EN.553.650 Computational Molecular Medicine
Project Proposal (Neurogenesis)
Ashwin De Silva, Ime Essien, Sai Koukuntla (Group 3)

Introduction We are interested in studying the genetic basis of embryonic neurogenesis. In particular, we wish to determine whether the dynamics of neuronal migration from the germinal zone to the cortical plate during mid-gestation are regulated by a non-linear function of gene expression. Identifying the collection of genes characterizing the neuronal activities in the germinal zone and the cortical plate would be important in understanding the mechanisms underlying early brain development.

Problem Formulation We consider a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ of n neurons where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iM}) \in \mathbb{R}^M$ is a vector containing expressions from M number of genes pertaining to the i^{th} neuron and $y_i \in \{0, 1\}$ indicates its location ($y_i = 0$ when the i $^{\text{th}}$ neuron belongs to the germinal zone and $y_i = 1$ if otherwise). Note that $n << M$ and most of the genes would not inform the location of the cell meaningfully. Given the dataset D, our objective is to determine the set of genes S ($|S| \leq M$) that contains discriminative information about the neuron's location and learn a binary classifier over S that captures the relationship between the gene expression and neuronal location.

Dataset For the purposes of this project, we use a single-cell RNA-seq dataset [1] on human embryonic neurogenesis which contains expression data from $M = 35543$ genes across $n = 15060$ cells.

Proposed Method Because the number of samples is significantly smaller than the number of features (in this case, the features are the gene expressions), we must first select features or combinations of features that are highly discriminative. To this end, we will use traditional methods (e.g. selecting features based on mutual information with labels, multiple rank tests with Bonferroni or B-H correction) and dimensionality reduction techniques such as principal component analysis (PCA). For our classifier, we will train a random forest (using the `xgboost` package) and a neural network (using `pytorch` library) to predict the label from our selected features. We plan to split the data into training, validation and testing sets and use K-fold cross-validation to assess which combination of features, hyperparameters, and model achieves the highest accuracy. We will then report the performance of the best 1-2 combinations on the test set. Finally, for these combinations, we will use the appropriate discriminant function (output unit value for neural network, # of trees with $Y = 1$ for random forest) to construct an ROC curve and determine the highest achievable specificity while maintaining 90% sensitivity.

References

[1] Damon Polioudakis, Luis de la Torre-Ubieta, Justin Langerman, Andrew G Elkins, Xu Shi, Jason L Stein, Celine K Vuong, Susanne Nichterwitz, Melinda Gevorgian, Carli K Opland, et al. A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron*, 103(5):785–801, 2019.