Davidson Causality

The Unreasonable Effectiveness of Mathematics in the Regulatory Genome

Sorin Istrail

Department of Computer Science
Brown University

May 20, 2021

Overview

- The Causal Gene Regulatory Networks (GRNs)
- The Glorious Sea Urchin
- The Genome
- The Transciptome
- The Logicome
- The Regulatory Genome and the Computer

- Davidson Causality: An Axiomatic Approach

- The Peter-Davidson Boolean Model

Seven papers and an essay

- Sorin Istrail, “Eric Davidson’s Regulatory Genome for Computer Science: Causality, Logic and Proof Principles of the Genomic cis-Regulatory Code”, Journal of Computational Biology, Vol 26, No. 17, pp. 635-684, 2019
- Sorin Istrail and Isabelle Peter, “How Does the Regulatory Genome Work?”, Journal of Computational Biology, Vol 26, No. 7, pp. 685-695, 2019
- Sorin Istrail “Eric Davidson: Master of the Universe” *Developmental Biology*, Apr 15;412(s suppl) S47-54, 2016
- Jongmin Nam, Ping Dong, Ryan Tarpine, Sorin Istrail, Eric H. Davidson, "Functional cis-regulatory genomics for systems biology", *In Proceedings of the National Academy of Sciences*, vol. 107, no. 8, pp. 3930-3935, 2010.
- Sorin Istrail, Smadar Ben-Tabou De-Leon, Eric H. Davidson, “The regulatory genome and the computer“ *Developmental Biology*, vol. 310, no. 2, pp. 187-195, 2007 (Homage to 50th anniversary of John von Neumann’s “The Computer and the Brain” book)
- Manoj P. Samanta, Waraporn Tongprasit, Sorin Istrail, R. Andrew Cameron, Qiang Tu, Eric H. Davidson, Viktor Stolc, “The Transcriptome of the Sea Urchin Embryo”, *Science*, vol. 314, no. 5801, pp. 960-962, 2006.
- Kim Worley, K. James Durbin, Yufeng Shen, Olivier Fedrigo, David Garfield, Ralph Haygood, Alexander Primus, Rahul Satija, Tonya Severson, Manuel L. Gonzalez-Garay, Andrew R. Jackson, Aleksandar Milosavljevic, Mark Tong, Christopher E. Killian, Brian T. Livingston, Fred H. Wilt, Nikki Adams, Robert Bell??, Seth Carbonneau, Rocky Cheung, Patrick Cormier, Bertrand Cosson, Jenifer Croce, Antonio Fernandez-Guerra, Anne-Marie Genevi??re, Manisha Goel, Hemant Kelkar, Julia Morales, Odile Mulner-Lorillon, Anthony J. Robertson, Jared V. Goldstone, Bryan Cole, David Epel, Bert Gold, Mark E. Hahn, Meredith Howard-Ashby, Mark Scally, John J. Stegeman, Erin L. Allgood, Jonah Cool, Kyle M. Judkins, Shawn S. McCafferty, Ashlan M. Musante, Robert A. Obar, Amanda P. Rawson, Blair J. Rossetti, Ian R. Gibbons, Matthew P. Hoffman, Andrew Leone, Sorin Istrail, Stefan C. Materna, Manoj P. Samanta, Viktor et al. Stolc, "The Genome of the Sea Urchin Strongylocentrotus purpuratus", *In Science*, vol. 314, no. 5801, pp. 941-952, 2006.
- Sorin Istrail, Eric Davidson, “Logic functions of the genomic cis-regulatory code”, *Proceedings of the National Academy of Sciences*, vol. 102, no. 14, pp. 4954-4959, 2005

Davidson's First Completeness Problem

***Q1*. What is the size of the complete cis-regulatory universe of all the published papers that present the cis-regulatory architectures of genes validated by cis-regulatory analysis (Davidson criterion). It was resolved via the "Cloning of Professor Davidson" problem, i.e., to automate through software the way Professor Davidson would read/scan a paper to find out whether it belongs to the cis-regulatory universe.**

- Sorin Istrail, "Eric Davidson's Regulatory Genome for Computer Science: Causality, Logic and Proof Principles of the Genomic cis-Regulatory Code", Journal of Computational Biology, Vol 26, No. 17, pp. 635-684, 2019

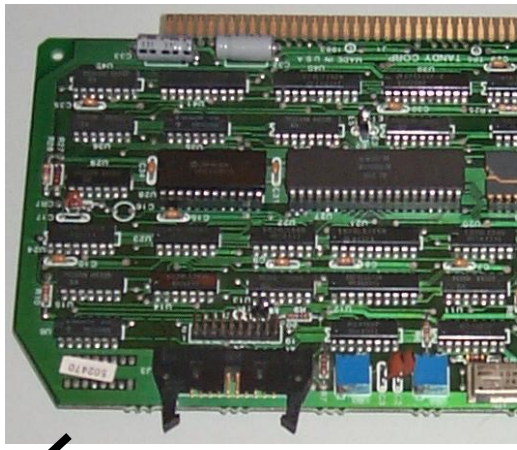
Davidson's Second Completeness Problem

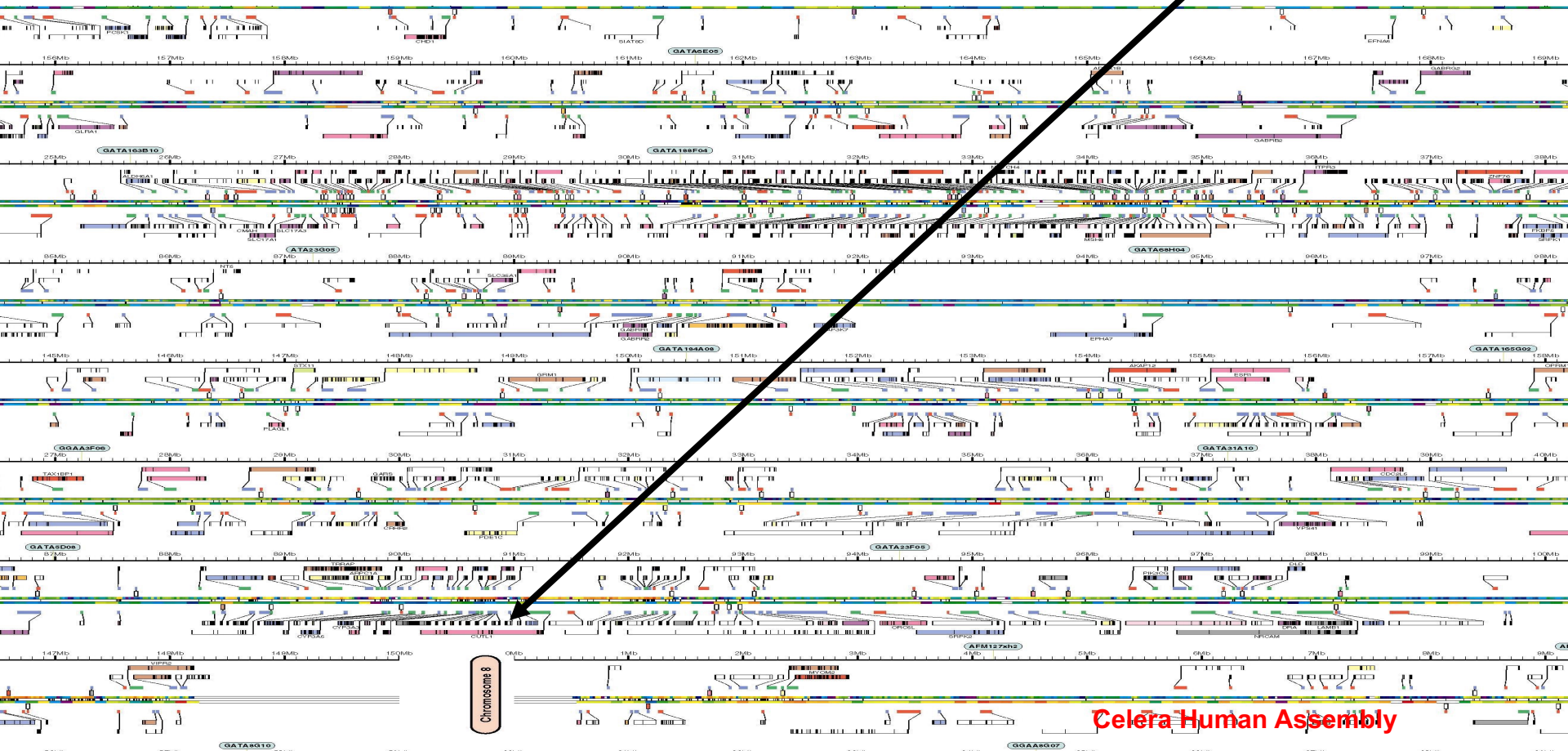
Q2. Can a formal computational model for GRNs be developed that is completely predictive of its experimental output?

The groundbreaking work of Isabelle Peter (Caltech) and Eric Davidson advanced a Boolean computational model called "GRN Network Equations" that is (almost) completely predictive of the experimental output. The extraordinary scale of this achievement applies to the sea urchin embryo GRN and is sufficient to predictably explain almost all the spacial regulatory transactions underlying a large portion of the pre-gastrulae embryonic process. Their 2015 book *Genomic Control Process: Development and Evolution* (Oxford: Academic Press/Elsevier 2015) presents in full detail the pioneering solution of the sea urchin embryo GRN via this (almost) completely predictive computational biology model.

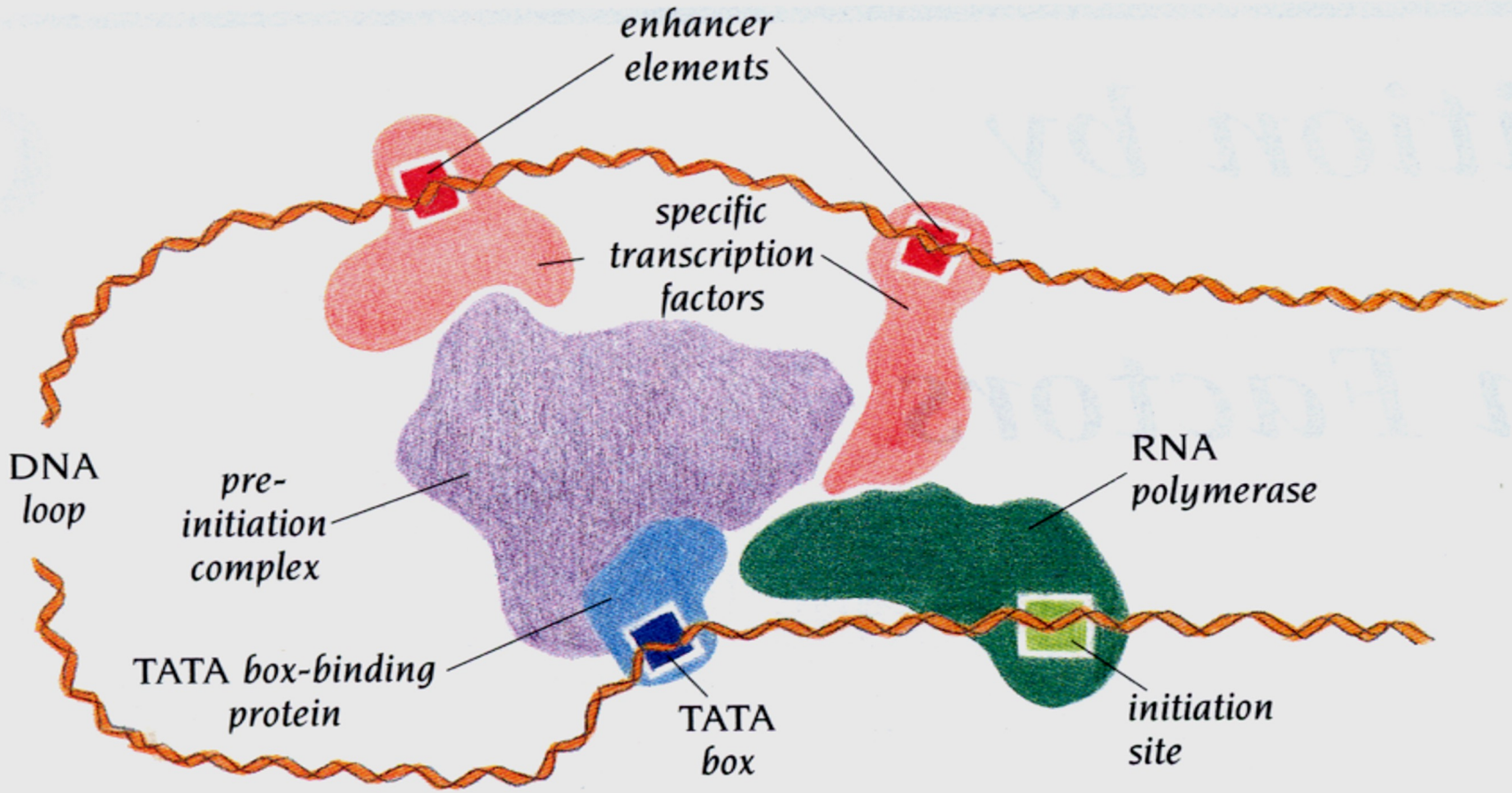
- Sorin Istrail, "Eric Davidson's Regulatory Genome for Computer Science: Causality, Logic and Proof Principles of the Genomic cis-Regulatory Code", Journal of Computational Biology, Vol 26, No. 17, pp. 635-684, 2019

Biology and Computer Science intertwined together into one.





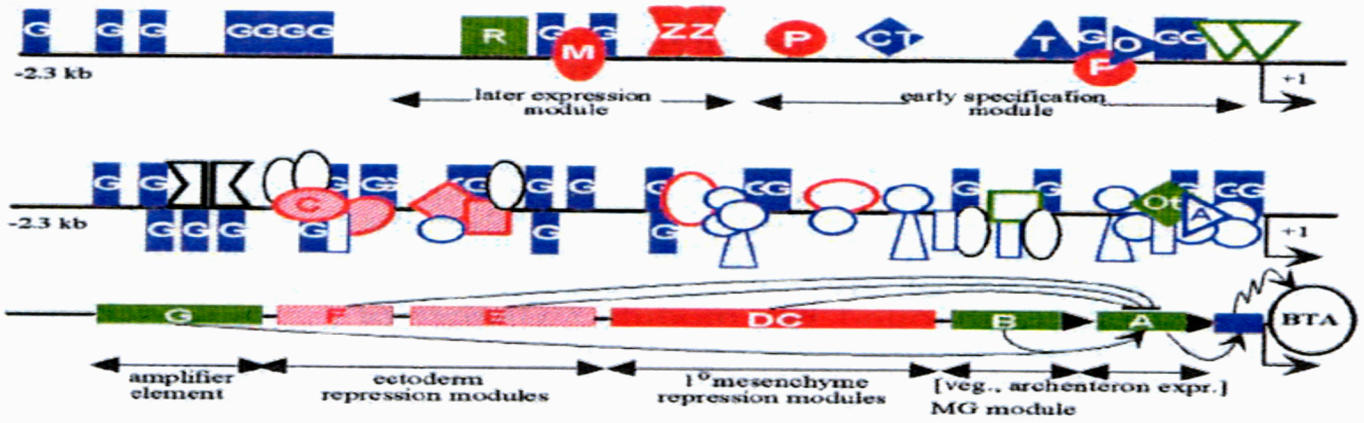
The Dogma


enhancer elements
specific transcription factors
DNA loop
pre-initiation complex
RNA polymerase
TATA box-binding protein
TATA box
initiation site

sea urchin

(A) *CyIIIa* complete system

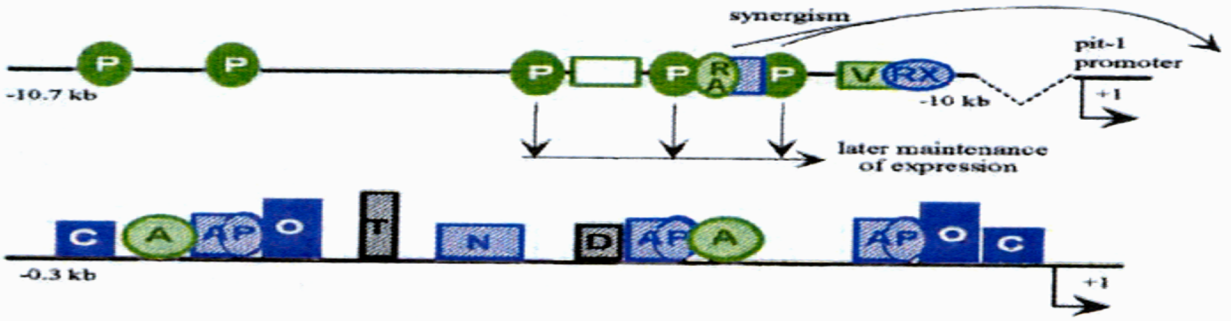
(B) *Endo16* complete embryonic system



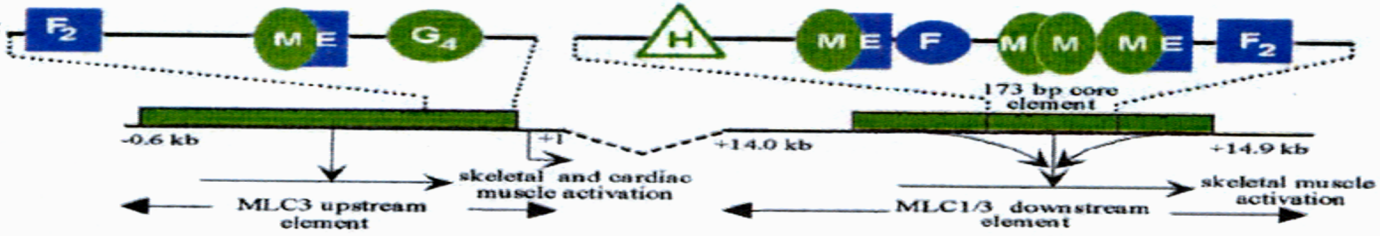
mammals

(C) mouse *pit-1* RA-response/ autoregulation element

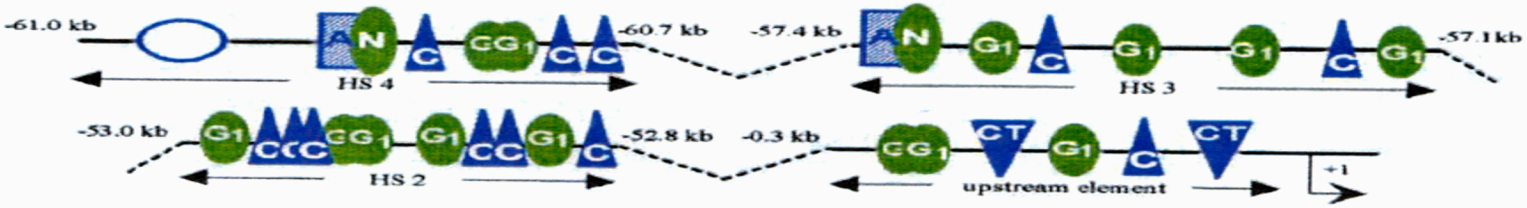
(D) mouse *IL-2* inducible element



(E) rat myosin light chain 1/3 elements



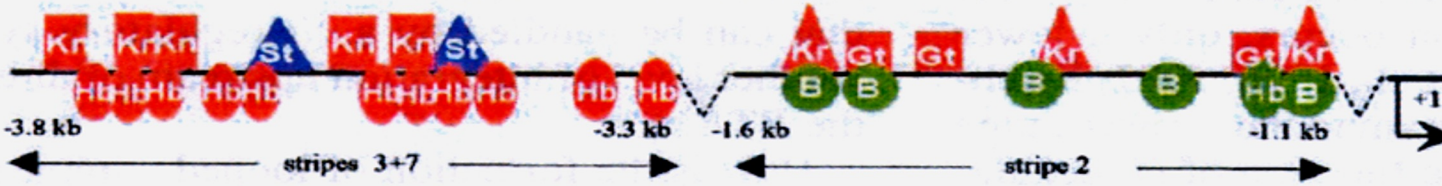
(F) human β-globin locus



Drosophila

(G) *ftz* zebra-stripe element

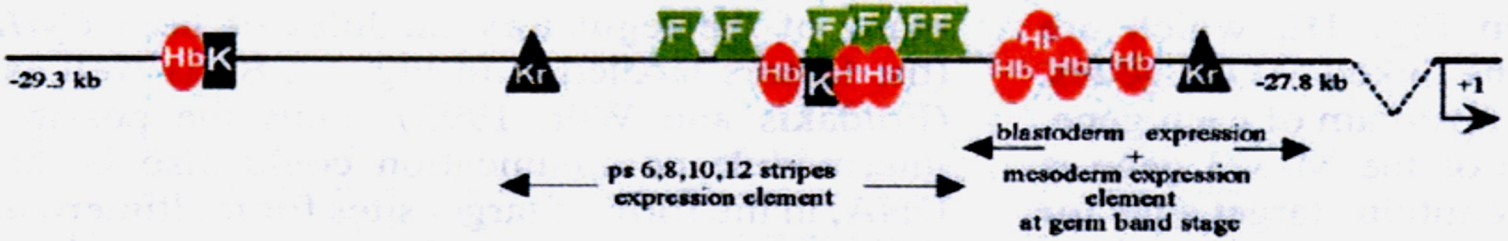
(H) *eve* stripe 3+7 and 2 elements



(I) *rho* lateral neurectoderm stripe element

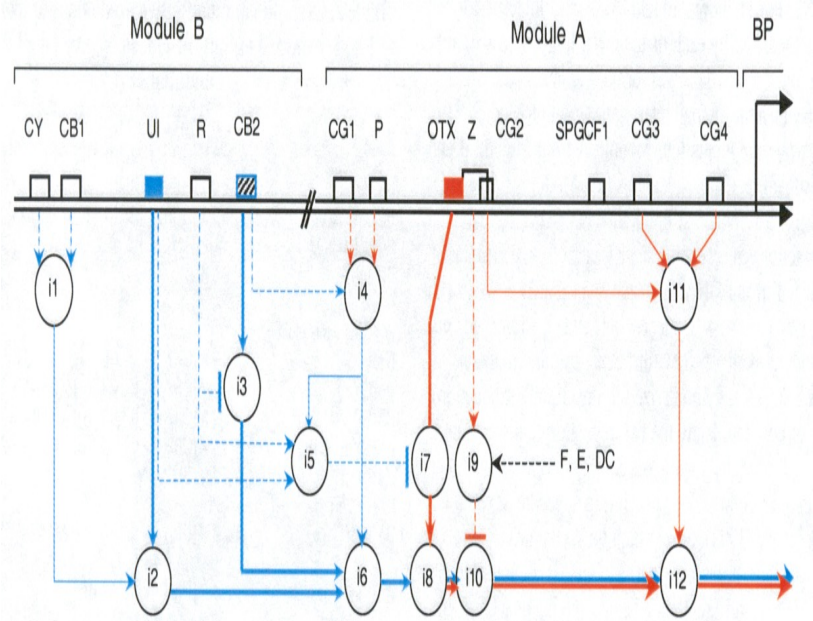
(J) *kni* posterior element

(K) PBX element of ubx gene

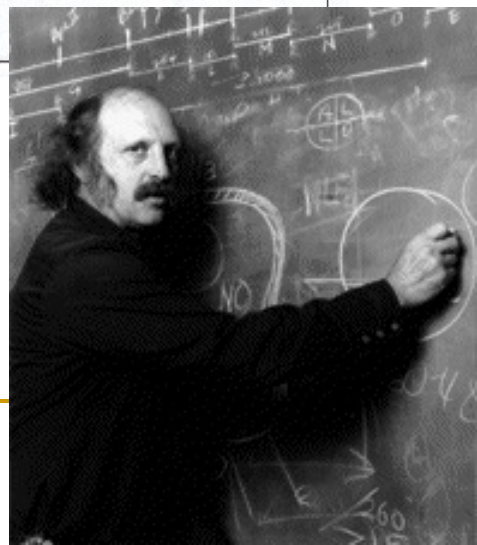


One gene, 30 years of study, 300 docs and postdocs

A Proposal for Nobel Prize



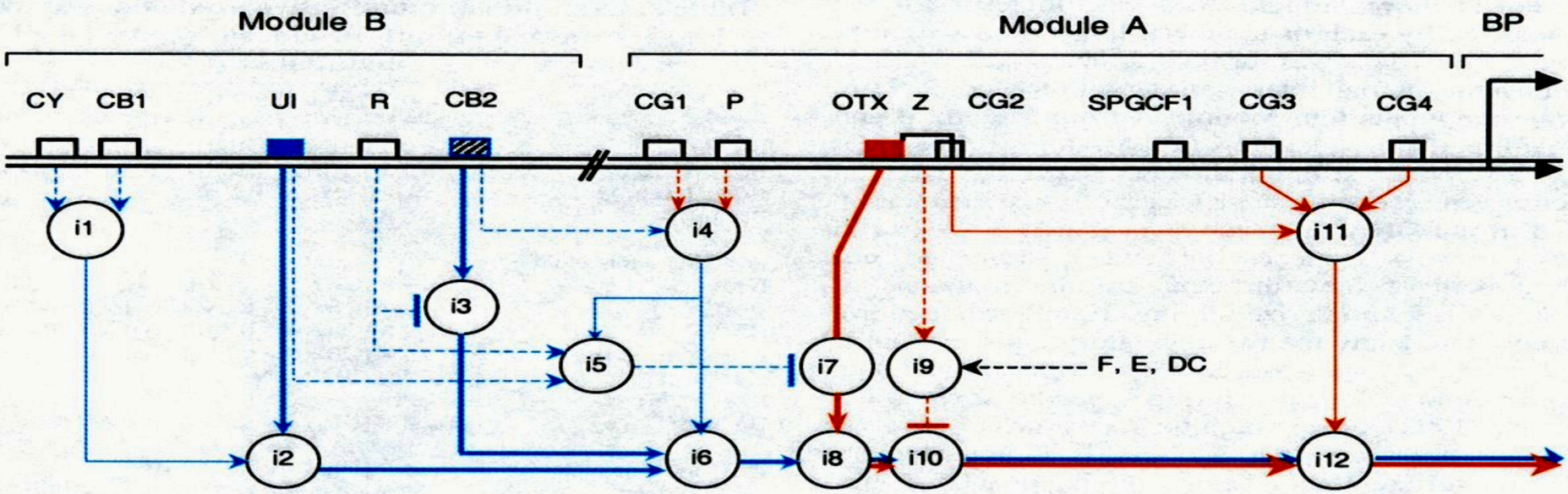
if CY & CB1	i1 = 1	if i5 = 0	i7 = OTX(t)
else	i1 = 0.5	else	i7 = 0
			i8 = i6 + i7
	i2 = i1 • UI(t)		
		if (F or E or DC) & Z	i9 = 1
		else	i9 = 0
if R	i3 = CB2(t)		
else	i3 = k • CB2(t) (1<k<2)	if i9=1	i10 = 0
		else	i10 = i8
if P & CG1 & CB2	i4 = 2		
else	i4 = 0	if (CG2 & CG3 & CG4)	i11 = 2
		else	i11 = 1
if UI(t)>threshold & R & i4≠0	i5 = 1		i12 = i11•i10
else	i5 = 0		
	i6 = i4 • (i2+i3)		



"Programs built into the DNA of every animal."

Eric H. Davidson

Genomic Regulatory Systems

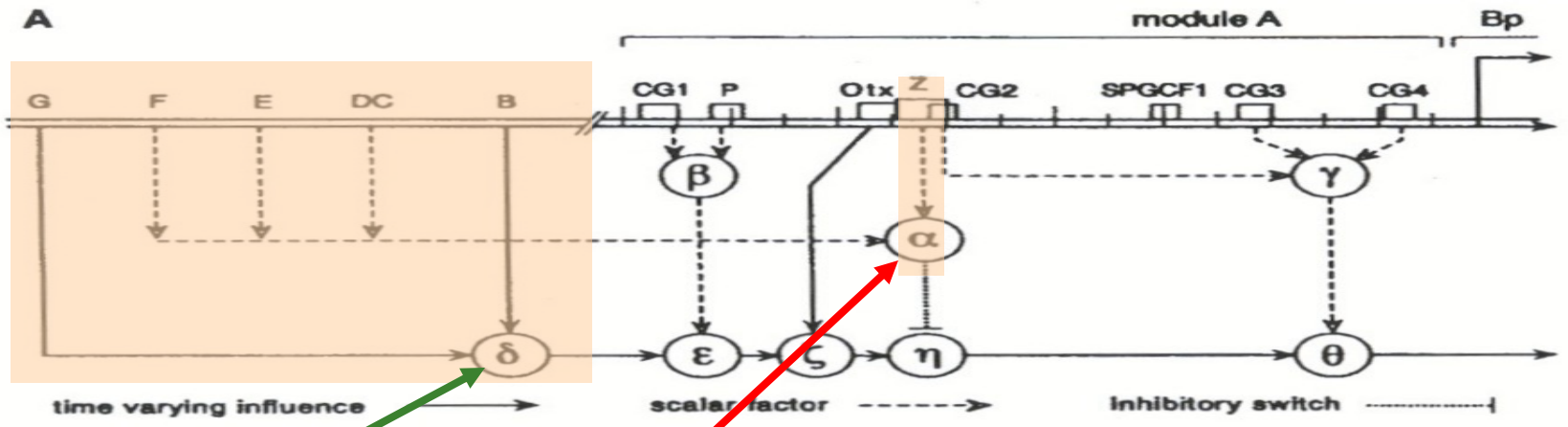


if CY & CB1	i1 = 1
else	i1 = 0.5
	i2 = i1 • UI(t)
if R	i3 = CB2(t)
else	i3 = k • CB2(t) (1<k<2)
if P & CG1 & CB2	i4 = 2
else	i4 = 0
if UI(t)>threshold & R & i4≠0	i5 = 1
else	i5 = 0
	i6 = i4 • (i2+i3)

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

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

From Left to Right...



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

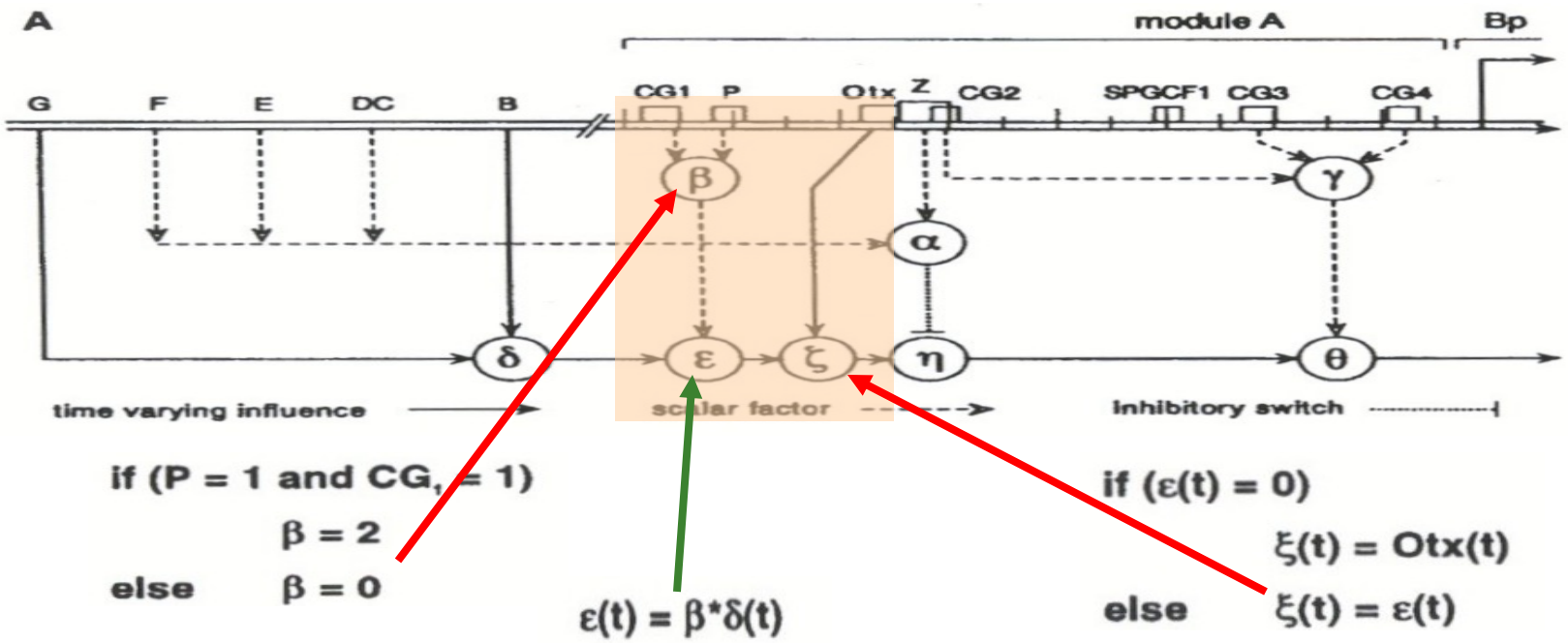
G and B are not analyzed in detail. Each initiates transcription.

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

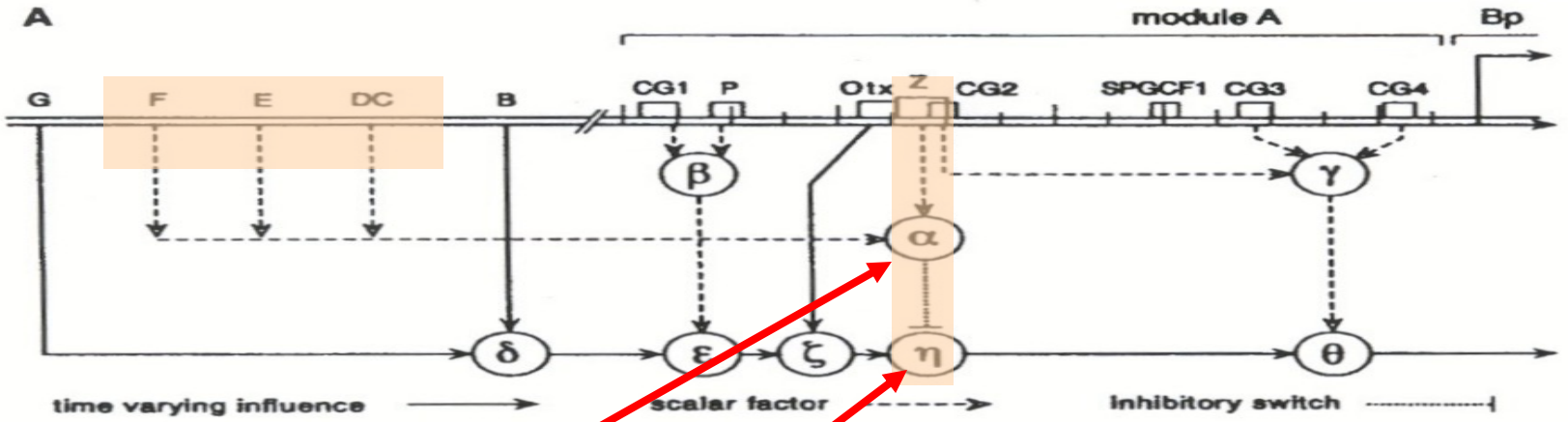
$\alpha = 1$

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



if (F = 1 or E = 1 or CD = 1) and (Z = 1)

$\alpha = 1$

else $\alpha = 0$

Repression functions of modules F, E, and DC mediated by Z site

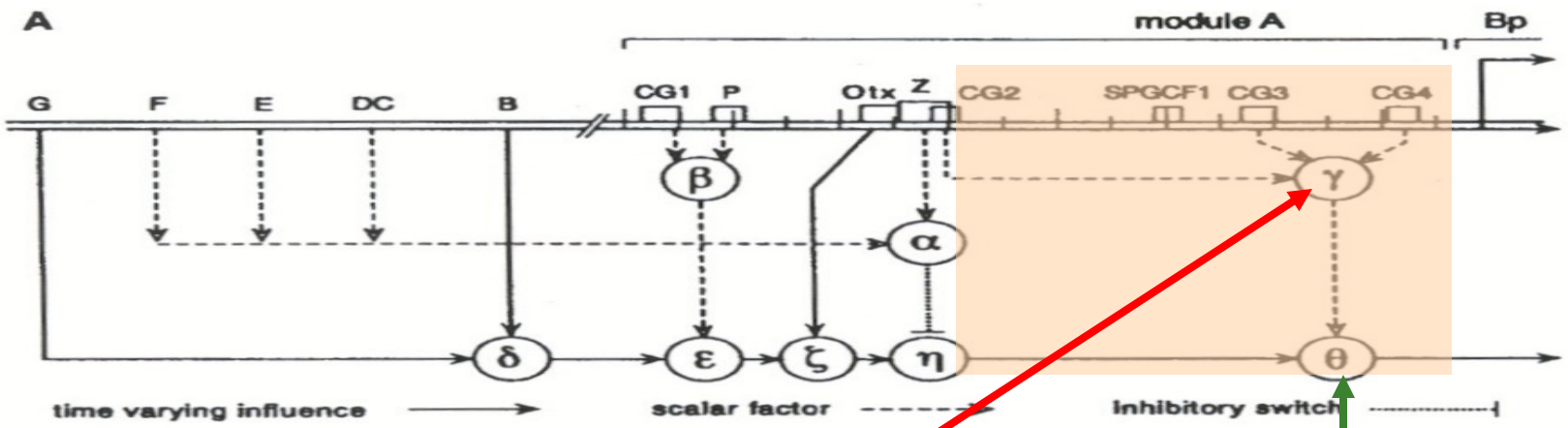
if ($\alpha = 1$)

$\eta(t) = 0$

else $\eta(t) = \xi(t)$

Repression function inoperative in endoderm but blocks activity elsewhere

CD, E, and F repress transcription totally



if ($CG_2 = 1$ and $CG_3 = 1$ and $CG_4 = 1$)

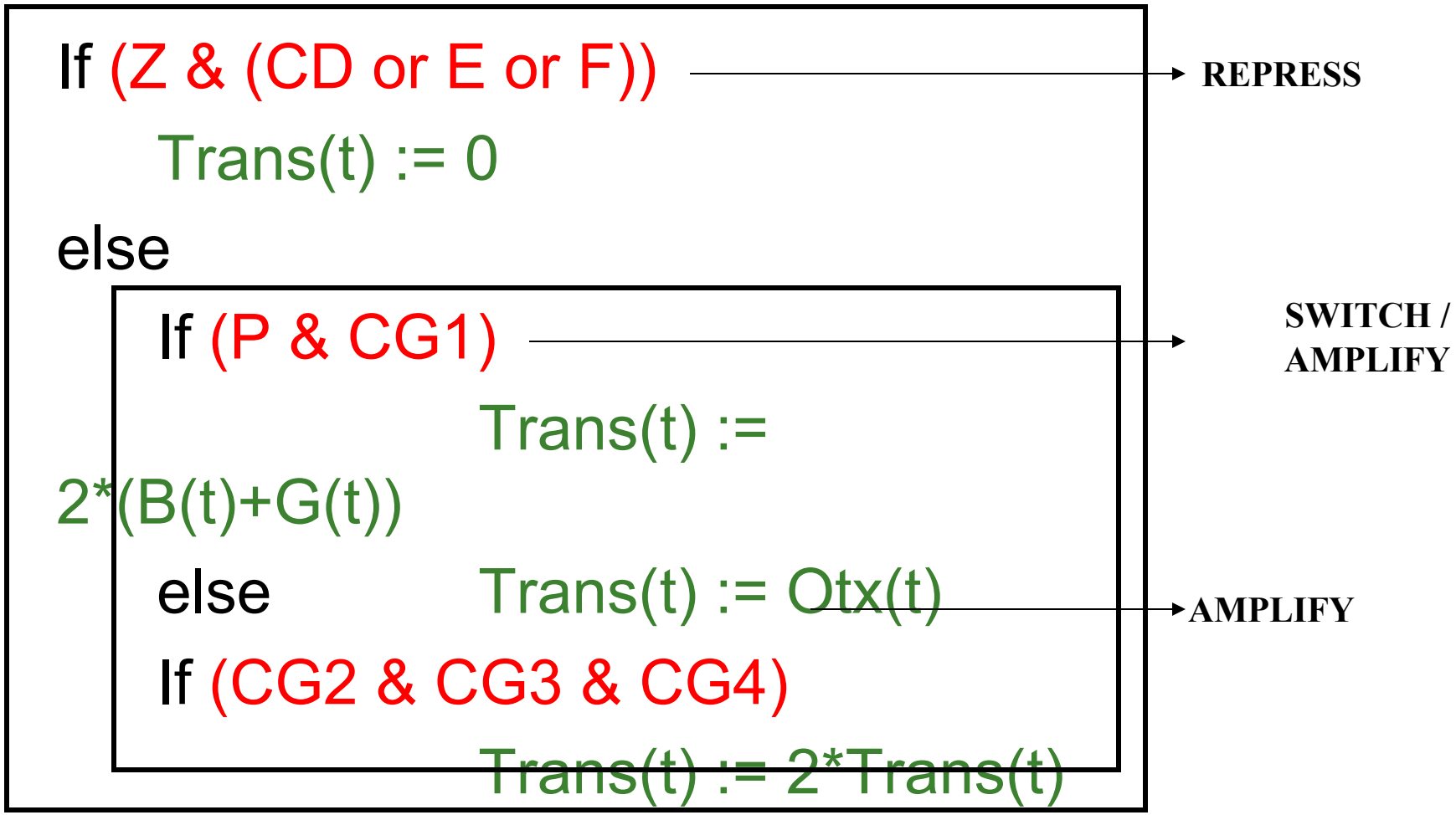
$\gamma = 2$

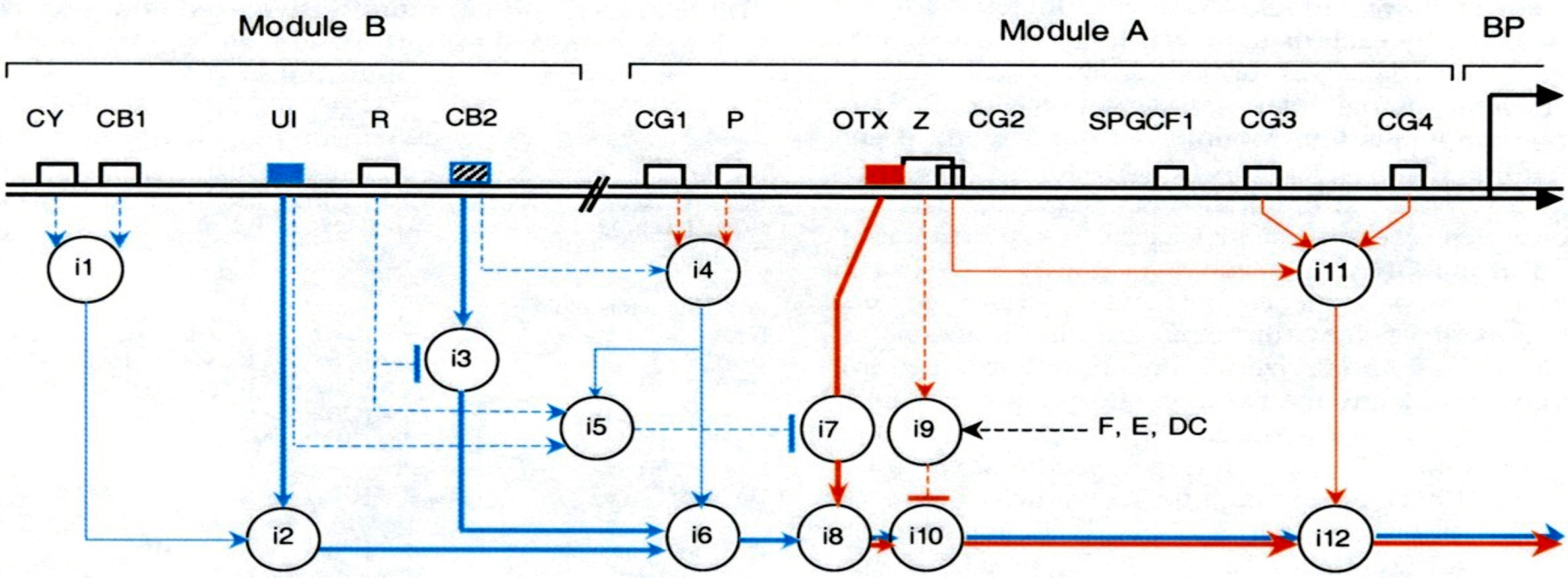
else $\gamma = 1$

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

$$\theta(t) = \gamma^{*}\eta(t)$$

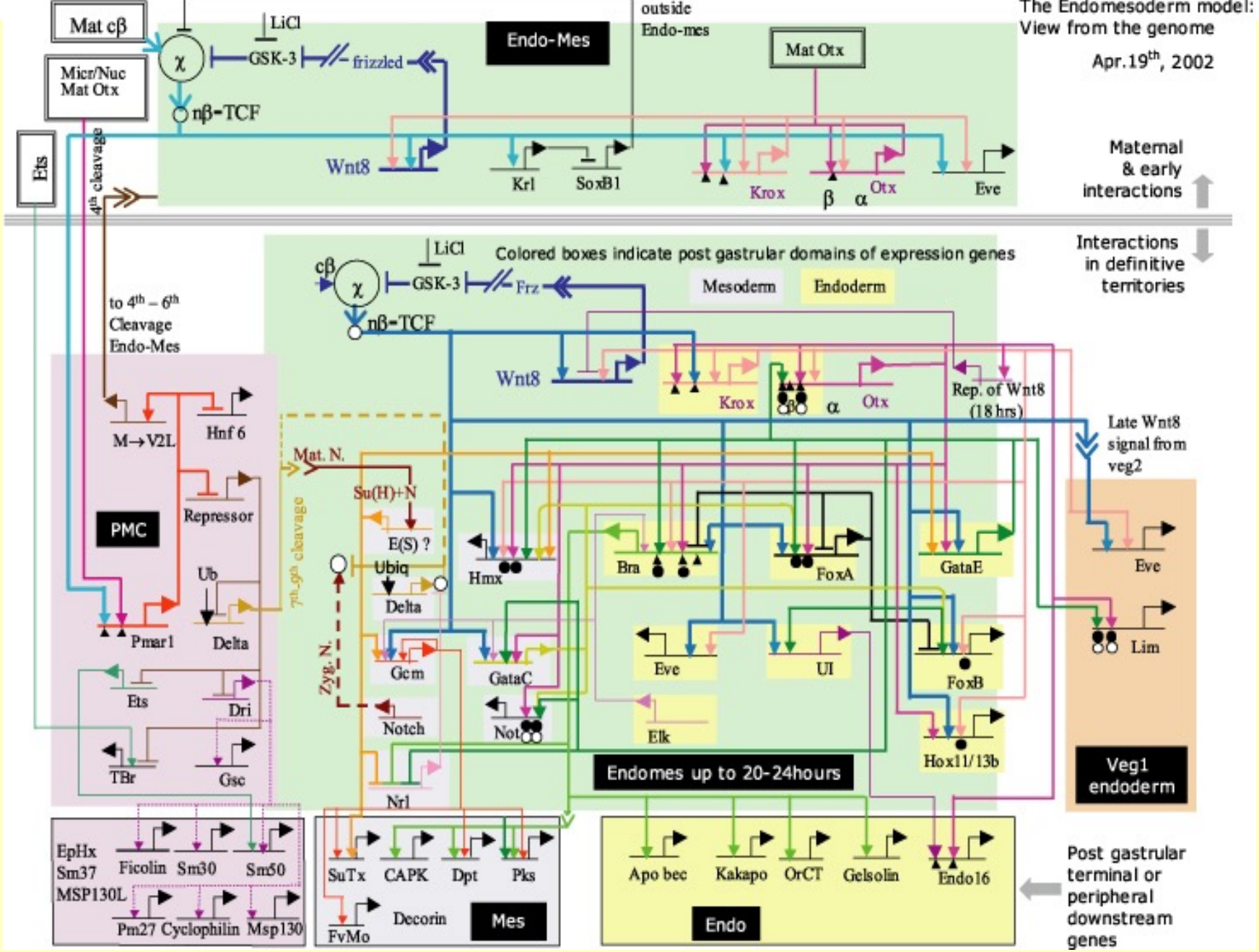
Summary of Program





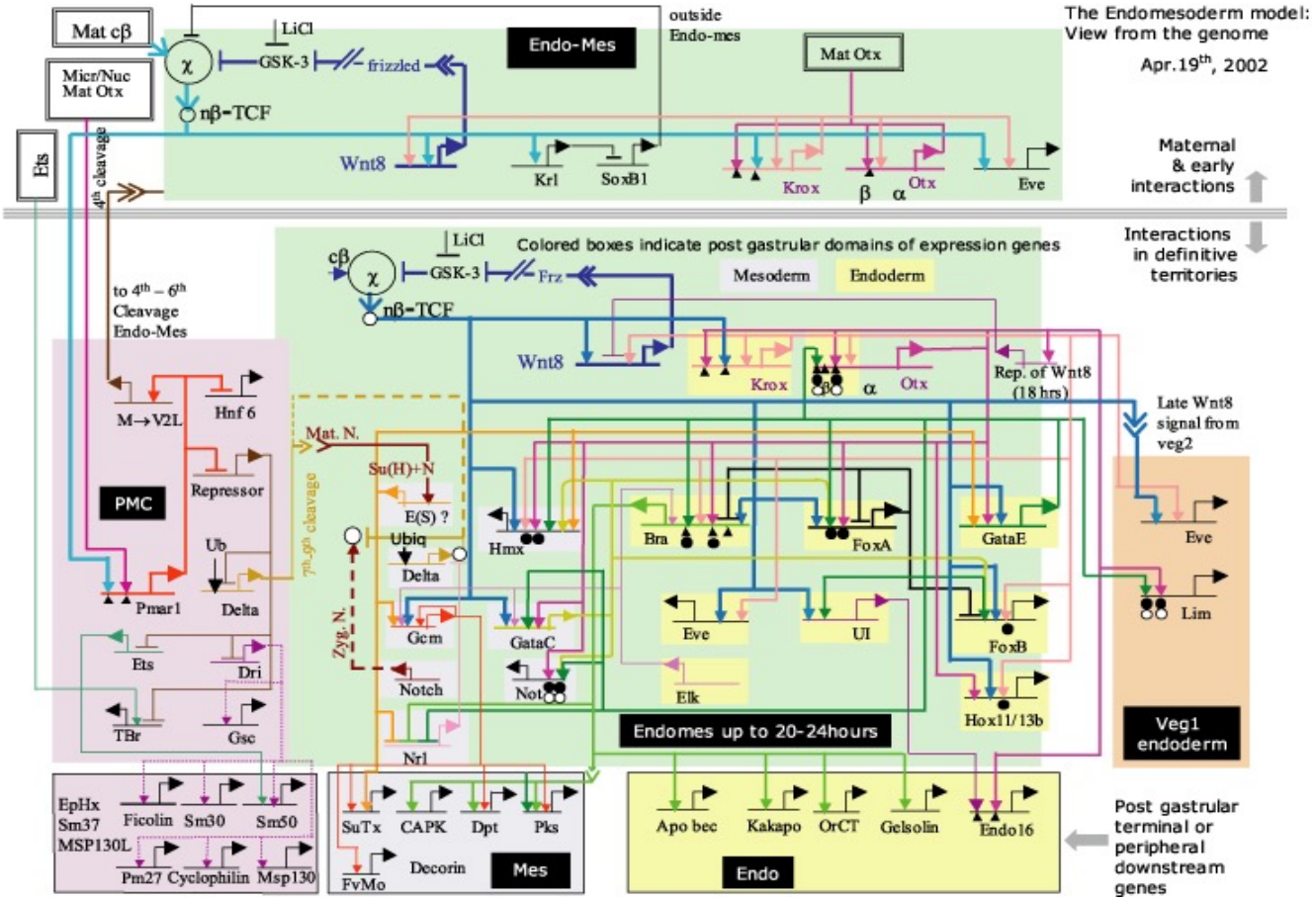
if CY & CB1	i1 = 1
else	i1 = 0.5
	i2 = i1 • UI(t)
if R	i3 = CB2(t)
else	i3 = k • CB2(t) (1<k<2)
if P & CG1 & CB2	i4 = 2
else	i4 = 0
if UI(t)>threshold & R & i4≠0	i5 = 1
else	i5 = 0
	i6 = i4 • (i2+i3)

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



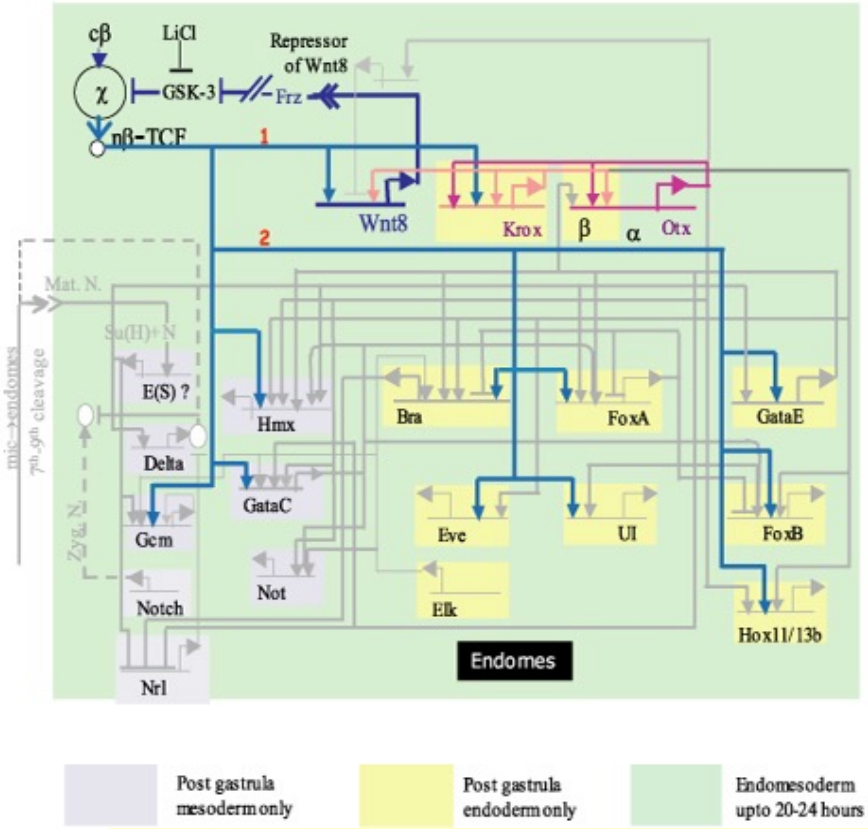
THE FIRST NETWORK

The View from the Genome



The View from the Nucleus

View from the nucleus: Endomesoderm nuclei to hatching blastula stage; the Wnt8/Tcf signalling loop and its genes.
Apr. 19th , 2002

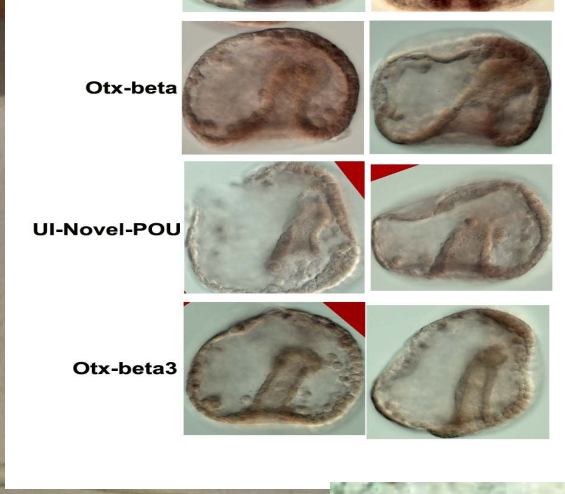


Notes:

1. β-catenin/Tcf input now produced by a zygotic signaling loop driven by Wnt8 expression in endmesoderm cells.
2. β-catenin/Tcf input required for expression of many regulatory genes that become active in the veg_2 endomesodermal territory during early-mid blastula stage.

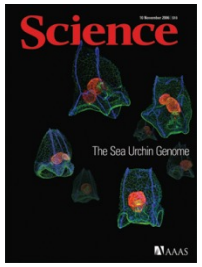
The Glorious Sea Urchin

Future Work



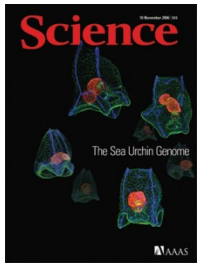
Caltech, Davidson Lab
October 2004

Overview



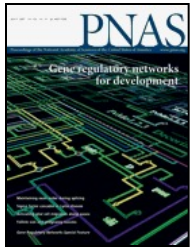
THE GENOME

E. Sodergren, G. Weinstock, E. Davidson, ..., S. Istrail, ...,
The Genome of the Sea Urchin *Strongylocentrotus purpuratus*,
Science, vol 314, pp. 941-952, **2006**



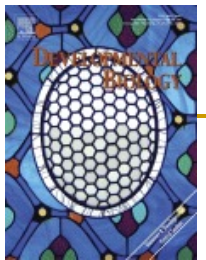
THE TRANSCIPTOME

M. Samanta, W. Tongprasit, S. Istrail, A. Cameron. Q. Tu, E. Davidson, V. Stolc
The transcriptome of the sea urchin embryo,
Science, vol 314, pp. 960-962, **2006**



THE LOGICOME

Sorin Istrail and Eric Davidson,
Logic functions of the genomic cis-regulatory code,
Proceedings of the National Academy of Sciences, vol. 102, No. 14, 4954-4959, **2005**



THE REGULATORY GENOME

S. Istrail, S. Ben-Tabou de-Leon, E.H. Davidson.
The Regulatory Genome and the Computer,
Developmental Biology, vol 310, pp. 187-195, **2007**

THE GENOME

THE GENOME OF THE SEA URCHIN STRONGYLOCENTROTUS PURPURATUS

E. Sodergren, G. Weinstock, E. Davidson, ..., S. Istrail, ..., The Genome of the Sea Urchin Strongylocentrotus purpuratus, Science, vol 314, pp. 941-952, 2006

We report the sequence and analysis of the 814-megabase genome of the sea urchin Strongylocentrotus purpuratus, a model for developmental and systems biology. The sequencing strategy combined whole-genome shotgun and bacterial artificial chromosome (BAC) sequences. This use of BAC clones, aided by a pooling strategy, overcame difficulties associated with high heterozygosity of the genome. The genome encodes about 23,300 genes, including many previously thought to be vertebrate innovations or known only outside the deuterostomes. This echinoderm genome provides an evolutionary outgroup for the chordates and yields insights into the evolution of deuterostomes.

THE TRANSCRIPTOME

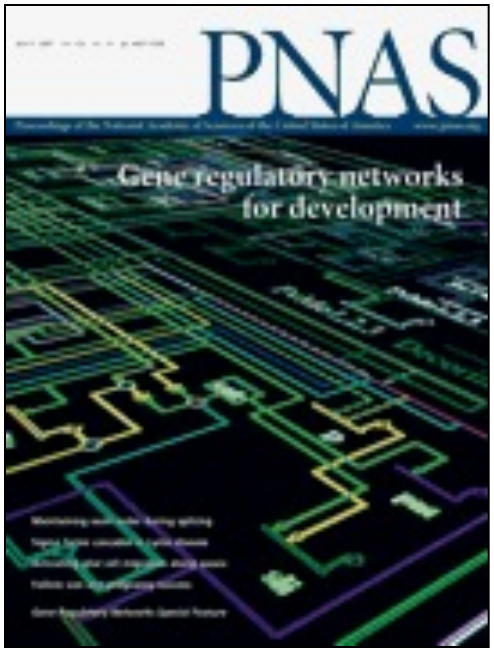
THE TRANSCRIPTOME OF THE SEA URCHIN EMBRYO

M. Samanta, W. Tongprasit, S. Istrail, A. Cameron. Q. Tu, E. Davidson, V. Stolc, The
transcriptome of the sea urchin embryo, Science, vol 314, pp. 960-962, 2006

ABSTRACT: The sea urchin Strongylocentrotus purpuratus is a model organism for study of the genomic control circuitry underlying embryonic development. We examined the complete repertoire of genes expressed in the S. purpuratus embryo, up to late gastrula stage, by means of high-resolution custom tiling arrays covering the whole genome. We detected complete spliced structures even for genes known to be expressed at low levels in only a few cells. At least 11,000 to 12,000 genes are used in embryogenesis. These include most of the genes encoding transcription factors and signaling proteins, as well as some classes of general cytoskeletal and metabolic proteins, but only a minor fraction of genes encoding immune functions and sensory receptors. Thousands of small asymmetric transcripts of unknown function were also detected in intergenic regions throughout the genome. The tiling array data were used to correct and authenticate several thousand gene models during the genome annotation process.

THE LOGICOME

LOGIC FUNCTIONS OF THE GENOMIC REGULATORY CODE



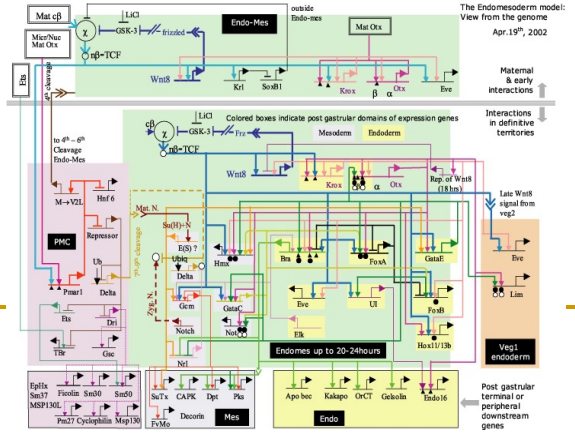
Sorin Istrail and Eric Davidson, Logic functions of the genomic cis-regulatory code, Proceedings of the National Academy of Sciences, vol. 102, No. 14, 4954-4959, 2005

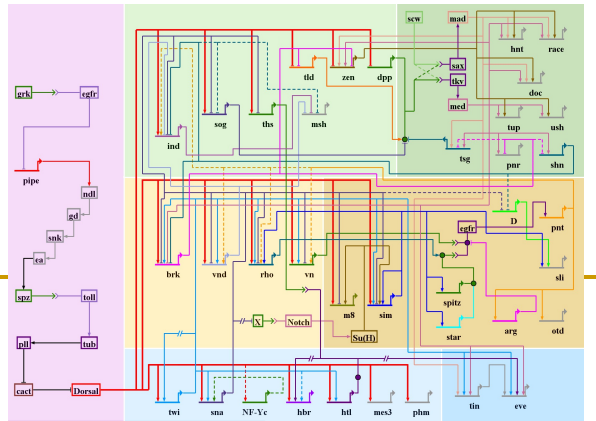
ABSTRACT: Cis-regulatory modules that control developmental gene expression process the regulatory inputs provided by the transcription factors for which they contain specific target sites. A prominent class of cis-regulatory processing functions can be modeled as logic operations. Many of these are combinatorial because they are mediated by multiple sites, although others are unitary. In this work, we illustrate the repertoire of cis-regulatory logic operations, as an approach toward a functional interpretation of the genomic regulatory code.

A Tale of Two Networks

Drosophila – Mike Levine

sea urchin –Eric Davidson

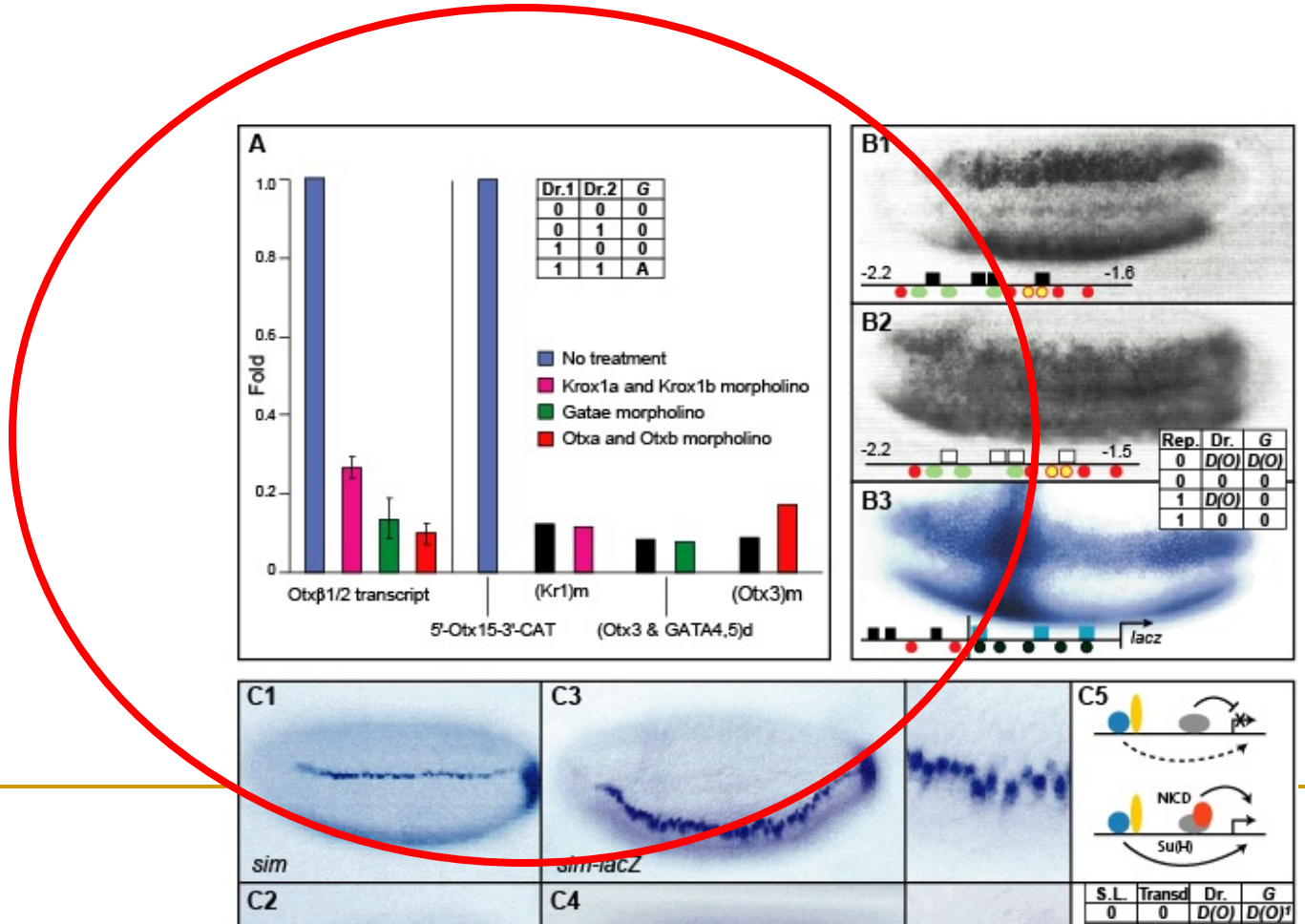




G-operators

Combinatorial Logic Functions

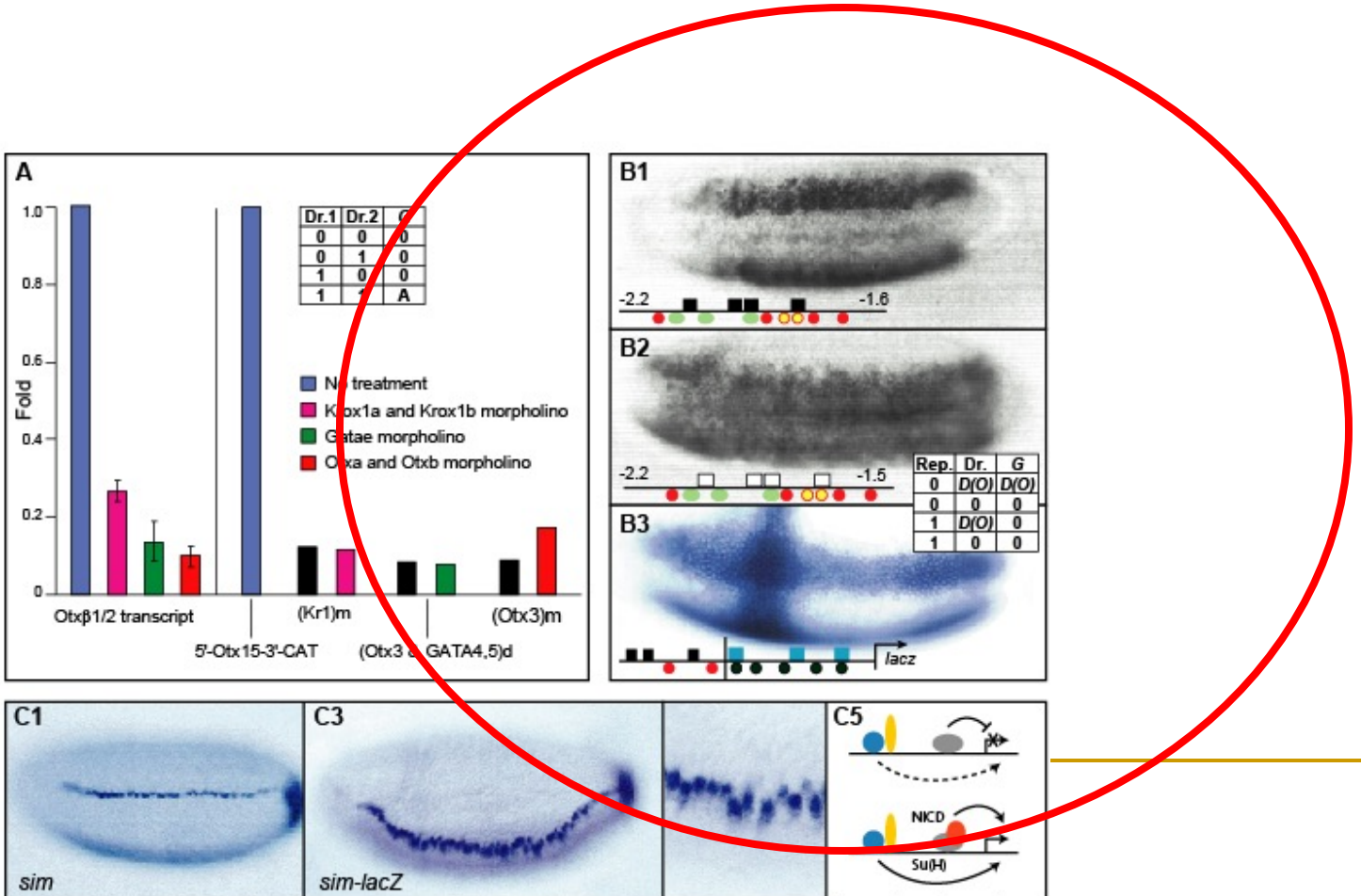
1.AND operators.



G-operators

Combinatorial Logic Functions

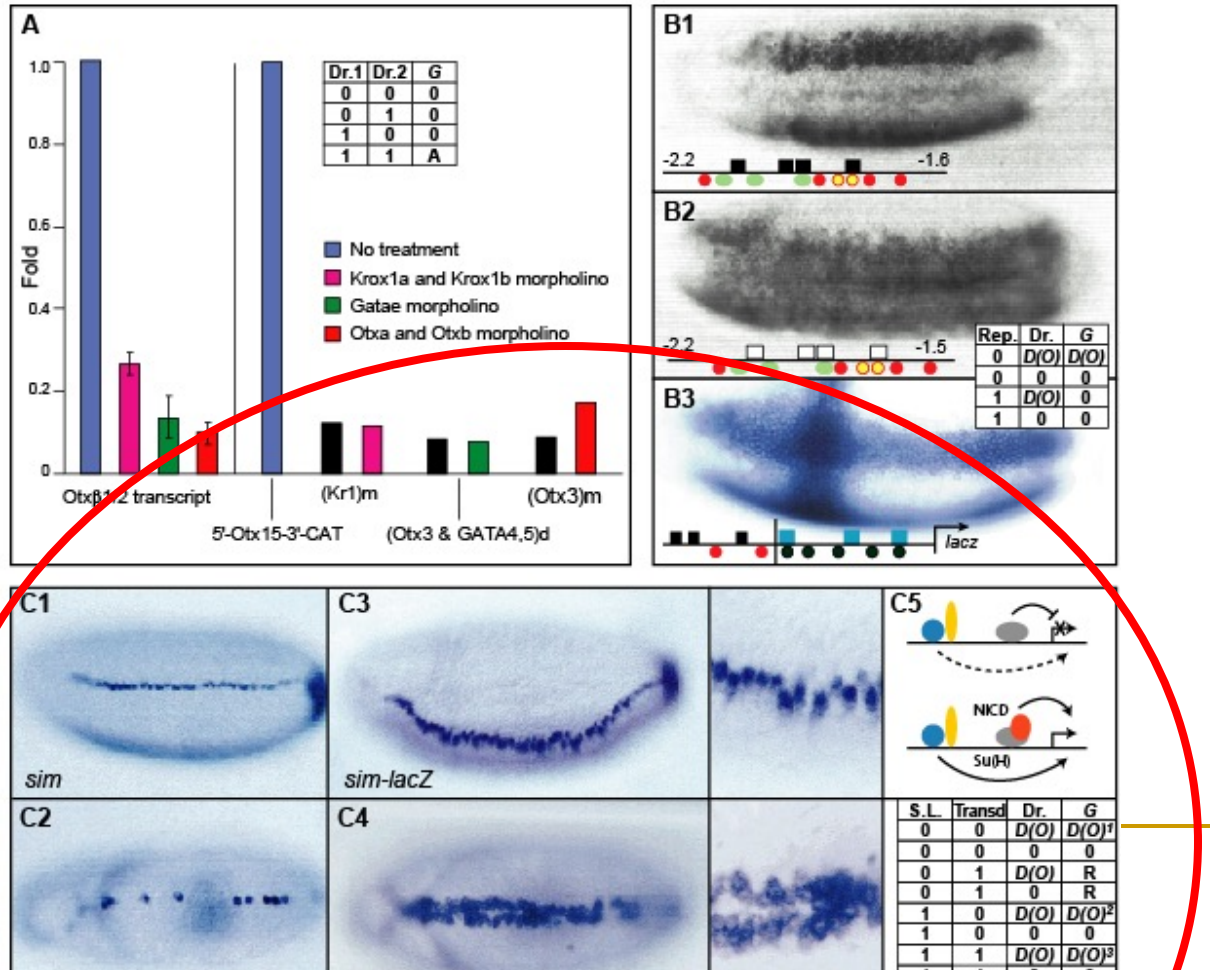
2. Short range repression binding within a CRM.



G-operators

Combinatorial Logic Functions

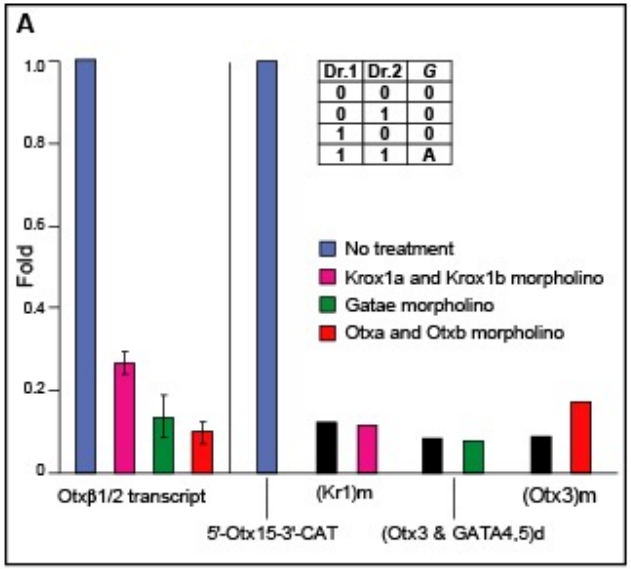
3. Signal mediated “toggle switch”.

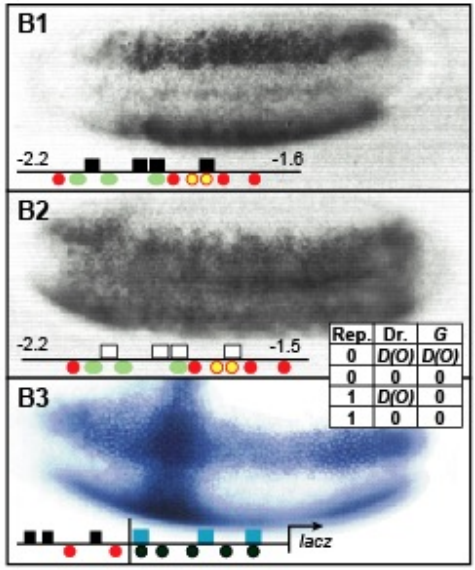


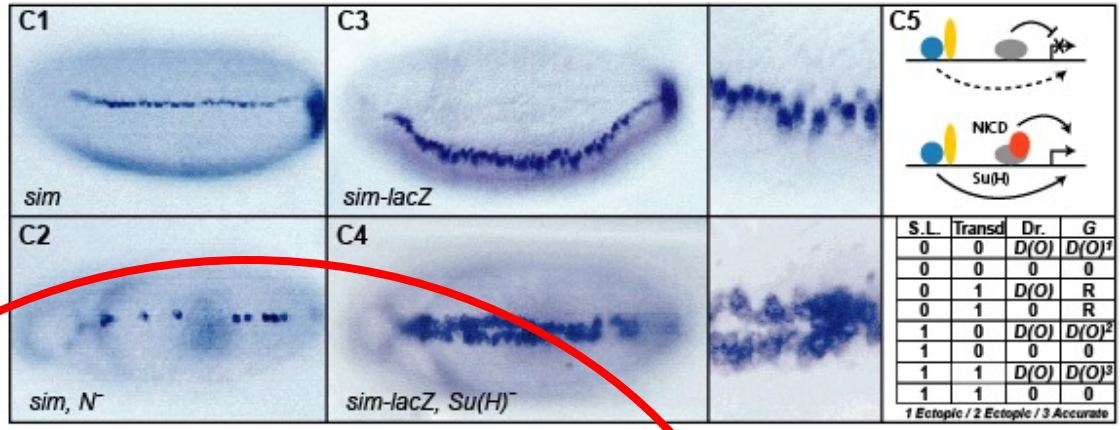
G-operators

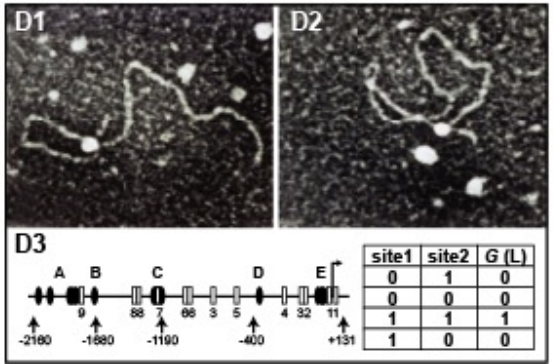
Combinatorial Logic Functions

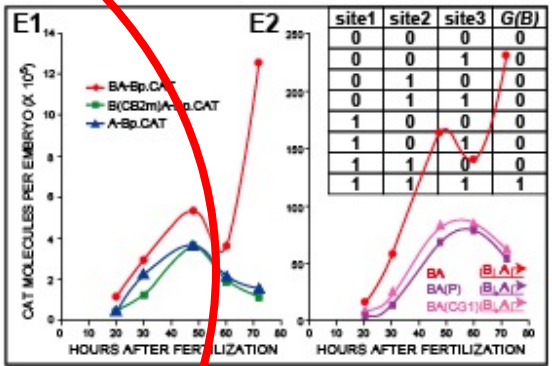
4. Essential DNA looping.







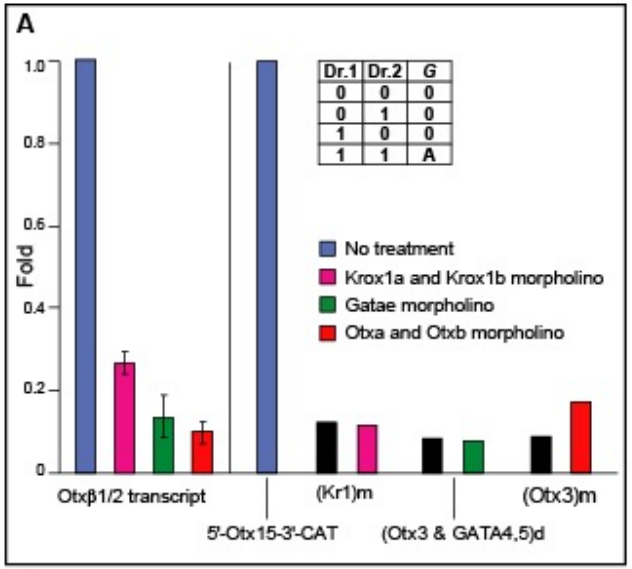


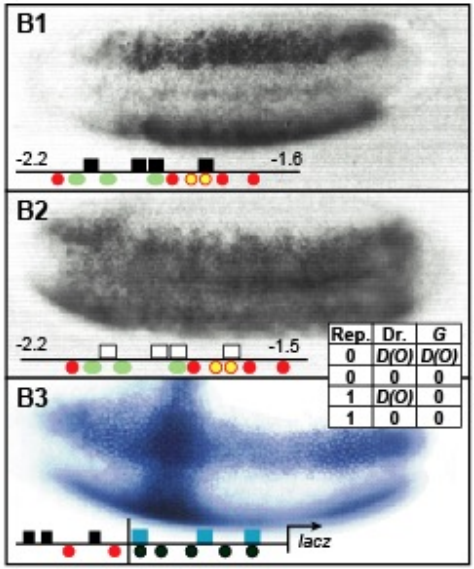


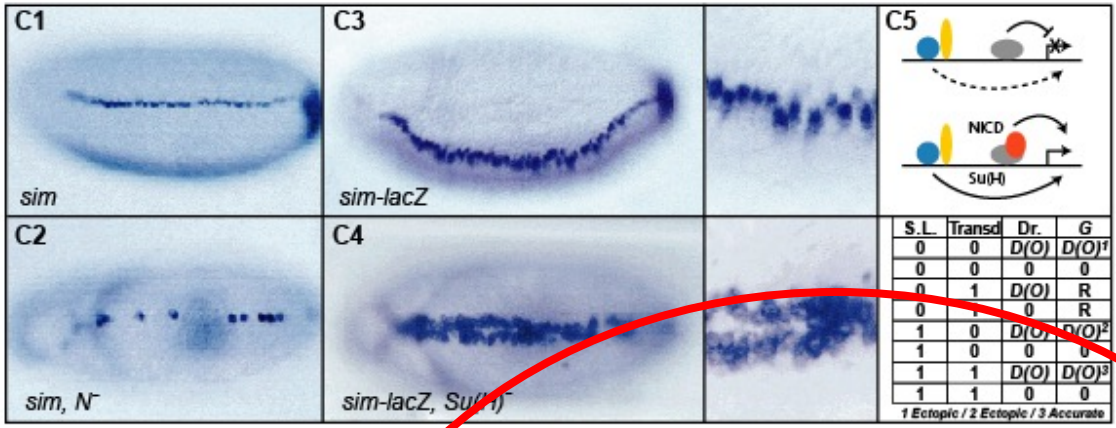
G-operators

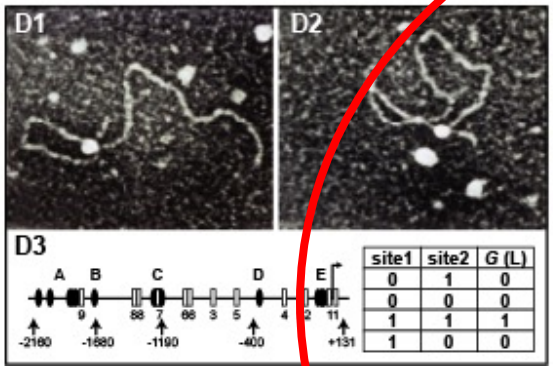
Combinatorial Logic Functions

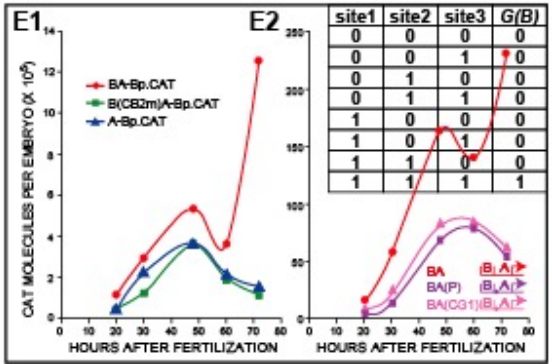
5. Module linker function.



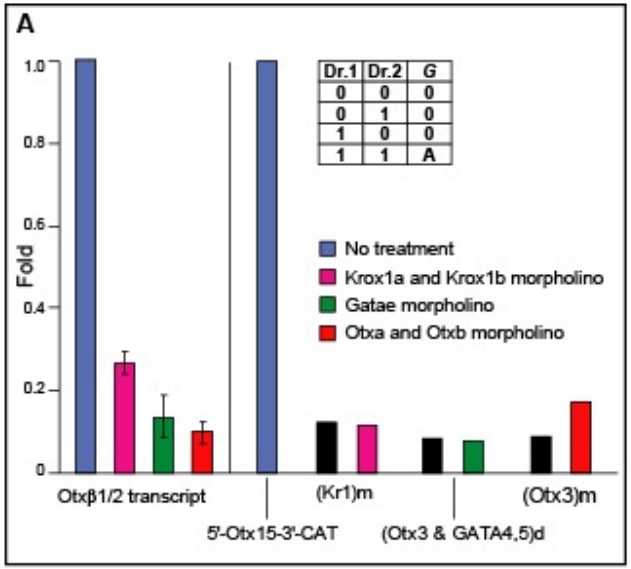


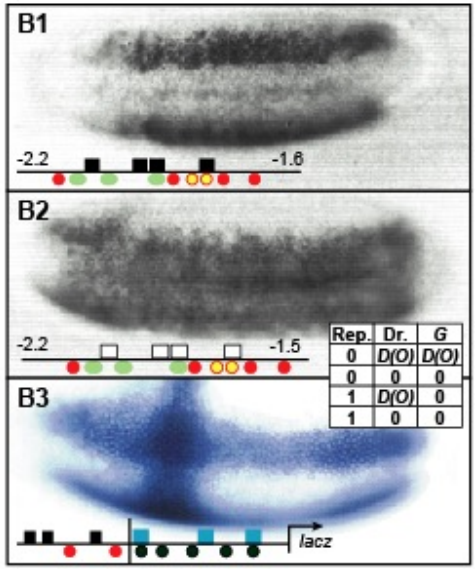


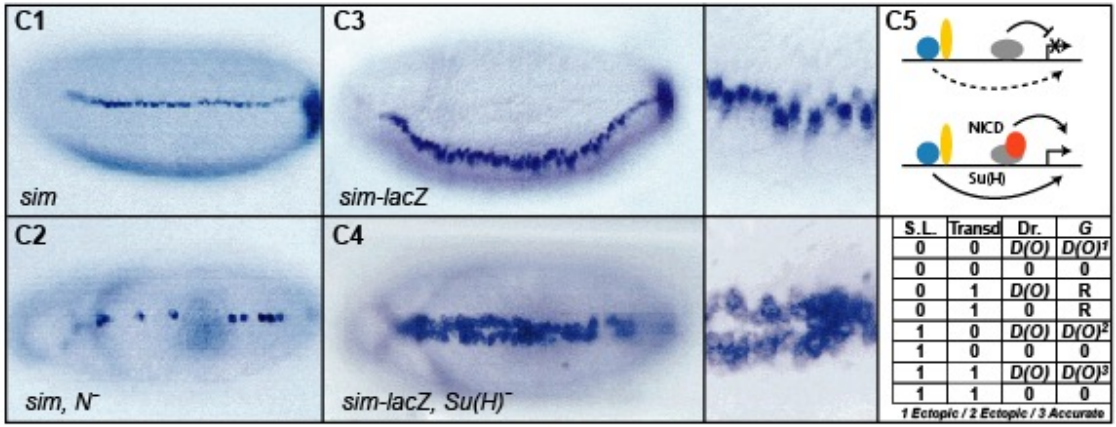


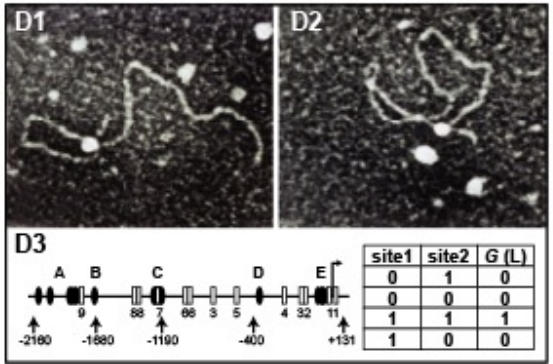


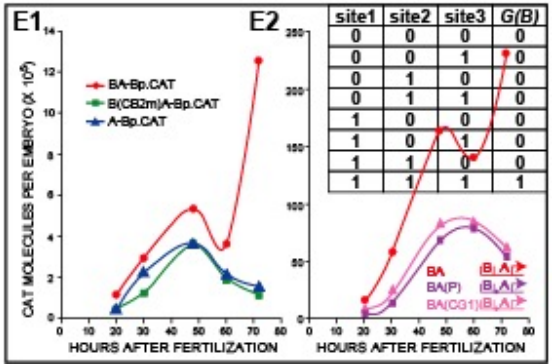
Logic functions of the genomic *cis*-regulatory code











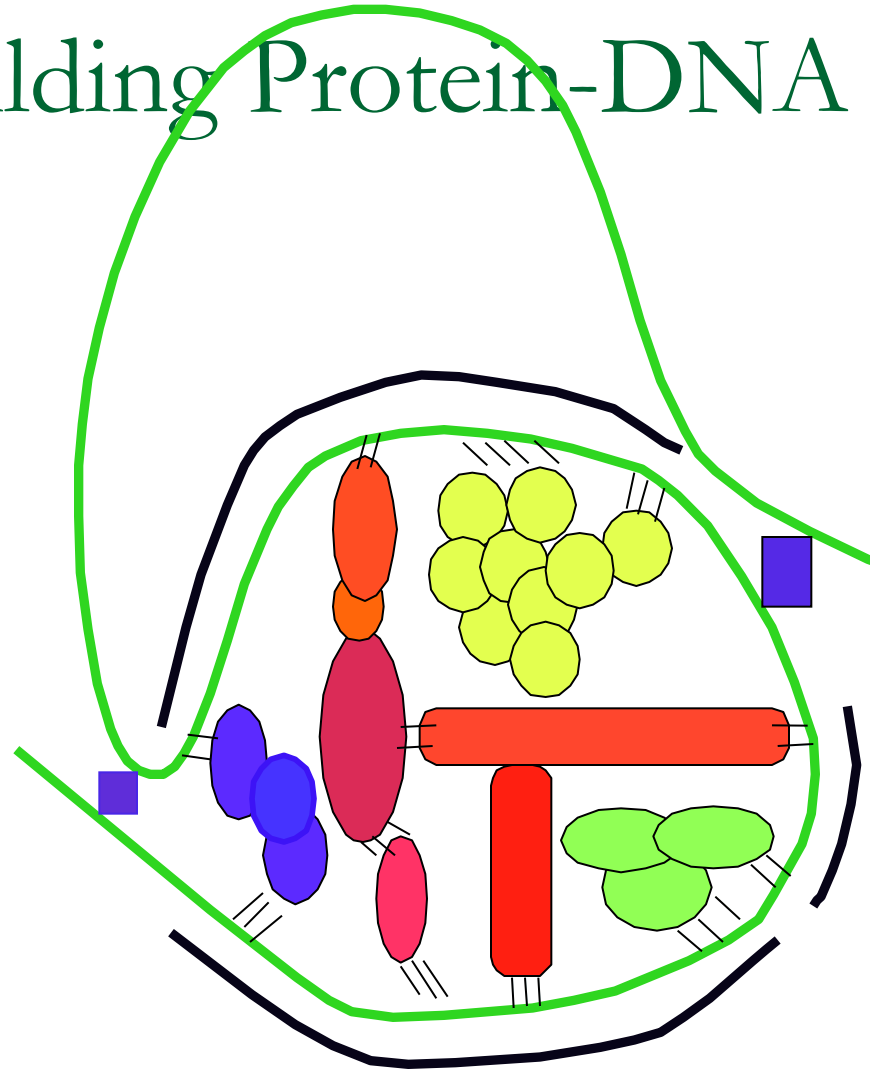
Functionally irreducible bulding blocks

- Two types:

LOGIC = e.g.,
"AND" regulatory and
switch functions, repression

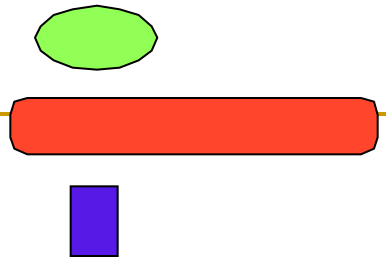
PROBABILISTIC =
constants in the
Bolouri-Davidson equation

Building Protein-DNA Assemblies



- DNA
- cismodule

- Cooperativity
- Linear-amp
- Gates
- Potentiality



- Inter-cismodule linkage
- Insulation
- Communication

building blocks of function

1. LOGIC: AND, OR1, OR2, XOR1, XOR2, GATE1, GATE2
2. CONTROL: Activation, Repression1, Repression2
3. CONTINUOUS-DISCRETE: Linear Amplification
4. ASSEMBLY: Positive Interactions (Cooperation, Communication), Negative Interaction (Insulation), Inter-cismodule Linkage, Potentiatlity

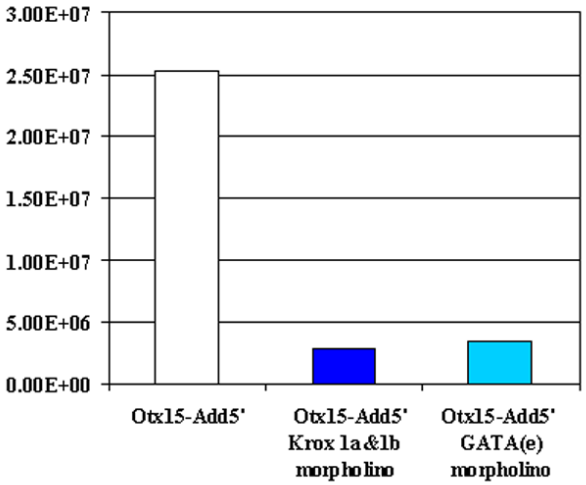
The 15 Gate Repertoire

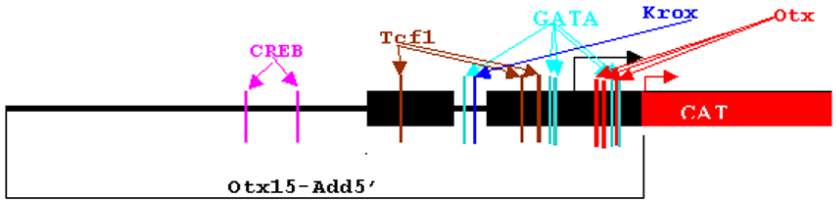
- SIMPLE ACTIVATION
- AND_LOGIC ACTIVATION
- OR_LOGIC ACTIVATION (2 types)
- COOPERATIVITY ACTIVATION
- EXCLUSIVE OR_LOGIC (2 types)
- SIGNAL-MEDIATED GATE (2 types)
- SHORT RANGE REPRESSION
- SILENCING
- INSULATION
- DNA LOOPING/COMMUNICATION
- LINEAR APMPLIFICATION
- INTERCISON LINKAGE

AND_LOGIC ACTIVATION

From unpublished data, C. –H. Yuh and E. Davidson

•Cartoon of expression construct consisting of a cison from the *otx* gene located about 15 kb upstream from the relevant transcription start site (*otx* element, β1/2 start site: Yuh *et al*., 2002), displaying inputs, BTA from the *endo16* gene, and reporter gene encoding CAT enzyme. Histogram shows control activity and near complete extinction of output when *either* GataE or Krox inputs are blocked by introduction of morpholino-substituted antisense oligonucleotides against the mRNAs encoding these factors.



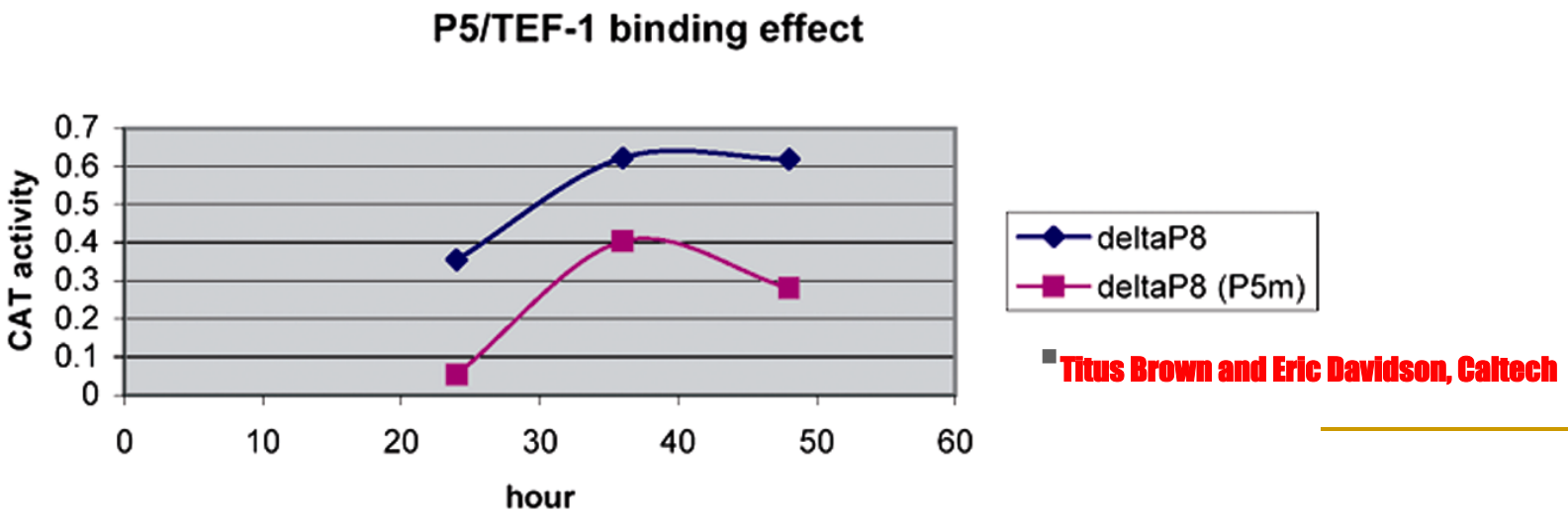


OR_LOGIC ACTIVATION:

DIVERSE *TRANS* INPUTS

- **Description:** Additive activation functions of two or more different *trans*-inputs, such that no one of them is essential, and some of the overall output of the cison is contributed by each; This form of activation function may (or may not) be combined with other kinds of activation function, mediated by other inputs.
- **Experimental indications:** Elimination of any one of the interactions included, by either mutation or *trans*-perturbation, partially decreases cison output, but spatial expression is not affected, nor is expression eliminated.

Unpublished data, C. T. Brown and E. Davidson

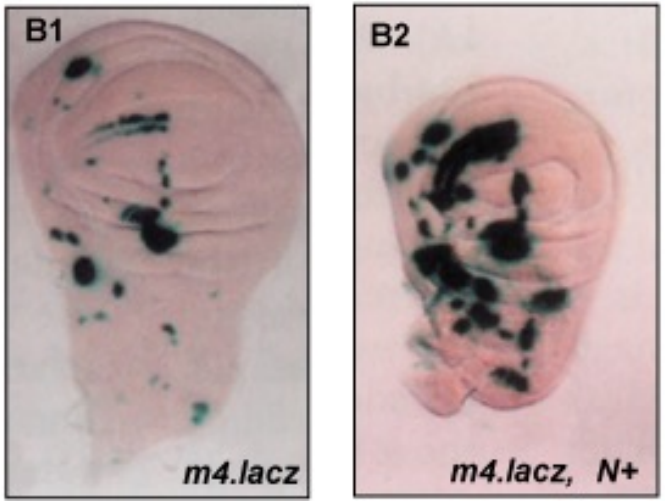


Titus Brown and Eric Davidson, Caltech

SIGNAL- MEDIATED GATES:

ADDITIONAL POSITIVE INPUT

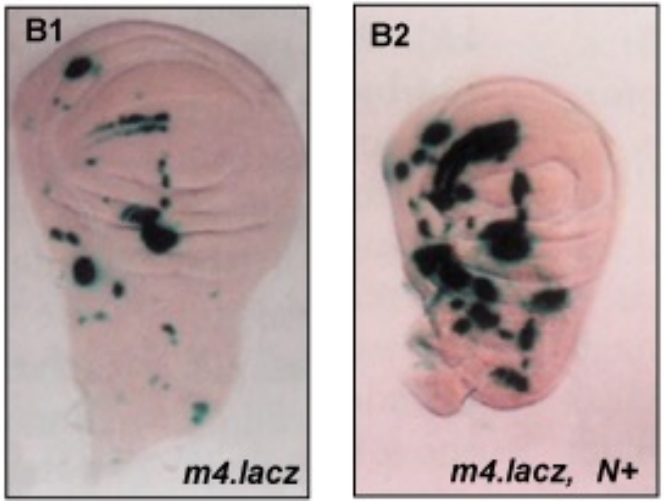
- **Description:** A signal from outside the cell mediated by a specific ligand –receptor interaction has the effect of causing the appearance in the cison of an additional activation function (AND,OR, or COOPERATIVE activation function). Without the signal this function would be absent and expression will not occur or will occur at a reduced rate, but the signal-mediated activation function is not sufficient in itself.



SIGNAL- MEDIATED GATES:

ADDITIONAL POSITIVE INPUT

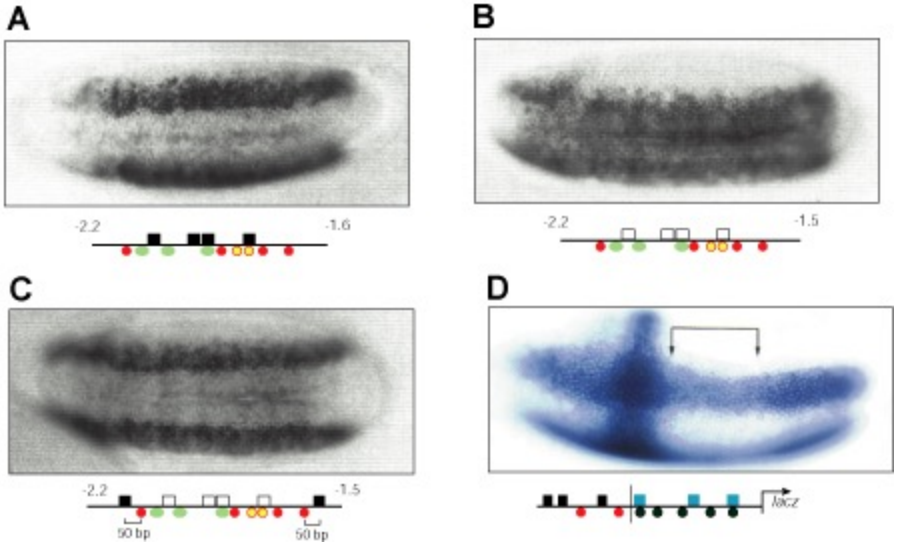
- **. Experimental indications:** Interference with presentation of ligand, or with function of receptor, or mutation of target site for signal-mediated activator, causes decrease or abolition of expression, but not ectopic expression. Overexpression of the signaling system increases expression, but does not cause ectopic



Expression of a construct including the Notch- response element of the *Espl* gene *m4* visualized by lacz reporter gene expression, in the proneural domains of the *Drosophila* wing; Left, control; Right, enhanced expression, in the same domains, caused by global expression of constitutively active Notch receptor. From Nelleson et al, 1999.

SHORT RANGE REPRESSION

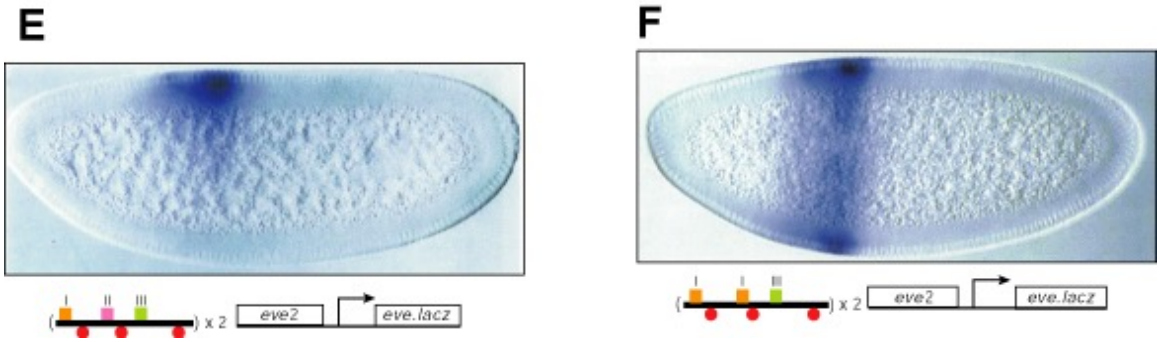
- **. Description of Function:** Dominant repressor bound within the cison cancels the positive output of transcriptional activators also bound within the cison, which in the absence of the factor mediating the repression, suffice to cause it to be active.



Experimental indications: **Mutation of the site where the repressor binds within the cison causes ectopic expression; if the site is moved to a location outside the cison or far from the activators within the cison the repression fails to occur; if the cison including the short range repression function is linked to another cison in an expression construct in which both utilize the same basal transcription apparatus, the repressor affects exclusively the expression of its own cison, and the other in the construct is expressed as it would be by itself.**

SILENCING

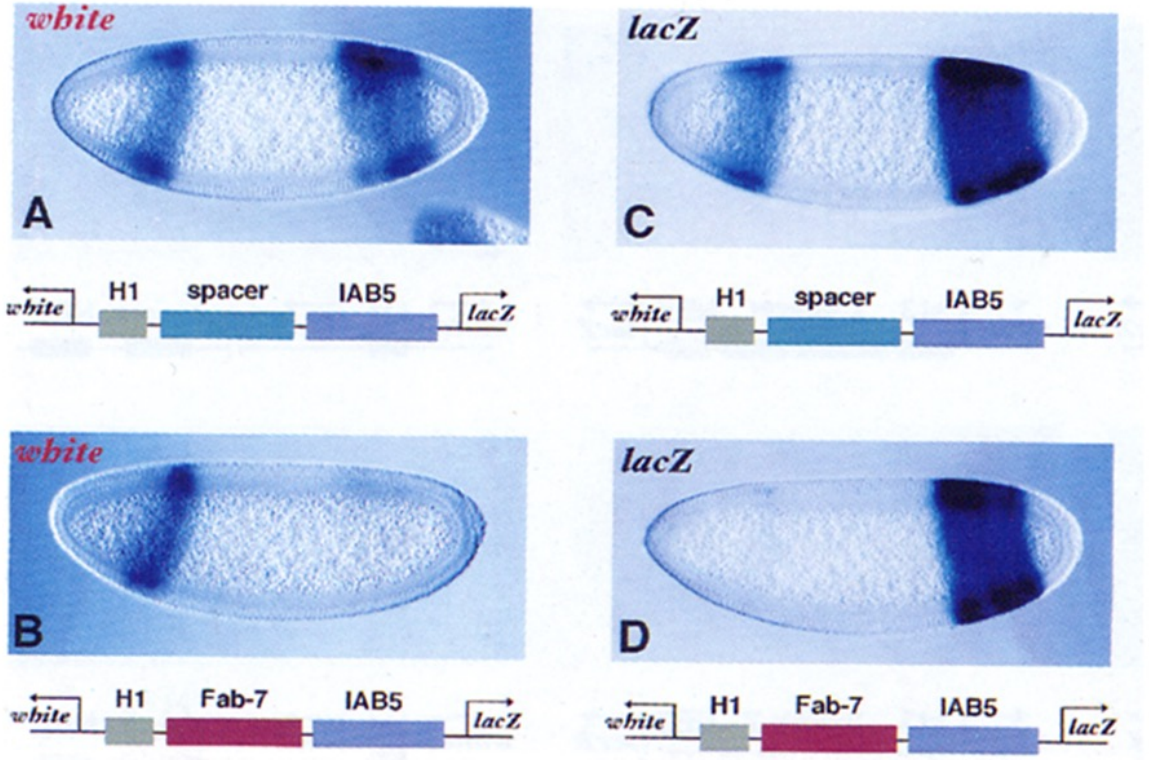
- **Description:** Binding of transcription factor that confers silencing function shuts down all cisons serviced by the basal transcription apparatus of the gene, irrespective of where the factor binds.



Expression of a compound expression construct containing a *zen* cison that includes a silence, distal, plus the *eve* stripe 2 element, proximal. (A) the silencer shuts down expression of the *zen* gene in the ventral region, and here it shuts down the associated *eve* element in the same region as well.(B), If the silencing function of the zen element is destroyed by mutation of a key site, the normal eve expression pattern is seen. From Cai et al, 1996.

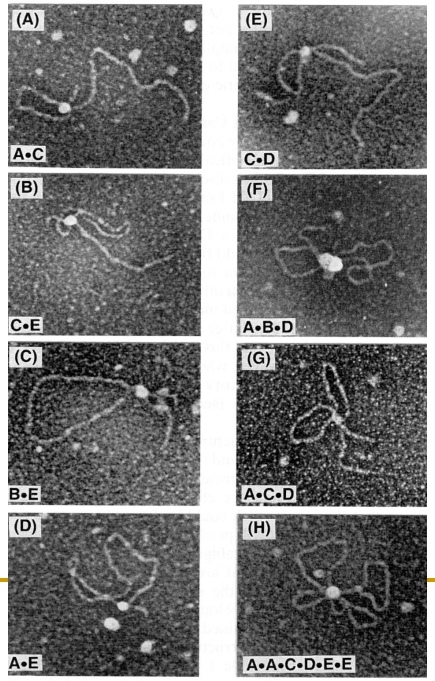
INSULATION

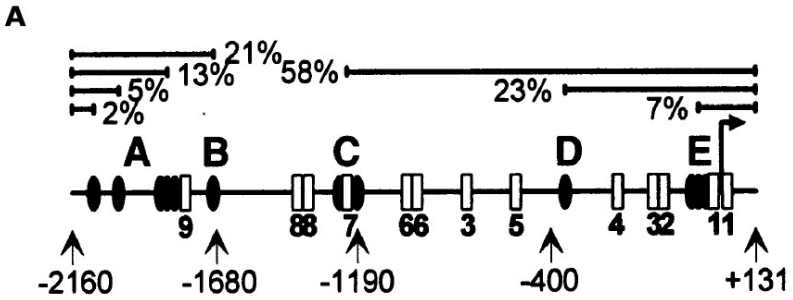
- **Description: Function is insulation of given basal transciption apparatus, or proximal cison plus the basal transcription apparatus, from given distal cisons.**



COMMUNICATION

- Decsription: Communication of regulatory function between cisons or between a cison and the basal transcription apparatus, by formation of loops that bring distant sites together so they can interact. It is mediated by sequence specific DNA binding factors which also bind to each other.





Four categories of operator functions: D, F, G, E

1. **D-operators**. Transcriptional activation operators: quantitative transcriptional activation functions mediated directly by driver target sites, considered to be directly proportional over much of its range to their values

2. **F-operators**. BTA control operators: operator functions mediated by sites that control interaction with the BTA, according to the intrinsic properties of the individual species of DNA binding proteins (and their cofactors)

3. **G-operators**. Combinatorial logic operators: operator functions that by definition depend on the participation of multiple CRM sites, i.e., on the combinatorial interaction of the proteins interacting at these sites (an their cofactors)

4. **E-operators**. External control operators: permissive or non-permissive operator functions mediated by sites outside CRM

F-operators
BTA control operators

1. CRM Silencers.

- Some CRM target sites bind repressors that silence the BTA, so that their effect is not limited to the CRM that includes these sites, but rather extend to any CRM utilizing that BTA for transcriptional expression.
- Silencers behave as unit dominant negative regulators.
- Silencers do not require interactions with nearby target sites where activator bind, but (via dedicated co-factors) they directly affect the BTA.
- In our models, if a silencer is present in a CRM and occupied, then

$$F(Oc(s,t)) = R$$

where R stands for "Repression" and by the rules below the result is to set the output of the regulatory system to zero

F-operators

BTA control operators (cont)

2. Communicators.

- The site function analyses carried out on the *endo16* and *cyIIIa* genes have both revealed CRM target sites that are required for a function that is mediated by other sites elsewhere to have an effect at the level of BTA execution
- In endo16 the spatial domain of the early phase of expression is confined to the future endoderm by repressors that bind at known target sites in upstream modules about a Kb away from module A, one of which is a Creb factor. All the repressor interactions are inutile, and ectopic expression occurs the same as if the repressor sites are absent, if a certain target site of Module A is mutated.
- In cyIIa, a cite for POU domain factor in the proximal module of that regulatory system is requiredfor the spatially controlled activating function or a more distal module to have any effect on expression.

F-operators

BTA control operators (cont)

2. Communicators (cont).

- Whether directly or indirectly, the proteins binding at these target sites cause regulatory inputs from outside the CRM (negative and positive respectively) to have an effect, which means interacting with the BTA, i.e., transferring or communicating the inputs to the BTA.
- Interacting with BTA may indeed be a general function of the proximally CRMs.
- In the above two examples (endo16 and cyllla), if the communicator sites are absent the regulatory values of the respective distal modules is zero, irrespective of whatever interactions occur there; that is, they behave in a Boolean manner.

F-operators

BTA control operators (cont)

3. Amplifiers.

- Another function discovered in the endo16 analysis was a linear amplification of a positive regulatory input from Module B by sites in Module A. In the quantitative kinetic study the expression of a construct including A and B modules plus the BTA is equal at all points in time to about 4x the output of a construct including only Module B plus the BTA.

- Directly or indirectly, this function of Module A requires communication of the Module B output and an effect on the level of BTA activity (but not de novo generation of a positive regulatory output in Module A, since the function is located in Module A, the output that is amplified comes from Module B, as shown by experiments)

E-operators

External Control Operators

Functions that are located outside the CRM butwhich control its activity cannot be enumerated here; for one thing it remains unclear exactly how major classes of external mechanism work, though it is completely clear that they do work. There are what could be summed together as "domain choice operators." The phrase denotes (non-exclusively):

1. **Alternative looping**, which brings into action a given CRM under given circumstances and other CRMs under other circumstances;

2. **Insulators** which prevent more proximal CRMs from interacting with BTA while permitting more distal ones to so interact

3. **Insulators** that in certain developmental circumstances transfer a given CRM to an inaccessible or sequestered chromatin domain

4. **Distant locus activators**, which if bound by the factors for which they contain target sites permit more proximal CRMs to be active, rpobably via loop formation with them

E-operators

External Control Operators (cont)

- Another class of external operators probably exists as well, and this is **external silencers**, which are not located within CRMs. Their target sites occur in the DNA flanking the CRM proper, not within the conserved sequence patches or complex clusters of diverse target sites that can be used to identify CRMs, or in any case, they are located in much smaller, much less diverse site clusters.

- Like other silencers, external silencers function in a dominantly negative, Boolean manner.

Summary

- Our main object has been to attempt a framework for interpretation of the *cis*-regulatory sequence code by linking target sites directly to a defined set of elemental functions, the integrated combination of which is the output of the molecular control element.
- Six basic properties, or principles of CRM function, that we claim apply to all CRMs, are taken into account this approach:

Axioms/Principles

1. The functional repertoire of each CRM is a constant, sequence based, feature of the species
2. The specific design of any CRM can be expressed in terms of its elemental functions
3. CRMs process continuously varying driver inputs
4. Many CRM processing functions can me modeled as logic operations
5. Occupancy is causal: CRM outputs are intrinsically conditional on site occupancies
6. In all cases where there is a qualitatively unique factor site interaction, the consequences of site mutation and of absence of occupancy due to absence of a unique DNA-binding factor are equal

$$O_{cis} = O_{trans}$$

The Genomic *cis*-Regulatory Code

- Yet, ultimately, the objective of "reading" the genomic code on inspection will require a more or less complete repertoire of such functions; this means that we will need to have for reference many more detailed *cis*-regulatory examinations in which the significance of each specific binding site has been determined.

- Comparatively speaking, the diversity or complexity of the genomic regulatory code is going to be greatly less than the diversity of the biochemical operations that execute each type of function (that is why it is properly referred as a "code").

The Genomic *cis*-Regulatory Code

- This is obvious from the fact that there are many different factors that in our terms carry out the same functions of repression, of activation, probably of looping, of signal transduction, and so forth; and even more different sets of factors that execute G class combinatorial functions such as AND logic.

- Tackling the genomic regulatory code head-on is liable to be a more direct avenue to learning what it says than dissection of the particular biochemistry operative in every different CRM. To reverse the argument, the mechanistic biochemical exploration of *cis*-regulatory function will indeed be much facilitated if it can be couched in terms of an elemental functional repertoire

Davidson Causality

Eric Davidson, "Genomics, "Discovery Science", Systems Biology and Causal Explanation: What Really Works?"
Perspectives in Biology and Medicine, Volume 58, Number 2, Spring 2015, pp. 165-181

Eric Davidson, "Systems Biology, Choices Arising", in Sara Green (Editor), "Philosophy of Systems Biology", pp. 69-78, Springer 2017

Citing Eric Davidson:

A total change in the epistemological landscape brought about by successful applications of the percepts of systems biology

Developmental biology thoughout the twentieth century suffered inescapably from Karl Popper's criticism of inductive scientific process, that all it takes is a single instance to the contrary to prove an inductive mechanisctic idea wrong.

I am not referring here to the majority predilection in twentieth century developmental biology to report exclusively empirical descriptions of embryonic events, some of which were very broad, but which no matter how detailed, could provide only factual background.

Citing Eric Davidson:

Description could be enormously indicative, but cannot complete the process of doing science, since it generates no causal explanations.

In the last century those experimental efforts aimed at revealing islands of causality in developmental biology research were invariably focused on single aspects of the overall process, on single genes, or on single instances, on the tractable minute bits of a system that could be satisfactorily taken down to brass tacks in focused experiments. … by isolating the unfathomable overall system what small pieces of it can be experimentally encompassed .. in “clean” situations.

Citing Eric Davidson:

Popper used the inductive assumption “all swans are white” as his paradigmatic example of the intrinsic problem in inductive logic, wherein the discovery of black swans in Australia provided an example of a previously unexpected falsification.

… who could be sure if another Australian black sewan would not turn up when further genes operating within the same system would be investigated?

Citing Eric Davidson:

But genomics has now changed everything!

The foundational principle of systems developmental biology, i.e., that all parts of the system must be included in mechanistic analysis, in principle and in practice offers a waterproof counter to the concern that it is extremely difficult or impossible to know if black swans, i.e., qualitatively different mechanisms, are lurking elsewhere than in the islands of phenomena thus far chosen for causal analysis.

Control of the developmental process is mediated primarily in and by the *regulatory genome,* and all the parts of the regulatory genome t are engaged in any give such process can now be determined exactly. Completed Gene Regulatory Networks (GRN) analyses present computational Demonstration that the sea urchin embryo GRN is sufficient to predictively explain almost all the spatial regulatory transactions of a large portion of the … embryonic process.

Citing Eric Davidson:

A set of urgent new epistemologival issues: a perversion of of scientific practice

".. The self-described field of "systems biology" has given protective cover, and beyond that, in government and institutional circles a pseudo rationale, to an enormous entreprise, the object of which is to obtain very large, solely observational data sets that are to be interpreted by ex post facto statistical correlations.

This type of activity has been elevated to the status of shining new universe of "discovery science" (an oxymoron if ever there was) which at last will supplant the "traditional" chains and bonds of prejudiced, expensive, slow, "hypothesis" testing by the means of the experimental method."

Citing Eric Davidson:

The epistemological issue that arises is not attenuated no matter how elegant the instrumentation, how clever the mathematics, not how massive the datasets. This is whether scientific causality can ever be established without perturbations of the behaviour of a system, without experimental tests of logical predictions of the results of perturbation of change in Condition, i.e., without “experimental hypothesis testing”.

Another fundamental epistemological problem that “discovery science” generates is an innate contradiction between conclusions based on statistical deduction from an unperturbed dataset, vs. conclusions based on tests of hypotheses generated by consideration of *prior scientific knowledge.*

Citing Eric Davidson:

An inbuilt scientific agnosticism and tolerance of ignorance of prior scientific is characteristic of "discovery science"; while true science intrinsically progresses by conceptually based operations executed *on* prior knowledge, be these operations deliberate revisions, or challenges, or confirmations, or extensions of the prior knowledge.

Eric Davidson's

Axiomatization of Causality

in the domain of molecular
developmental
gene regulatory networks
biology

Axiom 1

Rooted causality explanation is anchored in the genome.

Unrooted causality explanation is phenomenology, based on a small part of the whole system.

Explanation of system-wide control mechanism for a developmental process must begin with recognition of the elements of the genomic regulatory sequences.

Axiom 2

No sum of unrooted causality explanations equals a complete rooted explanation.

“Solution of the mechanism” is impossible if only a minor fraction of the components active in a process and of their interactions is involved in the analysis. No addition of fragmentary unrooted explanations (even infinitely many of them) will ever sum up to a complete rooted explanation.

The reason in the enormous combinatorial number of degrees of freedom with which the interactive control systems of development can be, and evolutionarily have been, assembled.

Axiom 3

Success measures must be assessed system wide.

The usefulness, success, failure, validation, or refutation of an explanation emerging from experimental systems developmental biology must be assessed system wide.

These tests must challenge the ability of the explanation to predict the behavior of the system as a whole, or the behavior of large sectors of it that include many individual components and their interactions.

Axiom 4

From sea of phenomenology to sea of causality: inverting the "islands in the sea" metaphor.

Phenomenology will inevitably result from research focused exclusively on a very small fraction of the components of the systems and their interactions, and thus unrooted causality explanation inevitably produces phenomenological information.

In the 20th century biology, there were many islands of causality floating in the vast sea of phenomenology. System-wide rooted causality explanation inverts this relationship, so that when successful it gives rise to a framework of causality, within which are always to be found islands of not yet understood phenomenology.

Axiom 5

High-resolution quantitative and qualitative observations are irreplaceable.

Problem solving within a tiny localized domain is a hopeless approach to system-wide explanation.

High-resolution quantitative and qualitative observations of transcriptional functions in time and space, and many other ‘‘descriptive’’ molecular, cell biology, and developmental aspects are of irreplaceable value as a starting point of a perturbation analysis, so long as they are system wide.

Axiom 6

Explanation /mechanism can only be revealed by deliberate experimental perturbations and predictive challenge of the system.

Considerations of secondary and tertiary effects are needed because of functional interactions within the system, as well as of the effects of multiple inputs at each node of the system.

Therefore, perturbation analysis in systems developmental biology demands the intellectual guidance provided by use of hypothesis at every step.

THE REGULATORY GENOME

THE REGULATORY GENOME AND THE COMPUTER

S. Istrail, S. Ben-Tabou de-Leon, E.H. Davidson. The Regulatory Genome and the Computer, Developmental Biology, vol 310, pp. 187-195, 2007

ABSTRACT: The definitive feature of the many thousand *cis*-regulatory control modules in an animal genome is their information processing capability. These modules are "wired" together in large networks that control major processes such as development; they constitute "genomic computers." Each control module receives multiple inputs in the form of the incident transcription factors which bind to them. The functions they execute upon these inputs can be reduced to basic AND, OR and NOT logic functions, which are also the unit logic functions of electronic computers. Here we consider the operating principles of the genomic computer, the product of evolution, in comparison to those of electronic computers. For example, in the genomic computer intra-machine communication occurs by means of diffusion (of transcription factors), while in electronic computers it occurs by electron transit along pre-organized wires. There follow fundamental differences in design principle in respect to the meaning of time, speed, multiplicity of processors, memory, robustness of computation and hardware and software. The genomic computer controls spatial gene expression in the development of the body plan, and its appearance in remote evolutionary time must be considered to have been a founding requirement for animal grade life.

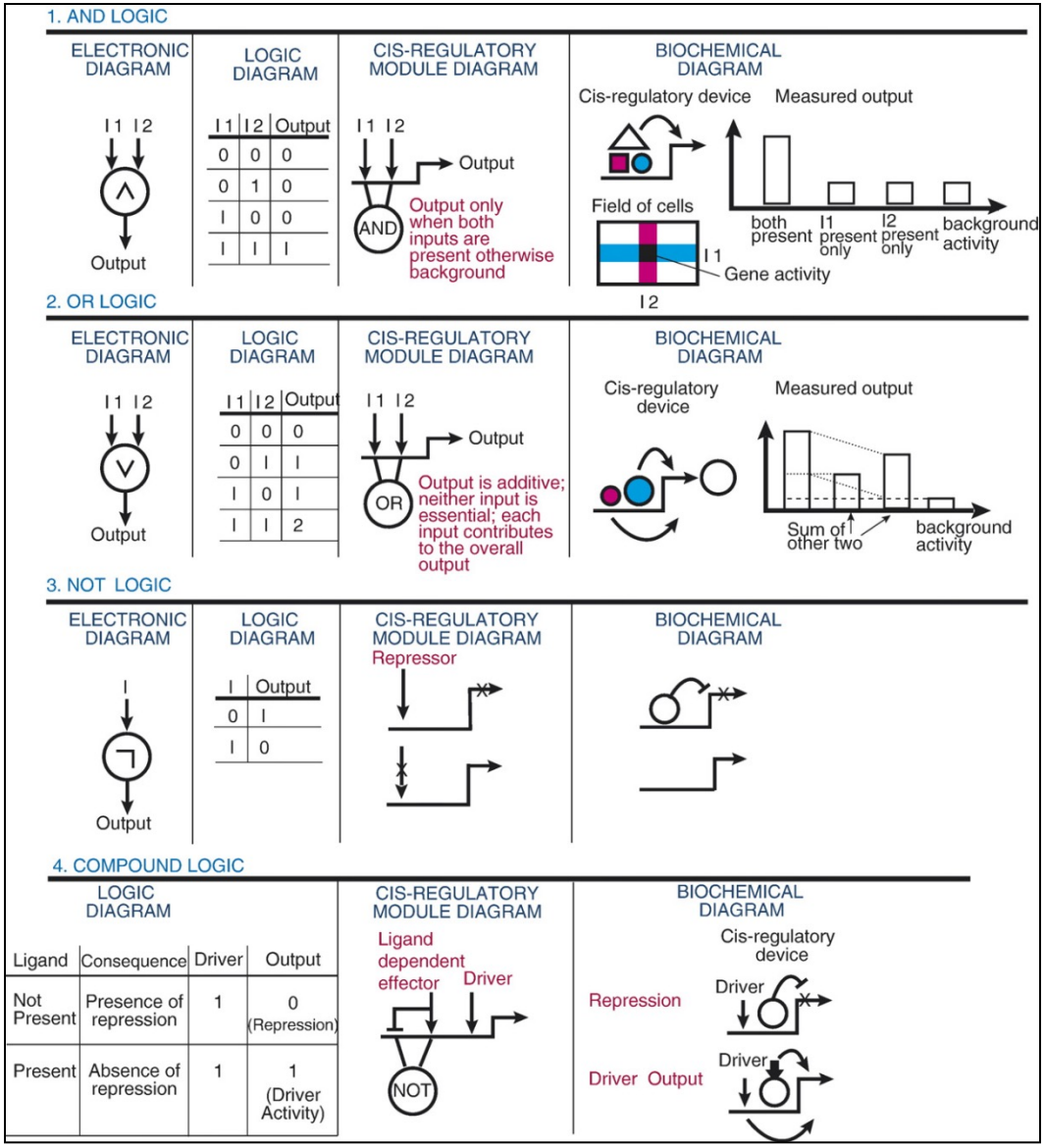
Our paper "The Regulatory Genome and the Computer" was written as an homage of the 50th anniversary of the publication of

"The Computer and the Brain"

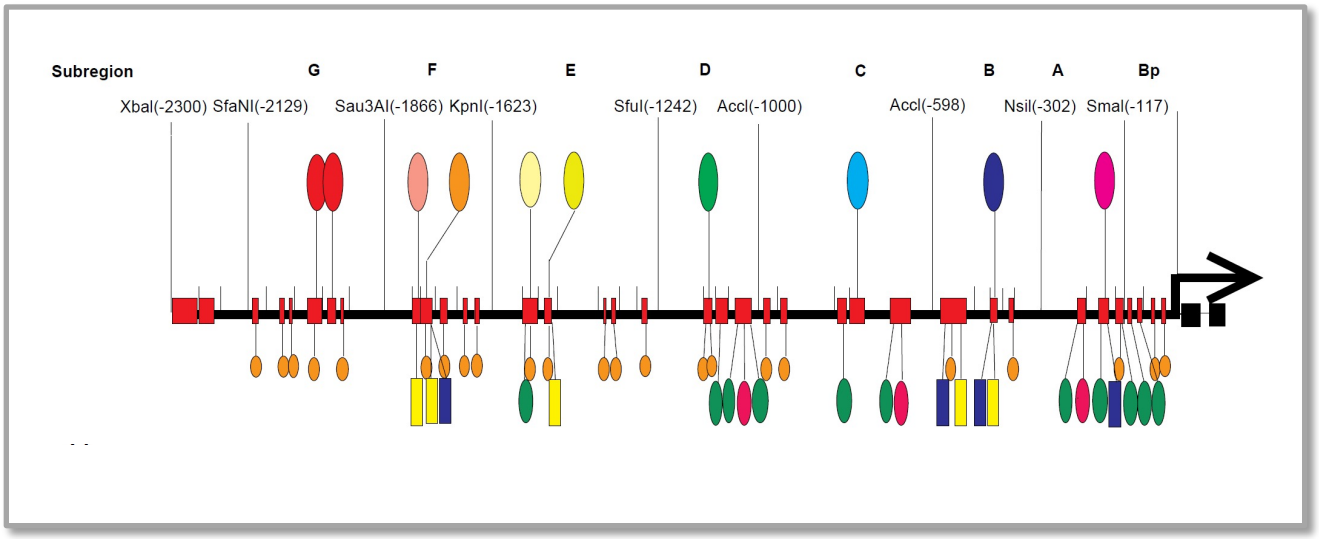
by John von Neumann

Yale University Press 1958

The Electronic Computer and the Regulatory Genome: basic logic operations executed by both



Regulatory states: combinations of transcription factors controlling spatial gene expression



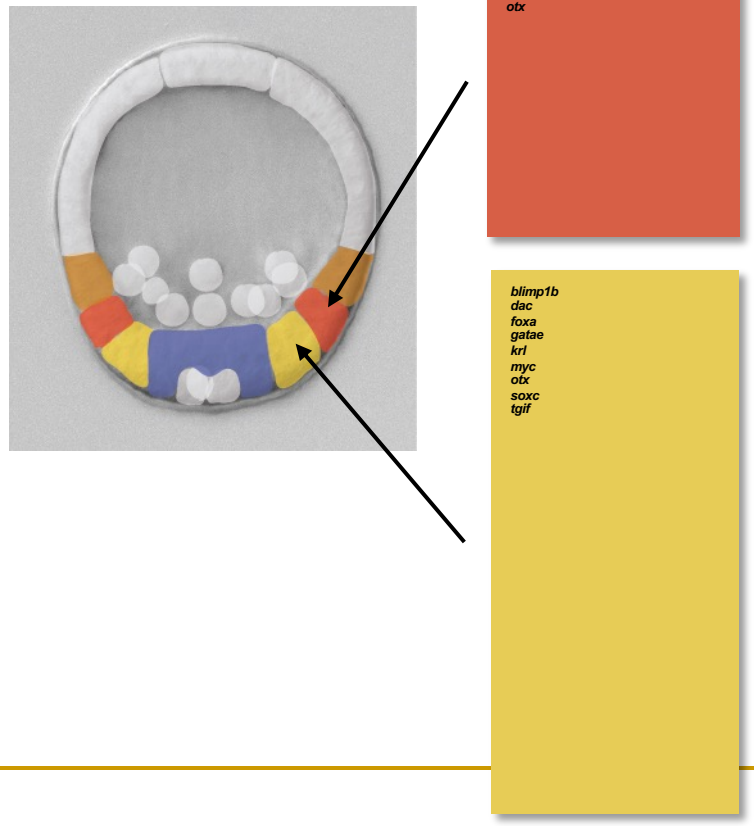
endo16 cis-regulatory sequence and transcriptional inputs

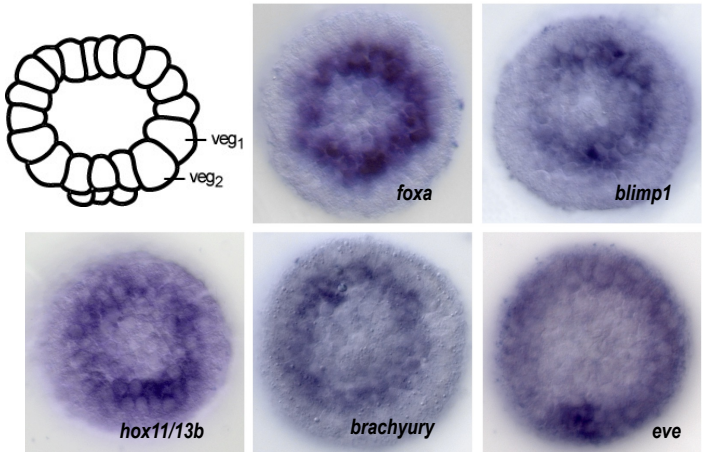
endo16 expression

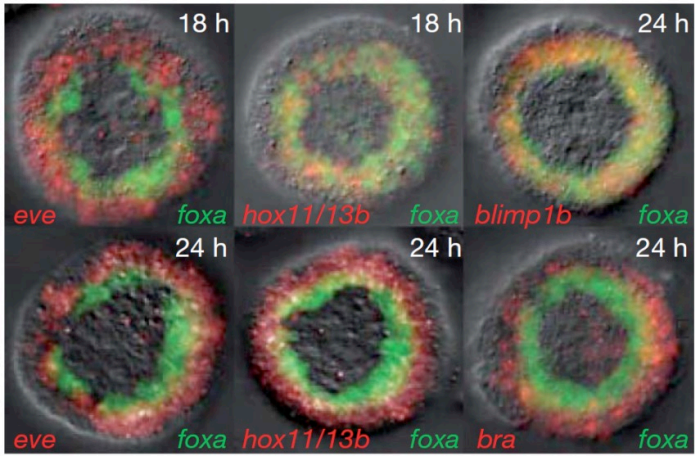
Yuh, Ransick, Martinez, Britten, Davidson, 1994 Mech. Dev.

Ransick & Davidson, 1998 Dev. Biol.

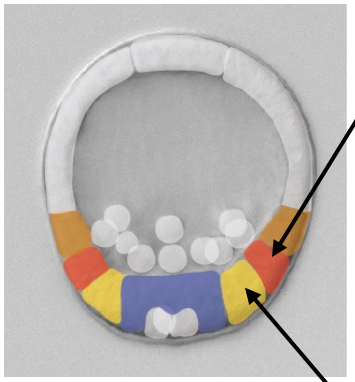
Experimental analysis of the endodermal GRN: identification of transcription factors





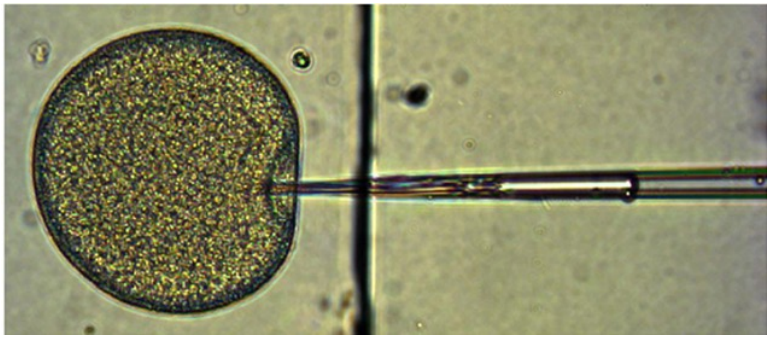


GRN analysis



brachyury
eve
hnf1
hox11/13b
otx

blimp1b
dac
foxa
gatae
krl
myc
otx
soxc
tgif



d 18 h

Input \ Output	*blimp1b*	*brachyury*	*brn1/2/4*	*dac*	*eve*	*foxa*	*gatae*	*hnf1*	*hox11/13b*	*krlz13*	*myc*	*otx-β*	*soxc*	*tgif*
Blimp1														+
Brachyury							+					+		
Dac														
Eve					-									
FoxA														
GataE														
Hnf1														
Hox11/13b	+	+				+								
Krl/Z13														+
Myc														
Otx														
SoxC													-	
Tgif	not expressed													

e 24 h

Input \ Output	*blimp1b*	*brachyury*	*brn1/2/4*	*dac*	*eve*	*foxa*	*gatae*	*hnf1*	*hox11/13b*	*krlz13*	*myc*	*otx-β*	*soxc*	*tgif*
Blimp1	-													+
Brachyury								+						+
Dac														
Eve	+							+	+					
FoxA														
GataE										-				
Hnf1														
Hox11/13b	+	+				+	+	+	-					+
Krl/Z13														
Myc											-			+
Otx														
SoxC													-	
Tgif														

f 27 h

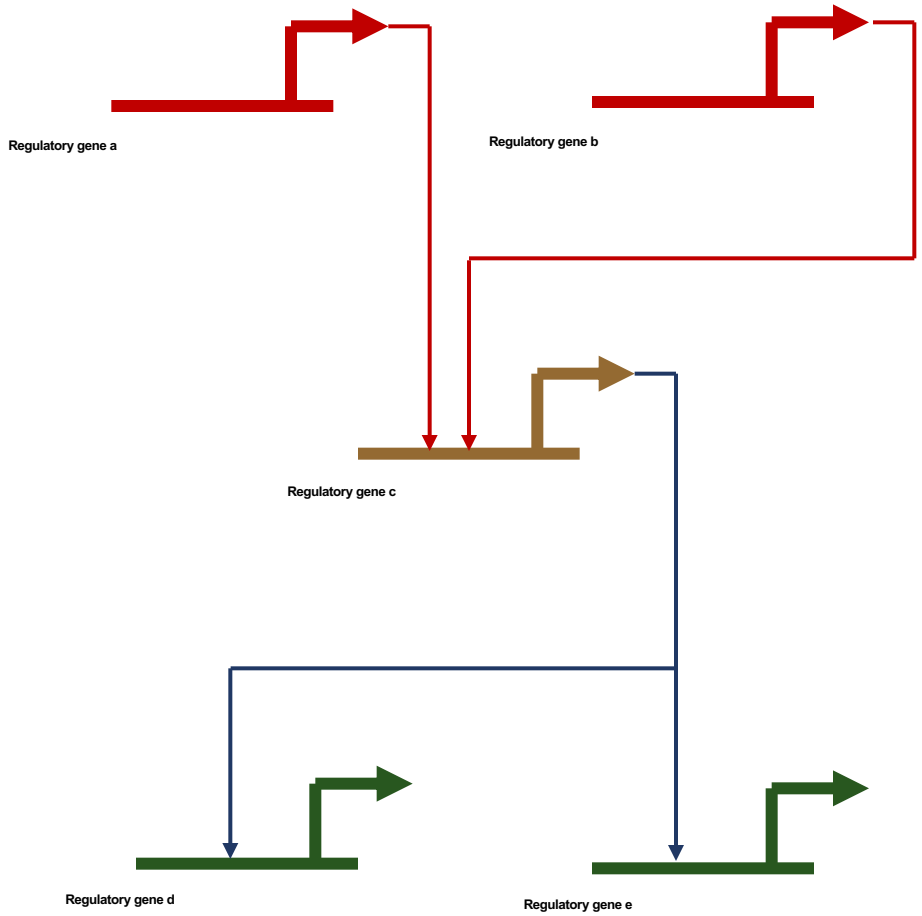
Input \ Output	*blimp1b*	*brachyury*	*brn1/2/4*	*dac*	*eve*	*foxa*	*gatae*	*hnf1*	*hox11/13b*	*krlz13*	*myc*	*otx-β*	*soxc*	*tgif*
Blimp1	-		+			+				-				+
Brachyury								+		-		+		+
Dac				-						-				
Eve	+		+				+	+	+					+
FoxA										-				
GataE			+							-		+		
Hnf1								-						
Hox11/13b	+	+				+	+	+	-					+
Krl/Z13			+											
Myc										-	-	-		
Otx			+											
SoxC													-	
Tgif														

1. Analysis of regulatory state expression: by qPCR, Nanostring, WMISH
2. Perturbation of transcription factors, analysis of change in expression of other genes
3. Cis-regulatory analysis
4. Modeling

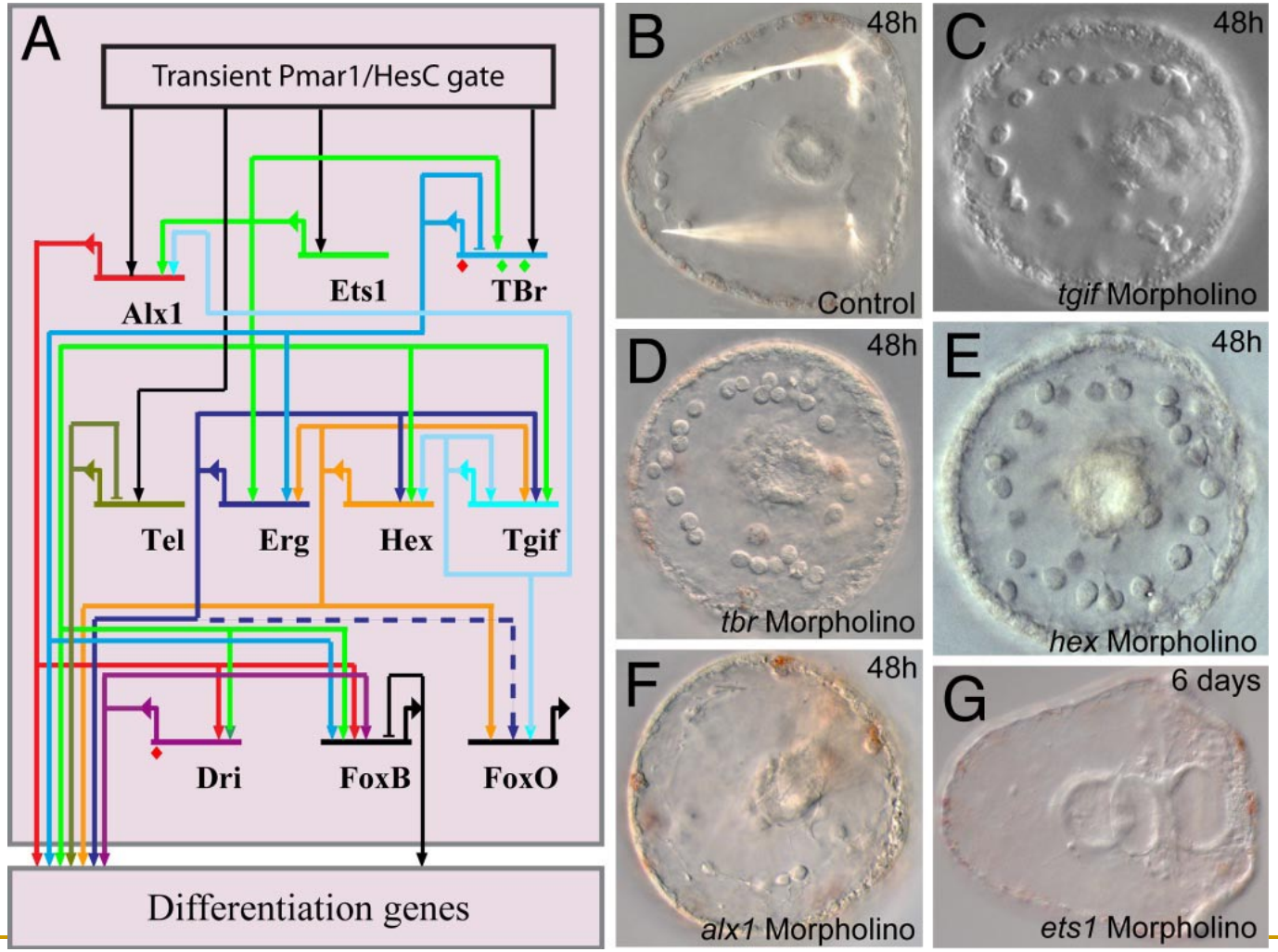
Introduction: Gene regulatory networks (Isabelle Peter)

Gene regulatory networks

- determine developmental cell fates by controlling spatial expression of regulatory states
- are composed of regulatory genes and the regulatory interactions among them
- Are encoded in the genome in regulatory genes and in *cis*-regulatory modules (regulatory interactions)
- Provide causal link between genomic sequence and development of body plan



Single gene versus network approaches (Isabelle Peter)



Oliveri et al., 2008

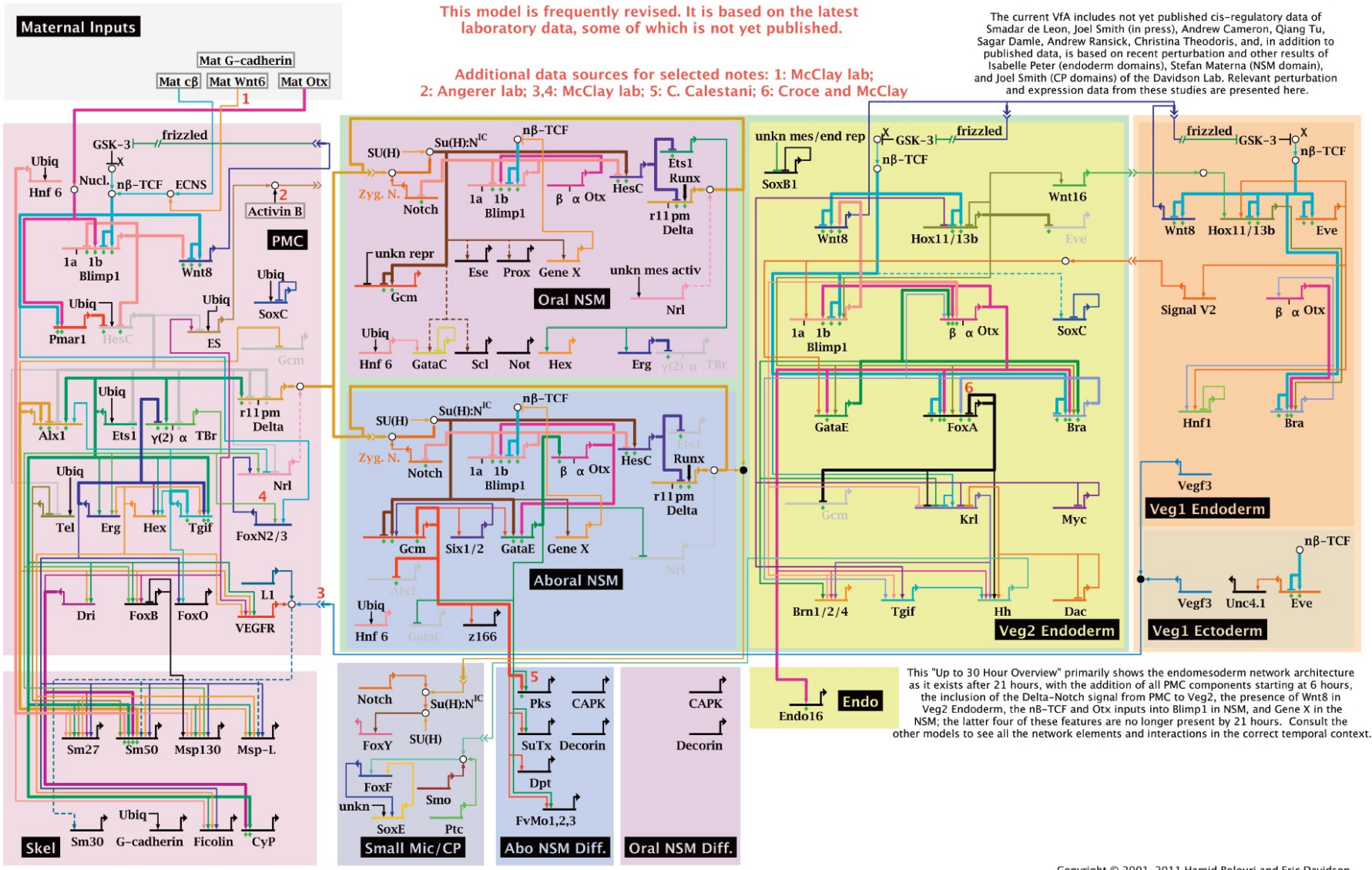
Endomesoderm Specification up to 30 Hours

September 27, 2011

This model is frequently revised. It is based on the latest laboratory data, some of which is not yet published.

Additional data sources for selected notes: 1: McClay lab; 2: Angerer lab; 3,4: McClay lab; 5: C. Calestani; 6: Croce and McClay

The current VfA includes not yet published cis-regulatory data of Smadar de Leon, Joel Smith (in press), Andrew Cameron, Qiang Tu, Sagar Damle, Andrew Ransick, Christina Theodoris, and, in addition to published data, is based on recent perturbation and other results of Isabelle Peter (endoderm domains), Stefan Materna (NSM domain), and Joel Smith (CP domains) of the Davidson Lab. Relevant perturbation and expression data from these studies are presented here.



This "Up to 30 Hour Overview" primarily shows the endomesoderm network architecture as it exists after 21 hours, with the addition of all PMC components starting at 6 hours, the inclusion of the Delta-Notch signal from PMC to Veg2, the presence of Wnt8 in Veg2 Endoderm, the nB-TCF and Otx inputs into Blimp1 in NSM, and Gene X in the NSM; the latter four of these features are no longer present by 21 hours. Consult the other models to see all the network elements and interactions in the correct temporal context.

Ubiq=ubiquitous; Mat = maternal; activ = activator; rep = repressor;
unkn = unknown; Nucl. = nuclearization; χ = β-catenin source;
nβ-TCF = nuclearized b-β-catenin-Tcf1; ES = early signal;
ECNS = early cytoplasmic nuclearization system; Zyg. N. = zygotic Notch

The Peter-Davidson Boolean Model

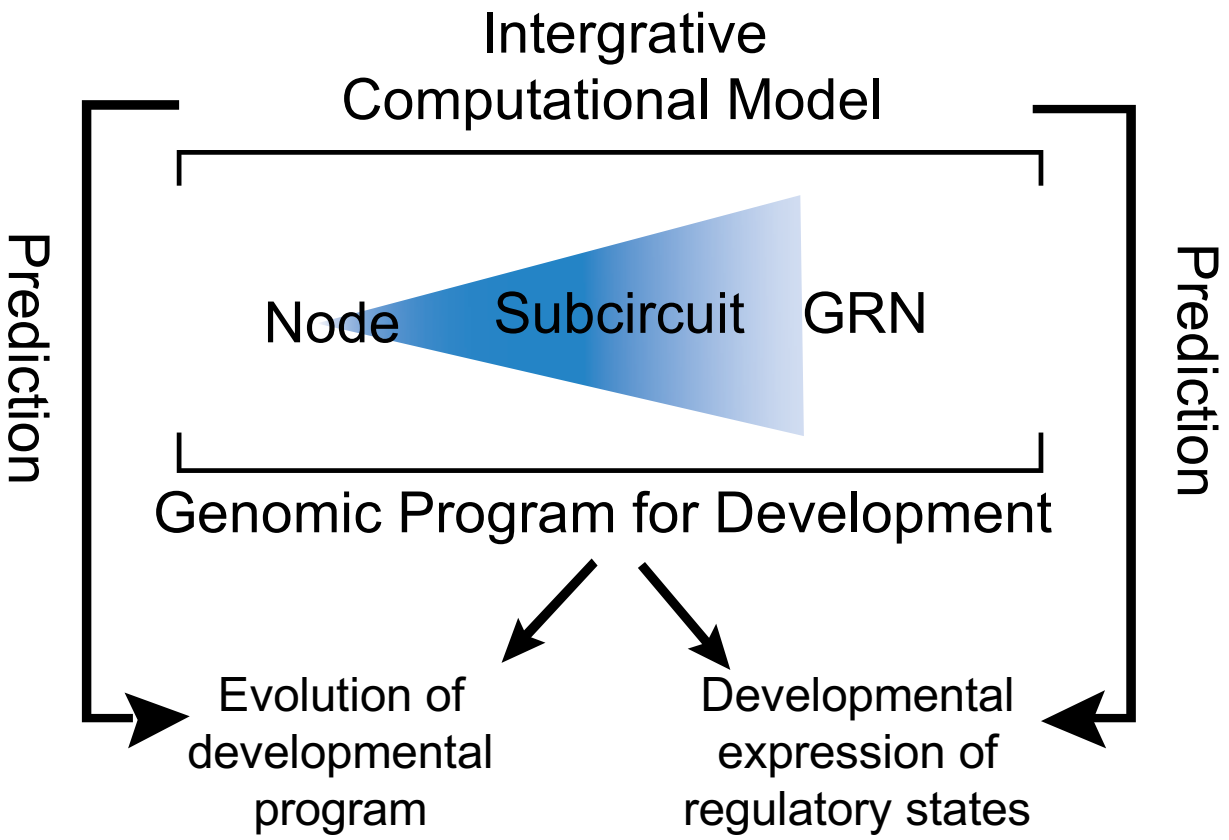
- Isabelle Peter and Eric Davidson, “Predictive Computation of Genomic Logic Processing Functions in Embryonic Development” Proceedings of the National Academy of Sciences, Vol 109, no. 41, 2012

The Peter-Davidson Boolean Model

Isabelle Peter and Eric Davidson’s book:

“Genomic Control Process: Development and Evolution”
Academic Press 2015

Isabelle Peter



Regulatory information encoded at the GRN level

Boolean modeling of Gene Regulatory Networks (Isabelle Peter)

GRN logic rules captured in cis-regulatory modules

Boolean Model: Vector Equations

GRN rules remain constant in development

Boolean Model: Vector Equations remain the same throughout computation

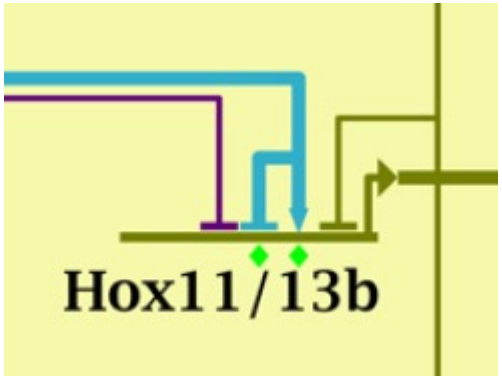
Expression of regulatory states changes in development

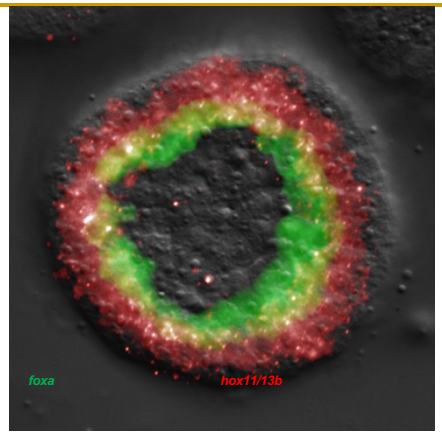
Boolean Model: Computed output of Vector Equations

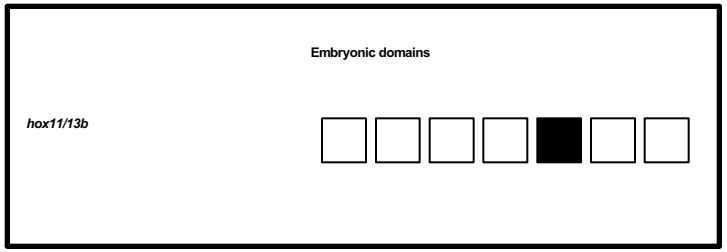
Boolean spatial gene expression (Isabelle Peter)

In the Boolean model, gene expression is represented as being

ON (=1) **if a gene is expressed at functional levels**
OFF (=0) **if a gene is not expressed**



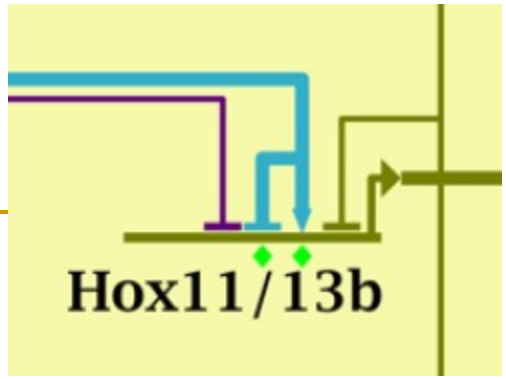




Testing the function of a large-scale GRN model (Isabelle Peter)

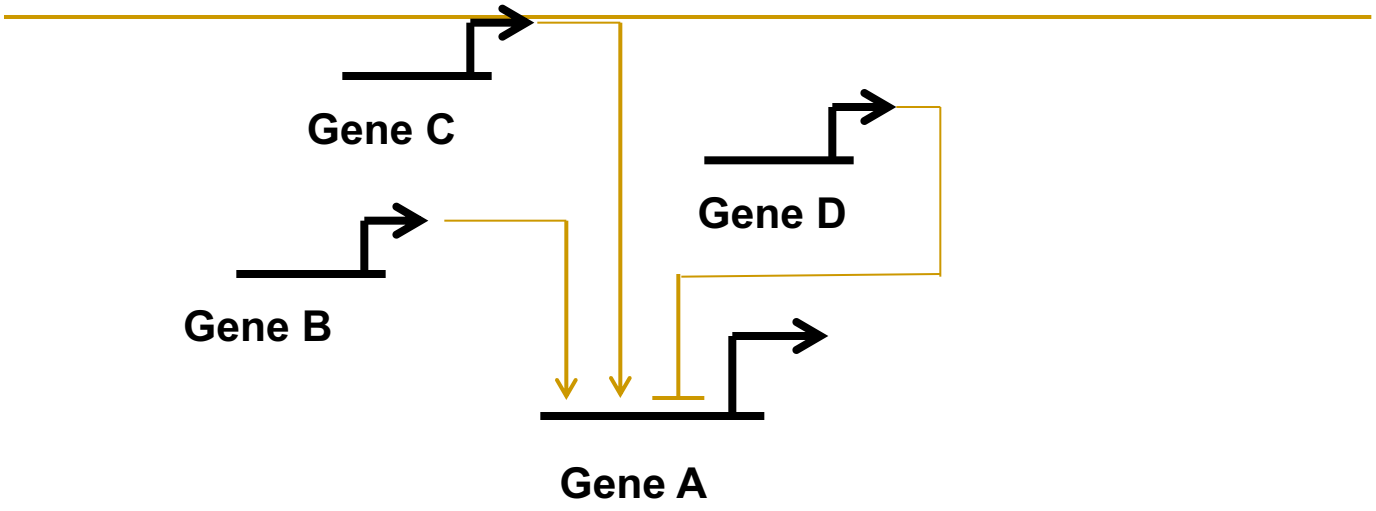
***Cis*-regulatory information for every gene in the network captured in vector equations:**

- **Type of transcription factor**
- **Activation or repression**
- **AND or OR logic**
- **Multiple *cis*-regulatory modules**
- **Logic between different modules**

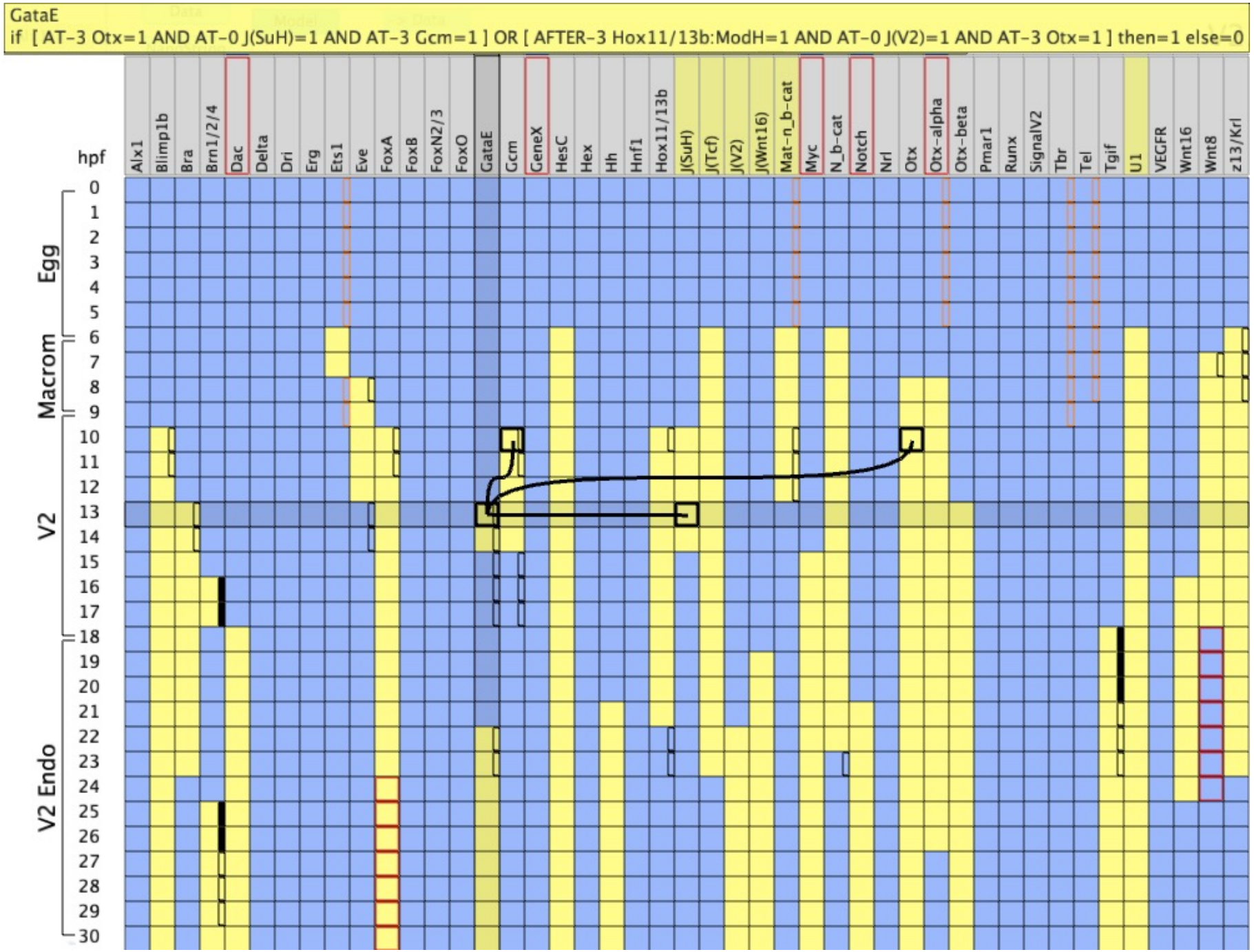


Vector Equations:
Computation of temporal dynamics (Isabelle Peter)

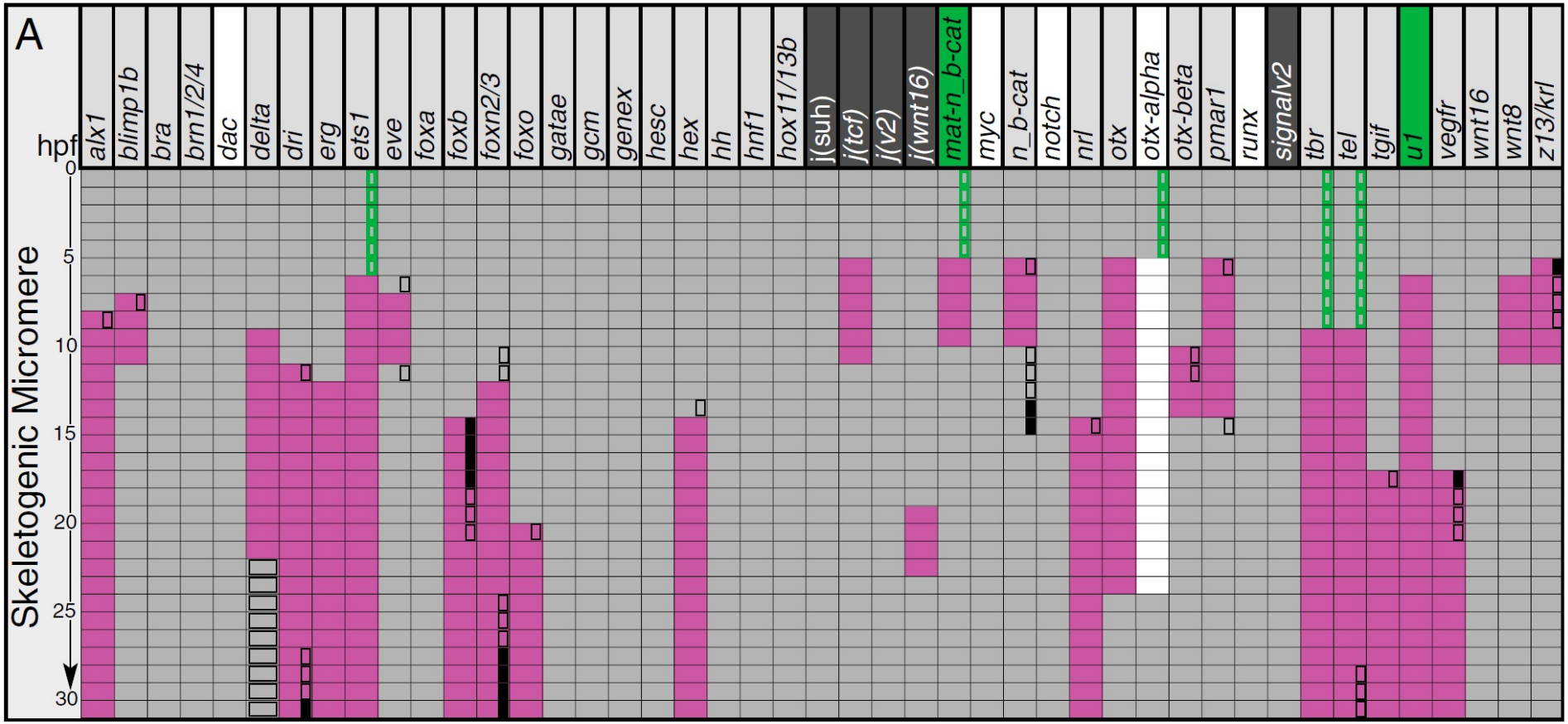
"Gene A = 1
if AT-3 Gene B = 1 AND AT-3 Gene C = 1 AND NOT AT-2 Gene D = 1,
else = 0"



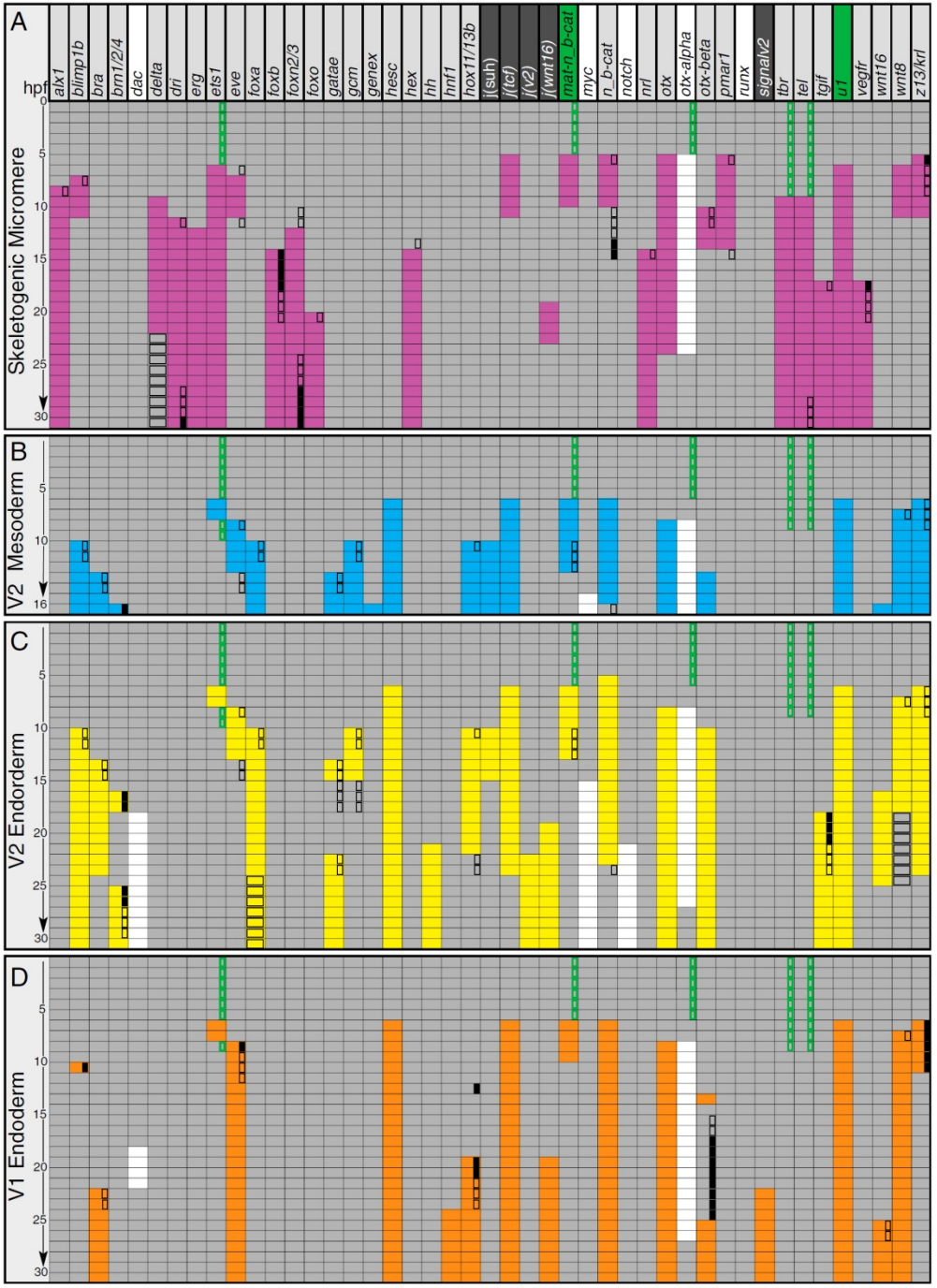
Gene		Vector Equation
alx1		if AT-2 *ets1*=1 AND AT-2 NOT *hesc*=1 then=1 else=0
blimp1b		if Mod1=1 OR Mod2=1 OR Mod3=1 then=1 else=0
	Mod1	if [AT-0 *j(tcf)*:ModH=1 OR AT-0 *j(tcf)*:ModL=1] AND PERM-0 *j(tcf)*:ModH=0 AND AT-2 *otx*=1 then=1 else=0
	Mod2	if AT-2 *hox11/13b*:ModH=1 AND AT-0 *j(tcf)*:ModH=1 AND PERM-0 *j(tcf)*:ModH=0 then=1 else=0
	Mod3	if AT-0 *j(v2)*=1 then=1 else=0
bra		if AT-3 *hox11/13b*=1 AND AT-0 *j(tcf)*=1 AND PERM-0 *j(tcf)*:ModH=0 AND AT-3 *otx*=1 AND [AT-3 *gatae*=1 OR AT-3 *gatae*=0] then=1 else=0
brn1/2/4		if AT-3 *otx*=1 AND AT-3 *blimp1b*=1 AND AT-3 *gatae*=1 AND [AT-3 *z13/krl*=1 OR AT-3 *z13/krl*=0] then=1 else=0
dac		if >17 AND IN V2 Endoderm then=1 else=0
delta		if [ModA=1 OR ModR11=1] AND AT-2 NOT *hesc*=1 then=1 else=0
	ModA	if AT-2 *runx*=1 AND AT-2 NOT *hesc*=1 then=1 else=0
	ModR11	if AT-3 *ets1*=1 then=1 else=0
dri		if AT-3 *alx1*=1 AND AT-3 *ets1*=1 then=1 else=0
erg		if AT-3 *ets1*=1 AND AT-3 *tbr*=1 OR AT-3 *hex*=1 then=1 else=0
ets1		if AT-0 *u1*=1 AND AT-2 NOT *hesc*=1 then=1 else=0
eve		if AT-2 *j(tcf)*=1 AND PERM-0 *j(tcf)*:ModH=0 AND PERM-3 [*hox11/13b*:ModH=1 AND *eve*=1] then=1 else=0
foxa		if Mod1=1 OR Mod2=1 OR Mod3=1 then=1 else=0
	Mod1	if [AT-0 *j(tcf)*:ModH=1 OR AT-0 *j(suh)*=1] AND PERM-0 *j(tcf)*:ModH=0 then=1 else=0
	Mod2	if AT-0 *j(tcf)*:ModH=1 AND AT-3 *hox11/13b*:ModH=1 AND AT-3 *otx*=1 AND [AT-3 *bra*=1 OR AT-3 *bra*=0] AND PERM-0 *j(tcf)*:ModH=0 then=1 else=0
	Mod3	if >23 AND IN V2 Endoderm then=1 else=0
foxb		if AT-3 *alx1*=1 AND AT-3 *dri*=1 AND AT-3 *ets1*=1 AND AT-3 *tbr*=1 then=1 else=0
foxn2/3		if AT-3 *tbr*=1 then=1 else=0
foxo		if AT-3 *tgif*=1 AND AT-3 *erg*=1 then=1 else=0
gatae		if [AT-3 *otx*=1 AND AT-0 *j(suh)*=1 AND AT-3 *gcm*=1] OR [AFTER-3 *hox11/13b*:ModH=1 AND AT-0 *j(v2)*=1 AND AT-3 *otx*=1] then=1 else=0
gcm		if ModE=1 OR ModG=1 then=1 else=0
	ModE	if AT-0 *j(suh)*=1 AND NOT *alx1*=1 AND PERM-0 *j(suh)*=0 then=1 else=0
	ModG	if AT-2 *gcm*=1 AND AT-2 NOT *foxa*=1 then=1 else=0
genex		if >15 AND AT-0 *j(suh)*=1 then=1 else=0
hesc		if [AT-0 *u1*=1 OR AT-0 *j(suh)*=1] AND PERM-0 *pmar1*=1 then=1 else=0
hex		if AT-2 *erg*=1 AND [AT-2 *ets1*=1 OR AT-2 *tgif*=1] then=1 else=0
hh		if AT-3 *dac*=1 AND AT-3 *foxa*=1 AND AT-3 *tgif*=1 AND AT-3 *otx*=1 AND [AT-3 *z13/krl*=1 OR AT-3 *z13/krl*=0] then=1 else=0
hnf1		if AT-2 *bra*=1 AND AT-2 *eve*=1 then=1 else=0
hox11/13b		if ModH=1 OR ModW=1 then=1 else=0
	ModH	if AT-0 *j(tcf)*:ModH=1 AND NOT [AT-3 *j(wnt16)*=1 AND AT-3 *hox11/13b*:ModH=1] AND


GataE
if [AT-3 Otx=1 AND AT-0 J(SuH)=1 AND AT-3 Gcm=1] OR [AFTER-3 Hox11/13b:ModH=1 AND AT-0 J(V2)=1 AND AT-3 Otx=1] then=1 else=0
hpf
Egg
Macrom
V2
V2 Endo
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
Alx1
Blimp1b
Bra
Brn1/2/4
Dac
Delta
Dri
Erg
Ets1
Eve
FoxA
FoxB
FoxN2/3
FoxO
GataE
Gcm
GeneX
HesC
Hex
Hh
Hnf1
Hox11/13b
J(SuH)
J(Tcf)
J(V2)
J(Wnt16)
Mat-n_b-cat
Myc
N_b-cat
Notch
Nrl
Otx
Otx-alpha
Otx-beta
Pmar1
Runx
SignalV2
Tbr
Tel
Tgif
U1
VEGFR
Wnt16
Wnt8
z13/Krl

Comparison model to data: skeletogenic micromeres



COMPUTATION OF
HOURLY SPATIAL
EXPRESSION vs DATA: OUT OF 2772
SPACE/TIME EXPRESSION DOMAINS VERY FEW
SIGNIFICANT DISCREPANCIES…



Peter, Faure, Davidson, 2012 PNAS

Conclusions (Isabelle Peter)

The Boolean model shows that the endomesoderm GRN contains sufficient regulatory information to recapitulate developmental gene expression

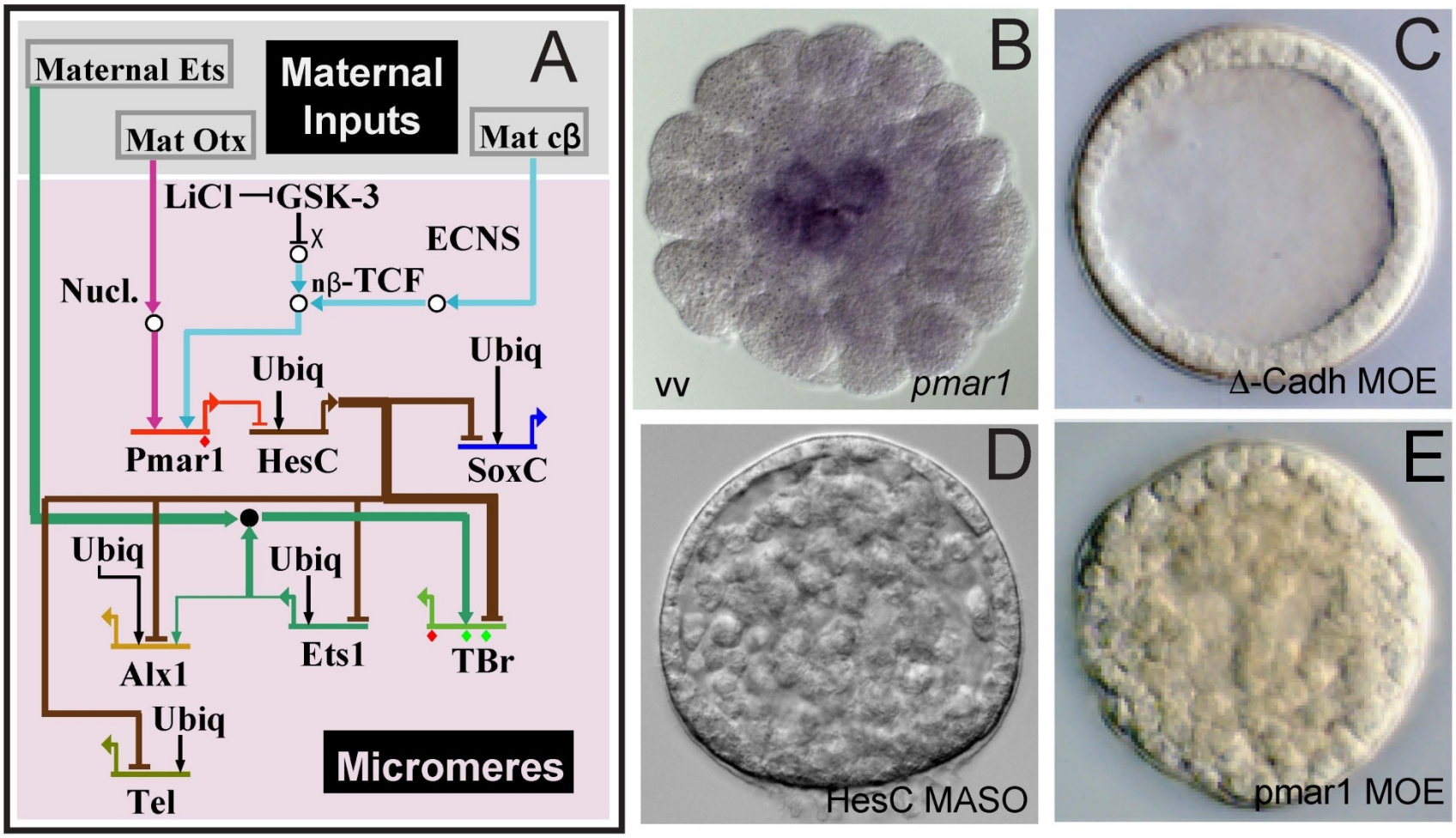
Almost complete experimental analysis of pre-gastrular GRNs

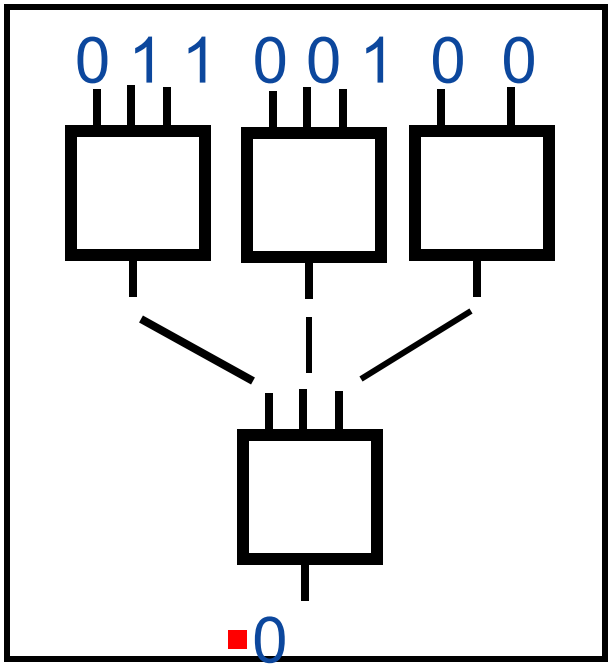
The Boolean model demonstrates that integration of *cis*-regulatory information produces correct system-level output

A proof of the GRN concept

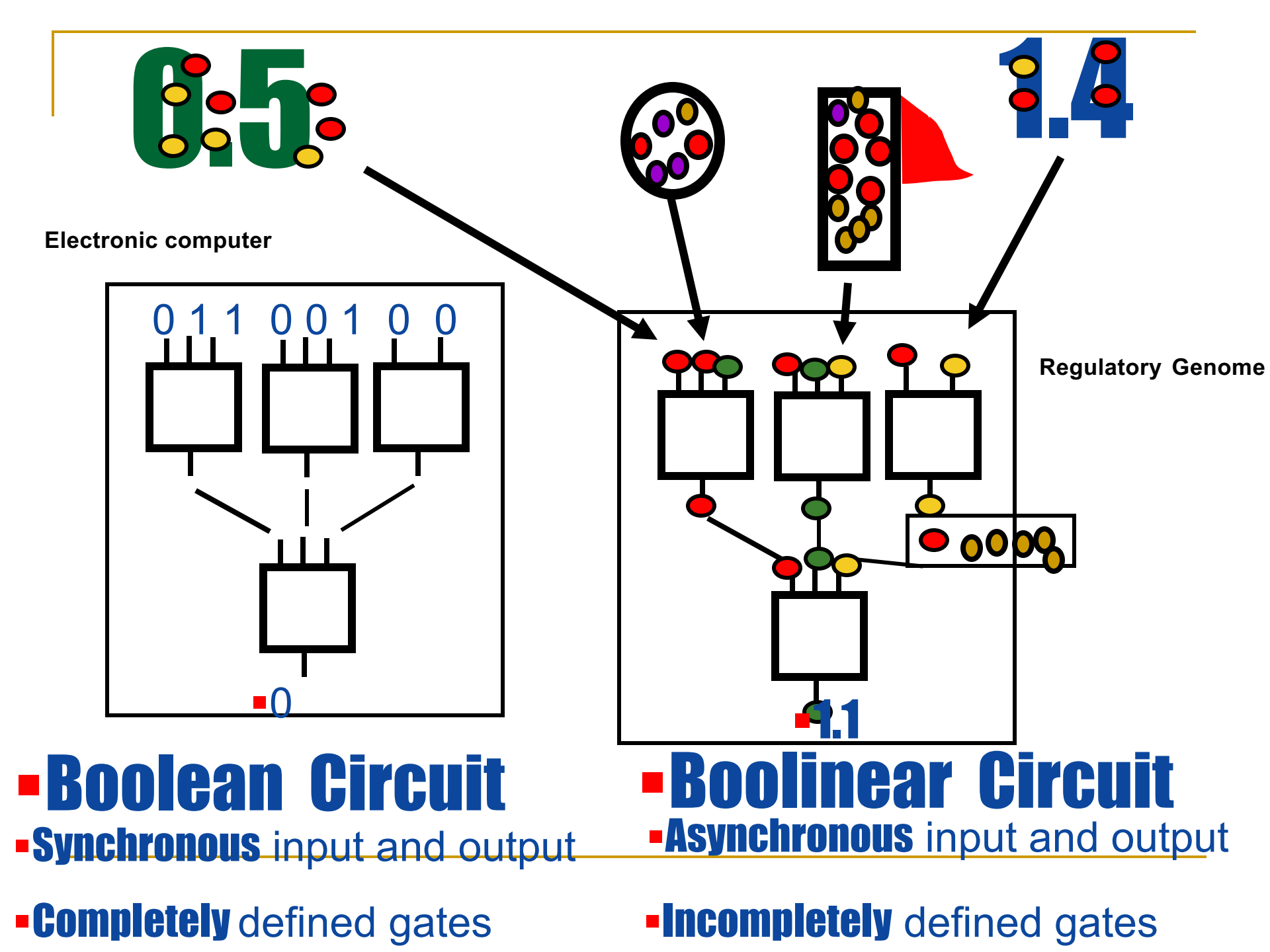
=>how does the model respond to perturbation?

Isabelle Peter



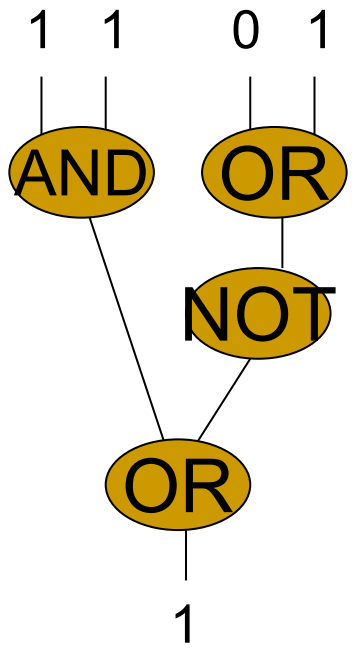


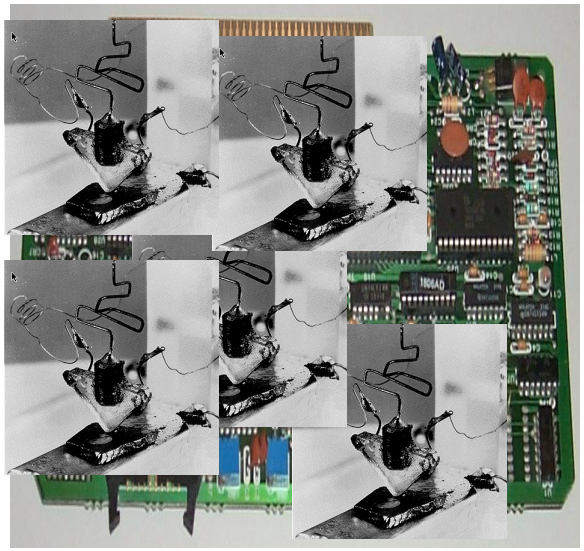
- **Boolean Circuit**
  - **Synchronous** input and output
  - **Completely** defined gates



- **Boolean Circuit**
  - **Synchronous** input and output
  - **Completely** defined gates
- **Boolinear Circuit**
  - **Asynchronous** input and output
  - **Incompletely** defined gates

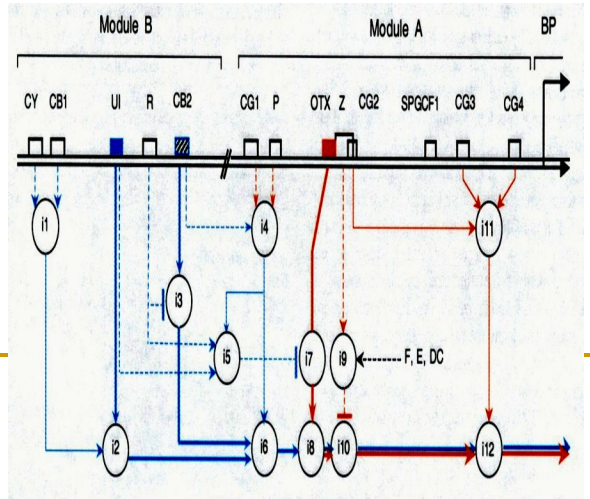
```
IF (x1 = 1
    AND
     x2= 1)
THEN
      .....
```

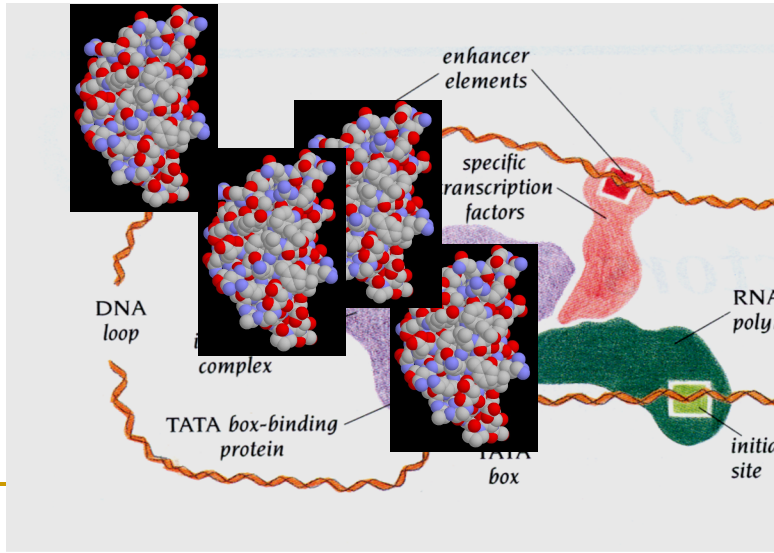




Software code **Logic workflow** **Hardware architecture**

GTAGGATTAAG
……
CATCCTAATTC
…….
GTATCTAGAAG
…….

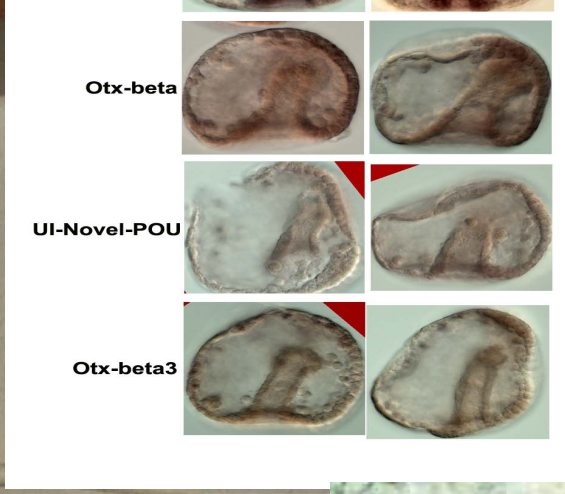




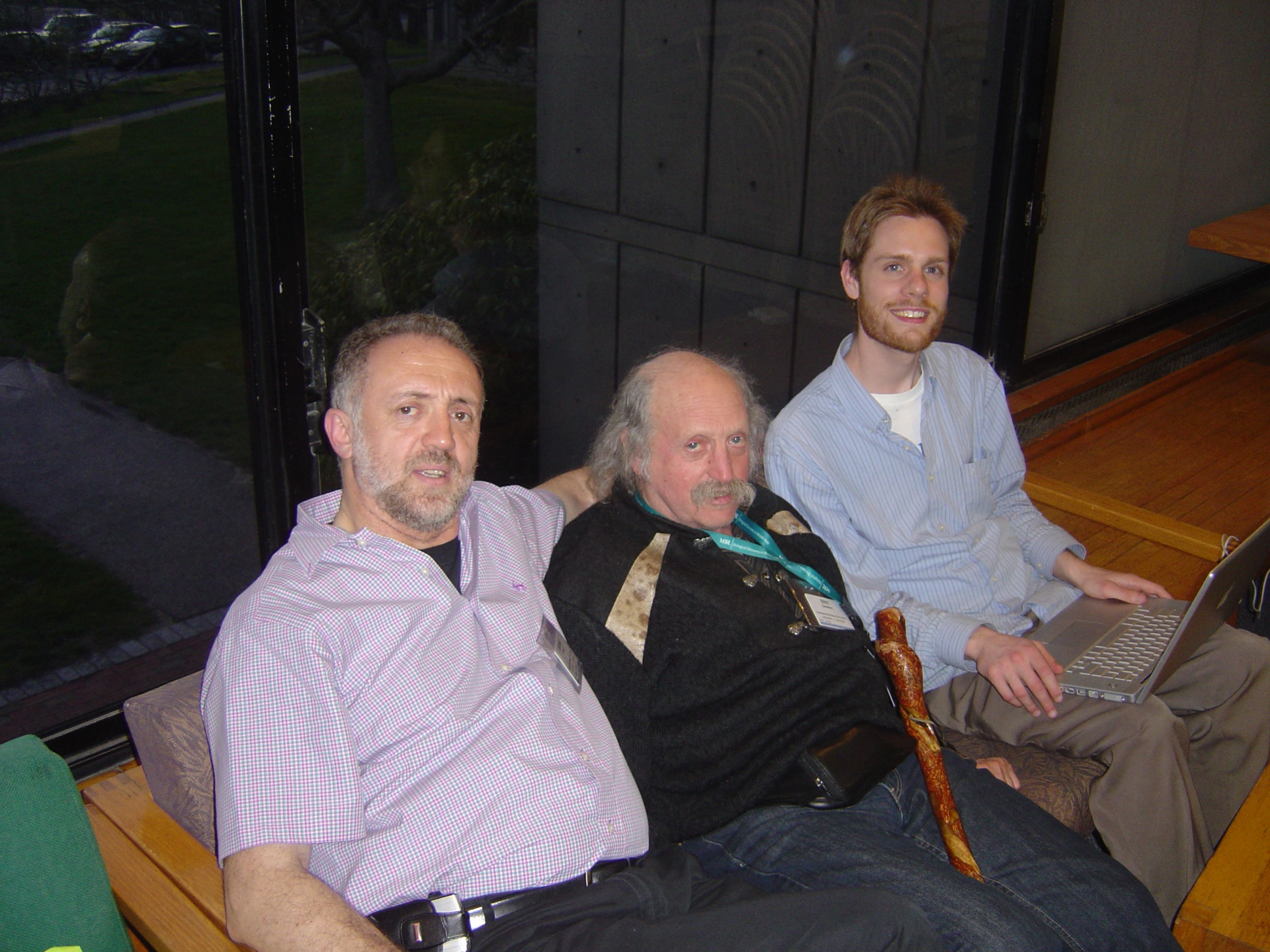
Eric Davidson's Axioms

- Axiom 1. "Serious people do things seriously"
- Axiom 2. "The answer to how to recognize the right place to crawl through the wall of thorns lies in the shape of the mental triangle composed of intuition, current knowledge, and logic"
- Axiom 3. "Logic Rules!"
- Axiom 4. "If we care about anything outside our current events, we need to care about those who will carry the brilliant torch forward"
- Axiom 5. "Life is short: Let's keep in touch!"
- Axiom 6. "Use it or lose it"
- Axiom 7. "Have inexhaustible optimism, inexhaustible curiosity, inexhaustible energy and inexhaustible honesty!" (Davidson, 2012)
- Axiom 8. And in the end, the love you take is equal to the love you make.

Future Work



Caltech, Davidson Lab
October 2004




SEQ ID
COORD
Resource Comp
BIOS


INVERTEBRATES
INVERTEBRATES
FOSSIL INVERTEBRATES








GASTRULATION