Introduction

A paper came out in the Journal of Computational Biology that used an interesting neural network architecture and model training strategy to select the most important features in the context of *cis-*regulatory elements in non-coding DNA sequences [1]. After analysis of the methods of this paper, we believe that the same techniques could be applied to other data sets to select important features. Our proposal is to take the same regularized neural network technique (deep feature selection) and apply it to several new data sets to see if the method does in fact translate to other domains.

This is essentially the task of feature selection. In a biological context, feature selection means extracting the elements in a biological sequence that contribute most substantially to some target. For example, perhaps the presence of “ACTGT” at positions 931-935 of a chromosome means that the patient may develop invasive ductal carcinoma. We will train models to pick up on these patterns and report the features (chromosome positions, for example) that contribute the most substantially.

Methods

We will need to collect a large set of labeled biological sequences, develop a high-quality classifier that takes these biological sequences and outputs the correct class, then perform the necessary modifications necessary to transform it from feedforward neural network to the deep feature selection model presented in the paper. At this point we are not certain if an off the shelf neural network package is capable of the required modifications, therefore this project will be broken into two stages. First we will develop the high quality neural network classifier, then we have a stretch goal of converting it to a deep feature selection model as presented in the paper. We will be using Keras and Tensorflow through Python to develop the neural networks.