

Genomic Cis-Regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene

Chiou-Hwa Yuh, Hamid Bolouri, Eric H. Davidson*

The genomic regulatory network that controls gene expression ultimately determines form and function in each species. The operational nature of the regulatory programming specified in cis-regulatory DNA sequence was determined from a detailed functional analysis of a sea urchin control element that directs the expression of a gene in the endoderm during development. Spatial expression and repression, and the changing rate of transcription of this gene, are mediated by a complex and extended cis-regulatory system. The system may be typical of developmental cis-regulatory apparatus. All of its activities are integrated in the proximal element, which contains seven target sites for DNA binding proteins. A quantitative computational model of this regulatory element was constructed that explicitly reveals the logical interrelations hard-wired into the DNA.

The genomic organization of cis-regulatory systems lies at the nexus of development and evolution. Regulated transcription of thousands of genes controls the mechanisms by which morphological form and differentiated cell function are spatially organized during development, and the sequences of the transcription factor target sites in the regulatory DNA of each of these genes determines the inputs to which it will respond (1). Reorganization of developmental cis-regulatory systems and of the networks in which they are linked must have played a major role in metazoan evolution, because differences in the genetically controlled developmental process underlie the particular morphologies and functional characteristics of diverse animals. For both developmental and evolutionary bioscience, understanding genomic cis-regulatory systems is a central necessity.

We now present an experimental analysis of the multiple functions of a well defined cis-regulatory element that controls the expression of a gene during the development of the sea urchin embryo. The outcome is a computational model of the element, in which the logical functions mediated through its DNA target site sequences are explicitly represented. The regulatory DNA sequences of the genome may specify thousands of such information-processing devices.

The *Endo16* cis-regulatory system. *Endo16* is a gene that encodes a polyfunc-

tional secreted protein (2) of the midgut in the late embryo and larva. Transcription of the gene is activated soon after the primordial endoderm lineages are specified (in late cleavage), long before the gut forms (3, 4). *Endo16* transcription is specifically repressed in the embryonic cell lineages that are adjacent to the primordial endoderm or vegetal plate, that is, in cells that will give rise to ectoderm above the vegetal plate and to skeletogenic mesenchyme below (5). In the late blastula, all cells of the vegetal plate express *Endo16*, and after invagination this gene is expressed throughout the archenteron (5, 6). During gastrulation, the gene is activated in an additional ring of prospective endoderm cells surrounding the blastopore; soon after this gene is activated, these cells invaginate as well to form the hindgut (6). *Endo16* expression is thus an excellent marker of endoderm cell fate specification, in both the initial and later phases of that process. Toward the end of embryogenesis, *Endo16* expression becomes confined to the differentiating cells of the midgut (4). Transcription is extinguished in the foregut and the delaminating mesoderm in the late gastrula, and thereafter in the hindgut; however, there is an increase in the rate of transcription in the midgut, where it can still be detected in advanced feeding larval stages.

Earlier results have indicated the functional and structural organization of the *Endo16* cis-regulatory system (Fig. 1) (5, 7, 8). When introduced into sea urchin eggs, the DNA sequence extending about 2300 base pairs upstream from the transcription start site is necessary and sufficient to recreate the expression of a linked reporter gene in the same developmental and spatial pattern as displayed by the endogenous

Endo16 gene (7). Within this cis-regulatory domain (Fig. 1A), target sites have been mapped for 15 different proteins that bind with high specificity, that is, $\geq 10^4$ times their affinity for synthetic double-stranded copolymer of deoxyinosine and deoxycytidine [poly(dI-dC)poly(dI-dC)] (7). Though some have been identified and cloned, most of these proteins are known only by their molecular mass, their DNA binding properties, and their site specificity.

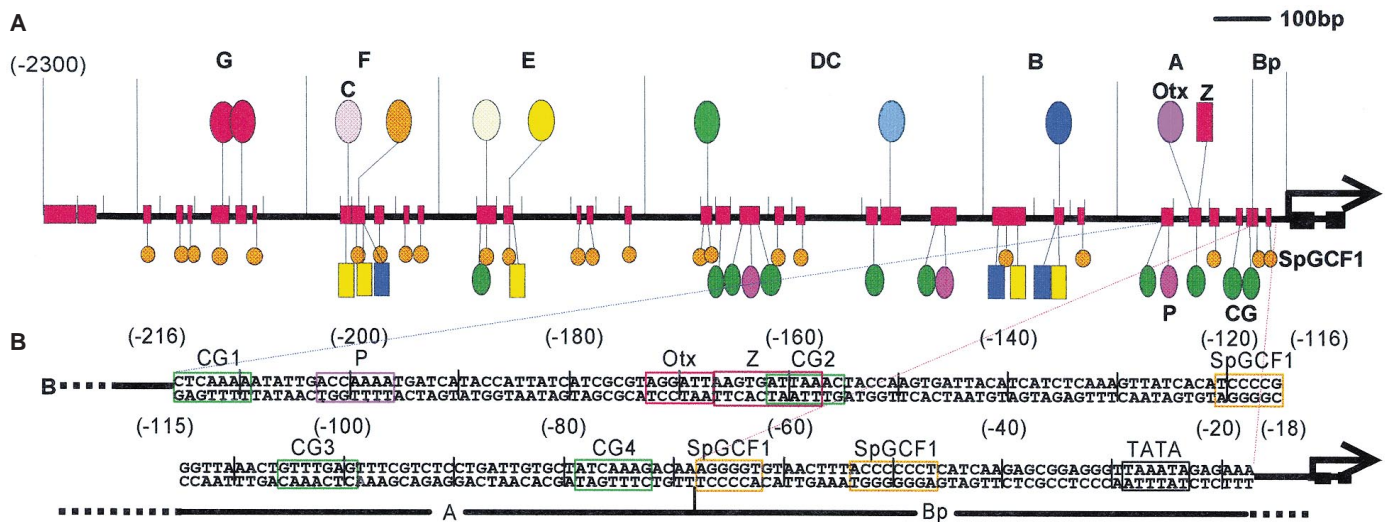
We have unraveled the functional organization of the 2300–base pair cis-regulatory system by determining the expression of constructs that include different subregions of the sequence, normal or mutated, or synthetic oligonucleotides representing versions of the specific target sites [see also (5, 8)]. Like other cis-regulatory systems that mediate complex developmental patterns of expression, the *Endo16* system is modular in organization (1, 9). That is, it consists of subelements of the DNA sequence, each of which can execute a certain regulatory function when included in a construct bearing either its own or a heterologous fragment of DNA on which the basal transcription apparatus will assemble. Each such subelement or regulatory module contains multiple target sites for DNA binding factors; there are typically four to eight different factors per module (1), and *Endo16* conforms to this expectation. The modular elements indicated by these experiments (5, 8) are indicated by the capital letters (G to A) in Fig. 1A. However, upstream of module B the boundaries of the subelements are as yet poorly defined.

When tested individually, the most distal element, module G, has the capacity to cause expression in the endoderm, as do modules B and A (5). However, their functions differ: Module G is relatively weak and appears to act throughout as an ancillary element; module B functions mainly in later development (5, 8), and after gastrulation it alone suffices to produce accurate midgut expression (5). Module A is probably responsible for initiating expression in the vegetal plate in the early embryo. In a construct that includes no other cis-regulatory subelements, the transcription-enhancing activity of module A rises early in development, but it then declines and disappears when expression is becoming confined to the midgut and module B becomes dominant (5, 8).

Under normal conditions, the central regions of the cis-regulatory system—that is, modules F, E, and DC (Fig. 1A)—have no inherent transcription-enhancing activity. Their role is to prevent ectopic *Endo16* expression in ectodermal cells descendant from blastomeres overlying the vegetal plate (modules F and E) and in skeletogenic cells (module DC). Thus, the positive reg-

C.-H. Yuh and E. H. Davidson are in the Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA. H. Bolouri is in the Engineering Research and Development Centre, University of Hertfordshire, Hatfield AL10 9AB, UK.

*To whom correspondence should be addressed. E-mail: davidson@mirsky.caltech.edu



C Module A functions:

Vegetal plate expression in early development:

Synergism with modules B and G enhancing endoderm expression in later development:

Repression in ectoderm (modules E and F) and skeletogenic mesenchyme (module DC):

Modules E, F and DC with LiCl treatment:

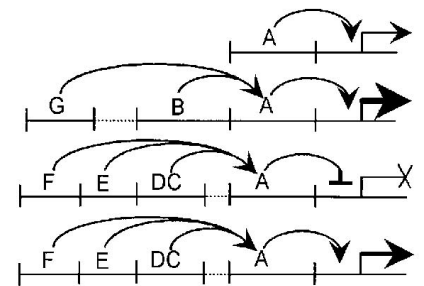


Fig. 1. *Endo16* cis-regulatory system and interactive roles of module A. (A) Diversity of protein binding sites and organization into modular subregions [modified from (7)]. Specific DNA binding sites are indicated as red blocks; modular subregions are denoted by letters G to A (Bp, basal promoter). Proteins binding at the target sites considered in this work are indicated: Otx, SpOtx-1 (12); SpGCF1 (14); the proteins CG, Z, and P, which are not yet cloned; and protein C [a CREB family protein (18)] in subregion F. Proteins for which sites occur in multiple regions of the DNA sequence (indicated by the black line) are shown beneath. (B) Sequence of module A and location of protein binding sites. Sites are indicated in the same colors as in (A). A fragment containing CG₃ and CG₄ sites as well as Bp has no endoderm-

specific activity and services other upstream cis-regulatory systems promiscuously; similarly, the *Endo16* cis-regulatory system functions specifically with heterologous promoters substituted for Bp (5, 8, 19). Boxed sequences indicate conserved core elements of the target sites (7, 12, 14), not the complete target site sequences. (C) Integrative and interactive functions of module A (5, 8). Module A communicates the output of all upstream modules to the basal transcription apparatus. It also initiates endoderm expression, increases the output of modules B and G, and is required for functions of the upstream modules F, E, and DC. These functions are repression of expression in nonendodermal domains and enhancement of expression in response to LiCl.

ulators that bind in modules G, B, and A are initially active in all of these domains as well as in the vegetal plate—that is, roughly in the whole bottom half of the embryo.

We previously discussed an interesting and useful effect of LiCl on the three repressor modules (5). This teratogen expands the domain of endoderm specification at the expense of the adjacent ectoderm (10, 11), and concomitantly, it expands the domain of *Endo16* expression (4). LiCl treatment abolishes the negative effect of the three repressor modules and instead causes them all to act as transcriptional stimulators. We have used this response, which is easy to assay quantitatively, in some of the following experiments.

In functional terms, module A interacts with all of the other *Endo16* cis-regulatory modules, and is either absolutely required for their operation or synergistically enhances their output. Moreover, it serves as a central switching unit, acting according to

inputs from the other modules. Our experiments delineate these encoded functions precisely. Module A functions are mediated through interactions at eight different target sites for DNA binding proteins (Fig. 1B). At least four different factors interact at these sites, only two of which have been cloned: an orthodenticle transcription factor family called *SpOtx-1* (12, 13) and a protein termed SpGCF1 (14). Because SpGCF1 multimerizes on binding to DNA, it may serve to mediate regionally specific DNA looping (14). There is one SpGCF1 site in module A, but two more occur downstream in the basal promoter (Bp) region (Fig. 1B). SpGCF1 sites occur commonly in sea urchin genes (7, 15, 16), and their function is usually manifested in gene transfer experiments as a weak stimulation of transcription (16, 17). Because the properties of this factor are known, the SpGCF1 sites of module A were not studied further. The module A sites labeled CG₁, CG₂,

CG₃, and CG₄ in Fig. 1B bind the same protein (7). However, the functional role of the CG₁ site differs from that of sites CG₂ to CG₄. The other two sites (P and Z in Fig. 1B) occur only within module A, as does the Otx site.

Recent evidence (8) had indicated that diverse and specific intermodule interactions are mediated by module A (Fig. 1C). First, module A communicates directly to the basal transcription apparatus (BTA) the status of the whole regulatory system, in that all the upstream modules work through it; in the normal endogenous arrangement they do not themselves interact directly with the BTA (8). Second, module A synergistically steps up the combined activity of modules B and G, boosting their output several fold; this becomes a particularly important function later in development. Third, module A is absolutely required either for the repressive function of modules F, E, and DC or for their LiCl response (8).

Considered together with its role in promoting early endoderm-specific gene expression, module A can be seen to execute a number of different regulatory functions. We sought to discern how these functions are programmed in the DNA sequence of module A, and to identify in precise terms the regulatory logic functions of module A.

Synergism and switch functions. When modules B and A are physically linked and joined to the Bp (BA construct), the transcriptional output is enhanced relative to the output of either module alone (Fig. 2A). This provides a classic example of synergism, in this case clearly mediated by interactions between module B and some elements of module A (8). Surprisingly, the output of the BA construct turns out to be exactly modeled by a simple linear amplification of the output of module B over the whole time course of module B activity. Thus, as also shown in Fig. 2A (8), the function executed by module A is to “multiply” the output of module B by a constant factor of about 4. Two questions then arise: (i) What element or elements of module A are specifically responsible for this function? (ii) Why does the combined BA output display the characteristics of the time course of module B, rather than some combination of the time courses generated by

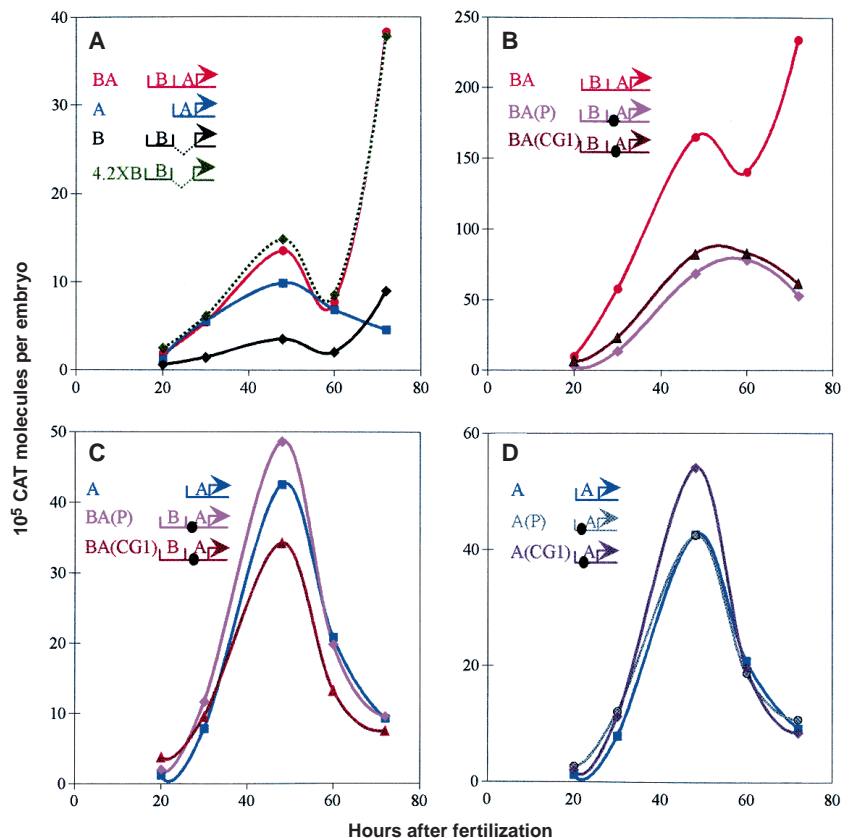
modules A and B when these are tested in isolation? The answer to both questions arises from a study of the effects of mutating the target sites designated CG_1 and P (Fig. 1B).

The experiments shown in Fig. 2, B to D, demonstrate that these sites provide the obligatory link between module B and module A and thence to the BTA, and that they mediate part of the synergistic enhancement of module B output. If either the CG_1 or P site in the BA construct is mutated, the output drops by about half during the early to middle period of development. However, the late rise in activity that is characteristic of both module B and the BA construct is completely absent (Fig. 2B). The BA(CG_1) and BA(P) constructs (we adopt the convention that a site in parentheses is mutated) now produce an output over time that is indistinguishable from that of module A alone (Fig. 2C). However, the same mutations have no discernible effect on the output of module A in itself (Fig. 2D). The following conclusions can be drawn: (i) Both the CG_1 and P sites are needed for function, in that a mutation of either produces the same quantitative effect (Fig. 2, B and C). (ii) The CG_1 and P sites constitute essential sites of module B interaction with module A, because in the absence of either,

the BA construct behaves exactly as if module B were not present (Fig. 2C). (iii) The CG_1 and P sites constitute the exclusive sites of interaction with module B, because no other mutations of module A sites have the effects shown in Fig. 2C. (iv) The CG_1 and P sites are dedicated to function (ii), because mutation of these sites has no effect whatsoever on the output of module A. (v) When linked to module A, module B does not communicate directly with the BTA except through the CG_1 and P sites of module A, confirming (8) on this point.

Furthermore, it is apparent (Fig. 2) that module A functions as a switch: When there is input from module B through the CG_1 and P sites, this input is amplified and transmitted to the BTA, and module A no longer has input to the amount of expression. Thus, the input is switched from that of module A to that of module B, even though module A displays the activity over much of the same time period. When there is no module B input, the output has the form of that generated by module A, completely lacking the late rise in the rate of expression. There is no module B input in the physical absence of module B (Fig. 2, A and D), when the CG_1 or P site of module A is mutated (Fig. 2B), or early in development (8). Module A becomes active first

Fig. 2. Roles of the CG_1 and P sites of module A. Constructs containing the indicated elements of the cis-regulatory sequence linked to the Bp (Fig. 1) and a CAT reporter gene were injected into fertilized eggs, and CAT enzyme activity was assayed on batches of 100 embryos per time point (5, 8). A given batch of eggs was used for each complete data set. **(A)** Time courses of modules A and B and of the BA construct, and demonstration that the time course of BA is a scalar amplification of that of B. The points represent averages from two similar experiments; in this and all further kinetic presentations, the smooth curves were generated by application of a standard derivative-matching (spline) algorithm. The green dotted line shows the time function generated by multiplying the data for the B construct by a factor of 4.2 at each measured time point [adapted from (8)]. **(B to D)** Time-course data generated by constructs in which CG_1 and P sites were mutated. Data are pooled from two or three experiments in each panel by normalizing to the 48-hour peak point (in the BA construct or module A) of the most active egg batch. The data were then averaged and the range or SD was calculated; the ordinates show CAT activities for the most active batch in each panel. The mutations used for the P target site (Fig. 1B) substituted an Eco RI site, 5'-GAATTC, for the sequence from (in base pairs) -208 to -198. The mutation used for the CG_1 target site substituted the same Eco RI sequence for the sequence from -216 to -209. Mutations, indicated schematically as black dots, were shown to abolish specific DNA-protein interactions. In (B), the effects of CG_1 and P mutations on output of BA are shown. Three data sets are pooled; SDs around the mean values plotted were $\pm 15\%$ of these values for the BA and BA(P) curves, and $\pm 30\%$ for the BA(CG_1) curve. In (C), BA(CG_1) and BA(P) are compared with wild-type module A. Two data sets are pooled; the range was $\pm 20\%$ of the values shown for module A, BA(P), and BA(CG_1) in this series of experiments. In (D), A(CG_1) and A(P) are compared with wild-type module A. Two data sets are



pooled; the range was $\pm 20\%$ of the mean points plotted for module A and $\pm 30\%$ for A(CG_1) and A(P).

and is probably responsible for installing expression in the vegetal plate shortly after endoderm specification. Thus, at the earliest quantitatively examined time point (20 hours), careful examination shows that there is no detectable output from module B when it alone is linked to the Bp, and the output of module A alone equals the output of either the BA (8) or GBA (5) construct [because the activity is relatively low at this point, it does not much affect the overall comparison of BA output and the calculated 4.2 times B output (Fig. 2A)]. Module A apparently contains a sort of “toggle switch” that either responds exclusively to the output of module B through the CG₁-P interaction, or, if this output is insignificant, transmits to the BTA the output of its own positive spatial regulator. Although not shown here, the rather low and almost constant output of module G is apparently added to that of module B in the complete cis-regulatory system, before the synergistic amplification performed by module A (5, 8).

Spatial and temporal patterns of expression generated by module A when isolated. The rise-and-fall time course of expression generated by module A when it is linked either to an SV40 (5, 8) or its own Bp (5, 8) (Fig. 2) depends for its form exclusively on interactions mediated by the Otx site (Fig. 3). This site is also necessary and sufficient to perform the early spatial regulatory function of module A, namely, to direct expression to the primordial endoderm lineages (as

well as to the surrounding cell tiers). Double-stranded oligonucleotides that included the Otx and Z sites were linked to a fragment bearing the BTA plus the CG₃ or CG₄ site (Fig. 1B). The CG₃-CG₄-Bp fragment itself has very low transcriptional activity and virtually no endoderm activity (5). In additional constructs, the Z and Otx sites were alternatively altered, and spatial and temporal activity were assessed. In the following, we use the convention that oligonucleotides included in constructs are indicated in italics (as above, mutations are indicated as parentheses around the affected sites).

In a typical embryo bearing the *OtxZ* construct and expressing chloramphenicol acetyltransferase (CAT) mRNA in endoderm cells (Fig. 3A), the Otx site alone suffices to generate endoderm expression, and the Z site is irrelevant. Furthermore, mutation of four base pairs in the core of the Otx site in an otherwise wild-type module A sequence [*A(Otx)* construct] abolishes its ability to promote expression in the endoderm (Fig. 3B). This mutation also destroys most of the transcriptional activity of module A. On the other hand, the *OtxZ* construct is able to produce a typical module A expression time course, although of lower amplitude (Fig. 3B; this result is also dependent only on the Otx site, and the Z site of the *OtxZ* oligonucleotide is again irrelevant). These experiments demonstrate discrete functions of module A, which are mediated exclusively by the Otx site.

It would appear that in the BA con-

struct, these same functions are “disconnected” when the regulatory switch inferred above instead “connects” the input of module B. Over the period measured, the kinetic output of the BA construct consists exclusively of the synergistically amplified input of module B, and no contribution from the module A time course can be detected (Fig. 2A). Although the Otx mutation abolishes the activity of module A when it is tested by itself [*A(Otx)* construct, Fig. 3B], it would then be predicted that this same mutation should not affect the expression of the BA construct over most of the period of measurement. This quantitative prediction is confirmed in the experiments summarized in Fig. 3C.

Transduction of input from modules F, E, and DC. Another function of module A is to mediate the activities of modules F, E, and DC, which in the normal embryo are responsible for confining expression to the endoderm by repressing the gene outside of this domain. Here, we used LiCl responsiveness as an index of module F function. As found earlier, enhanced expression resulting from LiCl treatment requires that both module A and one or more of modules F, E, and DC be present in the construct. Evidence with respect to module F is abstracted from (5) in the upper portion of Fig. 4. When linked to module F, module A generates a clear LiCl response, whereas modules B and G are blind to LiCl treatment, as is module A in the absence of module F. LiCl treatment causes an enhancement of

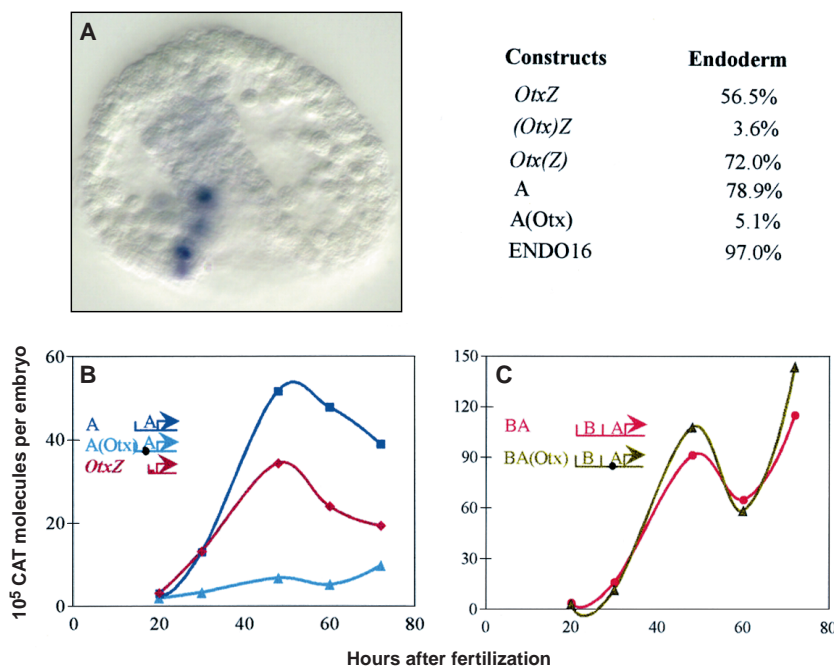


Fig. 3. Role of the Otx site in generating spatial and temporal expression of module A. **(A)** Spatial expression. Eggs were injected with a construct containing a synthetic oligonucleotide that includes an intact Otx site, plus Bp and CAT reporter sequences (construct *OtxZ*). The location of CAT mRNA expression was determined by whole-mount in situ hybridization at the gastrula stage. Cells expressing the construct are located in the gut endoderm. The ability of relevant constructs to specify expression in endoderm (80 to 200 embryos per construct) is shown at the right. Italics denote synthetic oligonucleotides included in the construct, together with the Bp and CAT reporter; parentheses denote that the site named was mutated, either in the normal module A sequence [*A(Otx)*] or in an oligonucleotide [*(Otx)Z*]. The *OtxZ* oligonucleotide includes the sequence from -175 to -157 (Fig. 1B), with 5'-TCGA (Xho I) and 5'-AGCT (Hind III) tags for directional cloning added on the 5' end of each strand. In the *(Otx)Z* oligonucleotide, the sequence ATTA in the core of the Otx site was changed to GCCG, and the Z target site sequence TGATTAA was changed to CAGCCGG (see Fig. 1B). In the *A(Otx)* mutation, a sequence including an Xba I site, TCTAGA, was substituted for the natural Otx site sequence GGATTA. For simplicity, the additional ectopic expression generated by all constructs lacking negative modules (5) is not shown. **(B)** Temporal and quantitative expression. Three data sets were pooled. SDs for wild-type module A and *A(Otx)* time courses were $\pm 25\%$ and $\pm 30\%$, respectively,

around the mean values shown. Only one experiment from the same batch of eggs is shown for the *OtxZ* construct. **(C)** Lack of effect of Otx mutation on expression of the BA construct. Three data sets were normalized and pooled to compare expression of BA and *BA(Otx)*. The outputs of these constructs are indistinguishable; SDs around the mean values shown are $\pm 30\%$.

expression by a factor of 2 to 3. This is attributable to the expansion of the spatial domain of expression at the expense of the ectoderm, as well as to the intensification of expression (5), was unequivocally observed (Fig. 4, bars indicating SDs on these measurements). The key sequence element of module F is a target site that binds a factor of the cyclic adenosine 3',5'-monophosphate response element-binding protein (CREB) family.

When oligonucleotide C, which includes this target site, is linked to module A (construct C-A), it confers LiCl sensitivity almost as well as does the whole of module F (Fig. 4). The interaction mediated by the C oligonucleotide also requires the Z site of module A, whereas the Otx site is irrelevant. This is shown both by mutations of the Z and Otx sites in an otherwise intact module A that has been linked to the C

oligonucleotide [C-A(Otx) and C-A(Z) constructs, Fig. 4] and by experiments in which all three sites are represented only as oligonucleotides [C-OtxZ, C-Otx(Z), and C-(Otx)Z constructs, Fig. 4]. These experiments identified the Z site of module A as the element specifically required for functional interactions with module F. Because the other repressor modules, E and DC, behave identically to F (5), we presume the Z site is used for all of these interactions. It thus appears that the obligatory and exclusive role of the Z site is to transduce the input of these upstream modules. In the absence of module A, these elements have no effect on the output of the cis-regulatory system (5) (Fig. 4).

Interaction with the basal transcription apparatus. An indication that the CG₃ and CG₄ sites are directly involved with interactions between module A and the adjacent BTA came from a comparison of the activities of the SV40 and endogenous promoters, linked to *Endo16* cis-regulatory elements (5, 8). When combined with a truncated version of module A lacking the CG₃ and CG₄ sites, the SV40 promoter was less active by a factor of ~2, but if an oligonucleotide bearing only these sites was inserted, its activity became indistinguishable from that of the endogenous promoter. However, this enhancement was not seen with module B, implying that module A contains an additional element that mediates interaction with CG₃ and CG₄ sites and is important for communication with the BTA. An obvious guess was that this site is the nearby CG₂ site (Fig. 1B). Mutation of the CG₂ site of the BA construct

caused its activity to decrease by half without affecting the characteristic shape of the time course (Fig. 5A).

The overall fourfold amplification of the output of module B by module A is attributable to the combined effects of the CG₁ and P sites (2× amplification) and the CG₂ through CG₄ sites (2× amplification). However, unlike the CG₁ and P sites, CG₂ is not dedicated to synergism with module B, because in contrast (Fig. 2D) the CG₂ site affects module A output by the same factor (Fig. 5B). Thus, the CG₂ site appears to process both positive upstream inputs. However, CG₂ function requires CG₃ and CG₄ sites as well: Mutation of either or both of these sites has exactly the same effect as does the CG₂ mutation on module A output (Fig. 5B). All three sites are evidently used for interaction with the adjacent BTA, in the process of which the level of expression is approximately doubled.

A computational model for module A functions. The experiments summarized (Figs. 2 to 5), together with earlier data (5, 8), specify dedicated functional roles for each of the seven module A target sites. In vivo, the occupancy of these sites will depend on the activity and concentration of the transcription factors that bind them, the null limit of which we have established by site mutations. Module A can be considered to execute a set of logic operations, according to site occupancy and to input from other upstream modules of the *Endo16* system. These operations are represented explicitly in the computational model of Fig. 6. This model was built and tested stepwise: Time-course data obtained in

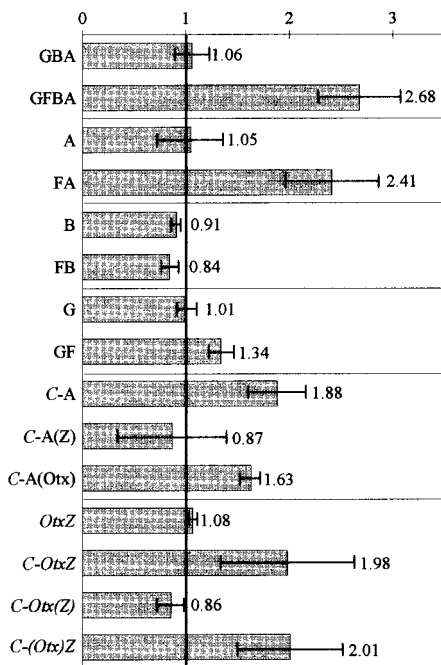


Fig. 4. Response of expression constructs to LiCl treatment of embryos. The shaded bars and numerical values give the mean ratios of CAT activity measured at 48 hours in samples of 100 LiCl-treated embryos to CAT activity in untreated embryos of the same batch; SDs are indicated by the terminated bars. If a construct lacks elements required to respond to LiCl, the ratio will be about 1. Constructs are indicated at the left; as above, capital letters indicate the subelements shown in Fig. 1A, parentheses denote mutations of the indicated target sites, and italics indicate oligonucleotides (see Fig. 3 for mutations and oligonucleotides). For the A(Z) mutation, a sequence including an Xba site, TCTAGA, was substituted for the sequence from -164 to -158 (Fig. 1B). The C oligonucleotide, 5'-GTGTGTGCGTGCTCTCAC-CTCA, includes the target site for a CREB factor binding in module F (20).

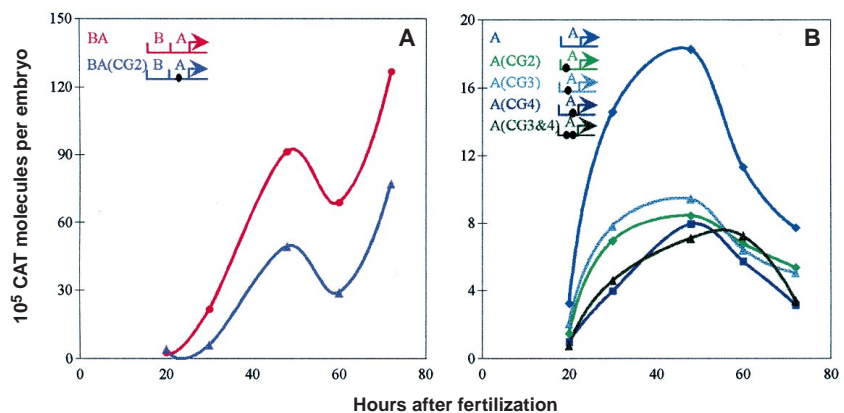


Fig. 5. Role of CG₂, CG₃, and CG₄ sites. **(A)** Effect of mutations of the CG₂ site on expression of the BA construct. An Xba I site, 5'-TCTAGA, was substituted for the natural sequence from -164 to -155. Although this mutation also destroys part of the Z target site, the latter does not affect expression unless one or more of modules F, E, or DC is present (see text). This mutation prevents binding of the CG factor to the CG₂ site. Two data sets were pooled; the range was about 18% around the mean values shown. **(B)** Effect of mutations of CG₂, CG₃, and CG₄ sites on module A activity. A(CG₃) and A(CG₄) were generated by substituting the same Xba I site as above for the natural sequence from -107 to -96 (CG₃ mutation) and from -79 to -66 (CG₄ mutation), thus destroying the ability of these sites to bind the CG protein. Three sets of data were pooled. The range was ±5 to 10% for A(CG₃), A(CG₄), and A(CG_{3&4}), and ±20% of the mean values shown for wild-type module A and A(CG₂).

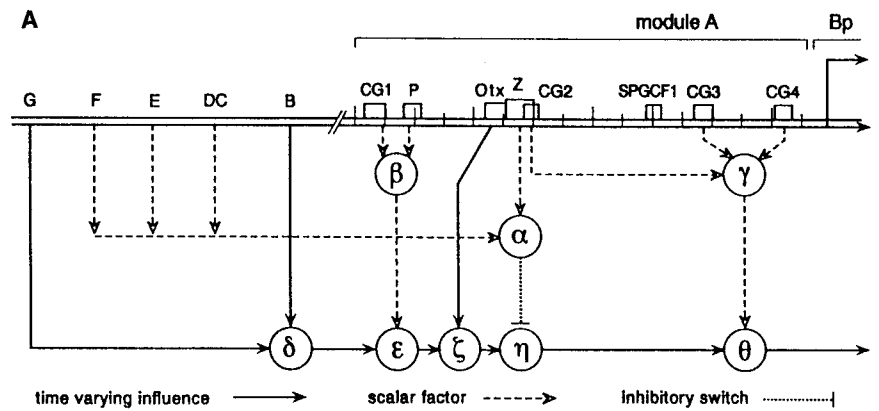
each series of experiments were tested against model output in the computer, essentially as in (8), and the logic statements were revised or new statements added progressively. The process by which the model was built engendered many of the experiments presented in Figs. 2 to 5, in that these experiments were designed to test predic-

tions of the model or to decide between alternatives.

The properties of module A enable the extensive upstream cis-regulatory apparatus of the *Endo16* gene to respond to the “instructions” (that is, the set of transcription factor activities) presented in each cell at each time in development. The model of

Fig. 6 interprets all of the multiple interactive roles of module A indicated in Fig. 1C: its synergistic amplification of module B input, its B versus A switch function, its mediation of the functional input from module F (and, by extension, modules E and DC), and its communication with the BTA.

Fig. 6. Computational model for module A regulatory functions. **(A)** Schematic diagram of interrelations and functions. Interrelations between upstream modules (G to B; Fig. 1A) and specific module A target sites demonstrated experimentally in this work, and among the module A target sites, are indicated beneath the line representing the DNA. The region from module G to module B is not to scale. Each circle or node represents the locus in the system of a specific quantitative operation, conditional on the state of the system; operations are specified for all relevant states in (B). Operations at each node are carried out on inputs designated by the arrows incident on each circle, and produce outputs designated by arrows emergent from each circle. Open arrowheads indicate inputs to the indicated node that are constant through time, the values of which are specified according to the logic sequence in (B); closed arrowheads indicate time-varying inputs (such as illustrated in Figs. 2, 3, and 5). The terminated bar indicates a Boolean repression function that under given conditions extinguishes activity at node η . **(B)** Logic sequence for operation of model shown in (A). The value 0 denotes that a given site or module site has been mutationally destroyed or is inactive because its factor (or factors) is missing or inactive; the value 1 indicates that the site or module is present and productively occupied by its cognate transcription factor. For the case of modules F, E, and DC, a Boolean representation is chosen because ectopic expression is essentially zero (beyond technical background) in ectoderm and mesenchyme when these modules (together with module A) are present in the construct (5); otherwise, ectopic expression occurs. Similarly, when they are present, LiCl response occurs; otherwise, it does not (5) (Fig. 4). Sites within module A are designated as above. The logic sequence specifies the values attained at each operation locus [circles in (A)], either as constants determined experimentally and conditional on the state of the relevant portions of the system, or in terms of time-varying, continuous inputs designated by the symbol (t) . With respect to these constants, the value $\beta = 2$ derives from the measurements showing that of the total synergism factor of 4 (Fig. 2A), the CG_2 - CG_4 system accounts for a factor of 2, so the remainder is asayed to CG_1 and P function; the value $\beta = 0$ derives from the experiment shown in Fig. 2C. The values $\gamma = 2$ or $\gamma = 1$ derive from the experiments of Fig. 5 (see text). See text for the several conditions when $\epsilon(t) = 0$, that is, when the input from module B approximates 0. The kinetic output of modules B and G and of the Otx site are represented as $B(t)$, $G(t)$, and $Otx(t)$, respectively. The input $B(t)$ can be observed as the CAT activity profile generated by module B over time in Fig. 2A [see also figure 2 of (8) and figure 3 of (5)]; $G(t)$ is shown in the same figures in (8) and (5). $Otx(t)$ is the time course generated by the construct *OtxZ* in Fig. 3B (21). It is assumed here that module G does not contribute significantly relative to module B (5), and it is included only for completeness. A second assumption of the model is that, as



B

- if (F = 1 or E = 1 or CD = 1) and (Z = 1)** Repression functions of modules F, E, and DC mediated by Z site
 - $\alpha = 1$
- else** $\alpha = 0$
- if (P = 1 and CG₁ = 1)** Both P and CG₁, needed for synergistic link with module B
 - $\beta = 2$
- else** $\beta = 0$
- if (CG₂ = 1 and CG₃ = 1 and CG₄ = 1)** Final step up of system output
 - $\gamma = 2$
- else** $\gamma = 1$
- $\delta(t) = B(t) + G(t)$ Positive input from modules B and G
- $\epsilon(t) = \beta * \delta(t)$ Synergistic amplification of module B output by CG₁-P subsystem
- if ($\epsilon(t) = 0$)** Switch determining whether Otx site in module A, or upstream modules (i.e., mainly module B), will control level of activity
 - $\zeta(t) = Otx(t)$
- else** $\zeta(t) = \epsilon(t)$
- if ($\alpha = 1$)** Repression function inoperative in endoderm but blocks activity elsewhere
 - $\eta(t) = 0$
- else** $\eta(t) = \zeta(t)$
- $\theta(t) = \gamma * \eta(t)$ Final output communicated to BTA

indicated in (5), LiCl response can indeed be used as a surrogate for the normal in vivo function of the F-Z system, and by extension the E-Z and DC-Z systems. These functions are normally to repress expression outside of the endoderm. However, no direct measurements of spatial repression were carried out in this work. The final output, $\theta(t)$, can be thought of as the factor by which, at any point in time, the endogenous transcriptional activity of the BTA is multiplied as a result of the interactions mediated by the cis-regulatory control system. Programming and analysis were carried out with MATLAB (MathWorks Inc.).

Figure 6 indicates that the DNA sequence of module A specifies what is essentially a hard-wired, analog computational device. The requirement for this logic device is that there are many different inputs to the regulatory system that must be sorted appropriately. It is to us a remarkable thought that every developmentally active gene in the organism may be equipped with devices of this nature. *Endo16* is a peripheral terminus of a regulatory gene network that includes all the genes encoding the transcription factors that direct its activity as well as the genes controlling them. But this network is to be considered not only a collection of genes, but also a network of linked regulatory devices that specify operational logic processes. This concept seems to be of basic importance in considering the operation of metazoan genomes in development as well as their origin and diversification in evolution.

The various functions mediated by module A are precisely encoded in the DNA sequence. Each target site sequence has a specific, dedicated function: Take the site away and the function is abolished; put it back in the form of a synthetic oligonucle-

otide and it reappears. Only one of the seven sites in this example is directly involved in spatial regulation. This is a warning, in that many current studies of developmental cis-regulatory organization focus exclusively on sites defined qualitatively in terms of spatial regulation. The other six sites of module A are all involved with the operation of the cis-regulatory system in itself. The properties of the module A regulatory apparatus enable it to process complex informational inputs and to support the modular, polyfunctional organization of the *Endo16* cis-regulatory system. Perhaps the main insight from this experimental exploration is that these system properties are all explicitly specified in the genomic DNA sequence.

REFERENCES AND NOTES

1. M. Arnone and E. H. Davidson, *Development* **124**, 1851 (1997).
2. M. Soltysik-Espanola, D. C. Klinzing, K. Pfarr, R. D. Burke, S. G. Ernst, *Dev. Biol.* **165**, 73 (1994).
3. C. Nocente-McGrath, C. A. Brenner, S. G. Ernst, *ibid.* **136**, 264 (1989).
4. A. Ransick, S. Ernst, R. J. Britten, E. H. Davidson, *Mech. Dev.* **42**, 117 (1993).
5. C.-H. Yuh and E. H. Davidson, *Development* **122**, 1069 (1996).
6. A. Ransick and E. H. Davidson, *Dev. Biol.*, in press.
7. C.-H. Yuh, A. Ransick, P. Martinez, R. J. Britten, E. H. Davidson, *Mech. Dev.* **47**, 165 (1994).
8. C.-H. Yuh, J. G. Moore, E. H. Davidson, *Development* **122**, 4045 (1996).
9. C. V. Kirchhamer, C.-H. Yuh, E. H. Davidson, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 9322 (1996).
10. S. Hörstadius, *Experimental Embryology of Echinoderms* (Clarendon, Oxford, 1973), Chapter 6.
11. R. A. Cameron and E. H. Davidson, *Dev. Biol.* **187**, 236 (1997).
12. L. Gan *et al.*, *ibid.* **167**, 517 (1995).
13. X. Li, C. K. Chuang, C. A. Mao, L. M. Angerer, W. H. Klein, *ibid.* **187**, 253 (1997).
14. R. W. Zeller *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 2989 (1995).
15. C. V. Kirchhamer and E. H. Davidson, *Development* **122**, 333 (1996).
16. K. Akasaka *et al.*, *J. Biol. Chem.* **269**, 20592 (1994).
17. J. A. Coffman, C. V. Kirchhamer, M. G. Harrington, E. H. Davidson, *Development* **124**, 4717 (1997).
18. C.-H. Yuh and E. H. Davidson, unpublished data.
19. C. V. Kirchhamer, L. D. Bogarad, E. H. Davidson, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13849 (1996).
20. L. F. Michael, J. L. Alcorn, E. Gao, C. R. Mendelson, *Mol. Endocrinol.* **10**, 159 (1996).
21. C.-H. Yuh, E. H. Davidson, H. Bolouri, data not shown.
22. We thank R. J. Britten and J. A. Coffman for perspicacious and critical reviews of this manuscript. We also thank many other colleagues who have discussed this work with us. Supported by a grant from the National Institute of Child Health and Human Development and by the Stowers Institute for Medical Research.

15 October 1997; accepted 26 January 1998

Discover a new sequence.

Visit the SCIENCE Online Web site and you just may find the key piece of information you need for your research. The fully searchable database of research abstracts and news summaries allows you to look through current and back issues of SCIENCE on the World Wide Web. Tap into the sequence below and see SCIENCE Online for yourself.

www.sciencemag.org

SCIENCE