Optimal Experimental Design (OED) of Biological Systems

Francis J. Alexander,[1,*] (falexander@bnl.gov), Ian Blaby[2], Edward Dougherty[3], Puhua Niu, 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.

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. Data derived from biological experiments 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.

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.

We have developed a new flexible analysis pipeline, **TRIMER**: Transcription Regulation Integrated with MEtabolic Regulation, enabling integrative systems modeling of transcription factor (TF) regulated metabolism. In this workflow, we adopt a Bayesian network (BN) inferred from large-scale gene-expression compendia, rather TF-gene conditional probabilities, which enables the incorporation of prior relational knowledge when modeling TF regulations that affect metabolism. Consequently, our modeling framework can take advantage of pathway knowledge and quantify the impact of extending our current knowledge regarding transcription regulation via future experiments on the objective (i.e., optimize metabolite yield). Based on the constructed BN, we can infer the probabilities of gene states of interest, and consequently predict genome-scale metabolic fluxes of mutants by TF knockouts. Additionally, we have developed a simulation framework to mimic the TF-regulated metabolic network, which is capable of generating both gene expression states and metabolic fluxes, thereby providing a fair evaluation platform for benchmarking models and predictions. Here, we present progress on these computational pipelines as well as their applicability to both simulated and actual experimental data.

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