scBoolSeq – scRNA-Seq data binarization and synthetic generation from Boolean dynamics

Gustavo Magaña López*[1], Laurence Calzone[2], Andrei Zinovyev[2], and Loïc Paulevé[1]

*gustavo.magana-lopez@labri.fr

[1] Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France

[2] Institut Curie, INSERM, U900, MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France

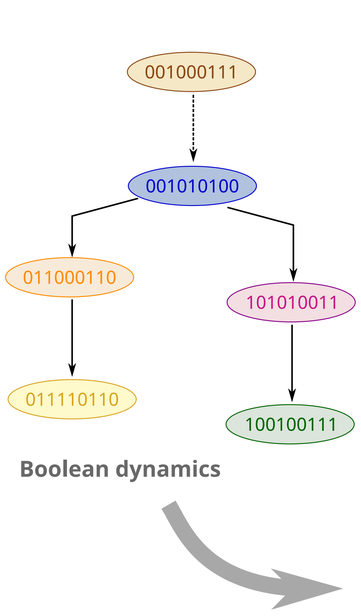
poster

github

PROJET FINANCÉ PAR L'ANR – PROJECT FUNDED BY THE ANR

Overview

- scBoolSeq is a Python package for **linking scRNA-Seq data and Boolean gene activation states**.
- It uses a reference dataset to:
  - **Binarize experimental data**
  - **Sample synthetic scRNA-Seq** from simulations of Boolean networks.
- **Synthetic scRNA-Seq** can serve as a **ground-truth baseline** for validating inference methods.



scRNA-seq counts

	Egr1	Junb	Bclaf1	Myc	Fil1	Gata2	Spi1
Cell_2039	3.892874	2.827419	1.273664	3.724108	1.328451	3.069076	1.344202
Cell_2304	1.248786	0.635362	0.000000	2.481673	2.954907	3.526111	3.858446
Cell_2395	3.057174	4.433269	2.263485	2.307164	2.574745	3.080573	0.000000
Cell_2698	0.836229	3.739484	3.175067	3.427367	2.953074	2.725792	0.711286
Cell_2401	1.937875	0.197981	2.125259	4.728801	2.002604	1.096830	1.468082

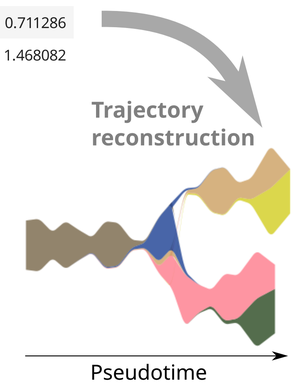
.simulate()

scBoolSeq

.binarize()

binary scRNA-seq counts

	Egr1	Junb	Bclaf1	Myc	Fil1	Gata2	Spi1
Cell_2039	0	0	1	0	1	0	0
Cell_2304	0	?	1	1	0	0	1
Cell_2395	0	1	1	?	?	1	1
Cell_2698	1	0	0	1	1	1	1
Cell_2401	?	1	1	0	0	0	0



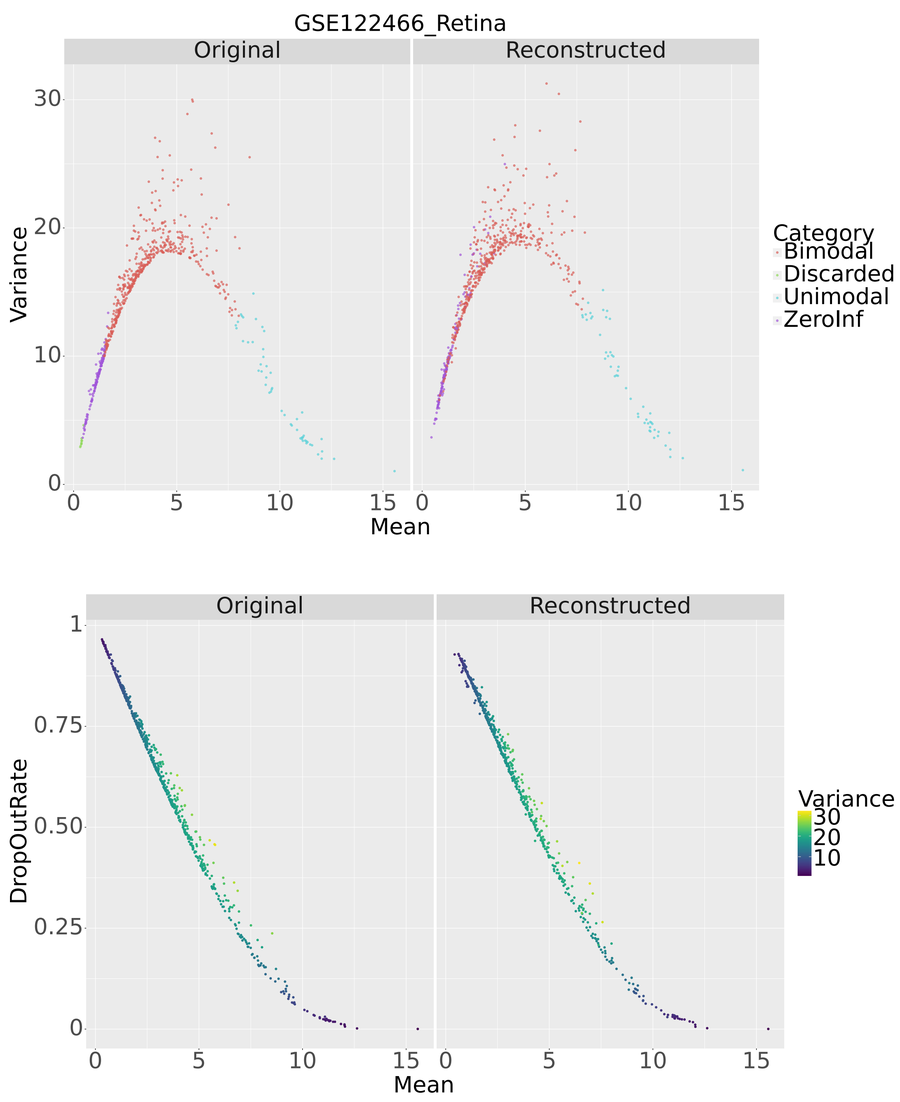
Binarization

Binarization is performed by comparing each count to the thresholds determined from the learned distributions on the reference dataset. It results in either a Boolean or undetermined value.

Minimal binarization example:

```
from scboolseq import scBoolSeq
reference = pd.read_csv("reference_scRNA_counts.csv")
scbool = scBoolSeq(data=reference)
scbool.fit() # compute binarization criteria
binarized = scbool.binarize(reference) # or other dataset
```

Sampling from the learned distributions with our algorithm reproduces the reference's profiles:



Total run-time (distribution learning + reconstruction) on specified dataset (688 genes over 5347 cells) is less than a minute.

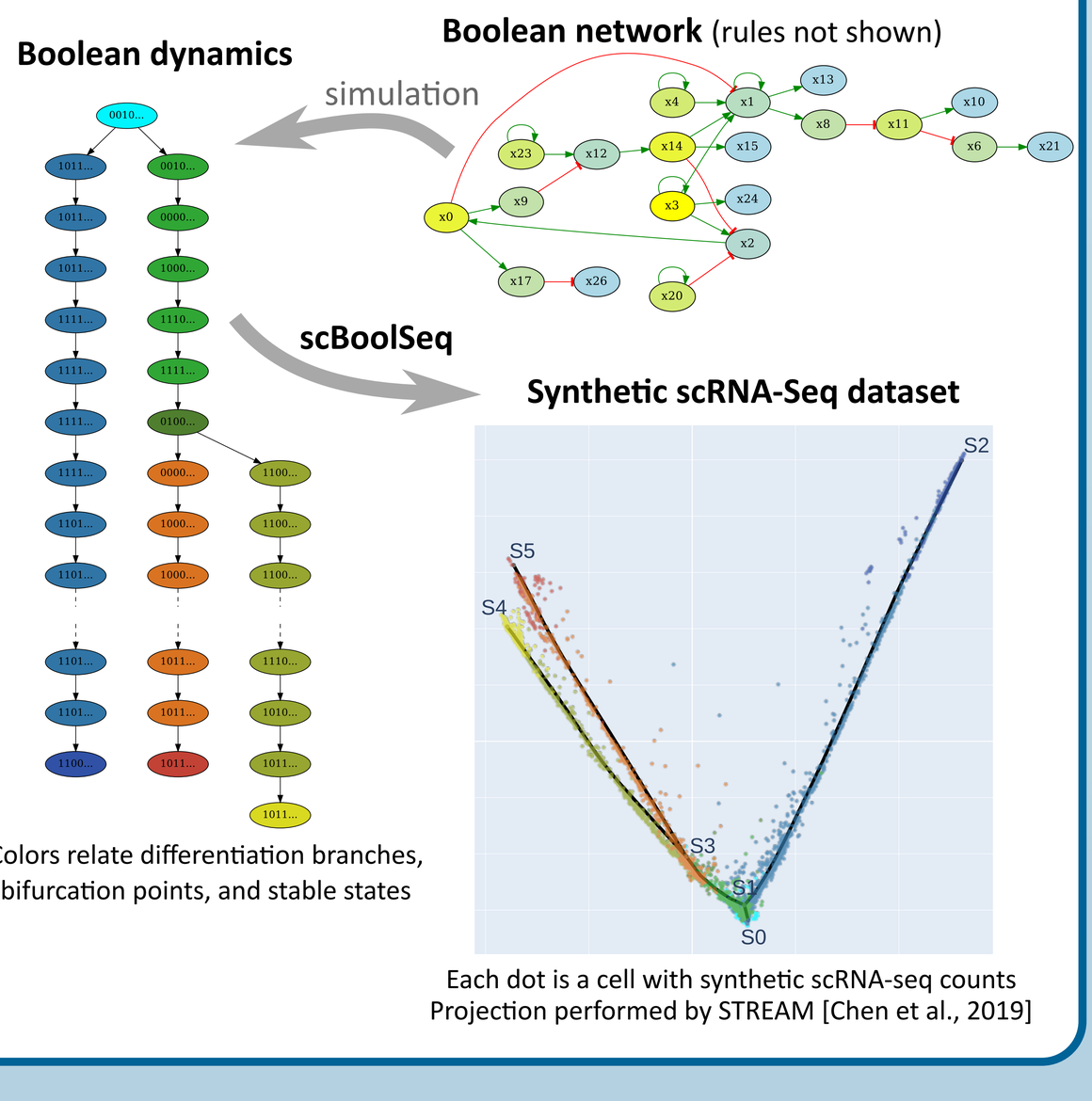
Synthetic scRNA-Seq Generation

The synthetic RNA-Seq data **from Boolean states** is generated by **biased sampling** from the estimated gene count distributions.

- Synthetic scRNA-Seq counts use biased sampling to **reflect the underlying Boolean value**.
- **Drop-outs are simulated** with probabilities that decay exponentially with the sampled expression value.

Minimal synthesis example:

```
from scboolseq import scBoolSeq
reference = pd.read_csv("reference_scRNA_counts.csv")
scbool = scBoolSeq(data=reference)
scbool.simulation_fit() # compute simulation criteria
syndata = scbool.simulate(boolean_trace)
```



Colors relate differentiation branches, bifurcation points, and stable states

Each dot is a cell with synthetic scRNA-seq counts
Projection performed by STREAM [Chen et al., 2019]