Bayesian Clustering and Topic Modeling for Analyzing Single-Cell RNA Sequencing Data

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

6.882 Project Proposal, March 21, 2017

Abstract

Tumors are composed of many sub-populations of cells, which exhibit different patterns of gene expression. In order to characterize sub-populations of tumor cells and identify functional gene modules, we propose applying Bayesian methods (parametric and non-parametric) for clustering and topic modeling to single-cell RNA-sequencing (scRNA-seq) data from tumors. We will compare these methods to standard methods for analyzing scRNA-seq data based their ability to identify known cell sub-populations and gene modules from the literature as well as their robustness to noise. We will furthermore determine if integrating the clustering model with the topic model results in superior performance in these tasks. Time permitting, we will attempt to extend these models to accommodate time-series data.

Overview

Using single-cell RNA-sequencing (scRNA-seq) data, it is possible to obtain the expression levels of 10,000+ genes for 1000+ cells in a single experiment. Given such data, we are primarily interested in answering the following questions:

1. Can Bayesian nonparametric topic models, such as the Hierarchical Dirichlet Process (HDP) and the Indian Buffet Process Compound Dirichlet Process (IBP-CDP), extract functional gene modules from this data? Do Bayesian nonparametric topic models perform as well as parametric topic models, such as Latent Dirichlet Allocation (LDA), on this data?

2. Can we get meaningful cell-type assignments by clustering cells, using finite or infinite mixture models, based on the output of the topic models?

3. How robust are these methods to dropout, a phenomenon commonly observed in real scRNA-seq data?

4. Can integrated models of cell clustering and topic modeling yield superior results in these tasks?

5. Can we extend any of these models to time-series data?

Risk: For a 6-7 week project, this list seems ambitious, but we have mitigated the risk in several ways. (1) Although we are primarily interested in the utility of nonparametric methods for topic (gene module) modeling, we are also using Latent Dirichlet Allocation so that subsequent steps can proceed even if the nonparametric methods do not work well. (2) If we become stuck on steps 4-5, we can submit a good report from objectives 1-3.

Tentative Schedule

2.a Apply LDA and HDP to scRNA-seq data - SK

2.b Apply IBP-CDP to scRNA-seq data - KY

2.c Evaluate biological relevance of gene modules from topic models - KY

2.d Evaluate robustness of topic models to dropout - SK

2.e Apply finite mixture model to cluster topic model output - SK

2.f Apply infinite mixture model to cluster topic model output - KY

2.g Integrate topic modeling and cell clustering - TBA

2.h Extend model to time-series data - TBA

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?