Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

SIMULATION AND VERIFICATION FOR COMPUTATIONAL MODELLING OF SIGNALLING PATHWAYS

Marta Kwiatkowska
Gethin Norman
David Parker
Oksana Tymchyshyn

School of Computer Science
University of Birmingham
Edgbaston, B15 2TT, UK

John Heath

CRUK Growth Factor Group
School of Biosciences
University of Birmingham
Edgbaston, B15 2TT, UK

Eamonn Gaffney

School of Mathematics
University of Birmingham
Edgbaston, B15 2TT, UK

ABSTRACT

Modelling of the dynamics of biochemical reaction networks typically proceeds by solving ordinary differential equations or stochastic simulation via the Gillespie algorithm. More recently, computational methods such as process algebra techniques have been successfully applied to the analysis of signalling pathways. One advantage of these is that they enable automatic verification of the models, via model checking, against qualitative and quantitative temporal logic specifications, for example, "what is the probability that the protein eventually degrades?". Such verification is exhaustive, that is, the analysis is carried out over all paths, producing exact quantitative measures. In this paper, we give an overview of the simulation, verification and differential equation approaches to modelling biochemical reaction networks. We discuss the advantages and disadvantages of the respective methods, using as an illustration a fragment of the FGF signalling pathway.

1 INTRODUCTION

Biological signalling processes control key responses in multicellular organisms such as cell multiplication, differentiation and movement. Many modelling frameworks have been put forth to advance the scientific understanding of these complex processes. Traditionally, one assumes that the time evolution of the number (or concentration) of molecules is continuous, leading to a set of coupled ordinary differential equations (usually non-linear) called reaction rate equations. An alternative, stochastic, approach views the system as a continuous time Markov process, and admits an efficient solution via stochastic simulation (Gillespie 1977). More recently, the observation that concurrency is present in these processes has led to the adoption of process algebra approaches developed for description and analysis of complex software systems in computer science. In particular, in (Regev and Shapiro 2002, Priami et al. 2001) the stochastic π-calculus has been proposed as particularly appropriate to model the dynamics of molecular processes.

Process-algebraic approaches view systems as networks composed of concurrent, interacting molecules or molecular ensembles, and can be applied at all levels of abstraction, molecular, cellular and tissue (Regev and Shapiro 2002). The "molecule-as-computation" paradigm embodied in process calculi is very attractive, since it offers a compact notation with a minimal repertoire of computational abstractions of molecular interactions that are supported by a formal reasoning framework. Thus, one can formulate a hypothesis about a specific signalling mechanism in terms of a π-calculus process, and benefit from computer assisted reasoning via *in silico genetics*, i.e., a series of experiments on the models performed by manipulating process descriptions, for example, the removal of a protein, each of which can be validated against experimental data and prioritised according to the potential of the discovery being predicted.

The stochastic π-calculus modelling framework supports not only Monte Carlo simulation to obtain time-evolution of molecular concentrations using tools such as BioSPI (Regev and Shapiro 2004) and SPiM (Cardelli and Phillips 2004), but also formal reasoning, for example, automatic verification via model checking. With the help of techniques such as probabilistic model checking (Rutten et al. 2004), one can obtain qualitative and quantitative answers to queries such as "does this reaction always lead to degradation?", "what is the probability that the protein eventually degrades?" and "what is the expected number of complexation reactions before relocation occurs?".

Naturally, each of the modelling frameworks and analysis techniques mentioned above has advantages and disadvantages, and it is important to understand these in order to decide on their applicability for a particular modelling or experimental context. In this paper, we give an overview of the main modelling and analysis approaches for signalling pathways, and discuss their respective strengths and weaknesses using as an illustration a fragment of a complex signalling pathway, the FGF (Fibroblast Growth Factor) pathway. FGF are a family of proteins which play an important role in cell signalling, e.g., wound healing. The dynamics of the FGF pathway are complex and not yet fully understood. Aspects of the full pathway were studied elsewhere using ODEs, e.g., (Yamada et al. 2004), and process calculi (Heath et al. 2006). Other simpler pathways have also been studied using process calculi approaches, e.g., ERK (Calder et al. 2006) and MAPK (Phillips and Cardelli 2005).

2 MODELLING FRAMEWORKS

We consider the problem of modelling a mixture of molecules from N different molecular species, which can interact through a number of reactions. We assume a spatially uniform mixture in a fixed volume V at constant pressure and temperature. In this section, we distinguish between two distinct modelling approaches, based on either a continuous or a discrete time evolution of the molecules; see e.g., (Wolkenhauer et al. 2004, Gillespie 1977).

2.1 The Continuous Deterministic Approach

One approach is to approximate the number of molecules of each species in V at time t by a *continuous* function, which is justifiable for large numbers of molecules. More precisely, this measures the concentration of each species in molar units (M) which corresponds to the number of molecules per unit volume (litre) divided by Avogadro's number ($N_A = 6.022e{+}23$).

Consider for example a reversible reaction between species A and B that can become bound (AB):

$$A + B \underset{k_2}{\overset{k_1}{\rightleftharpoons}} AB$$

where k_1 ($\mathrm{M}^{-1}\mathrm{s}^{-1}$) describes the velocity of the compound formation and k_2 (s^{-1}) is the velocity of the breakdown of the complex. The values are called kinetic rates and are derived from experimental data. Using the *principle of mass action*, the change in concentration is proportional to the kinetic rate and the amount of reactant species, and therefore we can represent the time evolution of the concentration $[AB]$ of the complex AB by the reaction rate equation:

$$\frac{d[AB](t)}{dt} = k_1 \cdot [A](t) \cdot [B](t) - k_2 \cdot [AB](t)$$

The solution of the derived set of ordinary differential equation in N-dimensional space gives the required time evolution of the concentrations. There are different types of biochemical reactions, which vary in the number of reactants, the type of reaction (reversible or irreversible) and the type of reactant (e.g., enzyme/substrate). The analysis of the enzyme-catalysed reaction can be simplified by the Michaelis-Menten kinetics.

Note that, although the underlying physical interpretation involves random collisions of molecules, the ODEs predict average population levels. Therefore, the model is *deterministic*, but only with respect to a perceived average of a process that is subject to random fluctuations. In the derivation of the differential equation, we assume a large number of molecules so that a process with discrete changes can be approximated by a continuous model. Mathematically, this corresponds to approximating a difference equation with a differential equation.

2.2 The Discrete Stochastic Approach

An alternative is to take a *discrete* view of the evolution of the system, where the occurrence of a reaction between molecules corresponds to a discrete event. It is argued that this is a more accurate representation of the physical system being modelled, particularly when dealing with small numbers of molecules. The evolution of such a model is inherently *stochastic*, representing the probability that there are n molecules of the ith species at time t, for each i. This is a discrete-state time homogeneous Markov process whose states are vectors of molecule counts and state changes are dependent on stochastic constants (determined from the rate constants) and the numbers of molecules of each species.

This approach is based on the *grand probability function* $P(\mathbf{x}, t)$ – the probability that, at time t, there will be $\mathbf{x}_S$ of species S, where $\mathbf{x}$ is a vector of molecular species populations and the solution can be formulated as a set of partial differential equations, known as the *chemical master equation* (Gillespie 1977). Returning to the simple reactions above, $\mathbf{x}$ is of the form $(\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_{AB})$ (the quantities of A, B and AB), and so denoting the complexation and decomplexations reactions by 1 and 2 respectively, we have:

$$\frac{\partial P(\mathbf{x}, t)}{\partial t} = \sum_{i=1}^{2} \big(a_i(\mathbf{x}-\mathbf{v}) P(\mathbf{x}-\mathbf{v}, t) - a_i(\mathbf{x}) P(\mathbf{x}, t) \big)$$

where $a_i(\mathbf{x}){\cdot}\partial t$ is the probability of, in state $\mathbf{x}$, reaction i occurring in the interval $(t{+}\partial t)$ (and can be derived from k_i and $\mathbf{x}$), and $\mathbf{v}_i$ is the stoichiometric vector defining the result of reaction i, i.e., $\mathbf{v}_1{=}(-1,-1,1)$ and $\mathbf{v}_2{=}(1,1,-1)$.

Under the assumption of constant state-dependent rates, the underlying Markov process is a continuous-time Markov chain, which assumes exponentially distributed reaction rates; this is justified since, if collision times are small compared to the times between collisions, molecules are moving chaotically, and a constant ratio of overall collisions lead to reactions (Wolkenhauer et al. 2004). The states of the resulting Markov chain are vectors representing interacting molecules, and transitions are determined by the rates combined with concentrations and are selected according to the usual race condition.

The approach described above can be classified as *population-based* since the model represents the number of each molecular species present, and does not consider interactions between individual molecules. However, we can extend the discrete stochastic approach to an *individual-based* model, where the state and behaviour of each molecule is represented separately. This approach is desirable if, for example, the assumptions of perfect diffusion and well stirred substance are dropped, or if we are interested in the behaviour of individual components. Clearly, though, this comes with the cost of increased complexity.

3 MODELLING FORMALISMS AND LANGUAGES

We now summarise a selection of different formalisms which have been proposed for the modelling of biochemical reactions. To do so, we introduce a running example: a fragment of the earlier studied (Heath et al. 2006) Fibroblast Growth Factor (FGF) pathway.

Figure 1 shows a graphical representation of the elements of the system that we consider. Figure 2 presents the set of reactions between the elements, which can be summarised as follows. An FGF protein (molecule) can bind to an FGF receptor (FGFR). When FGF and FGFR are bound, two different residues on FGFR can become phosphorylated which, subsequently, allow the signal transducing proteins Src and Grb2 to bind to FGFR. Each of these reactions is also reversible. Finally, when Src is bound, FGFR can be relocated, along with any components bound to it. The reaction rates given in Figure 2 are based on experimental observations from the literature.

3.1 SBML

SBML (SBML 2006) is a computer-readable language based on XML for representing models of biochemical reaction networks. SBML is intended as a standardised representation of models that can be shared, manipulated and analysed using tools available in the systems biology community. Models

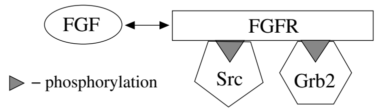


Figure 1: Graphical Representation of FGF and FGFR Interaction and Their Effect on Src and Grb2.

1: FGF binds/releases FGFR	
FGF + FGFR $\rightarrow$ FGFR:FGF	k_1=5e+8 $M^{-1}s^{-1}$
FGF + FGFR $\leftarrow$ FGFR:FGF	k_2=0.002 s^{-1}
2: Phosphorylation of FGFR (whilst FGFR:FGF)	
FGFR:FGF + FGFR1 $\rightarrow$ FGFR:FGF + FGFR1P	k_3=0.1 s^{-1}
FGFR:FGF + FGFR2 $\rightarrow$ FGFR:FGF + FGFR2P	k_4=0.1 s^{-1}
3: Dephosphorylation of FGFR	
FGFR1P $\rightarrow$ FGFR1	k_5=0.1 s^{-1}
FGFR2P $\rightarrow$ FGFR2	k_6=0.1 s^{-1}
4: Effectors bind phosphorylated FGFR	
SRC + FGFR1P $\rightarrow$ SRC:FGFR	k_7=1e+6 $M^{-1}s^{-1}$
SRC + FGFR1P $\leftarrow$ SRC:FGFR	k_8=0.02 s^{-1}
GRB2 + FGFR2P $\rightarrow$ GRB2:FGFR	k_9=1e+6 $M^{-1}s^{-1}$
GRB2 + FGFR2P $\leftarrow$ GRB2:FGFR	k_{10}=0.02 s^{-1}
5: Relocation of FGFR (whilst SRC:FGFR)	
SRC:FGFR $\rightarrow$ relocFGFR	k_{11}=1.1e-3 s^{-1}

Figure 2: Summary of the Reactions

are composed from components, which permit definition of reactant species, product species, descriptions of reaction equations using MathML expressions, and the specification of kinetic laws and parameters. Compartments are allowed, but not considered here. Figure 3 shows a fragment of the SBML description for the reactions in Figure 1, more specifically, for the first line of reaction **4**.

SBML is widely supported and facilitates interchange between different tools, e.g., it is supported by ODE tools such as SIGMOID and Cellerator (Shapiro et al. 2003), and stochastic simulation (e.g., Dizzy). Recently, automatic generation of a large fragment of the stochastic π-calculus from SBML has been implemented (Eccher 2006).

3.2 The Stochastic π-Calculus

Process calculi are formal languages for representing systems as networks of concurrent interacting processes, each operating according to explicitly given rules and combined in parallel. Such compositional descriptions of networks are compact and easy to manipulate. In order to model biochemical reactions, which occur at specified reaction rates, stochastic extensions of process calculi have been formulated. There are different dialects of calculi that differ in the synchronisation method used (e.g., channel or action-based, binary/multi-way) and types of operators.

The *stochastic π-calculus* was proposed in (Regev and Shapiro 2004) as a framework for modelling of biological processes, and a translation scheme was given for representing biochemical reactions in this formalism. The models of reaction networks induced from stochastic π-calculus process are continuous time Markov chains, and therefore the stochastic π-calculus should be viewed as a conve-

```
<listOfSpecies>
           ...
  <species id="FGFR_Ph1" initialConcentration="0".../>
  <species id="SRC" initialConcentration="N".../>
           ...
</listOfSpecies>
<reaction id="Reaction1" reversible="false">
  <listOfReactants>
    <speciesReference species="FGFR_Ph1" />
    <speciesReference species="SRC" />
  </listOfReactants>
  <listOfProducts>
    <speciesReference species="FGFR_SRC" />
  </listOfProducts>
  <kineticLaw>
    <math xmlns="http://www.w3.org/1998/Math/MathML">
      <apply> <times/>
        <ci>k7</ci> <ci>FGFR_Ph1</ci> <ci>SRC</ci>
      </apply>
    </math>
  </kineticLaw>
</reaction>
```

Figure 3: SBML Code Fragment

nient, compositional language for describing discrete-state stochastic models of Section 2.2. Two simulation systems that accept π-calculus process syntax are available, BioSPI based on (Regev and Shapiro 2004) and SPiM (Cardelli and Phillips 2004). Other stochastic process calculi include PEPA (Hillston 1996), which has been successfully applied to the modelling of small examples such as the Ras/Raf/ERK signalling pathway (Calder et al. 2006).

A fragment of the π-calculus code (in the textual format of BioSPI) relating to FGFR and its interactions with FGF and Src is shown in Figure 4. We encode an FGFR protein as the parallel composition of interacting subcomponents, each of which models a characteristic of the protein, for example the connection to an FGF protein (bound/unbound) or the state of a residue (phosphorylated/unphosphorylated). Subcomponents serve as both protein internal states and its interfaces through which the interactions with other proteins occur. FGF, Src and Grb2 are modelled as separate processes (omitted from Figure 4). Input, output, action prefix and choice are denoted by "?", "!", "," and ";" respectively. For further details, see e.g., (Regev and Shapiro 2004).

The stochastic π-calculus can be used for both population- and individual-based models, and can be automatically generated from SBML descriptions of biochemical networks, see (Eccher 2006) for a recent proposal. One disadvantage of the π-calculus is the restriction to (asymmetric) binary input/output communication, rather than multi-way interactions between processes (see for example the sequence of actions (reloc, reloc1, ...) in FGFR_SRC in Figure 4 to model FGFR relocation).

3.3 PRISM

PRISM (Hinton et al. 2006, PRISM 2006) is a probabilistic model checker: a tool for the formal verification of quantitative properties of stochastic systems. It supports construction

```
FGFR ::= FGFR_FGF0 | FGFR_Ph10 | ... .

FGFR_FGF0::= bind_fgf!{rel_fgf, reloc4}, FGFR_FGF1; % binding FGF
             reloc1?[] , true . % relocation
FGFR_FGF1::= rel_fgf?[] , FGFR_FGF0; % releasing FGF
              ph1?[] , FGFR_FGF1; % phosphorylation
             reloc1?[] , reloc4![] , true; % relocation
             ... .
FGFR_Ph10 ::= ph1![] , FGFR_Ph11 . % phosphorylation
FGFR_Ph11 ::= bind_src!{rel_src1, rel_src2} , FGFR_SRC; % binding Src
             dph1![] , FGFR_Ph11 . % dephosphorylation
FGFR_SRC ::= rel_src1?[] , FGFR_Ph11; % releasing Src
             dph1![] , rel_src2! [], FGFR_Ph10;
             % dephosphorylation (and releasing Src)
             reloc![] , reloc1![] , reloc2![] , true . % relocation
```

Figure 4: Stochastic π-calculus Code Fragment

and analysis of three types of models: continuous-time and discrete-time Markov chains and Markov decision processes. Quantitative properties such as "what is the probability that protein A relocates within 2 hours?" or "what is the expected number of complexations that occur before degradation?" can be expressed using temporal logics (in this case, CSL and its extensions). Values for properties are then computed automatically by the tool. PRISM also supports the stochastic process algebra PEPA (Hillston 1996).

Models to be analysed in PRISM are specified in a simple, state-based description language. The PRISM modelling language variant that corresponds to PEPA has expressive power similar to the stochastic process calculus, and therefore can be viewed as an alternative, compositional language for inducing discrete-state stochastic models of Section 2.2. However, it is based on multi-way *synchronisation* rather than binary channel communication. Figure 5 shows a fragment of the PRISM language code for our running example, relating to FGFR and its interactions with FGF and Src. Each system component is described by a separate module, whose state is represented by a number of finite-values variables. The stochastic behaviour of each component is described by a set of guarded commands. Modules can interact through synchronisation, which is achieved by annotating commands in two or more modules with the same label.

PRISM has already been successfully applied to the modelling and analysis of several biochemical reaction networks, such as the Ras/Raf/ERK signalling pathway (Calder et al. 2006, Calder et al. 2006), cyclin (PRISM 2006) and FGF (Heath et al. 2006). These case studies demonstrate the use of PRISM for both population- and individual-based models. Similarly to the π-calculus, PRISM models are also easy to modify at the level of individual molecules or ensembles, for example when formulating an alternative hypothesis for the mechanism under study, and can be manipulated via text processing tools. Based on (Eccher 2006), automated translation of PRISM models from SBML is feasible and would avoid the difficulties with binary synchronisation.

```
module fgfr

fgfr_fgf : [0..1] init 0; // FGF bound
fgfr_ph1: [0..1] init 0; // state receptor 1 phosphorylated
fgfr_src : [0..1] init 0; // Src bound
reloc_fgfr: [0..1] init 0; // FGFR relocated
                              ...
// binding and release of FGF
[bind_fgf] reloc_fgfr=0∧fgfr_fgf=0→k1 : (fgfr_fgf'=1);
[rel_fgf]  reloc_fgfr=0∧fgfr_fgf=1→k2 : (fgfr_fgf'=0);
// phosphorylation/dephosphorylation (release SRC under dephosphorylation)
[] reloc_fgfr=0∧fgfr_fgf=1∧fgfr_ph1=0 →k3 : (fgfr_ph1'=1);
[] reloc_fgfr=0∧fgfr_ph1=1∧fgfr_src=0 →k5 : (fgfr_ph1'=0);
[rel_src] reloc_fgfr=0∧fgfr_ph1=1∧fgfr_src=1 →
                              k5 : (fgfr_ph1'=0)∧(fgfr_src'=0);
// binding and release of Src
[bind_src] reloc_fgfr=0∧fgfr_ph1=1∧fgfr_src=0→k7 : (fgfr_src'=1);
[rel_src]  reloc_fgfr=0∧fgfr_src=1→k8 : (fgfr_src1'=0);
// relocation (caused by Src)
[] reloc_fgfr=0∧fgfr_src=1→1/(15*60) : (reloc_fgfr'=1);

endmodule
```

Figure 5: PRISM Language Description Fragment

4 ANALYSIS TECHNIQUES

4.1 Differential Equations

As described in Section 2.1, continuous deterministic models of biochemical reaction networks describe the time evolution of molecular concentrations as a set of coupled ordinary differential equations (ODEs).

To build the differential equation model for our running example, we used Cellerator (Shapiro et al. 2003), a Mathematica-based tool for generating, translating and solving complex signal transduction networks. Cellerator supports a convenient input of fundamental biochemical reactions with arrow-based notation to represent biochemical reactions. Examples of chemical formulae recognised by Cellerator include association, dissociation, synthesis and degradation, and conversion reactions. Reactions are automatically translated into differential equations based on the law of mass action or enzymatic kinetic models.

For illustration purposes Figure 6 shows a fragment of the ODEs automatically generated for our example, relating to FGFR and its interactions with FGF and Src. Cellerator also supports solution of these ODEs. Figure 7(a) shows the results generated for the concentration of relocated FGFR and Grb2 bound to FGFR over a time period of 4 hours. We assumed concentrations to be of the order of 10^{-5} M, which necessitates the rescaling of binary reaction rates by the same factor. The system of ODEs is solved for initial conditions of 10^{-5} M for FGF, FGFR, Src and Grb2.

ODE models are particularly suitable for studying events in a linear pathway mediated by sequential reactions. Indeed, it is often possible, for small systems of ODEs arising from simple biochemical networks, to utilise matched asymptotics and quasi-steady state approximations to develop accurate analytical approximations (Murray 1989). Our running example is sufficiently simple to be amenable to such an approach, but we do not illustrate this as such techniques are not extendable to large systems, which must be addressed using numerical techniques.

In particular, if we allow parallel molecular state changes, such as the formation of complexes of multiple proteins, the complexity of the model significantly increases. The number of different system-wide states that fully describe the interactions between different proteins increases exponentially with the number of participating molecules, as does the number of equations.

While there exist algorithms capable of solving these very large systems of ODEs arising from networks of biochemical reactions, they are generally far too inefficient. This is because the ODEs are usually very stiff, that is the underlying reactions possess disparate timescales. Stability requirements for the use of explicit ODE solvers with such problems enforce an extremely small timestep and thus an excessively prolonged runtime. Implicit schemes are even more prohibitive in terms of runtime. An attempt to circumvent such difficulties is illustrated in (Tokman 2006).

ODE models predict the time course of average values of concentrations or substance, but their applicability is limited to cases with large numbers of molecules (on which the continuous abstraction depends). As we demonstrate in the next section, averages may be misleading for small numbers of molecules. On the other hand, ODEs are capable of modelling complex dynamics, such as higher order biochemical reactions, and are less dependent on the strong assumptions of constant volume and temperature.

Many other techniques for analysing systems of ODEs are also applicable in this domain. Bifurcation theory, for example, studies the dramatic changes in the solution behaviour when some parameters undergo a small change, allowing a modeller to narrow significantly the search for key dynamical behaviour in the parameter space. Also of interest is the use of hybrid automata as a computational modelling formalism for biological systems (Piazza et al. 2005). Hybrid automata support a combination of continuous and discrete system dynamics.

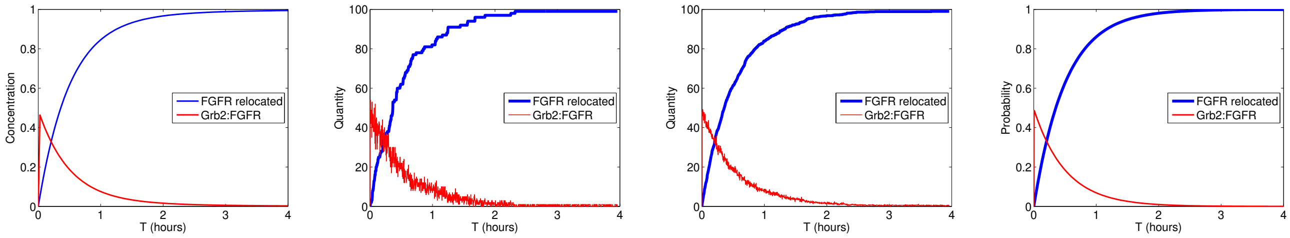
4.2 Simulation

For approaches based on discrete stochastic models, the most common analysis technique is the use of discrete event Monte Carlo simulation, which evolves the system over time in order to estimate the quantities of concentrations of specified complexes. This can be done directly from the syntactic description of the model, and corresponds to the algorithm of (Gillespie 1977) for population-based models. A number of simulation tools exist, including BioSPI (Regev and Shapiro 2004) and SPiM (Phillips and Cardelli 2005) for the stochastic π-calculus, and PRISM simulator for PEPA and PRISM models (Hinton et al. 2006).

$$
\begin{aligned}
Fgfr'_{0,0}(t) &= -\mathrm{bind_{fgf}}\cdot Fgf(t)\cdot Fgfr_{0,0}(t) + \mathrm{rel_{fgf}}\cdot Fgfr\_Fgf_{0,0}(t) + \mathrm{dph_1}\cdot Fgfr_{1,0}(t) + \mathrm{dph_1}\cdot Fgfr_{2,0}(t) \ldots \\
Fgfr'_{1,0}(t) &= -\mathrm{bind_{fgf}}\cdot Fgf(t)\cdot Fgfr_{1,0}(t) + \mathrm{rel_{fgf}}\cdot Fgfr\_Fgf_{1,0}(t) - \mathrm{dph_1}\cdot Fgfr_{1,0}(t) \\
&\quad -\mathrm{bind_{src}}\cdot Src(t)\cdot Fgfr_{1,0}(t) + \mathrm{rel_{src}}\cdot Fgfr_{2,0}(t)\ \ldots \\
Fgfr'_{0,1}(t) &= -\mathrm{bind_{fgf}}\cdot Fgf(t)\cdot Fgfr_{0,1}(t) + \mathrm{rel_{fgf}}\cdot Fgfr\_Fgf_{0,1}(t) + \mathrm{dph_1}\cdot Fgfr_{1,1}(t) + \mathrm{dph_1}\cdot Fgfr_{2,1}(t)\ \ldots \\
Fgfr'_{1,1}(t) &= -\mathrm{bind_{fgf}}\cdot Fgf(t)\cdot Fgfr_{1,1}(t) + \mathrm{rel_{fgf}}\cdot Fgfr\_Fgf_{1,1}(t) - \mathrm{dph_1}\cdot Fgfr_{1,1}(t) \\
&\quad -\mathrm{bind_{src}}\cdot Src(t)\cdot Fgfr_{1,1}(t) + \mathrm{rel_{src}}\cdot Fgfr_{2,1}(t)\ \ldots \\
Fgfr'_{2,0}(t) &= -\mathrm{bind_{fgf}}\cdot Fgf(t)\cdot Fgfr_{2,0}(t) + \mathrm{bind_{src}}\cdot Src(t)\cdot Fgfr_{1,0}(t) + \mathrm{rel_{fgf}}\cdot Fgfr\_Fgf_{2,0}(t) - \mathrm{rel_{src}}\cdot Fgfr_{2,0}(t) \\
&\quad -\mathrm{reloc}\cdot Fgfr\_Fgf_{2,0}(t) - \mathrm{dph_1}\cdot Fgfr\_Fgf_{2,0}(t)\ \ldots \\
Fgfr'_{0,2}(t) &= -\mathrm{bind_{fgf}}\cdot Fgf(t)\cdot Fgfr_{0,2}(t) + \mathrm{rel_{fgf}}\cdot Fgfr\_Fgf_{0,2}(t) + \mathrm{dph_1}\cdot Fgfr_{2,2}(t) + \mathrm{dph_1}\cdot Fgfr_{1,2}(t)\ \ldots \\
Fgfr'_{2,2}(t) &= -\mathrm{bind_{fgf}}\cdot Fgf(t)\cdot Fgfr_{2,2}(t) + \mathrm{rel_{fgf}}\cdot Fgfr\_Fgf_{2,2}(t) - \mathrm{rel_{src}}\cdot Fgfr_{2,2}(t) - \mathrm{reloc}\cdot Fgfr_{2,2}(t) \\
&\quad -\mathrm{dph_1}\cdot Fgfr_{2,2}(t) + \mathrm{bind_{src}}\cdot Src(t)\cdot Fgfr_{1,2}(t)\ \ldots \\
Fgfr'_{2,1}(t) &= -\mathrm{bind_{fgf}}\cdot Fgf(t)\cdot Fgfr_{2,1}(t) + \mathrm{bind_{src}}\cdot Src(t) Fgfr\_Fgf_{1,1}(t) + \mathrm{rel_{fgf}}\cdot Fgfr\_Fgf_{2,1}(t) - \mathrm{rel_{src}}\cdot Fgfr_{2,1}(t) \\
&\quad -\mathrm{reloc}\cdot Fgfr_{2,1}(t) - \mathrm{dph_1}\cdot Fgfr_{2,1}(t)\ \ldots \\
Fgfr'_{1,2}(t) &= -\mathrm{bind_{fgf}}\cdot Fgf(t)\cdot Fgfr_{1,2}(t) + \mathrm{rel_{src}}\cdot Fgfr\_Fgf_{2,2}(t) + \mathrm{rel_{fgf}}\cdot Fgfr\_Fgf_{1,2}(t) - \mathrm{dph_1}\cdot Fgfr_{1,2}(t) \\
&\quad -\mathrm{bind_{src}}\cdot Src(t)\cdot Fgfr_{1,2}(t)\ \ldots
\end{aligned}
$$

In the terms $Fgfr_{res_1,res_2}$ and $Fgfr\_Fgf_{res_1,res_2}$ the components res_1 and res_2 correspond to two independent residues of the protein: 0 (unphosphorylated), 1 (phosphorylated) and 2 (bound to Src or Grb2).

Figure 6: Fragment of the Automatically Generated ODEs



(a) ODEs (b) BioSPI (1 run) (c) BioSPI (10 runs) (d) PRISM

Figure 7: Results of Analysis of the Example: Concentration/Quantity of Two Forms of FGFR over Time

Although useful information about a model can be extracted from a single random run, to obtain more robust estimates of the system behaviour over time it is necessary to average over several simulation runs. These can then be compared with experimental outcomes. We used BioSPI as the simulation platform for the stochastic π-calculus model of the FGF fragment. The BioSPI system inputs the π-calculus code and performs simulations using the Gillespie algorithm, starting from a given initial state. Figures 7(b) and 7(c) show the results generated for the amount of relocated FGFR and Grb2 bound to FGFR over a time period of 4 hours, both for a single simulation run and averaged over 10 runs. In each case, we assume an initial population of 100 FGF, FGFR, Src and Grb2 molecules.

In the model for the running example above, Src-mediated endocytic internalization of FGFR was presumed to attenuate signalling by relocating and degrading receptor complex. Recent evidence suggests that FGF-stimulated signalling can be amplified by internalization (Ware et al. 1997, Frame 2004). Src can alter cell structure, in particular the actin cytoskeleton, resulting in changes of intracellular trafficking of Src and FGFR. Src might positively regulate FGFR signalling by recruiting non-active FGFR to the membrane. This can be modelled by adding the following schematic reaction to the model:

$$\mathrm{FGFR:Src} \quad \rightarrow \quad \mathrm{FGFR:Src} + \mathrm{FGFR} + \mathrm{Src}$$

and adapting the π-calculus model appropriately. We change the initial amount of Src from 100 to 10 molecules in the π-calculus model and from concentration 10^{-5} M to 10^{-6} M in the ODE model (all other initial conditions remain the same as before). Figure 8 shows plots of the amount of Grb2 bound to FGFR. In this case, the ODE result disagrees with averaged simulation runs from the π-calculus model. This is because the stochastic approach is more accurate when the number of molecules is small and the behaviour of the reaction system becomes non-continuous. The behaviour of the ODE model differs because Src cannot be totally degraded (the degradation is balanced by the formation of new Src), whereas in the stochastic model the random walk of Src, which starts at 10, can easily lead to 0.

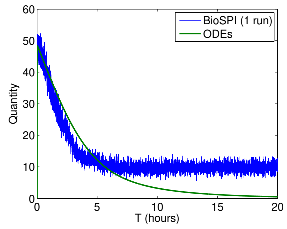


Figure 8: An Extended Example Demonstrating the Difference between ODE and Stochastic Modelling Results

Monte Carlo simulation techniques can be implemented efficiently. However, it is well known that the number of runs that need to be generated is quadratic in the inverse of the desired accuracy. Thus, obtaining accurate approximations can be costly. In particular, this is unavoidable if the quantities concerned are very small. In systems with considerably differing timescales, which is the case in biochemical networks, long-run average properties cannot be obtained with simulation. Alternative algorithms, e.g., as supported by StochSim (Novère and Shimizu 2001), can simulate individual molecules and their spatial arrangement.

4.3 Verification

For discrete-state stochastic models, an alternative to Monte Carlo simulation is *formal verification*, and in particular, in this context, probabilistic model checking. This approach proceeds by first building a representation of the underlying continuous-time Markov chain, usually in some compact form, followed by exhaustive exploration of the paths of the system in order to produce answers to *quantitative* temporal queries expressible in temporal logic. Note that this differs from simulation approaches, which can generate system trajectories from the syntax of model description. The properties include the probability of an event, transient probability of an event (i.e., at a particular time instant), long-run probability, or expectation.

Probabilistic model checking tools, such as PRISM, compute values for these quantitative properties using numerical solution algorithms, typically based on iterative methods. Usually computation reduces to the problem of solving a system of linear equations, for which well-known efficient iterative methods such as Gauss-Seidel exist. However for transient probabilities, an iterative method known as uniformisation is used, which is based on discretisation. For more information see e.g., (Rutten et al. 2004).

In Figure 7(d) we show, similarly to the previous two sections, experimental results for the amount of relocated FGFR and Grb2 bound to FGFR. Here, the results have been generated with PRISM for the case when there is one molecule of each species, and hence we have plotted the probability of FGFR being relocated and Grb2 being bound.

The main obstacle associated with probabilistic model checking (and formal verification in general) is the state-space explosion, i.e., that the parallel composition of N components (molecules) leads to systems whose state space is exponential in N. State-of-the-art techniques developed in the area enable the analysis of systems with billions of states. These include *symbolic* methods, using sophisticated data structures based on binary decision diagrams (BDDs) and techniques such as symmetry reduction (Kwiatkowska et al. 2006), which in the biological setting actually corresponds to employing the population based approach, as well as using an abstract notion of quantities (Calder et al. 2006).

Returning to our running example, the state space explosion problem can be seen when increasing the number of molecules of each type; for example, increasing this number from 1 to 5 leads to an increase in the state space from 22 to 4,568,094. On the other hand, using symmetry reduction or employing the population based approach, we have, for the case of 5 molecules of each type, a reduction in the state space from 4,568,094 to 63,756.

Figure 9 presents further results obtained with PRISM: both the expected number of reactions of a certain type and the expected time a complex is present by time T.

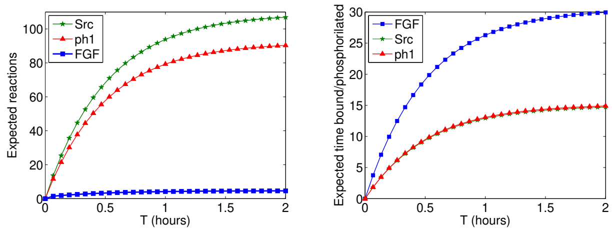


Figure 9: Further PRISM Results

Below we include a number of long-run properties analysed using probabilistic model checking with PRISM.

- The probability that FGF, Src or Grb2 is free when FGFR degrades equals 4.0e-7, 0 and 0.50660 respectively. The probability for Src is 0 because Src must be bound for FGFR to degrade. The values for FGF and Grb2 are justified as FGF binds quickly and is released slowly, whereas the binding and release of Grb2 happens at the same rate.
- The expected number of phosphorylations of type 1 and 2 before FGFR degrades equal 92.09 and 91.84. The expected number is higher for type 1 because type 1 must occur for FGFR to degrade.
- The expected time until FGFR degrades is 30.53 minutes and the expected time that FGF, Src or Grb2 spend bound to FGFR before degradation equals 30.53, 15.00 and 15.04 minutes respectively. This shows that FGF is bound for most of the time that FGFR is present and can be attributed to the fact that FGF binds quickly and is released slowly. Src and Grb2 spend roughly half the time bound because their complexation and decomplexation rates are the same. Grb2 is bound for slightly longer than Src because, the binding of Src causes degradation.

We demonstrate further quantitative properties that can be automatically verified using PRISM with the help of the full FGF pathway studied in (Heath et al. 2006), see also (PRISM 2006). The full pathway additionally includes the following elements: FRS2, Plc, Spry, Sos, Cbl and Shp2. We were able to verify, amongst others:

- “The expected time Grb2 spends bound to FRS2 before either degradation or relocation occurs.”

- "The expected number of times Grb2 binds to FRS2 before either degradation or relocation occurs."
- "The probability that each possible cause of degradation/relocation occurs first."

For illustration, the graph in Figure 10 shows the amount of Grb2 bound to FRS2 (not included in the running example). The plots show the result of our *in silico genetics* experimentation, that is, how the variation in quantity is affected by the removal of certain key components.

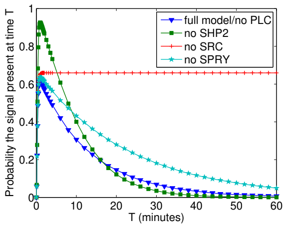


Figure 10: The Variation in the Amount of Grb2 Bound to FRS2 in the Full FGF Pathway Model

As illustrated above, automatic verification techniques can greatly enhance the range of analyses possible for biochemical networks. Verification is based on an exploration of the full model, and is therefore able to inspect the temporal relationships between events in fine detail, and in particular detect 'corner cases' such as unwanted deadlocks. Related work of interest in this area is a technique from the tool Simpathica (Antoniotti et al. 2003) which evaluates temporal logic queries against a set of system traces obtained through simulation. For quantitative verification, probabilistic model checking is, based on the quality of the data provided, (numerically) exact, as opposed to simulation which produces estimates, and can automatically identify scenarios that yield best/worst case answers. Note that discrete quantities (such as expected bindings) cannot be obtained with ODE models. However, the size of the resulting models is at present a limitation on applicability of automatic verification techniques.

5 CONCLUSIONS

In this paper we gave an overview of the ODE, simulation and verification approaches to the analysis of biochemical reaction networks. Such networks can be described in SBML, and the corresponding ODE or discrete-state models generated automatically, subject to certain restrictions. The ODE models are continuous and deterministic with respect to average concentrations, and while this admits complex dynamics and a broad range of solvers, the approach cannot handle small numbers of molecules and discrete quantities such as expected number of bindings. Discrete event simulation can be applied to generate time trajectories of approximate reactant quantities directly from their syntactic representation. This method is inefficient if the quantities are very small, and not feasible for long-run averages, though, on the other hand, it is easy to parallelise. Automatic verification techniques aim to produce a detailed analysis of the causal and temporal relationships between events in the model, which necessitates the construction of the full model and its systematic exploration. This approach supports a wide range of qualitative and quantitative temporal queries, is exact and can produce best/worst case answers and the corresponding scenarios. However, Monte Carlo simulation and ODEs can tackle a larger class of models. The size of the resulting models remains the main limitation of the automatic verification approaches, motivating the need for research into compositional reasoning.

ACKNOWLEDGMENTS

This work is part-sponsored by EPSRC grants GR/S46727, GR/S11107 and Integrative Biology (GR/S72023), Microsoft Research Cambridge contract MRL 2005-04, and by a programme from Cancer Research UK.

REFERENCES

Antoniotti, M., A. Policriti, N. Ugel, and B. Mishra. 2003. Model building and model checking for biochemical processes. *Cell Biochemistry and Biophysics* 38.

Calder, M., S. Gilmore, and J. Hillston. 2006. Modelling the influence of RKIP on the ERK signalling pathway using the stochastic process algebra PEPA. *Transactions on Computational Systems Biology* 4230.

Calder, M., V. Vyshemirsky, D. Gilbert, and R. Orton. 2006. Analysis of signalling pathways using continuous time Markov chains. *Transactions on Computational Systems Biology* 4220.

Cardelli, L., and A. Phillips. 2004. A correct abstract machine for the stochastic pi-calculus. In *Proceedings of BioConcur'04*.

Eccher, C. 2006. *Translation of Systems Biology Markup Language into process algebra*. Ph. D. thesis.

Frame, M. 2004. Newest findings on the oldest oncogene; how activated Src does it. *Journal of Cell Science* 117.

Gillespie, D. 1977. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry* 81 (25).

Heath, J., M. Kwiatkowska, G. Norman, D. Parker, and O. Tymchyshyn. 2006. Probabilistic model checking of complex biological pathways. In *Proc. International Conference on Computational Methods in Systems Biology*, Volume 4210 of *LNBI*: Springer.

Hillston, J. 1996. *A compositional approach to performance modelling*. Cambridge University Press, Cambridge.

Hinton, A., M. Kwiatkowska, G. Norman, and D. Parker. 2006. PRISM: A tool for automatic verification of probabilistic systems. In *Proc.12th. International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, Volume 3920 of *LNCS*: Springer.

Kwiatkowska, M., G. Norman, and D. Parker. 2006. Symmetry reduction for probabilistic model checking. In *Proc. 18th International Conference on Computer Aided Verification*, Volume 4144 of *LNCS*: Springer-Verlag.

Murray, J. 1989. *Mathematical biology*. Springer Verlag.

Novère, N. L., and T. Shimizu. 2001. Stochsim: modelling of stochastic biomolecular processes. *Bioinformatics* 17.

Phillips, A., and L. Cardelli. 2005. A graphical representation for the stochastic pi-calculus. In *Proceedings of Bioconcur'05*.

Piazza, C., M. Antoniotti, V. Mysore, A. Policriti, F. Winkler, and B. Mishra. 2005. Algorithmic algebraic model checking i: Challenges from systems biology. In *Proc. 17th International Conference on Computer Aided Verification*, Volume 3576 of *LNCS*.

Priami, C., A. Regev, E. Shapiro, and W. Silverman. 2001. Application of a stochastic name-passing calculus to representation and simulation of molecular processes. *Information Process Letters* 80.

PRISM 2006. ⟨www.cs.bham.ac.uk/dxp/prism⟩.

Regev, A., and E. Shapiro. 2002. Cells as computation. *Nature* 419.

Regev, A., and E. Shapiro. 2004. The pi-calculus as an abstraction for biomolecular systems. In *Modelling in Molecular Biology*: Springer.

Rutten, J., M. Kwiatkowska, G. Norman, and D. Parker. 2004. *Mathematical techniques for analyzing concurrent and probabilistic systems*, Volume 23 of *CRM Monograph Series*. American Mathematical Society.

SBML 2006. ⟨http://sbml.org/index.psp⟩.

Shapiro, B., A. Levchenko, E. Meyerowitz, B. Wold, and E. Mjolsness. 2003. Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations. *Bioinformatics* 19 (5).

Tokman, M. 2006. Efficient integration of large stiff systems of ODEs with exponential propagation iterative (EPI) methods. *Journal of Computational Physics* 213.

Ware, M., D. Tice, S. Parsons, and D. Lauffenburger. 1997. Overexpression of cellular Src in fibroblasts enhances endocytic internalization of Epidermal Growth Factor receptor. *Journal of Biological Chemistry* 272.

Wolkenhauer, O., M. Ullah, W. Kolch, and K. Cho. 2004. Modeling and simulation of intracellular dynamics: choosing an appropriate framework. *IEEE Transactions on Nanobioscience* 3.

Yamada, S., T. Taketomi, and A. Yoshimura. 2004. Model analysis of difference between EGF pathway and FGF pathway. *Biochemical and Biophysical Research Communications* 314.

AUTHOR BIOGRAPHIES

MARTA KWIATKOWSKA is Professor of Computer Science in the University of Birmingham, UK. Her research is mainly concerned with developing modelling frameworks and novel methods for analysing large complex systems, especially automatic verification techniques. She led development of the state-of-the-art probabilistic model checker PRISM and is on the Editorial Board of Transactions on Computational Systems Biology and Logical Methods in Computer Science. Email: ⟨mzk@cs.bham.ac.uk⟩, webpage: ⟨www.cs.bham.ac.uk/~mzk⟩.

GETHIN NORMAN is Research Fellow in the School of Computer Science in the University of Birmingham, UK, and holds a BSc from Oxford University and PhD from the University of Birmingham. He has made a substantial contribution to the development of the modelling and verification techniques that underpin PRISM and has carried out many modelling case studies, including verification of FGF. Email: ⟨gxn@cs.bham.ac.uk⟩, webpage: ⟨www.cs.bham.ac.uk/~gxn⟩.

DAVID PARKER is Research Fellow in the School of Computer Science in the University of Birmingham, UK, who obtained his BSc and PhD from the University of Birmingham. His PhD thesis was runner up in the 2003 BCS Distinguished Dissertation Awards. He implemented the PRISM model checker and is author of several novel techniques, such as symmetry reduction. Email: ⟨dxp@cs.bham.ac.uk⟩, webpage: ⟨www.cs.bham.ac.uk/~dxp⟩.

OKSANA TYMCHYSHYN is a PhD student in the School of Computer Science in the University of Birmingham, UK. Her research is part of the Integrative Biology project (www.integrativebiology.ox.ac.uk) and concerns modelling signalling pathways and their influence on colon crypt development. E-mail: ⟨oxt@cs.bham.ac.uk⟩, webpage: ⟨www.cs.bham.ac.uk/~oxt⟩.

JOHN HEATH is Professor and Head of School of Biosciences in the University of Birmingham, UK. His lab is working on the specificity and dynamics of the interaction between growth factors and their receptors, and especially growth factor signalling in cancer. Email: ⟨j.k.heath@bham.ac.uk⟩.

EAMONN GAFFNEY is Lecturer in the School of Mathematics in the University of Birmingham, UK. He works on mathematical models of tumours, signalling pathways and cell movement, and has contributed an ODE model of FGF. Email: ⟨eag@for.mat.bham.ac.uk⟩, webpage: ⟨web.mat.bham.ac.uk/E.A.Gaffney⟩.