [3]Texas A&M University, College Station, TX.

**Project Goals:**

The project's overall goal is to develop optimal experimental campaigns to achieve a particular objective, namely metabolite yield alteration. The optimal experiments will be designed by quantifying the cost of uncertainty in the current predictive model—a transcriptional regulatory network model that regulates metabolism—and selecting the experiments that are expected to maximally reduce the model uncertainty that affects the attainment of the aforementioned objective. This approach will serve as a proof of principle, demonstrating the significant potential of computationally guided biology in areas directly relevant to BER's missions.

Any future bio-economy likely will include a spectrum of engineered organisms. As sources of economically valuable products, prokaryotes offer many beneficial attributes (e.g., rapid growth and diverse metabolic capabilities), including the production of multiple value-added products that can offset the cost of bioenergy products. However, the biological complexity and diversity of these organisms impede development of genome-wide engineering strategies. Lack of knowledge about proteins that participate in or regulate given processes presents a barrier to predictive engineering. Consequently, despite recent molecular advances with Clustered Regularly Interspaced Short Palindromic Repeats associated (CRISPR Cas)-based tools, knowing what and how to engineer organisms to achieve a desired goal remains a bottleneck, resulting in many genome engineering projects that do not meet expected outcomes. Even with simple organisms such as prokaryotes, knowledge is highly uncertain and incomplete. Understanding how these systems respond to an intervention is even less exhaustive. Thus, such paucity of knowledge regarding complex biological systems requires robust optimization strategies.

To accelerate strain improvement strategies, significant paradigm shifts in life science research are needed, particularly a move toward tool development for optimizing and controlling highly uncertain systems. While computing infrastructures can assist bench scientists in designing experiments that can effectively fill knowledge gaps in biological networks, designing and implementing these infrastructures remain significant tasks. One valid and successful approach is to understand each component (regulation, transport, biochemical pathway, etc.) of the system via experiments that capture specific data points. However, data derived from biological experiments can be as complex as the organisms from which said information is collected. Data sets and experimental results are multifaceted, multidimensional, and originate from different sources (i.e., organisms), and interpretation often requires understanding and analyzing multiple fields of research. Consequently, engineered organisms may exhibit unanticipated outcomes. For example,

a system designed for elevated levels of a given metabolite may not differ significantly from the parent strain because of unknown pathway branch points or regulation (i.e., "metabolic buffering capacity").

Addressing these challenges requires a probabilistic framework for integrative modeling of heterogeneous omics data (especially transcriptomics and metabolomics data), quantification of the uncertainty affecting the objective (i.e., strain improvement to optimize metabolite yield), and designing the optimal experiment that can effectively reduce this objective-based uncertainty. The MOCU (mean objective cost of uncertainty) concept and the MOCU-based OED framework proposed in this project are well suited for overcoming these challenges.

This project exploits the team's collective expertise in systems biology, high-performance computing, mathematical modeling, and control of uncertain complex systems to: (1) take advantage of existing models and data, even when there is uncertainty, to robustly predict optimal experiments; and (2) employ an OED framework to optimize the outcome in an efficient manner (i.e., fewer experiments and less guesswork), where optimization is achieved by optimally (most favorably) improving knowledge about the model (or the microbial system represented by the model) relevant to the objective. To achieve this goal, the team is using a multidisciplinary approach involving two interlinked aims. Aim 1 identifies, adapt, and implement the necessary algorithms to make OED applicable to biological problems by reducing the cost of uncertainty in cellular metabolism. This is achieved by inferring the transcriptional regulatory network (TRN) from *E. coli* gene expression compendia using a Bayesian network (BN) and employing the gene expression values predicted by the BN under various control actions/conditions to infer their impacts on the metabolite yield. The metabolic outcomes will then be predicted through flux balance analysis (FBA) with proper constraints on the metabolic pathways regulated by the TRN modeled by the BN. Aim 2 will define, execute, and iterate genome-scale engineering approaches guided by MOCU and MOCU-based OED using the aforementioned models in Aim 1. Realization of Aim 1 will yield robust predictions, even in the face of uncertainty (i.e., incomplete information), informing the model. In Aim 2, the necessary genome manipulations guided by these predictions will be performed to quantitatively assess their success and iterate.

*This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER).*