1. Introduction

This is code for data visualization, statistical analysis and computation of the prototype described in "An integrated modeling framework from cells to organism based on a cohort of digital embryos", Paul Villoutreix, Julien Delile, Barbara Rizzi, Louise Duloquin, Thierry Savy, Paul Bourgine, René Doursat, Nadine Peyriéras.
The code reads in a set of arrays that contain information about the reconstructed embryos. It performs a number of preprocessing steps, such as temporal and spatial rescaling, computes groups of cells statistics and a prototype as the centroid of the cohort. It finally generates artificial cell lineages based on group-level statistics.

2. Requirements

All algorithms and analysis are implemented in MATLAB® (R2015a, The MathWorks, Natick, Massachusetts).

3. How to use

The basic steps are as follows:

1. Read in selections and L-arrays
2. Temporal and spatial rescaling
3. Embryo-level dynamics
4. Population level statistics
5. p-values for distribution approximation
6. R^2 values for within lineage correlations (mother/daughter, sister-sister)
7. Micro-level volume and surface area dynamics
8. Evaluation of the multi-level probabilistic model
9. Prototypical statistics with Kullback-Leibler divergence
10. Artificial lineages using population level statistics (measured or prototypical)

3.1. Organization of the data set

The data set consists in 5 digitally reconstructed embryos obtained from the BioEmergences workflow[1]. For each embryo the corresponding data is stored in two files called *selection* and *L-array*. All the files are gathered in the folder 'dataset'.

The *selection* file consists in a time step by time step description of each cell. It is stored as an array. Each line describe features of a cell at a given time step. It should be noted that the cell population is a number that describes the population a cell belong to: Small Micromeres (1), Large Micromeres (2), Macromeres (3), and Mesomeres (4). It is organised in the following way:

[1] Faure, *et al.*, "A workflow to process 3D+time microscopy images of developing organisms and reconstruct their cell lineage", Nature Communications, 2016.

selection:

cell identity	cell population	time step	predecessor identity	global id	volume (μm^3)	surface area (μm^2)

For example:

63537582	2	19	63536354	35	4505	1587
63537583	3	19	63536355	47	7842	2162
63537584	3	19	63536358	55	8079	2393
63537585	2	19	63536357	32	4047	1572
63537586	3	19	63536358	56	10628	2691
63537587	4	19	63536360	1	9518	2460
63534022	4	20	63537963	13	13857	3200
63534026	4	20	63541305	51	9074	2448
63534027	4	20	63537952	20	13123	3115
63540751	4	20	63540748	44	11214	2784
63540752	4	20	63537967	43	10633	2888

The *L-array* file consists in a global description of the cell lineage where cells are identified by global identities that remain constant between two consecutive mitoses. Sometimes mitoses are not always observed in the beginning or in the end of the cycle. We keep track of this observation with a Boolean value (0 if not observed and 1 if it is). It is stored as an array. Each line is a cell considered throughout its observed cell cycle.

L-array:

global id	cell population	initial time step	final time step	observed initial mitosis	observed final mitosis	generation	id mother	mean volume	Mean surface area

For example:

105	2	42	81	1	0	7	32	2222.61363	995.6818182
106	2	42	81	1	0	7	32	1655.84090	829.5227273
107	4	42	77	1	1	8	41	4026.55555	1443.027778
108	4	42	74	1	1	8	41	3671.84848	1405.878788
109	4	42	76	1	1	8	44	3633.82857	1376.257143
110	4	42	73	1	1	8	44	5348.5625	1771.03125
111	3	42	71	1	1	8	49	4439.8	1684.733333
112	3	42	70	1	1	8	49	4808.20689	1769.206897
113	4	42	74	1	1	8	51	4608.66666	1586.393939
114	4	42	74	1	1	8	51	5347.48484	1740.848485
115	3	42	73	1	1	8	53	5055.53125	1683.875
116	3	42	71	1	1	8	53	5753.6	1797.1
117	4	42	70	1	1	8	73	4620.96551	1705.344828

3.2. How to use the code

The basic set of instructions sufficient to generate all the figures associated with the probabilistic reconstruction is stored in the file 'main.m'. Therefore typing the command

```
main
```

will generate all the figures in the folder 'figures'. The steps of the script are described below.

The first step consists in adding the path to the various subfolders that contain the scripts and in creating a folder where the figures will be stored.

```
addpath('artificiallineages')
addpath('spatiotemporalrescaling')
addpath('statisticalanalysis')
addpath('visualization')
addpath('tools')
if ~exist('figures','dir'),
    mkdir('figures'),
end
```

The second step consists in loading the data sets (*selection* and *L-arrays*).

```
Data.selection = {};
Data.L = {};

for i = 1:5,
    Data.selection{i} = csvread(['dataset/selection',num2str(i),'.csv']);
    Data.L{i} = csvread(['dataset/L',num2str(i),'.csv']);
end

% Defining experimental parameter values
% Size of the cohort
Options.Nb = 5;
% Maximal time step considered for each embryo
Options.tmax = [81; 86; 90; 75; 126];
% Experimental time step in seconds
Options.deltat = [119 ; 133 ; 207 ; 220 ; 180 ];
% Experimental initial time of each experiment
Options.tinit = [13500 ; 20700 ; 15600; 18000; 15600 ];

% color code
Options.colors_selections = [[92 0 77];[255 121 255];[228 19 0];[13 173
209]]./255;
Options.colors_embryos =
[[35,91,190];[39,173,25];[238,110,33];[76,77,79];[140,129,100]]./255;
Options.colors_generations =
[[37,253,233];[119,181,254];[0,0,255];[102,0,153];[253,108,158];[205, 205,
13]]./255;

% names of the various subpopulations
Options.ident = {'Smic';'LMic';'Mac';'Mes'};

% number of cell in each subpopulation at generation 6
Options.ngen6 = [4,4,8,16];
```

```
% initial generation number in each population of each embryo
Options.g0 = [[6,6,6,6];[6,7,8,8];[6,6,6,6];[6,6,7,7];[6,6,7,7];];
```

The third step consists in performing spatial and temporal rescaling of the data sets

```
[Data.rescaledtime,~,~] = temporalrescaling(Data,Options);

[Data] = spatialrescaling(Data,Options);
```

The fourth step consists in computing and visualizing embryo-level dynamics

```
[Data.NbCells, Data.VolCells, Data.SurfCells] =
embryo_level_variables(Data, Options);
embryo_level_variables_visualization(Data, Options);
```

The fifth step consists in computing and visualizing population level statistics

```
[ pop_parameters ] = population_parameters( Data, Options );
population_parameters_visualization(pop_parameters, Options);
```

The sixth step consists in evaluating the statistical characteristics of the cell features within the lineage (p-values and R^2).

```
gaussianFit(Data, Options);
correlations(Data, Options);
```

The seventh step consists in computing the volume and surface area microdynamics.

```
microdynamics(Data, Options);
```

The eighth step consists in evaluating the multi-level probabilistic model.

```
modelevaluation(pop_parameters,Options);
```

The ninth step consists in computing and visualizing the prototypical population statistics using the symmetrized Kullback-Leibler divergence.

```
[proto_parameters] = prototypical_parameters(pop_parameters);
population_proto_parameters_visualization(proto_parameters, Options);
```

The tenth step consists in generating random cell lineages using cell populations statistical parameters for each embryo and for the prototype. The embryo-level statistics can then be visualized to confirm the accuracy of the model. The number of draws will increase the robustness of the statistics at the cost of an increased computing time.

```
% with statistics from one embryo
% embryo number
i = 3;
% number of randomly generated cell lineages - warning: increasing the
% number of draws increases highly the computation time - ndraw = 300 gives
% good results
```

```
ndraw = 3;
mainmodelstat(i,ndraw, Data, Options, pop_parameters);



% with prototypical statistics
% number of randomly generated cell lineages - warning: increasing the
% number of draws increases highly the computation time - ndraw = 300 gives
% good results
ndraw = 3;
prototype(ndraw, Data, Options, proto_parameters);
```