**SCUBA v1.0**

**1. Overview**

SCUBA stands for "Single-cell Clustering Using Bifurcation Analysis". SCUBA is a novel computational method for extracting lineage relationships from single-cell gene expression data, and modeling the dynamic changes associated with cell differentiation. SCUBA draws techniques from nonlinear dynamics and stochastic differential equation theories, providing a systematic framework for modeling complex processes involving multi-lineage specifications.

Reference: Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, Yuan GC. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proc Natl Acad Sci U S A. 2014 Dec 15. pii: 201408993. [Epub ahead of print]

**2. System Requirements**

SCUBA is independent of operating systems because it is written in Matlab. Basic requirement for running SCUBA includes **MATLAB** and the **Statistics toolbox**. The pseudotime estimation step requires two external Matlab packages which are publicly available: **drtoolbox** (which can be downloaded from http://lvdmaaten.github.io/drtoolbox/), and **ksegments** (which can be downloaded from http://lear.inrialpes.fr/~verbeek/software.php). Another option is to utilize the R package **princurve** (which can be downloaded from http://cran.r-project.org/web/packages/princurve/index.html) for instead of ksegments for principal curve analysis. In this case, both R and Matlab are required for running SCUBA.

**3. Usage**

Unzip the package. Change the current directory in Matlab to the folder containing the scripts.

The data for SCUBA analysis has to be placed in the folder 'sample\_data', in a folder specifying the dataset. The package comes with three datasets and their corresponding folders: 'guo2010', 'deng2014' and 'bendall2014'. Prepare the data in an appropriate format (.txt or .fcs) with a standardized name. See below for detailed description.

Run one of the three preprocessing scripts:

PCR\_preprocess.m --- for qPCR data. Data are tab-delimited text format. First row contains the cell ID. Second row contains the cell-stage information. The rest contains the gene expression data matrix. Example: guo2010Data.txt

RNAseq\_preprocess.m --- for RNAseq data. Data are tab-delimited text format. First row contains the cell ID. Second row contains the cell-stage information. The rest contains the gene expression data matrix. By default, the sequence reads are log2-transformed. Example: deng2014Data.txt

MassCytometry\_preprocess.m --- for MassCytometry data. Data are in the binary .fcs format for flow cytometry experiments. This preprocessing step contains a pseudotime estimation algorithm. Example: bendall2014Data.fcs

Each script takes 'dataset' as input, where 'dataset' is the name of the dataset, e.g. 'guo2010'. The preprocessed data are saved as a mat file in the intermediate directory.

Run SCUBA

This is the main function. It has two arguments: 'dataset' and 'cluster\_mode'.

'dataset' refers to the name of the dataset, which is also the name of the data folder.

'cluster\_mode' refers to the method for clustering. It has three options. 'original' -- using the Euclidean distance; 'pca' -- convert the data to principal components then apply Euclidean distance; 'pca2' --- similar to 'pca' but PCA analysis is based on samples in the final cell-stage (used in our paper).

SCUBA has two main steps. In the first step, we infer the cellular hierarchy, using a binary tree model. For simplicity, we only consider steady-state attractors. In the second step, we quantitatively model the dynamics in the reduced state space along each bifurcation direction, using a potential V(x) to characterize gene expression dynamics associated with each bifurcation event.

1. Inference of cellular hierarchy using dynamic clustering.

initial\_tree.m --- This function provides an initial estimate the cellular hierarchy, using a series of k-means clustering.

refine\_tree.m --- this function refines the tree structure by maximizing the penalized likelihood function (Equation 1 in the paper).

2. Bifurcation analysis:

bifurcation\_direction.m --- infer the direction associated with each bifurcation and project data along the bifurcation directions.

bifurcation\_analysis.m --- Infer the dynamical changes of gene expression patterns along the bifurcation direction by fitting a Fokker-Planck equation.

reductionSimulations.m ---Function to predict the effects of perturbing potential regulators in the lineage bias.

For each dataset, the results are deposited in the following three directories:

intermediate\_files, containing intermediate results from the analysis.

figures, containing jpg figures of the analysis.

results, containing the final results of the analysis.

**4. Examples:**

**Example 1:** Analysis of qPCR data in Guo et al. "Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst.". Dev Cell. 2010 Apr 20;18(4):675-85.

>> PCR\_preprocess('guo2010');

>> SCUBA('guo2010')

**Example 2:** Analysis of RNAseq data in Deng et al. " Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells." Science. 2014 Jan 10;343(6167):193-6.

>> RNAseq\_preprocess('deng2014');

>> SCUBA('deng2014')

**Example 3:** Analysis of Mass Cytometry data in Bendall et al. " Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development." Cell. 2014 Apr 24;157(3):714-25.

>> select\_marker\_names = {'CD10','CD117','CD179a','CD179b','CD19','CD20',...

'CD24','CD34','CD38','CD45','CD72','CD79b','HLADR','IgD','IgM-i','IgM-s','Kappa','Lambda'};

>> MassCytometry\_preprocess('bendall2014', select\_marker\_names, 'Rprincurve', 'CD34', 9)

>> SCUBA('bendall2014')

>> plotGeneProfiles('bendall2014', {'CD19', 'CD20', 'CD34', 'CD10', 'CD38'})