**SCUBA v1.0 beta**

**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. PNAS 2014; published ahead of print December 15, 2014, doi:10.1073/pnas.1408993111

**2. 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. Note that processing this dataset requires a machine with at least 40 Gb of memory.

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

SCUBA 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 can have three values. '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.