# Qualitative and Quantitative Formal Modeling of Biological Systems

Paolo Milazzo

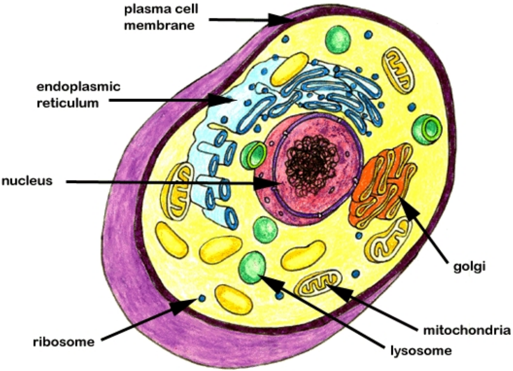
Dipartimento di Informatica, Università di Pisa, Italy

Pisa – June 21st, 2007

# Outline of the talk

1. Introduction
   - Cells are complex interactive systems
   - The EGF pathway and the *lac* operon
2. The Calculus of Looping Sequences (CLS)
   - Definition of CLS
   - The EGF pathway and the *lac* operon in CLS
3. Bisimulations in CLS
   - A labeled semantics for CLS
   - Bisimulations in CLS
   - Bisimulations applied to the CLS model of the *lac* operon
4. CLS variants
   - Stochastic CLS
   - LCLS
5. Future Work and References

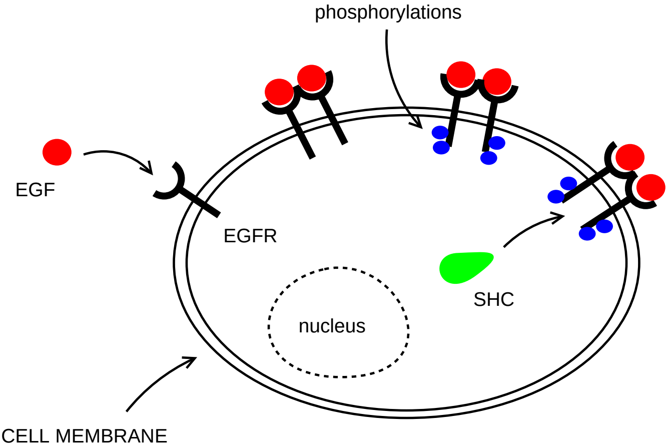
# Cells: complex systems of interactive components



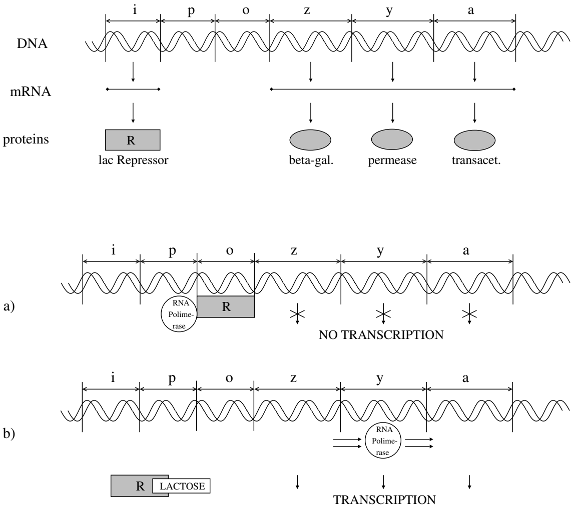
- Two classifications of cell:
  - procaryotic
  - eucaryotic
- Main actors:
  - membranes
  - proteins
  - DNA/RNA strands
- Interaction networks:
  - metabolic pathways
  - signaling pathways
  - gene regulatory networks

Computer Science can provide biologists with formalisms for the description of interactive systems and tools for their analysis.

# Examples of interaction networks: the EGF pathway



# Examples of interaction networks: the *lac* operon



# Outline of the talk

1. Introduction
   - Cells are complex interactive systems
   - The EGF pathway and the *lac* operon
2. The Calculus of Looping Sequences (CLS)
   - Definition of CLS
   - The EGF pathway and the *lac* operon in CLS
3. Bisimulations in CLS
   - A labeled semantics for CLS
   - Bisimulations in CLS
   - Bisimulations applied to the CLS model of the *lac* operon
4. CLS variants
   - Stochastic CLS
   - LCLS
5. Future Work and References

# 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:

$$\begin{array}{rcl} T & ::= & S \quad \big| \quad \left(S\right)^{L} \rfloor\, T \quad \big| \quad T \mid T \\ S & ::= & \epsilon \quad \big| \quad a \quad \big| \quad S \cdot S \end{array}$$

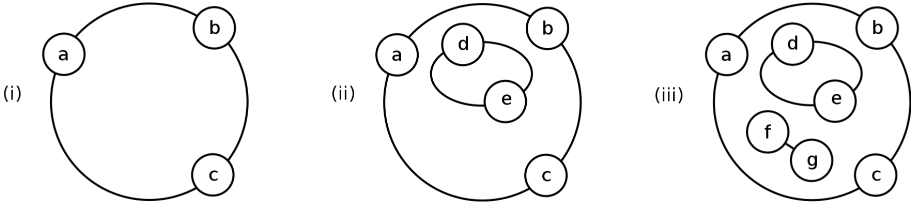
where $a$ is a generic element of $\mathcal{E}$, and $\epsilon$ is the empty sequence.

The operators are:

$S \cdot S$ : Sequencing
$\left(S\right)^{L}$ : Looping ($S$ is closed and it can rotate)
$T_1 \rfloor T_2$ : Containment ($T_1$ contains $T_2$)
$T|T$ : Parallel composition (juxtaposition)

Actually, looping and containment form a single binary operator $\left(S\right)^{L} \rfloor T$.

# Example of Terms



$$(i) \qquad \left(a \cdot b \cdot c\right)^{L} \rfloor \epsilon$$

$$(ii) \qquad \left(a \cdot b \cdot c\right)^{L} \rfloor \left(d \cdot e\right)^{L} \rfloor \epsilon$$

$$(iii) \qquad \left(a \cdot b \cdot c\right)^{L} \rfloor \left(f \cdot g \mid \left(d \cdot e\right)^{L} \rfloor \epsilon\right)$$

# 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:

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

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

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

We write $\equiv$ for $\equiv_T$.

# CLS Patterns

Let us consider variables of three kinds:

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

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

$$\begin{array}{rcl} P & ::= & SP \quad \big| \quad (SP)^L \rfloor P \quad \big| \quad P \mid P \quad \big| \quad X \\ SP & ::= & \epsilon \quad \big| \quad a \quad \big| \quad SP \cdot SP \quad \big| \quad x \quad \big| \quad \widetilde{x} \end{array}$$

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

The structural congruence relation $\equiv$ extends trivially to patterns

# Rewrite Rules

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

$\Sigma$ is the set of all possible instantiations $\sigma$

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 $P\sigma$.

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 $a \cdot c \cdot c \cdot a$

# Formal Semantics

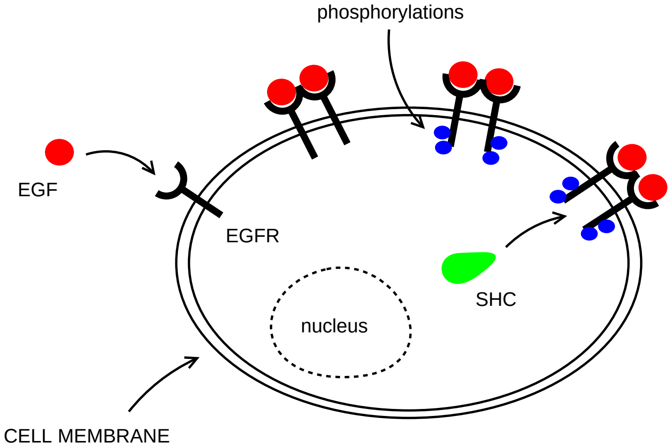
Given a set of rewrite rules $\mathcal{R}$, evolution of terms is described by the transition system given by the least relation → satisfying

$$\frac{P \mapsto P' \in \mathcal{R} \qquad P\sigma \not\equiv \epsilon}{P\sigma \rightarrow P'\sigma}$$

$$\frac{T \rightarrow T'}{T \mid T'' \rightarrow T' \mid T''} \qquad \frac{T \rightarrow T'}{(S)^L \rfloor T \rightarrow (S)^L \rfloor T'}$$

and closed under structural congruence $\equiv$.

# CLS modeling examples: the EGF pathway (1)



# CLS modeling examples: the EGF pathway (2)

First steps of the EGF signaling pathway up to the binding of the signal-receptor dimer to the SHC protein

- The EGFR,EGF and SHC proteins are modeled as the alphabet symbols *EGFR*, *EGF* and *SHC*, respectively
- The cell is modeled as a looping sequence (representing its external membrane):

$$EGF \mid EGF \mid \left(EGFR \cdot EGFR \cdot EGFR \cdot EGFR\right)^{L} \rfloor \left(SHC \mid SHC\right)$$

Rewrite rules modeling the first steps of the pathway:

$$EGF \mid \left(EGFR \cdot \widetilde{x}\right)^{L} \rfloor X \mapsto \left(CMPLX \cdot \widetilde{x}\right)^{L} \rfloor X \tag{R1}$$

$$\left(CMPLX \cdot \widetilde{x} \cdot CMPLX \cdot \widetilde{y}\right)^{L} \rfloor X \mapsto \left(DIM \cdot \widetilde{x} \cdot \widetilde{y}\right)^{L} \rfloor X \tag{R2}$$

$$\left(DIM \cdot \widetilde{x}\right)^{L} \rfloor X \mapsto \left(DIMp \cdot \widetilde{x}\right)^{L} \rfloor X \tag{R3}$$

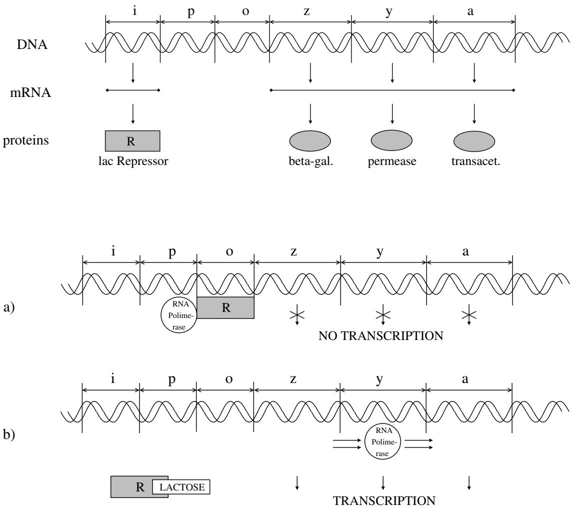
$$\left(DIMp \cdot \widetilde{x}\right)^{L} \rfloor \left(SHC \mid X\right) \mapsto \left(DIMpSHC \cdot \widetilde{x}\right)^{L} \rfloor X \tag{R4}$$

# CLS modeling examples: the EGFR pathway (2)

A possible evolution of the system:

$$
\begin{aligned}
& EGF \mid EGF \mid \big(EGFR \cdot EGFR \cdot EGFR \cdot EGFR\big)^{L} \rfloor (SHC \mid SHC) \\
\xrightarrow{(R1)}\ & EGF \mid \big(EGFR \cdot CMPLX \cdot EGFR \cdot EGFR\big)^{L} \rfloor (SHC \mid SHC) \\
\xrightarrow{(R1)}\ & \big(EGFR \cdot CMPLX \cdot EGFR \cdot CMPLX\big)^{L} \rfloor (SHC \mid SHC) \\
\xrightarrow{(R2)}\ & \big(EGFR \cdot DIM \cdot EGFR\big)^{L} \rfloor (SHC \mid SHC) \\
\xrightarrow{(R3)}\ & \big(EGFR \cdot DIMp \cdot EGFR\big)^{L} \rfloor (SHC \mid SHC) \\
\xrightarrow{(R4)}\ & \big(EGFR \cdot DIMpSHC \cdot EGFR\big)^{L} \rfloor SHC
\end{aligned}
$$

# CLS modeling examples: the *lac* operon (1)



# CLS modeling examples: the *lac* operon (2)

$$Ecoli \ ::= \ (m)^L \rfloor (lacI \cdot lacP \cdot lacO \cdot lacZ \cdot lacY \cdot lacA \mid polym)$$

Rules for DNA transcription/translation:

$$\begin{aligned}
lacI \cdot \widetilde{x} &\mapsto lacI' \cdot \widetilde{x} \mid repr && \text{(R1)}\\
polym \mid \widetilde{x} \cdot lacP \cdot \widetilde{y} &\mapsto \widetilde{x} \cdot PP \cdot \widetilde{y} && \text{(R2)}\\
\widetilde{x} \cdot PP \cdot lacO \cdot \widetilde{y} &\mapsto \widetilde{x} \cdot lacP \cdot PO \cdot \widetilde{y} && \text{(R3)}\\
\widetilde{x} \cdot PO \cdot lacZ \cdot \widetilde{y} &\mapsto \widetilde{x} \cdot lacO \cdot PZ \cdot \widetilde{y} && \text{(R4)}\\
\widetilde{x} \cdot PZ \cdot lacY \cdot \widetilde{y} &\mapsto \widetilde{x} \cdot lacZ \cdot PY \cdot \widetilde{y} \mid betagal && \text{(R5)}\\
\widetilde{x} \cdot PY \cdot lacA &\mapsto \widetilde{x} \cdot lacY \cdot PA \mid perm && \text{(R6)}\\
\widetilde{x} \cdot PA &\mapsto \widetilde{x} \cdot lacA \mid transac \mid polym && \text{(R7)}
\end{aligned}$$

# CLS modeling examples: the *lac* operon (3)

$$Ecoli \ ::= \ \left(m\right)^{L} \rfloor \left(lacI \cdot lacP \cdot lacO \cdot lacZ \cdot lacY \cdot lacA \mid polym\right)$$

Rules to describe the binding of the lac Repressor to gene o, and what happens when lactose is present in the environment of the bacterium:

$$\begin{aligned}
repr \mid \widetilde{x} \cdot lacO \cdot \widetilde{y} &\mapsto \widetilde{x} \cdot RO \cdot \widetilde{y} && \text{(R8)}\\
LACT \mid \left(m \cdot \widetilde{x}\right)^{L} \rfloor X &\mapsto \left(m \cdot \widetilde{x}\right)^{L} \rfloor \left(X \mid LACT\right) && \text{(R9)}\\
\widetilde{x} \cdot RO \cdot \widetilde{y} \mid LACT &\mapsto \widetilde{x} \cdot lacO \cdot \widetilde{y} \mid RLACT && \text{(R10)}\\
\left(\widetilde{x}\right)^{L} \rfloor \left(perm \mid X\right) &\mapsto \left(perm \cdot \widetilde{x}\right)^{L} \rfloor X && \text{(R11)}\\
LACT \mid \left(perm \cdot \widetilde{x}\right)^{L} \rfloor X &\mapsto \left(perm \cdot \widetilde{x}\right)^{L} \rfloor \left(LACT \mid X\right) && \text{(R12)}\\
betagal \mid LACT &\mapsto betagal \mid GLU \mid GAL && \text{(R13)}
\end{aligned}$$

# CLS modeling examples: the *lac* operon (4)

$$Ecoli \;::=\; (m)^{L} \rfloor (lacI \cdot lacP \cdot lacO \cdot lacZ \cdot lacY \cdot lacA \mid polym)$$

Example:

$$\begin{aligned}
& Ecoli|LACT|LACT \\
\to^{*}\; & (m)^{L} \rfloor (lacI' \cdot lacP \cdot lacO \cdot lacZ \cdot lacY \cdot lacA \mid polym \mid repr)|LACT|LACT \\
\to^{*}\; & (m)^{L} \rfloor (lacI' \cdot lacP \cdot RO \cdot lacZ \cdot lacY \cdot lacA \mid polym)|LACT|LACT \\
\to^{*}\; & (m)^{L} \rfloor (lacI' \cdot lacP \cdot lacO \cdot lacZ \cdot lacY \cdot lacA|polym|RLACT)|LACT \\
\to^{*}\; & (perm \cdot m)^{L} \rfloor (lacI'{-}A|betagal|transac|polym|RLACT)|LACT \\
\to^{*}\; & (perm \cdot m)^{L} \rfloor (lacI'{-}A|betagal|transac|polym|RLACT|GLU|GAL)
\end{aligned}$$

# Outline of the talk

1. Introduction
   - Cells are complex interactive systems
   - The EGF pathway and the *lac* operon
2. The Calculus of Looping Sequences (CLS)
   - Definition of CLS
   - The EGF pathway and the *lac* operon in CLS
3. Bisimulations in CLS
   - A labeled semantics for CLS
   - Bisimulations in CLS
   - Bisimulations applied to the CLS model of the *lac* operon
4. CLS variants
   - Stochastic CLS
   - LCLS
5. Future Work and References

# Bisimulations

Bisimilarity is widely accepted as the finest extensional behavioral equivalence one may impose on systems.

- Two systems are bisimilar if they can perform step by step the same interactions with the environment.
- Properties of a system can be verified by assessing the bisimilarity with a system known to enjoy them.

Bisimilarities need semantics based on labeled transition relations capturing the potential interactions with the environment.

- In process calculi, transitions are usually labeled with actions.
- In CLS labels are contexts in which rules can be applied.

# Labeled semantics (1)

**Contexts** $\mathcal{C}$ are given by the following grammar:

$$\mathcal{C} ::= \Box \quad | \quad \mathcal{C} \mid T \quad | \quad T \mid \mathcal{C} \quad | \quad (S)^L \rfloor \mathcal{C}$$

where $T \in \mathcal{T}$ and $S \in \mathcal{S}$. Context $\Box$ is called the *empty context*.

**Parallel Contexts** $\mathcal{C}_P$ are given by the following grammar:

$$\mathcal{C}_P ::= \Box \quad | \quad \mathcal{C}_P \mid T \quad | \quad T \mid \mathcal{C}_P.$$

where $T \in \mathcal{T}$.

$C[T]$ is context application and $C[C']$ is context composition.

# Labeled semantics (2)

Given a set of rewrite rules $\mathcal{R} \subseteq \Re$, the **labeled semantics** of CLS is the labeled transition system given by the following inference rules:

$$\text{(rule\_appl)}\ \frac{P \mapsto P' \in \mathcal{R} \quad C[T''] \equiv P\sigma \quad T'' \not\equiv \epsilon \quad \sigma \in \Sigma \quad C \in \mathcal{C}}{T'' \xrightarrow{C} P'\sigma}$$

$$\text{(cont)}\ \frac{T \xrightarrow{\Box} T'}{(S)^L \rfloor T \xrightarrow{\Box} (S)^L \rfloor T'} \qquad \text{(par)}\ \frac{T \xrightarrow{C} T' \quad C \in \mathcal{C}_P}{T \mid T'' \xrightarrow{C} T' \mid T''}$$

where the dual version of the *(par)* rule is omitted.

Rule (rule_appl) describes the (potential) application of a rule.

- $T'' \not\equiv \epsilon$ in the premise implies that $C$ cannot provide completely the left hand side of the rewrite rule.
- Example: let $R = a \mid b \mapsto c$, we have $a \xrightarrow{\Box \mid b} c$, but $\epsilon \overset{a|b}{\not\longrightarrow}$.

# Labeled semantics (3)

Given a set of rewrite rules $\mathcal{R} \subseteq \Re$, the **labeled semantics** of CLS is the labeled transition system given by the following inference rules:

$$\text{(rule\_appl)}\ \frac{P \mapsto P' \in \mathcal{R} \quad C[T''] \equiv T\sigma \quad T'' \not\equiv \epsilon \quad \sigma \in \Sigma \quad C \in \mathcal{C}}{T'' \xrightarrow{C} P'\sigma}$$

$$\text{(cont)}\ \frac{T \xrightarrow{\Box} T'}{(S)^L \rfloor T \xrightarrow{\Box} (S)^L \rfloor T'} \qquad \text{(par)}\ \frac{T \xrightarrow{C} T' \quad C \in \mathcal{C}_P}{T \mid T'' \xrightarrow{C} T' \mid T''}$$

where the dual version of the *(par)* rule is omitted.

Rule (cont) propagates $\Box$–labeled transitions from the inside to the outside of a looping sequence.

- Transition labeled with a non–empty context cannot be propagated.
- Example: let $R = a \mid b \mapsto c$, we have $a \xrightarrow{\Box \mid b} c$, but $(d)^L \rfloor a \not\xrightarrow{\Box|b}$.

# Labeled semantics (4)

Given a set of rewrite rules $\mathcal{R} \subseteq \Re$, the **labeled semantics** of CLS is the labeled transition system given by the following inference rules:

$$\text{(rule\_appl)}\ \frac{P \mapsto P' \in \mathcal{R} \quad C[T''] \equiv T\sigma \quad T'' \not\equiv \epsilon \quad \sigma \in \Sigma \quad C \in \mathcal{C}}{T'' \xrightarrow{C} P'\sigma}$$

$$\text{(cont)}\ \frac{T \xrightarrow{\Box} T'}{(S)^L \rfloor T \xrightarrow{\Box} (S)^L \rfloor T'} \qquad \text{(par)}\ \frac{T \xrightarrow{C} T' \quad C \in \mathcal{C}_P}{T \mid T'' \xrightarrow{C} T' \mid T''}$$

where the dual version of the *(par)* rule is omitted.

Rule (par) propagates transitions labeled with parallel contexts in parallel components.

- Example: let $R = (a)^L \rfloor b \mapsto c$, we have $b \xrightarrow{(a)^L \rfloor \Box} c$, but $b \mid d \overset{(a)^L \rfloor \Box}{\not\longrightarrow}$ because $R$ cannot be applied $(a)^L \rfloor (b \mid d)$

# Bisimulations in CLS (1)

A binary relation $R$ on terms is a **strong bisimulation** if, given $T_1, T_2$ such that $T_1 R T_2$, the two following conditions hold:

- $T_1 \xrightarrow{C} T_1' \implies \exists T_2'$ s.t. $T_2 \xrightarrow{C} T_2'$and $T_1' R T_2'$
- $T_2 \xrightarrow{C} T_2' \implies \exists T_1'$ s.t. $T_1 \xrightarrow{C} T_1'$ and $T_2' R T_1'$.

The *strong bisimilarity* $\sim$ is the largest of such relations.

A binary relation $R$ on terms is a **weak bisimulation** if, given $T_1, T_2$ such that $T_1 R T_2$, the two following conditions hold:

- $T_1 \xrightarrow{C} T_1' \implies \exists T_2'$ s.t. $T_2 \overset{C}{\Longrightarrow} T_2'$and $T_1' R T_2'$
- $T_2 \xrightarrow{C} T_2' \implies \exists T_1'$ s.t. $T_1 \overset{C}{\Longrightarrow} T_1'$ and $T_2' R T_1'$.

The *weak bisimilarity* $\approx$ is the largest of such relations.

**Theorem:** Strong and weak bisimilarities are congruences.

# Bisimulations in CLS (2)

Consider the following set of rewrite rules:

$$\mathcal{R} = \{ \quad a \mid b \mapsto c \quad , \quad d \mid b \mapsto e \quad , \quad e \mapsto e \quad , \quad c \mapsto e \quad , \quad f \mapsto a \quad \}$$

We have that $a \sim d$, because

$$a \xrightarrow{\square|b} c \xrightarrow{\square} e \xrightarrow{\square} e \xrightarrow{\square} \dots$$

$$d \xrightarrow{\square|b} e \xrightarrow{\square} e \xrightarrow{\square} \dots$$

and $f \approx d$, because

$$f \xrightarrow{\square} a \xrightarrow{\square|b} c \xrightarrow{\square} e \xrightarrow{\square} e \xrightarrow{\square} \dots$$

On the other hand, $f \not\sim e$ and $f \not\approx e$.

$$e \xrightarrow{\square} e \xrightarrow{\square} e \xrightarrow{\square} \dots$$

# Bisimulations in CLS (3)

Let us consider systems $(T, \mathcal{R})$...

A binary relation $R$ is a **strong bisimulation on systems** if, given $(T_1, \mathcal{R}_1)$ and $(T_2, \mathcal{R}_2)$ such that $(T_1, \mathcal{R}_1) R (T_2, \mathcal{R}_2)$:

- $\mathcal{R}_1 : T_1 \xrightarrow{C} T_1' \implies \exists T_2'$ s.t. $\mathcal{R}_2 : T_2 \xrightarrow{C} T_2'$ and $(T_1', \mathcal{R}_1) R (T_2', \mathcal{R}_2)$
- $\mathcal{R}_2 : T_2 \xrightarrow{C} T_2' \implies \exists T_1'$ s.t. $\mathcal{R}_1 : T_1 \xrightarrow{C} T_1'$ and $(R_2, T_2') R (\mathcal{R}_1, T_1')$.

The *strong bisimilarity on systems* $\sim$ is the largest of such relations.

A binary relation $R$ is a **weak bisimulation on systems** if, given $(T_1, \mathcal{R}_1)$ and $(T_2, \mathcal{R}_2)$ such that $(T_1, \mathcal{R}_1) R (T_2, \mathcal{R}_2)$:

- $\mathcal{R}_1 : T_1 \xrightarrow{C} T_1' \implies \exists T_2'$ s.t. $\mathcal{R}_2 : T_2 \overset{C}{\Longrightarrow} T_2'$ and $(T_1', \mathcal{R}_1) R (T_2', \mathcal{R}_2)$
- $\mathcal{R}_2 : T_2 \xrightarrow{C} T_2' \implies \exists T_1'$ s.t. $\mathcal{R}_1 : T_1 \overset{C}{\Longrightarrow} T_1'$ and $(T_2', \mathcal{R}_2) R (T_1', \mathcal{R}_1)$

The *weak bisimilarity on systems* $\approx$ is the largest of such relations.

Strong and weak bisimilarities on systems are NOT congruences.

# Bisimulations in CLS (4)

Consider the following sets of rewrite rules

$$\mathcal{R}_1 = \{a \mid b \mapsto c\} \qquad \mathcal{R}_2 = \{a \mid d \mapsto c\ ,\ \ b \mid e \mapsto c\}$$

We have that $\langle a, \mathcal{R}_1 \rangle \approx \langle e, \mathcal{R}_2 \rangle$ because

$$\mathcal{R}_1 : a \xrightarrow{\square | b} c \qquad \mathcal{R}_2 : e \xrightarrow{\square | b} c$$

and $\langle b, \mathcal{R}_1 \rangle \approx \langle d, \mathcal{R}_2 \rangle$, because

$$\mathcal{R}_1 : b \xrightarrow{\square | a} c \qquad \mathcal{R}_2 : d \xrightarrow{\square | a} c$$

but $\langle a \mid b, \mathcal{R}_1 \rangle \not\approx \langle e \mid d, \mathcal{R}_2 \rangle$, because

$$\mathcal{R}_1 : a \mid b \xrightarrow{\square} c \qquad \mathcal{R}_2 : c \mid d \overset{\square}{\not\to}$$

# Applying bisimulations to the *lac* operon (1)

$$Ecoli \ ::= \ (m)^{L} \rfloor (lacI \cdot lacP \cdot lacO \cdot lacZ \cdot lacY \cdot lacA \mid polym)$$

It can be easily proved that

$$\begin{gathered} lacI \cdot lacP \cdot lacO \cdot lacZ \cdot lacY \cdot lacA \\ \approx \\ lacP \cdot lacO \cdot lacZ \cdot lacY \cdot lacA \mid repr \end{gathered}$$

and since weak bisimularity is a congruence the former can be replaced by the latter in the model.

# Applying bisimulations to the *lac* operon (2)

By using the weak bisimilarity on systems we can prove that from the state in which the repressor is bound to the DNA we can reach a state in which the enzymes are synthesized only if lactose appears in the environment.

We replace rule

$$\widetilde{x} \cdot RO \cdot \widetilde{y} \mid LACT \;\mapsto \widetilde{x} \cdot lacO \cdot \widetilde{y} \mid RLACT \qquad \text{(R10)}$$

with

$$\begin{aligned} &\left(\widetilde{w}\right)^{L} \rfloor \left(\widetilde{x} \cdot RO \cdot \widetilde{y} \mid LACT \mid X\right) \mid START \;\mapsto \\ &\qquad\qquad \left(\widetilde{w}\right)^{L} \rfloor \left(\widetilde{x} \cdot lacO \cdot \widetilde{y} \mid RLACT \mid X\right) \qquad \text{(R10bis)} \end{aligned}$$

The obtained model is bisimilar to $(T_1, \mathcal{R})$ where $\mathcal{R}$ is

$$\begin{array}{llll} T_1 \mid LACT \;\mapsto\; T_2 \quad \text{(R1')} & \qquad & T_2 \mid START \;\mapsto\; T_3 \quad \text{(R3')} \\ T_2 \mid LACT \;\mapsto\; T_2 \quad \text{(R2')} & \qquad & T_3 \mid LACT \;\mapsto\; T_3 \quad \text{(R4')} \end{array}$$

that is a system satisfying the property.

# 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

Bisimilarities of Brane Calculi are preserved after translation into CLS

# Some variants of CLS

- Full–CLS
  - The looping operator can be applied to any term
  - Rule $a \mid b \mapsto c$ can be applied to $b \mid \left(a \cdot a \cdot a \cdot a\right)^{L} \rfloor d$

- CLS+
  - More realistic representation of the fluid nature of membranes: the looping operator can be applied to parallel compositions of sequences
  - Can be encoded into CLS

- Stochastic CLS
  - The application of a rule consumes a stochastic quantity of time

- LCLS (CLS with Links)
  - Description of protein–protein interactions at the domain level

# Outline of the talk

1. Introduction
   - Cells are complex interactive systems
   - The EGF pathway and the *lac* operon
2. The Calculus of Looping Sequences (CLS)
   - Definition of CLS
   - The EGF pathway and the *lac* operon in CLS
3. Bisimulations in CLS
   - A labeled semantics for CLS
   - Bisimulations in CLS
   - Bisimulations applied to the CLS model of the *lac* operon
4. CLS variants
   - Stochastic CLS
   - LCLS
5. Future Work and References

# Background: the kinetics of chemical reactions

Usual notation for chemical reactions:

$$\ell_1 S_1 + \ldots + \ell_\rho S_\rho \underset{k_{-1}}{\overset{k}{\rightleftharpoons}} \ell'_1 P_1 + \ldots + \ell'_\gamma P_\gamma$$

where:

- $S_i, P_i$ are molecules (reactants)
- $\ell_i, \ell'_i$ are stoichiometric coefficients
- $k, k_{-1}$ are the kinetic constants

The kinetics is described by the *law of mass action*:

$$\frac{d[P_i]}{dt} = \ell'_i \underbrace{k[S_1]^{\ell_1} \cdots [S_\rho]^{\ell_\rho}}_{reaction\ \ rate} \qquad \frac{d[S_i]}{dt} = \ell_i \underbrace{k_{-1}[P_1]^{\ell'_1} \cdots [P_\gamma]^{\ell'_\gamma}}_{reaction\ \ rate}$$

# Background: Gillespie's simulation algorithm

- represents a chemical solution as a multiset of molecules
- computes the reaction rate $a_\mu$ 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 $S_2$

reaction $R_1 : S_1 + S_2 \rightarrow 2S_1$ rate $a_1 = X_1 X_2 c_1$

reaction $R_2 : 2S_1 \rightarrow S_1 + S_2$ rate $a_2 = \frac{X_1(X_1-1)}{2} c_2$

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

- The time $t + \tau$ at which the next reaction will occur is randomly chosen with $\tau$ exponentially distributed with parameter $\sum_{\nu=1}^{M} a_\nu$;
- The reaction $R_\mu$ 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 $\tau$ and the chemical solution is updated.

# Stochastic CLS (1)

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

Two main problems:

- What is a reactant in Stochastic CLS?
  - A *subterm* of a term $T$ is a term $T' \not\equiv \epsilon$ such that $T \equiv C[T']$ for some context $C$
  - A *reactant* is an occurence of a subterm
- What happens with variables?
  - We consider rewrite rules containing variables as *rewrite rule schemata*
  - At each step we compute the set of ground rules that can be applied among those obtained by instantiating variables of the rewrite rule schama
  - We reduce the problem of defining the semantics with rule schemata to the simpler problem of defining the semantics with ground rules only

# Stochastic CLS (2)

Given a finite set of rewrite rule schemata $\mathcal{R}$, the semantics of Stochastic CLS is given by the following inference rule

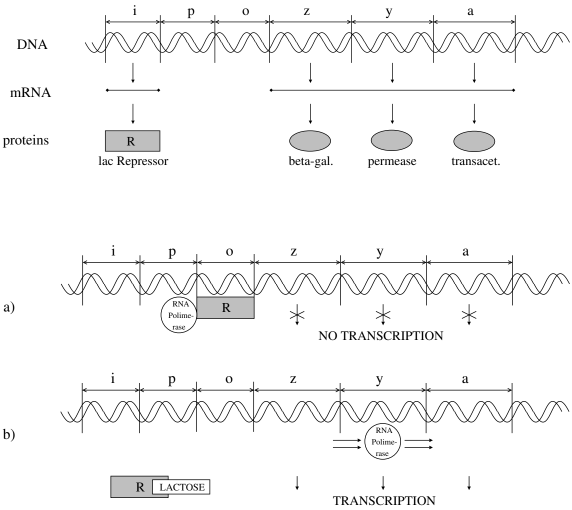
$$\frac{R = T_1 \overset{k}{\mapsto} T_2 \in AR(\mathcal{R}, T) \qquad T \equiv C[T_1]}{T \xrightarrow{R,k \cdot AC(R,T,C[T_2])} C[T_2]}$$

where:

- $AR(\mathcal{R}, T)$ is the set of ground rewrite rules obtained by schemata in $\mathcal{R}$ and applicable to $T$
- $AC(R, T, T')$ is the number of reactants in $T$ equivalent to the left–hand side of the ground rule $R$ and that allows obtaining term $T'$ after the application of $R$

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

# A Stochastic CLS model of the *lac* operon (1)



# A Stochastic CLS model of the *lac* operon (2)

Transcription of DNA, binding of lac Repressor to gene o, and interaction between lactose and lac Repressor:

$$lacI \cdot \widetilde{x} \overset{0.02}{\longmapsto} lacI \cdot \widetilde{x} \mid Irna \tag{S1}$$

$$Irna \overset{0.1}{\longmapsto} Irna \mid repr \tag{S2}$$

$$polym \mid \widetilde{x} \cdot lacP \cdot \widetilde{y} \overset{0.1}{\longmapsto} \widetilde{x} \cdot PP \cdot \widetilde{y} \tag{S3}$$

$$\widetilde{x} \cdot PP \cdot \widetilde{y} \overset{0.01}{\longmapsto} polym \mid \widetilde{x} \cdot lacP \cdot \widetilde{y} \tag{S4}$$

$$\widetilde{x} \cdot PP \cdot lacO \cdot \widetilde{y} \overset{20.0}{\longmapsto} polym \mid Rna \mid \widetilde{x} \cdot lacP \cdot lacO \cdot \widetilde{y} \tag{S5}$$

$$Rna \overset{0.1}{\longmapsto} Rna \mid betagal \mid perm \mid transac \tag{S6}$$

$$repr \mid \widetilde{x} \cdot lacO \cdot \widetilde{y} \overset{1.0}{\longmapsto} \widetilde{x} \cdot RO \cdot \widetilde{y} \tag{S7}$$

$$\widetilde{x} \cdot RO \cdot \widetilde{y} \overset{0.01}{\longmapsto} repr \mid \widetilde{x} \cdot lacO \cdot \widetilde{y} \tag{S8}$$

$$repr \mid LACT \overset{0.005}{\longmapsto} RLACT \tag{S9}$$

$$RLACT \overset{0.1}{\longmapsto} repr \mid LACT \tag{S10}$$

# A Stochastic CLS model of the *lac* operon (3)

The behaviour of the three enzymes for lactose degradation:

$$\left(\widetilde{x}\right)^{L} \rfloor \left(perm \mid X\right) \overset{0.1\cdot f_1}{\longmapsto} \left(perm \cdot \widetilde{x}\right)^{L} \rfloor X \tag{S11}$$

$$LACT \mid \left(perm \cdot \widetilde{x}\right)^{L} \rfloor X \overset{0.001\cdot f_2}{\longmapsto} \left(perm \cdot \widetilde{x}\right)^{L} \rfloor \left(LACT|X\right) \tag{S12}$$

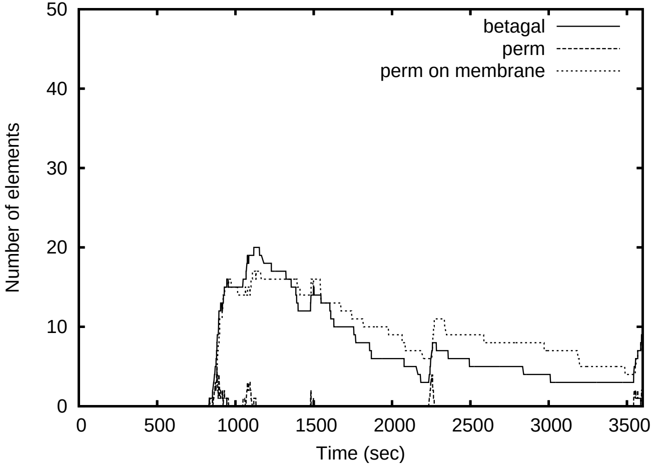
$$betagal \mid LACT \overset{0.001}{\longmapsto} betagal \mid GLU \mid GAL \tag{S13}$$

where $f_1(\sigma) = occ(perm, \sigma(X)) + 1,\ f_2(\sigma) = occ(perm, \sigma(\widetilde{x})) + 1$.

Degradation of all the proteins and mRNA involved in the process:

$$\begin{array}{rcllrcll}
perm & \overset{0.001}{\longmapsto} & \epsilon & \text{(S14)} & betagal & \overset{0.001}{\longmapsto} & \epsilon & \text{(S15)} \\
transac & \overset{0.001}{\longmapsto} & \epsilon & \text{(S16)} & repr & \overset{0.002}{\longmapsto} & \epsilon & \text{(S17)} \\
Irna & \overset{0.01}{\longmapsto} & \epsilon & \text{(S18)} & Rna & \overset{0.01}{\longmapsto} & \epsilon & \text{(S19)} \\
RLACT & \overset{0.002}{\longmapsto} & LACT & \text{(S20)} & & & &
\end{array}$$

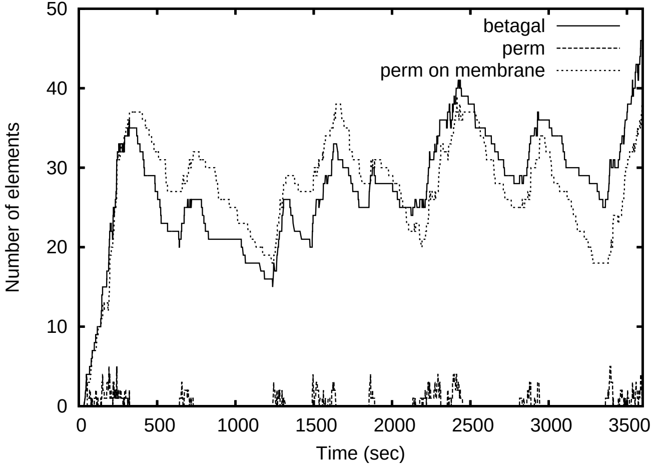
# Simulation results (1)



Production of enzymes in the absence of lactose

$$(m)^L \rfloor (lacI - A \mid 30 \times polym \mid 100 \times repr)$$

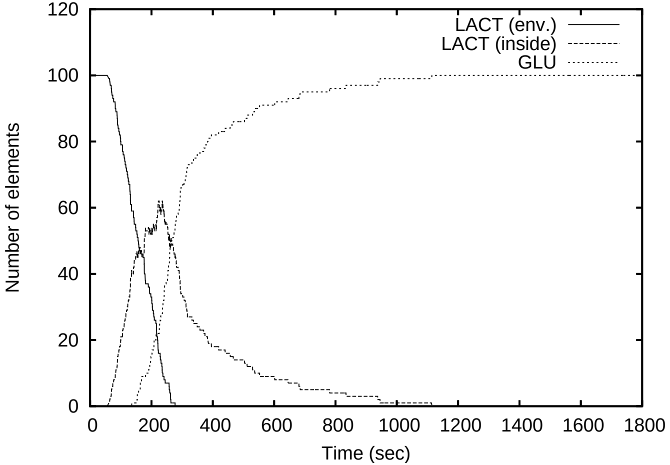
# Simulation results (2)



Production of enzymes in the presence of lactose

$$100 \times LACT \mid (m)^{L} \rfloor (lacI - A \mid 30 \times polym \mid 100 \times repr)$$

# Simulation results (3)



Degradation of lactose into glucose

$$100 \times LACT \mid (m)^{L} \rfloor (lacI - A \mid 30 \times polym \mid 100 \times repr)$$

# Outline of the talk

1. Introduction
   - Cells are complex interactive systems
   - The EGF pathway and the *lac* operon
2. The Calculus of Looping Sequences (CLS)
   - Definition of CLS
   - The EGF pathway and the *lac* operon in CLS
3. Bisimulations in CLS
   - A labeled semantics for CLS
   - Bisimulations in CLS
   - Bisimulations applied to the CLS model of the *lac* operon
4. CLS variants
   - Stochastic CLS
   - LCLS
5. Future Work and References

# Modeling proteins at the domain level

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: $a \cdot b^1 \cdot c \mid d \cdot e^1 \cdot f$

# Syntax of LCLS

**Terms** $T$ and **Sequences** $S$ of LCLS are given by the following grammar:

$$\begin{array}{rcl} T & ::= & S \quad \big| \quad \big(S\big)^{L} \rfloor\, T \quad \big| \quad T \mid T \\ S & ::= & \epsilon \quad \big| \quad a \quad \big| \quad a^{n} \quad \big| \quad S \cdot S \end{array}$$

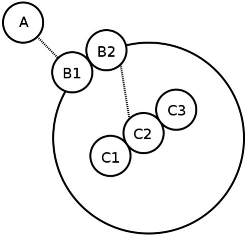
where $a$ is a generic element of $\mathcal{E}$, and $n$ is a natural number.

**Patterns** $P$ and **sequence patterns** $SP$ of LCLS are given by the following grammar:

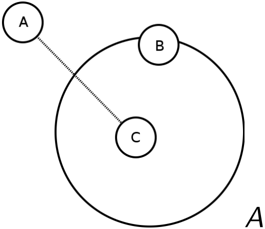
$$\begin{array}{rcl} P & ::= & SP \quad \big| \quad \big(SP\big)^{L} \rfloor\, P \quad \big| \quad P \mid P \quad \big| \quad X \\ SP & ::= & \epsilon \quad \big| \quad a \quad \big| \quad a^{n} \quad \big| \quad SP \cdot SP \quad \big| \quad \widetilde{x} \quad \big| \quad x \quad \big| \quad x^{n} \end{array}$$

where $a$ is an element of $\mathcal{E}$, $n$ is a natural number and $X, \widetilde{x}$ and $x$ are elements of $TV, SV$ and $\mathcal{X}$, respectively.

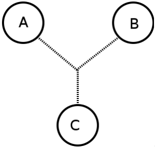
# Well–formedness of LCLS terms and patterns (1)



$A^1 \mid \left(B1^1 \cdot B2^2\right)^L \rfloor C1 \cdot C2^2 \cdot C3$ ✓



$A^1 \mid \left(B\right)^L \rfloor C^1$ ✗



$A^1 \mid B^1 \mid C^1$ ✗

# Well–formedness of LCLS terms and patterns (2)

An LCLS term (or pattern) is well–formed if and only if a label occurs no more than twice, and in the content of a looping sequence a label occours either zero or two times

Type system for well–formedness:

$$1.\ (\varnothing,\varnothing) \models \epsilon \qquad 2.\ (\varnothing,\varnothing) \models a \qquad 3.\ (\varnothing,\{n\}) \models a^n$$

$$4.\ (\varnothing,\varnothing) \models x \qquad 5.\ (\varnothing,\{n\}) \models x^n \qquad 6.\ (\varnothing,\varnothing) \models \widetilde{x} \qquad 7.\ (\varnothing,\varnothing) \models X$$

$$8.\ \frac{(N_1,N_1') \models SP_1 \quad (N_2,N_2') \models SP_2 \quad N_1 \cap N_2 = N_1' \cap N_2 = N_1 \cap N_2' = \varnothing}{(N_1 \cup N_2 \cup (N_1' \cap N_2'), (N_1' \cup N_2') \setminus (N_1' \cap N_2')) \models SP_1 \cdot SP_2}$$

$$9.\ \frac{(N_1,N_1') \models P_1 \quad (N_2,N_2') \models P_2 \quad N_1 \cap N_2 = N_1' \cap N_2 = N_1 \cap N_2' = \varnothing}{(N_1 \cup N_2 \cup (N_1' \cap N_2'), (N_1' \cup N_2') \setminus (N_1' \cap N_2')) \models P_1 \mid P_2}$$

$$10.\ \frac{(N_1,N_1') \models SP \quad (N_2,N_2') \models P \quad N_1 \cap N_2 = N_1' \cap N_2 = N_1 \cap N_2' = \varnothing \quad N_2' \subseteq N_1'}{(N_1 \cup N_2', N_1' \setminus N_2') \models \left(SP\right)^L \rfloor P}$$

# Application of rewrite rules

We would like to ensure that the application of a rewrite rule to a well–formed term preserves well–formedness

- not trivial: well–formedness can be easily violated
- e.g. the rewrite rule $a \mapsto a^1$ applied to $\big(b\big)^L \rfloor a$ produces $\big(b\big)^L \rfloor a^1$

A *compartment safe* rewrite rule is such that

- it does not add/remove occurrences of variables
- it does not moves variables from one compartment (content of a looping sequence) to another one

The application of a compartment safe rewrite rule preserves well–formedness

To apply a *compartment unsafe* rewrite rule we require that

- its patterns are CLOSED
- its variables are instantiated with CLOSED terms

# The semantics of LCLS

Given a set of compartment safe rewrite rules $\mathcal{R}^{CS}$ and a set of compartemnt unsafe rewrite rules $\mathcal{R}^{CU}$, the semantics of LCLS is given by the following rules

$$\text{(appCS)} \quad \frac{P_1 \mapsto P_2 \in \mathcal{R}^{CS} \quad P_1\sigma \not\equiv \epsilon \quad \sigma \in \Sigma \quad \alpha \in \mathcal{A}}{P_1\alpha\sigma \rightarrow P_2\alpha\sigma}$$

$$\text{(appCU)} \quad \frac{P_1 \mapsto P_2 \in \mathcal{R}^{CU} \quad P_1\sigma \not\equiv \epsilon \quad \sigma \in \Sigma_{wf} \quad \alpha \in \mathcal{A}}{P_1\alpha\sigma \rightarrow P_2\alpha\sigma}$$

$$\text{(par)} \quad \frac{T_1 \rightarrow T_1' \qquad L(T_1) \cap L(T_2) = \{n_1, \ldots, n_M\} \qquad n_1', \ldots, n_M' \text{ fresh}}{T_1 \mid T_2 \rightarrow T_1'\{^{n_1', \ldots, n_M'}/_{n_1, \ldots, n_M}\} \mid T_2}$$

$$\text{(cont)} \quad \frac{T \rightarrow T' \qquad L(S) \cap L(T') = \{n_1, \ldots, n_M\} \qquad n_1', \ldots, n_M' \text{ fresh}}{(S)^L \rfloor T \rightarrow (S)^L \rfloor T'\{^{n_1', \ldots, n_M'}/_{n_1, \ldots, n_M}\}}$$

where $\alpha$ is link renaming, $L(T)$ the set of links occurring twice in the top level compartment of $T$

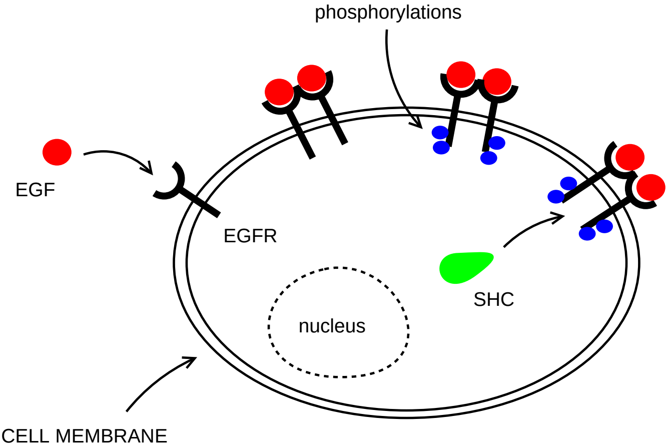
# Main theoretical result

**Theorem (Subject Reduction)**

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

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

# An LCLS model of the EGF pathway (1)



# An LCLS model of the EGF pathway (2)

We model the EGFR protein as the sequence $R_{E1} \cdot R_{E2} \cdot R_{I1} \cdot R_{I2}$

- $R_{E1}$ and $R_{E2}$ are two extra–cellular domains
- $R_{I1}$ and $R_{I2}$ are two intra–cellular domains

The rewrite rules of the model are

$$EGF \mid \left(R_{E1} \cdot \widetilde{x}\right)^{L} \rfloor X \;\mapsto\; EGF^{1} \mid \left(R_{E1}^{1} \cdot \widetilde{x}\right)^{L} \rfloor X \tag{R1}$$

$$\left(R_{E1}^{1} \cdot R_{E2} \cdot \widetilde{x} \cdot R_{E1}^{2} \cdot R_{E2} \cdot \widetilde{y}\right)^{L} \rfloor X \;\mapsto\; \left(R_{E1}^{1} \cdot R_{E2}^{3} \cdot \widetilde{x} \cdot R_{E1}^{2} \cdot R_{E2}^{3} \cdot \widetilde{y}\right)^{L} \rfloor X \tag{R2}$$

$$\left(R_{E2}^{1} \cdot R_{I1} \cdot \widetilde{x}\right)^{L} \rfloor X \;\mapsto\; \left(R_{E2}^{1} \cdot PR_{I1} \cdot \widetilde{x}\right)^{L} \rfloor X \tag{R3}$$

$$\begin{gathered}\left(R_{E2}^{1} \cdot PR_{I1} \cdot R_{I2} \cdot \widetilde{x} \cdot R_{E2}^{1} \cdot PR_{I1} \cdot R_{I2} \cdot \widetilde{y}\right)^{L} \rfloor (SHC \mid X) \;\mapsto \\ \left(R_{E2}^{1} \cdot PR_{I1} \cdot R_{I2}^{2} \cdot \widetilde{x} \cdot R_{E2}^{1} \cdot PR_{I1} \cdot R_{I2} \cdot \widetilde{y}\right)^{L} \rfloor (SHC^{2} \mid X)\end{gathered} \tag{R4}$$

# An LCLS model of the EGF pathway (3)

Let us write $EGFR$ for $R_{E1} \cdot R_{E2} \cdot R_{I1} \cdot R_{I2}$

A possible evolution of the system is

$$
\begin{array}{rl}
 & EGF \mid EGF \mid \left(EGFR \cdot EGFR \cdot EGFR\right)^{L} \rfloor \, (SHC \mid SHC) \\
\xrightarrow{(R1)} & EGF^{1} \mid EGF \mid \left(R_{E1}^{1} \cdot R_{E2} \cdot R_{I1} \cdot R_{I2} \cdot EGFR \cdot EGFR\right)^{L} \rfloor \, (SHC \mid SHC) \\
\xrightarrow{(R1)} & EGF^{1} \mid EGF^{2} \mid \left(R_{E1}^{1} \cdot R_{E2} \cdot R_{I1} \cdot R_{I2} \cdot EGFR \cdot R_{E1}^{2} \cdot R_{E2} \cdot R_{I1} \cdot R_{I2}\right)^{L} \rfloor \, (SHC \mid SHC) \\
\xrightarrow{(R2)} & EGF^{1} \mid EGF^{2} \mid \left(R_{E1}^{1} \cdot R_{E2}^{3} \cdot R_{I1} \cdot R_{I2} \cdot EGFR \cdot R_{E1}^{2} \cdot R_{E2}^{3} \cdot R_{I1} \cdot R_{I2}\right)^{L} \rfloor \, (SHC \mid SHC) \\
\xrightarrow{(R3)} & EGF^{1} \mid EGF^{2} \mid \left(R_{E1}^{1} \cdot R_{E2}^{3} \cdot PR_{I1} \cdot R_{I2} \cdot EGFR \cdot R_{E1}^{2} \cdot R_{E2}^{3} \cdot R_{I1} \cdot R_{I2}\right)^{L} \rfloor \, (SHC \mid SHC) \\
\xrightarrow{(R3)} & EGF^{1} \mid EGF^{2} \mid \left(R_{E1}^{1} \cdot R_{E2}^{3} \cdot PR_{I1} \cdot R_{I2} \cdot EGFR \cdot R_{E1}^{2} \cdot R_{E2}^{3} \cdot PR_{I1} \cdot R_{I2}\right)^{L} \rfloor \, (SHC \mid SHC) \\
\xrightarrow{(R4)} & EGF^{1} \mid EGF^{2} \mid \left(R_{E1}^{1} \cdot R_{E2}^{3} \cdot PR_{I1} \cdot R_{I2}^{4} \cdot EGFR \cdot R_{E1}^{2} \cdot R_{E2}^{3} \cdot PR_{I1} \cdot R_{I2}\right)^{L} \rfloor \, (SHC^{4} \mid SHC)
\end{array}
$$

# Outline of the talk

1. Introduction
   - Cells are complex interactive systems
   - The EGF pathway and the *lac* operon
2. The Calculus of Looping Sequences (CLS)
   - Definition of CLS
   - The EGF pathway and the *lac* operon in CLS
3. Bisimulations in CLS
   - A labeled semantics for CLS
   - Bisimulations in CLS
   - Bisimulations applied to the CLS model of the *lac* operon
4. CLS variants
   - Stochastic CLS
   - LCLS
5. Future Work and References

# Current and future work

We developed a prototype simulator based on Stochastic CLS to run the *lac* operon example

- currently, we are developing a complete and efficient simulator

In order to model cell divisions and differentiations, tissues, etc...

- we are developing a spatial extension of CLS in which terms are placed and can move in a 2D/3D space

Moreover,

- we are developing a translation of Kohn Molecular Interaction Maps into CLS

As future work:

- we plan to study other behavioural equivalences (traces, testing, . . . )
- we plan to use CLS to study (in collaboration with biologists) retinal cell develpment and differentiation

# References

- P. Milazzo. **Qualitative and Quantitative Formal Modeling of Biological Systems**, PhD Thesis, Università di Pisa.
- R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, P. Tiberi and A. Troina. **Stochastic CLS for the Modeling and Simulation of Biological Systems**. Submitted.
- R. Barbuti, A. Maggiolo-Schettini, P. Milazzo and A. Troina. **The Calculus of Looping Sequences for Modeling Biological Membranes**. Invited paper at the 8th Workshop on Membrane Computing (WMC8), LNCS, Springer, to appear.
- R. Barbuti, A. Maggiolo-Schettini and P. Milazzo. **Extending the Calculus of Looping Sequences to Model Protein Interaction at the Domain Level**. Int. Symposium on Bioinformatics Research and Applications (ISBRA'07), LNBI 4463, pages 638–649, Springer, 2006.
- R. Barbuti, A. Maggiolo-Schettini, P. Milazzo and A. Troina. **Bisimulation Congruences in the Calculus of Looping Sequences**. Int. Colloquium on Theor. Aspects of Computing (ICTAC'06), LNCS 4281, pages 93–107, Springer, 2006.
- R. Barbuti, A. Maggiolo-Schettini, P. Milazzo and A. Troina. **A Calculus of Looping Sequences for Modelling Microbiological Systems**. Fundamenta Informaticae, volume 72, pages 21–35, 2006.
- R. Barbuti, A. Maggiolo-Schettini, P. Milazzo and A. Troina. **An Alternative to Gillespie's Algorithm for Simulating Chemical Reactions**. Computational Methods in Systems Biology (CMSB'05), 2005.