Bayesian information sharing enhances detection of regulatory associations in rare cell types.

Machine Learning in Practice Reading Group

Duke B&B

February 23, 2024

Presented by Aditya Parekh

Introduction: Gene Regulatory Networks

- Gene regulatory networks (GRN) are graphs which model the regulation of gene expression [1]
- In a sense, cell type can be characterized by the GRNs present within the cell

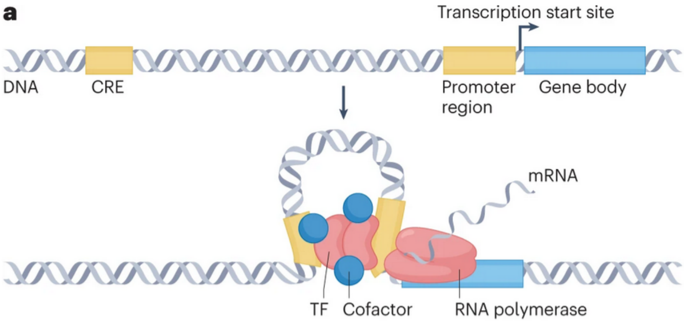

Figure: Gene expression is regulated by interactions between transcription factor proteins and DNA [1]

Introduction: Inference of GRNs

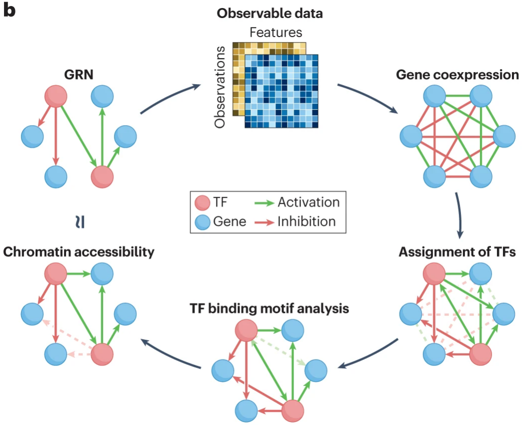

Figure: GRN inference workflow [1]

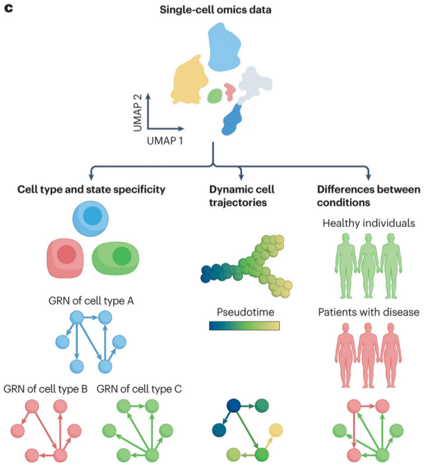

Figure: GRN inference from gene expression data [1]

Introduction: Inference of GRNs

Methods for GRN inference

- Many methods for inferring GRNs
- Two commonly used methods: GENIE3 [5] and PIDC [4]
- Can also simply consider the coexpression graph (edges weighted by pearson correlation)

Challenges

- Some cell types are rare - not much data available for GRN estimation in these cells
- Sparse data (gene dropout) - need more data for reliable coexpression estimation
- High-dimensional data - need more data

Intuition

- Exploit biological similarities between cell types to inform patterns of gene regulation
- In other words, share info on regulatory patterns between cell types

Introduction: shareNET

Central Problem

- Suppose we have a (black-box) method to construct a GRN
- Consider edge $e^{(c)} = (i, j, w_c)$ between gene i and gene j with weight w in cell type c
- How can we improve our estimation of w_c? (Suppose c is rare)

Approach

- Bayesian information sharing framework which shares information between different cell types
- Interaction patterns in common cell types inform patterns in rare cell types
- Patterns are adaptively learned from the data

Implementation

- Variational inference (again... dun dun dun)

Background: Variational Inference

Idea: approximate an intractable posterior distribution with a simpler distribution

- Suppose data X and we wish to estimate Z
- Posterior: $p(Z|X) \propto p(X|Z)p(Z)$
- Suppose posterior distribution $q(Z) = p(Z|X)$ is intractable
- Then approximate it with a simpler distribution $q^*(Z)$

Mean-Field Approximation: If $Z = (Z_1, \ldots, Z_K)$, then a possible choice is

$$q^*(Z) = \prod_{i=1}^{K} q_i^*(Z_i),$$

where q_i^* is the marginal distribution of Z_i.

Background: Variational Inference

How to find q^*? Let's do some math. The posterior is given by

$$p(Z|X) = \frac{p(X,Z)}{p(X)},$$

which implies

$$p(X) = \frac{p(X,Z)}{p(Z|X)}.$$

Now consider the marginal distribution $q(Z)$. Write

$$\begin{aligned} p(X) &= \frac{p(X,Z)}{p(Z|X)} \frac{q(Z)}{q(Z)} \\ &= \frac{p(X,Z)}{q(Z)} \frac{q(Z)}{p(Z|X)}. \end{aligned}$$

Background: Variational Inference

Taking the log on both sides, we obtain

$$\log p(X) = \log \frac{p(X,Z)}{q(Z)} + \log \frac{q(Z)}{p(Z|X)}.$$

Taking expectations with respect to q on both sides,

$$\begin{aligned}\mathbb{E}_q[\log p(X)] &= \mathbb{E}_q\left[\log \frac{p(X,Z)}{q(Z)}\right] + \mathbb{E}_q\left[\log \frac{q(Z)}{p(Z|X)}\right] \\ &= \mathbb{E}_q\left[\log \frac{p(X,Z)}{q(Z)}\right] + KL\left((q(Z)||p(Z|X)\right)\end{aligned}$$

Notice that the $KL \geq 0$, with equality obtained if and only if $q(Z) = p(Z|X)$.

Background: Variational Inference

Replacing q with the restricted class of distributions q^*, we have

$$\mathbb{E}_{q^*}\left[\log p(X)\right] = \mathbb{E}_{q^*}\left[\log \frac{p(X,Z)}{q^*(Z)}\right] + KL\left((q^*(Z)||p(Z|X)\right)$$

Thus, we choose q^* by minimizing the KL divergence between q^* and $p(Z|X)$.

Of course, this is impossible to do directly, but minimizing the KL divergence is equivalent to maximizing the quantity

$$\mathbb{E}_{q^*}\left[\log \frac{p(X,Z)}{q^*(Z)}\right] = \int q^*(Z) \log \frac{p(X,Z)}{q^*(Z)} dZ.$$

This quantity is known as the Evidence Lower Bound (ELBO).

Background: Variational Inference (Mean-Field)

If $q^* = \prod_i q_i^*(Z_i)$, then we can derive the optimal choice of q^* in closed form. Write

$$\begin{aligned}\mathbb{E}_{q^*}\left[\log\frac{p(X,Z)}{q^*(Z)}\right] &= \int \prod_{i=1}^{K} q_i^*(Z_i)\left[\log\frac{p(X,Z)}{\prod_{i=1}^{K} q_i^*(Z_i)}\right]dZ \\ &= \int \prod_{i=1}^{K} q_i^*(Z_i)\left[\log p(X,Z)\right]dZ - \int \prod_{i=1}^{K} q_i^*(Z_i)\sum_{i=1}^{K}\log q_i^*(Z_i)dZ\end{aligned}$$

To maximize the ELBO with respect to a single factor q_j^*, pick out the terms that depend only on q_j^*:

$$\begin{aligned}\mathbb{E}_{q^*}\left[\log\frac{p(X,Z)}{q^*(Z)}\right] &= \int q_j^*(Z_j)\left[\int \log p(X,Z)\prod_{i\neq j} q_i^*(Z_i)dZ_{i\neq j}\right]dZ_j - \int q_j^*(Z_j)\log q_j^*(Z_j)dZ_j + \text{const.} \\ &= \int q_j^*(Z_j)\left\{\mathbb{E}_{i\neq j}\log p(X,Z)\right\}dZ_j - \int q_j^*(Z_j)\log q_j^*(Z_j)dZ_j + \text{const.},\end{aligned}$$

where $\mathbb{E}_{i\neq j}\log p(X,Z)$ is an expectation with respect to the distribution $\prod_{i\neq j} q_i^*(Z_i)$.

Background: Variational Inference (Mean-Field)

Define a new distribution $\tilde{p}(X, Z_j) \propto \exp\{\mathbb{E}_{i \neq j} \log p(X, Z)\}$. Then we can rewrite the integral from the previous slide as

$$\begin{aligned}\mathbb{E}_{q^*}\left[\log \frac{p(X,Z)}{q^*(Z)}\right] &= \int q_j^*(Z_j)\tilde{p}(X,Z_j)dZ_j - \int q_j^*(Z_j)\log q_j^*(Z_j)dZ_j \\ &= -KL\left(q_j^*(Z_j)||\tilde{p}(X,Z_j)\right).\end{aligned}$$

Thus, maximizing the evidence lower bound is equivalent to minimizing the KL divergence between $q_j^*(Z_j)$ and $\tilde{p}(X, Z_j)$, which occurs precisely when

$$q_j^*(Z_j) = \tilde{p}(X, Z_j).$$

Hence the optimal form of the factor $q_j^*(Z_j)$ is given by

$$q_j^*(Z_j) = \frac{\exp\{\mathbb{E}_{i \neq j} \log p(X,Z)\}}{\int \exp\{\mathbb{E}_{i \neq j} \log p(X,Z)\}\, dZ_j}.$$

Background: Variation Inference

Helpful reading materials on variational inference:

- Pattern Recognition and Machine Learning (Bishop, 2006)
- Variational inference: A review for statisticians (Blei, 2017)
- `https://www.cs.cmu.edu/~epxing/Class/10708-20/lectures.html` **(highly recommended)**

The above derivation was taken pretty much straight out of Pattern Recognition and Machine Learning, Ch 10.1 [2]

Methods

Central idea: Use an information sharing framework to enhance accuracy of GRN inference.

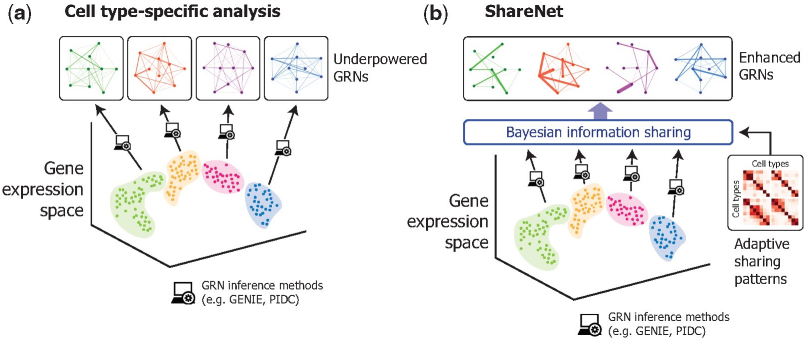
(a)
Cell type-specific analysis
Underpowered GRNs
Gene expression space
GRN inference methods (e.g. GENIE, PIDC)
(b)
ShareNet
Enhanced GRNs
Bayesian information sharing
Cell types
Cell types
Adaptive sharing patterns
Gene expression space
GRN inference methods (e.g. GENIE, PIDC)

Figure: Overview of method. [6]

By learning the interaction patterns shared across cell types, shareNET may be able to more accurately predict the presence of interaction edges within the GRN graph, as well as the interaction strength (edge weight).

Methods

Bayesian hierarchical model. The true edge weight interaction between gene i and gene j is given by z_{ij}. The collection of cell-type to cell-type sharing patterns are modeled as a Gaussian mixture with K mixing components

- $u_{ij} \sim \text{Categorical}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$
- $z_{ij} | u_{ij} \sim \mathcal{N}_C\left(\mu_{u_{ij}}, \Sigma_{u_{ij}}\right)$

Here, C is the number of cell-types under consideration. Given the true edge weight z_{ij}, our GRN inference algorithm produces a noisy observation of the edge weight in cell type c:

$$e_{ij}^{(c)} \sim \mathcal{N}\left(z_{ij}^{(c)}, \sigma_{ij}^{(c)2}\right),$$

where $z_{ij}^{(c)}$ is the c-th component of the vector z_{ij}.
In addition, the parameters $(\mu_k, \Sigma_k), k = 1, \ldots, K$ are drawn from a Normal-Wishart prior:

$$(\mu_k, \Sigma_k^{-1}) \sim NW\left(\mu_0, \beta_0, \mathbf{\Psi}, \nu\right).$$

Methods

Thus, the full hierarchical model is given as follows:

- $(\mu_k, \Sigma_k^{-1}) \sim NW(\mu_0, \beta_0, \mathbf{\Psi}, \nu)$
- $u_{ij} \sim \text{Categorical}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$
- $z_{ij} | u_{ij} \sim \mathcal{N}_C\left(\mu_{u_{ij}}, \Sigma_{u_{ij}}\right)$
- $e_{ij}^{(c)} | z_{ij} \sim \mathcal{N}\left(z_{ij}, \sigma_{ij}^{(c)2}\right)$

The cell type-to-cell type sharing pattern, contributed by the k-th mixture component, between cell type a and cell type b is encoded in the (a, b)-th entry of the covariance matrix Σ_k.

Methods

The authors also extended their model into a linear model predicting the gene expression levels of target genes based on the expression of regulator genes, using a framework called Bayesian Variable Selection originally developed for quantitative genetics [3].

Suppose we partition the genes into two categories: target genes and regulator genes. Let $\mathbf{X}^{(c)}$ be the expression levels of regulator genes in cell type c. Then $\mathbf{y}^{(c)}$ is the expression level of the target genes in cell type c. For a given target gene j, in a particular cell n, the expression level has distribution

$$y_{j,n}^{(c)} \sim \mathcal{N}\left(\sum_i \mathbf{X}_{i,n}^{(c)} \beta_{i,j}^{(c)} \gamma_{i,j}^{(c)}, \sigma_\epsilon^2\right),$$

where $\beta_{i,j}^{(c)}$ is the coefficient of the linear model describing the effect of regulator gene i on target gene j in cell type c, and $\gamma_{i,j}^{(c)}$ is a binary variable indicating whether or not gene i is to be included in the GRN for cell type c.

Methods

The additional parameters are modeled as

- $\gamma_{i,j}^{(c)} \sim \text{Bernoulli}\left(g(z_{ij}^{(c)})\right)$
- $\beta_{i,j} | \gamma_{i,j}^{(c)} \sim \begin{cases} \mathcal{N}\left(0, \sigma_\beta^2\right) & \text{if } \gamma_{i,j}^{(c)} = 1 \\ \delta_0(\beta_{i,j}^{(c)}) & \text{otherwise.} \end{cases}$

Thus the full model hierarchical model for Bayesian Variable Selection is

- $(\mu_k, \Sigma_k^{-1}) \sim NW(\mu_0, \beta_0, \mathbf{\Psi}, \nu)$
- $u_{ij} \sim \text{Categorical}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$
- $z_{ij} | u_{ij} \sim \mathcal{N}_C\left(\mu_{u_{ij}}, \Sigma_{u_{ij}}\right)$
- $\gamma_{i,j}^{(c)} \sim \text{Bernoulli}\left(g(z_{ij}^{(c)})\right)$
- $\beta_{i,j} | \gamma_{i,j}^{(c)} \sim \begin{cases} \mathcal{N}\left(0, \sigma_\beta^2\right) & \text{if } \gamma_{i,j}^{(c)} = 1 \\ \delta_0(\beta_{i,j}^{(c)}) & \text{otherwise.} \end{cases}$
- $y_{j,n}^{(c)} \sim \mathcal{N}\left(\sum_i \mathbf{X}_{i,n}^{(c)} \beta_{i,j}^{(c)} \gamma_{i,j}^{(c)}, \sigma_\epsilon^2\right)$

Methods

Returning to the shareNET model, the joint distribution is

$$p(\mathbf{u}, \mu, \Sigma, \mathbf{z}, \mathbf{e}) = \prod_{k=1}^{K} p(\mu_k, \Sigma_k^{-1}) \prod_{n=1}^{N} \prod_{k=1}^{K} p(z_n|\mu_k, \Sigma_k^{-1})^{u_{n,k}} \prod_{n=1}^{N} p(u_n) p(e_n|z_n, D_n^{-1}),$$

where $D_n^{-1} = \mathrm{diag}(1/\hat{\sigma}_1^{(1)2}, \dots, 1/\hat{\sigma}_n^{(C)2})$.
The goal is to compute the posterior:

$$\begin{aligned} p(\mathbf{z}, \mathbf{u}, \mu, \Sigma^{-1}|\mathbf{e}) &= \frac{p(\mathbf{u}, \mu, \Sigma, \mathbf{z}, \mathbf{e})}{p(\mathbf{e})} \\ &= \frac{\prod_k p(\mu_k, \Sigma_k^{-1}) \prod_n \prod_k p(z_n|\mu_k, \Sigma_k^{-1})^{u_{n,k}} \prod_n p(u_n) p(e_n|z_n, D_n^{-1})}{\int\int\int\int \prod_k p(\mu_k, \Sigma_k^{-1}) \prod_n \prod_k p(z_n|\mu_k, \Sigma_k^{-1})^{u_{n,k}} \prod_n p(u_n) p(e_n|z_n, D_n^{-1}) d\mu d\Sigma d\mathbf{u} d\mathbf{z}}, \end{aligned}$$

the denominator being obviously intractable.

Methods

The parameters of the shareNET model are $\mathbf{U} = (\mu, \Sigma, \mathbf{u}, \mathbf{z})$. The authors approximated the posterior distribution with the following mean-field approximation:

$$q(\mu, \Sigma, \mathbf{u}, \mathbf{z}) = \prod_{k=1}^{K} q(\mu_k, \Sigma_k) \prod_{n=1}^{N} q(z_n) q(u_n),$$

where

- $q(\mu_k, \Sigma_k^{-1}) = NW(\mu_k, \Sigma_k^{-1} | \tilde{\alpha}_k, \tilde{\beta}_0, \tilde{B}_k, \tilde{\nu}_k)$
- $q(z_n) = \mathcal{N}(z_n | \tilde{m}_n, \tilde{S}_n)$
- $q(u_n) = \text{Categorical}(\tilde{\psi}_n)$.

The parameters are estimated using CAVI.

Methods

Coordinate Ascent Variational Inference: the idea, as we have shown above, is that the optimal form of the variational factor q_j is given by

$$q_j^* \propto \exp\left\{\mathbb{E}_{i\neq j}\left[p(\mathbf{u}, \mu, \Sigma, \mathbf{z}, \mathbf{e})\right]\right\}.$$

We can then write

$$\log q_j^* = \mathbb{E}_{i\neq j}\left[p(\mathbf{u}, \mu, \Sigma, \mathbf{z}, \mathbf{e})\right] + \text{const.}$$

Writing out the expectation for each of the factors q_j in the ShareNET model, we solve the resulting system of equations to derive the iterative updates to the variational parameters $\tilde{\alpha}_k, \tilde{\beta}_0, \tilde{B}_k, \tilde{\nu}_k, \tilde{m}_n, \tilde{S}_n$, and $\tilde{\psi}_n$.

In a lot of ways, CAVI is like a deterministic version of Gibbs Sampling.

Methods

As an example of using CAVI, let's derive the variational updates for $q(\mathbf{u}_n)$ (section 2.2.5 of the supplementary material). We write

$$\begin{aligned}
\log q^*(\mathbf{u}_n) &= \mathbf{E}_{q(\mu,\Sigma^{-1},\mathbf{u}_{-n},\mathbf{z})}\left[\log p(\mu,\Sigma^{-1},\mathbf{z},\mathbf{u},\mathbf{e})\right] + \text{const.} \\
&= \log p(\mathbf{u}_n) + \mathbb{E}_{q(\mu,\Sigma^{-1},z)}\left[\sum_{k=1}^{K} u_{n,k}\log p\left(\mathbf{z}_n|\mu_k.\Sigma_k^{-1}\right)\right] + \text{const.} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} u_{n,k}\log\rho_{n,k} + \text{const.},
\end{aligned}$$

where

$$\log p_{n,k} = \log\frac{1}{K} + \frac{1}{2}\mathbb{E}_{q(\Sigma^{-1})}\left[\log|\Sigma_k^{-1}|\right] - \frac{C}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_{q(\mu,\Sigma^{-1},\mathbf{z})}\left[(\mathbf{z}_n-\mu_k)^T\Sigma_k^{-1}(\mathbf{z}_n-\mu_k)\right].$$

Methods

Exponentiating both sides, we have

$$q(\mathbf{u}_n) \propto \prod_{k=1}^{K} \rho_{n,k}^{u_{n,k}},$$

which is a categorical distribution when normalized. Therefore,

$$q(\mathbf{u}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \tilde{\phi}_{n,k}^{u_{n,k}},$$

where

$$\prod_{k=1}^{K} \tilde{\phi}_{n,k} = \frac{\rho_{n,k}}{\sum_{j=1}^{K} \rho_{n,j}}.$$

Note that this is not a closed-form solution, since $\rho_{n,k}$ depends on expectations computed with respect to $q(\mu, \Sigma^{-1}, \mathbf{z})$, Those expectations can be found once we've found the functional forms of the variational posteriors $q(\mathbf{z})$ and $q(\mu, \Sigma^{-1})$.

Section 6: Experimental Results

Three scRNA-seq datasets:

- BEELINE
- Tabula Moris
- mouse blood lineage

Estimate cell-type specific GRN networks built from these datasets using three methods and compare with reference networks from various biological databases.

Comparison: edge present or not?

Section 6: Experimental Results

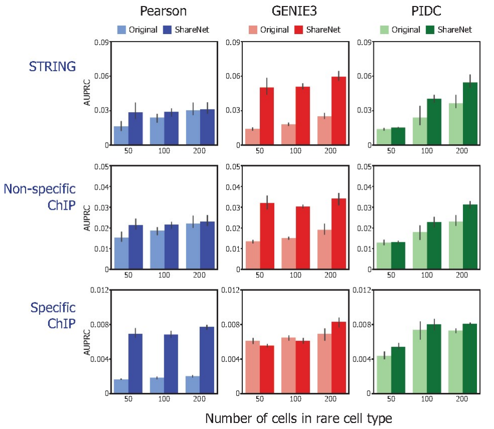

Figure: AAUPRC results for inferred GRN networks with and without shareNET on downsampled dendritic cells from the BEELINE dataset.

Section 6: Experimental Results

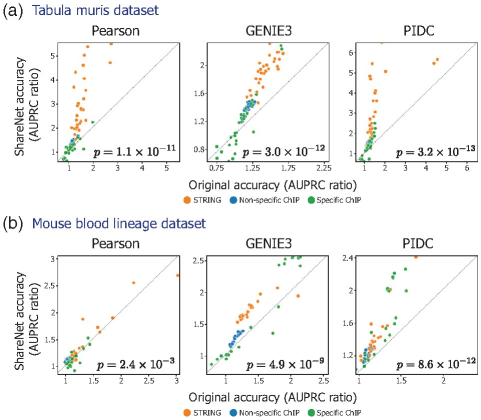

Figure: AUPRC results for inferred GRN networks with and without shareNET on Tabula muris and Mouse blood lineage datasets.

Section 6: Experimental Results

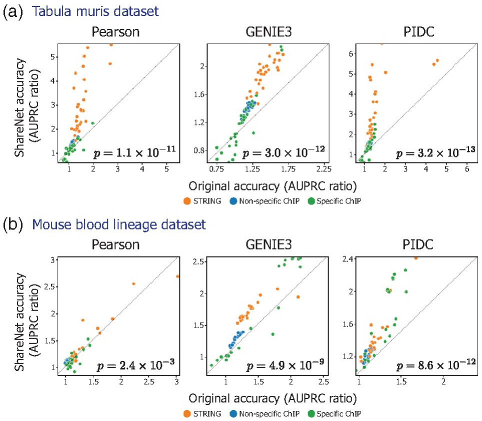

Figure: AUPRC results for inferred GRN networks with and without shareNET on Tabula muris and Mouse blood lineage datasets.

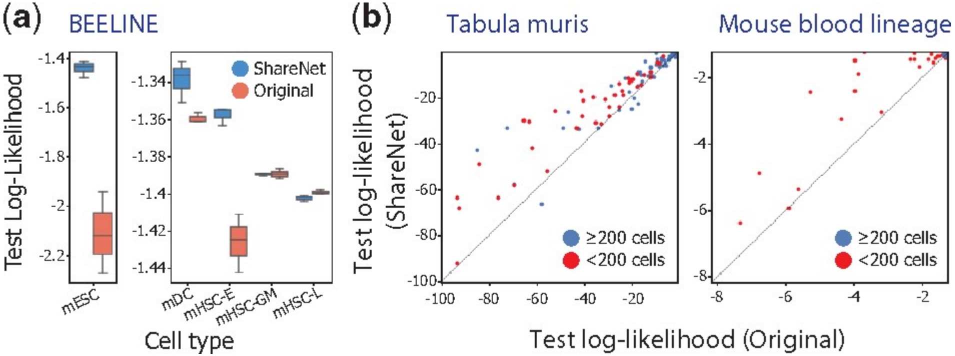

Figure: Holdout Likelihood Analysis results using BVS framework.

Section 6: Experimental Results

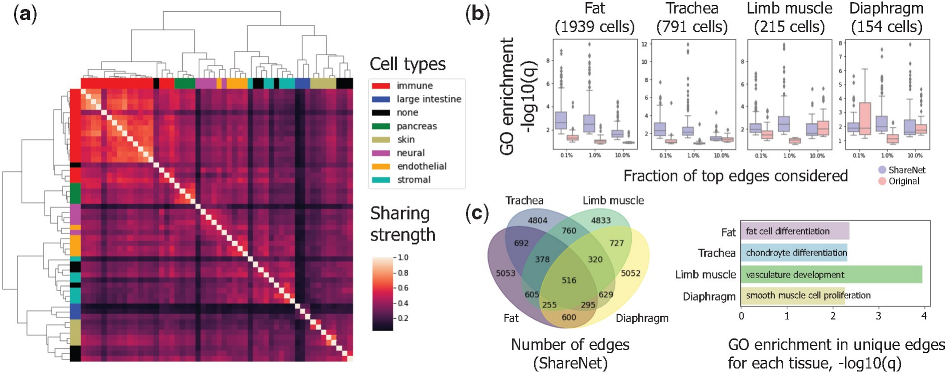

Figure: o(a) Plot of the Σ_k matrix corresponding to the highest mixture component. (b) GO Enrichment results of mesenchymal stem cells (MSCs) obained from four different tissues. (c) GO Enrichment results on top GRN edges present in one tissue but not in others.

Section 6: Experimental Results

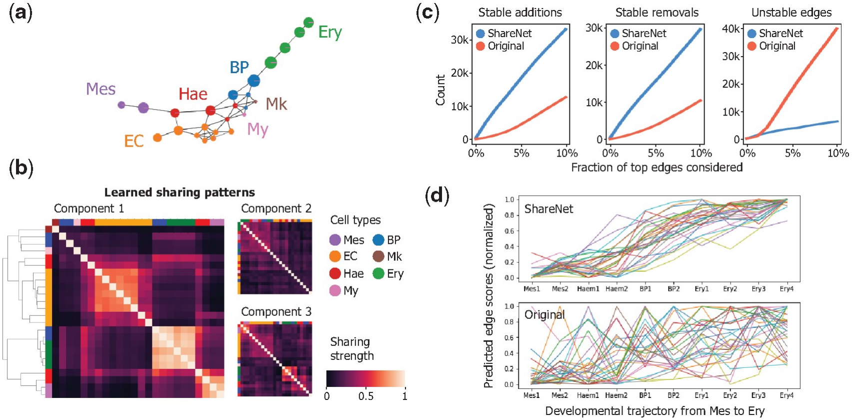

Figure: (a) Inferred developmental trajectory of cell types in the mouse blood linege dataset. (b) The Σ_k matrix corresponding to the highest mixture component in this dataset. (c) Quantification of stable vs. unstable edges along developmental trajectory. (d) Scores of stably added edges along developmental trajectory.

Section 7: Recommendations

Conclusions

- Elegant, fairly interpretable model for sharing biological information across cell types.
- Analysis seems to support that the shareNET framework improves accuracy of inferred GRNs
- Would have been nice to see some simulation results, since no ground truth in experiments

Recommendations

- Worth reading? Section 2 of the paper and section 3 of the supplementary materials may be useful to those looking to learn some more about variational inference.
- Worth implementing? For me, yes. The model is not overly complicated and could be instructional for how to implement Bayesian methods from scratch. The code is also available for download online.

References

[1] Pau Badia-i-Mompel et al. "Gene regulatory network inference in the era of single-cell multi-omics". In: *Nature Reviews Genetics* (2023), pp. 1–16.

[2] Christopher Bishop. "Pattern recognition and machine learning". In: *Springer google schola* 2 (2006), pp. 5–43.

[3] Peter Carbonetto and Matthew Stephens. "Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies". In: (2012).

[4] Thalia E Chan, Michael PH Stumpf, and Ann C Babtie. "Gene regulatory network inference from single-cell data using multivariate information measures". In: *Cell systems* 5.3 (2017), pp. 251–267.

[5] Vân Anh Huynh-Thu et al. "Inferring regulatory networks from expression data using tree-based methods". In: *PloS one* 5.9 (2010), e12776.

[6] Alexander P Wu et al. "Bayesian information sharing enhances detection of regulatory associations in rare cell types". In: *Bioinformatics* 37.Supplement_1 (2021), pp. i349–i357.