ComPlexUs
ORIGINAL RESEARCH PAPER
ComPlexUs 2003;1:54–64; DOI:10.1159/000070462

Transcriptional Control in *Drosophila*

John Reinitz[a] Shuling Hou[b] David H. Sharp[b,c]

[a]Department of Applied Mathematics and Statistics and The Center for Developmental Genetics, Stony Brook University, Stony Brook, N.Y.,
[b]Theoretical Division, and [c]Applied Physics Division, Los Alamos National Laboratory, Los Alamos, N. Mex., USA

Key Words
Transcriptional control · Transcription · System biology · Computational biology · *Drosophila* · Segmentation

Abstract
We present a new model of transcriptional control. A central goal of this model is to show how modular enhancers arise from groups of binding sites. The model has a three-layer organization. The first layer describes the binding of activators and repressors to the regulatory region of a gene and incorporates the effects of repression by competition and quenching. The second layer describes adapter molecules binding to DNA-bound activators, and incorporates the effect of direct repression. Finally, the activation of transcription is modeled by an Arrhenius mechanism in which activating adapters lower the activation energy barrier. We show that this model is testable against transcription data derived from early *Drosophila* embryos. We believe this model is sufficiently refined to give a realistic account of the physiological consequences of complex interactions of regulatory molecules. The present approach supplements and supports, in an essential way, the insights into multigenic regulation derived from work aimed at formulating logical design principles for regulatory networks.

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com
KARGER

Accessible online at:
www.karger.com/cpu

Dr. John Reinitz
Department of Applied Mathematics and Statistics and
The Center for Developmental Genetics, Stony Brook University
Stony Brook, NY 11794-3600 (USA)
Tel. +1 631 632 8352, E-Mail reinitz@ams.sunysb.edu

Synopsis

In simple prokaryotic organisms – in the bacterium *E. coli,* for instance – the control of transcription generally takes place through local molecular activity at a small number of binding sites. Regulation of the *lac* operon offers a classic and representative example. In an environment rich in the sugar lactose, *E. coli* expresses the *lac* operon, producing a handful of enzymes that allow lactose digestion. Conversely, in the absence of lactose, it is a single protein – the *lac* repressor – that acts to inhibit expression by binding to the site where RNA polymerase would otherwise initiate transcription. In this case, as in the regulation of most prokaryotic genes, activity at a single binding site can lead to direct phenotypic consequences.

In more complex organisms such as metazoans and other eukaryotes, the mechanisms of gene regulation are considerably more complicated. The control region of a eukaryotic gene often involves not just a handful of binding sites, but hundreds or even thousands, and significant control of transcription takes place through the binding of factors at many sites simultaneously. In this setting, in fact, identifiable regulatory functions cannot always be assigned to individual binding sites. Rather, functions tend to be associated with groups of sites that are operationally bound together by complex and highly nonlinear cooperative interactions.

Partly for this reason, the mechanisms of transcription control in eukaryotes largely remain unclear. Nevertheless, all is not confusion. In the genomes of many organisms, researchers have identified groups of binding sites that act as more or less coherent modules in controlling some aspects of gene expression. 'Modular enhancers' of this kind are not entire genes, but distinct stretches of DNA within the control region of a gene – typically 100–1,000 base pairs in length – that control expression in a specific tissue or stage of development. Because of their coherent

Introduction

Metazoan regulatory regions are extremely complex and qualitatively different from those of prokaryotes. In prokaryotes, transcriptional control is mediated by a small number (1–5) of binding sites, each of which has a well-defined function. By this we mean that the inactivation of a single site gives a clear-cut phenotypic effect that is readily interpretable. Although recent work has shown that this picture is not the whole story for certain prokaryotic genes [1, for example], the relative importance of these 'non-local' effects in prokaryotic genes is much smaller than in metazoan genes. For this reason, our basic picture of such systems is still based on the fundamental ideas first developed by Jacob and Monod for the *lac* operon [2] and elaborated by many others, notably Ptashne and Ackers [3, 4]. At the chemical level, the physiology of such control regions could be precisely described by classic methods of chemistry [5] because the regulation of such control regions could be faithfully reconstituted in assays in which only purified molecular species are used, i.e. assays in which no cell or nuclear extracts are used, so that the chemical composition of the assay is completely defined. Such experiments have shown that most repression in prokaryotes can be understood in terms of competition for overlapping binding sites, of which the classic example is the overlapping sites for binding *lac* repressor and RNA polymerase in the *lac* operon.

None of these conditions are realized in metazoan control regions. In contrast to prokaryotes, metazoan control regions are often extremely large, particularly in key developmental genes. Although data is incomplete, it would appear that 16 kilobases (kb) is a typical size [6], and that control regions of over 50 kb in size exist [7]. These control regions contain hundreds or thousands of binding sites. In many cases, mutagenesis of a single binding site has little or no discernible effect on expression. Major alterations in expression are seen only when many binding sites are disrupted at once, often by means of a deletion. This indicates that function is encoded by groups of binding sites, implying that the characterization of such groups is of considerable importance. To complicate matters further, there is no faithful in vitro assay for the regulation of RNA polymerase II. Indeed, it is now clear that a requirement for such an assay is the ability to reconstitute functional chromatin in vitro from pure substances. Such an assay faces experimental and theoretical difficulties since there is a strong body of evidence that chromatin in vivo is actively remodeled by a variety of mechanisms that involve the hydrolysis of ATP, so that chromatin is not an equilibrium structure.

The distinction outlined above between the description of prokaryotic and metazoan control regions reflects a corresponding distinction in our scientific understanding of these systems. In the case of prokaryotic genes, inferences about the functions of binding sites can be tested by in vitro chemical experiments, as discussed above. This places the transcription assay squarely within a large body of work which deals with chemical experiments using pure substances, and thus it establishes contact with those types of chemical experiments that form the underpinning of theoretical chemistry and its relationship to physics. Thus these experiments help to provide a rather *fundamental theory* of prokaryotic control regions, and support assertions such as: 'the regulation of bacterial promoters can be understood in terms of chemistry'. At our current level of understanding of the natural world, many scientific phenomena cannot be brought into *direct* contact with the fundamental laws of physics and chemistry. Nevertheless, many scientifically valid statements can be made about such phenomena, particularly in terms of how they relate to other observable phe-

function, modular enhancers offer some hope that gene regulation in the eukaryotes might be understood by identifying and characterizing a manageable number of such elements.

And in this paper, Reinitz, Hou and Sharp take a small but potentially significant step that should help to elucidate some of mechanisms by which modular enhancers achieve their specificity and control. They argue that the coherent function of a modular enhancer will most likely be revealed through a 'bottom up' approach – that is, through the detailed modeling of the interaction of numerous activators, repressors and other molecular factors. It is these elements, by acting in concert, that lead to the emergence of entities capable of coherent regulatory action. Reinitz et al. do not claim to have solved any specific problem, but rather propose a specific modeling strategy that can be applied widely in exploring transcription control of entire, intact genes, or of enhancer fragments. Most importantly, perhaps, the authors show – in the context of *Drosophila* embryos – how the model can be tested by comparison with empirical data that is now available. Because of this close tie to empirical reality, it should be possible to systematically correct and improve the model, and thereby gain important insights into the mechanisms of gene regulation in eukaryotic organisms.

To motivate a specific form for their model, the authors first review the most basic features of the molecular mechanisms of eukaryotic transcription control. To begin with, they point out, regulation takes place on all relevant spatial scales; that is, between genetic regions that are separated by any number of base pairs along the DNA. As in the prokaryotes, DNA-binding proteins sometimes act as repressors in a local fashion, by competing with activators for specific binding sites. But repression also takes place more globally. Repressors known as 'quenchers' act to block the activity of all activators bound

nomena. In physics, a body of statements pertaining to observable phenomena, made without reference to a more fundamental theory, is sometimes called a *phenomenological model*. For example, the laws of thermodynamics were originally formulated without reference to an underlying molecular theory, which in fact was not known at the time. An example from biology would be Mendelian genetics before the development of molecular cloning. The distinction between phenomenological models and 'fundamental' is of course a matter of degree, and changes with time. This distinction was for many years unimportant in biology because all biological models were phenomenological. This is no longer true, and distinguishing between these types of models is useful in order to make clear where and how scientific effort should be focussed. Much but not all of what is currently understood about metazoan control regions is phenomenological in nature, as we discuss below.

A striking feature of metazoan control regions (and those of other eukaryotes) is that they contain 'modular enhancers'. These are regions of 100–1,000 base pairs (bp) that control a particular domain of expression or expression in a particular tissue or developmental stage. They can act at distances up to 10 kb or more from the basal complex, and typically function if the orientation is reversed. They are often said to be 'additive' in the sense that if well separated, they can act independently. The phenomenological prominence of modular enhancers has rendered them of central importance for the practice of modern molecular biology, and a large number of laboratory manipulations depend on their phenomenology. In genes with relatively small control regions, such as *endo16* in the sea urchin *Strongylocentrotus purpuratus*, it is possible to construct models of the action of modular enhancers in terms of computational logic. These models describe the flow of information from *trans*-factors to the basal complex [8, 9], and can be formulated in terms of information obtained from careful inspection of in situ and kinetic data.

For the sake of specificity, we focus on control regions found in the fruit fly *Drosophila*, particularly those that act in the blastoderm stage embryo. This system permits highly informative experiments that clearly indicate that enhancers are not fundamental elements in the sense that genes are at the biological or binding sites are at the chemical level. At the chemical level, straightforward experiments exist to define a binding site and its spatial extent, while at the biological level two alleles can be assigned to the same locus or not by complementation. The operational definition of a modular enhancer does not define how the expression pattern of the enhancer relates to that of the intact gene, nor does it explain what the conditions for 'additivity' are or what happens when they are violated. Since many laboratories are attempting to understand the function of intact genes in terms of the behavior of modular enhancers, these are critical questions.

These points are well illustrated by certain experimental anomalies, such as the fact that an enhancer which faithfully mirrors the expression of the third *even-skipped (eve)* stripe in wild type is posteriorly derepressed in *knirps* mutants, but the native stripe is not [10, 11]. Even more perplexing is the observation that when a spacer element between the enhancers for the second and third *eve* stripes is eliminated, a novel expression pattern is produced in which the third stripe is reduced in intensity and fused with the second [12]. This observation caused Levine et al. [13] to propose that modular enhancers existed because short-range repressors ('quenchers'; see below) bound within a module were unable to affect other modules separated by spacers. This hypothesis addresses the reduced level of stripe 3 expression but not the fusion of stripes. We within a distance of roughly 50–100 base pairs. Some repressors also act in a truly global manner through the mechanism of gene silencing, in which a repressor can block an activator located anywhere on a gene. In view of these facts, a model of transcription should incorporate regulatory processes acting over many scales.

The authors also argue that transcription control can be viewed, roughly, as having a hierarchical structure of three layers. The first layer is the basic transcriptional machinery – RNA polymerase II and a host of associated factors that initiate transcription under appropriate conditions. A second layer involves a set of activators or repressors that bind to DNA and that act either locally or more globally to affect transcription. These activators and repressors, however, do not generally act directly on the basic transcriptional machinery, but require another set of factors to mediate the interaction. This third layer of control involves so-called adapter factors or co-regulators – molecules that do not bind directly to DNA, but interact with activators and repressors and help to transfer regulatory information to the basic transcriptional machinery. Reinitz and colleagues take this three-layer structure as a generic framework in building their model.

The model begins with the first layer of control, the basic transcriptional machinery itself. The authors suppose that the rate-limiting step in transcription is initiation, and represent this in a phenomenological way as an activation process. They suppose that diverse influences in operation around the basal complex (the location of transcription initiation on the gene) affect the value of an energy barrier, ΔA, and that thermal activation over this barrier initiates transcription. In thermal equilibrium at temperature T, the rate of initiation k is then proportional to $\exp(-\Delta A/RT)$, and increases as the barrier ΔA decreases (or vice versa). Because of the non-linear dependence on ΔA, Reinitz et al. point out, this expression naturally

believe that the hypothesis may well be correct but that a model will be required to test it in the complex phenomenological context in which quenching operates.

More generally, we believe that an explanation of the emergence and properties of enhancer modules must be derived from an analysis of the interaction of bound activators and repressors. Moreover, the rules by which enhancers combine to give the functional activity of an intact gene must be understood in order to connect enhancers with bona fide biological function. This is an important point, since enhancers are generally assayed fused to functionally inert bacterial or yeast genes. While a few very important functional assays have been performed with native gene product [14], lethality problems make these experiments very difficult. Much of the experimental work has started with intact genes and dissected them; we believe that it is natural for a theoretical approach to start with binding sites and work up towards larger composite entities. Such a theory must account for multiscale regulation and the mediation of adapter factors, as we now explain.

Regulation, including repression, takes place by many mechanisms simultaneously, and furthermore these mechanisms have different characteristic scales. For example, repression by competition requires binding sites to overlap and thus has a spatial scale in nucleotides of about the size of a binding site, ~10–20 bp. Another mechanism of repression, known as quenching, allows a repressor bound to DNA to block the action of activators bound less than 100 bp away [15]. A third mechanism known as silencing [16] enables a repressor to block the action of an activator over a large region of DNA bounded by insulator elements, except for large complex loci this usually means the whole gene. At larger scales still, not all repressive mechanisms have yet been characterized because of the lack of appropriate assays. The largest scale is clearly an entire chromosome (megabases), since all genes in all chromosomes are inactivated by chromosomal condensation during mitosis. While this last mechanism is not specific, its existence indicates that no mechanistic limitation to the scale of repression exists.

A picture is emerging of a three-layer structure for the metazoan transcription complex [17]. One layer is the transcription machinery itself, by which we mean the entire holoenzyme and associated TAFs. Another layer consists of activators and repressors that bind to sites in enhancers and elsewhere outside the basal complex. These two layers perhaps suffice for some simpler eukaryotes, such as yeast, where activators appear to act directly on members of the basal complex [18]. In metazoa, an extra layer mediates between the activators/repressors and the basal complex; members of this layer are known as 'adapter proteins', 'coactivators', or 'corepressors'. The key feature of these transcription factors is that they do not bind DNA directly and that they mediate the regulatory action of a DNA-bound ligand with the holoenzyme. The presence of these factors clearly increases both the flexibility and complexity of the transcriptional machinery. Although chemical details of the action of these factors are sketchy, they will need to be included in transcription models.

In considering such transcription models, we find it useful to identify three levels of description. In analogy to terminology used in structural biology, we classify models of transcription as being primary, secondary and tertiary as follows. A primary level model of transcription is one based solely on the sequence of the control region itself, without any notion of changes of state caused by binding of transcription factors. A tertiary model is one that includes the full chemical mechanism. In this paper we introduce the notion of a secondary level model of transcription that provides for the possibility of synergy between two transcription factors, as any two influences that act independently to decrease ΔA will together have more than an additive effect on the transcription rate k. To avoid unrealistic features, the authors also suppose that if ΔA were equal to zero, other factors in the cell would limit the transcription rate to some value R_0; likewise, that if ΔA became arbitrarily large, there would still be some 'leakage' rate of transcription. This feature enters the model naturally by letting $\Delta A/RT = \Theta_0 - QM$, which leads immediately to their Eq. 2, with the leakage rate of transcription being $R_0 \exp(-\Theta_0)$. In this formulation, M is the number of molecules present of some 'activating factor' (the identity of which depends, of course, on context), each molecule of which decreases the energy barrier by Q kcal/mole.

So far, Eq. 2 merely offers a way of relating the transcription rate of a gene (or gene fragment) to the quantity M of some generic 'activating factor'. Connecting this framework to reality – and especially to experimental data – means interpreting the activating factor as some real molecule, or set of molecules, relevant to the gene in question. Moreover, it is necessary to link M to the activity of numerous repressors, activators and other molecular factors – the basic elements of the regulatory apparatus. For many genes, researchers now know and have characterized a more or less complete list of binding sites and their associated transcription factors, and Reinitz et al. propose a specific scheme for incorporating such detailed knowledge into a testable model for transcription.

The authors start with the second layer of transcription control – the DNA binding proteins. Imagine the gene to have its bases numbered in order, starting at one, and increasing in the 5' to 3' direction. One can also number the binding sites along the gene (each will involve at least several bases or more), and represent the location and type of each site with the notation

scription. At this level, the promoter is viewed as a linear segment of DNA containing binding sites that may have ligands bound to them. The interactions of ligands with DNA are modeled by classic thermodynamics, but protein-protein interactions are modeled in a phenomenological manner, with functional forms derived from chemistry whenever possible. This amounts to using a generic average of adapters as a placeholder while awaiting further information and performing studies in a system (see below) in which the set of adapters is fixed. A disadvantage of this approach is that many important parts of the chemical mechanism are neglected. The advantage is that a large body of promoter-reporter experiments in molecular genetics give rise to data that are at the resolution of a secondary level model. Moreover, as more information about adapters becomes available this can be inserted into the mathematical framework.

On a more general level, this model will address the question of how groups of binding sites give rise to modular enhancers. Enhancers were originally defined in an operational manner as fragments of regulatory DNA that were sufficient to drive transcription – typically in an orientation-independent manner – while up to several kilobases distant from the transcription start site. While we believe the existence of such modules to be of great biological significance, we note that even the best-studied enhancers have probably not been fully characterized. First, for operational reasons, the boundary of an enhancer is always a restriction site, which is not physiologically significant in the native chromosome. Second, the practicalities of looking for enhancers ensure that all identified enhancers fall on a contiguous segment of DNA, but we are aware of no fundamental physiological reason why this should be so. Overcoming these limitations requires an understanding of how binding sites specify enhancers, since binding sites can be recognized from sequence in many cases.

The Model

General Considerations

Rather than model the holoenzyme directly, we take the point of view that transcription initiation is the rate-limiting step in transcription. We imagine that the presence of activating general factors (e.g. coactivators) catalyze initiation by reducing the energy barrier ΔA in a reaction with Arrhenius kinetics, so that the reaction rate k can be expressed as

$$k \propto \exp(-\Delta A/RT). \qquad (1)$$

The motivation for this representation is twofold. First, it is a reasonable minimal kinetic model for an enzymatically catalyzed reaction. Second, there is experimental data showing that certain combinations of transcription factors synergize in a greater than multiplicative fashion: if transcription is increased by a factor of a by protein 1 and b by protein 2, the two together increase transcription by more than a · b [19]. One way to achieve this effect is if factors 1 and 2 each decrease ΔA by a characteristic amount. Thus we imagine that each molecule of activating general factor decreases ΔA by Q kcal/mole. If ΔA is reduced to zero, other processes become limiting so that transcription has an overall maximum rate of R_0. We also imagine that in the absence of stimulating factors, transcription takes place at a very small leakage rate $R_0 \exp(-\Theta)$, where Θ is a positive number. Then the transcription rate is given by

$$d[\mathrm{mRNA}]/dt = R = \begin{cases} R_0 \exp[-(\Theta - QM)] \\ \text{iff } QM < \Theta \\ \\ R_0 \text{ otherwise,} \end{cases} \qquad (2)$$

where M is the average number of molecules of adapter factor that are present.

In the absence of exhaustive chemical information on the adapter factors, we cannot calculate with a vector of directly observed Ms for each adapter factor; $i[m_i, n_i; a]$. Here i identifies the binding site, m_i and n_i label the numbers of the first and last bases that belong to the site, and a identifies the type of ligand that can bind to the site. This scheme merely provides an ordered description of the various binding sites of a gene. The next step is to calculate how ligands bind to these sites. The authors calculate the fractional occupancy of a site i (the probability that a ligand of type a is bound to it) by their equations 3 and 4, the latter being appropriate if two binding sites overlap. The only factors entering here are the concentrations of the various ligands, v_a, and the binding affinities (or dissociation constants) K_i which reflect the likelihood that ligand a will bind to DNA at site i. As one would expect, $f_i[m_i, n_i; a]$ approaches one when either v_a or K_i becomes large.

Ligands may be either activators or repressors; however, only bound activators that are *not repressed* can contribute to the initiation of transcription. To include the effects of repression by quenching, the authors propose to calculate an 'effective fractional occupancy' for activators. The idea is to reduce the fractional occupancy of the various sites for activators by calculating the degree to which bound repressors block these activators even some distance away.

The authors do this in their Eq. 5. Suppose that an activator occupies site i with likelihood $f_i^{\mathrm{A}}[a; m_i, n_i]$. Other sites (nearby i or far away from it) may be occupied by quenchers Q. For each such occupied quenching site k, Eq. 5 reduces the fractional occupancy of site i by a multiplicative factor less than one. The size of the reduction produced by each site k depends on several factors. It increases with the quenching efficiency E_b of the quencher b, and with the likelihood $f_k^{\mathrm{Q}}[b; m_k, n_k]$ that site k is actually occupied by a quencher. The degree of repression also increases with the physical proximity of sites i and k. In particular, the effect of repression decreases with distance and

instead M merely provides a functional form. To link this picture to observables, we need to be able to calculate M in terms of the binding of ligands to DNA. This amounts to a chain of functional relationships that extend from ligands binding DNA to adapter factors binding to ligands and thus ultimately to the M in equation (2). Hence we now turn our attention to the other end of this chain and consider the binding of ligands to DNA.

We imagine that each gene is coordinatized by its sequence on the coding strand, with the first transcribed base assigned position 1. These sequence coordinates then increase in a 5' to 3' direction. Binding sites are represented as follows. Each binding site has an identifying index i together with certain properties. Each site is considered to bind a certain ligand a, and if a certain site can bind more than one ligand it would be represented as a set of sites that happened to be coextensive. The 5' boundary of the site lies at base m_i and the 3' boundary at base n_i, so that $n - m > 0$. Thus, when indexing a particular site we write $i[m_i, n_i; a]$ to express this set of properties. This representation also lends itself well to managing each site as a record in a relational database.

Each binding site $i[m_i, n_i; a]$ has associated with it a dissociation constant K_a. If ligand a is present in concentration ν^a and binding site i does not overlap or cooperate with other sites, the fractional occupancy of site i is given by

$$f_{i[m_i,n_i;a]} = \frac{K_i \nu^a}{1 + K_i \nu^a}. \qquad (3)$$

Other cases are straightforward generalizations. The most important one is competitive binding, which occurs when two sites $i[m_i, n_i; a]$ and $j[m_j, n_j; b]$ overlap, which happens if $m_j \leq n_i \leq n_j$ or if $m_j \leq m_i \leq n_j$. In that case,

$$f_{i[m_i,n_i;a]} = \frac{K_i \nu^a}{1 + K_i \nu^a + K_j \nu^b}. \qquad (4)$$

It is straightforward to generalize this further to cooperative or anticooperative binding.

Now we have a set of f_i's that are each a function of the concentration of one or more ligands. At this point we use external information about function to decide which ligands are activators. We denote the f_i associated with binding an activator A by $f^A_{i[m_i,n_i;a]}$. Certain other binding sites contain quenchers Q, and these sites are bound with occupancy $f^Q_{j[m_j,n_j;a]}$. Only unquenched activators are able to facilitate the association of adapter factors to the holoenzyme, and so we must correct the fractional occupancy of activator sites to account for quenching. The corrected fractional occupancy $\mathcal{F}^A_{i[m_i,n_i;a]}$ is then given by

$$\mathcal{F}^A_{i[a;m_i,n_i]} = f^A_{i[a;m_i,n_i]} \prod_k [1 - q(d_{ik}) E_b f^Q_{k[b;m_k,n_k]}]. \qquad (5)$$

Here $d_{ik} = m_i - m_k$, and $q(d_{ik})$ is an empirical function equal to 1 for $d_{ik} < 50$, 0 for $d_{ik} > 150$, with smooth interpolation in between (fig. 1) [20]. E_b is the quenching efficiency of quencher b, which must be between zero and one. The product is taken over all quenching sites k, but those that are far away do not affect the product since $q(d)$ will vanish for those terms and they will contribute a factor of one to the product. Note that this expression allows multiple quenching sites to contribute to repression by multiplying together many factors, each a little less than unity, to form a very small number.

Interaction with Adapter Factors

In the absence of a full tertiary model for activation or exhaustive assays for adapter factors (AFs), we use a minimal representation of the three functional layers of the transcription complex. We imagine that activators collectively form binding sites for activating adapter factors. Each activator may have a greater or lesser capacity to do this, and so we imagine that $1/C_a$ molecules of bound activator a form a binding site for AFs. vanishes for sites separated by more than 150 bp. [The authors include this latter effect through the function $q(d_{ik})$, the form of which is shown in their figure 1]. All told, Eq. 5 includes both the effect of repression by direct competition for a binding site (when i and k refer to the same site) as well as the quenching of activators within a neighborhood of any bound repressor.

At this point, the model reflects how activators bind to sites on DNA, and how some of these are rendered ineffective by repression. Reinitz et al. next aim to feed this information upward to the third layer of regulation, which involves so-called adapter factors or co-regulators – the 'activating factors' that mediate between the DNA-binding proteins and the basic transcription machinery. These factors will contribute directly to transcription initiation by acting at the basal complex (and their number M will enter into Eq. 2 to determine the transcription rate). To be specific, the authors suppose that a number $(1/C_a)$ of bound activators (at the lower layer) will together form a binding site for these higher-layer adapter (or 'activating') factors. Consequently, the number of such sites produced by activators of type a will be $C_a \Sigma_i \mathcal{F}^A_i[a; m_i, n_i]$, this being the total number of bound (and unrepressed) activators of type a divided by the number required to form one adapter-factor binding site $(1/C_a)$. Eq. 6 gives the total number N of such sites by summing over the various different species of activators. The number of bound adapter factors M (Eq. 7) is then simply given by the total number of binding sites N multiplied by the fractional occupancy of any such site, which is given by Eq. 8. Here $[AF]$ is the chemical concentration of the adapter factor, and K_{AF} is its associated binding affinity. The mathematical form of Eq. 8 merely provides a plausible function in which the fractional occupancy increases from 0 to 1 as the product $K_{AF}[AF]$ increases.

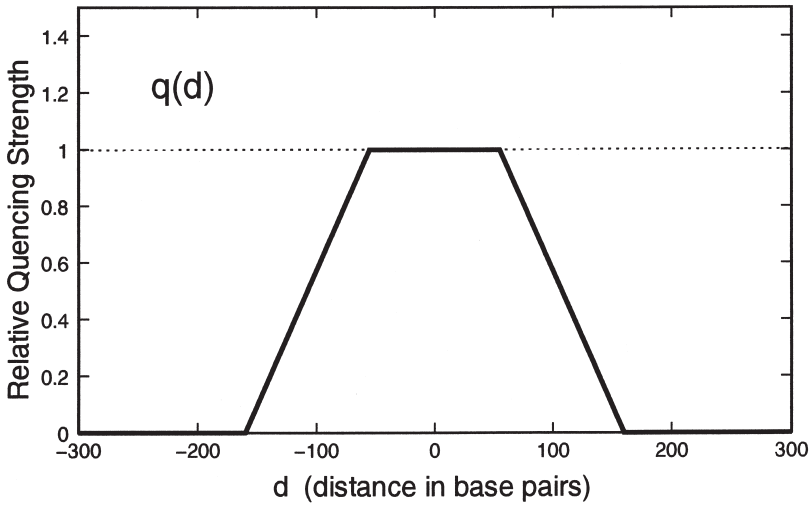


Fig. 1. The quenching function q(d).

Then the total number N of such sites is given by

$$N = \sum_a C_a \sum_i \mathcal{F}^A_{i[a;m_i,n_i]}. \quad (6)$$

Now we would like to know the number M of AFs bound to the surface of activators. That number will be

$$M = f^{AF} N, \quad (7)$$

where

$$f^{AF} = \frac{(K_{AF}[AF])^n}{1 + (K_{AF}[AF])^n}. \quad (8)$$

Here we use a Hill function of order n as a functional approximation of the presumably highly cooperative recruitment of adapter factors.

At this point we are almost ready to connect what we have written down to equation (2), which gives the transcription rate as a function of M. Before doing so, however, we must modify equation (6) to take into account the effects of direct repression. In direct repression, a quencher bound within 50–100 bp [21] from the basal complex prevents activators bound anywhere else on the control region from activating transcription. Since this phenomenon involves proteins that act elsewhere as quenchers and because the range of interaction is very close to that observed for classical quenching, we regard direct repression as 'quenching of the basal complex'. The idea is that quenchers act on whatever is in range. If only activators are within the critical distance, they are prevented from activating but activators elsewhere can still act. Since there is generally only one basal complex, quenching it will prevent it from receiving signals from any activators.

This picture suggests that the way to model direct repression is to calculate the effects of the quenchers on the basal complex itself, rather than on activators. This can be done by convolving f^{AF} with the activities of quenchers in a way analogous to equation (5). We do this by introducing a quenched AF occupancy $\mathcal{F}^{AF}$, which is given by

$$\mathcal{F}^{AF} = f^{AF} \prod_k [1 - q(d_k) E_b f^Q_{k[b;m_k,n_k]}]. \quad (9)$$

Here the terms have the same meaning as in equation (5), except that the quenchers operate on a single f^{AF} term rather than many f^A terms. Here d_k is the distance to the basal complex, and E_b has the same value for quenching or direct repression.

In principle, Eq. 7 completes the model and provides a means for calculating the transcription rate of a particular gene from knowledge of the concentrations and binding affinities of the various activators, repressors and adapter factors. However, as the authors point out, the model at this point has overlooked one other important mechanism of repression. While a repressor bound within the vicinity of an activator will dampen or even completely repress this particular activator, a repressor bound within a neighborhood of the basal complex of a gene will contribute to a global repression of transcription. This important mechanism was discovered experimentally and referred to as 'direct repression' by Michael Levine and colleagues (see reference 12 of the paper), and the authors include its consequences by calculating an 'effective' fractional occupancy for the adapter factor binding sites. Hence, Eq. 8 is naturally replaced by Eq. 9, which includes a factor less than one for each repressor that is bound within a neighborhood of the basal complex.

With this alteration, the model is complete. Using Eqs. 2 through 10, as well as detailed characterization of the binding sites and transcription factors relevant to a gene, one can calculate the transcription rate as it depends on a number of specific physical quantities. These include the binding affinities of activators and repressors with DNA binding sites, or of the adapter factors with the basal complex, the quenching strengths E_b of the various quencher species, the numbers C_a that describe how activators form binding sites for adapter factors, the parameters R_0 and Θ_0 that determine the maximum and leakage transcription rates, and so on. What will make this model useful, the authors suggest, is that all of these quantities can actually be measured with currently available techniques. 'Despite the complexity of the model,' they claim, 'it is completely defined by data that is already available or obtainable by existing experimental methods.'

To summarize, the quantity M which controls the transcription rate in equation (2) is given by

$$M = \mathcal{F}^{AF} N, \tag{10}$$

where $\mathcal{F}^{AF}$ is obtained from the direct repression equation (9), which in turn uses f^{AF} from equation (8) together with data about spacing of sites on the DNA. The term N in equation (10) represents the effects from activation and quenching. It does so because N is calculated from equation (6), which contains both the inherent activator strength C_a and the fractional occupancies $\mathcal{F}^A$ convolved with quenching, which in turn are calculated from the chemical occupancies f_i^{A} and f_i^{Q} of activators and quenchers, respectively. The calculation of $\mathcal{F}^A$ also utilizes the inherent strength E_a of the quenchers and their spacing on the DNA relative to activators. Each f_i is calculated from the chemical concentration of its ligand using the equilibrium binding affinity K_i.

We see that the model has a simple feedforward structure, although the functions that need to be evaluated are nonlinear and of considerable complexity. These functions capture a number of mechanisms operating simultaneously at different length scales on the DNA. The parameters appearing in the model fall into four classes according to the chemical species they are chiefly associated with.

(1) The binding affinities K_i depend on both sequence and ligand, but they are more tightly associated with sequence in the sense that they are the only parameters in the model whose number tends to increase in proportion to the length of DNA treated.

(2) The quenching and activation strengths E_a and C_a are properties of the quenching and activation proteins and grow in number with the number of such proteins considered.

(3) R_0 and Θ are associated with the basal promoter region and the basal complex.

(4) K_{AF} and $[AF]$ are presumably properties of the basal complex, but in the *Drosophila* blastoderm all of the factors lumped together as $[AF]$ are constant in space and time and so f_{AF} becomes a maternally controlled parameter.

Determining the Model Parameters

Despite the complexity of this model, it is completely defined by data already available or obtainable by existing experimental methods. Indeed, the 'secondary' level of approximation chosen is precisely that on which many experimental investigations of transcription are conducted. This is particularly true for the *Drosophila* blastoderm. A very limited number of genes are expressed in a spatially nonuniform manner in the blastoderm, and these genes have been unusually well characterized. In particular, the only genes that are expressed in a spatially controlled way as functions of anterior posterior position are the segmentation genes, and we have mapped their expression quantitatively at single-cell resolution [22].

It is noteworthy that all of the zygotic segmentation genes expressed at the blastoderm stage are not only transcription factors, but are also DNA-binding proteins. Other transcription factors that do not bind DNA (and many that do) are supplied maternally prior to gastrulation and are expressed in a spatially and temporally uniform manner. These factors include the AFs, which are maternally supplied and expressed in a spatially and temporally uniform manner in the blastoderm stage. This fixed population of adapter molecules supports the generic picture of their action used here.

There is, in addition, considerable data available from lines that have been transformed with P-elements. The experimental paradigm is to make an artificial construct containing a fragment of control region ligated to a basal promoter region and a reporter gene, typically bacterial *lac*Z. These constructs are stably inserted

To illustrate their point, the authors discuss an example of how specific data can be used to test the model. In the cells of the *Drosophila* blastoderm – an early stage in its embryonic development – only a handful of genes are expressed in a spatially non-uniform manner. These are the genes responsible for producing the segments that define the embryo's developing body plan. In this relatively simple setting, a variety of elegant and powerful experiments can be carried out to explore the expression of these genes (or fragments thereof, such as modular enhancers). In particular, researchers can create artificial genetic constructs 'containing a fragment of a control region ligated to a basal promoter and a reporter gene, typically bacterial lac Z'. With these lac Z transcripts, experimenters can flexibly explore the expression pattern of the gene (or fragment) in question under various conditions in the blastoderm – that is, in the presence of distinct combinations of the few proteins that control the expression of the segmentation genes. In this way, it is possible to measure both the transcription rate as well as the concentrations of the various transcription factors that should influence it. This provides much of the data required to test the model.

As a specific example, the authors refer to one existing data set that records the 'expression levels of segmentation gene proteins in a strip of 100 nuclei along the anterior-posterior axis.' Each of these 100 sets offers data on the level of lac Z expression as well as of its various regulatory proteins (activators and repressors). In their figure 2, Reinitz et al. display data of this kind obtained for the case of an enhancer for the even-skipped gene, which is controlled (except near the anterior tip) by just four proteins. As the authors point out, 'this is enough data to fully determine the parameters of the model.'

Some of these parameters (such as the binding affinities) can be measured in independent experiments; others are phe-

in chromosomal DNA, and there is a very rich set of transformant lines in which *lacZ* expression in the blastoderm is under the control of segmentation genes. Moreover, it is straightforward to generate more such lines to answer specific questions. These lines can be imaged quantitatively by fluorescent probes, and it is possible to do this while simultaneously imaging segmentation proteins by quantitative means.

We now outline specifically how quantitative data to test the model can be obtained, and give an example of a preliminary numerical data set. The full experimental background is given in the references. Colocalization of *lacZ* transcripts with protein permits the registration of this data with a preexisting quantitative data set of protein expression [22]. This data set contains the expression levels of segmentation gene proteins in a strip of 100 nuclei along the anterior-posterior (A–P) axis. The result of such registration is a set of 100 sets of values for the concentrations of regulators and *lacZ*, each in a different nucleus. We can assume that the *lacZ* transcription rate is proportional to the level of *lacZ* expression as long as the decay time of the message is short compared to the time required for changes in the transcription rate. The observed rapid loss of *lacZ* expression from large regions of the embryo indicates that this assumption is reasonable [23]. The data set contains protein concentrations from 8 different times during development, and *lacZ* expression can also be sampled at these different stages. Each spatial position at each time will correspond to a different combination of regulators, so that one can contemplate quantitatively assaying *lacZ* expression in a given construct under about 500 different combinations of concentrations of regulatory factors.

Although we have not yet begun our own *lacZ* assays, the model can be tested with available data using certain constructs. In particular, constructs that express only one or several stripes of the *even-skipped* seven striped expression pattern coextensively with the native stripes can be represented up to a proportionality

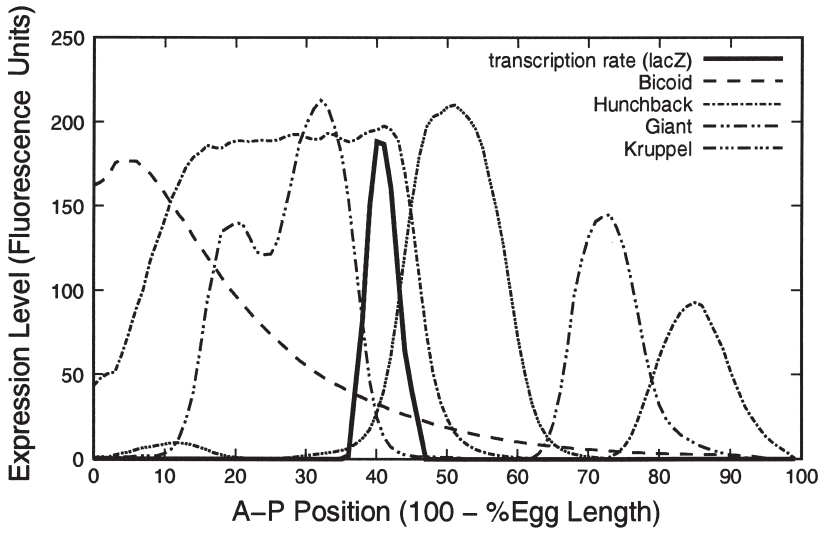
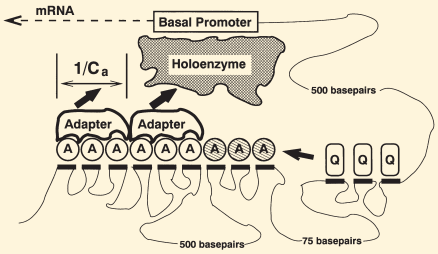


Fig. 2. Quantitative data for the analysis of even-skipped MSE2 (stripe 2 enhancer). Protein and lacZ expression levels are as labeled. This graph shows 100 sets of quantitative input-output combinations for the analysis of transcriptional control. lacZ levels were taken to be the same as even-skipped in stripe 2. See Myasnikova et al. [22] and http://urchin.spbcas.ru/FlyEx or http://flyex.ams.sunysb.edu/FlyEx for further information.



A schematic view of the model. The DNA upstream of the basal promoter is shown by a line, with thick areas to denote binding sites. Activators, shown as circles with the letter A, are bound to nine sites which are spaced at varying intervals along the DNA. The leftmost six activators can bind "adaptor" molecules. One adaptor molecule can bind to $1/C_a$ bound activators (with C_a equal to 1/3 in the diagram). The righthand three adapter molecules are quenched (shown by hatching) and cannot bind activators because there are three quenchers, the rectangles labeled Q, bound less than 100 basepairs away. The other activators are not affected by the quenchers because they are over 500 basepairs away. The polymerase holoenzme is bound to the basal promoter near the transcription start site.
It is not subjected to direct repression from the quenchers because they are bound more than 500 basepairs away. The bound adaptor factors reduce the Arrhenius activation energy by acting on the holoenzme. In the particular configuration shown there are two bound adaptor molecules, so M = 2.

nomenological. Values for these can be determined, however, by fitting the model in parallel to data for a number of distinct genetic constructs – some with pieces of DNA added or deleted, for example, or with mutations at particular binding sites. The strategy for using the model is straightforward. If one set of parameter values enables the model to fit the independent data simultaneously, this would suggest that the model incorporates the most important mechanisms of transcription control; if not, then it would suggest that some important features are missing.

How useful this model will be in exploring transcriptional control in *Drosophila* and in other eukaryotes remains to be seen, of course, and must await definitive

constant by the profile of the native stripe(s), which are already quantitatively mapped. An example of such data, which will be used for initial tests of the model, is given in figure 2. These data describe the expression of an enhancer for *even-skipped* stripe two, in which binding sites for the four proteins shown control expression throughout the embryo except for a very small region at the anterior tip [18, 24].

This is enough data to fully determine the parameters of this model. The parameters may be determined by minimizing the summed square deviations between observed *lacZ* expression and that given by the model. This might be accomplished by methods such as genetic algorithms or simulated annealing; we note that simulated annealing has been successful in solving a related problem in pattern formation [25–27].

Other data will determine the parameters even more precisely. Closely related constructs can be fit to the model as a group. Examples of such related constructs include those with site-directed mutations which delete, add, or change the affinity of binding sites, as well as insertions and deletions which change the spacing of sites and so on. Groups of constructs related by the operations above are used by experimentalists to define the functional properties of enhancers and binding sites. All of the perturbations mentioned above can be cleanly represented in the context of the model. A set of constructs related by mutation, insertion, and deletion can be fitted to the data as a group, simultaneously including several related constructs and their expression patterns and minimizing the summed squares for the whole group. This will afford powerful tests against experiment.

Discussion

The Explanatory Power of the Model

This model will be of value in elucidating the individual effects of different transcriptional mechanisms when acting in combination. It is these mechanisms which result in the spatio-temporal modulation of transcription rate that is responsible for patterned gene expression. In an intact gene, many mechanisms act in concert in a way that is still poorly understood, and this is also true of smaller constructs. The model will provide a systematic method for separating the contributions of different mechanisms acting simultaneously in a variety of different constructs. This will provide a substantial variety of cases for validation of the model in terms of the mechanisms included and the way they are represented. It may be that some parts of the model presented here will need modification when confronted with reality. Because of the rich set of data available, we are confident that the modification procedure will lead to a new model rather than to 'no model', at least while working in the blastoderm where AFs are fixed.

We regard the construction of this model as an ongoing process in the sense that transcriptional control involves many mechanisms. New mechanisms should be added to the model only after there is reason to believe that the model is functioning correctly for mechanisms already included. This approach naturally parallels experiment, since those mechanisms that are now well understood were characterized in relatively simple contexts. Nevertheless, a number of mechanisms known to be important in transcription are not represented in the model presented here. We briefly discuss these mechanisms and remark how they might be incorporated in the model while emphasizing the fact that the construction of specific model equations requires very careful comparison to specific experiments.

For example, many transcription factors can function as activators or repressors, depending on the available population of AFs, or by recruiting different sets of available AFs depending on nearby bound protein ligands [28]. The latter mechanism can be represented by adding a 'coactivation/corepression' term with its own distance function to equation (5). The former mechanism can be represented by representing specific adapter molecules in the model; although such information may not be available for some time since this mechanism does not operate in the blastoderm and the population of adapters is fixed. When such representation is possible, it provides a natural way to incorporate promoter competition [29]. Another possibility closer to the current framework would be to partition the quantity M in equation (2) by a fixed ratio between two competing basal complexes. Another area not yet represented are long-range interactions of various types. The activity of activators can attenuate over ranges of several kilobases, and silencing itself is a form of long-range repression. These and other mechanisms must be added in a stepwise manner, in each case by first modeling the experimental constructs used to elucidate the effect in the first place.

Over the longer term this model is focussed on the central problem of func-

attempts to fit real experimental data in a variety of settings. Nevertheless, in view of the many diverse factors that contribute to the regulation of transcription in eukaryotes, and their highly non-linear interactions, a model of this sort may well be indispensable in teasing out the contributions of one mechanism relative to another. Especially in view of its specific nature, and close contact with real data, the model here proposed possesses a distinctly attractive feature. It runs the risk and takes the chance of being wrong in a definitive way. Failure in this sense is also success, for it provides clues to systematic improvement and a path to deeper understanding.

Mark Buchanan

tional genomics. It is now widely accepted that understanding the function of a gene includes understanding the structure of its product and its transcriptional control. Transcriptional control in metazoans is a matter of spatio-temporal control, and this model is designed to provide a substantial step in solving what might be called the 'second half of the genetic code': How does a gene's sequence determine where it is expressed? In general, solving this problem also includes knowing which *trans*-factors are present in a cell, a bioinformatics and dynamics problem that we and others are addressing separately [22, 26, 30–33]. Constructs with regulatorily inert reporter transcripts separate these two questions, since the distribution of *trans*-factors can be regarded as fixed. The presence of adapter factors is a substantial complication, but in the blastoderm, where the set of adapter factors is uniform and determined by maternal input, it is likely that a model of the type presented here would be sufficiently accurate to predict the locations and actions of modular enhancers from knowledge of binding sites. With this information, the expression pattern itself can be predicted. Others [34] are solving the problem of predicting binding sites from sequence, and so the ultimate goal is the prediction of expression pattern directly from regulatory sequence.

Acknowledgments

J.R. was supported by grant RO1 RR07801 from the US National Institutes of Health, and S.H. and D.H.S were supported by the US Department of Energy under contract W-7405-ENG-36. We thank Johannes Jaeger and an anonymous reviewer for helpful comments on the manuscript.

References

1 Revet B, von Wilcken-Bergmann B, Bessert H, Barker A, Muller-Hill B: Four dimers of lambda repressor bound to two suitably spaced pairs of lambda operators form octamers and DNA loops over large distances. Curr Biol 1999;9:151–154.

2 Jacob F, Monod J: Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 1961;3:318–356.

3 Ackers GK, Johnson AD, Shea MA: Quantitative model for gene-regulation by lambda-phage repressor. Proc Natl Acad Sci USA 1982;79:1129–1130.

4 Ptashne M, Jeffrey A, Johnson AD, Maurer R, Meyer BJ, Pabo CO, Roberts TM, Sauer RT: How the lambda-repressor and cro work. Cell 1980;19:1–11.

5 Hawley DK, McClure WR: Mechanism of activation of transcription initiation from the lambda-PRM promoter. J Mol Biol 1982;157:493–525.

6 Fujioka M, Emi-Sarker Y, Yusibova GL, Goto T, Jaynes JB: Analysis of an *even-skipped* rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. Development 1999;126:2527–2538.

7 St. Johnston RD, Hoffmann FM, Blackman RK, Segal D, Grimaila R, Padgett RW, Irick HA, Gelbart WM: Molecular organization of the decapentaplegic gene in *Drosophila melanogaster*. Genes Dev 1990;4:1114–1127.

8 Yuh CH, Boluri H, Davidson EH: Genomic cis-regulatory logic: Functional analysis and computational model of a sea urchin gene control system. Science 1998;279:1896–1902.

9 Yuh CH, Boluri H, Davidson EH: Cis-regulatory logic in the *endo*16 gene: Switching from a specification to a differentiation mode of control. Development 2001;128:617–629.

10 Frasch M, Levine M: Complementary patterns of *even-skipped* and *fushi-tarazu* expression involve their differential regulation by a common set of segmentation genes in *Drosophila*. Genes Dev 1987;1:981–995.

11 Small S, Blair A, Levine M: Regulation of two pair-rule stripes by a single enhancer in the *Drosophila* embryo. Dev Biol 1996;175:314–324.

12 Small S, Arnosti DN, Levine M: Spacing ensures autonomous expression of different stripe enhancers in the *even-skipped* promoter. Development 1993; 119:767–772.

13 Gray S, Cai H, Barolo S, Levine M: Transcriptional repression in the *Drosophila* embryo. Philos Trans R Soc Lond B Biol Sci 1995;349:257–262.

14 Fujioka M, Jaynes JB, Goto T: Early *even-skipped* stripes act as morphogenetic gradients at the single cell level to establish *engrailed* expression. Development 1995;121:4371–4382.

15 Gray S, Szymanski P, Levine M: Short-range repression permits multiple enhancers to function autonomously within a complex promoter. Genes Dev 1994; 8:1829–1838.

16 Barolo S, Levine M: Hairy mediates dominant repression in the *Drosophila* embryo. EMBO J 1997;16: 2883–2891.

17 Lemon B, Tijian R: Orchestrated response: A symphony of transcription factors for gene control. Genes Dev 2000;14:2551–2569.

18 Wu YB, Reece RJ, Ptashne M: Quantitation of putative activator-target affinities predicts transcriptional activating potentials. EMBO J 1996;15:3951–3963.

19 Han K, Levine MS, Manley JL: Synergistic activation and repression of transcription by *Drosophila* homeobox proteins. Cell 1989;56:573–583.

20 Hewitt GF, Strunk BS, Margulies C, Priputin T, Wang XD, Amey R, Pabst BA, Kosman D, Reinitz J, Arnosti DN: Transcriptional repression by the *Drosophila* giant protein: Cis element positioning provides an alternative means of interpreting an effector gradient. Development 1999;126:1201–1210.

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

22 Myasnikova E, Samsonova A, Kozlov K, Samsonova M, Reinitz J: Registration of the expression patterns of *Drosophila* segmentation genes by two independent methods. Bioinformatics 2001;17:3–12.

23 Small S, Blair A, Levine M: Regulation of *even-skipped* stripe 2 in the *Drosophila* embryo. EMBO J 1992;11:4047–4057.

24 Andrioli LPM, Vasisht V, Theodosopoulou E, Oberstein A, Small S: Anterior repression of a *Drosophila* stripe enhancer requires three position-specific mechanisms. Development 2002;129:4931–4940.

25 Chu KW, Deng Y, Reinitz J: Parallel simulated annealing by mixing of states. J Comput Phys 1999;148: 646–662.

26 Reinitz J, Mjolsness E, Sharp DH: Cooperative control of positional information in *Drosophila* by *bicoid* and maternal *hunchback*. J Exp Zool 1995;271:47–56.

27 Sharp DH, Reinitz J: Prediction of mutant expression patterns using gene circuits. Biosystems 1998;47: 79–90.

28 Barolo S, Stone T, Bang AG, Posakony JW: Default repression and Notch signaling: Hairless acts as an adaptor to recruit the corepressors Groucho and dCtBP to Suppressor of Hairless. Genes and Development 2002;16:1964–1976.

29 Butler JE, Kadonaga JT: Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. Genes Dev 2001;15:2503–2508.

30 Kosman D, Small S, Reinitz J: Rapid preparation of a panel of polyclonal antibodies to *Drosophila* segmentation proteins. Dev Genes Evol 1998;208:290–294.

31 Reinitz J, Kosman D, Vanario-Alonso CE, Sharp D: Stripe forming architecture of the gap gene system. Dev Genet 1998;23:11–27.

32 Reinitz J, Sharp DH: Mechanism of formation of *eve* stripes. Mech Dev 1995;49:133–158.

33 von Dassow G, Meir E, Munro EM, Odell GM: The segment polarity network is a robust development module. Nature 2000;406:188–192.

34 Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. Proc Natl Acad Sci USA 2002;99:757–762.