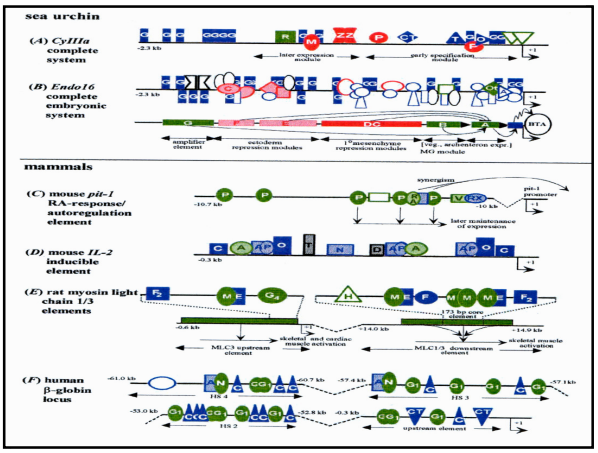
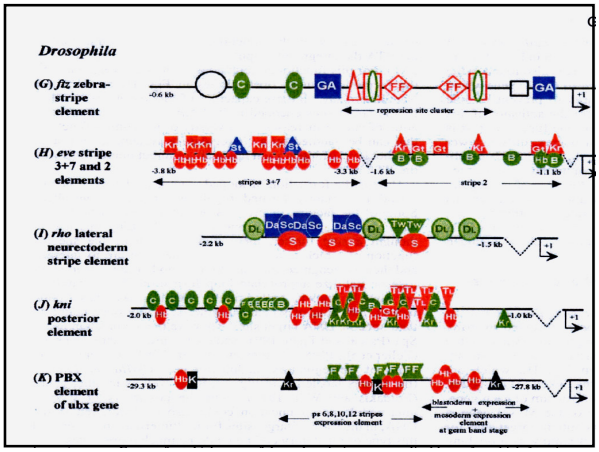
Realistic Models of Transcription

Endo 16

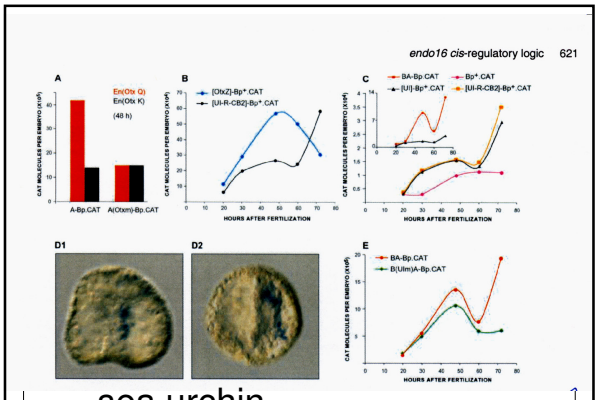




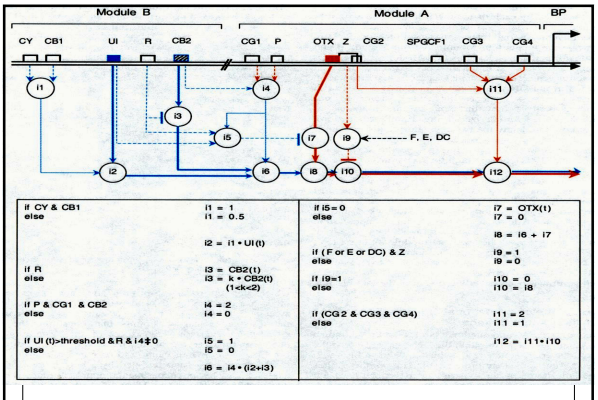
Davidson Experiments

Eric Davidson and colleagues (ED et al.) at Caltech have investigated gene regulation during development of Sea Urchin for the past 30 years

Their work culminated with the discovery of the exact quantitative and logical relationships among DNA elements that guide the expression of a gene (*endo16*)



sea urchin



The DNA program that regulates the expression of *endo16* in sea urchin

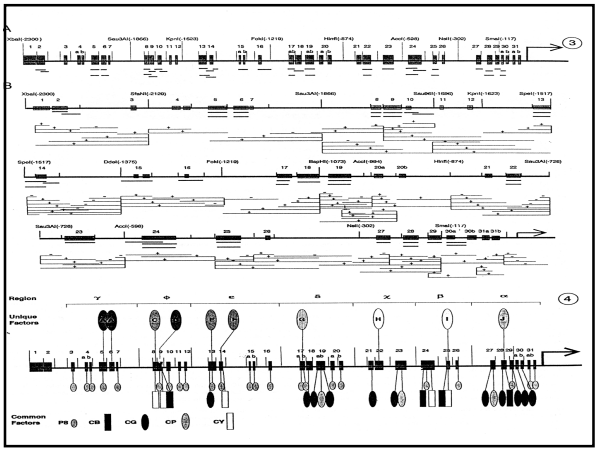
How Did They Do It?

It all started with choosing the right organism / genes and using the right technology:

- The organism of choice - *sea urchin* because of its simplicity (the embryos are see-through and have ~1000 cells)
- The gene of choice - *endo16* because it codes for a protein that has a visible phenotypic effect

The Search for k, the number of inputs

- DNA sequencing
- DNA construct assembly by gene fusion / DNA enzyme restriction / restriction fragment cloning
- Injection of exogenous DNA into and extraction of nuclear material from embryos
- DNA-protein interaction analysis through gel shift / oligonucleotide competition / affinity chromatography

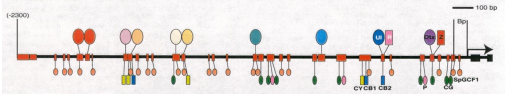


Step-by-Step Identification of Binding Sites

- Nuclear extracts were obtained @24h of development, when the gene is expressed fully
- Nested sets of probes were built out of given sub-fragments of the cis-region by successive restriction enzyme digestion
- The probes were exposed to the 24h nuclear extract
- DNA/protein complexes formed in these reactions were displayed using gel shift

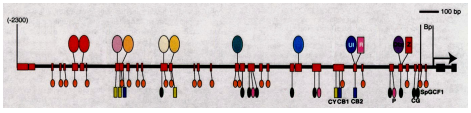
The *Endo16* Cis-region

These techniques were used to identify the complete cis-region, binding sites, and protein - DNA interactions of the region upstream of the *endo16 gene*



Endo16 cis-region: 2300 bp, ~40 binding sites, 16+ different trans. factors, ~60 protein-DNA interactions

The *Endo16* Cis-region



- Some proteins (under the DNA line above), like the looping protein SpCGF1, bind multiple sites all over the cis-region
- Other proteins (over the DNA line above) bind only in specific regions on the DNA

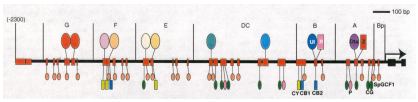
MODULARITY

"An experimental definition of a cis- regulatory **module** is a fragment of DNA containing multiple transcription factor target sites, which when tested in a gene transfer protocol produces some particular subelement of the overall pattern of expression of the gene."

Eric Davidson

Modularity of the *Endo16* Cis-region

- Upon closer examination, it was apparent that the binding sites on the cis-region were "somewhat" clustered
- Thus, ED et al. divided the cis-region into 7 different modules of clustered sites



Once the "players" in the cis-region have been identified, ED et al. went on to uncover their interplay

They asked: ***how do the parts of the cis-region fit in the whole picture?***

To answer this they had to break down the cis-region into smaller components and analyze their individual functions

The Technology: DNA-Expression Constructs

To measure the cis-region fragments' activity they developed the following techniques:

- tagging the fragments with a reporter gene (DNA constructs)
- injecting the constructs in the embryos
- observing the concentration of the reporter protein

DNA Constructs

DNA constructs were created by fusing a reporter gene to fragments of the gene's upstream DNA region (the proximal part) containing the basal promoter fragment

The DNA constructs were injected in the embryos and 75% of them successfully replicated clonally together with the host's DNA

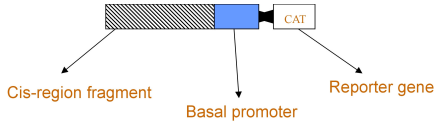
The CAT Reporter Protein

The reporter protein used was the CAT protein because:

- it is readily detectable
- it has a short half life (compared to the experiments' time-line), and
- its concentration is proportional to its coding gene's mRNA concentration

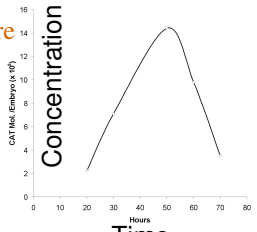
Expression Constructs

Each expression construct is a DNA sequence with three parts: cis-region fragment, basal promoter fragment, and a reporter gene.



Measuring Expression: CT graphs

- The CAT enzyme concentration is a measure of the activity of the cis-region fragment
- For each construct, CAT concentrations were observed @ 20, 30, 50, 60, and 70h in the embryos' development



CT-graph

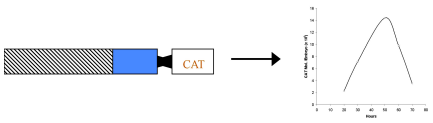
Experimental Framework

ED et al. performed numerous experiments as follows: in each experiment

- an expression construct representing a fragment of the cis-region were prepared,
- copies of it were injected in the embryos, and
- the resulting CT graphs (i.e. CAT concentration @ 20h, 30h, 50h, 60h, and 70h) were observed

Framework, contd.

Conceptually, each experiment assigns an a CT graph to an expression construct (if there is any transcriptional activity at all)



Syntax object

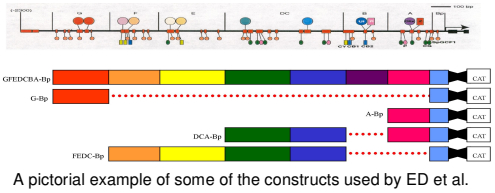
Semantics object

Tinkering

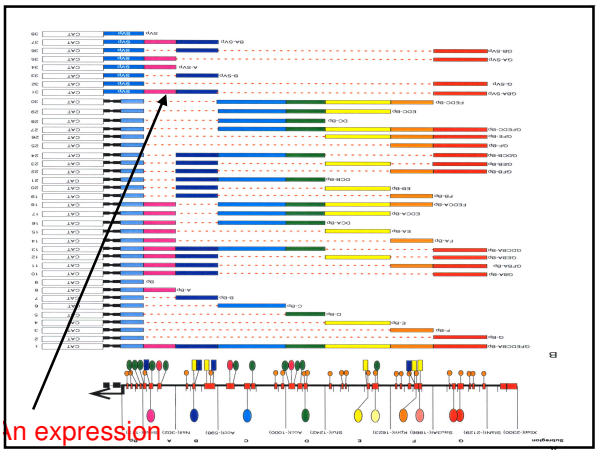
- Faced with the whole cis-region, the experimenters started tinkering by removing pieces, making expression constructs, and observing the CT graphs
- A natural way to break the cis-region was down the lines of the pre-identified modules
- A natural way of making constructs was to remove single or groups of modules

Tinkering, contd.

- ~40 constructs were made originally
- But there's a total of 2^7=128 possible constructs over 7 modules



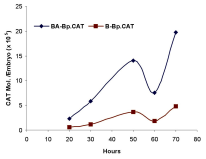
A pictorial example of some of the constructs used by ED et al.

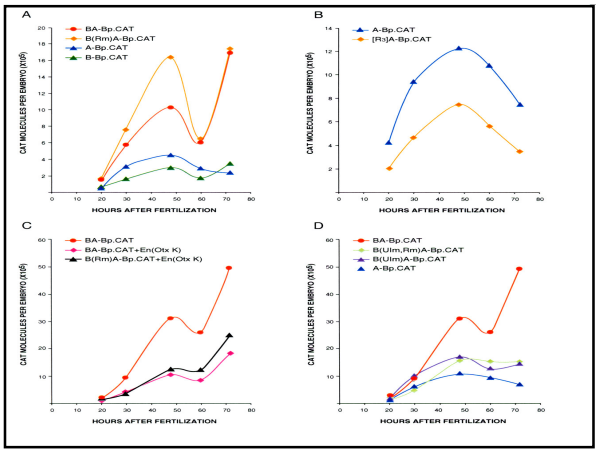


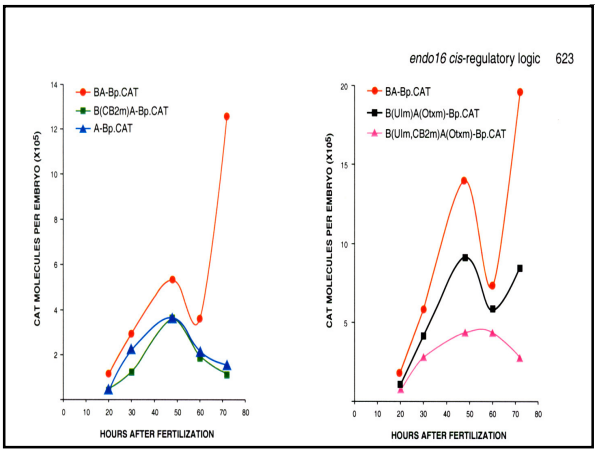
Differential Expression Theory:

It was noticed that only some constructs expressed CAT, and of those that did some had very similar CT graphs, when aligned:

- peaks at 50h, 70h, or both
- some curves looked parallel to each other
- the expression constructs of the similar curves had common sub-constructs







The conclusion drawn from the curve similarity was that the overall cis-region transcription can be decomposed into activities of its parts:

"The overall function of the *Endo16 cis*-regulatory system is the sum of the functions of the individual modules and of the specific interactions among them" (D4)

Refining the Experiments

- The tinkering continued on a finer scale: they added another dimension-mutation of individual binding sites
- A mutation was effectively an elimination of a binding site
- The resulting CT graphs, again, had similar characteristics
- Note: a total of 2^{40}~1000 billion experiments are necessary to cover the whole input

Summary of Results

endo16:

- Only some constructs result in transcription
- Simple relationships between CT graphs observed (similar absolute behavior, but for a constant multiplier)
- A few of the single binding site constructs induce transcription; they are called *kinetic drivers*
- Groups of binding sites act together to permit/prevent transcription downstream

Functional Calculus

We will introduce the following notation to describe the D-Inference:

- Let x and y be groups of contiguous binding sites from the cis-region, that have not been eliminated in the experiment
- Let xy be their union, and let F(z) be the CT graph of the construct z, where z is x or y

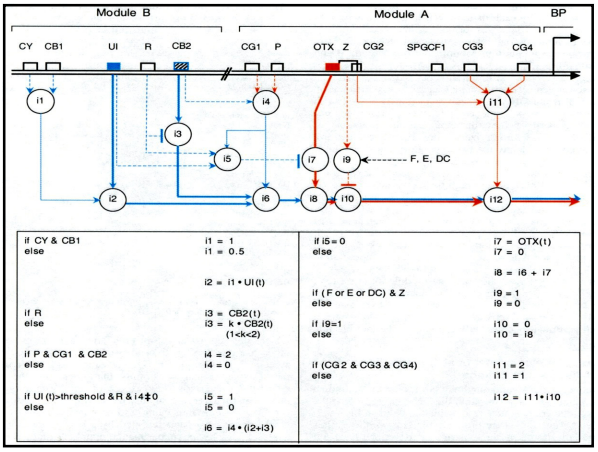
D-Inference Laws

- To relate constructs with sub-constructs through their CT graphs, ED et al. used a simple least squares modeling scheme (one free parameter):

  F(xy)= Lambda * G(F(x),F(y))

  where G(a,b) could be a, b, a+b, or a*b

- Out of the finite number of models on the right, the one that had "the best" fit (smallest rms. Error of model to reality) was chosen as "the model"

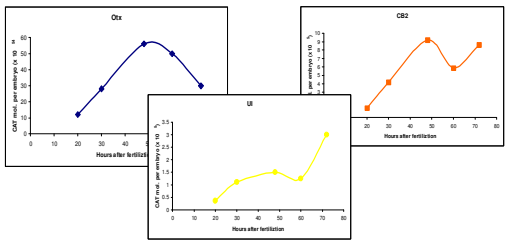


Vector Space of Kinetic Driver Dimension 3

The 3 kinetic drivers: F(UI), F(CB2), and F(OTX) for a Basis in the Functional space. Every other CT graph is a (restricted) linear combination of the Drivers

Kinetic Drivers

Single site interactions sufficient to initiate transcriptions.



There are three drivers in Endo16: Otx, UI and CB2.

Linear/Boolean Inferential Model

- The resulting transcription of the *endo16* cis-region is a linear combination of the CT graphs of the 3 kinetic drivers: F(UI), F(CB2), and F(OTX)
- This model predicts the exact transcription rate for any cis-trans interference of the cis-region

Transcription is a Linear/Boolean Combination of the Driver Signals

Another way to write their program is in a functional form:

$$R(t) = c_1 c_4 c_5 c_6 * F(UI) + c_2 c_4 c_5 c_6 * F(CB2) + c_3 c_5 c_6 * F(OTX), \text{ where}$$

$$c_1 = \begin{cases} 1, \text{ if } CY \wedge CB1 \\ \frac{1}{2}, \text{ otherwise} \end{cases}$$

$$c_2 = \begin{cases} 1, \text{ if } R \\ 1.5, \text{ otherwise} \end{cases}$$

$$c_3 = \begin{cases} 0, \text{ if } UI \wedge R \wedge P \wedge CG1 \wedge CB2 \\ 1, \text{ otherwise} \end{cases}$$

$$c_4 = \begin{cases} 2, \text{ if } P \wedge CG1 \wedge CB2 \\ 0, \text{ otherwise} \end{cases}$$

$$c_5 = \begin{cases} 0, \text{ if } (F \vee E \vee DC) \wedge Z \\ 1, \text{ otherwise} \end{cases}$$

$$c_6 = \begin{cases} 2, \text{ if } CG2 \wedge CG3 \wedge CG4 \\ 1, \text{ otherwise} \end{cases}$$

Where "if (X)" is true if binding site X is present and filled

D-Network of a Single Gene

The cis-region is an *information processing logic*, with inputs the states of the binding sites, and output a functional relationship of the driver signals

The processing elements, nodes or gates, are groups of binding sites which have two states: active and inactive, in each state exhibiting a different effect on the driver signals (factor multiplication)

The nodes of the network can be of different arrity

Inferring a Single Gene D-Network

Inferring a D-Network from a cis-region means finding the kinetic drivers and all the nodes

- If there are no constraints on the nodes we may need 2^k experiments, where k is # of binding sites
- But as ED et al. showed, the cis-region program is a function of its parts, and the parts are modular
- This top-down hierarchy, together with the small number of kinetic drivers, implies that in fact significantly fewer than 2^k experiments may suffice
- A viable assumption: the nodes are contiguous groups of binding sites

Networks the Davidson Way

How does ED extend the model of single gene transcription to gene networks?

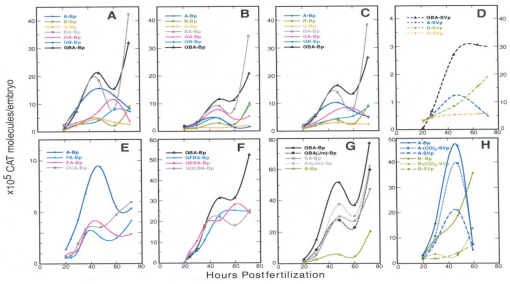
Three different levels of gene networks:

- single gene network (*endo16*)-predicting the transcription rates
- multiple gene network - view from the genome - specificity relationships
- peripheral gene network - view from the organism - phenotypic relationship

Endo 16 “””Inference in Detail: Module A

Constructs <=> CT Graphs

CT-Graphs That Were Used in the Inference



CT-graphs share common elements, like peaks, throughs points, etc. Axiom(intuition): the similarities are due to the modules in common to the corresponding constructs.

Modeling CT-Graph Similarity

- The CT-graphs were modeled as functions of other CT-graphs, i.e.

$$c_0 = \lambda \cdot f(c_1, c_2, \ldots, c_k)$$

"The procedure was to determine the closest possible match between an observed time course […] and a time course calculated by applying a mathematical operation to other observed time courses…" D5

- lambda is a free parameter in the model
- lambda was determined as the minimum least square fit to the model:

$$\lambda = \frac{\sum (c_{0_i}) \bullet f(c_{1_i}, c_{2_i}, \ldots, c_{k_i})}{\sum [f(c_{1_i}, c_{2_i}, \ldots, c_{k_i})]^2}$$

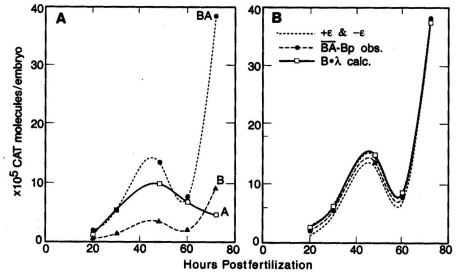
- The root-mean-square error was used to discriminate among the models:

$$\varepsilon = \sqrt{\frac{\sum [c_{0_i} - \lambda \bullet f(c_{1_i}, c_{2_i}, \ldots, c_{k_i})]^2}{N}}$$

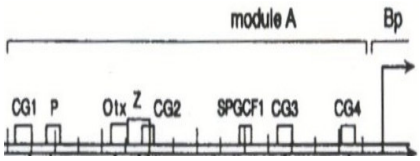
Example Models

Model†	ε (% max)*	λ‡	λ/function#
$\overline{BA}$=B•λ	0.227 (2%)	4.2	4.2
$\overline{BA}$=(B+A)•λ	9.07 (24%)	1.6	1.6
$\overline{BA}$=A•λ	6.49 (17%)	0.69	0.83
$\overline{GBA}$=$\overline{GB}$ •λ	1.99 (8%)	3.1	3.1
$\overline{GBA}$=$\overline{BA}$ •λ	4.35 (17%)	0.78	0.78
$\overline{GBA}$=$\overline{BA}$•G•λ	3.58 (14%)	0.39	0.62
$\overline{GBA}$=A•B•G•λ	4.65 (18%)	0.26	0.64
$\overline{GBA}$=$\overline{GB}$•A•λ	3.97 (15%)	0.50	0.70
$\overline{GBA}$=(G+B+A)•λ	4.40 (17%)	1.23	1.23
$\overline{GBA}$=B•A•λ	3.09 (12%)	0.59	0.77
$\overline{GBA}$=$\overline{GBA}$ (J_m)•λ	7.0 (9%)	1.42	1.42

Modeling $\overline{BA}$ as B*●

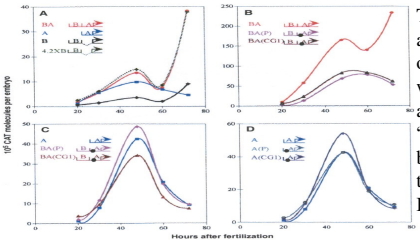


Module A Sites



- The sites in module A. SPGCF1 is a known looping protein, and does not contribute to the resulting expression anyhow.

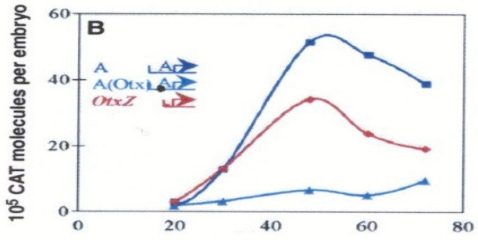
Functions of Sites in A



The role of **P** and **CG1** is that of **a switch**: when present and filled they "conduct" the behavior of B through A to the Bp.

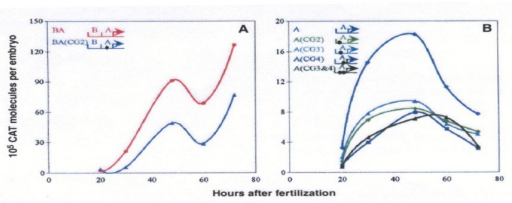
BA behaves as B (and a const. multiplier) when both P and CG1 are on. Eliminating either severs the link between A and B.

Identifying the Driver Site in A



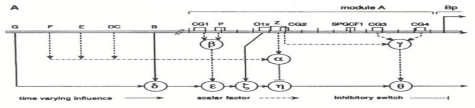
Module A with mutated Otx has no expression

The Rest of the Sites of A



The sites CG2, CG3, and CG4, when ALL present amplify the final expression two-fold. Eliminating any of them is enough to prevent amplification.

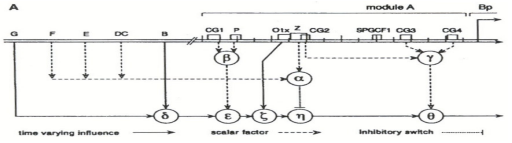
The Program of Module A



B

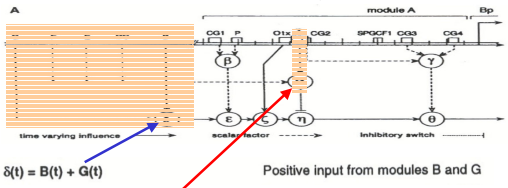
if (F = 1 or E = 1 or CD = 1) and (Z = 1)	Repression functions of modules F, E, and DC mediated by Z site
α = 1	
else α = 0	
if (P = 1 and CG$_1$ = 1)	Both P and CG$_1$ needed for synergistic link with module B
β = 2	
else β = 0	
if (CG$_2$ = 1 and CG$_3$ = 1 and CG$_4$ = 1)	Final step up of system output
γ = 2	
else γ = 1	
δ(t) = B(t) + G(t)	Positive input from modules B and G
ε(t) = β*δ(t)	Synergistic amplification of module B output by CG$_1$-P subsystem
if (ε(t) = 0)	Switch determining whether Otx site in module A, or upstream modules (i.e., mainly module B), will control level of activity
ξ(t) = Otx(t)	
else ξ(t) = ε(t)	
if (α = 1)	Repression function inoperative in endoderm but blocks activity elsewhere
η(t) = 0	
else η(t) = ξ(t)	
Θ(t) = γ*η(t)	Final output communicated to BTA

How Should We Interpret the Program



- Direction of execution: From left to right, following the arrows of the diagram. The transcription rate at any point is given by the greek letters at the very bottom of the diagram
- At every node we perform a logical decision or a factor multiplication
- The logical decisions are Boolean functions of individual site variables, which are 1 iff site is present and full, and 0 otherwise

From Left to Right...



$\delta(t) = B(t) + G(t)$ Positive input from modules B and G

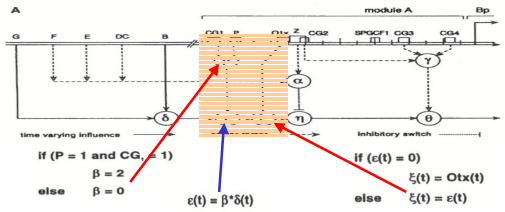
G and B are not analyzed in detail. Each initiates transcription.

if (F = 1 or E = 1 or CD = 1) and (Z = 1) Repression functions of modules F, E, and DC mediated by Z site

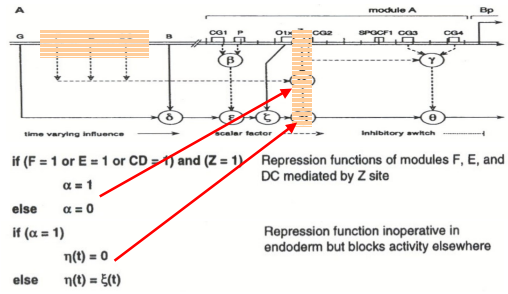
$\alpha = 1$

else $\alpha = 0$

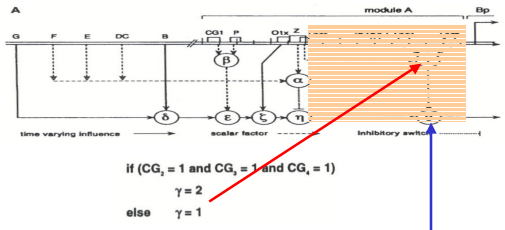
Any of CD, E, and F can repress expression, but only if Z is present and filled



P and CG1 are **a switch** that flips between Module A activity (early development) and Module B activity (late development), and also contribute a factor 2 amplification.When any one is not present, there is no transcription contribution from any module to the left of them



CD, E, and F repress transcription totally



The final transcription is amplified 2-fold if all three CG2, CG3, and CG4 are present and filled

$\Theta(t) = \gamma^*\eta(t)$

Summary of Program

If (Z & (CD or E or F)) → **REPRESS**

Trans(t) := 0

else

If (P & CG1) → **SWITCH / AMPLIFY**

Trans(t) := 2*(B(t)+G(t))

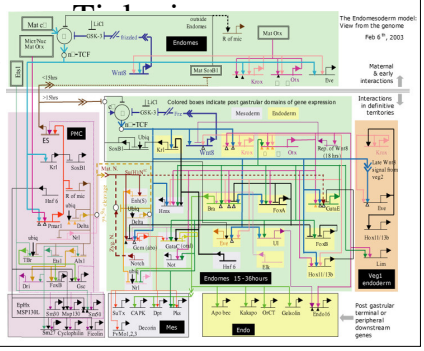
else Trans(t) := Otx(t)

If (CG2 & CG3 & CG4) → **AMPLIFY**

Trans(t) := 2*Trans(t)

4. Putting It All Together:

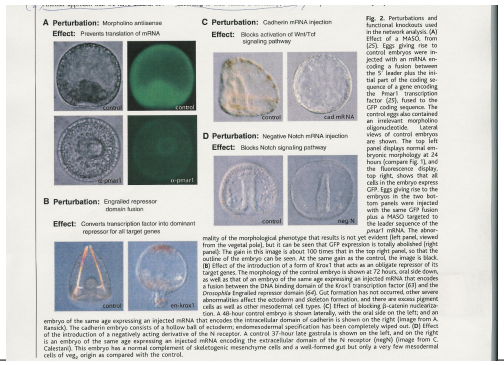
Davidson et al., 2002



Methods

- ~40 genes and trans factors
- Perturbation analysis
  - – types of perturbation:
    - Trans factor knockout
    - mRNA introduction
    - gene overexpression
    - Direct link rescue experiments
- Other data: spatial and temporal expression

Types of Perturbations Used



Goals

- Sequence-based gene network
- Uncover positive and negative regulatory relationships among the genes
- Group genes in gene batteries
- Identify domains of regulation and genes involved in corresponding development

Inference Procedure

0) Start with a small number of known regulatory genes and their regulatory relationships

2) Perturb regulatory expressions

2) Observe changes

3) Postulate relationships based on changes

4) Handle indirect influences