An introduction to BioModel Engineering, illustrated for signal transduction pathways

David Gilbert[1], Rainer Breitling[2], Monika Heiner[3], and Robin Donaldson[4]

[1] School of Information Science, Computing and Mathematics, Brunel University, UK
david.gilbert@brunel.ac.uk
[2] Groningen Bioinformatics Centre, University of Groningen, 9751 NN Haren, The Netherlands
[3] Department of Computer Science, Brandenburg University of Technology, Cottbus, Germany
[4] Bioinformatics Research Centre, University of Glasgow, Glasgow G12 8QQ, UK

Abstract. BioModel Engineering is the science of designing, constructing and analyzing computational models of biological systems. It is inspired by concepts from software engineering and computing science.

This paper illustrates a major theme in BioModel Engineering, namely that identifying a quantitative model of a dynamic system means building the structure, finding an initial state, and parameter fitting. In our approach, the structure is obtained by piecewise construction of models from modular parts, the initial state is obtained by analysis of the structure and parameter fitting comprises determining the rate parameters of the kinetic equations. We illustrate this with an example in the area of intracellular signalling pathways.

1 Introduction

BioModel Engineering takes place at the interface of computing science, mathematics, engineering and biology, and provides a systematic approach for designing, constructing and analyzing computational models of biological systems. Some of its central concepts are inspired by efficient software engineering strategies. BioModel Engineering does not aim at engineering biological systems per se, but rather aims at describing their structure and behavior, in particular at the level of intracellular molecular processes, using computational tools and techniques in a principled way.

In this paper we present some techniques for the systematic construction of models of biochemical systems from constituent building blocks, and how such models can be tuned to exhibit some desired behaviour, applying our approach to a signal transduction pathway. Our presentation is aimed at computing scientists and software engineers who want to learn how their skills can be successfully applied in modern systems biology.

After a brief introduction of the biological context, this paper illustrates a major theme in BioModel Engineering, namely that identifying a (qualitative) model means (1) finding the structure, (2) obtaining an initial state and

(3) parameter fitting. In our approach, the structure is obtained by piecewise construction of models from modular parts, the initial state which describes concentrations of species or numbers of molecules is obtained by analysis of the structure and parameter fitting comprises determining the rate parameters of the kinetic equations by reference to trusted data.

2 Biochemical Context

There are many networks of interacting components known to exist as part of the machinery of living organisms. Biochemical networks can be metabolic, regulatory or signal transduction networks.

In this paper we focus on signal transduction, which is the mechanism which enables a cell to sense changes in its environment and to make appropriate responses. The basis of this mechanism is the conversion of one kind of signal into another. Extracellular signaling molecules are detected at the cell membrane by being bound to specific trans-membrane receptors that face outwards from the membrane and trigger intracellular events, which may eventually effect transcriptional activities in the nucleus. The eventual outcome is an alteration in cellular activity including changes in the gene expression profiles of the responding cells. These events, and the molecules that they involve, are referred to as (intracellular) "signalling pathways"; they contribute to the control of processes such as proliferation, cell growth, movement, apoptosis, and inter-cellular communication. Many signal transduction processes are "signalling cascades" which comprise a series of enzymatic reactions in which the product of one reaction acts as the catalytic enzyme for the next. The effect can be amplification of the original signal, although in some cases, for example the MAP kinase cascade, the signal gain is modest [SEJGM02], suggesting that a main purpose is regulation [KCG05] which may be achieved by positive and negative feedback loops [BF00], although there may be some feedback redundancy with respect to receptor internalisation [OSG+08].

3 Modelling using building blocks based on enzymatic reactions

Formally, a quantitative model of a biochemical pathway can be described by a tuple $< T, M, K, R >$ where T is the topology (biochemical species and their connectivities), M an initial state describing concentrations or molecular numbers of species, K a set of kinetic equations and R a set of kinetic rate parameters. A qualitative model comprises at least the topology T with optionally an initial state.

In this section we discuss how signal transduction cascades can be modelled in a modular fashion using a building-block approach, which will generate the topology T and set of kinetic equations K of a model. We have shown in previous work [BGHO08] how building-block based construction can be achieved using

both a qualitative approach – Qualitative Petri nets, and quantitative approaches – Continuous Petri Nets and Ordinary Differential Equations (ODEs), but in this review we restrict our discussions to ODEs. In the following sections we will introduce techniques to generate an initial state, or in Petri net terminology *marking* and obtain a set of rate parameters for the model.

The basic building block of any biological dynamic system is the enzymatic reaction: the conversion of a substrate into a product catalysed by an enzyme. Such enzymatic reactions can be used to describe metabolic conversions, the activation of signalling molecules and even transport reactions between various subcellular compartments. Enzymes greatly accelerate reactions in one direction (often by factors of at least 10^6), and most reactions in biological systems do not occur at perceptible rates in the absence of enzymes. We can illustrate a simple enzymatic reaction involving one substrate A, one product B, and an enzyme E by

$$A \xrightarrow{E} B \; . \tag{1}$$

3.1 Basic kinetic descriptions

In general, there are two ways to describe the kinetics of enzymatic reactions: *Michaelis-Menten* and *Mass-action*.

Michaelis-Menten

The *Michaelis-Menten* equation for the basic enzymatic reaction is given in Equation MM

$$V = V_{\max} \times \frac{[A]}{K_M + [A]} \tag{MM}$$

where V is the reaction velocity, $V_{\max}$ is the maximum reaction velocity, and K_M, the *Michaelis constant*, is the concentration of the substrate at which the reaction rate is half its maximum value. The concentration of the substrate A is represented by $[A]$ in this rate equation. With the total enzyme concentration $[E_T]$ and the equation

$$k_{\text{cat}} = \frac{V_{\max}}{[E_T]} \tag{2}$$

we are able to write the differential equations describing the consumption of the substrate and production of the product as:

$$\frac{d[A]}{dt} = -\frac{d[B]}{dt} = -kcat \times [E_T] \times \frac{[A]}{(K_M + [A])} \tag{3}$$

The Michaelis-Menten equation only holds at the initial stage of a reaction before the concentration of the product is appreciable, and makes the following assumptions:

1. The concentration of product is (close to) zero.
2. No product reverts to the initial substrate.
3. The concentration of the enzyme is much less than the concentration of the substrate, i.e. $[E] \ll [A]$.

Although these are reasonable assumptions for enzyme assays in a test tube, assumptions 1 and 2 do not hold for most metabolic pathways *in vivo*, and none of the assumptions is correct for cellular signalling pathways.

Mass-action

A more detailed description using Mass-action kinetics can be given by taking into account the mechanism by which the enzyme acts, namely by forming a complex with the substrate, modifying the substrate to form the product, and a disassociation occurring to release the product, i.e. $A + E \rightleftharpoons AE \rightarrow B + E$. In order to take into account the kinetic properties of many enzymes, we associate *rate constants* with each reaction. Thus the enzyme E can combine with the substrate A to form the $A|E$ complex with rate constant k_1. The $A|E$ complex can dissociate to E and A with rate constant k_2, or form the product B with rate k_3:

$$A + E \underset{k_2}{\overset{k_1}{\rightleftarrows}} A|E \xrightarrow{k_3} B + E \ . \tag{MA1}$$

This simple mass-action model is related to the Michaelis-Menten Equation MM as described previously by the following constraints:

$$\frac{k_2 + k_3}{k_1} = K_M \tag{4}$$

where $k_3 = k_{\text{cat}} = \frac{V_{\text{max}}}{[E_T]}$ as in Equation 2.

We can derive a set of differential equations, see Equation 5 from the mass-action description given in Equation MA1:

$$\begin{aligned}
\tfrac{d[A]}{dt} &= -k_1 \times [A] \times [E] + k_2 \times [A|E] \\
\tfrac{d[A|E]}{dt} &= \ \ k_1 \times [A] \times [E] - k_2 \times [A|E] - k_3 \times [A|E] \\
\tfrac{d[B]}{dt} &= \qquad\qquad\qquad\qquad + k_3 \times [A|E] \\
\tfrac{d[E]}{dt} &= -k_1 \times [A] \times [E] + k_2 \times [A|E] + k_3 \times [A|E]
\end{aligned} \tag{5}$$

The mass-action model described in Equation MA1 assumes that almost none of the product reverts back to the original substrate, a condition that holds at the initial stage of a reaction before the concentration of the product is appreciable. This means that this type of mass-action model is a direct equivalent of the Michaelis-Menten equation, and will face the same limitations when applied to *in*

vivo signalling systems. However, the mass-action description offers much more flexibility and thus can be easily expanded to cover more general situations, for example Equations MA2 and MA3 below.

$$A + E \underset{k_2}{\overset{k_1}{\rightleftarrows}} A|E \xrightarrow{k_3'} B|E \underset{k_1'}{\overset{k_2'}{\rightleftarrows}} B + E \tag{MA2}$$

$$A + E \underset{k_2}{\overset{k_1}{\rightleftarrows}} A|E \underset{k_4'}{\overset{k_3'}{\rightleftarrows}} B|E \underset{k_1'}{\overset{k_2'}{\rightleftarrows}} B + E \tag{MA3}$$

3.2 Modelling one step in a signal transduction cascade

One step in a classical signal transduction cascade comprises the *phosphorylation* of a protein by an enzyme S which is termed a kinase, see Figure 1. It is the phosphorylated form R_p which can act as an enzyme to catalyse the phosphorylation of a further component in the cascades, see Figure 3(a).

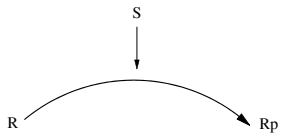

Fig. 1. Basic enzymatic step; R – signalling protein; Rp – phosphorylated form; S – kinase

We can model this reaction using any of the kinetic patterns introduced in Section 3.1 e.g., the Mass-action **MA1** pattern as follows, straightforwardly adapted from Equation MA1 by renaming in Equation 6, where R is a protein and R_p its phosphorylated form, S is a signal enzyme and $R|S$ the complex formed from R and S:

$$R + S \underset{k_2}{\overset{k_1}{\rightleftarrows}} R|S \xrightarrow{k_3} R_p + S \tag{6}$$

In order to ensure that such a single step is not a 'one shot' affair (i.e. to ensure that the substrate in the non-phosphorylated form is replenished and not exhausted), and hence that the signal can be deactivated where necessary, biological systems employ a phosphatase which is an enzyme promoting the dephosphorylation of a phosphorylated protein. This is depicted in Figure 2, which we are going to model by our Mass-action pattern.

Using the **MA1** pattern (Mass-action kinetics) we get Equation 7, where P is a phosphatase and k_n, kr_n are rate constants for the forward and reverse reactions respectively. In many cases it would also be justified to model the

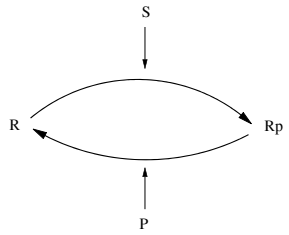

Fig. 2. Basic phosphorylation–dephosphorylation step; R – signalling protein; Rp – phosphorylated form; S – kinase; P – phosphatase

dephosphorylation as an un-catalysed first-order decay reaction, because detailed knowledge of phosphatase concentrations, specificities, and kinetic parameters is still lagging behind our understanding of the kinase enzymes.

$$\begin{aligned} R + S \underset{k_2}{\overset{k_1}{\rightleftarrows}} R|S \xrightarrow{k_3} R_p + S \\ R + P \xleftarrow{kr_3} R|P \underset{kr_2}{\overset{kr_1}{\rightleftarrows}} R_p + P \end{aligned} \tag{7}$$

3.3 Composing kinase cascades using building blocks

Once we have defined the building blocks, we can compose them by chaining together basic phosphorylation–dephosphorylation steps.

Vertical and horizontal composition Composition can be performed vertically as in Figure 3(a) to form a *signalling cascade*, where the signalling protein in the second stage is labelled RR and its phosphorylated form is labelled RR_p. Horizontal composition is illustrated in Figure 3(b) where a double phosphorylation step is described; the double phosphorylated form of a protein is subscripted by pp.

We can again use any of the kinetic patterns that were previously introduced in order to derive the models. For example, using **MA1** we can represent a two-stage cascade illustrated in Figure 3(a) by the following mass-action equations, ignoring for the sake of simplicity the dephosphorylation steps in the textual representation. The rate constants associated with the second stage are labelled kk_n. We would not expect the dephosphorylation rate constants to be related to the phosphorylation rate constants.

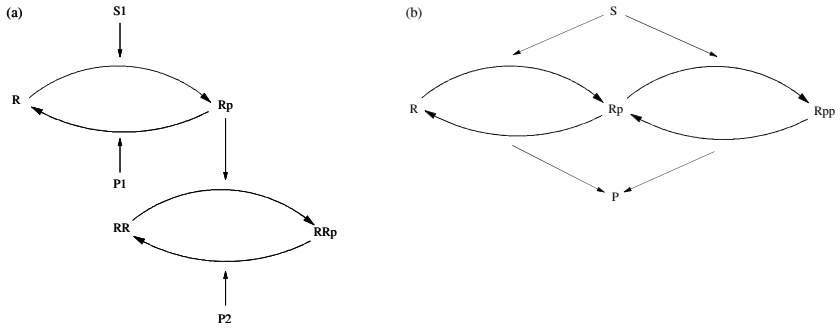

Fig. 3. (a) Vertical composition: Cascade formed by chaining two basic phosphorylation-dephosphorylation steps. (b) Horizontal composition: One stage cascade with a single to double phosphorylation step.

$$\begin{aligned} R + S_1 &\underset{k_2}{\overset{k_1}{\rightleftarrows}} R|S_1 \xrightarrow{k_3} R_p + S1 \\ RR + R_p &\underset{kk_2}{\overset{kk_1}{\rightleftarrows}} RR|R_p \xrightarrow{kk_3} RR_p + R_p \end{aligned} \tag{8}$$

The addition of a double phosphorylation step to a cascade layer is given in Figure 3(b), where both the single and double phosphorylation steps are catalysed by the same enzyme S; likewise, the two dephosphorylation steps are usually catalysed by the same phosphatase P. This system component can be described by Equation 9, if we apply the mass-action kinetics **MA1** and ignore again for the sake of simplicity the dephosphorylation steps in the textual representation. The rate constants associated with the double phosphorylation are labelled kp_n. Often, we can assume that the rate constants for the two steps of the double phosphorylation are similar to those for the single phosphorylation.

$$\begin{aligned} R + S &\underset{k_2}{\overset{k_1}{\rightleftarrows}} R|S \xrightarrow{k_3} R_p + S \\ R_p + S &\underset{kp_2}{\overset{kp_1}{\rightleftarrows}} R_p|S \xrightarrow{kp_3} R_{pp} + S \end{aligned} \tag{9}$$

3.4 Negative and positive feedback

Feedback in a signalling network can be achieved in several ways. For example, negative feedback can be implemented at the molecular level by sequestration

of the input signal S_1 by the product of the second stage RR_p. This system is sketched in Figure 4(a), and can be achieved by combining equations 8 and 10.

Similarly, positive feedback can also be achieved by the sequestration of the input signal S_1 by the product of the second stage, under the additional condition that the resulting $S_1|RR_p$ complex is a more active enzyme than S_1 alone. In this case we add Equation 11 to equations 8 and 10. The system is sketched in Figure 4(b).

$$S_1 + RR_p \underset{i_2}{\overset{i_1}{\rightleftarrows}} S_1|RR_p \tag{10}$$

$$R + S_1|RR_p \underset{kp_2}{\overset{kp_1}{\rightleftarrows}} R|S_1|RR_p \xrightarrow{kp_3} R_p + S_1|RR_p \tag{11}$$

Many other molecular mechanisms can be envisaged and are in fact observed in biological systems. All of these can be represented using the same basic formalism. For example, we can model an influence of RR_p on the phosphatase P_1, in which case the effects of positive and negative feedback are reversed, i.e. sequestration of P_1 by RR_p can cause positive feedback – see Figure 4(c). This can be achieved with Equations 8 and 12. Alternatively the situation where the $P_1|RR_p$ complex is more active than P_1 will cause negative feedback, Figure 4(d), and can be described by adding Equation 13 to Equations 8 and 12.

$$P_1 + RR_p \underset{i_2}{\overset{i_1}{\rightleftarrows}} P_1|RR_p \tag{12}$$

$$R_p + P_1|RR_p \underset{kr'_2}{\overset{kr'_1}{\rightleftarrows}} R_p|P_1|RR_p \xrightarrow{kr'_3} R + P_1|RR_p \tag{13}$$

4 Deriving initial states from qualitative descriptions

Recall that a quantitative model of a biochemical pathway can be described by a tuple comprising the topology, an initial marking or set of concentrations, a set of kinetic equations and a set of rate parameters. Using the approach described in the previous section we have shown how to construct an initial model with a topology and set of kinetic equations. In this section we outline an approach to generate the initial steady state, and in the following section we will describe one approach to obtain a set of rate parameters.

The most abstract representation of a biochemical network is *qualitative* and is minimally described by its topology, usually as a bipartite directed graph with nodes representing biochemical entities or reactions, or in Petri net terminology *places* and *transitions*. Petri nets are a well-established technique for representing biochemical networks, outlined e.g. in [MFD+03], [HGD08] and for a general introduction to Petri nets see [Mur89].

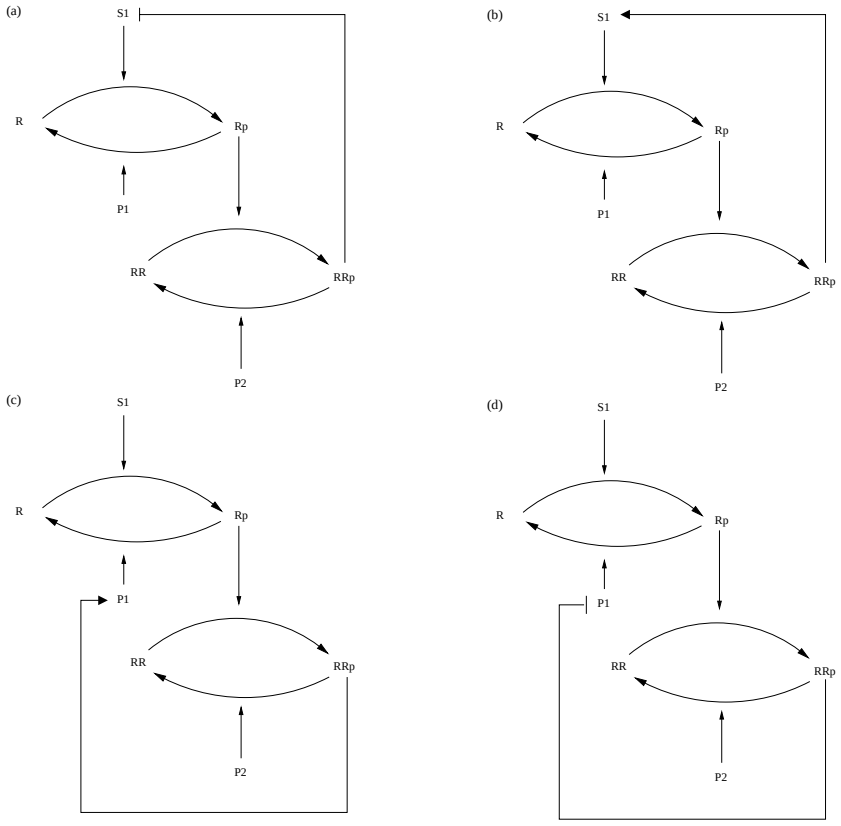

Fig. 4. Two-stage cascade with (a) negative feedback, (b) positive feedback; alternative two-stage cascade with (c) negative feedback, (d) positive feedback.

The qualitative description can be further enhanced by the abstract representation of discrete quantities of species, achieved in Petri nets by the use of tokens at places. These can represent the number of molecules, or the level of concentration, of a species, and a particular arrangement of tokens over a network is called a *marking*.

A P-invariant stands for a set of places, over which the weighted sum of tokens is constant, independently of any firing. So, P-invariants represent token-preserving sets of places. In the context of metabolic networks, P-invariants reflect substrate conservations, while in signal transduction networks P-invariants often correspond to the several states of a given species (protein or protein complex). A place belonging to a P-invariant is obviously bounded.

A T-invariant has two interpretations in the given biochemical context. The entries of a T-invariant represent a multiset of transitions which by their partially ordered firing reproduce a given marking, i.e. they occur basically one after

the other. The partial order sequence of the firing events of the T-invariant's transitions may contribute to a deeper understanding of the net behaviour.

The entries of a T-invariant may also be read as the relative transition firing rates of transitions, all of them occurring permanently and concurrently. This activity level corresponds to the steady state behaviour [PZHK05]. Independently of the chosen interpretation, the net representation of minimal T-invariants (the T-invariant's transitions plus their pre- and post-places and all arcs in between) characterize typically minimal self-contained subnetworks with an enclosed biological meaning.

In previous work [GH06] we have shown how to systematically generate initial markings from (unmarked) qualitative Petri net descriptions, which can then be used in corresponding quantitative models. This work took as a concrete example the RKIP pathway [CSK+03], which is a subset of the ERK signalling pathway.

Our approach was to systematically construct suitable initial states by P-invariants and then to check their suitability by T-invariants, which have to be feasible. In more detail, having initially created an unmarked place/transition Petri net, a systematic construction of the initial marking can be made by placing tokens on places, taking into consideration the following criteria:

- Each P-invariant needs at least one token.
- All (non-trivial) T-invariants should be feasible, meaning, the transitions, making up the T-invariant's multi-set can be fired in an appropriate order.
- Additionally, it is common sense to look for a minimal marking (as few tokens as possible), which guarantees the required behaviour.
- Within a P-invariant, choose the species with the most *inactive* (e.g. non-phosphorylated) or the *monomeric* (i.e. non-complexed) state.

In a previous paper [GH06] we created a discrete Petri net model of the RKIP pathway, and analysed the model to show that it enjoys several nice properties, among them boundedness, liveness, and reversibility. Moreover, the net is covered by P-invariants and T-invariants, all of them having sensible biological interpretation, and it fulfills several special functional properties, which have been expressed in temporal logic. Using reachability graph analysis we identified 13 strongly connected states out of 2048 theoretically possible ones, which permit self-reinitialization of the Petri net. From the viewpoint of the discrete model, all these 13 states are equivalent and could be taken as an initial state resulting in exactly the same total (discrete) system behaviour. We then transformed the discrete Petri net into a continuous model and demonstrated empirically that in the ODE model the 13 initial states, derived from the validated discrete model, result in the same (continuous) steady state. Moreover, none of the other 2035 possible initial states result in a steady state close to that derived using those identified by reachability graph analysis. This approach for steady state analysis was also successfully applied in another case study [GHL07] to a larger model of the core MAPK pathway created by Levchenko et al.[LBS00].

5 Parameter estimation using model checking

Model checking is an automated technique for the analysis of reactive systems to check whether properties, often expressed as formulas of temporal logic, hold for a model of the system. This technique was originally developed to check models of technical systems [Mer01], and has been more recently applied to biochemical networks e.g. [CF03]. In previous work [DG08] we have shown in detail how model checking based on temporal logic descriptions of behaviour can be used in parameter estimation; in this section we summarise the main results.

Temporal logic is well-suited to formally represent semi-quantitative descriptions given by biologists who are often unsure about exact values of biochemical species over time due to the nature of the wet-lab experimental technology, and will describe behaviour in a semi-quantitative manner. For example, "the concentration of the protein peaks within 2 to 5 minutes and then falls to less than 50% of the peak value within 60 minutes". A significant challenge is how to automatically build a model which conforms to semi-quantitative behaviour.

5.1 PLTLc

Linear-time Temporal Logic (LTL) [Pnu81] is the fragment of full Computational Tree Logic (CTL*) [CGP01] without path quantifiers, implicitly quantifying universally over all paths. LTL has been introduced in a probabilistic setting in [Bai98], and extended by numerical constraints over real value variables in [FR07]. PLTLc combines both extensions, complemented by the filter construct as used in Probabilistic Computational Tree Logic (PCTL) [HJ94] and Continuous Stochastic Logic (CSL) [AKVR96]. We start with the LTL with numerical constraints (LTLc) syntax:

$$\begin{aligned}\phi \ ::= \ & X\phi \mid G\phi \mid F\phi \mid \phi U\phi \mid \phi R\phi \mid \phi \vee \phi \mid \phi \wedge \phi \mid \neg\phi \mid \phi \rightarrow \phi \mid \\ & value = value \mid value \neq value \mid value > value \mid value \geq value \mid \\ & value < value \mid value \leq value \mid true \mid false\end{aligned}$$

Numerical constraints over free variables are defined in this logic through the inclusion of free variables denoted by $\$fVariable$ in the definition of *value*. Regular variables are read-only values which describe the behaviour of the model, whereas free variables are instantiated during the model checking process to the range of values for which the temporal logic property holds. Free variables are defined to have integer domains initialised to $[0 \rightarrow \infty)$ and describe protein concentrations, numbers of molecules and time. Constraints over free variables, which involve equality/inequality and relational operators, restrict the domain of the free variable.

PLTLc enhances LTLc by the inclusion of a probability operator and filter construct, and the probabilistic interpretation of the domains for the free variables. The top-level definition of PLTLc is:

$$\psi \ ::= \ \mathbf{P}_{\trianglelefteq x}[\,\phi\,] \mid \mathbf{P}_{\trianglelefteq x}[\,\phi\{SP\}\,]$$

where ϕ is an LTLc expression. SP is a State Proposition defined to be ϕ without any temporal operator (X, G, F, U, R), and containing *no* free variables without a loss of expressivity. Note that the square and curly brackets are part of PLTLc. Given that $\trianglelefteq \in \{>, \geq, <, \leq\}$, $P_{\trianglelefteq x}$ is any inequality comparison of the probability of the property holding true, for example $P_{\geq 0.5}$. We also permit the expression $P_{=?}$ returning the value of the probability of the property holding true. We disallow equality testing of the probability, $P_{=x}$ because of the representation of real values and the semantics of their equality.

The semantics of PLTLc is defined over a finite set of finite paths through the system's state space – stochastic or deterministic simulations, or time series data recorded in wet lab experiments. Let a path π be a finite sequence of states describing the behaviour of a biochemical system, $\pi = s_0, s_1, ..., s_n$ $(n < \infty)$ and π^i be the subsequence of π starting from state s_i, $i \leq n$, thus $\pi^i = s_i, s_{i+1}, ..., s_n$. Each path in the set of paths can be evaluated to a boolean value as to whether ϕ or $\phi\{SP\}$ holds. When all paths are evaluated, the number of true values in the set over the size of the set yields the overall probability of the PLTLc property. Hence for a stochastic model, where the set of paths is typically > 1, the probability is in the range $[0 \rightarrow 1]$ and calculated through Monte Carlo approximation, whereas a continuous model which contains a single path has a probability of either 0 or 1.

5.2 Distance Metrics

The distance between a model's behaviour M and the desired behaviour M_{des} with respect to some property ψ can be calculated using a distance metric.

Perhaps the simplest definition of the metric is the square difference between the model's probability of exhibiting some behavioural property ψ, $P(\psi)$ and desired probability $P_{des}(\psi)$:

$$d_{\psi}(M, M_{des}) = |P(\psi) - P_{des}(\psi)|^2$$

This approach works well in the stochastic world where the model exhibits many behaviours and the probability of the property is in the range $[0 \rightarrow 1]$. However, in the continuous world there is a single behaviour and the probability is either 0 or 1, thus the metric is too coarse grained to be used in a search algorithm in the continuous world. To be useful in the search algorithm, the distance metric should return a value which indicates whether altering the current model has caused its behaviour to be closer to the desired behaviour, therefore providing a gradient for the search algorithm to ascend. We have defined such a distance metric for continuous models using a residual sum of squares function over probabilistic domains of free variables - for more details see [DG08].

5.3 Computational System

We implemented a computational system called the Monte Carlo Model Checker with a Genetic Algorithm, MC2(GA). The purpose of this computational system is to estimate the parameters of a model to make it exhibit desired behavioural properties. A genetic algorithm is used to move models through parameter space to minimise their distance to the desired behaviour, checked using a model checker.

Each model in our MC2(GA) system has a fixed structure and is represented by a chromosome, which is a set of kinetic rate constant values to be estimated (the model's genes) within predefined ranges. The chromosome could equally include initial concentrations/masses.

In the initial generation, a population of models is created by assigning to each model random values within the ranges for the kinetic rate constants. Each model in the population is evaluated to a fitness value related to the distance of its behaviour to the desired behaviour, hence a model with a smaller distance to the desired behaviour has a higher fitness. Our approach is to vary models' kinetic rate constant values in order to maximise their fitness values.

5.4 Case Study: MAPK Pathway

The EGF signal transduction pathway conveys Epidermal Growth Factor signals from the cell membrane to the nucleus via the MAP Kinase cascade [KCG05]. The core MAPK cascade can be stimulated by both Epidermal Growth Factor (EGF) as well as Nerve Growth Factor (NGF). The reaction of the cell to EGF stimulation is cell proliferation, however the response to NGF is cell differentiation. The EGF signal transduction pathway produces transient Ras, MEK and ERK activation whereas NGF stimulation produces sustained activation. The underlying differences of the models describing EGF and NGF stimulation is of key interest to biochemists.

Work reported in [BF00], which attempted to discover the quantitative differences in initial concentrations and kinetic rate constants between models of these pathways with fixed topology. The authors in that original paper varied the initial concentrations and kinetic rate constants within biochemically sensible ranges. Simulation was performed with the model using each parameter value in the range and the output was manually inspected for sustained Ras, MEK and ERK activation. A result of this work was the finding that a 40-fold increase in the kinetic rate constant of SOS dephosphorylation can change the behaviour of the model from transient activation to sustained activation. In our approach, reported in [DG08] we showed that this analysis could be improved by constructing a formal definition of the desired behaviour in temporal logic, and using model checking of the desired behaviour to replace the manual inspection of the simulation outputs.

Characterising the Desired Pathway Behaviour The behaviour of sustained Ras, MEK and ERK activation arising from NGF stimulation observed

in wet-lab experiment was described in rather informal statements in the original paper [BF00].

"The level of RasGTP rapidly reaches a maximum of up to 20% of total Ras within 2 min [then] the level of RasGTP is sustained at around 8% of total Ras."

Similar statements were made about sustained MEK and ERK activation. We formalised these statements using semi-quantitative PLTLc such that a model could be automatically checked for these behaviours using the MC2(PLTLc) model checker. We formalised these statements in a way to account for biological error by relaxing the constraints, for example that the stable level of RasGTP is 8% to between 5% and 10%:

sustained Ras: Active Ras peaks within 2 minutes to a maximum of 20% of total Ras and is stable between 5% and 10% from at least 15 minutes

$$\mathbf{P}_{=?}\ [\ (\ d(\text{active Ras})\ >\ 0\ )\ \wedge\ (\ d(\text{active Ras})\ >\ 0\ )\ U\ (\ \text{time}\ \leq\ 2\ \wedge\ \text{active Ras}\ \geq\ 0.15{*}\text{total Ras}\ \wedge\ \text{active Ras}\ \leq\ 0.2{*}\text{total Ras}\ \wedge\ d(\text{active Ras})\ <\ 0\ \wedge\ (\ d(\text{active Ras})\ <\ 0\ \wedge\ \text{time}\ <\ 15\ )\ U\ (\ G(\ (\text{active Ras})\ \geq\ 0.05{*}\text{total Ras}\ \wedge\ \text{active Ras}\ \leq\ 0.10{*}\text{total Ras}\ )\ )\ )\ ]$$

where the protein RasGTP is found in isolation and in two complexes, thus active Ras $= RasGTP + Ras\_Raf + Ras\_GAP$ and total Ras $= RasGTP + Ras\_Raf + Ras\_GAP + RasGDP + Ras\_ShcGS$.

Genetic Algorithm We first implemented a fitness function for use in MC2(GA) to describe how close a model is to sustained activation. The descriptions of sustained Ras, MEK and ERK activation given earlier were not particularly helpful in the continuous setting due to the probability being simply 0 or 1. A fitness function based on a description which includes free variables allows greater expressivity using the probabilistic domains. Hence, we have rewritten these descriptions of sustained behaviours using free variables, for example:

sustained Ras with free variables: Active Ras peaks within 2 minutes to a maximum of 20% of total Ras and is stable between any value in $\$RasTail1$ and any value in $\$RasTail2$ from at least 15 minutes

$$\mathbf{P}_{=?}\ [\ (\ d(\text{active Ras})\ >\ 0\ )\ \wedge\ (\ d(\text{active Ras})\ >\ 0\ )\ U\ (\ \text{time}\ \leq\ 2\ \wedge\ \text{active Ras}\ \geq\ 0.15{*}\text{total Ras}\ \wedge\ \text{active Ras}\ \leq\ 0.2{*}\text{total Ras}\ \wedge\ d(\text{active Ras})\ <\ 0\ \wedge\ (\ d(\text{active Ras})\ <\ 0\ \wedge\ \text{time}\ <\ 15\ )\ U\ (\ G(\ \text{active Ras}\ \geq\ \$RasTail1\ \wedge\ \text{active Ras}\ \leq\ \$RasTail2\ )\ )\ )\ ]$$

We then applied our computational system to find novel parameter sets which exhibit the desired behaviour. We estimated the values of a set of 16 critical parameters identified as being potential candidates for modification by individually scanning all parameters and model checking the resulting simulation outputs. We also applied MC2(GA) to the critical parameters without V_28, to assess whether V_28 is crucial to achieving sustained activation. We found that if the critical parameters are estimated with V_28, then the convergence is quicker and the best model returned was fitter. The best model returned when estimating

the critical parameters had fitness value 1, whereas with V_28 removed the best model returned had a fitness value approximately 0.93.

Figure 5 shows the output of one of the best model returned when estimating the critical parameters with and without V_28. Both behaviours showed good similarity (visually and in terms of fitness value) to the behaviour of the NGF signalling pathway outlined in the original paper. We also found that we can achieve a model with fitness value 1 through a 16-fold increase of V_28, compared with the original paper's 40-fold increase, if we also vary the other critical parameters.

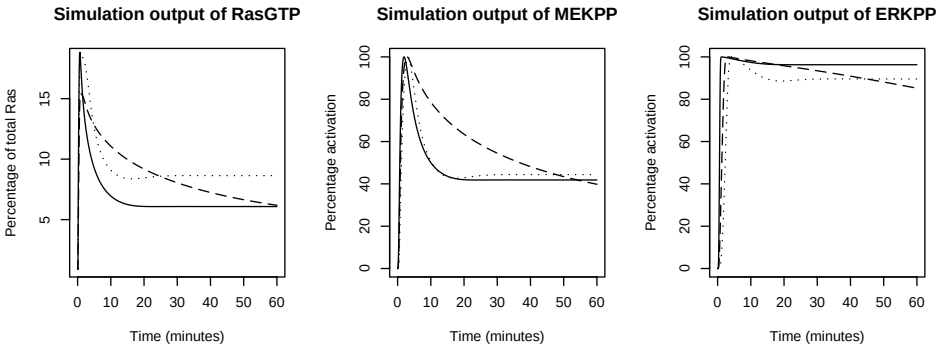

Fig. 5. The original model of the NGF signalling pathway (dotted) compared with the best model returned when varying the critical parameters (solid) and when varying the critical parameters without V_28 (dashed). The best model returned when varying the critical parameters only required a 16-fold increase in V_28 to achieve fitness value 1.

6 Conclusions

In this paper we have introduced the area of BioModel Engineering which is the science of designing, constructing and analyzing computational models of biological systems. We have illustrated some of the essential activities which BioModel Engineering encompasses - namely the construction of models of biological systems within a rigorous design framework, and techniques for the identification of initial start states, and the determination of rate parameters. We have presented these concepts in a practical manner, by way of an example in the area of intracellular signalling pathways. Our method is based on a modular building-block approach to the construction of network topology and associated biochemical equations, combined with analytical techniques from Petri nets for the determination of suitable start-states, and a novel model checking approach to drive the fitting of kinetic parameters.

Acknowledgments

DR was supported for his work on model checking by the SIMAP project which is funded by the European Commission framework 6 STREP programme.

References

[AKVR96] A. Aziz, K. Sanwal, V. Singhal, and R. K. Brayton. Verifying Continuous-Time Markov Chains. In Rajeev Alur and Thomas A. Henzinger, editors, *Eighth International Conference on Computer Aided Verification CAV*, volume 1102, pages 269–276, New Brunswick, NJ, USA, / 1996. Springer Verlag.

[Bai98] C. Baier. *On Algorithmic Verification Methods for Probabilistic Systems*. Habilitation thesis, University of Mannheim, 1998.

[BF00] F.A. Brightman and D.A. Fell. Differential feedback regulation of the mapk cascade underlies the quantitative differences in egf and ngf signalling in pc12 cells. *FEBS Lett 482*, pages 169–174, 2000.

[BGHO08] R. Breitling, D. Gilbert, M. Heiner, and R. J. Orton. A structured approach for the engineering of biochemical network models, illustrated for signalling pathways. *Briefings in Bioinformatics*, 9(5):404–421, 2008.

[CF03] N. Chabrier and F. Fages. Symbolic model checking of biochemical networks. In *Proc. CMSB 2003*, pages 149–162. LNCS 2602, Springer, 2003.

[CGP01] E.M. Clarke, O. Grumberg, and D.A. Peled. *Model checking*. MIT Press 1999, third printing, 2001.

[CSK+03] K.-H. Cho, S.-Y. Shin, H.-W. Kim, O. Wolkenhauer, B. McFerran, and W. Kolch. Mathematical modeling of the influence of RKIP on the ERK signaling pathway. *Lecture Notes in Computer Science*, 2602:127–141, 2003.

[DG08] R. Donaldson and D. Gilbert. A model checking approach to the parameter estimation of biochemical pathways. In M. Heiner and A. M. Uhrmacher, editors, *CMSB*, volume 5307 of *Lecture Notes in Computer Science*, pages 269–287. Springer, 2008.

[FR07] F. Fages and A. Rizk. On the analysis of numerical data time series in temporal logic. In *Proc. CMSB 2007*. LNCS/LNBI 4695, Springer, 2007.

[GH06] D. Gilbert and M. Heiner. From Petri nets to differential equations - an integrative approach for biochemical network analysis. In *Proc. ICATPN 2006*, pages 181–200. LNCS 4024, Springer, 2006.

[GHL07] D. Gilbert, M. Heiner, and S. Lehrack. A unifying framework for modelling and analysing biochemical pathways using Petri nets. In *Proc. CMSB 2007*. LNCS/LNBI 4695, Springer, 2007.

[HGD08] M. Heiner, D. Gilbert, and R. Donaldson. Petri Nets for Systems and Synthetic Biology. In *Schools on Formal Methods (SFM)*, pages 215–264. Springer LNCS 5016, 2008.

[HJ94] H. Hansson and B. Jonsson. A Logic for Reasoning about Time and Reliability. *Formal Aspects of Computing*, 6(5):512–535, 1994.

[KCG05] W. Kolch, M. Calder, and D. Gilbert. When kinases meet mathematics: the systems biology of MAPK signalling. *FEBS Lett 579*, pages 1891–1895, 2005.

[LBS00] A. Levchenko, J. Bruck, and P.W. Sternberg. Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties. *Proc Natl Acad Sci USA*, 97(11):5818–5823, 2000.

[Mer01] S. Merz. Model checking: A tutorial overview. *Lecture Notes in Computer Science*, 2067:3–38, 2001.

[MFD+03] H. Matsuno, S. Fujita, A. Doi, M. Nagasaki, and S. Miyano. Towards Pathway Modelling and Simulation. In *Proc. 24th ICATPN, LNCS 2679*, pages 3–22, 2003.

[Mur89] T. Murata. Petri nets: Properties, analysis and applications. *Proc.of the IEEE 77*, 4:541–580, 1989.

[OSG+08] R. Orton, O. E. Sturm, A. Gormand, W. Kolch, and D. Gilbert. Computational modelling reveals feedback redundancy within the epidermal growth factor receptor/extracellular-signal regulated kinase signalling pathway. *Systems Biology*, 2:173–183, 2008.

[Pnu81] A. Pnueli. The Temporal Semantics of Concurrent Programs. *Theor. Comput. Sci.*, 13:45–60, 1981.

[PZHK05] L. Popova-Zeugmann, M. Heiner, and I. Koch. Time Petri Nets for Modelling and Analysis of Biochemical Networks. *Fundamenta Informaticae*, 67:149–162, 2005.

[SEJGM02] B. Schoeberl, C. Eichler-Jonsson, E.D. Gilles, and G. Muller. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature Biotechnology*, 20:370–375, 2002.