Data-driven design and engineering of biomolecules: mRNA and DNA

Sanjan TP Gupta[1,2*](sgupta78@wisc.edu), Gina C Gordon[1,3], Xiangyang Liu[2,4], Srivatsan Raman[2,4], Brian F Pfleger[1,3], and **Jennifer L Reed**[1,2]

[1]Department of Chemical and Biological Engineering, UW Madison, WI – 53706; [2]Great Lakes Bioenergy Research Center, Madison, WI – 53726; [3]Microbiology Doctoral Training Program, University of Wisconsin-Madison, Madison, WI - 53706 [4]Department of Biochemistry, UW Madison, WI – 53706

Project Goals:

This project aims to employ machine learning principles to understand the design principles of biomolecular function by building empirical models using sequence-activity measurements, and in turn, use these models to accelerate the design-build-test cycle.

Abstract:

The advent of high-throughput technologies coupled with dropping costs of sequencing has led to the generation of new biological datasets paving the way for data-driven engineering of biomolecules. The current work describes case studies of engineering two different classes of biomolecules (mRNA and DNA) with applications to metabolic engineering and synthetic biology.

In the first study, ML models were built for predicting the mRNA half-lives in cyanobacteria - a photosynthetic microbe that can convert CO_2 into a variety of chemicals. A set of 28 sequence and structure based features (such as GC content, predicted RBS strength, and minimum free energy based on RNA folding) were compiled for the 3,238 genes found in *Synechococcus* sp. PCC 7002. Half-live values were measured for the corresponding mRNA transcripts using a rifampicin based transcription arrest assay and used as the target variable to be predicted based on the feature values. Analyzing the importance of various features used for building the model revealed that stable transcripts have higher normalized expression levels, higher translation rates, and are less likely to be found in an operon. Later, counts of 3 to 8 lettered sequence motifs in the 5' and 3' UTRs (untranslated regions) were used as features to build half-life predictors using a variant of random forest approach. These models were able to predict the half-lives accurately (with a spearman rank coefficient of 0.88 under 10-fold cross validation) and helped reveal of set of putative sequence motifs that could be used to enhance the stability of any gene of interest.

The second study looks at building ML models to predict the inducibility of genes under the control of *de novo* promoters with potential applications to genetic circuit design and development of biosensors for detecting intra-cellular metabolites. Using one-hot encoding and support vector regression, quantitative models were built to accurately predict the fold-induction ratios for a given operator sequence corresponding to three different prokaryotic transcription factors – PmeR, TtgR,

and NalC. These models helped reveal general insights into sequence determinants of promoter activity.

Insights and recommendations generated from these quantitative biology studies will in turn be beneficial for accelerating the bioengineering pipeline as well as improving the success rate for future rounds of biomolecular design.

References

1. Liu X *et al* (2019) De novo design of programmable inducible promoters. Nucleic Acids Research doi:10.1093/nar/gkz772
2. Gordon GC *et al* Genome-wide analysis of cyanobacteria RNA decay reveals highly stable transcripts encoding photosynthesis genes. (In revision)

This work was funded by the U.S. Department of Energy Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-SC0018409).