RESEARCH

A fast validation test of gene regulatory network models via the Fokker-Planck equation

Natalia López-Paleta[1] · Eduardo Moreno-Barbosa[1] · Jorge Velázquez-Castro[1]

Abstract

Since Waddington proposed the concept of the "epigenetic landscape" in 1957, researchers have developed various methodologies to represent it in diverse processes. Studying the epigenetic landscape provides valuable qualitative information regarding cell development and the stability of phenotypic and morphogenetic patterns. Although Waddington's original idea was a visual metaphor, a contemporary perspective relates it to the landscape formed by the basins of attraction of a dynamical system describing the temporal evolution of protein concentrations driven by a gene regulatory network. Transitions among these attractors can be driven by stochastic perturbations, with the cell state more likely to transition to the nearest attractor or to the one that presents the path of least resistance. In this study, we define the epigenetic landscape using the free energy potential obtained from the solution of the Fokker-Planck equation on the regulatory network. Specifically, we obtained a numerical approximate solution of the Fokker-Planck equation describing the *Arabidopsis thaliana* flower morphogenesis process. We observed good agreement between the coexpression matrix obtained from the Fokker-Planck equation and the experimental coexpression matrix. This paper proposes a method for obtaining this landscape by solving the Fokker-Planck equation (FPE) associated with a dynamical system describing the temporal evolution of protein concentrations involved in the process of interest. As these systems are high-dimensional and analytical solutions are often unfeasible, we propose a gamma mixture model to solve the FPE, transforming this problem into an optimization problem. This methodology can enhance the analysis of gene regulatory networks by directly relating theoretical mathematical models with experimental observations of coexpression matrices, thus providing a discriminating technique for competing models.

Keywords Gene regulatory network · Dynamical model · Fokker-Planck equation · Microarray

✉ Jorge Velázquez-Castro
jorge.velazquezcastro@correo.buap.mx

Natalia López-Paleta
nlpaleta23@gmail.com

Eduardo Moreno-Barbosa
emoreno@fcfm.buap.mx

[1] Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla, Puebla 72570, Puebla, México

1 Introduction

In recent years, dynamic systems theory has been able to make inferences about the phenotypic characteristics of some species from gene regulatory models [1–7]. Establishing a relationship between these phenotypic characteristics and genetic expression will enable the prediction of phenotypic effects due to changes in gene network reconfiguration or alterations in expression factors. Thus, establishing this relationship is a fundamental problem of systems biology and will facilitate the design and evaluation of particular and targeted therapies based on control over certain expression factors. To achieve this goal, a first problem is determining the regulatory network created by the interaction between the different genes and proteins in the cell and its dynamics.

Knowledge of the structure of a particular genetic regulation network (GRN) allows the determination of the state of health and function of cells. Therefore, it can be used as a diagnostic tool. Several methods have been developed to infer GRN from gene expression data [8–18]. In the last few years, also several machine learning techniques have been developed [19–23] to infer GNRs. Each methodology can provide different results owing to the mathematical assumptions used to reconstruct the interaction network from the data. Thus, this is an active area of research, and there is no consensus on a standard methodology for inferring GRNs [24]. Its even known that most methods provide an approximate picture of the underlying network [24]. Despite these shortcomings, there is a large body of knowledge of GRNs for different systems. Thus, it is convenient to evaluate the predictions of a network and compare them with the observed data. Gene expression databases [25, 26] are currently used to identify correlations between gene activation. Conversely, the properties and behavior of dynamical gene regulatory networks have been studied using theoretical models [27–31]. Two types of dynamic models are most frequently used: Boolean and continuous models. Boolean models represent the system state, with 1 s and 0 s representing the activation or non-activation of each gene involved. The discrete-time evolution rules determine the dynamics. Continuous models provide a framework for describing the temporal dynamics of gene product concentrations in a continuous-time context. These models are underpinned by the principles governing reaction rates among various biological components. We propose that the contemporary interpretation of the epigenetic landscape serves as a critical nexus connecting empirical gene expression data with theoretical dynamical models of gene regulatory networks (GRNs). This integration may enhance our comprehension of the intricate regulatory mechanisms governing gene expression and its broader biological implications.

Morphogenetic evolution follows from the complex interactions between genes in the network and was represented by C.H. Waddington in 1957 for what he termed the epigenetic landscape [32]. Waddington proposed that the morphogenesis process can be represented metaphorically as a ball rolling downhill through a landscape of mountains and valleys. Currently, efforts are being made to formalize the concept of the epigenetic landscape as an analogy of Lyapunov functions or energy potentials. The dynamic system attractors are at the bottom of the basins, and the steady state in each attractor characterizes each type of cell or its phenotypic condition [33].

These systems have as many dimensions as genes involved in the network; therefore, for realistic systems, the number of dimensions is high. The high dimensionality of these systems has presented challenges in analyzing this type of phenomenon and deducing the epigenetic landscape [34]. Another challenge is the inherently stochastic nature of these systems, which necessitates the development of efficient techniques for finding solutions. Recently, a methodology was proposed to tackle this problem effectively [35, 36], achieving

good results. This work aims to find a stationary probability distribution of concentrations, thus complementing the previously mentioned studies.

This study proposes a protocol to obtain a representation of the epigenetic landscape of a gene regulatory network described by a continuous dynamical system of protein concentrations. Specifically, the research aims to elucidate the epigenetic landscape associated with *Arabidopsis thaliana*'s flower morphogenesis process. The genetic regulation network that describes this process is already known, comprising 12 nodes representing the genes involved in the process, and the discrete dynamics describing their relationships are already well studied [37, 38].

A continuous time model for the *Arabidopsis thaliana*'s flower morphogenesis is proposed to generalize the discrete dynamics presented in [37, 38]. The Fokker-Planck equation (FPE) associated with the model is constructed, and a gamma mixture model is utilized to estimate its stationary solution. The stationary solution is subsequently employed to evaluate correlations in the genetic expressions that should be observed in an experimental setting [38, 39]. The free energy related to the stationary solution of the FPE is identified as the epigenetic landscape. This methodology will facilitate establishing a clear relationship between experimental data and theory, enabling the discrimination of different model proposals and thus allowing for the inference of a theoretical model from experimental results (see Fig. 1). Moreover, it would constitute a significant advancement towards inferring the phenotypic consequences resulting from changes in the genotype through theoretical means.

2 Methods

2.1 Gene Regulatory Network

Here, we will work with a well-studied GRN for which experimental data on gene coexpression are available. The data will allow us to test the building process of the model and its solution procedure. We construct a minimal continues model for the *Arabidopsis thaliana*'s flower morphogenesis from minimal information of the interaction between involved genes. The initial information is if a particular gene promotes or inhibits the activation of other genes, but a quantitative measure of the intensity of this interaction is unknown. A GRN that describes the qualitative relation between genes involved in AT flower morphology is proposed in [37]. In this network, each node represents a gene that takes part in such a developmental process. The name and order of those genes are shown in Table 1, and the network is shown in Fig. 2. With this limited information, proposing a Boolean model [40] is the first step to translating these gene relations into a GRN model. Mendoza and Állvarez Buylla presented a Boolean model representing the temporal evolution of this system, which is given by

$$x_i(t+1) = \mathbf{H}\left(\sum_{j=1}^{N} w_{ij} x_j(t) - \theta_i\right),\tag{1}$$

where $\mathbf{H}$ is a step function defined by

$$\mathbf{H}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}\tag{2}$$

where w_{ij} equals the intensity of the interaction of the j-th over the i-th gene.

Those gene states can be active or inactive and are represented by values 1 and 0, respectively. So $x_i = 1$ if the weighted sum of all genes that regulate it exceeds a threshold value θ_i.
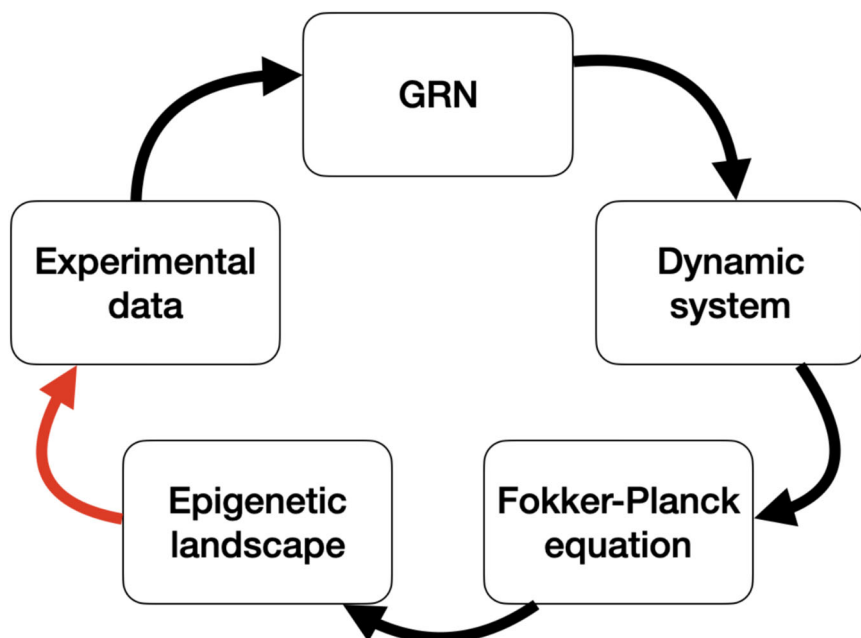
Fig. 1 Generally, the study of cell development processes starts from the experimental data available to create a GRN. Subsequently, a GRN is proposed based on data. A dynamic system can be proposed based on the GRN. A FPE can be derived from the dynamic system. The epigenetic landscape can be associated with the solution of FPE. The epigenetic landscape can be used to find simulated experimental data from the model. This data can be used to validate or discriminate the model

In addition, in [37], a genetic search algorithm [41] was used to find integer values for w_{ij} and θ_i that lead to at least four stationary states corresponding to 4 phenotypic stages in the developmental process of the *Arabidopsis thaliana*'s flower [42]. The weight matrix W and the threshold vector θ, whose entries correspond to the parameters w_{ij} and θ_i, were found to be

$$
W = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-2 & -1 & 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 5 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & -2 & 1 & -2 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 & -2 \\
0 & 0 & 4 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\qquad
\theta = \begin{bmatrix}
0 \\ 0 \\ 3 \\ -1 \\ 1 \\ 0 \\ 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0
\end{bmatrix}
$$

Table 1 List of the genes involved in the *Arabidopsis thaliana* morphogenesis

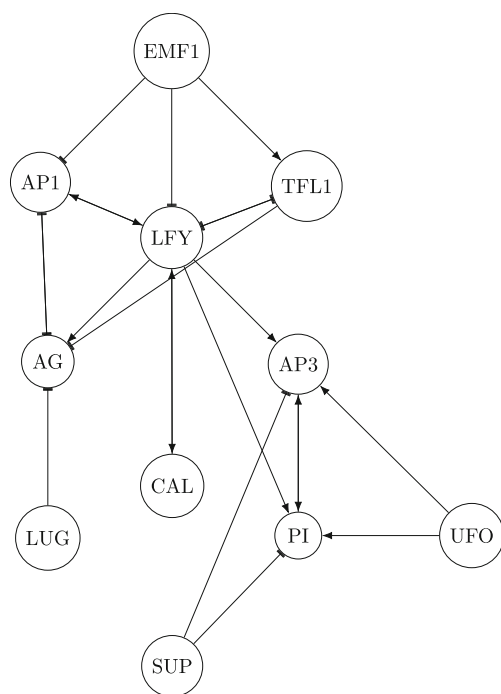| 1	EMF1	3	LFY	5	CAL	7	UFO	9	AG	11	PI
2	TFL1	4	AP1	6	LUG	8	BFU	10	AP3	12	SUP

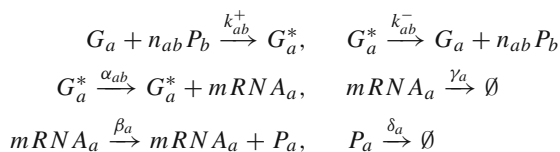**Fig. 2** Gene regulatory network of *Arabidopsis thaliana* morphogenesis

In this matrix, the negative signs indicate the repression of the i-th row gene due to the j-th column gene.

2.2 Continuous model

From the discrete time GRN dynamics (1), a continuous time model describing the protein concentrations involved in the morphogenesis of the AT flower was built. The model is based on a system of ordinary differential equations describing the evolution of the protein concentrations involved in the developmental process.

Each node of the GRN represents a gene, and we assumed that each gene G_a is activated by protein P_b, then it transcribes mRNA, and in turn, a corresponding protein P_a is translated. That is to say, for each one of the 12 nodes from the GNR, two differential equations are proposed. One of these equations will describe the temporal evolution of the protein concentration associated with each gene, and the other the transcribed mRNA.

The reaction scheme assumed for transcription of mRNA is given by
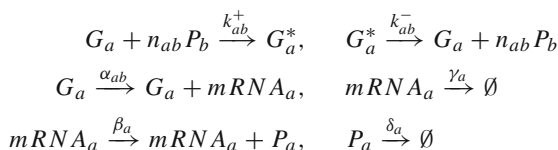
$$G_a + n_{ab} P_b \xrightarrow{k_{ab}^+} G_a^*, \qquad G_a^* \xrightarrow{k_{ab}^-} G_a + n_{ab} P_b$$

$$G_a^* \xrightarrow{\alpha_{ab}} G_a^* + mRNA_a, \qquad mRNA_a \xrightarrow{\gamma_a} \emptyset$$

$$mRNA_a \xrightarrow{\beta_a} mRNA_a + P_a, \qquad P_a \xrightarrow{\delta_a} \emptyset$$

The first two reactions are faster than the others, then active gene G_a^* concentration can be approximated at its quasi-steady state value. Then, the differential equations for the mRNA concentration m_a and protein P_a concentration are

$$\frac{dm_a}{dt} = \frac{\alpha_{ab} k_{ab} p_b^{n_{ab}}}{1 + k_{ab} p_b^{n_{ab}}} - \gamma_a m_a \tag{3}$$

$$\frac{dp_a}{dt} = \beta_a m_a - \delta_a p_a \tag{4}$$

where $k_{ab} = k_{ab}^+/k_{ab}^-$. These equations describe the activation of the protein (A) production by means of the corresponding mRNA by the concentration of protein B.

On the other hand, the basic reaction scheme of a repressor is

$$G_a + n_{ab} P_b \xrightarrow{k_{ab}^+} G_a^*, \qquad G_a^* \xrightarrow{k_{ab}^-} G_a + n_{ab} P_b$$

$$G_a \xrightarrow{\alpha_{ab}} G_a + mRNA_a, \qquad mRNA_a \xrightarrow{\gamma_a} \emptyset$$

$$mRNA_a \xrightarrow{\beta_a} mRNA_a + P_a, \qquad P_a \xrightarrow{\delta_a} \emptyset$$

At the quasi-steady state of the first two reactions, the corresponding differential equations describing the protein P_a and $mRNA$ concentrations are

$$\frac{dm_a}{dt} = \frac{\alpha_{ab}}{1 + k_{ab} p_b^{n_{ab}}} - \gamma_a m_a \tag{5}$$

$$\frac{dp_a}{dt} = \beta_a m_a - \delta_a p_a \tag{6}$$

These equations describe the repression of a protein (A) production via the corresponding mRNA by a protein B.

In these equations, p_a is the concentration of protein A, m_a is the concentration of mRNA associated with protein A, p_b is the concentration of protein B, which regulates protein A production, and m_b is the concentration of mRNA associated with protein B. Parameters α_{ab}, k_{ab}, and n_{ab} describe the way protein B concentration alters protein A concentration, while γ_a and δ_a are the mRNA and protein degradation rate, respectively.

This approach would yield two equations for each node in the GRN: one for the protein and one for the corresponding mRNA. However, to reduce the system's dimensionality, an approximation is made based on the difference between the time scales of both processes. Specifically, mRNA production rapidly reaches its stationary state while protein production continues to evolve. Namely, we assume $\frac{dm}{dt} = 0$ and so

$$0 = \frac{\alpha_{ab} k_{ab} p_b^{n_{ab}}}{1 + k_{ab} p_b^{n_{ab}}} + \alpha_{a0} - \gamma_a m_a \tag{7}$$

Therefore,

$$\Rightarrow m_a = \frac{\alpha_{ab} k_{ab} p_b^{n_{ab}}}{\gamma_a (1 + k_{ab} p_b^{n_{ab}})} + \frac{\alpha_{a0}}{\gamma_a} \tag{8}$$

Substituting in the equation for the protein production,

$$\frac{dp_a}{dt} = \frac{\beta_a}{\gamma_a} \frac{\alpha_{ab} k_{ab} p_b^{n_{ab}}}{1 + k_{ab} p_b^{n_{ab}}} + \frac{\beta_a \alpha_{a0}}{\gamma_a} - \delta_a p_a \tag{9}$$

Similarly, for the protein repression equation,

$$\frac{dp_a}{dt} = \frac{\beta_a}{\gamma_a} \frac{\alpha_{ab}}{1 + k_{ab} p_b^{n_{ab}}} + \frac{\beta_a \alpha_{a0}}{\gamma_a} - \delta_a p_a \tag{10}$$

We now have an equation for each one of the 12 proteins involved in the morphogenesis process.

We now need to relate the Boolean model parameters w_{ij} and θ_i to the parameters of the reaction schemes for a repressor. w_{ij} corresponds to the sensitivity to activation or repression of gene i from another gene j. It is natural to associate it with the activation rate k_{ij}^+. If w_{ij} is positive, then it indicates that j activates i, and if w_{ij} is negative, then j represses i. On the other hand, θ is described as the threshold value of the interaction from other genes to be activated. Because the activation rate must be greater than the inactivation rate of the gene to reach its activation state, it is possible to relate the absolute value of theta with the inactivation rate k_{ij}^- and its sign indicating if gene i has a basal synthesis rate corresponding for negative θ or if positive, there is not basal synthesis rate. As there are some values of θ equal to 0 and the deactivation rate should be greater than 0, we assume the deactivation rate has a basal value plus the absolute value of θ. Due to the lack of more information in the Boolean model and to avoid introducing fictitious information, we choose a parsimonious model, setting all other parameters to 1.

Another issue that must be addressed is that equations (9) and (10) describe a protein B modifying the production of a protein A; however, the production of a protein may depend on more than one protein. Thus, we must use a more general Hill function that considers the joint action of multiple proteins.

Thus, we employ the Hill functions

$$\frac{dp_i}{dt} = \frac{a}{a + \sum_j k_{i,j} p_j^n} \tag{11}$$

and

$$\frac{dp_i}{dt} = \frac{\sum_j k_{i,j} p_j^n}{a + \sum_j k_{i,j} p_j^n} \tag{12}$$

for the protein concentration repression and activation, respectively [43].

On the other hand, the a parameters, called the *activation coefficients*, can be interpreted as the j-th protein concentration needed to activate the i-th protein production; therefore, they are analogous to θ parameters derived in [37].

Based on the above, we propose an ODEs system representing the continuous dynamics of the proteins concentration involved in the morphogenesis of the AT flower shown in (13).

$$\frac{dp_1}{dt} = \alpha_{1,0} - \delta_1 p_1$$

$$\frac{dp_2}{dt} = \frac{1 + |\theta_2| + |w_{2,1}| p_1}{1 + |\theta_2| + |w_{2,1}| p_1 + |w_{2,3}| p_3} + \alpha_{2,0} - \delta_2 p_2$$

$$\frac{dp_3}{dt} = \frac{|w_{3,4}| p_4 + |w_{3,5}| p_5}{1 + |\theta_3| + |w_{3,1}| p_1 + |w_{3,2}| p_2 + |w_{3,4}| p_4 + |w_{3,5}| p_5} + \alpha_{3,0} - \delta_3 p_3$$

$$\frac{dp_4}{dt} = \frac{1 + |\theta_4| + w_{4,3}|p_3}{1 + |\theta_4| + |w_{4,1}|p_1 + |w_{4,3}|p_3 + |w_{4,9}|p_9} + \alpha_{4,0} - \delta_4 p_4$$

$$\frac{dp_5}{dt} = \frac{|w_{5,3}|p_3}{1 + |\theta_5| + |w_{5,3}|p_3} + \alpha_{5,0} - \delta_5 p_5$$

$$\frac{dp_6}{dt} = \alpha_{6,0} - \delta_6 p_6$$

$$\frac{dp_7}{dt} = \alpha_{7,0} - \delta_7 p_7$$

$$\frac{dp_8}{dt} = \frac{|w_{8,10}|p_{10} + |w_{8,11}|p_{11}}{1 + |\theta_8| + |w_{8,10}|p_{10} + |w_{8,11}|p_{11}} + \alpha_{8,0} - \delta_8 p_8$$

$$\frac{dp_9}{dt} = \frac{1 + |\theta_9| + |w_{9,3}|p_3}{1 + |\theta_9| + |w_{9,2}|p_2 + |w_{9,3}|p_3 + |w_{9,4}|p_4 + |w_{9,6}|p_6} + \alpha_{9,0} - \delta_9 p_9$$

$$\frac{dp_{10}}{dt} = \frac{1 + |\theta_{10}| + |w_{10,3}|p_3 + |w_{10,7}|p_7 + |w_{10,8}|p_8}{1 + |\theta_{10}| + |w_{10,3}|p_3 + |w_{10,7}|p_7 + |w_{10,8}|p_8 + |w_{10,12}|p_{12}} + \alpha_{10,0} - \delta_{10} p_{10}$$

$$\frac{dp_{11}}{dt} = \frac{1 + |\theta_{11}| + |w_{11,3}|p_3 + |w_{11,7}|p_7 + |w_{11,8}|p_8}{1 + |\theta_{11}| + |w_{11,3}|p_3 + |w_{11,7}|p_7 + |w_{11,8}|p_8 + |w_{11,12}|p_{12}} + \alpha_{11,0} - \delta_{11} p_{11}$$

$$\frac{dp_{12}}{dt} = \alpha_{12,0} - \delta_{12} p_{12} \tag{13}$$

2.3 Experimental data

Arabidopsis thaliana serves as a model organism, and experimental data regarding its processes are abundant and readily available. In particular, the work of Obayashi et al. [44] provides access to condition-independent coexpression data derived from publicly available RNA-seq and microarray data.

Each microarray analysis offers information about coexpression levels for thousands of genes simultaneously. Projects as in [44] contain large collections of microarray data, which allows us to access information about changes in transcript levels in these datasets even when such microarray was not intended for the same purpose.

It is possible to search at the project's site, coexpression data for each pair of genes of interest, and then calculate the Pearson correlation coefficient between them, given by

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

This was carried out for each pair of genes involved in the process of interest in Table 1. These correlation values were arranged in matrix M_e, as illustrated in Fig. 3, and will be compared with the results obtained from the theoretical model.

2.4 Solution to the Fokker-Planck equation

The variations in measurements of coexpression levels result from the system being influenced by many other variables and factors not explicitly taken into account. Thus, we must find a solution to the associated stochastic model with extrinsic fluctuations to obtain inferred correlations between protein concentrations from the theoretical model. Furthermore, stationarity can be assumed based on the experimental data.

Recall the multivariate chemical Langevin equation [45]

$$dp_i = A_i(\mathbf{p}, t)dt + \sum_j^d c_{ij}(\mathbf{p}, t)\eta_j(t)dt \tag{14}$$

that describes the changes in the system states through time with fluctuations. For system, (13) can write $A_i(\mathbf{p}, t) = B_i(\mathbf{p}, t) - \delta_i p_i$, where B_i represent the positive part of the equations. Also, if η_j are all independent identical Gaussian white noise processes, then $\sum_j^d c_{ij}(\mathbf{p}, t)\eta_j(t) = \frac{1}{\sqrt{\Omega}}(\sqrt{B_i(\mathbf{p}, t) + \delta_i p_i})\eta$. Here, η is a Gaussian white noise process identical to η_i, and Ω is the system size [46–48]. Thus, the chemical Langevin equation for the system (13) can be expressed as

$$dp_i = (B_i(\mathbf{p}, t) - \delta_i p_i)dt + \frac{1}{\sqrt{\Omega}}(\sqrt{B_i(\mathbf{p}, t) + \delta_i p_i})\eta dt \tag{15}$$

This set of equations is equivalent to the Fokker-Planck equation (FPE) [45, 49]

$$\frac{\partial P}{\partial t} = -\sum_j \frac{\partial}{\partial p_j}\left[(B_j(\mathbf{p}, t) - \delta_j p_j)P\right] + \frac{1}{2}\sum_{j,k} \frac{\partial^2}{\partial p_j \partial p_k}(\Gamma_j(\mathbf{p}, t)P) \tag{16}$$

where $\Gamma_j(\mathbf{p}, t) = \frac{1}{\Omega}(B_i(\mathbf{p}, t) + \delta_i p_i)$. It is worth noting that we have used the Ito interpretation of integration to go from (15) to (16); thus, in this representation, the drift term is not affected by the noise term Γ_j.

To find the epigenetic landscape of the system of interest, we solve the multivariate FPE (16) associated with the dynamics given by the proposed ODE system. This equation describes the probability density function of temporal evolution for our system. So, by solving it, we obtain the probability value P for each of the points in the state space of the system $\mathbf{y}$, after a given time t. This equation is related to the epigenetic landscape since, in its stationary version, the local maxima of the solution, which are the most probable states, correspond to the attractors in the energy potential and vice versa.

In the same spirit of previous results about the type of distributions a GRN presents [35, 50], we propose a gamma mixture model as the solution of the stationary FPE

$$\hat{\mathcal{P}} = \sum_{i=1}^n A_i P_i \tag{17}$$

with

$$P_i = \prod_j^{12} \frac{\beta_{ij}^{\alpha_{ij}}}{\Gamma(\alpha_{ij})^{\alpha_i}} p_j^{\alpha_{ij}-1} \exp(-\beta_{ij}p_j).$$

Here, n is the number of stationary states, the α_{ij}'s and β_{ij}'s are the shapes and rates of each gamma distribution, and $p'_j s$ are the concentrations of the different proteins.

Expression (17) is an ansatz for the solution to the stationary FPE

$$0 = -\sum_j \frac{\partial}{\partial p_j}\left[(B_j(\mathbf{p}, t) - \delta_j p_j)P\right] + \frac{1}{2}\sum_{j,k} \frac{\partial^2}{\partial p_j \partial p_k}(\Gamma_j(\mathbf{p}, t)P) \tag{18}$$

We use the least square weighted residual method [51] to find the parameters n, A_i, $y_{j,0}$, and $\sigma_{i,j}$ that solve equation (18); in this way, the problem of solving the stationary FPE becomes an optimization problem. That is, we seek to minimize

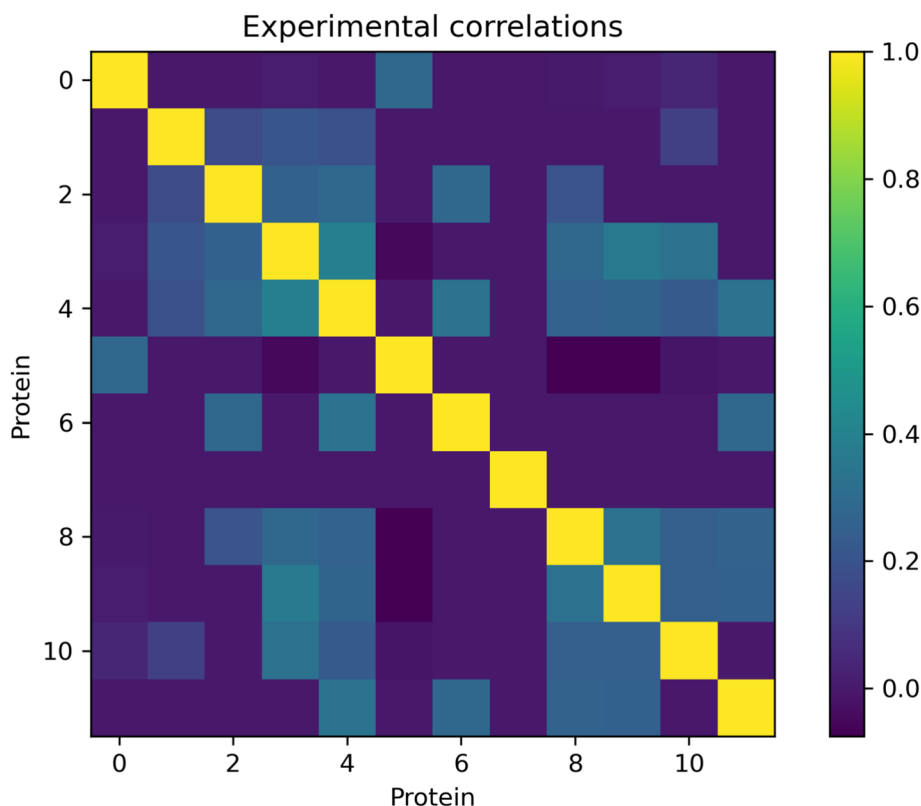$$\int_D \|R(\mathbf{p})\|dD \tag{19}$$

Fig. 3 Correlation matrix obtained from experimental data. Missing data is set to zero

where $R(\mathbf{p})$ is the residual in the equation

$$R(\mathbf{p}) = -\sum_{j} \frac{\partial}{\partial p_j} \left[\left(B_j(\mathbf{p}, t) - \delta_j p_j \right) \hat{P} \right] + \frac{1}{2} \sum_{j,k} \frac{\partial^2}{\partial p_j \partial p_k} (\Gamma_j(\mathbf{p}, t) \hat{P}). \qquad (20)$$

The integral (19) is solved using Monte Carlo quadrature/collocation [52] with $N = 10^6$ collocation points. The quadrature is computed each time a new set of parameters requires its evaluation by the algorithm that seeks its minimization.

Due to the relatively slow decay of the gamma distribution's tails, calculating the integral (19) can lead to wide variations across samples. This occurs because fluctuations in the sampling points can be considerable within a 12-dimensional space. To mitigate this problem, a Gaussian mixture model was initially used as a proxy distribution. This distribution helps identify regions close to the dynamical system's attractors and the variances of the fluctuations around them (Fig. 4). The centers and variances of the normal distributions within the Gaussian mixture model can then be used as a starting point for fine-tuning the gamma mixture model. The model fitting was performed in three stages: (1) a broad parameter space search using Monte Carlo methods with a Gaussian mixture model to locate the attractor regions and their variances; (2) a refined search for the position and variance of the system's attractor regions using gradient descent-based minimization methods; and (3) the position and variance found in the previous step were used to initialize the parameters of the

Gamma mixture model, followed by fine-tuning using gradient descent-based minimization algorithms. Each stage is detailed further below.

In the first stage, we employ the tree-structured parsen estimator (TPE) algorithm [53] to address the optimization problem. TPE is a variant of the Bayesian method [54, 55] that has demonstrated notable success in hyperparameter optimization [56, 57]. For this purpose, we utilized the package Optuna 4.1.0 in Python 3.12. In the subsequent stage, the Adam optimization algorithm [58] was employed to find the parameters of the Gaussian mixture model (MGM) that minimize the residual (19). The Tensorflow 2.18 and Tensorflow Probability 0.25 libraries in Python 3.11 were utilized to leverage automatic differentiation capabilities [59]. In this stage and the following one, the Tensorflow Probability library was used to represent both the Gaussian mixture distribution and the gamma mixture distribution for the next stage. Finally, the last stage commenced with a gamma mixture, initializing the modes of the mixture distribution with the locations from the Gaussian model and adjusting the variances of the gamma distributions based on the variances of the Gaussian mixture model. Subsequently, the Adam optimization algorithm was employed to fine-tune the gamma mixture. The algorithm was stopped when no significant decrease in the residual was observed, achieving an average value per collocation point of 10^{-6}. The resulting values and the respective locations of the attractors are presented in Table 2.

Table 2 Parameters for the gamma mixure model, location of the four attractors, and mean concentration of the proteins of activated genes in each attractor

| Parameter		Value
α_{1j} (shape)	=	(1, 2.25, 2.22, 1, 2.23, 2.24, 2.24, 2.23, 2.24, 2.24, 2.26, 1)
β_{1j} (rate)	=	(0.91, 0.64, 0.64, 0.89, 0.64, 0.64, 0.64, 0.64, 0.64, 0.64, 0.64, 0.91)
Mode $(\alpha - 1)/\beta$		
(location)	=	(0, 1.93, 1.88, 0, 1.89, 1.91, 1.92, 1.90, 1.92, 1.91, 1.94, 0)
Mean conc		
(active genes)	=	1.9
α_{2j} (shape)	=	(5.83, 5.79, 1, 5.82, 5.82, 5.74, 1, 5.66, 1, 5.69, 1, 5.78)
β_{2j} (rate)	=	(1.88, 1.87, 0.95, 1.86, 1.87, 1.86, 0.93, 1.85, 0.94, 1.85, 0.94, 1.87)
Mode $(\alpha - 1)/\beta$		
(location)	=	(2.56, 2.55, 0, 2.58, 2.57, 2.54, 0, 2.51, 0, 2.52, 0, 2.55)
Mean conc		
(active genes)	=	2.5
α_{3j} (shape)	=	(1, 1, 2.21, 2.21, 2.22, 2.22, 2.21, 1, 2.21, 1, 2.22, 2.22)
β_{3j} (rate)	=	(0.92, 0.93, 0.67, 0.67, 0.67, 0.67, 0.67, 0.94, 0.67, 0.94, 0.66, 0.67)
Mode $(\alpha - 1)/\beta$		
(location)	=	(0, 0, 1.79, 1.79, 1.82, 1.81, 1.81, 0, 1.79, 0, 1.83, 1.81)
Mean conc		
(active genes)	=	1.8
α_{4j} (shape)	=	(2.21, 1, 2.21, 2.23, 1, 1, 1, 2.19, 2.24, 2.24, 1, 2.22)
β_{4j} (rate)	=	(0.63, 0.89, 0.63, 0.63, 0.9, 0.92, 0.89, 0.64, 0.63, 0.63, 0.89, 0.63)
Mode $(\alpha - 1)/\beta$		
(location)	=	(1.91, 0, 1.91, 1.94, 0, 0, 0, 1.86, 1.96, 1.95, 0 1.91)
Mean conc		
(active genes)	=	1.9
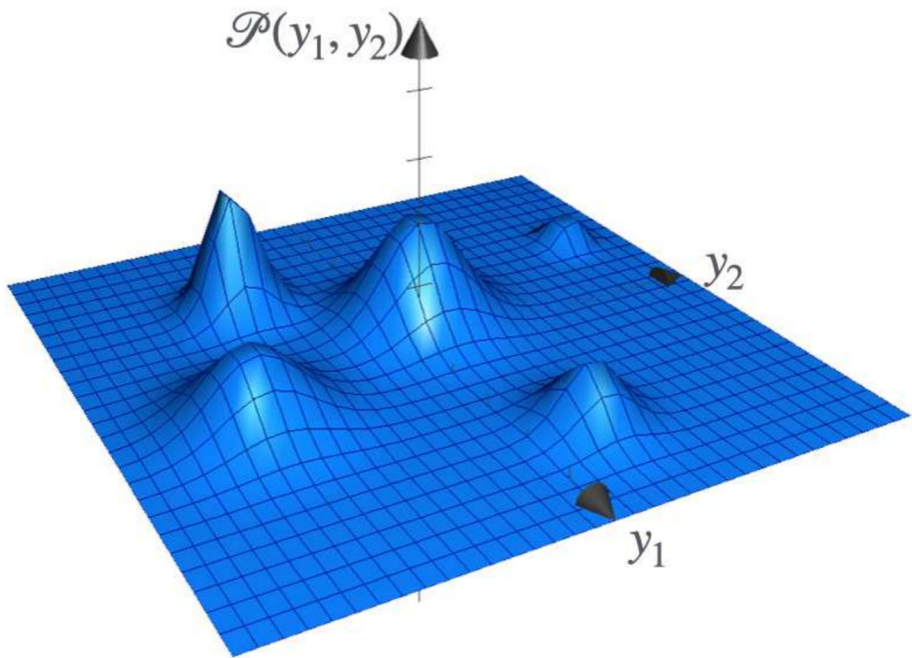
Fig. 4 Graph of the function (17) when $d = 2$ and $n = 5$. The maxima correspond to the stationary states of the system. For the system of interest, d has to be equal to 12

Upon estimation of the FPE solution P, it becomes feasible to identify the epigenetic landscape as $U(y) = -\ln P(y)$.

We can now calculate the correlations between pairs of proteins and compare them with experimental observations. It is worth noticing that, in this case, we are not able to use traditional hypothesis tests like (Kolmogorov-Smirnov, Wald-Wolfowitz) for a single variable, neither multivariate generalizations [60–63] because experimental data is normally reported as coexpression matrices rather than concentration samples.

We take a sample of 10^5 points from the estimated $\hat{\mathcal{P}}$ distribution. From this sample, the Pearson correlation was calculated for each pair of variables, and a matrix of correlations M_m may be derived. This matrix can be compared directly with the matrix M_e shown in Fig. 3 obtained from gene expression measures for model validation and comparison.

A straightforward comparison is achieved by calculating the Euclidean distance between both matrices, that is,

$$d(M, N) = \sqrt{\sum_{s=1}^{t^2}(m_s - n_s)^2} \tag{21}$$

where M and N are matrices both with dimension $t \times t$ and m_s and n_s are their respective entries.

3 Results

The theoretical correlation matrix M_m that was deduced from the solution to the FPE $\hat{\mathcal{P}}$ is represented in Fig. 5.

Then, the Euclidean distance

$$D = d(M_e, M_m) = 2.68 \tag{22}$$

was measured between the experimental data matrix and the theoretical correlations matrix obtained from the model. To make this value meaningful, we must calculate how significant this distance is compared to an uninformative correlation matrix.

To do so, two sets of distances were derived. The first one is taken from samples of the distances between two random matrices with the same characteristics of a correlation matrix; the second one is taken from a random sample of distances between the experimental matrix M_e and random matrices of correlations.

This allows us to infer the corresponding distance distributions shown in Fig. 6. The distance $d(M_e, M_m)$ was compared with both distributions, showing that up to this test, the model cannot be discarded.
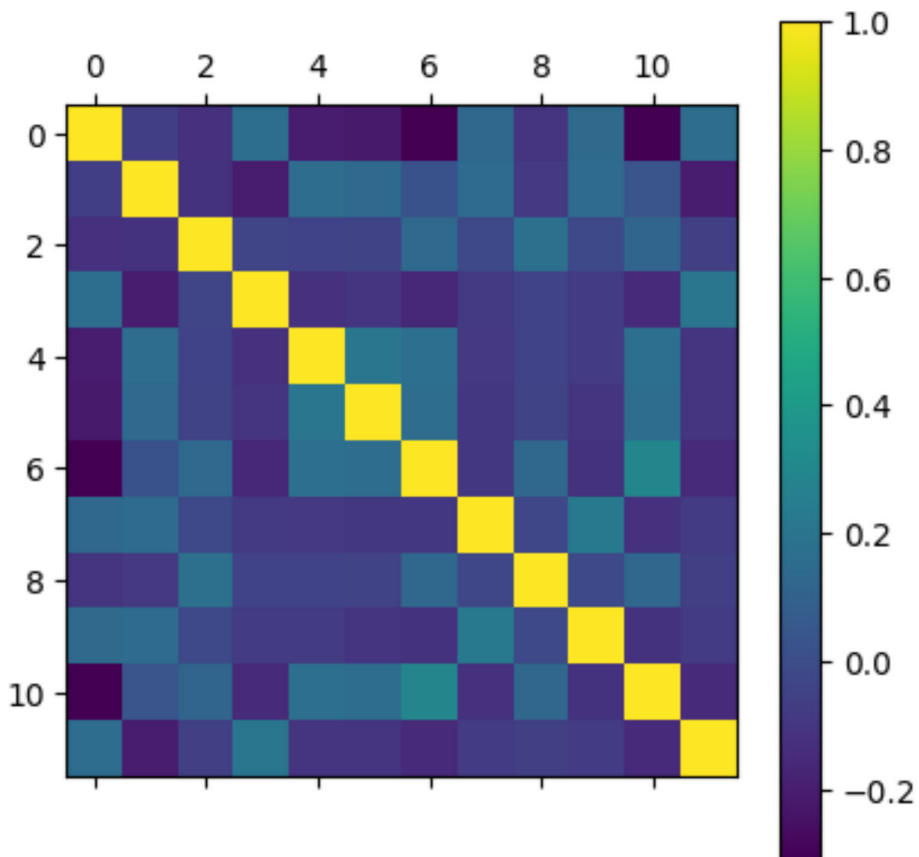
Fig. 5 Deduced correlations from the gene regulatory network. Numbers correspond to gene labels
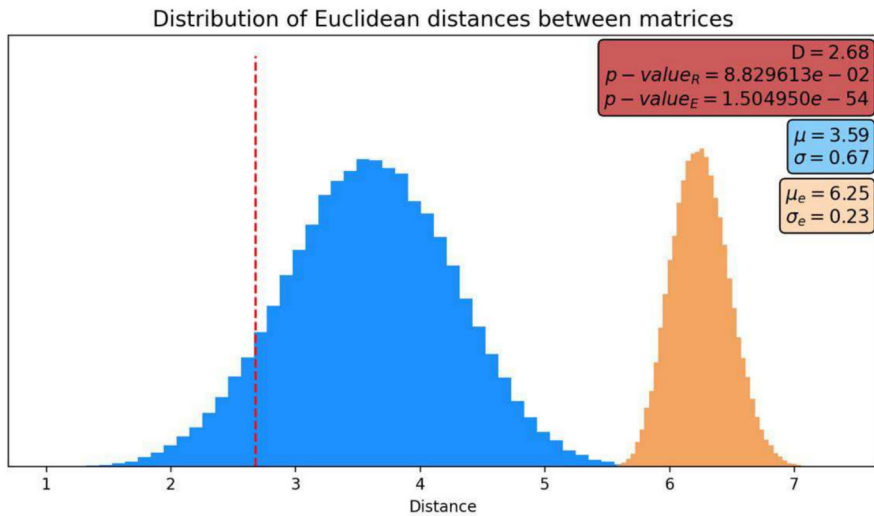
Fig. 6 Distance between the model correlation matrix and the experimental data matrix. The orange histogram shows the distribution of the distances between the experimental data matrix and a random matrix of the same characteristics. The blue histogram shows the distributions of the distances between two random matrices of the same characteristics as the experimental data matrix

The epigenetic landscape can provide relevant information regarding cellular processes. Specifically, the depth of each potential well offers insights into escape probability and mean escape time. The distance between wells yields information concerning the transition probability between distinct attractors. Thus, the most probable transitions between attractors and, in particular, cell processes could be determined, as well as the relative time spent in each cell state.

4 Discussion

Adequate knowledge of the epigenetic landscape is a relevant problem in systems biology and epigenetics. Various approaches have been proposed and implemented [64–67]. In this study, we propose a practical methodology for estimating it for systems with many dimensions. This method involves deriving the Fokker-Planck equation (FPE) that describes the dynamics of a cellular process from previous knowledge of its GRN. The epigenetic landscape $U(y)$ is associated with the stationary solution $P_{ss}(y)$ of the FPE as $U(y) = -\ln P_{ss}(y)$ [66]. A gamma mixture model is proposed as a solution for P_{ss} where the search for relevant parameters takes place in three stages. In the first stage, a global exploration is conducted using a Monte Carlo algorithm to search for the attractor regions. In the second stage, a Gaussian mixture model is employed as a proxy distribution to estimate the location of the attractors and the variance of the dynamic system around them. The Gaussian mixture is used because its quadrature points are not widely spread out, and the fluctuations in each evaluation are small, unlike with a gamma distribution. Finally, the location and variance found are used to initialize the parameter search for the gamma mixture model. Gradient-based optimization algorithms utilizing automatic differentiation are employed in the last two stages. The FPE solution can be used to validate the underlying model using experimental co-expression data. A Monte Carlo-Hastings algorithm was employed to sample from the

derived distribution, and correlations between protein concentrations were directly compared with correlations observed in the experimental co-expression data. Similarly, the solution of the FPE can be used to infer missing gene expression data from existing experimental data. To illustrate this procedure, we found the epigenetic landscape of the FPE for the *Arabidopsis thaliana* flower morphogenesis process. Furthermore, the estimation of the relative Manhattan distances between potential wells of the epigenetic landscape can be utilized to determine the most probable transition between cell states, and their relative heights can provide the relative duration spent in each attractor. For the *Arabidopsis thaliana* regulatory network analyzed herein, this duration is correlated with the time spent in each developmental phase. It is worth noting that gene regulatory networks inferred from Boolean models are inherently limited in the information they capture, a consequence of the models' simplicity. Therefore, these models are not meant to provide a detailed description of a GRN; rather, their aim is to highlight general relationships among protein concentrations across the network's distinct attractors. On the other hand, it should be noted that mRNA and protein production are bursty, and therefore, the present approach using the Langevin chemical equation based on white noise, for such complex processes like mRNA and protein production, is an approximation. It has recently been observed that stochastic models based on compound Poisson processes provide a good description of mRNA and protein expression levels [68, 69]. A continuous approximation of the chemical master equation could be employed in the future, along with the techniques described here, to find more reliable probability distributions for protein concentrations.

Author Contributions Conceptualization: J.V.; methodology: N.L., E.M., and J.V.; formal analysis and investigation: N.L., J.V.; writing—original draft preparation: N.L.; writing—review and editing: E.M., J.V.; funding acquisition: J.V.; resources: E.M.; supervision: E.M., J.V. All authors read and approved the final manuscript.

Data Availability The dataset of gene coexpression for *Arabidopsis thaliana* is available at https://doi.org/10.1093/pcp/pcac041 or https://atted.jp/.

Declarations

References

1. Alberghina, L., Westernhoff, H.V.: System Biology: Definitions and Perspectives. Springer, Germany (2008)

2. Roces, M.E.: Modeling Methods for Medical Systems Biology: Regulatory Dynamics Underlying the Emergence of Disease Processes. Advances in Experimental Medicine and Biology Ser, vol. v.1069. Springer, Cham (2018)

3. Balleza, E., Alvarez-Buylla, E.R., Chaos, A., Kauffman, S., Shmulevich, I., Aldana, M.: Critical dynamics in genetic regulatory networks: examples from four kingdoms. PLoS One **3**(6), 1–10 (2008). https://doi.org/10.1371/journal.pone.0002456

4. Azpeitia, E., Benítez, M., Padilla-Longoria, P., Espinosa-Soto, C., Alvarez-Buylla, E.R.: Dynamic network-based epistasis analysis: Boolean examples. Front. Plant Sci. **2**, 4 (2011). https://doi.org/10.3389/fpls.2011.00092

5. Barrio, R.A., Romero-Arias, J.R., Noguez, M.A., Azpeitia, E., Ortiz-Gutiérrez, E., Hernández-Hernández, V., Cortes-Poza, Y., Álvarez-Buylla, E.R.: Cell patterns emerge from coupled chemical and physical fields with cell proliferation dynamics: the *Arabidopsis thaliana* root as a study system. PLoS Comput. Biol. **9**(5), 1003026 (2013). https://doi.org/10.1371/journal.pcbi.1003026. Publisher: Public Library of Science

6. Resendis-Antonio, O., Arellano-Villavicencio, J., Vázquez-Jiménez, A., Oropeza-Valdéz, J., Padrón-Manrique, C., Prado-García, H., Tovar, A.: Intratumoral heterogeneity and metabolic cross-feeding in a three-dimensional model of breast cancer: an in silico perspective. In review (2024). https://doi.org/10.21203/rs.3.rs-4864972/v1. https://www.researchsquare.com/article/rs-4864972/v1

7. Pérez, B., Torre-Villalvazo, I., Wilson-Verdugo, M., Lau-Corona, D., Muciño-Olmos, E., Coutiño-Hernández, D., Noriega-López, L., Resendis-Antonio, O., Valdés, V.J., Torres, N., Tovar, A.R.: Epigenetic reprogramming of H3K4me3 in adipose-derived stem cells by HFS diet consumption leads to a disturbed transcriptomic profile in adipocytes. Am. J. Physiol.-Endocrinol. Metab. (2024). https://doi.org/10.1152/ajpendo.00093.2024

8. Gennarino, V.A., D'Angelo, G., Dharmalingam, G., Fernandez, S., Russolillo, G., Sanges, R., Mutarelli, M., Belcastro, V., Ballabio, A., Verde, P., Sardiello, M., Banfi, S.: Identification of microRNA-regulated gene networks by expression analysis of target genes. Genome Res. **22**(6), 1163–1172 (2012). https://doi.org/10.1101/gr.130435.111

9. Marsico, A., Huska, M.R., Lasserre, J., Hu, H., Vucicevic, D., Musahl, A., Orom, U., Vingron, M.: PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. Genome Biol. **14**(8), 84 (2013). https://doi.org/10.1186/gb-2013-14-8-r84

10. Griffon, A., Barbier, Q., Dalino, J., Helden, J., Spicuglia, S., Ballester, B.: Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. Nucleic Acids Res. **43**(4), 27 (2015). https://doi.org/10.1093/nar/gku1280

11. Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., Oord, J., Atak, Z.K., Wouters, J., Aerts, S.: SCENIC: single-cell regulatory network inference and clustering. Nat. Methods **14**(11), 1083–1086 (2017). https://doi.org/10.1038/nmeth.4463

12. Azizi, E., Prabhakaran, S., Carr, A., Pe'er, D.: Bayesian inference for single-cell clustering and imputing. Genomics Comput. Biol. **3**, 46 (2017). https://doi.org/10.18547/gcb.2017.vol3.iss1.e46

13. Åström, K.J., Murray, R.M.: Feedback Systems: An Introduction for Scientists and Engineers. Princeton University Press, Princeton (2008)

14. Azpeitia, E., Alvarez-Buylla, E.R.: A complex systems approach to *Arabidopsis* root stem-cell niche developmental mechanisms: from molecules, to networks, to morphogenesis. Plant Mol. Biol. **80**(4–5), 351–363 (2012). https://doi.org/10.1007/s11103-012-9954-6

15. Siahpirani, A.F., Knaack, S., Chasman, D., Seirup, M., Sridharan, R., Stewart, R., Thomson, J., Roy, S.: Dynamic regulatory module networks for inference of cell type–specific transcriptional networks. Genome Res. **32**(7), 1367–1384 (2022). https://doi.org/10.1101/gr.276542.121

16. Skok Gibbs, C., Jackson, C.A., Saldi, G.-A., Tjärnberg, A., Shah, A., Watters, A., De Veaux, N., Tchourine, K., Yi, R., Hamamsy, T., Castro, D.M., Carriero, N., Gorissen, B.L., Gresham, D., Miraldi, E.R., Bonneau, R.: High-performance single-cell gene regulatory network inference at scale: the Inferelator 3.0. Bioinformatics **38**(9), 2519–2528 (2022). https://doi.org/10.1093/bioinformatics/btac117

17. Ventre, E., Herbach, U., Espinasse, T., Benoit, G., Gandrillon, O.: One model fits all: combining inference and simulation of gene regulatory networks. PLOS Comput. Biol. **19**(3), 1010962 (2023). https://doi.org/10.1371/journal.pcbi.1010962

18. Mbebi, A.J., Nikoloski, Z.: Gene regulatory network inference using mixed-norms regularized multivariate model with covariance selection. PLoS Comput. Biol. **19**(7), 1010832 (2023). https://doi.org/10.1371/journal.pcbi.1010832

19. Wang, J., Ma, A., Ma, Q., Xu, D., Joshi, T.: Inductive inference of gene regulatory network using supervised and semi-supervised graph neural networks. Comput. Struct. Biotechnol. J. **18**, 3335–3343 (2020). https://doi.org/10.1016/j.csbj.2020.10.022

20. Ji, R., Geng, Y., Quan, X.: Inferring gene regulatory networks with graph convolutional network based on causal feature reconstruction. Sci. Rep. **14**(1), 21342 (2024). https://doi.org/10.1038/s41598-024-71864-8. Publisher: Nature Publishing Group

21. Yuan, Q., Duren, Z.: Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. Nat. Biotechnol. 1–11 (2024). https://doi.org/10.1038/s41587-024-02182-7

22. Gao, Z., Su, Y., Xia, J., Cao, R.-F., Ding, Y., Zheng, C.-H., Wei, P.-J.: DeepFGRN: inference of gene regulatory network with regulation type based on directed graph embedding. Brief. Bioinform. **25**(3), 143 (2024). https://doi.org/10.1093/bib/bbae143

23. Oropeza-Valdez, J.J., Padron-Manrique, C., Vázquez-Jiménez, A., Soberon, X., Resendis-Antonio, O.: Exploring metabolic anomalies in COVID-19 and post-COVID-19: a machine learning approach with explainable artificial intelligence. Front. Mol. Biosci. **11** (2024). https://doi.org/10.3389/fmolb.2024.1429281

24. Chen, S., Mar, J.C.: Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinformat. **19**(1), 232 (2018). https://doi.org/10.1186/s12859-018-2217-z

25. Edgar, R., Domrachev, M., Lash, A.E.: Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30**(1), 207–210 (2002). https://doi.org/10.1093/nar/30.1.207

26. Wang, J., Zhuang, J., Iyer, S., Lin, X.-Y., Greven, M.C., Kim, B.-H., Moore, J., Pierce, B.G., Dong, X., Virgil, D., Birney, E., Hung, J.-H., Weng, Z.: Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res. **41**(Database issue), 171–176 (2013). https://doi.org/10.1093/nar/gks1221

27. Alvarez-Buylla, E.R., Azpeitia, E., Barrio, R., Benítez, M., Padilla-Longoria, P.: From ABC genes to regulatory networks, epigenetic landscapes and flower morphogenesis: making biological sense of theoretical approaches. Semin. Cell Dev. Biol. **21**(1), 108–117 (2009). https://doi.org/10.1016/j.semcdb.2009.11.010

28. Yaghoobi, H., Haghipour, S., Hamzeiy, H., Asadi-Khiavi, M.: A review of modeling techniques for genetic regulatory networks. J. Med. Signals Sens. **2**(1), 61–70 (2012)

29. Vijesh, N., Chakrabarti, S.K., Sreekumar, J.: Modeling of gene regulatory networks: a review. J. Biomed. Sci. Eng. **6**(2), 223–231 (2013). https://doi.org/10.4236/jbise.2013.62A027

30. Cussat-Blanc, S., Harrington, K., Banzhaf, W.: Artificial gene regulatory networks—a review. Artif. Life **24**(4), 296–328 (2019). https://doi.org/10.1162/artl_a_00267

31. Gómez-Schiavon, M., Montejano-Montelongo, I., Orozco-Ruiz, F.S., Sotomayor-Vivas, C.: The art of modeling gene regulatory circuits. NPJ Syst. Biol. Appl. **10**(1), 1–6 (2024). https://doi.org/10.1038/s41540-024-00380-2

32. Waddington, C.H.: The Strategy of the Genes. A Discussion of Some Aspects of Theoretical Biology. George Allen & Unwin, Bristol (1957)

33. Kauffman, S.A.: The Origins of Order: Self-organization and Selection in Evolution. Oxford University Press, Oxford (1993)

34. Villarreal, C., Padilla-Longoria, P., Buylla, M.E.: General theory of genotype to phenotype mapping: derivation of epigenetic landscapes from n-node complex gene regulatory networks. Phys. Rev. Lett. **11**(109), 118102 (2012)

35. Sukys, A., Öcal, K., Grima, R.: Approximating solutions of the chemical master equation using neural networks. iScience **25**(9), 105010 (2022). https://doi.org/10.1016/j.isci.2022.105010

36. Jia, C., Grima, R.: Holimap: an accurate and efficient method for solving stochastic gene network dynamics. Nat. Commun. **15**(1), 6557 (2024). https://doi.org/10.1038/s41467-024-50716-z

37. Mendoza, L., Állvarez-Buylla, E.R.: Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis. J. Theor. Biol. **193**(2), 307–19 (1998)

38. Buylla, E.R., Chaos, A., Aldana, M., Benítes, M.: Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. PLoS One **11**(3), 1–13 (2008)

39. Barrio, R., Hernández-Machado, A., Varea, C., Romero-Arias, J.R., Álvarez-Buylla, E.: Flower development as an interplay between dynamical physical fields and genetic networks. PLoS One **5**(10), 1–9 (2010). https://doi.org/10.1371/journal.pone.0013523

40. Thakar, J.: Pillars of biology: Boolean modeling of gene-regulatory networks. J. Theor. Biol. **578**, 111682 (2024). https://doi.org/10.1016/j.jtbi.2023.111682

41. Katoch, S., Chauhan, S.S., Kumar, V.: A review on genetic algorithm: past, present, and future. Multimed. Tools Appl. **80**(5), 8091–8126 (2021). https://doi.org/10.1007/s11042-020-10139-6

42. Coen, E.S., Meyerowitz, E.M.: The war of the whorls: genetic interactions controlling flower development. Nature **353**(6339), 31–37 (1991). https://doi.org/10.1038/353031a0

43. Alon, U.: An Introduction to Systems Biology: Design Principles of Biological Circuits, 2nd edition. CRC Press, Boca Raton, Fla (2019)

44. Obayashi, T., Hibara, H., Kagaya, Y., Aoki, Y., Kinoshita, K.: ATTED-II v11: a plant gene coexpression database using a sample balancing technique by subagging of principal components. Plant Cell Physiol. **63**(6), 869–881 (2022). https://doi.org/10.1093/pcp/pcac041

45. Del Vecchio, D., Murray, R.M.: Biomolecular Feedback Systems. Princeton University Press, Princeton (2015)

46. Anderson, D.F., Kurtz, T.G.: Stochastic Analysis of Biochemical Systems. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16895-1 . https://link.springer.com/10.1007/978-3-319-16895-1

47. Jia, C., Wang, L.Y., Yin, G.G., Zhang, M.Q.: Single-cell stochastic gene expression kinetics with coupled positive-plus-negative feedback. Phys. Rev. E **100**, 052406 (2019). https://doi.org/10.1103/PhysRevE.100.052406

48. Chen, X., Jia, C.: Limit theorems for generalized density-dependent Markov chains and bursty stochastic gene regulatory networks. J. Math. Biol. **80**(4), 959–994 (2020). https://doi.org/10.1007/s00285-019-01445-1

49. Gardiner, C.W.: Handbook of Stochastic Methods: For Physics, Chemistry and Natural Sciences, 2nd edition. Springer, Berlin Heidelberg (1997)

50. Wang, X., Li, Y., Jia, C.: Poisson representation: a bridge between discrete and continuous models of stochastic gene regulatory networks. J. R. Soc. Interface **20**(208), 20230467 (2023). https://doi.org/10.1098/rsif.2023.0467

51. Chakraverty, S., Mahato, N., Karunakar, P., Rao, T.D.: Weighted Residual Methods, pp. 31–43. John Wiley & Sons, Ltd. (2019). Chap. 3. https://doi.org/10.1002/9781119423461.ch3. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119423461.ch3

52. E, W., Yu, B.: The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. Commun. Math. Stat. **6**(1), 1–12 (2018). https://doi.org/10.1007/s40304-018-0127-z

53. Watanabe, S.: Tree-structured parzen estimator: understanding its algorithm components and their roles for better empirical performance. arXiv. [cs] (2023). https://doi.org/10.48550/arXiv.2304.11127. arXiv:2304.11127

54. Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., Freitas, N.: Taking the human out of the loop: a review of Bayesian optimization. Proc. IEEE **104**(1), 148–175 (2016). https://doi.org/10.1109/JPROC.2015.2494218

55. Garnett, R.: Bayesian Optimization. Cambridge University Press, Cambridge (2023). https://doi.org/10.1017/9781108348973. https://www.cambridge.org/core/books/bayesian-optimization/11AED383B208E7F22A4CE1B5BCBADB44

56. Bergstra, J., Yamins, D., Cox, D.D.: Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. PMLR **28**(1), 115–123 (2013)

57. Watanabe, S., Awad, N., Onishi, M., Hutter, F.: Speeding up multi-objective hyperparameter optimization by task similarity-based meta-learning for the tree-structured Parzen estimator. IJCAI **4**, 4380–4388 (2023). https://doi.org/10.24963/ijcai.2023/487. ISSN: 1045-0823. https://www.ijcai.org/proceedings/2023/487

58. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv. [cs] (2017). https://doi.org/10.48550/arXiv.1412.6980. arXiv:1412.6980

59. Linnainmaa, S.: Taylor expansion of the accumulated rounding error. BIT Numer. Math. **16**(2), 146–160 (1976). https://doi.org/10.1007/BF01931367

60. Hotelling, H.: A generalized t test and measure of multivariate dispersion. (1951). https://api.semanticscholar.org/CorpusID:44845865

61. Friedman, J.H., Rafsky, L.C.: Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Ann. Stat. **7**(4), 697–717 (1979). https://doi.org/10.1214/aos/1176344722

62. Bickel, P.J.: A distribution free version of the Smirnov two sample test in the p-variate case. Ann. Math. Stat. **40**(1), 1–23 (1969). https://doi.org/10.1214/aoms/1177697800

63. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. J. Mach. Learn. Res. **13**(25), 723–773 (2012)

64. Li, C., Wang, J.: Landscape and flux reveal a new global view and physical quantification of mammalian cell cycle. Proc. Natl. Acad. Sci. U.S.A. **111**(39), 14130–14135 (2014). https://doi.org/10.1073/pnas.1408628111

65. Zhou, P., Li, T.: Construction of the landscape for multi-stable systems: potential landscape, quasi-potential, A-type integral and beyond. J. Chem. Phys. **144**(9), 094109 (2016). https://doi.org/10.1063/1.4943096

66. Ye, L., Song, Z., Li, C.: Landscape and flux quantify the stochastic transition dynamics for p53 cell fate decision. J. Chem. Phys. **154**(2), 025101 (2021). https://doi.org/10.1063/5.0030558

67. Kang, X., Li, C.: A dimension reduction approach for energy landscape: identifying intermediate states in metabolism-EMT network. Adv. Sci. **8**(10), 2003133 (2021). https://doi.org/10.1002/advs.202003133. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/advs.202003133

68. Jedrak, J., Ochab-Marcinek, A.: Time-dependent solutions for a stochastic model of gene expression with molecule production in the form of a compound poisson process. Phys. Rev. E **94**(3), 032401 (2016). https://doi.org/10.1103/PhysRevE.94.032401

69. Jia, C., Zhang, M.Q., Qian, H.: Emergent Lévy behavior in single-cell stochastic gene expression. Phys. Rev. E **96**(4), 040402 (2017). https://doi.org/10.1103/PhysRevE.96.040402

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.