Scalable Approach to Data Driven Transcriptome Dynamics Modeling

Alexandr Koryachko, Samiul Haque, and Cranos Williams
Department of Electrical and Computer Engineering, NCSU
890 Oval Drive, 27606, Raleigh, USA
(akoryac, shaque2, cmwilli5)@ncsu.edu

Abstract

The evolving field of biological experimentation allows for the collection of various types of data describing different aspects of gene regulation inside a living cell. However, most of the gene expression dynamic modeling approaches limit their choice of data to a time course, which leads to infeasible requirements on the number of sampling time points to estimate the multitude of biologically relevant parameters. Thus, the model scope and parameter identifiability have to be sacrificed to approximate transciptome dynamics based on a typical number of time course samples. In this paper, we propose a scalable framework for building a model of transcriptome dynamics by aggregating a collection of experimental data available and suggest the types of additional experimentation to supplement the time course efficiently. The described approach is capable of increasing model descriptive and predictive power when additional data become available.

keywords: Gene Expression, Mathematical Modeling, Model Selection, Experimental Design, Nonlinear Dynamics, ODEs.

1 Introduction

Living organisms develop and respond to stimuli through a set of regulations on a molecular level. The regulation rules are hard-written in a genome and implemented through the action of transcription factors which modulate gene activity according to a given condition. Various types of experiments are performed to gain insight into that machinery to modify organisms in novel and strategic ways. Transcriptome abundance measurements are a widely utilized technique to estimate the change of gene activity over time or under a condition of interest. Methods of various complexity have been used to analyze the transcriptome (gene expression) data [24, 13]. Out of those methods the systems of Ordinary Differential Equations (ODEs) present the most descriptive way of representing transcriptome dynamics over time within a cell.

Ordinary Differential Equations (ODEs) gain a growing interest as a tool for modeling gene expression dynamics [24], yet a typical limitation of 2 to 8 time samples per time course [29] is still a barrier for a wide use in practical applications [2]. This limitation also leads to formulation of phenomenological models like linear models [4], Standardized Qualitative Dynamical Systems (SQUAD) models [20], or nonlinear basis functions models [8] with a small number of biologically irrelevant parameters rather than using mathematical constructs based on molecular kinetics like in S-Systems [23] or Hill-function kinetics based models [15]. Moreover, a wide range of proposed ODE structures for modeling transcriptome activity makes the choice of an appropriate mathematical representation challenging due to a lack of specific requirements for experimental data in the corresponding papers.

A number of studies have successfully applied ODEs to model transcriptome dynamics given a sufficient amount of information in terms of gene regulatory network graph and/or the results of various types of experiments which complemented the time course data [9, 4, 31]. Despite the findings facilitated by such modeling and the potential of building on previous results by collecting additional data, the cases of gradual model evolution are rather an exception than a rule. One such exception is the circadian clock effect in plants, which has been the subject of a number of ODE models [19, 18], continuously improved over time by the addition of new feedback loops [17], post-transcriptional and post-translational regulation [25], and mutant expression data [26]. In each case the addition of new data allowed for greater descriptive and predictive power [3]. However, each iteration required a reformulation of the previous model structure to incorporate new experimental results, making the process of model improvement long and not intuitive.

In this paper, we propose a methodology for dynamic model building which allows for a gradual increase in model complexity when new experimental data become available. In Section 2 we summarize the commonly used ODE structures into levels of mathematical complexity where each new level extends the

978-1-943436-11-8 / copyright ISCA, BICOB 2018
March 19-21, 2018, Las Vegas, Nevada, USA

previous one based on additional data and propose the types of experiments allowing for an efficient transition between the levels. In Section 3 we propose criteria for data sufficiency at a given level of model complexity and an algorithm for aggregating the available experimental datasets. Thus, the resulting model will represent the outcomes of relevant experiments in a set of uniquely identifiable parameters, provide insights into transcriptome properties if the parameters are biologically relevant, and allow for gene expression predictions in a wide range of conditions combinations.

2 Model Formulation

2.1 Basic Model

Gene expression can be thought of as a balance between the rate of gene transcription and the rate of the corresponding mRNA degradation. Assuming both rates constant at a steady state, one can model gene expression dynamics with the following ODE:

$$\frac{dx}{dt} = a - bx, \tag{1}$$

where x represents gene expression, a represents the transcription rate ($a > 0$), and b represents the mRNA decay rate ($b > 0$). With a steady state assumption (i.e. $dx/dt = 0$) only one gene expression measurement x_{ss} would suffice to initiate the model building process and estimate the ratio of the rates at a steady state ($a/b = x_{ss}$).

Resolving between a and b requires additional experimentation. Time course data, the most common source of information in modeling approaches, can be used for this purpose if it captures a sufficient amount of gene expression dynamics. This scenario is rarely the case due to the typical sparseness of biological data. Zak et. al. proposed solving this problem by measuring the decay rate separately [35]. Barenco et. al. obtained direct mRNA decay rate measurements to constrain the tumor suppressor transcription factor p53 model while fitting it to the time course data [1]. Decay rate values may also be available in the literature [22, 32]. However, the reported values should be used with caution since decay rates are known to be condition specific [6]. Moreover, most experimental protocols are invasive and might heavily affect cellular physiology [21].

2.2 Transcription Factor Effect

Gene regulation in a cell is modulated through the activity of transcription factors. Assuming that transcription factors affect a common target gene x independently, this modulation can be reflected in Equation (1) as follows:

$$\frac{dx}{dt} = a f_1(x_1) f_2(x_2) \cdots f_R(x_R) - bx, \tag{2}$$

where a is a scaling coefficient, x_r ($r = 1, 2, \ldots, R$) is the expression of one of R transcription factors regulating x, and $f_r(x_r)$ is the regulator influence function which is equal to 1 when no regulation occurs, greater than 1 for activators, and between 0 and 1 for inhibitors. Influence function parameter estimation is heavily affected by the ability to differentiate between regulators' expression patterns based on sparse and noisy time course samples. Additional sampling time points or replicates do not guarantee sufficient resolution improvements. Thus, time course data should be supplemented with additional information to estimate the influence coefficients. Experiments where target expression is measured while regulator expression is manipulated can reveal this information.

Regulator knock-out mutant experiments [36, 14] can uniquely define a linear approximation of the influence function $f_r(x_r) = 1 + c_r x_r$. If transcription factor x_r is an activator with a measured wild-type expression x_r^{WT}, then target gene expression measurements in wild-type (x^{WT}) and mutant (x^{M_A}) conditions allow to approximate the regulator-target dependence with a line (Figure 1A):

$$x = x^{M_A} + \frac{x^{WT} - x^{M_A}}{x_r^{WT}} x_r,$$

which leads to a constant impact factor c_r^A approximation by associating x^{M_A} with the scaling coefficient a and rewriting the dependence in a form of the linear influence function:

$$f_{r_A}^{(lin)}(x_r) = 1 + \underbrace{\frac{x^{WT} - x^{M_A}}{x_r^{WT} x^{M_A}}}_{c_r^A} x_r.$$

However, a linear construct is expected to approximate the influence in a range that does not extend far beyond the regulator's wild type gene expression value. Otherwise, unrealistically high target expression is expected in case of activators and negative expression in case of inhibitors.

Hill-function approximation [7, 15] presents another, more biologically relevant, way of representing regulator influence (Figure 1B):

$$x = x^{M_A} + \left(x^{max} - x^{M_A}\right) \frac{x_r^l}{x_r^l + K^l} = x^{M_A} \cdot f_{r_A}^{(hill)}.$$

Here the target expression value under activator's influence is bounded. The bound estimate x^{max} can

be obtained through overexpression experiments [27] if the regulator's overexpression value is at least several fold larger than x_r^{WT}. The dissociation constant K can be estimated using knock-out mutant (x^{M_A}) and overexpression (x^{max}) experiment values. The regulator's protein affinity l can be obtained through additional experiments, for example, through fluorescence correlation spectroscopy in plants [5]. Hence, each additional parameter in the regulator influence function requires an experiment to estimate it.

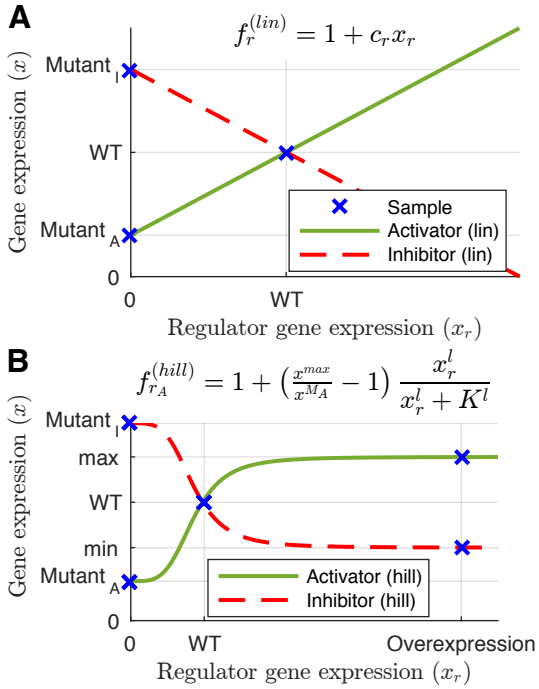

Figure 1: Influence function $f_r(x_r)$ under (A) Linear and (B) Hill-function approximation assumptions.

2.3 Condition Induced Effects

A host of transcriptome research studies are interested in mechanisms governing organism's response to a certain condition like biotic or abiotic stress in plants [12] or pathogen infection in single cells or animals [9, 34]. Equation (2) would be sufficient in describing gene expression dynamics over time under such condition if the full set of regulators is known, which is almost never the case at the current stage. Thus, the model has to account for unknown factors:

$$\frac{dx}{dt} = a f_u(t) \prod_{r=1}^{R} f_r(x_r) - bx, \tag{3}$$

where R is the number of known regulators, and $f_u(t)$ is a function aggregating the currently unknown influencing factors which change their activity under a condition of interest. An example of such influencing factor could be a change in a currently unknown condition induced transcription factor which binds to the target gene's promoter. $f_u(t)$ takes positive values and turns into 1 in wild-type conditions. The shape of $f_u(t)$ can be obtained using Gaussian process approximation [10]. Another approach would be to represent the unknown effect as a continuous shift to a new condition induced equilibrium:

$$u(t) = u_T \frac{1}{(\tau/t)^r + 1}, \tag{4}$$

where u_T represents an impact coefficient ($u_T > -1$), r quantifies how fast the transition between wild-type and condition induced steady states occurs ($r > 0$), and τ accounts for the transition delay (Figure 2). Because $u(t)$ turns to 0 when no condition is applied, an adjustment $f_u(t) = 1 + u(t)$ is needed to represent the unknown regulatory effect function. Parameters shaping $u(t)$ can be estimated by fitting the model to time course data under wild-type and the condition of interest.

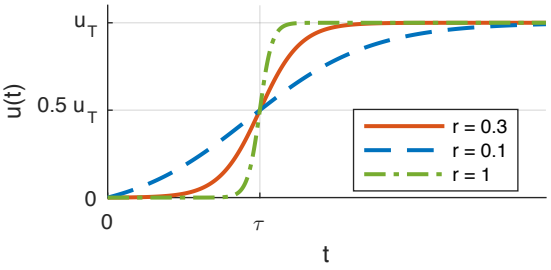

Figure 2: Sigmoid function approximation of unknown influencing factors effects. u_T – scale coefficient, r – rate coefficient , and τ – delay coefficient.

Additional experiments can help shaping the response to different levels of the applied condition if the condition levels are quantifiable and the sigmoid function is used. In such case the parameters shaping $u(t)$ are affected by the condition level S. We will concentrate on the condition dependence of the magnitude parameter u_T while the condition dependence of other two parameters from Equation (4) can be quantified in a similar fashion.

Wild-type and condition induced gene expression values allow for a linear approximation through a range of condition levels, which might, in some cases, be significantly far from reality. A more reliable approximation can be obtained by sampling gene expression at intermediate condition levels. However, transcriptome measurements are resource consuming, so the condition levels should be chosen in an efficient manner to produce maximum information with minimum experimentation. If the organism of interest exhibits a

quantifiable change in size, shape, or other easily accessible physiological parameters under the condition of interest, a faster and less expensive procedure of phenotyping can be used under a set of intermediate condition levels (e.g. micronutrient content level or pathogen load) to judge whether linear approximation captures the condition effect. Phenotyping results can also give clues on which condition levels to choose for the consequent transcriptome measurement experiments. Figure 3 shows a hypothetical example of selecting the most informative sampling point and the magnitude response function based on the results of phenotyping experiments.

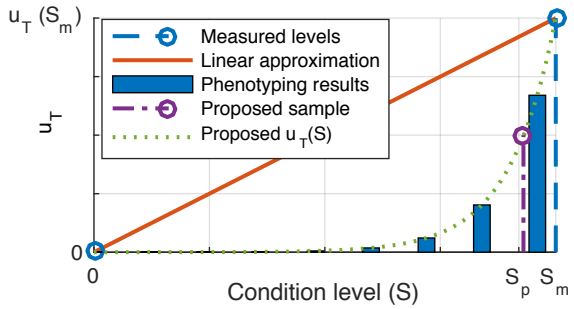

Figure 3: Condition level effect dependence modeling. S_m- measured condition level of the initial experiment, S_p-proposed condition level for gene expression measurements based on phenotyping experiments.

3 Model Fitting

Guidelines for the additional experimentation proposed in the previous section illustrate ways for increasing model complexity and, thus, its descriptive and predictive power. However, determining whether the collected data is sufficient for a given model complexity and combining various types of datasets to train the model are not trivial tasks considering that each type of experimentation has an associated measurement noise. We propose parameter identifiability analysis as a criteria for data sufficiency when scaling a model up and Bayesian inference for model parameter optimization.

3.1 Model Scalability Assessment

We require all parameters to be uniquely identifiable in the scaled up model representation to accept it. A parameter is considered non-identifiable if any deviation in its value produce an equally good model fit through the corresponding adjustments in other parameters. For example, any value of decay rate b can be compensated with a corresponding value of the transcription rate a in Equation (1) if x_{ss} is the only non wild type measurement at hand. Thus, parameter non-identifiability indicates a lack of data support for a given model structure. Several methods such as Differential Algebra Identifiability of Systems (DAISY) [30], Exact Arithmetic Rank (EAR) [11], and Profile Likelihood (PL) [28] have been used to detect non-identifiable parameters. Among these methods, PL is the only one that relies on experimental data in its identifiability analysis. We propose using the results of PL analysis for model discrimination when increasing model complexity based on additional data.

3.2 Parameter Estimation

Bayesian inference methods aggregate different sources of data by shaping prior distributions of the corresponding parameters before fitting a model to the corresponding time course. Each experiment that we proposed allows for obtaining mean and standard deviation estimates for a specific parameter. An assumption on experimental error distribution (e.g. Gaussian or Poisson) would allow to construct the corresponding prior distribution. The model fitting algorithm will sample parameter values from the corresponding prior distributions while minimizing the sum of squared differences between the time course data and gene expression pattern produced by the model. Parameter values from the regions that are far from the experimental measurements are highly unlikely to be sampled which would ensure that the model describes both the time course data and the results of the additional experiments.

Due to a nonlinear nature of the differential equations governing gene expression dynamics we suggest using the latest generation of Bayesian inference based parameter estimation algorithm, namely Differential Evolution Adaptive Metropolis (DREAM) software package [33]. It has been demonstrated that DREAM outperforms similar software in nonlinear, multimodal, and high dimensional problems [16].

4 Conclusion

In this paper, we presented a methodology for sequential increase in gene expression dynamic model complexity by aggregating different types of experimental data. The methodology provides a flexible framework for accumulating the existing knowledge of a biological process of interest at the transcriptome level and proposes efficient ways for expanding this knowledge through additional experimentation. This paper aims to facilitate modeling efforts in the studies where time course experiments have been implemented and the key regulatory connections have been identified.

References

[1] Martino Barenco, Daniela Tomescu, Daniel Brewer, Robin Callard, Jaroslav Stark, and Michael Hubank. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome biology*, 7(3):R25, 2006.

[2] Daniel Brewer, Martino Barenco, Robin Callard, Michael Hubank, and Jaroslav Stark. Fitting ordinary differential equations to short time course data. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 366(1865):519–544, 2008.

[3] Nora Bujdoso and Seth J Davis. Mathematical modeling of an oscillating gene circuit to unravel the circadian clock network of arabidopsis thaliana. *Frontiers in Plant Science*, 4, 2013.

[4] Javier Carrera, Guillermo Rodrigo, Alfonso Jaramillo, Santiago F Elena, et al. Reverse-engineering the arabidopsis thaliana transcriptional network under changing environmental conditions. *Genome Biol*, 10(9):R96, 2009.

[5] Natalie M Clark, Elizabeth Hinde, Cara M Winter, Adam P Fisher, Giuseppe Crosti, Ikram Blilou, Enrico Gratton, Philip N Benfey, and Rosangela Sozzani. Tracking transcription factor mobility and interaction in arabidopsis roots with fluorescence correlation spectroscopy. *Elife*, 5:e14770, 2016.

[6] Nicole L Garneau, Jeffrey Wilusz, and Carol J Wilusz. The highways and byways of mrna decay. *Nature reviews Molecular cell biology*, 8(2):113–126, 2007.

[7] Rudolf Gesztelyi, Judit Zsuga, Adam Kemeny-Beke, Balazs Varga, Bela Juhasz, and Arpad Tosaki. The hill equation and the origin of quantitative pharmacology. *Archive for history of exact sciences*, pages 427–438, 2012.

[8] Mika Gustafsson, Michael Hörnquist, Jesper Lundström, Johan Björkegren, and Jesper Tegnér. Reverse engineering of gene networks with LASSO and nonlinear basis functions. *Annals of the New York Academy of Sciences*, 1158(1):265–275, 2009.

[9] Reinhard Guthke, Ulrich Möller, Martin Hoffmann, Frank Thies, and Susanne Töpfer. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics*, 21(8):1626–1634, 2005.

[10] Ruirui Ji, Xinxin Zhang, and Xiaomei Yan. Modelling transcriptional regulation with fractional order differential equation using gaussian process. In *Control Conference (CCC), 2016 35th Chinese*, pages 9366–9370. IEEE, 2016.

[11] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946, 2012.

[12] Joachim Kilian, Dion Whitehead, Jakub Horak, Dierk Wanke, Stefan Weinl, Oliver Batistic, Cecilia D'Angelo, Erich Bornberg-Bauer, Jörg Kudla, and Klaus Harter. The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *The Plant Journal*, 50(2):347–363, 2007.

[13] Alexandr Koryachko, Anna Matthiadis, Joel J Ducoste, James Tuck, Terri A Long, and Cranos Williams. Computational approaches to identify regulators of plant stress response using high-throughput gene expression data. *Current Plant Biology*, 3:20–29, 2015.

[14] Alexandr Koryachko, Anna Matthiadis, Durreshahwar Muhammad, Jessica Foret, Siobhan M Brady, Joel J Ducoste, James Tuck, Terri A Long, and Cranos Williams. Clustering and differential alignment algorithm: Identification of early stage regulators in the arabidopsis thaliana iron deficiency response. *PloS one*, 10(8):e0136591, 2015.

[15] Jan Krumsiek, Sebastian Pölsterl, Dominik Wittmann, and Fabian Theis. Odefy-from discrete to continuous models. *BMC bioinformatics*, 11(1):233, 2010.

[16] Eric Laloy and Jasper A Vrugt. High-dimensional posterior exploration of hydrologic models using multiple-try dream (zs) and high-performance computing. *Water Resources Research*, 48(1), 2012.

[17] James CW Locke, László Kozma-Bognár, Peter D Gould, Balázs Fehér, Eva Kevei, Ferenc Nagy, Matthew S Turner, Anthony Hall, and Andrew J Millar. Experimental validation of a predicted feedback loop in the multi-oscillator clock of arabidopsis thaliana. *Molecular systems biology*, 2(1):59, 2006.

[18] James CW Locke, Megan M Southern, László Kozma-Bognár, Victoria Hibberd, Paul E Brown, Matthew S Turner, and Andrew J Millar.

Extension of a genetic network model by iterative experimentation and mathematical analysis. *Molecular systems biology*, 1(1), 2005.

[19] JCW Locke, AJ Millar, and MS Turner. Modelling genetic networks with noisy and varied experimental data: the circadian clock in arabidopsis thaliana. *Journal of theoretical biology*, 234(3):383–393, 2005.

[20] Luis Mendoza and Ioannis Xenarios. A method for the generation of standardized qualitative dynamical systems of regulatory networks. *Theoretical Biology and Medical Modelling*, 3:13, 2006.

[21] Sarah E Munchel, Ryan K Shultzaberger, Naoki Takizawa, and Karsten Weis. Dynamic profiling of mrna turnover reveals gene-specific and system-wide regulation of mrna decay. *Molecular biology of the cell*, 22(15):2787–2795, 2011.

[22] Reena Narsai, Katharine A Howell, A Harvey Millar, Nicholas O'Toole, Ian Small, and James Whelan. Genome-wide analysis of mrna decay rates and their determinants in arabidopsis thaliana. *The Plant Cell Online*, 19(11):3418–3436, 2007.

[23] Leon Palafox, Nasimul Noman, and Hitoshi Iba. Reverse engineering of gene regulatory networks using dissipative particle swarm optimization. *Evolutionary Computation, IEEE Transactions on*, 17(4):577–587, 2013.

[24] Chanda Panse, Dr Kshirsagar, et al. Survey on modelling methods applicable to gene regulatory network. *arXiv preprint arXiv:1310.2361*, 2013.

[25] Alexandra Pokhilko, Sarah K Hodge, Kevin Stratford, Kirsten Knox, Kieron D Edwards, Adrian W Thomson, Takeshi Mizuno, and Andrew J Millar. Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. *Molecular Systems Biology*, 6(1), 2010.

[26] Alexandra Pokhilko, Paloma Mas, and Andrew J Millar. Modelling the widespread effects of toc1 signalling on the plant circadian clock and its outputs. *BMC systems biology*, 7(1):23, 2013.

[27] Gregory Prelich. Gene overexpression: uses, mechanisms, and interpretation. *Genetics*, 190(3):841–854, 2012.

[28] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[29] Bruce A Rosa, Ji Zhang, Ian T Major, Wensheng Qin, and Jin Chen. Optimal timepoint sampling in high-throughput gene expression experiments. *Bioinformatics*, 28(21):2773–2781, 2012.

[30] Maria Pia Saccomani, Stefania Audoly, and Leontina D'Angiò. Parameter identifiability of nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.

[31] Yara-Elena Sanchez-Corrales, Elena R Alvarez-Buylla, and Luis Mendoza. The arabidopsis thaliana flower organ specification gene regulatory network determines a robust differentiation process. *Journal of theoretical biology*, 264(3):971–983, 2010.

[32] Kate Sidaway-Lee, Maria J Costa, David A Rand, Bärbel Finkenstadt, and Steven Penfield. Direct measurement of transcription rates reveals multiple mechanisms for configuration of the arabidopsis ambient temperature response. *Genome biology*, 15(3):R45, 2014.

[33] Jasper A Vrugt, CJF Ter Braak, CGH Diks, Bruce A Robinson, James M Hyman, and Dave Higdon. Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation*, 10(3):273–290, 2009.

[34] Shuang Wu, Zhi-Ping Liu, Xing Qiu, and Hulin Wu. Modeling genome-wide dynamic regulatory network in mouse lungs with influenza infection using high-dimensional ordinary differential equations. *PloS one*, 9(5):e95276, 2014.

[35] Daniel E Zak, Gregory E Gonye, James S Schwaber, and Francis J Doyle. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in-silico network. *Genome research*, 13(11):2396–2405, 2003.

[36] Jie Zhang, Bing Liu, Mengshu Li, Dongru Feng, Honglei Jin, Peng Wang, Jun Liu, Feng Xiong, Jinfa Wang, and Hong-Bin Wang. The bhlh transcription factor bhlh104 interacts with iaa-leucine resistant3 and modulates iron homeostasis in arabidopsis. *The Plant Cell*, 27(3):787–805, 2015.