Enzyme Classification using Machine Learning Techniques

Kenny Workman, Zhe Ji, Pino K. Andrew Snider

Introduction

The use of machine learning and statistical analysis to classify enzyme function based on observed structure and properties is an emerging area of interest. The demand for *in silico* modeling of biochemical processes, as a tool for both therapeutic development and synthetic biology, places an emphasis on extracting catalytic behavior of unknown proteins to develop computational predictors of dogmatic biological behavior. Such models have the ability to discover new enzymes from protein databases and reveal certain structural components that are key to understanding unknown catalytic behavior. Many efforts to spearhead a proof of concept model are well established in the literature. Gupta and colleagues achieved 66.39% accuracy with an artificial neural network predicting on 45 sample observations ranging from mass to peptide sequence, and a 86.49% accuracy with a discrete, binary decision tree called a C5.0 (Gupta 16). Cai and colleagues similarly built several machine learning models to predict enzymatic function for a given reaction, achieving over 90% accuracy with neural networks, decision trees, and random forest prediction (Cai 1175). These models demonstrate the proficiency of predictive algorithms, with a special emphasis on deep learning and decision trees, to identify emergent patterns from vast data sets generated by the three dimensional structure and chemical properties of catalytic proteins.The results presented were able to come close to the benchmarks established in the literature by using recurrent analysis on peptide sequences; a method to predict catalytic behavior that exploits the conserved evolutionary homology of protein structure. Various attempts to analyze structural data and molecular properties with convolutional techniques resulted in similar accuracy as the neural networks described previously, almost certainly due to data pipeline bottlenecks and hardware limitations that will be explored.

Methods

Two distinct vector spaces were prepared to train the convolutional neural networks - a structural space with active-site and cofactor atom information, and a whole-structure residue space with amino acid information. Both data pipelines were designed to evaluate the efficacy of their respective observations in predicting enzymatic function, however in the end, variation in model performance between the two vector spaces was rather minute. Structural enzyme tensors were constructed in arrays of dimension 64 x 64 x 64 x 18. Pocket atoms were extracted to fill a volumetric cube of length 64, and 18 channels were used to identify structural function of atoms (ie. catalytic, binding, etc.) and the atom type. A similar volumetric tensor was constructed to represent residual structure, with 21 channels used instead of 18, with 20 for

amino acids and 1 for nucleic acids. Enzyme labels were converted into 6 dimensional arrays using one-hot encoding.

The architecture for the convolutional neural network was built in Keras, using two Conv3D layers of 32 and 64 filters respectively, a kernel dimension of 3 and a relu activation function. Max pooling and a dropout of 0.5 were implemented following each convolutional layer. Two Dense layer of dimensions 64 and 6, with relu and softmax activation functions respectively, concluded the design. The choice of stochastic gradient descent or Adam made negligible difference in model performance and categorical cross-entropy was used to evaluate loss.

Data for the recurrent neural network was prepared by taking the first 50 amino acids from each of the enzymes and one-hot encoding the sequence into an array. The architecture consisted of only two layers, a Dense layer of dimension 20 and LSTM layer of dimension 2, optimized with Adam and loss evaluated with categorical cross-entropy.

Results

Observations from 5,000 enzymes were extracted from Protein DataBase (PDB) to train the CNN in a 70-30 training-validation split. For both datasets, a test accuracy of 0.55 was achieved before the validation accuracy began to deviate significantly. Test accuracy beyond this point eventually overfitted to a perfect value. Despite the use of a fit generator data stream to greatly reduce memory load during training, the large size of individual data points limited the training set. Increasing the size of training data beyond the threshold used caused memory errors on the tensorflow distribution used on the training machine (v1.12.0), an issue that could have been fixed if a custom package was built from source to more efficiently distribute the memory of the specific hardware. It is likely that a training set that spanned the extent of the PDB would have achieved similar metrics to the ANNs described in the literature, if not greater do to the convolutional design that exploits the structural determinants of catalytic behavior.

Sequences from 222 enzymes were used to train the RNN model in an 80-20 training-validation split. An accuracy of 0.70 was achieved using this technique.

Discussion

One primary takeaway from these models is the importance of designing intelligent solutions to machine learning problems. Engineering model architecture and data pipelines with purpose and application in mind is essential to avoid treating these complicated algorithms as “black boxes”. Despite the perceived efficacy of volumetric matrix data, selecting structural enzyme information as it pertains to known catalytic behavior of active sights and understanding the importance of shape in protein function, the simpler solution prevailed, not to mention with far less training hours and lines of code. By identifying the powerful patterns among homologous protein sequences, we let evolution do much of the heavy lifting in the data processing, using a simple recurrent network to fill in the remaining pieces.

The numbers achieved by the more successful network were shy of the academic state of the art by a mere 7%, a neural network that achieved accuracy of 0.776 under Amidi and colleagues (Amidi 6). Given the resources, information, and experience at our disposal, we are greatly pleased by the combined results of our efforts.

The application of recurrent analysis of homologous sequence data to genetics and synthetic biology would be interesting. Identifying other emergent properties of unknown proteins based on identified gene information would greatly accelerate *in silico* models of systems biology, and will be logical development of this research.

Works cited

Amidi, Afshine, *et al*. *PeerJ.* 2018, *6*, e4750

Gupta, et al. "Protein Classification Using Machine Learning and Statistical Techniques: A Comparative Analysis." *ArXiv.org*, 18 Jan. 2019, arxiv.org/abs/1901.06152.

Cai, Y., Yang, H., Li, W., Liu, G., Lee, P. W., & Tang, Y. (2018). *Multiclassification Prediction of Enzymatic Reactions for Oxidoreductases and Hydrolases Using Reaction Fingerprints and Machine Learning Methods. Journal of Chemical Information and Modeling, 58(6), 1169–1181.*doi:10.1021/acs.jcim.7b00656