SPRUCE - Single-cell Pairwise Relationships Untangled by Composite ETM

Park Lab

12:56:21 PM, Jul 04, 2022

Methods

Data origins and preprocessing

We generated a mixuture dataset combining three different recent breast cancer studies. The first one is a breast cancer dataset consisting of 100k cells from a single-cell atlas of human breast cancers $Wu et al. [2021]. The secod dataset consited of 48k cells from a single-cell atlas of the healthy breast tissues (Bhat-Nakshatri et al. [2021]).The third dataset is 6k subset of breast cancer CD4 and CD8 T-cells from a pan-cancer atlas of tumour-infiltrating T cells profiled across 21 cancer types and 316 donors (Zheng et al. [2021]). The total number of cells in the combined dataset was 155913. We filtered out genes detected in less than 3 cells along with mitochondrial gene and spike genes, which lead to 20265 genes in the final dataset.

The SPRUCE Model

The SPRUCE model consists of two steps of topic modeling - first, we model cell topic to assign each cell into type/state, and then we use these cell topics to construct a set of neighbours and model interaction topic, which assigns each neighbour pair to unique interaction state.

Cell topic analysis

The cell topic modeling extends on the ideas of LDA. Consider a sample of cells $i_1, ..., i_N$ and a list of genes $g_1, ..., g_G$, where $x_{i1}, x_{i2}, ..., x_{iG}$ are raw count data for G genes in cell i. We assume

that each cell count vector X_i was generated from a multinomial distribution parameterized by a gene expression frequency matrix with an element ρ_{ig} of a gene g in the cell i.

$$p(\mathbf{x}_i|\rho_i) = \prod_g \rho_{ig}^{X_{ig}}$$

We introduce Dirichlet prior on the ρ and parameterize the Dirichlet as a generalized linear model (GLM). Exploiting the conjugacy between the multinomial and Dirichlet, we integrate out the unknown parameters ρ.

$$p(\rho_i|\lambda_i) = \frac{\Gamma(\sum_g \lambda_{ig})}{\prod_g \Gamma(\lambda_{ig})} \prod_{g \in \mathsf{genes}} \rho_{ig}^{\lambda_{ig}-1}$$

$$\lambda_{ig} = \exp\left(\sum_{t=1}^{T} \theta_{it}(\beta_{tg} + b_g)\right)$$

The topic proportion θ_{it} for cell i is drawn from logistic normal distribution with model hyperparameters δ_{it} and we assume topic proportions within a simplex, namely $\sum_{t \in \mathsf{topics}} \theta_{it} = 1$.

$$\delta_{it} \sim N(0, I); \theta_{it} = softmax(\delta_{it}) \tag{1}$$

The marginal likelihood

$$p(\mathbf{x}_i|\cdot) = \frac{\Gamma(\sum_g \lambda_{ig})}{\sum_g \Gamma(\lambda_{ig})} \frac{\Gamma(\sum_g \lambda_{ig} + X_{ig})}{\sum_g \Gamma(\lambda_{ig} + X_{ig})}$$

The marginal likelihood of each cell is an intractable problem because it involves integral over the topic proportion. Variational inference techniques can be used to approximate this type of intractable integrals.

Let $p_\theta(z \mid x)$ be a true posterior and $q_\phi(z \mid x)$ be an approximate posterior.

$$
\begin{aligned}
D_{KL}(q_\phi \mid\mid p_\theta) &= E_{q_\phi}[log \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)}] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z \mid x)] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ \frac{p_\theta(z,x)}{p_\theta(x)}] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z,x)] + E_{q_\phi}[log\ p_\theta(x)] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z,x)] + \int q_\phi(z \mid x) log\ p_\theta(x) dz \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z,x)] + log\ p_\theta(x) \int q_\phi(z \mid x) dz \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z,x)] + log\ p_\theta(x) \\
log\ p_\theta(x) &= -E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(z,x)] + D_{KL}(q_\phi \mid\mid p_\theta)
\end{aligned}
\tag{2}
$$

Here, $D_{KL}(q_\phi \mid\mid p_\theta)$ is intractable but it is always ≥ 0, we can use this property to remove D_{KL} from the equation and the marginal log likelihood $log\ p_\theta(x)$ will be at least $\geq -E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(z,x)]$. Since this term is the lower bound on the evidence,it is called as Evidence Lower Bound(ELBO). We maximize the marginal log likelihood by maximizing the ELBO and indirectly minimize the KL divergence.

$$
\begin{aligned}
ELBO &= -E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(z,x)] \\
&= -E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(x \mid z)] + E_{q_\phi}[log\ p_\theta(z)] \\
&= E_{q_\phi}[log\ p_\theta(x \mid z)] - E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(z)] \\
ELBO &= E_{q_\phi}[log\ p_\theta(x \mid z)] - E_{q_\phi}[log \frac{q_\phi(z \mid x)}{p_\theta(z)}]
\end{aligned}
\tag{3}
$$

Here, $E_{q_\phi}[log\ p_\theta(x \mid z)]$ is an expected reconstruction error and $E_{q_\phi}[log \frac{q_\phi(z|x)}{p_\theta(z)}]$ is KL Divergence between approximate posterior and the prior.

Approximate posterior and prior

The approximate posterior $q_\phi(z \mid x)$ is a Gaussian variational distribution $q(\delta_i; i_n, v) = N(\mu, \Sigma)$

whose mean and variance are constructed form a neural network parameterized by v. The network takes raw count data of a cell i_n for G genes and outputs a mean and variance of δ_i. The prior $p_\theta(z) = N(0, I)$. The KL divergence between these two form of Gaussians exist in closed form and given as-

$$\begin{aligned} D_{KL}(q_\phi \mid\mid p_\theta) &= E_{q_\phi}[log\frac{q_\phi(z \mid x)}{p_\theta(z)}] \\ &= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z)] \\ &= 1/2 \sum_d (1 + log(\Sigma) - \mu^2 - \Sigma) \end{aligned} \tag{4}$$

In addition to the latent state KL divergence, we take into account the uncertainty of β_{tg} parameters:

$$\beta_{tg} \sim \mathcal{N}(0, 1)$$

Total Expected log-likelihood Lower-bound (ELBO):

$$\frac{J}{n} = \frac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{x}_i | \theta_i(\mathbf{z}_i), \beta) \tag{5}$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} D_{\mathsf{KL}} \left(q(z_{it}) \| p(z_{it}) \right) \tag{6}$$

$$+ \frac{1}{n} \sum_{t=1}^{T} \sum_{g=1}^{G} D_{\mathsf{KL}} \left(q(\beta_{tg}) \| p(\beta_{tg}) \right) \tag{7}$$

$$\tag{8}$$

Neighbour cells calculation

A set of neighbour cells were calculated for each cell using the topic assignment from cell topic model. For each cell, five neighbours from each topic were calculated using ANNOY. In total, we generated 159 neighbours for each cell - 32 topics and 5 neighbours from each topic. - removed self neighbour pair. 18 topics with cell count less than 100 were removed during generating annoy model list. Topics were not removed during generating neighbours, only selected topics were used to create a model list.

Ligan receptor data augmentation

For each cell pair, we selected ligand and receptors genes from celltalkDB database and incorporated known interactions from the database using the transformation of spaces as shown in the figure.

Interaction topic analysis

Multinomial-Dirichlet:

$$p(\mathbf{y}_i|\mathbf{q}_i) = \frac{(\sum_g Y_{ig})!}{\prod_g Y_{ig}!} \prod_g q_{ig}^{Y_{ig}}$$

$$\mathbf{q}_i \sim \mathsf{Dir}(\mathbf{q}_i|\rho_i) = \frac{\Gamma(\sum_g \rho_{ig})}{\prod_g \Gamma(\rho_{ig})} \prod_g q_{ig}^{\rho_{ig}-1}$$

Single-cell generative model:

$$p(\mathbf{x}_j|\cdot) = \frac{\Gamma(\sum_g \lambda_{jg})}{\sum_g \Gamma(\lambda_{jg})} \frac{\Gamma(\sum_g \lambda_{jg} + X_{jg})}{\sum_g \Gamma(\lambda_{jg} + X_{jg})}$$

where

$$\lambda_{jg} = \lambda_0 \exp\left(\sum_{t=1}^{T} \theta_{jt}(\beta_{tg} + \delta_g)\right)$$

$$\lambda_0 = \exp(\tilde{\lambda}_0)$$

$\sum_t \theta_{jt} = 1$

Bayesian regularization of the model parameters

$$\beta_{tg} \sim \mathcal{N}(0, 1)$$

Total Expected log-likelihood Lower-bound (ELBO):

$$\frac{J}{n} = \frac{1}{n}\sum_{i=1}^{n} \log p(\mathbf{x}_i|\theta_i(\mathbf{z}_i), \beta) \tag{9}$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} D_{\mathsf{KL}}\left(q(z_{it})\|p(z_{it})\right) \tag{10}$$

$$+ \frac{1}{n}\sum_{t=1}^{T}\sum_{l=1}^{L} D_{\mathsf{KL}}\left(q(\beta_{tl})\|p(\beta_{tl})\right) \tag{11}$$

$$+ \frac{1}{n}\sum_{t=1}^{T}\sum_{r=1}^{R} D_{\mathsf{KL}}\left(q(\beta_{tr})\|p(\beta_{tr})\right) \tag{12}$$

Results

Probabilistic topic models identify resident cell types and cancer subtypes

Cell-cell interaction topics reveal new cancer types

Different interaction topics induce subtype-specific gene-gene networks

Each interaction topic show disjoint differential expression genes

- State:24 Ligand S100A9 in cancer and receptor CD68 in neighbouring B cells/myeloid cells. S100A9 is a calcium-binding protein that is associated with inflammation and expressed not only in myeloid cells but also in some tumours. S100A9 expressed in ER−PgR− breast cancers induces inflammatory cytokines and is associated with an impaired overall survival. British Journal of Cancer volume 113, pages1234–1243 (2015) Ligand HSPA1A in cancer LncRNA HOTAIR enhances breast cancer radioresistance through facilitating HSPA1A expression via sequestering miR-449b-5p
- State:22 B2M is abnormally expressed in many cancer types. B2M binds to CD3D,CD3G,KLRD1, etc. The chemokine receptor CXCR4 has been found to be a prognostic marker in various types of cancer, including breast cancer.

SUPPLEMENTAL

Total number of cells from different datasets-

```
GSE164898            48495


T-cells              35214
Cancer Epithelial    24489
Myeloid               9675
Endothelial           7605
CAFs                  6573
PVL                   5423
Normal Epithelial     4355
Plasmablasts          3524
B-cells               3206 --> 101149


GSE156728-CD4         3063
GSE156728-CD8         4291 --> 6269
```

Removed cell topics during generating neighbour model list.

```
h35       57
h11       50
h10       39
h8        36
h44       34
h42       20
h29       18
h3        16
h5        13
h16       12
h36        7
```

```
h25     6
h47     6
h41     5
h49     4
h15     3
h18     2
h13     2
```

References

Poornima Bhat-Nakshatri, Hongyu Gao, Liu Sheng, Patrick C McGuire, Xiaoling Xuei, Jun Wan, Yunlong Liu, Sandra K Althouse, Austyn Colter, George Sandusky, et al. A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. *Cell Reports Medicine*, 2(3):100219, 2021.

Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R Torpy, Nenad Bartonicek, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9):1334–1347, 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.