BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Iterative approach to model identification of biological networks

Kapil G Gadkar, Rudiyanto Gunawan and Francis J Doyle III*

Address: Department of Chemical Engineering, University of California Santa Barbara, CA, USA

Email: Kapil G Gadkar - gadkar@engr.ucsb.edu; Rudiyanto Gunawan - gunawan@engr.ucsb.edu; Francis J Doyle* - doyle@engr.ucsb.edu

* Corresponding author

Published: 20 June 2005

BMC Bioinformatics 2005, **6**:155 doi:10.1186/1471-2105-6-155

Received: 03 February 2005

Accepted: 20 June 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/155

Abstract

Background: Recent advances in molecular biology techniques provide an opportunity for developing detailed mathematical models of biological processes. An iterative scheme is introduced for model identification using available system knowledge and experimental measurements.

Results: The scheme includes a state regulator algorithm that provides estimates of all system unknowns (concentrations of the system components and the reaction rates of their interconversion). The full system information is used for estimation of the model parameters. An optimal experiment design using the parameter identifiability and D-optimality criteria is formulated to provide "rich" experimental data for maximizing the accuracy of the parameter estimates in subsequent iterations. The importance of model identifiability tests for optimal measurement selection is also considered. The iterative scheme is tested on a model for the caspase function in apoptosis where it is demonstrated that model accuracy improves with each iteration. Optimal experiment design was determined to be critical for model identification.

Conclusion: The proposed algorithm has general application to modeling a wide range of cellular processes, which include gene regulation networks, signal transduction and metabolic networks.

Background

A systems level understanding of highly complex biological systems requires an integration of experimental techniques and computational research [1]. Current molecular biology techniques can generate high-throughput quantitative data that support *in silico* research using mathematical models [2]. These models can be used to simulate and study the dynamic interactions among the components of cellular systems as well as the systems' responses to external perturbations and signals. Such tools offer enormous potential for understanding cellular functions at the organism level [3]. Mathematical models also serve as test beds for generating hypotheses and designing experiments to test them [4]. Furthermore, they provide bases for model-based product and process design applications. Useful insights and predictions have been obtained for several biological systems from computational modeling and analysis. A few examples include the metabolic network analysis of *Escherichia coli* growth on glucose and acetate [5], the MAP kinase signaling pathways [6] and caspase function in apoptosis [7], and bifurcation analysis of cell cycle in *Saccharomyces cerevesiae* [8].

An iterative process for model development and the testing of hypotheses has been proposed by many researchers in the field and was recently highlighted by Kitano [1]. A qualitative approach of this process is described in [2]. In addition, Rabitz and co-workers [9] have recently developed an iterative method for closed loop parameter identification in biochemical reaction networks. A global

inversion algorithm was used to extract the parameter estimates that minimize the differences between model prediction and experimental data. Unfortunately, global search methods typically have high computational requirements, and thus, do not scale very well with the system size. In this work, a quantitative model identification is developed that effciently obtains parameter estimates and facilitates scalability to very large network sizes. A proposed strategy is described in Section 2, with an emphasis on the modeling element. The modeling strategy is decomposed into three main steps: (1) determining the connectivity of the biological network and the interactions of the sub-components, (2) formulating the kinetics of inter-conversion among the subcomponents, and (3) estimating the parameters in the rate equations. To the authors' knowledge, this work represents the first documented example of multiple iterations for model refinement using such a framework in systems biology.

The parameter estimation from experimental data remains the bottleneck in the model development [4]. Banga and coworkers [10] have compared several advanced deterministic and stochastic global optimization methods for parameter identification from available experimental data. It was observed that the traditional gradient-based optimization methods often failed to arrive at the global optimal solutions. Deterministic methods [11-13] can achieve global optimality for certain classes of problems, but there is no guarantee of convergence in finite time [14]. Stochastic strategies [14-16] can locate the parameter region containing the global solution with relatively better effciency, but global optimality is not guaranteed. Furthermore, both methods suffer from the large computational burden required, even for moderately sized problems. Moreover, the validity of model with the estimated parameters over the entire operating space remains to be determined.

Parameter identifiability tests should be performed prior to the estimation process to ensure that the parameter estimation problem is well-posed. Further, the identifiability tests assist in selection of optimal measurements. Several researchers [17-19] have developed methods to determine whether a parameter is "identifiable *a priori*", i.e., identifiable from a given experiment design using the available measurements. A similar concept known as "practical identifiability" is concerned with the achievable accuracy of the parameter estimates. The confidence interval for the model parameters are determined using the Fisher Information Matrix (FIM) [20,21]. Doyle and coworkers [22] have performed model identifiability studies for a gene regulatory network using gene expression data, in which the identifiability of the parameters was found to be strongly dependent on the driving function.

The final step in the iterative model development process is the design of "optimal experiments" that would provide rich experimental data for improving the parameter estimates. Experiments can also be designed for discrimination among competing model structures that translates to selection between multiple proposed mechanisms of cellular function. Asprey and Macchietto [23] have developed a strategy of optimal experiment design for model structural identifiability. The strategy was used to identify the kinetics of the reactions in the fermentation of *Saccharomyces cerevesiae*. Kremling and co-workers [24] propose several strategies for model discrimination to identify the correct reaction mechanism of a test metabolic network. Banga and coworkers [25] have formulated the optimal design problem, using a scalar function of the Fisher Information Matrix (FIM) as the performance index, for parameter estimation of nonlinear dynamic systems.

In this work, an iterative procedure for model identification is proposed and applied to the caspase-dependent apoptosis system. An optimal measurement set is determined using the Fisher Information Matrix (FIM). The parameter estimation from partial measurements is decoupled into two parts. First, the available measurements are used to estimate the profiles of all unmeasured concentrations and reaction rates using a State Regulator Problem (SRP) formulation. In the second part the concentration and rate estimates are used to determine the model parameter values. The SRP formulation in this work is an extension of the dynamic Flux Balance Analysis (dFBA) approach developed by Doyle and coworkers [26]. Finally, the model-based experiment design uses parameter identifiability and D-optimality criteria to obtain the optimal experimental procedure that would generate the most informative data for model refinement in the next iteration.

Results

The iterative scheme for model identification is shown in Figure 1. The optimal set of measurements is determined *a priori*. For an effcient model identification, a significant fraction of the unknown model parameters should be identifiable. In the case of poor identifiability, a higher number of measurements would be motivated. Also, the model complexity could be reduced to decrease the number of parameters; but this does not guarantee identifiability of the reduced number of parameters. Alternatively, a richer protocol (e.g. perturbation sequence [22]) might yield improved identifiability. In this work, selection of the *a priori* optimal measurement set is restricted to the "preliminary" experiment design that may be suboptimal.

The model connectivity and reaction mechanisms are developed from existing biological knowledge and are

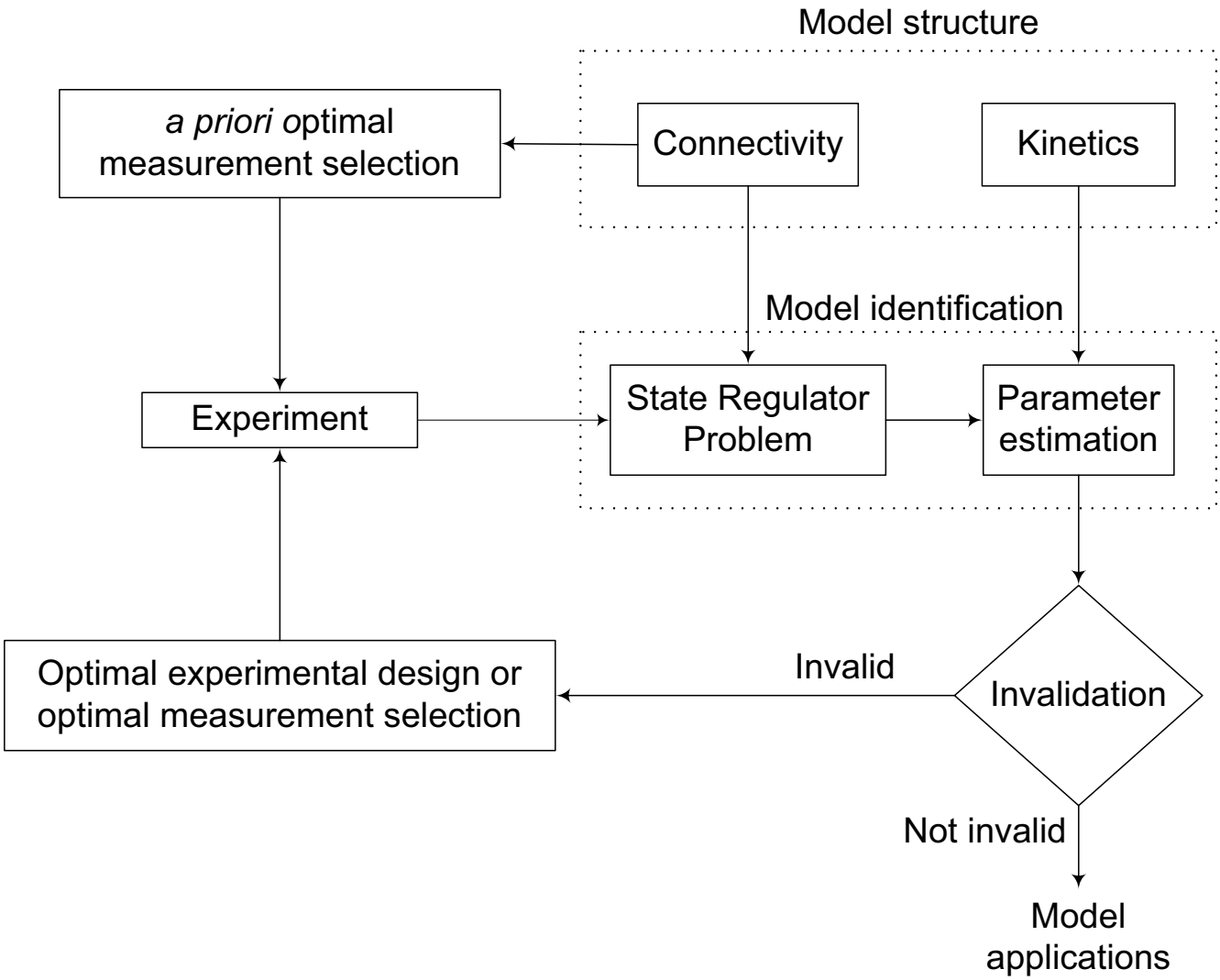

Figure 1
Iterative scheme for model identification.

assumed to be known. The network connectivity, along with the partial measurements (optimal), is used in the State Regulator Problem (SRP) to obtain estimates of all system unknowns (unmeasured concentrations and reaction rates). Here it is important to note that the kinetics of the reaction rates are not used in the SRP algorithm. Next, the full estimates of the concentrations and the reaction rates are used for estimating the parameters in the kinetic model. This decouples the model identification into two parts such that the parameters involved in the kinetic equation of each reaction are independently determined as opposed to simultaneously estimating full model parameters from limited measurements. Next, the model invalidation test, which is a critical step in model development and the last "quality control" step before the desired application [27-29], is performed. Invalidity of the model could be determined by comparing model predictions with experimental data that is not used in the SRP algorithm. Further, model invalidity can occur if the model predictions conflict with documented biological knowledge of the system. In case of an invalid model, the model parameters are refined in subsequent iterations using the information obtained from the optimal experiment or by expanding the measurement set. The process of model identification is repeated in an iterative manner until an "acceptable" model is obtained.

System

The mathematical model considered in this paper has the following structure:

$$\dot{x} = Ax + Br + C \quad (1)$$

where

$$r = f(x, p), \quad (2)$$

x represents the protein/metabolite/gene concentrations, r the reaction rates, and p the model parameters. This is a very general nonlinear state space model in the variable x. The matrices A and C describe degradation and auto-generation respectively, whereas the matrix B represents the stoichiometry of the biological network. The kinetics among the proteins/metabolites/genes interactions show up in the reaction rates r. The aforementioned model is a continuous time invariant affine system from which a discrete version can be derived by standard techniques using a zero-order hold [30]. The resulting discrete model equation is represented as:

$$x(k+1) = Ax(k) + Br(k) + C \quad (3)$$

where

$$\hat{A} = e^{A\Delta T}$$
$$\hat{B} = (e^{A\Delta T} - I)A^{-1}B$$
$$\hat{C} = (e^{A\Delta T} - I)A^{-1}C$$

The discrete version of the model is used in the SRP estimation algorithm.

Theory

Step 1: Determination of measurement set

The optimal measurement set consists of species whose concentration measurements would have maximum benefit for model identification, *e.g*, parameter identifiability and accuracy. In this work, the measurement set is determined such that the model parameters can be estimated accurately. The assessment of parameter identifiability in a model is crucial prior to parameter estimation from experimental data [31]. Identifiability is closely linked with parametric sensitivity analysis through the Fisher Information Matrix (FIM) [27]. The unidentifiable parameters are determined using the orthogonal procedure proposed by MacAuley and coworkers [19]. Here, a scaled sensitivity coeffcient matrix ($\bar{Z}$) shown below is computed:

$$\bar{\mathbf{Z}} \equiv \begin{bmatrix} \hat{Z}(1) \\ \hat{Z}(2) \\ \vdots \\ \hat{Z}(\mathrm{N}) \end{bmatrix} \qquad \hat{Z}(c) \equiv \begin{bmatrix} \frac{\hat{p}_1}{\eta_1}\frac{\partial \eta_1}{\partial p_1}|\, t = t_c & \cdots & \frac{\hat{p}_k}{\eta_1}\frac{\partial \eta_1}{\partial p_k}|\, t = t_c \\ \vdots & \ddots & \vdots \\ \frac{\hat{p}_1}{\eta_m}\frac{\partial \eta_m}{\partial p_1}|\, t = t_c & \cdots & \frac{\hat{p}_k}{\eta_m}\frac{\partial \eta_m}{\partial p_k}|\, t = t_c \end{bmatrix} \quad (4)$$

where $\{p_1, \ldots, p_k\}$ are the model parameters, $\{\eta_1, \ldots, \eta_m\}$ are the response variables which include all possible measurable quantities, $\{t_1, t_2, \ldots, t_N\}$ are the sampling times for the measurements, and $\hat{p}_i$ is the "initial" parameter value that is either the guess values of the parameter or the value obtained from literature. The orthogonal method is a geometric based approach where the number of identifiable parameters correlates with the rank of the orthogonalization of the scaled sensitivity matrix. The parameters corresponding to the columns of orthogonalized sensitivity matrix are deemed unindentifiable if the norms are smaller than a given tolerance. The details of the orthogonal method are not included here for the sake of brevity.

The next step is to obtain a measurement set that maximizes the expected accuracy in the identifiable parameters (practical identifiability). The Fisher Information Matrix (FIM) along with the Cramer-Rao theorem are used to determine a measurement set such that the estimated parameters have minimum variance. A detailed description of the procedure and its theoretical foundation can be found in [32]. Assuming that the measurement errors are additive and Gaussian, the FIM is given by [33]:

$$\mathrm{FIM} = \mathbf{J}^{\mathrm{T}}\mathbf{W}\mathbf{J} \quad (5)$$

where $\mathbf{W}$ is the inverse of the measurement error covariance matrix and $\mathbf{J}$ denotes the sensitivity coefficient matrix for the measured response variables:

$$\mathbf{J} \equiv \begin{bmatrix} \frac{\partial \eta_1^{measured}}{\partial p_1}|_{t=t_1} & \cdots & \frac{\partial \eta_1^{measured}}{\partial p_r}|_{t=t_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial \eta_h^{measured}}{\partial p_1}|_{t=t_1} & \cdots & \frac{\partial \eta_h^{measured}}{\partial p_r}|_{t=t_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial \eta_1^{measured}}{\partial p_1}|_{t=t_N} & \cdots & \frac{\partial \eta_1^{measured}}{\partial p_r}|_{t=t_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial \eta_h^{measured}}{\partial p_1}|_{t=t_N} & \cdots & \frac{\partial \eta_h^{measured}}{\partial p_r}|_{t=t_N} \end{bmatrix} \quad (6)$$

The quantity $\{p_1, \ldots, p_r\}$ denotes the identifiable parameter vector and $\{\eta_1^{measured}, \ldots, \eta_h^{measured}\}$ denotes the measured response variable vector.

The Cramer-Rao inequality establishes a lower bound on the variance of the identifiable parameters given by:

$$\sigma^2(p_i) \geq (\mathrm{FIM}^{-1})_{ii} \quad (7)$$

The 95% confidence interval (CI) for a parameter is given by:

$$\mathrm{CI} = \hat{p}_i \pm 1.96\sigma(p_i) \quad (8)$$

In Equation (8) the lower bound of the variance is used. Symmetry of the confidence region about the nominal value is assumed. This results in the following definition of the percentage deviation from the nominal value:

$$\%\ \text{error} = \frac{1.96\sigma(p_i)}{p_i} \times 100\% \quad (9)$$

The optimal measurement set is chosen such that the sum of the percentage error (E) for all the identifiable parameters is minimized. In this work, the optimal set is determined by a brute-force search over all combinations of measurement sets subject to restrictions that may be imposed by the system. Doyle and co-workers have developed effcient rational algorithms to determine the optimal measurement set with minimum computational burden [34]. The confidence intervals for non-identifiable parameters are infinitely large and hence are eliminated from the analysis. Identifiability for these parameters can be obtained only by a change in the experimental design or by the selection of an alternative model structure.

Step 2: State Estimation Algorithm

Generally, it is not possible to measure all time-varying components in a metabolic or signaling network. However, there are several techniques from systems engineering to estimate the behavior of unmeasured components given partial measurements of other system constituents. Bastin and Dochain have used an adaptive nonlinear observer for estimation of specific growth rate and biomass concentration [35]. Given accurate models, Extended Kalman Filters (EKF) have had success in several biological applications [36-38]. Artificial Neural Networks (ANN) have also found applications where dynamic models are not available [39-41].

In this work, an extension of Dynamic Flux Balance Analysis (dFBA) [26] is developed to estimate unmeasured concentration and reaction rate trajectories given partial measurement sets. The premise of this approach is straightforward: cellular processes have evolved regulatory structures that optimally use cellular resources. This premise translates into two postulates; (1) network flows are managed to minimize internal accumulation and (2) networks are managed to minimize the number of edges carrying flux at any given time. These two requirements are analogous to a classic problem in automatic control, namely, the State Regulator Problem (SRP). The SRP based estimator uses the measurement set selected from Step 1 to estimate unknown concentration and reaction rate trajectories via a constrained convex programming problem. The SRP estimator constrained by the key measurements captures the optimal cellular behavior of the system.

Estimates of the reaction rates at time step k and protein/gene/metabolite concentrations at time step $k + 1$ are determined by the SRP and the discrete mass balance equations. The SRP must be solved at each sampling interval to obtain estimates of the unknown rates and concentrations. Consider the model (Equation 3); the discrete mass balance equations over a p step horizon are given by:

$$\mathcal{X}_{k+1} = \mathcal{S}^x \mathrm{x}(k) + \mathcal{S}^r \mathcal{R}_k + \mathcal{S}^c \quad (10)$$

Where the matrices $\mathcal{X}_{k+1}$, $\mathcal{R}_k$, $\mathcal{S}_x$ and $\mathcal{S}_c$ are defined as

$$\mathcal{X}_{k+1} = \begin{bmatrix} \mathbf{x}(k+1) \\ \mathbf{x}(k+2) \\ \vdots \\ \mathbf{x}(k+p) \end{bmatrix} \quad \mathcal{R}_k = \begin{bmatrix} \mathbf{r}(k) \\ \mathbf{r}(k+1) \\ \vdots \\ \mathbf{r}(k+p-1) \end{bmatrix} \quad \mathcal{S}^x = \begin{bmatrix} \hat{\mathbf{A}} \\ \hat{\mathbf{A}}^2 \\ \vdots \\ \hat{\mathbf{A}}^p \end{bmatrix} \quad \mathcal{S}^c = \begin{bmatrix} \hat{\mathbf{C}} \\ (\hat{\mathbf{A}}+\mathbf{I})\hat{\mathbf{C}} \\ \vdots \\ \sum_{i=0}^{p-1} \hat{\mathbf{A}}^i \hat{\mathbf{C}} \end{bmatrix}$$

and the matrix $\mathcal{S}_r$ is given by

$$\mathcal{S}^r = \begin{bmatrix} \hat{\mathbf{B}} & 0 & \cdots 0 \\ \hat{\mathbf{A}}\hat{\mathbf{B}} & \hat{\mathbf{B}} & \cdots 0 \\ \vdots & \vdots & \ddots \vdots \\ \hat{\mathbf{A}}^{p-1}\hat{\mathbf{B}} & \hat{\mathbf{A}}^{p-2}\hat{\mathbf{B}} & \cdots \hat{\mathbf{B}} \end{bmatrix}$$

The SRP estimator is a Quadratic Program (QP) with two cost terms, the cost of intermediate accumulation and the cost of operating a network reaction. The SRP penalizes intracellular metabolite or protein accumulation, but does not explicitly forbid it. Moreover, because reactions introduce an additional cost, the SRP only utilizes those reactions required to satisfy the mass balances and thermodynamic constraints. Formally, the estimation problem is given by:

$$\min_{\mathcal{R}_k} \left[\mathcal{X}_{k+1}^T \mathbf{W}_\mathbf{X} \mathcal{X}_{k+1} + \mathcal{R}_k^\mathrm{T} \mathbf{W}_\mathbf{R} \mathcal{R}_k \right] \quad (11)$$

subject to:

$$\mathcal{X}_{k+1} \geq \mathbf{0} \quad (12)$$

$$\alpha_r(k) \le \mathcal{R}_k \le \beta_r(k) \quad (13)$$

$$\mathcal{X}^*_{k+1} - \Delta^X_{tolerance} \le \Xi^X \mathcal{X}_{k+1} \le \mathcal{X}^*_{k+1} + \Delta^X_{tolerance} \quad (14)$$

The SRP problem is subject to non-negativity constraints (Equation 12), flux-directionality constraints (Equation 13) and constraints imposed by the measurement set. Specifically, constraint of Equation 14 forces state estimates belonging to the measurement set to equal the corresponding measured value. The quantities $\mathcal{X}^*_{k+1}$ denote measurements that may have been corrupted with noise. $\Delta^X_{tolerance}$ specifies the tolerance around the measurement within which the estimate is constrained to lie (incorporated to avoid numerical inconsistencies that may arise due to noisy measurements). The term Ξ^X defines the measurement set. The matrix $\mathbf{W_x}$ defines the cost of intermediate accumulation whereas the matrix $\mathbf{W_R}$ represents the reaction cost:

$$\mathbf{W_X} = \begin{bmatrix} w_x & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_x \end{bmatrix} \quad \mathbf{W_R} = \begin{bmatrix} w_r & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_r \end{bmatrix} \quad (15)$$

For the caspase system considered in this work, the w_x and w_r are taken to be the order of magnitude of the inverse of the maximum value of the corresponding state or rate. This requires only approximate information regarding ranges of the protein concentrations and identification of the slow versus the fast reactions:

$$(w_x)_{ii} = \text{order of magnitude}\left[\frac{1}{\max(x_i)}\right] \quad (16)$$

$$(w_r)_{ii} = \text{order of magnitude}\left[\frac{1}{\max(r_i)}\right] \quad (17)$$

Step 3: Parameter estimation

The estimates of the concentration profiles and the reaction rates allow efficient determination of the parameter values by decoupling the full parameter estimation into multiple sets. Each set consists of parameters associated with one reaction rate. The parameters are obtained by minimizing the difference between the estimates of each reaction rate and that predicted by the kinetics $r(x, p)$, which is a function of the concentrations. In case of the first iteration the minimization follows:

$$\min_p \left[(r_i - r_i(x,p))^T (r_i - r_i(x,p)) \right] \ \forall \ i = 1, \cdots, N_R \quad (18)$$

where r_i are the individual reaction rates and N_R is the total number of reactions. The kinetic parameters associated with a reaction rate equation are determined independently from those with other reactions, *i.e.*, the parameter estimation is decoupled with respect to each reaction.

For subsequent iterations, the Bayesian estimation formulation in [42] is used. In this formulation, in addition to the difference between the estimates of the reaction rates and the model predictions, the deviations of parameter values from those obtained after the previous iteration are minimized. The formulation can be represented as:

$$\min_p \quad \Big[(r_i - r_i(x,p))^T V_\in^{-1} (r_i - r_i(x,p)) + (p - p_0)^T V_p^{-1} (p - p_0) \Big] \ \forall \ i = 1, \cdots, N_R \quad (19)$$

In the above equations, $\hat{r}_i$ is the estimate of the i^{th} reaction rate and $\hat{x}$ is the estimates of the concentrations obtained from the SRP algorithm, $r_i(\hat{x}, p)$ is the predicted rate of the i^{th} reaction from the kinetics in Equation 2, p are the parameters associated with the i^{th} reaction rate, p_0 are the parameter values obtained from the previous iteration, and V_ε and V_p are the variances of the estimates of the reaction rates and the prior parameter estimates. The parameter variances are determined using the Fisher Information Matrix (Equation 7). The variances of the non-identifiable parameters are infinite and penalty for deviations for these parameters are not considered in Equation 19. The variance for the estimates of the reaction rates can be determined from the expected noise in the measurements from which the estimates are obtained.

Step 4: Model invalidation tests

Given the iterative nature of this framework, a termination criterion must be established. Poolla et al. [29] have shown that for certain experimental data, it is not possible to confirm whether the model is really valid; however, one can conclude whether the model is not contradicted by the given data. Model (in)validation tests are usually based on the difference between the simulated and measured output and some statistics about these differences. Typical statistics for the model errors include maximum absolute value, mean value and variance [28]. In this work, model invalidity is tested by determining the model prediction errors using the estimated parameters. This error is calculated as:

$$\bar{E}_{x_i}(k) = \frac{\left| x_i^{predicted}(k) - x_i^{measured}(k) \right|}{\max_j \left[x_i^{measured}(j) \right]} \times 100\% \quad (20)$$

To implement this test, experimental data that was not used in the SRP algorithm is required. The statistic used is the maximum and mean value of the errors for the measured states. When the prediction errors are below a certain desired value, the iterations are terminated.

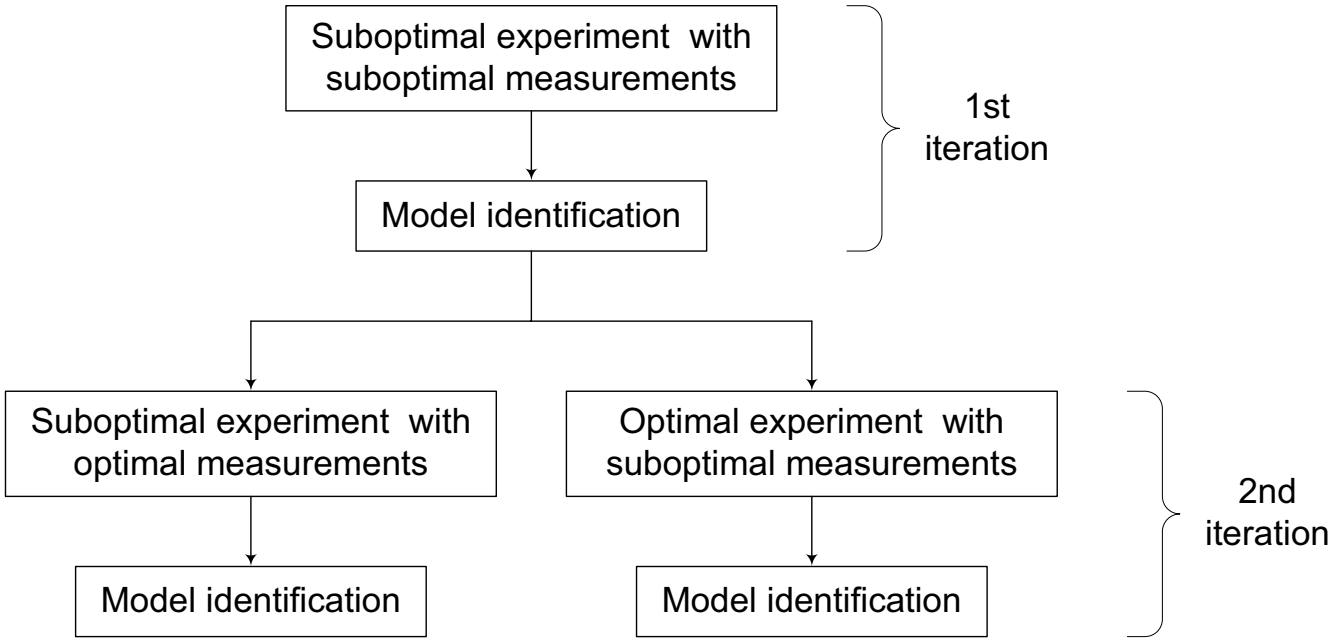

Figure 2
Sequence of simulations for model identification to test efficiency of optimal experiment design and optimal measurement selection.

Step 5: Model-based optimal experiment design

The optimal experiment design determines the optimal experiment to be performed for the next iteration such that there is maximum information content in the measurements. This would maximize the accuracy of the estimated parameters. The model-based optimal experiment design uses the Fisher Information Matrix as a measure of the amount of information contained in a given set of measurements about the model parameters [43]. The optimization searches through the space of experimental conditions or some parameterizations of the experimental protocol. For example, an optimal ligand input can be parameterized into a time series profile such that the optimization variables are the levels of ligand at different times (usually equally spaced in time). Naturally, the optimization will be restricted by the limitations in the experimental conditions and apparatus.

There exist several FIM-based optimality measures that quantify the overall informativeness of the measurements [32]. Among these, parameter identifiability and the D-optimality are the most widely used measures. For accurate model identification it is critical that maximum number of the parameters be estimated accurately. Thus, maximizing the number of identifiable parameters is the primary criterion proposed for determining the next experimental design. The orthogonal procedure proposed by McAuley and co-workers [19] is used to determine the number of identifiable parameters. There can be multiple experimental designs with the same maximum number of identifiable parameters. The selection among these is done so as to maximize the informativeness of measurement data. For this purpose, the D-optimality criteria is proposed. The use of D-optimality translates to minimizing the confidence interval of all the identifiable parameter estimates. The optimal experiment design criterion is shown as follows:

$$\max \quad \det(\mathrm{FIM}) \qquad (21)$$
$$\text{s.t.} \left[\max_{E \in \mathbf{E}} \; r \right]$$

where **E** denote the feasible experimental conditions (defined by constraints in experiments), FIM is given in Equation 5 and r denotes the number of identifiable parameters. Thus, the identifiability is maximized in the sense that the hyper-dimensional confidence interval is minimized. For experimental designs with the maximum number of identifiable parameters, the one with the highest determinant of the Fisher Information Matrix is selected as the optimal design for the next experiment.

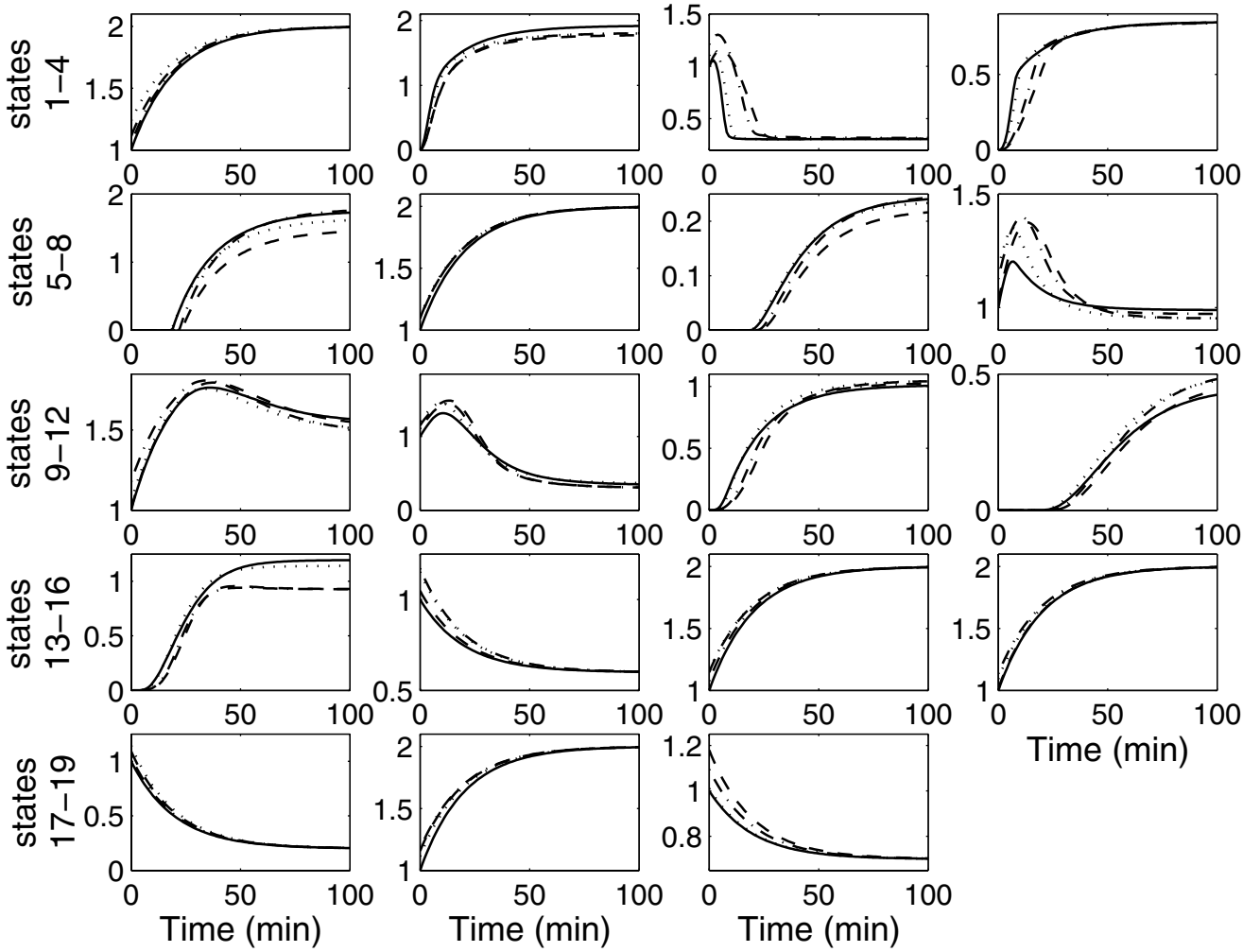

Figure 3
Prediction profiles of the 19 protein concentrations for the test experiment for the caspase system. Solid line: real system; dashed line: prediction with estimated parameters after first iteration (suboptimal experiment with suboptimal measurements); dash-dotted line: prediction with estimated parameters after second iteration (suboptimal experiment with optimal measurements); dotted line: prediction with estimated parameters after second iteration (optimal experiment with suboptimal measurements)

A case study

The proposed iterative model identification is applied to the function of caspase-8 and caspase-9 in apoptosis. Caspase enzymes are at the core of the cell's suicide machinery. These enzymes are activated either by an external signal or by stress, and activated enzymes will then dismantle the cells. Varner and co-workers have developed a model for the caspase function in apoptosis [7]. The model describes the key elements of receptor-mediated and stress-induced caspase activations. The model consists of 19 states (enzymes) and 11 reactions with 27 parameters (11 rate constants and 16 saturation constants; see Appendix for additional details of the model). The *in silico* experiments in this study use the Varner model as the "actual" system. Measurements are assumed to be obtained from this "actual" system corrupted with up to 10% noise. The iteration starts using a "initial" parameter set, generated by perturbing the parameter values of the Varner model (considered as "exact") by 70–100%. The external and stress signals that activate the caspase system are considered as the manipulated variables in Step 5. Model refinement is performed either by

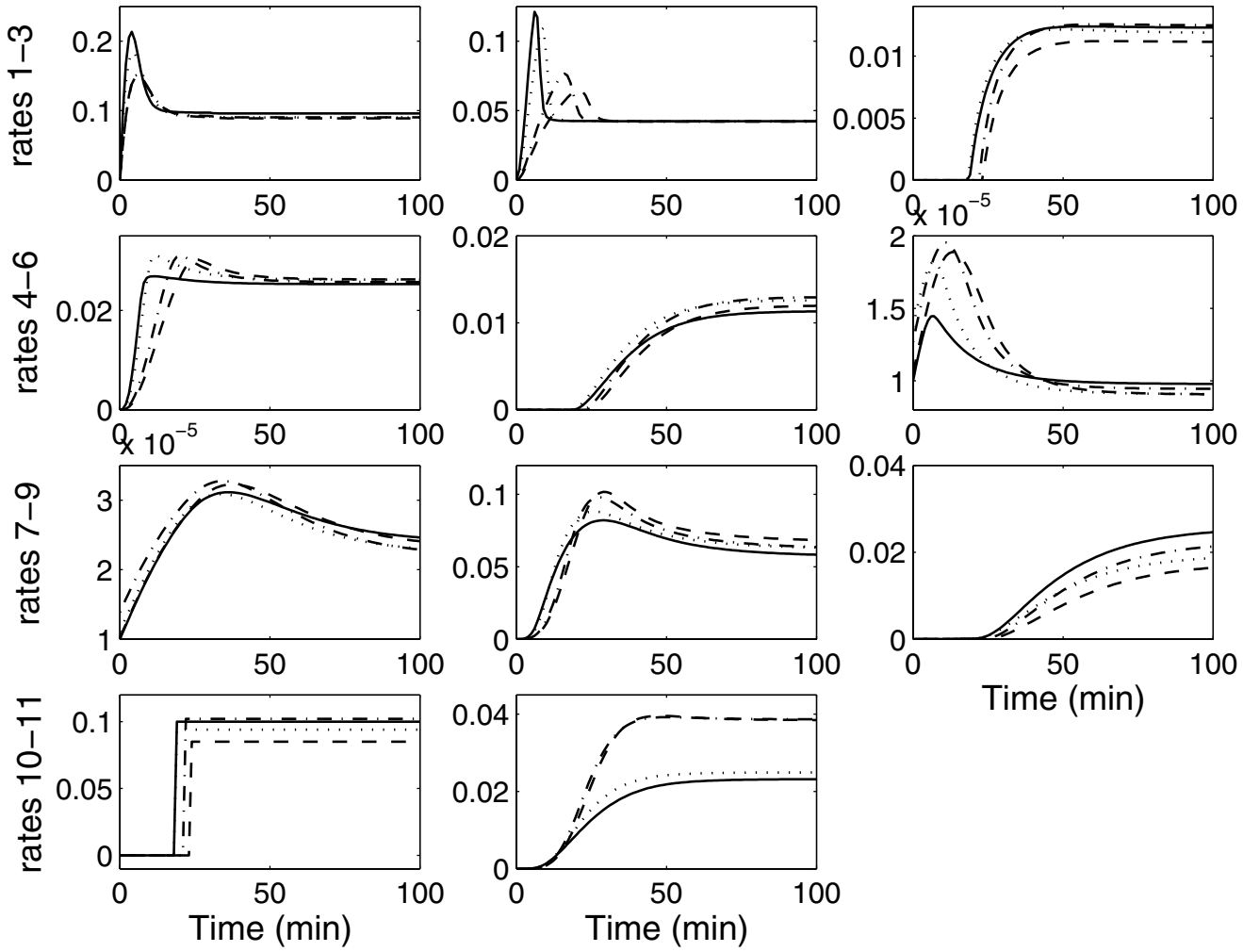

Figure 4
Prediction profiles of the 11 reaction rates for the test experiment for the caspase system. Solid line: real system; dashed line: prediction with estimated parameters after first iteration (suboptimal experiment with suboptimal measurements); dash-dotted line: prediction with estimated parameters after second iteration (suboptimal experiment with optimal measurements); dotted line: prediction with estimated parameters after second iteration (optimal experiment with suboptimal measurements)

determination of an optimal experiment or by optimal refinement of the measurement set. The performance of each of these criteria for model refinement requires to be tested. Therefore, in the first iteration, a "preliminary" suboptimal experiment with a suboptimal measurement set is considered. Moreover, in most cases of model identification, it is expected that preliminary experimental data is available. This data is usually obtained from a suboptimal experiment design and does not include the optimal set of measurements. The second iteration is performed in two ways; one by improving the measurement set and second by improving the experiment design. This tests the performance of both refinement criteria. The sequence of events is shown in Figure 2. It should be noted that it would be best to use optimal experiment design along with the optimal measurement set. However, it may not always be possible to do so due to feasibility issues specific to the particular system. Hence, this approach is not considered in this work. Model identification under less constrained conditions using a similar framework is included in [44].

After the first iteration, it is observed that there is a significant improvement in the predictions with the estimated

parameters for both the protein concentrations and the reaction rates, as shown in Figures 3 and 4. The high errors with the "initial" parameters demonstrate that there is no bias in the results based on the starting guess values for the parameters and that there is indeed an improvement in prediction of both the protein concentrations and the reaction rates. However, the improvement is not suffcient as observed from the invalidation test (see Methods). This warrants a second iteration.

In general, the model predictions improve with the second iteration, as shown again in Figures 3 and 4. However, it is observed that the predictions are better for the case with the optimal experiment design, in spite of a suboptimal measurement set. This is due to the fact that model performance depends strongly on the accuracy of the estimated parameters. Using the suboptimal experiment, only 14 of the 27 parameters were identifiable. The optimal measurement set simply improved the confidence in these 14 parameters. On the other hand, the optimal experiment increased parameter identifiability to 18 parameters even with the suboptimal measurement set. These result indicates that performance of model identification is strongly linked with parameter identifiability.

Table 1: Experimental procedures used in model identification for the caspase system. Both receptor and stress signals are increased from zero to their maximum value in 30 minutes after which they are held constant (units same as in Varner model [7]).

	Maximum receptor signal	Maximum stress signal
Preliminary Experiment	0.24	0
Optimal Experiment	0.09	0.045
Test Experiment	0.15	0.030

Table 2: Measurement sets for the SRP estimator for the caspase system.

	States measured						
Optimal set	2	3	5	7	10	11	12
Suboptimal set	2	3	4	5	7	10	12

Discussion

In the proposed algorithm, the network topology and the mechanism of interactions in the pathway are assumed to be known. Several approaches have been proposed in literature to determine the network connectivity from experimental data [45-47]. In the case of unknown connectivity, these approaches should be used prior to the proposed model identification. The mismatch between the model and the actual network can appear in two different aspects of the algorithm. First, the SRP step can fail because there exist no feasible state and flux estimates that satisfy the measurement constraints. Such scenario arises mainly due to the network topology mismatch, and has low probability to occur due to the large degree of freedoms in a typical biological system. Alternatively, a well-designed model (in)validation step catches the mismatch between the model and the real system, *e.g.*, an independent measurement set contradicts the model prediction. Also, if multiple models are proposed for a particular biological process, model discrimination methods [24,48] can be used to identify the correct model structure. The effect of incorrect connectivity and/or mechanism would depend on the degree of mismatch and is case dependent.

The fact that cellular processes are carried out in an optimal manner lends tremendous promise to the success of this approach. In case the assumed optimality does not represent the *in vivo* behavior, the estimates from SRP may be inaccurate. However, the measurements (through the constraints in Equation 14) can attenuate this problem by restricting part of the estimates to match the observations. The parameter estimates may also be inaccurate if the real system deviates considerably from the assumed optimal behavior and this deviation is not captured by the measurements. The model identification framework has applicability to all systems that could be represented in the form shown in Equations 1 and 2. All biological processes are complex, interconnected networks. A feature common to these processes is that they have a fixed connectivity. The proposed algorithm for model development could be applied to metabolic networks, signaling processes and gene networks.

The computation burden for solving the SRP is minimal as it involves only a quadratic programming problem. The parameter estimation is also not computationally intensive due to the decoupling facilitated by SRP. A global optimization algorithm can also be used in the parameter estimation instead of the gradient search method to avoid convergence to local minima. However, the computation burden of parameter estimation will increase. To avoid local minima and high computation cost, the first few iterates (1 or 2) can utilize a global optimization method, while the remaining iterations can implement a gradient search algorithm. Due to the iterative nature of the approach, errors in parameter estimates can be tolerated as the corrections will be made in the next iteration with the optimal experiment. The optimal measurement selection is performed by a brute force search in this work. For very large systems the computation burden for this process grows exponentially. Computationally efficient

Table 3: Confidence intervals for the model parameters of the caspase system. Case 1: suboptimal experiment with optimal measurement set; Case 2: suboptimal experiment with suboptimal measurement set; Case 3: optimal experiment with suboptimal measurement set.

No.	Case 1		Case 2		Case 3	
	CI	% E	CI	% E	CI	% E
1	1.05 ± 0.25	23.31	1.05 ± 0.26	24.43	0.55 ± 0.04	07.18
2	1.65 ± 0.06	03.69	1.65 ± 0.06	03.50	0.95 ± 0.02	02.20
3	0.52 ± 0.01	02.45	0.52 ± 0.01	02.45	0.29 ± 0.04	15.53
4	1.69 ± 6e-3	00.34	1.69 ± 5e-3	00.32	1.69 ± 0.54	32.23
5	0.62 ± 0.01	01.81	0.62 ± 0.01	01.81	0.62 ± 0.14	22.38
6	NI	-	NI	-	NI	-
7	NI	-	NI	-	NI	-
8	0.87 ± 0.11	12.51	0.87 ± 0.11	12.48	1.33 ± 0.18	13.28
9	0.91 ± 0.25	27.87	0.91 ± 0.87	95.69	0.75 ± 0.44	57.85
10	0.15 ± 2e-3	01.49	0.15 ± 2e-3	01.49	0.09 ± 6e-4	00.70
11	0.23 ± 8e-3	03.39	0.23 ± 0.02	06.67	0.23 ± 0.04	17.69
12	2.12 ± 0.56	26.23	2.12 ± 0.59	27.59	17.8 ± 1.60	08.99
13	0.13 ± 2e-3	01.18	0.13 ± 2e-3	01.17	0.10 ± 8e-4	00.81
14	NI	-	NI	-	NI	-
15	NI	-	NI	-	14.7 ± 4.39	30.55
16	NI	-	NI	-	128 ± 30.8	24.03
17	NI	-	NI	-	NI	-
18	NI	-	NI	-	NI	-
19	NI	-	NI	-	127 ± 3.38	2.66
20	NI	-	NI	-	NI	-
21	(3.21 ± 0.06) × 1e3	01.76	(3.21 ± 3.14) × 1e3	97.51	(3.22 ± 1.16) × 1e3	36.05
22	NI	-	NI	-	NI	-
23	NI	-	NI	-	NI	-
24	NI	-	NI	-	NI	-
25	NI	-	NI	-	(3.09 ± 1.05)	33.95
26	0.79 ± 0.12	14.57	0.79 ± 0.12	14.52	1.04 ± 0.22	21.07
27	8.65 ± 0.31	03.60	NI	-	8.62 ± 3.62	42.04

algorithms for optimal measurement selection [34] can be used. Efficient measurement selection algorithms and the decoupling of parameter estimation for individual reaction rates into separate optimization problems result in good scalability properties of the proposed algorithm for large scale systems. One limitation of the approach is that it is dependent on the weights in Equation 15 for the minimization of the cellular resources. The choice of weights used in this work has provided accurate results but it requires information of the order of magnitude of the concentrations and rates of the system under study [44]. Efficient schemes for determining the weights for metabolic networks has been developed by Varner and coworkers [49].

Conclusion

An iterative methodology for model identification from experimental data is developed in this paper. Identifiability tests are performed for an optimal measurement set selection for a given experimental design. The optimal measurements represent maximum information such that the model identification process is maximally benefitted. The model identification process is decoupled into two parts. In the first, the measurements are used to estimate all the unmeasured quantities of the system. This is achieved using the State Regulator Problem (SRP) formulation which is based on the assumption that the cell is an optimal strategist and uses its resources in an optimal manner. The SRP algorithm developed in this work has shown promising results. The average errors in the estimates for a significant fraction of the unmeasured responses is less than 10%. The accuracy of the estimates obtained by the SRP decreases with decrease in the information content due to suboptimal measurement set. In the second part, the full state and rate estimates are used to determine the model parameters. The decoupling relaxes considerable computation burden compared to estimating all model parameters simultaneously from the limited measurements. In the final step, a model-based experiment design determines the optimal experimental

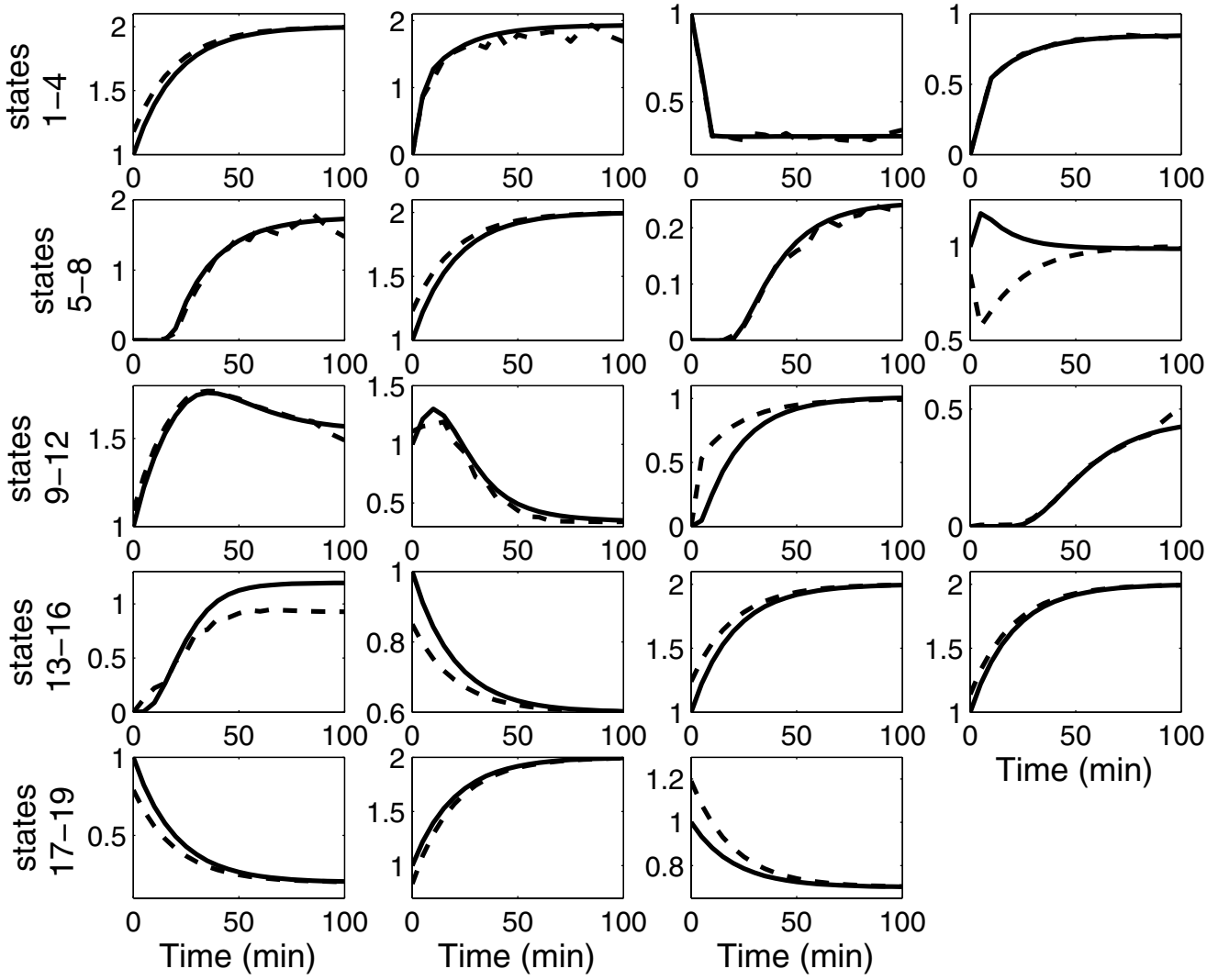

Figure 5
Profiles of the 19 protein concentrations for the caspase system. Solid line: actual system; dashed line: estimate by the SRP algorithm with the suboptimal experiment with suboptimal measurement set (first iteration)

procedure that generates the most informative measurements for the next iteration. A strong dependence is observed between parameter identifiability and model performance. Thus, it is critical that the experiment design and measurement set be chosen such that maximum number of parameters are identifiable.

Tools developed for quantitative analysis of the dynamics of cellular pathways have tremendous potential in improving the predictive capabilities of biological systems especially in cases where experimental data is available but the kinetic parameters of the pathway reactions are unknown. The model developing tools are used for a host of applications and systems analysis.

The measurement selection algorithm presented in this work is freely available as part of a model analysis and development toolkit, BioSens [50].

Methods

Measurement set selection

The measurement selection analysis is performed using the "initial" parameter set for the "preliminary" experimental conditions shown in Table 1. A sampling time of 5 minute is assumed for a total simulation time of 100

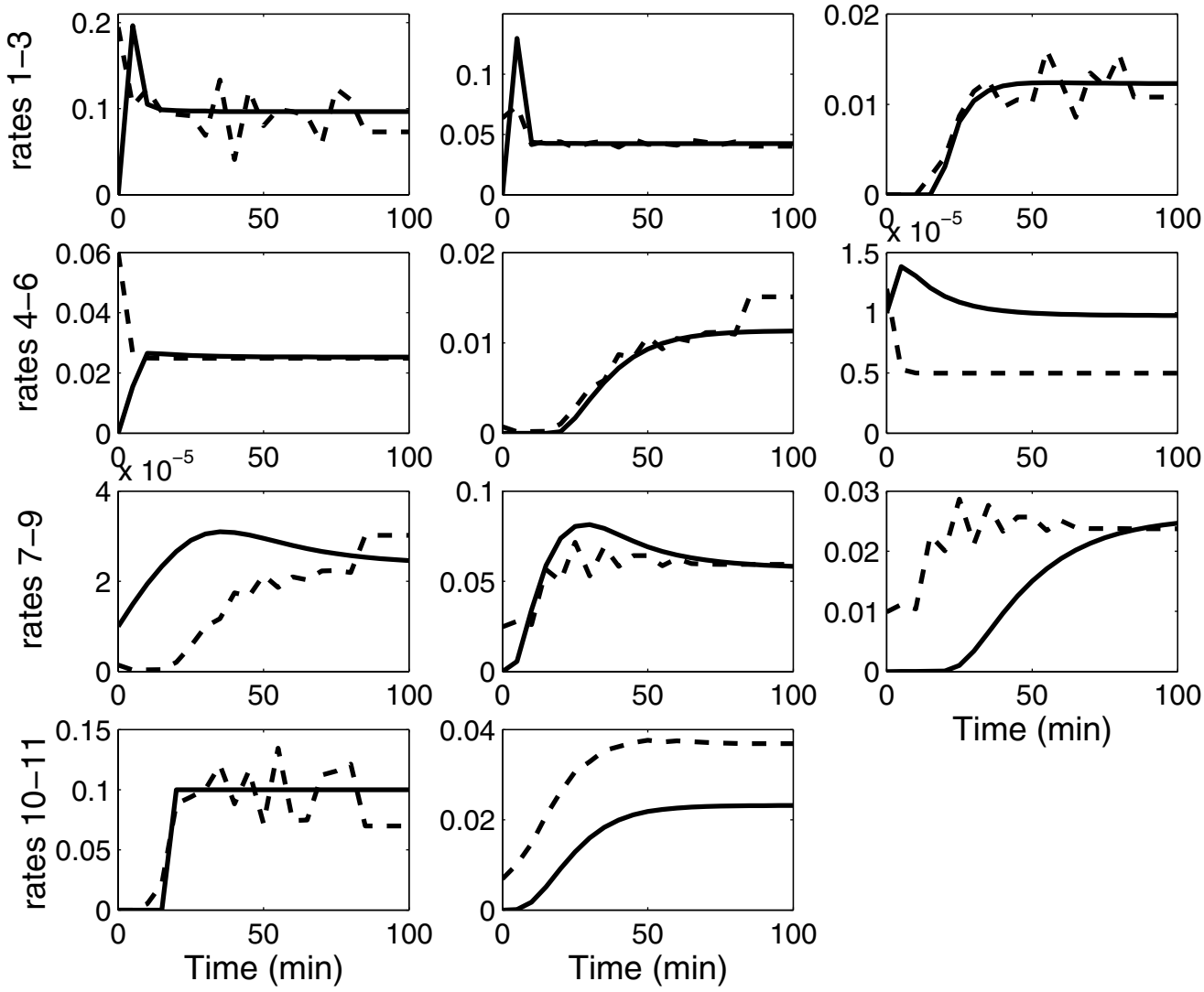

Figure 6
Profiles of the 11 reaction rates for the caspase system. Solid line: actual system; dashed line: estimate by the SRP algorithm with the suboptimal experiment with suboptimal measurement set (first iteration)

minutes. Using the orthogonal procedure [19], the non-identifiable parameters are eliminated (13 of the 27 parameters). Perturbations of the non-identifiable parameters have no noticeable effect upon system dynamics for the given experimental protocol or have a strong correlation with the perturbation of one or more identifiable parameters. The non-identifiable parameters include the rate constants for auto-activation of the procaspases (parameters 5 and 6). The auto-activation is orders of magnitude lower compared to the activation by initiator. The small contribution of the auto-activation cannot be independently captured by the measurements and hence leads to non-identifiability. All other non-identifiable parameters are reaction saturation constants; the dynamics of which are not captured by the measurements at 5 minute sampling time for the given measurement noise. Equation 8 is used to estimate a bound on the confidence interval of the identifiable parameters and the deviation from the nominal value is calculated using Equation 9. A measurement set of 7 protein concentrations is assumed

Table 4: Estimation error for the protein concentrations in caspase system.

Protein (x_i)	First iteration[a]		Second iteration[b]		Second iteration[c]	
	Max. E_{x_i}	Avg. E_{x_i}	Max. E_{x_i}	Avg. E_{x_i}	Max. E_{x_i}	Avg. E_{x_i}
1	08.93	01.91	08.93	01.91	08.93	01.91
2	12.49	04.74	10.54	03.55	12.04	04.32
3	03.64	01.38	04.02	01.31	03.27	01.40
4	02.16	00.82	02.38	00.78	01.94	00.83
5	14.63	03.95	12.46	03.35	13.87	03.62
6	11.54	02.47	11.54	02.47	11.54	02.47
7	09.04	02.58	07.48	02.07	08.47	02.27
8	51.10	10.65	12.65	04.40	51.63	11.24
9	05.11	01.46	05.11	01.58	05.09	01.49
10	10.22	04.18	08.45	02.35	09.02	03.65
11	48.30	09.41	07.79	02.74	50.75	10.49
12	18.56	02.76	17.67	02.59	18.19	02.69
13	22.27	14.66	22.10	15.40	13.13	03.45
14	15.00	03.21	15.00	03.21	15.00	03.21
15	12.04	02.58	12.04	02.58	12.04	02.58
16	07.02	01.50	07.02	01.50	07.02	01.50
17	21.00	04.50	21.00	04.50	21.00	04.50
18	08.42	01.80	08.42	01.80	08.42	01.80
19	19.00	04.07	19.00	04.07	19.00	04.07

[a] suboptimal experiment with suboptimal measurement set
[b] suboptimal experiment with optimal measurement set
[c] optimal experiment with suboptimal measurement set

Table 5: Estimation error for the reaction rates in caspase system.

Reaction rate (r_i)	First iteration[a]		Second iteration[b]		Second iteration[c]	
	Max. E_{r_i}	Avg. E_{r_i}	Max. E_{r_i}	Avg. E_{r_i}	Max. E_{r_i}	Avg. E_{r_i}
1	98.97	15.98	103.3	15.91	67.05	14.14
2	49.09	05.68	48.76	05.68	45.11	06.54
3	30.86	11.60	30.46	11.64	29.49	11.93
4	225.9	14.86	38.71	18.08	221.1	16.25
5	34.32	10.44	33.49	09.79	33.87	10.30
6	61.73	38.69	66.30	41.42	65.93	38.27
7	79.40	37.65	80.00	38.12	81.26	37.20
8	35.03	10.47	36.67	10.75	36.03	11.34
9	112.1	38.78	109.9	37.32	112.5	36.55
10	34.50	17.92	36.56	17.92	29.43	10.44
11	76.59	62.44	72.72	60.33	29.76	10.03

[a] suboptimal experiment with suboptimal measurement set
[b] suboptimal experiment with optimal measurement set
[c] optimal experiment with suboptimal measurement set

to be available. No measurements of the reaction rates are available. The choice of the measurement set was such that the maximum confidence was obtained for the identifiable parameters. The choice was made by a rigorous brute force search among all possible combinations. The optimal measurement set is shown in Table 2 and the confidence intervals are shown in Table 3. All the identifiable parameters have a confidence window with percentage error less than 30%. Further reduction in the percentage errors would require assuming more measurements in

Table 6: Error in the model predictions for the protein concentrations using the "initial" parameters, the estimated parameters after the first iteration, and the estimated parameters after second iteration for the "test" experiment of the caspase system.

(x_i)	Initial parameters		First iteration[a]		Second iteration[b]		Second iteration[c]	
	Max. E_{r_i}	Avg. E_{r_i}	Max. E_{r_i}	Avg. E_{r_i}	Max. E_{r_i}	Avg. E_{r_i}	Max. E_{r_i}	Avg. E_{r_i}
1	11.21	02.26	05.66	01.14	02.54	00.51	12.04	02.43
2*	38.08	15.37	16.42	07.58	15.92	06.44	06.02	04.55
3*	90.59	16.37	72.90	09.57	78.27	09.91	32.07	02.11
4*	53.58	09.69	45.41	05.91	40.71	04.19	15.79	01.31
5*	75.88	27.09	21.47	13.56	16.54	02.81	06.40	03.96
6	01.74	00.35	04.82	00.97	04.76	00.96	05.52	01.11
7*	69.30	31.54	15.13	09.34	08.36	02.43	02.87	01.35
8	40.52	30.01	21.11	05.55	20.80	04.79	14.21	03.94
9	16.20	10.53	01.95	00.83	10.05	02.95	03.01	01.99
10*	56.81	38.13	17.30	06.43	12.93	05.32	15.92	03.01
11*	47.03	34.93	24.23	06.28	18.50	03.95	05.18	03.79
12*	66.80	36.63	07.88	03.04	13.61	04.96	12.22	06.98
13	53.52	39.26	22.14	14.01	22.25	13.75	04.66	02.79
14	11.13	02.24	04.75	00.96	17.06	03.44	15.18	03.06
15	11.69	02.36	07.36	01.48	03.80	00.77	07.90	01.59
16	05.84	01.18	00.72	00.15	06.79	01.37	04.65	00.94
17	10.47	02.11	09.19	01.85	03.77	00.76	14.38	02.90
18	10.61	02.14	07.92	01.60	08.75	01.77	05.66	01.14
19	13.13	02.65	17.94	03.62	09.46	01.91	01.10	00.22

[a]suboptimal experiment with suboptimal measurement set
[b]suboptimal experiment with optimal measurement set
[c]optimal experiment with suboptimal measurement set
*measured protein concentration for the invalidation test

Table 7: Error in the model predictions for the reaction rates using the "initial" parameters, the estimated parameters after the first iteration, and the estimated parameters after second iteration for the "test" experiment of the caspase system.

(r_i)	Initial parameters		First iteration[a]		Second iteration[b]		Second iteration[c]	
	Max. E_{r_i}	Avg. E_{r_i}	Max. E_{r_i}	Avg. E_{r_i}	Max. E_{r_i}	Avg. E_{r_i}	Max. E_{r_i}	Avg. E_{r_i}
1	76.90	08.74	39.88	04.68	39.23	04.08	15.83	03.08
2	90.69	07.75	80.38	05.66	74.15	05.66	48.14	02.59
3	98.85	59.01	48.39	10.98	39.17	03.33	07.83	02.33
4	92.70	39.32	72.30	11.25	63.99	08.22	20.95	05.45
5	77.79	46.55	14.52	04.71	14.38	07.99	12.95	08.80
6	87.73	60.74	43.40	10.46	43.64	09.30	26.27	07.14
7	32.22	20.58	3.92	01.57	12.40	05.10	05.13	03.52
8	74.11	23.75	28.21	13.53	23.36	09.19	11.27	04.83
9	65.88	37.41	33.29	20.39	14.47	09.46	24.21	12.18
10	100.0	53.42	100.00	16.39	100.0	04.53	06.00	04.87
11	36.65	16.09	80.79	53.88	79.09	55.00	13.76	08.04

[a]suboptimal experiment with suboptimal measurement set
[b]suboptimal experiment with optimal measurement set
[c]optimal experiment with suboptimal measurement set

addition to the current 7 measurements, a faster sampling of the available measurements or a new experimental protocol.

Model identification

1st iteration

The "preliminary" experiment (Table 1) with a suboptimal measurement set (Table 2) is used for obtaining the

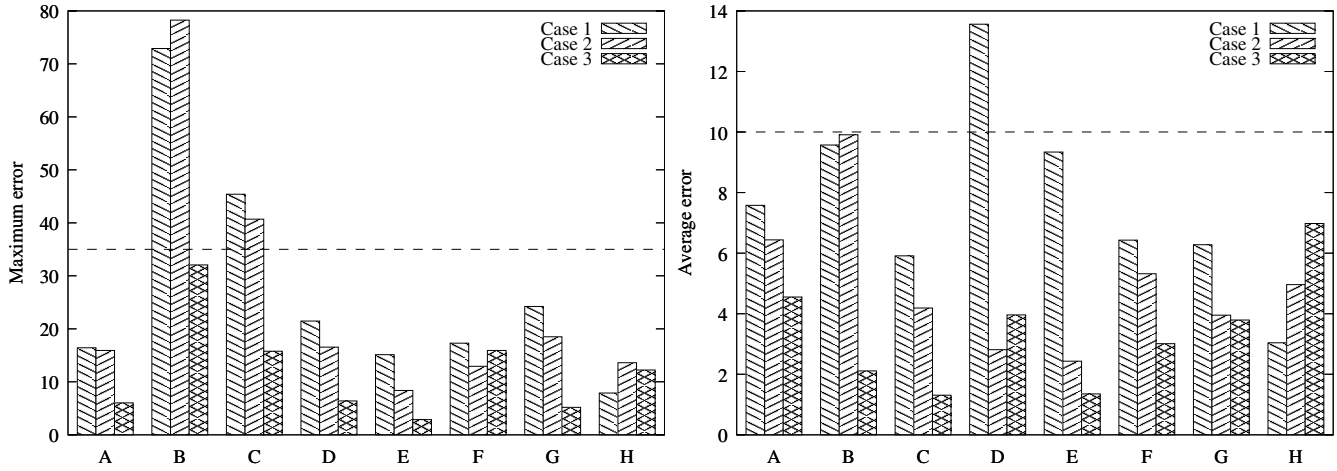

Figure 7
Maximum and average prediction errors for the measured protein concentrations. Case 1: first iteration (suboptimal experiment with suboptimal measurement); Case 2: second iteration (suboptimal experiment with optimal measurements); Case 3: second iteration (optimal experiment with suboptimal experiment). The measured protein concentrations are (A) FAS/FASL complex; (B) FADD; (C) FAS/FASL-FADD complex; (D) cytochrome c; (E) Apaf-1-cytochrome c complex; (F) executioner procaspase; (G) caspase-8; (H) caspase-9.

estimates of the unknown concentration and reaction rate trajectories using the SRP algorithm. The sampling time is taken to be 5 minutes with a prediction horizon of 4. A higher prediction horizon showed no appreciable change in the estimates. The initial condition of the protein concentrations is assumed to be equal to the corresponding "actual" system corrupted by up to 25% relative error. Measurements are obtained from the "actual" system with up to 10% noise. The tolerance of the concentration measurements (14) is taken to be 5%. Figures 5 and 6 show estimated versus "actual" profiles for protein and reaction rate trajectories respectively.

The estimation error is determined by calculating the difference between the estimated and "actual" value for a rate/state at time k scaled by the maximum value over the entire simulation. Equation 22 shows the estimation error for concentration (the equation for reaction rates is identical). The scaling is done with respect to the maximum value in order to prevent misleading analysis at low concentrations or reaction rates:

$$E_{x_i}(k) = \frac{\left| x_i^{estimate}(k) - x_i^{actual}(k) \right|}{\max_j [x_i^{actual}(j)]} \times 100\% \qquad (22)$$

The estimation errors (Equation 22) are given in Tables 4 and 5. Overall, it is observed that the estimates are fairly accurate and the system dynamics are captured. The average errors are less than 15% for all the state estimates and for most of the reaction rate estimates. Poor estimates, especially during the initial sampling times are mainly caused by the mismatch in the initial conditions. Further, the noise in the measurements results in fluctuations in some of the estimates. It should be noted that the estimation errors would not be available in real situations because the "actual" profiles are unknown. These are included here as a proof of concept.

The estimates are then used to determine the model parameters by solving the optimization problem in Equation 18. The optimization to determine the parameters is a nonlinear program for the caspase system in which the model equations for the reaction rates are nonlinear with respect to the parameters. The nonlinear optimization is solved using the MATLAB routine *fmincon* which employs a gradient descent search method. As a starting point for the search, the "initial" parameter values are used. The optimization is solved for each reaction rate separately to obtain all the parameters in the model equations. Figures 3 and 4 compare the prediction profiles of the protein concentration and the reaction rates obtained with the estimated parameters with the "actual" profiles. These profiles are for an experiment condition ("test" conditions in Table 1) that is different from the one used for model identification. As a model invalidation test, the prediction errors are calculated using Equation 20. A threshold of 35% for the maximum error and 10% for the average error can be considered to be stringent. Figure 7 shows the result for the measured protein concentrations. It is observed that after the first iteration, the threshold is

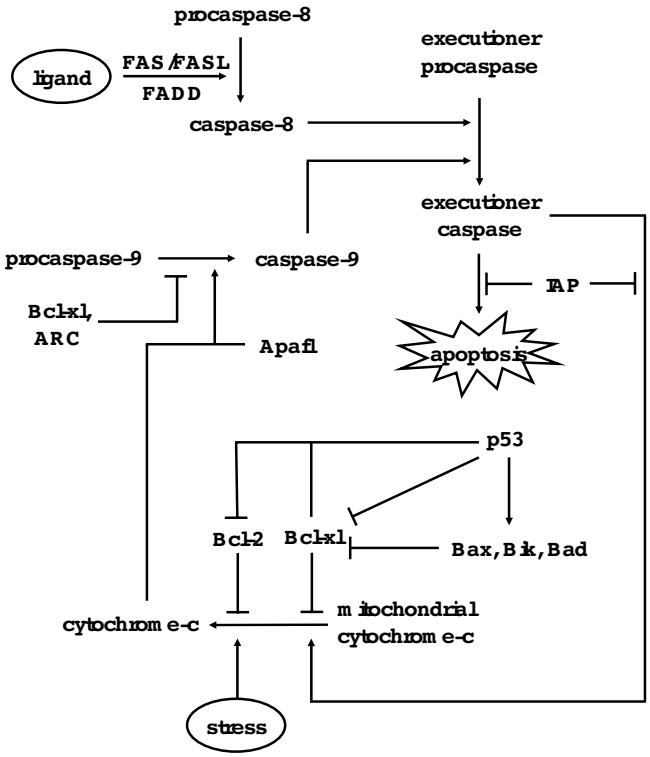

Figure 8
Caspase-dependent apoptosis mechanism. The model includes two triggers for the activation of cell suicide mechanism, extracellular death ligand and stress-related factor [7]. The cell death occurs when executioner caspase is activated by caspase-8 (ligand effector) or caspase-9 (stress-related effector).

Table 8: Nomenclature in caspase-dependent apoptosis model.

No.	Protein complex name	No.	Protein complex name
1	total receptor ligands	11	caspase-8
2	clustered FAS/FASL complex	12	caspase-9
3	FADD	13	executioner caspase
4	FAS/FASL-FADD complex	14	decoy protein
5	cytochrome c	15	decoy protein
6	Apaf-1	16	decoy protein
7	Apaf-1-cytochrome c complex	17	activator protein
8	procaspase-8	18	Bcl-2
9	procaspase-9	19	Bcl-x_L
10	executioner procaspase		

Table 9: Parameter values in caspase-dependent apoptosis model.

No.	Parameter	No.	Parameter	No.	Parameter	No.	Parameter
1	k_1	8	k_{83a}	15	K_H	22	K_K
2	k_a	9	k_{93a}	16	K_I	23	K_L
3	k_h	10	α_{CE}	17	K_J	24	K_N
4	k_{8za1}	11	k_u	18	K_C	25	K_O
5	k_{9za1}	12	K_S	19	K_D	26	K_P
6	k_{8za2}	13	K_A	20	K_F	27	K_R
7	k_{9za2}	14	K_B	21	K_G		

violated by the errors in the predictions of the FADD, FAS/FASL-FADD complex, and the cytochrome c. Tables 6 and 7 show the prediction errors for all the protein concentrations and reaction rates using the estimated parameters and also the "initial" parameters set. Again it is important to note that all these would not be available but are included here as a proof of concept.

2nd iteration

The second iteration is performed in two ways suggested in Figure 2. The first case involves using suboptimal experiment design with the optimal measurement set; whereas, the second case uses the optimal experiment with the suboptimal measurements. The model obtained after the first iteration is used to identify the optimal experiment that generate maximum information. For the selection of the optimal experiment design it is assumed that the measurement set remains unchanged. For the caspase system, the receptor and stress signals are parameterized such that both signals start at time 0 with constant rate injections reaching the maximum level in 30 minutes. The design variables are the final levels of the receptor and stress signals of the caspase activation. The search was constrained over a range of 0–0.4 for the receptor signal and 0–0.05 for the stress signal. A brute force search results in an optimal experiment (Table 1) with maximum information content for 18 identifiable parameters. The confidence interval for the identifiable parameters (Equation 8) and the deviations from the "nominal" parameter values (Equation 9) are shown in Table 3. Here it should be noted that the "nominal" parameters are the values obtained after the first iteration. The optimal levels for the signals suggest an optimal experiment with low receptor concentrations and high stress signal.

In each of the two cases, the model identification procedure is repeated in a similar manner as in the first iteration. The errors in the estimates of the protein concentrations and the reaction rates for both cases are shown in Tables 4 and 5 respectively. It is observed that the optimal measurement set improves the estimates of the states for which the information content increases. For example, the optimal measurement set includes caspase-8 (state 11), a state that is not included in the suboptimal measurements. The measurement of the caspase-8 improves the estimates of both the caspase-8 (state 11)

and the procaspase-8 (state 8). The estimates are used to refine the parameter estimates using Equation 19.

Figure 7 shows the maximum and average prediction errors for the measured concentrations for the "test" experiment for the two cases in the second iteration. With the optimal measurements, although the overall errors are reduced, the threshold values are still violated. However, the predictions with parameters obtained from the optimal experiment reduce all the errors below the threshold. This support the termination of the iterative process with an acceptable model. The model prediction for all the concentrations and reaction rates for the "test" experiment are shown in Figures 3 and 4 and the prediction errors in Tables 6 and 7.

Authors' contributions

KG performed the model identification work and drafted the manuscript. KG and RG performed the optimal experiment design and measurement selection work. FJD conceived of the study and participated in its design and co-ordination. All authors have read and approved the final manuscript.

Appendix

The model of the caspase activated apoptosis proposed by Varner and co-workers [7] consists of 19 states (protein concentrations) and 11 reaction rates. Figure 8 gives the schematic of the apoptosis mechanism.

The model Equations can be represented as:

$$
\begin{aligned}
\dot{x}_1 &= \Omega_1 - \mu x_1 \\
\dot{x}_2 &= r_1 - \mu x_2 \\
\dot{x}_3 &= \Omega_3 - 2r_2 - \mu x_3 \\
\dot{x}_4 &= r_2 - \mu x_3 \\
\dot{x}_5 &= r_{10} - r_3 - \mu x_5 \\
\dot{x}_6 &= \Omega_6 - \mu x_6 \\
\dot{x}_7 &= r_3 - \mu x_7 \\
\dot{x}_8 &= \Omega_8 - 2r_4 - 2r_6 - \mu x_8 \\
\dot{x}_9 &= \Omega_9 - 2r_5 - 2r_7 - \mu x_9 \\
\dot{x}_{10} &= \Omega_{10} - r_8 - r_9 - \mu x_{10} \\
\dot{x}_{11} &= 2r_4 + 2r_6 - \mu x_{11} \\
\dot{x}_{12} &= 2r_5 + 2r_7 - \mu x_{12} \\
\dot{x}_{13} &= r_8 + r_9 - r_{11} - \mu x_{13} \\
\dot{x}_k &= \Omega_k - \mu x_k \qquad k = 14,\ 15,\ldots,19
\end{aligned}
$$

where x_i denotes the ith protein concentration, r_j denotes the jth reaction rate, Ω_k denotes the rate of synthesis of the protein k and μ denotes the protein complex degradation rate. The reaction rates are as follows:

$$
\begin{aligned}
r_1 &= \frac{k_l(x_1 - x_2)L}{K_S^{-1} + L} \\
r_2 &= \frac{k_a x_3 x_2}{(1 + K_A x_3 + K_A K_B x_3^2} - \frac{x_4}{K_A K_B x_3} \\
r_3 &= \frac{k_h x_5 x_6}{1 + K_H x_5 + K_I \dfrac{x_{19}}{1 + K_J x_{17}}} - \frac{x_7}{K_H} \\
r_4 &= \frac{k_{8za1} x_8^2 x_4}{K_C^{-1}K_D^{-1} + K_D^{-1}x_8 + x_8^2 + K_F K_C^{-1}K_D^{-1}x_{15} + K_G K_D^{-1}x_8 x_{15}} \\
r_5 &= \frac{k_{9zal} x_9^2 x_7}{K_K^{-1}K_L^{-1} + K_L^{-1}x_9 + x_9^2 + K_N K_K^{-1}K_L^{-1}x_{16} + K_O K_L^{-1}x_9 x_{16}} \\
r_6 &= k_{8za2} x_8^2 \\
r_7 &= k_{9za2} x_9^2 \\
r_8 &= \frac{k_{83a} x_{10} x_{11}}{K_P^{-1} + K_R K_P^{-1} x_{14} + x_{10}} \\
r_9 &= \frac{k_{93a} x_{10} x_{12}}{K_P^{-1} + K_R K_P^{-1} x_{14} + x_{10}} \\
r_{10} &= \alpha_{CE}[\upsilon(x_{13}, x_{18}) + \upsilon(X, x_{18}) \\
r_{11} &= k_u x_{13} \frac{[IAPs]}{1 + K_U[IAPs]}
\end{aligned}
$$

where

L = free ligand concentration (receptor)

$$
\upsilon(x_{13}, x_{18}) = \begin{Bmatrix} 1 \ \forall \ \frac{x_{13}}{x_{18}} > 0.25 \\ 0 \ \forall \ \frac{x_{13}}{x_{18}} \le 0.25 \end{Bmatrix}
$$

$$
\upsilon(X, x_{18}) = \begin{Bmatrix} 1 \ \forall \ \frac{X}{x_{18}} > 0.025 \\ 0 \ \forall \ \frac{X}{x_{18}} \le 0.025 \end{Bmatrix}
$$

X = chemical/nutritional factor (stress)

$$
\frac{[IAPs]}{1 + K_U[IAPs]} = 0.1765
$$

Tables 8 and 9 represent the nomenclature of the protein complexes and parameter values in the apoptosis model, respectively [7].

Acknowledgements

This work is supported by National Science Foundation (BES-0000961), by the Institute for Collaborative Biotechnologies through grant DAAD19-03-D-0004 from the U.S. Army Research Office, and by the DARPA BioComp program.

References

1. Kitano H: **Computational Systems Biology.** *Nature* 2002, **420:**206-210.
2. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Network.** *Science* 2001, **292(5518):**929-934.
3. Zhu H, Huang S, Dhar P: **The Next Step in Systems Biology: Simulating the Temporospatial Dynamics of Molecular Networks.** *Bioessays* 2003, **26:**68-72.
4. 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* 2003, **79:**726-739.
5. Edwards JS, Ibarra RU, Palsson BO: ***In silico* Predictions of *Escherichia coli* Metabolic Capabilities are Consistent with Experimental Data.** *Nat Biotechnol* 2001, **19:**125-130.
6. Schöberl B, Jonsson CE, Gilles ED, Müller G: **Computational Modeling of the Dynamics of the MAP Kinase Cascade Activated by Surface and Internalized EGF Receptors.** *Nat Biotechnol* 2002, **20:**370-375.
7. Fussenegger M, Bailey JE, Varner J: **A Mathematical Model of Caspase Function in Apoptosis.** *Nat Biotechnol* 2000, **18:**768-774.
8. Chen KC, Csikasz-Nagy A, Gyorffy B, Val J, Novak B, Tyson JJ: **Kinetic Analysis of a Molecular Model of the Budding Yeast Cell Cycle.** *Mol Bio Cell* 2000, **11:**369-391.
9. Feng XJ, Rabitz H: **Optimal identification of biochemical reaction networks.** *Biophys J* 2004, **86(3):**1270-1281.
10. Moles CG, Mendes P, Banga JR: **Parameter Estimation in Biochemical Pathways: A Comparison of Global Optimization Methods.** *Genome Res* 2003, **13:**2467-2474.
11. Esposito WR, Floudas CA: **Global Optimization for the Parameter Estimation of Differential-Algebraic Systems.** *Ind Eng Chem Res* 2000, **39:**1291-1310.
12. Grossmann IE: *Global Optimization in Engineering Design* Dordrecht, Netherlands: Kluwer Academic Publishers; 1996.
13. Papamichail I, Adjiman CS: **A Rigorous Global Optimization Algorithm for Problems with Ordinary Differential Equations.** *J Global Optim* 2002, **24:**1-33.
14. Guss C, Boender E, Romeijn HE: *Handbook of Global Optimization* Dordrecht, Netherlands: Kluwer Academic Publishers 1995 chap. Stochastic methods:829-869.
15. Ali MM, Storey C, Törn A: **Application of Stochastic Global Optimization Algorithms to Practical Problems.** *J Optim Theory Appl* 1997, **95:**545-563.
16. Törn A, Ali MM, Viitanen S: **Stochastic Global Optimzation: Problem Classes and Solution Techniques.** *J Global Optim* 1999, **14:**437-447.
17. Audoly S, Bellu G, D'Angio L, Saccomani MP, Cobelli C: **Global Identifiability of Nonlinear Models of Biological Systems.** *IEEE Trans Biomed Eng* 2001, **48:**55-65.
18. Jacquez JA, Perry T: **Parameter estimation: local identifiability of parameters.** *Amer J Physiol* 1990, **258:**E727-E736.
19. Yao KZ, Shaw BM, Kou B, McAuley KB, Bacon DW: **Modeling Ethylene/Butene Copolymerization with Multi-Site Catalysts: Parameter Estimability and Experimental Design.** *Polymer Reaction Eng* 2003, **11:**563-588.
20. Landaw EM, DiStefano III JJ: **Multiexponential, Multicompartmental, and Noncompartmental Modeling. II. Data Analysis and Statistical Considerations.** *Amer J Physiol* 1984, **246:**R665-R677.
21. Petersen B, Gernaey K, Vanrolleghem PA: **Practical Identifiability of Model Parameters by Combined Respirometric-Titrimetric Measurements.** *Water Science Tech* 2001, **43:**347-355.
22. Zak DE, Gonye GE, Schwaber J, Doyle III FJ: **Importance of Input Perturbations and Stochastic Gene Expression in the Reverse Engineering of Genetic Regulatory Networks: Insights from an Identifiability Analysis of an in Silico Network.** *Genome Res* 2003, **13:**2396-2405.
23. Asprey SP, Macchietto S: **Statistical Tools for Optimal Dynamic Model Building.** *Comput Chem Eng* 2000, **24:**1261-1267.
24. Kremling A, Fischer S, Gadkar K, Doyle III FJ, Sauter T, Bullinger E, Allgower F, Gilles ED: **A Benchmark for Methods in Reverse Engineering and Model Discrimination: Problem Formulation and Solutions.** *Genome Res* 2004, **14:**1773-1785.
25. Banga JR, Versyck KJ, Impe JFV: **Computation of Optimal Identification Experiments for Nonlinear Dynamic Process Models: A Stochastic Global Optimization Approach.** *Ind Eng Chem Res* 2002, **41:**2425-2430.
26. Mahadevan R, Edwards JS, Doyle III FJ: **Dynamic Flux Balance Analysis of Diauxic Growth in.** *E coli Biophys J* 2002, **83:**1331-1340.
27. Ljung L: *System Identification: Theory for the User* Englewood Cliffs, NJ: Prentice Hall; 1999.
28. Ljung L, Guo L: **The Role of Model Validation for Assessing the Size of the Unmodeled Dynamics.** *IEEE Trans Automat Contr* 1997, **42:**1230-1239.
29. Poolla K, Khargonekar P, Tikku A, Krause J, Nagpal K: **A Time-Domain Approach to Model Validation.** *IEEE Trans Automat Contr* 1994, **39:**951-959.
30. Brogan WL: *Modern Control Theory* Upper Saddle River, NJ: Prentice Hall; 1991.
31. Vajda S, Rabitz H, Walter E, Lecourtier Y: **Qualitative and Quantitative Indentifiability Analysis of Non-Linear Chemical Kinetic Models.** *Chem Eng Commun* 1989, **83:**191-219.
32. Emery AF, Nenarokomov AV: **Optimal Experimental Design.** *Meas Sci Technol* 1998, **9:**864-876.
33. Beck JV, Arnold KJ: *Parameter Estimation in Engineering and Science* Toronto, Canada: John Wiley & Sons; 1977.
34. Gadkar KG, Gunawan R, Doyle III FJ: **Heuristic Methods for Measurement Selection for Identification of Biological Systems.** *Int Conf Systems Biology, Heidelberg, Germany* 2004.
35. Bastin G, Dochain D: *On-Line Estimation and Adaptive Control of Bioreactors* Amsterdam: Elsevier; 1990.
36. Albiol J, Robusté J, Casas C, Poch M: **Biomass Estimation in Plant Cell Cultures Using Extended Kalman Filter.** *Biotechnol Prog* 1993, **9:**174-178.
37. Gee DA, Ramirez WF: **On-Line State Estimation and Parameter Identification for Batch Fermentation.** *Biotechnol Prog* 1996, **12:**132-140.
38. Stephanopoulos G, San KY: **Studies on on-Line Biorector Identification. I. Theory.** *Biotechnol Bioeng* 1984, **26:**1176-1188.
39. Glassey J, Ignova M, Ward AC, Montague GA, Morris AJ: **Bioprocess Supervision: Neural Networks and Knowledge Based Systems.** *J Biotechnol* 1997, **52:**201-205.
40. Karim MN, Rivera SL: **Comparision of Feed-Forward and Recurrent Neural Networks for Bioprocess State Estimation.** *Comp Chem Eng* 1992, **16(Suppl):**S369-S377.
41. Simutis R, Lübbert A: **Exploratory Analysis of Bioprocesses Using Artificial Neural Network-Based Methods.** *Biotechnol Prog* 1997, **13:**479-487.
42. Gunawan R, Jung MYL, Seebauer EG, Braatz RD: **Maximum a Posteriori Estimation of Transient Enhanced Diffusion Energetics.** *AIChE J* 2003, **49:**2114-2123.
43. Cover TM, Thomas JA: *Elements of Information Theory* John Wiley & Sons; 1991.
44. Gadkar KG, Varner J, Doyle III FJ: **Model Identification of Signal Transduction Networks from Data Using a State Regulator Problem.** *IEE Systems Biology* 2005, **2:**.
45. Wagner A: **How to reconstruct a large genetic network from *n* gene perturbations in fewer than n^2 easy steps.** *Bioinformatics* 2001, **17(12):**1183-1197.

46. Gardner TS, Bernardo DD, Lorenz D, Collins JJ: **Inferring genetic networks and identifying compound model of action via expression profiling.** *Science* 2003, **301:**102-105.
47. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells.** *Nat Genet* 2005, **37(4**382-390 [http://dx.doi.org/10.1038/ng1532].
48. Chen BH, Asprey SP: **On the Design of Optimally Informative Deynamic Experiments for Model Discrimination in Multiresponse Nonlinear Situations.** *Ind Eng Chem Res* 2003, **42:**1379-1390.
49. Frey AD, Kallio PT, Gadkar KG, Varner J: **Dynamic Flux Balance Analysis of VHb Expression in *Escherichia coli* MG1655.** *Biotechnol Bioeng* 2005. accepted
50. Taylor S, Gunawan R, Doyle III FJ: **BioSens 2.0: Sensitivity Analysis Toolkit.** 2005 [http://www.chemengr.ucsb.edu/~ceweb/faculty/doyle/biosens/BioSens.htm].