Inferring Boolean networks from single-cell human embryo datasets

proof of concept with trophectoderm maturation

Mathieu Bolteau

LS2N - ComBi team

Supervisors: Jérémie Bourdon, Carito Guziolowski, Laurent David

Friday, June 16th, 2022

Motivations

Poor success of *in vitro* fertilization (IVF)

→

Need to **better understand preimplantation development** (especially cell fate transition)

Motivations

Poor success of *in vitro* fertilization (IVF)

→

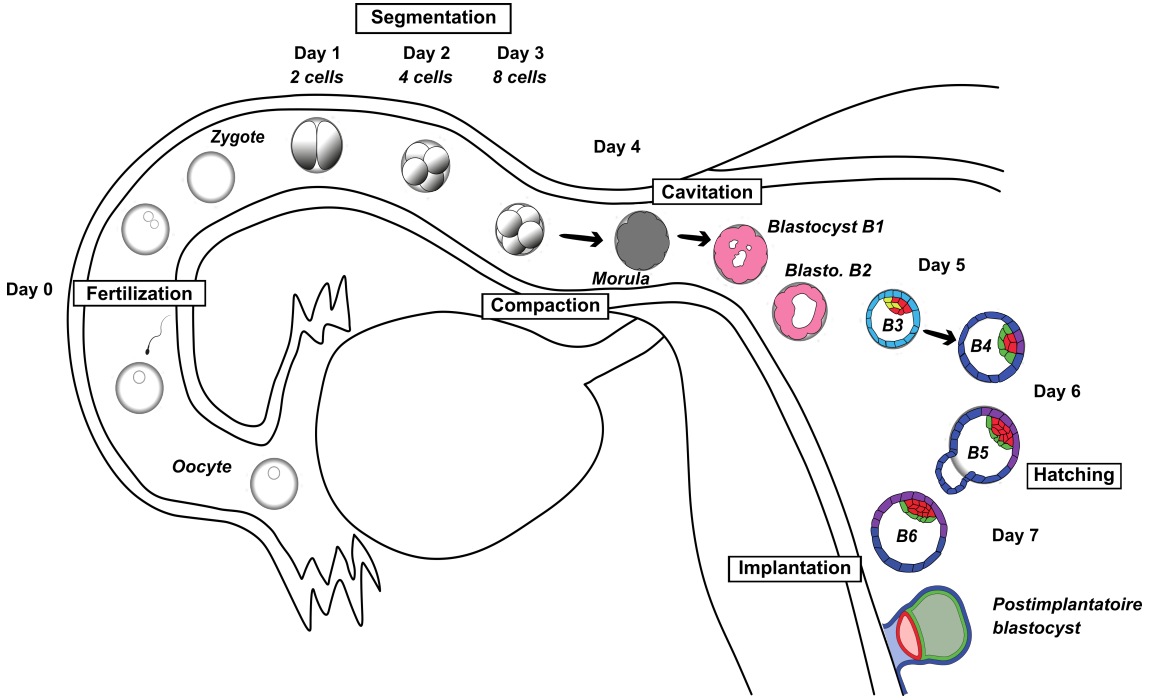
Need to **better understand preimplantation development** (especially cell fate transition)

Research on human embryos is **limited** (experiments, law, ethics)

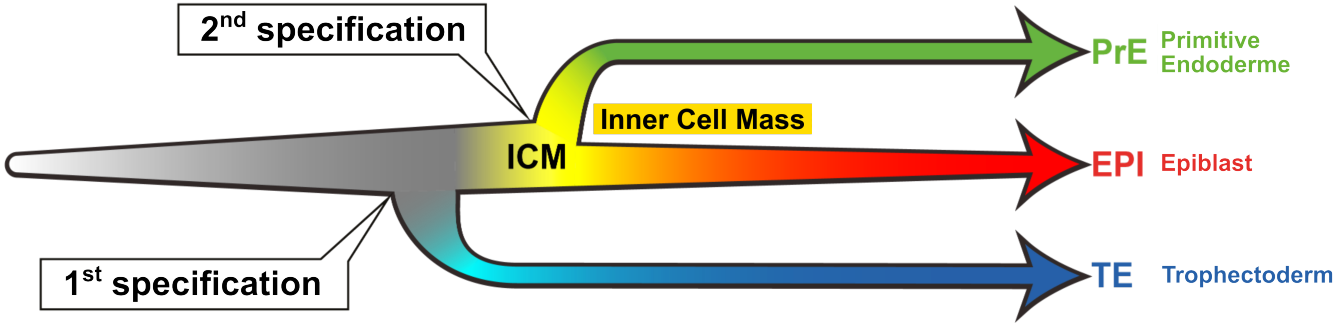
↓

***In silico* predictive model of the cell fate transition** during the human preimplantation development

Human embryonic development

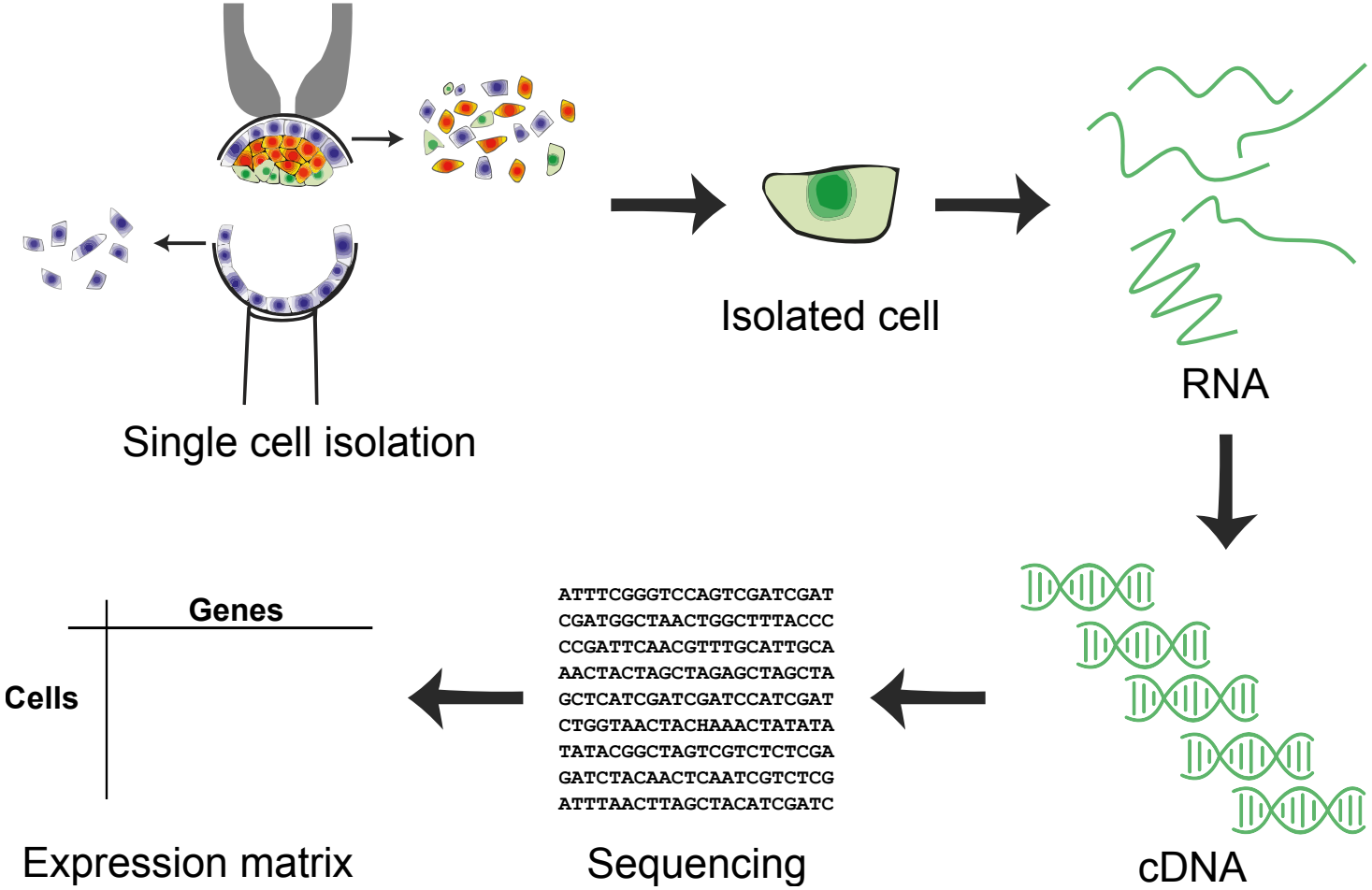


Adapted from Meistermann, Ph.D. thesis (2020).



Adapted from Meistermann D.

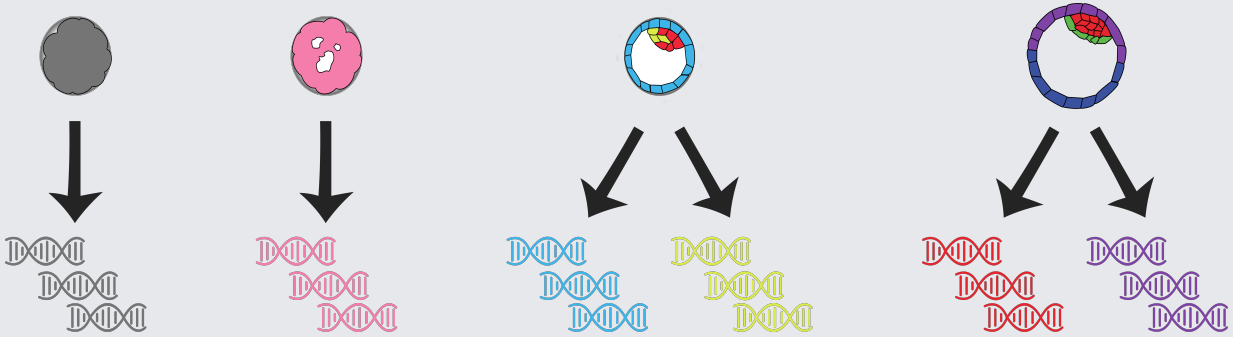
Single cell RNA sequencing



Data

scRNAseq data from multi-stage embryos

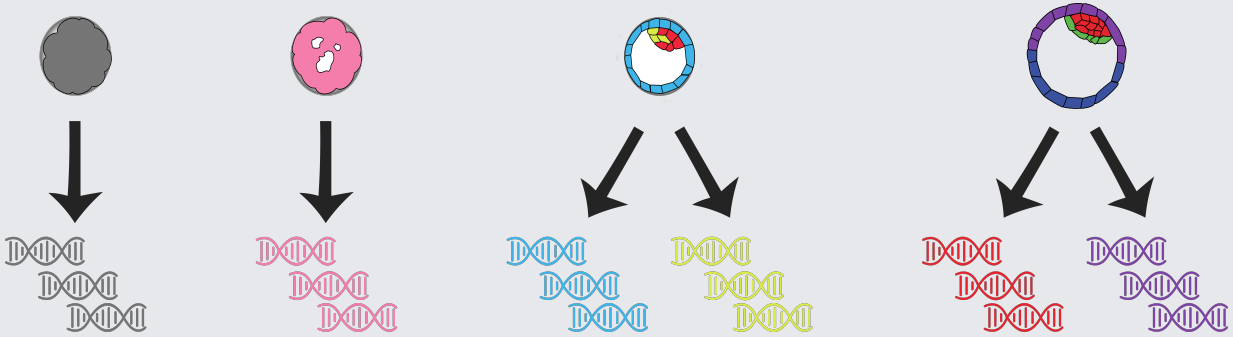
128 embryos at different developmental stages



Data

scRNAseq data from multi-stage embryos

128 embryos at different developmental stages



Expression Matrix

Cells \ Genes	5S_rRNA	5_8S_rRNA	7SK	A1BG	A1BG-AS1	A1CF	...	ZZZ3
E4.9.1209	7	1	0	0	0	0	...	95
E4.6.609	0							
E6.13.1377	1							
E5.10.936	0							
E5.5.100	0							
E6.11.1064	0							
...	...							
E6.14.1409	0							

$\sim$20,000 genes X $\sim$1,700 cells

Goal: Boolean models of embryonic development stages

Challenges

- Different developmental stages: high variability
- High dimensional data
- Unavailable perturbations

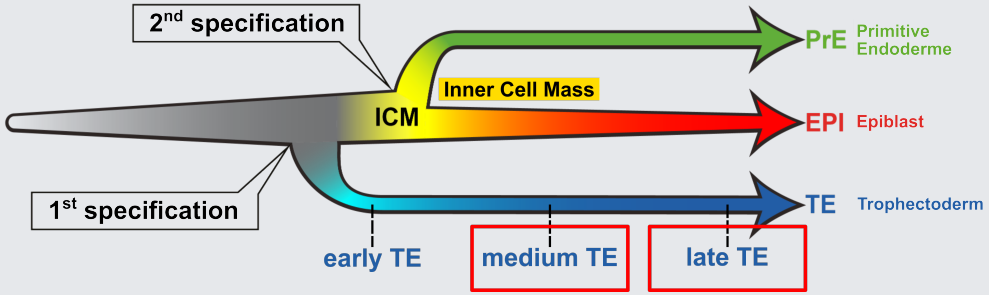
Goal: Boolean models of embryonic development stages

Challenges

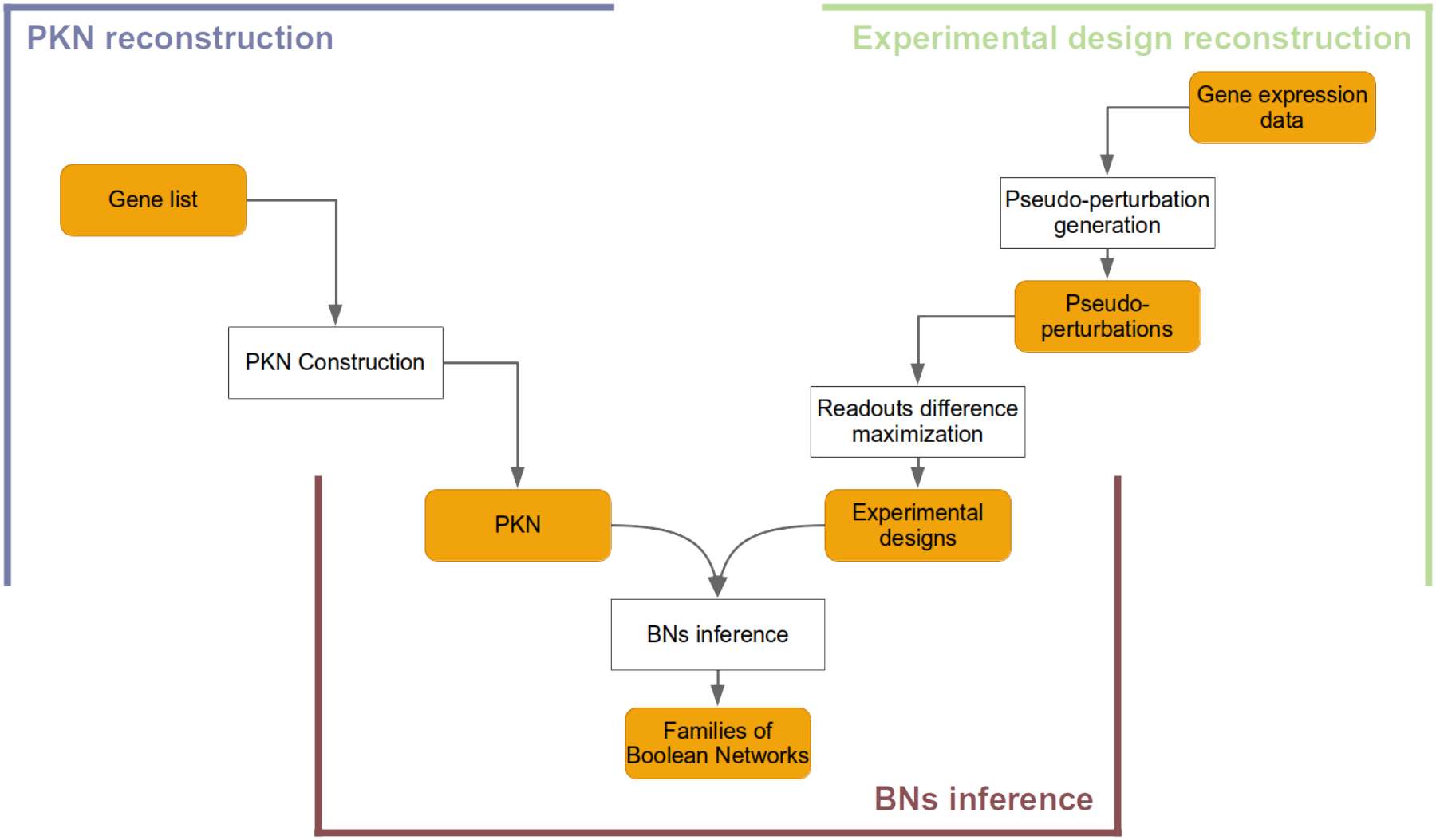
- Different developmental stages: high variability
- High dimensional data
- Unavailable perturbations

Proposed solution

- Focus on two developmental stages: medium (M^{TE}) and late (L^{TE}) TE
- Build specific network models for each stage
- Identify regulatory mechanisms that differentiate both models

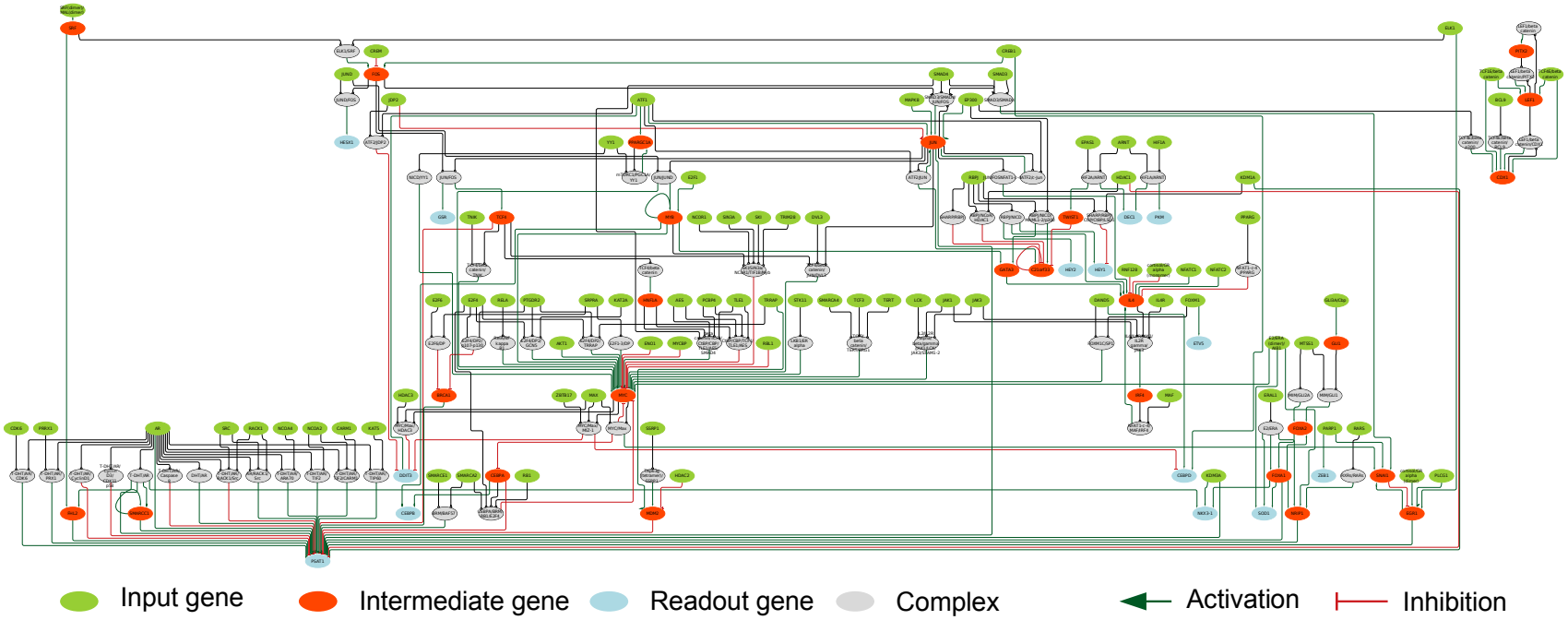


Pipeline


PKN reconstruction
Experimental design reconstruction
Gene expression data
Gene list
Pseudo-perturbation generation
Pseudo-perturbations
PKN Construction
Readouts difference maximization
PKN
Experimental designs
BNs inference
Families of Boolean Networks
BNs inference

Step 1: Prior-Knowledge Network (PKN) reconstruction

- PKN reconstructed from 438 TFs identified in embryonic development, using the PathwayCommons database (pyBRAvo) [Lefebvre et al. *Database*, 2021]
- Labeled (activation/inhibition) and oriented graph
- Nodes: genes, protein-complexes – Edges: transcription regulation
- 192 nodes : inputs (84), intermediates (27), readouts (14) and 285 edges



Programming paradigms: Semantic Web & Python

Step 2: Experimental design reconstruction

Idea

Extract pseudo-perturbation experiments from scRNAseq data given the PKN structure (Step 1)

Data preprocessing of the expression matrix

- Binarization of input + intermediate genes (5 genes).
  Basic approach: gene is expressed (1) if at least 2 reads are present; else it is absent (0).
- Normalization of readout genes

Programming paradigms: Answer Set Programming (ASP) & Python

Step 2: Experimental design reconstruction

Example

Inputs + intermediates | Readouts

Cell	A	B	C	D	E	F	G	H	Class
1	1	1	0	1	0	0.8	0.4	0.6	M^{TE}
2	1	0	1	0	0	0.2	0.5	0.3	M^{TE}
3	0	1	0	1	1	0.8	0.3	0.9	M^{TE}
4	1	1	1	0	1	0.6	0.1	0.2	L^{TE}
5	1	0	0	1	1	0.7	0.8	0.5	L^{TE}
6	0	0	1	0	1	0.7	0.2	0.3	L^{TE}

Logic program (k=3)

Cell	A	C	D	F	G	H	Class
1	1	0	1	0.8	0.4	0.6	M^{TE}
2	1	1	0	0.2	0.5	0.3	M^{TE}

Cell	A	C	D	F	G	H	Class
4	1	1	0	0.6	0.1	0.2	L^{TE}
5	1	0	1	0.7	0.8	0.5	L^{TE}

Example output:

- 3 selected genes: A, C, D ($k = 3$)
- Matching cells: (1,5), (2,4) ← pseudo-perturbations
- Optimal number of matching cells: 2

Step 2: Experimental design reconstruction

Example

Inputs + intermediates | Readouts

Cell	A	B	C	D	E	F	G	H	Class
1	1	1	0	1	0	0.8	0.4	0.6	M^{TE}
2	1	0	1	0	0	0.2	0.5	0.3	M^{TE}
3	0	1	0	1	1	0.8	0.3	0.9	M^{TE}
4	1	1	1	0	1	0.6	0.1	0.2	L^{TE}
5	1	0	0	1	1	0.7	0.8	0.5	L^{TE}
6	0	0	1	0	1	0.7	0.2	0.3	L^{TE}

Logic program (k=3)

Cell	A	C	D	F	G	H	Class
1	1	0	1	0.8	0.4	0.6	M^{TE}
2	1	1	0	0.2	0.5	0.3	M^{TE}

Cell	A	C	D	F	G	H	Class
4	1	1	0	0.6	0.1	0.2	L^{TE}
5	1	0	1	0.7	0.8	0.5	L^{TE}

Example output:

- 3 selected genes: A, C, D ($k = 3$)
- Matching cells: (1,5), (2,4) ← pseudo-perturbations
- Optimal number of matching cells: 2

Real case study results

- **20 pseudo-perturbations** found in 65h (sub-optimal solution)

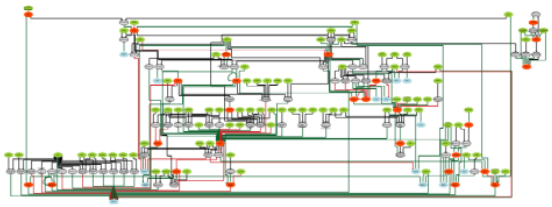
Step 3: Boolean Networks (BNs) inference

Cell	I_1	...	I_n	In_1	...	R_1	...	R_2	Class
1	1	1	1	0	1	0.6	0.1	0.2	M^{TE}
...	...	...	...	...	...	...	...	...	M^{TE}
20	0	0	1	0	1	0.7	0.2	0.3	M^{TE}

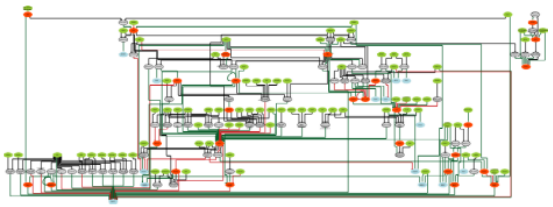
Experimental design for Class M^{TE}

Cell	I_1	...	I_n	In_1	...	R_1	...	R_2	Class
1	1	0	1	0	0	0.2	0.5	0.3	L^{TE}
...	...	...	...	...	...	...	...	...	L^{TE}
20	1	0	0	1	1	0.7	0.8	0.5	L^{TE}

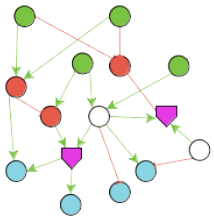
Experimental design for Class L^{TE}



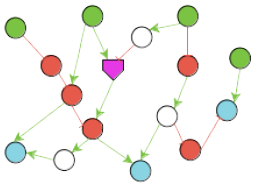
PKN



Learning Boolean Networks
(Caspo)



Boolean Networks
for Class M^{TE}

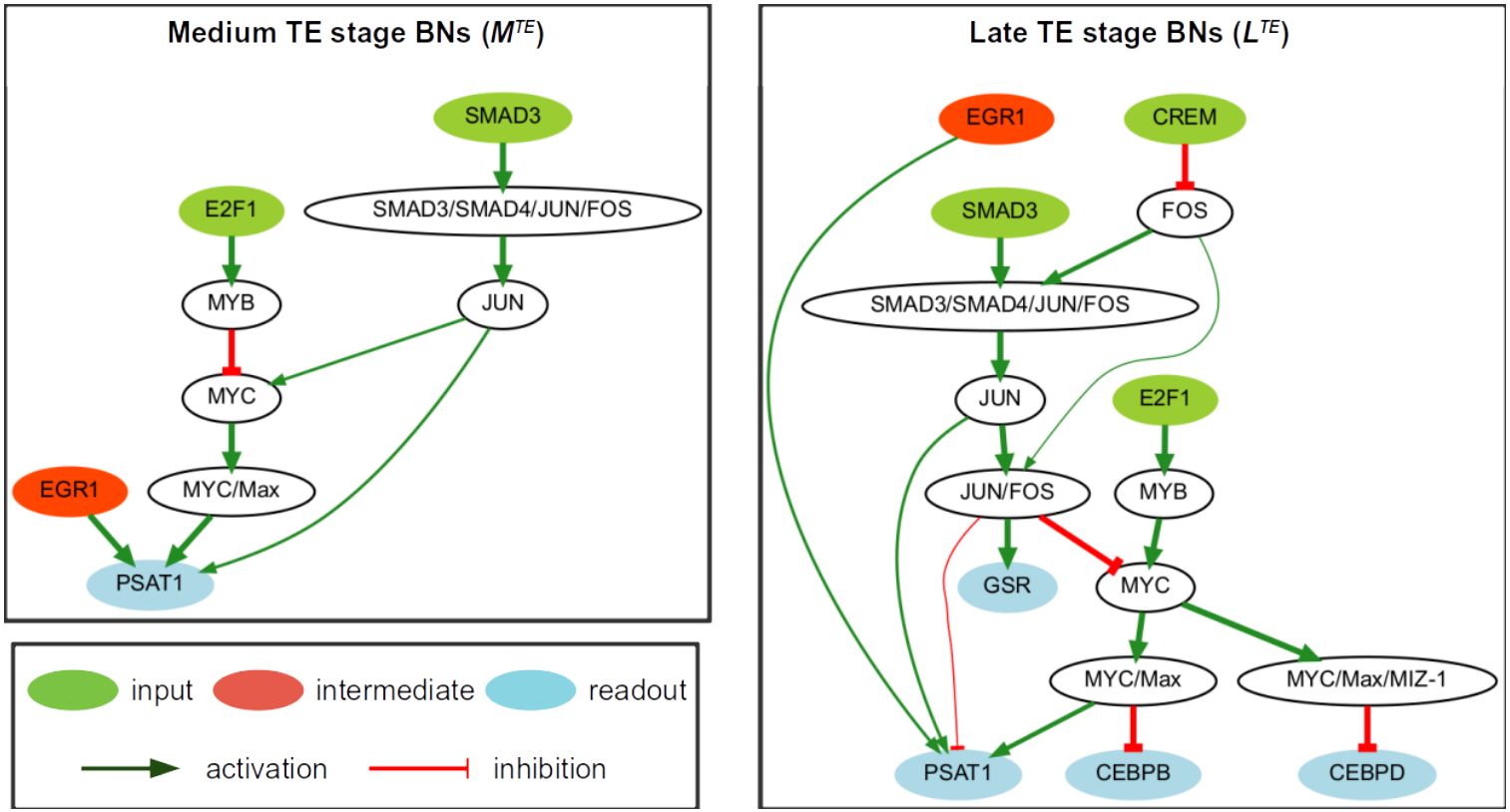


Boolean Networks
for Class L^{TE}

Caspo [Guziolowski et al. *Bioinformatics*, 2013.]

Programming paradigms: Answer Set Programming (ASP) & Python

Step 3: Boolean Networks (BNs) inference



- Selected L^{TE} cells agree on a BN structure that is more complex (size 15 compared to 8) than M^{TE} cells.
- Opposite regulatory paths for gene PSAT1 from genes E2F1 and SMAD3 proposed by M^{TE} and L^{TE} BNs.
- Hypothesis: M^{TE} stage seems more stable than L^{TE} one, concording to the transition of L^{TE} to another stage.

Methodological contribution

- Proposed a method that learns Boolean networks of 2 stages: medium and late TE using scRNASeq data and Prior Knowledge.
- First findings are in agreement with biological literature

Next straight-forward directions

- Improve the computational time of the method to select genes $\rightarrow$ 92 pseudo-perturbations in $\sim$7h
- Study the robustness of our method when using more prior knowledge (larger PKN)
- Study the impact of different discretization methods
- Evaluate the difference between the BNs found

Aknowledgements

- Jérémie Bourdon – LS2N, Nantes Université
- Carito Guziolowski – LS2N, Centrale Nantes
- Laurent David – CR2TI, CHU de Nantes, Nantes Université
- ANR AIBY4 & ANR BOOSTIVF