Paper Presentation

Weize Xu

HuaZhong Agricultural University

July 1, 2018

Paper information

Title:
scEpath: Energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data

Authors:
Jin, MacLean, Peng, and Nie

Journal and year:
Bioinformatics, 2018

Abstract

Motivation

Single-cell RNA-sequencing (scRNA-seq) offers unprecedented resolution for studying cellular decision-making processes. Robust inference of cell state transition paths and probabilities is an important yet challenging step in the analysis of these data.

Results

- A robust algorithm that calculates energy landscapes and probabilistic directed graphs in order to reconstruct developmental trajectories.
- Identified marker genes and gene expression patterns associated with cell state transitions.
- scEpath allows us to identify common and specific temporal dynamics and transcriptional factor programs along branched lineages, as well as the transition probabilities that control cell fates.

Background

Tasks

- Identification of functionally relevant (sub)populations of cells.
- Cell state transitions along developmental or other trajectories.
- Hierarchical lineage relationships (e.g. stem cell differentiation) and pseudo-temporal ordering.

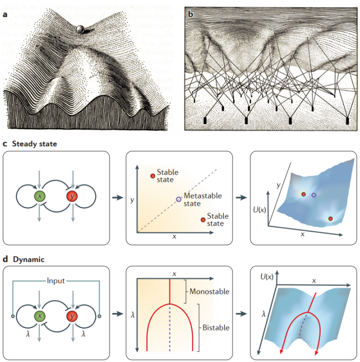
Former research address these tasks

- MST(minimal span tree) based method, (Monocle[9], TSCAN[4])
- Reverse graph embedding (Monocle2[7])
- Diffusion-like random walks (DPT[3])
- Neighborhood-based cell state transitions (Mpath[2])
- Probabilistic graphical model (TASIC[8])

Basic Idea

Waddington landscape

The metaphorical epigenetic landscape conceived by Waddington is frequently used to depict or describe cell fate decision-making processes.[6]



Mapping the quantitative energy landscape of single-cell dynamical processes using statistical physics modeling, such that we can obtain transition probabilities between cell states, reconstructed lineages and pseudotemporal ordering of cells.

Methods

Preprocessing

Input: $X = (x_{ij})$
x_{ij} denote the expression of j-th gene/transcript in i-th cell. Can be *TPM*, *FPKM* or *UMI* values.

Pseudocount:
$X \leftarrow log_2(X + 1)$

Construction of gene-gene interaction network

Adjacency matrix: $A = (a_{ij})$

where a_{ij} takes value 1 or 0 depending on the presence whether node i and j are linked or not:

$$a_{ij} = \begin{cases} 1, & \text{if } |cor(x_i, x_k)| > \tau \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Where τ is the threshold parameter. $cor(x_i, x_j)$ is the Spearman correction between expression profiles of $x_i = (x_{i1}, xi2, ..., xim)$ and $x_j = (x_{k1}, x_{k2}, ..., x_{km})$.

Selection of parameter τ

Most biological networks's node connectivities follow a power law[1]:

$$p(k) \sim k^{-\gamma}$$

Plot the $log10(p(k))$ vs $log10(k)$, a straight line is indicative of scale-free topology. Here, use the R^2 as the measurement of how well a network satisfies a scale-free topology.

Calculation of single cell energy (scEnergy)

A statistical physics-based approach to quantitatively measure developmental states of single cells by deriving energy landscapes from single cell transcriptome data.
The states of a cell j containing n genes is represented by a random vector $Y_j = \left(Y_{1j}, Y_{2j}, ..., Y_{nj}\right)$, where Y_{ij} indicates the expression of gene i in cell j. Y_{ij} is modeled by Boltzmann-Gibbs distribution:

$$p_j(y) = \frac{e^{-E_j(y)}}{\sum_{i=1}^{m} e^{-E_i(y)}}$$

Where $p_j(y)$ is the probability that system will be in a cell state j with the gene expression pattern y, $E_j(y)$ is the scEnergy of cell j and m is the number of states accessible to the system. e.g. the number of cells.

Calculation of single cell energy (scEnergy)

If the energy of a gene depends on its expression, then it should also depend on the expression levels of genes that are closely interacting with it.

The scEnergy of a cell j with the expression pattern y was given by:

$$E_j(y) = \sum_{i=1}^{n} E(y) = -\sum_{i=1}^{n} y_{ij} \ln \frac{y_{ij}}{\sum_{k \in N(i)} y_{kj}} y_{ij} \tag{2}$$

Where, $N(i)$ is the neighborhood of node i in the network. And, define $E_{ij} = 0$, when $y_{ij} = 0$.

Calculation of single cell energy (scEnergy)

Expression rescaling with:

$$y_{ij} = (x_{ij} - x_{.j}^{min})/(x_{.j}^{max} - x_{.j}^{min})$$

Where, $x_{.j}^{min}$ and $x_{.j}^{max}$ are the minimum and maximum of the expression in the cell j.
Moreover, we define the normalized scEnergy (taking values between 0 and 1) as:

$$\hat{E}_j(y) = \frac{\left(E_j(y)/\overline{E}(y)\right)^2}{1 + \left(E_j(y)/\overline{E}(y)\right)^2} \tag{3}$$

Where, $\overline{E}(y)$ is the average scEnergy across all the cell.