Modelling Biological Regulatory Networks using Formal Methods

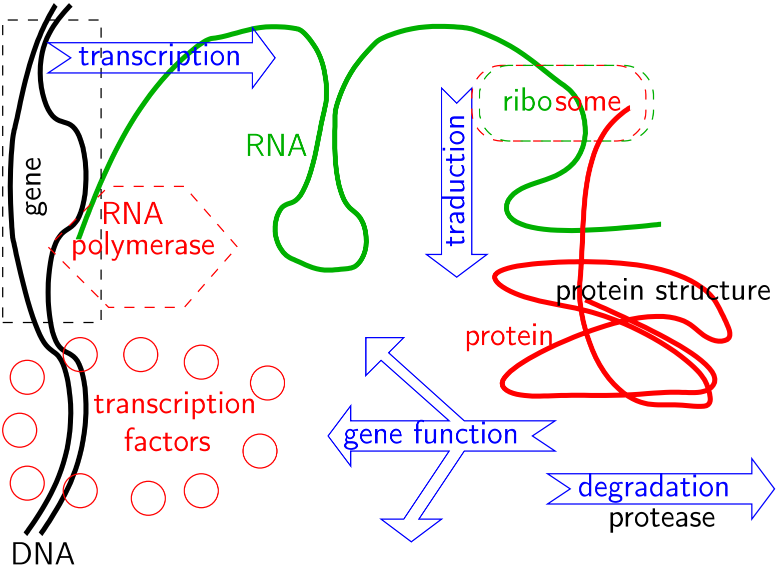
Gilles Bernot

Université Côte d'Azur, CNRS, I3S, France

Menu

- DNA, RNA, proteins and chemical kinetics of regulatory genes
- Discrete models for regulatory networks
- Hand made identification of parameters
- Regulatory networks and temporal logic
- The TotemBioNet approach
- Extracting interesting experiments from models
- Complex vs. complicated. . .

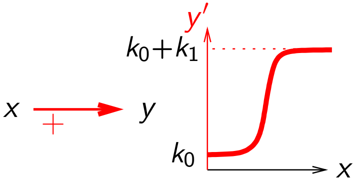
DNA, RNA, proteins and genes



Chemical kinetics of regulatory genes

Regulatory genes = Genes whose products regulate other genes

From concentration levels to production rates:



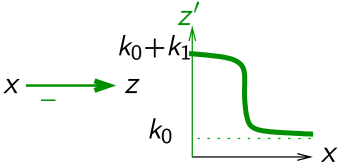
$$\frac{dy}{dt} = k_0^y + k_1^y . f_x^y(x) + \cdots - \gamma_y . y$$

k_0^y = minimal level of production

f_x^y: increasing sigmoid function, calibrated from 0 to 1

γ_y = degradation rate

Presence of an activator = Absence of an inhibitor



Similarly:

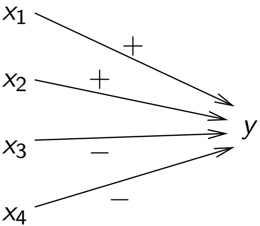
$$\frac{dz}{dt} = k_0^z + k_1^z . f_x^z(x) + \cdots - \gamma_z . z$$

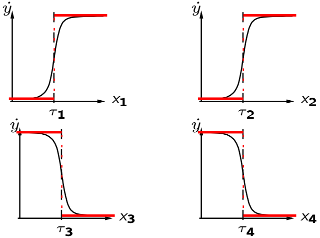
f_x^z is now *decreasing*

Experimental capabilities: hopeless to measure all k_i^v, f_u^v, γ_v!

First simplification: piecewise linear

Approximate sigmoids as step functions:





$$\frac{dy}{dt} = k_0 + k_1.\mathbb{1}_{x_1\geqslant\tau_1} + k_2.\mathbb{1}_{x_2\geqslant\tau_2} + k_3.\mathbb{1}_{x_3<\tau_3} + k_4.\mathbb{1}_{x_4<\tau_4} - \gamma.y$$

Solutions of the form $Ce^{-\gamma t} + \frac{\Sigma\mathbb{1}k_i}{\gamma}$ whose $\lim_{t\to\infty}$ is $\frac{\Sigma\mathbb{1}k_i}{\gamma}$

As many such equations as genes in the interaction graph

In each hypercube, all the trajectories have a unique *attractive point*, which can be outside de hypercube

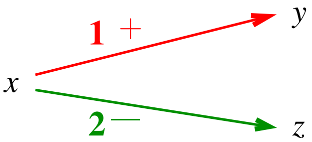
Experimental capabilities: hopeless to measure all k_i^v, τ_i^v, γ_v + does not capture non deterministic behaviours. . .

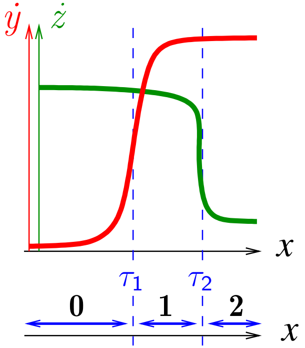
Menu

- DNA, RNA, proteins and chemical kinetics of regulatory genes
- **Discrete models for regulatory networks**
- Hand made identification of parameters
- Regulatory networks and temporal logic
- The TotemBioNet approach
- Extracting interesting experiments from models
- Complex vs. complicated. . .

Multivalued Regulatory Graphs

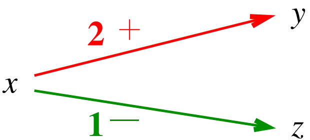
Homogeneous intervals w.r.t. the action of the gene on the network

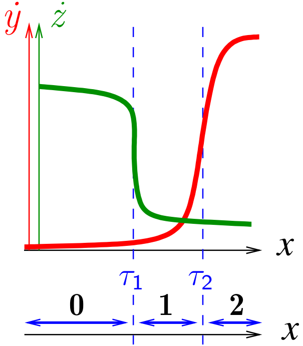




Multivalued Regulatory Graphs

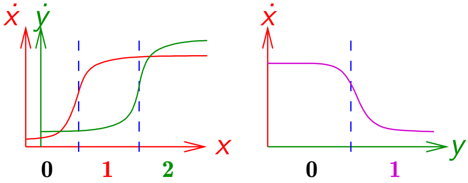
Homogeneous intervals w.r.t. the action of the gene on the network



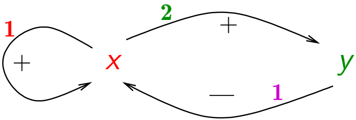


Only the relative order of the τ_i matters!

Thomas (& Snoussi) regulatory networks



In each state,
a variable v tries to
go toward the interval
numbered $K_{v,\omega}$:
the one containing $\frac{\Sigma \mathbb{1} k_i}{\gamma}$



No help :	K_x	K_y
x helps :	$K_{x,x}$	$K_{y,x}$
Absent y helps :	$K_{x,\overline{y}}$	
Both :	$K_{x,x\overline{y}}$	

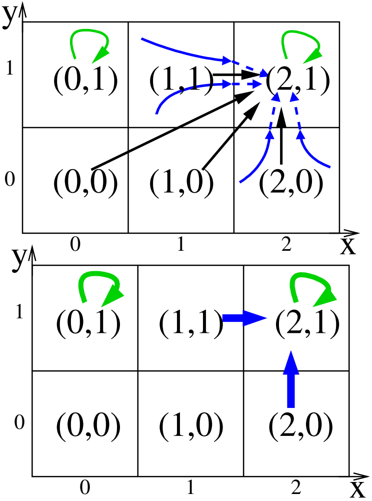
(x,y)	*Focal Point*
(0,0)	$(K_{x,\overline{y}}, K_y)$
(0,1)	(K_x, K_y)
(1,0)	$(K_{x,x\overline{y}}, K_y)$
(1,1)	$(K_{x,x}, K_y)$
(2,0)	$(K_{x,x\overline{y}}, K_{y,x})$
(2,1)	$(K_{x,x}, K_{y,x})$

Presence of an activator = Absence of an inhibitor = **A resource**

State Graphs

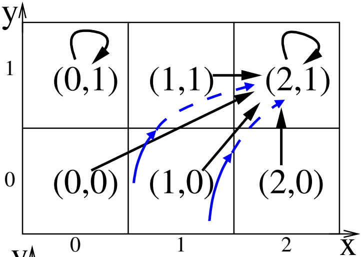
(x,y)	*Focal Point*
(0,0)	$(K_{x,\bar{y}}, K_y)$=(2,1)
(0,1)	(K_x, K_y)=(0,1)
(1,0)	$(K_{x,x\bar{y}}, K_y)$=(2,1)
(1,1)	$(K_{x,x}, K_y)$=(2,1)
(2,0)	$(K_{x,x\bar{y}}, K_{y,x})$=(2,1)
(2,1)	$(K_{x,x}, K_{y,x})$=(2,1)

Note: arbitrary values of the *K*...

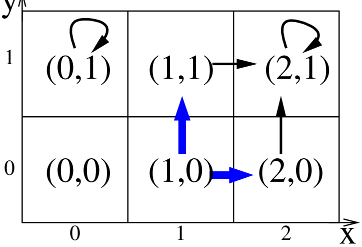


State Graphs

(x,y)	*Focal Point*
(0,0)	$(K_{x,\bar{y}}, K_y)$=(2,1)
(0,1)	(K_x, K_y)=(0,1)
(1,0)	$(K_{x,x\bar{y}}, K_y)$=(2,1)
(1,1)	$(K_{x,x}, K_y)$=(2,1)
(2,0)	$(K_{x,x\bar{y}}, K_{y,x})$=(2,1)
(2,1)	$(K_{x,x}, K_{y,x})$=(2,1)



"desynchronization" ⟶

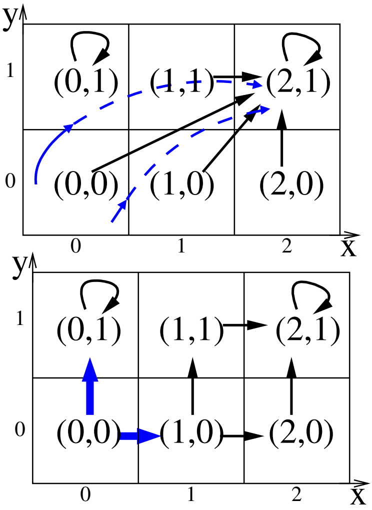


State Graphs

(x,y)	*Focal Point*
$(0,0)$	$(K_{x,\overline{y}}, K_y)$=(2,1)
$(0,1)$	(K_x, K_y)=(0,1)
$(1,0)$	$(K_{x,x\overline{y}}, K_y)$=(2,1)
$(1,1)$	$(K_{x,x}, K_y)$=(2,1)
$(2,0)$	$(K_{x,x\overline{y}}, K_{y,x})$=(2,1)
$(2,1)$	$(K_{x,x}, K_{y,x})$=(2,1)

"desynchronization" ⟶

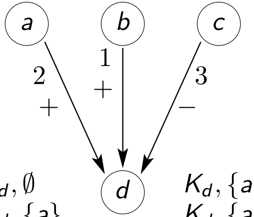
by **units** of Manhattan distance



Thomas parameters: exponential number

2^i parameters
where i is the in-degree of the gene

$\prod_{genes} (o+1)^{2^i}$ possible parameter values
where o is the out degree of each gene



$K_d, \emptyset$
$K_d, \{a\}$
$K_d, \{b\}$
$K_d, \{c\}$

$K_d, \{a, b\}$
$K_d, \{a, c\}$
$K_d, \{b, c\}$
$K_d, \{a, b, c\}$

Yeast≈7000 genes Human≈25000 genes Rice≈40000 genes

The main problem

Exhaustively identify the sets of (integer) parameters that cope with known behaviours from biological experiments

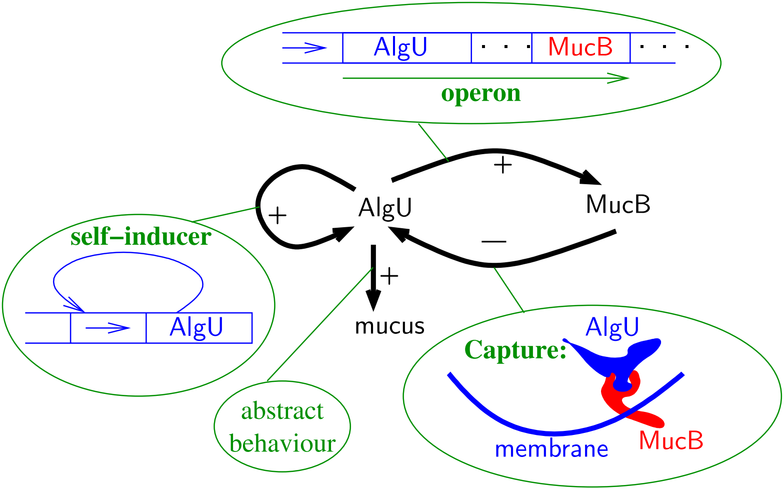
Solution = perform reverse engineering *via* formal logic

- 2003: enumeration + CTL + model checking *(Bernot,Comet,Pérès,Richard)*
- 2005: path derivatives + model checking *(Batt,De Jong)*
- 2005: PROLOG with constraints *(Trilling,Corblin,Fanchon)*
- 2007: symbolic execution + LTL *(Mateus,Le Gall,Comet)*
- 2011: traces + enumeration + CTL + model checking *(Siebert,Bockmayr)*
- 2014: Process Hitting *(Paulevé,Roux,Magnin,Folschette)*
- 2014 (tool): CoLoMoTo *(collectif)*
- 2015: genetically modified Hoare logic + constraint solving *(Bernot,Comet,Roux,Khalis,Richard)*
- 2020 (tool): TotemBioNet *(Collavizza)*

Menu

- DNA, RNA, proteins and chemical kinetics of regulatory genes
- Discrete models for regulatory networks
- **Hand made identification of parameters**
- Regulatory networks and temporal logic
- The TotemBioNet approach
- Extracting interesting experiments from models
- Complex vs. complicated. . .

Mucus production in *P. aeruginosa*

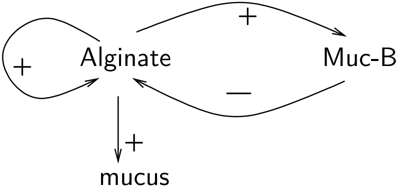


Static Graph *v.s.* Dynamic Behaviour

Difficulty to predict the result of combined regulations

Difficulty to measure the strength of a given regulation

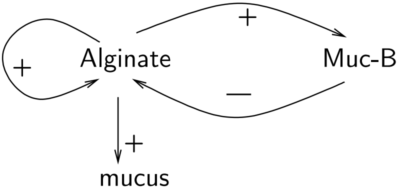
Example of “competitor” circuits

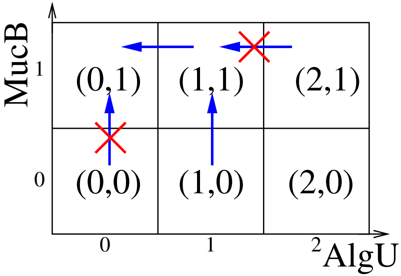


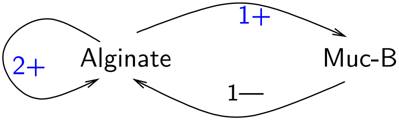
Multistationarity ?
Homeostasy ?

Many underlying qualitative models: ≈ 700 qualitative behaviours

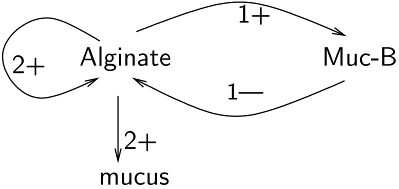
Strongly connected + Tresholds

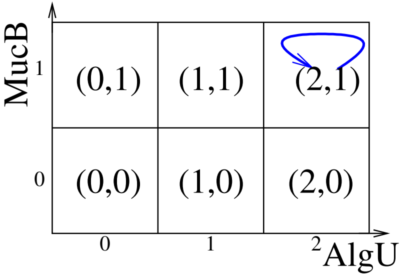






Stable states

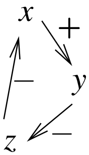


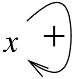


$$K_{AlgU,AlgU} = 2 \quad \text{and} \quad K_{MucB,AlgU} = 1$$

Multistationarity vs. positive cycles

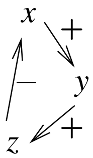
- A cycle in the interaction graph is *positive* if it contains an *even* number of inhibitions
- ***Theorem:*** *if the state graph exhibits several attraction basins then there is at least one positive cycle in the interaction graph*
- Was a conjecture from the 70's to 2004; proved by Adrien Richard and Jean-Paul Comet
(and by Christophe Soulé for the continuous case)

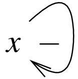




Oscillations vs. negative cycles

- A cycle in the interaction graph is *negative* if it contains a *odd* number of inhibitions
- ***Theorem:*** *if the state graph exhibits an homeostasy (stable oscillations) then there is at least one negative cycle in the interaction graph*
- Was a conjecture from the 70's to ≈2010.
  True within the global graph
  (but Counter-examples have been found for *local* graphs: A. Richard, J.-P. Comet, P. Ruet)

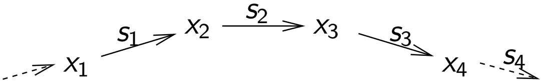




These theorems are very useful in practice when modelling biological examples

Caracteristic state of a cycle

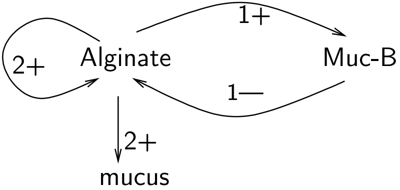
Helps characterizing the saddle point (resp. center of the oscillations) of the behaviour “driven” by a positive (resp. negative) cycle.



s_i means treshold $s_i - 1 \mid s_i$

Whatever the sign of $x_i \to x_{i+1}$, for some set of resources ω one should have $K_{x_{i+1},\omega} < s_{i+1} \leqslant K_{x_{i+1},\omega x_i}$, all along the cycle

Example:



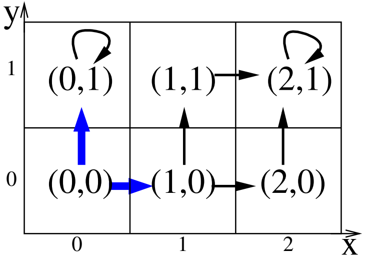
Knowledge: *Oscollations of Alginate and MucB have been observed*
Consequence: $K_{MucB} < 1$ and $K_{MucB,Alginate} \geqslant 1$ and $K_{Alginate} < 1$ and $K_{Alginate,Alginate\ MucB} \geqslant 1$

Knowledge: *Producing or not producing mucus are stable phenotypes*
Consequence: $K_{Alginate} < 2$ and $K_{Alginate,Alginate\ MucB} \geqslant 2$

Menu

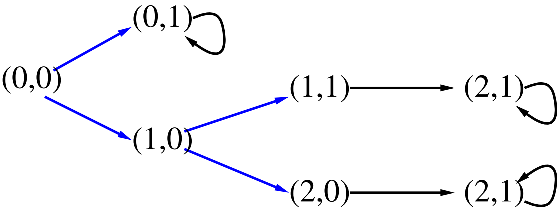
- DNA, RNA, proteins and chemical kinetics of regulatory genes
- Discrete models for regulatory networks
- Hand made identification of parameters
- **Regulatory networks and temporal logic**
- The TotemBioNet approach
- Extracting interesting experiments from models
- Complex vs. complicated. . .

Time has a tree structure. . .



As many possible state graphs as possible parameter sets. . . (huge number)

. . . from each initial state:



CTL = Computation Tree Logic

Atoms = comparaisons : (x=2) (y>0) . . .

Logical connectives: $(\varphi_1 \ \wedge \ \varphi_2) \quad (\varphi_1 \implies \varphi_2) \quad \cdots$

Temporal modalities: made of 2 characters

first character	second character
A = for **A**ll path choices	X = ne**X**t state
	F = for some **F**uture state
E = there **E**xist a choice	G = for all future states (**G**lobally)
	U = **U**ntil

$AX(y = 1)$: the concentration level of y belongs to the interval 1 in all states directly following the considered initial state.

$EG(x = 0)$: there exists at least one path from the considered initial state where x always belongs to its lower interval.

Temporal Connectives of CTL

neXt state:

$EX\varphi$: φ can be satisfied in a next state
$AX\varphi$: φ is always satisfied in the next states

eventually in the Future:

$EF\varphi$: φ can be satisfied in the future
$AF\varphi$: φ will be satisfied at some state in the future

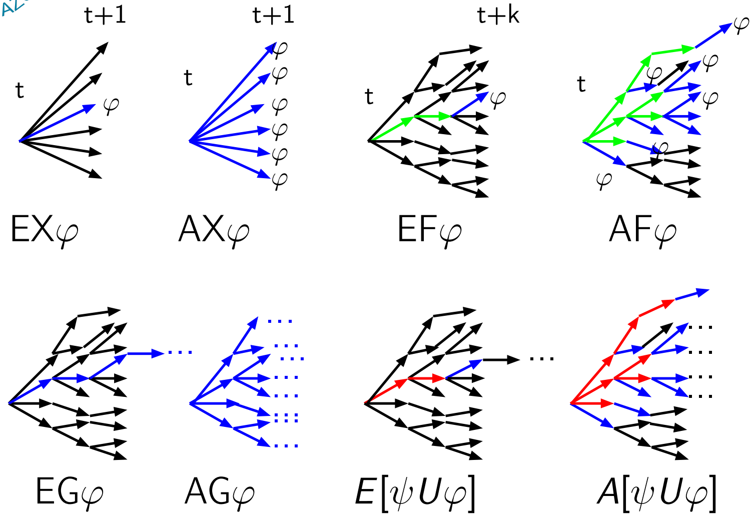
Globally:

$EG\varphi$: φ can be an invariant in the future
$AG\varphi$: φ is necessarilly an invariant in the future

Until:

$E[\psi U \varphi]$: there exist a path where ψ is satisfied until a state where φ is satisfied
$A[\psi U \varphi]$: ψ is always satisfied until some state where φ is satisfied

Semantics of Temporal Connectives

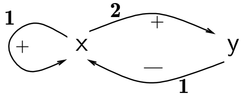


EXφ AXφ EFφ AFφ

EGφ AGφ $E[\psi U \varphi]$ $A[\psi U \varphi]$

$(after \longrightarrow : \varphi \ , \ after \longrightarrow : \psi \ )$

CTL to encode Biological Properties



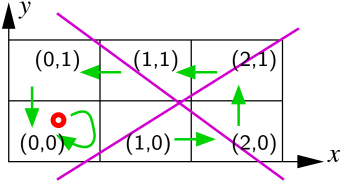
Common properties:

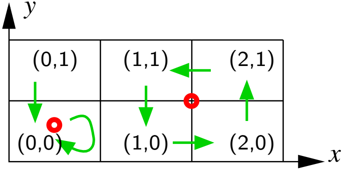
"functionality" of a sub-graph

Special role of "feedback loops"

– positive: *multistationnarity* (even number of —)

– negative: *homeostasy* (odd number of —)





Characteristic properties: $\begin{cases} (x = 2) \Longrightarrow AG(\neg(x = 0)) \\ (x = 0) \Longrightarrow AG(\neg(x = 2)) \end{cases}$

They express *"the positive feedback loop is functional"*

(satisfaction of these formulas relies on the parameters $K_{\dots}$)

Model Checking

- Efficiently computes all the states of a state graph which satisfy a given formula: $\{\ \eta \ \mid \ M \ \models_\eta \ \varphi\ \}$.
- Efficiently select the models which globally satisfy a given formula.

Intensively used:

- to find the set of **all** possible discrete parameter values
- to check models under construction w.r.t. **known behaviours** (one often gets an empty set of parameter values!)
- and to prove the **consistency** of a biological **hypothesis**

Menu

- DNA, RNA, proteins and chemical kinetics of regulatory genes
- Discrete models for regulatory networks
- Hand made identification of parameters
- Regulatory networks and temporal logic
- **The TotemBioNet approach**
- Extracting interesting experiments from models
- Complex vs. complicated. . .

TotemBioNet

Takes as input:

- an interaction graph
- some constraints on the parameters, if available
- a set of temporal formulas (CTL or similar)
- some experimentally observed paths, if available

Provides as output:

- The *exhaustive* set of correct parameter settings that satisfy the input information

using sophisticated enumeration strategies in order to reduce the number of proofs by model checking.

TotemBioNet methodology

Most of the time, the set of correct parameter settings is either *empty* or *huge*

If *empty:* good news! research goes on

- reconsider biological “knowledge”
- reconsider its temporal logic encoding

If *huge:*

- randomly take one or two correct parameter settings
- randomly extract a few paths in the state graph
- most of the time, the biologist has an “obvious” reason to reject some paths
- encode the reason in temporal logic and start again. . .

. . . until the number of parameter settings becomes low and no more “obviously bad paths” are found.

Menu

- DNA, RNA, proteins and chemical kinetics of regulatory genes
- Discrete models for regulatory networks
- Hand made identification of parameters
- Regulatory networks and temporal logic
- The TotemBioNet approach
- **Extracting interesting experiments from models**
- Complex vs. complicated. . .

Simplifications driven by the hypothesis

Biologists spend money and time for experiments because they have a **hypothesis** φ in mind that they want to test. . .

. . . Successive simplified views of the studied biological object and of the hypothesis:

$$\begin{array}{c} \text{Model} \\ M_1 \\ \text{satisfies} \\ \varphi_1 \end{array} \iff \begin{array}{c} \text{Model} \\ M_2 \\ \text{satisfies} \\ \varphi_2 \end{array} \iff \begin{array}{c} \text{Model} \\ M_3 \\ \text{satisfies} \\ \varphi_3 \end{array} \iff \dots$$

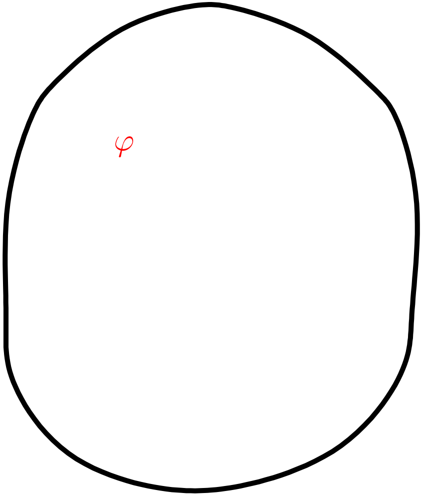
Node removing / Expression level folding / Node fusion / *etc.*

"Kleenex" models: hypothesis dependant models

Generation of biological experiments (1)

Set of all the formulas:

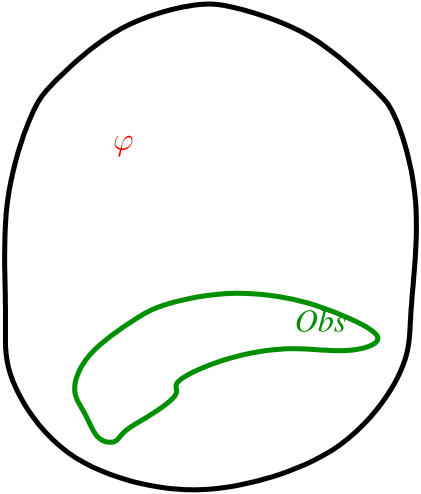
φ = hypothesis



Generation of biological experiments (2)

Set of all the formulas:

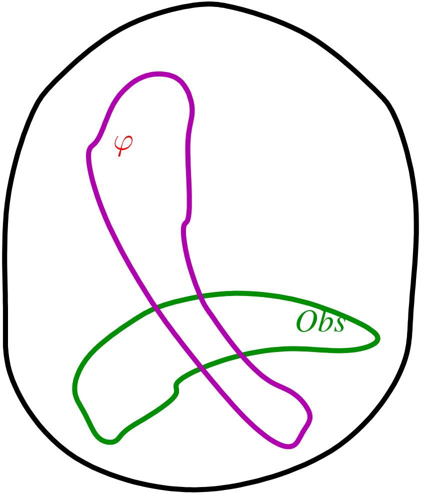
φ = hypothesis
Obs = possible experiments



Generation of biological experiments (3)

Set of all the formulas:

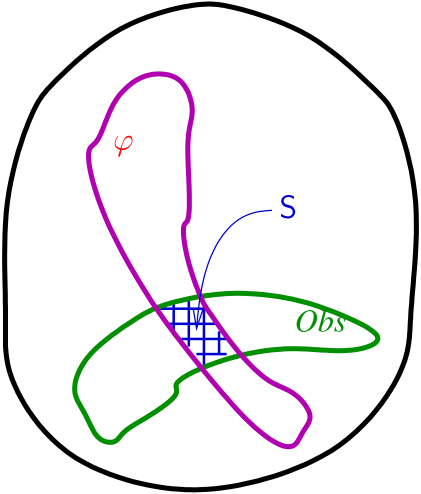
φ = hypothesis
Obs = possible experiments
$Th(\varphi) = \varphi$ inferences



Generation of biological experiments (4)

Set of all the formulas:

φ = hypothesis
Obs = possible experiments
$Th(\varphi)$ = φ inferences
S = sensible experiments



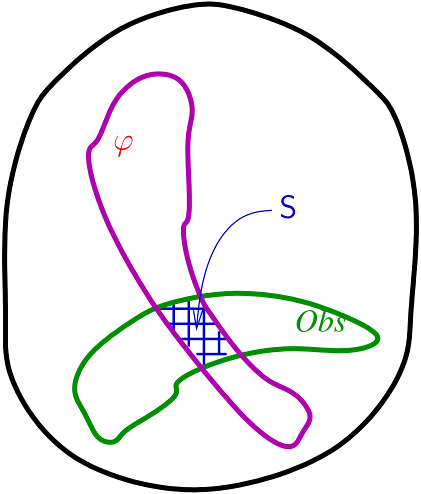
Generation of biological experiments (5)

Set of all the formulas:

φ = hypothesis
Obs = possible experiments
$Th(\varphi) = \varphi$ inferences
S = sensible experiments

Refutability:

$$S \Longrightarrow \varphi\ ?$$



Generation of biological experiments

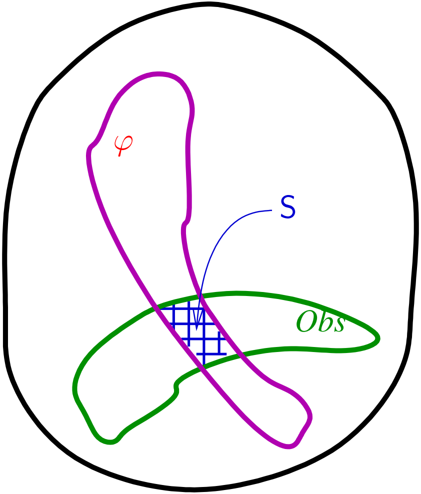
Set of all the formulas:

φ = hypothesis
Obs = possible experiments
$Th(\varphi)$ = φ inferences
S = sensible experiments

Refutability:

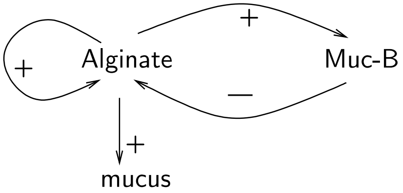
$$S \Longrightarrow \varphi \;?$$

Best refutations:
Choice of experiments in S ?
. . . optimisations



How to validate a multistationnarity

$\mathcal{M}$: *(unknown thresholds)*



$$\Phi: \begin{cases} (Alginate = 2) \Longrightarrow AG(Alginate = 2) & (hypothesis) \\ (Alginate = 0) \Longrightarrow AG(Alginate < 2) & (knowledge) \end{cases}$$

Assume that only *mucus* can be observed:

Lemma: $AG(Alginate = 2) \Longleftrightarrow AFAG(mucus = 1)$

(... formal proof by computer ...)

$\rightarrow$ $\boxed{\text{To validate: } (Alginate = 2) \Longrightarrow AFAG(mucus = 1)}$

$(Alginate = 2) \Longrightarrow AFAG(mucus = 1)$

$A \Longrightarrow B$	*true*	*false*
true	true	false
false	true	true

Karl Popper:

to validate = to try to refute

thus A=false is useless

experiments must begin with a pulse

The pulse forces the bacteria to reach the initial state $Alginate = 2$.
If the state is not directly controlable we need to prove lemmas:

$$(something\ reachable) \Longrightarrow (Alginate = 2)$$

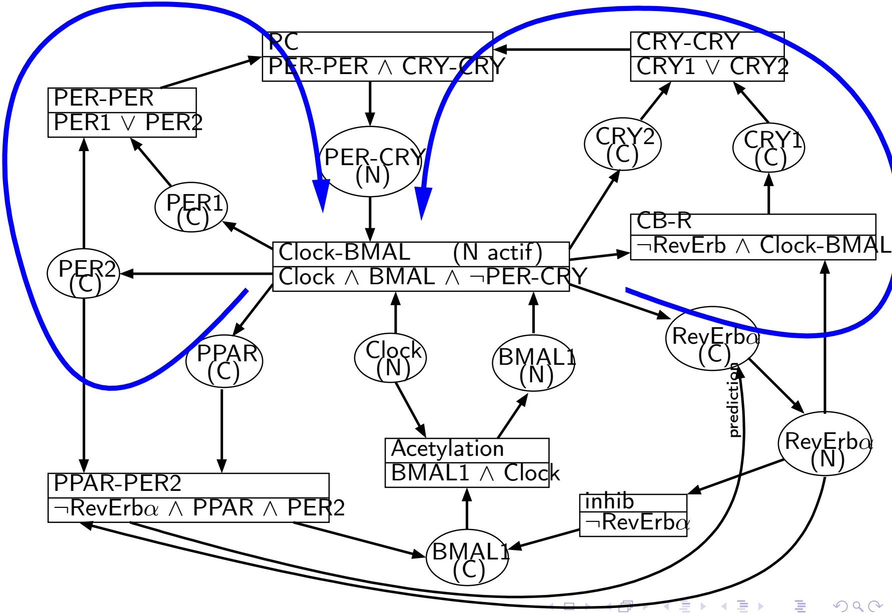
General form of a test:

$$(something\ \underline{reachable}) \Longrightarrow (something\ \underline{observable})$$

Menu

- DNA, RNA, proteins and chemical kinetics of regulatory genes
- Discrete models for regulatory networks
- Hand made identification of parameters
- Regulatory networks and temporal logic
- The TotemBioNet approach
- Extracting interesting experiments from models
- **Complex vs. complicated. . .**

Circadian clock interaction graph



The target question

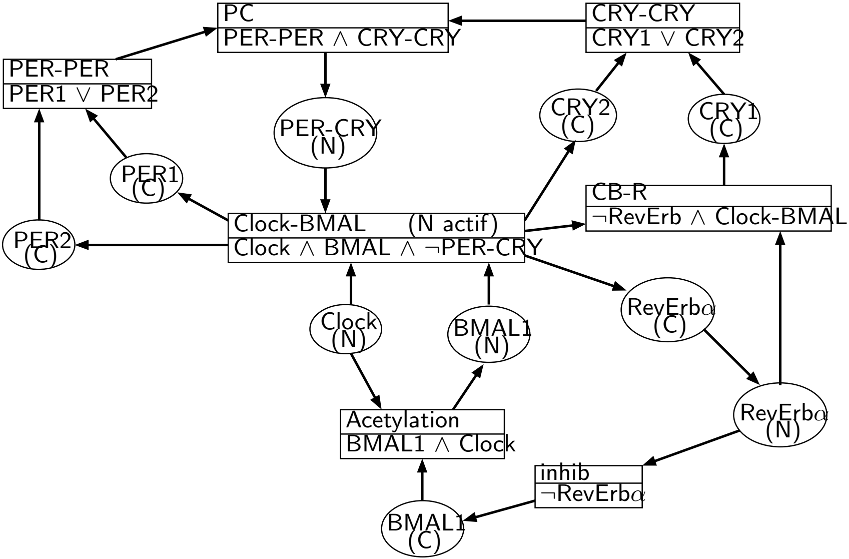
Impact of the day length on the persistence of the circadian circle ?

$\Longrightarrow$ framework with time delays:

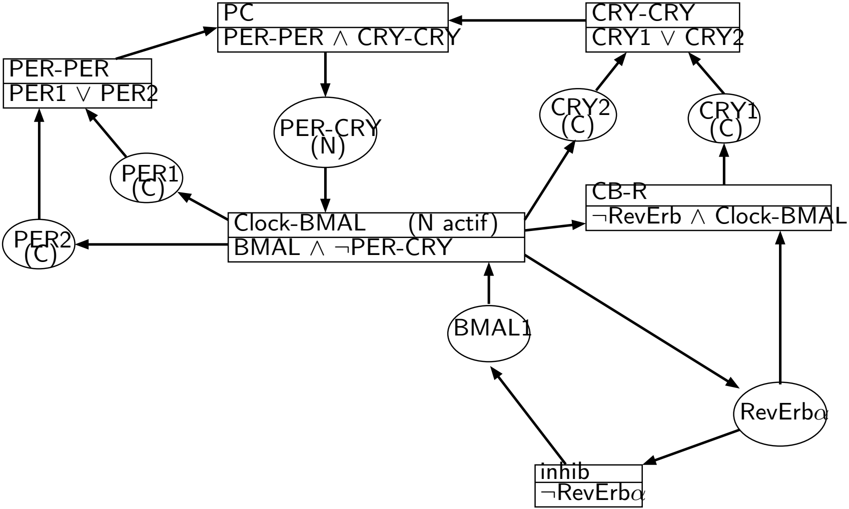
- mainly replace the integer $K_{x,\omega}$ by real numbers $C_{x,\omega,n}$, called *celerities*, where n is the current state of x
- notice that $C_{x,\omega,n} > 0$ if $K_{x,\omega} > n$ and a few other logical properties
- extension of temporal logic with delays: $AF_{[t_1,t_2]}$ and so on

Decidability is lost but the identification of parameters remains "almost" automatic with such constant speeds $C_{x,\omega,n}$ (constraint solving on intervals)

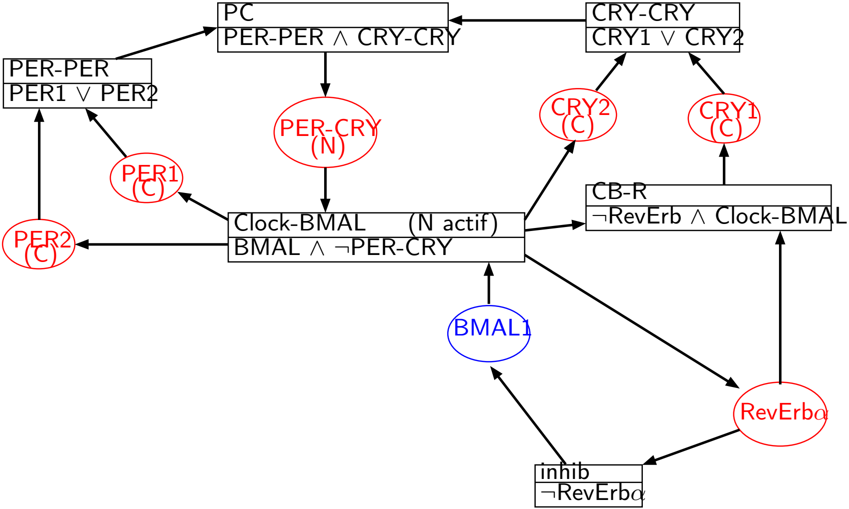
Fold levels and remove PPAR



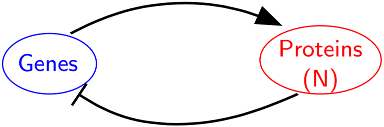
Remove Clock and “tunnel” pathways



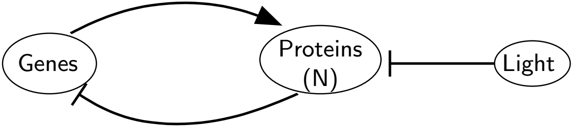
Separate inhibitors/activators of Clock-BMAL



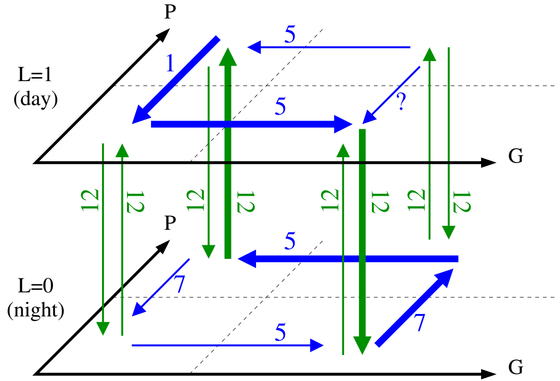
Fusion of all inhibitors



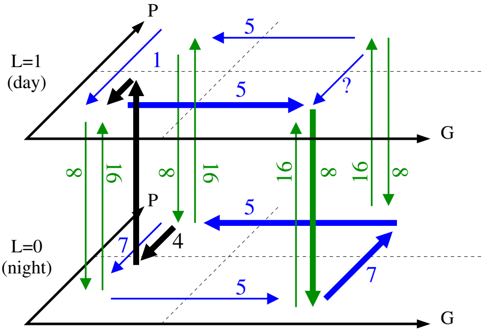
and Light prevents PER-CRY to enter the nucleus:



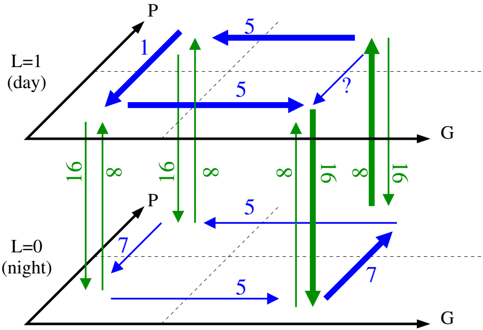
12 hours model



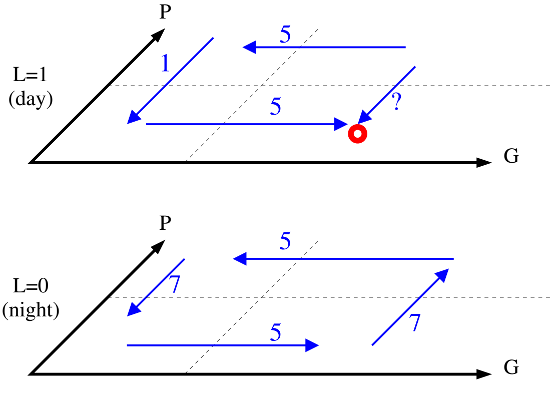
Winter model



Summer model



Jet lag + training



Take Home Messages

Make explicit the hypotheses that motivate the biologist

A far as possible formalize them to get a computer aided approach

Behavioural *properties* are as much important as *models*

Mathematical models are not reality: let's use this freedom !
(several views of a same biological object)

Modelling is significant only with respect to the considered experimental *reachability* and *observability* (for refutability)

Formal proofs can suggest wet experiments

"Kleenex" models help understanding main behaviours

Specialized qualitative approaches can make complex models simple