# SCODE Documentation

## About the algorithm:

This method uses linear ordinary differential equations to represent how a regulatory network results in observed gene expression dynamics. SCODE relies on a specific relational expression that can be estimated efficiently using linear regression. In combination with dimension reduction, this approach leads to a considerable reduction in the time complexity of the algorithm.

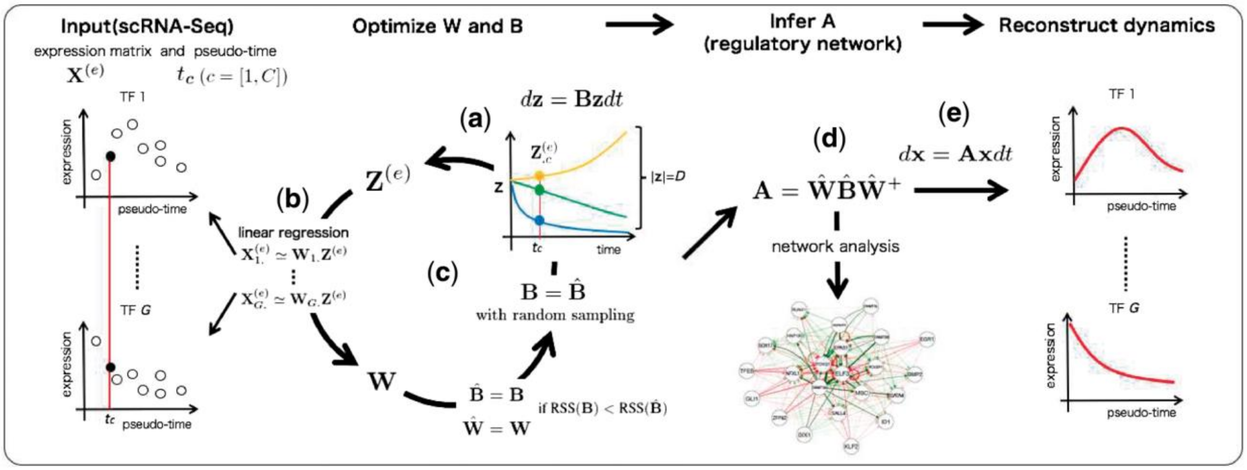
## About pseudotime:

The analysis of RNA-Seq data from individual differentiating cell enables us to reconstruct the differentiation process and the degree of differentiation (in pseudo-time) of each cell. Such analyses can reveal detailed expression dynamics and functional relationships for differentiation. The pseudo-time can be regarded as time information and therefore single-cell RNA-Seq data are time-course data.

In analyses by scRNA-Seq, the reconstruction of cellular differentiation processes attracts attention as a novel approach to revealing differentiation mechanisms. The differentiation process can be reconstructed using dimensional reduction and stochastic processes, and the degree of differentiation (in pseudo-time) of each cell is characterized by the position in the reconstructed process. By investigating the expression pattern in pseudo-time, genes can be clustered into multiple groups with different biological functions. In addition, scRNA-Seq also enables the calculation of accurate correlations of expression between genes because scRNA-Seq can distinguish the detailed states of individual cells without contamination from multiple cell types. The accurate co-expression pattern of each cell type (progenitor cells and multiple types of differentiated cells) can reveal the key regulatory factors for lineage programming. In this way, expression dynamics in pseudo-time and accurate relationships among genes can be inferred from scRNA-Seq data. For the next step in differentiation analyses using scRNA-Seq, it is important to reveal the regulatory interactions among genes that bring about the observed expression dynamics during differentiation, namely, gene regulatory network (GRN) inference from scRNA-Seq data. Pseudo-time can be regarded as time information, and hence, scRNA-Seq performed on cells undergoing differentiation can be regarded as time-course expression data at a high temporal resolution.

## About ODE:

Ordinary differential equations have been used to describe regulatory network and expression dynamics. ODEs can describe continuous variables over continuous time and the underlying physical phenomena, and therefore they are suitable for inferring GRN from scRNA-Seq during differentiation. SCODE is a highly efficient optimization algorithm that is used to describe regulatory networks and expression dynamics with linear ODEs for scRNA-Seq performed on differentiating cells by integrating the transformation of linear ODEs and linear regression. Linear ODEs can be transformed from fixed-parameter linear ODEs if they satisfy a relational expression. The relational expression can be estimated analytically and efficiently by linear regression. SCODE uses a small number of factors to reconstruct expression dynamics, which results in a marked reduction of time complexity.


Input(scRNA-Seq)
expression matrix and pseudo-time
$\mathbf{X}^{(e)}$
$t_c \ (c = [1, C])$
TF 1
expression
pseudo-time
TF $G$
expression
$t_c$
pseudo-time
Optimize W and B
(a)
$d\mathbf{z} = \mathbf{B}\mathbf{z}dt$
$\mathbf{Z}^{(e)}$
$\mathbf{Z}^{(e)}_{.c}$
z
$t_c$
time
$|\mathbf{z}|=D$
(b)
linear regression
$\mathbf{X}^{(e)}_{1.} \simeq \mathbf{W}_{1.}\mathbf{Z}^{(e)}$
$\mathbf{X}^{(e)}_{G.} \simeq \mathbf{W}_{G.}\mathbf{Z}^{(e)}$
(c)
$\mathbf{B} = \hat{\mathbf{B}}$
with random sampling
W
$\hat{\mathbf{B}} = \mathbf{B}$
$\hat{\mathbf{W}} = \mathbf{W}$
if $\mathrm{RSS}(\mathbf{B}) < \mathrm{RSS}(\hat{\mathbf{B}})$
Infer A
(regulatory network)
(d)
$\mathbf{A} = \hat{\mathbf{W}}\hat{\mathbf{B}}\hat{\mathbf{W}}^{+}$
network analysis
Reconstruct dynamics
(e)
$d\mathbf{x} = \mathbf{A}\mathbf{x}dt$
TF 1
expression
pseudo-time
TF $G$
expression
pseudo-time

**(a)** Sample $Z^e$ from the ODE of z (**b**) Estimate W based on linear regression. (**c**) Optimize B iteratively. (**d**) Infer A from optimized W and B. (**e**) The expression dynamics can be reconstructed from the optimized ODE of x.

### Describing regulatory networks and expression dynamics with linear ODEs

First, describe transcription factor expression dynamics throughout differentiation with linear ODEs: *d*x = Ax*dt*, where x is a vector of length G (G is the number of TFs) that denotes the expression of TFs and A is a square matrix with dimensions equal to G that denotes the regulatory network among TFs. Infer the TF regulatory network by optimizing A such that the ODE can successfully describe the observed expression data.

The observed expression data consist of a G x C matrix ($X^{(e)}$ ), where C is the number of cells. So objective is to optimize A such that *d*x = Ax*dt* can properly represent $X^{(e)}$ at a corresponding time point. Here, A contains G x G parameters.

### Deriving A from a linear ODE transformation

Consider the linear ODE: *dz* = Bz*dt*, where z is a vector of length G and B is a known square matrix. If we know a matrix W that satisfies x=Wz, can derive the ODE of x by transforming the ODE of z as follows: $dz = \mathrm{B}zdt \rightarrow dz = \mathrm{BW}^{-1}\mathrm{W}zdt \rightarrow \mathrm{W}dz = \mathrm{WBW}^{-1}\mathrm{W}zdt \rightarrow dx = \mathrm{WBW}^{-1}xdt$

### Estimating W using linear regression

To infer A, have to estimate a matrix W that satisfies x = Wz. W can be optimized by linear regression for each TF, and A can be efficiently calculated from $\mathrm{WBW}^{-1}$

### Dimension reduction of z

The basic idea of reduction is that the patterns of expression dynamics are limited, and expression dynamics can be reconstructed with a small number of patterns.

### Optimizing B

Evaluate the appropriateness of the matrix B with the residual sum of squares. We assume B is a diagonal matrix and optimize B by random sampling and iterative optimization so that the RSS decreases.

### Parameter optimization in SCODE

1. Initialize a diagonal matrix B randomly, and set B_hat to B.
2. Generalize $Z^{(e)}$ from the ODE of z determined from B.
3. Optimize W based on linear regression, and calculate RSS(B, W)
4. If RSS(B,W)<RSS(B_hat, W_hat), we update B_hat with B and W_hat with W.
5. Set B to B_hat.
6. Sample i and B uniformly.
7. Return to step 2 until reaching maximum iteration.
8. After iterative optimization, A is inferred from WBW.

### Input files

1. A G x C matrix of expression data where each row corresponds to each gene and each column corresponds to each cell.
2. A time point (pseudo-time data) containing the time point data of each cell. The first column is the information of a cell (e.g. index of a cell, experimental time point), and the second column is the time parameter (normalized from 0 to 1.0).

### Output files

1. A G x G matrix, which corresponds to inferred regulatory network. Aij represents the regulatory relationship from transcription factor j to transcription factor i. G is the number of transcription factors.
2. A G x D matrix, which corresponds to W of linear regression. D is the number of z.
3. The residual sum of squares of linear regression.

## References

https://academic.oup.com/bioinformatics/article/33/15/2314/3100331

https://github.com/hmatsu1226/SCODE