Bayesian Clustering and Topic Modeling of Gene Expression Data

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

6.882 Project Proposal, March 23, 2017

Overview

Tumors are composed of different sub-populations of cells, and these sub-populations exhibit shared patterns of gene expression. With contemporary sequencing machines, it is possible to obtain the expression levels of 10,000+ genes for 1000+ cells in a single experiment. In this project, we aim to apply Bayesian methods for clustering and topic modeling to discover meaningful cell clusters and gene modules from single-cell RNA-sequencing (scRNA-seq) datasets from tumors. We first intend to explore the use of a hierarchical topic model, Latent Dirichlet Allocation (LDA) [6], as well as two non-parametric topic models, Hierarchical Dirichlet Processes (HDP) [7] and the Indian Buffet Process Compound Dirichlet Process (IBP-CDP) [8] for analyzing scRNA-seq data. We will evaluate these models based on their ability to discover meaningful functional gene modules. Subsequently, we plan to explore the use of mixture models to cluster cells based on the output of the topic models, which we will evaluate based on their ability to differentiate cell types (e.g. immune, cancer, non-cancerous). Finally, we will train a combined clustering-topic model (e.g. MGCTM [5]) to see if it outperforms the individual models. 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 study extensions of these models appropriate for time-series scRNA-seq data [9].

Prior research has applied basic measures of statistical distance to cluster cell groups or gene modules based on samples of single-cell RNA-sequencing (scRNA-seq) data taken from tumors [1]. Good clustering of scRNA-seq data often has biologically meaningful results, and as such, a wide variety of correlation based methods have been used in prior work to derive meaning from scRNA-seq data [2, 3, 4]. We believe that Bayesian methods for clustering and topic modeling can be more illuminating than these previous methods, due to their consideration of higher-order relationships within the data.

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

1. Do Bayesian topic models such as LDA, HDP and IBP-CDP work well on scRNA-seq data? Can we extract functional gene modules from the topics?
   Topics are to documents what gene modules are to cells; therefore, we will apply these topic models to scRNA-seq datasets to discover genes that tend to be co-expressed in cells, which suggests they are functionally related. We are interested in using LDA, as it is the top-performing parametric topic model and its implementation is available online [6]. We are interested in HDP because it is a non-parametric extension of LDA, and it might be more flexible as we would not need to designate the number of gene modules beforehand [7]. Its implementation is also available online. Finally, we are interested in IBP-CDP because it offers the nonparametric flexibility of HDP, plus it overcomes the assumptions of HDP that gene modules active in many cells must also be highly expressed [8]. Its implementation is not available online, but we plan to ask the authors for their code.

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? How does this compare to existing methods?
   Just as document clustering can be done on topic proportions, we plan to cluster cells based on their gene module proportions. We plan to use a finite Gaussian mixture model, as well as an infinite Gaussian mixture model based on the Dirichlet Process.

3. How robust are these methods to data dropout?
   Real scRNA-seq data suffers from dropout (i.e. a gene that was actually expressed in a cell has zero count for that cell, due to noise and limitations in experimental detection). We hypothesize that clustering based on gene module proportions will be more robust to dropout than clustering based on individual gene counts.

4. Do integrated models of cell clustering and topic modeling yield superior results in these tasks?
   We plan to compose our best topic model and best clustering method into an integrated model, following the example of the MGCTM [5].

5. Can we extend any of these models to time-series data?
   We hope to extend our models to time-series data, for example, using Hidden Markov Models or Dependent Dirichlet Process [9].

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] [CO]** Integrate topic modeling and cell clustering
   This task will be planned and divided once the authors have a clearer idea of the most successful models from the earlier steps. Both authors are expected to deliberate on which models will be most appropriate to integrate based on the outcomes of previous steps in this project. Both authors will contribute to the implementation. More will be announced in the progress report.
8. **[05/05] [CO]** Extend model to time-series data
   Similar to the above, this task will be planned and divided once the authors have a clearer idea of the most successful models from the earlier steps. Both authors are expected to deliberate on which models will be most appropriate to extend to time-series data based on the outcomes of previous steps in this project. Both authors will contribute to the implementation. More will be announced in the progress report.

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

[6] Latent Dirichlet Allocation. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[7] Hierarchical Dirichlet Processes http://people.eecs.berkeley.edu/~jordan/papers/hdp.pdf

[8] The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling http://www.cs.columbia.edu/~blei/papers/WilliamsonWangHellerBlei2010.pdf

[9] A Time-Series DDP for Functional Proteomics Profiles https://www.ma.utexas.edu/users/pmueller/pap/NM12.pdf