Published in final edited form as:

SIAM J Appl Math. 2013 March 1; 73(2): 804–826. doi:10.1137/120887588.

TWO-LAYER MATHEMATICAL MODELING OF GENE EXPRESSION: INCORPORATING DNA-LEVEL INFORMATION AND SYSTEM DYNAMICS*

JACQUELINE M. DRESCH[†], MARC A. THOMPSON[‡], DAVID N. ARNOSTI[§], and CHICHIA CHIU[¶]

JACQUELINE M. DRESCH: jdresch@g.hmc.edu; MARC A. THOMPSON: thom1068@msu.edu; DAVID N. ARNOSTI: arnosti@msu.edu; CHICHIA CHIU: chiu@math.msu.edu

[†]Department of Mathematics, Harvey Mudd College, Claremont, CA 91711. This author's work was partly supported by a Teaching and Research Postdoctoral Fellowship at Harvey Mudd College

[‡]Department of Bioengineering, North Carolina Agricultural and Technical State University, Greensboro, NC27411

[§]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824. This author's work was partly supported by NIH grant GM056976

[¶]Department of Mathematics, Michigan State University, East Lansing, MI 48824

Abstract

High-throughput genome sequencing and transcriptome analysis have provided researchers with a quantitative basis for detailed modeling of gene expression using a wide variety of mathematical models. Two of the most commonly employed approaches used to model eukaryotic gene regulation are systems of differential equations, which describe time-dependent interactions of gene networks, and thermodynamic equilibrium approaches that can explore DNA-level transcriptional regulation. To combine the strengths of these approaches, we have constructed a new two-layer mathematical model that provides a dynamical description of gene regulatory systems, using detailed DNA-based information, as well as spatial and temporal transcription factor concentration data. We also developed a semi-implicit numerical algorithm for solving the model equations and demonstrate here the efficiency of this algorithm through stability and convergence analyses. To test the model, we used it together with the semi-implicit algorithm to simulate a *Drosophila* gene regulatory circuit that drives development in the dorsal-ventral axis of the blastoderm-stage embryo, involving three genes. For model validation, we have done both mathematical and statistical comparisons between the experimental data and the model's simulated data. Where protein and *cis*-regulatory information is available, our two-layer model provides a method for recapitulating and predicting dynamic aspects of eukaryotic transcriptional systems that will greatly improve our understanding of gene regulation at a global level.

*This work supported in part by NSF grant DMS-0839966. <http://www.siam.org/journals/siap/x-x/88758.html>

Keywords

reaction-diffusion equation; thermodynamic equilibrium; transcriptional regulation

1. Introduction

All living organisms have genes that are differentially activated and repressed to regulate life processes. Gene regulation drives expression of genes at precise levels, spatial locations, and times. Thus, the understanding of the process of gene regulation underlies central questions in modern biology, ranging from bacterial growth to human disease pathways. At the initial level of transcriptional regulation, we seek to understand the operation of regions of DNA that control gene expression (Figure 1). These DNA regions are referred to generally as *cis*-regulatory elements, or in animals, “enhancers” when their action involves regulation in a distance- and orientation-independent manner [9]. In bacteria, *cis*-regulatory elements typically have a small number of binding sites for transcription factors (TFs) that bind to genes to control their expression, and the protein-DNA and protein-protein interactions are often fairly well understood. In eukaryotes, however, we lack a general understanding of how groups of TFs interact with each other, with the chromatin template present in all eukaryotic cells, and with the general transcriptional machinery that includes the RNA polymerase and many associated factors.

A second key area of study is networks of genes, including layers of regulatory events that drive the synchronized expression of tissue-specific genes during the process of development, as well as the reaction of cells and whole systems to environmental perturbations. Transcriptional networks underlie many of these interactions, although additional events that include subsequent levels of gene expression (mRNA metabolism, translational and post-translational regulation, signaling) play important roles. Both of these facets of gene expression have been the target of many empirical studies that focus on particular aspects of eukaryotic regulation within a given system, but we lack general unifying concepts and tools to explain how DNA regulatory sequences give rise to the complex spatio-temporal expression of genes.

The development of mathematical models of gene expression has been one answer to this challenge. In this paper we focus on modeling gene expression during the early embryonic (blastoderm) stage of development in the eukaryotic organism *Drosophila melanogaster*, the fruit fly. Because of the strong genomic and genetic tools available in this organism, the application of mathematical tools is particularly powerful, but we believe that similar approaches will be useful in other systems to provide biologists with a great deal of information about gene expression, providing a basis for hypotheses about disease, development, and evolution.

In eukaryotic organisms, transcriptional *cis*-regulatory information is often dispersed with respect to the gene being regulated, with some elements located over one million base pairs from the start of the transcription unit [27]. Thus, one initial application of mathematical modeling has been to identify possible *cis*-regulatory elements, based on the predicted activity of putative binding sites located in candidate regions. As complementary

experimental methods have come to the fore, it is becoming easier to find where regulatory sequences are located based on chromatin modifications and maps of *in vivo* binding by TFs, but the quantitative output of a particular *cis*-regulatory element cannot generally be derived from these sorts of measurements. Here, mathematical models can also be quite useful in interpreting the significance of genetic variation that may quantitatively affect the activity of *cis*-regulatory elements, a property which is important in the context of development, disease, population variation, and the evolution of new traits. Mathematical models are capable of using available DNA sequence information and quantitative expression data to estimate parameter values related to activator and repressor protein efficiency or cooperativity.

In the field of gene regulation, three different mathematical models have been applied most frequently: Boolean models, thermodynamic-based (also termed “fractional occupancy”) models, and differential equation models [6]. More recently, a fourth approach, which involves stochastic dynamical models together with a modified Gillespie algorithm, has been proposed for this problem [18, 26]. These models all have inherent strengths and weaknesses involving the amount of input data, the level of complexity, and the time-frame in which they can model gene expression. Boolean models as well as stochastic dynamical models have the benefit of being able to use large amounts of data available from genome-wide arrays to identify and study the activity of specific gene interaction networks, but they do not in general focus on details of the *cis*-regulatory elements (see [39] for an exception). Boolean models, by their nature, also do not seek to incorporate the dynamics or specific levels of particular components [33, 39, 2]. On the other hand, stochastic dynamical models incorporate the noise effects and the temporal changes in gene expression, but have been shown to be computationally efficient only when solved in the steady state [18]. Thermodynamic-based models were derived to specifically account for features of *cis*-regulatory elements; utilize information about DNA sequence, protein binding specificity, and TF concentrations using the laws of biophysical thermodynamics; and often include terms for specific TF interactions and distance-dependent DNA binding events. As more data about sequences of *cis*-regulatory elements and regulatory protein concentrations becomes available, fractional occupancy models may potentially permit us to decipher the *cis*-regulatory grammar underlying enhancer architecture [1, 11, 40, 25, 35, 19, 16, 23].

Thermodynamic-based models are static in nature. They may use information from slightly varying time-frames, but their outputs are only snapshots of a system. Boolean models often incorporate some time-dependence, although they only represent time, state, or space with a discrete set of values. Whether these two model types are adding insight into which gene products affect other genes or whether they are predicting expression, these conclusions are made for specific time points or over fixed discrete time intervals. Differential equation models and probabilistic dynamical models have allowed researchers to model transcriptional regulation over continuous time intervals to study the behavior of subtly evolving gene or protein network structure. These models can simulate continuous variations in gene expression over time as well as diffusion and decay of proteins and mRNAs. This dynamic movie of gene expression can be extremely useful in developing visual representations of small networks of genes. The primary drawback to these models is

that, when applied to larger data sets or larger gene networks, the number of parameters rapidly increases, and setting these parameters can be computationally infeasible. Also, in past applications, the intricate details of transcriptional regulation, such as specific protein binding sites and affinities that thermodynamic-based models offer, were not included [21, 7, 29, 34, 38, 24, 20, 13, 30, 10]. In eukaryotic systems, models have taken into account the positive or negative contributions of TFs to gene expression, but none have included the DNA sequence information that would allow a dynamic quantitative model of gene expression to incorporate terms for binding specificity and TF interactions [24, 30, 10]. This paper aims to address this gap in our understanding.

Simple two-layer models have been applied and proven successful in bacterial systems, such as that of the *lac* operon in *E. coli* [34]. The difficulty of applying such models to eukaryotic systems is the increased complexity of *cis*-regulatory modules in addition to the difficulty of collecting high quality data sets. With advances in high throughput genome sequencing and transcriptome analysis, we have gained crucial insights and data sets that make such a model possible in a complicated eukaryotic system. Here, we describe the steps taken in building a two-layer model of eukaryotic transcriptional regulation and its numerical implementation, as well as point out some of the advantages to using this approach.

2. Thermodynamic-based modeling

Thermodynamic-based, or fractional occupancy, models aim to predict gene expression output from the DNA sequence of the gene's regulatory region and the concentration of proteins (i.e., TFs) involved in regulation. The term "thermodynamic" is often misinterpreted; it refers only to the fact that the model has been derived with thermodynamic equilibrium assumptions. Historically, these calculations have been made using techniques derived from statistical physics which involve computing all possible states of TFs binding to a promoter and relating these states to gene expression [36, 32, 11, 25, 40, 35, 19, 16, 23].

2.1. Rules/properties

1. Assume that the expression level of the gene (measured as the concentration of mRNA) is proportional to the fraction of successful states. Here, a state refers to a particular arrangement of transcription factors bound to their corresponding binding sites on the enhancer. A "successful" state is any state where the bound TFs are communicating with the basal machinery to turn the gene on, in other words, a state where mRNA is being produced. Thus, the general formula will be

$$[\text{mRNA}] \propto \sum_{i \in G} \frac{S_i}{S}, \quad (2.1)$$

where $G = \{\text{all successful states}\}$, S_i is the contribution from state i , and S is the sum of contributions of all possible states of the enhancer. The state of the enhancer with nothing bound is defined to have contribution 1.

2. Each state consists of a product of protein concentrations and binding affinities (i.e., a product of terms of the form $K_{TF} [TF]$), derived from the assumption that the

system is in thermodynamic equilibrium, using simple chemical stoichiometric and equilibrium equations.

3. Weights are assigned to states according to how successful a state becomes due to protein-protein interactions (i.e., TF-TF interactions and TF-RNA polymerase interactions).
 - i. The cooperativity of two proteins simultaneously bound is assumed to affect the state's weight by modifying the binding energy of that state. Thus, a parameter C , representing the cooperative interaction between the TFs when they bind simultaneously, is multiplied by the binding affinities of the bound proteins, $CK_A K_B$. The state i with A and B bound cooperatively is given by $S_i = CK_A K_B [A] [B]$. In general,

$$S_i = C_i \prod_{j=1}^{n_i} K_j [\text{TF}_j], \quad (2.2)$$

where n_i is the number of binding sites bound by a TF in state i , K_j is the binding affinity of binding site j , $[\text{TF}_j]$ is the protein concentration corresponding to the TF bound at binding site j , and C_i is the cooperativity parameter representing the product of all cooperative interactions between the TFs bound in state i . Recall that the state of the enhancer with nothing bound has contribution 1. We will refer to this state as S_1 , and hence $S_1 = 1$ by definition.

- ii. The quenching (or repressive) efficiency of a bound repressor on an adjacently bound activator is assumed to affect a state only in terms of whether or not it is successful. Note that one can depict a state with activators and repressors bound adjacently as being divided into two substates, one successful state when the activators have the ability to activate the gene and one unsuccessful state when the repressors are blocking that ability. Since this is simply the division of a single state, the Q parameter is always in the interval $[0, 1]$, representing the proportion of time that the state is successful. So, in general,

$$[\text{mRNA}] \propto \sum_{i \in G} \frac{S_i}{S} Q_i, \quad (2.3)$$

where $0 \leq Q_i \leq 1$ is the quenching parameter representing the product of all repressive interactions between the activators and repressors bound in state i , causing the proportion of time the state is successful to be reduced.

Therefore, the general formula for calculating the level of gene expression using a thermodynamic-based model is

$$\sum_{i \in G} \frac{S_i}{S} Q_i = \frac{1}{S} \sum_{i \in G} S_i Q_i = \frac{1}{\sum_{i=1}^N S_i} \sum_{i \in G} S_i Q_i \quad (2.4)$$

$$= \frac{1}{1 + \sum_{i=2}^N \left(C_i \prod_{j=1}^{n_i} K_j [\text{TF}_j] \right)} \sum_{i \in G} \left(C_i \prod_{j=1}^{n_i} K_j [\text{TF}_j] \right) Q_i. \quad (2.5)$$

Property 1—Given the assumptions listed above, the general formula for calculating the level of gene expression using a thermodynamic-based model always results in an expression level between 0 and 1. In other words,

$$0 \leq \sum_{i \in G} \frac{S_i}{S} Q_i \leq 1. \quad (2.6)$$

Proof: Property 1 follows directly from the facts that $0 \leq \sum_{i \in G} \frac{S_i}{S} \leq 1$, and $\frac{S_i}{S} \geq 0$ and $0 \leq Q_i \leq 1$ for all i .

Example 1—Consider the case of an enhancer, E , which contains three binding sites, two for transcriptional activators, $A1$ and $A2$, and one for a transcriptional repressor, R . The order in which the binding sites are arranged on the enhancer is $A1, R, A2$, which is illustrated in Figure 2. The assumption is that cooperative and repressive interactions take place only between adjacently bound proteins. The expression level is then calculated as

$$[mRNA] \propto \sum_{i \in G} \frac{S_i}{S} Q_i = \frac{1}{S} \sum_{i \in G} S_i Q_i, \quad (2.7)$$

where $G = \{2, 4, 5, 6, 7, 8\}$, $S = 1 + \sum_{i=2}^8 S_i$, and each S_i and Q_i term is defined in Table 2.1.

Note that the above example (and any other following the assumptions given in section 2.1) results in an equation of the form

$$[mRNA]_a \propto \frac{\Phi_a(Y)}{1 + \Psi_a(Y)}, \quad (2.8)$$

where a represents the gene of interest, Y is a vector of concentration values, and Φ_a and Ψ_a are polynomials with coefficients depending on the parameter values. These parameter values include the binding affinities, K_p ; cooperativity values, C_{pq} ; and quenching parameters, Q_{pq} , for $p, q \in PR$.

This simplified form of the equation used in thermodynamic-based modeling will be used for the remainder of the paper.

3. Dynamic modeling: Problem description

As mentioned in section 1, modeling of eukaryotic gene regulation has focused on either time-dependent interactions of gene networks [24, 30, 10] or equilibrium approaches that can explore the *cis*-regulatory grammar of transcriptional enhancers [35, 25, 40, 23, 16]. These models rely on systems of differential equations or thermodynamic-based descriptions, which can be used to understand either the dynamics of a system or DNA-level regulatory processes. For many real-world settings, both aspects need to be considered in order to simulate the dynamic changes in gene expression taking place early in development.

To combine the strengths of these approaches, we have developed a novel model that joins them methods to provide a dynamical description of gene regulatory systems, using detailed DNA-based information as well as spatial TF concentration data at varying time points. Our two-layer modeling approach uses a thermodynamic-based model as the mRNA synthesis term in our differential equation, representing the rate of mRNA production. The differential equation incorporates all other terms, representing decay and diffusion of mRNA and protein as well as protein production.

We begin constructing a model by considering the general reaction-diffusion equation:

$$\frac{dY}{dt} = R(Y) + D\Delta Y \quad (3.1)$$

for $Y \in \mathbb{R}^m$, where D is a diagonal matrix containing diffusion constants and Δ is the one-dimensional (1-D) Laplacian operator. The 1-D simplifying assumption here is based on biological observations that gradients of TFs in the *Drosophila* embryo tend to run in a 1-D axis, either dorsal-ventral or anterior-posterior. Note that this is a PDE due to its dependence on both space and time. This equation is a starting point for a realistic model of gene regulation during development. It can incorporate mRNA and protein concentrations, their spatial diffusion and decay over time, and the intricate relationships between these concentrations can. In the setting of the blastoderm-stage embryo, cell membranes have not yet separated the nuclei from each other, so proteins and mRNAs are able to diffuse in this syncytial environment. mRNA production, the process of transcription, is heavily regulated by protein (TF) binding, and protein production, the process of translation, is dependent on mRNA production. It is therefore essential to employ an equation that allows for diffusion, decay, and an easily modified synthesis term that can incorporate natural feedforward and feedback loops.

We first discretize (3.1) in space to change the PDE system to an ODE system. Unlike the standard practice in numerical analysis where the continuous spatial variable is divided into mesh points in the domain, we fix our discretization points as the positions of the nuclei. This treatment is based on the assumption that diffusion takes place between nuclei and that, over the time period we are modeling, the positions of the nuclei are fixed. This leads to the following ODE:

$$\frac{dy_{a,i}}{dt} = S_{a,i}(Y) + D_a(y_{a,i+1} - 2y_{a,i} + y_{a,i-1}) - \lambda_a y_{a,i} \quad (3.2)$$

with the zero flux boundary conditions

$$y_{a,0}=y_{a,1}, \quad (3.3)$$

$$y_{a,n+1}=y_{a,n}, \quad (3.4)$$

where $Y \in \mathbb{R}^m$, a represents the specific mRNA or protein that the equation corresponds to, and D_a and λ_a are the corresponding diffusion and decay rates. We will assume that we are modeling a given set of mRNAs, $MR = \{m_1, m_2, \dots, m_\eta\}$, and a given set of proteins, $PR = \{p_1, p_2, \dots, p_\eta\}$, where $\eta = \frac{m}{2\mu}$. So, for all $a \in MR$, a represents a specific type of mRNA, and for all $a \in PR$, a represents a specific type of protein. i represents the nucleus (spatial) location.

The most important term in (3.2) is the synthesis term, $S_a(Y)$. For each nucleus i , $S_{a,i} : \mathbb{R}^m \rightarrow \mathbb{R}$. The function S_a represents the vector of these functions for the protein or mRNA, a , over all nuclei considered; hence $S_a : \mathbb{R}^m \rightarrow \mathbb{R}^\mu$ for $\mu = m$. This synthesis term holds all of the DNA-level information about transcription when $a \in MR$ and models translation when $a \in PR$. It is also the term that controls the overall system behavior, hence affecting the convergence and stability criteria for any numerical solver implemented.

First consider when $a \in MR$. For these terms, $S_a(Y)$ represents the scaled output from a thermodynamic-based model of the enhancer for gene a at each nucleus, i . Recall from section 1 that this function models the TF-DNA binding and TF-TF interactions regulating gene a 's expression level in a given cell. Thus, this term in the ODE is a rational function of the protein concentrations in nucleus i and takes the form

$$S_{a,i}(Y) = \chi_a \frac{\Phi_{a,i}(Y)}{1 + \Psi_{a,i}(Y)}, \quad (3.5)$$

where χ_a is the scaling constant associated with gene a and $\Phi_{a,i}$ and $\Psi_{a,i}$ are polynomials with coefficients depending on the parameter values. These parameter values include the binding affinities, K_p ; cooperativity values, C_{pq} ; and quenching parameters, Q_{pq} , for $p, q \in PR$.

Next consider when $a \in PR$. For these terms, $S_a(Y)$ represents translation. Here, we operate under the simplified assumption that newly translated proteins are produced at a rate proportional to the concentration of the corresponding mRNA in the vicinity of nucleus i . Thus, this term in the ODE is the product of the translation parameter, τ , and the mRNA concentration for the gene corresponding to protein $a = p_k$:

$$S_a(Y) = \tau Y_{m_k}, \quad (3.6)$$

where m_k denotes the mRNA for the gene corresponding to protein p_k . During implementation, the value of the translation parameter is fixed throughout the simulation,

and, regardless of the number of different simulations run, this value is bounded above and below.

Figure 3 illustrates the biological processes corresponding to each term in the ODE system for each case, $a \in MR$ and $a \in PR$.

4. Development of a semi-implicit numerical method

4.1. Construction

We begin with (3.2) for a given mRNA or protein, a , in the vicinity of nucleus i . To solve the ODE system, it is natural to use an implicit scheme for linear terms (diffusion and decay) and an explicit scheme for nonlinear terms (synthesis). This treatment allows us to exploit the simple structure created by discretizing diffusion. Approximating the left-hand side with a semi-implicit finite difference gives us

$$\frac{y_{a,i}^{n+1} - y_{a,i}^n}{h} = S_{a,i}(Y^n) = D_a(y_{a,i+1}^{n+1} - 2y_{a,i}^{n+1} + y_{a,i-1}^{n+1}) - \lambda_a y_{a,i}^{n+1}, \quad (4.1)$$

where h is the size of each time step and where superscripts n and $n+1$ denote the solution at that given time step. This can be rewritten as

$$-hD_y y_{a,i-1}^{n+1} + y_{a,i}^{n+1}(1 + 2hD_a + h\lambda_a) - hD_a y_{a,i+1}^{n+1} = y_{a,i}^n + hS_{a,i}(Y^n). \quad (4.2)$$

Now, considering all nuclei in the 1-D strip of interest and the zero flux boundary conditions, for a given mRNA or protein a , we get the equation

$$(I + hM_a)Y_a^{n+1} = Y_a^n + hS_a(Y^n), \quad (4.3)$$

where I is the identity matrix and M_a is the following matrix:

$$M_a(i, j) = \begin{cases} 2D_a + \lambda_a & \text{if } i=j, \\ -D_a & \text{if } i=j+1 \text{ or } i=j-1, \\ 0 & \text{otherwise.} \end{cases} \quad (4.4)$$

Note that M_a is tridiagonal, diagonally dominant, symmetric, and positive semidefinite. Thus, the existence of the inverse of $I + hM_a$ is guaranteed, and the algorithm to compute it is computationally very inexpensive. Hence, the solver is defined by the following iteration:

$$Y_a^{n+1} = (I + hM_a)^{-1}(Y_a^n + hS_a(Y^n)). \quad (4.5)$$

Note that this equation holds for each mRNA or protein, a . To obtain a general formula for the entire system, we can view our system as a block-system containing $2\eta = \frac{m}{\mu}$ of these blocks and define our solver in the following way:

$$Y^{n+1} = (I + hM)^{-1}(Y^n + hS(Y^n)), \quad (4.6)$$

where

$$Y^n = [Y_{m_1}^n, \dots, Y_{m_\eta}^n, Y_{p_1}^n, \dots, Y_{p_\eta}^n]^T, \\ S(Y^n) = [S_{m_1}(Y^n), \dots, S_{m_\eta}(Y^n), S_{p_1}(Y^n), \dots, S_{p_\eta}(Y^n)]^T,$$

and

$$M = \begin{bmatrix} M_{m_1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & M_{m_\eta} & 0 \\ \vdots & & 0 & M_{p_1} \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \dots & & 0 & M_{p_\eta} \end{bmatrix}.$$

Note that M is a block tridiagonal matrix. The computational complexity for solving (4.5) or (4.6) is of the same order as an explicit method.

4.2. Stability

Proofs of the lemmas and theorem stated in this section are available in the supplemental materials (section 7).

Lemma 1—Given a set of fixed parameter values and a fixed $a \in MR \cup PR$, there exists a constant, c_a , such that, for all $n < \infty$, $\|S_a(Y^n)\|_2 \leq c_a \|Y^n\|_2$.

Note that we are only interested in numerically solving our system in a finite time interval, $[0, T]$. Our solution is not guaranteed to be bounded for all time, but it is bounded on the finite time interval in which we are interested, as shown in the following lemmas.

Lemma 2—Given $T > h > 0$ and a set of fixed parameter values, there exists a constant, c , such that, for all $n \leq \frac{T}{h}$, $\|Y^n\|_2 \leq (1 + ch)^n \|Y^0\|_2$.

Lemma 3—Suppose that $Y^0 \in \mathbb{R}^m$ is a vector of initially normalized mRNA and protein concentrations, and that $\tilde{Y}^0 \in \mathbb{R}^m$ is a perturbed vector of initial values. Given $T > h > 0$, a set of fixed parameter values, and a fixed $a \in MR \cup PR$, there exists a constant, c_a , such that, $\forall n \leq \frac{T}{h}$, if Y^n and $\tilde{Y}^n$ are the corresponding numerical solutions at time step n , then $\|S_a(Y^n) - S_a(\tilde{Y}^n)\|_2 \leq c_a \|Y^n - \tilde{Y}^n\|_2$.

Theorem 4.1—Given a set of fixed parameter values, the numerical solution given by (4.6) is stable (in the sense of Lyapunov) [14] in a finite time interval $[0, T]$.

Knowing that our solver is stable gives us an idea of how robust the numerical solver is. We have found a constant, e^{cT} , which measures the effect of small perturbations in the initial data on the numerical solution. Now, we would like to determine how well our numerical solver actually approximates the true solution of the system and whether we can improve our approximations by decreasing the step-size, h . For this, we need to determine whether our solver is convergent, and we will use the stability we have just shown to help in proving convergence.

4.3. Convergence

A fundamental theorem of numerical solvers states that if a method is consistent and 0-stable, then the method is convergent [3, Theorem 3.1]. Since we have shown stability (in the sense of Lyapunov) in the previous section, a stronger condition than 0-stability, to prove convergence of the numerical solution given by (4.6) it suffices to show that the method is consistent.

Theorem 4.2—Given a set of fixed parameter values and a finite time interval $[0, T]$, the numerical solution given by (4.6) is consistent.

The proof of this theorem is available in section 7.

Now that we have shown stability and consistency of our numerical solver, it follows that the solver is convergent. Thus, as we take smaller and smaller step-sizes, i.e., as $h \rightarrow 0$, the numerical solution given by (4.6) converges to the true solution of (3.2).

5. Numerical simulations

Using the numerical solver as defined by (4.6), along with an appropriate parameter estimation algorithm, CMAES, we were able to fit our model, as described in (3.2)–(3.5), to expression data corresponding to a dynamically changing system involved in early eukaryotic development (system depicted in Figure 4) [22].

We used an 80 minute time window corresponding to the period of *Drosophila* development from late nuclear cycle 13, when the genes *twi*, *sna*, and *rho* begin to be expressed, to late nuclear cycle 14, when *twi*, *sna*, and *rho* expression patterns reach their final pregastrulation states. For fitting, we began with previously published quantitative data for mRNA concentrations corresponding to late cycle 14 ($t = 80$ min) [40]. 20 nuclear positions were used at approximately 50% egg length (anterior-posterior) from the ventral-midline to approximately 50% egg height (ventral-dorsal). We combined this with a qualitative assessment of blue-white *in situ* hybridization images generated in the Arnosti lab at Michigan State University at time points corresponding to late cycle 13 and mid-cycle 14 ($t = 0$ and $t = 30$ min in model simulations). We then linearly interpolated these three sets of data to create a set of data corresponding to 16 time points, in 5 minute intervals, to use for fitting model parameters. Initial mRNA concentrations used for *twi*, *sna*, and *rho* were those corresponding to the data at $t = 0$ min, and initial protein concentrations were set to 0 for Twi and Sna. Dorsal protein concentrations corresponding to late cycle 14 were also extracted from previously published data [40]. For Dorsal protein concentrations at earlier

time points, we followed the assumption that Dorsal concentration gradients follow a Gaussian curve with constant width throughout embryogenesis, although Dorsal levels are varying during each nuclear cycle [28]. We operated under the simplifying assumption that in late cycle 13 ($t = 0$ min) Dorsal concentration levels are approximately 30% of those which they reach in late cycle 14 ($t = 80$ min). Enhancer sequences were taken from the REDfly database [17], and the MAST program (part of the MEME suite) was used to predict the position and strength of binding sites [8].

We implemented the model and semi-implicit solver in C++ (source code available upon request) and combined the code with the covariance matrix adaptation evolution strategy (CMAES) written by Hansen in C (see [22]). In our model formulation, we assume that repressors act only on the most proximal activator proteins, so that when a state has a TF bound between a repressor and a particular activator, the closest TF is the only factor to experience the effects of the repressor. A repressor can interact with two TFs that are bound on either side. Similarly, we assume that cooperativity occurs only between adjacently bound TFs. These assumptions, as depicted in Figure 5, are made to simplify the calculations, but other scenarios are possible. The objective function used for parameter fitting was the root mean square error (RMSE) between mRNA concentrations obtained from model simulations and the experimental data. Figure 6 illustrates the ability of the model to successfully replicate the dynamic expression of a simple *Drosophila* gene regulatory circuit.

Comparison of the model's predicted results with the experimental data is a critical step in validation of the model. To measure the "goodness of fit," different thermodynamic-based modeling studies have used different functions and statistics. The two most common measures are RMSE and CC (Pearson correlation coefficient) [16, 15, 40, 35, 23]. These measures can be used to determine the quality of fit using a very quantitative pointwise comparison at each nuclear position and time point (RMSE) or a more qualitative comparison of the expression patterns at each time point (CC). The first can often be too stringent due to the inherent noise in immunofluorescent images, a common property of many quantitative data sets. On the other hand, a qualitative comparison is not adequate since our model aims to predict varying levels of expression over time. To avoid these two extreme cases, we used a third statistical method which was used previously to compare model predictions to expression patterns obtained experimentally, the Kolmogorov–Smirnov test (K-S test) for two independent samples [12]. The K-S test has the major advantage of being nonparametric and distribution-free, making no a priori assumptions regarding the distribution of the data sets. It tests the similarity between two data samples based on the null hypothesis: The distribution of data in sample 1 is consistent with the distribution of data in sample 2. Each of these three measures was calculated to compare the experimental data with the model prediction data shown in Figure 6 as well as model prediction data from a "poor" run of the model (one that resulted in an RMSE of 0.1883). The resulting values are shown in Table 5.1. Note that higher CC values correspond to samples with similarly shaped patterns, lower RMSE values correspond to samples with the smallest amount of pointwise absolute differences, and D values obtained from the K-S test need to be compared to the critical value, 0.1075, to determine significance. This value corresponds to a confidence

level $\alpha = 0.05$ and a sample size of 320. Any D values above the critical value of 0.1075 correspond to data sets in which the null hypothesis should be rejected (i.e., the distribution of data in sample 1 is not consistent with the distribution of data in sample 2).

One should first recognize that all three goodness-of-fit measures result in the same general conclusion: The expression patterns predicted as a result of the run of the model in which a lower overall RMSE was obtained (0.0616, referred to as the “Good Fit”) better fit the data than those resulting from runs for which a higher overall RMSE was obtained (0.1883, referred to as the “Poor Fit”). However, on a gene-by-gene basis, the results differ. First, using CC alone, one may believe that in the case of the Poor Fit, only the *rho* expression pattern is fitting the data poorly, since each of the other genes resulted in a $CC \approx 0.9$. This same conclusion would be drawn using the K-S test alone, since the D value obtained in the case of Poor Fit on *rho*, 0.1666, is the only D value obtained for these two data sets in which the null hypothesis should be rejected (i.e., the distribution of the experimental data for *rho* is not consistent with the model predictions for *rho*). If, instead, the RMSE is used to measure the goodness-of-fit for each individual gene, one can clearly see from the table that each gene’s expression fits the data much better in the case of the Good Fit compared to the Poor Fit (0.0518 vs. 0.1887, 0.0611 vs. 0.1697, and 0.0705 vs. 0.2034).

We believe that all three measures combined can lead to a deeper understanding of the predictive power of the model and the difference between these two scenarios (“Good” vs. “Poor” Fits). Both *twi* and *sna* genes have patterns of expression that are highest in the ventral regions of the embryo and monotonically decrease along the ventral-dorsal axis. Each of these genes’ expression is driven by Dorsal and/or Twist, which also have patterns of expression that are highest in the ventral regions of the embryo and monotonically decrease along the ventral-dorsal axis. Thus, there is a large variety of parameter values with the ability to predict *twi* and *sna* expression patterns with approximately correct shapes, although the timing and magnitude of their expression peaks may be incorrect. This may result in high CC values and low D values, although the predicted patterns are not precise enough to drive the proper expression of those genes that they regulate (i.e., *rho*). In the case of the “Good Fit,” which was depicted in Figure 6, all three goodness-of-fit measures result in the conclusion that all three genes’ expression patterns were correctly predicted by the model (i.e., resulted in high CCs, low RMSEs, and cannot be determined to be significantly different from the experimental data using the K-S test).

6. Concluding remarks

This model, as described in (3.2)–(3.5), is the first in a eukaryotic system to include a transcriptional term that takes into account DNA sequence information and a translational term to relate mRNA and protein concentrations. All simulation results shown were obtained using the numerical solver given by (4.6), which we have shown is both stable and convergent in the finite time interval in which simulations were run. Each simulation took under one minute to generate a numerical estimate for the 80 minute time period of interest, and parameter estimation using an evolutionary strategy took less than 24 hours. This is a notable improvement in computational speed over previous methods that did not incorporate any DNA-level information [24, 4, 5]. However, one should note that it has been reported

that a steady-state solution using a Bayesian approach has been computed in approximately “15 hours for 4283 genes,” while also incorporating *cis*-regulatory information such as “protein-DNA binding affinity” [26]. When comparing absolute runtime, one must consider the advantages and disadvantages of these two very different modeling approaches. Although we have modeled only a very small number of genes, the 24 hours of computational time allowed us to approximate not only a steady-state solution, but a solution fit to real experimental data that incorporates both protein-DNA binding and protein-protein interactions and illustrates the system dynamics over a crucial time period in development.

The results shown support our earlier claim that by incorporating data on DNA sequence we are able to successfully model context-specific features of enhancers, as well as replicate the dynamic expression of a simple *Drosophila* gene regulatory circuit that drives development in the dorsal-ventral axis of the blastoderm embryo, involving Dorsal, Twist, Snail, and *rhomboid*. When using such a model, our results have also highlighted the need for multiple measures of fitness. To truly understand a model’s predictions for an intricate network of genes, one must combine multiple measures of fitness to gain insight into what is driving the difference between a relatively good and relatively poor fit. Above all, where protein and *cis*-regulatory information is available, this model provides a powerful method to recapitulate and predict dynamic aspects of eukaryotic transcriptional systems that will greatly improve our understanding of gene regulation at a global level.

References

1. Ackers GK, Johnson AD, Shea MA. Quantitative model for gene regulation by λ phage repressor. *Proc Natl Acad Sci USA*. 1982; 79:1129–1133. [PubMed: 6461856]
2. Albert R, Othmer HG. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *J Theoret Biol*. 2003; 223:1–18. [PubMed: 12782112]
3. Ascher, UM.; Petzold, LR. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM; Philadelphia: 1998.
4. Ashyraliyev M, Jaeger J, Blom JG. Parameter estimation and determinability analysis applied to *Drosophila* gap gene circuits. *BMC Syst Biol*. 2008; 2:83. [PubMed: 18817540]
5. Ashyraliyev M, Siggins K, Janssens H, Blom JG, Akam M, Jaeger J. Gene circuit analysis of the terminal gap gene huckebein. *PLoS Comput Biol*. 2009; 5:e1000696.
6. Ay A, Arnosti DN. Mathematical modeling of gene expression: A guide for the perplexed biologist. *Crit Rev Biochem Mol Biol*. 2011; 46:137–151. [PubMed: 21417596]
7. Babloyantz A, Sanglier M. Chemical instabilities of “all-or-none” type in beta-galactosidase induction and active transport. *FEBS Lett*. 1972; 23:364–366. [PubMed: 11946654]
8. Bailey TL, Gribskov M. Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics*. 1998; 14:48–54. [PubMed: 9520501]
9. Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote sv40 DNA sequences. *Cell*. 1981; 27:299–308. [PubMed: 6277502]
10. Bieler J, Pozzorini C, Naef F. Whole-embryo modeling of early segmentation in *Drosophila* identifies robust and fragile expression domains. *Biophys J*. 2011; 101:287–296. [PubMed: 21767480]
11. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Kuhlman T, Phillips R. Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev*. 2005; 15:125–135. [PubMed: 15797195]

12. Chiu C, Fakhouri W, Liu N, Dayringer E, Dresch JM, Arnosti DN. A two-scale mathematical model for DNA transcription. *Math Biosci.* 2012; 236:132–140. [PubMed: 22343054]
13. Coppey M, Berezhkovskii AM, Kim Y, Boettiger AN, Shvartsman SY. Modeling the bicoid gradient: Diffusion and reversible nuclear trapping of a stable protein. *Dev Biol.* 2007; 312:623–630. [PubMed: 18001703]
14. Deuflhard, P.; Bornemann, F. *Scientific Computing with Ordinary Differential Equations.* Springer; New York: 2002. p. 124 and 3.2
15. Dresch JM, Liu X, Arnosti DN, Ay A. Thermodynamic modeling of transcription: Sensitivity analysis differentiates biological mechanism from mathematical model-induced effects. *BMC Syst Biol.* 2010; 4:142. [PubMed: 20969803]
16. Fakhouri WD, Ay A, Sayal R, Dresch J, Dayringer E, Arnosti DN. Deciphering a transcriptional regulatory code: Modeling short-range repression in the *Drosophila* embryo. *Molec Syst Biol.* 2010; 6:341. [PubMed: 20087339]
17. Gallo SM, Li L, Hu Z, Halfon MS. REDfly: A regulatory element database for *Drosophila*. *Bioinformatics.* 2006; 22:381–383. current version online at <http://redfly.ccr.buffalo.edu/>. [PubMed: 16303794]
18. Gelenbe E. Steady-state solution of probabilistic gene regulatory networks. *Phys Rev E.* 2007; 76:031903.
19. Gertz J, Siggia ED, Cohen BA. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature.* 2009; 457:215–218. [PubMed: 19029883]
20. Gregor T, Bialek W, de Ruyter van Steveninck RR, Tank DW, Wieschaus EF. Diffusion and scaling during early embryonic pattern formation. *Proc Natl Acad Sci USA.* 2005; 102:18403–18407. [PubMed: 16352710]
21. Griffith JA. Mathematics of cellular control processes. *J Theoret Biol.* 1968; 20:202–216. [PubMed: 5727239]
22. Hansen, N.; Ostermeier, A. Convergence properties of evolution strategies with the derandomized covariance matrix adaptation: The $(\mu/\mu, \lambda)$ -ES, in EUFIT'97. Proceedings of the 5th European Congress on Intelligent Techniques and Soft Computing; 1997; p. 650-654.
23. He X, Samee MA, Blatti C, Sinha S. Thermodynamics-based models of transcriptional regulation by enhancers: The roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol.* 2010; 6:e1000935. [PubMed: 20862354]
24. Jaeger J, Surkova S, Blagov M, Janssens H, Kosman D, Kozlov KN, Manu, Myasnikova E, Vanario-Alonso CE, Samsonova M, Sharp D, Reinitz J. Dynamic control of positional information in the early *Drosophila* embryo. *Nature.* 2004; 430:368–371. [PubMed: 15254541]
25. Janssens H, Hou S, Jaeger J, Kim A, Myasnikova E, Sharp D, Reinitz J. Quantitative and predictive model of transcriptional control of the *Drosophila* melanogaster even skipped gene. *Nature Genet.* 2006; 38:1159–1165. [PubMed: 16980977]
26. Kim H, Gelenbe E. Stochastic gene expression modeling with hill function for switch-like gene expression. *IEEE/ACM TCBB.* 2012; 9:973–979. [PubMed: 22144531]
27. Lettice LA, Horikoshi T, Heaney SJH, van Baren MJ, van der Linde HC, Breedveld GJ, Joosse M, Akarsu N, Oostra BA, Endo N, Shibata M, Suzuki M, Takahashi E, Shinka T, Nakahori Y, Ayusawa D, Nakabayashi K, Scherer SW, Heutink P, Hill RE, Noji S. Disruption of a long-range cis-acting regulator for *Shh* causes preaxial polydactyly. *Proc Natl Acad Sci USA.* 2002; 99:7548–7553. [PubMed: 12032320]
28. Liberman LM, Reeves GT, Stathopoulos A. Quantitative imaging of the dorsal nuclear gradient reveals limitations to threshold-dependent patterning in *Drosophila*. *Proc Natl Acad Sci USA.* 2009; 106:22317–22322. [PubMed: 20018754]
29. Nicolis, G.; Prigogine, I. *Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order through Fluctuations.* Wiley; New York: 1977.
30. Papatsenko D, Levine M. The *Drosophila* gap gene network is composed of two parallel toggle switches. *PLoS One.* 2011; 6:e21145. [PubMed: 21747931]
31. Petersen KB, Pedersen MS. *The Matrix Cookbook, Informatics and Mathematical Modeling.* Technical University of Denmark. 2008 section 10.4.
32. Reinitz J, Hou S, Sharp DH. Transcriptional control in *Drosophila*. *ComplexUs.* 2003; 1:54–64.

33. Sanchez L, Thieffry D. A logical analysis of the *Drosophila* gap-gene system. *J Theoret Biol.* 2001; 211:115–141. [PubMed: 11419955]
34. Santillan M, Mackey MC. Influence of catabolite repression and inducer exclusion on the bistable behavior of the lac operon. *Biophys J.* 2004; 86:1282–1292. [PubMed: 14990461]
35. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature.* 2008; 451:535–540. [PubMed: 18172436]
36. Shea MA, Ackers GK. The or control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Molec Biol.* 1985; 181:211–230. [PubMed: 3157005]
37. Stoer, J.; Bulirsch, R. *Introduction to Numerical Analysis.* 3. Springer; 2002. section 4.4
38. von Dassow G, Meir E, Munro EM, Odell GM. The segment polarity network is a robust developmental module. *Nature.* 2000; 406:188–192. [PubMed: 10910359]
39. Yuh CH, Bolouri H, Davidson EH. Cis-regulatory logic in the endo16 gene: Switching from a specification to a differentiation mode of control. *Development.* 2001; 128:617–629. [PubMed: 11171388]
40. Zinzen RP, Senger K, Levine M, Papatsenko D. Computational models for neurogenic gene expression in the *Drosophila* embryo. *Current Biol.* 2006; 16:1358–1365.

Fig. 1.

Illustration of a cis-regulatory element. The horizontal line represents a sequence of DNA. The arrow corresponds to the transcription start site, the nucleotide at which copying of DNA into mRNA commences. The blue region, just upstream of the transcription start site, depicts the region of DNA where regulatory proteins bind to control gene expression, the *cis*-regulatory element or enhancer. Green circles and red squares depict TFs (activators and repressors, respectively).

Fig. 2.

Illustration of enhancer region. The blue horizontal line represents the enhancer region, where regulatory proteins bind to control gene expression. Green circles and red squares depict TFs (activators and repressors, respectively). The order in which the binding sites are arranged on the enhancer is $A1$, R , $A2$.

Fig. 3.

Illustrations of the biological processes represented in the ODE system. Top: The equation corresponding to the case when $a \in MR$, representing the rate of change of mRNA concentration. An illustration is included for each term in the equation: the synthesis of new mRNA through transcription, mRNA diffusion, and mRNA decay. Bottom: The equation corresponding to the case when $a \in PR$, representing the rate of change of protein concentration. An illustration is included for each term in the equation: the synthesis of new protein through translation, protein diffusion, and protein decay.

Fig. 4.

Dynamic transcriptional network. On the left, each node corresponds to a TF or gene involved in the network. Dorsal (DI), present in the early *Drosophila* embryo, acts as the input to the network, activating (depicted by an arrow) all other genes in the network. As twist (Twi) is expressed, it in turn activates snail (Sna) and rhomboid (rho). Snail then represses (depicted by a blunt arrow) rhomboid. On the right, the corresponding final expression patterns in late nuclear cycle 14 *Drosophila* embryos are shown.

Fig. 5.

Assumptions on TF interactions. During model implementation, repressors are assumed to act only on the most proximal activator proteins bound. Therefore, we assume that the quenching activity of a repressor affects only an adjacently bound activator. Thus, the model considers only the quenching depicted here by solid black lines. The possible quenching activity shown by the dashed gray line in the top example is not included during modeling implementation. Similarly, cooperativity is assumed to occur between adjacently bound TFs only. Again, during implementation the model considers only the cooperativity depicted here by solid black lines, not the possible cooperative activity shown by the dashed gray line in the bottom example.

Fig. 6.

Data vs. model predictions. Eight time points are shown. Each pair of graphs depicts the data used to fit model parameters (left) and the model predictions after parameter fitting (right) for the given time point. Observe that the model is very successful in fitting the general dynamics and expression pattern trends of this system. The corresponding RMSE for this prediction is 0.0616.

Table 2.1

State contributions.

State (i)	C_i	S_i	Q_i
1	1	1	N/A
2	1	$K_{A1} [A1]$	1
3	1	$K_R [R]$	N/A
4	1	$K_{A2} [A2]$	1
5	1	$K_{A1}K_R [A1] [R]$	$Q_{A1,R}$
6	1	$K_RK_{A2} [R] [A2]$	$Q_{R,A2}$
7	$C_{A1,A2}$	$C_{A1,A2}K_{A1}K_{A2} [A1] [A2]$	1
8	1	$K_{A1}K_RK_{A2} [A1] [R] [A2]$	$Q_{A1,R}Q_{R,A2}$

Table 5.1

Statistical measures for goodness of fit corresponding to each of the three genes' expression patterns. These measures are shown in this table for two different runs of the model, one that resulted in a good fit and one that resulted in a poor fit. All of these measures were calculated by comparing the model output to the experimental data.

Goodness-of-fit measures				
	Gene	Pearson correlation coefficient	D value from the K-S test	RMSE
Good Fit	<i>twi</i>	0.9804	0.0126	0.0518
	<i>sna</i>	0.9793	0.0247	0.0611
	<i>rho</i>	0.9413	0.0670	0.0705
Poor Fit	<i>twi</i>	0.9139	0.0350	0.1887
	<i>sna</i>	0.8992	0.0811	0.1697
	<i>rho</i>	0.3187	0.1666	0.2034