# Optimization of Deep Learning Neural Networks to Yield an Accurate Predictor of Post-Translational Protein Modifications in Threonine and Serine Phosphorylation Datasets

Vijay Anand, Ziping Liu, Andrew Gow
Rutgers, The State University of New Jersey

The post-translational modifications (PTMs) of proteins are critical factors in the activation, regulation, and transportation of key components of cellular regulation. The identification of such modifications involves complex biochemical measurements. Artificial Neural Networks (ANNs) represent versatile tools to allow computer-based approaches to simplify protein studies. ANN architectures can be used to assess the presence of PTMs, given the amino acid sequence and/or other details of an observed protein. The purpose of this project was to develop and optimize two separate 3-Layer Deep Learning ANNs using serine and threonine phosphorylation datasets as model systems. Data was taken from the Phospho.ELM database and split into two separate datasets, which were each split into a training set and a testing set, at a ratio of 70:30, respectively. A cost versus iteration/age graph of the training set and the Mathew Correlation Coefficient (MCC) of the testing set were used to monitor performance; changes were made to the hyper-parameters between runs to optimize the ANN. For the serine dataset, the ANN achieved an MCC of ~0.438 with a mini-batch size of 512, learning rate of 0.05, L2 regularization of 0.02, 180 iterations and an amino acid window size of 40. For the threonine dataset, the ANN achieved an MCC of ~0.442 with a mini-batch size of 64, learning rate of 0.05, L2 regularization of 0.002, 95 iterations and an amino acid window size of 40. Both optimized ANNs contained 2 Hidden Layers, 50 neurons each, which used Rectified Linear Unit (ReLU) activation functions. Notably during optimization with the threonine dataset, full-batch descent yielded a slowly decreasing cost function plateauing at ~0.683, whereas mini-batch descent showed a significant decrease in the cost to ~0.007; the serine dataset experienced a similar pattern. These data demonstrate that 3-Layer Deep Learning ANNs can learn trends in example PTMs, though issues such as overfitting may require further modifications to the ANNs. Funded by NIH HLO8661 and R25ES020721 and the SURF Fellow Program.