ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

**Cite this article:** Ehsan Elahi F, Hasan A. 2018 A method for estimating Hill function-based dynamic models of gene regulatory networks. *R. Soc. open sci.* **5**: 171226.
http://dx.doi.org/10.1098/rsos.171226

Received: 28 August 2017
Accepted: 25 January 2018

**Subject Category:**
Genetics

**Subject Areas:**
systems biology/computational biology

**Keywords:**
gene regulatory networks, parameter estimation, optimization, *Escherichia coli*

**Author for correspondence:**
Ammar Hasan
e-mail: ammar.hasan@seecs.edu.pk

THE ROYAL SOCIETY PUBLISHING

# A method for estimating Hill function-based dynamic models of gene regulatory networks

Faizan Ehsan Elahi and Ammar Hasan

National University of Sciences and Technology (NUST), H-12, 44000, Islamabad, Pakistan

AH, 0000-0003-2755-8410

Gene regulatory networks (GRNs) are quite large and complex. To better understand and analyse GRNs, mathematical models are being employed. Different types of models, such as logical, continuous and stochastic models, can be used to describe GRNs. In this paper, we present a new approach to identify continuous models, because they are more suitable for large number of genes and quantitative analysis. One of the most promising techniques for identifying continuous models of GRNs is based on Hill functions and the generalized profiling method (GPM). The advantage of this approach is low computational cost and insensitivity to initial conditions. In the GPM, a constrained nonlinear optimization problem has to be solved that is usually underdetermined. In this paper, we propose a new optimization approach in which we reformulate the optimization problem such that constraints are embedded implicitly in the cost function. Moreover, we propose to split the unknown parameter in two sets based on the structure of Hill functions. These two sets are estimated separately to resolve the issue of the underdetermined problem. As a case study, we apply the proposed technique on the SOS response in *Escherichia coli* and compare the results with the existing literature.

## 1. Introduction

Gene expression is a process that transcribes information of genes and translates it into functional gene products. These products are proteins that play a central role in performing numerous cellular functions. Spatio-temporal expression of these products is controlled by close interaction of different genes with each other, which is commonly known as gene regulatory network (GRN). The large and complex nature of GRNs limits our ability to understand them intuitively [1,2]. Therefore, mathematical models are constructed to get better insight into and understanding of these complex phenomena.

Mathematical models that describe GRNs are of three major types: logical models, stochastic models and continuous models [1]. Logical models [3], e.g. Petri nets, Boolean and Bayesian networks, can only describe qualitative behaviour of GRNs. Stochastic and continuous models, represented by ordinary differential equations (ODEs), can describe the dynamic behaviour of GRNs quantitatively [1,2]. GRNs like any natural system involve randomness and this randomness becomes significant especially when the numbers of molecules are small [4,5]. Single-molecule level models [6] and stochastic differential equations [7] are applied to describe quantitative behaviour with randomness. Stochastic models are complex, computationally expensive and suitable only for a small number of molecules [8]. ODEs, on the other hand, are relatively simple, well studied and can easily simulate the quantitative behaviour of GRNs. These attributes make ODEs preferable whenever randomness is negligible. ODEs are mainly of two types, i.e. linear and nonlinear. Analytical solution of linear ODEs is found easily, but they can be used only in establishing the qualitative behaviour of GRNs [1,9]. GRNs, like most of the physical phenomena, are nonlinear. Therefore, nonlinear ODEs are used in modelling and quantitative analysis of GRNs. Moreover, nonlinear ODEs can easily simulate regulation and feedback effects of genes.

The mathematical model has unknown parameters whose value has to be obtained through experimental data of the biological system. The dynamics of a biological process are captured using time-series experiments that sample the process for measuring the concentration of gene products at different times. These dynamic time-series data can be used for parameter estimation of the ODE model [10]. Data for gene expressions are usually sparse [1,10], i.e. measurements are taken at only a few time points because of limitations of experimental set-up and high expense associated with these experiments. Sparseness of the data forms an underdetermined problem, i.e. the number of equations generated from the experimental data are small compared with the number of unknown parameters of ODEs. Besides sparseness, these measurements suffer from noise that may cause inaccuracies in parameter estimation. Methods for estimating ODE parameters from time-series data can be classified into three broader categories: classical, discretization and collocation methods. Classical techniques are based on first-order Taylor series expansion. First-order expansion for replacing nonlinear structures is only suitable for small durations and mild nonlinearities [11]. Discretization methods, e.g. [12], employ direct numerical solution of ODEs for data fitting. Therefore, ODE parameters are estimated along with initial conditions, which results in an increased number of unknowns. As discretization methods use numerical solutions they are computationally intensive. Moreover, an inherent disadvantage of these methods is that they can result in inaccurate parameter estimation because of high sensitivity with initial conditions [13,14].

Collocation methods [11,15] use basis systems such as polynomials, splines and Fourier basis, etc. Instead of fitting the experimental data directly to the numerical solution of ODEs, data are fitted to a function of basis systems. This function includes coefficients that are determined first and then parameters of ODEs are estimated through these coefficients. Thus, these methods avoid computationally expensive numerical solutions. Another important advantage of these methods is the estimation of initial conditions as a side product of the procedure. On the other hand, solution of nonlinear ODEs is very sensitive to initial conditions and can lead to inaccurate estimates of parameters. The most common variant of collocation methods is the generalized profiling method (GPM) [11]. The GPM employs cascaded optimization. The outer optimization is for parameters of ODEs. The inner optimization is for coefficients of the function of the basis system. The GPM produces better estimates than some of the other variants of collocation methods [15,16]. Detailed discussion of the GPM is given in §2.2

As stated earlier, the optimization problem for GRNs is usually underdetermined because of sparse data [10] and large number of unknowns involved in modelling with ODEs [2]. Moreover, it is a constrained optimization problem due to different reasons such as protein concentration cannot be negative. Nonlinear optimization in collocation methods can be solved by different algorithms. Global optimization algorithms, e.g. particle swarm [17], are very attractive for their convergence to a global solution. However, these algorithms use sampling approach, i.e. they need a large number of function evaluations at each optimization iteration. Function evaluation is computationally expensive in collocation methods because of the cascaded optimization structure. Therefore, algorithms based on a sampling approach are not suitable. Moreover, global optimization algorithms are suitable for a small number of unknowns. Researchers are focusing to improve these two aspects of global optimization algorithms, (e.g. [18,19]). The trust-region-reflective algorithm [20] and the Levenberg–Marquardt algorithm [21] are two other popular techniques of nonlinear optimization. A limitation of the trust-region-reflective algorithm is its inability to deal with underdetermined problems, which is usually the case in GRNs. Levenberg–Marquardt cannot deal with constrained optimization [21].

In this paper, we propose a new optimization approach to solve an underdetermined optimization problem with constraints. This approach is based on Hill functions. Hill functions are considered suitable for building GRN models with ODEs [1,2]. They can quantify activation and inhibition effects of genes. Hill functions are mainly composed of two types of parameters: threshold and cooperativity. Based on the structure of Hill functions, we propose separation of parameters of ODEs into two sets, i.e. threshold and cooperativity parameters. Splitting unknowns into two sets helps in avoiding underdeterminedness. These two sets of parameters are estimated separately with a suitable solver. The two-step estimation is iterated until change in parameters becomes insignificant. Details are given in §3. We take the SOS response in *Escherichia coli* as a case study for the proposed technique and compare the results with work reported in the literature. The case study is discussed in §4.

# 2. Background

The complete process of parameter estimation of a Hill function-based ODE model from experimental genome data is shown as a block diagram in figure 1. Firstly, gene network structure is obtained either from well-established literature on the subject, experimental observations or by applying reverse engineering techniques (e.g. BANJO [22], ARACNE [23] and TSNI [9]). Network structure is used to construct a Hill function-based ODE model of the GRN with unknown parameters. The parameters of the ODE model are estimated by using the experimental data. Our work is focused on parameter estimation based on the generalized profiling method (GPM) [11]. In the GPM, an optimization problem has to be solved to calculate the most optimal parameters. Lastly, the identified mathematical model can be used for dynamic simulations of the GRN. In the following subsections, we provide details of the Hill function-based modelling and the GPM.

## 2.1. Hill function-based ordinary differential equation model of gene regulatory networks

ODEs belong to the category of continuous mathematical models. Concentrations of gene products are considered as the state variables. The rate of change of these concentrations is expressed as a function of state variables. The standard form is as follows:

$$\dot{x}_i = f_i(x_1, x_2, \ldots, x_N), \quad i = 1, \ldots, N, \tag{2.1}$$

where $x_i$ denotes the concentration of the product of gene $i$ and $N$ is the total number of genes in a GRN. The rate of change of a state $\dot{x}_i$ is described as some mathematical function $f_i(.)$ of all the states. The concentration of any gene product $x_i$ is non-negative. Hence the model should be such that $x_i \geq 0$ for $i = 1, \ldots, N$ for all practically possible initial conditions [24].

For modelling of GRNs with ODEs, Hill or sigmoidal functions are employed in the literature [25,26]. Hill curves are preferred because they can adopt sigmoidal shape [2,27] with suitable parameters. Mendes [27] has suggested the following function:

$$\dot{x}_i = P_i \prod_{j \in \mathbb{I}_i} h^-(x_j, Q_{i,j}, R_{i,j}) \prod_{k \in \mathbb{A}_i} h^+(x_k, Q_{i,k}, R_{i,k}) - S_i x_i, \quad i = 1, \ldots, N, \tag{2.2}$$

where

$$h^-(x_j, Q_{i,j}, R_{i,j}) := \frac{{Q_{i,j}}^{R_{i,j}}}{({x_j}^{R_{i,j}} + {Q_{i,j}}^{R_{i,j}})} \tag{2.3}$$

and

$$h^+(x_k, Q_{i,k}, R_{i,k}) := 1 + \frac{{x_k}^{R_{i,k}}}{({x_k}^{R_{i,k}} + {Q_{i,k}}^{R_{i,k}})}, \tag{2.4}$$

where $x_i$ denotes concentration of the $i$th gene product, $\mathbb{I}_i$ is a subset of $\{1, \ldots, N\}$ that denotes the set of all inhibiting gene products for the $i$th gene, $\mathbb{A}_i$ denotes the set of all activating gene products for the $i$th gene, $P_i$ is the synthesis rate constant, $S_i$ is the degradation rate constant, $h^-$ is the inhibiting Hill function, $h^+$ is the activating Hill function, $Q_{i,j}$ and $Q_{i,k}$ denote the threshold parameters of Hill functions, and the exponents $R_{i,j}$ and $R_{i,k}$ denote the cooperative parameters. Threshold parameter is the value of the concentration after which significant effect of inhibitor or activator is observed. Cooperative parameter controls how sharply the level transition occurs across the threshold. Activator and inhibitor Hill functions are depicted graphically for different values of the cooperative parameter in figure 2.

If the GRN composed of $N$ genes has $M$ interconnections, then the complete Mendes model of the form (2.2) requires $2 * (M + N)$ parameters to fully describe the GRN, which includes $N$ synthesis rate

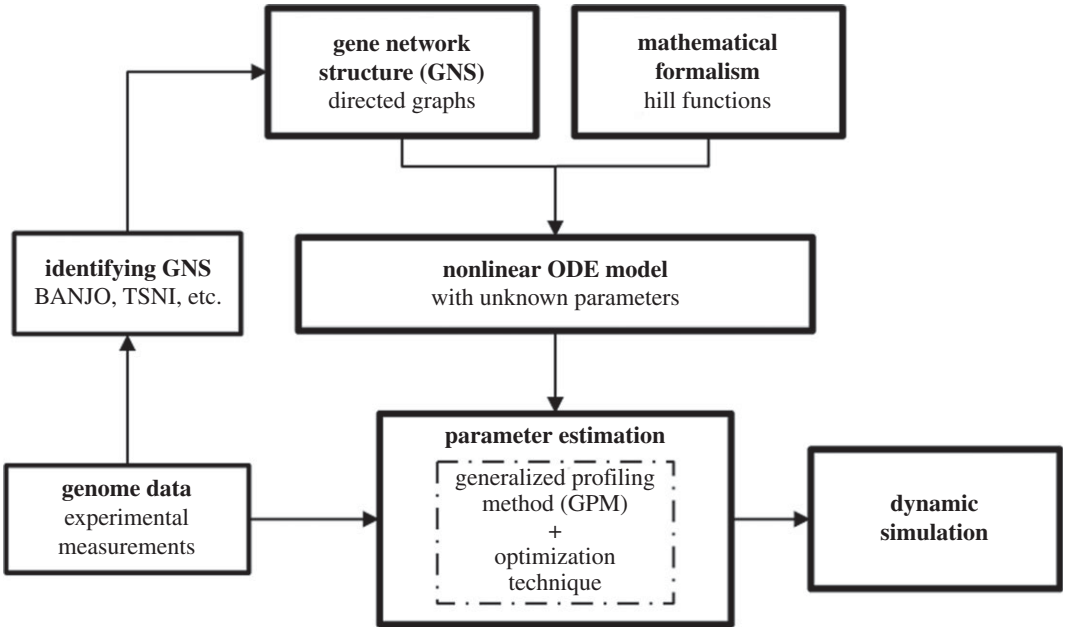


**Figure 1.** Block diagram describing the estimation of GRNs.

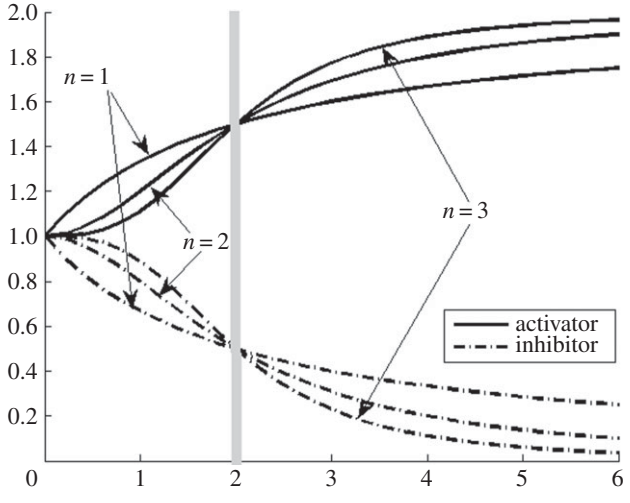


**Figure 2.** Activator and inhibitor functions are shown for fixed-threshold parameter $Q_{i,j} = Q_{i,k} = 2$, which is marked by a thick vertical line, while three different curves correspond to three different cooperative parameters $R_{i,j} = R_{i,k} = 1, 2, 3$. Increase in the cooperative parameter causes more rapid change across the threshold.

constants, $N$ degradation rate constants, $M$ threshold parameters and $M$ cooperative parameters . We can write the model (2.2) in compact form as shown below:

$$\dot{x}(t) = f_i(X(t), \theta) \quad i = 1 \ldots N, \tag{2.5}$$

where $X = [x_1 \ldots x_N]^{\mathrm{T}}$ denotes state variables and $\Theta = [P_1 \ldots P_N \;\; Q_{1,2} \ldots Q_{N,N-1} \;\; R_{1,2} \ldots \; R_{N,N-1} \; S_1 \ldots S_N]^{\mathrm{T}}$ is the vector of all the parameters. The parameters $\Theta$ have to be estimated from the experimental data.

It may be noted that all the parameters $\Theta$ should be positive. In a real GRN it is not possible to have negative synthesis rate constants $P_i$ or degradation rate constants $S_i$. Similarly, the concentration thresholds $Q_{i,j}$ or the exponents $R_{i,j}$ of the Hill functions cannot be negative. The necessity of positive parameters should be incorporated in the estimation problem.

## 2.2. Generalized profiling method for estimation of parameters

Estimation is a challenging task because of non-availability of an analytical solution of (2.2) and numerical methods are expensive. Collocation methods avoid these difficulties by using polynomial regression. One of the efficient collocation methods is the generalized profiling method (GPM), which provides accurate estimation with low computational load [11]. In the GPM, β-splines are used as polynomial regression of states and their derivatives. The β-splines are preferred for interpolation as they are differentiable. Further details on β-splines can be found in [28]. The states and their derivatives are defined in terms of β-splines as follows:

$$x_i(t) = c_i^{\mathrm{T}}\phi(t) \quad i = 1 \ldots N$$

and

$$\dot{x}_i(t) = c_i^{\mathrm{T}}\dot{\phi}(t) \quad i = 1 \ldots N,$$

where $c$ is column vector of coefficients, while $\phi(t)$ denotes β-spline basis systems. Estimation is achieved in two cascaded optimization steps. The inner optimization determines the coefficient vector $c := [c_1^{\mathrm{T}} \ldots c_N^{\mathrm{T}}]^{\mathrm{T}}$ for a given fixed value of parameters $\theta$ such that the following objective function is minimized:

$$\hat{c}(\theta) = \arg\min_{c} \left\{ \sum_{t \in \mathbb{T}_{\mathrm{E}}} \sum_{i=1}^{N} (y_i(t) - c_i^{\mathrm{T}}\phi(t))^2 + \sum_{i=1}^{N} \lambda_i \int_{t_0}^{t_f} (c_i^{\mathrm{T}}\dot{\phi}(t) - f_i(c, \phi(t), \theta))^2 \, \mathrm{d}t \right\}, \tag{2.6}$$

where $y_i(t)$ is the experimentally observed concentration of the product of the $i$th gene, $t_0$ is the starting time of the experiment, $t_f$ is the ending time of the experiment, $\mathbb{T}_{\mathrm{E}}$ is the set of all times at which experimental measurements are available and $\lambda_i$ is the weighting parameter. The first term of (2.6) minimizes the sum of squared residuals between the data $y_i(t)$ and the state of the model $x_i(t) = c_i^{\mathrm{T}}\phi(t)$, whereas the second term penalizes the residuals between derivatives from nonlinear functions $f_i(.)$ in (2.5) and their estimates from β-splines. Weighting factor $\lambda_i$ is used as a smoothening parameter.

The outer optimization determines the estimate of ODEs parameters $\hat{\Theta}$ by minimizing the following objective function:

$$\hat{\Theta} = \arg\min_{\theta} \left\{ \sum_{t \in \mathbb{T}_{\mathrm{E}}} \sum_{i=1}^{N} (y_i(t) - \hat{c}_i^{\mathrm{T}}(\Theta)\phi(t))^2 \right\}, \tag{2.7}$$

where $\hat{c}$ is the optimal value of coefficient vector $c$ such that the objective function in (2.6) is minimized. The above-stated objective function is an implicit function of ODE parameters $\Theta$. Further details on the GPM can be found in [29].

As mentioned in §2.1, the parameters $\Theta$ should be positive. Therefore, the above optimization problems are to be solved with the following constraint:

$$\Theta \geq 0. \tag{2.8}$$

# 3. Proposed method

The nonlinear optimizations in the generalized profiling method in §2.2 have to be solved by numerical optimization algorithms. Having the constraint (2.8) in the optimization problem makes it more complex to solve. Some solvers, such as Lavenberg–Marquardt [21], also do not handle constraints. We propose a reformulation of the mathematical model that allows us to avoid the constraints. Let $P_i = e^{p_i}$, $Q_{i,j} = e^{q_{i,j}}$, $R_{i,j} = e^{r_{i,j}}$ and $S_i = e^{s_i}$. The model (2.2) can be rewritten as follows:

$$\dot{x}_i = e^{p_i} \prod_{j \in \mathbb{I}_i} h^{-}(x_j, e^{q_{i,j}}, e^{r_{i,j}}) \prod_{k \in \mathbb{A}_i} h^{+}(x_k, e^{q_{i,k}}, e^{r_{i,k}}) - e^{s_i} x_i, \quad i = 1, \ldots, N. \tag{3.1}$$

In the above model, the unknown parameters are $\theta = [p_1 \ldots p_N \; q_{1,2} \ldots q_{N,N-1} \; r_{1,2} \ldots r_{N,N-1} \; s_1 \ldots s_N]^{\mathrm{T}}$. The benefit of this change of variables is that we do not have to place any constraints on $\theta$. Whether the elements of $\theta$ are negative or positive, the elements of $\Theta$ will always be positive because of the exponential function.

As mentioned earlier, the number of time samples in experimental data are usually fewer than the number of unknown parameters, i.e. $2(M + N)$, in the ODE model (2.2) or (3.1). Therefore, the nonlinear

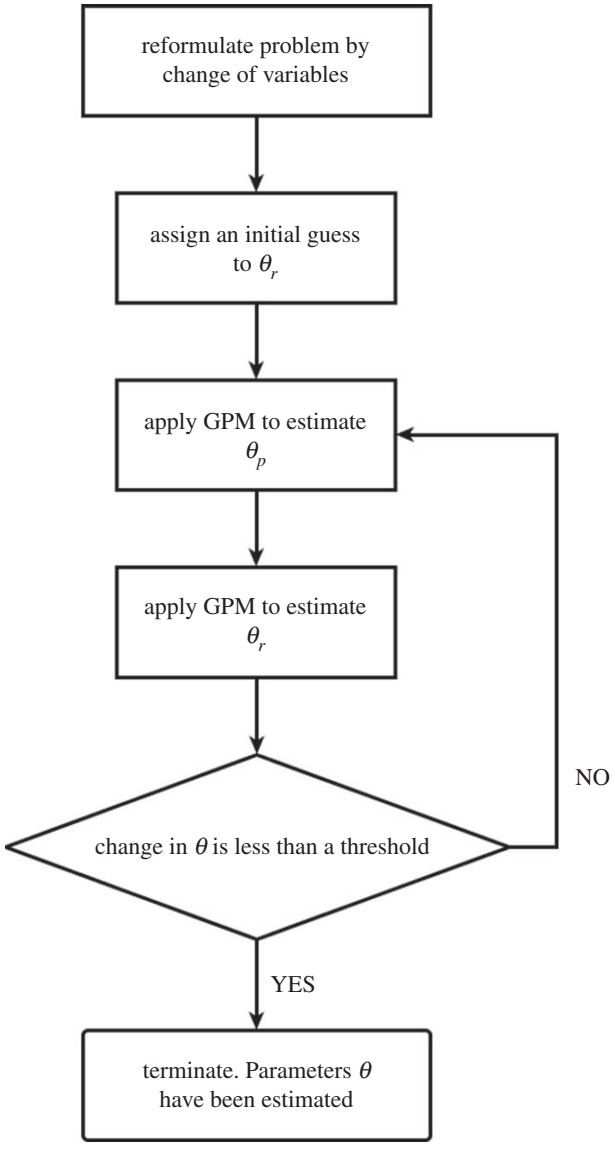


**Figure 3.** Flow chart for the proposed scheme.

optimizations in the generalized profiling method in §2.2 are underdetermined. This could result in poor estimates of the parameters. Some of the optimization techniques, such as the trust-region-reflective method [20], are more prone to inaccuracy in case of underdeterminedness.

To solve the issue of underdeterminedness, we propose a new approach. The idea is based on the structure of Hill functions. The cooperative parameters $R_{i,j} = e^{r_{i,j}}$ in the Hill functions control the sharpness of the switching action around the threshold parameters. Therefore, cooperative parameters only have prominent impact near the switching of Hill functions. We propose to split the parameters $\theta$ to be estimated into two sets $\theta_r = [r_{1,2} \dots r_{N,N-1}]^{\mathrm{T}}$, which includes only the cooperative parameters $r_{i,j}$, and $\theta_p = [p_1 \dots p_N \; q_{1,2} \dots q_{N.N-1} \; s_1 \dots s_N]^{\mathrm{T}}$, which includes all the rest of the unknown parameters. We propose to initially assume a value of $\theta_r$, e.g. 2 for all $r_{i,j}$, and estimate only the parameters $\theta_p$. As the number of unknown parameters are reduced, it helps to improve the issue of underdeterminedness. However, as $\theta_r$ was initially guessed, we then fix the value of $\theta_p$ and solve the optimization problem to estimate $\theta_r$. We repeat this process until the estimates of all the parameters converge to a value.

The complete proposed scheme is illustrated with a flow chart in figure 3. Initially the estimation problem is reformulated with the change of variables to avoid constraints. An initial value of $\theta_r$ is assumed. The GPM method is applied to only estimate the parameters $\theta_p$, which has $M + 2N$ elements. Using the recent optimal values of $\theta_p$, the GPM method is applied to only estimate the parameters

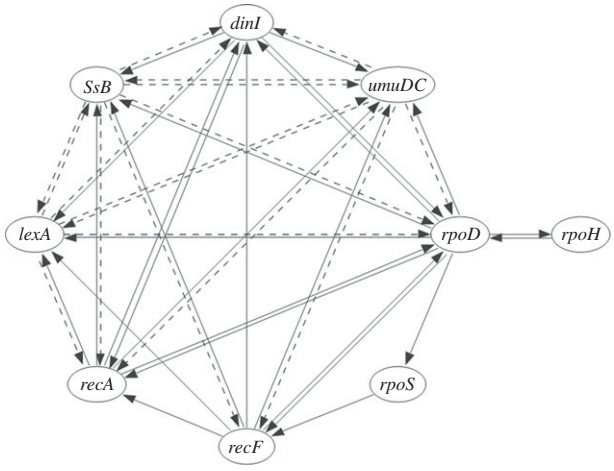


**Figure 4.** Network interconnection of nine genes SOS response in *Escherichia coli*. Solid lines denote activating interactions, dotted lines denote inhibiting interactions and arrow heads show the direction of effect. Reproduced with permission from [9,30].

$\theta_r$, which has $M$ elements. If the change in all the parameters $\theta$ is less than some threshold, then the estimation process is complete; otherwise repeat the process.

# 4. Case study: SOS response in *Escherichia coli*

SOS response in *Escherichia coli* is an inducible DNA repair system that enables bacteria to survive under severe DNA damage. In the SOS response, nearly 40 genes are directly regulated by *recA* and *lexA*, while tens of other genes are indirectly controlled [30]. Nine genes named *lexA*, *recA*, *recF*, *rpoD*, *rpoS*, *dinI*, *umuDC*, *rpoH* and *ssB* are considered to be directly participating in the SOS response. Established network connections among these nine genes are known in the literature [9,30,31]. Network structure of these nine genes is shown as in figure 4.

## 4.1. Hill function-based model of the SOS response

The ODE model of the SOS response in *Escherichia coli* is constructed based on the network structure given in figure 4 [9,30]. The network has nine genes and 43 interconnections, i.e. $N = 9$ and $M = 43$. The Hill function-based ODEs for the nine gene product concentrations are given below:

$$\begin{aligned}
\dot{lexA} &= P_1 * h^+(recA, Q_1, R_1) * h^+(recF, Q_2, R_2) * h^+(rpoD, Q_3, R_3) \\
&\quad * h^+(dinI, Q_4, R_4) * h^-(ssB, Q_5, R_5) * h^-(umuDC, Q_6, R_6) - S_1 * lexA, \\
\dot{recA} &= P_2 * h^+(recF, Q_7, R_7) * h^+(rpoD, Q_8, R_8) * h^+(dinI, Q_9, R_9) \\
&\quad * h^-(lexA, Q_{10}, R_{10}) * h^-(ssB, Q_{11}, R_{11}) * h^-(umuDC, Q_{12}, R_{12}) - S_2 * recA, \\
\dot{recF} &= P_3 * h^+(rpoD, Q_{13}, R_{13}) * h^+(rpoS, Q_{14}, R_{14}) * h^-(ssB, Q_{15}, R_{15}) \\
&\quad * h^-(umuDC, Q_{16}, R_{16}) - S_3 * recF, \\
\dot{rpoS} &= P_4 * h^+(rpoD, Q_{17}, R_{17}) - S_4 * rpoS, \\
\dot{rpoD} &= P_5 * h^+(recF, Q_{18}, R_{18}) * h^+(recA, Q_{19}, R_{19}) * h^+(dinI, Q_{20}, R_{20}) \\
&\quad * h^+(rpoH, Q_{21}, R_{21}) * h^-(ssB, Q_{22}, R_{22}) * h^-(lexA, Q_{23}, R_{23}) \\
&\quad * h^-(umuDC, Q_{24}, R_{24}) - S_5 * rpoD, \\
\dot{umuDC} &= P_6 * h^+(rpoD, Q_{25}, R_{25}) * h^+(recF, Q_{26}, R_{26}) * h^+(recA, Q_{27}, R_{27}) \\
&\quad * h^-(lexA, Q_{28}, R_{28}) * h^-(dinI, Q_{29}, R_{29}) * h^-(ssB, Q_{30}, R_{30}) - S_6 * umuDC, \\
\dot{dinI} &= P_7 * h^+(rpoD, Q_{31}, R_{31}) * h^+(recF, Q_{32}, R_{32}) * h^+(recA, Q_{33}, R_{33}) \\
&\quad * h^-(lexA, Q_{34}, R_{34}) * h^-(ssB, Q_{35}, R_{35}) * h^-(umuDC, Q_{36}, R_{36}) - S_7 * dinI,
\end{aligned}$$

**Table 1.** SOS response model parameters estimated by the proposed scheme.

| parameter | estimated value | parameter | estimated value |
|---|---|---|---|
| $P_1$ | 0.82645 | $S_1$ | 0.56596 |
| $P_2$ | 3.0312 | $S_2$ | 0.06717 |
| $P_3$ | 26.844 | $S_3$ | 1.151 |
| $P_4$ | 0.22028 | $S_4$ | 0.03181 |
| $P_5$ | 22.807 | $S_5$ | 0.97227 |
| $P_6$ | 30.438 | $S_6$ | 0.037697 |
| $P_7$ | 5.7443 | $S_7$ | 0.024872 |
| $P_8$ | 9.6769 | $S_8$ | 0.60515 |
| $P_9$ | 1809.6 | $S_9$ | 258.52 |
| $Q_1$ | 1.6273 | $R_1$ | 1.6354 |
| $Q_2$ | 0.58209 | $R_2$ | 1.6833 |
| $Q_3$ | 0.65218 | $R_3$ | 1.8982 |
| $Q_4$ | 14.988 | $R_4$ | 2.3757 |
| $Q_5$ | 17.705 | $R_5$ | 2.2303 |
| $Q_6$ | 272.67 | $R_6$ | 1.5392 |
| $Q_7$ | 11.981 | $R_7$ | 2.674 |
| $Q_8$ | 56.946 | $R_8$ | 3.5452 |
| $Q_9$ | 4.9391 | $R_9$ | 1.4626 |
| $Q_{10}$ | 34.691 | $R_{10}$ | 1.3153 |
| $Q_{11}$ | 4.779 | $R_{11}$ | 1.8499 |
| $Q_{12}$ | 18.819 | $R_{12}$ | 3.0102 |
| $Q_{13}$ | 61.511 | $R_{13}$ | 4.9999 |
| $Q_{14}$ | 1.3297 | $R_{14}$ | 1.1567 |
| $Q_{15}$ | 8.4207 | $R_{15}$ | 1.8104 |
| $Q_{16}$ | 11.72 | $R_{16}$ | 2.117 |
| $Q_{17}$ | 14.013 | $R_{17}$ | 2.049 |
| $Q_{18}$ | 11.867 | $R_{18}$ | 1.8588 |
| $Q_{19}$ | 9.5381 | $R_{19}$ | 2.2855 |
| $Q_{20}$ | 1.7108 | $R_{20}$ | 3.3169 |
| $Q_{21}$ | 4.6155 | $R_{21}$ | 2.3944 |
| $Q_{22}$ | 23.779 | $R_{22}$ | 2.9639 |
| $Q_{23}$ | 39.778 | $R_{23}$ | 1.7128 |
| $Q_{24}$ | 3.0976 | $R_{24}$ | 2.0704 |
| $Q_{25}$ | 1.5401 | $R_{25}$ | 0.60901 |
| $Q_{26}$ | 0.9762 | $R_{26}$ | 1.3191 |
| $Q_{27}$ | 13.822 | $R_{27}$ | 4.0316 |
| $Q_{28}$ | $1.3686 * 10^6$ | $R_{28}$ | 1.8207 |
| $Q_{29}$ | 0.50903 | $R_{29}$ | 1.9367 |
| $Q_{30}$ | 150.6 | $R_{30}$ | 1.3081 |
| $Q_{31}$ | 9.3893 | $R_{31}$ | 1.9267 |
| $Q_{32}$ | 134.06 | $R_{32}$ | 1.9118 |

(*Continued*.)

**Table 1.** (*Continued*.)

| parameter | estimated value | parameter | estimated value |
|---|---|---|---|
| $Q_{33}$ | 15.778 | $R_{33}$ | 2.3397 |
| $Q_{34}$ | 2.4818 | $R_{34}$ | 1.9367 |
| $Q_{35}$ | 8.9587 | $R_{35}$ | 1.7976 |
| $Q_{36}$ | 40.718 | $R_{36}$ | 1.3055 |
| $Q_{37}$ | 14.545 | $R_{37}$ | 2.6053 |
| $Q_{38}$ | 376.12 | $R_{38}$ | 1.1819 |
| $Q_{39}$ | 351.21 | $R_{39}$ | 1.8328 |
| $Q_{40}$ | 27.191 | $R_{40}$ | 1.9916 |
| $Q_{41}$ | 133.91 | $R_{41}$ | 2.9068 |
| $Q_{42}$ | 9.7889 | $R_{42}$ | 2.0699 |
| $Q_{43}$ | 14.225 | $R_{43}$ | 2.041 |

$$\begin{aligned}\dot{ssB} &= P_8 * h^+(dinI, Q_{37}, R_{37}) * h^+(rpoD, Q_{38}, R_{38}) \\ &\quad * h^+(recF, Q_{39}, R_{39}) * h^+(recA, Q_{40}, R_{40}) \\ &\quad * h^-(umuDC, Q_{41}, R_{41}) * h^-(lexA, Q_{42}, R_{42}) - S_8 * ssB, \\ \dot{rpoH} &= P_9 * h^+(rpoD, Q_{43}, R_{43}) - S_9 * rpoH.\end{aligned}$$

A total of $2 * (43 + 9) = 104$ parameters are needed to fully describe the above model. We can write the above model in compact form as shown below:

$$\dot{x}_i = f_i(X(t), \theta) \quad i = 1 \ldots 9, \tag{4.1}$$

where $X := [x_1 \ldots x_9]^{\mathrm{T}} = [lexA\ recA\ recF\ rpoS\ rpoD\ umuD\ dinI\ ssB\ rpoH]^{\mathrm{T}}$ denotes state variables, and $\Theta = [P_1 \ldots P_9\ Q_1 \ldots Q_{43}\ R_1 \ldots R_{43}\ S_1 \ldots S_9]^{\mathrm{T}}$ is the set of unknown parameters. The value of $\Theta$ has to be estimated from the experimental genome data.

## 4.2. Results of the proposed method

In this section, the results of estimation by the proposed method are presented and compared with the results reported in the literature. SOS is a well-studied response, and a lot of literature and experimental datasets are available. Many Microbe Microarrays Database M3D [32] provides datasets for the SOS response under different perturbations, like ultraviolet light or some antibiotic. M3D database provides datasets in two formats, i.e. time course and steady-state measurements. Both forms of data are employed in this work. Time course measurements have been used for estimation of the model parameters. Experimental measurements of steady-state values have been used for validation.

Parameters are estimated with an initial guess of $r_{i,j} = 2$. For the GPM, the weighting factor $\lambda_i$ is chosen as 150 for $i = 1, \ldots, N$. The basis system is chosen as β-splines of order 4 with 100 knots, which is the number of joints in the spline function. The choice of $\lambda_i$, order and knots is subjective and chosen based on the required smoothness. The algorithm converged in four iterations of the loop in figure 3 with a total execution time of approximately 82 minutes on a PC with a Core i5-650 processor. The Matlab code is available at [33]. Estimated parameters are given in table 1.

The estimated model is simulated to generate time course evolution of gene product concentrations. The results are compared with experimental data and the model reported in [31]. Baralla *et al.* [31] also investigated the same subpart of the SOS response using the same dataset of M3D. They have applied the particle swarm algorithm for optimization and a direct numerical solution of the ODE model for data fitting. It can be seen in figure 5 that dynamic simulation with the model estimated by the proposed method is much more accurate than that of [31], especially for *recF*, *rpoD*, *ssB* and *rpoH*.

For the purpose of validation, the model estimated by the proposed method has also been used to calculate the steady-state values of concentrations. The steady-state values are given in table 2 along with the experimentally observed values and the values obtained by the estimated model in [31]. The table also gives the sum of squared error between the experimental data and values obtained from estimated models. It can be seen that the total error in the values obtained by the model estimated by the proposed method is much smaller than that of [31].

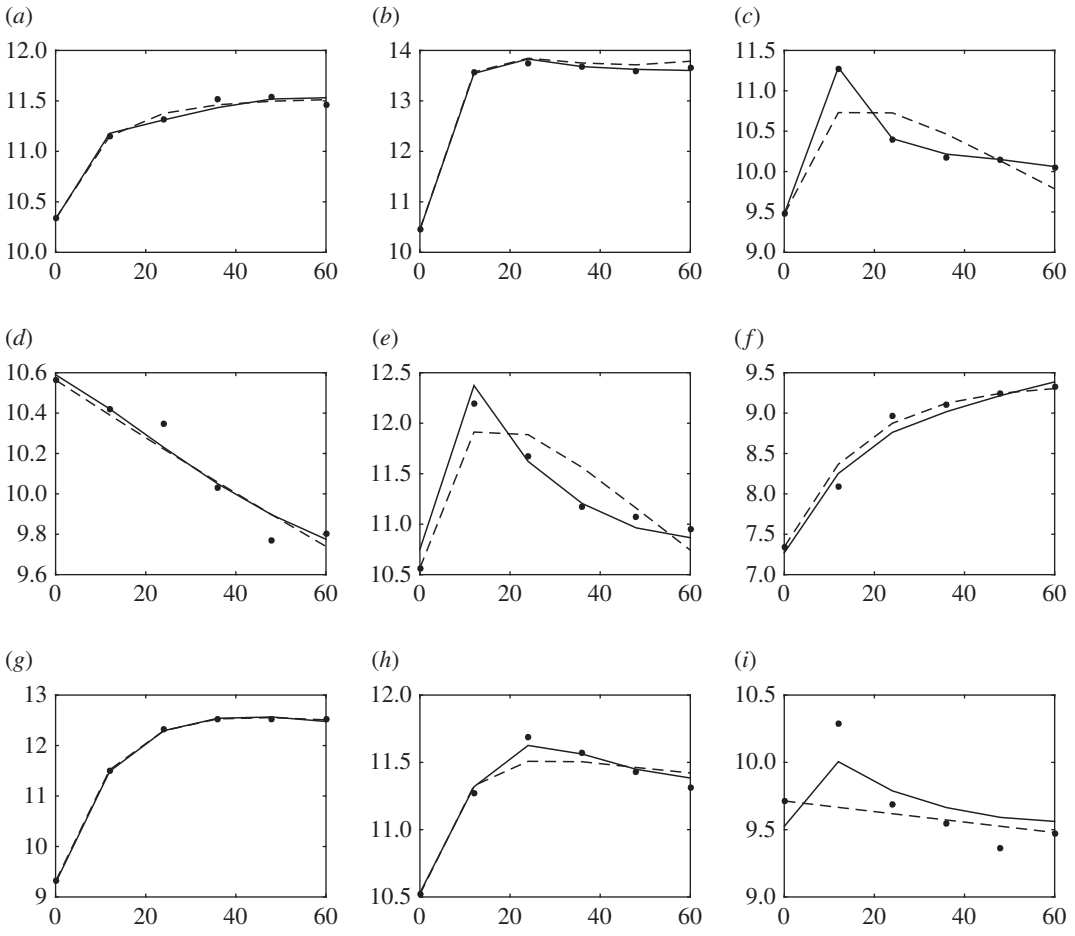


**Figure 5.** Comparison of experimental and simulated data concentration of gene products: experimental (dots), simulation of the model estimated by the proposed method (solid lines) and simulation of the model estimated in [31] (dashed lines). All the *x*-axes are in time (s) and all the *y*-axes are in concentration (mM). (*a*) *lexA*, (*b*) *recA*, (*c*) *recF*, (*d*) *rpoS*, (*e*) *rpoD*, (*f*) *umuD*, (*g*) *dinI*, (*h*) *ssB* and (*i*) *rpoH*.

**Table 2.** Comparison of steady-state values obtained by the model estimated by the proposed method and the model estimated in [31].

| gene | experimental (mM) | values obtained by the proposed method (mM) | values obtained by [31] (mM) |
|---|---|---|---|
| *lexA* | 11.471 | 11.192 | 11.4599 |
| *recA* | 11.795 | 13.326 | 16.492 |
| *recF* | 8.975 | 9.8327 | 4.87727 |
| *rpoS* | 10.383 | 9.6427 | 3.46706 |
| *rpoD* | 9.4618 | 11.322 | 5.20705 |
| *umuD* | 7.8192 | 9.6932 | 9.35587 |
| *dinI* | 9.9169 | 11.542 | 12.1365 |
| *ssB* | 10.213 | 11.334 | 11.2432 |
| *rpoH* | 8.5983 | 9.6991 | 1.12474 |
| sum of squared errors | — | 15.787 | 169.03 |

# 5. Conclusion

Obtaining a mathematical model of GRNs requires estimation of model parameters from the experimental data. To estimate the optimal values of parameters an optimization problem has to be

solved. Usually the experimental data have much fewer measurements than the number of parameters which could result in an underdetermined optimization problem. Moreover, depending on how the optimization problem is posed, constraints have to be incorporated to find practically feasible values of parameters, which could result in numerical issues. In this paper, we have proposed a new approach to reformulate the estimation problem such that constraints are not required. Based on the structure of Hill functions, we also suggest to split the number of unknowns into two sets, which are estimated one at a time. This helps to alleviate the issue of underdeterminedness.

The proposed technique is applied to estimate the model of the SOS response in *Escherichia coli*. The results are compared with the existing literature and experimental data. Both dynamic and steady-state results show that the proposed technique provides more accurate estimates of the unknown model parameters.

Data accessibility. The Matlab code of the proposed algorithm is available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.ht047 [33].

Authors' contributions. A.H. conceived the research idea. F.E.E. did the coding and simulations. Both co-authors analysed and interpreted the results. Both co-authors drafted the manuscript and gave their final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. We received no funding for this study.

# References


1. Karlebach G, Shamir R. 2008 Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* **9**, 770–780. (doi:10.1038/nrm2503)
2. De Jong H. 2002 Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* **9**, 67–103. (doi:10.1089/10665270252833208)
3. Friedman N, Linial M, Nachman I, Pe'er D. 2000 Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601–620. (doi:10.1089/106652700750050961)
4. Ross IL, Browne CM, Hume DA. 1994 Transcription of individual genes in eukaryotic cells occurs randomly and infrequently. *Immunol. Cell Biol.* **72**, 177–185. (doi:10.1038/icb.1994.26)
5. Bae K, Lee C, Hardin PE, Edery I. 2000 dCLOCK is present in limiting amounts and likely mediates daily interactions between the dCLOCK–CYC transcription factor and the PER–TIM complex. *J. Neurosci.* **20**, 1746–1753.
6. McAdams HH, Arkin A. 1997 Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA* **94**, 814–819. (doi:10.1073/pnas.94.3.814)
7. Arkin A, Ross J, McAdams HH. 1998 Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells. *Genetics* **149**, 1633–1648.
8. Vijesh N, Chakrabarti SK, Sreekumar J. 2013 Modeling of gene regulatory networks: a review. *J. Biomed. Sci. Eng.* **6**, 223–231. (doi:10.4236/jbise.2013.62A027)
9. Bansal M, Della Gatta G, Di Bernardo D. 2006 Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics* **22**, 815–822. (doi:10.1093/bioinformatics/btl003)
10. Bar-Joseph Z, Gitter A, Simon I. 2012 Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **13**, 552–564. (doi:10.1038/nrg3244)
11. Ramsay JO, Hooker G, Campbell D, Cao J. 2007 Parameter estimation for differential equations: a generalized smoothing approach. *J. R. Stat. Soc.: Ser. B* (*Stat. Methodol.*) **69**, 741–796. (doi:10.1111/j.1467-9868.2007.00610.x)
12. Biegler L, Damiano J, Blau G. 1986 Nonlinear parameter estimation: a case study comparison. *AIChE J.* **32**, 29–45. (doi:10.1002/aic.690320105)
13. Marlin TE. 2000 *Process control*. New York, NY: McGraw-Hill.
14. Xu J. 2010 Robust estimation for differential equations, time series analysis on climate change and MCMC simulation of duration-of-load problem. MS thesis, Simon Fraser University, Burnaby, Canada.
15. Varah J. 1982 A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Stat. Comput.* **3**, 28–46. (doi:10.1137/0903003)
16. Cao J, Qi X, Zhao H. 2012 Modeling gene regulation networks using ordinary differential equations. In *Next generation microarray bioinformatics: methods and protocols* (eds J Wang, AC Tan, T Tian), pp. 185–197. New York, NY: Humana Press.
17. Kennedy J. 2011 Particle swarm optimization. In *Encyclopedia of machine learning* (eds C Sammut, GI Webb), pp. 760–766. Berlin, Germany: Springer.
18. Fernández-Martínez JL, Mukerji T, García-Gonzalo E. 2010 Particle Swarm Optimization in high dimensional spaces. In *Int. Conf. on Swarm Intelligence, Beijing, China, 12–15 June*, pp. 496–503. Berlin, Germany: Springer.
19. Parno MD, Fowler KR, Hemker T. 2009 *Framework for particle swarm optimization with surrogate functions*. Darmstadt, Germany: Darmstadt Technical University.
20. Coleman TF, Li Y. 1996 An interior trust region approach for nonlinear minimization subject to bounds. *SIAM. J. Optim.* **6**, 418–445. (doi:10.1137/0806023)
21. Moré JJ. 1978 The Levenberg-Marquardt algorithm: implementation and theory. In *Numerical analysis* (ed. GA Watson), pp. 105–116. Berlin, Germany: Springer.
22. Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED. 2004 Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* **20**, 3594–3603. (doi:10.1093/bioinformatics/bth448)
23. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. 2005 Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* **37**, 382–390. (doi:10.1038/ng1532)
24. Tyson JJ, Othmer HG. 1978 The dynamics of feedback control circuits in biochemical pathways. *Prog. Theor. Biol* **5**, 1–62. (doi:10.1016/B978-0-12-543105-7.50008-7)
25. Ptashne M, Switch AG. 1992 *Phage lambda and higher organisms*. Cambridge, MA: Cell & Blackwell Scientific.
26. Yagil G, Yagil E. 1971 On the relation between effector concentration and the rate of induced enzyme synthesis. *Biophys. J.* **11**, 11–27. (doi:10.1016/S0006-3495(71)86192-1)
27. Mendes P, Sha W, Ye K. 2003 Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics* **19**, ii122–ii129. (doi:10.1093/bioinformatics/btg1016)
28. Gallier JH. 2000 *Curves and surfaces in geometric modeling: theory and algorithms*. Los Altos, CA: Morgan Kaufmann.
29. Cao J, Ramsay JO. 2007 Parameter cascades and profiling in functional data analysis. *Comput. Stat.* **22**, 335–351. (doi:10.1007/s00180-007-0044-1)
30. Gardner TS, Di Bernardo D, Lorenz D, Collins JJ. 2003 Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105. (doi:10.1126/science.1081900)
31. Baralla A, Cavaliere M, de la Fuente A. 2008 Modeling and parameter estimation of the SOS response network in *E. coli*. MS thesis, University of Trento, Trento, Italy.
32. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS. 2008 Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic. Acids Res.* **36**, D866–D870. (doi:10.1093/nar/gkm815)
33. Ehsan Elahi F, Hasan A. 2018 Data from: A method for estimating Hill function-based dynamic models of gene regulatory networks. Dryad Digital Repository. (doi:10.5061/dryad.ht047)