Formal Methods and Systems Biology: The Calculus of Looping Sequences

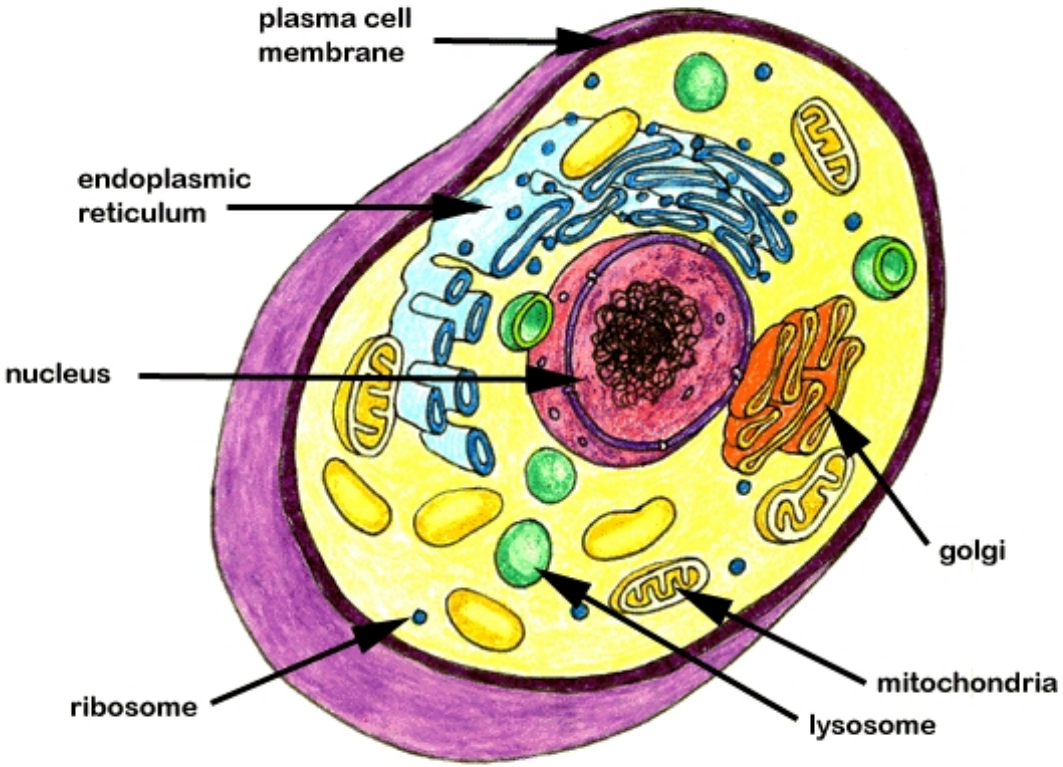
Paolo Milazzo

Dipartimento di Informatica, Università di Pisa, Italy

Verona – January 22, 2008

Outline of the talk

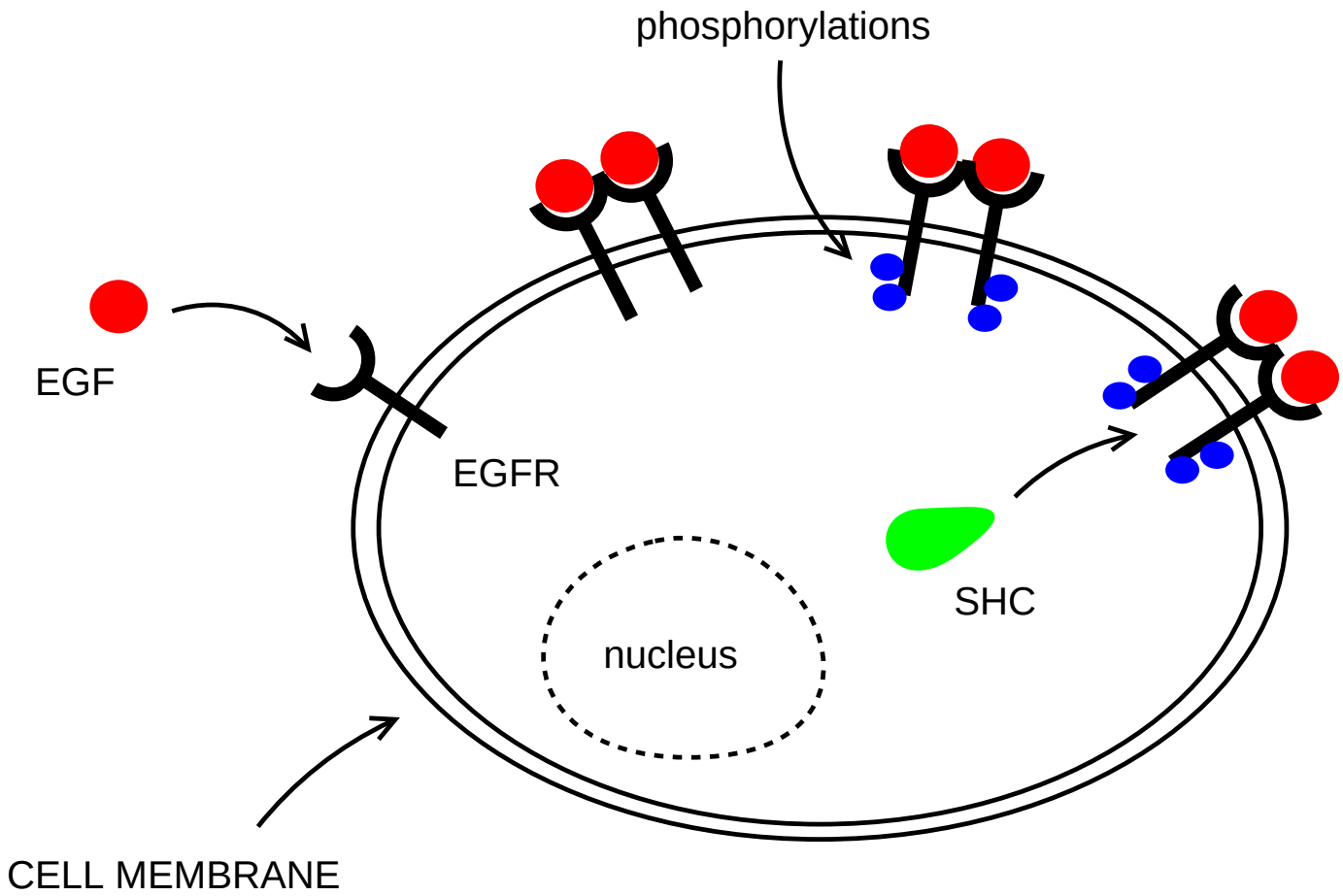
Cells: complex systems of interactive components



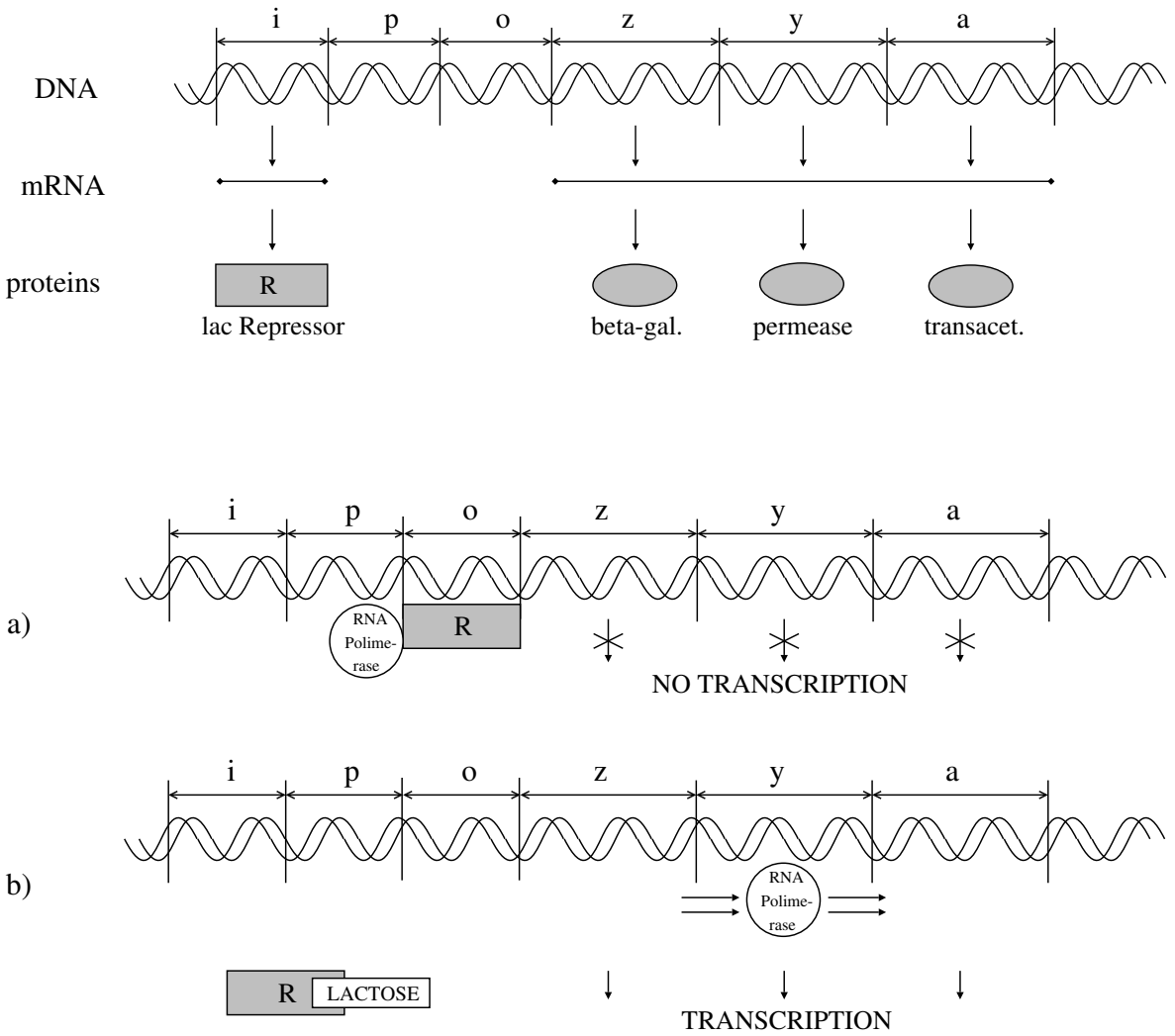
- Two classifications of cell:
  - procaryotic
  - eucaryotic
- Main actors:
  - membranes
  - proteins
  - DNA/RNA
  - ions, macromolecules,...
- Interaction networks:
  - metabolic pathways
  - signaling pathways
  - gene regulatory networks

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

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:

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

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

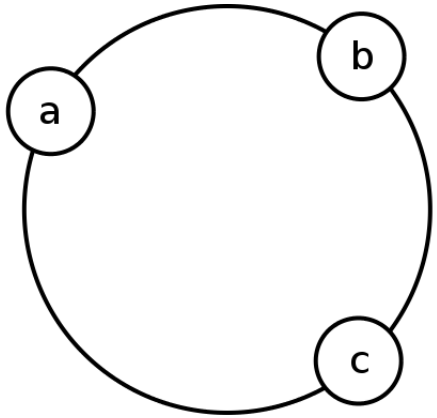
The operators are:

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

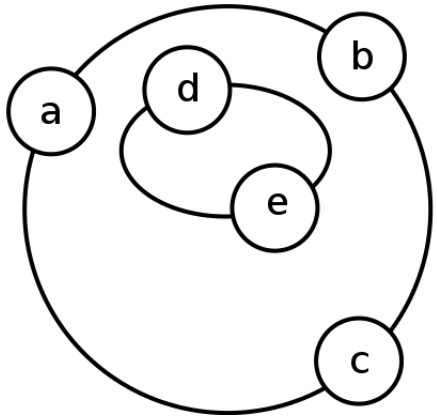
Actually, looping and containment form a single binary operator $(S)^L \rfloor T$.

Examples of Terms

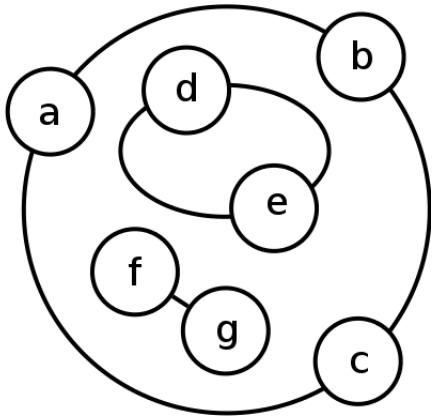
(i)



(ii)



(iii)



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

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

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

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:

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

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

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

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

CLS Patterns

Let us consider variables of three kinds:

- term variables $(X, Y, Z, \ldots)$
- 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:

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

where a is a generic element of $\mathcal{E}$, ϵ 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.

Σ is the set of all possible instantiations σ

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

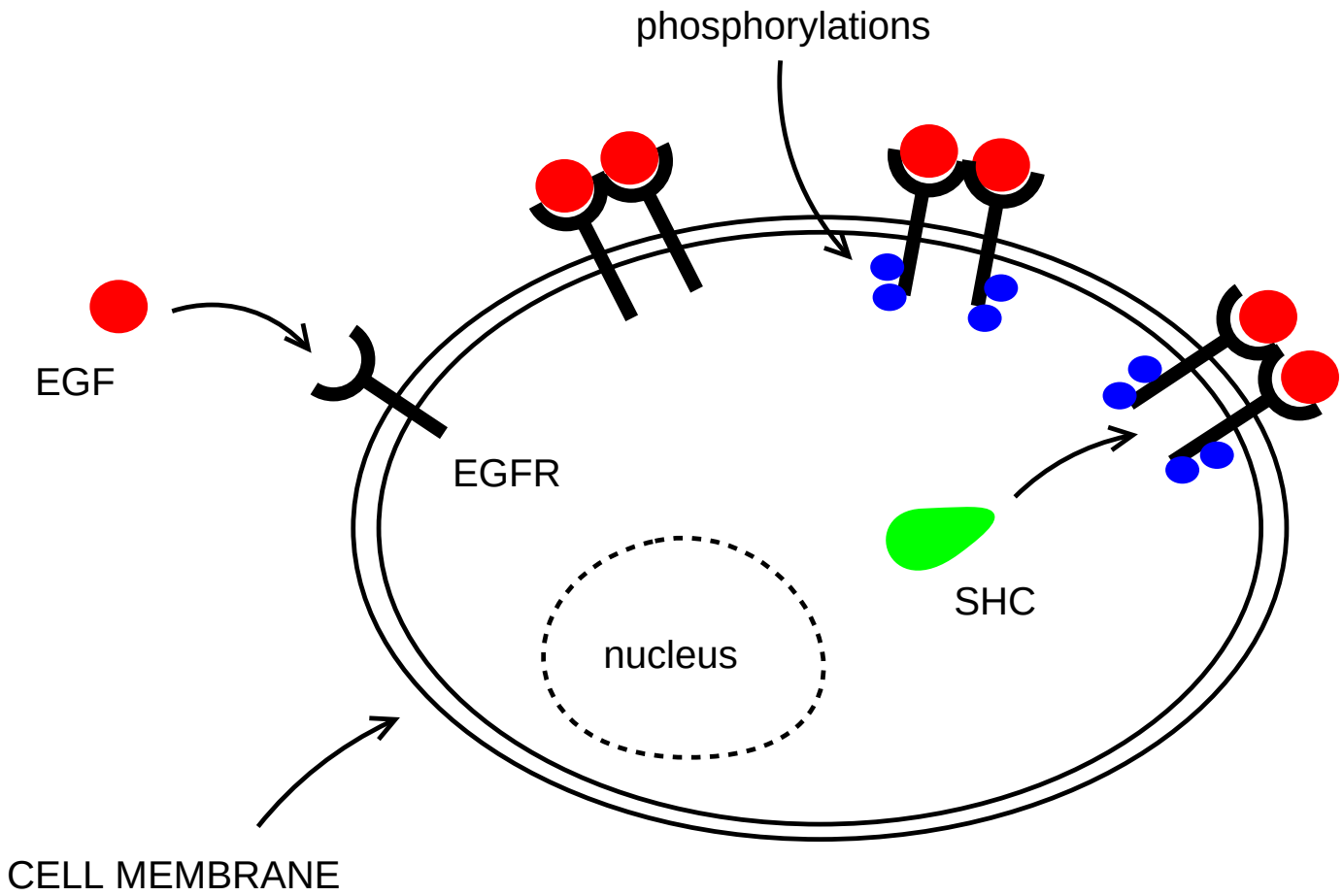
Given a set of rewrite rules $\mathcal{R}$, evolution of terms is described by the transition system given by the least relation $\rightarrow$ satisfying

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

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

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:

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

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

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

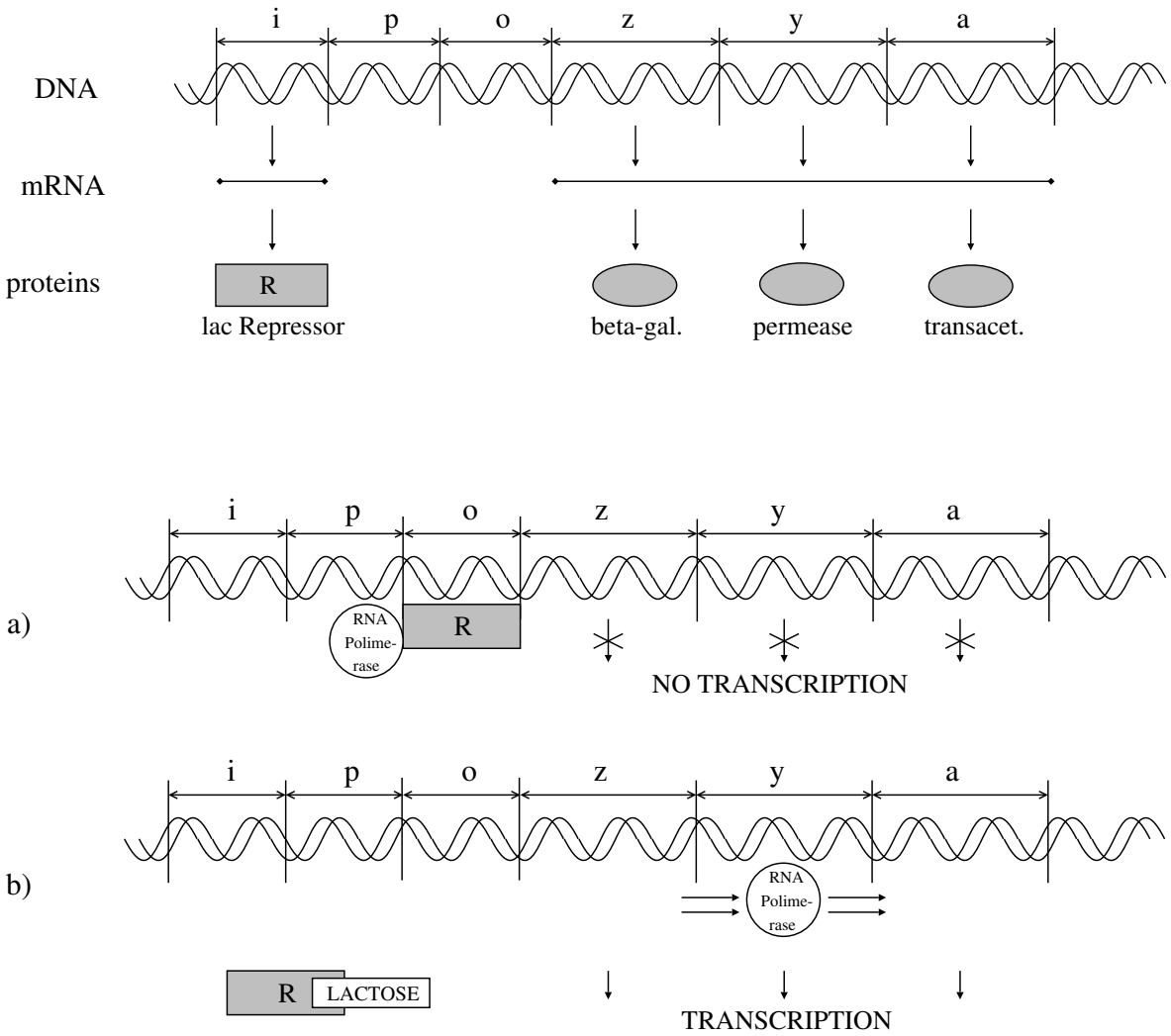
$$\left(DIMp \cdot \widetilde{x}\right)^{L} \rfloor \left(SHC \mid X\right) \;\mapsto\; \left(DIMpSHC \cdot \widetilde{x}\right)^{L} \rfloor X \qquad \text{(R4)}$$

CLS modeling examples: the EGFR pathway (2)

A possible evolution of the system:

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

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



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

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

Rules for DNA transcription/translation:

$$lacI \cdot \widetilde{x} \;\mapsto\; lacI' \cdot \widetilde{x} \mid repr \qquad \text{(R1)}$$

$$polym \mid \widetilde{x} \cdot lacP \cdot \widetilde{y} \;\mapsto\; \widetilde{x} \cdot PP \cdot \widetilde{y} \qquad \text{(R2)}$$

$$\widetilde{x} \cdot PP \cdot lacO \cdot \widetilde{y} \;\mapsto\; \widetilde{x} \cdot lacP \cdot PO \cdot \widetilde{y} \qquad \text{(R3)}$$

$$\widetilde{x} \cdot PO \cdot lacZ \cdot \widetilde{y} \;\mapsto\; \widetilde{x} \cdot lacO \cdot PZ \cdot \widetilde{y} \qquad \text{(R4)}$$

$$\widetilde{x} \cdot PZ \cdot lacY \cdot \widetilde{y} \;\mapsto\; \widetilde{x} \cdot lacZ \cdot PY \cdot \widetilde{y} \mid betagal \qquad \text{(R5)}$$

$$\widetilde{x} \cdot PY \cdot lacA \;\mapsto\; \widetilde{x} \cdot lacY \cdot PA \mid perm \qquad \text{(R6)}$$

$$\widetilde{x} \cdot PA \;\mapsto\; \widetilde{x} \cdot lacA \mid transac \mid polym \qquad \text{(R7)}$$

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

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

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:

$$repr \mid \widetilde{x} \cdot lacO \cdot \widetilde{y} \ \mapsto \widetilde{x} \cdot RO \cdot \widetilde{y} \qquad \text{(R8)}$$

$$LACT \mid (m \cdot \widetilde{x})^L \rfloor X \ \mapsto (m \cdot \widetilde{x})^L \rfloor (X \mid LACT) \qquad \text{(R9)}$$

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

$$(\widetilde{x})^L \rfloor (perm \mid X) \ \mapsto (perm \cdot \widetilde{x})^L \rfloor X \qquad \text{(R11)}$$

$$LACT \mid (perm \cdot \widetilde{x})^L \rfloor X \ \mapsto (perm \cdot \widetilde{x})^L \rfloor (LACT \mid X) \qquad \text{(R12)}$$

$$betagal \mid LACT \ \mapsto betagal \mid GLU \mid GAL \qquad \text{(R13)}$$

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:

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

Outline of the talk

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

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

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

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

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:

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

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

where $\mathcal{C}_P$ are contexts that do not include $\left(S\right)^{L} \rfloor \mathcal{C}$ and the dual version of the *(par)* rule is omitted.

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:

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

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:

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

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

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

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

and $f \approx d$, because

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

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

$$e \xrightarrow{\Box} e \xrightarrow{\Box} e \xrightarrow{\Box} \dots$$

Bisimulations in CLS (3)

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

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 $(T_1, \mathcal{R}_1) R (T_2, \mathcal{R}_2)$:

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

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 $(T_1, \mathcal{R}_1) R (T_2, \mathcal{R}_2)$:

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

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

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

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

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

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

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

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

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

Applying bisimulations to the *lac* operon

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

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

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

$$T_1 \mid LACT \mapsto T_2 \qquad \text{(R1')} \qquad\qquad T_2 \mid START \mapsto T_3 \qquad \text{(R3')}$$
$$T_2 \mid LACT \mapsto T_2 \qquad \text{(R2')} \qquad\qquad T_3 \mid LACT \mapsto T_3 \qquad \text{(R4')}$$

that is a system satisfying the property.

Some variants of CLS

- Full–CLS
  - The looping operator can be applied to any term
  - Terms such as $\left(a \mid \left(b\right)^{L} \rfloor c\right)^{L} \rfloor d$ are allowed
- 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

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)
- ℓ_i, ℓ'_i are stoichiometric coefficients
- k, k_{-1} are the kinetic constants

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

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

Background: Gillespie's simulation algorithm

- represents a chemical solution as a multiset of molecules
- 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

reaction $R_1 : S_1 + S_2 \rightarrow 2S_1$ rate $a_1 = \binom{X_1}{1}\binom{X_2}{1}k_1 = X_1 X_2 k_1$

reaction $R_2 : 2S_1 \rightarrow S_1 + S_2$ rate $a_2 = \binom{X_1}{2}k_2 = \frac{X_1(X_1-1)}{2}k_2$

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 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 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 a rule $(a)^L \rfloor (b \mid X) \mapsto (c)^L \rfloor X$ as a reaction between a molecule a on a membrane and *any* molecule b contained in the membrane.
  - The semantics has to count how many times b occurs in the instantiation of X

Stochastic CLS (2)

Let us assume the syntax of Full–CLS...

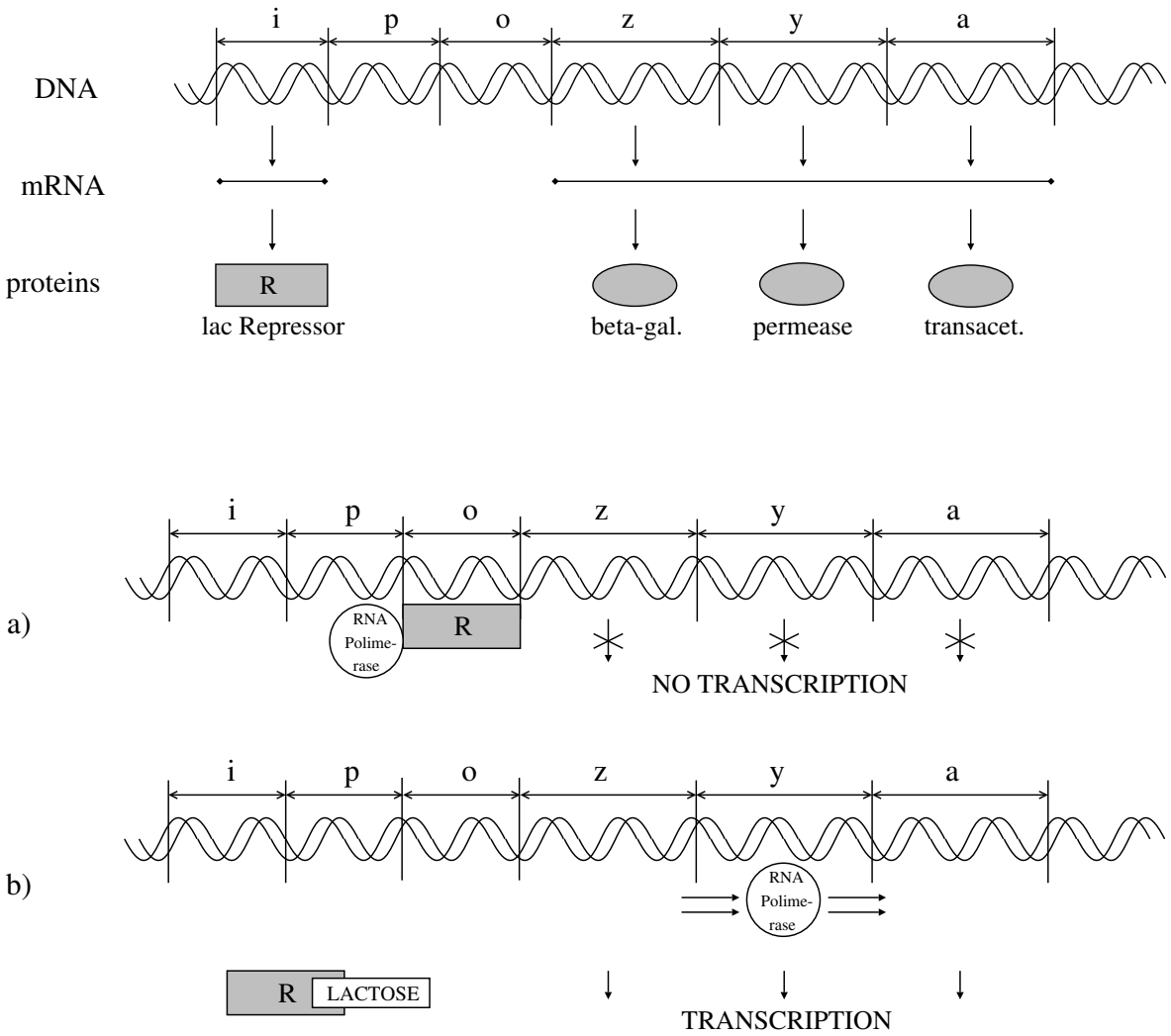
Given a finite set of stochastic rewrite rules $\mathcal{R}$, the semantics of Stochastic CLS is the least transition relation $\xrightarrow{R,T,r,b}$ closed wrt $\equiv$ and satisfying by the following inference rules:

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

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

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:

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

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

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

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

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

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

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

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

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

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

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

The behaviour of the three enzymes for lactose degradation:

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

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

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

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

$$perm \overset{0.001}{\mapsto} \epsilon \tag{S14}$$

$$betagal \overset{0.001}{\mapsto} \epsilon \tag{S15}$$

$$transac \overset{0.001}{\mapsto} \epsilon \tag{S16}$$

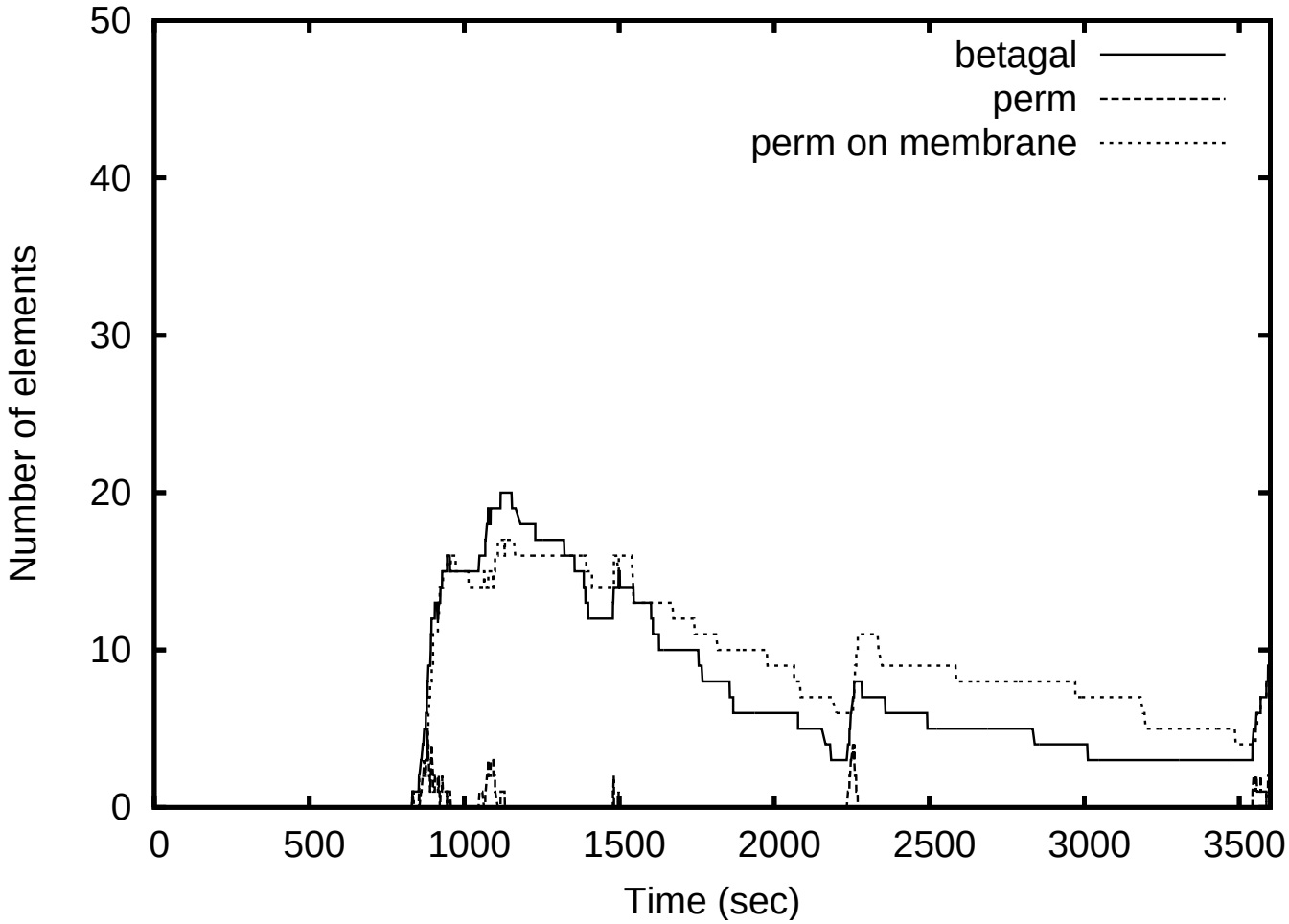
$$repr \overset{0.002}{\mapsto} \epsilon \tag{S17}$$

$$Irna \overset{0.01}{\mapsto} \epsilon \tag{S18}$$

$$Rna \overset{0.01}{\mapsto} \epsilon \tag{S19}$$

$$RLACT \overset{0.002}{\mapsto} LACT \tag{S20}$$

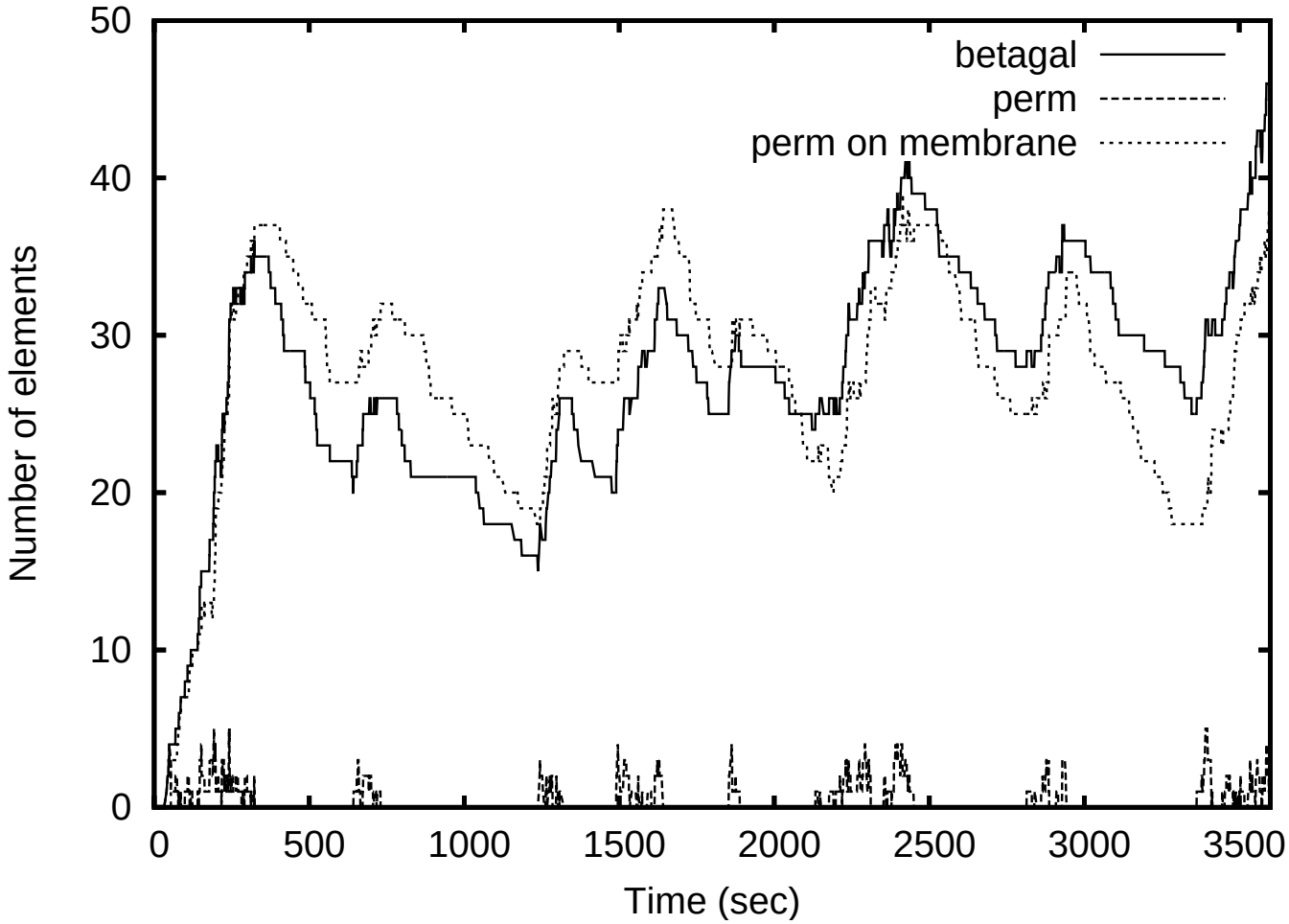
Simulation results (1)



Production of enzymes in the absence of lactose

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

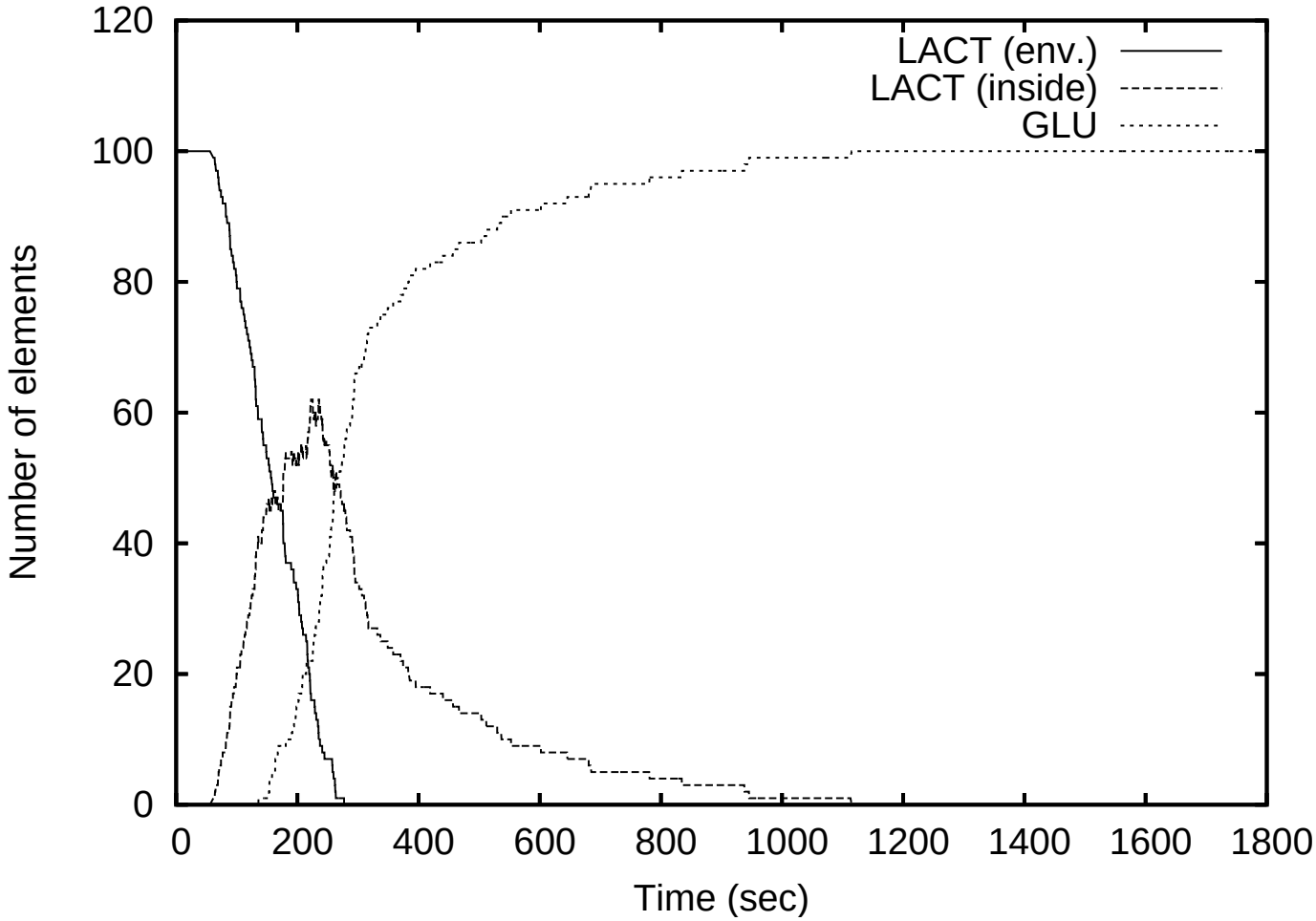
Simulation results (2)



Production of enzymes in the presence of lactose

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

Simulation results (3)



Degradation of lactose into glucose

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

Outline of the talk

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:

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

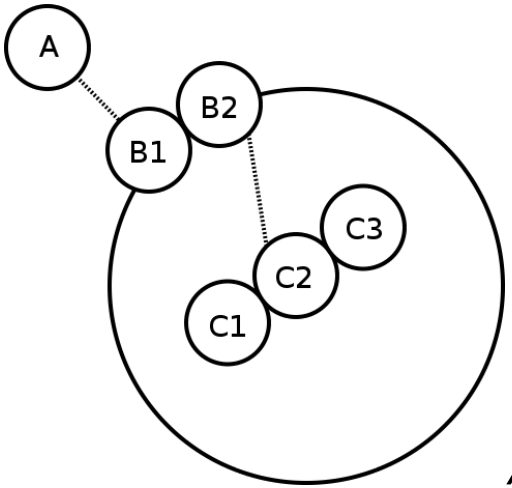
where a is a generic element of $\mathcal{E}$, and n is a natural number.

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

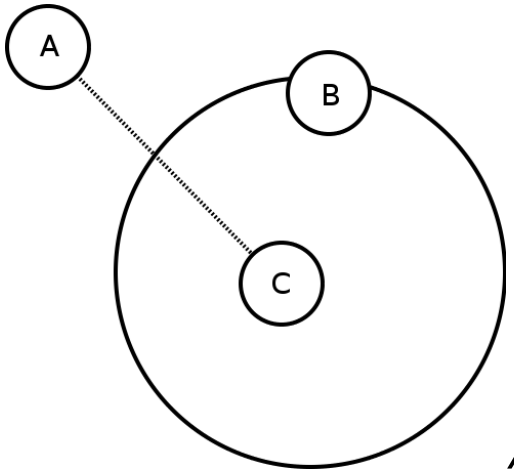
$$\begin{array}{rcl} P & ::= & SP \mid (SP)^L \rfloor P \mid P \mid P \mid X \\ SP & ::= & \epsilon \mid a \mid a^n \mid SP \cdot SP \mid \widetilde{x} \mid x \mid x^n \end{array}$$

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.

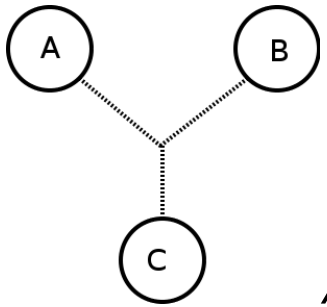
Well–formedness of LCLS terms and patterns (1)



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



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



$$A^1 \mid B^1 \mid C^1$$ ✗

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 occurs either zero or two times

Type system for well–formedness:

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

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

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

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

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

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 $\left(b\right)^L \rfloor a$ produces $\left(b\right)^L \rfloor a^1$

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

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

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

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

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

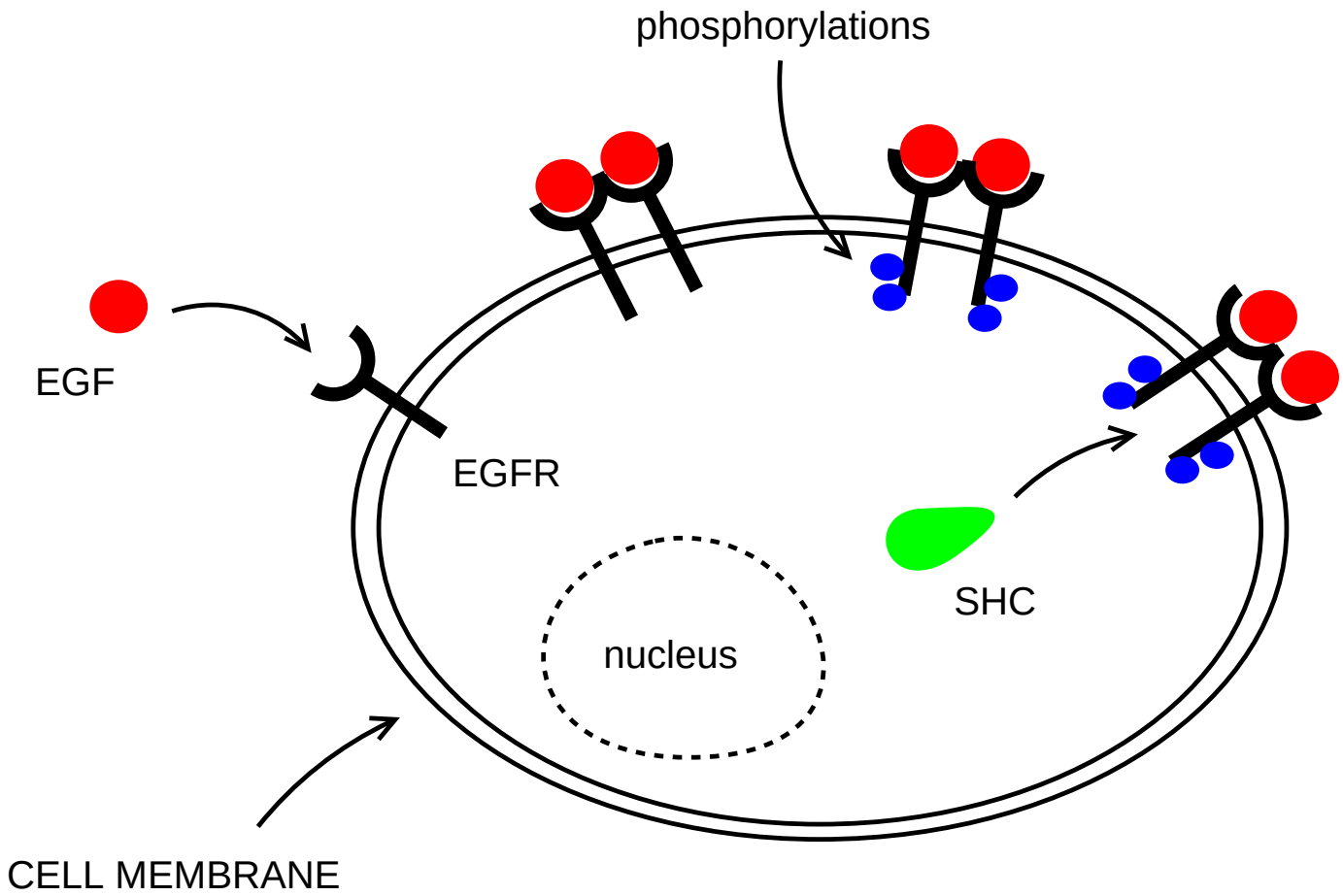
where α is link renaming, $L(T)$ the set of links occurring twice in the top level compartment of T

Main theoretical result

Theorem (Subject Reduction)
Given a set of well–formed rewrite rules $\mathcal{R}$ and a well–formed term T

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

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

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

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

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

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

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

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

Outline of the talk

Current and future work

We developed a stochastic simulator based on Stochastic CLS

- currently, we are developing an intermediate language for stochastic simulation of biological systems (sSMSR)
- high level formalisms (Stochastic CLS, π–calculus, etc...) can be translated into sSMSR
- we plan to develop analysis and verification techniques for sSMSR

In order to model cell division and differentiation, tissues, etc...

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

We are translating Kohn's Molecular Interaction Maps into CLS

Moreover:

- we plan to study other behavioural equivalences (traces, testing, ...)
- we plan to use CLS to study (in collaboration with biologists) retinal cell development 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. **Bisimulations in Calculi Modelling Membranes**. Formal Aspects of Computing, in press.
- R. Barbuti, G. Caravagna, A. Maggiolo-Schettini and P. Milazzo. **An Intermediate Language for the Simulation of Biological Systems**. Int. Workshop From Biology To Concurrency and back (FBTC'07), ENTCS 194 (3), pages 19–34, 2008.
- 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 4860, pages 54–76, Springer, 2007.
- 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. **A Calculus of Looping Sequences for Modelling Microbiological Systems**. Fundamenta Informaticae, volume 72, pages 21–35, 2006.