Boolean networks as a framework to model human preimplantation development

Mathieu Bolteau[1]
Jérémie Bourdon[1] Laurent David[2] Carito Guziolowski[1]

[1]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

[2]Nantes Université, CHU Nantes, INSERM, Center for Research in Transplantation and Translational Immunology, UMR 1064, F-44000 Nantes, France

ISMB/ECCB 2023

Thursday, July 27th

Motivations

Need to **better understand preimplantation development**

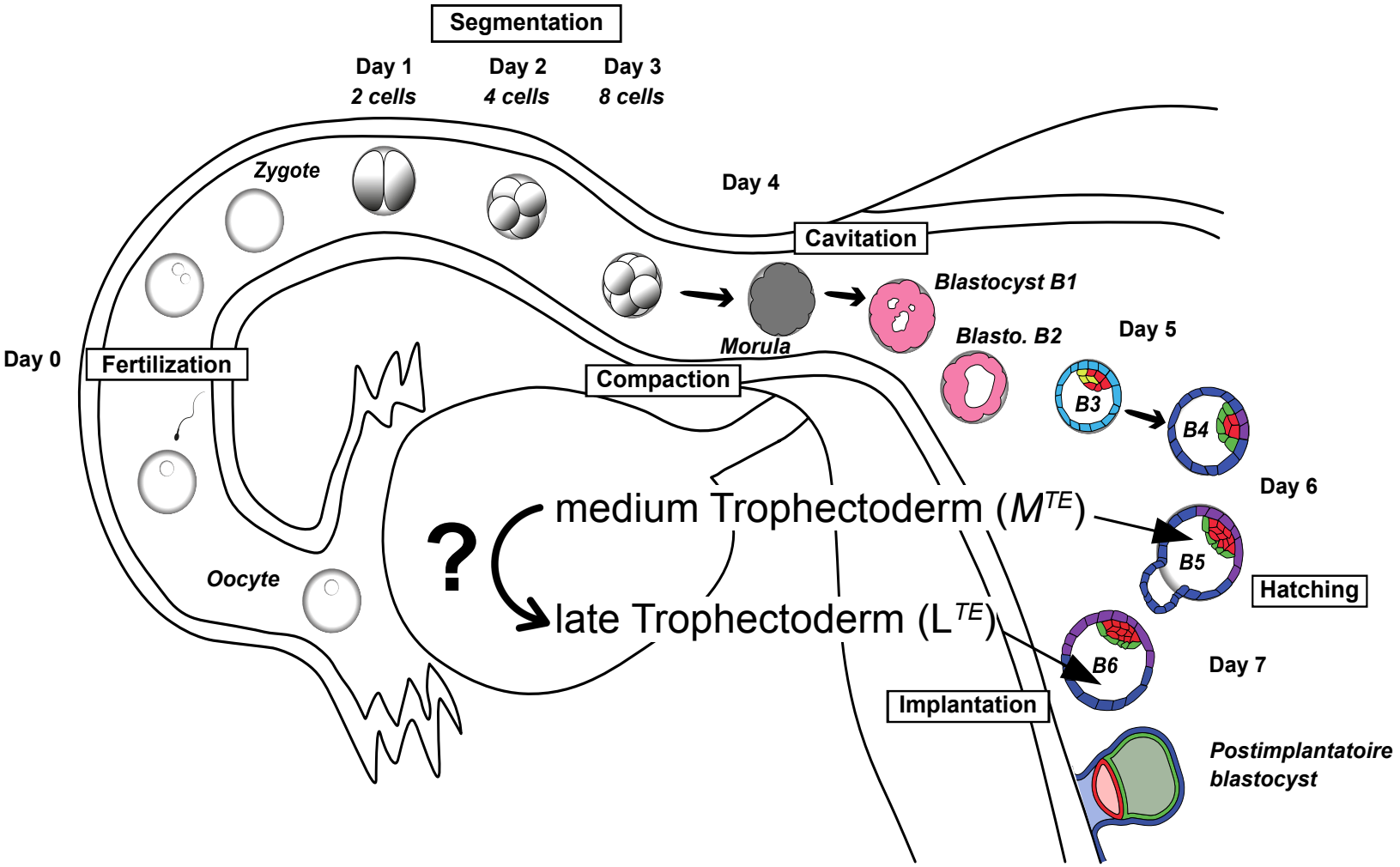
Motivations

Need to **better understand preimplantation development**

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

In silico predictive model of human embryonic development

Human embryonic development



Adapted from Meistermann, *Ph.D. thesis*, 2020

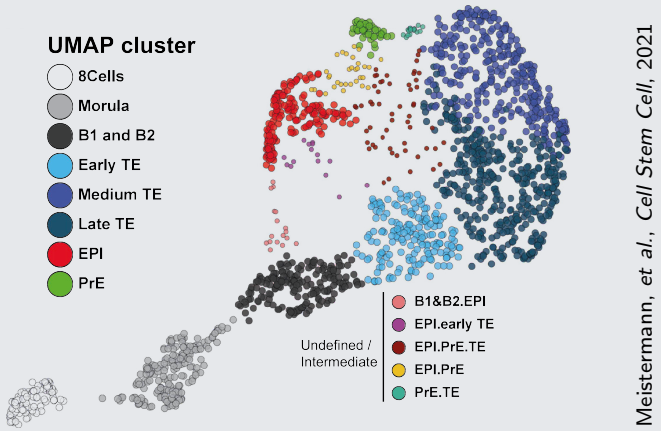
Background [Meistermann, *et al.*, *Cell Stem Cell*, 2021]

scRNAseq data from multiple stage embryos

Expression of $\sim 20,000$ genes in $\sim 1,700$ cells from 128 multi-stage embryos

Previous results (in house)

- **Clustering** of cells
- Identification of **gene modules** $\rightarrow$ 438 transcription factors (TFs)
- **Pseudotime** evolution of cells at different developmental stages



State of the art – modeling of single-cell data

Data analysis

- Statistical, *e.g.* weighted correlation network analysis (WGCNA [Langfelder & Horvath, *BMC Bioinformatics*, 2008])
- Machine learning, *e.g.* reverse graph embedding (pseudotime [Qiu *et al.*, *Nature Methods*, 2017]), uniform manifold approximation and projection (UMAP [McInnes *et al.*, *arXiv preprint*, 2018])

Network inference

- Correlation, *e.g.* gene regulatory network (GRN) inference (SCENIC [Aibar *et al.*, *Nat Methods*, 2017])

Modeling

- Dynamic Boolean models requires average of gene expression and prior knowledge (BoNesis [Chevalier *et al.*, *ICTAI*, 2019])
- Mouse embryo development computational models requires genetic perturbations and knockdowns [Dunn *et al.*, *EMBO journal*, 2019]

Goal: Boolean models of embryonic developmental stages

Challenges

- Single cell data specifities: sparsity and redundancy
- High dimensional data: $\sim 20,000$ genes for $\sim 1,700$ cells
- Unavailable perturbations

Goal: Boolean models of embryonic developmental stages

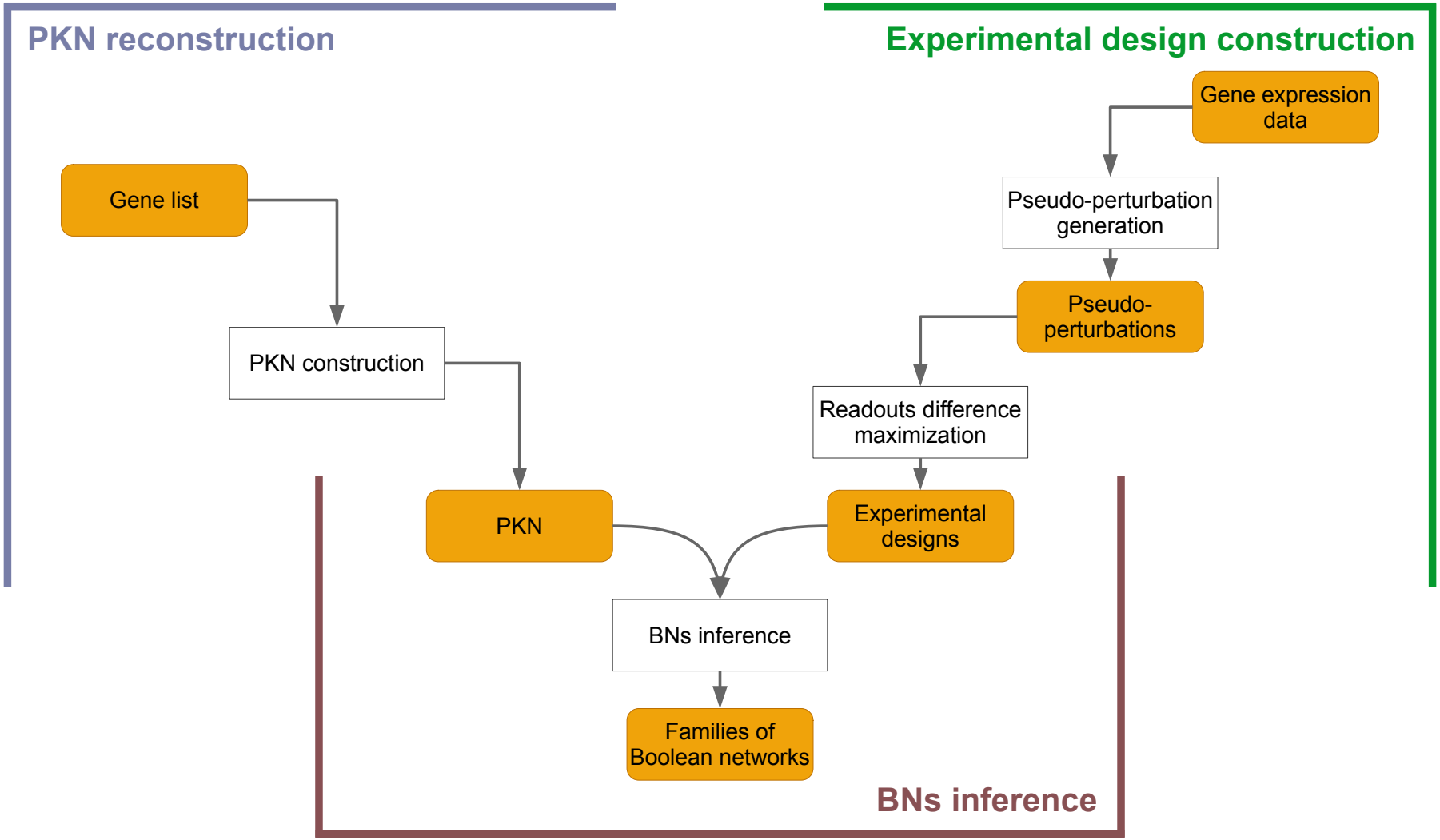
Challenges

- Single cell data specifities: sparsity and redundancy
- High dimensional data: $\sim 20,000$ genes for $\sim 1,700$ cells
- Unavailable perturbations

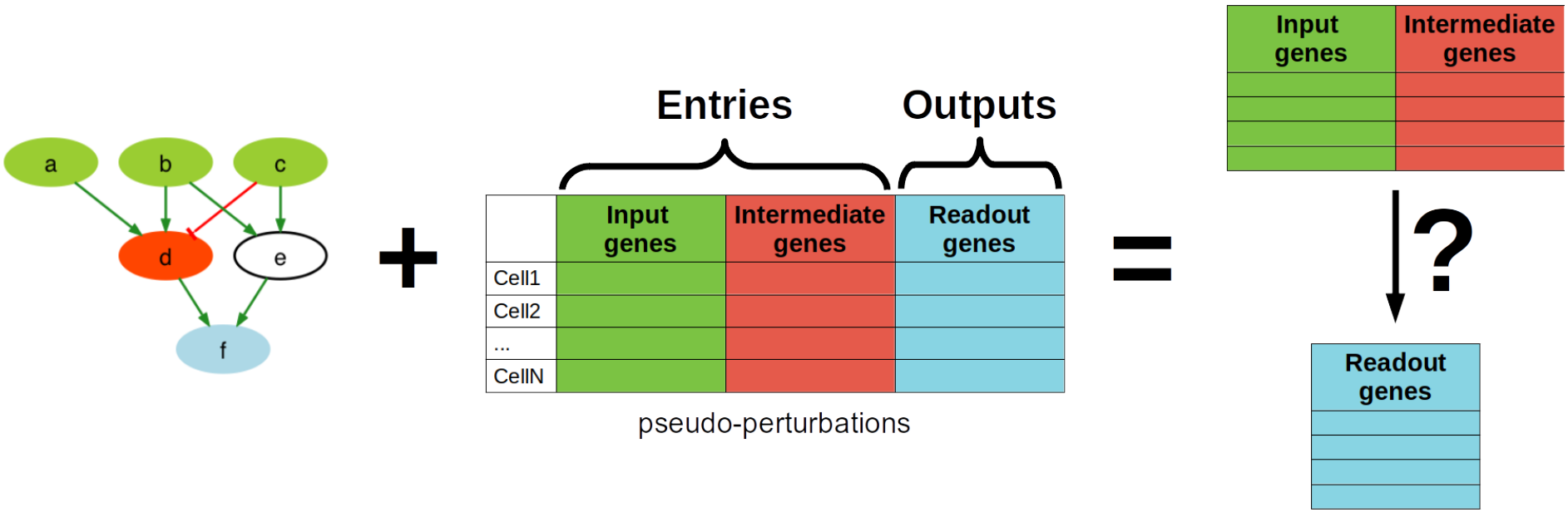
Proposed solution

- Distinguish between **two developmental stages**
- Build **families of network models** for each stage
- Identify **regulatory mechanisms** that differentiate both models and representing multiple cells
- Application on medium (M^{TE}) and late (L^{TE}) trophectoderm stages

Pipeline


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

Learning predictive models



Prior-Knowledge Network

Experimental design

Predictive model

- Signed and directed causal interactions among genes
- Gene expression for a developmental stage

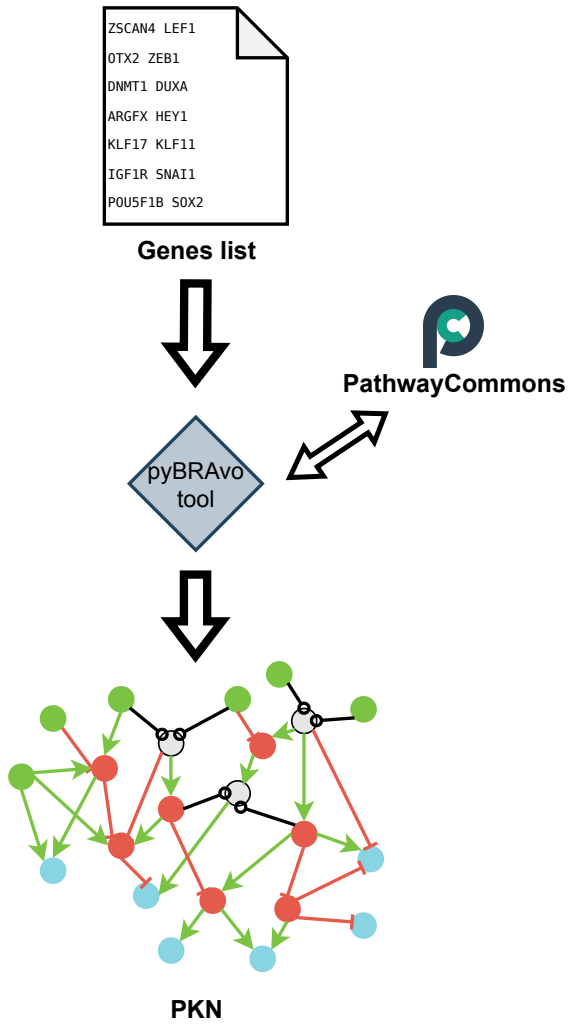
Step 1. PKN reconstruction

Query on PathwayCommons database,
via (in house) pyBRAvo tool [Lefebvre *et al.*, *Database*, 2021]

Output PKN

- Labeled (activation/inhibition) and oriented graph
- Nodes: genes (inputs + intermediates + readouts), protein-complexes
- Edges: Transcription regulation

→ input and intermediate genes: entry for experimental design (Step 2)
→ readout genes: output for experimental design (Step 2)



Step 2. Experimental design reconstruction

Idea

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

Data preprocessing

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

Step 2.Pseudo-perturbation generation

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	A
2	1	0	1	0	0	0.2	0.5	0.3	A
3	1	1	0	1	1	0.8	0.3	0.9	A
4	1	1	1	0	1	0.6	0.1	0.2	B
5	1	0	0	1	1	0.7	0.8	0.5	B
6	0	0	1	0	1	0.7	0.2	0.3	B

Logic program in ASP (k=3)

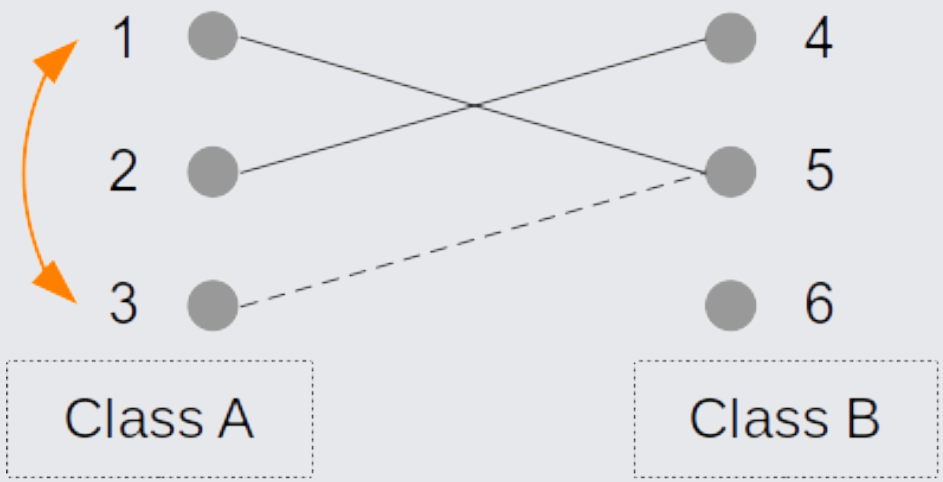
Cell	A	C	D	F	G	H	Class
1	1	0	1	0.8	0.4	0.6	A
2	1	1	0	0.2	0.5	0.3	A

Cell	A	C	D	F	G	H	Class
5	1	0	1	0.7	0.8	0.5	B
4	1	1	0	0.6	0.1	0.2	B

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

Step 2. Maximizing the readout difference

Redundancy



N. of matching cells: 2 (max)

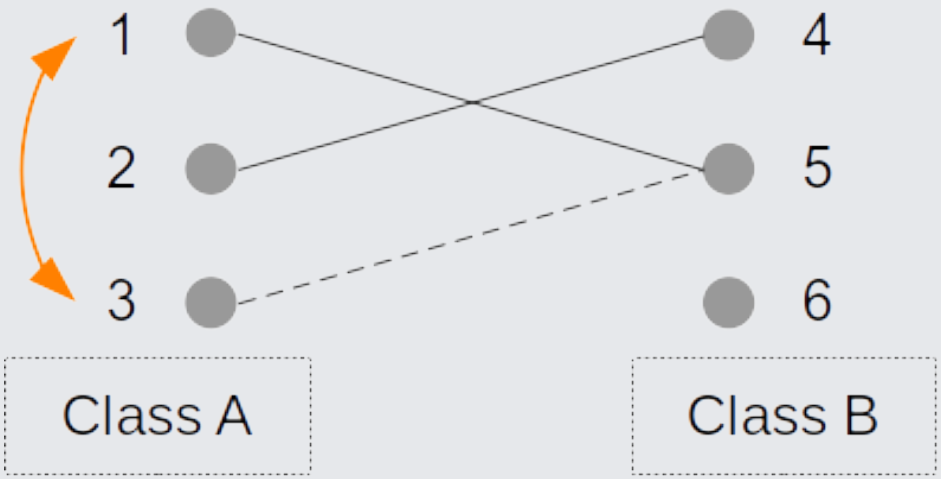
Solution 1 : (1,5), (2,4)

Solution 2 : (3,5), (2,4)

Cell representativity:
Class A : 3 out of 3 (100%)
Class B : 2 out of 3 (66%)

Step 2. Maximizing the readout difference

Redundancy



N. of matching cells: 2 (max)

Solution 1 : (1,5), (2,4)

Solution 2 : (3,5), (2,4)

Cell representativity:
Class A : 3 out of 3 (100%)
Class B : 2 out of 3 (66%)

Readout difference maximization

diff(1,5) = |0.8-0.7| + |0.4-0.8| + |0.6-0.5| = 0.6

Cell	A	C	D	F	G	H	Class
1	1	0	1	0.8	0.4	0.6	A

Cell	A	C	D	F	G	H	Class
5	1	0	1	0.7	0.8	0.5	B

diff(3,5) = |0.8-0.7| + |0.3-0.8| + |0.9-0.5| = 0.9

Cell	A	C	D	F	G	H	Class
3	1	0	1	0.8	0.3	0.9	A

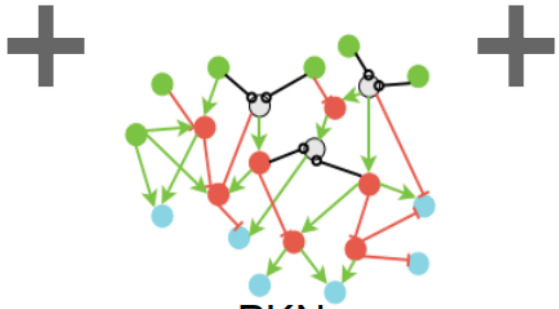
Step 3. BNs inference using Caspo [Guziolowski *et al.*, *Bioinformatics*, 2013]

Cell	A	C	D	F	G	H	Class
3	1	0	1	0.8	0.3	0.9	A
2	1	1	0	0.2	0.5	0.3	A

Experimental design for Class A

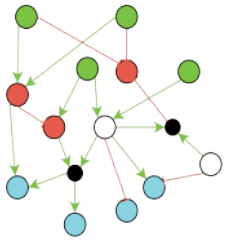
Cell	A	C	D	F	G	H	Class
5	1	0	1	0.7	0.8	0.5	B
4	1	1	0	0.6	0.1	0.2	B

Experimental design for Class B

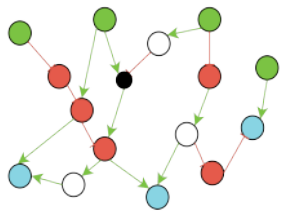


PKN

Learning Boolean Networks (Caspo)



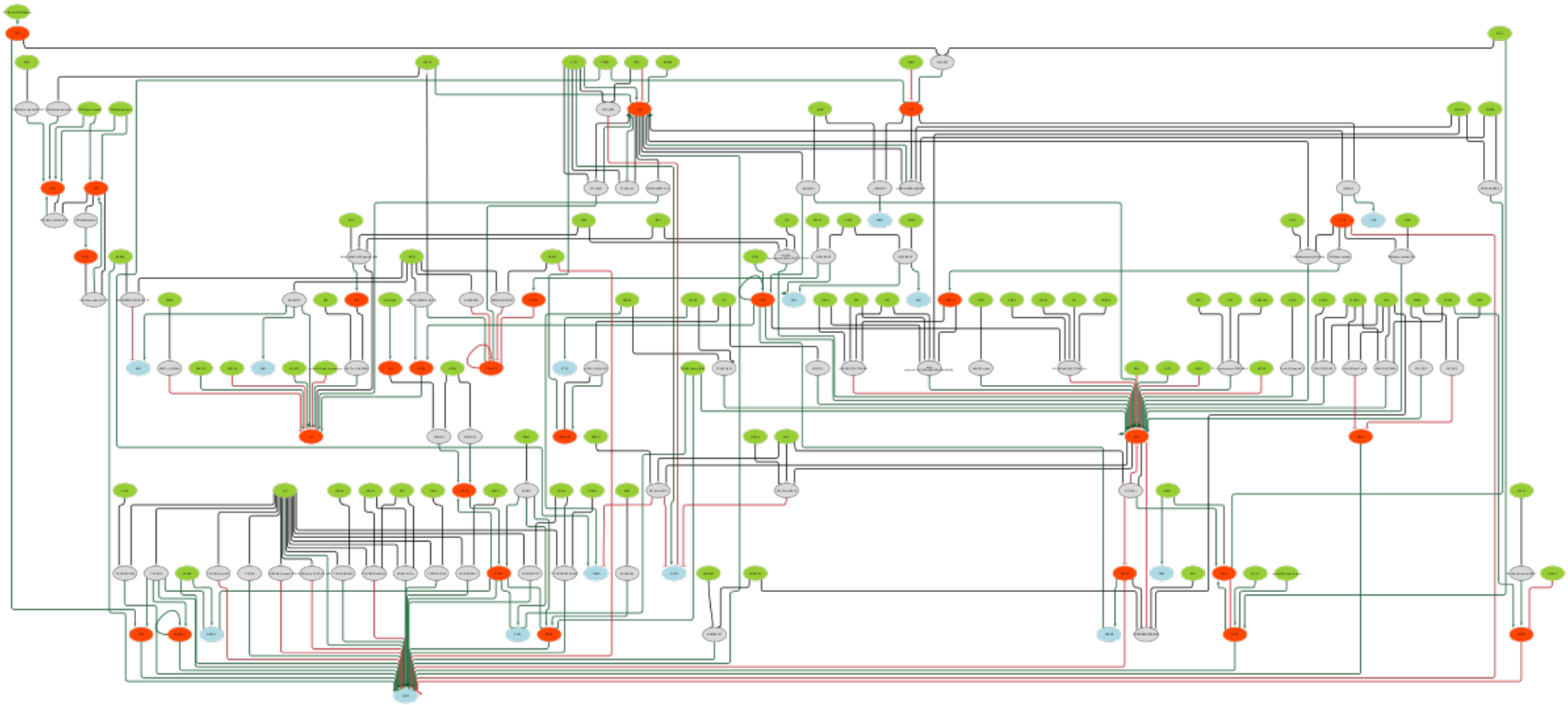
Familiy of Boolean Networks for Class A



Familiy of Boolean Networks for Class B

Reconstructed PKN

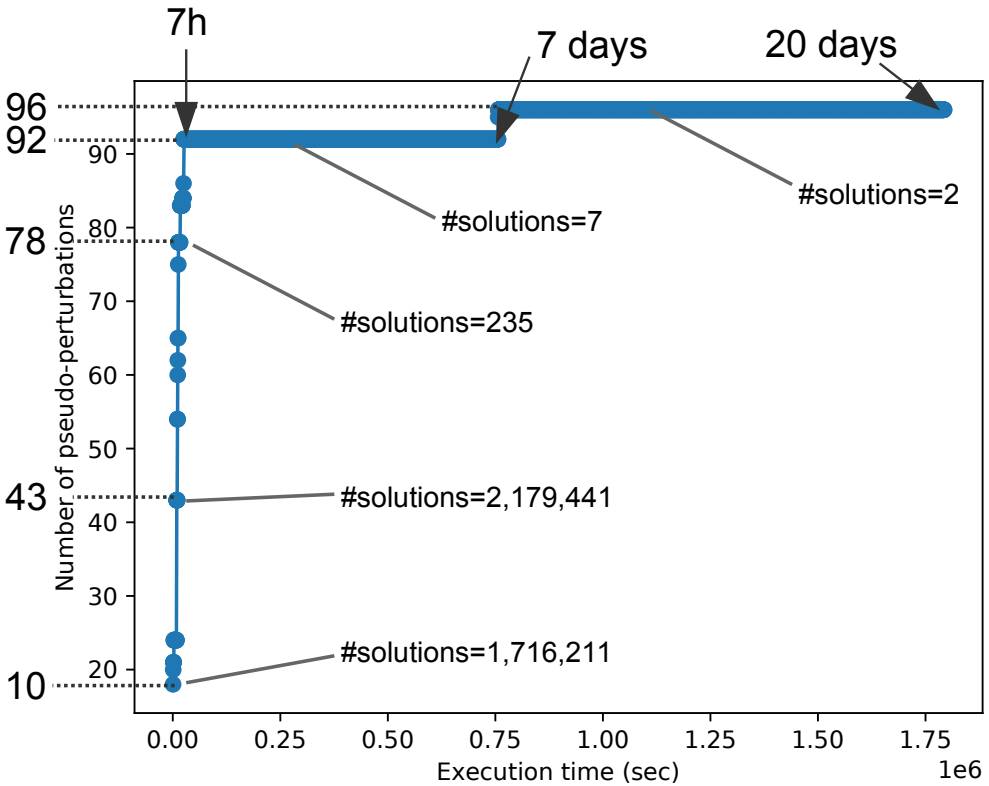
- From 438 TFs
- 233 nodes : inputs (85), intermediates (36), readouts (19)
- 369 edges



Pseudo-perturbations search

Inputs

- $\#M^{TE}$ *cells* $= 348$
- $\#L^{TE}$ *cells* $= 332$
- $k = 10$: 10 genes selected from 121 input and intermediate genes
- Complexity: 8.01×10^{34793}

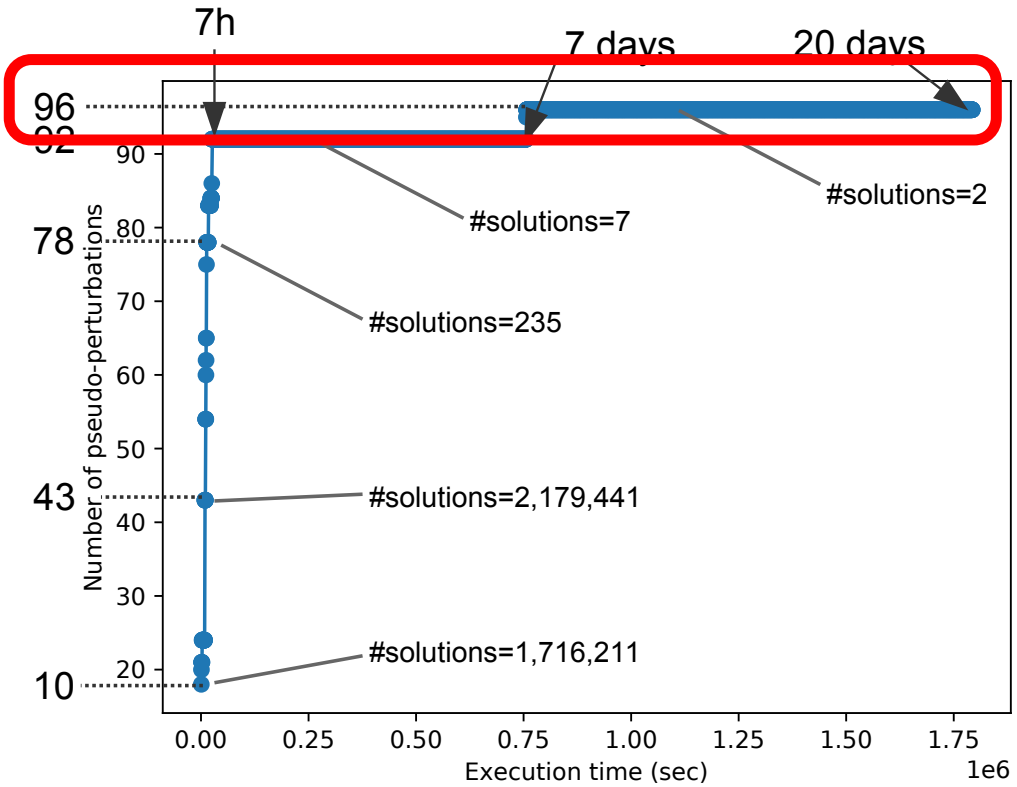


Convergence of the number of pseudo-pertubations over time.

Pseudo-perturbations search

Inputs

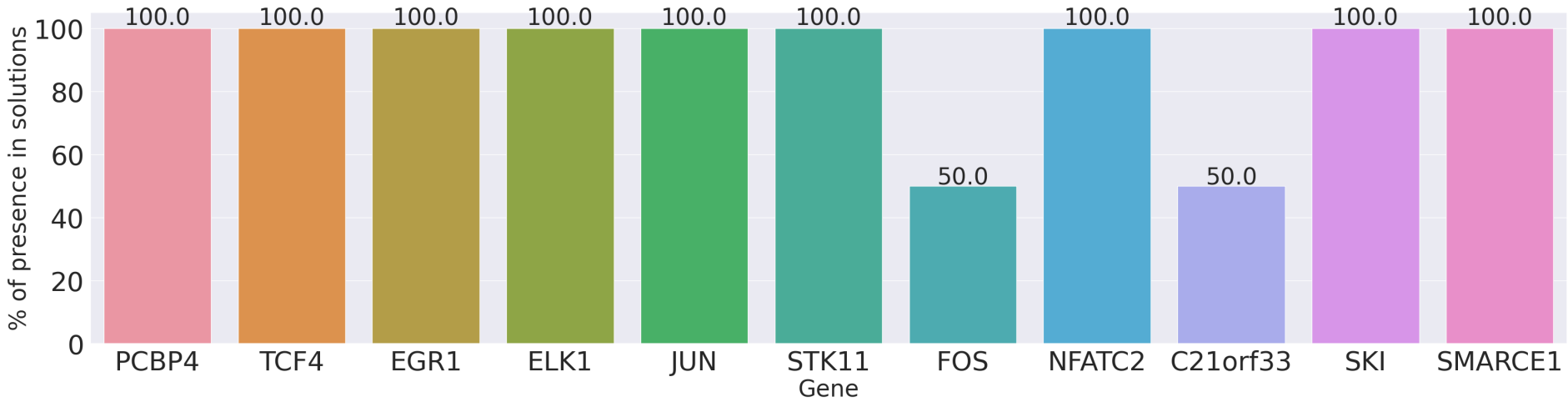
- $\#M^{TE}$ *cells* $= 348$
- $\#L^{TE}$ *cells* $= 332$
- $k = 10$: 10 genes selected from 121 input and intermediate genes
- Complexity: 8.01×10^{34793}



Convergence of the number of pseudo-pertubations over time.

96 pseudo-perturbations sub-optimal solution

- Number of solution = 2
- Different genes in solutions = 11



11 characteristic genes
to have the same Boolean behavior in M^{TE} and L^{TE}

Cell representativity (redundancies)

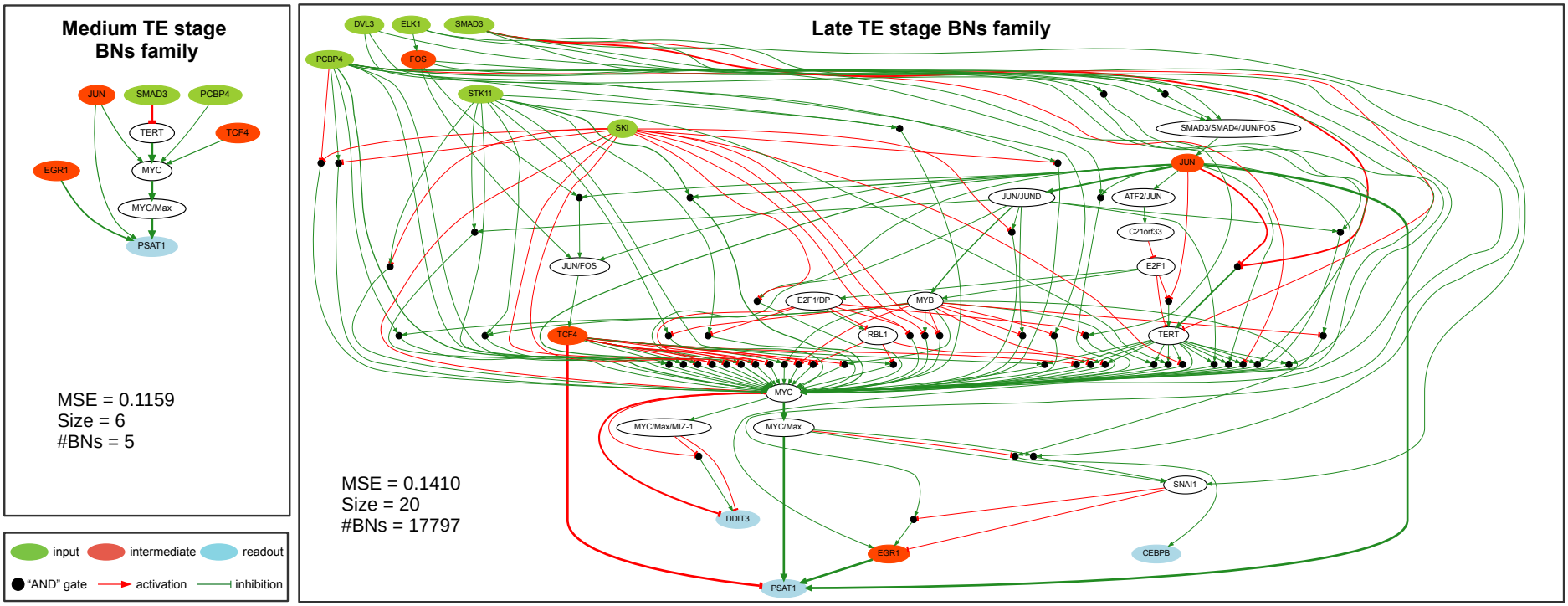
Solution	M^{TE} (%)	L^{TE} (%)	Total (%)
1	**266 (76%)**	**246 (74%)**	**512 (75%)**
2	235 (68%)	248 (75%)	483 (71%)

- $\#M^{TE}\,cells = 348$
- $\#L^{TE}\,cells = 332$
- $\#Total\ cells = 680$

- 96 pseudo-perturbations

On average, 73% of representativity for the total number of cell at each stage.

Inferred BNs families for solution 1



- Greater BNs variability for L^{TE} $\rightarrow$ Gain of function
- L^{TE} seems more unstable (number of BNs) $\rightarrow$ transition from L^{TE} to another stage

Conclusion

Pseudo-perturbation generation

- Efficient algorithm to select cells and genes to generate pseudo-perturbations $\rightarrow$ 92 pseudo-perturbations in 7h
- Robustness of the generated solutions $\rightarrow$ from $+2$ millions of solutions to only 2
- Discovery of 11 genes whose on/off values remain identical for 96 cells across 2 classes
- Expression of 11 genes across 96 cells are representative of the cell populations (*e.g.* 72% in M^{TE} and 73% in L^{TE})
- Our method deals with single cell data and its specificities (redundancy and sparsity)

General method

- A method that learns Boolean networks of 2 stages using scRNAseq data and PKN
- Mechanisms of TF-gene regulations distinguishing 2 developmental stages
- Overall approach achieves a good computational time ($\sim$ 1 day)
- Complementarity with the state of the art
  - Boolean models without using perturbations
  - Method taking into account the diverse states of cell population

Aknowledgements

- Jérémie Bourdon @LS2N, Nantes University
- Carito Guziolowski @LS2N, Centrale Nantes
- Laurent David @CR2TI, Nantes University Hospital, Nantes University
- ANR AIBY4 & ANR BOOSTIVF