A DATA-DRIVEN APPROACH FOR EXPLOITING ENZYME PROMISCUITY AS A MEANS TO PREDICT NOVEL BIOCHEMICAL REACTIONS

Sanjan TP Gupta, Department of Chemical and Biological Engineering, UW-Madison; Great Lakes Bioenergy Research Center, USA
sgupta78@wisc.edu
Parameswaran Ramanathan, Department of Electrical and Computer Engineering, USA
Jennifer L. Reed, Department of Chemical and Biological Engineering, UW-Madison, USA
Great Lakes Bioenergy Research Center, USA

Key Words: machine learning, chemo-informatics, novel reactions, synthetic biology

Systems metabolic engineering has been widely used to produce chemicals of high commercial value from low cost substrates. But this process has challenges for some applications, such as harnessing lignocellulosic biomass for biofuel and biochemical production, due to our limited metabolic knowledgebase. With current advances in protein engineering, it is possible to exploit substrate promiscuity of enzymes to enable novel biochemical reactions. Nevertheless, performing experiments to determine what substrates an enzyme can act on can be time consuming and it is not always clear what potential substrates to test. So, the current work aims to employ machine learning approaches for identifying novel substrates and in turn, predicting novel reactions that are more promising than the putative reactions predicted simply based on compound similarity measures (e.g., Tanimoto coefficient). A highly accurate (up to 88.3%) machine learning model was developed to identify candidate substrates for *alcohol dehydrogenase* (ADH) using a dataset consisting of 23 metabolites (with 8 of them being known positives) and 46 chemo-informatics based molecular descriptors (e.g., topology, stereochemistry, and electronic features). In addition, support vector regression proved to be a useful method for estimating enzyme kinetics (characterized by Michaelis-Menten constants, K_m and V_{max}) for a variety of oxidoreductases that are typically found in biofuel biosynthesis pathways. Such machine learning methods can be applied to other classes of enzymes and hence, used as a tool to expand the knowledgebase of metabolic reactions paving the way for next generation of metabolic/ pathway engineering.

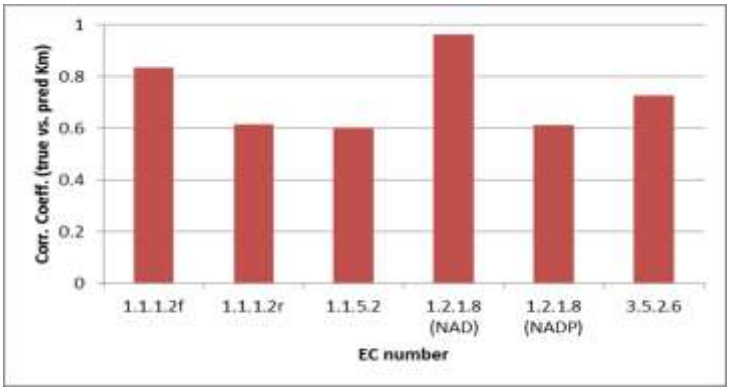

Figure 1 – Correlation coefficient b/w true and model predicted K_M values for different enzyme-substrate datasets [f - forward reaction; r - reverse reaction]