SCNet: Automatic Multi-Channel Genome Network Inference from Single-Cell RNA Sequences

Shiyin Wang (wangshiy16@mails.tsinghua.edu.cn)

December 23, 2019

Outline

1 Introduction
- Single-Cell RNA Sequencing
- Mathematics Preliminaries

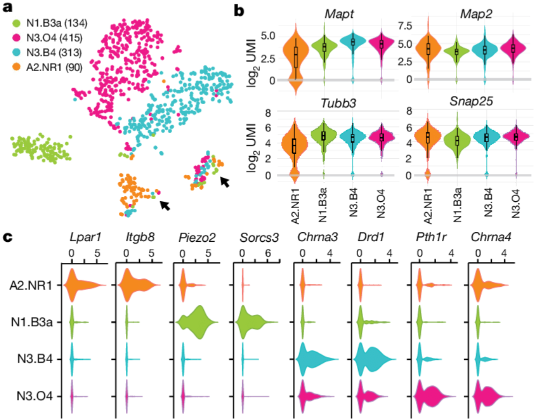
2 Methods
- Probabilistic models
- Optimization

3 Results

4 Future Work

5 Q & A

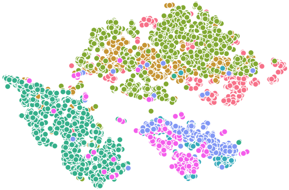
What is single-cell RNA sequencing?



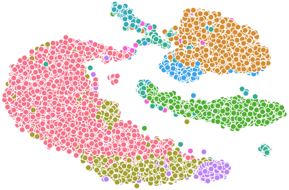
Tsunemoto, Rachel, et al. "Diverse reprogramming codes for neuronal identity." Nature 557.7705 (2018): 375.

Datasets

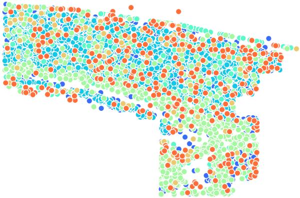
Cortex (3005 cells, 55g genes, 7 types)



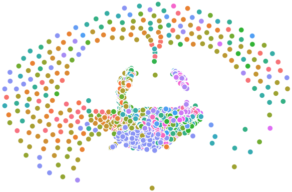
PBMC (11990 cells, 3345 genes, 9 types)



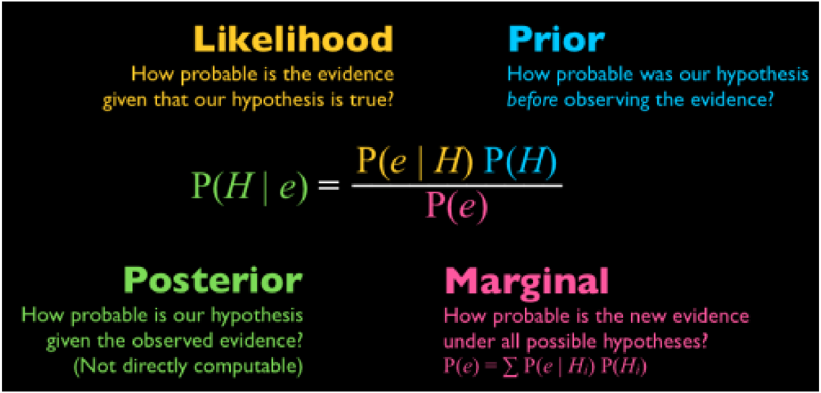
osmFISH (4530 cells, 33 genes, 6 types)



Hemato (4016 cells, 7397 genes, 101 types)



Bayesian Methods



Likelihood

How probable is the evidence given that our hypothesis is true?

Prior

How probable was our hypothesis *before* observing the evidence?

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

Posterior

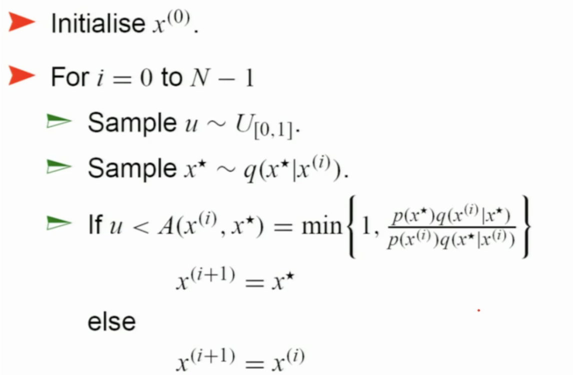
How probable is our hypothesis given the observed evidence? (Not directly computable)

Marginal

How probable is the new evidence under all possible hypotheses?

$P(e) = \sum P(e \mid H_i)\, P(H_i)$

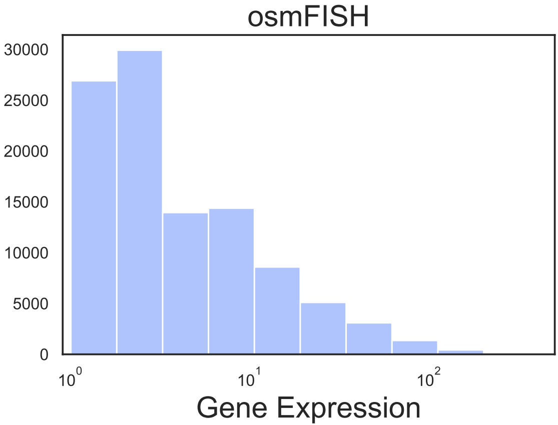
Markov Chain Monte Carlo



```
- Initialise $x^{(0)}$.
- For $i = 0$ to $N-1$
    - Sample $u \sim U_{[0,1]}$.
    - Sample $x^{\star} \sim q(x^{\star}|x^{(i)})$.
    - If $u < A(x^{(i)}, x^{\star}) = \min\left\{1, \frac{p(x^{\star})q(x^{(i)}|x^{\star})}{p(x^{(i)})q(x^{\star}|x^{(i)})}\right\}$
            $x^{(i+1)} = x^{\star}$
      else
            $x^{(i+1)} = x^{(i)}$
```

Slide in CPSC 540, taught in 2013 at UBC by Nando de Freitas

Choose Log-Scale to Model Gene Expression



Bayesian Hierarchical Linear Model

X and Y are the expression profiles of two genes. The edge weights in the desired network correspond to the regression coefficient *k* in the equation.

$$\begin{aligned} & log(Y+1) = klog(X+1) + \epsilon \\ & k \sim N(\beta, \sigma) \\ & \epsilon \sim N(0, \gamma^2) \end{aligned} \tag{1}$$

This model estimates k from single-cell RNA sequencing records.

Define Transition Probability through Edge Weights

Once we have a weighted network, we can define the association probability between two nodes X and Y by a very trivial model.

$$Pr(X \to Y) = 1 - (1 - w_{X,Y}) \Pi_{a \in V} (1 - w_{X,a} w_{a,Y}) \Pi_{a \in V, b \in V} (1 - w_{X,a} w_{a,b} w_{b,Y}) \tag{2}$$

For simplicity, I expanded the search for three steps. Walk length is flexible to choose.

Maximal Likelihood Optimization

Bayesian hierarchical linear model and transition probability explain the network dynamics from two different angles. Now we can combine them together to infer the edge weights of networks (V, E, W).

$$maximize\ L(W;_k, \Sigma_k) = \sum_{v_1 \in V} \sum_{v_2 \in V} Pr_{N(_k, \Sigma_k)}(w = Pr(v_1 \rightarrow v_2)) \tag{3}$$

Add regulation term $\lambda|V|$ to restrain the number of edges in the network.

$$maximize\ L(W;_k, \Sigma_k) = \sum_{v_1 \in V} \sum_{v_2 \in V} Pr_{N(_k, \Sigma_k)}(w = Pr(v_1 \rightarrow v_2)) - \lambda|V| \tag{4}$$

Data Preprocessing - Normalization

1. Retained the top genes ordered by variance as in [Lopez *et al.*, 2018][1]
2. Normalize genes to standard Gaussian distribution $N(0,1)$

Reason for normalization to standard deviation: So that the regression coefficient in $Y = kX + \epsilon$ is $\hat{k} = \frac{cov(X,Y)}{var(X)} = cov(X,Y)$, which is exchangeable and bidirectional.

[1] Lopez, R., Regier, J., Cole, M.B. et al. Deep generative modeling for single-cell transcriptomics. Nat Methods 15, 1053–1058 (2018) doi:10.1038/s41592-018-0229-2

Why use Poisson Distribution

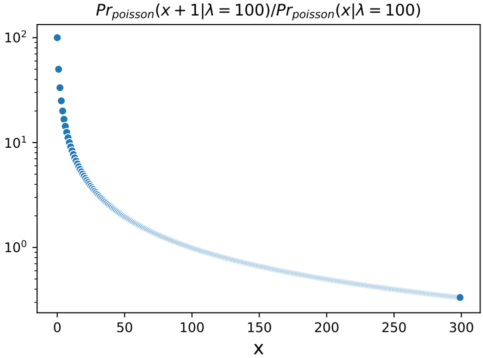


Figure: This fraction changes rapidly when x is far away from λ, while not significant when x is close to λ. By adjusting parameter λ, we can guide the sparsity of the network.

Choose Log-Scale to Model Gene Expression

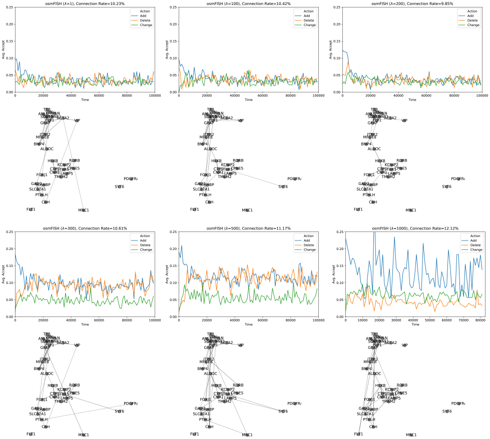


Figure: The convergence curves and results on the osmFISH dataset.

Possible Directions

- Theoretical analysis and case study to compare with other methods that derive networks from covariance matrix directly
- Integrate existing knowledge from protein-protein interaction networks (STRING, OmniPATH, ConsensusPath, etc) as priors
- Design better probability models
- Make interactive transition videos

Resources

- 10x Genomics: Datasets providing single cell and spatial views of biological systems (`https://www.10xgenomics.com`)

Questions & Answers