Time-series and Hierarchical Modeling for sc-RNA Expression Levels

Skanda Koppula (skoppula@mit.edu), Karren Yang (karren@mit.edu)

6.882 Project Proposal, March 10, 2017

Cancer cells implanted into mice exhibit patterned levels of RNA expression. The expression in each cancer cell varies over time, and the cells may naturally diverge in their expression during this time. Previous work has looked into an LDA-based clustering of expression vectors for particular time slices, resulting in interesting and biologically meaningful cell grouping shifts over time.

This project is interested in constructing the first generative model of this process. Specifically, we look to answer some of the following questions:

1. Can we apply a topic model based probabalistic model for a collection of expression vectors for a particular time slice? Will a topic model based solution yield meaningful cell type assignments?
2. How can we apply time-series probabilistic models such as Chinese Restaurant to model cell types over time?
3. How can we model noise in expression level observations in our probabilistic models? Is it possible to use our model to establish estimates for missing coordinates in expression vectors (a common problem in high-throughput expression measurement machines)?
4. Is it possible to emulate expression data with a simple sampling processes (for the purposes of testing any models we create)?

These were a few lines of inquiry we were interested in pursuing. The authors have access to an sc-RNA expression level dataset through their research lab.

1. Referenced paper, outside/related work
2. Outline work to be done
3. 4 steps per team member, internal deadlines
4. Risks: what might be more difficult than planned, thoughts to mitigate these
5. How to evaluate our methods?