

Applications of Machine Learning in Enzyme Design

Jonah Nichols

May 18, 2022

1 Methods

1.1 Data Acquisition

RheaDB was utilized to find ester hydrolysis reactions according to the Enzyme Classification (EC) standards. From selected reactions, Rhea’s cross references were utilized to obtain sequences for the various chemical reactions from UniprotKB. For every sequence-reaction pair, the Morgan Fingerprint Difference of the reaction was created using RDKit’s built-in fingerprinting function. Due to a lack of known directional data, unknown directional enzymes were assumed to catalyse the left-to-right version of the reaction. Next the fingerprint was converted into a PyTorch tensor for import into the machine learning algorithm.

For every sequence, a one-hot encoding of the sequence was built to create a suitable output for the reaction. These sequences were then converted into PyTorch tensors for input into the machine learning algorithm paired up with its corresponding reaction. These pairs were split randomly into a ratio of 75% training and 25% test datasets

1.2 Neural Network Design

Given inexperience in the subject, the neural network was designed to a very low degree of confidence. The input layer was the reaction fingerprint previously generated. Three

repetitions of linear layers followed by ReLU activation layers expanded the fingerprint out to 7192 bits, corresponding to the longest sequence found. A final linear layer presented the output of the model.

1.3 Training and Analysis

For training, the inputs and outputs were batched to 64 samples and shuffled to limit overfitting. The L1 Loss function was used to determine the error of the model. Stochastic gradient descent was utilized to optimize the model. The training sequence was run for 100 epochs to get an optimized model.

Due to time constraints, the model’s learning capabilities and pattern recognition were not determined.

2 Results

Of the 9890 sequences queried from UniprotKB with the enzyme class of 3.1.x.x, 19084 reaction-sequence pairs were generated and used for training in the machine learning model.

In a visual examination of the output the neural network gave, it became apparent that the neural network had not learned much of the data. The output given was to be a one-hot encoded sequence in number form, to be converted back to a proper sequence. Unfortunately the output received was a sequence of numbers between 12-14 and then a gradual descent to 7’s, presumably the ”pad” character created to sync

3 Discussion

Most likely

While the model has unfortunately not succeeded in learning anything meaningful, this work does need to be continued due to both the commercial and ecological value of note.