SPRUCE - Single-cell Pairwise Relationships Untangled by Composite ETM

Sishir Subedi and Yongjin Park

05:59:10 PM, Jul 14, 2022

SUMMARY

INTRODUCTION

RESULTS

The SPRUCE model

SPRUCE utilizes composite embedded topic models (ETM) to identify cell types and define its subtypes based on cell-cell communication in tumour microenvironment (TME). The input for the model is a single-cell data, which contains multiple cell types present in TME accross different data sets. Let X_{iG} be raw count gene expression data for G genes in cell i. We modeled X_{iG} with multinomial distribution parameterized by a normalized gene expression frequency which achieves a scale-invariant property across different cells, batches, and data sets. Further, we introduce Dirichlet prior on the gene frequency and parameterize the Dirichlet as a generalized linear model(GLM) with linear combinations of topic specific probabilities.

The model consists of two ETMs each with a pair of encoder-decoder networks. The first ETM, cell topic model, takes gene expression single-cell data as input and models cell topic and topic specific gene probabilities. Next, we assigned a cell topic to each cell based on the highest topic proportion from the cell topic model. The topic assignment was used to construct a set of neighbours for

each cell such that one cell is paired with five nearest neighbours from each topic. The ligand and receptor gene expression data from each cell pair were transformed to each other's space by interaction database celltalkdb. The transformed ligand and receptor data were treated as two independent modules with its own encoder and decoder in the model. The latent variables with encoded information from two modules were combined by taking its average to obtained a final interaction topic variables as mixture of experts from ligand and recetor latent space. The second ETM, interaction topic model, uses transformed ligand-receptor data from neighbouring cell pairs as input and models combined interaction topic for each cell pair and separate topic specifc ligand and receptor gene probabilities.

SPRUCE identifies resident cell types and cancer subtypes

To assign each cell to a cell topic we applied our purposed cell-topic model to a mixture dataset from tumour microenvironment of breast cancer, which consisted 155913 cells and 20265 genes. After unsupervised training of the model, all cells were mapped in the latent cell topic space. Clustering on latent variables was performed using kmeans and clusters were mapped to known cell populations using majority rule. Annotations for cell type was assigned to each cell based on assignment from previous studies and in case of its unavailability we used reference-based cell type identification method SingleR with a tumor micro-environment reference dataset from CHETAH (add reference). The latent cell topics with assigned cell type were visualized with UMAP which shows distint clusters of all major types of cells including epithelial, immune, and cancer cells presnt in tumor microenvironment (Figure 1a). The estimated cell topic proportions show that the resident cell types have similar topic proportions. However, cancer cells have different mixture of topic proportions which shows that the model identified many distinct cell topics of cancer cells(Figure 1b). Furthermore, the model optimized the gene frequency for each topic which is biologically interpretable in understanding which genes are driving each cell topic (Figure 1c). Here, cell topics for different cancer cells show disjoint set of genes with high frequencies.

Cell-cell interaction topics generated by SPRUCE reveals new cancer types

The interaction topic model uses transformend ligand-receptor gene expression data from neighbouring cell pairs and models interaction topic for each cell pair and identifies ligand and receptor

genes enriched in each interaction topic. The model identified seven major patterns of interaction topic for cell pairs with cancer cell (Figure 2a). Each interaction topic represent unique cell-cell interaction state based on the gene expression levels ligands and receptor genes between them. The cell type distribution of neighbouring cells of cancer cells in each topic show enrichment of specific cell types (Figure 2b). For example, topic 22 and 24 shows enrichment of immune cells where topic 22 has higher proportion of T-cells like neighbours while topic 24 has higher proportion of Monocyte/B-cells like neighbours. Similarly, topic 18 shows enrichment of cancer-associated fibroblasts and perivascular like cell types. The top ligand and receptor genes in each topic identified by the model corresponds with the cell types enriched in the respective topic. Similarly, these top ligand and receptor genes are highly expressed in cancer and non-cancer cell pairs.

Different interaction topics induce subtype-specific gene-gene networks

Each interaction topic show disjoint differential expression genes

METHOD DETAILS

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.

Cell topic model

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.

Generating neighbour cells

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 python package ANN.

A annoy model was created for each topic with total cell count more than 100. In total, we generated 159 neighbour cell pairs for each cell - 32 topics and 5 neighbours from each topic, excluding self neighbour pair.

Ligand receptor data augmentation

For each cell pair generated by neighbour analysis, we transformed ligand and receptors raw gene expression data into each others space using a binary interaction matrx generated from publicly available celltalkDB database (Shao et al. [2021]).

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):

$$\frac{J}{n} = \frac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{x}_i | \theta_i(\mathbf{z}_i), \beta) \tag{2}$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} D_{\mathsf{KL}} \left(q(z_{it}) \| p(z_{it}) \right) \tag{3}$$

$$+ \frac{1}{n} \sum_{t=1}^{T} \sum_{l=1}^{L} D_{\mathsf{KL}} \left(q(\beta_{tl}) \| p(\beta_{tl}) \right) \tag{4}$$

$$+ \frac{1}{n} \sum_{t=1}^{T} \sum_{r=1}^{R} D_{\mathsf{KL}} \left(q(\beta_{tr}) \| p(\beta_{tr}) \right) \tag{5}$$

SUPPLEMENTAL

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

$$
\begin{aligned}
D_{KL}(q_\phi \,||\, p_\theta) &= E_{q_\phi}[log\frac{q_\phi(z \mid x)}{p_\theta(z \mid x)}] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z \mid x)] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ \frac{p_\theta(z,x)}{p_\theta(x)}] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z,x)] + E_{q_\phi}[log\ p_\theta(x)] \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z,x)] + \int q_\phi(z \mid x) log\ p_\theta(x) dz \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z,x)] + log\ p_\theta(x) \int q_\phi(z \mid x) dz \\
&= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z,x)] + log\ p_\theta(x) \\
log\ p_\theta(x) &= -E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(z,x)] + D_{KL}(q_\phi \,||\, p_\theta)
\end{aligned}
\tag{6}
$$

Here, $D_{KL}(q_\phi \,||\, p_\theta)$ 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.

$$
\begin{aligned}
ELBO &= -E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(z,x)] \\
&= -E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(x \mid z)] + E_{q_\phi}[log\ p_\theta(z)] \\
&= E_{q_\phi}[log\ p_\theta(x \mid z)] - E_{q_\phi}[log\ q_\phi(z \mid x)] + E_{q_\phi}[log\ p_\theta(z)] \\
ELBO &= E_{q_\phi}[log\ p_\theta(x \mid z)] - E_{q_\phi}[log\frac{q_\phi(z \mid x)}{p_\theta(z)}]
\end{aligned}
\tag{7}
$$

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-

$$\begin{aligned} D_{KL}(q_\phi \mid\mid p_\theta) &= E_{q_\phi}[log \frac{q_\phi(z \mid x)}{p_\theta(z)}] \\ &= E_{q_\phi}[log\ q_\phi(z \mid x)] - E_{q_\phi}[log\ p_\theta(z)] \\ &= 1/2 \sum_d (1 + log(\Sigma) - \mu^2 - \Sigma) \end{aligned} \tag{8}$$

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):

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

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

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

$$\tag{12}$$

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
```

References

ANNOY library. https://github.com/spotify/annoy. Accessed: 2022-03-01.

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.

Xin Shao, Jie Liao, Chengyu Li, Xiaoyan Lu, Junyun Cheng, and Xiaohui Fan. Celltalkdb: a manually curated database of ligand–receptor interactions in humans and mice. *Briefings in bioinformatics*, 22(4):bbaa269, 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.