EE 377 Project Proposal

Govinda Kamath and Jesse Zhang

26 January 2016

1 Motivation

With recent advancements in RNA sequencing (RNA-seq) technologies, scientists have enjoyed a surplus of data at the resolution of individual cells. One of the most interesting problems in biology is that of determining the differentiation patterns of cells, such as the process of a stem cell evolving (or differentiating) into a heart cell. A recent experiment [1] has produced a single-cell RNA-seq dataset consisting of 271 human primary myoblasts collected at four different times. The authors attempt to order these cells according to a so-called "pseudotime" or a latent variable that quantifies how differentiated a given cell is. Attempts of studying cell differentiation have been performed by various groups using various methods. For example, Mueller et al. [2] attempts to order *batches* of sampled cells according to the distributions of certain transcripts (or features) within clusters. For this project, we hope to combine the notion of measuring the distance between cells based on distributions of certain transcripts with the concept of "pseudotime." We hope that by doing this, we can obtain a more biologically meaningful (and statistically sound) ordering of the 271 primary myoblasts. The success of this project could lead to meaningful insights in modeling cell differentiation; these techniques can perhaps be applied to other single-cell datasets studying the growth of cancer, for example.

2 Approach

The problem is as follows: for each of T sampling times, we are given n length-d vectors (representing n cells each of which are described with d features or transcripts/genes). We want to get a partial ordering on the cells such that a cell that comes later is further along the differentiation process. We do not know *a priori* the number of clusters and the number of end states. In other words, a single cell has a non-zero probability of taking multiple differentiation paths, ending at one of multiple unique end states.

We want to model the differentiation process as a latent continuous-time Markov chain (CTMC) where states represent cell types and transition rates represent times required to transition from one cell state to another. We would want to do inference on such a model.

One potential approach taken in [3] is to cluster the Tn cells into k clusters and then draw a minimum-spanning tree through the k centroids. We hope that over the course of this quarter, we will be able to apply what we learn to estimating both the structure of the CTMC and the values of the parameters. It would also be interesting to identify the relationship between number of samples and model complexity.

3 Desired results

We hope to find a good ordering of the 271 cells, and we evaluate how "good" our results are using the authors' results and by looking at gene patterns. In particular, we hope to provide theoretical justification for our results based on clustering using chosen distance metrics and running a minimum-spanning tree on the cluster centroids to give us a partial ordering [3]. Because we do not fully understand the theoretical

implications of this project just yet, we could potentially fail to model the problem in a tractable but reasonable manner. For example, perhaps a CTMC will turn out to be suboptimal for the dataset and we will have to consider other models.

4 References

1. Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells." Nature biotechnology 32, no. 4 (2014): 381-386.

2. Mueller, Jonas, Tommi Jaakkola, and David Gifford. "Modeling Trends in Distributions." arXiv preprint arXiv:1511.04486 (2015).

3. Ntranos, Vasilis, Govinda M. Kamath, Jesse Zhang, Lior Pachter, and N. Tse David. "Fast and accurate single-cell RNA-Seq analysis by clustering of transcript-compatibility counts." bioRxiv (2016): 036863.