Developmental Biology **246,** 2–13 (2002)
doi:10.1006/dbio.2002.0617, available online at http://www.idealibrary.com on IDEAL®

REVIEW

Modeling DNA Sequence-Based *cis*-Regulatory Gene Networks

Hamid Bolouri* and Eric H. Davidson†,[1]

**Science & Technology Research Centre, University of Hertfordshire, Hatfield, Hertfordshire AL10 9AB, United Kingdom; †Division of Biology, California Institute of Technology, Pasadena, California 91125*

Gene network analysis requires computationally based models which represent the functional architecture of regulatory interactions, and which provide directly testable predictions. The type of model that is useful is constrained by the particular features of developmentally active *cis*-regulatory systems. These systems function by processing diverse regulatory inputs, generating novel regulatory outputs. A computational model which explicitly accommodates this basic concept was developed earlier for the *cis*-regulatory system of the *endo16* gene of the sea urchin. This model represents the genetically mandated logic functions that the system executes, but also shows how time-varying kinetic inputs are processed in different circumstances into particular kinetic outputs. The same basic design features can be utilized to construct models that connect the large number of *cis*-regulatory elements constituting developmental gene networks. The ultimate aim of the network models discussed here is to represent the regulatory relationships among the genomic control systems of the genes in the network, and to state their functional meaning. The target site sequences of the *cis*-regulatory elements of these genes constitute the physical basis of the network architecture. Useful models for developmental regulatory networks must represent the genetic logic by which the system operates, but must also be capable of explaining the real time dynamics of *cis*-regulatory response as kinetic input and output data become available. Most importantly, however, such models must display in a direct and transparent manner fundamental network design features such as intra- and intercellular feedback circuitry; the sources of parallel inputs into each *cis*-regulatory element; gene battery organization; and use of repressive spatial inputs in specification and boundary formation. Successful network models lead to direct tests of key architectural features by targeted *cis*-regulatory analysis.

Key Words: **gene regulatory network; model; *endo16* gene; regulatory genomics.**

Introduction

cis-Regulatory systems can be thought of in very different ways. For example, *cis*-regulatory elements can be regarded as pieces of DNA sequence that contain clustered arrays of brief but specific target site sequences recognized by DNA-binding proteins. Or the same elements can be considered as staging platforms for the assembly of enormous multiprotein machines, of which the DNA-binding parts serve merely as anchors, and the functional meaning lies in the biochemical transactions that cause transcription or its repression. The first sort of view leads off into computational genomics, the second into the kind of structure/function biochemistry that explains how cells work. But development entails another agenda: to understand why it happens, we need to think about *cis*-regulatory systems in such a way that we can grasp the overall logic of their control functions. The basic concept that fits this bill is to think of *cis*-regulatory elements as genetically hardwired information processors, and of networks of these elements as systems of linked information processors. Each such element receives informational inputs that determine its activity, and it produces an informational output in the form of the regulatory instructions that it conveys to the basal transcription apparatus.

The genetic regulatory apparatus resides unchanged in every cell throughout the life cycle. What it does depends

[1] To whom correspondence should be addressed. Fax: (626) 793-3047. E-mail: davidson@caltech.edu.

0012-1606/02 $35.00

on the diverse inputs it receives at each point in time and space during development. The inputs are carried into the system by the transcription factors that bind at its target site sequences, for their presentation and activation depend on external circumstances: signaling, prior state, lineage, location, time, and so forth. Part of this information flow depends on the *cis*-regulatory transactions which govern the genes encoding the relevant transcription factors; part on other events, such as transduction of extracellular signals. The primary requirement of models for developmental gene regulatory networks is to specify the flow of regulatory information; that is, to represent specifically both the source and targets of inputs into each *cis*-regulatory element in the network.

Every *cis*-regulatory element carries out some processing of its input information, the essential regulatory function that underlies all developmental processes. The output is never identical to the input. For one thing, the inputs are always multiple, while when the element is active in any given cell, the output is a unique function that in some manner informs the basal transcription apparatus of a given gene how frequently to initiate transcription, or imposes on it a state of silence. *cis*-Regulatory information processing is specifically important in development because development depends fundamentally on spatial as well as temporal control of gene expression. During specification, the job of the regulatory apparatus is to integrate the information that will decide which genes will be expressed where (and where not). It is probably true, in general, that these decisions result from logic functions carried out by *cis*-regulatory elements controlling regulatory genes, such that incident spatial information is converted to a new spatial regulatory output. For example, a given *cis*-regulatory element might mandate expression of the gene it controls only where two differently positioned regulatory inputs overlap, resulting in the appearance of a new transcription factor in the particular spot in the embryo; or it might control expression through the interplay between positive and negative spatial inputs, each positioned differently than is the outcome of this interplay (for multiple examples and review, see Davidson, 2001 and earlier discussions in Davidson, 1990; Arnone and Davidson, 1997). It follows that a second requirement for models that are useful in dealing with *cis*-regulatory systems that control development is that they represent the logic operations which transform the inputs into the outputs of each element. Ultimately, this statement must refer to all the properties of the inputs, i.e., their temporal kinetics as well as their spatial location in the organism; their amplitude as well as their sign, i.e., their positive or repressive import.

Thinking about *cis*-regulatory systems from an informational point of view leads smoothly to the mutatable, measurable, regulatory properties of genomic DNA. The DNA sequence specifies the inputs each *cis*-regulatory element should listen to, and the information processing functions that it is capable of. So each input portrayed in a *cis*-regulatory model indicates a target site sequence, and these sequences can be recognized and tested functionally by mutation and gene transfer. Models that directly handle *cis*-regulatory information flow are in the end most illuminating and most useful in practical terms because they are DNA sequence-based models.

Illustration of Principle: The Endo16 cis-Regulatory Model

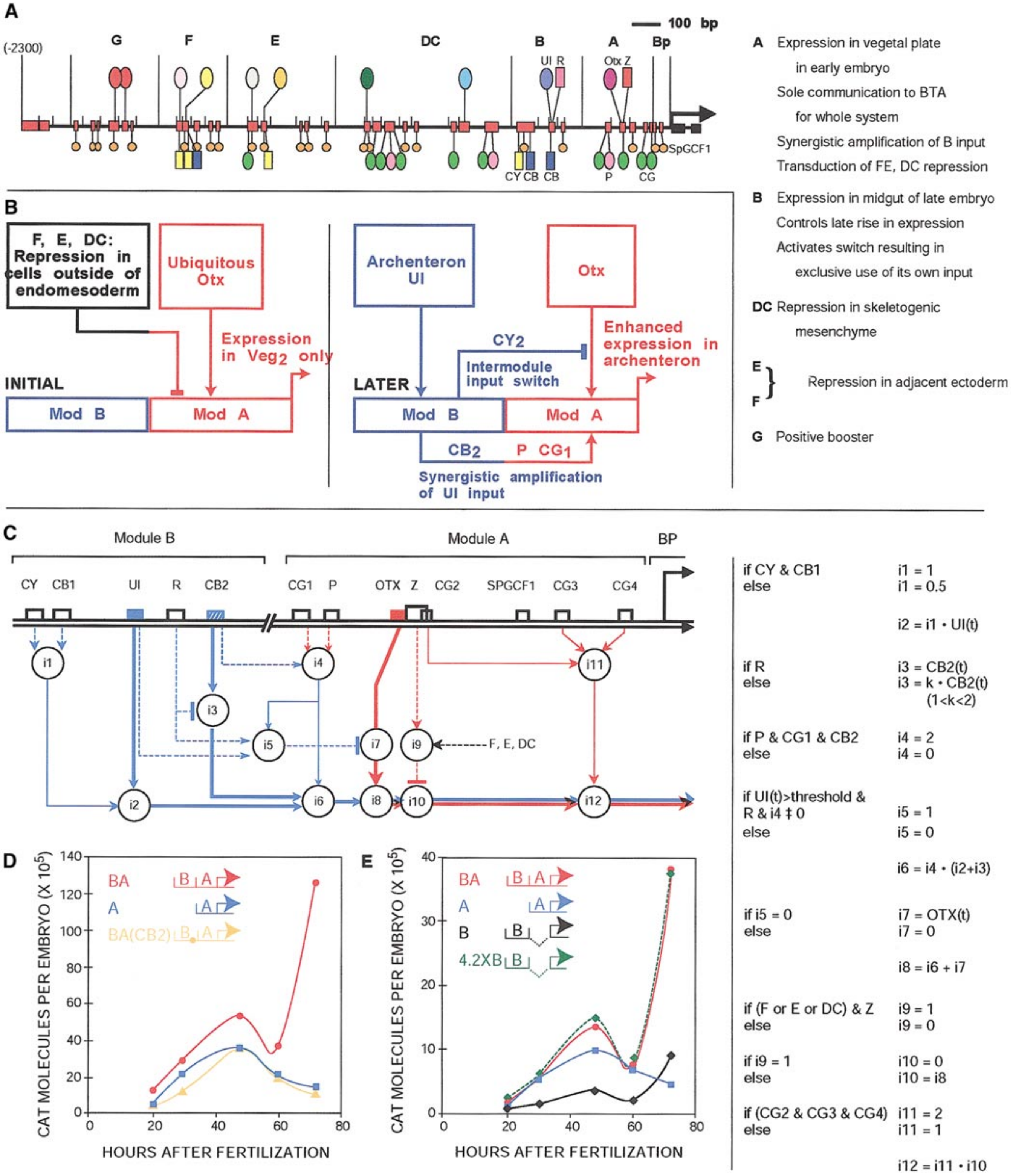
The *cis*-regulatory system of the developmentally regulated *endo16* gene of the sea urchin has been studied in perhaps greater depth than any other. This gene has a modestly complex pattern of expression during embryogenesis (Nocente-McGrath *et al.*, 1989; Godin *et al.*, 1996; Ransick *et al.*, 1993). It is activated in the vegetal plate of the embryo, specifically in the veg$_2$ lineage, at about eight cleavage; veg$_2$ consists of the progeny of eight sixth cleavage founder cells, and from it derives most of the endoderm, and all except the skeletogenic cell types of the embryonic mesoderm. The *endo16* gene is transcribed in this endomesodermal progenitor field until gastrulation, during which it is expressed throughout the invaginating archenteron but no longer in the mesodermal domain; then, as the gut becomes regionalized, expression is extinguished in the foregut and hindgut but accelerated in the midgut, its definitive locus of expression, and there it continues to be expressed in the feeding larva (Arnone *et al.*, 1997). The *endo16* gene encodes large and probably polyfunctional proteins which are secreted into the lumen of the midgut (Soltysik-Espanola *et al.*, 1994; Godin *et al.*, 1996). It is important to keep in mind for what follows that there is nothing unusual about this gene: it is a garden variety, cell type-specific differentiation gene, and its *cis*-regulatory system is no more complex than that of many other developmentally regulated genes (Arnone and Davidson, 1997).

The *cis*-regulatory system that controls *endo16* expression is about 2300 bp in length, and it consists of several clusters of target sites that turn out to execute distinct functions, and can be thought of as separable modular regulatory elements. These are indicated in Fig. 1A. The basal transcription apparatus (Bp in Fig. 1A) is entirely promiscuous and has no regulatory activity on its own. We use it in our laboratory to service regulatory elements expressed in every domain of the embryo. Linked to mutated variants of the *endo16 Ci*-regulatory apparatus, or specific fragments, or partially synthetic versions thereof, the readout from this Bp has revealed the function of every subregion of the *endo16* system and of every target site within those modules examined in detail (Yuh and Davidson, 1996; Yuh *et al.*, 1996, 1998, 2001). The upstream regions F-C in the protein binding map of Fig. 1A mediate repression outside of the veg$_2$ lineage during the specification period of development, i.e., the cleavage-blastula stage. Spatial repressors which bind in these regions prohibit expression in the overlying ectoderm and in the skeletogenic cell precursors in the center of the vegetal plate, as indicated. But in order for these repression functions to work, a specific target site is also required in the most proximal region, Module A.

Modules B and A carry out many interesting regulatory functions, and indeed they provide a first-rate illustration of the concept of *cis*-regulatory information processing. They also presented us with a novel challenge: how to design a *cis*-regulatory model that could accommodate the results of a fine-scale experimental analysis, and could explain exactly how Modules A and B work in terms of the individual functional operations mediated by their target sites. The basic features of the computational DNA sequence-based model to which we were led in the course of this work have turned out to be useful for systems that consist of networks of *cis*-regulatory elements, as well as for single *cis*-regulatory systems.

Modules B and A together have 17 target sites for factors that recognize and bind specifically at given DNA sequence motifs ("specific" here means $\geq 10^4$-fold preference for the target site vs other double-stranded DNA sequence). What is probably an architectural protein, SpGCF1, interacts at five of these sites (small orange circles in Fig. 1); this protein is capable of multimerizing and looping the DNA to which it is bound. The 12 other target sites are serviced by 9 different transcription factors (colored ovals in Fig. 1). Each species of interaction in the *endo16* system has a distinct and measurable functional meaning, as has also been seen in other *cis*-regulatory systems that have been looked at in this level of detail (Kirchhamer and Davidson, 1996; see reviews of Arnone and Davidson, 1997; Davidson, 2001). The roles of Modules A and B are summarized diagrammatically in Fig. 1B. Here, some major regulatory inputs and intermodular interactions are indicated by arrows, which essentially symbolize the flow of regulatory information. The diagram shows that Module A directs expression early, driven by the input called "Otx" (the name of the responsible transcription factor), while Module B takes over later. As expression is confined to the gut and further along, to the midgut, the system is driven by the input at the "UI" site. Module B turns off the input from Otx in Module A when its input from UI rises. This switch function confers the benefit that it frees the system from reliance on spatial repression in order to keep *endo16* off except in the endomesoderm, initially a necessity because of the widespread early activity of its Otx driver (Li *et al.*, 1999; Yuh *et al.*, 2001). The UI binding protein is in contrast apparently a dedicated endoderm factor. But it is not a very strong activator by itself, and the level of the UI input must be greatly stepped up, through another action of Module A, as also indicated in Fig. 1B. For some purposes, the level of knowledge represented in Fig. 1B will suffice. But Fig. 1B is entirely a black box model, which tells us nothing about how the functions shown are generated or programmed. The *cis*-regulatory system it represents is a genetic control element: to understand it, we need to know the meaning of the target sites shown in Fig. 1A in respect to the functions

FIG. 1. Information processing in the *endo16 cis*-regulatory system. (A) Map of protein interactions in the *endo16 cis*-regulatory system. The 2300-bp DNA sequence indicated as the horizontal line is necessary and sufficient to provide accurate expression of a reporter construct. Proteins that bind at unique locations are shown above the line, and proteins that bind at several locations are indicated below(Yuh *et al.*, 1994). Different colors indicate distinct proteins. "G–A" indicate the functional regions or modules as indicated to the right (from Davidson, 2001; adapted from Yuh and Davidson, 1994). (B) Diagram indicating major functions of Modules A and B, i.e., the B and A regions of (A). Arrows indicate positive inputs and outputs; barred lines indicate repressive interactions. Inputs and activators of Module B are shown in blue, and Module A is in red. (C) Computational model of regulatory function for Modules B and A. The regulatory DNA of *endo16* is shown as a horizontal strip at the top of the diagram. The individual binding sites are indicated by labeled boxes. Module B and its effects are shown in blue; Module A and its effects are shown in red. Logic interactions (I) are indicated by numbered circles. Each represents a specific regulatory interaction modeled as a logic operation. Note the two types of regulatory input: time-varying interactions (colored boxes), which determine the temporal and also spatial pattern of *endo16* expression, and time invariant interactions (open boxes), which affect the level of expression and control intrasystem output and input traffic. In the diagram, interactions that can be modeled as Boolean are shown as dashed lines; those which are scalar as thin solid lines; those which are time-varying quantitative inputs as heavy solid lines. The individual logic interactions are defined in the set of statements below the diagram. Here, statements of the form "If X," where X is the name of a target site, means that this site is present and occupied by the respective factor. If the site has been mutated (or if the factors were inactivated or eliminated), this is denoted by zero; or as the alternative ("else") to the site being present and occupied. The statements afford testable predictions of the output for any given mutation or alteration of the system (from Davidson, 2001; adapted from Yuh *et al.*, 2001). (D) Kinetic experiment, showing output of transgenes injected into sea urchin eggs (the ordinate shows reporter gene expression, as CAT molecules per embryo) as a function of time after fertilization. Each point is obtained from a lysate of 100 individual embryos and all points were obtained from the same batch of fertilized eggs. For details, error estimates, and biological aspects see Yuh *et al.* (2001). The red curve displays the output of a construct consisting of the normal BA sequence associated with the basal promoter (Bp of part A) plus the reporter gene assembly; the yellow curve the same except for mutation of the core of the CB2 site (see C); the blue curve, Module A alone with same Bp and reporter (from Yuh *et al.*, 2001 and The Company of Biologists Ltd.). (E) Kinetic experiment as in (D) but with a different batch of eggs and some different constructs: the red curve again shows the BA control for this batch of eggs; the blue curve again shows Module A alone; the black curve shows the output of Module B alone; the dashed green curve is the black curve multiplied arithmetically by the factor 4.2 (from Davidson, 2001); adapted from Yuh *et al.*, 1998).



A Expression in vegetal plate in early embryo
Sole communication to BTA for whole system
Synergistic amplification of B input
Transduction of FE, DC repression

B Expression in midgut of late embryo
Controls late rise in expression
Activates switch resulting in exclusive use of its own input

DC Repression in skeletogenic mesenchyme

E, **F** Repression in adjacent ectoderm

G Positive booster

```
if CY & CB1                   i1 = 1
else                          i1 = 0.5

                              i2 = i1 • UI(t)

if R                          i3 = CB2(t)
else                          i3 = k • CB2(t)
                                   (1<k<2)

if P & CG1 & CB2              i4 = 2
else                          i4 = 0

if UI(t)>threshold &
R & i4 ‡ 0                    i5 = 1
else                          i5 = 0

                              i6 = i4 • (i2+i3)

if i5 = 0                     i7 = OTX(t)
else                          i7 = 0

                              i8 = i6 + i7

if (F or E or DC) & Z         i9 = 1
else                          i9 = 0

if i9 = 1                     i10 = 0
else                          i10 = i8

if (CG2 & CG3 & CG4)          i11 = 2
else                          i11 = 1

                              i12 = i11 • i10
```

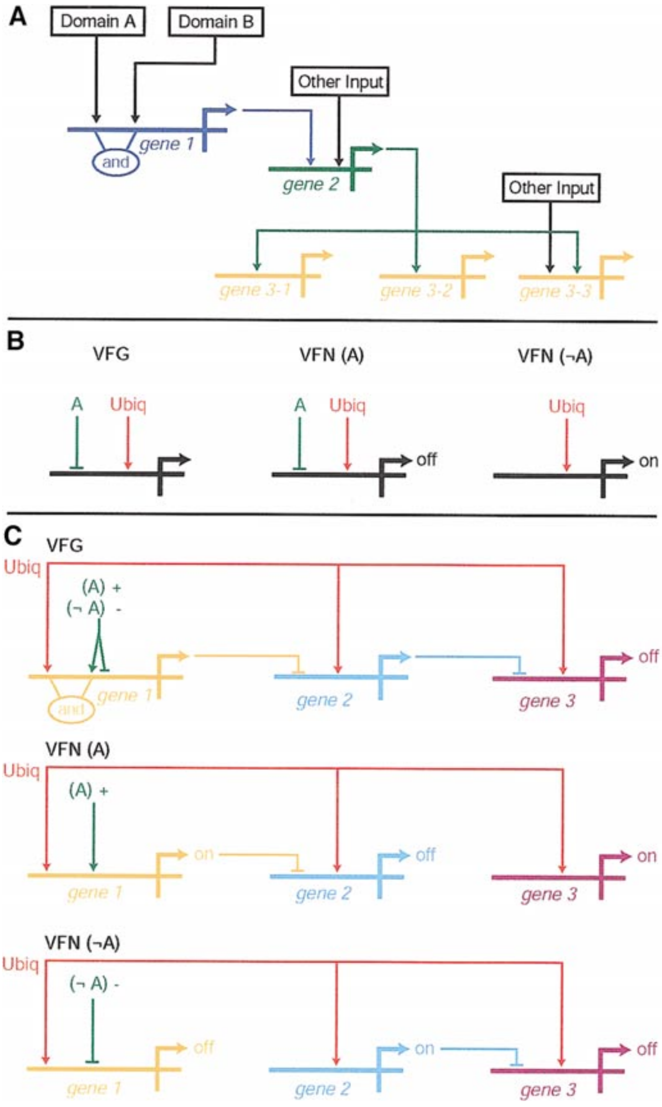


FIG. 2. Model elements for DNA sequence-based *cis*-regulatory network. (A) A network subelement consisting of five genes of which genes 1 and 2 encode transcription factors and genes 3-1 through 3-3 encode differentiation proteins. Thick horizontal lines represent *cis*-regulatory elements of these genes, the outputs of which are symbolized by bent rightward arrows. The targets of each of the regulatory genes are indicated by the thin solid lines (network linkages). The "and" indicates that both inputs are required (at productive levels) for the gene to be transcribed; otherwise, we may assume "or" logic, meaning that the gene will run given either input. Positive regulatory inputs are indicated by downward arrows. Domains A and B represent different spatial regions of the embryo which intersect, and where they interact the system runs. (B) Illustration of concept of VFG vs VFN. Symbolism as above. (A) represents a given spatial domain and (¬A, or not A) the remainder of the embryo; the barred line indicates a repressive interaction. (C) A three-gene network element, shown at the top as a VFG; in the center as a VFN in domain (A); at the bottom as a VFN in domain (¬A). The green input represents a regulator which acts as a repressor (−) unless it is modified, e.g., by a signal transduction event. This is the case in domain (A), when this module acts as a positive regulator (+). Gene 1 requires both positive inputs in order for it to be expressed ("and" logic).

indicated in Fig. 1B. This is the purpose of the model shown in Fig. 1C.

The target sites of Modules B and A are indicated in Fig. 1C by the boxes on the line representing the DNA at the top, blue for Module B and red for Module A. The arrows beneath lead from the target sites to the logic operations indicated in the circles and stated explicitly in the accompanying table. Our interest here is in the properties of the *endo16* model itself; see Yuh *et al.* (2001) for details, analysis, developmental implications, and experimental verification.

Logic Operations

The first of these properties is that the model specifies logic operations by which its inputs are processed and the altered value of these inputs carried forward. The simplest and most common of these is "and" logic: of the eight "if" statements in Fig. 1C (table), the first, third, sixth, and eighth are of this kind. The gist of these statements is that when all the conditions connected by "and" signs are met, then the indicated operation on the value of the regulatory output at that node in the internal network will take place; otherwise, something else will happen. There is a direct physical implication. This is that the proteins binding at the respective sites are together necessary for the function to occur (only specific DNA-binding proteins are directly relevant to a DNA sequence-based model). Or put another way, these proteins are in each case all obligate participants in a functional complex. So, a prediction for the biochemist would be that these proteins physically interact with one another. Their interactions could be cooperative: this would depend on their effective binding constants in the presence of one another, as opposed to singly. The implication of a physical complex is particularly clear and unequivocal in the case of the third "if" statement. The three sites referred to therein, CB_2, CG_1 and P, are required in order for a functional linkage between Modules B and A to exist, so that if any one of them is mutated, the transcriptional system becomes blind to the presence of Module B, even though the Module B DNA fragment remains physically linked to Module A. This outcome is illustrated in the kinetic experiment reproduced in Fig. 1D: the red curve shows the output of the complete BA system (i.e., of the native B-A-Bp fragment, associated with a reporter gene, the measured activity of which is expressed on the ordinate). The signature of Module B is its late rise in level of expression. This feature is entirely absent in the kinetics generated by Module A alone, as shown by the blue curve. The orange curve shows the output of the complete BA system, except that a few base pairs in the CB_2 target had been mutated. The result is that the output is the same as that of Module A alone, as if Module B did not even exist in the injected construct. The same happens if the CG_1 or P sites are mutated (Yuh *et al.*, 1998, 2001); hence protein binding to all three sites must be needed for the "linking" function to work.

"And" logic does not necessarily imply an all-or-nothing output, as is essentially the case in Fig. 1D. For example, the first and last of the "if" statements both describe elemental functions which cause an approximately twofold quantitative step-up of output at their respective nodes. In the first case (see Fig. 1C), if either or both the sites CY and CB_1 are absent, the output of the transcriptional driver of Module B, i.e., the transcription factor binding at the UI site, is about half its value as when they are present. Similarly, the CG_2, CG_3, and CG_4 sites contribute about a twofold step-up of whatever is the output of the whole upstream system at node i11. There is no unique mechanism implied, though the possibilities are constrained: for example, the proteins binding at the CY and CB_1 sites could together cooperatively improve UI site occupancy, or they could form a more effective "platform" with the UI-binding protein for an off-the-DNA coactivator; the protein binding the CG site could function architecturally by multimerization, thereby accounting for the need for all the sites; and so forth. The point is that the model describes the functions that are mediated by each site, conditional on the inputs presented. It does not attempt to describe the biochemistry of the proteins that actually contribute those functions. But, it is not hard to think of biochemical interpretations (that could of course be tested). Some of the logic statements in the *endo16* model are more abstract, if no less necessary and exact, for example, those which describe the intermodule switch shown in Fig. 1B. These are the fourth and fifth "if" statements. The fourth describes the conditions in which the switch turns off input from the Otx driver, i.e., the Otx transcription factor that binds at its target site in Module A. These conditions are: that the R site is occupied, where R is a site in Module B where a protein binds that is essential for this internal repression function (i.e., the site is not mutated, and the R binding protein is also present); that there is an input greater than some threshold from the UI site of Module B; and that Module B is linked to Module A in the normal fashion, dependent on the CB_2-CG_1-P system (node i4 of the diagram). The fifth statement describes what happens when these conditions are all met: the input from the Otx site (at i7) is wiped out, so that its value becomes zero. These statements are uniquely required by the experimental observations of Yuh *et al.* (1996, 1998, 2001). For example, in the intact system there is no kinetic input from Module A after blastula stage; while if the R site is mutated, the output is the sum of the outputs of Modules B and A, i.e., the output is much greater than normal at the stage when Otx input is highest, but not later when the Otx input has disappeared (see the blue curve in Fig. 1D). But how this switch actually works is not immediately implied, and hence the term "abstract." Perhaps the protein binding at R has a domain which is capable of interfering with the activation function of Otx, and this domain is exposed only when the UI protein binds at the adjacent site. Again, our point is that the object of the model is just to specify the consequences mediated by each target site according to the circumstances which determine the inputs at each site. The model states the functional meaning of the genetic sequences at which transcription factors bind, in terms of *cis*-regulatory information processing.

Continuous and Boolean Functions

The filled-in boxes from which heavy solid lines extend into the model in Fig. 1 indicate transcriptional regulatory inputs for which the amplitude varies over time (as well as in space). These kinetic inputs can be thought of as the time courses described by the concentrations (actually activities) of the transcription factors which bind at these sites, i.e., the UI, CB_2, and Otx sites. These inputs are processed in the manner specified in the model, and the outputs at any point on the heavy solid lines of the model are also kinetic. The final output for any given external condition is of course also a continuous kinetic function, as illustrated, for instance, for the three different cases in Fig. 1D.

A second class of interactions is indicated by sites denoted by open boxes from which emerge thin dashed lines. It is likely that the proteins that bind at these sites are always present in excess, or at least there is no evidence to the contrary. The significance of these sites is defined experimentally by what happens if they are mutated, and the functions that they give rise to can be regarded as Boolean operations: they occur or not, given the set of transcription factors that are present, depending on whether the sites are intact or not. The Boolean parts of the mechanism provide essential parts of the processing capacities of the overall system, but on their own they can produce no output unless there are inputs into the parts of the system that receive the kinetic drivers.

A third class of operations mediated by the *endo16* control system is indicated by open boxes from which extend thin solid lines. These indicate scalar operations on the inputs at the indicated nodes. An example is shown in Fig. 1E, where the scalar operation is multiplication by a constant factor of a kinetic input. For example, Fig. 1C shows that the output of Module B (at node i6) is that at i2 plus that at i3, i.e., the sum of the inputs of the kinetic drivers of Module B [UI(t) and CB_2(t)] × the value at i4. This operation depends on an intact CB_2-CG_1-P linkage subsystem (see Fig. 1C, table). When these conditions are met, the result is a scalar multiplication of Module B input into the system by a factor of about two. A further scalar multiplication of about this magnitude occurs if the CG_2, CG_3, and CG_4 sites are present. The overall amplification is illustrated kinetically in Fig. 1E. Here, the blue and black curves show the kinetic outputs of Modules B and A alone, the red curve (as in Fig. 1D) the output of the system when these two DNA fragments are in their normal configurations, joined together in *cis*. The dotted line shows what would be the arithmetic result of multiplying the output of Module B alone by a constant factor close to four at every point in time. This result closely approximates the measured output of the BA construct, illustrating the scalar amplification specified in the model. Scalar amplifications could be mediated by favorable architectural structure (DNA looping, nucleosome phasing, etc.) or even by the enzymatic activity of cofactors that use the proteins bound at the required target sites as anchors or platforms.

The *endo16* model is not a kinetic model per se, i.e., it does not consist of a set of time-based differential equations describing kinetic reactions. Instead, it describes the logic functions mediated by the DNA target sites. But nor is it a Boolean model, of which the output is either one state or another state. It is a model set up for processing kinetic inputs, and delivering kinetic outputs, because the key regulatory inputs into a transcriptional control system usually vary continuously over time. If one steps away from the individual *cis*-regulatory systems in order to survey whole regulatory networks that cause genes to be expressed in given spatial domains of an embryo and not in others, the character of the control system may superficially look more Boolean. But, in reality, neither are such networks Boolean, for each of the constituent *cis*-regulatory elements is like the *endo16* system: its job is to process incident inputs which by nature are continuous functions of developmental time.

Models for Networks of cis-Regulatory Elements: Symbolism and Significance

All major processes in animal development that we know about are driven forward by networks of regulatory genes, i.e., genes encoding DNA-binding transcription factors (see Davidson, 2001 for review). For convenience, these processes are customarily divided up into discrete packages, such as development of limbs from limb buds in tetrapods; or wings from wing imaginal discs in *Drosophila*; or specification and formation of the embryonic endomesoderm in the sea urchin embryo; or formation of the notochord in an ascidian embryo. Of course, none of these are really discrete developmental events, and the regulatory networks that control these processes are connected into other networks that control prior and surrounding processes in both temporal and spatial senses. However, in finding the reasonable boundaries of a gene regulatory network, we can make use of the concept that in bilaterian animals development of either embryonic structures or adult body parts begins with specification of a field of progenitor cells that will give rise to the part or structure. The beginning of the process for which the network displays the genetic program can be considered the installation of a specific transcriptional state in the progenitor field (Davidson, 2001). The end of the process, modeled at the termini of the network, is activation of differentiation gene batteries. In between, lies the major portion of the network, that which explains the regulatory interactions by which the transcriptional territories of the structure or body part are set up.

General Purposes of DNA Sequence-Based Network Models

The object of a DNA sequence-based gene regulatory network model for a developmental process is to state both the key inputs into the *cis*-regulatory systems of all the genes in the network, and those outputs which affect any other genes in the network. When complete, such models will be enormously informative: they will explain why each gene runs where and when it does, and why not elsewhere; how the spatial territories are progressively set up; and why the differentiated functions mounted on the structure are deployed where and when they are. But even incomplete such models are very useful, for so long as they are correct the interactions found always have some functional meaning, some explanatory power. The inputs into the *cis*-regulatory systems in these models link their control functions to their genomic DNA sequences. Like the *endo16* model, *cis*-regulatory network models are not intended to deal with what happens off the DNA, i.e., the cell biology and biochemistry of the transactions that create three-dimensional structures and differentiated cell function.

These models, then, are constructed of the *cis*-regulatory systems of the relevant genes and their inputs and outputs. Each *cis*-regulatory system is indicated in a bird's eye view that includes only its connections with other genes in the model. Of course if one were to zoom in on any given *cis*-regulatory element, it would appear more or less like the *endo16* model in Fig. 1C. Zooming out again, as if Modules A and B of the *endo16* gene were portrayed at the network level, apart from the spatial repressors, the only inputs that would likely be shown are those of the continuous drivers that bind at UI and Otx sites. Though the positive inputs in a *cis*-regulatory network model in general have dynamic properties, spatial repression of transcription must be treated somewhat differently. Repression is dominant in transcriptional control systems (Zhang and Levine, 1999; Davidson, 2001). When and where transcriptional silencers are brought into play the state of the target *cis*-regulatory element is simply “off.”

To illustrate the symbolism that we have found useful for *cis*-regulatory network models, an example is shown in Fig. 2A. Here, genes 1 and 2 encode transcription factors, and genes 3-1, 3-2, and 3-3 encode differentiation proteins. These are the terminal output of the network element shown, in that their expression affects no other genes in the network. Inputs are shown by the vertical downward arrows, and outputs by bent arrows. Two inputs that originate in different spatial domains of the embryo, but outside the network subelement in the diagram, are shown in boxes. Gene 1 carries out a spatial specification function: it is activated where the two spatial domains A and B intersect.

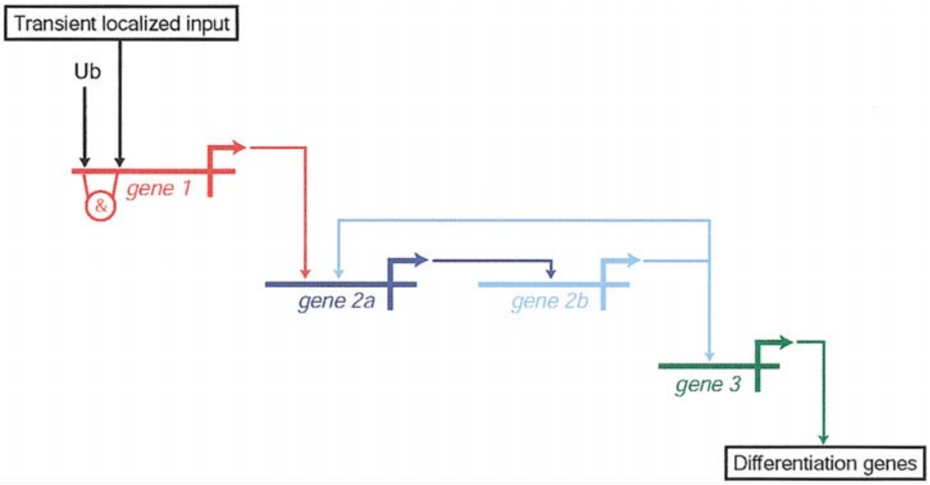


FIG. 3. A vectorial developmental process. Each stage is shown at a separate level, top to bottom, and all genes shown encode transcriptional regulators. The different levels represent distinct functional "layers" of the developmental process. In the initial stage, gene 1 undergoes a specification event. The "and" sign indicates that this *cis*-regulatory element requires both the input from a ubiquitous activator and the transient specification input (cf. Figs. 1 and 2 for examples of *cis*-regulatory "and" logic). In the second stage, genes 2a and 2b are activated. Gene 2a responds to gene 1, then activates gene 2b. Gene 2b cross-regulates gene 2a: the predicted target sites are, in gene 2a, sites for the gene 1 product and the gene 2b product; in gene 2b, sites for the gene 2a product. There only the target sites linking the genes in the diagram are shown: in life each such *cis*-regulatory element will respond to multiple inputs (e.g., see Arnone and Davidson, 1997). At the final stage shown, gene 3 is activated by gene 2b. Gene 3 is the dedicated controller of a differentiation gene battery (as in Fig. 2A), not shown.

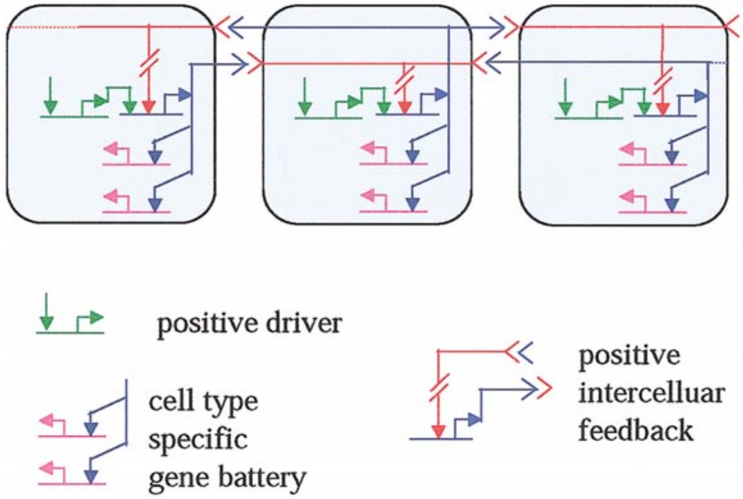

positive driver
cell type specific gene battery
positive intercelluar feedback

FIG. 4. Schematic of regulatory relations underlying the "community effect" phenomenon. The blue boxes represent adjacent cells, which intercommunicate by ligand (blue arrows)–receptor (red V-forms) interactions. The receptors are shown present constitutively and the ligands are present as the result of transcription from the gene colored blue, with the understanding that this could be an indirect effect, i.e., there could be an intermediate regulatory gene intervening between the signal transduction input (red arrow) and the ligand-encoding gene. This gene (also perhaps via an intermediate regulatory gene) is responding to a cell type-specific regulatory driver input (green). Downstream (directly or indirectly) of the same regulatory system lies a differentiation gene battery (blue and purple). The effect of the system is that it requires the intercellular interactions for expression of the differentiation gene battery; or put the other way around, the ligand–receptor interaction ensures the continuing cell type-specific pattern of gene expression. At the same time, the system is remarkably robust to variations in the absolute level of activity of the driver (green) gene.

This is a common device used in spatial regulatory programming in development, which depends on the use of *cis*-regulatory "and" logic. Gene 1 then activates gene 2. It then turns on the differentiation gene battery represented by genes 3-1, 3-2, and 3-3, all in the same domain initially defined by the intersection of the regulatory inputs present in domains A and B (though gene 3-3, using "or" logic, also runs in some additional domain). Note that each of the regulatory elements in Fig. 2A has a different kind of function. Gene 1 is responsible for transcriptional interpretation of the initial spatial inputs into the system, while gene 2 determines the presence of the factor controlling the differentiation gene battery.

The small model element illustrated in Fig. 2A has the key attributes required of DNA sequence-based network models. First, it represents the way a developmental transcriptional state is set up in a given spatial domain, according to the particular inputs that the hardwired *cis*-regulatory system will respond to. It shows why the differentiation gene battery is ultimately expressed where it is, in such a way that one can trace the regulatory "lines of authority." Second, it provides a direct set of testable predictions that can be refuted or proved by *cis*-regulatory analysis. For each input arrowhead is a specific prediction of one or more target site sequences for the indicated regulatory factor, that mediates a predicted function.

Genomic and Nuclear "Views"

A useful concept for DNA sequence-level network models is the distinction between the "view from the genome" (VFG) and the "view from the nucleus" (VFN), introduced by Arnone and Davidson (1997). The VFG shows all the immediately relevant interactions of which the *cis*-regulatory systems included are capable, whenever or wherever they occur. This is the view that is directly required for predicting what target sites will be present in the genomic sequence of any given *cis*-regulatory element. The VFNs focus on those sites that are occupied by the indicated inputs in any given nucleus at any given time, i.e., it shows what the regulatory system is doing at the time and in the place that each VFN snapshot is taken. The VFG is the sum of all the VFNs, over all time and space (just as the genome encodes developmental instructions for each gene over all time and space). A very simple illustration is shown in Fig. 2B, where two spatial domains of an embryo are indicated in parentheses at the top, *viz* domain (A), and the rest of the embryo (i.e., not A or $\neg$A). We see in the VFG that a certain *cis*-regulatory element is subject to activation by a ubiquitous regulator, providing that a (dominant) repressor for which there is also a target site is not also present. The repressor is present in domain A. The VFNs show the inputs the gene receives in the two domains, and the resulting "off" and "on" states. A slightly more complex example is shown in Fig. 2C, drawn from a real sea urchin embryo network element. There is a ubiquitous positive activator, "Ubiq," to which all three genes are capable of responding. However, gene 1 also requires a second positive input in order to be expressed ("and" *cis*-regulatory logic). This regulator acts positively in domain A, but in the absence of a signal which affects its behavior, i.e., in domain $\neg$A, it acts as a dominant repressor. Genes 1 and 2 both encode transcriptional repressors, while gene 3 encodes a protein that is not a transcription factor. The purpose of this network element is to regulate the spatial presence or absence of the gene 3 product. The VFG shows that gene 1 can repress gene 2, and gene 2 can repress gene 3, and it indicates all the possible inputs to which all three genes are able to respond. The result, how the system works in development, is shown in the VFNs: in domain A, gene 1 is expressed, thereby preventing gene 2 from repressing gene 3, so gene 3 is expressed in this domain. Elsewhere, gene 3 is silent. In general terms, the VFG tells how all the developmental possibilities are encoded in the genomic sequence; this view is genetic and time-independent. The VFNs tell us what is actually going on in any given developmental situation that we choose to look at.

We turn now to some of the kinds of process that can be represented in DNA sequence-based network models constructed according to these principles, and further implications that arise.

Inexorable Progression: Coding a Unidirectional Succession of Regulatory States

At a regulatory network level, development is very unlike most physiological processes, such as response to a metabolite or lack thereof, to a toxin, to the advent of a pathogen, or a nervous impulse, and so forth. Where they involve changes in gene expression, these all require self-limiting, regulatory transients that soon return to the initial state; they are homeostatic, and they flicker on and off depending on circumstances. Often, specific physiological responses represent the high point of the repertoire of differentiated functions of the cell that execute them. The developmental process is quite different. It begins with the specification of undifferentiated cells, the result of which is to produce in these cells a new transcriptional regulatory state, but then it at once moves forward into further regulatory states. Though there are some bizarre exceptions known in which given differentiated cells dedifferentiate and then redifferentiate as something else (often in the context of metamorphosis), developmental processes in complex animals are almost always unidirectional and progressive. They move forward, and never under natural conditions reverse directions. Only by perturbing the expression of regulatory genes that define a state through which the process traverses can the natural progression be altered.

Many different examples support the generalization that an initial state of specification always soon resolves into a more solid regulatory condition. This progression is caused by regulatory network processes which are also determined directly by the sequence structure of key *cis*-regulatory elements, just as is the initial process of specification. The outcome is that expression of genes activated in the course of specification becomes independent of the transient inputs which triggered it in the first place. The pattern of expression is now maintained through new regulatory interactions within the network. In the simplest cases, the cell lineage proceeds as directly as it can to installation of a process of terminal differentiation. This requires further regulatory states, in particular the activation of controllers of differentiation gene batteries, and of the downstream genes thereof. But in complex animals, the regulatory processes that control morphogenesis of body parts typically have a much longer series of successive states interspersed between specification and terminal differentiation. Each such state is defined by expression of a new set of transcriptional regulators, in a particular spatial domain and/or at a particular temporal stage (see Davidson, 2001 for review). We can use the term “network depth” to indicate the number of these successive states.

A causal explanation for the generally unidirectional quality of development emerges from the structure of some commonly encountered *ci*s-regulatory network subelements. A glance at Fig. 3 will serve for an illustration. Here, gene 1, encoding a regulatory protein, is participating in a typical specification function, responding to a transient localized input (like gene 2 in Fig. 2A). The input could be, for example, a localized maternal factor in an embryonic blastomere, or a transcription factor activated in consequence of signaling from adjacent cells. The *cis*-regulatory element of gene 1 operates by “and” logic, needing also an input from a ubiquitous activator in order for a positive function to result. The next state forward is shown in Fig. 3 at the next level of the diagram: this is the stabilization step, the state lock-down. Following specification (activation of gene 1), the protein it encodes is translated, accumulates, and transits into the nucleus. So after some delay required by these processes, gene 2a is activated. This in turn drives a sister gene, gene 2b, into transcriptional activity (with a further similar delay in real time). Gene 2b then participates in a stabilizing loop: its product in turn serves to activate expression of gene 2a, so that genes 2a and 2b become locked in a stimulatory embrace, so to speak. Genes 2a and 2b generate a robust and resilient device which ensures the operation of gene 3, in the next and final state generated by this network element. Gene 3 controls expression of a differentiation gene battery. Even after the transient input disappears and gene 1 becomes silent, genes 2a and 2b provide to the system a continuing memory of its advent.

One might reasonably ask why gene 2a, for example, does not perform the role of gene 3, i.e., why it does not itself operate the differentiation gene battery. But the reason why systems such as diagrammed here exist is that *cis*-regulatory networks are the products of evolutionary processes, not of a systems designer’s drawing board. Gene 3 plus its downstream structural genes represent a commonly conserved kind of regulatory cassette or network subunit, i.e., the differentiation gene battery (Davidson *et al.*, 2001). It is the role of genes 1, 2a, and 2b in the network subelement shown to install the function of this gene battery into the network by linking in the activity of its controller, gene 3 (i.e., rather than to rebuild the battery by substituting gene 2a for gene 3). The point is that the same differentiation gene battery is likely to be linked into other developmental subsystems as well.

A one-line summary of the import of Fig. 3 is that it shows how a set of genomic *cis*-regulatory elements can mandate a unidirectional series of successive developmental regulatory states.

Another regulatory network program device which acts to drive development forward by locking down the result of a specification event is the “community effect.” This is a term invented by J. Gurdon (Gurdon, 1988; Gurdon *et al.*, 1993), for a mechanism discovered in *Xenopus* embryos in which adjacent cells within a territory that have been specified to a mesodermal fate signal to one another. Disaggregation of these cells causes a severe attenuation in mesodermal gene expression. Similar observations have been reported from many other systems in which disaggregation of normally tightly joined cells results in catastrophic decrease in state-specific gene expression levels, from mouse liver parenchymal cells (Clayton and Darnell, 1983) to sea urchin embryos (Hurley *et al.*, 1989). These results can be combined with the now commonly seen expression of known signaling ligands throughout entire developmental territories (for examples in sea urchins, *Drosophila*, and mammalian development, see Davidson, 2001): the suspicion arises that “community effects” are far more widespread than the few well-studied cases. Perhaps it is generally true that multicellular spatial domains of gene expression consist of cells linked to one another by ligand–receptor interactions, which in turn help to drive state-specific gene expression.

The meaning of such a community effect mechanism at the *cis*-regulatory network level is shown in a general way in Fig. 4. Many other developmental phenomena can be similarly represented, for example, the repressive signaling interactions frequently observed across boundaries that separate domains of gene expression. Because there are many different kinds of signal transduction system, which affect transcriptional activity in different ways, it is not useful to consider this at a more detailed level. The point is anyway made: the community effect is another regulatory device which confers a vectorial direction on the developmental process, definitive functions of which are encoded in the genomic DNA sequence. It takes the cells out of the initial conditions of alternative transcriptional possibility that they confronted initially in the specification processes, and locks them into a stable state of gene expression, this time by means of intercellular reinforcement.

Concluding Remarks

cis-Regulatory network models serve as the developmental biologist’s essential organizer for getting causal relationships between genes straight. They are essential because there are too many genes and too many relationships that are too complex to deal with in any other way. But these models live in a strange place with respect to the traditional domains of bioscience. They are not in themselves chemical reaction models, though physical–chemical principles provide the stoichiometry and kinetics for the DNA–protein interactions they include; they are not in themselves genetic models though they specify the way the genetically inherited regulatory program works; they are not in themselves genomics models though their key physical elements are genomic target site sequence elements. The DNA sequence-based regulatory models we discuss here fall between all these stools, and perhaps that is why precedents have been lacking.

Any real set of relationships can be (and must be) looked at from different angles, at different focal lengths, or from what we term different "views"; so also for useful models of such relationships. Here, we have dealt with four different kinds of views. If we wish to focus in on a given *cis*-regulatory element, with the intent of understanding its internal organization, its means of processing continuous inputs, its switch functions, then we require a model which conveys the "worm's eye" view illustrated by our *endo16* model (summarized in Fig. 1). If instead we focus out, so as to be able to perceive the architecture of the interconnections between many such *cis*-regulatory elements, we need rather the kind of "bird's eye" view illustrated elementarily in the diagrams of Figs. 2–4. If we want to know all the *cis*-regulatory target sites required to explain when and where each gene in a network is expressed, so that we can "read" the regulatory DNA code, then the "view from the genome" is what is required. If we want to know what is going on in any given cell at any given time, then we need the "view from the nucleus." A strength of the kind of model we have been working with is that no transformations are needed to transit from one "view" to another, only a change in observational focus.

In the end, DNA sequence-based *cis*-regulatory models will permit closure to be brought on developmental gene network analysis, something not too common in the bioscience of our time. Developmental gene networks generate progressive changes in state, in time and space. Closure with respect to this particular problem will consist of understanding the identity of the regulatory interactions mediated by the DNA target site sequences; and in learning their logical consequences. The model enables decisive experimental tests for each interaction, by predicting its identity (i.e., the identity of the transcription factor target site, in terms of which regulatory gene it is a target of). These are all predictions that can be directly challenged by asking experimentally if the predicted target site is or is not present in the predicted *cis*-regulatory element. The logic relations in the model are also predictions: they say what will happen in terms of amount or location of gene expression, if the target site is mutated or if expression of the *trans* factor that binds there is experimentally altered. Furthermore, the model provides the opportunity to test its system properties "in silico," and to determine constraints on inputs, given its logic structure. Perhaps most important of all, the process of constructing the model is also the process of determining how the system works experimentally. Its construction is an interactive, information-dependent process. So modeling is understanding, step by step, or at least such has been our experience.

ACKNOWLEDGMENTS

We thank the reviewers of this manuscript, Dr. Jonathan P. Rast, and Prof. Ellen Rothenberg of Caltech. This work was supported by grants from the NIH (GM-61005) to E.H.D. and H.B. and from UK BBSRC (310/BIO12024) to H.B.

REFERENCES

Arnone, M., and Davidson, E. H. (1997). The hardwiring of development: Organization and function of genomic regulatory systems. *Development* **124,** 1851–1864.

Arnone, M. I., Bogarad, L. D., Collazo, A., Kirchhamer, C. V., Cameron, R. A., Rast, J. P., Gregorians, A., and Davidson, E. H. (1997). Green fluorescent protein in the sea urchin: New experimental approaches to transcriptional regulatory analysis in embryos and larvae. *Development* **124,** 4649–4659.

Clayton, D. F., and Darnell, J., Jr. (1983). Changes in liver-specific compared to common gene transcription during primary culture of mouse hepatocytes. *Mol. Cell. Biol.* **3,** 1552–1561.

Davidson, E. H. (1990). How embryos work: A comparative view of diverse modes of cell fate specification. *Development* **108,** 365–389.

Davidson, E. H. (2001). "Genomic Regulatory Systems: Development and Evolution." Academic Press, San Diego, CA.

Godin, R. E., Urry, L. A., and Ernst, S. G. (1996). Alternative splicing of the *Endo16* transcript produces differentially expressed mRNAs during sea urchin gastrulation. *Dev. Biol.* **179,** 148–159.

Gurdon, J. B. (1988). A community effect in animal development. *Nature* **336,** 772–774.

Gurdon, J. B., Lemaire, P., and Kato, K. (1993). Community effects and related phenomena in development. *Cell* **75,** 831–834.

Hurley, D. L., Angerer, L. M., and Angerer, R. C. (1989). Altered expression of spatially regulated embryonic genes in the progeny of separated sea urchin blastomeres. *Development* **106,** 567–579.

Kirchhamer, C. V., and Davidson, E. H. (1996). Spatial and temporal information processing in the sea urchin embryo: Modular and intramodular organization of the *CyIIIa* gene *cis*-regulatory system. *Development* **122,** 333–348.

Li, X., Wikramanayake, A. H., and Klein, W. H. (1999). Requirement of SpOtx in cell fate decisions in the sea urchin embryo and possible role as a mediator of β-catenin signaling. *Dev. Biol.* **212,** 425–439.

Nocente-McGrath, C., Brenner, C. A., and Ernst, S. G. (1989). *Endo16*, a lineage-specific protein of the sea urchin embryo, is first expressed just prior to gastrulation. *Dev. Biol.* **136,** 264–272.

Ransick, A., Ernst, S., Britten, R. J., and Davidson, E. H. (1993). Whole mount *in situ* hybridization shows *Endo16* to be a marker for the vegetal plate territory in sea urchin embryos. *Mech. Dev.* **42,** 117–1124.

Soltysik-Espanola, M., Klinzing, D. C., Pfarr, K., Burke, R. D., and Ernst, S. G. (1994). *Endo16*, a large multidomain protein found on the surface and ECM of endodermal cells during sea urchin gastrulation, binds calcium. *Dev. Biol.* **165,** 73–85.

Yuh, C.-H., and Davidson, E. H. (1996). Modular *cis*-regulatory organization of *Endo16*, a gut-specific gene of the sea urchin embryo. *Development* **122,** 1069–1082.

Yuh, C.-H., Ransick, A., Martinez, P., Britten, R. J., and Davidson, E. H. (1994). Complexity and organization of DNA-protein interactions in the 5′ regulatory region of an endoderm-specific marker gene in the sea urchin embryo. *Mech. Dev.* **47,** 165–186.

Yuh, C.-H., Moore, J. G., and Davidson, E. H. (1996). Quantitative functional interrelations within the *cis*-regulatory system of the *S. purpuratus Endo16* gene. *Development* **122,** 4045–4056.

Yuh, C.-H., Bolouri, H., and Davidson, E. H. (1998). Genomic *cis*-regulatory logic: Functional analysis and computational model of a sea urchin gene control system. *Science* **279,** 1896–1902.

Yuh, C.-H., Bolouri, H., and Davidson, E. H. (2001). *cis*-Regulatory logic in the *endo16* gene: Switching from a specification to a differentiation mode of control. *Development* **128,** 617–628.

Zhang, H., and Levine, M. (1999). Groucho and dCtBP mediate separate pathways of transcriptional repression in the *Drosophila* embryo. *Proc. Natl. Acad. Sci. USA* **96,** 535–540.

Received for publication December 18, 2001
Revised February 7, 2002
Accepted February 7, 2002