Bayesian Clustering and Topic Modeling of Single-Cell RNA

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

6.882 Project Proposal, March 22, 2017

Overview

Tumors are composed of different sub-populations of cells. These sub-populations exhibit shared patterns of gene expression, and prior research has applied basic measures of statistical distance to cluster cell groups based on samples of single-cell RNA-sequencing (scRNA-seq) data taken from tumors [1]. With contemporary sequencing machines, it is possible to obtain the expression levels of 10,000+ genes for 1000+ cells in a single experiment. Good clustering of sc-RNA data often has biologically meaningful results, and as such, a wide variety of correlation based methods have been used in prior art to derive meaning from scRNA-seq data [2, 3, 4].

In this project, we aim to apply Bayesian methods to study scRNA-seq datasets from the author's lab. Specifically, we intend to explore the use of a hierarchical topic model and combined clustering-topic model (e.g. MGCTM [5]) to analyze both real and simulated scRNA-seq data. We will compare the results of these methods with standard methods used in prior work, and evaluate these methods' robustness to noise. Time permitting, we will explore the use of non-parametric models, and study time-series sc-RNA data.

In summary, and in order of importance, we are primarily interested in answering the following questions:

1. Can we get meaningful cell-type assignments by clustering cells, using finite or infinite mixture models, based on the output of the topic models? How does this compare to existing methods?
2. How robust are these methods to data dropout (commonly observed in real scRNA-seq data)?
3. Do integrated models of cell clustering and topic modeling yield superior results in these tasks?
4. Can 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 nonparametric topic models perform as well as parametric topic models on this data?
5. Can we extend any of these models to time-series data?

Tentative Schedule and Task Breakdown

1. **[04/02] [SK]** Apply LDA and HDP to scRNA-seq data
2. **[04/08] [KY]** Apply IBP-CDP to scRNA-seq data
3. **[04/10] [KY]** Evaluate biological relevance of gene modules from topic models
4. **[04/13] [SK]** Evaluate robustness of topic models to dropout
5. **[04/22] [SK]** Apply finite mixture model to cluster topic model output
6. **[04/25] [KY]** Apply infinite mixture model to cluster topic model output
7. **[05/05] [TBD]** Integrate topic modeling and cell clustering (MGCTM)
8. **[05/05] [TBD]** Extend model to time-series data

Risk and Evaluation

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.

In addition to standards methods to evaluate model fit (e.g. likelihood of held-out data, model visualization, etc.), we can cross-reference prior work, and, when possible, cross-check the biological interpretation of our results.

References

[1] Spearman-based hierarchical clustering of scRNA-seq. http://www.nature.com/nsmb/journal/v24/n3/fig_tab/nsmb.3365_SF1.html

[2] Exploiting single-cell expression to characterize co-expression replicability. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0964-6

[3] SC3 - consensus clustering of single-cell RNA-Seq data. http://biorxiv.org/content/early/2016/09/02/036558

[4] SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004575

[5] Integrating Document Clustering and Topic Modeling. https://arxiv.org/pdf/1309.6874.pdf