Abstract

In pharmaceutical research and development, enzymes play an important role in the synthesis of drugs and drug-related molecules. For higher efficiency and increased production, it is important to optimize the yield of these enzymes, a task often addressed by protein engineering and design. This process of enzyme optimization however can become tedious with the vast options of mutations for each single protein.

To improve the process of enzyme optimization, sequence homology and machine learning methods can be used. These greatly reduce the manual effort of protein redesign and can assist in finding the most fit enzyme for the given task, increasing the efficiency of the overall drug development pipeline.

With this aim in mind this thesis summarizes a selection of existing methods and their possible application to enzyme optimization. Testing two predictive models with varying complexity on 4 datasets in an attempt to optimize absorption, enantioselectivity, or thermostability found only a modest correlation between actual target values and their predicted values: mean Pearson's R 0.20775 and 0.5188. Comparing probability patterns of protein sequence embeddings led to a 0.815 correlation score with the BLOSUM62 substitution matrix, confirming the language model's intuition about natural frequency of different types of mutations.

While the results of the predictive models are not promising for a significant increase in efficiency with help of computational approaches, the models presented leave room for improvement and hint at a possible increase in predictive performance by fine-tuning the sequence representation for the task at hand.

Keywords: enzyme optimization, directed evolution, protein redesign, pharmaceutics, machine learning, sequence homology