

MaBoSS

Markovian Boolean Stochastic Simulator

Algorithm: Boolean Kinetic Monte-Carlo

G. Stoll (NSERM, UMR 1138, Eq. G. Kroemer, Paris, France)

E.Viara (SYSRA, Yerres, France),

L. Calzone (INSERM U900/Institut Curie, Paris, France)

3rd CoLoMoTo meeting

Motivation for continuous time algorithm

Synchronized / asynchronous algorithms are defined on discrete time steps.

=> Problems:

- Comparison between model and experimental results mainly on final states (when does a biological experimental system reach its final state?)
- Difficulty of modeling transient effects (e.g. in cell cycle)
- Difficulty of interpreting non-deterministic trajectories
- Difficulty of implementing different time scales of events (e.g. a phosphorylation is faster than a transcription)

=> General idea: fill the gap between ODE and discrete time Boolean modeling with **continuous time Boolean modeling**

Basic properties

Principles:

1. The state of each node is given by a Boolean number (0 or 1) (referred to as **node state**); $S_i \in \{0, 1\}, i = 1, \dots, n$
2. The state of the network is given by the set of node states (referred to as **network state**); \mathbf{S}
3. The update of a node state is based on the signs and the logic linking the incoming arrows of this node;
4. Time is parameterized by a real number;
5. Time evolution is stochastic.

→ Mathematical model: **continuous time Markov process**. $s(t)$

Continuous time Markov process

A model that is represented by a continuous time Markov process is defined by:

- Initial condition: probability distribution over network states at time 0
- Transition rates: $\rho_{(\mathbf{s}' \rightarrow \mathbf{s})}(t) \in [0, \infty[$
=> They can be interpreted as the inverse of mean time of the transition.

Transition rates in MaBoSS

A transition rate is non-zero only if the transition from S to S' consists of flipping one single node (generalization of asynchronous Boolean dynamics).

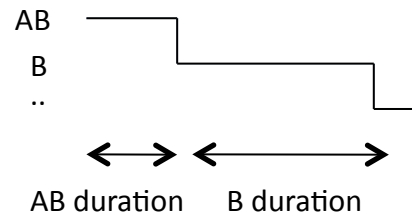
Transition rates are defined with specific language within MaBoSS software: conditional rate of activation/inhibition for each node (operators: AND (&), OR (|), NOT(!), if (? :), +, -, *, /)

```
node ROS
{
  logic = (!NFkB) & (MPT | RIP1K );
  rate_up = @logic ? 0.01 : 0.0;
  rate_down = @logic ? 0.0 : 1.0;
}
```

Algorithm in MaBoSS

Kinetic Monte-Carlo or Gillespie algorithm is a method for producing realizations of a continuous time Markov process.

A realization is a set of network states with their respective duration.



MaBoSS (**Markov Boolean Stochastic Simulator**) is a software that applies this algorithm (**Boolean Kinetic Monte-Carlo**) to a Boolean network, where transition rates are given within a specific language.

MaBoSS can also simulate discrete time asynchronous dynamics.

Basic outputs

- Time dependent instantaneous probabilities: $\mathbf{P} [s(t) = \mathbf{S}]$
- Set of indecomposable stationary distributions (characterization of asymptotic behavior)
 - A continuous time Markov process always converges to a stationary distribution
 - *Stationary distribution*: Markov process with constant instantaneous probabilities (different initial condition, same transition rates). Could be interpreted in term of cell population.
 - *Indecomposable stationary distribution*: stationary distribution that is not a linear combination of different stationary distributions. Could be interpreted in term of cell sub-populations.

Data processing of simulation algorithm

- Estimated probabilities on time window (should provide Δt).

$$\mathbf{P} [s(\tau) = \mathbf{S}] \equiv \frac{1}{\Delta t} \int_{\tau \Delta t}^{(\tau+1)\Delta t} dt \mathbf{P} [s(t) = \mathbf{S}]$$

- Estimated set of indecomposable stationary distributions by clustering time average over each trajectories.

Large network (nb of nodes is a compilation parameter of MaBoSS) \rightarrow huge number of network states \rightarrow global characterization:

Entropy, transition entropy, Hamming distance distribution.

Global characterizations

- **Entropy:** $H(\tau)$
0 if one single network state has probability 1,
#nodes if all network states have identical probabilities.
- **Transition entropy:** $TH(\tau)$
0 if only one node can flip,
 $\log_2(\text{\#nodes})$ if all nodes can flip with the same rate.
Measures how deterministic the time evolution is, at single cell level.

Consequences (definition of *signatures*):

- If the model converges to a fixed point, entropy and transition converges to 0.
- If the model converges to a cycle, only transition entropy converges to 0

MaBoSS practical use (1)

- 1) Define a model within MaBoSS language

Here model of p53/
MDM2 interaction
(Abou-Jaoudé)

```
node p53
{
  logic = NOT Mdm2N;
  rate_up = @logic ? 1.0 : 0.0 ;
  rate_down = ((NOT @logic) AND NOT p53_h) ? 1.0 : 0.0 ;
}
node p53_h
{
  logic = NOT Mdm2N;
  rate_up = (@logic AND p53) ? 1.0 : 0.0;
  rate_down = (@logic ? 0.0 : 1.0);
}
node Mdm2C
{
  logic = $case_a ? p53_h : p53;
  rate_up = @logic ? 1.0 : 0.0 ;
  rate_down = @logic ? 0.1 : 1.0 ;
}
node Mdm2N
{
  logic_p53 = $case_a ? p53 : p53_h;
  rate_up = (@logic_p53 AND Mdm2C) ? $KMn_pMC :
    ((@logic_p53 ? $KMn_p : 0.0) + (Mdm2C ? $KMn_MC : 0.0) +
    ((NOT @logic_p53 AND NOT Mdm2C) ? $KMn : 0.0));

  rate_down = (@logic_p53 AND Mdm2C) ? (1-$KMn_pMC) :
    ((@logic_p53 ? (1-$KMn_p) : 0.0) + (Mdm2C ? (1-$KMn_MC) : 0.0) +
    ((NOT @logic_p53 AND NOT Mdm2C) ? (1-$KMn) : 0.0));
}
```

MaBoSS practical use (2)

2) Set simulation parameters (default values provided in reference card):

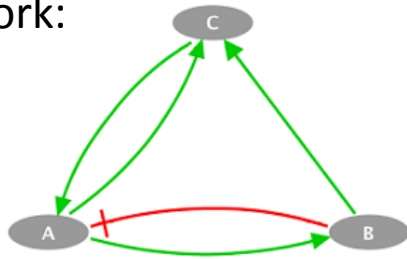
- Number of trajectories
- Maximum time
- Time window (for time window probabilities)
- Internal nodes (not used for computing probabilities and entropies)
- Nodes initial condition (random if not specified)
- Number of trajectories for stationary distribution estimates
- Threshold for stationary distribution estimates (clustering algorithm)

3) Run MaBoSS: command line C++ software. The output is composed of:

- .csv file with time window probabilities (with error), entropy, transition entropy (with error)
- .csv file with set of indecomposable stationary distribution (with error)

Example 1: toy model

Network:



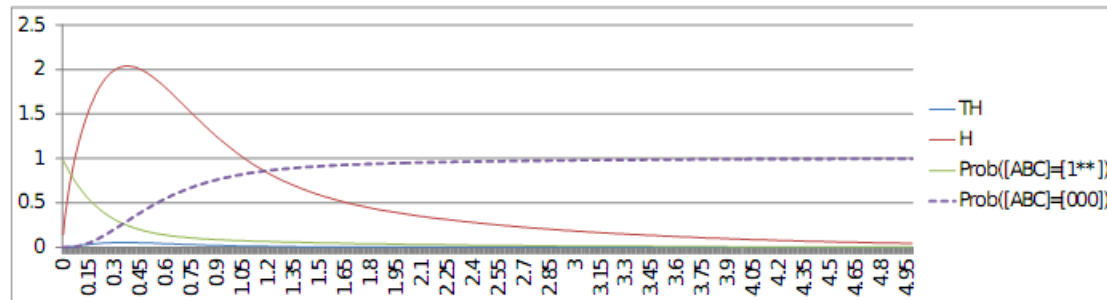
```

Node A {
rate_up=(C AND (NOT B)) ? $Au : 0.0;
rate_down= B ? $Ad : 0.0;
}

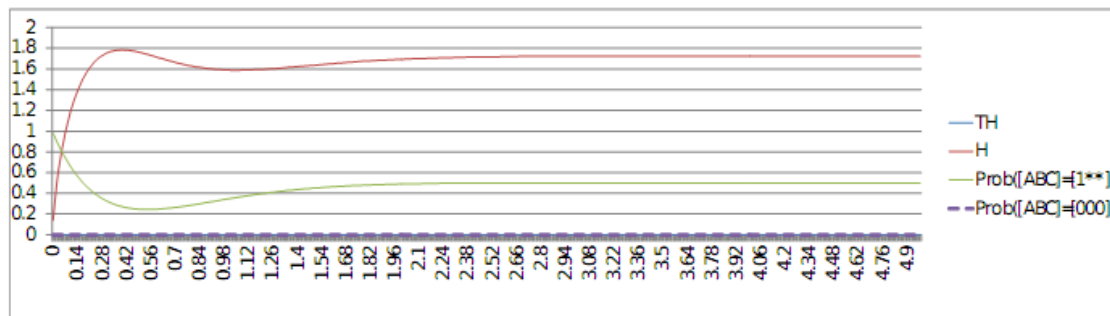
Node B {
rate_up= A ? $Au : 0.0;
rate_down = A ? 0.0 : $Ad ;
}

Node C{
rate_up=0.0;
rate_down=((NOT A) AND (NOT B)) ? $escape : 0.0 ;
}
    
```

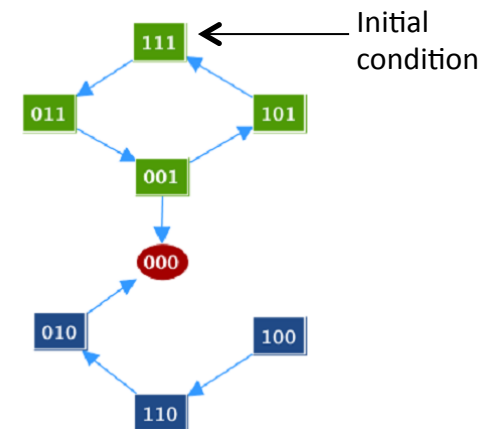
Normal transition rates: fixed point



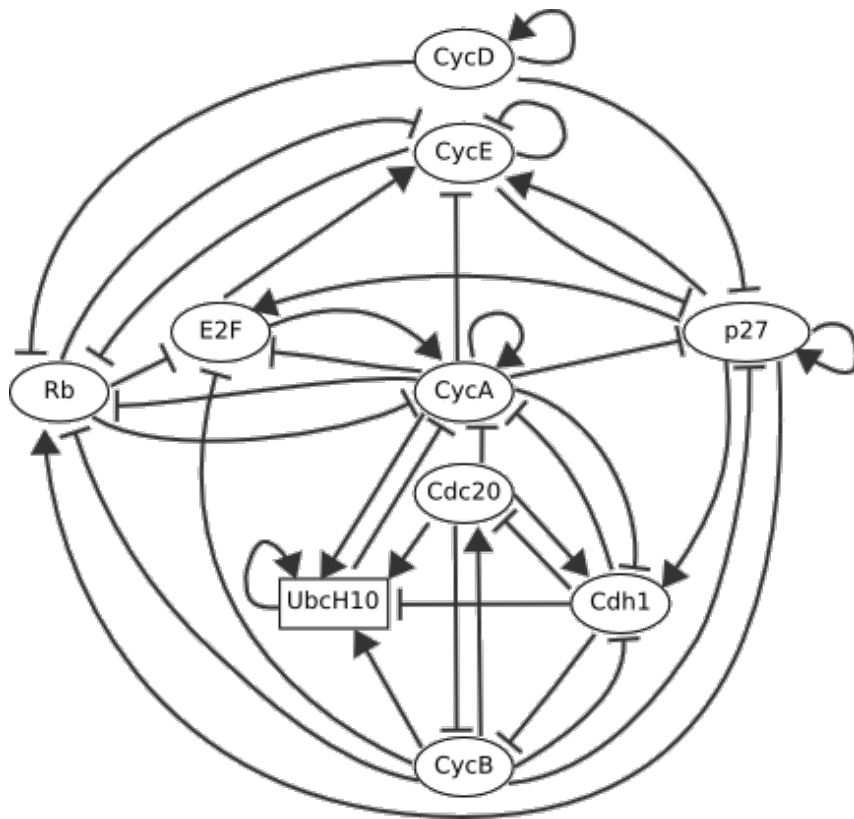
Extremely slow transition 001 → 000: pseudo-cycle



Graph of non-zero transition rates: (transition graph)



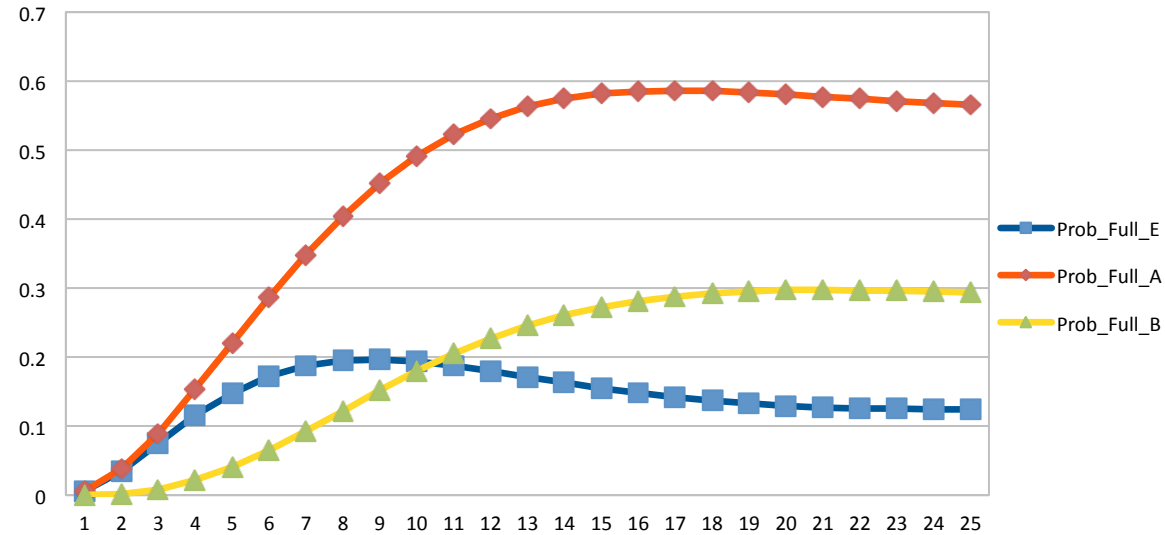
Example with mammalian cell-cycle model



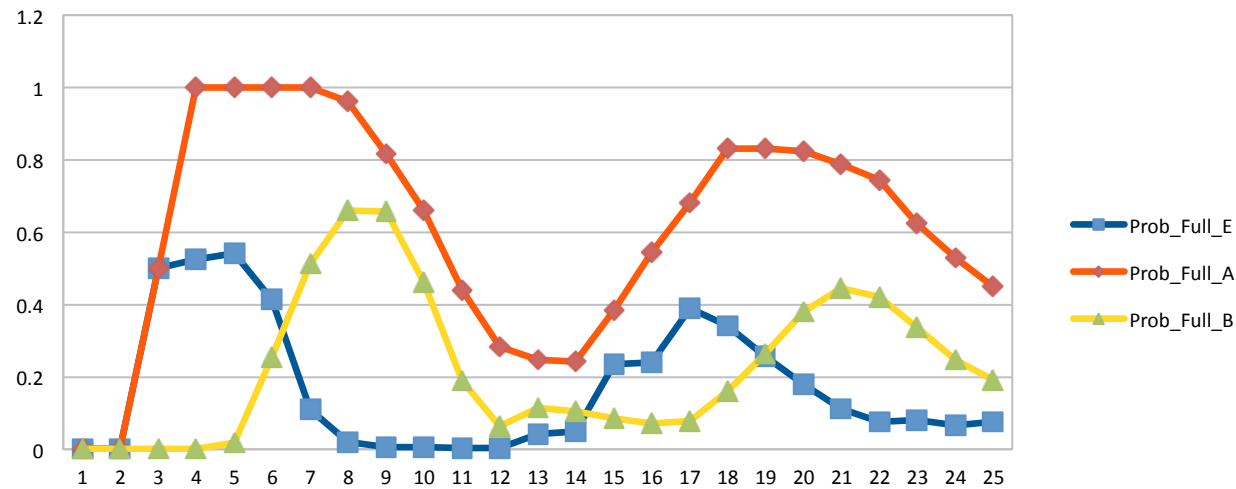
- Model from GINsim repository.
- Same logical rules
- Priority rules replaced by two scales of transition rates: fast=10, slow =1
- No synchronized update added

Trajectories with G1 initial condition (cycl_D on)

Cyclins in continuous modeling (damped oscillation)

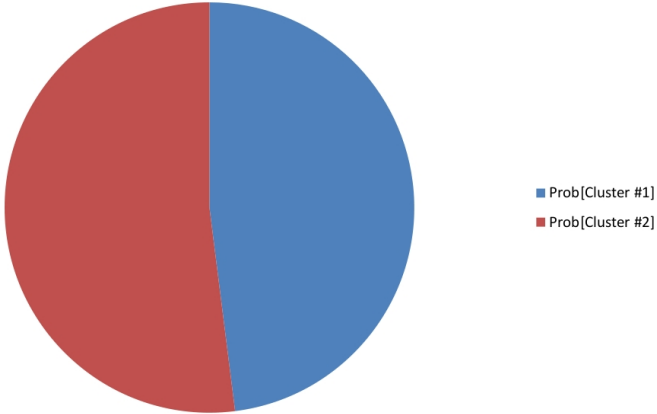


Cyclins in discrete modeling

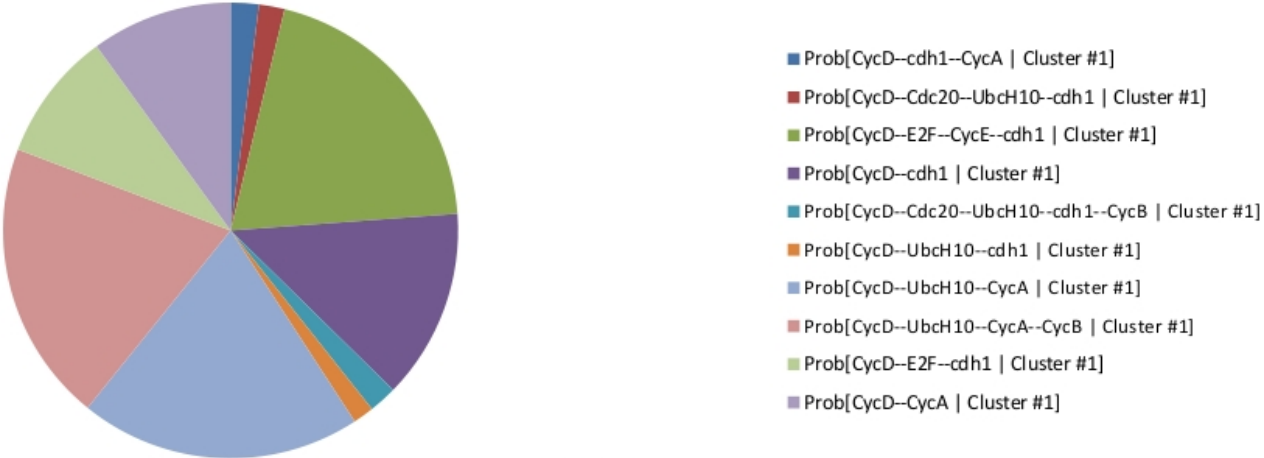


Stationary distribution with random initial condition

Two indecomposable stationary distributions:



Desynchronized proliferating cell population:



Fixed point, representing the G1 phase:

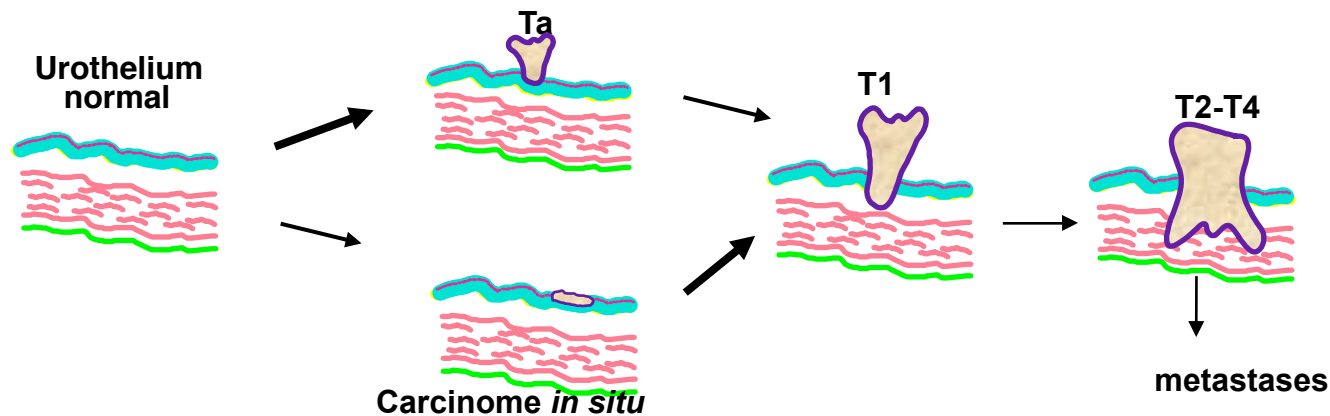


Application on bladder tumorigenesis

Bladder cancer

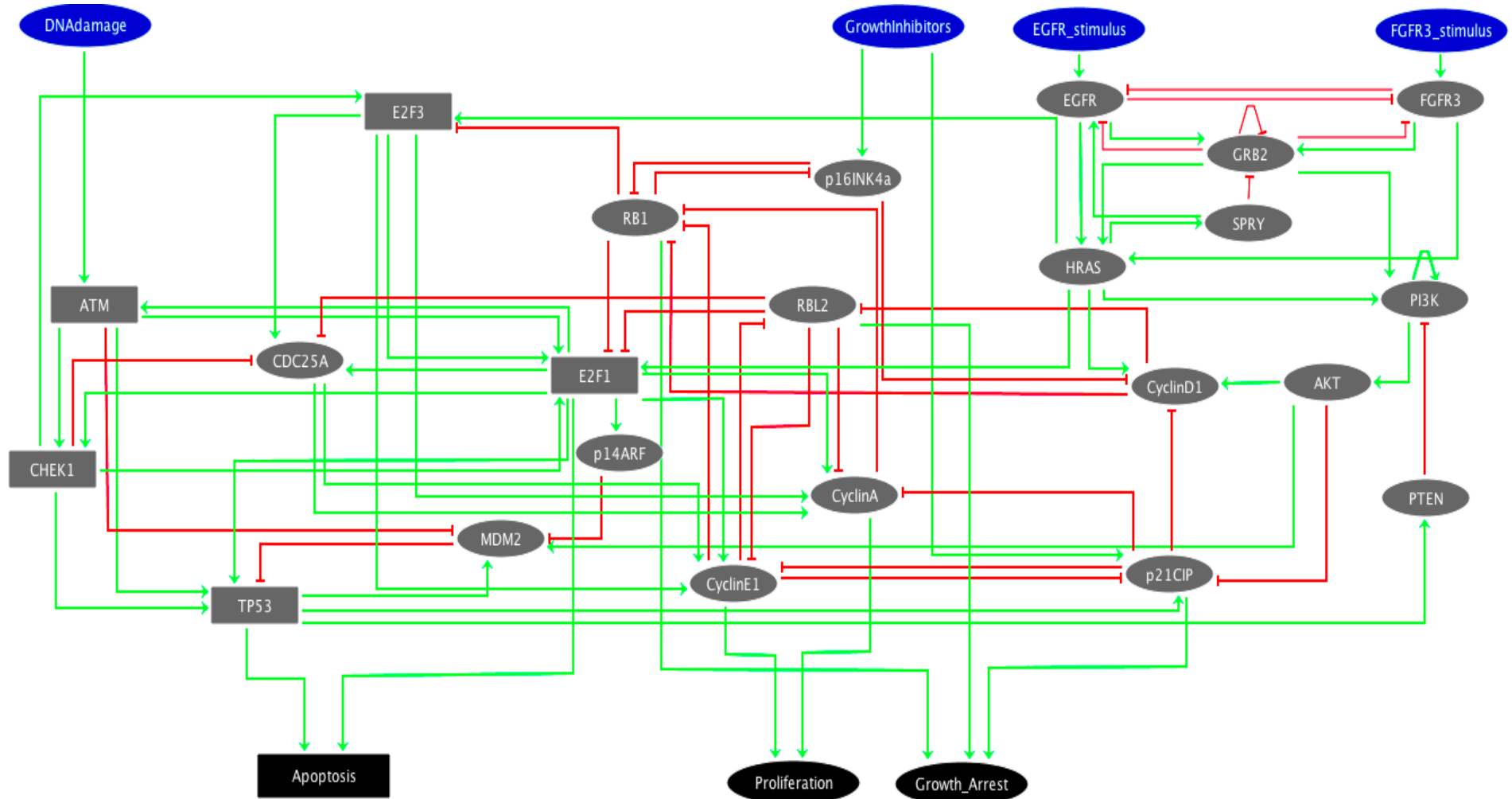
There are two paths that lead to aggressive tumours in bladder cancers:

- Ta pathway associated to FGFR3 mutations (less invasive)
- CIS pathway associated to TP53 mutations and no FGFR3 mutations (more invasive)



=> How do tumour cells become invasive in both FGFR3 mutated and non mutated cells?

The regulatory graph



The rules

Node	Value	Logical function	Comment
DNA damage	0/1	Constant (input)	
GrowthInhibitors	0/1	Constant (input)	
EGFR_stimulus	0/1	Constant (input)	
FGFR3_stimulus	0/1	Constant (input)	
EGFR signaling pathway			
EGFR	1	(EGFR_stimulus SPRY) & !FGFR3 & !GRB2	
FGFR3	1	FGFR3_stimulus & !EGFR & !GRB2	
GRB2	1	(FGFR3 & !GRB2 & !SPRY) EGFR	
SPRY	1	RAS	
RAS	1	EGFR FGFR3 GRB2	
PI3K	1	GRB2 & RAS & !PTEN	
AKT	1	PI3K	
PTEN	1	TP53	
CyclinD1	1	(RAS AKT) & !p16INK4a & !p21CIP	
p16INK4a	1	GrowthInhibitors & !RB1	
p14ARF	1	E2F1	
RB1	1	!CyclinD1 & !CyclinE1 & !p16INK4a & !CyclinA	
RBL2	1	!CyclinD1 & !CyclinE1	
p21CIP	1	(GrowthInhibitors TP53) & !CyclinE1 & !AKT	
CDC25A	1	(E2F1 E2F3) & !CHEK1_2 & !RBL2:1	
CyclinE1	1	CDC25A & (E2F1 E2F3) & !RBL2 & !p21CIP	
CyclinA	1	(E2F1 E2F3) & CDC25A & !p21CIP & !RBL2	
E2F1	1	((!(CHEK1_2:2 & ATM:2) & (RAS E2F3:1 E2F3:2)) (CHEK1_2:2 & ATM:2 & !RAS & E2F3:1)) & !RB1 & !RBL2	
	2	(RAS E2F3:2) & CHEK1_2:2 & ATM:2 & !RB1 & !RBL2	
E2F3	1	RAS & !RB1 & !CHEK1_2:2	
	2	RAS & !RB1 & CHEK1_2:2	
ATM	1	DNA damage & !E2F1	
	2	DNA damage & E2F1	
CHEK1_2	1	ATM & !E2F1	
	2	ATM & E2F1	
MDM2	1	(TP53 AKT) & !p14ARF & !ATM	
TP53	1	((ATM & CHEK1_2) E2F1:2) & !MDM2	
Cell cycle and apoptosis			
Apoptosis	1	!E2F1 & TP53	
	2	E2F1:1 E2F1:2	
Proliferation	1	CyclinE1 CyclinA	
GrowthArrest	1	p21CIP RB1 RBL2	

The solutions: fixed points, cyclic attractor (purple)

EGFR_stimulus	FGFR3_stimulus	DNA_damage	Growth_inhibitor	Proliferation	Apoptosis	Growth_arrest	EGFR	FGFR3	HRAS	E2F1	E2F3	CyclinD1	CylinE1	CylinA	CDC25A	P16INK4a	RB1	RBL2	p21CIP	ATM	CHEK1	MDM2	TP53	p14ARF	PTEN	PI3K	AKT	GRB2	SRPY	Number of states
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1
0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	1	0	1	0	0	0	0	1
0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	1	0	1	0	0	0	0	1
0	1	0	0	1	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
0	1	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1
0	1	0	1	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	1	0	1	0	0	0	0	1
0	1	1	1	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	1	0	1	0	0	0	0	1
0	1	1	1	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	1	0	1	0	0	0	0	1
1	0	0	0	*	0	*	*	0	*	0/1	*	0	*	*	*	0	*	*	0	0	0	*	0	*	0	*	*	*	*	*
1	0	0	0	*	0	*	*	0	*	0	*	*	*	*	*	0	*	0	0	0	0	*	0	*	0	*	*	*	*	*
1	0	0	0	*	0	*	*	0	*	0/1	*	*	*	*	*	0	*	0	0	0	0	*	0	*	0	*	*	*	*	*
1	0	0	0	*	0	*	*	0	*	0/1	*	*	*	*	1	0	*	0	0	0	0	*	0	*	0	*	*	*	*	*
1	0	0	1	0	0	1	*	0	*	0	0/1	0	0	0	0	1	0	1	*	0	0	*	0	0	0	*	*	*	*	*
1	0	1	0	0	1	1	*	0	*	0	0	0	0	0	0	0	1	1	1	1	1	0	1	0	1	0	0	*	*	*
1	0	1	1	0	1	1	*	0	*	0	0/1	0	0	0	0	1	0	1	1	1	1	0	1	0	1	0	0	*	*	*
1	0	1	1	0	1	1	*	0	*	0	0/1	0	0	0	0	1	0	1	1	1	1	0	1	0	1	0	0	*	*	*
1	1	0	0	1	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
1	1	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1
1	1	0	1	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1
1	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	1	0	1	0	0	0	0	1
1	1	1	1	0	1	1	0	1	1	0	1	0	0	0	0	1	0	1	1	1	1	0	1	0	1	0	0	0	0	1
1	1	1	1	0	1	1	0	1	1	0	1	0	0	0	0	1	0	1	1	1	1	0	1	0	1	0	0	0	0	1
1	1	1	1	0	1	1	0	1	1	0	1	0	0	0	0	1	0	1	1	1	1	0	1	0	1	0	0	0	0	1

184320

512

16

16

32

1

1

1

1

1

1

1

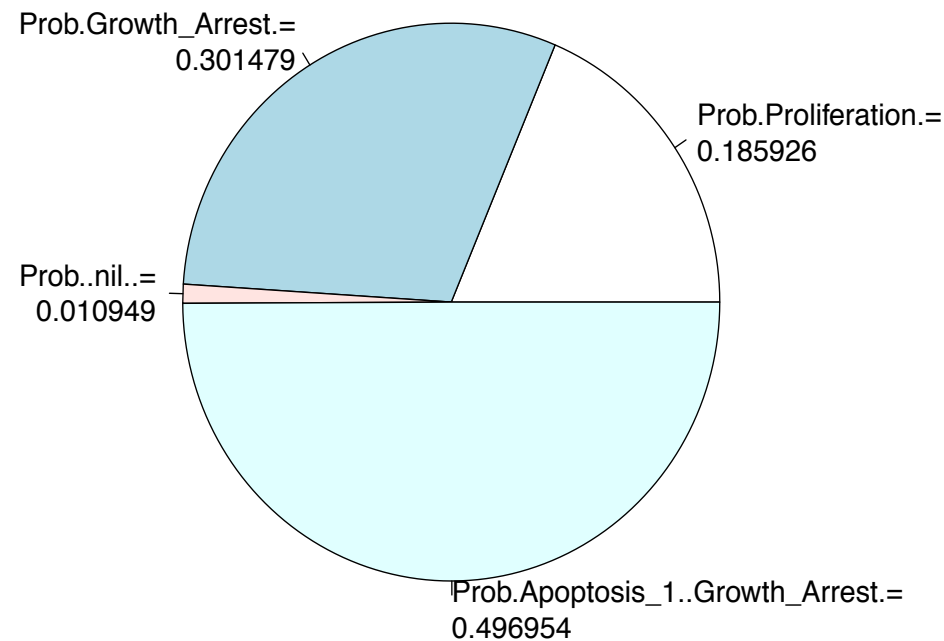
MaBoSS solutions

Rate of transcriptional influence: 0.04 (need to add mRNA node in 10 cases)

Rate of post-transcriptional influence: 1

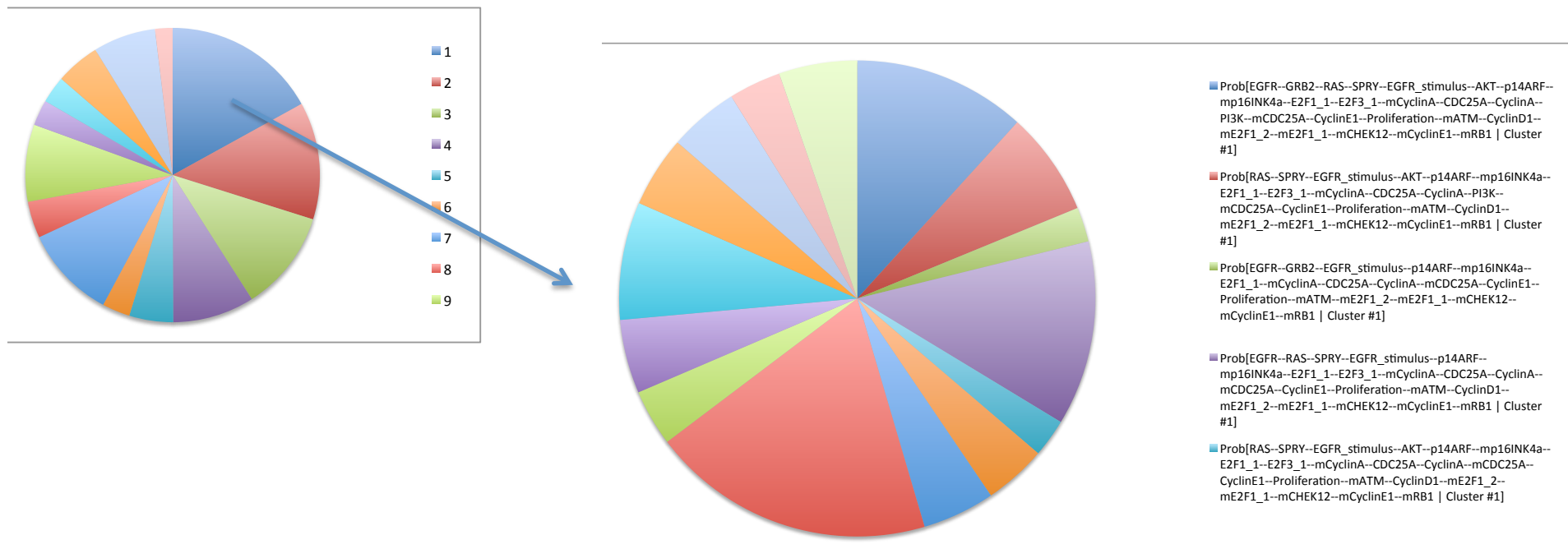
Time for reaching stationary behaviour: 140

Stationary distribution: (asymptotic behavior of cell population)



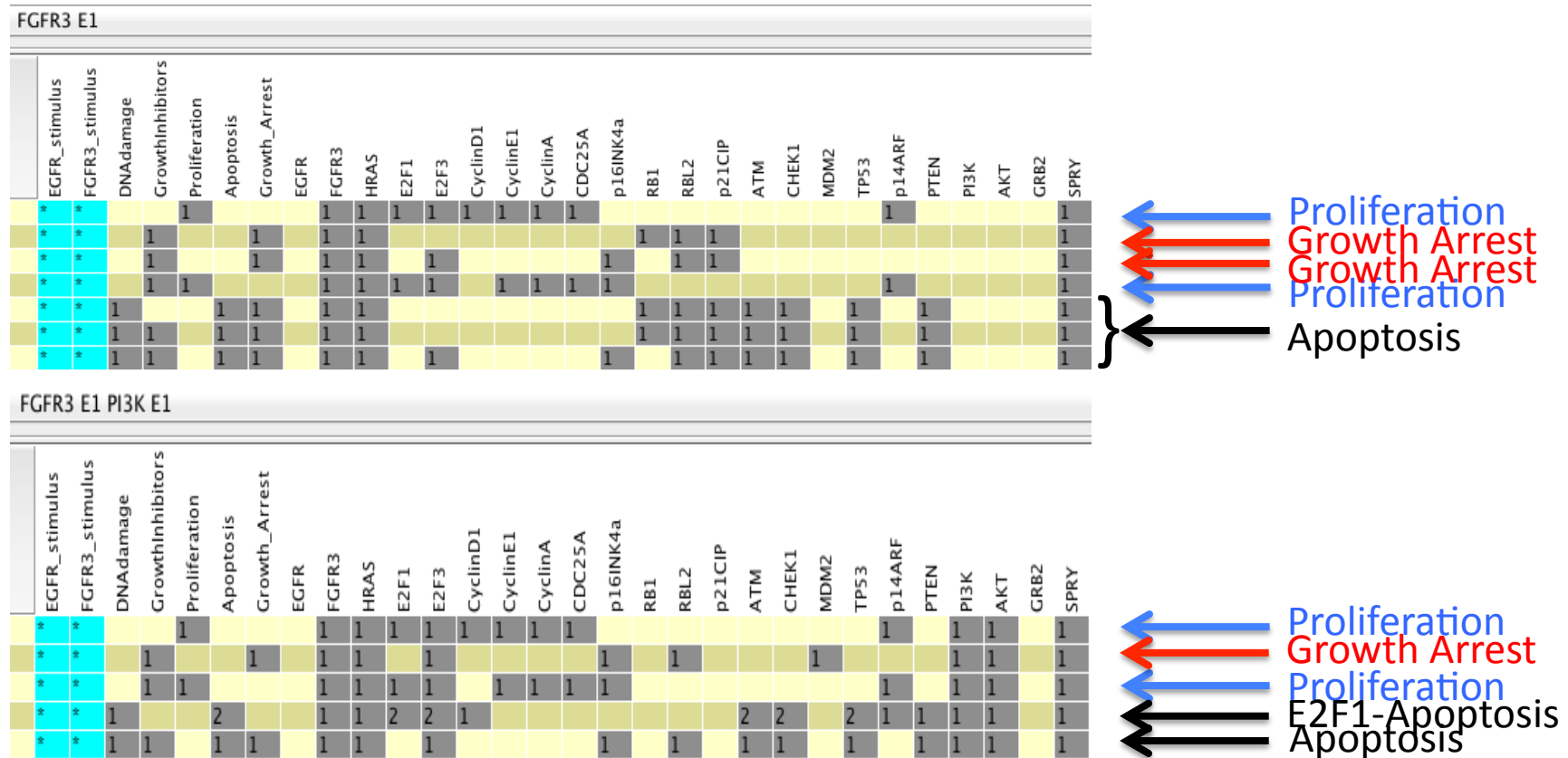
MaBoSS solutions

- Stationary distribution: decomposition in fixed point/cyclic attractors.
- Random initial condition -> 66 fixed points (version with mRNA nodes!)
- Search for cyclic attractors (different than fixed points) with initial condition: EGFR.stimulus=1, FGFR3.stimulus=0, FGFR3=0
=> 9 attractors (estimation!)



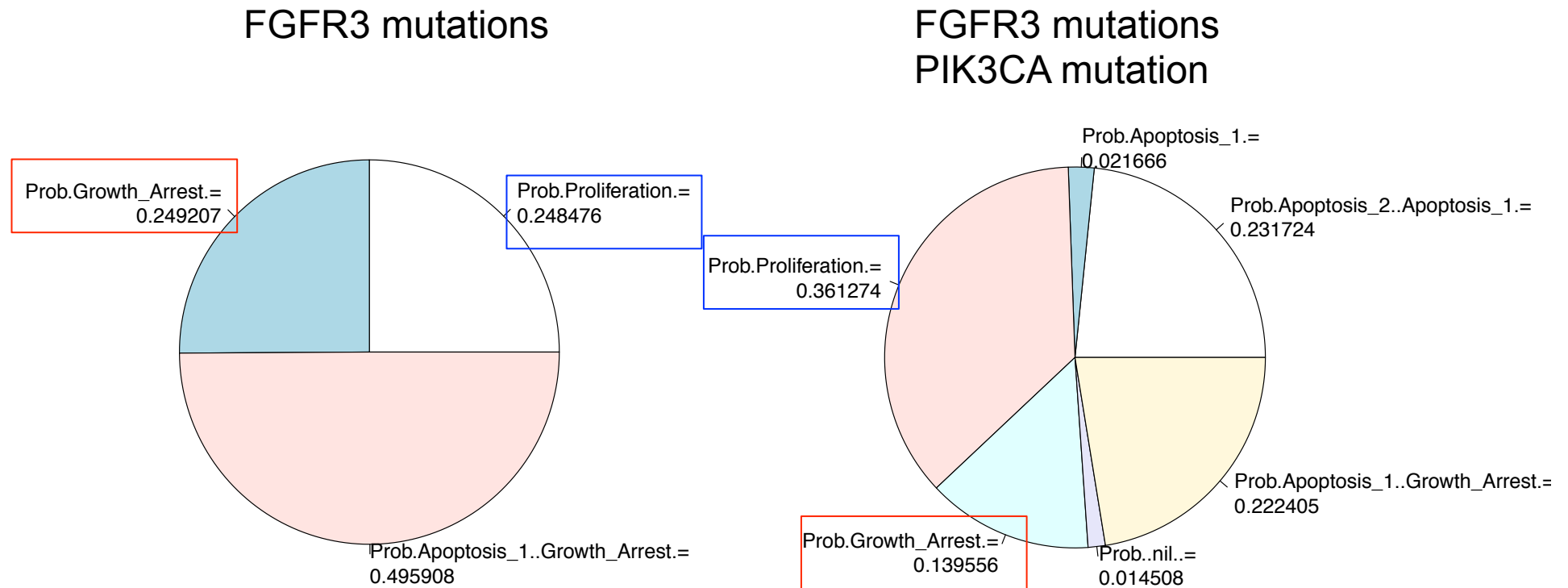
Use MaBoSS to answer questions quantitatively

Q: Why are mutations of *PIK3CA* often observed with *FGFR3* mutations?



⇒ Difficult to conclude anything

Q: *Why are mutations of PIK3CA often observed with FGFR3 mutations?*



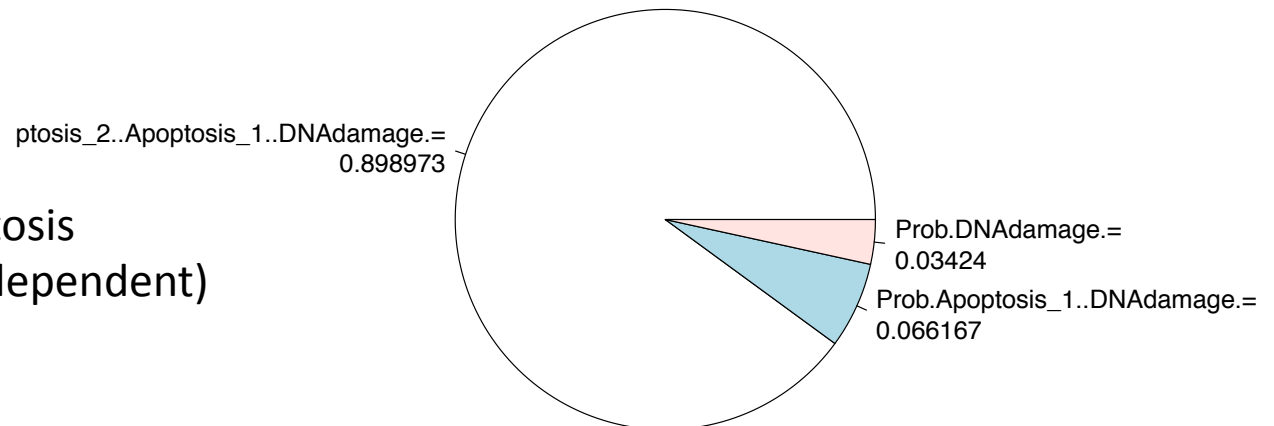
⇒ Probability of “proliferation” is increased and probability of “growth arrest” is decreased in the double mutant.

It is “advantageous” to the cancer cell to mutate PIK3CA in a FGFR3 mutated cell

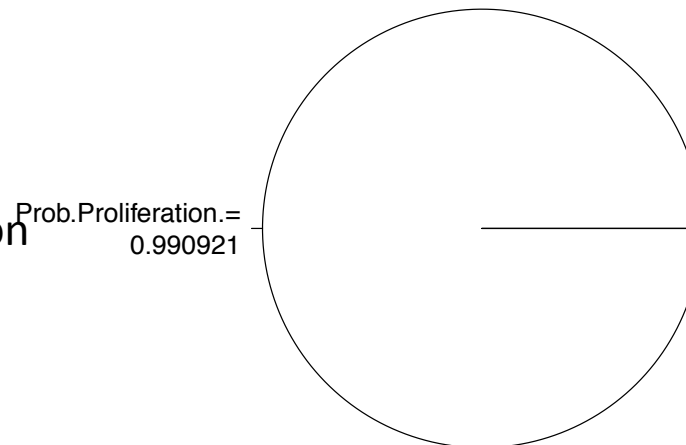
Prediction of the model

Third deletion of CDKN2A (p16INK4a + p14ARF) eliminates growth arrest in absence of DNA damage

With DNA damage, only apoptosis
(both E2F1-dependent and independent)



Without DNA damage, only proliferation
⇒ Uncontrolled growth
⇒ Very invasive tumours



Verification in the data

- 162 tumours
- 12 tumours are FGFR3-mutated & PIK3CA-mutated and 2 are invasive, 2 non-invasive
- The 2 invasive tumours are CDKN2A deleted
- Correlation between invasiveness and CDKN2A HD in FGFR3 & PIK3CA mutated tumours:
 - p-value=0.1 in our dataset (CIT)
 - p-value=0.09 in another public dataset (Lindgren et al.)

⇒ Prediction not conclusive?

⇒ Not enough data?

⇒ The two alterations (FGFR3 & PIK3CA) may not be that frequent / biased view (only in non-invasive)

⇒ FGFR3 mutations and CDKN2A homozygous deletions are highly correlated to invasiveness

Linking model solutions to data

Simple method

1. Data discretization of transcriptomics data

$$\forall \text{ gene } k : BoolExpr_k = Expr_k - Avg_k(P)$$

P = set of samples (patients)

	Pat1	Pat2	Pat3	Pat4	Pat5
FGFR3	10.68	6.09	12.99	7.85	14.58
PIK3CA	10.06	9.61	10.72	9.84	10.22
MDM2	13.00	11.59	12.01	12.57	15.99
SPRY1	10.57	10.88	10.24	9.59	11.41
CDC25A	8.44	9.78	9.76	9.94	10.27



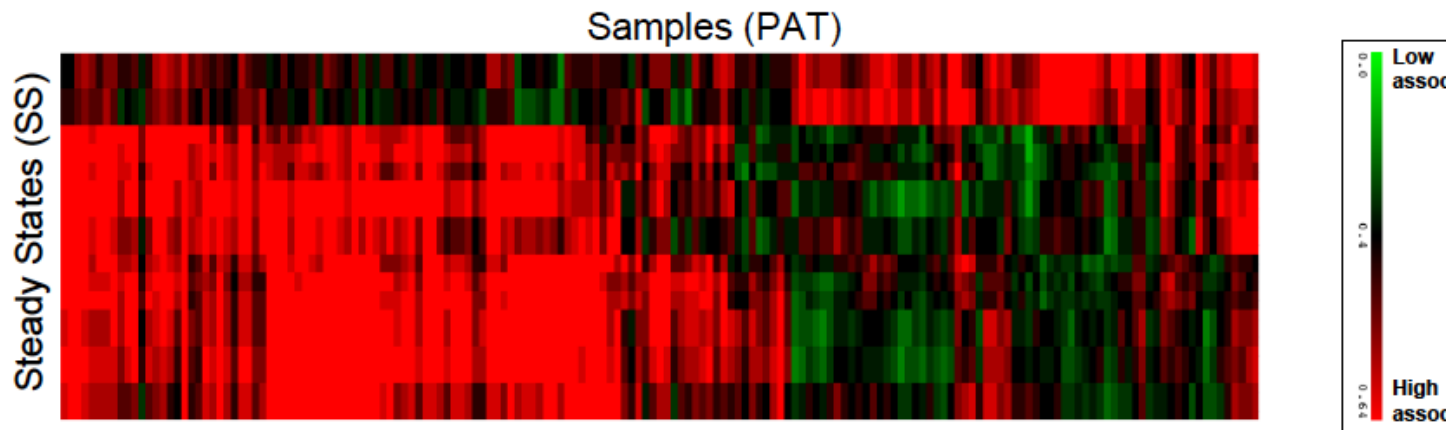
	Pat1	Pat2	Pat3	Pat4	Pat5
FGFR3	0	0	1	0	1
PIK3CA	0	0	1	0	1
MDM2	1	0	0	1	1
SPRY1	1	1	0	0	1
CDC25A	0	1	1	1	1

Linking model solutions to data

2. Association score: how close a sample is to a model stable state

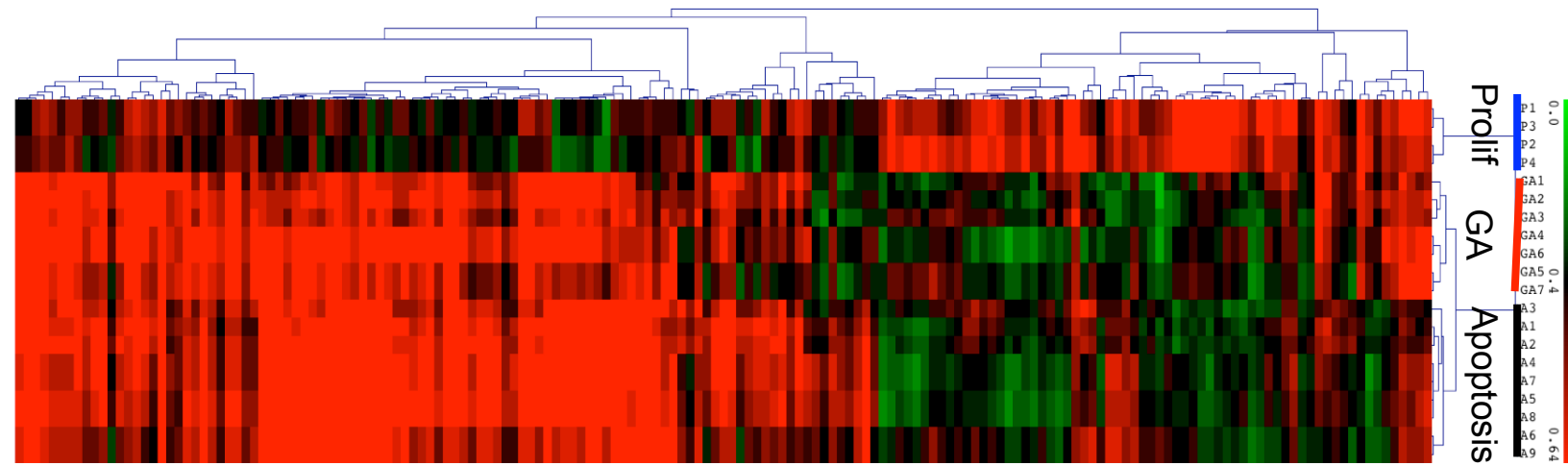
$$S_{ij} = \frac{\sum_{k=1}^N (SS_i[k] = PAT_j[k])}{N}$$

$N = \text{set of model variables (genes)}$



Linking model solutions to data

3. Hierarchical clustering of the scoring matrix (using MeV software)



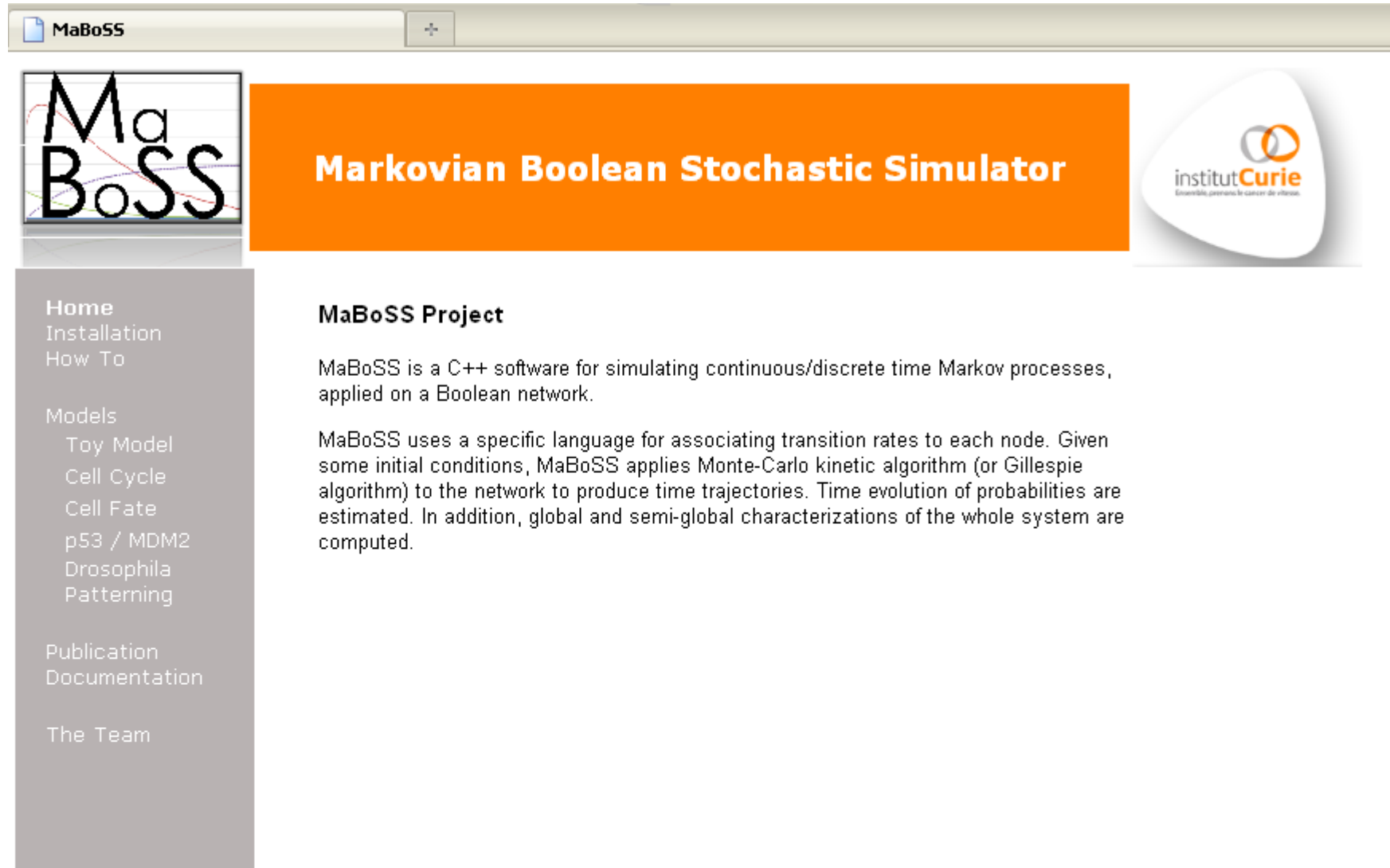
Mostly non-invasive
Mostly FGFR3-non-mutated

Mostly invasive
Mostly FGFR3-mutated
Related to KI67 expression
(proliferation marker)

Linking model solutions to data


- Refine association score
- How to use the information obtained?
 - Model validation
 - Identification of clusters of patients
 - Personalized patient prediction according to the clinical data compared to the clusters
- How far can we go with additional data?
 - Mutations (genome)
 - (phospho-)proteome
- How to integrate all data types into a “consensus” score
- Robust binarization?

<https://maboss.curie.fr>



MaBoSS

Markovian Boolean Stochastic Simulator



MaBoSS Project

MaBoSS is a C++ software for simulating continuous/discrete time Markov processes, applied on a Boolean network.

MaBoSS uses a specific language for associating transition rates to each node. Given some initial conditions, MaBoSS applies Monte-Carlo kinetic algorithm (or Gillespie algorithm) to the network to produce time trajectories. Time evolution of probabilities are estimated. In addition, global and semi-global characterizations of the whole system are computed.

- Home
- Installation
- How To
- Models
 - Toy Model
 - Cell Cycle
 - Cell Fate
 - p53 / MDM2
 - Drosophila Patterning
- Publication
- Documentation
- The Team

Future plans

- Apply MaBoSS to other pathways (autophagy, mitophagy, etc.)
- Translate MaBoSS language (e.g. GINsim \leftrightarrow MaBoSS)
- Implement in other environment (GINsim, R, user-friendly interface)
- MaBoSSpp: dynamical population modeling.

Acknowledgement

- MaBoSS
 - Eric Viara (developer)
- Bladder Model
 - Francois Radvanyi (biologist)
 - Sandra Rebouissou (biologist)
 - Elisabeth Remy (modeller)
 - Claudine Chaouiya (modeller)
- Model to Data
 - Luca Grieco
 - Elisabeth Remy
- Thanks for (constant) discussions
 - Aurelien Naldi
 - Denis Thieffry
 - Claudine Chaouiya

Global characterizations

- **Entropy:** $H(\tau)$
0 if one single network state has probability 1,
#nodes if all network states have identical probabilities.
- **Transition entropy:** $TH(\tau)$
0 if only one node can flip,
 $\log_2(\text{\#nodes})$ if all nodes can flip with the same rate.
Measures how deterministic the time evolution is, at single cell level.

Consequences (definition of *signatures*):

- If the model converges to a fixed point, entropy and transition converges to 0.
- If the model converges to a cycle, only transition entropy converges to 0

Semi-global characterization: Hamming distance distribution

- Need to provide a reference network state, S_{ref}
- Compute the probability that the model has HD different node states compares to S_{ref} , $P[HD]$

=> Semi-global characterization because HD is not bigger than #nodes.

Stochastic process

Consider a network of n nodes. \mathbf{S} is a Boolean state, i. e.

$$S_i \in \{0, 1\} \text{ for } i = 1 \dots n$$

A *stochastic process* is defined by a set of random variables (defined on the same probability space) indexed by a real parameter,

$$\mathcal{S}(t) \quad t \in I \subset \mathbb{R}$$

The process is given by the set of all possible joint probabilities,

$$p(\{\mathcal{S}(t) = \mathbf{S}(t)\})$$

Markov process

Definition: conditional probabilities in the future, related to the present and the past, depend only on the present, i. e.

$$p \left[\mathcal{S}(t_i) = \mathbf{s}^{(i)} \mid \mathcal{S}(t_1) = \mathbf{s}^{(1)}, \mathcal{S}(t_2) = \mathbf{s}^{(2)} \dots \mathcal{S}(t_{i-1}) = \mathbf{s}^{(i-1)} \right]$$
$$= p \left[\mathcal{S}(t_i) = \mathbf{s}^{(i)} \mid \mathcal{S}(t_{i-1}) = \mathbf{s}^{(i-1)} \right]$$

Transition rates in continuous time Markov process

A *continuous time* Markov process is completely defined by:

- Transition rates: $W(\mathbf{S} \rightarrow \mathbf{S}')$
- Initial condition: $p[\mathcal{S}(t_m) = \mathbf{S}]$

In that case, instantaneous probabilities $p[\mathcal{S}(t) = \mathbf{S}]$ solutions of a *master equation*:

$$\frac{d}{dt} p[\mathcal{S}(t) = \mathbf{S}] = \sum_{\mathbf{S}'} \{ W(\mathbf{S}' \rightarrow \mathbf{S}) p[\mathcal{S}(t) = \mathbf{S}'] - W(\mathbf{S} \rightarrow \mathbf{S}') p[\mathcal{S}(t) = \mathbf{S}] \}$$

Data analysis of Markov process realizations: entropies

Entropy is a global characterization of a probability distribution. Small entropy: probability is concentrated in few states.

Define entropy on time window:

$$H(\tau) = - \sum_{\mathbf{S}} \log_2 (p [\mathcal{S}(\tau) = \mathbf{S}]) p [\mathcal{S}(\tau) = \mathbf{S}]$$

Define transition entropy on time window:

$$TH(\tau) = - \sum_{\mathbf{S}} p [\mathcal{S}(\tau) = \mathbf{S}] \sum_{\mathbf{S}'} \log_2 (p_{\mathbf{S} \rightarrow \mathbf{S}'}) p_{\mathbf{S} \rightarrow \mathbf{S}'}$$

$p_{\mathbf{S} \rightarrow \mathbf{S}'}$ is the probability of the transition $\mathbf{S} \rightarrow \mathbf{S}'$.

Transition entropy measure how non-deterministic is the process. It gives an estimation of behavior at single cell level.