Data-driven dynamical modelling of a pathogen-infected plant gene regulatory network: a comparative analysis

Mathias Foo[a,*], Leander Dony[b,c,d], Fei He[e]

[a]*School of Engineering, University of Warwick, CV4 7AL, Coventry UK*

[b]*Institute of Computational Biology, Helmholtz Munich, 85764 Neuherberg, Germany*

[c]*Department of Translational Psychiatry, Max Planck Institute of Psychiatry, and International Max Planck Research School for Translational Psychiatry (IMPRS-TP), 80804 Munich, Germany*

[d]*TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany*

[e]*Centre for Computational Science and Mathematical Modelling, Coventry University, CV1 2JH, Coventry, UK*

*Corresponding author

Email address: M.Foo@warwick.ac.uk (Mathias Foo), leander.dony@helmholtz-munich.de (Leander Dony), Fei.He@coventry.ac.uk (Fei He)

Abstract

Recent advances in synthetic biology have enabled the design of genetic feedback control circuits that could be implemented to build resilient plants against pathogen attacks. To facilitate the proper design of these genetic feedback control circuits, an accurate model that is able to capture the vital dynamical behaviour of the pathogen-infected plant is required. In this study, using a data-driven modelling approach, we develop and compare four dynamical models (i.e. linear, Michaelis-Menten with Hill coefficient (Hill Function), standard S-System and extended S-System) of a pathogen-infected plant gene regulatory network (GRN). These models are then assessed across several criteria, i.e. ease of identifying the type of gene regulation, the predictive capability, Akaike Information Criterion (AIC) and the robustness to parameter uncertainty to determine its viability of balancing between biological complexity and accuracy when modelling the pathogen-infected plant GRN. Using our defined ranking score, we obtain the following insights to the modelling of GRN. Our analyses show that despite commonly used and provide biological relevance, the Hill Function model ranks the lowest while the extended S-System model ranks highest in the overall comparison. Interestingly, the performance of the linear model is more consistent throughout the comparison, making it the preferred model for this pathogen-infected plant GRN when considering data-driven modelling approach.

Keywords: Data-Driven Modelling, Gene Regulatory Network, Linear Model, Hill Function Model, S-System Model, Synthetic Biology

1. Introduction

One of the common fungal pathogens that infects plant is the *Botrytis cinerea*. When infection occurs, the interactions between the pathogen and the host plant often lead to the host plant succumb to diseases. This is because pathogen often disrupts the host defense mechanism through secretion of a range of proteins, small RNAs and metabolites to aid their colonisation Williamson et al., 2007; Jamir et al., 2007; Weiberg et al., 2013; Jones and Dangl, 2006). Advances in the area of molecular biology have provided plant synthetic biologists means of improving plant resilience through the use of synthetic feedback control circuits (see *e.g.,* (Aoki et al., 2019)) to restore the regulation that is affected by the pathogen attack (Foo et al., 2018a). Pathogen affected genes involved in defence tend to have their expression levels compromised, leading to their reduced functional ability (Ng et al., 2018; Sood et al., 2021). The synthetic feedback control circuits would sense the changes in the expression level of pathogen affected genes, where the genes *cis*-regulatory elements are modified resulting in changes in their regulations and expression levels (see (Gherman, 2018) and references therein) and regulate appropriate transcription factor to allow the compromised expression levels to be controlled thereby enabling plant to recover their defence functionality.

To facilitate the design of these synthetic feedback control circuits, an accurate dynamical model depicting the gene regulatory network (GRN) involved in the plant defense mechanism is required. In our previous study (Foo et al., 2018a), equipped with the temporal data of gene expressions (Windram et al., 2012) and the knowledge of the interacting genes involved in plant defence (Gherman, 2018), a linear dynamical model is developed using a data-driven modelling approach to model the pathogen-infected plant GRN with good accuracy and subsequently used to design and develop a framework of engineering resilience plant using synthetic genetic feedback control circuits. The reason the linear model is considered in (Foo et al., 2018a) is due to the design of the proposed synthetic genetic controller carried out in the frequency domain using tools from linear control theory.

In this study, as a follow up to (Foo et al., 2018a), we aim to answer the following question: "*When using the data-driven modelling approach, given the temporal data and knowledge about the pathogen-infected plant GRN interaction, is the linear model the most viable model to facilitate the design of synthetic feedback control and if not what is the alternate candidate model*?"

Since the advancement in the area of Systems Biology, GRN modelling has been extensively studied (see the review paper by (Schlitt and Brazma, 2009) and references therein). According to (Schlitt and Brazma, 2009), most of the models described in those studies can be categorised into four main classes in the order of increased complexity --- part list model (*e.g.,* description of the GRN component), topology model (*e.g.,* directed graph model), control logic model (*e.g.,* Boolean function model) and dynamical model (*e.g.,* differential equation model). The models developed here are often based on first principles, *i.e.,* using the biological understanding of the interacting components.

With the access to high throughput data at molecular level becoming available, attention turns to another branch of modelling approach called reverse engineering (Tegner et al., 2003; Hache et al., 2009), where models are developed in the attempt to fit those data using various methods such as correlation-based method, Bayesian networks, regression analysis, information theoretical approaches, Gaussian graphical models, dynamic differential equations, etc (Penfold and Wild, 2011; Chai et al., 2014; Villaverde et al., 2013; Vinciotti et al., 2016). In a reverse engineering approach, usually there is no assumption about the model structure and the interacting components.

The development in this area often parallels the development of GRN network inference algorithms, where the types and directions of regulation between components are inferred directly from data (see the review paper (Emmert-Streib et al., 2012) and references therein). As a note, in the area of systems and control engineering (He et al., 2016), the reverse engineering approach is also known as system identification or data-driven modelling.

Here, we would like to make several remarks to provide readers the main scope of this study. First, this study is not about comparing network inference algorithms, hence the discussion on this topic is beyond the scope of this study. Interested readers can see the following review papers (Long et al., 2008; Den Broeck et al., 2020) for more details. Second, unlike typical reverse engineering (data-driven modelling)

approaches that assume almost no prior knowledge about the GRNs and the model structures, here we have some knowledge about the interacting genes and we have a set of candidate model structures of interest to be compared. Thus, our `network inference' approach will be simpler with the focus on identifying the regulation type. Third, our study is system specific, *i.e.*, a pathogen-infected plant GRN, and the main goal is to answer the key question posted above, *i.e.*, the suitability of the linear dynamical model (Foo et al., 2018a) in modelling a pathogen-infected plant GRN.

To the best of our knowledge, the only comparative study of different dynamical models for plant-specific GRN has been carried out in (Wang et al., 2014), where the authors compared several dynamical models for the GRN involved in plant flowering time. Different from that study, our study focuses on the data-driven modelling approach and uses different quantitative metrics for model comparison. In our comparative analysis, in addition to the linear model given in (Foo et al., 2018a), we consider the Michaelis-Menten model with Hill coefficient and two S-System based models. The choice of these three models are motivated by their capabilities in modelling GRN (see *e.g.*, Karlebach and Shamir, 2008; Chowdhury and Chetty, 2016; Foo et al., 2020)). From here onward, we will use the notation *Hill Function* model for Michaelis-Menten with Hill coefficient model (see *e.g.*, Santillan, 2008)).

The manuscript is organised in the following manner. In Section 2, we present the pathogen-infected plant GRN used as our case study. The main results on the comparative analysis of the four GRN models are presented and discussed in detail in Section 3. In Section 4, we analysed the applicability of our approach to other GRN. Finally, the discussion and conclusions are provided in Section 5.

2. System description

The plant GRN involved in the defence against pathogen attack and used in this study is adapted from (Windram et al., 2012), where a subnetwork of nine genes – hereinafter termed 9GRN (Foo et al., 2018a) – has been identified to be involved in the defence against *Botrytis cinerea*, as shown in Fig. 1. In Fig. 1, while the direction of regulation between genes in 9GRN is known, the type of regulation (*i.e.*, activation or inhibition) is not entirely known. Among these nine genes, seven of them are directly affected by the pathogen, as indicated by the yellow hexagon. *CHE* and *ATML1* are part of the circadian clock genes as their oscillatory profiles are influenced by external

light, as indicated by the red lightning. Moreover, the gene *CHE* has been identified to be an important gene in the plant defence mechanism and when it is affected by the pathogen, its expression level would decrease thereby reducing its defence capability (Windram et al., 2012; Gherman, 2018). Therefore, it is imperative that the expression level of *CHE* being kept high and thus the role of the synthetic feedback control circuitry is to ensure its expression level stay high when under pathogen attack.

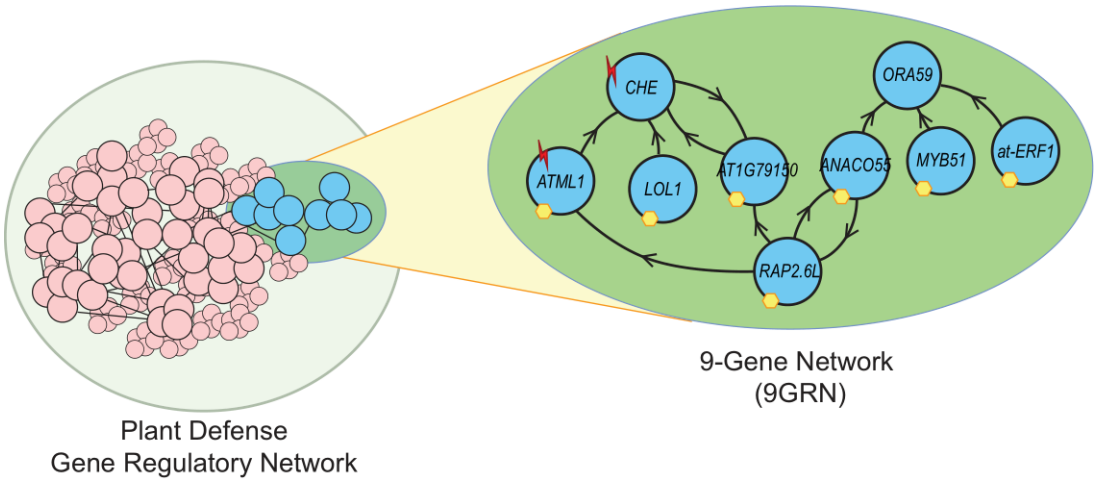

Fig. 1. Plant (Arabidopsis) gene regulatory network (termed 9GRN) involved in the defence response to *Botrytis cinerea* adapted from (Foo et al., 2018a). The yellow hexagon symbol represents genes that have been identified to be directly affected by *Botrytis cinerea*. Red lightning symbol represents genes that are light regulated. The directional arrows indicate the influence of one gene to another despite its regulation type unknown.

3. Comparative analysis of the 9GRN models

3.1. Comparison criteria

In this comparative study, the four dynamical models of 9GRN will be evaluated across the following criteria.

- Criterion I: Ease of identifying regulation type.
- Criterion II: Predictive capability.
- Criterion III: Quality of data fit using Akaike weights based on Akaike Information Criterion (AIC).
- Criterion IV: Robustness to parameter uncertainties.

These four criteria are chosen following typical model evaluation techniques (see *e.g.*, (Turchin, 2003) and references therein) that considers metric such as model

prediction error, model quality amidst complexity and model robustness to uncertainties. For more details see (Turchin, 2003).

3.2. Model structures for 9GRN

The general structure for all these four models are given as follows:

Linear model: This linear model is the one used in (Foo et al., 2018a).

$$\frac{dX_i}{dt} = \sum_{j=1}^{n_i^P} \alpha_{i,j} X_j - \beta_i X_i + B_{S,i} + c_i W + \gamma_i L_I \tag{1}$$

where X_i is the expression level of ith gene, n_i^P is the number of genes involved in regulating X_i, α is the production rate, β is the degradation rate, B_S is the gene basal level while c and γ parameterised the external input from pathogen W and light L_I, respectively. For more details on how each of the terms in (1) are derived, see (Foo et al., 2018a).

Standard S-System model: The standard S-System model developed from the field of biochemical system theory was initially proposed in (Savageau, 1969) to model metabolic pathways. Over the course of its development (see *e.g.*, (Savageau, 2001; Voit et al., 2015) and references therein), this model has been used to model GRN with good accuracy (Maki et al., 2000) and it has the following form.

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{i,j}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{i,j}} \tag{2}$$

where α is the production rate, β is the degradation rate, n and m are respectively the total number of dependent and independent variables and $g_{i,j}$ and $h_{i,j}$ are exponents associated with the production and degradation processes, respectively. Note that the standard S-System model structure does not have provision to account for gene basal level and the external input, and these variables are incorporated directly as part of the independent variables.

Extended S-System model: This model was proposed in (Foo et al., 2020) to individually account for the effect of gene basal expression and external input, instead

of being part of the independent variables, and was shown to accurately describe the plant circadian system compared to the standard S-System model. The extended S-System model has the following form.

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n_i^P} X_k^{g_{i,j}} - \sum_{j=1}^{n_i^D} \beta_{i,j} X_i \left(\prod_{k=1}^{n} X_k^{h_{i,j,k}} \right) + \sum_{j=1}^{n_i^E} \gamma_{i,j} U_{i,j} \tag{3}$$

where α, β and γ are the reaction rate constants associated with production, degradation, and external input regulation (*e.g.*, light, perturbation, basal level, etc), respectively. Like the standard S-System, $g_{i,j}$ represents the exponent related to production while $h_{i,j}$ represents the exponent related to degradation. n_i^P, n_i^D and n_i^E are the number of genetic components involved in the respective production, degradation, and external input regulation of X_i. $U_{i,j}$ encapsulates the effect of those aforementioned external regulations on X_i. As a note, the Hill Function model is commonly used to model GRN instead of the S-System model. Nevertheless, the analysis in (Foo et al., 2020) has shown the efficacy of the S-System model structure in modelling GRN, thus warranting modellers with two alternate model candidates that can be considered when attempting to model GRN.

Hill Function model: Conventionally, this model has been widely used to model GRN (see *e.g.*, (Bolouri and Davidson, 2002; Rue and Garcia-Ojalvo, 2013) and references therein) and it has the following form.

$$\frac{dX_i}{dt} = \sum_{j=1}^{n_i^P} \alpha_{i,j} \left[f_A\left(X_j, W, L_I\right) + f_R\left(X_j, W, L_I\right) \right] - \beta_i X_i + B_{S,i} \tag{4}$$

where like before, α and β are the production and degradation rate, respectively, B_S is the basal level, f_A and f_R are respectively, the activator and inhibitor type of regulation. Both have different forms, and they are usually modelled as $f_A = X^q/(K^q + X^q)$ and $f_R = 1/(K^q + X^q)$, where K is the Michaelis-Menten kinetic constant and q is the Hill coefficient. Note that here, the regulations are modelled as a summation of

successive regulations, and they could also be modelled as a product of successive regulations.

One immediate observation from these four model structures is that the Hill Function model structure requires the knowledge of the regulation type when deriving the ordinary differential equations (ODE) for each gene, thus making this model not suitable for reverse engineering (Youseph et al., 2015). If the regulation type is unknown, extra steps (discussed in Section 3.3) are required to construct the best fitting Hill Function model structure. Since the Hill Function model requires extra steps in identifying the regulation types, it can often incur additional computational load.

On the other hand, for the linear and the two S-System based models, the sign of the production rate $\alpha_{i,j}$ (for linear model) and the exponent associated with the production rate $g_{i,j}$ (for S-System based model) estimated from data can directly inform us the type of regulation for each gene, where a positive value denotes activation while a negative value denotes inhibition. For the S-System models, there are also approaches being developed that can be used to estimate those parameters in an efficient and fast manner (Wang et al., 2010). Moreover, for the linear and extended S-System models, the positive or negative regulation of the external inputs can also be inferred through the sign of the estimated parameters (*i.e.*, c's and γ's, respectively).

3.3. Detailed ordinary differential equations (ODEs) model of 9GRN

We use subscripts L, SS, ES and HF in the model parameters to represent the linear, standard S-System, extended S-System and Hill Function models, respectively. In order to avoid overloading of variables, the following numbers are used to denote the genes in 9GRN. 1: *ORA59*, 2: *MYB51*, 3: *LOL1*, 4: *AT1G79150*, 5: *ANAC055*, 6: *a-ERF-1*, 7: *ATML1*, 8: *CHE* and 9: *RAP2.6L*.

Linear model: The corresponding ODEs following (1) are given as follow, which is the same linear model used in (Foo et al., 2018a),

$$\frac{dG_1}{dt} = \alpha_{L,1,1}G_2 + \alpha_{L,1,2}G_5 + \alpha_{L,1,3}G_6 - \beta_{L,1}G_1 + B_{L,1}$$

$$\frac{dG_2}{dt} = -\beta_{L,2}G_2 + B_{L,2} + c_{L,2}W$$

$$\frac{dG_3}{dt} = -\beta_{L,3}G_3 + B_{L,3} + c_{L,3}W$$
$$\frac{dG_4}{dt} = \alpha_{L,4,1}G_8 + \alpha_{L,4,2}G_9 - \beta_{L,4}G_4 + B_{L,4} + c_{L,4}W$$
$$\frac{dG_5}{dt} = \alpha_{L,5,1}G_9 - \beta_{L,5}G_5 + B_{L,5} + c_{L,5}W$$
$$\frac{dG_6}{dt} = -\beta_{L,6}G_6 + B_{L,6} + c_{L,6}W$$
$$\frac{dG_7}{dt} = \alpha_{L,7,1}G_9 - \beta_{L,7}G_7 + B_{L,7} + c_{L,7}W + \gamma_{L,7}L_I$$
$$\frac{dG_8}{dt} = \alpha_{L,8,1}G_3 + \alpha_{L,8,2}G_4 + \alpha_{L,8,3}G_7 - \beta_{L,8}G_8 + B_{L,8} + \gamma_{L,8}L_I$$
$$\frac{dG_9}{dt} = \alpha_{L,9,1}G_5 - \beta_{L,9}G_9 + B_{L,9} + c_{L,9}W \quad (5)$$

Standard S-System model: Following (2), the corresponding ODEs for 9GRN are given as follow.

$$\frac{dG_1}{dt} = \alpha_{SS,1}G_2^{g_{SS,1,1}}G_5^{g_{SS,1,2}}G_6^{g_{SS,1,3}} - \beta_{SS,1}G_1$$
$$\frac{dG_2}{dt} = \alpha_{SS,2}W^{g_{SS,2,1}} - \beta_{SS,2}G_2$$
$$\frac{dG_3}{dt} = \alpha_{SS,3}W^{g_{SS,3,1}} - \beta_{SS,3}G_3$$
$$\frac{dG_4}{dt} = \alpha_{SS,4}G_8^{g_{SS,4,1}}G_9^{g_{SS,4,2}}W^{g_{SS,4,3}} - \beta_{SS,4}G_4$$
$$\frac{dG_5}{dt} = \alpha_{SS,5}G_9^{g_{SS,5,1}}W^{g_{SS,5,2}} - \beta_{SS,5}G_5$$
$$\frac{dG_6}{dt} = \alpha_{SS,6}W^{g_{SS,6,1}} - \beta_{SS,6}G_6$$
$$\frac{dG_7}{dt} = \alpha_{SS,7}G_9^{g_{SS,7,1}}W^{g_{SS,7,2}}L_I^{g_{SS,7,3}} - \beta_{SS,7}G_7$$
$$\frac{dG_8}{dt} = \alpha_{SS,8}G_3^{g_{SS,8,1}}G_4^{g_{SS,8,2}}G_7^{g_{SS,8,3}}L_I^{g_{SS,8,4}} - \beta_{SS,8}G_8$$
$$\frac{dG_9}{dt} = \alpha_{SS,9}G_5^{g_{SS,9,1}}W^{g_{SS,9,2}} - \beta_{SS,9}G_9 \quad (6)$$

Note again that the two external variables W, which represents the effect of *Botrytis cinerea* inoculation and L_I, which represents the effect of light are considered as the independent variables.

Extended S-System model: Following (3), we arrive at the following ODEs for the 9GRN,

$$\begin{aligned}
\frac{dG_1}{dt} &= \alpha_{ES,1} G_2^{g_{ES,1,1}} G_5^{g_{ES,1,2}} G_6^{g_{ES,1,3}} - \beta_{ES,1} G_1 + \gamma_{ES,1,1} \\
\frac{dG_2}{dt} &= -\beta_{ES,2} G_2 + \gamma_{ES,2,1} + \gamma_{ES,2,2} W \\
\frac{dG_3}{dt} &= -\beta_{ES,3} G_3 + \gamma_{ES,3,1} + \gamma_{ES,3,2} W \\
\frac{dG_4}{dt} &= \alpha_{ES,4} G_8^{g_{ES,4,1}} G_9^{g_{ES,4,2}} - \beta_{ES,4} G_4 + \gamma_{ES,4,1} + \gamma_{ES,4,2} W \\
\frac{dG_5}{dt} &= \alpha_{ES,5} G_9^{g_{ES,5,1}} - \beta_{ES,5} G_5 + \gamma_{ES,5,1} + \gamma_{ES,5,2} W \\
\frac{dG_6}{dt} &= -\beta_{ES,6} G_6 + \gamma_{ES,6,1} + \gamma_{ES,6,2} W \\
\frac{dG_7}{dt} &= \alpha_{ES,7} G_9^{g_{ES,7,1}} - \beta_{ES,7} G_7 + \gamma_{ES,7,1} + \gamma_{ES,7,2} W + \gamma_{ES,7,3} L_I \\
\frac{dG_8}{dt} &= \alpha_{ES,8} G_3^{g_{ES,8,1}} G_4^{g_{ES,8,2}} G_7^{g_{ES,8,3}} - \beta_{ES,8} G_8 + \gamma_{ES,8,1} + \gamma_{ES,8,2} L_I \\
\frac{dG_9}{dt} &= \alpha_{ES,9} G_5^{g_{ES,9,1}} - \beta_{ES,9} G_9 + \gamma_{ES,9,1} + \gamma_{ES,9,2} W
\end{aligned} \tag{7}$$

As a remark, despite the regulation type is unknown, the ODEs for these three models can still be written down as depicted in (5), (6) and (7), as the regulation type can be inferred through the sign of the estimated parameters. Also, for the S-System based models, we set $h_{i,j} = 1$ to reduce the amount of parameters that need to be estimated.

Hill Function model: Unlike the previous three models, the ODEs of the Hill Function model can only be written down when the type of regulation is known. To facilitate the

derivation of these ODEs, we need to employ additional steps to infer those regulation types.

GRN network inference and parameter estimation using Hill Function ODEs can be a challenging problem, as repeatedly solving the ODEs via numerical integration can be computationally expensive. In this study, we employ our recently proposed parametric gradient-matching method (see Supplementary Text Section S1.3, Algorithm I and (Dony et al., 2019)) as the GRN inference approach, which incorporates dynamics information and computational efficient. It is an inference approach based on parametric Hill-Function nonlinear ODEs representation of a GRN (Babtie et al., 2014). The approach significantly reduced the computational cost of repeatedly solving the candidate ODEs via a two-step gradient matching. It first employs a Gaussian process to interpolate each time-course gene expression data. Then, the parameters of the ODEs are optimised by minimising the difference between interpolated derivatives and the right-hand-side of the ODEs. In such a way, the ODEs do not need to be solved explicitly, thereby reducing the computational cost. For more details of the method, see (Dony et al., 2019; Babtie et al., 2014). We note that there are copious of other similar methods to identify regulation type of the Hill Function model in a GRN (see e.g., (Aijo and Bonneau, 2016; Saint-Antoine and Singh, 2020)). As the main goal of this work is to perform comparative analysis of the GRN models and not on the network inference algorithm, we will treat the identified regulation type from our network inference algorithm as the correct regulation for our comparative analyses. The summary of the identified regulation types is given in Table 1.

Table 1

Identified regulation types for interaction within 9GRN following the parametric gradient-matching approach. The (+) and (–) signs indicate the activation and inhibition regulation types respectively. The signs for W and L_I indicate that this gene is positively or negatively regulated by those external inputs (see Fig. 1).

Number	Gene	Regulation Types
1	*ORA59*	*MYB51* (+), *ANAC055* (-), *a-ERF-1* (-)
2	*MYB51*	W (+)
3	*LOL1*	W (+)
4	*AT1G79150*	*CHE* (+), *RAP2.6L* (-), W (-)
5	*ANAC055*	*RAP2.6L* (+), W (+)
6	*a-ERF-1*	W (+)
7	*ATML1*	*RAP2.6L* (-),W (-), L_I (+)
8	*CHE*	*LOL1* (+), *AT1G79150* (+), *ATML1* (+), L_I (+)
9	*RAP2.6L*	*ANAC055* (+), W (+)

With that, the corresponding ODEs are given as follow.

$$\frac{dG_1}{dt} = \frac{\alpha_{HF,1,1}G_2^2}{K_{HF,1,1}+G_2^2} + \frac{\alpha_{HF,1,2}}{K_{HF,1,2}+G_5^2} + \frac{\alpha_{HF,1,3}}{K_{HF,1,3}+G_6^2} - \beta_{HF,1}G_1 + B_{HF,1}$$

$$\frac{dG_2}{dt} = \frac{\alpha_{HF,2,1}W^2}{K_{HF,2,1}+W^2} - \beta_{HF,2}G_2 + B_{HF,2}$$

$$\frac{dG_3}{dt} = \frac{\alpha_{HF,3,1}W^2}{K_{HF,3,1}+W^2} - \beta_{HF,3}G_3 + B_{HF,3}$$

$$\frac{dG_4}{dt} = \frac{\alpha_{HF,4,1}G_8^2}{K_{HF,4,1}+G_8^2} + \frac{\alpha_{HF,4,2}}{K_{HF,4,2}+G_9^2} + \frac{\alpha_{HF,4,3}}{K_{HF,4,3}+W^2} - \beta_{HF,4}G_4 + B_{HF,4}$$

$$\frac{dG_5}{dt} = \frac{\alpha_{HF,5,1}G_9^2}{K_{HF,5,1}+G_9^2} + \frac{\alpha_{HF,5,2}W^2}{K_{HF,5,2}+W^2} - \beta_{HF,5}G_5 + B_{HF,5}$$

$$\frac{dG_6}{dt} = \frac{\alpha_{HF,6,1}W^2}{K_{HF,6,1}+W^2} - \beta_{HF,6}G_6 + B_{HF,6}$$

$$\frac{dG_7}{dt} = \frac{\alpha_{HF,7,1}}{K_{HF,7,1}+G_9^2} + \frac{\alpha_{HF,7,2}}{K_{HF,7,2}+W^2} + \frac{\alpha_{HF,7,3}}{K_{HF,7,3}+L_I^2} - \beta_{HF,7}G_7 + B_{HF,7}$$

$$\frac{dG_8}{dt} = \frac{\alpha_{HF,8,1}G_3^2}{K_{HF,8,1}+G_3^2} + \frac{\alpha_{HF,8,2}G_4^2}{K_{HF,8,2}+G_4^2} + \frac{\alpha_{HF,8,3}G_7^2}{K_{HF,8,3}+G_8^2} + \frac{\alpha_{HF,8,4}L_I^2}{K_{HF,8,4}+L_I^2} - \beta_{HF,8}G_8 + B_{HF,8}$$

$$\frac{dG_9}{dt} = \frac{\alpha_{HF,9,1}G_5^2}{K_{HF,9,1}+G_5^2} + \frac{\alpha_{HF,9,2}W^2}{K_{HF,9,2}+W^2} - \beta_{HF,9}G_9 + B_{HF,9} \quad (8)$$

In all the four models, the infection of *Botrytis cinerea* is modelled as a step function with gradual increase from time 48 to 72 hours, *i.e.*, the time inoculation occurs. Mathematically, this is modelled as

$$W = \begin{cases} 0 & 0 \le t < 48 \\ \frac{1}{24}t - 2 & 48 \le t \le 72 \\ 1 & t > 72 \end{cases} \quad (9)$$

For the light regulated genes, these genes are affected by the duration of photoperiod of light. In (Windram et al., 2012), the experiment was carried out under 16 hours of light and 8 hours of dark. The resulting genes in response to this photoperiod duration behave in a sinusoidal manner with its peak between 8 to 10 hours at the first instance

of light. In view of this, the effect of light is modelled as a sinusoidal signal that peaks at around 9 hours at the first stance of light and its expression is given by

$$L_I = \sin\left(\frac{2\pi t}{T_P} + \phi\right) + B_L \tag{10}$$

where $B_L = 1.0001$ is the expression base level, $\phi = \pi/6$ radian is the phase shift and $T_P = 24$ hours is the period of the sinusoid. The reason for setting $B_L = 1.0001$ is to avoid L_I becoming zero, which can be problematic when it is used in modelling 9GRN using standard S-System.

3.4. Parameter estimation

The data used in this study is taken from (Foo et al., 2018a). For the linear and two S-System based models, the parameters of the corresponding models were fitted to the experimental data set by minimising the weighted mean squared error (WMSE) between the simulated and experimental data, *i.e.*, by finding

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}}\, \mathcal{W}(G(t), \hat{G}(t, \Theta) \tag{11}$$

where

$$\mathcal{W}\left(G(t), \hat{G}(t, \Theta)\right) = \frac{1}{n_G}\frac{1}{n_T}\sum_{i=1}^{n_G}\sum_{j=1}^{n_T}\left(\frac{G_i(t_j) - \hat{G}_i(t_j, \Theta}{\mathcal{A}_i}\right)^2 \tag{12}$$

with

$$\mathcal{A}_i = \max_{1 \le j \le n_T} G_i(t_j) \tag{13}$$

where G represents the gene component, t denotes the time index, $n_G = 9$ is the number of gene components and $n_T = 48$ is the number of data point used. Given that the amplitude of different gene components is different, to allay any bias during the optimisation procedure for model parameter fitting, we introduce the weights $\mathcal{A}$ in (12) where we normalise each time series to its maximum value. The MATLAB function `fminsearch` that employs Nelder-Mead simplex algorithm was used to minimise (11). The estimated parameters for these three models are given in Tables S1 to S3. Note that the estimated model parameters of the linear model is somewhat different than the one provided in (Foo et al., 2018a). This is because in this study, instead of directly

using the estimated parameters from (Foo et al., 2018a), they are used as the initial value for the optimisation to determine whether any further improvement in terms of the WMSE can be achieved. The estimated parameters given in Table S1 are very close to the one estimated in (Foo et al., 2018a) suggesting a high confidence level in the estimated parameters for the linear model.

The parameters associated with the Hill Function model are given in Table S4. These parameters have been estimated together with the inference algorithm (see Supplementary Text, Section S1.3) via the gradient-matching method (see (Dony et al., 2019) and its Supplementary Material for more details).

Table 2

Identified regulation types based on estimated parameters of 9GRN using linear (L), standard S-System (SS) and extended S-System models (ES). The (**+**) and (**–**) signs indicate the activation and inhibition regulation types, respectively. The signs for W and L_I indicates that this gene is positive/negative regulated by those external inputs (see Fig. 1). The regulation type for Hill Function model (HF) shown in Table 1 is also listed for ease of comparison. Identified regulation types that are different are highlighted in grey.

Number	Gene	Regulation Types
1	*ORA59*	L: *MYB51* (+), *ANAC055* (-), *a-ERF-1* (+)
		SS: *MYB51* (+), *ANAC055* (-), *a-ERF-1* (+)
		ES: *MYB51* (+), *ANAC055* (-), *a-ERF-1* (+)
		HF: *MYB51* (+), *ANAC055* (-), *a-ERF-1* (-)
2	*MYB51*	L: W (+)
		SS: W (+)
		ES: W (+)
		HF: W (+)
3	*LOL1*	L: W (-)
		SS: W (-)
		ES: W (+)
		HF: W (+)
4	*AT1G79150*	L: *CHE* (+), *RAP2.6L* (-), W (+)
		SS: *CHE* (+), *RAP2.6L* (-), W (-)
		ES: *CHE* (+), *RAP2.6L* (+), W (+)
		HF: *CHE* (+), *RAP2.6L* (-), W (-)
5	*ANAC055*	L: *RAP2.6L* (+), W (+)
		SS: *RAP2.6L* (+), W (+)
		ES: *RAP2.6L* (+), W (+)
		HF: *RAP2.6L* (+), W (+)
6	*a-ERF-1*	L: W (+)
		SS: W (+)
		ES: W (+)
		HF: W (+)
7	*ATML1*	L: *RAP2.6L* (-), W (+), L_I (+)
		SS: *RAP2.6L* (-), W (+), L_I (+)
		ES: *RAP2.6L* (-), W (+), L_I (+)
		HF: *RAP2.6L* (-), W (+), L_I (+)
8	*CHE*	L: *LOL1* (+), *AT1G79150* (+), *ATML1* (+), L_I (+)
		SS: *LOL1* (+), *AT1G79150* (+), *ATML1* (+), L_I (+)
		ES: *LOL1* (+), *AT1G79150* (+), *ATML1* (+), L_I (+)
		HF: *LOL1* (+), *AT1G79150* (+), *ATML1* (+), L_I (+)
9	*RAP2.6L*	L: *ANAC055* (+), W (+)
		SS: *ANAC055* (+), W (+)
		ES: *ANAC055* (+), W (+)
		HF: *ANAC055* (+), W (+)

The identified regulation types for these three models are given in Table 2. We have also included the regulation types inferred from Hill Function model in this table for ease of comparison. In general, there is a general consensus on the identified regulation types shown in Table 2 apart from genes *ORA59*, *LOL1* and *AT1G79150*. Specifically, for gene *ORA59*, only the inferred regulation type for *a-ERF-1* when using

the Hill Function model is different from the other three models. For *ORA59*, the time series shows an increasing trend, which is consistent with the increasing trend of *a-ERF-1*, suggesting a higher possibility of a positive regulation, which agrees with the three models rather than the Hill Function model. For gene *LOL1*, there is difference in the inferred pathogen regulation type with the linear and standard S-System models identified negative regulation, while extended S-System and Hill Function models identified positive regulation. Lastly, for gene *AT1G79150*, the inferred regulation types for *RAP2.6L* and pathogen are different across all four models. A detail look at the time series data for these genes *LOL1* and *AT1G79150* (Fig. S1) suggests that the difficulty in identifying these regulation types is attributed to the almost plateau nature of these two gene expression levels.

Using the identified parameters given in Tables S1 to S4, we compared the predictive capability of the models with the experimental data on a set of data that is not used in parameter estimation exercise and the result are shown in Fig. 2. As a quantitative measure, we calculated the WMSE, using (12), and they are shown in Table 3.

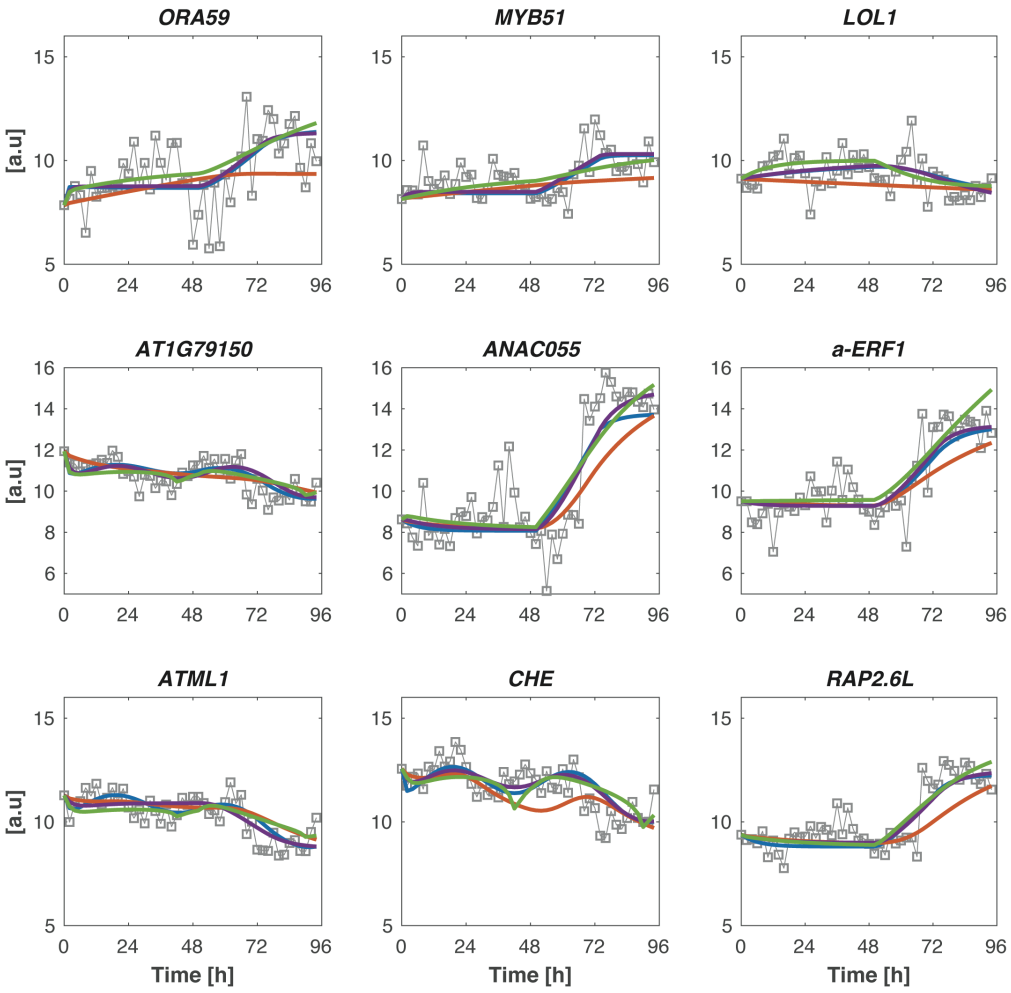

Fig. 2. Comparison of the models against experimental data set that is not used in the parameter estimation exercise. Solid grey with 'square': Experimental data. Solid blue: Linear model. Solid red: Hill Function model. Solid green: Standard S-System model. Solid purple: Extended S-System model.

Table 3

Average total WMSE for both 'training' and 'validation' data sets for 9GRN, which is calculated by taking the average sum of the individual WMSE given in Tables S5 and S6. The 'training' data set refers to the data that is used in parameter estimation exercise, while the 'validation' data set refers to the data that is not used in the parameter exercise.

Model	Average Total WMSE (training)	Average Total WMSE (validation)
Linear	0.00267	0.00543
Hill Function	0.00614	0.00784
Standard S-System	0.00359	0.00606
Extended S-System	0.00256	0.00516

The results shown in Fig. 2 and Table 3 show that all four models are able to pick up the general trend of the data well. Specifically, the linear and two S-System models perform really well with relatively smaller total WMSE compared to the Hill Function model. For the Hill Function model, there are several instances where the model falls short in terms of realising the correct amplitude levels (*e.g.*, *ORA59* and *MYB51*), which is also reflected in the individual WMSE shown in Tables S5 and S6.

To further test the performance of these models, we compare qualitatively the dynamics of these four models against mutant data set, where two different genes, i.e., Δnac and $\Delta rap2.6l$ have been mutated. Fig. 3 shows the predictive capability of the four models against the mutant data. In general, all models pick up the correct trend of the mutant behaviours albeit the two S-System models have difference in the amplitude. For instance, gene ORA59 from the extended S-System model has higher expression level under both knockdown mutants. Similarly, gene RAP2.6L from standard S-System model has lower expression level under $\Delta anac055$. Nevertheless, all the models are able to predict the mutant behaviours qualitatively well. Readers who are interested in the quantitative mutant analysis can refer to Supplementary Text.

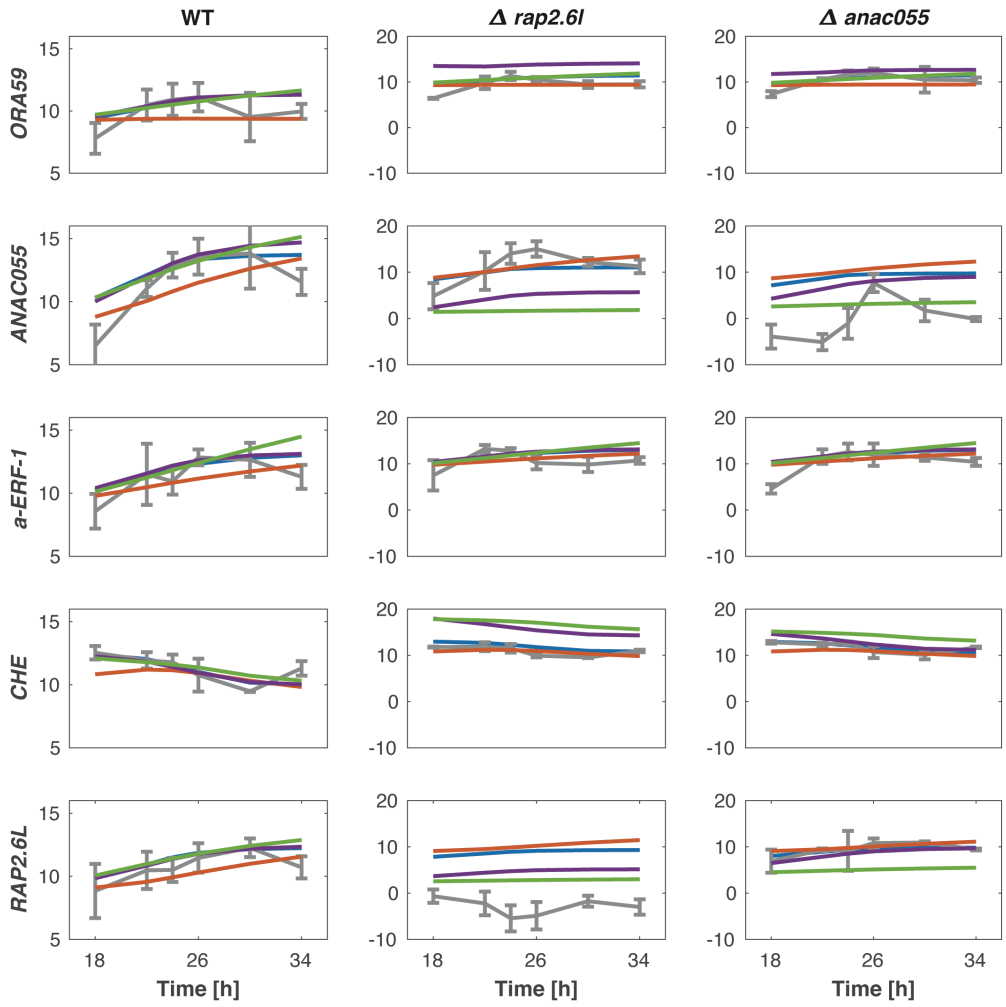

Fig. 3. Comparison of the models against mutant experimental data set. For the simulation of the mutant, we reduce the production rate associated with the knockdown gene by 20%. Solid grey with error bar: Experimental data. Solid blue: Linear model. Solid red: Hill Function model. Solid green: Standard S-System model. Solid purple: Extended S-System model.

3.5. Assessing model quality using Akaike weight based on Akaike Information Criterion (AIC)

While the WMSE and the mutant analysis provide respectively the quantitative and qualitative approaches of the performance of the model, these approaches however do not reflect fully the quality of fit given the different model structures employed and the number of parameters used. In order to quantify the relative quality of the model fits to the experimental training data obtained with the four models considered, we employed the widely-used Akaike Information Criterion (AIC), which

calculates the best approximating model to a given dataset with respect to Kullback-Leibler information loss (Burnham and Anderson, 2002, 2004).

For a given model, the AIC is defined as

$$\mathrm{AIC} = -2\ln\left(\hat{\mathcal{L}}\right) + 2K_{\Theta} \tag{14}$$

where $\hat{\mathcal{L}}$ is the maximised log-likelihood and K_{Θ} is the number of model parameters. Consider that the optimal parameter estimates for all four 9GRN models were acquired through the minimisation of weighted least squares cost function, it can be shown that (Banks and Joyner, 2017)

$$\ln\left(\hat{\mathcal{L}}\right) = -\frac{n_G n_T}{2}\ln(2\pi + 1) - n_T \sum_{i=1}^{n_G} \ln(\mathcal{A}_i) - \frac{n_G n_T}{2}\ln\left(\mathcal{W}\left(G(t), \hat{G}(t,\Theta)\right)\right) \tag{15}$$

with n_G is the number of genes, n_T is the number of data points in the time series, $\mathcal{A}_i$s as defined in (13) are the cost function weights and $\mathcal{W}(G(t), \hat{G}(t,\Theta))$ as defined in (12).

By denoting $\mathrm{AIC_i}$ as the AIC value of the ith model, these four 9GRN models are ranked by their AIC differences calculation, *i.e.*,

$$\Delta_i(\mathrm{AIC}) = \mathrm{AIC}_i - \min_{1\le i\le 4} \mathrm{AIC}_i \tag{16}$$

and finally, the corresponding Akaike weights can be calculated as follow:

$$w_i(\mathrm{AIC}) = \frac{\exp\left(-\frac{1}{2}\Delta_i(\mathrm{AIC})\right)}{\sum_{i=1}^{4}\exp(-\frac{1}{2}\Delta_i(\mathrm{AIC}))} \tag{17}$$

We can interpret this Akaike weight, $w_i(\mathrm{AIC})$ as the probability that the ith model is the best from the perspective of minimising K-L information loss, given the set of candidate models and the data. In addition, the strength of evidence that favours model i over model j is quantified by the ratio $w_i(\mathrm{AIC})/w_j(\mathrm{AIC})$ (Burnham and Anderson, 2002, 2004; Wagenmakers and Farrell, 2004; Banks and Joyner, 2017).

Finally, since n_G, n_T and $\mathcal{A}_i$ in (15) are fixed across the respective GRN models, the expression of the AIC value (*i.e.*, (14)) can be further simplified to

$$\mathrm{AIC} = n_G n_T \ln\left(\mathcal{W}\left(G(t), \hat{G}(t,\Theta)\right)\right) + 2(K_\Theta + 1) \tag{18}$$

where (18) is then used to compute the AIC differences $\Delta_i(\mathrm{AIC})$ and Akaike weights $w_i(\mathrm{AIC})$ of a given 9GRN model.

The AIC criterion in Table 4 indicates that the two most viable candidate models (in the sense of K-L divergence) are the extended S-System model and the linear model with their Akaike weight of $w_{ES}(\mathrm{AIC}) = 0.9836$ and $w_L(\mathrm{AIC}) = 0.0164$, respectively. Between these two models, the ratio of $w_{ES}(\mathrm{AIC})/w_L(\mathrm{AIC}) \approx 60$ suggests that the extended S-System model is 60 times more likely to be the viable model candidate compared to the linear model. On the other hand, the Akaike weights also exclude the Hill Function and the standard S-System models as the viable models given their Akaike weights are close to zero.

Table 4

Ranking model fits to experimental data based on AIC weights for 9GRN. The notation L, HF, SS, ES, denote the linear, Hill Function, Standard S-System and Extended S-System models, respectively. Here $n_G = 9$, $n_T = 48$, K_Θ is the number of parameters in the model, $\mathcal{W}\left(G(t), \hat{G}(t,\Theta)\right)$ is the WMSE best fit to the data set used for parameter estimation, $\Delta_i(\mathrm{AIC})$ is the AIC differences and $w_i(\mathrm{AIC})$ is the Akaike weights for each model.

Model	L	HF	SS	ES
K_Θ	38	58	38	44
$\mathcal{W}\left(G(t), \hat{G}(t,\Theta)\right)$	0.00267	0.00614	0.00359	0.00256
$\Delta_i(\mathrm{AIC})$	8.182	397.506	134.696	0
$w_i(\mathrm{AIC})$	0.0164	4.14×10^{-89}	5.55×10^{-30}	0.9836

3.6. Robustness of the models to parameter uncertainties

In practice, the estimated parameters of the model are subjected to uncertainties (*e.g.*, intrinsic noise, modelling error, etc). To test the robustness of these four models, we perform a global sensitivity analysis, where all the parameters of the model are simultaneously varied in a random manner in each simulation. In this study, we assume that the uncertainties account for the parameters to vary $\pm 30\%$ (see *e.g.*, (Acker et al., 1982; Transtrum and Qiu, 2012; Paulino et al., 2019)) from its nominal

value. To ensure an unbiased sampling of the parameter values, we adopted the *Latin Hypercube Sampling* approach (see *e.g.*, (Marino et al., 2008; Sheikholeslami and Razavi, 2017)) to randomly generate a parameter set that is within $\pm 30\%$ of the original value of each parameter for each simulation.

In the Latin Hypercube Sampling approach, each model parameter is first discretised into N_s evenly spaced intervals from the defined lower and upper bounds. As we are varying the parameter within $\pm 30\%$, this results in N_s evenly spaced interval between 0.7× to 1.3× the nominal parameter. Here, we choose $N_s = 1000$, and this results in a total number of $(1000 \times K_\Theta)$ randomly combined parameter sets, where K_Θ is the number of parameters in each of the four models. We run a total number of 10000 simulations for each of the four models, where in each simulation, we sample only once from this total number of randomly combined parameter sets. Due to the non-repetitive nature of this sampling approach, not only the biased sampling can be averted, an extensive sampling within the model parameter range of interest can also be covered (Marino et al., 2008).

Following (Wang et al., 2014), we compute the Mean Relative Error (MRE) given by

$$\text{Mean Relative Error (MRE)} = \frac{1}{n_G}\sum_{i=1}^{n_G}\sum_{j=1}^{n_T}\left|\frac{G_i(t_j) - \hat{G}_i(t_j, \Theta)}{G_i(t_j)}\right| \tag{19}$$

as a quantitative metric to evaluate the model response to parameter uncertainties using the same notation as (12). To determine the robustness of the models, we collate the number of simulations (over 10000), where the MREs are within 4× the nominal MRE value. The choice of 4× is based on the observation over 10000 simulations that the performance of the models is deemed acceptable. To compare the robustness of the model, a model is considered more robust than the other if the number of simulations within 4× nominal MRE value is higher in the former than the latter.

Table 5 shows the MRE values for all four models. Defining N_{SIM} as the number of simulations for each model where the MRE values are within 4× nominal MRE value. The results show that the Hill Function and the extended S-System models has the largest and smallest N_{SIM} values, respectively suggesting these models respectively being relatively the most and least robust to parameter uncertainty. Also, we notice that the N_{SIM} values for the two S-System based models are comparative smaller,

which is expected given that the exponent term tends to be sensitive to uncertainties (Rinon et al., 2019).

Table 5

Nominal MRE for each model and the number of simulations across 10000 that has the MRE values within 4× the nominal MRE value. The notation L, HF, SS, ES, denote the linear, Hill Function, Standard S-System and Extended S-System models, respectively. N_{SIM} denotes the number of MRE within 4× nominal MRE value.

Model	L	HF	SS	ES
Nominal MRE	0.0523	0.0704	0.0516	0.0611
N_{SIM}	2285	9271	1082	731

To see how the N_{SIM} values are distributed, we plot the histogram in Fig. S2, and the histogram shows that the majority of the MRE values are distributed close to the nominal MRE value indicating these models are more robust than anticipated. To further investigate this, we plot the lower and upper bound of each model simulated using the parameter sets within N_{SIM} that produce the largest and smallest MRE value, and these plots are shown in Figs. S3 to S6. Interestingly, majority of the genes are robust to parameter uncertainty where their uncertainty bounds are narrow apart from a handful of genes (*e.g.*, *ORA59*, *ANAC055* and *CHE*), where we observe a wider uncertainty bound. Moreover, despite having the largest N_{SIM} value, the Hill Function model has four genes with wide uncertainty bounds compared to the same four genes for the other three models. This suggests that while the larger N_{SIM} in Hill Function model is most probably attributed to the genes with narrow uncertainty bound, it comes at the expense of reduced robustness in other genes such as *CHE*, which has the widest uncertainty bound.

4. Extension to other gene regulatory networks

Our analysis using 9GRN has suggested the linear and extended S-System models being the viable model candidate for pathogen-infected plant GRN. To investigate whether the above analysis has wider applicability, we repeat the above analysis on three other GRNs, *i.e.*, the DREAM3 and DREAM4 gene regulatory

networks (Stolovitzky et al., 2007, 2009), and the plant circadian gene regulatory network (De Caluwe et al., 2016).

4.1. DREAM3 and DREAM4 gene regulatory networks

The DREAM3 and DREAM4 networks are *in silico* gene regulatory networks established for public challenges related to the development of network inference algorithms from experimental data, which typically consists of realistic temporal data (Marbach et al., 2009) of each gene in the network. Despite being *in silico* networks, they are a subset of an actual network from organisms, such as *E. coli* and *S. cerevisiae*, hence a good representation of the biological system. The two networks used in this study are shown in Fig. 4.

Just like the 9GRN, both the DREAM GRNs consist of 10 genes in which their interactions are known but their regulation types are unknown. This property makes them suitable for us to repeat the above analysis with one small change to the Hill Function model. Instead of employing the parametric gradient-matching method (see Algorithm I (Dony et al., 2019) in Supplementary Text Section S1.3) to determine the regulation types, we use the alternate approach, *i.e.*, using the inferred regulation types from the linear model, given its better performance (see Tables S8 and S9).

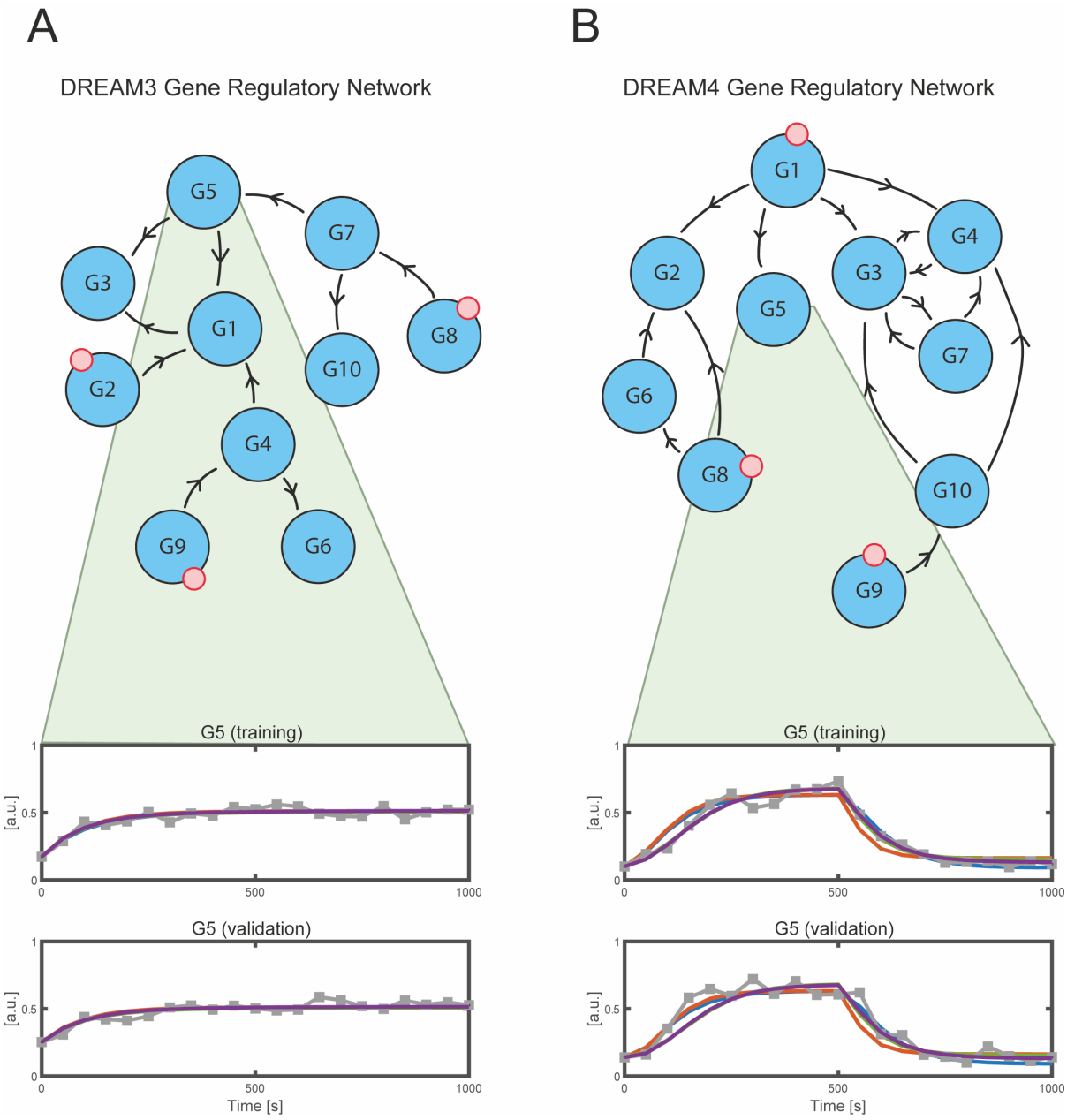

Fig. 4. (A) DREAM3 and (B) DREAM4 gene regulatory networks. Like the 9GRN, the information about the known directional arrows indicates the influence of one gene to another while the regulation types are unknown. The red circle denotes the gene that is directly affected by external perturbation. The comparison between the four model structures and the data for Gene 5 for both networks are shown for illustration. Solid grey with 'square': Experimental data. Solid blue: Linear model. Solid red: Hill Function model. Solid green: Standard S-System model. Solid purple: Extended S-System model. For the comparison for all the genes, see Figures S7 and S8 for DREAM3 and Figures S14 and S15 for DREAM4.

The results of the analysis for the DREAM3 and DREAM4 GRNs are shown in Figures S7 to S13 and Figures S14 to S20, respectively. As an illustration, the comparison of the four model structures on gene G5 for both DREAM3 and DREAM4 GRNs is shown in Fig. 4. In addition, the summary of the comparison against the same four criteria is given in Tables S18 and S26 for DREAM3 and DREAM4, respectively. These DREAM3 and DREAM4 results in the same conclusion, *i.e.*, the linear and extended S-System models are the most viable model structures describing these two DREAM GRNs.

4.2. Plant circadian gene regulatory network

Here, we consider the plant circadian GRN (De Caluwe et al., 2016) to analyse the effect of mis-identification of the regulation types on the four model structures. The plant circadian GRN is the ideal candidate GRN for this illustration given the recent discovery of the change in regulation types. The plant circadian GRN is shown in Fig. 5. Previously, it was thought that gene P97 is activated by gene CL (red arrow head in Fig. 5) and most of the plant circadian GRN models that have been constructed are based on the knowledge of this regulation. However, this regulation is in fact an inhibition (green bar head in Fig. 5) following the discovery made in (Adams et al., 2015). With this change of regulation, all the previous plant circadian GRN models have to be revised. Here, we illustrate the impact of this change in regulation on the four model structures. As shown in Fig. 5, all four model structures are able to reproduce the experimental data of P97, suggesting the robustness of these model structures against mis-identification of the regulation types (Supplementary Text Section S1.4).

The robustness of the model structures against mis-identification of the regulation type has partially been demonstrated with the Hill Function model. In Table 2, we highlighted the different regulation types identified between the Hill Function model and the other three models. An alternate Hill Function model was then proposed, and the alternate Hill Function model was also able to reproduce experimental data as shown in Tables S8 and S9.

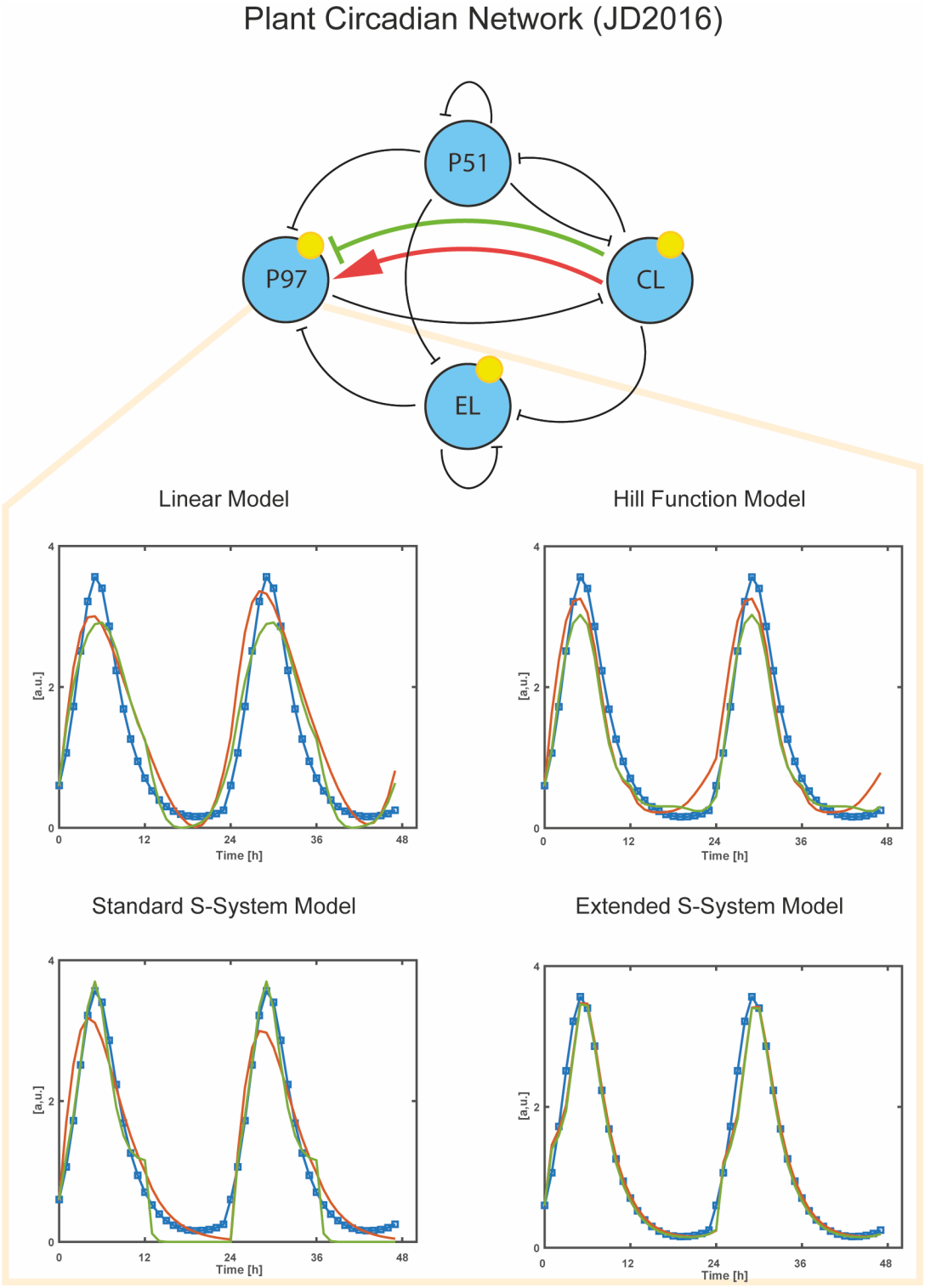

Fig. 5. Plant circadian network of JD2016 (De Caluwe et al., 2016). The arrow and bar heads represent activation and inhibition regulations, respectively. The yellow circle represents genes that are light-regulated. It is initially thought that genes LC activates P97 as indicated by the red arrow head. Recent findings in plant circadian literature show that LC in fact is inhibiting P97 as indicated by the green bar head. Simulated P97 expression given by the four model structures are shown. The solid blue with 'square': Experimental data. Solid red: CL activates P97. Solid green: CL inhibits P97.

5. Discussion and Conclusion

In this study, we have compared four dynamical models of 9GRN obtained using a data-driven modelling approach in terms of four criteria, namely their ease of identifying regulation type, predictive capability, quality of data fit based on AIC and robustness to parameter uncertainties.

The linear and the two S-System based models have a general model structure that can facilitate the identification of the regulation types directly from data through the sign of the estimated parameters. In contrast, due to the requirement of different functions for different regulation types for the Hill Function model (Section 3.2), additional steps are required to ensure the most viable regulation types when identifying them from data, making this model the least favoured in terms of Criterion I. Furthermore, despite the identified regulation types given in Table 2 showing a consensus, when comparing the difference in the identified regulation types, the linear and two S-System based models have more common agreement compared to the Hill Function model for *e.g.*, in gene *ORA59*.

In terms of the model predictive capability, the linear and extended S-System models rank higher in terms of their smaller WMSE value both in the training and validation data set compared to the standard System and Hill Function models (Table 3) suggesting Criterion II is in favour of these two models. In terms of mutant analysis, between the linear and extended S-System models, the former qualitatively better predicts the mutant behaviours (Fig. 3) than the latter.

For Criterion III, the analysis of AIC weights (Table 4) suggests the linear and extended S-System models are the two most viable candidate models compared to the standard S-System and Hill Function models. Nevertheless, the extended S-System model is 60 times more likely to be the candidate model compared to the linear model given its larger AIC weight, $w_{ES}(\mathrm{AIC})$, which suggests the extended S-System in the most favoured model for Criterion III.

For the last criterion, the analyses using Latin Hypercube Sampling and MRE (Table 5) indicate that the Hill Function model has the largest N_{SIM}, suggesting this model is relatively robust against parameter uncertainty compared to the other three models. Interestingly, when analysing the histogram of the MRE distributions (Fig. S2) and the lower and upper bounds uncertainty plots (Figs. S3-S6), the width of the

uncertainty bounds are smaller and similar across the linear and the two S-System based models. On the other hand, despite the Hill Function model having narrow uncertainty bounds across most of the genes in 9GRN, some genes (*e.g.*, *CHE* and *ORA59*) have the widest uncertainty bound across all four models. This suggests that the large N_{SIM} of the Hill Function model are attributed to the narrow bounds of most genes but at the expense of wide bounds on certain genes like *CHE*.

Table 6

Summary of the model performance across four criteria. For Criteria II to IV, the model is ranked in bracket with '1' being the most favoured model and '4' being the least favoured model according to the metric used in the comparison. The notation L, HF, SS, ES, denote the linear, Hill Function, Standard S-System and Extended S-System models, respectively. For Criterion II, the notation 'Tra.' and 'Val.' represent training and validation, respectively. The Total Rank Score (TRS) is the sum of the ranking number across Criteria II to IV given in the bracket.

Criterion	L	HF	SS	ES
I	Easy due to its general model structure	Difficult due to extra steps required to determine the relevant function	Easy due to its general model structure	Easy due to its general model structure
II	Tra. WMSE = 0.00267 (2) Val. WMSE = 0.00543 (2)	Tra. WMSE = 0.00601 (4) Val. WMSE = 0.00768 (4)	Tra. WMSE = 0.00359 (3) Val. WMSE = 0.00606 (3)	Tra. WMSE = 0.00256 (1) Val. WMSE = 0.00516 (1)
III	$w_L(\mathrm{AIC}) = 0.0164$ (2)	$w_{HF}(\mathrm{AIC}) \approx 0$ (2)	$w_{SS}(\mathrm{AIC}) \approx 0$ (2)	$w_{ES}(\mathrm{AIC}) = 0.9836$ (1)
IV	$N_{SIM} = 2285$ (2)	$N_{SIM} = 9271$ (1)	$N_{SIM} = 1082$ (3)	$N_{SIM} = 731$ (4)
TRS	8	13	12	7

Table 6 summarises the performance of all four models across the four criteria. For Criteria II to IV, we provide the associated ranking in each criterion with '1' being the most favoured model and '4' being the least favoured model based on the metrics used to compare them. We then calculated the Total Rank Score (TRS), which is the sum of the ranking number given in bracket with the smallest and largest scores represent the most and least favoured models, respectively.

The extended S-System model scores the smallest TRS, followed closely by the linear model, while the Hill-Function model scores the largest TRS. While the linear model scores a lower TRS compared to the extended S-System model, the linear model performs consistently across all criteria with rankings of '2' compared to the extended S-System model. Based on this consistency, we surmise that the linear model is a more viable candidate model for constructing this 9GRN using a data-driven modelling approach.

The applicability of our analysis to other GRNs is demonstrated in three other GRNs. For the DREAM3 and DREAM4 GRNs, we obtain similar conclusion as the

9GRN. For the plant circadian GRN, we demonstrated the robustness of the model against mis-identification of regulation types.

The finding from our comparative analysis in principle agrees with the finding from (Wang et al., 2014), where in that study, the standard S-System model is found to be a more viable model compared to the Michaelis-Menten (Hill coefficient, n is set to 1 in (Wang et al., 2014)) and mass-action model for describing plant flowering time regulatory network. Our analysis extends that finding by comparing two additional models, *i.e.*, the linear and extended S-System models and explore a different plant regulatory network. More interestingly, between the standard and extended S-System models in our study, our analysis shows that the latter model outperforms the former model across the given criteria. This is expected given that the extended S-System model considers the external input as being a separate term instead of grouping them as part of the dependent variables. This thus provides more degree of freedom for the external input to influence the model dynamics, which could improve the accuracy of the model (Foo et al., 2020).

Our finding that the least viable model being the Hill Function model may seem surprising given its wide usage in modelling GRN. In a review work by (Kim and Tyson, 2020), it has been reported that the Michaelis-Menten rate law of the Hill Function has been often misused without ensuring the valid operating condition in many previous studies. The same review (and references therein) and our previous studies (Foo et al., 2018b, 2020) also highlighted issues pertaining to the identifiability of the Hill Function parameters. These two points accentuated the underlying challenge in using Hill Function model, which could possibly be the reason for its poor viability. One may potentially argue that the choice of network inference algorithm to obtain the Hill Function models (such as the one used in this study) may influence the analysis and the results. As such, we derive an alternate Hill Function model with the regulation type following the linear model and found that despite showing some improvement in Criteria II and III, the overall performance of the model is still ranked behind the linear and extended S-System model (see Tables S8 and S9).

Returning to our main question posed for this study – "*In using the data-driven modelling approach, what is the most viable model given the temporal data and knowledge about the 9GRN interaction*?" While traditionally Hill Function model has been the model of choice due to its biological relevance and interpretability (see (Youseph et al., 2015)), our comparative analysis seems to tip the balance towards

the linear model being the preferred choice of model for 9GRN suggesting the linear model used for genetic control design suggested in (Foo et al., 2018a) is a viable one. This analysis also advocates some of the previous works (see e.g., (Foo and Kim, 2014; Foo et al., 2017)) on the use of linear models in designing controller for other plant gene regulatory networks. Our results also suggest that when considering data-driven modelling approach, the extended S-System model can be a good alternative for modelling GRN as compared to the commonly used Hill Function model. We note that there is a need to ensure intelligibility and biological interpretability when using the extended S-System model. Nevertheless, our suggestion concurs with the findings made in (Vilela et al., 2008), where in that study, the authors concluded the increasing trend of using the S-System model structure in describing biological temporal data is attributed to the unique property of the S-System model structure being able to strike a balance between model intelligibility and interpretability.

Funding

This work was supported by a grant from The Royal Society via research grant RGS/R2/180195 to MF. LD acknowledges support by the Joachim Herz Foundation.

Availability

All the MATLAB simulation codes are available at https://github.com/mathiasfoo/9grncomparison

Authors contribution

MF and FH conceived the study. MF and LD performed the simulation and analysed the data. MF and FH drafted the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

MF would like to thank Dr. Xun Tang from Louisiana State University for providing assistance with the global sensitivity analysis using Latin Hypercube Sampling.

References

Acker, G., Johnson, A., Shea, M., 1982. Quantitative model for gene regulation by lambda phage repressor. Proceedings of the National Academy of Sciences 79, 1129–1133. doi:10.1073/pnas.79.4.1129.

Adams, S., Manfield, I., Stockley, P., Carre, I.A., 2015. Revised morning loops of the Arabidopsis circadian clock based on analyses of direct regulatory interactions. PLoS One 10, e0143943. doi:10.1371/journal.pone.0143943.

Aijo, T., Bonneau, R., 2016. Biophysically motivated regulatory network inference: progress and prospects. Human Heredity 81, 62–77. doi:10.1159/000446614.

Aoki, S., Lillacci, G., Gupta, A., Baumschlager, A., Schweingruber, D., Khammash, M., 2019. A universal biomolecular integral feedback controller for robust perfect adaptation. Nature 570, 533–537. doi:10.1038/s41586-019-1321-1.

Babtie, A.C., Kirk, P., Stumpf, M.P.H., 2014. Topological sensitivity analysis for systems biology. Proceedings of the National Academy of Sciences 111, 18507–18512. doi:10.1073/pnas.1414026112.

Banks, H.T., Joyner, M.L., 2017. AIC under the framework of least squares estimation. Applied Mathematics Letters 74, 33–45. doi:10.1016/j.aml.2017.05.005.

Bolouri, H., Davidson, E., 2002. Modeling transcriptional regulatory networks. Bioessays 24, 1118–1129. doi:10.1002/bies.10189.

Burnham, K., Anderson, D., 2002. Information and Likelihood Theory: A Practical Information-Theoretic Approach. Springer-Verlag, New York.

Burnham, K., Anderson, D., 2004. Multimodel inference: Understanding AIC and BIC in model selection. Sociology Methods and Research 33, 261–304. doi:10.1177/0049124104268644

Chai, L.E., Loh, S.K., Low, S.T., Mohamad, M.S., Deris, S., Zakaria, Z., 2014. A review on the computational approaches for gene regulatory network construction. Computers in Biology and Medicine 48, 55–65.

Chowdhury, A., Chetty, M., 2016. Reconstruction of large-scale gene regulatory network using S-System model. Evolutionary Computation in Gene Regulatory Network Research 2016, 185–210. doi:10.1002/9781119079453.ch8

De Caluwe, J., Xiao, Q., Hermans, C., Verbruggen, N., Leloup, J.C., Gonze, D., 2016. A compact model for the complex plant circadian clock. Frontiers in Plant Science 7. doi:10.3389/fpls.2016.00074.

Den Broeck, L., Gordon, M., Inze, D., Williams, C., Sozzani, R., 2020. Gene regulatory network inference: connecting plant biology and mathematical modeling. Frontiers in Genetics 11, 457. doi:10.3389/fgene.2020.00457.

Dony, L., He, F., Stumpf, M.P.H., 2019. Parametric and non-parametric gradient matching for network inference: a comparison. BMC Bioinformatics 20, 1–12. doi:10.1186/s12859-018-2590-7.

Emmert-Streib, F., Glazko, G., Altay, G., de Matos Simoes, R., 2012. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Frontiers in Genetics 3, 8. doi:10.3389/fgene.2012.00008.

Foo, M., Bates, D.G., Akman, O.E., 2020. A simplified modelling framework facilitates more complex representations of plant circadian clocks. PLoS Computational Biology 16, e1007671. doi:10.1371/journal.pcbi.1007671

Foo, M., Gherman, I., Denby, K.J., Bates, D.G., 2017. Control strategies for mitigating the effect of external perturbations on gene regulatory networks. Proceedings of IFAC

World Congress, 9-14 July 2017, Toulouse, France 50, 12647–12656. doi:10.1016./j.ifacol.2017.08.2237

Foo, M., Gherman, I., Zhang, P., Bates, D.G., Denby, K.J., 2018a. A framework for engineering stress resilient plants using genetic feedback control and regulatory network rewiring. ACS Synthetic Biology 7, 1553–1564. doi:10.1021/acssynbio.8b00037.

Foo, M., Kim, J., Bates, D.G., 2018b. Modelling and control of gene regulatory networks for perturbation mitigation. IEEE/ACM Transactions on Computational Biology and Bioinformatics 16, 583–595. doi:10.1109/TCBB.2017.2771775.

Foo, M., Kim, P.J., 2014. Modeling and control design of plant circadian system for flowering time in Arabidopsis. Proceedings of IEEE Conference on Control Applications, 8-10 October 2014, Antibes, France, 1687–1692. doi:10.1109/CCA.2014.6981555.

Gherman, I., 2018. Engineering stress resilient plants using gene regulatory network rewiring. Ph.D. thesis. University of Warwick.

Hache, H., Lehrach, H., Herwiga, R., 2009. Reverse engineering of gene regulatory networks: a comparative study. EURASIP Journal on Bioinformatics and Systems Biology 2009, 1–12. doi:10.1155/2009/617281

He, F., Murabito, E., Westerhoff, H.V., 2016. Synthetic biology and regulatory networks: where metabolic systems biology meets control engineering. Journal of the Royal Society Interface 13, 20151046. doi:10.1098/rsif.2015.1046

Jamir, Y., Guo, M., Oh, H.S., Petnicki-Ocwieja, T., Chen, S., Tang, X., Dickman, M.B., Collmer, A., Alfano, J.R., 2007. Identification of Pseudomonas syringae type III effectors that can suppress programmed cell death in plants and yeast. Plant Journal 37, 554–565. doi:10.1046/j.1365-313X.2003.01982.x

Jones, J.D.G., Dangl, J.L., 2006. The plant immune system. Nature 444, 323–329. doi:10.1038/nature05286.

Karlebach, G., Shamir, R., 2008. Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology 9, 770–780. doi:10.1038/nrm2503.

Kim, J., Tyson, J., 2020. Misuse of the Michaelis-Menten rate law for protein interaction networks and its remedy. PLoS Computational Biology 16, e1008258. doi:10.1371/journal.pcbi.1008258.

Long, T., Brady, S., Benfey, P., 2008. Systems approaches to identifying gene regulatory networks in plants. Annual Review of Cell and Developmental Biology 24, 81–103. doi:10.1146/annurev.cellbio.24.110707.175408

Maki, Y., Tominaga, D., Okamoto, M., Watanabe, S., Eguchi, Y., 2000. Development of a system for the inference of large scale genetic networks. Biocomputing 2001, 446–458. doi:10.1142/9789814447362_0044

Marbach, D., Schaffter, T., Mattiussi, T., Floreano, D., 2009. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. Journal of Computational Biology 16, 229–239. doi:10.1089/cmb.2008.09TT.

Marino, S., Hogue, I., Ray, C., Kirschner, D., 2008. A methodology for performing global uncertainty and sensitivity analysis in systems biology. Journal of Theoretical Biology 254, 178–196. doi:10.1016/j.jtbi.2008.04.011

Ng, D.W.K., Abeysinghe, J.K., Kamali, M., 2018. Regulating the regulators: the control of transcription factors in plant defense signaling. International Journal of Molecular Sciences 19, 3737. doi:10.3390/ijms19123737

Paulino, N., Foo, M., Kim, J., Bates, D., 2019. Robustness analysis of a nucleic acid controller for a dynamic biomolecular process using the structured singular value. Journal of Process Control 78, 34–44. doi:10.1016/j.jprocont.2019.02.009

Penfold, C.A., Wild, D.L., 2011. How to infer gene networks from expression profiles, revisited. Interface Focus 1, 857–870. doi:10.1098/rsfs.2011.0053

Rinon, J., Mendoza, R., Mendoza, V., 2019. Parameter estimation of an S-System model using hybrid genetic algorithm with the aid of sensitivity analysis. Proceedings of Philippines Computing Science Congress, Manila, Philippines, 28-30 March 2019 , 94–102

Rue, P., Garcia-Ojalvo, J., 2013. Modeling gene expression in time and space. Annual Review of Biophysics 42, 605–627. doi:10.1146/annurev-biophys-083012-130335.

Saint-Antoine, M.M., Singh, A., 2020. Network inference in systems biology: recent developments, challenges, and applications. Current Opinion in Biotechnology 63, 89–98. doi:10.1016/j.copbio.2019.12.002

Santillan, M., 2008. On the use of the Hill Functions in mathematical models of gene regulatory networks. Mathematical Modelling of Natural Phenomena 3. doi:10.1051/mmnp:2008056

Savageau, M.A., 1969. Biochemical systems analysis ii. the steady state solutions for an n-pool system using a power-law approximation. Journal of Theoretical Biology 25, 370–379. doi:10.1016/S0022-5193(69)80027-5

Savageau, M.A., 2001. Design principles for elementary gene circuits: elements, methods, and examples. Chaos 11, 142–159. doi:10.1063/1.1349892

Schlitt, T., Brazma, A., 2009. Current approaches to gene regulatory network modelling. BMC Bioinformatics 8, S9. doi:10.1186/1471-2105-8-S6-S9

Sheikholeslami, R., Razavi, S., 2017. Progressive latin hypercube sampling: an efficient approach for robust sampling-based analysis of environmental models. Environmental Modelling and Software 93, 109–126. doi:10.1016/j.envsoft.2017.03.010.

Sood, M., Kapoor, D., Kumar, V., Kalia, N., Bhardwaj, R., Sidhu, G.P., Sharma, A., 2021. Mechanisms of plant defense under pathogen stress: a review. Current Protein and Peptide Science 22, 376–395. doi:10.2174/1389203722666210125122827

Stolovitzky, G., Monroe, D., Califano, A., 2007. Dialogue on reverse engineering assessment and methods: The DREAM of high throughput pathway inference. Annals of the New York Academy of Sciences 1115, 1–22. doi:10.1196/annals.1407.021

Stolovitzky, G., Prill, R., Califano, A., 2009. Lessons from the DREAM2 challenges: a community effort to assess biological network inference. Annals of the New York Academy of Sciences 1158, 159–195. doi:10.1111/j.1749-6632.2009.04497.x

Tegner, J., Yeung, M., Hasty, J., Collins, J., 2003. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. Proceedings of the National Academy of Sciences 100, 5944–5949. doi:10.1073/pnas.0933416100

Transtrum, M., Qiu, P., 2012. Optimal experiment selection for parameter estimation in biological differential equation models. BMC Bioinformatics 13, 181. doi:10.1186/1471-2105-13-181

Turchin, P., 2003. Complex Population Dynamics. Princeton University Press.

Vilela, M., Chou, I.C., Susana, S., Tereza, A., Vasconcelos, R., Voit, E.O., Almeida, J.S., 2008. Parameter optimization in S-System models. BMC Systems Biology 2, 35. doi:10.1186/1752-0509-2-35

Villaverde, A.F., Ross, J., Banga, J.R., 2013. Reverse engineering cellular networks with information theoretic methods. Cells 2, 306–329. doi: 10.3390/cells2020306

Vinciotti, V., Augugliaro, L., Abbruzzo, A., Wit, E.C., 2016. Model selection for factorial gaussian graphical models with an application to dynamic regulatory networks. Statistical Applications in Genetics and Molecular Biology 15, 193–212. doi: 10.1515/sagmb-2014-0075

Voit, E.O., Martens, H.A., Omholt, S.W., 2015. 150 years of mass action law. PLoS Computational Biology 11, e1004012. doi:10.1371/journal.pcbi.1004012

Wagenmakers, E.J., Farrell, S., 2004. AIC model selection using Akaike weights. Psychonomin Bulletin and Review 11, 192–196. doi:10.3758/BF03206482

Wang, C.C.N., Chang, P.C., Ng, K.L., Chang, C.M., Sheu, P.C.Y., Tsai, J.J.P., 2014. A model comparison study of the flowering time regulatory network in Arabidopsis. BMC Systems Biology 8. doi:10.1186/1752-0509-8-15

Wang, H., Qian, L., Dougherty, E., 2010. Inference of gene regulatory networks using S-System: a unified approach. IET Systems Biology 4, 145–156. doi:10.1049/iet-syb.2008.0175

Weiberg, A., Wang, M., Lin, F.M., Zhao, H., Kaloshian, I., Huang, H.D., Jin, H., 2013. Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. Science 342, 118–123. doi:10.1126/science.1239705

Williamson, B., Tudzynski, B., Tudzynski, P., Van Kan, J.A.L., 2007. Botrytis cinerea: the cause of grey mould disease. Molecular Plant Pathology 8, 561–580. doi:10.1111/j.1364-3703.2007.00417.x

Windram, O., Madhou, P., McHattie, S., Hill, C., Hickman, R., Cooke, E., Jenkins, D.J., Penfold, C.A., Baxter, L., Breeze, E., Kiddle, S.J., Rhodes, J., Atwell, S., Kliebenstein, D.J., Kim, Y.S., Stegle, O., Borgwardt, K., Zhang, C., Tabrett, A., Legaie, R., Moore, J., Finkenstadt, B., Wild, D.L., Mead, A., Rand, D., Beynon, J., Ott, S., Buchanan-Wollaston, V., Denby, K.J., 2012. Arabidopsis defense against Botrytis cinerea: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. Plant Cell 24, 3530–3557. doi:10.1105/tpc.112.102046.

Youseph, A., Chetty, M., Karmakar, G., 2015. Gene regulatory network inference using Michaelis-Menten kinetics. Proceedings of IEEE Congress of Evolutionary Computation, 25-28 May 2015, Sendai, Japan, 2392–2397. doi:10.1109/CEC.2015.7257181