

Formal Modeling and Simulation of Biological Systems

G. Caravagna
Department of Computer Science
University of Pisa

Imperial College
London, June 2009

" Perché la vita é troppo importante e complessa
per lasciarla studiare solo ai biologi. " (V.Manca)

Index

- 1 A gentle introduction to Systems Biology
- 2 The Calculus Of Looping Sequences
 - Stochastic CLS
 - Spatial CLS
 - CLS with Links
- 3 Formal Modeling Biological Systems With Delays

Index

1 A gentle introduction to Systems Biology

2 The Calculus Of Looping Sequences

- Stochastic CLS
- Spatial CLS
- CLS with Links

3 Formal Modeling Biological Systems With Delays

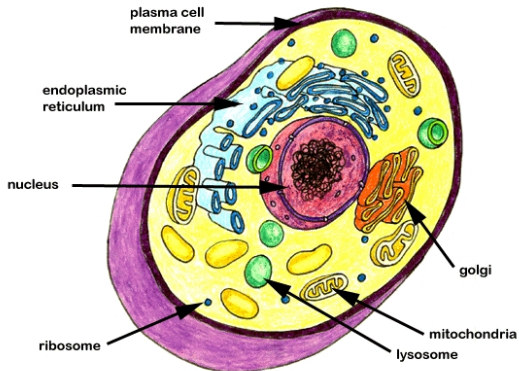
Some “Spontaneous” Questions

- ① What is Systems Biology ?
- ② How is it possible to simulate a real biological system ?
- ③ What can Computer Science do ?

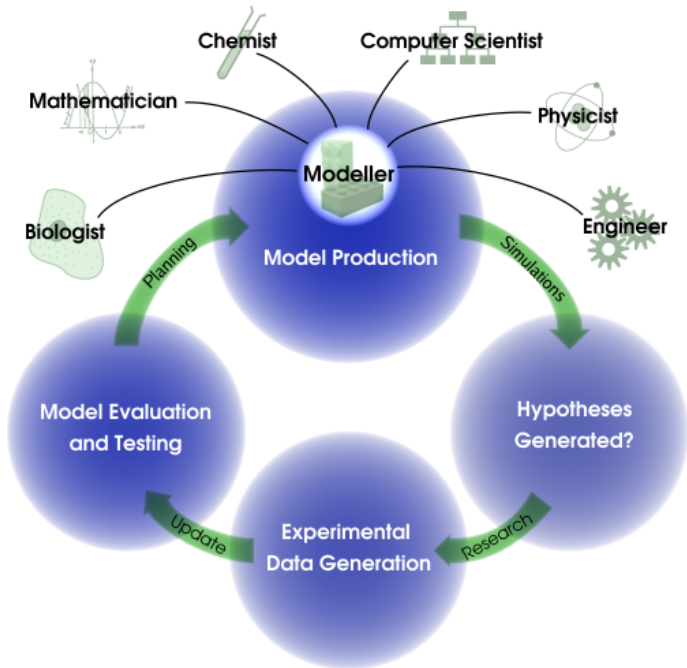
Question #1

What is Systems Biology ?

Cells: complex systems of interactive components



- Two classifications of cell:
 - ▶ prokaryotic
 - ▶ eukaryotic
- Main actors:
 - ▶ membranes
 - ▶ proteins
 - ▶ DNA/RNA
 - ▶ ions, macromolecules, . . .
- Interaction networks:
 - ▶ metabolic pathways
 - ▶ signaling pathways
 - ▶ gene regulatory networks



Question #2

How is it possible to simulate a real biological system ?

Methods from Mathematics

Deterministic Models as *differential equations*:

- describe the variations of *concentrations*;
- the terms of the equations *model* the observed events.

Advantages:

- well-founded theory (since Newton, Leibiniz);
- many analysis techniques;

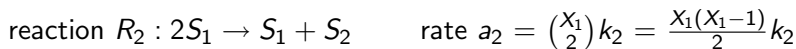
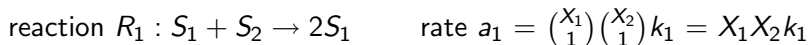
Disadvantages:

- a continuous approach (concentrations) which is not suitable for systems dealing with small concentrations.

Background: Gillespie's simulation algorithm

- represents a chemical solution as a multiset of molecules
- each chemical reaction is associated with a kinetic constant
- computes the reaction rate a_μ by multiplying the kinetic constant by the number of possible combinations of reactants

Example: chemical solution with X_1 molecules S_1 and X_2 molecules S_2



Given a set of reactions $\{R_1, \dots, R_M\}$ and a current time t

- The time $t + \tau$ at which the next reaction will occur is randomly chosen with τ exponentially distributed with parameter $\sum_{\nu=1}^M a_\nu$;
- The reaction R_μ that has to occur at time $t + \tau$ is randomly chosen with probability $\frac{a_\mu}{\sum_{\nu=1}^M a_\nu}$.

At each step t is incremented by τ and the chemical solution is updated.

Stochastic Simulation: motivations

Advantages:

- is *exact* (number of molecules)
- under *reasonable* conditions
 - ▶ “small” number of molecules;
 - ▶ oscillating behaviours;

stochastic models exhibit behaviors observable in the real biological systems but not in the deterministic counterpart (a more precise matching with reality).

Disadvantages:

- 1 as it is exact the simulation is slower than the deterministic one;
- 2 serious scalability problem (w.r.t. size of a model and speed of reactions);

Question #3

What can Computer Science do?

Historical Background 1/2

- Formalisms adopted from CS:
 - ▶ L-Systems: parallel rewriting system to model the growth processes of plant development
 - ▶ Petri Nets.
 - ▶ Stochastic π -calculus (Priami) to model biochemical reactions.

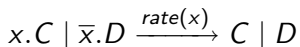
Historical Background 1/2

- Formalisms adopted from CS:
 - ▶ L-Systems: parallel rewriting system to model the growth processes of plant development
 - ▶ Petri Nets.
 - ▶ Stochastic π -calculus (Priami) to model biochemical reactions.

Main idea: given a reaction $A + B \xrightarrow{k} C + D$

- build processes $A \equiv x.C$ and $B \equiv \bar{x}.D$;
- associate each channel a rate ($rate(x) = k$);

the communication on channel x models the firing of the reaction.



Historical Background 2/2

- Formalisms to explicitly model biological structures:

- ▶ DNA/RNA strands;
- ▶ cellular membranes;
- ▶ whole cells;

and events

- ▶ complexation / de-complexation;
- ▶ endocytosis / exocytosis / ...;

such as:

- ▶ Bio Ambients, Brane Calculi (Cardelli et al.);
- ▶ Bio PEPA (Gilmore et al.);
- ▶ k calculus (Laneve et al.);
- ▶ Calculus Of Looping Sequences (Milazzo et al.);

Index

- 1 A gentle introduction to Systems Biology
- 2 The Calculus Of Looping Sequences**
 - Stochastic CLS
 - Spatial CLS
 - CLS with Links
- 3 Formal Modeling Biological Systems With Delays

Outline of the talk

- 1 A gentle introduction to Systems Biology
- 2 The Calculus Of Looping Sequences
 - Stochastic CLS
 - Spatial CLS
 - CLS with Links
- 3 Formal Modeling Biological Systems With Delays

The Calculus of Looping Sequences (CLS)

We assume an alphabet \mathcal{E} . **Terms** T and **Sequences** S of CLS are given by the following grammar:

$$\begin{aligned} T & ::= S \mid (S)^L \mid T \mid T \\ S & ::= \epsilon \mid a \mid S \cdot S \end{aligned}$$

where a is a generic element of \mathcal{E} , and ϵ is the empty sequence.

The operators are:

$S \cdot S$: Sequencing

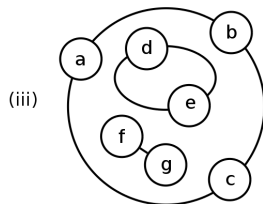
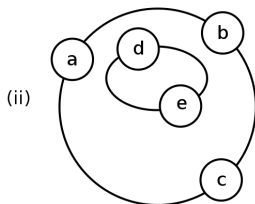
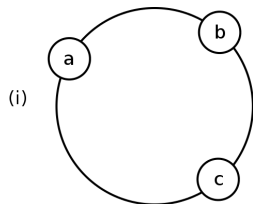
$(S)^L$: Looping (S is closed and it can rotate)

$T_1 \mid T_2$: Containment (T_1 contains T_2)

$T \mid T$: Parallel composition (juxtaposition)

Actually, looping and containment form a single binary operator $(S)^L \mid T$.

Examples of Terms



$$(i) \quad (a \cdot b \cdot c)^L \mid \epsilon$$

$$(ii) \quad (a \cdot b \cdot c)^L \mid (d \cdot e)^L \mid \epsilon$$

$$(iii) \quad (a \cdot b \cdot c)^L \mid (f \cdot g \mid (d \cdot e)^L \mid \epsilon)$$

Structural Congruence

The **Structural Congruence** relations \equiv_S and \equiv_T are the least congruence relations on sequences and on terms, respectively, satisfying the following rules:

$$S_1 \cdot (S_2 \cdot S_3) \equiv_S (S_1 \cdot S_2) \cdot S_3 \quad S \cdot \epsilon \equiv_S \epsilon \cdot S \equiv_S S$$

$$T_1 \mid T_2 \equiv_T T_2 \mid T_1 \quad T_1 \mid (T_2 \mid T_3) \equiv_T (T_1 \mid T_2) \mid T_3$$

$$T \mid \epsilon \equiv_T T \quad (S_1 \cdot S_2)^L \rfloor T \equiv_T (S_2 \cdot S_1)^L \rfloor T$$

We write \equiv for \equiv_T .

CLS Patterns

Let us consider variables of three kinds:

- term variables (X, Y, Z, \dots)
- sequence variables ($\tilde{x}, \tilde{y}, \tilde{z}, \dots$)
- element variables (x, y, z, \dots)

Patterns P and **Sequence Patterns** SP of CLS extend CLS terms and sequences with variables:

$$\begin{array}{l} P ::= SP \mid (SP)^L \mid P \mid P \mid X \\ SP ::= \epsilon \mid a \mid SP \cdot SP \mid x \mid \tilde{x} \end{array}$$

where a is a generic element of \mathcal{E} , ϵ is the empty sequence, and x, \tilde{x} and X are generic element, sequence and term variables

The structural congruence relation \equiv extends trivially to patterns

Rewrite Rules

A **Rewrite Rule** is a pair (P, P') , denoted $P \mapsto P'$, where:

- P, P' are patterns
- variables in P' are a subset of those in P

A rule $P \mapsto P'$ can be applied to all terms that are instantiations of P .

Example: $a \cdot x \cdot a \mapsto b \cdot x \cdot b$

- can be applied to $a \cdot c \cdot a$ (producing $b \cdot c \cdot b$)
- cannot be applied to $a \cdot c \cdot c \cdot a$

Example: $(a \cdot \tilde{x})^L \rfloor (b \mid X) \mapsto (c \cdot \tilde{x})^L \rfloor X$

- can be applied to $(a \cdot a \cdot a)^L \rfloor (b \mid b \mid (a)^L \rfloor b)$
- the result is either $(c \cdot a \cdot a)^L \rfloor (b \mid (a)^L \rfloor b)$ or $(a \cdot a \cdot a)^L \rfloor (b \mid b \mid (c)^L \rfloor \epsilon)$

Formal Semantics

$P\sigma$ denotes the term obtained by replacing any variable in T with the corresponding term, sequence or element.

Σ is the set of all possible instantiations σ

Given a set of rewrite rules \mathcal{R} , evolution of terms is described by the transition system given by the least relation \rightarrow satisfying

$$\frac{P \mapsto P' \in \mathcal{R} \quad P\sigma \neq \epsilon \quad \sigma \in \Sigma}{P\sigma \rightarrow P'\sigma}$$
$$\frac{T \rightarrow T'}{T \mid T'' \rightarrow T' \mid T''} \quad \frac{T \rightarrow T'}{(S)^L \rfloor T \rightarrow (S)^L \rfloor T'}$$

and closed under structural congruence \equiv .

Some Theoretical Results

CLS is Turing complete

- A Turing machine encoded into a CLS term and a single rewrite rule

Formalisms capable of describing membranes can be encoded into CLS

- Brane Calculi
- P Systems

We defined the classical equivalence relations (**Theorem:** Strong and weak bisimilarities are congruences).

Remark: These kind of results (i.e. notions of similarity) can be defined also for biologically-inspired formalisms, is this enough?

i.e. is bisimulation always satisfactory in a biological context?

Finally, this is not stochastic, we need to define the Stochastic CLS.

Index

- 1 A gentle introduction to Systems Biology
- 2 The Calculus Of Looping Sequences
 - Stochastic CLS
 - Spatial CLS
 - CLS with Links
- 3 Formal Modeling Biological Systems With Delays

Stochastic CLS incorporates Gillespie's stochastic framework into the semantics of CLS

- Rewrite rules are enriched with kinetic constants

What is a reactant in Stochastic CLS?

- A *reactant combination* is an occurrence (up to \equiv) of a left hand side of a rewrite rule

Example: The application rate of $a \mid b \xrightarrow{k} c$ to $a \mid a \mid b \mid b$ is $6k$

Example: The application rate of $(a \cdot \tilde{x})^L \mid (b \mid X) \xrightarrow{k} (c \cdot \tilde{x})^L \mid X$ to $(a \cdot a \cdot a)^L \mid (b \mid b) \mid (a \cdot a)^L \mid b$ is

- $6k$, with $(c \cdot a \cdot a)^L \mid b \mid (a \cdot a)^L \mid b$ as result
- $+ 2k$, with $(a \cdot a \cdot a)^L \mid (b \mid b) \mid (c \cdot a)^L \mid \epsilon$ as result
- $= 8k$

Stochastic CLS (2)

Given a finite set of stochastic rewrite rules \mathcal{R} , the semantics of Stochastic CLS is the least transition relation $\xrightarrow{R, T, r, b}$ closed wrt \equiv and satisfying by the following inference rules:

$$\frac{R : P_L \xrightarrow{k} P_R \in \mathcal{R} \quad \sigma \in \Sigma}{P_L \sigma \xrightarrow{R, P_L \sigma, k \cdot \text{comb}(P_L, \sigma), 1} P_R \sigma} \quad \frac{T_1 \xrightarrow{R, T, r, b} T_2}{T_1 \mid T_3 \xrightarrow{R, T, r, b \cdot \text{binom}(T, T_1, T_3)} T_2 \mid T_3}$$

$$\frac{T_1 \xrightarrow{R, T, r, b} T_2}{(T_1)^L \mid T_3 \xrightarrow{R, (T_1)^L \mid T_3, r \cdot b, 1} (T_2)^L \mid T_3} \quad \frac{T_1 \xrightarrow{R, T, r, b} T_2}{(T_3)^L \mid T_1 \xrightarrow{R, (T_3)^L \mid T_1, r \cdot b, 1} (T_3)^L \mid T_2}$$

The transition system obtained can be easily transformed into a *Continuous Time Markov Chain*

A Stochastic CLS model of the Quorum Sensing (1)

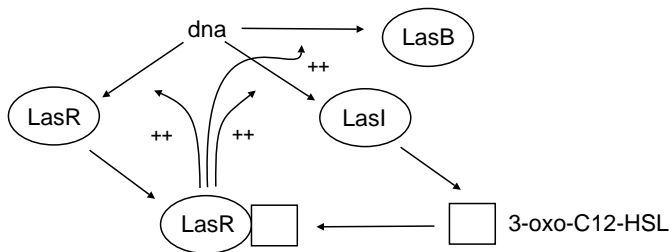
It is recognised that many bacteria have the ability of modulating their gene expressions according to their population density. This process is called *quorum sensing*.

- A diffusible small molecules (called *autoinducers*)
- One or more transcriptional activator proteins (*R-proteins*) located within the cell
- The autoinducer can cross freely the cellular membrane
- The R-protein by itself is not active without the autoinducer. The autoinducer molecule can bind the R-protein to form an *autoinducer/R-protein* complex.
- The *autoinducer/R-protein* complex binds the DNA enhancing the transcription of specific genes.
- These genes regulate both the production of specific behavioural traits and the production of the autoinducer and of the R-protein.

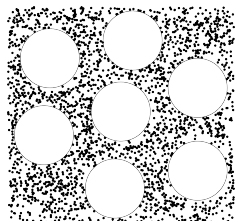
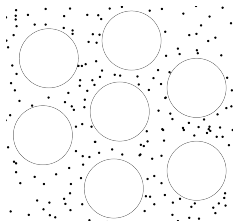
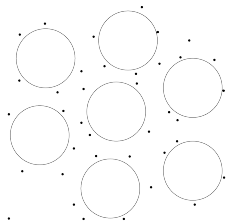
A Stochastic CLS model of the Quorum Sensing (2)

At low cell density, the autoinducer is synthesized at basal levels and diffuses in the environment where it is diluted. With high cell density the concentration of the autoinducer increases. Beyond a threshold the autoinducer is produced autocatalytically.

The autocatalytic production results in a dramatic increase of product concentration.

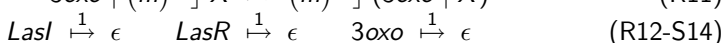
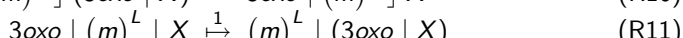
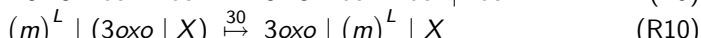
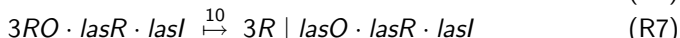
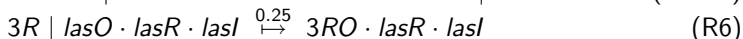
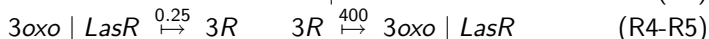
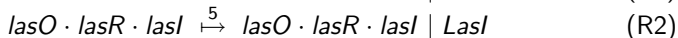


A Stochastic CLS model of the Quorum Sensing (3)

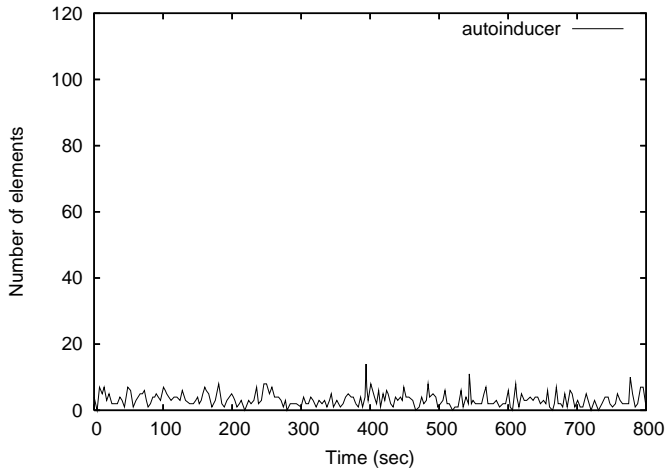


A Stochastic CLS model of the Quorum Sensing (4)

The behaviour of a single bacterium:

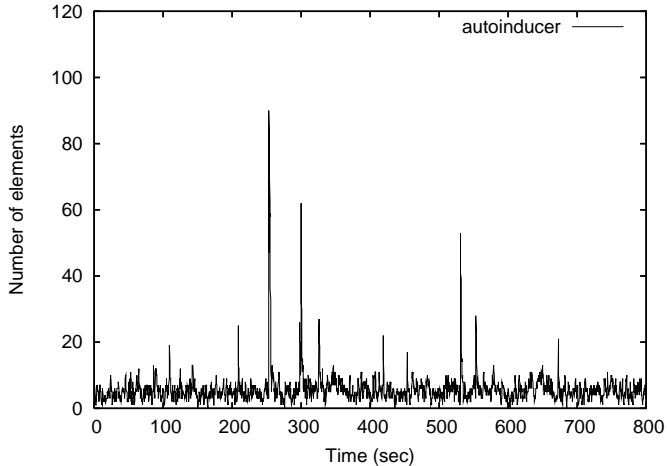


Simulation results (1)



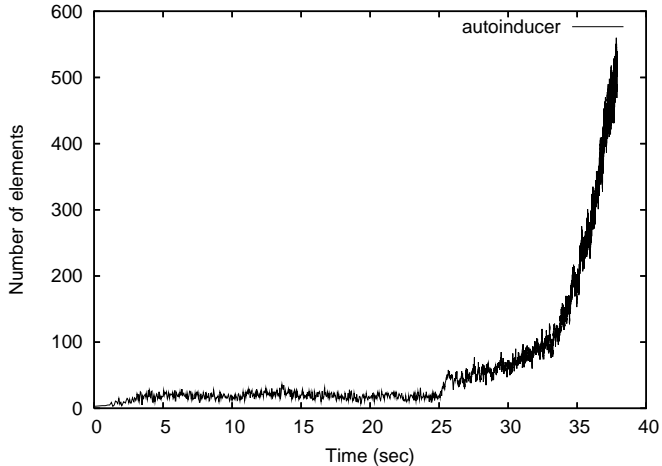
Production of the autoinducer by a single bacterium

Simulation results (2)



Production of the autoinducer by a population of five bacteria

Simulation results (3)



Production of the autoinducer by a population of twenty bacteria

Index

- 1 A gentle introduction to Systems Biology
- 2 **The Calculus Of Looping Sequences**
 - Stochastic CLS
 - **Spatial CLS**
 - CLS with Links
- 3 Formal Modeling Biological Systems With Delays

Spatial CLS

The spatial organization of elements may affect system dynamics

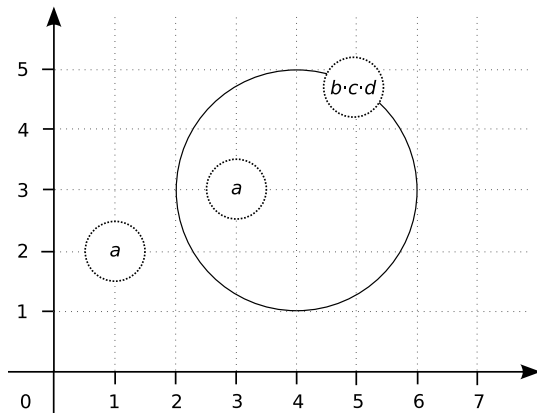
- reaction-diffusion system
- molecular crowding

We developed Spatial CLS by extending the Calculus of Looping Sequences

Elements of Spatial CLS are spheres in a continuous space

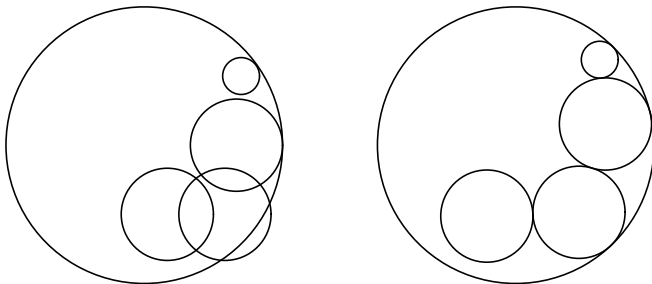
- the containment hierarchy is reflected in the spheres
- elements can move autonomously
- interactions can depend on the spatial information of elements (position, radius, ecc.)
- rewrite rules are endowed with rates

Example of Spatial CLS term



$$T = (a)_{[(1,2),m_1],0.5} \mid ((b \cdot c \cdot d)_{\cdot,0.5})_{[(4,3),m_2],2}^L \mid (a)_{[(-1,0),m_3],0.5}$$

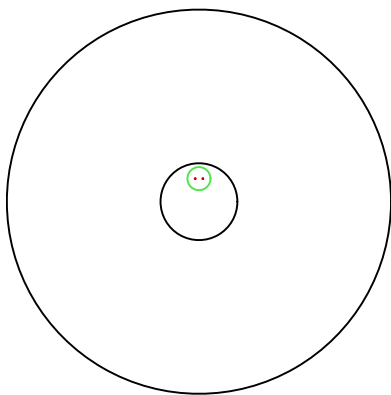
Resolving space conflicts



Elements push each other

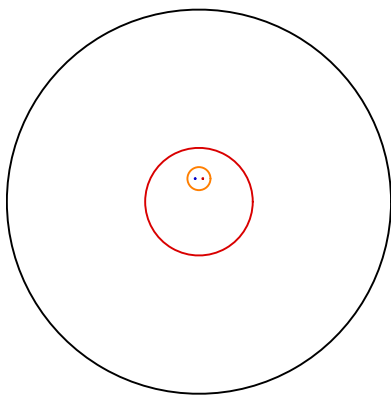
- the pushing effect is modeled with a system of differential equations
- the rearranged state corresponds to its equilibrium state

Simulation of cell proliferation



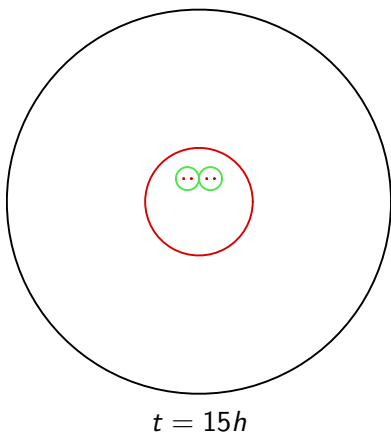
$t = 0h$

Simulation of cell proliferation

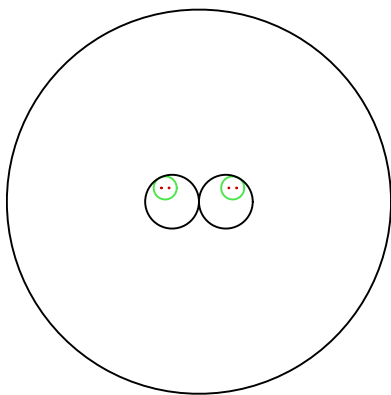


$t = 6h$

Simulation of cell proliferation

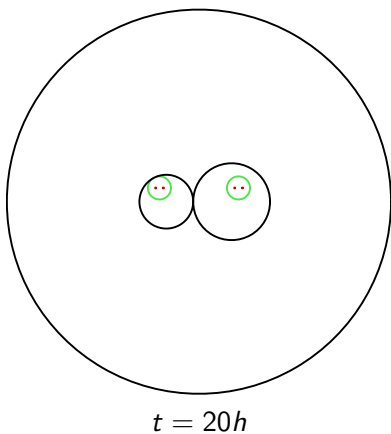


Simulation of cell proliferation

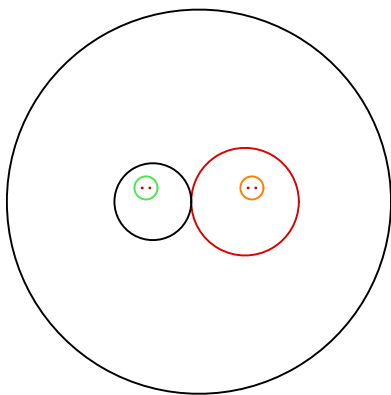


$t = 16h$

Simulation of cell proliferation

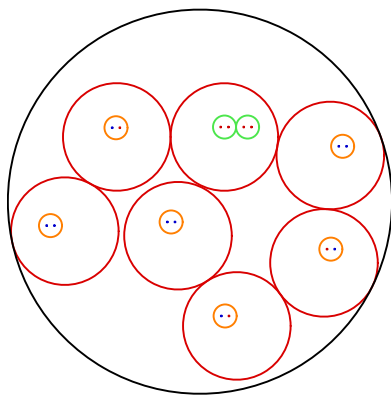


Simulation of cell proliferation



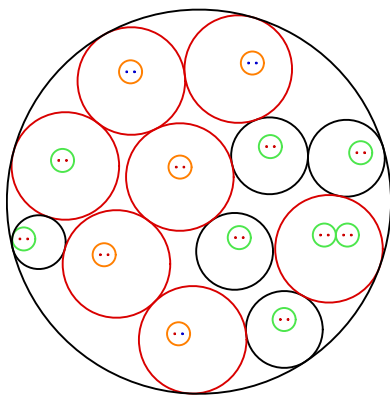
$t = 25.66h$

Simulation of cell proliferation



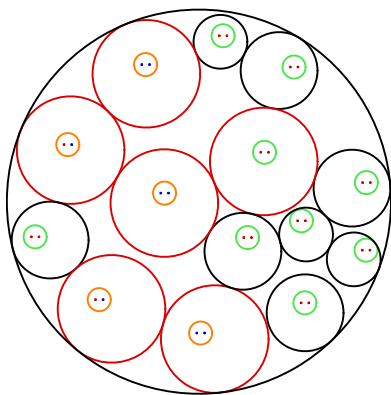
$t = 88h$

Simulation of cell proliferation



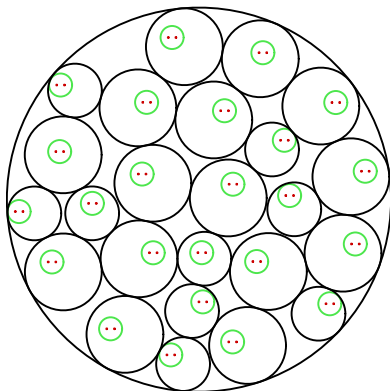
$t = 98h$

Simulation of cell proliferation



$t = 102h$

Simulation of cell proliferation



$t = 141h$

Index

- 1 A gentle introduction to Systems Biology
- 2 **The Calculus Of Looping Sequences**
 - Stochastic CLS
 - Spatial CLS
 - **CLS with Links**
- 3 Formal Modeling Biological Systems With Delays

Modeling proteins at the domain level

In proteins there are places (domains) where bindings to other molecules can occur

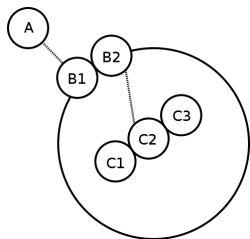
To model a protein at the domain level in CLS it would be natural to use a sequence with one symbol for each domain

The binding between two elements of two different sequences cannot be expressed in CLS

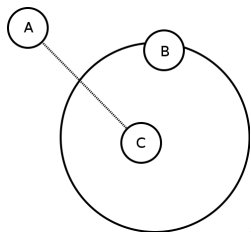
LCLS extends CLS with labels on basic symbols

- two symbols with the same label represent domains that are bound to each other
- example: $a \cdot b^1 \cdot c \mid d \cdot e^1 \cdot f$

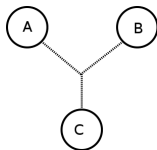
Well-formed LCLS terms and patterns



$$A^1 \mid (B1^1 \cdot B2^2)^L \mid C1 \cdot C2^2 \cdot C3 \quad \checkmark$$



$$A^1 \mid (B)^L \mid C^1 \quad \times$$



$$A^1 \mid B^1 \mid C^1 \quad \times$$

Checking well-formedness via typechecking

A term is well-formed iff a label occurs no more than twice, and two occurrences of a label are always in the same compartment.

We define $(N_1, N_2) \models T$ where

- $N_1 \subset \mathbb{N}$ set of labels used twice;
- $N_2 \subset \mathbb{N}$ set of labels used once in the top-level compartment of T ;

Hence T is well-formed iff T has a type.

The application of a rule may introduce inconsistencies, we define compartment safe rules so that we get

Theorem (Subject Reduction)

Given a set of well-formed rewrite rules \mathcal{R} and a well-formed term T

$$T \rightarrow T' \quad \Longrightarrow \quad T' \text{ well-formed}$$

Index

- 1 A gentle introduction to Systems Biology
- 2 The Calculus Of Looping Sequences
 - Stochastic CLS
 - Spatial CLS
 - CLS with Links
- 3 Formal Modeling Biological Systems With Delays

Delays as abstractions

Typically, it may be that:

- some data is missing (i.e. kinetic constants cannot be measured);
- a complex dynamic with partial knowledge of the sub-events;

The average time to complete the event can be seen as a delay (+ the stochastic time).

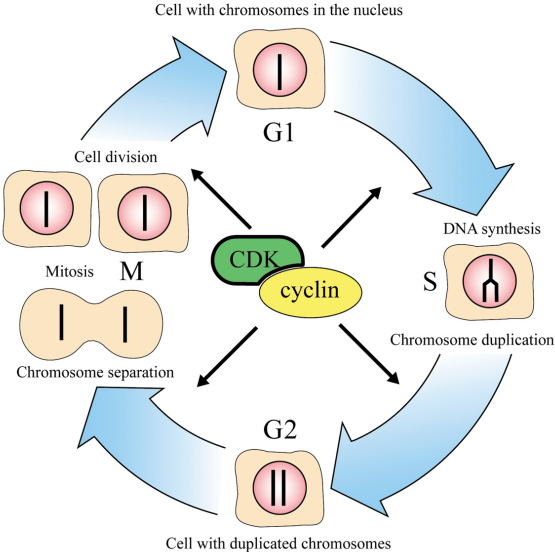
A much more complex scenario:

- deterministic models become DDEs;
- the SSA becomes a non-markovian process;
- different interpretations of delays;

My Ph.D. topic is "Formal Modeling Biological Systems With Delays".

Delays as abstractions: the cell cycle

The Cell Cycle



People in Pisa

- Roberto Barbuti, Professor
- Andrea Maggiolo-Schettini, Professor
- Paolo Milazzo, Research Fellow
- Giulio Caravagna, Ph.D. Student
- Giovanni Pardini, Ph.D. Student
- Aureliano Rama, Ph.D. Student
- Guido Scatena, IMT Ph.D. Student

Topics:

- formal modeling biological systems (Kohn MIM, CLS, spatiality, delays);
- probabilistic analysis of models;
- evolutionary models;
- Natural Computing (PE Systems, P Automata).

References

R.Barbuti, G.Caravagna, A.Maggiolo-Schettini, P.Milazzo and G.Pardini " **The Calculus of Looping Sequences**" Chapter in: M.Bernardo, P.Degano and G.Zavattaro (Eds.): Formal Methods for Computational Systems Biology (SFM 2008), Springer LNCS 5016, pages 387-423, 2008.

R. Barbuti, A. Maggiolo-Schettini, P. Milazzo and A. Troina. " **Bisimulations in Calculi Modelling Membranes**". Formal Aspects of Computing 20, 351-377, 2008.

R. Barbuti, A. Maggiolo-Schettini, P. Milazzo and A. Troina. " **Spatial Calculus of Looping Sequences.** ". ENTCS, 229(1): 21-39 (2009)

R.Barbuti, G.Caravagna, A.Maggiolo-Schettini and P.Milazzo " **On the Interpretation of Delays in Delay Stochastic Simulation of Biological Systems**" Submitted.

R.Barbuti, G.Caravagna, A.Maggiolo-Schettini and P.Milazzo " **A Delay Stochastic Simulation Algorithm with Delayed Propensity Functions**" Draft.

G. Caravagna " **Formal Modeling of Biological Systems With Delays**" Ph.D. Thesis Proposal, Department of Computer Science, University of Pisa, 2009