Quantitative analysis of biochemical signalling pathways

[Extended Abstract]

Jane Hillston
Laboratory for Foundations of Computer Science
University of Edinburgh
Edinburgh, UK
jane.hillston@ed.ac.uk

ABSTRACT

This paper gives a short introduction to some of the issues of modelling biochemical signalling pathways in the context of systems biology.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences; G.3 [**Mathematics of Computing**]: Probability and Statistics; G.1.7 [**Mathematics of Computing**]: Numerical Analysis—*ordinary differential equations*

Keywords

Systems biology, biochemical signalling pathways, stochastic process algebra

1. INTRODUCTION

Recent biological advances mean that much more is now known about the components within cells and the interactions between them. As a consequence a new endeavour within the field of biology, *systems biology* has emerged. In this approach the objective is to understand the processes by which a cell achieves its many functions. This is in contrast to much of the preceding work in cell biology which has been focussed on identifying the components of the cell and categorising the role that they play. Within systems biology modelling plays a crucial role, allowing biologists to develop hypotheses about how the behaviour they observe arises. This is shown schematically in Figure 1. In a wet lab the biologist conducts experiments on a natural system and makes observations of biological phenomena (proliferation, movement or death for example). In order to understand the observations a formal model is constructed and used to make inferences about the underlying biological system that gave rise to the phenomenon. This leads to a hypothesis about how the natural system works, which in turn is tested by further experimentation.

SMCTools '07, October 26, 2007, Nantes, France

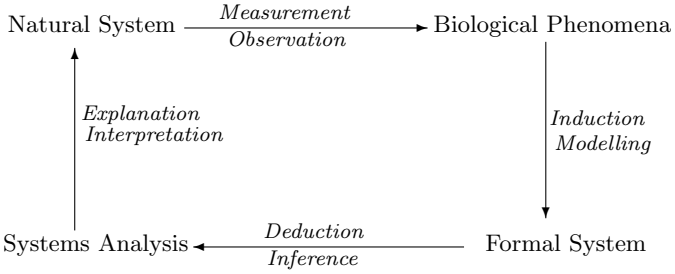

Figure 1: Schematic view of systems biology

Previously the formal system in which the biological system was modelled has most often been a mathematical representation, predominantly a set of nonlinear ordinary differential equations. However, in the last few years there has been some exploration of the use of formal modelling techniques, such as those used in theoretical computer science, for constructing models of biological processes. In particular formalisms previously used for Markovian-based performance analysis of computer systems have been applied to a number of systems [8, 19, 2, 3, 11, 12].

2. MODELLING PATHWAYS

Systems biology can be applied to biological systems at all scales from intra-cellular to complete organisms or populations of organisms. In this paper we will focus on intracellular systems and particularly *signal transduction* pathways. Here the entities involved are biochemical *species*, most commonly proteins. In signal transduction pathways external stimuli initiate messages that are carried through a cell via a cascade of biochemical reactions. The message is instigated by a molecule attaching to a receptor on the cell membrane, and is propagated through a series of protein accumulations. Increasing protein concentration broadcasts the information that an event has occurred. The message is "received" by a concentration-dependent response or reaction. This is illustrated in Figure 2.

In a signal transduction network, the delay between events is determined by the delay while signal molecule concentrations accumulate or decline sufficiently to trigger the next reaction. The accumulation of protein is a stochastic process affected by several factors in the cell (temperature, pH, etc.). In the mid 1970s Gillespie used Newtonian physics and thermodynamics to arrive at a stochastic model of biochemical reactions based on a form often termed the *propensity func-*

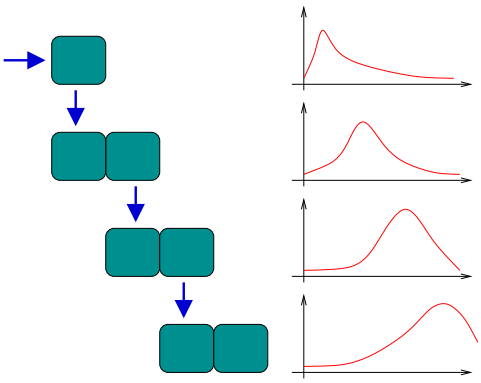

Figure 2: Cell signalling

tion that gives the probability a_μ of reaction μ occurring in time interval $(t, t + \mathrm{d}t)$.

$$a_\mu \mathrm{d}t = h_\mu c_\mu \mathrm{d}t$$

where the M reaction mechanisms which are at play are given an arbitrary index μ $(1 \leq \mu \leq M)$, h_μ denotes the number of possible combinations of reactant molecules involved in reaction μ, and c_μ is a stochastic rate constant [9]. This rate constant c_μ is dependent on the radii of the molecules involved in the reaction, and their average relative velocities — a property that is itself a direct function of the temperature of the system and the individual molecular masses. These quantities are basic chemical properties which for most systems are either well known or easily measurable. Thus, for a given chemical system, the propensity functions, a_μ can be easily determined.

The stochastic formulation proceeds by considering the *grand probability function* $\Pr(\mathbf{X}; t)$. This is the probability that at time t in the volume V there will be X_i molecules of species S_i, where $\mathbf{X} \equiv (X_1, X_2, \ldots X_N)$ is a vector of molecular species populations. By considering a discrete infinitesimal time interval $(t, t + \mathrm{d}t)$ in which either 0 or 1 reactions occur we see that there exist only $M + 1$ distinct configurations at time t that can lead to the state $\mathbf{X}$ at time $t + \mathrm{d}t$.

$$\begin{aligned}
&\Pr(\mathbf{X}; t + \mathrm{d}t) \\
&\quad = \Pr(\mathbf{X}; t)(1 - \sum_{\mu=1}^{M} a_\mu(\mathbf{X})\mathrm{d}t) \\
&\quad + \sum_{\mu=1}^{M} \Pr(\mathbf{X} - \mathbf{v}_\mu; t)(\sum_{\mu=1}^{M} \Pr(\mathbf{X} - \mathbf{v}_\mu; t) a_\mu(\mathbf{X} - \mathbf{v}_\mu)\mathrm{d}t)
\end{aligned}$$

where $\mathbf{v}_\mu$ is a *stoichiometric vector* defining the result of reaction μ on state vector $\mathbf{X}$, i.e. $\mathbf{X} \to \mathbf{X} + \mathbf{v}_\mu$ after an occurrence of reaction μ.

The Chapman-Kolmogorov equation for this system is known as the *Chemical Master Equation*:

$$\frac{\partial \Pr(\mathbf{X}; t)}{\partial t} = \sum_{\mu=1}^{M} a_\mu(\mathbf{X} - \mathbf{v}_\mu) \Pr(\mathbf{X} - \mathbf{v}_\mu; t) - a_\mu(\mathbf{X}) \Pr(\mathbf{X}; t)$$

The Chemical Master Equation can be solved analytically for only a few very simple systems, and numerical solutions are usually prohibitively difficult. Thus this CTMC is generally analysed using Gillespie's *stochastic simulation algorithm* or its variants [10].

However most models developed in systems biology do not consider the stochastic nature of the reactions which underlie the process, but are instead based on deterministic assumptions, representing the system in terms of continuously varying concentrations of species in a system of nonlinear ordinary differential equations (ODEs). Furthermore modelling biochemical reactions using deterministic rate laws has proven extremely successful for many years. This deterministic approach has at its core the *law of mass action*, an empirical law giving a simple relation between reaction rates and molecular component concentrations. Given knowledge of initial molecular concentrations, the law of mass action provides a complete picture of the component concentrations at all future time points.

These two approaches can seem to be in opposition. Stochastic simulation uses a discrete state space, stochastic model based on the individual elements in the system, molecules. Models can be computationally expensive to solve and many trajectories are needed in order to get statistically significant results. In contrast, ODE approaches use a continuous state space, deterministic model based on the population view of the system, distinguishing populations of molecules in terms of concentrations. Many efficient solvers for ODEs are available and only one solution is needed due to the deterministic nature of the model. It is known that when populations are large, i.e. there are large number of molecules, stochastic simulation tends to the same results as the ODE models. So these should be viewed as alternatives, each having a valuable role to play in appropriate circumstances.

3. SPA MODELS

Work on applying formal system description techniques from computer science to biochemical signalling pathways was initially stimulated by [11, 20, 19]. In particular, the desire to incorporate the stochastic nature of biochemical reactions has led many researchers to examine the use of formalisms previously used for performance analysis, such as stochastic Petri nets, stochastic process algebras and G-networks. Following the influential work by Regev *et al.* [20, 19], there has been much work in which the stochastic π-calculus is used to model biological systems, for example [6, 17, 18]. This work is based on a correspondence between molecules and processes. Each *molecule* in a signalling pathway is represented by a component in the process algebra representation. The local states of the component correspond to the physical changes which a molecule will undergo in the course of biochemical reactions (e.g. phosphorylation, ubiquitination and complex formation). Thus, if a complex C is formed from molecules A and B, two process algebra components A and B will interact ("communicate") and one will evolve into a C, the other becoming null. In order to represent a system with populations of molecules, many copies of the process algebra components are needed. This leads to underlying CTMC models with enormous state spaces. Even when the symmetries in the model are exploited to carry out aggregation the only possible solution technique is generally simulation based on Gillespie's algorithm [9].

Recent work on PEPA has investigated a more abstract way of mapping biochemical signalling pathways into a process algebra [3]. Rather than a correspondence between

molecules and components, we have proposed a correspondence between *species* or *subpathways* and components (c.f. modelling classes rather than individual objects). Now the components in the process algebra model capture a pattern of behaviour of a whole set of molecules, rather than the identical behaviour of thousands of molecules having to be represented individually. The local states of the components now correspond to the concentrations of species which are represented in the ODEs but discretised into a number of *"levels"*. We refer to the number of levels of concentration distinguished in a model as the *granularity* of the model.

Our motivations for seeking more abstraction in process algebra models for systems biology are:

- Process algebra-based analyses such as comparing models (e.g. for equivalence or simulation) and model checking are only possible is the state space is not prohibitively large.
- The data that we have available to parameterise models is sometimes *speculative* rather than precise. This suggests that it can be useful to use *semiquantitative* models rather quantitative ones.

From such PEPA models we are able to generate underlying mathematical models, suitable for analysis, in a number of different ways. The usual semantics of PEPA gives rise to a CTMC which can be solved numerically (if state space size does not prohibit it) [13]. Here each state of the CTMC corresponds to a discrete level of concentration for each chemical species in the pathway. For PEPA models based on modelling species and their concentrations [4] showed how a model with two levels of concentration could be used to generate a set of nonlinear ODEs. Moreover, such models can also be used to derive CTMCs in which molecules are represented individually, suitable for Gillespie simulation (see Figure 3). In [14], Hillston showed how a set of nonlinear ODEs can be derived from more general PEPA models. These underlying mathematical models have different strengths offering different forms of analysis. The relationship between Gillespie-style molecular simulations and ODEs has been known for some time. Moreover recently the relationship between the CTMCs with levels of concentrations, such as arise from PEPA models, and ODEs derived from PEPA models has also been recently established.

Even within our abstract approach to modelling there are alternative ways of expressing the model [3]. We distinguish these as *reagent-centric* and *pathway-centric*. In a reagent-centric model we treat each distinct reagent or species in the pathway as a distinct component type as described above. The component definition then captures the possible reactions that the reagent may be involved in. The local states of the components correspond to differing levels of concentration and the process definition records the impact of each reaction type on the concentration of the reagent—it will either increase the concentration, moving it up a level, decrease it, moving the state down a level, or leave it unchanged. In a pathway-centric model we focus instead on the transformations which a reagent or species with non-zero initial concentration may undergo through the course of a pathway (phosphorylation, complex formation etc.). Each such subpathway is then represented as a distinct component in the model. Local states now correspond to the states which the physical entity may find itself in through the subpathway. Differing levels of concentration are represented in the global state by having differing multiples of components of a particular pathway type.

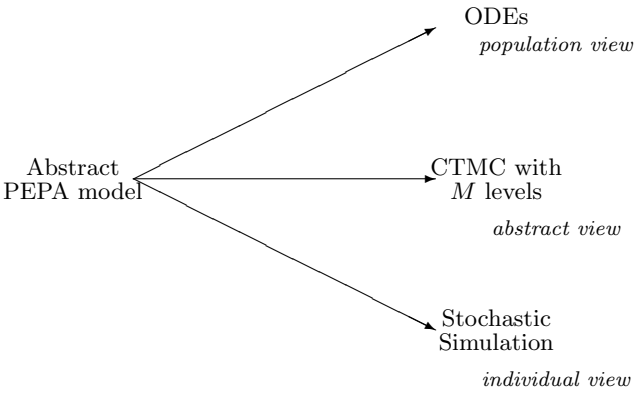

Figure 3: Alternative modelling approaches: a single PEPA description of a system may be used to derive alternative mathematical representations offering different analysis possibilities.

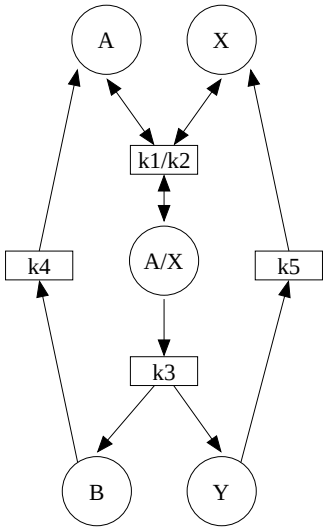

Figure 4: Small synthetic example pathway

As a small example we consider the pathway shown in Figure 4.

This is comprised of the following kinetic reactions :

$$\begin{aligned} A + X &\underset{k2}{\overset{k1}{\rightleftharpoons}} A/X \overset{k3}{\rightarrow} B + Y \\ B &\overset{k4}{\rightarrow} A \\ Y &\overset{k5}{\rightarrow} X \end{aligned}$$

We assume an initial positive concentration of reagents A and X, all other reagents initially being absent.

Reagent-centric model. Modelled in the reagent-centric style with the coarsest possible granularity, i.e. just two levels, the pathway in Figure 4 is represented by the following declarations:

$$
\begin{aligned}
A_H &\stackrel{def}{=} (k1react, k1).A_L \\
A_L &\stackrel{def}{=} (k2react, k2).A_H + (k4react, k4).A_H \\
X_H &\stackrel{def}{=} (k1react, k1).X_L \\
X_L &\stackrel{def}{=} (k2react, k2).X_H + (k5react, k5).X_H \\
A/X_H &\stackrel{def}{=} (k2react, k2).A/X_L + (k3react, k3).A/X_L \\
A/X_L &\stackrel{def}{=} (k1react, k1).A/X_H \\
B_H &\stackrel{def}{=} (k4react, k4).B_L \\
B_L &\stackrel{def}{=} (k3react, k3).B_H \\
Y_H &\stackrel{def}{=} (k5react, k5).Y_L \\
Y_L &\stackrel{def}{=} (k3react, k3).Y_H
\end{aligned}
$$

The complete model is the interaction of these components constrained by cooperation to share the appropriate actions:

$$
\begin{aligned}
(((A_H &\underset{\{k1react,k2react\}}{\bowtie} X_H) \underset{\{k1react,k2react\}}{\bowtie} A/X_L) \\
&\underset{\{k3react,k4react\}}{\bowtie} B_L) \underset{\{k3react,k5react\}}{\bowtie} Y_L
\end{aligned}
$$

More details of this style of representation can be found in [3].

Pathway-centric model. Modelled in the pathway-centric style the pathway in Figure 4 is represented by the following declarations:

$$
\begin{aligned}
A &= (k1react, \ell_1).A/X \\
A/X &= (k2react, \ell_2).A + (k3react, \ell_3).B \\
B &= (k4react, \ell_4).A \\
X &= (k1react, 1).X/A \\
X/A &= (k2react, \ell_2).X + (k3react, \ell_3).Y \\
Y &= (k5react, \ell_5).X
\end{aligned}
$$

where we have two distinct subpathways, corresponding to A and X respectively. We also need the following system equation to complete the model :

$$
\begin{aligned}
(A[n_{1_1}] \parallel A/X[n_{1_2}] \parallel B[n_{1_3}]) &\underset{\{k2react,k3react\}}{\overset{\{k1react\}}{\bowtie}} \\
&(X[n_{2_1}] \parallel X/A[n_{2_2}] \parallel Y[n_{2_3}])
\end{aligned}
$$

where having $\{k1react\}$ over the cooperation combinator denotes that sychronisation on this action type follows a *mass action* kinetics rather than the usual *bounded capacity* kinetics used in PEPA (see [7] for details). The activity rates ℓ_i are chosen to reflect the granularity of the discretisation, which is determined by the total number of copies of each subpathway type (i.e. $n_{1_1} + n_{1_2} + n_{1_3}$ and $n_{2_1} + n_{2_2} + n_{2_3}$ in this example).

4. RELATING THE MODELS

In a recent paper [7] we have investigated the relationship between the ODEs which can be derived from a PEPA description and the CTMC models which are generated by models in the pathway-centric style with increasing levels of discretisation. We can consider a sequence of CTMCs representing the system: as we increase the level of discretisation in the model we move to the next CTMC in the sequence. In [7] we show that in the limit this sequence of CTMCs converges to the same behaviour as the ODEs derived from the PEPA model (with only two levels of discretisation). This is based on an earlier result by Kurtz [15] in which the author shows that when certain conditions are satisfied a sequence of CTMCs converge to a set of ODEs.

4.1 Kurtz's Theorem

Kurtz's Theorem states that, under certain assumptions, the solutions provided by a set of ODEs can be regarded as the limit of a sequence of "pure jump" Markov processes. As a special case of this general result, Kurtz shows how to obtain the ODEs as the limit of a sequence of *density dependent* CTMCs, which model discrete numbers of elements in their different states [15]. The density dependent condition means that the rates of the CTMCs may depend on a scaled representation of states. For instance, when states represent number of individuals and are normalized with respect to volume or area, then the rates depend on population densities. Instead in the case of the Markov chains derived from PEPA models we are interested in studying their behaviour when the number of levels increases. Therefore the rates do not contain information on area or volume but on the number of levels $N + 1$.

Definition 1. A family of CTMCs is called *density dependent* if and only if there exists a continuous function $f(x, l)$, $x \in \mathbb{R}^h$, $l \in \mathbb{Z}^h$, such that the infinitesimal generators of X_N are given by :

$$
q_{k,k+l} = Nf\left(\frac{k}{N}, l\right), \quad l \neq 0
$$

with $q_{k,k+l}$ denoting an entry of the infinitesimal generator of X_N, k a numerical state vector and l a *transition vector* that contains the modifications for each state of each species (i.e. the number of copies to add or substract) when the transition is taken.

In [15] Kurtz shows that the ODE system $\frac{dX(t)}{dt} = F(X)$ defined by :

$$
F(x) = \sum_l lf\,(x, l)
$$

is the solution of the limit of X_N when N tends to infinity, in the sense that :

$$
\begin{aligned}
&\lim_{N\to\infty} \frac{X_N(0)}{N} = X(0) \quad \Longrightarrow \\
&\qquad \forall \delta > 0 \quad \lim_{N\to\infty} \mathbb{P}\left(\sup_{s\leq t} \left|\frac{X_N(s)}{N} - X(s)\right| > \delta\right) = 0
\end{aligned}
$$

The limit expresses that the probability for X_N to take a trajectory different from X tends to 0 when N tends to infinity. The result is based on the assumption that the following conditions are met :

There exists an open set $E \subset \mathbb{R}^h$ such that $X(t) \in E$ and

$$
\exists M, \forall x, y \in E \quad |F(x) - F(y)| < M\,|x - y| \tag{1}
$$

$$
\sup_{x\in E} \sum_l |l| f(x, l) < \infty \tag{2}
$$

$$
\lim_{d\to\infty} \sup_{x\in E} \sum_{|l|>d} |l| f(x, l) = 0 \tag{3}
$$

These conditions can be understood as follows:

(1) This says that the function F is Lipschitz continuous, imposing a certain degree of *smoothness* on the function;

(2) This imposes that for each transition the rate of change is bounded;

(3) This ensures that there is a bound for the whole state space which means that the impact of transitions remains bounded.

It is important to note that Kurtz's result does not tell us about the relationship between the Markov chain with N levels and the system of ODEs. However, it does tell us that in the limit, as N tends to infinity, the agreement between the Markov chain and the system of ODEs is complete, in the sense that the behaviour of the two with respect to the state variables will be identical. We can regard this as saying that for density dependent Markov chains the stochasticity is such that when there are large numbers of entities the variability balances in such a way that the process tends to a deterministic limit.

In the context of PEPA models of biochemical signalling pathways we show that Kurtz's theorem holds by proving that the CTMCs generated from PEPA models are density dependent. Furthermore, we show how the deterministic distribution obtained as the limit of the sequence of CTMCs is related to the solution of the system of ODEs derived syntactically from the corresponding PEPA model. Full details can be found in [7].

4.2 Related work

Whilst a significant body of work is developing on modelling biochemical systems with stochastic process algebras and related formalisms (e.g. [11, 20, 19, 6, 17, 18]) in most cases the modelling is carried out at a more detailed and less abstract level than the work on PEPA. Consequently, most analysis of such systems is carried out using Gillespie's stochastic simulation and similar approaches [10, 18]. To the best of our knowledge no other authors have considered the relationship between ODE and CTMC models in the context of stochastic process algebras. The original relationship between the two was established by Kurtz in 1970 [15], and considered in the context of chemical reactions in 1972 [16]. The mapping from process algebra models to ODEs has recently been considered by Cardelli [5] and Bortolussi and Policriti [1], but not the relationship with a CTMC.

5. ACKNOWLEDGMENTS

Jane Hillston is an EPSRC Advanced Research Fellow supported by grant EP/c543696/01. This work was developed in collaboration with Nil Geisweiller, Stephen Gilmore and Marco Stenico.

6. REFERENCES

[1] L. Bortolussi and A. Policriti. Stochastic concurrent constraint programming and differential equations. In *Proceedings of Fifth Workshop on Quantitative Aspects of Programming Languages*. Braga, Portugal, March 2007.

[2] M. Calder, A. Duguid, S. Gilmore, and J. Hillston. Stronger computational modelling of signalling pathways using both continuous and discrete-state methods. In *Computational Methods in Systems Biology*, pages 63–77, Trento, Italy, October 2006.

[3] M. Calder, S. Gilmore, and J. Hillston. Modelling the influence of RKIP on the ERK signaling pathway using the stochastic process algebra PEPA. *BioConcur*, 2004.

[4] M. Calder, S. Gilmore, and J. Hillston. Automatically deriving ODEs from process algebra models of signalling pathways. In *Computational Methods in Systems Biology*, pages 204–215, 2005.

[5] L. Cardelli. From processes to ODEs by chemistry, unpublished manuscript, 2006

[6] D. Chiarugi, M. Curti, P. Degano, and R. Marangoni. VICE: A VIrtual CEll. In *Computational Methods in Systems Biology*, Paris, France, apr 2004. Springer.

[7] N. Geisweiller, J. Hillston, and M. Stenico. Relating continuous and discrete PEPA models of signalling pathways. In press.

[8] E. Gelenbe. Steady-state solution of probabilistic gene regulatory networks. Submitted for publication, 2007.

[9] D. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340 – 2361, 1977.

[10] D. Gillespie and L. Petzold. Numerical simulation for biochemical kinetics. 2006.

[11] P. Goss and J. Peccoud. Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. *Proceedings of National Academy of Science, USA*, 95(12):7650–6755, jun 1998.

[12] J. Heath, M. Kwiatkowska, G. Norman, D. Parker, and O. Tymchyshyn. Probabilistic model checking of complex biological pathways. In *Computational Methods in Systems Biology*, pages 32–47, Trento, Italy, October 2006.

[13] J. Hillston. *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.

[14] J. Hillston. Fluid flow approximation of PEPA models. In *Proc. of 2nd Intl Conference on Quantitative Evaluation of Systems (QEST)*, pages 33–42, Torino, Italy, September 2005. IEEE Computer Society Press.

[15] T. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Appl. Prob.*, vol. 7:pages 49–58, 1970.

[16] T. Kurtz. The relationship between stochastic and deterministic models for chemical reactions. *Journal of Chemical Physics* vol. 57(7): pages 2976–2978, 1972.

[17] C. Kuttler and J. Niehren. Gene regulation in the pi calculus: simulating cooperativity at the lambda switch. *Transactions on Computational Systems Biology*, LNCS 4230:24–55, 2006.

[18] P. Lecca, C. Priami, C. Laudanna, and G. Constantin. A Biospi model of lymphocyte-endothelial interactions in inflamed brain venules. In *Pacific Symposium of Biocomputing*, pages 521–532, 2004.

[19] C. Priami, A. Regev, W. Silverman, and E. Shapiro. Application of a stochastic name passing process calculus to representation and simulation of molecular processes. *Information Processing Letters*, 80:25–31, 2001.

[20] A. Regev. Computational systems biology: a calculus for biomolecular knowledge, 2002.