CS 601.482 – Machine Learning: Deep Learning – Project Proposal

Team Members: Kaushik Srinivasan, Nathan Vallapureddy

We are planning to use Deep Learning to classify SNP-genes with their associated tissue specificity. We have a substantial genetic corpus, but our ability to identify which tissue (e.g. brain, blood, skin) the regions affect is minimal. If we can identify patterns in genetic code and its associated tissue specificity, we can strongly map the impact of SNPs to their associated effects.

Our input data consists of genetic coordinates, which specify the chromosome number and the location and labels which contain the tissue specific to the coding region. The labels associated with the genetic code have been identified to a high accuracy by the GTEx consortium using eQTL analysis provided by GWAS studies. We will take (hyperparameter) n base pairs starting from our variant, and identify the coding region's tissue specificity. We plan to vary the input genome by splitting it into k-mers (k is a hyperparameter). We initially plan to encode the ATCG sequence and have categorical inputs, however this is subject to change. We will initially perform a binary classification of tissues (i.e. brain and non-brain). Depending on the success of our results, we will attempt to perform multi-class labelling.

The specific genome assembly reference we plan to use is Human Genome v19 (hg19 - GRC37), and will be obtaining and manipulating input using samtools/bedtools/BowTie.

In general, Kaushik will be focusing on obtaining the data and generating graphs from the output model. Nathan will be generating the neural net model and analyzing the data outputs. We will both equally write up the report. This is not final and there will be regions of the project where our responsibilities overlap.

We will use GTEx data from and work alongside Alexis Battle's Lab and will receive council from her PhD student Yuan He.