Journal of Bioinformatics and Computational Biology
Vol. 10, No. 4 (2012) 1231001 (29 pages)

DOI: 10.1142/S0219720012310014

Imperial College Press
www.icpress.co.uk

MODELING AND ANALYSIS OF BIOPATHWAYS DYNAMICS

BING LIU* and P. S. THIAGARAJAN†

Department of Computer Science
National University of Singapore
Computing 1, 13 Computing Drive
Singapore 117417
**liubing@comp.nus.edu.sg*
†*thiagu@comp.nus.edu.sg*

Received 10 October 2011
Revised 15 February 2012
Accepted 15 February 2012
Published 3 April 2012

Cellular processes are governed and coordinated by a multitude of biopathways. A pathway can be viewed as a complex network of biochemical reactions. The dynamics of this network largely determines the functioning of the pathway. Hence the modeling and analysis of biochemical networks dynamics is an important problem and is an active area of research. Here we review quantitative models of biochemical networks based on ordinary differential equations (ODEs). We mainly focus on the parameter estimation and sensitivity analysis problems and survey the current methods for tackling them. In this context we also highlight a recently developed probabilistic approximation technique using which these two problems can be considerably simplified.

Keywords: Systems biology; pathway modeling; parameter estimation; model analysis.

1. Introduction

Cellular processes are driven by networks of biochemical reactions. These networks are often termed *biopathways*, and they can be loosely classified into signaling pathways, metabolic pathways, and gene regulatory networks. Cells rely on the tight coordination of these pathways to achieve proper functioning. Here we mainly focus on signaling pathways, though the models and techniques we present can be applied to other types of networks as well. With the help of signaling pathways, a cell senses changes in its environment or internal state. This information is then passed on via cascades of biochemical reactions to the appropriate mechanisms which respond by modifying the metabolic and transcriptional activities. This in turn modifies the behavior of the cell.

Consequently, the *dynamics* of biopathways play a crucial role in determining cellular functions. A nice example is the biopathway controlling the circadian

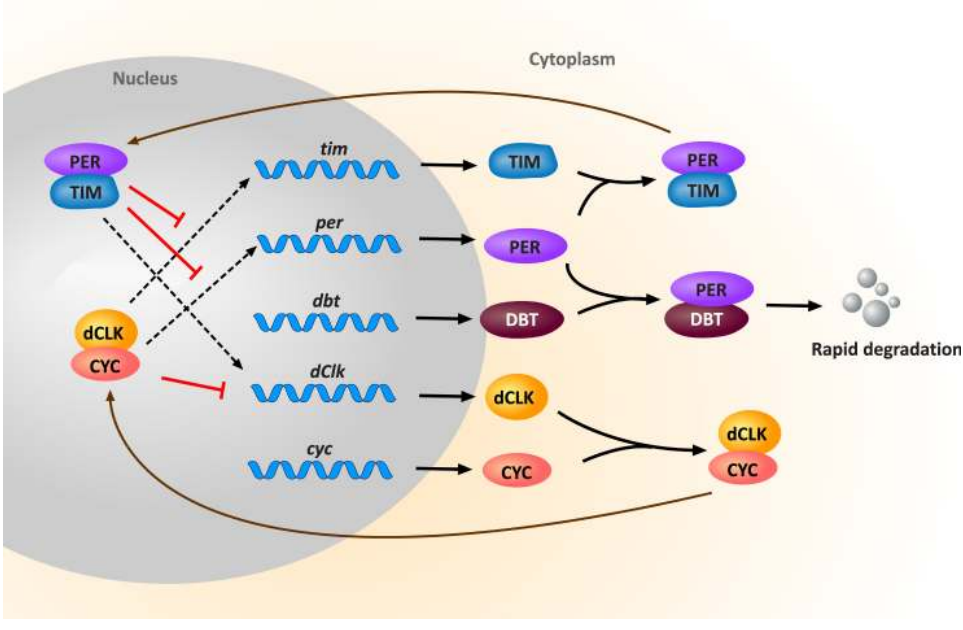


Fig. 1. The *Drosophila* circadian rhythm pathway model.[2]

rhythm.[1] It arises from the oscillatory expression levels of a number of genes and the periods of the oscillations roughly equal 24 h. Figure 1 depicts the *Drosophila* circadian rhythm pathway composed of interlocking feedback loops that regulate the concentrations of the relevant transcription factors which in turn control the expression levels of many other genes which then leads to physiological rhythms.[2]

Other well-known examples are: the apoptosis pathway inducing programmed cell death,[3] the EGF-NGF pathway determining the choice between cell differentiation and cell proliferation,[4] the Wnt signaling pathway governing the expressions of developmental genes,[5] and the NF-κB pathway which regulates inflammatory responses.[6] There are literally hundreds of such pathways associated with basic cellular functions and many diseases arise due to the malfunctioning of signaling pathways.[7,8] For example, a dysfunction in the apoptosis pathway can lead to cancer.[9] Misregulation of the Wnt pathway can lead to a variety of degenerative diseases such as Alzheimer's disease.[10]

Thus the systematic study of the dynamics of biopathways is a crucial task and has given rise to a rich body of research. Our goal here is to highlight a restricted but significant portion of this research.

1.1. *Plan and scope of the paper*

A rich variety of mathematical formalisms have been developed to study the dynamics of biopathways. It is next to impossible to address even a significant portion of them in

a tutorial paper. Hence we shall focus here on one major mathematical formalism, namely Ordinary Differential Equations (ODEs). Even here, a thorough survey of all the facets, including recent developments, is too formidable a task. Hence we shall concentrate on a major barrier that ODE-based approaches must overcome, namely, the parameter estimation problem. In addition, from among the many analysis techniques that have been — and are being — developed for ODEs-based models, we shall address the key technique of sensitivity analysis. Following this, we shall describe a novel probabilistic approximation technique that we have developed — in collaboration with David Hsu[11] — using which both the parameter estimation and sensitivity analysis tasks can be considerably eased.

In the next section we introduce the notions of model construction, calibration, validation and analysis. In Sec. 3, we provide a brief overview of the prevalent modeling formalisms. In Sec. 4, we discuss model calibration, also known as parameter estimation. This is followed by a brief discussion of model validation in Sec. 5. In Sec. 6, we turn to important model analysis techniques with the main focus on sensitivity analysis. Finally, we present a recently developed probabilistic approximation method by which a system of ODEs modeling a biochemical network is approximated as a dynamic Bayesian network,[11]. In the concluding section, we summarize the contents of the paper and sketch some future research directions.

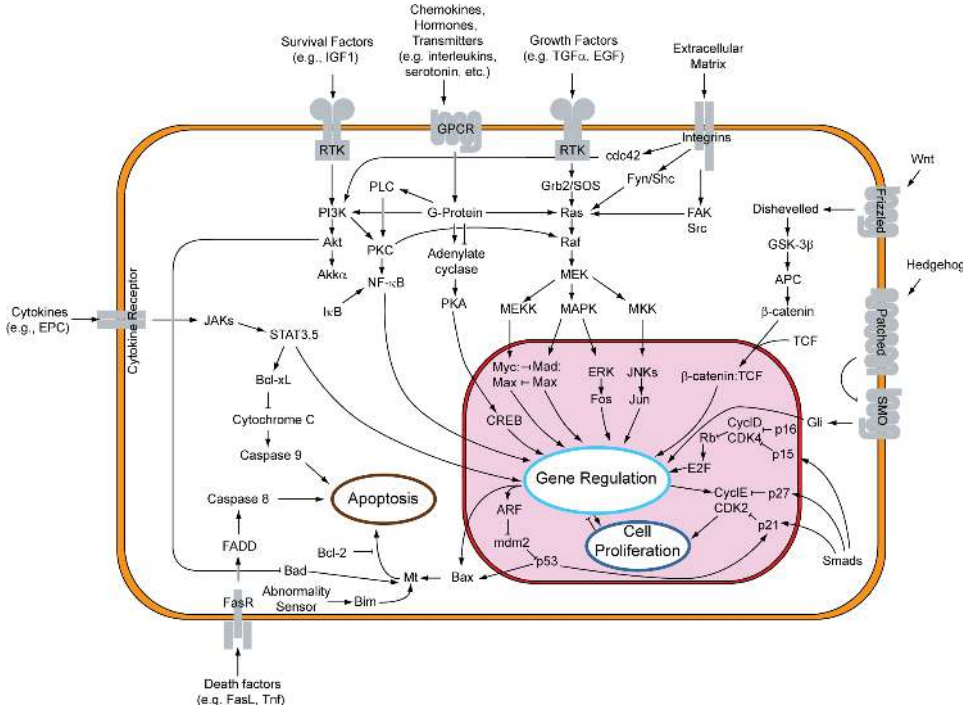


Fig. 2. Overview of some of the important signaling pathways.[8]

2. Biopathway Modeling

A variety of modeling approaches have been proposed in recent years to study the dynamics of signaling pathways.[12,13] The Boolean network model is often used in qualitative studies,[14] while typical quantitative formalisms are ODEs,[15] Petri nets of various kinds,[2,16] and process algebra−based languages such as κ and PRISM.[17,18] Regardless of the type of the formalism used, a typical modeling effort involves the following steps:

(1) **Model construction.** Fix the model scope and accordingly build the network of the major players and their interactions.
(2) **Model calibration.** Divide the available experimental data relevant to the dynamics into training data and test data. Then calibrate (i.e. estimate) the unknown model parameters so that the calibrated model reproduces the training data well.
(3) **Model validation.** Confirm the predictive abilities of the calibrated model by matching the behaviors produced by the calibrated model to the test data. An obvious requirement is that the model must be validated using data that was not used for training it.
(4) **Model analysis.** Perform various analysis tasks on the validated model in order to gain biological insights and generate interesting hypotheses.

In Step 1, an initial model is usually constructed based on the literature and pathway databases such as Reactome.[19] In collaborative efforts, one also depends crucially on the guidance of biologists. In Steps 2 and 3, the experimental data will often consist of the time series of species concentrations. Step 2 is also known as the parameter estimation step, which we discuss in more detail in Sec. 4. In practice, model construction will involve a cyclic workflow. In Step 2, if one is unable to estimate parameters which fit the training data well, one may have to go back to Step 1 and refine the model structure and add further structural details that had been left out. Similarly, for Step 3, if the model cannot be validated, one will have to go back to Step 1 and refine the model. One could also try to acquire more experimental data concerning the structure and dynamics. If we still cannot get through Step 2 and Step 3 this could be the basis for proposing missing links, cross-talks, feedback loops, etc. which can then guide further experiments.

3. Modeling Formalisms

We now review some well-established quantitative models. They can be *stochastic* or non-stochastic and in particular, *deterministic*. In practice, deterministic models are used more frequently due to their simplicity and scalability. However, when the concentrations of species are low, the variability of reaction processes will increase and may significantly influence the system's behavior. For example, the development of phage λ infected *Escherichia coli* cells is determined by a switch point. Two *small*

quantities of proteins competitively control this switch. As a result, the developmental outcome is probabilistic and a deterministic model in this setting leads to misleading conclusions.[20]

3.1. *Stochastic models*

Stochastic modeling involves tracking the number of molecules of each type in a chemical mix, where these numbers change as the mix evolves through chemical reactions. A basic assumption here is that one is working with a well-stirred mixture in a fixed volume and that the system is in thermodynamic (but not chemical) equilibrium. Consequently one need not track the locations and velocities of the individual molecules.

In stochastic modeling, one describes the state of the system by a vector $X(t) = (X_1(t), X_2(t), \ldots, X_N(t))$, where $X_i(t)$ is the *number* of molecules of species i at time t. The states of the system evolve according to the sequence of reactions that takes place starting from the initial state $X(0)$. Which reaction will occur next and when it will occur are both determined probabilistically. This induces, for each time point, a probability distribution over the possible system states. The Chemical Master Equation (CME), a stochastic differential equation, can be used to capture the time evolution of this probability distribution.[12] However, except for trivially small systems, the CME does not admit analytical solutions. Hence one must resort to stochastic simulations. This can be computationally very expensive due to the fact that the CME specifies a probability density function over the system states and the number of system states will be in general exponential in the number of species. Gillespie's algorithm and its variants are often used for stochastic simulations.[21–25] By examining the statistical properties of the resulting set of trajectories one may infer — approximately — properties of the stochastic dynamics specified by the CME. In the recent past, efficient techniques to numerically solve the CME directly have also been studied.[26]

A number of languages have been developed to describe the stochastic dynamics of a biochemical network. A prominent example is the κ language.[17,27] It can compactly and appealingly describe large biopathways. The basic idea is to use *agents* to describe the players in the pathway such as proteins, protein complexes and genes respectively. Each agent will have a number of sites with associated internal states that can represent the status of post-translational modifications (e.g. phosphorylation) or bindings with other agents. One then formulates *rules* to specify how the states of the agents are modified by the reactions. The rules are sufficiently expressive to capture bindings, dissociations, and modifications to the state of a site as well as the production and degradation of molecular species.

As an example, consider a system consisting of a protein P with two phosphorylatable sites x and y and an enzyme E with a binding site z. Correspondingly, there will be two agents, $P(x, y)$ and $E(z)$. The internal states of sites x and y will be denoted as 'u' (unphosphorylated), and 'p' (phosphorylated). Initially $P(x \sim u, y \sim u)$

where $x \sim u$ says that the site x of $P(x, y)$ is currently unphosphorylated. Suppose E can bind to P at x or y and catalyze the phosphorylation the site and then unbind from P. This is captured by the following rules:

$$E(z), P(x) \leftrightarrow E(z!1), P(x!1) \tag{1}$$
$$E(z!1), P(x \sim u!1) \rightarrow E(z!1), \quad P(x \sim p!1) \tag{2}$$
$$E(z), P(y) \leftrightarrow E(z!1), P(y!1) \tag{3}$$
$$E(z!1), P(y \sim u!1) \rightarrow E(z!1), \quad P(y \sim p!1) \tag{4}$$

where "!" denotes the binding and "1" is used to identify the binding pair. Note that the rules follow the "*don't care, don't write*" convention, e.g. in rule (1), site y and the internal state of site x are left unspecified in P's interface.

Given the initial number of copies of each agent, the time evolution of the system can be generated through stochastic simulations using a rule-based variant of Gillespie's algorithm. For further details we refer the reader to Danos *et al.*[28]

In the stochastic modeling approaches discussed above, the dynamics of a biochemical network is given by an underlying Continuous Time Markov Chain (CTMC). Thus a κ model may be viewed as a succinct and understandable description of a large and complex CTMC. Languages such as PRISM and PEPA take a similar approach but with a less elaborate syntax.[29,30]

PRISM was originally created to aid the formal verification of CTMC models arising in various domains.[29] Subsequently it has also been used to model and analyze biopathways.[18,31] The modeling language uses *variables* and *modules*. In biopathway applications, the values of variables will represent the discrete concentration levels of species. A module contains a number of variables together with update rules for modifying them. Each rule describes how the values of variables involved in a reaction are updated under particular conditions. PRISM allows dynamical properties of the system under study to be specified as formulas taken from various temporal logics such as Linear Temporal Logic (LTL),[32] Probabilistic Continuous Temporal Logic (PCTL),[33] and Continuous Stochastic Logic (CSL).[34] The PRISM tool will then verify (or falsify) in an automated fashion whether the PRISM model satisfies the property.[35] Here is an example of a PCTL formula:

$$(A < 2) \Rightarrow \mathbf{P}_{>0.2}[true\mathbf{U}^{[0,4]}(AB = 3)] \tag{5}$$

It says "if protein A's concentration level is currently less than 2, then the probability of the complex AB's concentration level being 3 within the next 4 seconds is greater than 0.2."

A common limitation of stochastic models is scalability. Since stochastic simulations are computationally intensive, realistic pathways cannot be handled. For instance, verifying the PRISM model of the ERK pathway, which consists of just 11 species, requires the computational power of a 90-node grid.[36] An equally serious problem is that in stochastic modeling (as in the ODE-based approaches we shall soon address), each reaction will have a rate constant associated with it. Roughly

speaking this constant captures the probability (often described as an exponential distribution over time) of the reaction occurring at time Δt starting t. In all the approaches we have mentioned, these rate constants are assumed to be known. If they are not, which is often the case, the simulations are carried out after fixing the rate constant values more or less arbitrarily. Encouragingly, this fundamental parameter estimation problem is beginning to be tackled.[37–40]

3.2. *Deterministic models*

We now turn to non-stochastic models. The first one to be considered, namely ODEs, will play a prominent role in the subsequent parts of the paper.

3.2.1. *Ordinary differential equations*

An implicit assumption in this setting is that in the pathway under study, all the molecular species in the pathway are abundantly available. The idea then is to use an ODE to capture the concentration level changes of a molecular species using the reactions it takes part in as a reactant or product. The formulation of the ODE is guided by the kinetic law governing the reaction.[15] For example, assuming that the reactant molecules are spatially homogeneous, the mass action law states that the rate of a reaction is proportional to the concentrations of reacting species. A reversible binding process of two species can now be described using mass action law as follows:

$$A + B \underset{v_2}{\overset{v_1}{\rightleftharpoons}} AB \tag{6}$$

where A and B are substrates, AB denotes the formed complex, and v_1 and v_2 represents the association rate and dissociation rate respectively. By the mass action law, we have:

$$\begin{aligned} v_1 &= k_1 \cdot A \cdot B \\ v_2 &= k_2 \cdot AB \end{aligned}$$

where k_1 and k_2 are so-called rate constants. This leads to the system of ODEs:

$$\begin{aligned} \frac{dA}{dt} &= -k_1 \cdot A \cdot B + k_2 \cdot AB \\ \frac{dB}{dt} &= -k_1 \cdot A \cdot B + k_2 \cdot AB \\ \frac{dAB}{dt} &= k_1 \cdot A \cdot B - k_2 \cdot AB \end{aligned}$$

On the other hand, the enzyme catalyzed reactions such as protein phosphorylation are often modeled using Michaelis–Menten kinetics. Equation (7) below shows the reaction network involving a simple enzyme catalyzation.

$$S + E \overset{v}{\rightarrow} P + E \tag{7}$$

where S denotes substrate, E denotes enzyme, P denotes product, and v is the reaction rate. Under suitable assumptions, the kinetics of this reaction scheme is expressed by the Michaelis−Menten equation:

$$v = \frac{k \cdot S \cdot E}{K_m + S} \tag{8}$$

where k and K_m are constants. This then leads to the ODE:

$$\frac{dP}{dt} = \frac{k \cdot S \cdot E}{K_m + S} \tag{9}$$

Depending on the biochemical context one may also use many other kinetic laws such Hill equation, etc.[13] Consequently, a biopathway can be modeled as a system of ODEs of the form:

$$\frac{dx_i}{dt} = f_i(\mathbf{x}(t), \mathbf{p}) \tag{10}$$

where the vector $\mathbf{x}(t)$ represents the concentrations of species at time t, and the vector $\mathbf{p}$ refer to the rate constants of the reactions. Given the initial values of the variables, the values of the rate constants and suitable continuity assumptions, a system of ODEs will have a unique solution.[41] Hence in principle ODE-based models can be used to predict system behavior by solving a standard *initial value problem*. However, the ODE systems describing biopathway dynamics will be high-dimensional and nonlinear and hence they will not admit closed form solutions. Instead, one will have to resort to numerical integration methods to get approximate solutions. A standard approach is to use a *finite difference method* to numerically approximate the solution of a system of differential equations. To illustrate the idea consider:

$$x'(t) = \lim_{\delta \to 0} \frac{x(t+\delta) - x(t)}{\delta}, \tag{11}$$

then a reasonable approximation of the derivative at t would be

$$x'(t) \approx \frac{x(t+\delta) - x(t)}{\delta} \tag{12}$$

for a sufficiently small δ. Since $x'(t)$ is known, given the initial condition $x(0)$, we can iteratively compute $x(t)$ for any t as follows:

$$x(t+\delta) = x(t) + \delta \cdot x'(t) \tag{13}$$

This is the so-called *explicit Euler's method*. δ must be very small to ensure high accuracy as well as stability. Accordingly, if T is the maximal time point of interest then $\frac{T}{\delta}$, the required number of simulation steps will be a large number. This is especially so for *stiff* ODE systems, in which the fast rates of changes of some variables — in comparison to others — require the step size to be very small. Many ODE models of biochemical networks will be high dimensional and stiff and hence

numerically solving them is a computationally demanding task. In the past decades, many advanced ODE solvers have been developed to improve the performance of numerical integration. Different solvers are usually specialized for better performance on specific classes of ODEs. To deal with the ODE-based biopathway models, methods such as Runge−Kutta and LSODA are often used.[42,43]

A number of techniques have been proposed to reduce the complexity of ODE-based models. An obvious tactic is to include in the model only the species necessary for the analysis task at hand. One can further simplify the model through abstractions justified by suitable assumptions. In fact, the Michaelis−Menten equation of the simple enzyme catalyzation reaction above (Eq. (9)) is obtained by abstracting mass action kinetics of the reaction scheme shown in Fig. 4(a) by assuming (i) the concentration of substrate is much larger than the concentration of enzyme and (ii) the reversible pair of reactions proceed much faster than the irreversible reaction.[13] This idea has been extended to deal with any kinetic law that can be written as a fraction of two polynomials.[44] As a result, a complex rate equations such as the one below:

$$v_{\text{original}} = \frac{V\left(\frac{[F16bP]}{K_{F16bP}} - \frac{[DHAP][GAP]}{K_{F16bP}K_{Keq}}\right)}{1 + \frac{[F16bP]}{K_{F16bP}} + \frac{[DHAP]}{K_{DHAP}} + \frac{[GAP]}{K_{GAP}} + \frac{[F16bP][GAP]}{K_{F16bP}K_{GAP}} + \frac{[DHAP][GAP]}{K_{DHAP}K_{GAP}}} \tag{14}$$

can be simplified as:

$$v_{\text{simplified}} = \frac{K_2[F16bP]}{1 + K_1[F16bP]} \tag{15}$$

Another technique is to consider restricted class of ODEs. Specifically, piecewise-affine systems of ODEs have been used to model gene regulatory networks.[12,45] The restriction to piecewise-affine differential equations is guided by a number of factors. Firstly, the rate of activation of a gene often follows a steep sigmoidal curve,[12] much like a step function. Secondly, this class of differential equations allows the qualitative properties of the system such as reachability, stability, and oscillations to be analyzed without performing numerically integration.[47,48] Third, one does not need to know the exact values of the rate constants in order to qualitatively analyze the system. Further exploitation of multi-affine systems can be found in Batt *et al.*[45]

3.2.2. *Petri nets*

Petri nets, proposed by Carl Adam Petri, is a mathematical model for the representation and analysis of distributed computing systems.[49] A Petri net consists of three primitive elements — places, transitions, and directed arcs. In the context of biopathway modeling, places will denote species while transitions represent the biochemical reactions. The places are connected to the transitions (and vice versa) via directed arcs to form the network. In the graphical representation, places are drawn as circles; transitions are denoted by bars or boxes. The input places of a

transition are the places from which there is an arc to it; its output places are those to which an arc goes from it.

Places may carry a non-negative number of tokens, which are represented as block dots inside the corresponding place. A distribution of tokens over the places of a net is called a *marking* denoting the current state of the system. A transition is enabled to fire at a marking of all its input places carry at least one token at the marking. When a transition fires, it consumes one token from each of is input places and adds one token to each of its output places.

Example 1. Figure 3 shows a Petri net model of the enzyme catalysis system. In this example, the places E, S, P denote the enzyme, product and substrate, respectively. The transition T represents the enzyme catalyzed reaction. The number of tokens depicts the concentration level of a species. The initial marking is shown in the left panel of Fig. 3. Transition T is enabled. After firing T once, the resulting marking is shown in the right panel of Fig. 3.

Many variants of Petri nets have been developed over the years to fit modeling requirements arising in different fields.[50] Timed Petri nets, stochastic Petri nets, hybrid Petri nets, functional Petri nets, and hybrid functional Petri nets have been deployed for describing the dynamics of biopathways.[51] For instance, Ruths *et al.* studied a MAPK and AKT signaling network downstream from EGFR in two breast tumor cell lines using a stochastic Petri net model.[16] Bonzanni *et al.* used a coarse-grained quantitative Petri net to mimic the multicellular process of *Caenorhabditis elegans* vulval development.[52] The hybrid functional Petri net is a powerful extension which can capture both the discrete and continuous features of pathway dynamics. This variant has been implemented in a software tool called Cell Illustrator,[53] which has been used to analyze a number of biopathways.[54,55]

4. Model Calibration

The next important step in the model construction process is model calibration. The quantitative formalisms we have described above will induce a large number of parameters, namely the rate constants associated with the reactions and the initial concentrations of the various species. Usually, the values of only a few of them will be available in literature or can be directly measured experimentally. The remaining

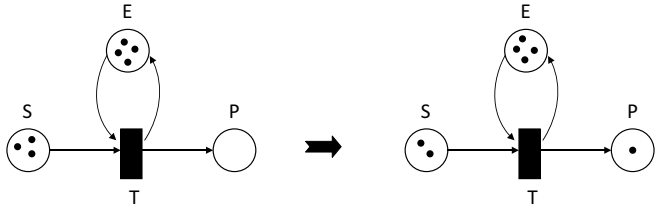


Fig. 3. A Petri net example of the enzyme catalysis system.

ones will have to be estimated using experimental data. This is the parameter estimation problem and it arises in both stochastic and deterministic settings. Here, we shall mainly focus on this problem in the context of the ODEs formalism. As pointed out earlier, less work has been done on this problem in stochastic setting.

The goal of parameter estimation is to compute the values of unknown parameters so that the resulting model can reproduce the experimental observations. A common approach is to iteratively optimize the agreement between the model prediction and available experimental data with the help of search techniques. Typically, the goodness-of-fit of a parameter combination is evaluated by the weighted sum of square error between model prediction and experimental data captured by the following objective function:

$$f_{obj}(\mathbf{p}) = \sum_{i,j} \omega_i (x_{i,j} - y_{i,j}(\mathbf{p}))^2 \tag{16}$$

where $\mathbf{p}$ is the parameter set being tested, $x_{i,j}$ is the experimental observation of the concentration of species x_i at time point j, $y_{i,j}(\mathbf{p})$ is the corresponding prediction generated using $\mathbf{p}$, and ω_i is the normalization factor for x_i which is usually the inverse of the maximum value of x_i.

In order to find the parameter set $\mathbf{p}_{opt}$ that has the minimum objective value, an optimization algorithm repeatedly executes two steps: (1) guess the values of the parameters; (2) evaluate the goodness-of-fit of the guesses. For step (1), guesses may be generated randomly in the first round but later guesses are usually guided by the results of previous rounds. For step (2), to get the value of $y_{i,j}$ in Eq. (16), one will have to simulate the ODE system up to the maximum time point for which experimental observations are available. Consequently, for high-dimensional models parameter estimation is a difficult problem with no guarantees of success. The algorithm is terminated if a sufficiently good fit to data has been achieved or if the computational resources allocated for the task (typically running time) have been exhausted.

A critical issue is how to make "clever" new guesses based on guesses that have already been evaluated. In other words, how to search the solution space so that the optimal solution — or a good approximation of it — can be found as fast as possible. Different algorithms use different search methods. For instance, the Steepest Descent[56] method follows the direction of steepest descent on the hyper surface of the objective function. The Levenberg−Marquardt method combines this idea with the Newton methods.[57] The Hooke and Jeeves (HJ) method remembers the descent directions of previous searches and suggests a new direction to search.[58] These methods are usually classified as the *local methods*. In practice, they converge quite fast. However, they suffer from the problem of getting trapped in local minima and often return suboptimal solutions.[59]

To overcome this, a number of global methods have been proposed. For example, algorithms such as the Genetic and Evolutionary Strategy (ES) algorithms try to maintain a population of candidate solutions.[60,61] In this population, biologically

begin
  Initialize parent population $\mathbf{P}_\mu = \{\mathbf{p}_1, \ldots, \mathbf{p}_\mu\}$
  repeat
    for $i \leftarrow$ *1 to* λ **do**
      $\mathcal{S} \leftarrow \mathbf{P}_\mu$
      Randomly select parents $\mathbf{p}_{c1}, \mathbf{p}_{c2} \in \mathbf{P}_\mu$
      $\mathbf{p}_{new} \leftarrow \text{Recombine}(\mathbf{p}_{c1}, \mathbf{p}_{c2})$
      $\mathbf{p}_{new} \leftarrow \text{Mutate}(\mathbf{p}_{new})$
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathbf{p}_{new}\}$
      Sort($\mathcal{S}$)
      $\mathbf{P}_\mu \leftarrow$ Select first μ from $\mathcal{S}$
    end
  until Stopping Criteria ;
end

Algorithm 1. $(\mu + \lambda)$-ES.

inspired operations such as recombination, crossover and mutation are performed on selected members to produce the next *generation* of solutions. The idea of ES is illustrated in Algorithm 1. In each iteration, λ children solutions will be generated from μ parent solutions. Each child is obtained from two parents by random crossover of the parameters, or by using the midpoints of their respective parameter values. The best μ solutions from the combined set of parent and children solutions are selected to be the parents for the next generation. The entire process then repeats itself until no better solutions can be found, or the specified maximum number of generations is reached.

Another global method called Particle Swarm Optimization (PSO) is inspired by a flock of birds or a school of fish searching for food.[62] It maintains a population (*swarm*) of candidate solutions whose members are called *particles*. Each particle has a position and velocity associated with it. The particles are moved around in the solution space iteratively by changing their positions and velocities. In each round of the algorithm, the movement of a particle is determined by its velocity which in turn is guided by its best known position as well as the entire swarm's best known position (relative to the global objective function).

Comparisons of the global methods can be found in Moles *et al.*[59] A variant of ES called Stochastic Ranking Evolutionary Strategy (SRES) appears to outperform many commonly used global methods.[63] Further attempts to improve the performance of SRES have been reported in recent papers.[64,65]

For signaling pathway models, Koh *et al.* proposed a decompositional approach for high-dimensional models.[66] By exploiting the structure of a pathway and the distribution of available experimental data, the global model is decomposed into components and parameter estimation is performed for each component separately using the SRES method mentioned above. The key point is that a component model will be often much smaller than the global model. However, independent estimates

from different components for a shared parameter may be in conflict. To reconcile these conflicts, each component is represented as a factor graph, a standard probabilistic graphical model. The resulting factor graphs are then combined using the probabilistic inference technique called belief propagation to obtain the maximally likely parameter values that are globally consistent.[67] The belief propagation technique has also been used by Koh *et al.* to update parameter estimates when new experimental data becomes available.[68]

4.1. *Model identifiability*

In case it is not — even theoretically — possible to estimate a unique set of parameters from the given observations, the model is said to be *non-identifiable*. In the present context, it is tempting to assume that there is a unique set of "true" parameters associated with a biopathway model and to expect that a parameter estimation procedure will find them. In practice, this will be possible only for small idealized models. For large models accompanied by limited and noisy experimental observations, multiple sets of parameter values will fit the training data.[69–71] This is so since the landscape of the high-dimensional search space may have many optimal valleys rather than a broad funnel leading to a single optimum.[72]

Inadequate and noisy experimental data is a major cause of non-identifiability in practical settings. For instance, we recall the following simple system of ODEs describing the reversible binding of two species shown in Sec. 3.2.1.

$$\begin{aligned}\frac{dA}{dt} &= -k_1 \cdot A \cdot B + k_2 \cdot AB \\ \frac{dB}{dt} &= -k_1 \cdot A \cdot B + k_2 \cdot AB \\ \frac{dAB}{dt} &= k_1 \cdot A \cdot B - k_2 \cdot AB\end{aligned}$$

If the only data available is regarding the steady state values of A, B, and AB, then we can only determine the ratio of k_1 and k_2 and not the individual values of k_1 and k_2.

Thus, in order achieve identifiability, we need to measure more species, with higher sampling frequency, under more stimulation conditions, and as accurately as possible. This can be a very expensive and even infeasible proposition. To mitigate this problem, strategies for optimally selecting species and the time points at which to measure their concentrations have been proposed.[73,74] It is also worth noting that many non-identifiable models can be partially identifiable.[75] In other words, a subset of the parameters may be identifiable and the resulting model might yield meaningful information about the system dynamics. For example, crucial insights about the ErbB signaling have been gained using a partially identified model.[76] Finally, as a pragmatic approach, one can maintain maximal likelihood estimates of parameter values that are consistent with the current model and its experimental data. One can

then update these estimates in a principled manner using Bayesian inferencing methods when new data becomes available.[68]

5. Model Validation

In many existing approaches, the model construction process is terminated after the calibration step. However, it is important to assess whether the model has been "over-fitted" in that the calibrated model can only predict the training data and not much else. To ensure this, test data must be available consisting of experimental observations that were *not* used for estimating the parameters. Apart from quantitative data, it can also include qualitative observations such as oscillations and bistability.[77] Using such data, the calibrated model must be simulated to see if it produces good fitness with the test data.

6. Model Analysis

Assuming that an ODE-based model has been constructed, calibrated and validated, we turn to some of the major analysis methods that can be applied.

6.1. *Sensitivity analysis*

Sensitivity analysis studies how variations in the *input* affects the *output* of the model.[78] Here the input can be the initial state or parameters of the model and the output the time profile of a chosen species. Sensitivity analysis is a powerful technique and can be used for[79]: (i) drug target selection,[80] (ii) biomarker selection,[81] (iii) experiment design,[65] (iv) model reduction,[82] and (v) robustness analysis.[83]

The basic idea is to define the sensitivity coefficient s_{ij} to be the normalized first order derivative of the model output o_i with respect to the model parameter p_j:

$$s_{ij} := \frac{\partial o_i}{\partial p_j} \cdot \frac{p_j}{o_i} \equiv \frac{\partial \ln(o_i)}{\partial \ln(p_j)} \tag{17}$$

Centered difference approximation techniques can be employed to compute sensitivity coefficients s_{ij} as follows[84]:

$$s_{ij} = \frac{\partial o_i}{\partial p_j} \cdot \frac{p_j}{o_i} \approx \frac{o_i(p_j + \Delta p_j) - o_i(p_j - \Delta p_j)}{2\Delta p_j} \cdot \frac{p_j}{o_i} \tag{18}$$

Often p_j is a rate constant or the initial concentration of a species and o_i is a characteristic of the system response. For instance, we may define o_i to be the transient concentration of a particular species (usually the endpoint of signal transduction) at a specific time point t.[85] In this case, the sensitivity s_{ij} will become time dependent and can be denoted as $s_{ij}(t)$. One may then plot $s_{ij}(t)$ and further investigate how the sensitivities evolve over time.[86] Furthermore, depending on the dynamical properties of the system being studied, many other characteristics of the output response can be used such as: the amplitude and time of the response peak,

the duration of the response,[87] the integration of the response curve,[88] the amplitude, period and phase of oscillation,[86,89,90] the steady-state levels,[91] and the deviation from the observations.[92]

Local sensitivity analysis as described earlier, assesses the effects of perturbations within a small local region around a specific point in parameter space. However, as discussed earlier, the values of many parameters have to be estimated from noisy and limited data. Hence it is possible for local sensitivity analysis to yield different conclusions based on different sets of estimated values. In addition, changes in cellular environments may induce variations of model parameters which in turn will lead to different local sensitivities. Therefore, it is important to do sensitivity analysis in a more *global* manner by exploring the effects of perturbations within a large region of parameter space.

6.1.1. *Global sensitivity analysis*

Various global methods have been recently developed for biopathway models.[78,79] These methods assess the overall effects of parameters on the model output by simultaneously perturbing all the parameters within a parameter space. A common Monte Carlo scheme adopted by many of them can be described as follows:

(1) draw a representative number of samples from the parameter space,
(2) simulate the system for each sampled combination of parameters,
(3) derive the global sensitivities of parameters by a statistical or information theoretic analysis of the simulation results.

In step (3), the global sensitivities are measured in different ways depending on the method used. For instance, the partial rank correlation coefficient (PRCC) analysis calculates the global sensitivities from the Pearson correlation coefficients between model output and input parameters.[93] The global sensitivities calculated by Bentele *et al.* is a weighted average of the local sensitivities of sampled values of parameters, where the weights are determined by a Boltzmann distribution function of the error between model simulation and experimental data.[82] Sobol's method estimates the partial variances of the model output for input parameters and defines the global sensitivities as the ratio of the related partial variances to the overall variance of the model output.[94] In multi-parametric sensitivity analysis (MPSA), the sampled parameter sets are classified into two classes based on the objective value of each sample, which measures the error between experimental data and prediction generated by selected parameters.[92] The global sensitivities are then evaluated as the Kolmogorov−Smirnov statistic of cumulative frequency curves of the parameter values associated with the two classes. There have also been attempts to derive global sensitivities via information theoretic analysis. For example, Ludtke *et al.* treated the pathway system as a 'communication channel' and quantified the associations between input parameters and model output by decomposing their mutual information.[95] More information on global sensitivity analysis can be found in Saltelli's book.[78]

For large models, step (1) of the above scheme will require a large number of samples. Consequently, carrying out global sensitivity analysis is very time consuming. To get around of this, efficient sampling methods have been proposed. For instance, Latin hypercube sampling (LHS) is a method requiring fewer samples while guaranteeing that individual parameter ranges are evenly covered.[96] Instead of random sampling, heuristic sampling using optimization algorithms has also been used.[97] Finally, Zhang and Rundell have advocated the reuse of the computational effort invested during parameter estimation to improve the performance of global sensitivity analysis.[96]

6.2. *Perturbation optimization*

The control of cellular mechanisms by means of genetic modifications or drug treatment is an important goal. Many applications are based on such a strategy ranging from therapeutic[98] to metabolic engineering[99] and synthetic biology.[100] For example, L-threonine, an amino acid widely used in cosmetics and pharmacy, has been produced from *E. coli* through biosynthetic pathways.[101] The productivity of such substances can be improved by mutating genes encoding pathway components. To achieve this goal, one has to determine which genes to mutate. The number of possibilities is large, and it will be impossible to test them one by one. Instead, one must use computational models, on which *in silico* perturbation effects can be simulated and examined. We term this kind of model analysis *perturbation optimization*.

Mathematically, perturbation optimization is a combinatorial optimization problem: maximize $f(x)$ subject to $\mathbf{c}(x)$, where the decision variable x denotes a perturbation, the objective function f quantifies the changes in the outputs w.r.t to the perturbation, and $\mathbf{c}$ is a set of constraints specifying the requirements that must be met to ensure cells continue to survive and have proper functioning. A perturbation can be the mutation of a set of genes leading to changes of initial conditions or kinetic parameters in the model. To combat the combinatorial explosion of solution space for large models, many optimization methods have been used in recent years including linear programming, bilevel optimization, mixed integer nonlinear programming, and dynamic optimization.[102]

6.3. *Model checking*

The verification technique called model checking is a powerful automated method to verify that critical hardware and software computing systems behave as intended.[35] The model checking framework has been applied — mainly in stochastic settings — to biopathways models.[103–105]

Briefly, the model checking procedure operates as follows. Given a model $\mathcal{M}$ with initial state s, a model checker decides if a property written as a temporal logic formula ϕ is satisfied, denoted as $\mathcal{M}, s \models \phi$. This can be done by: (1) constructing a finite labeled state-transition system corresponding to $\mathcal{M}$ in which each state

represents a possible configuration and each transition represents an evolution of the system from one configuration to another and (2) verify whether ϕ is satisfied by exhaustively exploring the state-transition system. In stochastic settings, both the model and the property to be verified will be cast in a probabilistic language. Specifically $\mathcal{M}$ will be a CTMC and a probabilistic model checker will decide whether $\mathcal{M}$ satisfies ϕ with probability at least ρ. However, the state space explosion phenomenon (the number of states to be explored is exponential in the number of variables in the system) limits their use to small models. A more pragmatic approach is to use a statistical or Monte Carlo framework.[106,107] Here, a finite number of sample trajectories are drawn from the model. Each sample trajectory is evaluated to determine whether it satisfies ϕ, and the number of satisfying and non-satisfying traces is used to determine whether $\mathcal{M}, s \models P_{\geq\rho}(\phi)$. This approach scales well and though the results are approximate, bounds on the probability of producing an incorrect answer can be provided.

7. A Probabilistic Approximation Approach

Our goal here is to present an approximation technique using which an ODE-based model of a biopathway can be represented as a Dynamic Bayesian Network (DBN). As pointed out earlier, ODE systems will not admit closed-form solutions and hence one will have to resort to numerical simulations to perform analysis. In particular, tasks such as parameter estimation and sensitivity analysis and perturbation analysis will entail a large number of simulations. Second, the experimental data used for training and testing models will be often cell population−based and have limited precision. Hence there will be a large gap between the precision obtained through numerical simulations and that of experimental data. Further, to simulate the model and compare with such data, one must resort to Monte Carlo methods to ensure that sufficiently many values from the distribution of parameters and initial concentrations are being sampled. As a result, model calibration, validation, and analysis will require the *repeated* generation of a large number of numerical simulations. Consequently it is very difficult to handle large ODE-based pathway models. To address these challenges we have (in collaboration with our colleague David Hsu) developed the means for approximating the dynamics of an ODE-based pathway model as a DBN.[108] Using the DBN approximation, one can then efficiently perform parameter estimation and other analysis tasks by exploiting standard Bayesian inferencing techniques. Indeed, as the case studies we discuss below show, the one time cost of constructing the DBN can be easily amortized by carrying out multiple analysis tasks on the DBN approximation.

7.1. *Dynamic Bayesian networks*

Before proceeding to the approximation procedure, we first recall the notion of a DBN.[109] It consists of a directed acyclic graph in which the nodes are grouped into

layers with each layer representing a time point. In the simple version of DBNs we need, there will only be a finite number of layers corresponding to the time points $\{0, 1, \ldots, T\}$ and the nodes in layer $t-1$ will be connected to only nodes in the layer t. The DBN will have a set of random variables $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ associated with it. For each $X \in \mathcal{X}$ and each time point t, there will be a unique node X^t in the time layer t which will be used to capture the value assumed by X at time t. Further the connectivity structure between two adjacent time slices will be invariant in the sense if there is an edge from X^{t-1} to Y^t then there will be an edge from $X^{t'-1}$ to $Y^{t'}$ for every $t' \in \{1, 2, \ldots, T\}$. Next, if there is an edge from node X^{t-1} to Y^t then X^{t-1} is said to be a parent node of Y^t. Finally, each node will have a conditional probability table (CPT) associated with it to specify the local probabilistic dynamics of the DBN. If $\{Z_1^{t-1}, Z_2^{t-1}, \ldots, Z_k^{t-1}\}$ is the set of parent nodes of X^t then a typical entry in the CPT of X^t will be of the form $Pr(X^t = v \mid Z_1^{t-1} = v_1, Z_2^{t-1} = v_2, \ldots, Z_k^{t-1} = v_k) = p$. This will denote that the conditional probability of X^t assuming the value v (i.e. X assuming the value v at time t) is p given that the value assumed by Z_j^{t-1} (i.e. the value assumed by Z_j at time $t-1$) is v_j for $1 \leq j \leq k$.

Two successive time slices of a DBN are shown in Fig. 4(c). The annotations will become clear once we explain below the procedure for deriving a DBN from an ODE system.

7.2. *The approximation procedure*

We first discretize the time domain and the value space. This is motivated by the fact that the values of the variables are experimentally observed only for a finite number of time points and that too with only limited precision. Hence we assume the dynamics is of interest only for discrete time points, $\{0, 1, \ldots, T\}$. Assuming that the minimum and maximum values of the variables and rate constants are known, we next partition the range of each variable x_i into a set of intervals $\mathbf{I}_i$ according to the precision of observations. The range of each rate constant r_j is also discretized into a finite set of intervals $\mathbf{I}_j$. The initial values as well as the rate constants of the ODE system are described as distributions (usually uniform) over certain intervals defined by the discretization. For unknown parameters, they are assumed to be uniformly distributed over *all* their intervals.

The second step is to sample the initial states of the system sufficiently many times, according to the initial distribution and generate a trajectory by numerical integration for each sampled initial state. The resulting set of trajectories is then treated as an approximation of the dynamics of ODE system.

A key idea is to compactly store the generated set of trajectories as a DBN. This is achieved by means of a simple counting procedure that exploits the network structure. In the present setting, there will be one node x_i^t (r_j^t) corresponding to each variable x_i (rate constant r_j) to capture in which interval the value of x_i (r_j) falls at time t. Next, $Pa(x_i^t)$, the set of parent nodes of the node x_i^t is determined as follows. The node $x_k^{t-1}(r_j^{t-1})$ will be in $Pa(x_i^t)$ iff $x_k(r_j)$ appears in the equation for x_i. On the

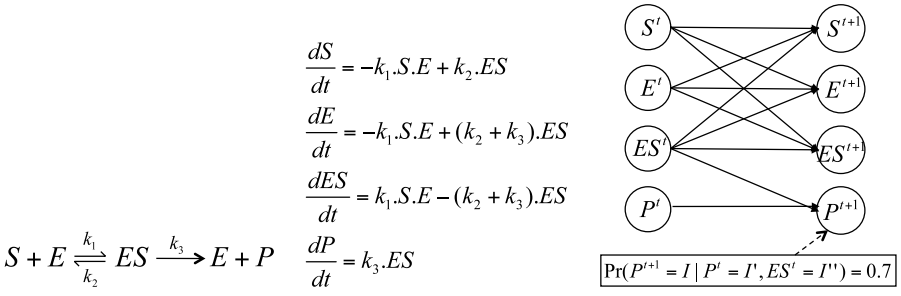


Fig. 4. (a) The enzyme catalytic reaction network. (b) The ODE model. (c) The DBN approximation for two successive time slices.

other hand, r_j^{t-1} will be the only parent of the parameter node r_j^t since the rate constant values will not change once their initial values have been fixed.

Example 2. In Fig. 4, we show a simple enzymatic reaction network, its ODE model and the structure of its DBN approximation for two successive time points.

As indicated in Fig. 4(c), a typical entry in the CPT of x_i^t will be of the form $\Pr(x_i^t = I | z_1^{t-1} = I_1, z_2^{t-1} = I_2, \ldots, z_l^{t-1} = I_l) = p$ with $Pa(x_i^t) = \{z_1^{t-1}, z_2^{t-1}, \ldots, z_l^{t-1}\}$ being the set of parents of x_i^t. Such an entry means that p is the probability that the value of x_i falls in the interval I at time t, given that the value of z_u was in I_u at time $t-1$ for each z_u^{t-1} in $Pa(x_i^t)$. The probability p is calculated through simple counting: Suppose N trajectories are generated. Record the number of the trajectories from this collection for which their value of z_u fell in the interval I_u for each z_u in $\{z_1, z_2, \ldots, z_l\}$ at time $t-1$. Suppose this number is J. Then determine for how many of these J trajectories, the value of x_i fell in the interval I at time t. If this number is J', then p is set to be $\frac{J'}{J}$.

For the ODE models of biochemical networks, one may assume that the vector field defined by the ODE system is a C^1 (continuously differentiable) function on a compact space. This is due to the fact that the concentration levels of the species are restricted to a bounded set of values and the kinetics of the biochemical reactions (i.e. the vector field of the ODE system) are captured by laws such as mass law and Michaelis−Menton described by (low degree) polynomials or rational functions which are naturally C^1 functions. As a result, the solution to the system of ODEs will exist. Further it will be continuous and hence measurable. As a result, the discretization and initial distribution will let us represent the generated family of trajectories as a finite state Markov chain. Supposing there are n variables and m rate constants. Then each state of this Markov chain will be of the form $((I_1, I_2, \ldots, I_n, I_{n+1}, I_{n+2}, \ldots, I_{n+m}), t)$. The probability of this state holding at time point t' will be 0 if $t \neq t'$. On the other hand the probability of this state holding at t will be the (well-defined) probability measure of the set of trajectories which have the value of

x_i (r_j) falling in the interval I_i (I_{n+j}) at t. This Markov chain will however be of size exponential in n. Hence we impose independence assumptions obtained from the network structure and derive the DBN as a factored — and much more succinct — representation of this Markov chain. A more detailed account of this construction can be found in Liu *et al.*[11]

It is worth noting that our approximation technique applies to ODEs whose vector fields are C^1 functions over compact spaces. In this sense it can be applied in other settings as well. However we have so far focused on models of biochemical networks, a setting in which these restrictions are naturally satisfied.

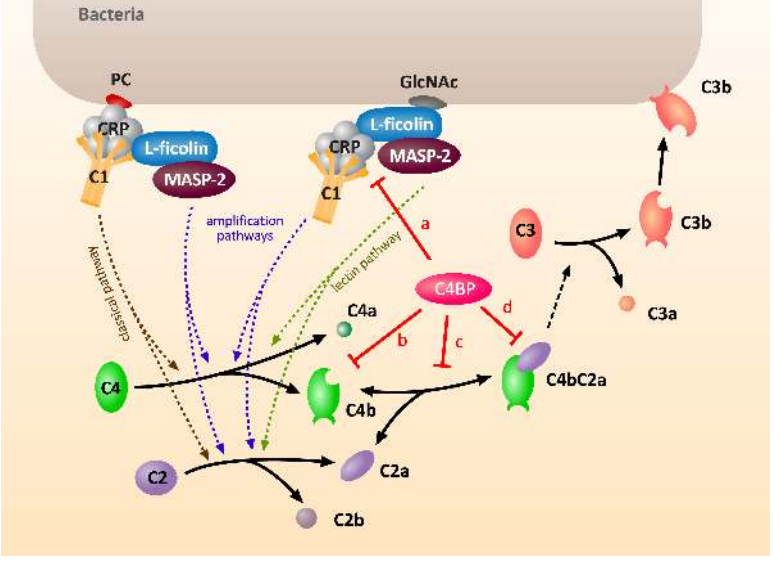
Since the trajectories are grouped together through the discretization, this method bridges the gap between the accuracy of the results obtained by ODE simulation and the limited precision of experimental data used for model development. More crucially, many interesting pathway properties can be analyzed efficiently through standard Bayesian inference techniques, instead of resorting to large scale numerical simulations. In particular we can perform:

- **Probabilistic inference.** Given initial state as evidence, Bayesian inference techniques such as the Factored Frontier algorithm can be used to approximately but efficiently infer the marginal probability of each species' concentration at a given time point.[110]
- **Parameter estimation.** The DBN approximation enables a two-stage parameter estimation method. In the first stage, one infers the marginal distributions of the species at different points in the DBN. The mean values of each marginal distribution are then computed and compared with the time series training data. Standard optimization methods are then used for searching in the *discretized* parameter space, resulting in a maximum likelihood estimate of a combination of intervals of parameter values. In the second stage, by treating the result of the first stage as the *drastically reduced* search space, one can further estimate — if necessary — point values for unknown parameters.
- **Global sensitivity analysis.** To perform global sensitivity analysis, Monte Carlo samples are drawn from the discretized parameter space. Simulation trajectories are approximated by the mean of marginal distributions inferred from the DBN by presenting the selected combination of intervals of parameter values as evidence.

The DBN approximation framework has been used to successfully study a number of large biochemical networks. Here we consider two models taken from the BioModels database,[111] namely, the EGF-NGF signaling pathway and the segmentation clock network.[11] Although the parameter values for these models are known, to mimic realistic biopathways models, in each case we designated a subset of the parameters as 'unknown', and constructed the DBN approximation accordingly. Specifically, 20 of the 48 parameters for the EGF-NGF model and 40 of the 75 parameters for the segmentation clock network were singled out to be unknown. After discretizing time domain and the ranges of each variable and unknown parameters, approximately

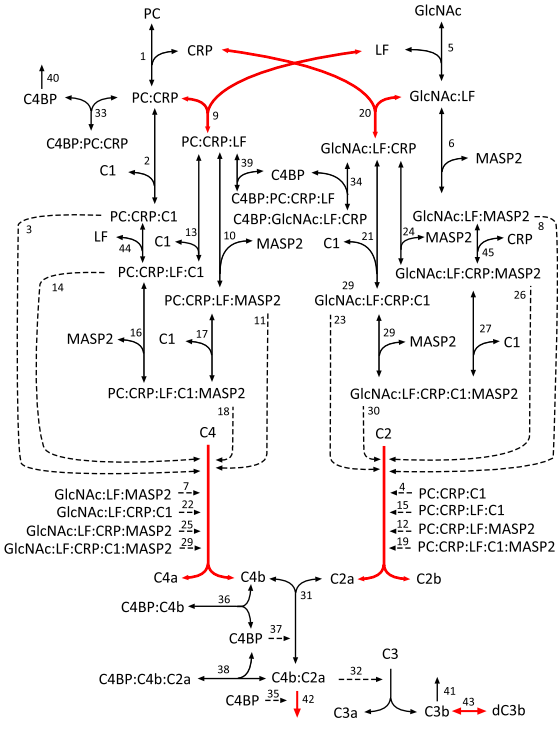
three million trajectories were generated for each model to build the DBNs. We then synthesized experimental time series data and divided them into the training and test data sets. The unknown parameters were then estimated using the training data and the resulting models were validated using the test data. We also performed global sensitivity analysis using synthesized data. For the EGF-NGF model it took around 4 h to construct the DBN approximation while the total running time of performing parameter estimation and global sensitivity analysis was reduced (in comparison to ODE-based methods) from 23 h to 0.6 h. For the clock segmentation network model it took around 3.5 h to construct the DBN approximation while the total running time of performing parameter estimation and global sensitivity analysis was reduced from 82 h to 3.3 h. More details can be found in Liu *et al.*[11]

Our method has also been used in a "live" setting. In collaboration with biologists and clinicians it was used to study the complement system, which is the frontline defense mechanism of the human immune system.[112] The activation of complement system is necessary for the clearance of bacteria and apoptotic cells. However, insufficient or excessive complement activation will lead to immune-related diseases. We built a computational model involving the enhancement and suppression mechanisms that regulate complement activity. The schematic representation and reaction network diagram of the model is shown in Fig. 5. It consists of 42 species, 45 reactions and 85 kinetic parameters with 71 of the parameters being unknown. Based on *in vivo* experimental data, the DBN approximation method was used to estimate



(a)

Fig. 5. (a) Simplified schematic representation of the complement system. (b) Reaction network diagram of the ODEs model.



(b)

Fig. 5. (*Continued*)

the unknown rate constants (model calibration). The model was then validated with the help of previously published data. By performing global sensitivity analysis on the DBN approximation as well as *in silico* perturbation experiments on the validated ODE model, interesting hypotheses about the regulatory mechanisms of the complement system were derived, which were then experimentally confirmed. Specifically, the combined computational and experimental study highlighted the importance of infection-mediated microenvironmental perturbations, which alter the pH and calcium levels. It also revealed that the inhibitor, C4BP, induces differential inhibition on the classical and lectin complement pathways and acts mainly by facilitating the decay of the C3 convertase. These results helped to elucidate the regulatory mechanisms of the complement system and can potentially contribute to the development of complement-based immunomodulation therapies. The resulting ODE-based model has been added to the BioModels database.[111]

These results show that this probabilistic approximation method achieves a good tradeoff between efficiency and accuracy. More importantly it can handle large pathways models that most of the current ODE-based approaches will not be able to cope with.

8. Conclusion

Computational modeling and analysis is essential for understanding biopathway dynamics at the system level. Here we have reviewed the prevalent stochastic and deterministic quantitative models for representing pathways. With the focus on ODE-based models, we have then discussed the methods for estimating unknown model parameters and described three important analysis techniques, namely, sensitivity analysis, perturbation optimization and model checking. After highlighting the challenges faced when dealing with ODE-based models we have presented a promising probabilistic approximation method.

We are currently improving and augmenting this approximation technique in a number of ways. Firstly, the construction of the DBN approximation for large pathway models is computationally intensive. However, many components of this construction can be easily parallelized. Thus motivated, research is underway to mapping this process onto Graphics Processing Units (GPUs). A preliminary study of the Thrombin-MLC pathway which has 105 species (we assumed 163 parameters to be unknown) shows that the GPU implementation scales well with the problem size. In comparison, a 10-PC cluster failed to deal with this large model.[113] We are also beginning to study the means for implementing the parameter estimation and sensitivity analysis tasks on the GPU platform. Finally, work is underway for doing probabilistic model checking on the DBN approximations.

From a broader perspective, the material presented here is a core component of computational systems biology. Although this field is making excellent progress many challenges remain. A key one is the inherently incremental and incomplete nature of the model construction process. Biologists will continue to decipher new cellular mechanisms and additional experimental data will be constantly generated. In this light, a computational model is at best a consistent record of what is currently known about a particular process. Thus incorporating new knowledge and data into an existing model without having to redo model construction, calibration, and validation processes is a fundamental challenge. A recent attempt has been made to integrate new data to an existing model using belief propagation techniques.[68] More advanced approaches that can integrate component models when cross talks are discovered and for assimilating changes made to the pathway structure need to be explored.

Another challenge is to develop modeling frameworks that can integrate stochastic and deterministic features. This will considerably expand the range of applicability of modeling techniques. Current techniques force one to choose stochastic or deterministic methods in a mutually exclusive fashion. Yet another fundamental challenge is to develop multi-scale modeling and analysis methods using which one can connect the dynamics of molecular mechanisms governing single cell behaviors to multi-cell functionalities. For instance, it is known that noise and stochasticity in gene expressions and other cellular events can cause genetically identical cells to exhibit variability in cellular states.[114] Such variations can have significant

impact on cellular functioning and phenotype.[115] Hence it is vital to bridge the gap between the single cell dynamics and the emergent functionalities of a collection of cells. It is worth noting in this context that initial attempts are being made with systems such as cell death and differentiation.[3,52]

References

1. Bell-Pedersen D, Cassone VM, Earnest DJ, Golden SS, Hardin PE, Thomas TL, Zoran MJ, Circadian rhythms from multiple oscillators: Lessons from diverse organisms, *Nat Rev Genet* **6**:544−556, 2005.
2. Matsuno H, Tanaka Y, Aoshima H, Doi A, Matsui M, Miyano S, Biopathways representation and simulation on hybrid functional Petri net, *In Silico Biol* **3**:389−404, 2003.
3. Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK, Non-genetic origins of cell-to-cell variability in trail-induced apoptosis, *Nature* **459**:428−432, 2009.
4. Kholodenko BN, Untangling the signalling wires, *Nat Cell Biol* **9**:247−249, 2007.
5. Logan CY, Nusse R, The Wnt signaling pathway in development and disease, *Annu Rev Cell Dev Biol* **20**:781−810, 2004.
6. Egan LJ, Toruner M, NF-kappaB signaling: Pros and cons of altering NF-kappaB as a therapeutic approach, *Ann N Y Acad Sci* **1072**:114−122, 2006.
7. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P, *Molecular Biology of the Cell*, 5th edn. Garland Science, New York, USA, 2008.
8. Hanahan D, Weinberg RA, The hallmarks of cancer, *Cell* **100**:57−70, 2000.
9. Lowe SW, Lin AW, Apoptosis in cancer, *Carcinogenesis* **21**:485−495, 2000.
10. De Ferrari GV, Inestrosa NC, Wnt signaling function in Alzheimer's disease, *Brain Res Brain Res Rev* **33**:1−12, 2000.
11. Liu B, Thiagarajan PS, Hsu D, Probabilistic approximations of ODEs based bio-pathway dynamics, *Theor Comput Sci* **412**:2188−2206, 2011.
12. de Jong H, Modeling and simulation of genetic regulatory systems: A literature review, *J Comput Biol* **9**:67−103, 2002.
13. Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H, *Systems Biology in Practice: Concepts, Implementation and Application*, Wiley-VCH, 2005.
14. Thakar J, Pilione M, Kirimanjeswara G, Harvill ET, Albert R, Modeling systems-level regulation of host immune responses, *PLoS Comput Biol* **3**:e109, 2007.
15. Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK, Physicochemical modelling of cell signalling pathways, *Nat Cell Biol* **8**:1195−1203, 2006.
16. Ruths D, Muller M, Tseng JT, Nakhleh L, Ram PT, The signaling Petri net-based simulator: A non-parametric strategy for characterizing the dynamics of cell-specific signaling networks, *PLoS Comput Biol* **4**:1—15, 2008.
17. Danos V, Feret J, Fontana W, Harmer R, Krivine J, Rule-based modelling of cellular signalling, in Caires L, Vasconcelos VT (Eds.), *Proc. 18th Int. Conf. Concurrency Theory (CONCUR'07)*, *Lecture Notes in Computer Science*, Springer, pp. 17−41, 2007.
18. Heath J, Kwiatkowska M, Norman G, Parker D, Tymchyshyn O, Probabilistic model checking of complex biological pathways, *Theor Comput Sci* **319**:239−257, 2008.
19. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L, Reactome: A knowledgebase of biological pathways, *Nucleic Acids Res* **33**:D428−D432, 2005.
20. Arkin A, Ross J, McAdams HH, Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells, *Genetics* **149**:1633−1648, 1998.

21. Gillespie D, Exact stochastic simulation of coupled chemical reactions, *J Chem Phys* **81**:2340−2361, 1977.
22. Wilkinson DJ, *Stochastic Modelling for Systems Biology*, Taylor & Francis, Boca Raton, 2006.
23. Munsky B, Khammash M, A multiple time interval finite state projection algorithm for the solution to the chemical master equation, *J Chem Phys* **226**:818−835, 2007.
24. Gillespie D, Petzold L, Improved leap-size selection for accelerated stochastic simulation, *J Chem Phys* **119**:8229−C8234, 2003.
25. Haseltine E, Rawlings J, On the origins of approximations for stochastic chemical kinetics, *J Chem Phys* **123**:164115, 2005.
26. Wolf V, Goel R, Mateescu M, Henzinger T, Solving the chemical master equation using sliding windows, *BMC Syst Biol* **5**:1−19, 2010.
27. Feret J, Danos V, Krivine J, Harmer R, Fontana W, Internal coarse-graining of molecular systems, *Proc Natl Acad Sci USA* **106**:6453−6458, 2009.
28. Danos V, Feret J, Fontana W, Krivine J, Abstract interpretation of cellular signalling networks, in Logozzo F, Peled D, Zuck LD (Eds.), *Proc. 9th Int. Conf. Verification, Model Checking, and Abstract Interpretation* (*VMCAI'08*), *Lecture Notes in Computer Science*, Springer, pp. 83−97, 2008.
29. Kwiatkowska MZ, Norman G, Parker D, PRISM: Probabilistic symbolic model checker, in Field T, Harrison PG, Bradley JT, Harder U (Eds.), *Proc. 12th Int. Conf. Computer Performance Evaluations, Modelling Techniques and Tools* (*TOOLS'02*), *Lecture Notes in Computer Science*, Springer, pp. 200−204, 2002.
30. Hillston J, *A Compositional Approach to Performance Modelling*, University Press, 1996.
31. Kwiatkowska MZ, Heath JK, Biological pathways as communicating computer systems, *J Cell Sci* **122**:2793−2800, 2009.
32. Pnueli A, The temporal logic of programs, in *FOCS*, IEEE, pp. 46−57, 1977.
33. Hansson H, Jonsson B, A logic for reasoning about time and reliability, *Formal Asp Comput* **6**:512−535, 1994.
34. Aziz A, Sanwal K, Singhal V, Brayton R, Model checking continuous time Markov chains, *ACM T Comput Log* **1**:162−170, 2000.
35. Clarke EM, Grumberg O, Peled DA, *Model Checking*, MIT Press, 1999.
36. Calder M, Vyshemirsky V, Gilbert D, Orton R, Analysis of signalling pathways using the PRISM model checker, in *Proc Comput Meth Syst Biol* (*CMSB'05*), pp. 179−190, 2005.
37. Tianhai Tian T, Xu S, Gao J, Burrage K, Simulated maximum likelihood method for estimating kinetic rates in gene expression, *Bioinformatics* **23**:84−91, 2007.
38. Reinker S, Altman RM, Timmer J, Parameter estimation in stochastic biochemical reactions, *IET Syst Biol* **153**:168−178, 2006.
39. Poovathingal SK, Gunawan R, Global parameter estimation of stochastic biochemical systems, *BMC Bioinfo* **11**:1−12, 2010.
40. Wang Y, Christley S, Mjolsness E, Xie X, Parameter inference for discretely observed stochastic kinetic models using stochastic gradient descent, *BMC Syst Biol* **4**:1−16, 2010.
41. Hirsch MW, Smale S, Devaney RL, *Differential Equations, Dynamical Systems and an Introduction to Chaos*, Elsevier, 2004.
42. Hindmarsh AC, ODEPACK, a systematized collection of ODE solvers, *Sci Comput* **1**:55−64, 1983.
43. Petzold L, Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations, *SIAM J Sci Stat Comput* **4**:136−148, 1983.
44. Schmidt H, Madsen MFMF, Dano S, Cedersund G, Complexity reduction of biochemical rate expressions, *Bioinformatics* **24**:848−854, 2008.

45. Batt G, Belta C, Weiss R, Temporal logic analysis of gene networks under parameter uncertainty, *IEEE Trans Circuits Syst I/Automat Control (Special Issue on Systems Biology)* **53**:215−229, 2008.
46. de Jong H, Gouzé JL, Hernandez C, Page M, Sari T, Geiselmann J, Qualitative simulation of genetic regulatory networks using piecewise-linear models, *Bull Math Biol* **66**:301−340, 2004.
47. Batt G, Ropers D, de Jong H, Geiselmann J, Page M, Schneider D, Qualitative analysis and verification of hybrid models of genetic regulatory networks: Nutritional stress response in *Escherichia coli*, in *Proc. 8th Int. Workshop on Hybrid Systems: Computation and Control (HSCC'05)*, *Lecture Notes in Computer Science*, Springer-Verlag, pp. 134−150, 2005.
48. Machina A, Ponosov A, Stability of stationary solutions of piecewise affine differential equations describing gene regulatory networks, *J Math Anal Appl* **380**:736−749, 2011.
49. Petri CA, *Kommunikation mit automaten*. Ph.D. Thesis, University of Bonn, 1962.
50. Kordic V (Ed.), *Petri Net, Theory and Applications*, InTech, 2008.
51. Koch I, Reisg W, Schreiber F (Eds.), *Modeling in Systems Biology: The Petri Net Approach*, Springer, New York, 2011.
52. Bonzanni N, Krepska E, Feenstra KA, Fokkink W, Kielmann T, Bal H, Heringa J, Executing multicellular differentiation: Quantitative predictive modelling of *C. elegans* vulval development, *Bioinformatics* **25**:2049−2056, 2009.
53. Nagasaki M, Saito A, Jeong E, Li C, Kojima K, Ikeda Y, Miyano S, Cell Illustrator 4.0: A computational platform for systems biology, *In Silico Biol* **10**:0002, 2010.
54. Tasaki S, Nagasaki M, Kozuka-Hata H, Semba K, Gotoh N, Hattori S, Inoue J, Yamamoto T, Miyano S, Sugano S, Oyama M, Phosphoproteomics-based modeling defines the regulatory mechanism underlying aberrant EGFR signaling, *PLoS One* **5**: e13926, 2010.
55. Do JH, Nagasaki M, Miyano S, The systems approach to the prespore-specific activation of sigma factor SigF in *Bacillus subtilis*, *Biosystems* **100**:178−184, 2010.
56. Fogel D, Fogel L, Atmar J, Meta-evolutionary programming, in *Proc. 25th Asiloma Conf. on Signals, Systems and Computers*, IEEE Computer Society, pp. 540−545, 1992.
57. Levenberg K, A method for the solution of certain nonlinear problems in least squares, *Quart Appl Math* **2**:164−168, 1994.
58. Hooke R, Jeeves TA, "Direct search" solution of numerical and statistical problems, *J ACM* **8**:212−229, 1961.
59. Moles CG, Mendes P, Banga JR, Parameter estimation in biochemical pathways: A comparison of global optimization methods, *Genome Res* **13**:2467−2474, 2003.
60. Mitchell M, *An Introduction to Genetic Algorithms*, MIT Press, 1995.
61. Beyer HG, Schwefel HP, Evolution strategies — a comprehensive introduction, *Nat Comput* **1**:3−52, 2002.
62. Kennedy J, Eberhart R, Particle swarm optimization, in *Proc. 4th IEEE International Conf. on Neural Networks*, pp. 1942−1948, 1995.
63. Runarsson T, Yao X, Stochastic ranking for constrained evolutionary optimization, *IEEE T Evolut Comput* **4**:284−294, 2000.
64. Kleinstein SH, Bottino D, Lett GS, Nonunifrom sampling for global optimization of kinetic rate constants in biological pathways, in *Proc. 2006 Winter Simulation Conference* (*WSC'06*), pp. 1161−1166, 2006.
65. Rodriguez-Fernandez M, Mendes P, Banga JR, A hybrid approach for efficient and robust parameter estimation in biochemical pathways, *Biosystems* **83**:248−265, 2006.
66. Koh G, Teong HFC, Clement MV, Hsu D, Thiagarajan PS, A decompositional approach to parameter estimation in pathway modeling: A case study of the Akt and MAPK pathways and their crosstalk, *Bioinformatics* **22**:e271−e280, 2006.

67. Koh G, Tucker-Kellogg L, Hsu D, Thiagarajan PS, Globally consistent pathway parameter estimates through belief propagation, in *Proc. 7th Int. Workshop on Algorithms in Bioinformatics* (*WABI'07*), pp. 420−430, 2007.
68. Koh G, Hsu D, Thiagarajan PS, Incremental signaling pathway modeling by data integration, in Berger B (Ed.), *Proc. 14th Int. Conf. Research in Computational Molecular Biology (RECOMB'10)*, *Lecture Notes in Computer Science*, Springer, pp. 281−296, 2010.
69. Brown K, Hill C, Calero G, Myers C, Lee K, Sethna J, Cerione R, The statistical mechanics of complex signaling networks: Nerve growth factor signaling, *Phys Biol* **1**:184−195, 2004.
70. Lipniacki T, Paszek P, Brasier A, Luxon B, Kimmel M, Mathematical model of NFκB regulatory module, *J Theor Biol* **228**:195−15, 2004.
71. Gutenkunst R, Waterfall J, Casey F, Brown K, Myers C, Sethna J, Universally sloppy parameter sensitivities in systems biology models, *Plos Comput Biol* **3**:1871−1878, 2007.
72. Lodhi H, Muggleton S (Eds.), *Elements of Computational Systems Biology*, John Wiley and Sons, New York, 2009.
73. Banga JR, Balsa-Canto E, Parameter estimation and optimal experimental design, *Essays Biochem* **45**:195−10, 2008.
74. Bandara S, Sclöder J, Eils R, Bock H, Meyer T, Optimal experimental design for parameter estimation of a cell signaling model, *PLoS Comput Biol* **5**:1−2, 2009.
75. Birtwistle MR, Hatakeyama M, Yumoto N, Ogunnaike BA, Hoek JB, Kholodenko BN, Ligand-dependent responses of the ErbB signaling network: Experimental and modeling analyses, *Phys Biol* **3**:1−16, 2007.
76. Chen WW, Schoeberl B, Jasper PJ, Niepel M, Nielsen UB, Lauffenburger DA, Sorger PK, Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data, *Mol Syst Biol* **5**:1−19, 2009.
77. Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK, Physicochemical modelling of cell signalling pathways, *Nat Cell Biol* **8**:1195−1203, 2006.
78. Saltelli A, *Global Sensitivity Analysis: The Primer*, John Wiley, Chichester, 2008.
79. van Riel NA, Dynamic modelling and analysis of biochemical networks: Mechanism-based models and model-based experiments, *Brief Bioinform* **7**:364−374, 2006.
80. Cascante M, Boros LG, Comin-Anduix B, de Atauri P, Centelles JJ, Lee PWN, Metabolic control analysis in drug discovery and disease, *Nat Biotechnol* **20**:243−249, 2002.
81. de Pillis LG, Radunskaya AE, Wiseman CL, A validated mathematical model of cell-mediated immune response to tumor growth, *Cancer Res* **65**:7950−7958, 2005.
82. Bentele M, Lavrik I, Ulrich M, Stober S, Heermann D, Kalthoff H, Krammer P, Eils R, Mathematical modeling reveals threshold mechanism in CD95-induced apoptosis, *J Cell Biol* **166**:839−851, 2004.
83. von Dassow G, Meir E, Munro EM, Odell GM, The segment polarity network is a robust developmental module, *Nature* **406**:188−192, 2000.
84. Gunawan R, Cao Y, Petzold L, Doyle FJ 3rd, Sensitivity analysis of discrete stochastic systems, *Biophys J* **88**:2530−2540, 2005.
85. Birtwistle MR, Hatakeyama M, Yumoto N, Ogunnaike BA, Hoek JB, Kholodenko BN, Ligand-dependent responses of the ErbB signaling network: Experimental and modeling analyses, *Mol Syst Biol* **144**:1−16, 2007.
86. Gunawan R, Doyle FJ 3rd, Isochron-based phase response analysis of circadian rhythms, *Biophys J* **91**:2131−2141, 2006.
87. Schilling M, Maiwald T, Hengl S, Winter D, Kreutz C, Kolch W, Lehmann WD, Timmer J, Klingmüller U, Theoretical and experimental analysis links isoform-specific ERK signalling to cell fate decisions, *Mol Syst Biol* **5**:334, 2009.

88. Swameye I, Muller TG, Timmer J, Sandra O, Klingmuller U, Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling, *Proc Natl Acad Sci USA* **100**:1028−1033, 2003.
89. van Stiphout RGPM, van Riel NAW, Verhoog PJ, Hilbers PAJ, Nicolay K, Jeneson JAL, Computational model of excitable cell indicates ATP free energy dynamics in response to calcium oscillations are undampened by cytosolic ATP buffers, *Syst Biol* (*Stevenage*) **153**:405−408, 2006.
90. Schoeberl B, Eichler-Jonsson C, Gilles ED, Müller G, Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors, *Nat Biotechnol* **20**:370−375, 2002.
91. Feng Xj, Rabitz H, Optimal identification of biochemical reaction networks, *Biophys J* **86**:1270−1281, 2004.
92. Cho KH, Shin SY, Kolch W, Wolkenhauer O, Experimental design in systems biology, based on parameter sensitivity analysis using a Monte Carlo method: A case study for the TNFα-mediated NF-κB signal transduction pathway, *Simulation* **79**:726−739, 2003.
93. Draper NR, Smith H, *Applied Regression Analysis*, Wiley, New York, 1981.
94. Sobol IM, Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, *Math Comput Simulat* **55**:271−280, 2001.
95. Lüdtke N, Panzeri S, Brown M, Broomhead DS, Knowles J, Montemurro MA, Kell DB, Information-theoretic sensitivity analysis: A general method for credit assignment in complex networks, *J R Soc Interface* **5**:223−235, 2008.
96. Zhang Y, Rundell A, Comparative study of parameter sensitivity analyses of the TCR-activated Erk-MAPK signalling pathway, *Syst Biol (Stevenage)* **153**:201−211, 2006.
97. Sahle S, Mendes P, and U Kummer SH, A new strategy for assessing sensitivities in biochemical models, *Phil Trans R Soc A* **366**:3619−3631, 2008.
98. Khosla C, Keasling JD, Metabolic engineering for drug discovery and development, *Nat Rev Drug Discov* **2**:1019−1025, 2003.
99. Raab RM, Tyo K, Stephanopoulos G, Metabolic engineering, *Adv Biochem Eng Biotechnol* **100**:1−17, 2005.
100. Andrianantoandro E, Basu S, Karig DK, Weiss R, Synthetic biology: New engineering rules for an emerging discipline, *Mol Syst Biol* **2**:2006.0028, 2006.
101. Lee KH, Park JH, Kim TY, Kim HU, Lee SY, Systems metabolic engineering of Escherichia coli for I-threonine production, *Mol Syst Biol* **3**:149, 2007.
102. Banga JR, Optimization in computational systems biology, *BMC Syst Biol* **2**:1−7, 2008.
103. Antoniotti M, Policriti A, Ugel N, Mishra B, Model building and model checking for biochemical processes, *Cell Biochem Biophys* **38**:271−286, 2003.
104. Chabrier-Rivier N, Chiaverini M, Danos V, Fages F, Schachter V, Modeling and querying biomolecular interaction networks, *Theor Comput Sci* **325**:25−44, 2004.
105. Gong H, and Anvesh Komuravelli PZ, Faede JR, Clarke EM, Analysis and verification of the HMGB1 signaling pathway, *BMC Bioinform* **11**(Suppl 7):1−13, 2010.
106. Clarke EM, Faeder JR, Langmead CJ, Harris LA, Jha SK, Legay A, Statistical model checking in BioLab: Applications to the automated analysis of T-Cell receptor signaling pathway, in Heiner M, Uhrmacher AM (Eds.), *Proc. 6th Int. Conf. Computational Methods in Systems Biology* (*CMSB'08*), *Lecture Notes in Computer Science*, Springer, pp. 231−250, 2008.
107. Donaldson R, Gilbert D, A Monte Carlo model checker for probabilistic LTL with numerical constraints. Technical report, University of Glasgow, 2008.
108. Liu B, Thiagarajan PS, Hsu D, Probabilistic approximations of signaling pathway dynamics, in Degano P, Gorrieri R (Eds.), *Proc. 7th Int. Conf. Computational Methods*

in Systems Biology (*CMSB'09*), *Lecture Notes in Computer Science*, Springer, pp. 251–265, 2009.
109. Murphy KP, *Dynamic Bayesian Networks: Representation, Inference and Learning.* Ph.D. Thesis, University of California, Berkeley, 2002.
110. Murphy KP, Weiss Y, The factored frontier algorithm for approximate inference in DBNs, in *Proc. 17th Int. Conf, Uncertainty in Artificial Intelligence* (*UAI'01*), San Francisco, CA, USA, pp. 378–385, 2001.
111. Le Novere N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep J, Hucka M, BioModels Database: A free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems, *Nucleic Acids Res* **34**:D689–D691, 2006.
112. Ricklin D, Hajishengallis G, Yang K, Lambris JD, Complement: A key system for immune surveillance and homeostasis, *Nat Immunol* **11**:785–797, 2010.
113. Liu B, Hagiescu A, Palaniappan SK, Chattopadhyay B, Cui Z, Wong WF, Thiagarajan PS, Approximate probabilistic analysis of biopathway dynamics, [Submitted].
114. Sigal A, Milo R, Cohen A, Geva-Zatorsky N, Klein Y, Liron Y, Rosenfeld N, Danon T, Perzov N, Alon U, Variability and memory of protein levels in human cells, *Nature* **444**:643–646, 2006.
115. Raj A, van Oudenaarden A, Nature, nurture, or chance: Stochastic gene expression and its consequences, *Cell* **135**:216–226, 2008.

Bing Liu received his B.Comp from School of Computing, National University of Singapore (NUS), Singapore (2006) and his Ph.D in Computational Biology from NUS Graduate School for Integrative Sciences and Engineering, NUS (2011). He is currently a Research Fellow in the Computer Science Department, NUS. His research interests include computational systems biology and high-performance computing for computational biology.

P. S. Thiagarajan received his B.Tech (Electrical Engineering) from the Indian Institute of Technology (IIT), Madras (1970) and his Ph.D. in Computer Science from Rice University, Houston, Texas, USA (1973). He is a Professor in the Computer Science Department of National University of Singapore. He has served on the editorial boards of the journals *Theoretical Computer Science* and *International Journal on Foundations of Computing.* He currently serves on the editorial boards of *The Real-Time Systems Journal* and *Transactions on Petri nets and Other Models of Concurrency.* He is a Fellow of the Indian Academy of Sciences and the Indian National Academy of Sciences. His research interests include computational systems biology, hybrid automata, and quantitative verification methods.