# CS 224d Project Proposal

Govinda Kamath and Jesse Zhang

21 April 2016

## 1 Problem Description

In a single cell RNA-seq experiment, one is interested in recovering the different types of cells present in a tissue. RNA transcripts become protein (which are the functional units of the cell), and therefore the reads (observed subsequences of a constant length) obtained from these transcripts are related the cell's type. The purpose of this project is to explore if a neural network can learn a relationship between reads and cell type by treating reads as sentences of $k$-mers (length $k$ strings of $A, C, T, G$).

## 2 Data

The data will be reads obtained from sequencing experiments where the cell type is known. We will use data from a popular recent single cell RNA-seq experiment by [3], which studied mouse brain cells. This data set consists of 3005 cells, of which are of two main types: astrocytes and interneurons. There seems to be enough data, as there are around $1,884,135,00e0$ reads in total, which should allow us to train models reasonably well.

If the model works well, we will try the approach on other single cell assays like single cell ATAC-seq [1].

## 3 Methodology

Supervised training of the neural network will be performed by

1. Mapping reads to vectors using methods such as word2vec
2. Feeding these vectors into the network using known cell type as the predicted label
3. Cross validating to select model hyper-parameters such as network structure and activation function.

After training (and use of a development set to set hyper-parameters), the model will be tested on a set of held-out data. Each read from a test cell will be fed into the model, and the cell will be classified as the most frequent output type.

## 4 Related work

Usually in this setting, people cluster using a reference, which is a set of all possible RNA sequences that can be seen in the cell. There has been some interest in this problem recently [2]. A disadvantage of this however is that the reference is most likely incomplete, and there may be many rare transcripts that we do not really know of. Hence we propose to work reference-less in this project. There is very little work done in that direction. The ultimate goal of such a project would be to do unsupervised reference-less clustering, but we plan to work on the easier problem of classification here.

While not much work has been done on treating the genome as a natural language, we will use NLP techniques taught in class to generate vectors from reads (such as word2vec and GloVe).

## 5 Evaluation plan

We will evaluate out results by looking at the classification error rate of the network compared to simpler models such as logistic regression on a bag of words model (which is essentially trying to use per-cell $k$-mer counts to cluster.).

## References

[1] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015.

[2] Vasilis Ntranos, Govinda M Kamath, Jesse Zhang, Lior Pachter, and David N Tse. Fast and accurate single-cell rna-seq analysis by clustering of transcript-compatibility counts. *Genome Biology, in press*, page 036863, 2016.

[3] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.