Critical Reviews in Biochemistry and Molecular Biology, 2011; 46(2): 137–151

ISSN 1040-9238 print/ISSN 1549-7798 online
DOI: 10.3109/10409238.2011.556597

informa
healthcare

REVIEW ARTICLE

Mathematical modeling of gene expression: a guide for the perplexed biologist

Ahmet Ay[1,2] and David N. Arnosti[3]

[1]Department of Biology, Colgate University, Hamilton, NY, USA, [2]Department of Mathematics, Colgate University, Hamilton, NY, USA, and [3]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA

Abstract

The detailed analysis of transcriptional networks holds a key for understanding central biological processes, and interest in this field has exploded due to new large-scale data acquisition techniques. Mathematical modeling can provide essential insights, but the diversity of modeling approaches can be a daunting prospect to investigators new to this area. For those interested in beginning a transcriptional mathematical modeling project, we provide here an overview of major types of models and their applications to transcriptional networks. In this discussion of recent literature on thermodynamic, Boolean, and differential equation models, we focus on considerations critical for choosing and validating a modeling approach that will be useful for quantitative understanding of biological systems.

Keywords: Gene regulation, thermodynamic models, differential equation models, Boolean models, model selection, parameter estimation, sensitivity analysis, Drosophila melanogaster

Introduction

A confluence of high-throughput experimental techniques, expanding genomic information and a focus on systems biology, has lent new momentum to mathematical modeling of biological systems. Modeling gene regulation is central to such efforts because gene expression is at the nexus of many biological processes, and subtle changes of regulatory protein levels or links can underlie human diseases, population differences, and the evolution of morphological novelties (Carroll et al., 2001). Despite our expanding knowledge of the biochemistry of gene regulation, we lack a quantitative understanding of this process at a molecular level. Until recently, the preponderance of mechanistic gene regulatory studies have been empirically focused on the activity of individual transcriptional components, rather than generating an integrated picture of a system. In a recent shift, large-scale biological datasets have now provided a quantitative basis for systems biology studies. These data include complete genome sequences for many organisms, identification of many proteins and RNAs involved in the regulatory processes inside the nucleus, dynamic measurements of expression levels for many genes, and *in vivo* occupancy of the DNA by transcription factors, cofactors, and nucleosomes. However, even with these extensive datasets, the quantitative understanding of gene regulation is far from comprehensive, since the available data usually gives only an average of many cell states or a few snapshots of dynamic systems. Thus, obtaining a complete operational picture using solely experimental approaches is challenging. Mathematical modeling provides an alternative path for this key problem, offering new approaches that incorporate detailed dynamics of sets of biochemical interactions.

Mathematical modeling has been applied to biological systems for decades, but with respect to gene expression, too few molecular components have been known to build useful, predictive models. New efforts have been greatly aided by much more extensive "parts lists" of DNA sequences and proteins, as well as considerably

Address for Correspondence: David N. Arnosti, Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824-1319, USA. Tel: 517-432-5504. Fax: 517-353-9334. E-mail: arnosti@msu.edu

(Received 11 November 2010; revised 16 January 2011; accepted 18 January 2011)

enhanced computational power. These improvements make possible the use of diverse mathematical modeling methods for different biological problems. As more biologists venture into systems-level studies, a general understanding of diverse modeling approaches related to gene expression is necessary to facilitate close collaborations between experimentalists and modelers. Here, we focus on models developed for eukaryotic systems, discussing common approaches and their applications, and summarize goals, challenges, and future directions.

Gene regulatory models generally employ either statistical or analytical approaches. Both approaches can be highly effective in providing non-intuitive insights into gene regulatory systems. The first method is especially appropriate for datasets comprising the transcriptome, representing the expression levels of thousands of genes. Graph-based probabilistic models such as neural, Boolean, and Bayesian networks are used to represent regulatory interactions (Friedman et al., 2000; de Jong, 2002; Friedman, 2004; Filkov, 2005; Markowetz and Spang, 2007; Hecker et al., 2009). Under different conditions, statistical correlations highlight which groups of genes function together, indicating possible regulatory relations. As part of this analysis, shared motifs involved in transcriptional control can be identified. The ultimate goal of this approach is to discern regulatory networks that underlie the given data. Such statistical approaches provide a big picture covering large fractions of genes in an organism, but they cannot explain complex relations between transcription factors, polymerase, and other regulatory proteins, or the fine details of enhancer architecture. Since gene array data has been available for some time, statistical approaches are fairly well-developed and have been discussed in recent reviews (de Jong, 2002; Friedman, 2004; Markowetz and Spang, 2007; Hecker et al., 2009). In contrast, there is less uniformity in the types of analytical approaches applied for modeling gene regulation, and their value is less generally appreciated, especially in the area of DNA sequence-based modeling. Here we review the second, analytical approach, which generally focuses on expression of a small number of genes, and is represented by a variety of distinct mathematical models. The models can include terms relating to binding of transcription factors and RNA polymerase to the DNA, cooperative, and inhibitory interactions between transcription factors, mRNA and protein degradation, and mRNA translation rate. Unlike some statistical methods, for these approaches we need an extensive knowledge of system components and hypotheses about the system structure. Three major classes of mathematical models have been applied in such cases: thermodynamic, Boolean, and differential equation-based models. These models have been used to summarize experimental data (Yuh et al., 2001), to infer new relations from complex experimental data, guiding the researcher to new testable hypotheses (Jaeger et al., 2004a) and to find properties of the system that are hard to measure directly but can lead to accurate modeling of novel elements (Fakhouri et al., 2010). Several general features characterize these models. Most often models are deterministic, that is, the change to an independent variable has a predictable, reproducible impact on dependent variables, but they can also be structured as stochastic models to capture the erratic behavior of many biological systems influenced by intrinsic or extrinsic noise. Modeling approaches can also be categorized as discrete and continuous. Discrete forms, such as Boolean models, represent time, state, or space as a discrete set of values, simplifying calculations, although differential equation-based models utilize continuous values to provide a "smoother" representation of dynamical changes.

Here, we discuss the structure and applications of three major classes of models: thermodynamic, differential equation-based, and Boolean. The choice of which model to use usually depends on the system and problem under consideration. If successful, the model chosen should fit the existing data and give new biological insights on a system, not merely recapitulate what is already known.

Thermodynamic models

This modeling approach seeks to extract information about gene regulation from the sequences of *cis*-regulatory regions and the measured or inferred binding of sequence-specific transcription factors to these elements. That is, given a promoter and some well-characterized transcription factors, one strives to predict how a gene will be activated or repressed. These models predict how different combinations of binding sites on a regulatory region function together to give diverse temporal and spatial expression outputs, making the specific assumption that gene activity is proportional to the level of bound activators and inversely proportional to the level of bound repressors. Thermodynamic (also termed fractional occupancy) models are based on simple biophysical descriptions of DNA–protein interactions and statistical physics (Figure 1A). Current implementation ignores additional processes such as chromatin structure and modification, or DNA methylation, and does not independently treat recruitment of cofactors or the general transcription machinery, although these aspects may be incorporated in future models. This simplification does not appear to be a fatal flaw; the relative success of these models suggests that those events downstream of the primary DNA/protein interactions might therefore play lesser roles in determining the relationship between enhancer architecture and gene expression.

There are two basic steps to implementing such models. First, all possible states of the enhancer are listed, based on potential transcription factor–DNA interactions, with a statistical weight assigned to each state (Figure 1A). The probability of a gene firing is calculated as the fraction of the "successful" states, that is, those with preponderance of activators bound. For a simple regulatory region

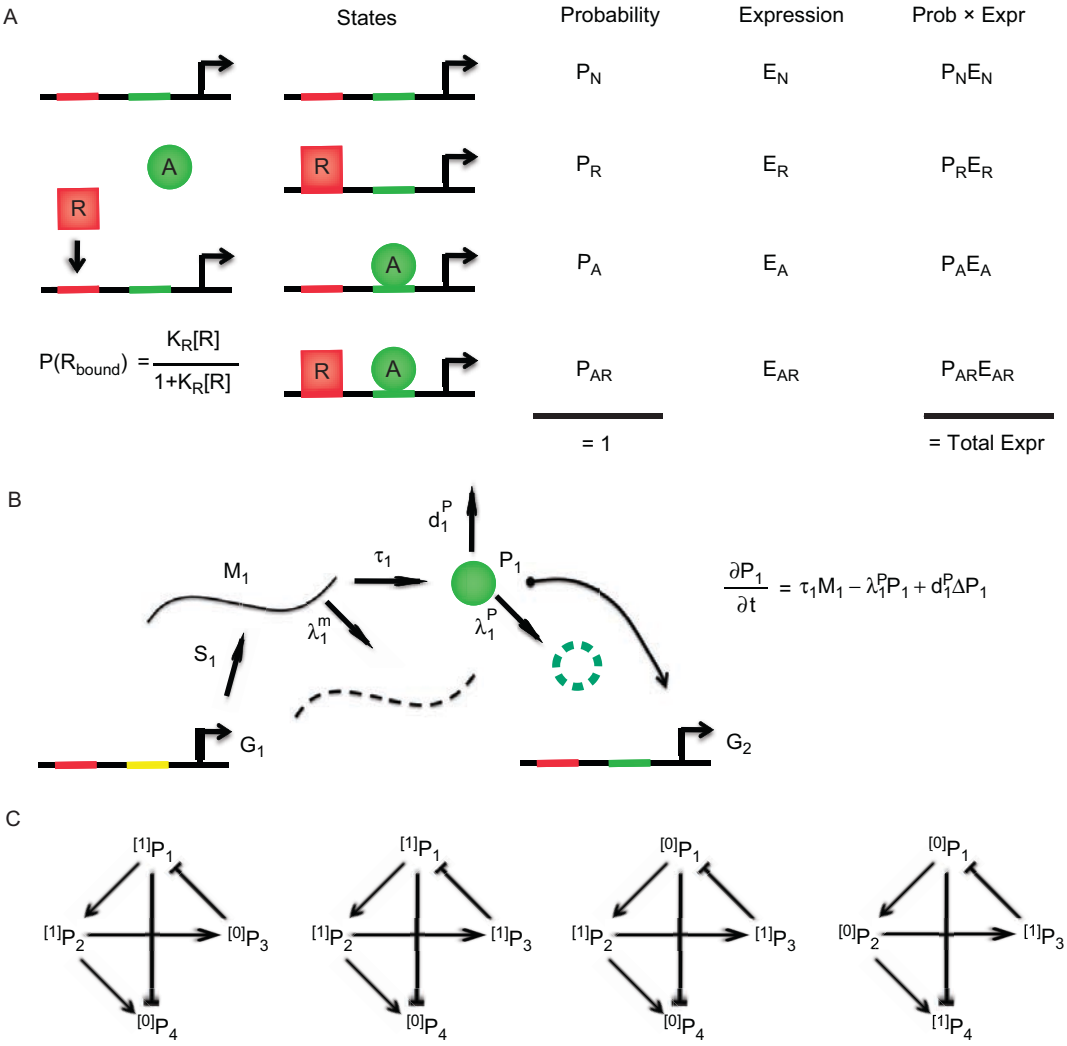

Figure 1. Analytical modeling approaches used in gene regulation studies. (A) Thermodynamic or fractional occupancy model of gene expression. The first column shows a simplified enhancer region with two binding sites for a repressor (R) and an activator (A). The mathematical formulation represents binding efficiency for the repressor site. In the second column, all four possible states of this enhancer region are shown. The third column represents the probability of this state occurring, which is not simply one-fourth but rather a function of the protein concentration and quality of the binding site(s). The fourth column indicates the efficiencies with which a particular state drives gene expression. This may be a simple additive expression of activators minus repressors, or a more complex expression. The last column represents the total expressions coming from each state (the probability that a state will occur multiplied by the potential of this configuration of proteins) and their summation, which provides a measure of the total output of the *cis*-element. (B) Differential equation model of gene expression. In this case, regulatory relationship between two genes is depicted. Synthesis of gene 1 (G_1) involves expression of mRNA (M_1) and translation of protein (P_1), which regulates gene 2 (G_2). Both mRNA and protein are subjects to turnover and protein is subject to diffusion. mRNA and protein synthesis, degradation, and diffusion events are shown at left. This process can be modeled with reaction diffusion equations as shown at right. Each molecular constituent is assigned such an equation. (C) Boolean model of gene expression. The network describing the regulatory relationships among four proteins is shown; the directed arrows show activation, and the blunt arrows show repression. Starting from initial state, three temporal steps are demonstrated. In this model, protein turnover occurs in one time interval and repression is assumed to be dominant over activation. Here, superscripts [1] and [0] indicate active or inactive states, respectively.

consisting of one binding site, there will be just two states, bound and unbound, although an element with four sites will have 16 states. A statistical weight for a state is calculated using the concentration of transcription factors and binding affinity of these factors to their sites on the DNA. For abundant proteins binding to high affinity sites, the weight will be much greater than cases where the transcription factor is scarce or the binding site is weak. The probability of each state can be calculated by dividing the statistical weight of the state by the sum of the statistical weight of all possible states. This calculation process can incorporate properties known to affect transcription. For example, cooperative and competitive interactions between transcription factors and inhibitory effects of repressors on activators can be explicitly added to the model by assigning higher or lower weights. The second step in thermodynamic modeling is to calculate gene expression output from each state. States with high

activator occupancy are likely to induce high expression, although repressor occupancy might result in low expression. Different approaches have been employed to convert occupancy to gene expression. As discussed below, one can model gene expression output as proportional to the binding probability of the RNA polymerase or weighted sum of the transcription factors (Bintu et al., 2005a, b; Segal et al., 2008; Fakhouri et al., 2010; He et al., 2010).

How has this approach been implemented in different areas? The theoretical underpinnings of thermodynamic modeling have been explored first and foremost in prokaryotic systems. Since the regulatory regions are generally small, binding to few transcriptional factors, simple bacterial systems provide a tractable setting for quantitative studies. The *lac* operon in *Escherichia coli* and the lysis/lysogeny switch of phage lambda are two examples that have been treated (von Hippel et al., 1974; Ackers et al., 1982; Shea and Ackers, 1985; Buchler et al., 2003; Vilar and Leibler, 2003; Bintu et al., 2005a). Additional promoters and configurations are considered in Bintu et al. (2005a, b). Zhou and Su generalized the results of Bintu et al. (2005a) to derive a single formula calculating transcriptional probability for all simple regulatory configurations. The model is available as a Python module, "tCal," which allows the user to easily build and configure transcription models of target genes (Zhou and Su, 2008). Although the use of thermodynamic modeling in simple prokaryotic systems has helped researchers establish and improve this modeling approach, the findings of these studies are usually not directly extendable to eukaryotic systems due to the fundamental differences in gene regulation mechanisms (Struhl, 1999).

In eukaryotes, complex *cis*-regulatory regions lend themselves to thermodynamic modeling, as this method offers the greatest potential to predict the function of different combinations of transcription factor-binding sites. Recent uses of thermodynamic models in yeast and *Drosophila* illustrate the possibilities and limitations of this approach (Granek and Clarke, 2005; Janssens et al., 2006; Zinzen et al., 2006; Segal et al., 2008; Gertz et al., 2009; Gertz and Cohen, 2009; Fakhouri et al., 2010; He et al., 2010).

To identify possible regulatory motifs in a set of promoters, thermodynamic analysis can detect degenerate binding sites that simple pattern searching may overlook. Granek and Clarke (2005) applied thermodynamic modeling to detect transcription factor targets in the yeast genome by using concentrations of transcription factors and their binding site preferences represented by position weight matrices (PWM). Their algorithm, GOMER, is unique in its ability to identify putative competitive and cooperative interactions between transcription factors from relatively sparse datasets, an approach that is difficult to carry out with machine-learning approaches. Using GOMER, they identified Fkh2 and Mcm1 targets controlling the expression of cell cycle-regulated genes, and analyzed the role of cooperativity in this process. They further investigated the role of competition between the Ndt80 and Sum1 transcription factors in distinguishing between mitotic and meiotic programs of gene regulation. The algorithm also predicted genome-wide binding of Rap1, which was confirmed by chromatin immunoprecipitation. These studies focused on yeast, but this algorithm could be applied to other regulatory systems.

The above approach focused on analysis of endogenous sequences binding particular transcription factors, providing a view of only a very small fraction of the potential arrangements that these regulatory factors might adopt on a promoter. In contrast, Gertz and Cohen analyzed large collections of synthetic promoters in yeast bound by random combinations of three to four transcription factors known to co-regulate genes in that organism (Gertz et al., 2009; Gertz and Cohen, 2009). They tested over 2800 promoters, not an exhaustive list of possible configurations, but several orders of magnitude greater than that afforded by only considering sets of co-regulated genes in a genome. The quantitative output of each promoter was assayed by means of a fluorescent reporter, and the activities were fit by a thermodynamic model (Shea and Ackers, 1985; Gertz et al., 2009; Gertz and Cohen, 2009). Their model, which was able to explain 44–59% of the variance in the gene expression driven by different promoter architectures, took into account cooperativity between transcription factor-binding sites, and the effects of weak binding sites. They then used the model to predict novel targets in the genome, including new targets of Mig1 that had been overlooked due to the low binding affinities of the sites. These two approaches are directed at unlocking a general transcriptional "grammar," which may be applicable to new genetic regulatory arrangements found in different species.

At the other end of the spectrum, thermodynamic modeling has also been applied to discover the detailed functioning of a single, complex regulatory region. Reinitz and colleagues modeled the activity of a 1.7-kb promoter proximal region of the *Drosophila melanogaster* even-skipped (*eve*) gene, which is expressed in seven stripes in the embryo. This region directs blastoderm expression of stripe 2, as well as weak expression of stripe 7. After careful observation of the expression directed by this DNA fragment, the authors incorporated the spatial and temporal expression levels of transcription factors regulating this gene into a thermodynamic model. Using only experimentally determined 17 binding sites for four transcription factors (those found using DNAse I footprinting), they were unable to recreate the expression patterns produced by reporter gene. However, when they included an additional set of bioinformatically predicted binding sites for three additional transcription factors, the model was able to fit the data. An important conclusion of this study is that widely dispersed binding sites may operate together to generate enhancer-like outputs, suggesting that not all developmental regulatory elements exist as compact modules. They extended the analysis by correctly predicting alterations in patterns induced

by mutation of specific binding sites or loss of specific transcription factors. This modeling effort provided quantitative support for a new picture of *cis*-regulatory regions, but the parameters found in such a study cannot be readily used for other enhancer regions, which limits the model's broader application. Modeling this single enhancer region containing potentially dozens of binding sites poses a challenge for parameter estimation, due to compensation between parameters. This effect arises because the data used in such a study is limited, thus particularly when there are many parameters it is likely that many combinations—reflecting completely different biological scenarios—will generate the same result. For example, an enhancer with a strong site for activator A and a weak site for activator B might be equivalent to one with a weak A and a strong B site. More experimental data would be essential to identify biologically correct values.

On a much larger scale, Segal and colleagues conducted a study that took advantage of high-quality quantitative data available in the *Drosophila* blastoderm embryo, extending the approach of Reinitz to 59 different enhancers (Segal et al., 2008). The dataset incorporated spatial expression data for eight transcription factors and expression of the target genes in mid-blastoderm embryos. Their model incorporated parameters for concentration scaling, homotypic (but not heterotypic) cooperative binding, and the expression contribution for each transcription factor. Unlike the *eve* promoter study, this model made no attempt to incorporate "quenching," that is the distance effect of short-range repressors, a critical feature of these proteins. Despite these simplifications, reasonable predictions are obtained for many of the enhancers. The study predicted that weak protein-binding sites contained within *cis*-regulatory modules make important contributions to total enhancer activity, as do homotypic cooperative interactions, which can provide sharper patterns at lower input concentrations. Their model generally predicts the expression patterns of the earlier expressed gap genes well, but is less successful with later expression patterns of pair-rule genes, possibly because heterotypic cooperative interactions and distance-dependent quenching are not considered. Both of these features are known to play key roles in many settings (Simpson-Brose et al., 1994; Szymanski and Levine, 1995; Arnosti et al., 1996).

A distinct approach to thermodynamic modeling was taken by Papatsenko and colleagues, who focused on gene regulatory rules relevant to enhancers driving neurogenic gene expression in the *Drosophila* embryo (Zinzen et al., 2006). The *rho*, *vnd*, and *vn* enhancers are regulated by two transcriptional activators Dorsal (Dl) and Twist (Twi), and one repressor Snail (Sna). Differences in the regulatory regions for these genes lead to slight differences in expression patterns in dorsal and ventral regions. This study applied thermodynamic modeling to *in silico* conceptual regulatory elements containing key core blocks of Dorsal–Twist–Snail (DTS) sites, rather than endogenous sequences used by Segal and Reinitz. Their model was able to produce patterns similar to those of the endogenous *rho*, *vnd*, and *vn* genes, and suggested that the structural features such as differences in cooperativity between transcription factors and numbers of DTS modules could explain the differences in expression between these genes. Parameter comparisons indicated that *rho* models require 5–10-fold higher Dl–Twi cooperativity than *vnd*, as well as higher Twi–Twi cooperativity, whereas models for *vnd* require more DTS modules and higher Sna–Sna cooperativity than do those for *rho*. Phylogenetic comparisons were employed to validate these conclusions: spacing between factor-binding sites is generally conserved, and the number of DTS modules in *vnd* is always more than in *rho*. Unlike the other examples discussed above, the function of modeled DNA sequences was not directly tested; however, most of the results of this article were consistent with earlier qualitative studies (Ip et al., 1992; Szymanski and Levine, 1995).

To reap the benefits of analyzing highly defined elements in a physiologically redundant context, a combination approach was recently employed in a study that modeled synthetic regulatory elements. Quantitative *in vivo* expression data was obtained from 27 synthetic enhancers devised to test features affecting repression in early *Drosophila* embryos (Fakhouri et al., 2010). Levels of reporter gene activity were measured by confocal laser scanning imaging of over 900 embryos, and quantitative differences resulting from subtle changes in enhancer structure were noted. To simplify the analysis, this study focused on specific features affecting repressors, so the arrangement and number of activator sites was held constant. Significantly, this application of modeling provided insights that were not apparent from the analysis of individual embryos, most notably a nonlinear function describing quenching effects of short-range repressors, and similar susceptibility to quenching of different activators. Extending these insights to the endogenous *rho* enhancer, the study showed that parameters learned from synthetic enhancers are directly applicable to natural enhancers, highlighting important features of the architecture of this enhancer. Earlier studies were based on analysis of structurally diverse enhancer sequences, making it difficult to identify important features of binding site composition in enhancers, knowledge that is a key to understanding enhancer evolution (Ludwig and Kreitman, 1995; Crocker et al., 2008). By focusing on a well-defined set of similar elements, Arnosti and colleagues were able to employ a model with a tractable number of parameters and robust estimation of these parameters.

These recent examples illustrate the applications of thermodynamic modeling in diverse contexts. Despite the incorporation of quantitative information about DNA sequence, transcription factor abundance, and binding affinity, this approach still neglects major features of the transcription process, such as nucleosome effects, orientation of binding sites, proximity to transcription start site, and chromatin modifications. Thermodynamic

models simplify these complexities by considering the process in up to three distinct layers; namely, the binding of transcription factors, the subsequent recruitment of cofactors, and the facilitation of transcription by these cofactors.

This process is illustrated by three steps in the model of Janssens et al. (2006): fractional occupancy of transcription factors (including the correction of activator occupancies due to quenching by short-range repressors, recruitment of cofactors (termed "adapters"), and calculation of transcription rate, here represented by an Arrhenius expression. In the first layer of their model, transcription factors bind to the DNA independently (i.e. no cooperative binding), and occupancy of activators is reduced when short-range repressors bind and quench them. Repression is represented by a multiplicative term, so that several repressors can act on the same activator, serially reducing its activity. Just as for activators, the potencies of repressors (or "scaling factor") are taken into account as free parameters. The second layer of this model describes cofactor recruitment by transcription factors, a crude simplification of the process, where each activator has a constant potential to recruit cofactors and all cofactors are equivalent. The third layer describes activation of transcription, in which cofactors lower the activation energy barrier, described by an Arrhenius expression. The model assumes cooperative effects between activators, generating a nonlinear activation response; at low levels this activity corresponds to observed biological properties of gene switches, but in this representation the signal increases exponentially as more cofactors are recruited, therefore an arbitrary maximum threshold level is set to limit transcription. The activation of transcription can also be described by other expressions to give sigmoidal shaped responses, such as logistic functions.

Other studies use a variation on this three-layer approach; Segal et al. (2008) allows homotypic cooperativity between transcription factors but leaves out distance-dependent repression. The second level involves summation of expression contributions of transcription factors (a parameterized feature), and although they do not refer to cofactor recruitment, it is logically parallel to the Janssens et al.'s (2006) study. The third layer, the calculation of transcription is represented by a sigmoidal function. Fakhouri et al. (2010) incorporate two additional features known to play important roles for these enhancers, namely short-range repression and heterotypic cooperativity. Not all models utilize a three-step approach. Zinzen et al. (2006) models only the first layer of transcription, the binding of transcription factors to the DNA, then assumes that the transcriptional level is linearly correlated to the level of active states, making an *ad hoc* assumption that an active enhancer must have at least one Dorsal and one Twist activator bound, and no Snail repressor. Cooperative binding for transcription factors is included, in contrast to the treatment of Janssens et al. (2006). Gertz et al. (2009) follow a similar approach, as do Granek and Clarke (2005), who also included weight functions for cooperativity and competition.

The diversity of implementation of these layers in thermodynamic models indicates the relatively undeveloped state of affairs optimizing such models; no study systematically considers the effect of different formulations on overall model robustness and accuracy. An additional major challenge in thermodynamic modeling is a prosaic yet fundamental one; the definition of functional binding sites. Transcription factors can tolerate high sequence variability, which gives a high flexibility to gene regulation; however, this flexibility makes the detection of binding sites a complicated task. The number of known binding sites experimentally is limited, and bioinformatic techniques do not guarantee accurate detection of binding sites, which limits modeling effectiveness. However, the knowledge gap in functional binding sites is beginning to close, as comprehensive surveys of binding preferences are implemented (Noyes et al., 2008; Zhu et al., 2009; Jaeger et al., 2010). Even so, *in vivo* binding often does not match with predictions because of poorly understood context-specific effects. In addition, the activities of transcription factors can exhibit context-dependent effects; for example, the Hunchback (Hb) protein can function as either an activator or repressor, depending on the enhancer bound (Schulz and Tautz, 1994; Simpson-Brose et al., 1994). This context dependency is not taken into account in recent studies, however; Hb is taken as an activator in Janssens et al. (2006) and a repressor in Segal et al. (2008). To better understand context effects, ChIP-sequence and transcriptome experiments may help to supply the necessary genomic information about *in vivo* binding and function (MacArthur et al., 2009).

Despite its shortcomings, for detailed analysis of transcriptional *cis*-element function, thermodynamic modeling represents the most biophysically informed approach that promises to decipher gene regulation at the DNA level. Current simplifications and unknown features limit its predictive power, but more powerful and complex models may be generated using better datasets such as *in vivo* transcription factors occupancy. Data limitations should not prevent mathematicians from creating new approaches, which can be tested on synthetic datasets and used to guide experimentalists. For a truly global understanding, thermodynamic models should be connected to network level modeling studies, a subject of future inquiries.

Differential equation models

Thermodynamic models are especially valuable at capturing the detailed, quasi-equilibrium activity of well-defined transcriptional elements. However, many biological problems demand a model that can represent a multicomponent, temporally evolving dynamic system. Here, differential equation models come to the fore. Regulatory networks can be represented by differential equations, in which a set of molecules such as mRNAs

and proteins interact by explicit rules defined in terms of rate equations. These equations specify the levels of each protein or mRNA as a function of the other components as the system evolves. These models usually include time- and/or space-dependent variables such as protein and mRNA concentrations, and parameters such as production and degradation rates (Figure 1B).

Differential equation models can be divided into two main groups: those using ordinary differential equations (ODE), which depend on a single variable such as time, and those using partial differential equations (PDE), which involve multiple variables such as time and space. ODEs are a well-studied field of mathematics; although they are generally hard to solve analytically (i.e. finding formulas that express the solutions as explicit functions), approximations of the solutions can be found by a variety of numerical methods, and convenient software tools are freely available. PDEs are also well-studied analytically and numerically, but PDE theory is more complex and computations are demanding. The difficulty of finding analytical solutions means that here, too, numerical simulations are the main analysis tools.

The initial gene regulatory systems, which the ODE models have been applied, are bacterial operons such as *lac* and tryptophan (*trp*). Each of these operons consists of structural genes and a small regulatory DNA region that controls gene expression by binding to transcriptional regulators, RNA polymerase, and in the case of *trp*, ribosomal interactions with leader mRNA. These operons have been studied extensively experimentally and quantitatively. Over 40 years ago, Goodwin developed the first mathematical model for operon dynamics, and then Griffith developed a more comprehensive analysis of simple inducible and repressible gene regulatory networks (Goodwin, 1965; Griffith, 1968a, b). The beauty of the problem attracted many additional researchers who developed more complicated models that took into account additional relevant processes to understand the dynamics of the *lac* and *trp* operons (Babloyantz and Sanglier, 1972; Nicolis and Prigogine, 1977; Bliss et al., 1982; Sinha, 1988; Sen and Liu, 1989; Yanofsky and Horn, 1994; Wong et al., 1997; Maffahy and Simeonov, 1999; Santillán and Mackey, 2001; Vilar et al., 2003; Yildirim and Mackey, 2003; Mackey et al., 2004; Santillán, 2008). These earlier studies are not necessarily directly generalizable to other biological systems, but they lay out general principles for analysis of regulatory regions in bacteria and eukaryotes. An advanced example is provided by Santillán and Mackey (2004), who presented a model of *lac* operon dynamics that combines the DNA level thermodynamic modeling with the transcription factor level differential equation modeling. The DNA level features such as known operators and the cooperativity among them were described by thermodynamic approach, and protein level features such as degradation and translation were described by ODE. Such models have not been developed for eukaryotic systems, but the direction shown by these prokaryotic studies is very valuable.

Modelers have made extensive use of differential equations for well-studied biological systems such as embryo patterning, and population and infection dynamics. The use of these models in eukaryotic gene regulatory networks is more recent, however, and the framework they provide is not familiar to many biologists who work in this field. As discussed below, however, high-quality datasets such as the segmentation network in *Drosophila* provide an excellent opportunity to make the use of these models (von Dassow et al., 2000; Jaeger et al., 2004a; Gregor et al., 2005).

Differential equation models have been applied to dynamic eukaryotic regulatory networks of varying levels of complexity, ranging from simple descriptions of a diffusible morphogen (e.g. Bicoid) to complex gene regulatory networks that incorporate cell-to-cell signaling. Anterior–posterior patterning of *D. melanogaster* is controlled by a gene regulatory cascade involving maternal, gap, pair-rule, and segment polarity genes (Rivera-Pomar and Jackle, 1996). This process is one of the best-studied developmental systems, with extensive information derived from genetics, genomics, and molecular biology about regulatory relationships, *cis*-regulatory elements, and signaling pathways. However, although extensive, these experimental studies have not been sufficient to provide a complete, quantitative, picture of this patterning process. Differential equation-based mathematical models have been used to provide a deeper level of understanding of this system. Here, we briefly describe the use of these models in settings of increasing complexity to make predictions about temporal and spatial changes in eukaryotic regulatory networks.

Morphogens are diffusible substances that trigger differential developmental responses depending on threshold concentrations. The Bicoid (Bcd) morphogen of *Drosophila* diffuses from the anterior region of the early blastoderm embryo to form a gradient that shapes the anterior–posterior axis (Ephrussi and St. Johnston, 2004; McGregor, 2005). The shape and stability of this gradient are usually assumed to be the result of localized production, diffusion, and degradation. Gregor and colleagues used differential equation-based reaction diffusion models to investigate the formation of this gradient in *Drosophila* species that have embryos of different sizes (Gregor et al., 2005). Their model describes the change in Bcd concentration over time as the protein diffuses and decays. Experimental measurements revealed minimal differences in diffusion constants, thus the model indicated that the nearly identical Bcd-driven patterns in very different sized embryos result from species-specific differences in the lifetime of the Bcd protein. Recent live-imaging experiments show that Bcd undergoes rapid nucleocytoplasmic shuttling and equilibrates between the cytoplasmic and nuclear compartments during rounds of mitosis (Gregor et al., 2007). Shvartsman and colleagues proposed a new ODE model to test whether the exponential shape of the Bcd gradient was strongly influenced by this process. Their model incorporates

constant localized production at the anterior pole of the embryo, and diffusion and nucleocytoplasmic shuttling in the presence of the growing number of nuclei (Coppey et al., 2007). The model predicts that nuclei do not contribute significantly to the shape of the Bcd gradient; rather the Bcd gradient is established before the nuclei migrate to the periphery of the blastoderm stage embryo and remains stable during subsequent nuclear divisions. The analysis of Shvartsman and colleagues suggests that the nuclear Bcd profile is robust; the model parameters do not have to be fine-tuned, and local defects in nuclear density should generate only local defects in the nuclear Bcd profile. Despite these important predictions, this model does not account for scaling of the gradient with the size of the embryo as the previously mentioned studies did.

The former studies focused on a single transcription factor, but differential equation models have also been applied to networks of interacting transcription factors. In several recent studies, Reinitz and colleagues modeled dynamic changes in the *Drosophila* gap gene network (comprising the gap proteins encoded by hunchback (*hb*), Kruppel (*Kr*), knirps (*kni*) and giant (*gt*), maternal factors bicoid (*bcd*), and caudal (*cad*) and the zygotic gene tailless (*tll*)) 1 h prior to cellularization and the pair-rule gene *eve* (Reinitz and Sharp, 1995; Jaeger et al., 2004a, b). Their model used reaction diffusion equations that incorporate synthesis, decay, and diffusion, and was based on a high-quality dataset describing concentrations of these proteins in the blastoderm embryo. This model reproduced gap gene expression with high accuracy, and agreed with earlier mutant and reporter gene studies. The study also suggested new regulatory interactions, such as the activation of *Kr* by Cad, and clarified the regulatory effects of Hb on *Kr*, Kr on *kni*, and Gt on *kni*. Some previously reported regulatory interactions were not necessary for good model fitting. This analysis suggests that although maternal factors drive initial activation of gap genes, the positioning and maintenance of gap gene boundaries depend mostly on interactions among gap genes (Jaeger et al., 2004a, b). Interestingly, diffusion is not critical for observing the dynamic shifts in gap gene expression (Jaeger et al., 2004a). This model did not satisfactorily predict the effects of null mutants, probably because of oversimplifications. The one-dimensional framework used assumes anterior–posterior genes are regulated independently of dorsal–ventral patterning networks, although this assumption is not entirely true (Keränen et al., 2006; Luengo Hendriks et al., 2006). In addition, the movement of nuclei, not included in the model, can also affect gap gene regulation. This facet is considered in a three-dimensional treatment of the problem (Keränen et al., 2006; Luengo Hendriks et al., 2006).

The differential equation-based model is also suitable for more complex settings that involve cell-to-cell communication and signaling cascades. Barkai and colleagues used reaction diffusion equations to describe how the TGF-β pathway regulates dorsal patterning in *Drosophila* (Eldar et al., 2002). Their model includes the TGF-β signaling molecules Scw and Dpp, the Dpp inhibitor Sog, and the protease Tld that cleaves Sog. Equations are included to account for the formation of the Dpp/Scw–Sog complex, diffusion of Sog, Dpp/Scw, and Dpp/Scw–Sog, and cleavage of Sog by Tld, both when Sog is free or in a complex. One interesting feature of this system is that except for *dpp*, the genes involved are recessive; thus, one-half dose is adequate to generate correct activities—this is the hallmark of a robust system. In the process of conducting 66,000 simulations, with parameters for rate constants and protein concentrations ranging over four orders of magnitude, they observed that only 198 yielded parameter sets that are robust to 2-fold changes in Sog, Tld, and Dpp/Scw, and showed the wild-type pattern. They found that robust networks could have a wide range of possibilities for most parameters, with two restrictions; cleavage of Sog by Tld is facilitated by the formation of the Sog–Dpp/Scw complex and the Dpp/Scw complexed to Sog is diffusible, although free Dpp/Scw is not. Their modeling suggested the transport of the Scw and Dpp into the dorsal midline by Sog, the inhibitor, was key to robustness.

In this former case, a great deal of experimental information about how factors interact was built into the model. Models can also be used to uncover such information. von Dassow and colleagues analyzed the establishment of segment polarity in the *Drosophila* embryo, in which a very stable differentiation state is determined by cell-to-cell interactions involving the Wnt and Hedgehog pathways (von Dassow et al., 2000). Their model had 48 parameters for binding rates, cooperativity coefficients, and half-lives of proteins and mRNAs; for the most part, the real values were unknown. Given realistic initial conditions, and using the known interactions in this network, the model did not reproduce the activity of the segment polarity genes and their products (von Dassow et al., 2000). However, by adding two new interactions, a positive feedback loop in the Wnt pathway and a negative interaction in the Hh pathway, they found many parameter sets that enable the model to reproduce the known robust behavior of the system. They also showed that robustness is not highly dependent on a single network topology; models with additional links and components maintain this property as long as the core topology remains the same. This robustness is also manifested by insensitivity to initial conditions, which the authors argue permits this circuitry to be easily adaptable to other systems or contexts.

Differential equation approaches are uniquely suited to capture the dynamical nature of biological systems. However, these models have important limitations. The quality and quantity of data needed to construct these models make them difficult to apply to poorly characterized systems. Addition of a new protein to a network may have profound effects, which may nonetheless be missed because of overfitting of the incomplete model (von Dassow et al., 2000). A significantly improved result is noted

between earlier and later efforts of Reinitz and colleagues, in which the same modeling and optimization techniques were used but increased data quality, allowed for much lower error levels and more precise parameter estimations (Reinitz and Sharp, 1995; Jaeger et al., 2004a, b). To make these models tractable, they are usually applied to smaller modules derived from a large regulatory network.

Even when extensive data is available for modeling, the often large number of parameters poses substantial computational challenges. For this reason, it is hard to scale up this approach to analyze complex regulatory networks with hundreds of interacting molecules. Although these systems are best treated with statistical methods, improvements in computational techniques may ameliorate of this problem (Friedman et al., 2000; de Jong, 2002; Friedman, 2004; Filkov, 2005; Markowetz and Spang, 2007; Hecker et al., 2009). Differential equation models also generally do not consider very fine-scale effects such as translational regulation or the sequence of transcriptional *cis*-regulatory elements. For this reason, these models do not provide insight into enhancer structure and organization, such as that provided by thermodynamic models. As such, differential equation models generally occupy a middle ground, providing an inroad into biological systems of moderate to high complexity, without the extreme detail of thermodynamic avenues, but with a reasonable ability to describe dynamical aspects that are lacking with other approaches.

Boolean models

Biological processes such as bacterial competence, apoptosis, and gene transcription often show on/off switch-like behavior. Boolean models, which represent regulatory relations as logic gates, can capture and describe this behavior. In this approach, the entities in the system such as mRNAs and proteins usually have two states on (1) or off (0) (Figure 1C). Logic gates such as "and," "or," and "not" are used to define the associations between the entities. For a gene that is regulated by two transcription factors, the AND function implies that the gene is transcribed only if both are bound, OR implies that the gene is transcribed if one of them is bound, and NOT implies that the gene is not transcribed if both are bound.

For any biological system where interactions between its elements are well-described, Boolean modeling can be used to combine qualitative experimental observations in a logical structure or to simulate the dynamic behavior of the system. Due to their simple nature, these models circumvent the need for quantitative details about the reactions of the biological systems, which makes Boolean models easy to analyze analytically, implement computationally, and extend to large-scale biological systems. Boolean models can thus provide a good starting point for systems where the details of the network are not clear; variants of the same network can be easily created and analyzed. Despite their simplicity, they can provide insights into the fundamental nature of the underlying system.

Boolean approaches to model gene regulation have been employed in a variety of settings (Sánchez and Thieffry, 2001; Yuh et al., 2001; Albert and Othmer, 2003). As discussed in the previous section, the *Drosophila* gap gene network has been studied by reaction diffusion models (Jaeger et al., 2004a, b). Sánchez and Thieffry (2001) took a Boolean approach to analyze the same network, simulating qualitative gap gene expression patterns in wild-type and mutant backgrounds. Typical for this approach, their Boolean model sums the regulatory inputs for a target and transforms them into logical output. To select parameter values, they dynamically analyzed the gap gene network, running iterative cycles in which the output of one run is fed back into the model for the next run, and arbitrarily accepted the smallest parameter values that generate correct expression states for wild-type and mutant phenotypes. To simplify the system further, they divided the embryo into four domains along the anterior–posterior axis, depending on the concentration levels of the maternal factors. Based on known experimental relationships and their modeling, they assigned different functional threshold levels to the proteins involved in the network; for instance, they assumed that Cad would activate Kni at the first threshold and Gt at the second. This study illustrated the ways a gap gene network functions to generate different patterns in response to the maternally provided Bcd, Cad, and Hb transcription factors, and provided insight on the most crucial interactions in the gap gene network, threshold levels for regulatory interactions, and the importance of cross regulation between gap genes in this network. For example, although cross-inhibition between gap genes was suggested as a critical mechanism for creating gap gene expression patterns, their analysis suggested that cross inhibitory interactions between *gt* and *Kr* but not other genes is critical (Rivera-Pomar and Jackle, 1996).

Although Boolean and differential equation-based models provide comparable level of understanding of the gap gene network, there are key differences (Sánchez and Thieffry, 2001; Jaeger et al., 2004a, b). The Boolean approach taken here discretizes the continuous protein concentrations of the embryo into just four functional thresholds, corresponding to location along the anterior–posterior axis. This simplification, although computationally advantageous, and capable of being implemented with low-resolution data, makes impossible the detailed modeling of gap gene network features, such as boundary sharpening. Comparing these two studies, there are several differences in predictions. First, a repressive feedback loop between *kni* and *hb* is reported as essential in Jaeger et al. (2004a, b) but not in Sánchez and Thieffry (2001), possibly due to the fact that the latter study did not take the posteriorly acting *tll* and *hb* repressor into consideration. Neglecting

this repressive circuitry is a choice made regarding the extent of genetics that the authors wished to model, and not related to the type of model selected. Second, the differential equation-based approach suggested that autoactivation is a critical component for sharpening gap domain boundaries, but because of the thresholding mentioned above, the Boolean model of Sánchez and Thieffry (2001) could not detect it. Finally, the Boolean study suggested that Hb may have both positive and negative regulatory roles, a possibility that was excluded by the formulation of the differential equation model. Thus, the main distinctions between these approaches appear to arise not from the overall approach but from the details of implementation.

As discussed above, the differential equation analysis of the *Drosophila* segment polarity network suggested that its robustness is due to the network's topology. A simpler Boolean approach, used to analyze this network, recapitulated the main conclusions of the earlier study, including an accurate prediction of the dynamics of this network (Albert and Othmer, 2003). Here, some simplifying assumptions are taken: inhibitors are always dominant over activators, mRNAs are translated into proteins in one time step, mRNAs decay completely in one time step if not transcribed, and proteins disappear after one time step if their mRNA is not present. A more refined two-step approach, in which the proteins degrade in two steps, did not change the main conclusions (Chaves et al., 2005). The study sought to find all possible steady states of the network, using early patterns of segment polarity genes as initial states and the stable later patterns as final states. Implementing this model, they found that after only six time steps, the expression pattern stabilized in a time invariant spatial pattern, a property of the endogenous gene circuit. The model's performance was measured by its prediction of spatial and temporal gene expression levels in the network, represented as present (1) or absent (0). They found 10 solutions, leading to six distinct steady states, one of which corresponded to wild-type patterns and two to known mutant patterns with either no stripes or broadened stripes. The existence of three additional steady states suggests that this network can produce patterns that are not accessed during normal development, but might be utilized in some other context. The assessment of potential initial conditions for each steady state indicated that the segment polarity network is robust and can correct errors in the initial expression patterns.

Their model gives several insights into the design of the segment polarity network. First of all, it suggests that the *wingless* gene is a key element in the network, and its initiation in the right pattern at the right time is vital. Although the studies of Albert and Othmer and von Dassow agree on robustness of the segment polarity network, they have employed slightly different networks, due to opposing observations on *en* inhibition (Cadigan et al., 1994; Aza-Blanc et al., 1997). The models differ in the implementation of inhibitory effects; the differential equation model allows inhibitory effects to particularly reduce the level of activation, but in the Boolean model the inhibitory effects are dominant and complete. This difference resulted in a large number of patterns with very broad *en* and *wg* stripes for even wild-type initial gene expression patterns in differential equation treatment of the network (von Dassow et al., 2000).

A distinct application of Boolean modeling to gene regulation was to model transcription of the *endo16* gene at the DNA level (Yuh et al., 2001). The *endo16* gene, which is expressed in the embryonic and larval midgut of the sea urchin, has a complex regulatory region that controls spatial and temporal expression. This gene, which has been analyzed in great detail experimentally, has served as a paradigm for how developmental enhancers process regulatory information. Extensive mutational analysis of this gene provided the experimental foundation for Davidson and colleagues to propose a Boolean model, describing the interactions between the regulatory elements of this gene (Yuh et al., 1998, 2001). A promoter proximal module A initiates early gene expression in the vegetal plate. Once gut differentiation begins, a more distal module B becomes the primary operating unit, transmitting its regulatory input to module A, which amplifies this input to drive the expression of *endo16*. The Boolean model also incorporated repressive contributions from additional modules DC, E, and F to module A, and complex interactions between module A and module B. Their Boolean model describes an internal switch in the *endo16* enhancer region, which moves the control from module A to module B. Their model not only allowed them to summarize the interactions and explain the control of expression changes in the *endo16* gene throughout embryogenesis, but also provided many testable predictions and predicted output of mutant regulatory elements.

Boolean models offer a simple and computationally facile approach to modeling gene regulation; however, the simplicity of these models can take a toll in accuracy of the results. If a system depends crucially on fine details of reaction rates, or concentrations of mRNAs or proteins, then Boolean models may fail to describe the system. As an example, if a gene is negatively regulating its own production, a Boolean model would generate oscillatory behavior, although in reality such a process may normally lead to steady state. For ambitious investigators, it is tempting to turn to a modeling approach that employs detailed biophysical descriptions relating to protein–DNA interactions, molecular turnover, or diffusion, with the thought that harnessing such detailed information must be advantageous. The record of Boolean models suggests that this simpler approach can be used for investigative purposes, especially for systems with poorly described parameters. Boolean modeling provides a mechanism to rapidly explore a wide variety of networks with the caveat that success here can be heavily affected by network architecture.

Operator's manual

Sensitivity analysis

Most biological modeling problems reduce to an inverse problem, where parameters in the model must be estimated. The models we examined here invariably contain unknown parameters. Parameters are often estimated to fit models to experimental data; this process has been used to infer biological features such as cooperativity, diffusion, and degradation rates. However, it is critical to realize that the parameter estimation process is influenced not only by the biological system but also by the form of the model itself. To understand how the structure of a model can influence the results, one must determine how strongly model outputs react to changes in parameter inputs. Those parameters that are highly sensitive are likely to be better fit and provide useful biological insights than those of low sensitivity. This process is termed sensitivity analysis, and it represents an essential step in implementing a model. Such analysis has been applied to diverse models, including differential equation and thermodynamic models (Gutenkunst et al., 2007; Ingalls, 2008; Dresch et al., 2010).

Sensitivity analysis can rely on either local or global approaches. Local methods focus on a specific point in parameter space and measure responses of the model to local parameter alterations, one parameter at a time. Local techniques usually use partial derivatives to determine how the model output changes with respect to small variations of a particular parameter (Erb and Michaels, 1999; Reeves and Fraser, 2009). Despite their simple formulation, easy implementation, and interpretation, local methods do not provide a complete picture of parameter space, and these methods do not account for interactions between parameters, resulting in an underestimation of true model sensitivities in nonlinear models (Frey and Patil, 2002; Tang et al., 2006; Marino et al., 2008; Ziehn and Tomlin, 2008). Due to the large uncertainty of parameter values in many biological studies, local methods are not robust, so global sensitivity techniques are preferred. These methods of analysis try to capture the entire parameter space simultaneously, exploring multiple parameter values at the same time. Global approaches vary values over a larger parameter space and have the ability to quantify parameter interactions (Frey and Patil, 2002; Tang et al., 2006; Marino et al., 2008; Ziehn and Tomlin, 2008). A common element in all global sensitivity methods is the exploration of the full parameter space and quantification of model's sensitivity to perturbations in each parameter.

A recent study examined thermodynamic models used for mathematical modeling of gene expression, and applied sensitivity analysis to determine the significance of estimated parameter values for thermodynamic models (Dresch et al., 2010). Both biological effects and mathematical structures contributed to the observed differential sensitivity of the parameters. In one case, the parameters that were thought to prove the system's dependence on activator–activator cooperativity were found to be relatively insensitive, indicating that the biological relevance of this aspect of the model is weak. Thus, modelers should recognize a need for sensitivity analysis; discerning the differential effects distinct parameters have on the model output can help modelers to choose the model formulation that best fits their biological system, as well as to highlight which experiments will be most informative for modeling. Although sensitivity analysis may be useful in determining a correct parameter estimation strategy, there is not a single generally accepted approach for the diverse models and systems under consideration. As we discuss below, there are different parameter estimation methods that are customized for different problem types (Mendes and Kell, 1998).

Parameter estimation

The choice of a suitable parameter estimation technique and its validation determines whether and how efficiently a problem can be solved. Surprisingly, this issue has not been systematically explored in recent efforts of transcriptional modeling (Janssens et al., 2006; Zinzen et al., 2006; Segal et al., 2008; He et al., 2010). To estimate the parameters that fit a model to experimental data, one uses an objective function to measure the model performance, often the sum of squares of the residuals between the model's prediction and experimental data. In brief, parameter estimation algorithms start with an initial guess and iteratively generate new estimates seeking to minimize the error, until they stop at a solution. Different parameter estimation techniques employ distinct iteration processes, depending on the objective function and constraints of the parameters. In biological models, objective functions and the constraints are often nonlinear so that the objective function might have many local optima. One strives to identify the global optimum solution among the set of all possible solutions.

Parameter estimation approaches can use either local or global techniques. Local techniques include the conjugate gradient method, Newton's method, and the Nelder–Mead simplex method (Nocedal and Wright, 1999; Madsen et al., 2004). These local techniques rely on either gradient-based approaches, which require the calculation of the objective function's derivative or its approximation, or direct search methods that are based on a comparison of the objective function at the vertices of a simplex, a geometrical shape that has N+1 vertices and the edges between them in an N-dimensional space, in the space of parameters. Since simplex methods only require the evaluation of the function at some parameter values, they are more suitable for problems with discontinuous, nondifferentiable, or nonlinear objective functions. In these methods, the search usually continues iteratively by shrinking the simplex size (if possible) toward the lowest scoring parameter set until one of the stopping criteria of

the problem is satisfied. Local parameter estimation techniques often find local rather than global optima, unless there is good prior knowledge about the parameters involved (Mendes and Kell, 1998). To overcome this problem, local parameter estimation techniques can be used repeatedly, starting from different initial parameter values, but this approach is not very efficient (Mendes and Kell, 2001).

Global parameter estimation techniques are better than their local counterparts in finding the global optimum of the system and have been shown to be more suitable for biological systems (Mendes and Kell, 1998, 2001; Moles et al., 2003). Global parameter estimation techniques include deterministic strategies such as the branch and bound method, and interval optimization, as well as stochastic strategies such as genetic algorithms, simulated annealing, and evolutionary strategies (Weise, 2009). Deterministic methods, although computationally more expensive, are superior at finding the global optimum. Stochastic models are less likely to find the global optimum, but they can reach the vicinity of the real solution in a reasonable amount of time. Banga and colleagues compared a number of parameter estimation algorithms for their efficiency and reliability, including deterministic and stochastic methods. In this case, the stochastic evolutionary strategy approach functioned best on their continuous problem with a high number of unknowns (Moles et al., 2003). In another comparison of parameter estimation methods, Mendes and Kell (1998) tested a model of the irreversible inhibition of HIV proteinase, which involved 20 parameters. The global-simulated annealing algorithm functioned the best; however, a local parameter estimation technique that ran 750-fold faster achieved comparable results in this case. However, the success of local methods is not guaranteed, so researchers often prefer to use a more robust global algorithm for biological problems. To facilitate computation, hybrid models combining global and local techniques have been used frequently (Gursky et al., 2004). In addition, many global methods have parallel versions that speed the computational process.

The demands on modeling are only going to grow, as biologists seek to assimilate new genome-scale data. Parameter estimation algorithms must be robust independent of starting point, computationally efficient, and accurate, that is, not sensitive to errors in the data. To select the best approach, it is critical to have more comparative studies that test the performance of different parameter estimation approaches including with respect to gene regulation models. Previous studies suggest that there is no single best algorithm that works well for all problems (Mendes and Kell, 1998). Initial studies have begun to compare the performance of parameter estimation techniques for gene regulation models. The differential equation model of the *Drosophila* gap gene network has been used to compare the performance of simulated annealing versus evolutionary strategy methods. The latter approach was more computationally efficient and produced similar results (Jaeger et al., 2004a, b; Perkins et al., 2006; Fomekong-Nanfack et al., 2007; Jostins and Jaeger, 2010). Other studies have compared genetic algorithms, simulated annealing, and evolutionary strategy, but such an analysis is far from standard in most studies (Fakhouri et al., 2010). Clearly, more effort is required in this area.

Model selection

Currently, global studies have generated massive genomic, transcriptomic, and proteomic datasets, and analysis of these diverse data groups using quantitative modeling will be key for comprehensive understanding of biological systems. Efficient selection of the most appropriate model is critical to tap the information in these datasets. We consider here several mathematical approaches that, in various guises, have been applied to model gene expression (Table 1). As noted above, the most popular general methods are statistical; they can efficiently draw attention to correlations of biological significance. For a more detailed mechanistic view, one may choose one of the three general approaches discussed here. To proceed, one must have a good grasp of the main system components and if possible, their interactions. Differential equation models are suited to the dynamic nature of biological systems, but these models are not easy to apply to large-scale networks (von Dassow et al., 2000; Jaeger et al., 2004a). Thermodynamic models are particularly useful to explore features of enhancer architecture, which is treated as a black box by differential equation models. In the future, it may be possible to combine differential equation and thermodynamic models to achieve dynamic understanding of systems that include *cis*-regulatory details. Boolean models also allow simple, computationally efficient, and scalable treatment of regulatory modules although at the cost of less detail (Sánchez and Thieffry, 2001; Yuh et al., 2001).

Whatever the model selected, high-quality data is key in quantitative analysis of biological systems. Models restricted by limited data can usually only reproduce the experimental observations without providing any new insights. This requirement makes the modeling studies extremely challenging, since the data collection from biological systems is usually expensive, limited, and subject to noise. A new challenge is also how to merge large-scale "omic" data types. Statistical approaches have successfully merged bioinformatic and transcriptomic data to discern regulatory networks, but apart from thermodynamic analysis of DNA sequences and protein levels, the modeling approaches described here have not yet integrated all available types of data (Tavazoie et al., 1999; Segal et al., 2003).

The era of systems biology has brought together people from diverse disciplines for collaborative studies. Quantitative modeling, including mathematical modeling of gene expression, calls for appropriate

Table 1. Summary of the properties of mathematical models of gene regulation.

	Thermodynamic	Differential equation	Boolean
1. Context in which the models are applied	Models gene regulation at the DNA level, taking into account quality and arrangement of binding sites. Employs biophysical descriptions to describe protein–DNA interactions.	Models regulatory interactions without taking DNA sequences into consideration; DNA level events are treated as a black box.	Usually has been applied to model gene regulation without taking DNA sequence into consideration.
2. Nature of experimental data required	Quality and quantity of data needed is high, usually quantitative spatial or temporal data.	Quality and quantity of data needed is high, usually quantitative spatial and/or temporal data.	Qualitative experimental observations might be sufficient. Models can provide a good starting point when structure of a network is not clear.
3. Describing system dynamics	By itself, cannot describe dynamical nature of biological systems. System is assumed to be at quasi-equilibrium.	Is well-suited to capture the dynamic nature of biological systems.	Can be used to capture the dynamic nature of biological systems.
4. Discrete or continuous?	No time component; represents state or space as continuous.	Represents time, state, or space as continuous values.	Represents time, state, or space as discrete values.
5. Quality of the results	Depending on model implementation and data, results can be better quality.	Depending on model implementation and data, results can be better quality.	Tend to produce lower quality results. The results may be inaccurate if the behavior of the system depends on the fine details.
6. Model analysis and its extension	Easy to analyze analytically and implement computationally; however, hard to scale up to study large number of regulatory elements.	Not easy to analyze analytically or implement computationally. Extension to large-scale biological systems is challenging.	Easy to analyze analytically, implement computationally, and extend to large-scale biological systems.

Modeling approaches are compared according to the areas in which they are used, the quality of the experimental data needed, utility in describing the dynamic nature of biological systems, their discrete or continuous natures, the quality of the results, and ease of model analysis and extension to larger systems.

applications that are easy to understand, code, and implement by researchers. Computational efficacy is just one important criterion for model selection, affecting model applicability and extendability. Successfully implemented models should also be consistent with the data, reflect essential system properties, and should help answer specific questions about the system to generate new hypothesis. To attain these goals, modeling studies require tight coordination between experimentalists and modelers that begins before the execution of large-scale projects. Possible model types can be tested before experiments are performed, using synthetic datasets to measure model applicability. In this way, models can be redesigned to reduce bias toward certain features of the system and equalize sensitivity of the various parameters. After the experimental data has been obtained and the model parameters extracted, sensitivity analysis can once again be used to provide additional information about the robustness of biological conclusions. Although successful modeling requires substantial multidisciplinary efforts and represents a leap into the unknown for many biologists, this approach provides unique and powerful avenues to obtain a comprehensive understanding of biological systems.

Acknowledgement

We would like to thank Jackie Dresch, Jacob Clifford, Rupinder Sayal, Yerzhan Suleimenov, and members of the Arnosti laboratory for thoughtful discussions.

Declaration of interest

The authors declare to have no conflict of interest. The authors alone are responsible for the content and the writing of the manuscript. This project was supported by a Special Projects Grant from the MSU Foundation and NIH GM 56976 to D.N.A., and fellowships from the MSU Quantitative Biology Initiative and Gene Expression in Development and Disease focus group to A.A.

References

Ackers GK, Johnson AD, Shea MA. 1982. Quantitative model for gene regulation by lambda phage repressor. Proc Natl Acad Sci USA 79:1129–1133.

Albert R, Othmer HG. 2003. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. J Theor Biol 223:1–18.

Arnosti DN, Gray S, Barolo S, Zhou J, Levine M. 1996. The gap protein knirps mediates both quenching and direct repression in the *Drosophila* embryo. EMBO J 15:3659–3666.

Aza-Blanc P, Ramírez-Weber FA, Laget MP, Schwartz C, Kornberg TB. 1997. Proteolysis that is inhibited by hedgehog targets *Cubitus interruptus* protein to the nucleus and converts it to a repressor. Cell 89:1043–1053.

Babloyantz A, Sanglier M. 1972. Chemical instabilities of "all-or-none" type in beta-galactosidase induction and active transport. FEBS Lett 23:364–366.

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R. 2005a. Transcriptional regulation by the numbers: models. Curr Opin Genet Dev 15:116–124.

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Kuhlman T, Phillips R. 2005b. Transcriptional regulation by the numbers: applications. Curr Opin Genet Dev 15:125–135.

Bliss RD, Painter PR, Marr AG. 1982. Role of feedback inhibition in stabilizing the classical operon. J Theor Biol 97:177–193.

Buchler NE, Gerland U, Hwa T. 2003. On schemes of combinatorial transcription logic. Proc Natl Acad Sci USA 100:5136–5141.

Cadigan KM, Grossniklaus U, Gehring WJ. 1994. Localized expression of sloppy paired protein maintains the polarity of *Drosophila* parasegments. Genes Dev 8:899–913.

Carroll SB, Grenier JK, Weatherbee SD. 2001. From DNA to Diversity. MA, USA: Blackwell Science.

Chaves M, Albert R, Sontag ED. 2005. Robustness and fragility of Boolean models for genetic regulatory networks. J Theor Biol 235:431–449.

Coppey M, Berezhkovskii AM, Kim Y, Boettiger AN, Shvartsman SY. 2007. Modeling the bicoid gradient: diffusion and reversible nuclear trapping of a stable protein. Dev Biol 312:623–630.

Crocker J, Tamori Y, Erives A. 2008. Evolution acts on enhancer organization to fine-tune gradient threshold readouts. PLoS Biol 6:e263.

de Jong H. 2002. Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol 9:67–103.

Dresch JM, Liu X, Arnosti DN, Ay A. 2010. Thermodynamic modeling of transcription: sensitivity analysis differentiates biological mechanism from mathematical model-induced effects. BMC Syst Biol 4:142.

Eldar A, Dorfman R, Weiss D, Ashe H, Shilo BZ, Barkai N. 2002. Robustness of the BMP morphogen gradient in *Drosophila* embryonic patterning. Nature 419:304–308.

Ephrussi A, St Johnston D. 2004. Seeing is believing: the bicoid morphogen gradient matures. Cell 116:143–152.

Erb RS, Michaels GS. 1999. Sensitivity of biological models to errors in parameter estimates. Pac Symp Biocomput 4:53–64.

Fakhouri WD, Ay A, Sayal R, Dresch J, Dayringer E, Arnosti DN. 2010. Deciphering a transcriptional regulatory code: modeling short-range repression in the *Drosophila* embryo. Mol Syst Biol 6:341.

Filkov V. 2005. Identifying gene regulatory networks from gene expression data. In: Aluru S, ed. Handbook of Computational Molecular Biology. FL, USA: Chapman & Hall/CRC Press, pp. 708–736.

Fomekong-Nanfack Y, Kaandorp JA, Blom J. 2007. Efficient parameter estimation for spatio-temporal models of pattern formation: case study of *Drosophila melanogaster*. Bioinformatics 23:3356–3363.

Frey HC, Patil SR. 2002. Identification and review of sensitivity analysis methods. Risk Anal 22:553–578.

Friedman N. 2004. Inferring cellular networks using probabilistic graphical models. Science 303:799–805.

Friedman N, Linial M, Nachman I, Pe'er D. 2000. Using Bayesian networks to analyze expression data. J Comput Biol 7:601–620.

Gertz J, Cohen BA. 2009. Environment-specific combinatorial *cis*-regulation in synthetic promoters. Mol Syst Biol 5:244.

Gertz J, Siggia ED, Cohen BA. 2009. Analysis of combinatorial *cis*-regulation in synthetic and genomic promoters. Nature 457:215–218.

Goodwin BC. 1965. Oscillatory behavior in enzymatic control processes. Adv Enzyme Regul 3:425–438.

Granek JA, Clarke ND. 2005. Explicit equilibrium modeling of transcription-factor binding and gene regulation. Genome Biol 6:R87.

Gregor T, Bialek W, de Ruyter van Steveninck RR, Tank DW, Wieschaus EF. 2005. Diffusion and scaling during early embryonic pattern formation. Proc Natl Acad Sci USA 102:18403–18407.

Gregor T, Wieschaus EF, McGregor AP, Bialek W, Tank DW. 2007. Stability and nuclear dynamics of the bicoid morphogen gradient. Cell 130:141–152.

Griffith JS. 1968a. Mathematics of cellular control processes. I. Negative feedback to one gene. J Theor Biol 20:202–208.

Griffith JS. 1968b. Mathematics of cellular control processes. II. Positive feedback to one gene. J Theor Biol 20:209–216.

Gursky VV, Jaeger J, Kozlov KN, Reinitz J, Samsonov AM. 2004. Pattern formation and nuclear divisions are uncoupled in *Drosophila* segmentation: comparison of spatially discrete and continuous models. Physica D 193:286–302.

Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. 2007. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput Biol 3:1871–1878.

He X, Samee MA, Blatti C, Sinha S. 2010. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput Biol 6:e1000935.

Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. 2009. Gene regulatory network inference: data integration in dynamic models—a review. BioSystems 96:86–103.

Ingalls B. 2008. Sensitivity analysis: from model parameters to system behaviour. Essays Biochem 45:177–193.

Ip YT, Park RE, Kosman D, Bier E, Levine M. 1992. The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the *Drosophila* embryo. Genes Dev 6:1728–1739.

Jaeger J, Surkova S, Blagov M, Janssens H, Kosman D, Kozlov KN, Manu, Myasnikova E, Vanario-Alonso CE, Samsonova M, Sharp DH, Reinitz J. 2004a. Dynamic control of positional information in the early *Drosophila* embryo. Nature 430:368–371.

Jaeger J, Blagov M, Kosman D, Kozlov KN, Manu, Myasnikova E, Surkova S, Vanario-Alonso CE, Samsonova M, Sharp DH, Reinitz J. 2004b. Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster*. Genetics 167:1721–1737.

Jaeger SA, Chan ET, Berger MF, Stottmann R, Hughes TR, Bulyk ML. 2010. Conservation and regulatory associations of a wide affinity range of mouse transcription factor binding sites. Genomics 95:185–195.

Janssens H, Hou S, Jaeger J, Kim AR, Myasnikova E, Sharp D, Reinitz J. 2006. Quantitative and predictive model of transcriptional control of the *Drosophila melanogaster* even skipped gene. Nat Genet 38:1159–1165.

Jostins L, Jaeger J. 2010. Reverse engineering a gene network using an asynchronous parallel evolution strategy. BMC Syst Biol 4:17.

Keränen SV, Fowlkes CC, Luengo Hendriks CL, Sudar D, Knowles DW, Malik J, Biggin MD. 2006. Three-dimensional morphology and gene expression in the *Drosophila* blastoderm at cellular resolution II: dynamics. Genome Biol 7:R124.

Ludwig MZ, Kreitman M. 1995. Evolutionary dynamics of the enhancer region of even-skipped in *Drosophila*. Mol Biol Evol 12:1002–1011.

Luengo Hendriks CL, Keränen SV, Fowlkes CC, Simirenko L, Weber GH, DePace AH, Henriquez C, Kaszuba DW, Hamann B, Eisen MB, Malik J, Sudar D, Biggin MD, Knowles DW. 2006. Three-dimensional morphology and gene expression in the *Drosophila* blastoderm at cellular resolution I: data acquisition pipeline. Genome Biol 7:R123.

MacArthur S, Li XY, Li J, Brown JB, Chu HC, Zeng L, Grondona BP, Hechmer A, Simirenko L, Keränen SV, Knowles DW, Stapleton M, Bickel P, Biggin MD, Eisen MB. 2009. Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biol 10:R80.

Mackey MC, Santillán M, Yildirim N. 2004. Modeling operon dynamics: the tryptophan and lactose operons as paradigms. C R Biol 327:211–224.

Madsen K, Nielsen HB, Tingleff O. 2004. Methods for Non-Linear Least Squares Problems. Informatics and Mathematical Modelling, Technical University of Denmark [Online] Available at: http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3215 Accessed on November 7, 2010.

Maffahy JM, Simeonov E. 1999. Stability analysis for a mathematical model of the *lac* operon. Q Appl Math 57:37–53.

Marino S, Hogue IB, Ray CJ, Kirschner DE. 2008. A methodology for performing global uncertainty and sensitivity analysis in systems biology. J Theor Biol 254:178–196.

Markowetz F, Spang R. 2007. Inferring cellular networks—a review. BMC Bioinformatics 8 (Suppl 6):S5.

McGregor AP. 2005. How to get ahead: the origin, evolution and function of bicoid. Bioessays 27:904–913.
Mendes P, Kell D. 1998. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Bioinformatics 14:869–883.
Mendes P, Kell DB. 2001. MEG (Model Extender for Gepasi): a program for the modelling of complex, heterogeneous, cellular systems. Bioinformatics 17:288–289.
Moles CG, Mendes P, Banga JR. 2003. Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome Res 13:2467–2474.
Nicolis G, Prigogine I. 1977. Self-Organization in Nonequilibrium Systems. From DIssipative Structures to Order through Fluctuations. NY, USA: Wiley.
Nocedal J, Wright SJ. 1999. Numerical Optimization. NY, USA: Springer.
Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. 2008. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell 133:1277–1289.
Perkins TJ, Jaeger J, Reinitz J, Glass L. 2006. Reverse engineering the gap gene network of *Drosophila melanogaster*. PLoS Comput Biol 2:e51.
Reeves GT, Fraser SE. 2009. Biological systems from an engineer's point of view. PLoS Biol 7:e21.
Reinitz J, Sharp DH. 1995. Mechanism of eve stripe formation. Mech Dev 49:133–158.
Rivera-Pomar R, Jackle H. 1996. From gradients to stripes in *Drosophila* embryogenesis: filling in the gaps. Trends Genet 12:478–483.
Sánchez L, Thieffry D. 2001. A logical analysis of the *Drosophila* gap-gene system. J Theor Biol 211:115–141.
Santillán M. 2008. Bistable behavior in a model of the lac operon in *Escherichia coli* with variable growth rate. Biophys J 94:2065–2081.
Santillán M, Mackey MC. 2001. Dynamic behavior in mathematical models of the tryptophan operon. Chaos 11:261–268.
Santillán M, Mackey MC. 2004. Influence of catabolite repression and inducer exclusion on the bistable behavior of the lac operon. Biophys J 86:1282–1292.
Schulz C, Tautz D. 1994. Autonomous concentration-dependent activation and repression of Krüppel by hunchback in the *Drosophila* embryo. Development 120:3043–3049.
Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. 2008. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. Nature 451:535–540.
Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34:166–176.
Sen AK, Liu WM. 1989. Dynamic analysis of genetic control and regulation of amino acid synthesis: the tryptophan operon in *Escherichia coli*. Biotechnol Bioeng 35:185–194.
Shea MA, Ackers GK. 1985. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. J Mol Biol 181:211–230.
Simpson-Brose M, Treisman J, Desplan C. 1994. Synergy between the hunchback and bicoid morphogens is required for anterior patterning in *Drosophila*. Cell 78:855–865.
Sinha S. 1988. Theoretical study of tryptophan operon: application in microbial technology. Biotechnol Bioeng 31:117–124.
Struhl K. 1999. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell 98:1–4.
Szymanski P, Levine M. 1995. Multiple modes of dorsal-bHLH transcriptional synergy in the *Drosophila* embryo. EMBO J 14:2229–2238.
Tang Y, Reed P, Wagener T, van Werkhoven K. 2006. Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. Hyrdol Earth Syst Sci Discuss 3:3333–3395.
Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. 1999. Systematic determination of genetic network architecture. Nat Genet 22:281–285.
Vilar JM, Guet CC, Leibler S. 2003. Modeling network dynamics: the lac operon, a case study. J Cell Biol 161:471–476.
Vilar JM, Leibler S. 2003. DNA looping and physical constraints on transcription regulation. J Mol Biol 331:981–989.
von Dassow G, Meir E, Munro EM, Odell GM. 2000. The segment polarity network is a robust developmental module. Nature 406:188–192.
von Hippel PH, Revzin A, Gross CA, Wang AC. 1974. Non-specific DNA binding of genome regulating proteins as a biological control mechanism: I. The lac operon: equilibrium aspects. Proc Natl Acad Sci USA 71:4808–4812.
Weise T. 2009. Global Optimization Algorithms—Theory and Application. [Online] Available at: http://www.it-weise.de/projects/book.pdf
Wong P, Gladney S, Keasling JD. 1997. Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. Biotechnol Prog 13:132–143.
Yanofsky C, Horn V. 1994. Role of regulatory features of the trp operon of *Escherichia coli* in mediating a response to a nutritional shift. J Bacteriol 176:6245–6254.
Yildirim N, Mackey MC. 2003. Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data. Biophys J 84:2841–2851.
Yuh CH, Bolouri H, Davidson EH. 2001. *Cis*-regulatory logic in the *endo16* gene: switching from a specification to a differentiation mode of control. Development 128:617–629.
Yuh CH, Bolouri H, Davidson EH. 1998. Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279:1896–1902.
Zhou X, Su Z. 2008. tCal: transcriptional probability calculator using thermodynamic model. Bioinformatics 24:2639–2640.
Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M, Philippakis AA, Hu Y, De Masi F, Pacek M, Rolfs A, Murthy T, Labaer J, Bulyk ML. 2009. High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res 19:556–566.
Ziehn T, Tomlin AS. 2008. A global sensitivity study of sulfur chemistry in a premixed methane flame model using HDMR. Int J Chem Kinet 40:742–753.
Zinzen RP, Senger K, Levine M, Papatsenko D. 2006. Computational models for neurogenic gene expression in the *Drosophila* embryo. Curr Biol 16:1358–1365.

Editor: Michael M. Cox