Machine learning techniques help accelerate enzyme engineering: a case study with glycoside hydrolases

Sanjan TP Gupta [1,2*] (sgupta78@wisc.edu), Evan Glasgow[2,3], Brian G Fox[2,3], Parmesh Ramanathan[4], 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]Department of Biochemistry, UW Madison, WI – 53706; [4]Department of Electrical and Computer Engineering, UW-Madison, WI – 53706

http://reedlab.che.wisc.edu

Project Goals:

The field of protein engineering primarily aims at improving the functional properties of a protein such as thermo-stability, binding affinity, and/or catalytic activity and has become a vital step in engineering industrial enzymes. This project aims to accelerate the protein engineering workflow using machine learning based approaches. In this regard, the broad goals of the project are:

i) To build an empirical model of protein fitness landscape to accurately predict the functional properties of an enzyme based on its protein sequence
ii) To intelligently navigate the above protein fitness landscape to design novel synthetic proteins with superior functional properties.

Abstract:

In this work, we have developed a machine learning (ML) based approach called 'MLProScape' to build an accurate model of the protein fitness landscape (ProScape) and then, use this model to design synthetic protein designs with superior functional properties. Unlike approaches using directed evolution, which requires experimentally screening millions of protein variants, our method requires substantially lesser number of variants (on the order of tens to hundreds) to be screened experimentally, owing to the power of statistical inference. MLProScape consists of three major steps - i) numerically encode a protein sequence using amino acid based physio-chemical properties as features, ii) build a machine learning based model using a subset of the most highly informative features, and iii) identify synthetic designs with improved characteristics.

To demonstrate our proposed workflow, we applied MLProScape to engineer the catalytic activity of glycoside hydrolases (enzymes that can break down cellulosic and hemi-cellulosic polysaccharides). We used experimentally measured specific activities for a diverse set of glycoside hydrolases to train the ML statistical models. The resulting elastic net regression based models have a high predictive power (with correlation coefficient and R^2 values as high as 0.896 and 0.714, respectively, between the predicted and experimentally measured specific activities

under 5-fold cross validation) surpassing previously published ML based enzyme engineering studies [1, 2]. Moreover, the use of position specific features helps us identify amino acid positions distal to the active site that might play a key role in modulating the activity level. Lastly, this ML based workflow is capable of modeling complex design criteria, such as optimizing the protein sequence for hydrolyzing multiple substrates simultaneously, as well as, to account for other desirable traits such as high stability and better *in vivo* expression levels.

References/Publications:

1. Fox RJ, Davis SC, Mundorff EC, Newman LM, Gavrilovic V, Ma SK, et al. Improving catalytic function by ProSAR-driven enzyme evolution. Nat Biotechnol. 2007; 25: 338–344. doi:10.1038/nbt1286
2. Srinivasulu Y, Wang J-R, Hsu K-T, Tsai M-J, Charoenkwan P, Huang W-L, et al. Characterizing informative sequence descriptors and predicting binding affinities of heterodimeric protein complexes. BMC Bioinformatics. 2015; 16: S14. doi:10.1186/1471-2105-16-S18-S14
3. Gupta STP, Glasgow E, Fox BG, Ramanathan P, Reed JL. MLProScape: Machine learning guided approach for engineering enzymes faster. (*In preparation*)

Funding statement:

This work was funded by the U.S. Department of Energy Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494).