Single-cell network mixed-membership community detection by XXX

Sishir Subedi[*] Yongjin P. Park[†]

04:11:15 PM, Feb 15, 2022

Background

The advancement in single-cell RNA-sequencing (scRNA-seq) has emerged as a new frontier in transcriptomics and contributed to our understanding of complex disease biology. The ability to quantify gene expression levels at a single-cell resolution provides a framework to uncover novel cell types, interactions, and dynamics of cellular systems during disease progression.[Nomura, 2021] There are numerous popular tools to analyze gene expression data from single-cell. Most of these tools incorporate a common workflow that includes data normalization, filtering, and representation in lower dimensions for various downstream analyses such as clustering, differential expression, and cell type identification.[Zappia and Theis, 2021] This multi-step process has limitations (TODO:explain), and efforts are ongoing to develop a streamlined and robust computational method to model gene expression levels directly from raw count data.

Autoencoder is an unsupervised machine learning method based on neural networks architecture and has been utilized in many areas of single-cell analysis, such as denoising and clustering.[Eraslan et al., 2019, Geddes et al. [2019]]. These studies have shown that autoencoder-based techniques capture the essential biological signals from sparse and heterogeneous single-cell data by efficiently representing the data in lower dimensions.

Results

[*] Bioinformatics Program, The University of British Columbia

[†] Department of Statistics, The University of British Columbia, ypp@stat.ubc.ca

Discussion

Conclusions

Methods

Datasets

- toy data?
- breast cancer - Chung et al. [2017]
- peripheral blood mononuclear cells (pbmc) - Freytag et al. [2018]
- tcells - Zheng et al. [2021]

The Embedded Latent Space model

We have a sample of cells $c_1, ..., c_N$ and a list of genes $g_1, ..., g_D$, where $c_n g_1, ..., c_n g_D$ are raw count data for D genes in cell c_n.

Likelihood

$$p(c_{nd} \mid \delta_n, \beta) = \sum_k \theta_{nk} \beta_{k,c_{nd}} \tag{1}$$

How to construct a feature-incidence matrix

- Train ETM model
- Use latent dimension to find neighbouring cells
- For each neighbourhood, construct an interaction matrix where rows are neighbouring cell pairs and edges are ligand-receptor pairs from known database
- calculate ligand-receptor interaction score

$$f(c_i, c_j, l_x, r_y) = max(e_{c_i l_x} \times e_{c_j r_y}, e_{c_i r_y} \times e_{c_j l_x})$$

$e_{c_i l_x}$ is expression of ligand x in cell i, $e_{c_j r_y}$ is expression of receptor y in cell j, c_i and c_j are neighbouring cells from ETM model, and l_x and r_y are ligand-receptor pairs from known database

- X_{gi}

Clustering the rows of an incidence matrix

Notations

- Y_{eg}: a feature g's contribution to an edge e, $Y \geq 0$
- Z_{ek}: a latent variable for an edge e
- $p(Z_{ek} = \pi)$, where $\pi = 1$ if and only if the edge e belongs to the cluster k; otherwise, $\pi = 0$
- λ_k : parameter vector for a cluster k

Likelihood

$$\begin{aligned} p(Y, Z|\lambda, \pi) &= \prod_g \sum_k p(y_g, z_g = k) \\ &= \prod_g \sum_k p(y_g \mid z_g = k) p(z_g = k) \\ &= \prod_g \sum_k Poisson(y_g \mid \lambda_k^y) \pi_k \\ &= \prod_g \sum_k \prod_e Poisson(y_{eg} \mid \lambda_{ek}^y) \pi_k \end{aligned} \tag{2}$$

Log-likelihood

$$\begin{aligned} log(p(Y, Z|\lambda, \pi)) &= \sum_g log(\sum_k p(y_g, z_g = k)) \\ &\geq \sum_g \sum_k q(z_g = k) log(\frac{p(y_g, z_g = k)}{q(z_g = k)}) \\ &= \sum_g \sum_k q(z_g = k) log(p(y_g, z_g = k)) - q(z_g = k) log(q(z_g = k)) \end{aligned} \tag{3}$$

EM algorithm (like forward-backward of HMM):

1. Step 1. Estimate z given λ
2. Step 2. Estimate λ given z

References

Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, et al. Single-cell rna-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications*, 8(1):1–12, 2017.

Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):1–14, 2019.

Saskia Freytag, Luyi Tian, Ingrid Lönnstedt, Milica Ng, and Melanie Bahlo. Comparison of clustering tools in r for medium-sized 10x genomics single-cell rna-sequencing data. *F1000Research*, 7, 2018.

Thomas A Geddes, Taiyun Kim, Lihao Nan, James G Burchfield, Jean YH Yang, Dacheng Tao, and Pengyi Yang. Autoencoder-based cluster ensembles for single-cell rna-seq data analysis. *BMC bioinformatics*, 20(19):1–11, 2019.

Seitaro Nomura. Single-cell genomics to understand disease pathogenesis. *Journal of Human Genetics*, 66(1):75–84, 2021.

Luke Zappia and Fabian J Theis. Over 1000 tools reveal trends in the single-cell rna-seq analysis landscape. *Genome biology*, 22(1):1–18, 2021.

Liangtao Zheng, Shishang Qin, Wen Si, Anqiang Wang, Baocai Xing, Ranran Gao, Xianwen Ren, Li Wang, Xiaojiang Wu, Ji Zhang, et al. Pan-cancer single-cell landscape of tumor-infiltrating t cells. *Science*, 374(6574):abe6474, 2021.