Probabilistic Modelling and Verification of Biological Systems

Paolo Milazzo

milazzo@di.unipi.it

Ph.D. Thesis Proposal
Pisa – February 14, 2005

Biological systems

Biology is "the science that studies living organisms". It includes:

- the analysis of molecular interactions at the level of proteins, enzymes, etc;
- the study of cells and tissues;
- the study of the origin, development and distribution of animals and plants.

Common characteristics of almost all the fields of biology:

- systems composed by a huge number of (often simple) interactive elements;
- systems exhibit very complex overall behaviors.

These characteristics suggest the application to biology of models originally developed to describe concurrent interactive software systems.

Modelling biological systems

Advantages of modelling biology with formalisms for concurrent interactive systems:

- systems can be described *precisely*: it is possible to compute all the reachable states;
- systems can be described *compositionally*: once the invidual behaviors of some subsystems has been understood, it is possible to predict the behavior of the whole system;
- *simulators* can be developed;
- *automatic analysis techniques* for systems of software components can be applied or adapted to verify properties of biological systems.

At the moment, the main application fields are biochemistry and cellular biology.

Qualitative and quantitative aspects

In biological systems there are both qualitative and quantitative aspects to consider:

qualitative aspects are related to state dependent properties (such as reachability);

quantitative aspects are related to time and space dependent properties.

Aim of the proposed thesis:

> *to develop formal models and verification techniques for biological systems by considering both qualitative and quantitative aspects.*

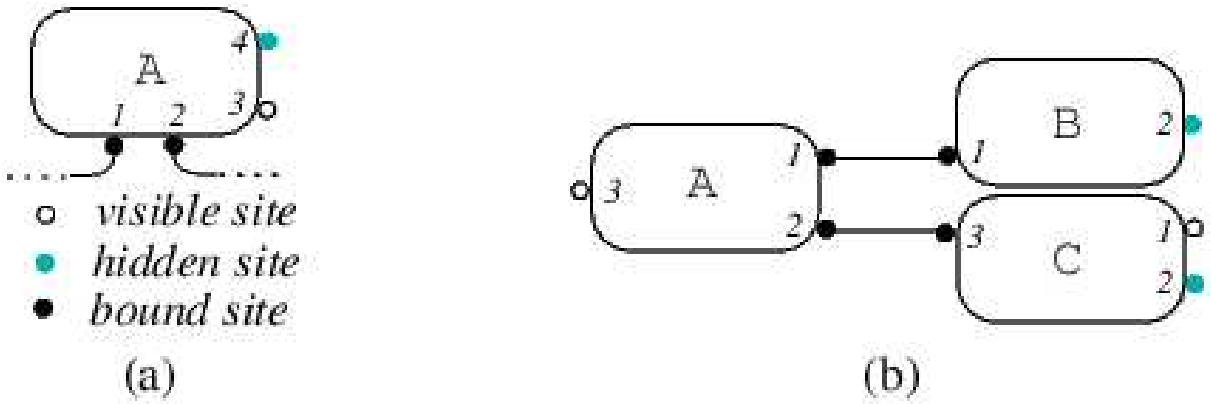
Most of the existing models are based on the π–calculus process algebra.

Examples of models: the κ–calculus

Developed by Danos and Laneve in 2003, the κ–calculus

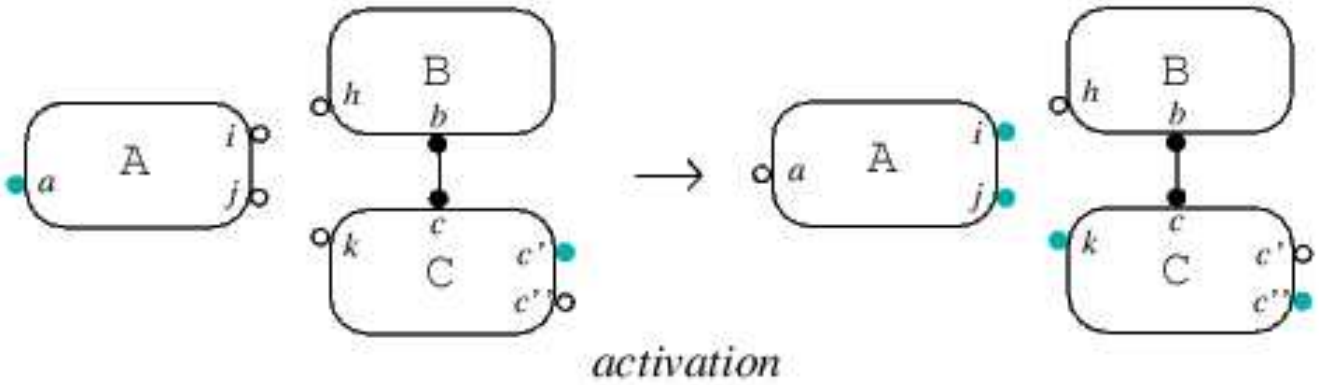
- describes formally *protein–protein interactions*;
- is enriched with a very intuitive graphical notation;
- has been encoded into the π–calculus.

In the κ–calculus proteins (a) and complexes (b) are represented by graphs with sites:

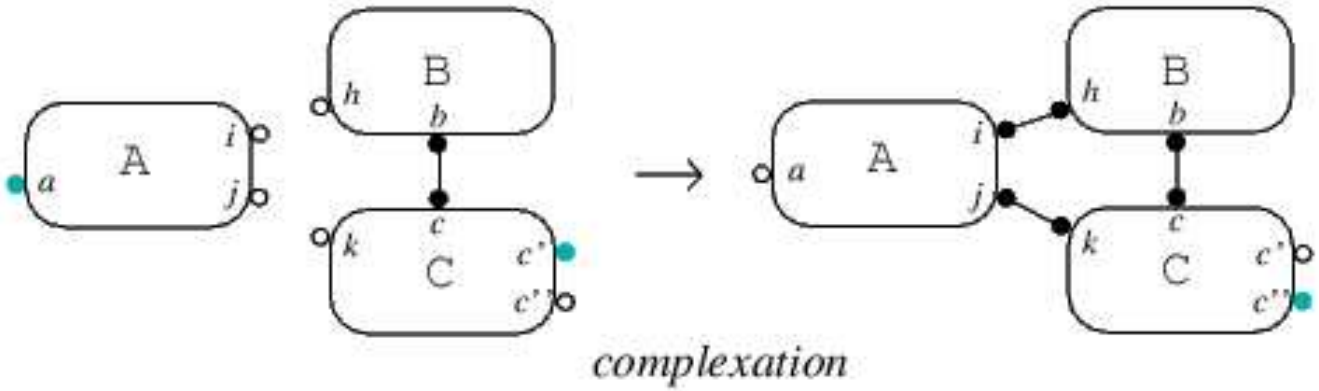


Examples of models: the κ–calculus

Reactions (protein–protein interactions) are represented by graph–rewriting rules:



activation



complexation

Examples of models: BioAmbients

Developed by Regev, Panina, Silverman, Cardelli and Shapiro in 2003, the BioAmbients calculus

- is inspired by both the π–calculus and the Mobile Ambients calculus;
- describes biochemical systems with *compartments* (such as membranes).

Example in BioAmbients of interaction of a cell with a vescicle containing a molecule:

$$\texttt{cell[accept } v \texttt{ | golgi[merge+}g\texttt{.PROTEIN]] |}$$
$$\texttt{vesc[enter } v\texttt{.merge-}g\texttt{.MOLECULE]}$$
$$\downarrow$$
$$\texttt{cell[vesc[merge-}g\texttt{.MOLECULE] |}$$
$$\texttt{golgi[merge+}g\texttt{.PROTEIN]]}$$
$$\downarrow$$
$$\texttt{cell[golgi[MOLECULE | PROTEIN]]}$$

Examples of models: the stochastic π–calculus

Regev and Shapiro in 2001 proposed to describe (*quantitatively*) molecular systems as processes in the stochastic π–calculus.

Inspired by Gillespie's stochastic algorithm for simulating chemical reactions (1977).

- *Reactions* are represented by *communications* on π–calculus channels;
- a *rate constant* is associated with each communication channel;
- the time of the next communication is *exponentially distributed* with the rate of the channel as parameter.

We proposed an alternative to Gillespie's algorithm that considers discrete time steps (to appear).

Discussion on biological systems and models

Most of the formal models for biology deal with molecular and cellular systems.
Common characteristics of many biological systems (not only molecular ones) are:

- there is a *huge number of instances* of a *few kinds* of elements;
- interactions occur among *a small* number (often couples) of elements;
- interactions involving different elements may occur *concurrently*.

Molecular systems are an example, another example are populations of living organisms.

Populations and sympatric speciation

"Speciation" is the separation of one species into two (such as humans and chimpanzees)

- easy to explain if the two populations live in separate environments (*allopatric speciation*);
- more difficult to explain if they live in the same environment (*sympatric speciation*).

Some hypothesis on the reason of sympatric speciation have been formulated by physicists and biologists, for instance *sexual selection*.

This is another example of biological system.

Multiset Rewriting (MSR)

The Multiset Rewriting (MSR) model is a simple formalism that can be used to describe biological systems.

In the MSR model:

- *multisets are states* of a computation;
- *transitions* between states are performed by applying *rewriting rules*.

Example of multiset rewriting rule: $\{A, A\} \rightarrow \{B\}$

Example of rule application:
$\{A, A, A, B\}$ is rewritten into $\{A, B, B\}$

Multiset Rewriting (MSR)

Formal definitions: let Σ be an alphabet for multisets.

Definition (MSR System). A *MSR System* is a pair $(M, \mathcal{R})$ where M is a multiset of elements in Σ and $\mathcal{R} = \{M_1 \to M_1', \ldots, M_n \to M_n'\}$ is a finite set of rewriting rules.

At each step one rule in $\mathcal{R}$ is chosen (non deterministically) and it is applied to M.

Definition (Semantics). The *semantics of MSR* is a transition system in which states are MSR systems and transitions are described by the following inference rule:

$$\frac{M_i \to M_i' \in \mathcal{R} \quad M_i \subseteq M}{(M, \mathcal{R}) \to ((M \setminus M_i) \cup M_i', \mathcal{R})}$$

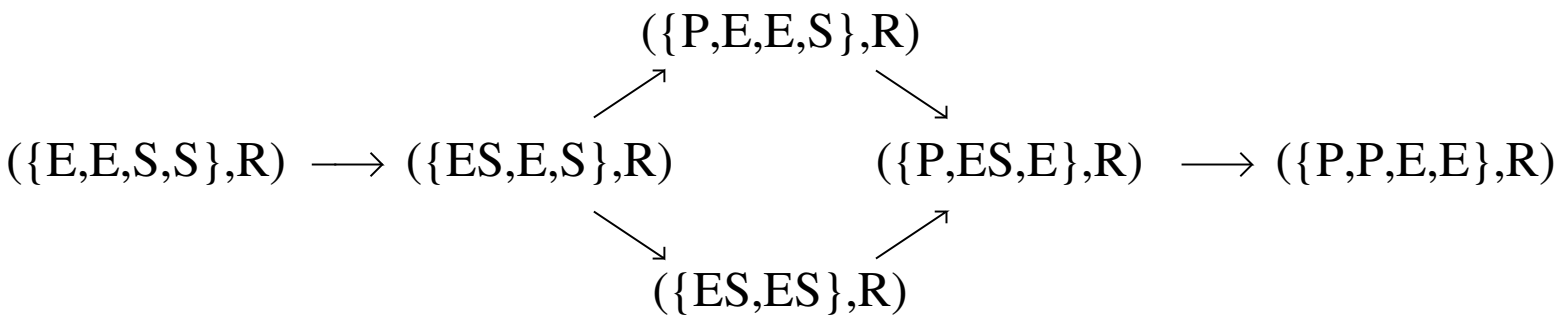
A simple example from biochemistry

Irreversible single substrate reaction catalyzed by enzyme E:

$$\mathcal{R} = \{\{S, E\} \rightarrow \{SE\}, \{SE\} \rightarrow \{P, E\}\}$$

(S = substrate, P = product)

Initial system = $(\{E, E, S, S\}, \mathcal{R})$. We obtain the following transition system:



Quantitative extensions of MSR

Non determinism in MSR makes this model suitable for describing only qualitative (time independent) aspects of biological systems.

We introduce three extensions of MSR in which rules are applied with probabilities (representing speeds of events):

- The *Stochastic Multiset Rewriting (SMSR)*, in which the times of rule applications are exponentially distributed random variables;
- The *Probabilistic Multiset Rewriting (PMSR)*, in which at each step each rule can be applied with a probability;
- The *Parallel Probabilistic Multiset Rewriting (PPMSR)*, in which at each step a subset of the rule is chosen and applied with a probability.

Stochastic Multiset Rewriting (SMSR)

The SMSR model:

- is aimed at studying systems in which the duration of each action is *exponentially distributed*;
- extends MSR rules with *stochastic rate functions*.

Example of stochastic rewriting rule: $\{A, A\} \rightarrow_f^S \{B\}$.

Example of rule application to multiset $\{A, A, A, B\}$:

- there is a probability p that the result is $\{A, B, B\}$ and the elapsed time is τ;
- p is given by an exponential distribution with parameters $f(\{A, A, A, B\})$ and τ.

Stochastic Multiset Rewriting (SMSR)

The probability of application of a rule $R_i \in \mathcal{R}$ to the multiset M with a duration of τ time units is:

$$P(M, \tau, i) = \exp\left(-\sum_{j=1}^{n} f_j(M)\tau\right) \cdot \frac{f_i(M)}{\sum_{j=1}^{n} f_j(M)}$$

Definition (Semantics). The *semantics of SMSR* is a probabilistic transition system in which states are pairs $\langle (M, \mathcal{R}), t \rangle$, where $(M, \mathcal{R})$ is a SMSR Systems and t is a clock, and transitions are described by the following inference rule:

$$\frac{M_i \rightarrow^{S}_{f_i} M_i' \in \mathcal{R} \quad M_i \subseteq M}{\langle (M, \mathcal{R}), t \rangle \xrightarrow{P(M,\tau,i)} \langle ((M \setminus M_i) \cup M_i', \mathcal{R}), t + \tau \rangle}$$

Probabilistic Multiset Rewriting (PMSR)

The PMSR model:

- is aimed at studying systems in which actions are performed with a *speed*;
- extends MSR rules with *probability functions*.

Example of probabilistic rewriting rule: $\{A, A\} \rightarrow_f \{B\}$.

Discrete time steps:

- a time interval Δt has to be chosen such that *each speed multiplied by* Δt *becomes smaller than one*;
- the speed multiplied by Δt is used as probability of a rule;
- at each step a rule is chosen randomly, and it is applied with its probability.

Probabilistic Multiset Rewriting (PMSR)

The probability that $R_i \in \mathcal{R}$ is applied to the multiset M is

$$P(M, i) = \frac{1}{n} \cdot f_i(M)$$

and the probability that no rules are applied is

$$P_0(M) = 1 - \sum_{i=1}^{n} P(M, i)$$

where n is the number of rules and the time spent is Δt.

Definition (Semantics). The semantics of PMSR Systems is the probabilistic transition system in which states are PMSR Systems and transitions are described by the following inference rules:

$$\frac{M_i \rightarrow_{f_i} M_i' \in \mathcal{R} \quad M_i \subseteq M}{(M, \mathcal{R}) \xrightarrow{P(M,i)} ((M \setminus M_i) \cup M_i', \mathcal{R})} \qquad \frac{}{(M, \mathcal{R}) \xrightarrow{P_0(M)} (M, \mathcal{R})}$$

Parallel Prob. Multiset Rewriting (PPMSR)

The PPMSR model:

- is based on the same ideas of PMSR;
- every step is performed *by applying simultaneously a subset of the rules* chosen with their probabilities.

Example of probabilistic rewriting rule: $\{A, A\} \rightarrow_f^P \{B\}$.
Problems:

- more than one subset of rules may rewrite M into M';
- some rules could not be applied simultaneously to M, for instance:

$$M = \{A, B\} \qquad \mathcal{R} = \{\{A, B\} \rightarrow_{f_1}^P \{C\}, \{A\} \rightarrow_{f_2}^P \{D\}\}$$

Parallel Prob. Multiset Rewriting (PPMSR)

It is possible to compute the probability $P(M, M', \mathcal{R})$ of M to be rewritten into M' in one step.

When the two described problems do not occur, it holds:

$$P(M, M', \mathcal{R}) = \widetilde{f_1}(M) \times \ldots \times \widetilde{f_n}(M)$$

where $\widetilde{f_i} = f_i$ if R_i is applied, $\widetilde{f_i} = 1 - f_i$ otherwise.

Definition (Semantics). The semantics of PPMSR is a probabilistic transition system in which states are PPMSR Systems and transitions are described by the following inference rule:

$$\frac{}{(M, \mathcal{R}) \xrightarrow{P(M,M',\mathcal{R})} (M', \mathcal{R})}$$

Comparing SMSR, PMSR and PPMSR

In the particular case of biochemical systems, the probabilities of PMSR approximates the ones of SMSR (to appear).

We believe that, given the same system $(M, \mathcal{R})$ in PMSR and PPMSR, the average number of applications of each rule is the same in both cases.

A deeper comparison is a future work.

Modelling molecular biology

Usual representation of chemical reactions:

$$\ell_1 S_1 + \ell_2 S_2 \underset{k_{-1}}{\overset{k}{\rightleftharpoons}} \ell_3 S_3 + \ell_4 S_4$$

Example:

$$2A \underset{k_{-1}}{\overset{k}{\rightleftharpoons}} B$$

For the *law of mass action* the speed of production of B is

$$\frac{dB_+}{dt} = k[A]^2$$

and the speed of destruction of B is

$$\frac{dB_-}{dt} = k_{-1}[B]$$

Molecular biology as PMSR

The idea is to choose a small time interval Δt such that

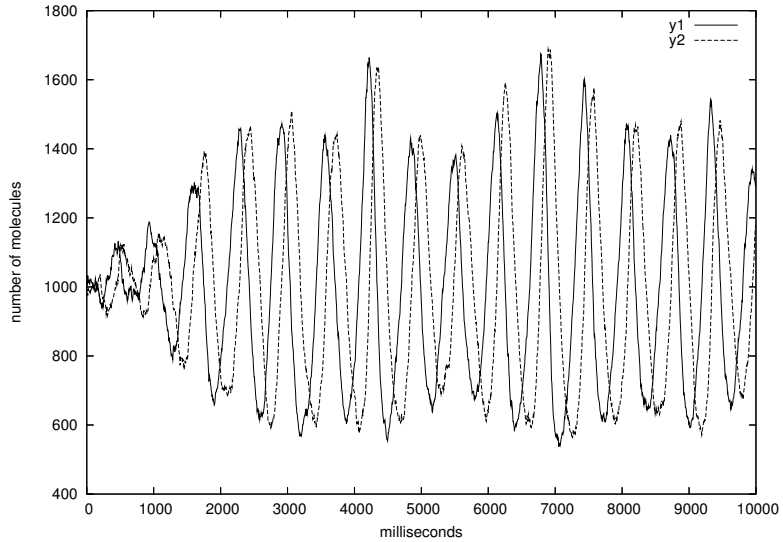
$$0 < k[S_1]^{\ell_1} \cdots [S_\rho]^{\ell_\rho} \Delta t \leq 1 \,.$$

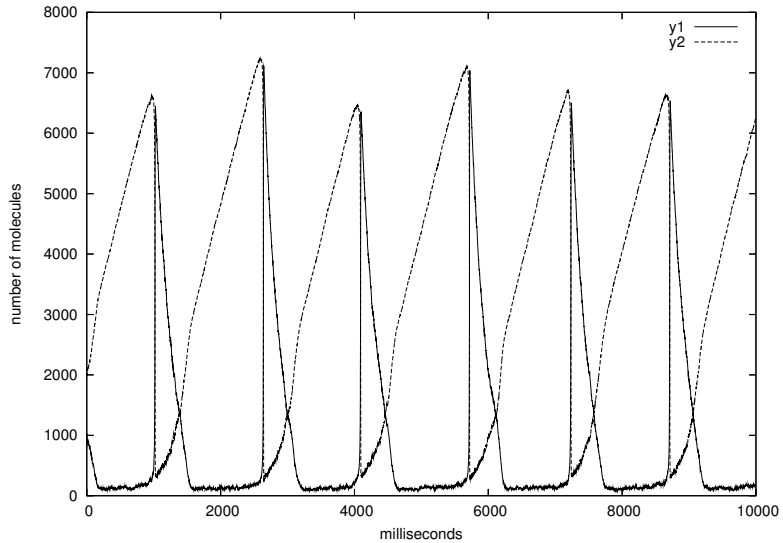
and then to use such multiplication as probability function, namely:

$$f(M) = \begin{cases} k[S_1]^{\ell_1} \cdots [S_\rho]^{\ell_\rho} \Delta t & \textit{if the rule can be applied} \\ 0 & \textit{otherwise} \end{cases}$$

We implemented a simulator based on this idea.

Lotka and Brusselator reactions





Lotka reactions

$$Y_1 \xrightarrow{k_1} 2Y_1$$

$$Y_1 + Y_2 \xrightarrow{k_2} 2Y_2$$

$$Y_2 \xrightarrow{k_3} Z$$

Brusselator reactions

$$X \xrightarrow{k_1} X + Y_1$$

$$Y_1 \xrightarrow{k_2} Y_2$$

$$2Y_1 + Y_2 \xrightarrow{k_3} 3Y_1$$

$$Y_1 \xrightarrow{k_4} Z$$

Sorbitol Dehydrogenase

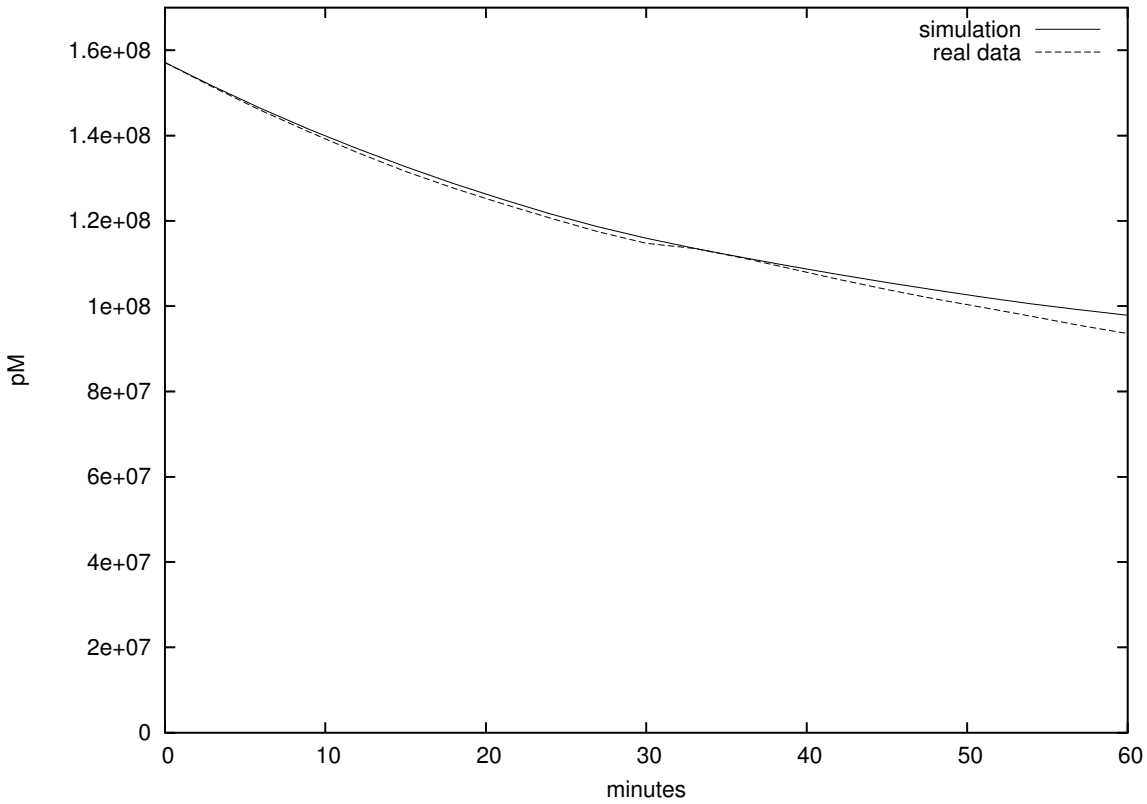
Real enzymatic activity involving Fructose and Sorbitol.

$$E + NADH \underset{k_2}{\overset{k_1}{\rightleftharpoons}} ENADH \qquad E \xrightarrow{k_7} E_i$$

$$ENADH + F \underset{k_4}{\overset{k_3}{\rightleftharpoons}} ENAD^+ + S \qquad ENAD^+ \underset{k_6}{\overset{k_5}{\rightleftharpoons}} E + NAD^+$$

$$\begin{aligned}
k_1 &= 6.2 \times 10^{-6}\, s^{-1} pM^{-1} \\
k_2 &= 33\, s^{-1} \\
k_3 &= 2.2 \times 10^{-9}\, s^{-1} pM^{-1} \\
k_4 &= 7.9 \times 10^{-9}\, s^{-1} pM^{-1} \\
k_5 &= 227\, s^{-1} \\
k_6 &= 6.1 \times 10^{-7}\, s^{-1} pM^{-1} \\
k_7 &= 1.9 \times 10^{-3}\, s^{-1}
\end{aligned}$$

Sorbitol Dehydrogenase – simulation 1



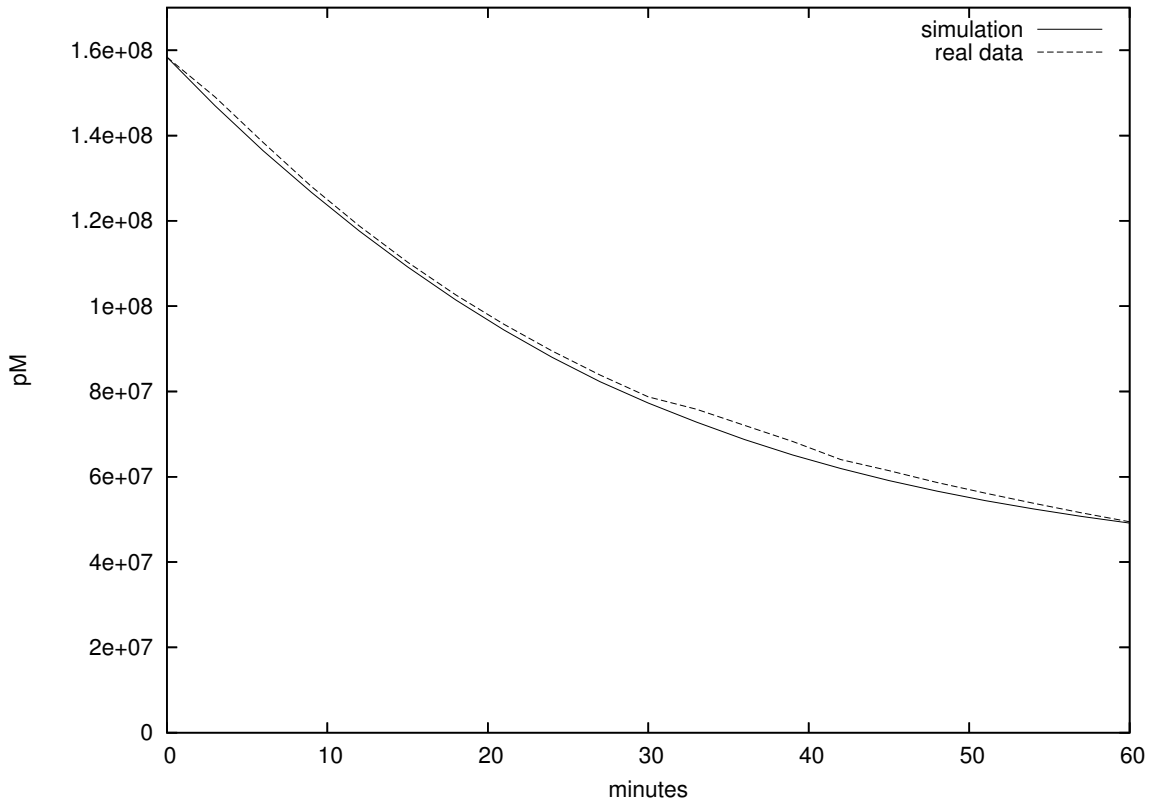
simulation (solid line) vs experiments (dashed line)

Initial (picomolar) concentrations

$$[E] = 210 \qquad [F] = 4 \times 10^{11} \qquad [NADH] = 1.6 \times 10^{8}$$

$$[S] = [NAD^{+}] = [ENADH] = [ENAD^{+}] = 0$$

Sorbitol Dehydrogenase – simulation 2



simulation (solid line) vs experiments (dashed line)

Initial (picomolar) concentrations

$$[E] = 430 \qquad [F] = 4 \times 10^{11} \qquad [NADH] = 1.6 \times 10^{8}$$

$$[S] = [NAD^{+}] = [ENADH] = [ENAD^{+}] = 0$$

Future work (on specification of systems)

About SMSR, PMSR and PPMSR:

- compare the three models precisely;
- investigate their expressiveness.

Moreover, we would like to develop an algebraic formalism (a calculus) for biochemical systems in order to study *compositional aspects* of systems.

New case studies:

- enzymatic cascades;
- sympatric speciation.

Future work (on verification of systems)

We used the probabilistic model checker PRISM to verify simple properties of the Sorbitol Dehydrogenase (CS&P'04, to apper in Fundamenta Informaticae).

Model Checking is a technique for deciding whether a property is satisfied by a finite state machine model. It consists in:

- a model description language;
- a language for properties (tipically a temporal logic);
- an efficient state exploration engine.

We would like to develop efficient model checking techniques for (general) biological systems.

Future work (on verification of systems)

Examples of typical biochemical properties that can be verified by a probabilistic model checker are:

- Given a solution S, there exists a pathway for synthesizing a molecule m.
- A molecule m can reach a concentration greater than τ with probability p.

We believe that with our discrete time models it is possible to define model checking based techniques for answering questions like:

- Given an expected behavior of a solution (in terms of a temporal logic formula), which should be the kinetic constants of the (given) involved reactions?

Future work (on implementations)

We would like to improve our simulator:

- by making it more efficient;
- by implementing a user friendly graphical interface.

Conclusions

The aim of the proposed thesis is to develop formal models and verification techniques for biological systems.

In this talk we have:

- discussed the application to biology of formal models for concurrency;
- presented three extensions of the MSR model for quantitatively describe biological systems;
- shown the results of simulation of biochemical solutions described as PMSR systems;
- proposed as future work the definition of a calculus for biochemical systems and the development of new model checking techniques.

References

R. Barbuti, A. Maggiolo–Schettini, P. Milazzo and A. Troina. **An alternative to Gillespie's algorithm for simulating chemical reactions**. 3rd International Workshop on Computational Methods in Systems Biology (CMSB'05). April, 2005 (to appear).

R. Barbuti, S. Cataudella, A. Maggiolo–Schettini, P. Milazzo and A. Troina. **A probabilistic model for molecular systems**. To appear in *Fundamenta Informaticae*. A preliminary version appeared in Proc. of Concurrency Specification & Programming (CS&P'04). September, 2004.