JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 26, Number 7, 2019
Mary Ann Liebert, Inc.
Pp. 653–684
DOI: 10.1089/cmb.2019.0144

Eric Davidson's Regulatory Genome for Computer Science: Causality, Logic, and Proof Principles of the Genomic *cis*-Regulatory Code

SORIN ISTRAIL

I think that it is a relatively good approximation to truth which is much too complicated to allow anything but approximations that mathematical ideas originate in empirics. But, once they are conceived, the subject begins to live a peculiar life of its own and is governed by almost entirely aesthetical motivations. In other words, at a great distance from its empirical source, or after much ''abstract'' inbreeding, a mathematical subject is in danger of degeneration. Whenever this stage is reached the only remedy seems to me to be the rejuvenating return to the source: the reinjection of more or less directly empirical ideas.—John von Neumann (1947).

Development of Western science is based on two great achievements: the invention of the formal logical system (In Euclidean geometry) by the Greek philosophers and the discovery of the possibility to find out causal relationships by systematic experiment (during Renaissance).—Albert Einstein (1953)

Considering evolution of body plans in terms of network circuitry, as a history of assembly of grades of network organization, is to transform this vexed subject into a prospectus for laboratory research.—Eric Davidson (2006)

Causality isn't everything, it's the only thing.—(a paraphrase on football coach Vince Lombardi's quote)

1. ABSTRACT

In this article, we discuss several computer science problems, inspired by our 15-year-long collaboration with Prof. Eric Davidson, focusing on computer science contributions to the study of the regulatory genome. Our joint study was inspired by his lifetime trailblazing research program rooted in causal gene regulatory networks (GRNs), system completeness, genomic Boolean logic, and genomically encoded regulatory information. We present first four inspiring questions that Eric Davidson asked, and the follow-up, namely, seven technical problems, fully or partially resolved with the methods of computer science. At the center, and unifying the intellectual backbone of those technical challenges, stands ''Causality.'' Our collaboration produced the causality-

Department of Computer Science, Center for Computational Molecular Biology, Brown University, Providence, Rhode Island.

inferred cis*GRN-Lexicon* database, containing the *cis*-regulatory architecture (CRA) of 600+ transcription factor (TF)-encoding genes and other regulatory genes, in eight species: human, mouse, fruit fly, sea urchin, nematode, rat, chicken, and zebrafish. These CRAs are causality-inferred regulatory regions of genes, derived experimentally through the experimental method called "*cis*-regulatory analysis" (also known as the "Davidson criteria"). In this research program, causality challenges for computer science show up in two components: (1) how to define data structures that represent the causality-inferred, by the Davidson criteria, DNA structure data and to define a versatile software system to host them; and (2) how to identify by automated software for text analysis the experimental technical articles applying the Davidson criteria to the analysis to genes. We next present the cis*GRN-Lexicon* Meta-Analysis (Part I). We conclude the article with some reflections on epistemology and philosophy themes concerning the role of causality, logic, and proof in the emerging elegant mathematical theory and practice of the regulatory genome.

It is challenging to explain what "explanation" is, and to understand what "understanding" is, when the technical task is to "prove" system-level causality completeness of a 50-gene causal GRN. Within the Peter-Davidson Boolean GRN model, the Peter-Davidson completeness "theorem" provides a seminal answer: *Experimental causality system completeness = Computational exact prediction completeness*.

The article is organized as follows. Section 2 is dedicated to our Prof. Eric Davidson. Section 3 gives a brief introduction for computer scientists to the regulatory genome and its information processing operations in terms similar to the electronic computer. Section 4 proposes to honor Eric Davidson's life-long scientific work on the regulatory genome by naming a most fundamental time unit constant after him. Section 5 presents four grand challenge questions that Eric Davidson asked, and seven follow-up problems inspired by the first two questions, which we fully or partially solved together. Central to the mentioned solutions is our construction of the *cis*GRN-Lexcion, the database of causally inferred CRA of 600+ regulatory genes in eight species. Section 6 presents Part I of the *cis*GRN-Lexcion Meta-Analysis, coached as "rules" of the genomic *cis*-regulatory code. Section 7 is devoted to reflections on epistemological and philosophical themes: causality, logic, and proof in the elegant mathematical modeling of the regulatory genome. We present here the "Davidsonian Causal Systems Biology Axioms," which guide us toward understanding of the meaning of "proving" causality completeness, for a complex experimental system, by exact computational predictions.

Keywords: causal systems biology, causality, Eric Davidson's regulatory genome, gene regulatory networks, genomic *cis*-regulatory code.

2. ERIC DAVIDSON OUR PROFESSOR AND MENTOR

In his seminal books *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution* (Davidson, 2006) and *Genomic Control Process: Development and Evolution* (coauthored with Isabelle Peter) (Peter and Davidson, 2015), Eric Davidson, the foremost experimentalist of regulatory genomics, who passed away in 2015, forcefully reminds us that in the scientific method, *causality* is essential; all other approaches are just distractions. In contrast, Davidson, a notoriously elegant writer, offers devastating criticism of the posterior biology approaches all too impatiently employed today: measure first expression of thousands of genes and then computationally infer biology. The past century's luminaries of mathematical statistics taught us in no uncertain terms that causality cannot be inferred from statistical tables. Davidson aligns with them, adding to their argument a practical dose of reality. The exquisite regulatory mechanisms, *locked down by evolution*, can only be revealed through systematic experimental perturbations. In the absence of the ocean deep prior biology knowledge, no amount of clustering statistics, or other skinny deep dives, would be able to infer causality in biology. Like his mentor Max Delbruck, and with the sea urchin genome in hand, Eric Davidson becomes the leading liberator of quantitative principles of cell regulation, trapped in the qualitative descriptive world of biology without genomic sequence.

Prof. Davidson's legacy consisted of 400+ articles and he mentored ~300 PhDs, postdocs, and faculty in his laboratory in the Division of Biology at California Institute of Technology. He also authored six books; only the last one was coauthored, with Isabelle Peter, one of his most important research collaborators. In the Davidson Lab, she led several of the major experimental GRN projects, and coauthored the seminal Peter-Davidson Boolean GRN Equations Model (Peter et al., 2012) covered in full detail in their book *Genomic Control Process: Development and Evolution* (Peter and Davidson, 2015). Our beloved teacher and mentor Eric Davidson united us biologists, physicists, biochemists, engineers, mathematicians, and computer scientists, like in his CalTech Laboratory in a research renaissance movement toward the quest for the functional meaning of DNA. From such research will ultimately come, by experimental demonstration, the revelation of the much-sought laws of regulatory biology.

3. THE REGULATORY GENOME IS A "COMPUTER": BRIEF PRIMER FOR COMPUTER SCIENTISTS

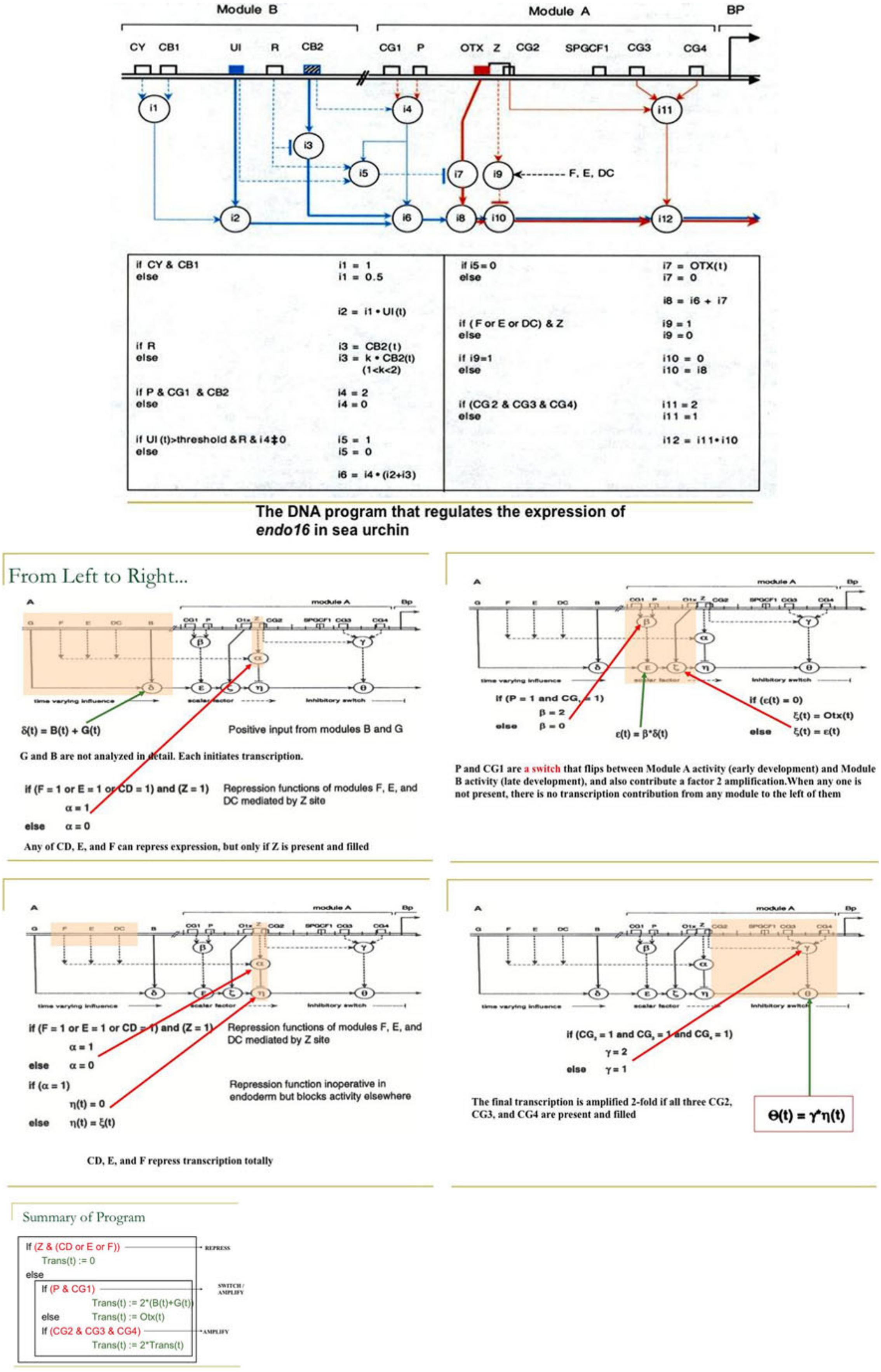
The regulatory genome is a complex information processing system (Istrail et al., 2007). The way it ''computes'' is through a symbiosis of analog and digital principles similar to electronic computers. We exemplify the type of ''computation'' performed in the regulatory genome through a single gene and its information processing structure, the *endo16* gene (Fig. 1), the gene that has a vital role in the gut of the sea urchin. Its information processing architecture and the pseudo code of its logical operation were identified by Yuh et al. (1998), and this complete causal explanation of mechanism is one of the towering achievements of the regulatory genome. In the regulatory genome, mathematical logic, especially Boolean logic, emerges naturally in a variety of fundamental components. We present here how the standard Boolean logic gates are defined with their biochemistry structures, abstracted, as in computer science, to input–output (Fig. 2). We then show how the regulatory genome does information processing using its regulatory genes, called TFs, and the overall workflow of genomic import that makes computation possible. Section 7.3 presents further analogies of the regulatory genome computing principles with respect to those of electronic computer.

Here are the basic biological components that implement the ''computation'' apparatus. The genetic code explains *how* to synthesize a protein, but it does not explain *when*. This is of critical importance, because every cell contains a copy of the entire genome of that organism—cells in the eye contain not only instructions on how to make the eye but also how to make the heart as well (and vice versa). However, cells in the eye ''know'' which orders they should be carrying out; they ''know'' they are part of the eye so they only develop the parts of the eye, not parts of the heart. The process of determining which proteins are synthesized and which are not is called *gene regulation*. There are many types of gene regulation; we focus here on a particular one called *cis*-regulation.

cis-Regulation is caused by TFs binding to specific sequences in the genome that trigger the activation or repression of gene expression. TFs are a specific type of protein, so they themselves are encoded by genes, which must be regulated by other TFs. These chains of regulation yield a graph or network, known as a GRN. Unlike the rules governing the translation of genetic material into protein, which have been understood well for nearly 50 years, the rules governing gene expression are still only conceived of at the level of general principles. It remains difficult not only to determine the effect a given region has on gene expression, but also even to recognize a *cis*-regulatory region at all.

The locations where TFs bind are called transcription factor binding sites (TFBSs). These binding sites are rarely found in sequence that encodes proteins. More often, they are found in noncoding sequence near the gene that they regulate, which is then referred to as the *cis*-regulatory region of that gene. TFBSs are not found scattered uniformly throughout these regions; usually they are found in clusters called *cis*-regulatory modules (CRMs). These clusters are called modules not only because they look modular—in fact their function is modular too. Each CRM performs a specific self-contained regulatory function, which its TFBSs work together to carry out. The *cis*-regulatory region of the gene *endo16*, for example, contains six CRMs, named Modules A, B, DC, E, F, and G (Fig. 3). Module A ensures that the gene is expressed initially during development, Module B causes activation at a later stage, Module DC represses expression in a region where the gene should not be expressed, and the other modules perform similar but independent regulatory functions. One can even see that Module DC contains two clusters of TFBSs, but these two clusters comprise one single CRM, not two. It is the function that defines the boundaries of a module, not the locations of its TFBSs.

Each individual TFBS performs a particular regulatory function, whether activation, repression, or any of the other possible functions that we will introduce later. The regulatory function of a CRM is a combination of the individual TFBS functions. The rules of how the overall function is represented in DNA comprise the *cis*-regulatory code. To crack this code, we need three types of knowledge: (1) identification of binding



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)		

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

From Left to Right...

δ(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)
α = 1
else α = 0

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

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

if (P = 1 and CG$_1$ = 1)
β = 2
else β = 0

ε(t) = β*δ(t)

if (ε(t) = 0)
ξ(t) = Otx(t)
else ξ(t) = ε(t)

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)
α = 1
else α = 0
if (α = 1)
η(t) = 0
else η(t) = ξ(t)

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

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)
γ = 2
else γ = 1

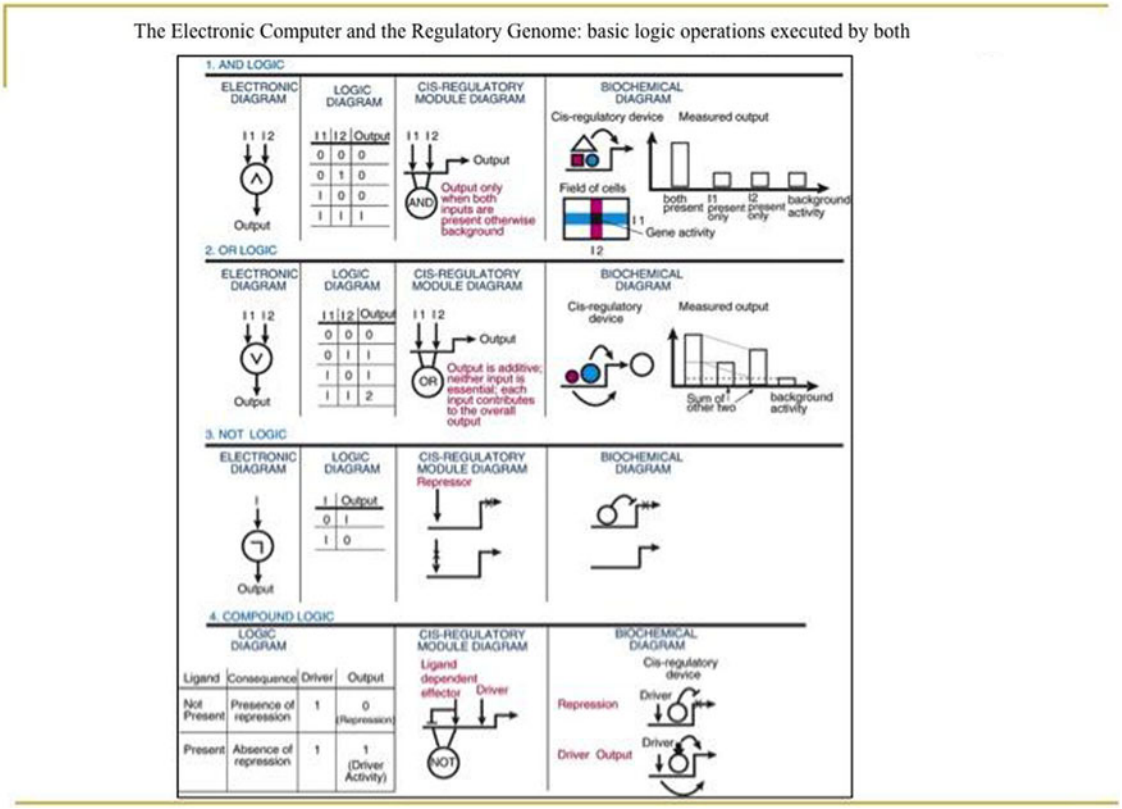
The final transcription is amplified 2-fold if all three CG2, CG3, and CG4 are present and filled

Θ(t) = γ*η(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)
```

FIG. 1. *cis*-Regulatory information processing by the endo16 gene, the *First Gene* of the regulatory genome.



I1	I2	Output
0	0	0
0	1	0
1	0	0
1	1	1

I1	I2	Output
0	0	0
0	1	1
1	0	1
1	1	2

I	Output
0	1
1	0

Ligand	Consequence	Driver	Output
Not Present	Presence of repression	1	0 (Repression)
Present	Absence of repression	1	1 (Driver Activity)

FIG. 2. Boolean logic gates and their biochemistries.

sites and their TFs (the *Identification Problem*); (2) interpretation of individual site functions (the *Interpretation Problem*); and (3) rules for combining the individual functions to infer overall output (the *Combining-Rules Problem*) (Davidson, 2006).

> *‘‘***Real-Time Kinetics of Sea Urchin Embryo Development.** *The basic metric of regulatory progress in development is the time interval between activation of a given regulatory gene and the activation of an immediate downstream target regulatory gene. This interval, which we shall term the step time, is a function of the basic kinetics of the molecular processes of transcription, RNA turnover, protein synthesis and turnover, transcription factor-DNA interaction, and cis-regulatory activity. In the sea urchin embryo, which develops at 15 C, biosynthetic processes are much slower than, for example, in Drosophila, and it requires several hours for successive changes of regulatory states to occur. In 2003, Bolouri and Davidson (5) modeled these kinetics for sea urchin embryos living at 15 C, using a large set of kinetic parameters previously measured for this system. This model was based on a first-principles treatment of cis-regulatory occupancy, a probabilistic mathematical argument that the rate of transcription relative to the measured maximum possible rate depends on the cis-regulatory occupancy, and standard synthesis/turnover kinetics (*Fig. S6*). The step time that emerged was* $\sim 3\,h$*. Many subsequent direct observations on S. purpuratus embryos made in the course of GRN analysis confirmed that this canonical computation approximates reality for specific cases (e.g., refs. 2, 4).’’* (Peter et al., 2012)

4. A MODEST PROPOSAL: A REGULATORY GENOME CONSTANT *ONE DAVIDSONIAN* 1*D**

The regulatory genome operates through its genes, the TF encoding genes, which operate in specific time intervals. The entire system is a distributed asynchronous system; it is timed in the sea urchin at

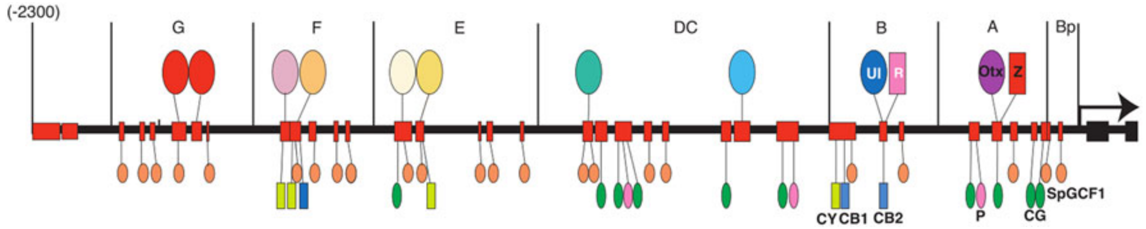


FIG. 3. *cis*-Regulatory region of *endo16* [from Yuh et al. (2001)]. The red rectangles represent the locations of TFBSs. The ovals and rectangles linked to the sequence (thick line) by thin lines represent the TFs. TFs above the sequence are those that bind uniquely in a single region of the sequence; TFs below bind in multiple locations. TF, transcription factor; TFBSs, transcription factor binding sites.

3 hours, a constant that represents *the time from gene activation to activation of an immediate downstream target regulatory gene*. The constant is universal to all animals, but the value of the constant is different from animal to animal. This time constant became fundamental for the Peter-Davidson Boolean GRN Equations model of sea urchin and its embodiment, mathematically, as a computational engine. See the ‘‘Real-time kinetics of sea urchin embryo development’’ box for more details.

We propose to call ‘‘the time step’’: one **Davidsonian**, and denote it by *D**.

Surely, the star exponent in the notation is meant to resemble an impressionistic sea urchin!

5. ERIC DAVIDSON’S CHALLENGES TO US: FOUR BIG QUESTIONS

We will present the four grand-challenge questions asked by Eric Davidson that inspired our work of the past 10 years. We solved Questions 1 and 2. Question 3 was solved by the seminal Peter-Davidson Boolean GRN Model. Question 4 remains a wide open problem, essentially of experimental import, with major impact toward GRN reengineering.

5.1. *Question 1: Causality-Based* cis-*Regulatory Information*

> *Comparatively speaking, the diversity or complexity of the genomic regulatory code will be significantly less than the diversity of the biochemical operations that execute each type of function (called a ‘‘code’’). Tackling the genomic regulatory code head on is liable to be a more direct avenue to learning what it says than by 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 CRM repertoire*. (Istrail and Davidson, 2005).

Question 1. What is *cis*-regulatory information? What is the nature of the *cis*-regulatory evidence? What is the nature of the causal explanation of gene expression? What is the causal-based CRA of regulatory genes (GRN-genes), and of off-network genes (off-GRN genes)? What are the rules of the genomic *cis*-regulatory code?

Several other related questions are as follows. How can we represent the *cis*-regulatory evidence in models (computational representations)? How to represent the DNA structure of the regulatory regions of genes annotated with the CRA structural motifs? What are the ‘‘logic functions’’ of the information processors that are the CRMs and how are they ‘‘hardwired’’ in the CRM’s DNA? How do we represent the regulatory genome in a *cis*GRN-Browser devoted to regulatory genomics? How to represent GRNs in BioTapestry views and the CRAs of the *cis*GRN-Browser in an integrated way into the BioTapestry-*cis*GRN-Browser Logic Map?

The *cis*-regulatory analysis experimental procedure, also known as the Davidson criteria: the nature of the causal *cis*-regulatory evidence

The spgcm gene of *Strongylocentrotus purpuratus* species of sea urchin was shown to be required for pigment cell specification (Ransick and Davidson, 2006). The experimental steps that revealed this causal relationship give a good introduction to the nature of *cis*-regulatory evidence. The glial cells missing regulatory gene (spgcm) is the target of the Delta/Notch (i.e., Delta/Notch binds to the regulatory region of spgcm) signaling required for specification of the mesodermal precursor of pigment cells. Microinjection of an spgcm antisense morpholino oligonucleotide results in larvae without pigment cells, thus confirming that the spgcm gene is required for pigment cell specification. Three CRMs have been identified. When they are incorporated in expression constructs, they recapitulate the early expression patterns of this gene, but the expression of such constructs is severely disrupted when coexpressed with the suppressor dn-Su(H). This confirms that spgcm is a direct target of canonical N signaling mediated through Su(H) inputs. In a series of *cis*-regulatory analyses by mutation of consensus Su(H) sites, a conserved motif paired site was identified. This was in the middle of a module and functions by driving expression in a certain spatial domain (SMC precursor) that received the Delta signal, but also repressing expression in ectopic (i.e., other) domains that lack this signal. With respect to its logic function, these sites of the Su(H) provide the CRA of a on–off switch.

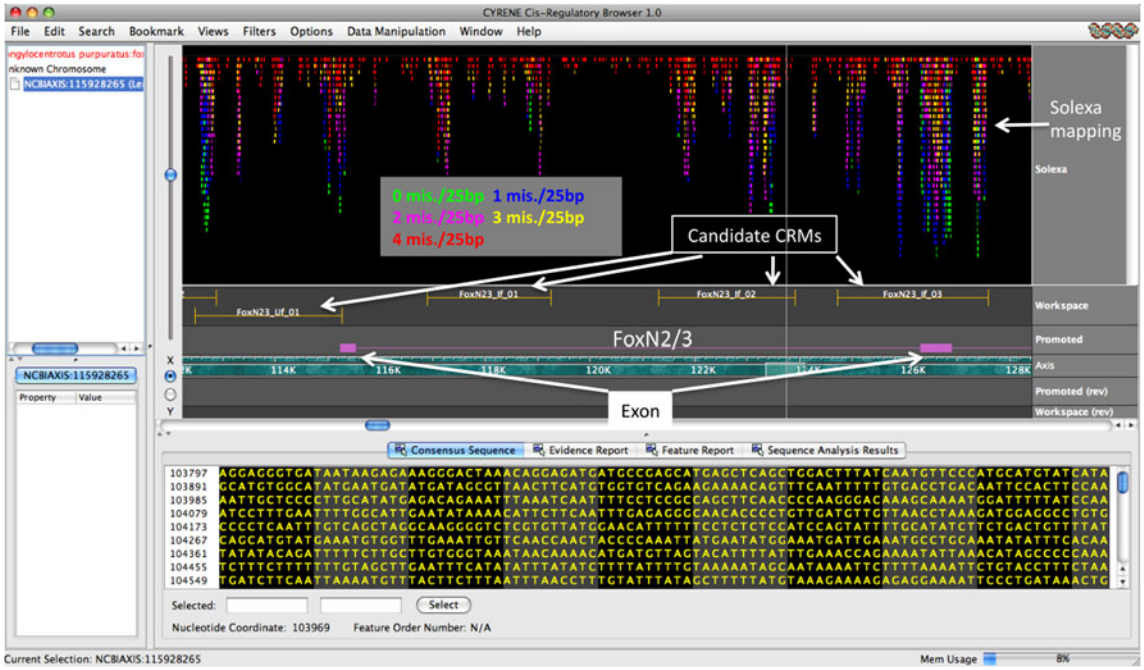
BioTapestry is a Davidson Lab software system designed around the concept of a developmental network model, and is intended to deal with large-scale models with consistency and clarity. It is capable of representing systems that exhibit increasing complexity over time, such as the genetic regulatory network controlling endomesoderm development in sea urchin embryos (Davidson et al., 2005; Longabaugh, 2012).

Problem 1. Building the cisGRN database of genomic cis-regulatory information: the ***cis*GRN-Lexicon** *is hosted in the* ***cis*GRN-Browser** *a full genome browser dedicated to the regulatory genome.*

The database is capturing the complex nature of the *cis*-regulatory evidence from many kinds of regulatory information across technologies (Tarpine, 2012). The database contains the following components: genomic structure and organization, comparative genomics, *cis*-regulatory analysis of expression constructs and mutational analysis, spatial expression profiles, and logic functions of the genomic *cis*-regulatory code. The *cis*GRN-Lexicon is a database of CRAs of regulatory genes components of GRNs, genes validated individually by *cis*-regulatory analysis (Istrail et al., 2010). It contains the following components: (1) genomic structure and organization, (2) comparative genomics, (3) *cis*-regulatory analysis of expression constructs and mutational analysis, and (4) spatial expression profiles. The *cis*GRN-Lexicon is hosted in the software system, *cis*GRN-Browser shown in Figure 4.

5.1.1. Computational representation of cis-*regulatory information in gene regulatory networks: Views.* The causality-inferred DNA architecture data obtained through the *cis*-regulatory analysis are stored in data structures specifically designed to capture in an effective way data for the molecular biologist and computational biologist. In the *cis*GRN-Browser and in the BioTapestry and *cis*GRN-Browser Logic Map, these computational representations are called **Views**. There are the classical Views, the View from the genome (VFG) and the View from the nucleus (Davidson, 2006). In addition we developed the software visualizations for the following Views: the View from information processing, the View from BioTapestry-*cis*GRN-Browser Logic Map, and the View from sequence conservation.

VFG (Davidson et al., 2002). A View of the network (Fig. 5) that considers at once all interactions occurring at all nodes of the network at all times and in all relevant spatial domains is referred to as the VFG (Bolouri and Davidson, 2002). This static View presents all the components at once, and it is only at this level that the regulatory logic emerges. In the VFG (sea urchin), the description is given by the genes (horizontal arrows) and by the interactions between TFs and CRMs. These directed edges, called *linkages*, are represented as directed arrows or directed flat ends, signifying activation and repression, respectively. For each such component, a reproducible *cis*-regulatory analysis experiment exists that validates its representation in the model. Predictions expressed in the models logic, for example,



```
103797 AGGAGGGTGATAATAAGAGAAAGGACTAAAACAGGAGATGATGCCGAGCATGAGCTCAGCTGGACTTTATCAATGTTCCCATGCATGTATCATA
103891 GCATGTGGCATATGAATGATATGATAGCGTTAACTTCATGTGGTGTCAGAAGAAAACAGTTTCAATTTTTGTGACCTGACAATTCCACTTCCAA
103985 AATTGCTCCCCTTGCATATGAGACAGAAATTTAAATCAATTTTCCTCCGCCAGCTTCAACCCCAAGGGACAAAGCAAAATGGATTTTTATCCAA
104079 ATCCTTTGAATTTTGGCATTGAATATAAAACATTCTTCAATTTGAGAGGGCAACACCCCTGTTGATGTTGTTAACCTAAAGATGGAGGCCTGTG
104173 CCCCTCAATTTGTCAGCTAGGCAAGGGGTCTCGTGTTATGGAACATTTTTTCCTCTCTCCATCCAGTATTTTGCATATCTTCTGACTGTTTTAT
104267 CAGCATGTATGAAATGTGGTTTGAAATTGTTCAACCAACTACCCCAAATTATGAATATGGAAATGATTGAAATGCCTGCAAATATATTTCACAA
104361 TATATACAGATTTTTCTTGCTTGTGGGTAAATAACAAAACATGATGTTAGTACATTTTATTTGAAACCAGAAAATATTAAACATAGCCCCCAAA
104455 TCTTTCTTTTTTTGTAGCTTGAATTTCATATATTTATATCTTTTATTTTGTAAAAATAGCAATAAAATTCTTTTAAAATTCTGTACCTTTCTAA
104549 TGATCTTCAATTAAAATGTTTACTTCTTTAATTTAACCTTTGTATTTATAGCTTTTTATGTAAAGAAAAGAGAGGAAAATTCCCTGATAAACTG
```

Selected:

Nucleotide Coordinate: 103969 Feature Order Number: N/A

Current Selection: NCBIAXIS:115928265

FIG. 4. The *cis*GRN-Browser: a full genome browser dedicated to the regulatory genome.

mutation of this site X should imply loss of function of gene Y, are testable, that is, falsifiable by *cis*-regulatory analysis experiments.

5.1.2. View from the genome.

Problem 2. Building the **genome assembly of sea urchin at 3× coverage**. We solved this problem using the Celera Genome Assembler software (Sea Urchin Genome Sequencing Consortium, et al., 2006).

Problem 3. Building the **transcriptome assembly of sea urchin**. We were part of the team that built in 2006 the first high-resolution transcriptome map of the sea urchin embryo, a research collaboration between Caltech, NASA, and Brown University (Samanta et al., 2006).

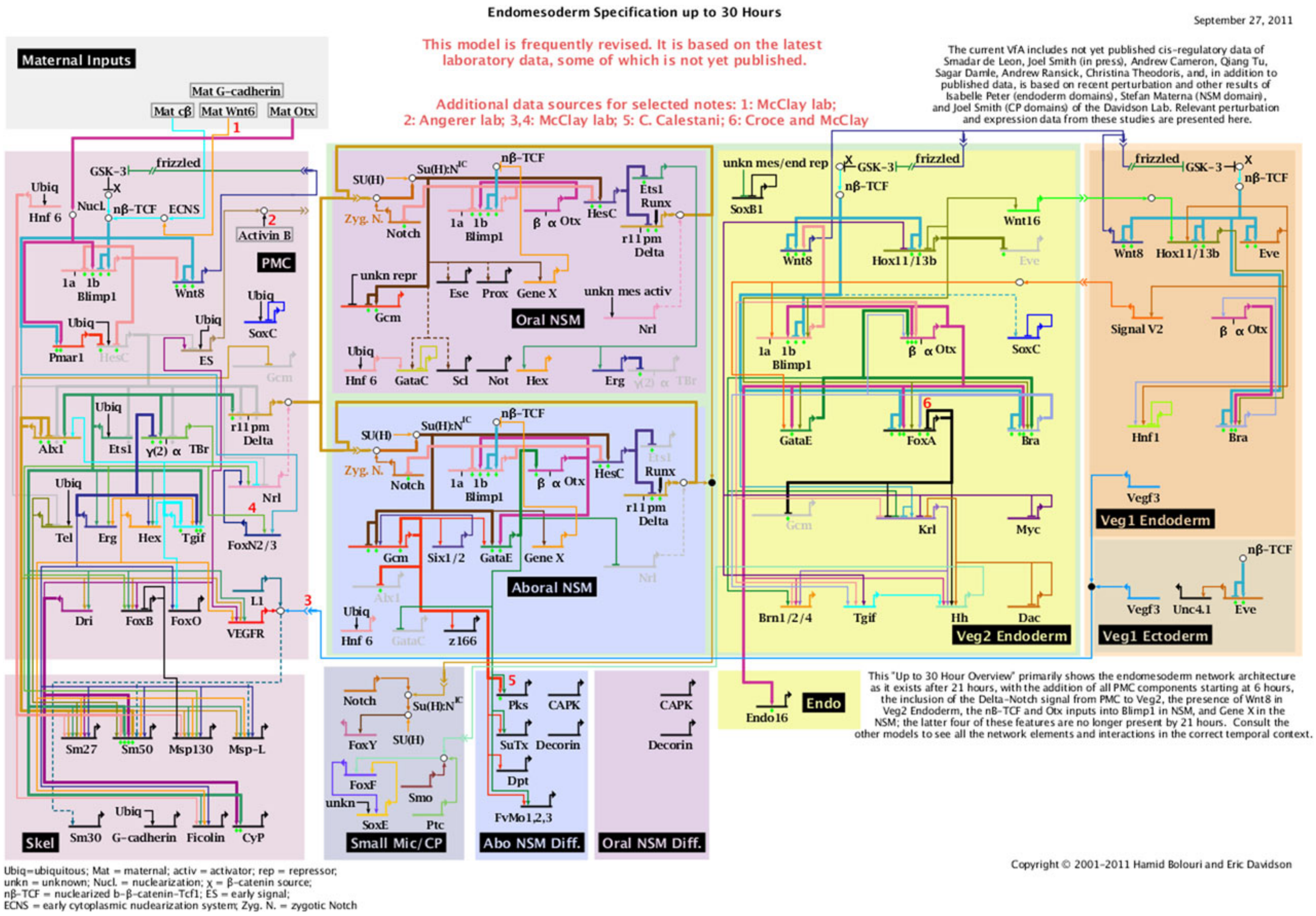
5.1.3. View from the nucleus. A view that specifies only the interactions occurring in a given regulatory state, that is, in a given nucleus at a given moment in time, is called the view from the nucleus (Bolouri and Davidson, 2002; Davidson et al., 2002). This is the developmental biologist's view of the network.

5.1.4. View from information processing and the view from the logic functions.

> *Quantitative models should lead smoothly through the information processing view to the mutable, measurable, regulatory properties of genomic DNA. Models that handle cis-regulatory information flow are in the end the most illuminating and most useful in practical terms because they are DNA sequence-based models.* (Davidson, 2001)

CRMs act as information processors. The work on the sea urchin endo16 gene provided the first detailed quantitative insight into the information-processing view of the genomic regulatory systems (Yuh et al., 2001). The interpretation of the *cis*-regulatory sequence code, by linking target sites directly to a defined set of elemental functions, was pursued formally in Istrail and Davidson (2005); the article provided a first repertoire of logical functions for CRMs. The CRM processes information as an integrated combination of logic functions that is the output of this modular control element. An approach toward a functional interpretation of the genomic regulatory code is to start cataloging the repertoire of *cis*-regulatory operations. It turns out that a prominent class of *cis*-regulatory processing functions can be modeled as logic operations.

A quantitative model verified by kinetic measurements of output showed that logic statements (Fig. 2) represent accurately the functional contributions of those sites at which factors other than the drivers bind.



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

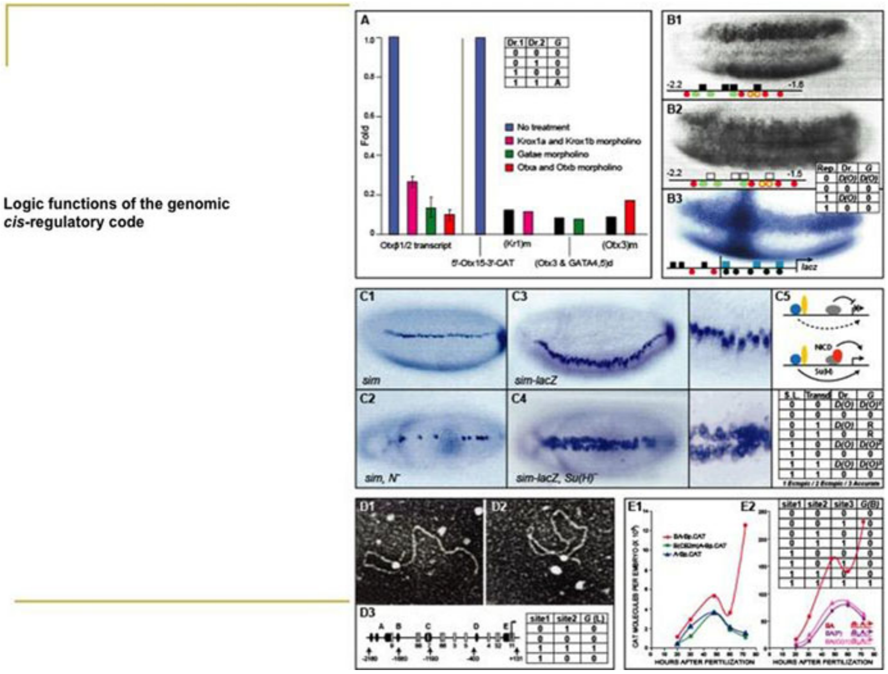
Copyright © 2001-2011 Hamid Bolouri and Eric Davidson

FIG. 5. The sea urchin endomesoderm gene regulatory network.

It was demonstrated that the conditional logic functions executed by these sites in combination explicitly represent the input-processing capabilities of this whole CRM. The endo16 analysis, of course, illuminated only functions operating in that control system. Additional such functions are evident in another sea urchin gene that was the subject of a similar analysis, the cyIIIa gene (Yuh et al., 1998). Many diverse *cis*-regulatory activities, more or less well known, can similarly be treated as operations that determine how driver inputs are used in each given CRM (Istrail and Davidson, 2005).

Problem 4. Building a first set of **logic functions** *of the genomic cis-regulatory code* (Istrail and Davidson, 2005).

Figure 6 shows a class of logic functions, the *Combinatorial logic functions (G operators)* of Istrail and Davidson (2005). They are of the following types: (1) **AND operators**. *cis*-Regulatory analyses often reveal that diverse sites must be occupied in order for significant expression to occur. In development, this device is used to ensure that a gene is activated only in a subdomain (spatial and/or temporal) where two generally noncoincident inputs overlap. In the absence of either factor, there is no expression, even if the other factor is present at the normal level; if either site is destroyed, there is no expression, even if the other remains intact (Yuh et al., 2004). In the input/output table (truth table) shown in Figure 6, the output is considered qualitatively as Activation (A) when both factors are present above threshold (th); if they are not, the output is considered insignificant activity. (2) **Short-range repressor binding within a CRM**. Here the CRM contains target sites for a transcriptional repressor and these site(s) must be within, or loosely adjacent to, a CRM that also contains sites for an activating driver (Gray et al., 1994; see truth table in Fig. 3). The repressor is dominant, so that the activators function only in its absence; otherwise the output is nil. However, the effect of these short-range repressors is limited to the cancellation of the activation functions of that CRM with which they are associated. (3) **Signal-mediated toggle switch**. As reviewed in Barolo and Posakony (2002), many developmentally active intercellular signaling systems used in processes of fate specification operate in a Janus manner: when the signal ligand is presented and the DNA-binding TF that mediates signal transduction is also present in the CRM, a coactivating driver is permitted to stimulate transcriptional expression; if the ligand is absent, however, the same transducing factor acts as a dominant repressor. (4) **Essential DNA looping**, present in some CRMs (perhaps all that are located far from the basal transcription apparatus [BTA]), that contain sites for DNA-binding looping proteins (Mastrangelo et al., 1991; Su et al., 1991; Zeller et al., 1995). (5) **Module linker function**, revealed in the endo16 analysis, where linkage of the A and B CRMs of this gene required three different DNA-binding proteins (see the endo16 gene flowchart in Fig. 1).


Logic functions of the genomic *cis*-regulatory code

Dr.1	Dr.2	G
0	0	0
0	1	0
1	0	0
1	1	A

Rep.	Dr.	G
0	D(O)	D(O)
0	0	0
1	D(O)	0
1	0	0

S.L.	Transd.	Dr.	G
0	0	D(O)	D(O)^1
0	0	0	0
0	1	D(O)	R
0	1	0	R
1	0	D(O)	D(O)^2
1	0	0	0
1	1	D(O)	D(O)^3
1	1	0	0

1 Ectopic / 2 Ectopic / 3 Accurate

site1	site2	G (L)
0	1	0
0	0	0
1	1	1
1	0	0

site1	site2	site3	G(B)
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

FIG. 6. Logic functions: AND logic, short-range repression, toggle switch, DNA looping.

The logical analysis of information processing at the level of single gene, subcircuit, and full network is at its beginning. Detailed mathematical analysis of these principles reveals insights into design of such circuitry, fault tolerance, and self-repair principles necessary for experimental intervention eventually to reprogram such control devices. Although our work is based on experimental results in the Davidson and Levine Labs, a wealth of contributions in developmental biology hold the key to advances through multiorganism comparative gene networks: Xenopus (Koide et al., 2005), B cell developmental pathways (Singh et al., 2005), *C. elegans* (Maduro and Rothman, 2002; Inoue et al., 2005), and T cell development (Rothenberg and Anderson, 2002).

Problem 5. Building the **transcription factors translation table** *across species.*

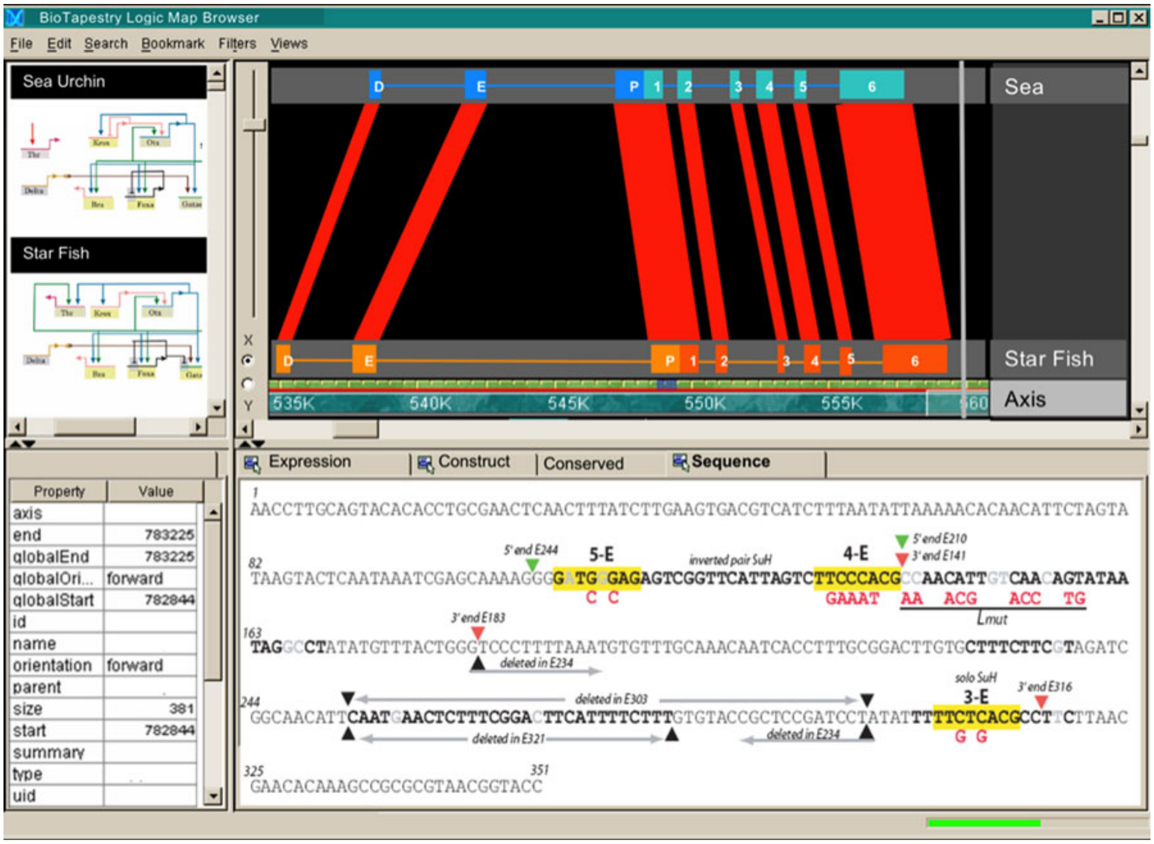
Identifying homologous TFs across species requires careful consideration of their CRA, CRMs, and translation of their names (i.e., the infamous ''Gene Naming Problem''). This problem was solved in collaboration with Derek Aguiar (Tarpine, 2012). This tool is of fundamental importance to obtaining the rules of the genomic *cis*-regulatory code regarding the TFs inputs ''signature'' of organ-specific CRMs across species.

5.1.5. The view from the BioTapestry and cis*GRN-Browser logic map.*

Problem 6. Building the **BioTapestry and *cis*GRN-Browser Logic Map**, *that is, the Regulatory Genomics Logic Map Browser.*

The project was based on the development of a next-generation genome browser incorporating model building, annotation, and visualization capabilities for gene regulatory systems and networks. In its present capabilities, it presents genomic structural views, spatial expression views, expression constructs views, and information processing views, obtained through mathematical analysis of the logical principles of genomic regulatory systems and networks (Tarpine, 2012). Figures 7 and 8 show views from the BioTapestry and *cis*GRN-Browser Logic Map.

BioTapestry is the state-of-the-art genomic GRN tools system for displaying GRNs and predictive regulatory genomic information, high-throughput expression analysis, and network perturbation and



Property	Value
axis	
end	783225
globalEnd	783225
globalOri...	forward
globalStart	782844
id	
name	
orientation	forward
parent	
size	381
start	782844
summary	
type	
uid	

```
1
AACCTTGCAGTACACACCTGCGAACTCAACTTTATCTTGAAGTGACGTCATCTTTAATATTAAAAACACAACATTCTAGTA

82
TAAGTACTCAATAAATCGAGCAAAAGGGGATGGGAGAGTCGGTTCATTAGTCTTCCCACGCCAACATTGTCAACAGTATAA

163
TAGGCCTATATGTTTACTGGGTCCCTTTTAAATGTGTTTGCAAACAATCACCTTTGCGGACTTGTGCTTTCTTCGTAGATC

244
GGCAACATTCAATGAACTCTTTCGGACTTCATTTTCTTTGTGTACCGCTCCGATCCTATATTTTTCTCACGCCTTCTTAAC

325                       351
GAACACAAAGCCGCGCGTAACGGTACC
```

FIG. 7. BioTapestry and *cis*GRN-Browser logic map: the view from the genomic sequence.

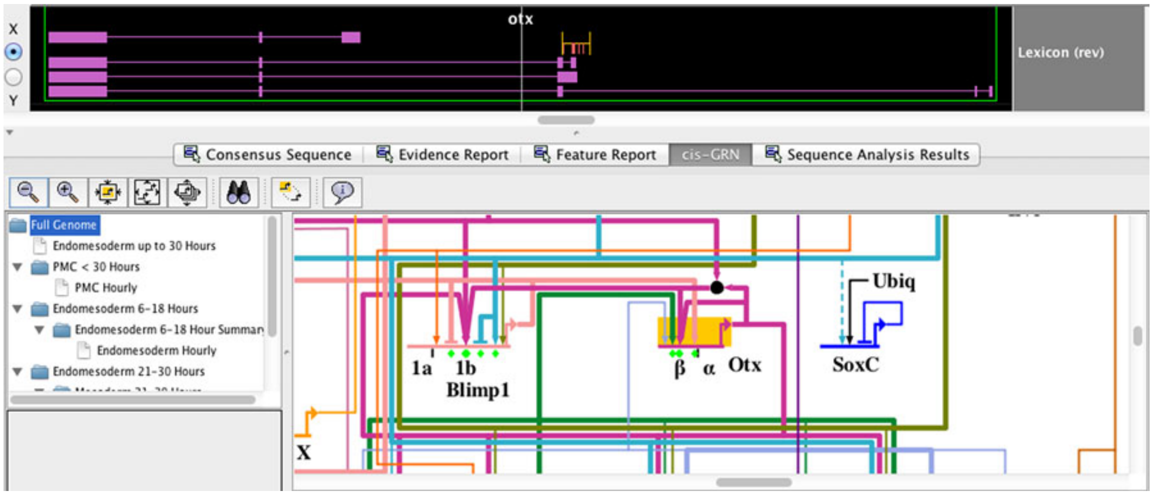


FIG. 8. The BioTapestry and *cis*GRN-Browser logic map.

inference analysis. The BioTapestry software system was developed in the Davidson Lab and helps automate some aspects of the wet laboratory work on genome analysis and discovery of CRA. In particular, it allows to incorporate model-building capabilities for GRNs, annotation, and visualization. Our research project was done in collaboration with Bill Longabaugh, the software developer of the BioTapestry GRN software visualization system (Longabaugh et al., 2005).

5.2. *Question 2: Semantics of causality as written in experimental articles.*

> *Congratulations, Ryan! This is indeed very useful. I did not believe that it can be done.*
> (Davidson, 2011)

Question 2. What is the dimension (size) of the complete *cis*-regulatory universe of all the published articles that present the CRAs of genes validated by *cis*-regulatory analysis (Davidson criteria)?

The work presented in this section will be developed fully in our article in preparation (Davidson et al., in preparation). This question bears a machine learning—literature extraction flavor, but apparently, it is beyond the scope of current machine learning methods then in 2011, and now as well. It focuses on text analysis of all articles in all journals that publish causality-based experimental procedures molecular biology articles, describing genes and their CRAs validated by the experimental method of *cis*-regulatory analysis; these articles also contain the DNA sequence data validated this way. This experimental procedure is the most rigorous experimental validation of causality for CRA of gene regulatory regions; it involves mutagenesis of regulatory DNA followed by gene loss of function, in vivo (see the box ‘‘Nature of the *cis*-regulatory evidence’’ above).

5.2.1. cis*GRN-Lexicon ontology search engine.* With the *cis*GRN-Browser and *cis*GRN-Lexicon efficiently allowing our small army of annotators we had in 2009–2011 to input and store *cis*-regulatory information, the key bottleneck becomes finding relevant journal articles for the annotators to read. One of the key goals of the *cis*GRN-Lexicon project is to be complete: to contain (nearly) all of the information available in the literature. Some help is found through the other regulatory databases. More help is found through surveys, reviews, and books, such as Davidson (2006). But the other databases are incomplete, and even books only claim to give explanatory examples, not to be comprehensive indexes. Therefore, we began to develop the *cis*GRN-Lexicon Ontology Search Engine (CLOSE) to search the literature to automatically detect relevant articles (which we call *cis*-regulatory articles). We call this the *cis*-regulatory article problem:

Problem 7. ***cis*-Regulatory article problem:** Find the most journal publications with information relevant to the *cis*-Lexicon while minimizing false positives.

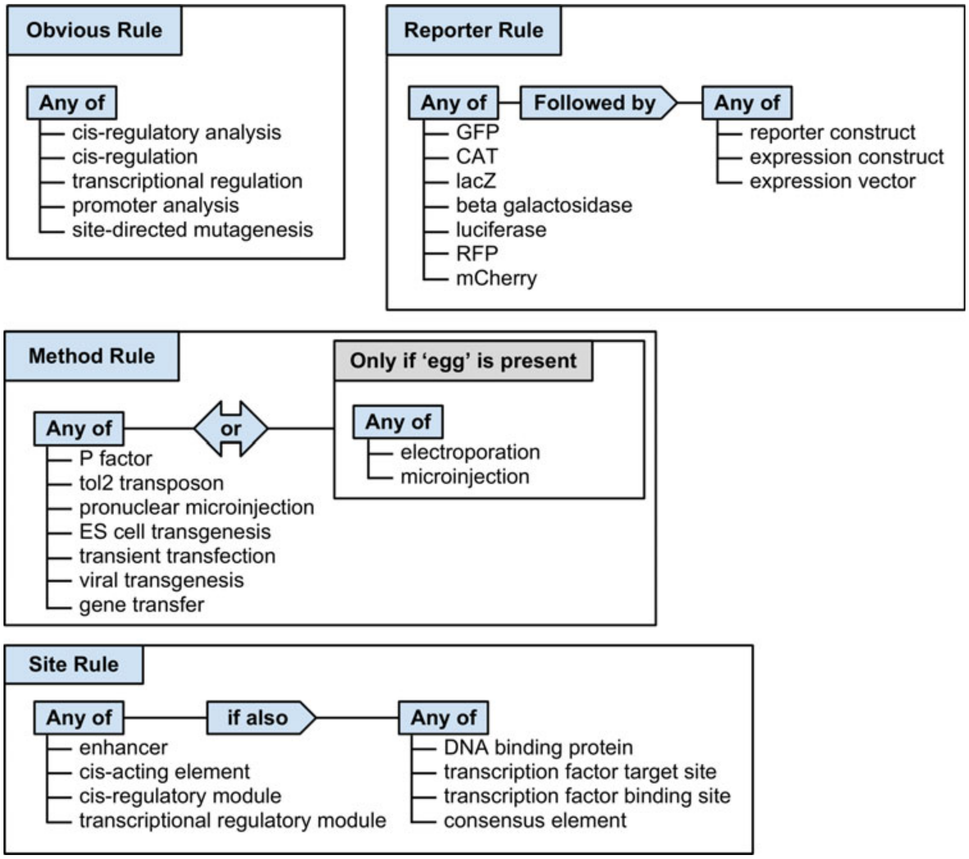


FIG. 9. The CLOSE four original Davidson rules, as suggested by Eric Davidson (later extended with more terms and an additional rule: the action rule, matching terms ''activation,'' ''repression,'' ''boosting,'' etc. CLOSE, *cis*GRN-Lexicon Ontology Search Engine.

Finding ''the most'' is necessary for Lexicon completeness, while minimizing false positives is necessary to allow annotators to actually examine every article. As no approach is perfect, any technique will require a (possibly arbitrary) balance between these two constraints. We have found the F-measure, the weighted harmonic mean of precision and recall, to be a useful measure of success (Manning, 2009). By varying the weighting, we can emphasize precision over recall or vice versa. This begs the question: how do we calculate precision or recall? For precision, we make the naive assumption that all novel articles returned are irrelevant. This is actually a reasonable approximation because we expect only (roughly) one thousand out of a million articles to be relevant. With this assumption, increasing precision means minimizing the size of the returned set. To estimate recall, we simply see how many articles from our training set (which we call CYRENE articles) are recovered.

5.2.2. Semantics of experimentally derived causality as written in articles. There are three challenges (Tarpine, 2012):

- Articles are written in very unstructured format.
- Concepts are semantically promiscuous: notoriously ambiguous, with multiple meanings on related but different areas of molecular biology.
- Cryptic nature of key names involved, the notoriously difficult ''gene naming problem.''

It took us 4 years to resolve this problem through the ''cloning'' of Prof. Davidson approach, that is, to automate through software the way Prof. Davidson would read/scan an article to find out whether it belongs to universe of the experimentally validated articles by the Davidson criteria (Tarpine, 2012).

5.3. Question 3: Proof of system completeness of the gene regulatory network.

> *The emerging paradigm is that the causal functional regulatory mechanisms, revealed through systematic perturbation of DNA, are governed by information processing principles. The nature of this evidence is described by a variety of components: genomic structure of sites and modules, comparative genomics of regulatory regions of species at the right evolutionary distance, mutational analysis of expression constructs, spatial expression. In addition, as cis-regulatory modules are not only genomic structures but also information processors, a repertoire of logic functions building blocks is emerging; they are functionally irreducible protein-DNA assemblies of transcription factors binding to subregions of cis-regulatory modules.* (Istrail and Davidson, 2005).

Question 3: How can we prove the "system completeness" of a causality-focused genomic-based biological system? Can a formal mathematical-computational model for GRNs be developed that is (nearly) completely predictive of its experimental output? Can we design a Programming Logic Language to express the input–output information processing, that is, "computation" performed by each regulatory gene, and their composition together, the "computation" performed together by the whole set of regulatory genes of a GRN?

System-level understanding of a biological system requires completeness: determining all components of the system of interacting parts and their interactions. According to the Davidsonian axioms of causal systems biology, Section 7.1, this means we need to have identified (1) that we have all the components, that is, all the genes; (2) that we have all the linkages, that is, interactions between genes; and (3) each interaction is causality-direct, that is, each linkage has a direct-causality experimental perturbation that shows direct cause and effect for the linkage.

The first seminal moment in genomics for such system-level completeness and understanding was the construction, by computational biology methods, of the first complete genomes of human and sea urchin and other animals. Solving completeness problems is one of the technically most difficult problems. From mathematical logic (capturing all *true* propositions) of logical theories to Internet searches (building Siri's database of all search questions asked on Google) require some of the deepest and most powerful methods, in fact they define what seminal achievements are.

In Eric Davidson's research program a seminal question was how to design a mathematical logic and computational model. That is, a *computational engine* that is information processing-equivalent to the GRNs.

In such a model, the proof of system completeness, de facto, the causality-completeness of the GRNs, is then equated with complete computational predictability of the experimental output; to witness Davidson's exceedingly high scientific standards, the mentioned test should hold for all of the most extraordinary sea urchin regulatory genomics experiments of the past decade. There are deep foundational components involved in this question, including the relationship between a sense of "effective causality"—that is provable through a perturbation experimental technology that is reproducible in different laboratories—and its completeness as witnessed by complete computational predictability.

The groundbreaking work of Isabelle Peter and Eric Davidson (Peter et al., 2012; Faure et al., 2013; Peter and Davidson, 2015) advanced a Boolean computational model called the *Boolean Model of GRN Network Equations* that is (almost) completely predictive of the experimental output. Therefore, the first experimental GRN, the Endomesoderm network (Fig. 5), of exceptional scale and complexity containing about 50 genes, causally explaining the early hours of the sea urchin embryo, was shown **causally systems biology (near) complete** by the near complete predictability of its computational engine by computational biology methods. The extraordinary scale of this achievement, a once-in-a-lifetime 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 pregastrulae embryonic process. The 2015 seminal book by Isabelle Peter and Eric Davidson, *Genomic Control Process: Development and Evolution* (Peter and Davidson, 2015), presents in full detail the pioneering solution of the sea urchin embryo GRN through this (almost) completely predictive computational biology model.

The computational biology achievement of the Peter-Davidson Boolean Model of GRN's computational engine joins the few other seminal achievements in computational biology at the absolute top of the genome era: the Smith–Waterman local alignment algorithm, the BLAST genomic database searching algorithm, and the Celera Genome Assembly algorithm.

5.4. Question 4: Re-engineering by reprogramming the code.

> *The time is almost upon us when we will be able to build cis-regulatory modules and network subcircuits in the laboratory and test their developmental operation in living systems.* (Davidson, 2006)

Question 4. How can we develop a new theory and practice of ''tinkering the regulatory DNA regions'' or reprogramming or reengineering of the CRMs? How tinkering can preserve functionality in novel ways? Or, is the functionality unique? What are the possibilities and limitations?

This is a very challenging experimental problem, which needs guidance from the abstract modeling and the logical theory of the GRN, the Boolean GRN modes of network equations. It was shown that the endo16 CRMs are ''compositional'' and result of the endo16 gene expression is a sum of two nonlinear functions, each one for each module. Although the CRMs and GRNs are information processing systems that operate asynchronous and nondeterministic, in some components sequential and in others, in parallel, they are de facto *deterministic* processing units.

6. THE *cis*GRN-LEXICON META-ANALYSIS, PART I: WHAT ARE THE RULES OF THE GENOMIC *CIS*-REGULATORY CODE?

Algorithmic approaches fail when attempting to cracking the *cis*-regulatory code. Predictive algorithms will only succeed when they utilize databases of accurate experimentally derived CRA. Some databases do exist, such as ORegAnno, REDfly, TRANSFAC, and TRED (Matys et al., 2006; Griffith et al., 2007; Jiang et al., 2007; Gallo et al., 2010), but they accept TFBS annotations resulting from experiments such as DNase I footprinting, gel shifts, and ChIP-chip/seq. All of these techniques suffer from limited resolution—they report a region wherein a factor likely binds. Chromatin immunoprecipitation (ChIP) methods, in particular, are known to be noisy; many of the putative regions it identifies will not in reality contain a binding site (Euskirchen et al., 2007; Johnson et al., 2008). Zinzen et al. (2009), for example, reported that motifs recognized by PWMs were found within 100 bp of only '' ~60%–80%'' of their ChIP peaks. Balmer and Blomhoff (2009) searched the literature for retinoic acid binding sites and found 81 ''tested and verified'' experimentally, from which, upon close examination, at least 22 (27%) appeared to be spurious. The results of techniques such as these cannot be taken as conclusive proof of the existence of regulatory regions or the lack thereof. They also cannot detect the regulatory function of the TFBSs or CRMs they do find. Details on the most popular *cis*-regulatory databases are as follows:

Other databases suffer from similar problems. For accurate conclusions to be drawn about the properties of regulatory regions and the distinctions between random clusters of binding site sequences as opposed to real regulatory modules, accurate data need to be utilized. If any progress is to be made regarding predicting the actual function of CRMs (whether they activate or repress, and in what time and location), this information must be recorded for previously known modules as well. Davidson (2006) suggested that major reasons for the *cis*-regulatory logic code to not yet be understood are the lack of ''sufficiently useful, discriminatory, and general target site databases'' and that ''the decisive importance of particular site and factor combinations has only been sporadically recorded.''

Existing *cis*-regulatory databases usually do contain some type of annotation describing the ''quality'' of each element, this often does not give enough detail as to the type of experiment. Only experiments that pass what we call the Davidson Criteria yield results suitable for entry into the *cis*-Lexicon:

Davidson criteria: TFBSs must be functionally authenticated by site-specific mutagenesis, conducted in vivo, and followed by gene transfer and functional test (Istrail et al., 2010).

Experiments fulfilling this criterion prove the causal links between genes in the GRN. They find the precise means by which one gene regulates another, which is through TFs binding to their sites in the *cis*-regulatory region of the target gene. Other techniques can only prove correlation or association.

6.1. The cis*GRN-Lexicon: a database for* cis*-regulatory information*

The *cis*GRN-Lexicon is a database of *cis*-regulatory information. Although other databases have been built in the past by various groups, these have all suffered the drawbacks discussed in the previous section.

The only place this information can be found is in the journal articles themselves, so for a period of 4 years, we employed a small army of undergraduate biologists to read these articles, determine which meet our standards, and input the data into the *cis*GRN-Lexicon through the *cis*GRN-Browser. The *cis*GRN-Lexicon contains >730 TFs binding >2300 sites in the regulatory regions of >570 target genes. The next box presents the list of types of annotations of the *cis*GRN-Lexicon.

The *cis*GRN-Lexicon contains the following types of annotations:

CRM coordinates: It is not clear how to formulate a definition for the boundaries of a CRM. Sequence conservation often extends beyond the functional binding sites, but this may only be due to evolutionary selection against large insertions or deletions within regulatory sequence (Cameron et al., 2005). Since it is unknown whether the precise boundaries are significant, the boundaries given in the article are stored in the Lexicon along with a note as to how they were determined (whether by restriction sites, sequence conservation, or otherwise). The *general concept* of CRM boundaries is undoubtedly important, as the sites inside a CRM work together yet are functionally independent from sites in other modules.

TFBS coordinates: Knowing the TFBS coordinates implies knowing both location and sequence. The location specifies the relationship with other sites that work in combination, as well as the distance to the transcription start site (which affects how the binding factor interacts with the transcription apparatus). The sequence of individual sites aids in defining models as to the general type of site a given TF binds to.

TFBS regulatory function: TFBSs can be annotated with the regulatory functions that they fulfill: activation, repression, signal response, DNA looping, etc. The precise choices available are terms from the *cis*-Regulatory Ontology (CRO), which was designed by examining typical *cis*-regulatory analysis articles and distilling the various terms describing the same fundamental phenomena into a controlled vocabulary setting.

TFBS binding factors: The binding factor (or factors, in the case of a complex) is specified for each site. To avoid the ‘‘Gene Naming Problem’’ (Tarpine and Istrail, 2009) where the precise identity of a gene is unknown because it is known by a set of names (sometimes overlapping with a set of names from a different gene), the NCBI GeneID is stored. This also allows for efficient algorithmic processing.

TF families: TFs exist in a hierarchy. There are multiple ways to classify them, and we chose a modified TRANSFAC system (Fig. 10) to classify all of the TFs in the *cis*-Lexicon (both target genes and *cis*-regulatory inputs). For meta-analysis (Section 6), for TFs that are not found in the *cis*-Lexicon enough times to generate reliable statistics, we plan to combine data of TFs within the same family.

Sequence conservation: Occasionally articles note that *cis*-regulatory sequence, whether for individual binding sites or for entire modules, is conserved across species. Annotators can quickly record this knowledge in the *cis*-Lexicon. Sometimes the additional species may not have their whole genomes sequenced yet, such as opossum and elephant. Keeping these data in the Lexicon allows for the sequence to be annotated in those species when their full genomes are sequenced in the future.

Target gene function: An open question concerning *cis*-regulatory regions is whether their architecture is fundamentally different between different types of genes—are the regulatory regions of TF encoding genes different from those of housekeeping genes? To allow this type of analysis, annotators note in the Lexicon the type of each target gene.

Given our limited time and resources, we decided to focus on collecting the regulatory information of TF-encoding genes in eight particular species only, for the current time: human, mouse, fruit fly, sea urchin, nematode, rat, chicken, and zebrafish, with the highest priority on the first five species. When completeness (Section 6) of TF regulatory regions in these species is reached, then our focus will move to a new type of gene.

The following is the *cis*-Regulatory Ontology (CRO) box with its formally defined vocabulary.

The CRO has the following controlled vocabulary:

Repression: It indicates that mutating the TFBS increases gene expression or produces ectopic expression. Repressors may act ''long range'' when the repression effect may target more than one enhancer, or ''short range'' when repression affects only neighboring activators (Gray et al., 1994; Courey and Jia, 2001). The function of repression applies in cases where the repressors interact with the BTA either directly or indirectly (Nakao and Ishizawa, 1994).

Activation: It indicates that mutation decreases gene expression. An activator TFBS may act over a large genomic distance or short. See Latchman (2008) for further discussion of some of the many ways a TF can accomplish activation.

Signal response: It indicates that the TF has been shown to be activated by a ligand such as a hormone (phosphorylation is not included) (Barolo and Posakony, 2002).

DNA looping: It indicates that the binding factor is involved in a protein–protein interaction with another binding factor some distance away that causes the DNA to form one or more loops. This looping brings distant regulatory elements closer to each other and to the BTA (Zeller et al., 1995).

Booster: It indicates that the TFBS does not increase gene expression on its own but can augment activation by other TFBSs.

Input into AND logic: It indicates that the TFBS can activate gene expression only when two or more cooperating TFBSs are bound (Istrail and Davidson, 2005; Istrail et al., 2007).

Input into OR logic: It indicates that the TFBS can activate gene expression when either or both of two or more cooperating TFBSs are bound (Istrail and Davidson, 2005; Istrail et al., 2007).

Linker: It indicates that a TFBS is responsible for communicating between CRMs [such as the CB2, CG1, or P sites in modules A and B of endo16 (Yuh et al., 2001)]—mutating the TFBS prevents the functions of the independent modules from combining.

Driver: It indicates that this TFBS is the primary determining factor of gene expression. The binding factor appears only in certain developmental situations and thus is the key input for directing gene expression. TFBSs that are not drivers usually bind ubiquitous factors (Smith and Davidson, 2008).

Communication with BTA: It indicates that the sites are directly involved with interactions with the BTA (many sites are only indirectly involved—they use other sites as mediators).

Insulator: It indicates that the TFBS causes *cis*-regulatory elements to be kept separate from one another. Insulators can separate the function of *cis*-regulatory elements of different genes as well as act as a barricade to keep active segments of DNA free of histones and remain active (West et al., 2002).

6.2. Implementation

We implemented the *cis*GRN-Lexicon using Apache Derby, an open source relational database. Since Apache Derby is implemented entirely in Java, the *cis*GRN-Browser remains entirely cross-platform. Derby can be run either in embedded mode, where the database is stored and accessed locally, or as a network client, where the database is stored remotely and accessed through a server. This allows the *cis*GRN-Lexicon to be packaged with the *cis*GRN-Browser for ease of access or to be stored in one central location so users of the *cis*GRN-Browser around the world will see an updated database from the moment a change is made.

6.3. Completeness

Completeness of the *cis*-Lexicon is critical in order for correct conclusions to be drawn. There are two kinds of completeness:

1. Biological completeness, where we know all *cis*-regulatory information.
2. Literature completeness, where we have captured all information available in the literature.

Achieving completeness of the first type is unfortunately not possible, as this information is simply unavailable. Therefore, we aim to be as close to literature completeness as possible, and this is what we refer to when we use the term ''completeness'' without qualification. We use several methods to judge completeness, including (1) TF counts per species, (2) inspection by domain experts, (3) consulting literature reviews, (4) literature searches (CLOSE; see Section 5.2), and (5) other regulatory databases.

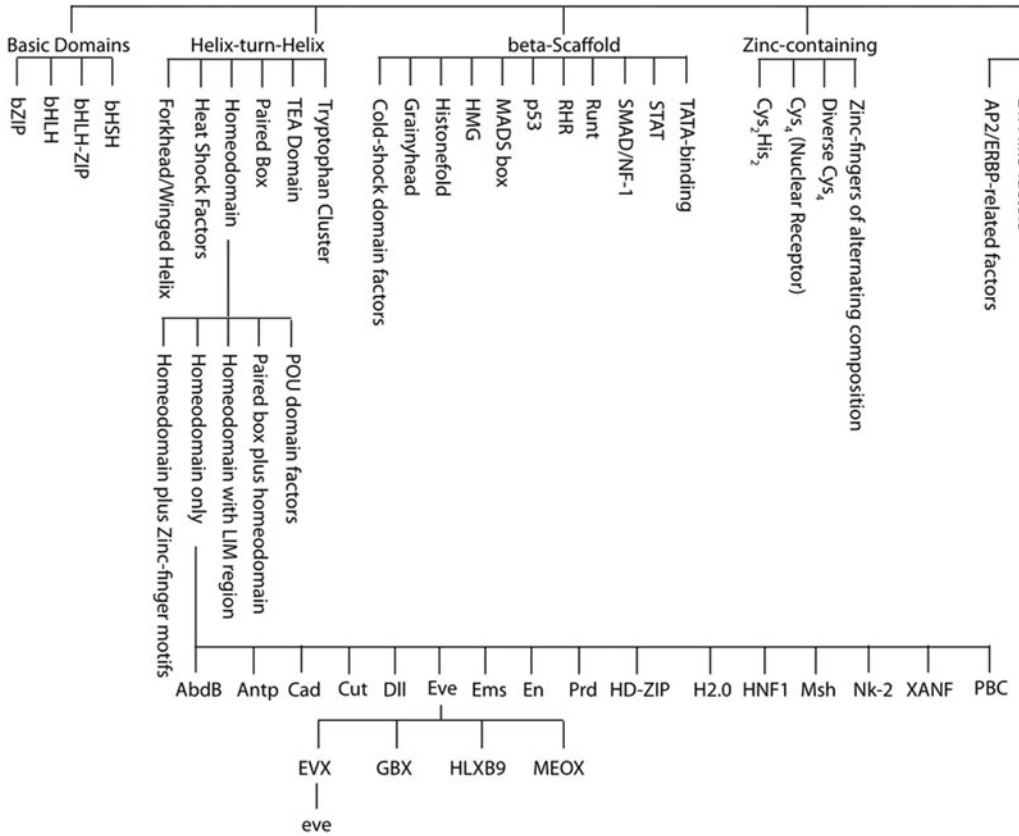


FIG. 10. Transcription factor hierarchy.

A pessimistic estimate of literature completeness can be given by evaluating biological completeness. This is possible because the set of all TFs is known for species whose genomes have been sequenced. There are standard tools for recognizing which genes encode TFs, even if the genes have not been studied experimentally.

We have carried out routine reviews of the Lexicon by domain experts, who verify that genes are classified correctly and that the most well-known modules are in place. Although one might expect that the best-understood modules are the most likely to have their literature found and entered into the Lexicon, actually the reverse is often true: the most popular genes tend to have many articles discussing them that it can be difficult to find the original articles where the *cis*-regulatory analysis was performed. Recognizing them is especially difficult with automated methods like CLOSE (Section 5.2). For this reason, completeness tests based on these genes are less biased that one would expect.

Many of the articles that discuss but do not actually perform *cis*-regulatory analysis cite several articles that do carry it out. We used such articles to test the Lexicon as well.

CLOSE is a key method for testing completeness. We routinely take samples of 100–500 articles from CLOSE results and examine them to see how many discuss *cis*-regulatory analysis of genes not yet in the Lexicon. Recently only 1%–2% of the articles have contained novel information. This is a strong sign that there are few genes remaining to be found.

Our annotators have studied in detail existing databases such as REDfly and TRANSFAC to look for genes not yet in the Lexicon. This is a tedious process because even when the evidence given in those databases does not meet our criteria, they check whether newer literature has been published on those genes.

All of these methods give evidence to the *cis*-Lexicon being nearly literature complete.

6.4. A first visit to the cis*GRN-Lexicon meta-analysis*

We performed several preliminary analyses of the data in the *cis*-Lexicon to extract properties of *cis*-regulatory regions that distinguish them from chance clusters of TFBSs (which, due to the size of the genome, occur thousands of times in the genome of any complex species). Without a clear understanding of

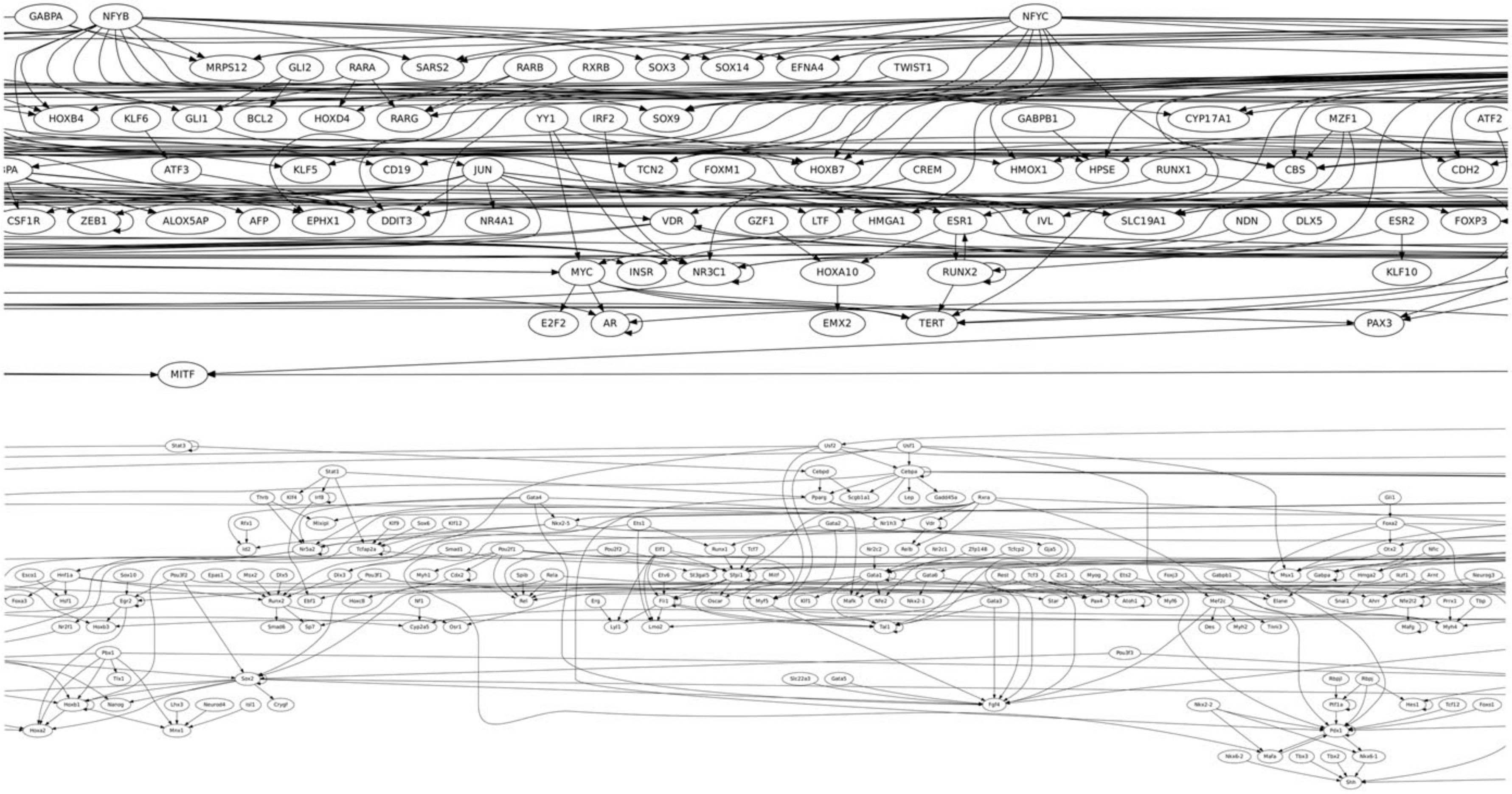


FIG. 11. Excerpts from the transcription factor connectivity graphs for the human, mouse, and fruit fly genomes; each vertex represents a gene, and we draw a directed edge between two vertices if the first gene regulates the second.

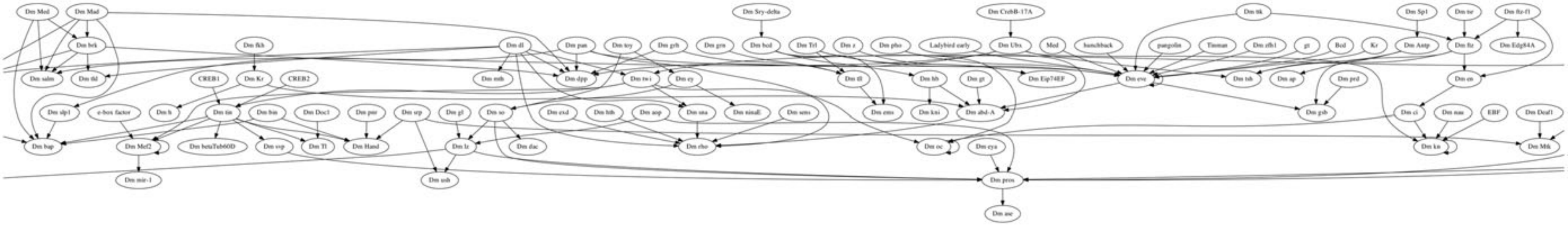


FIG. 11. (*Continued*)

what distinguishes true regulatory regions, past work has involved looking for clusters of TFBSs or conserved sequence, which are neither sufficient nor necessary.

6.5. Transcription factor co-occurrence

Different TFs whose sites occur within the same CRM belong to the same *regulatory state*. The regulatory state of a cell at a given time is the set of TFs that are expressed in the cell at that time. It is called the regulatory state since future gene regulation in the cell is determined almost entirely by which TFs are present. Cells that adopt different fates (e.g., becoming part of the eye vs. becoming part of the heart) do this due to a difference in regulatory state. Not all states possible in theory occur in practice. If there are 20,000 genes in a species, for example, this does not mean that $2^{20,000}$ regulatory states are involved in the development of an organism of that species. In reality, a much smaller set of states occur, determined largely by the possible ways that the TF proteins (and their cofactors) can interact with each other to influence transcription.

To discover TFs that are part of the same regulatory state, we searched the *cis*GRN-Lexicon for TFs that bind within the same CRMs. An idea of the complexity of this information alone can be seen by inspecting the human, mouse, and fruit fly connectivity graphs, shown in Figure 11.

These graphs, which also represent the ''View from the Genome'' described in Davidson (2006), contain a vertex for each gene and draw a directed edge between two vertices if the first gene regulates the second, Figure 11. These graphs are not GRNs, but rather the union of multiple GRNs, showing the relationships between genes at various times and locations in the development of an organism. False conclusions could be drawn if these graphs were interpreted like GRNs. For example, if there is an edge from Gene A to Gene B and another edge from Gene B to Gene C, this does not mean necessarily that the expression of Gene A affects the expression of Gene C. It may be that the two edges in the graph are from different times or locations, and at no time does Gene A influence Gene C. But connectivity graphs are useful for visualizing the complexity of the regulatory genome and the completeness of the *cis*GRN-Lexicon.

The most common pairs of TFs often included SP1, SP3, and their homologs, which are ubiquitously expressed in mammalian cells and are known to regulate genes involved in almost all types of cellular processes by interacting with a variety of proteins (Lin et al., 2004). Binding sites for these factors are not informative, so we do not present or discuss them. The most common pairs not involving SP1 or SP3 are given in Table 1.

The number of binding sites of each TF was not considered—for example, if a CRM contained two TFBSs for Mad and three for Med, this was counted as a single occurrence of the pair Mad–Med. We examined both the number of CRMs that a pair was found in as well as the number of unique target genes, although these were generally the same. The only exception was murine Ptf1a and Rbpj, which was found regulating two genes but in three CRMs (one CRM of Pdx1 and two CRMs of Ptf1a itself). There were many more examples of pairs found exactly twice ($\sim$30 more), but they are not presented here to save space (since they are not strong evidence of TF cooperation). However, we note that several well-known examples of TFs that combine to form regulatory complexes currently have only two examples in the *cis*GRN-Lexicon [e.g., dl-twi (Zeitlinger et al., 2007) and PAX6-SOX2 (Kamachi et al., 2001)]. This does not mean that the *cis*-Lexicon is missing information—for those two examples, the complexes were recognized by high-throughput methods and careful experimental analysis of a single gene, respectively. It

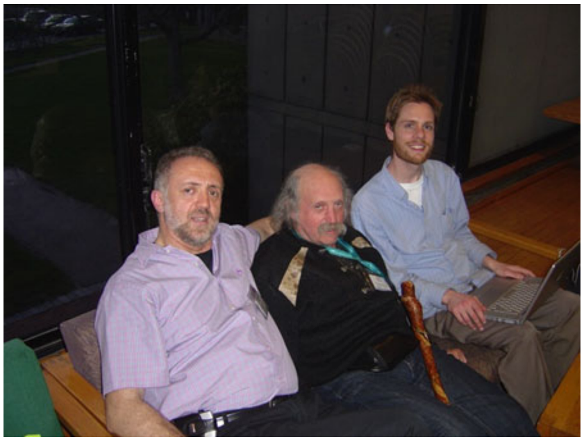


FIG. 12. From right to left: Ryan Tarpine, Eric Davidson, and Sorin Istrail.

Table 1. Pairs of Transcription Factors Observed to Bind Within the Same *cis*-Regulatory Modules

TF1	*TF2*	*No. of CRMs*	*No. of target genes*
Mm Pou5f1 (18999)	Sox2 (20674)	4	4
Dm Mad (33529)	Med (43725)	3	3
Mm Hoxb1 (15407)	Pbx1 (18514)	3	3
Dm exd (32567)	hth (41273)	3	3
Hs HNF4A (3172)	HNF1A (6927)	3	3
Dm brk (31665)	Mad (33529)	3	3
Mm Nkx2-2 (18088)	Pdx1 (18609)	3	3
Hs CEBPB (1051)	DBP (1628)	3	3
Mm Ptf1a (19213)	Rbpj (19664)	3	2
Rn Usf2 (81817)	Usf1 (83586)	2	2
Hs USF1 (7391)	USF2 (7392)	2	2
Hs RXRA (6256)	NR2F2 (7026)	2	2
Hs NFYA/B/C (4800)	USF1 (7391)	2	2
Dm dl (35047)	twi (37655)	2	2
Gg PAX6 (395943)	SOX2 (396105)	2	2

Species' names are abbreviated (*M. musculus* is *Mm*, etc.), and the unique NCBI GeneID for each gene is given in parentheses.

CRM, *cis*-regulatory module; TF, transcription factor.

is also interesting to note that TFs that often bind within the same CRM do not necessarily work together—it is possible that their binding sites overlap so that they *compete* for occupancy to carry out their regulatory function, such as brk and mad (Kirkpatrick et al., 2001).

6.6. Intertranscription factor binding sites spacing

Not all TFs part of the same regulatory state necessarily interact with each other directly. For example, the five TFs that bind to the 10 sites of Module A of endo16 do not all interact. TFs that directly interact with each other are said to be *cooperative*. One example of such a pair already known to work together is Dorsal and Twist or Snail (Zeitlinger et al., 2007). TFs that bind cooperatively tend to bind a specific distance apart from each other. This allows the proteins to interact without any need for DNA looping [see Zeller et al. (1995) for a review]. This can only occur when the binding sites have the specific correct distance between them. If the sites are too far apart, the proteins cannot come into contact with each other

Table 2. Pairs of Transcription Factors Whose Sites Are Observed to Bind Nearby Each Other Within the Same *cis*-Regulatory Modules

TF1	*TF2*	*No. of occurrences*	*No. of target genes*
Mm Pou5f1 (18999)	Sox2 (20674)	4	4
Dm Mad (33529)	Med (43725)	3	3
Dm exd (32567)	hth (41273)	3	3
Mm Ptf1a (19213)	Rbpj (19664)	3	2
Mm Pbx1 (18514)	Pknox1 (18771)	3	2
Mm Hoxb1 (15407)	Pknox1 (18771)	3	2
Gg PAX6 (395943)	SOX2 (396105)	2	2
Hs PDX1 (3651)	HNF1A (6927)	2	2
Hs PDX1 (3651)	NEUROD1 (4760)	2	2
Hs HNF4A (3172)	HNF1A (6927)	2	2
Hs GABPA (2551)	NFYA/B/C (4802)	2	2
Mm Pou2f1 (18986)	Sox2 (20674)	2	2
Mm Pou2f1 (18986)	Sfpi1 (20375)	2	2
Mm Gata1 (14460)	Tcfcp2 (21422)	2	2

Species' names are abbreviated (*M. musculus* is *Mm*, etc.), and the unique NCBI GeneID for each gene is given in parentheses.

without some sort of looping; if the sites are too close, the TFs cannot bind simultaneously. Consistency of distance is strong evidence that two TFs do in fact work cooperatively. Even if two factors often bind within the same CRM, if the distance between them appears random, it is not clear whether they interact through DNA looping or whether they do not interact at all.

The analysis we discussed in the previous section permitted the binding sites of the two factors to be situated anywhere within the CRM. To detect cooperativity, we scanned the *cis*-Lexicon to find examples of pairs of TFs whose sites are consistently a nearly constant distance apart. Since binding sites are input into the *cis*GRN-Lexicon exactly as they are given in the literature, their boundaries are not consistent. For example, one article might give a 4 bp binding site, whereas another will present an 8 bp binding site. Different binding sites may represent different parts of the overall binding motif. To allow for these types of differences, we simply searched for pairs of binding sites for which the distance between them was less than 20 bp. This is the maximum typical distance that allows neighboring bound proteins to interact. The results are summarized in Table 2.

Unlike the TF co-occurrence analysis mentioned, in which there were very many pairs that were found regulating exactly two target genes, the set of results here was much smaller (as to be expected). All pairs found regulating two or more target genes are given in Table 2.

6.7. Transcription factor binding site multiplicity

In some CRMs, a single TF binds at many different sites, such as Kni in the stripes 3 + 7 regulatory module of the *Drosophila melanogaster* gene *eve* (12 binding sites) and Gata in the Otx15 module of the *S. purpuratus* gene *otx* (5 binding sites). Other TFs bind only once within a CRM, such as Slp in *DmDll* and Brn1/2/4 in *SpEndo16* (Davidson, 2006). We searched the *cis*-Lexicon to recognize the TFs that tend to bind at many sites within a single CRM, only once, or have no clear pattern. Some of the results are given in Table 3. Most TFs appeared to bind only once or twice per CRM.

6.8. Interspecies analysis

The mentioned three analyses treated every gene as a unique entity. This is the most conservative type of analysis, but not the most powerful. More correlations can be recognized if the same genes in *multiple species* are grouped together. For example, the TFs HNF1A and HNF4A in human have been observed to bind near each other twice in human, and the TFs Hnf1a and Hnf4a have been observed to bind near each other once in rat. Separately, these observations are not very significant. When pooled together to yield three observations of the same pair of TFs, the evidence is much stronger.

Determining genes in different species to be the ''same'' is nontrivial. In general, when two genes are said to be the same, what is meant is that the genes are ''orthologous'': they are descended evolutionarily from a common ancestor. Seeing the names HNF1A and Hnf1a might lead one to expect that genes that are the same will have the same name. This is often not the case, as in the trio of genes *Rbpj* (in mouse), *Su(H)* (in *Drosophila*), and *lag-1* (in *C. elegans*) (this gene is also called *RBPJ* in human and *Su(h)* in the sea urchin). Many genes have interesting scientific histories behind them and are often named after the

Table 3. Categorization of Various Transcription Factors According to the Number of Binding Sites Found in Single *cis*-Regulatory Modules

Transcription factor	*No. of CRMs*	*No. of genes*	*TFBSs (CRMs)*	*Category*
Hs JUN (3725)	16	15	2 (2); 1 (14)	Single
Mm Rxra (20181)	11	10	2 (2); 1 (9)	Single
Mm Gata1 (14460)	11	8	2 (6); 1 (5)	Single
Hs HNF4A (3172)	10	9	2 (1); 1 (9)	Single
Mm Sfpi1 (20375)	8	7	3 (3); 2 (1); 1 (4)	Varies
Dm dl (35047)	7	7	3–4 (4); 2 (3)	Multiple
Dm Ubx (42034)	6	5	4–12 (3); 1–2 (3)	Varies
Dm Su(H) (34881)	6	6	3–7 (5); 1 (1)	Multiple
Dm Mad (33529)	5	5	9 (1); 4 (1); 1–2 (3)	Varies
Dm srp (41944)	4	3	5 (1); 3 (3)	Multiple

TFBS, transcription factor binding site.

Table 4. Interspecies Transcription Factor Cooccurrence Analysis

		No. of CRMs	*No. of target genes*	*No. of species*
Hnf4a	Hnf1a	5	5	3
Pou5f1	Sox2	5	5	2
Nr1h3	Rxra	4	4	2
Mad/Smad1	Med/Smad4	4	4	2
exd/Pbx1	hth/Meis1	4	4	2
Hoxb1	Pbx1	4	4	2
Srf	Creb1	3	3	3
Srf	Egr1	3	3	3
Srf	Elk1	3	3	2
Gata4	Nkx2-5	3	3	2
Pbx1	Pknox1	3	3	2
Rxra	Nr2f2	3	3	2
Hoxb1	Pknox1	3	3	2
Hnf4a	Nr2f2	3	3	2

phenotypes they cause when mutated, which varies from species to species even when the basic function of the gene is the same. Oftentimes a journal article will give synonyms for the gene or genes under study, such as ''CEH-22/tinman/Nkx2.5,'' ''Wnt/MAPK,'' ''POP-1/TCF,'' and ''SYS-1/beta-catenin'' mentioned in Lam et al. (2006).

Databases that attempt to record orthology relationships, such as NCBI Homologene (Davidson et al., in preparation) and InParanoid (Ostlund et al., 2010), are based on automated sequence comparisons that cannot take into account the functional relationships between related genes. Thus, one will often find mentioned in articles that *tinman* and *Nkx2.5* are synonyms, but one will fail to find this relationship in databases. This is due to the fact that there is a family of related TFs, and if one considers only the sequence, one will judge another factor in *Drosophila* other than *tinman* to be more closely related to *Nkx2.5*, and *tinman* to be more closely related to something other than *Nkx2.5*. The reason that these two genes are considered synonyms despite the difference in sequence is due to their shared function: they are both involved in the development of the heart. There is no perfect solution to this problem other than carefully recording all synonyms reported in the literature. Although we have been doing this as part of the *cis*-Lexicon annotations, this has not been a focus, and our records are not complete enough to be relied upon. The articles we use to annotate the *cis*-Lexicon are not consistent in citing synonyms since they are not necessary. For this reason, we have relied on NCBI HomoloGene as an imperfect but reasonably complete source of interspecies synonyms.

Results from interspecies TF co-occurrence analysis are given in Table 4. This table only shows pairs of TFs seen in at least three CRMs and in *multiple species*. Results from interspecies inter-TFBS spacing analysis are presented in Table 5. Again, this table only shows pairs of TFs found in *multiple species*.

More surprising results came from the interspecies TFBS multiplicity analysis. TFs known to bind multiply in one species tended to bind singly in other species, such as Su(H) in *Drosophila*, which binds multiply (as already noted), whereas the mammalian homolog Rbpj tends to bind singly in mouse and human (one CRM in mouse contains three sites for Rbpj, whereas the other five known CRMs in mouse and

Table 5. Interspecies Intertranscription Factor Binding Site Spacing Analysis

		No. of CRMs	*No. of target genes*	*No. of species*
hth/Meis1	exd/Pbx1	4	4	2
Su(H)/Rbpj	da/Tcf12	4	2	2
Hnf4a	Hnf1a	3	3	2
Srf	Creb1	3	2	3
Gata4	Nkx2-5	4	2	2
Tcf3	Hnf1a	2	2	2
Smad2/3/4	Fos/Jun	2	2	2
Gata1	Fli1	2	2	2

human contain only one site). Similarly, although Dl usually binds multiply as mentioned earlier, its mammalian homolog Rela has only a single site in 11 of the 12 CRMs discovered in human, mouse, and rat. The other TF found to bind multiply, srp, does not have homologs binding in the *cis*-Lexicon. We did not observe any TF to consistently bind multiply across different species.

7. THE UNREASONABLE EFFECTIVENESS OF MATHEMATICS IN THE REGULATORY GENOME

We present in this section some epistemological and philosophical reflections about models of experimental systems and their underlying mathematical and computational sciences structures emerging in the regulatory genome. Causality, logic, and proof are the topics discussed here.

Causal experimental systems biology was ''axiomatized'' by Eric Davidson through the concept of ''rooted causality'' explanation that is genome-based. We present his axioms as a guidance for what system-wide causal explanations are and what ''understanding'' of ''complete'' causality means. It is in this system-wide causal understanding where the fundamental question of system completeness rests. Only in the past 25 years, a rigorous computational and mathematical sciences-based theory of causality has emerged. It is very pleasing to see that in the light of this new theory, the Davidson's models have been causal models to the fullest extent by the present mathematical theory, starting from the seminal Britten–Davidson 1969 article.

We next present some remarks on logic and computation through von Neumann's ''the computer and the brain'' analysis and remark that although almost all computers of today have the *von Neumann architecture* as hardware organization, the regulatory genome as a natural automaton computes in totally different ways as compared with the other natural automata, the brain (in von Neumann's modeling of it), and for that matter, in totally different ways as well as the ways employed by the artificial automata, the electronic computer. Both natural automata, the regulatory genome and the brain, compute through a mixture of digital and analog computing principles, but digital computations are dominant, and, therefore: *They are digital ''computers''!*

We conclude this section with some remarks on the results of the Peter-Davidson Boolean GRN Model (Peter et al., 2012), and comment about obtaining a mathematically provable ''completeness theorem: GRN is Solved,'' that is, complete rooted causality explanation ''proved'' with full mathematical rigor through near complete computational predictability.

7.1. Davidsonian causal systems biology axioms

Eric Davidson's research program centered on ''causality'' in the molecular biology of GRNs. As the name ''gene regulatory networks'' is used in the literature for many other types of molecular biology models, a more faithful name for Eric Davidson's seminal life work models would be ''Causal Gene Regulatory Networks,'' or even more specific, ''Causal Transcription Factors Gene Regulatory Networks.''

To get a glance at the basic unit of ''causality,'' as represented in the today mathematical models of causality, we can look at the Endomesoderm network (Fig. 5): there we have genes (TF encoding genes) that are represented by ''bent arrows,'' with their first part, a flat line, indicating the regulatory region of the gene, and then the ''arrow head'' representing the gene expression output of the gene. Such gene arrows go and target the regulatory regions or some other genes; the arrowhead is displayed there in the regulatory regions of those genes. Such ''arrows'' define direct causality between a gene and a target gene. The elegant theory of causality is captured by such directed graphs, ''causal graphs'' where the directed edges correspond to direct cause-and-effect edges.

The following axioms for ''Causal Systems Biology'' models were presented in Davidson (2015):

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.

7.2. Mathematical models of causality and the Davidson gene regulatory network model

Causality is one of the deepest concepts in science and mathematics and computer science. It is also one of the most fundamental concepts in philosophy. Although its study has a very long timeline, only in the past 25 years the rigorous theory of causality emerged, and computer science methods played an essential role in this very exciting development. Excellent texts about causality models and their mathematical and computer science foundations are Pearl (2000), Glymour et al. (2000), Shipley (2016), Williamson (2005), and Simon (1957).

Judea Pearl won the Turing Award (also known as the Nobel Prize for Computing) for his pioneering work on computational and mathematical models for causality. Such models are very recent, they did not exist 25 years ago. Pearls's book "Causality" (Pearl, 2000) provided the seminal foundation for this new, exciting, and quite controversial, mathematical, statistical, and computer science research area. Pearl's theory of causality presents the "Ladder of Causation" containing three levels of causation (Pearl, 2018):

1. Level 1: *Association*. "What if I see..?." Correlation, observing, and seeing data "correlation is no causation."
2. Level 2: *Intervention*. "What if I do…? How?" Doing, intervening, perturbations.
3. Level 3: *Counterfactuals* "What if I had done …?Why?" Imagining, retrospection, understanding.

Eric Davidson's research program focused on causality and GRN "complete" causality, in the decisive sense of complete computational predictability, aligns perfectly and fully on all three levels of the *Ladder of Causality* in the state-of-the-art mathematical theory of today (Pearl, 2018). Pearl's theory of causality restates again and again, in powerful rigorous ways the warning that "correlation is not causation," that is, association, the Level 1 of causality, is all you can achieve with computational and mathematical analytics. To move to Levels 2 and 3, you need to do experimental perturbations. In the same spirit, Eric Davidson provides a strong critique of today big data, that is, Level 1 only approaches.

We cite next some of Eric Davidson's prophetic reflections, in his own words, on causality principles.

- **The 1969 Britten-Davidson Model:** "*where causality in embryonic development must lie.*"
- *In 1969 Roy Britten and I formulated a gene regulatory network(GRN) model called "A theory of gene regulation for higher cells" (Britten and Davidson, 1969), which was the initial foruner of modern GRN models for control of development. This model was system biology in essence: it was a distributed gene interaction model in which*

regulatory genes had multiple targets and target genes were expressed in an integrated, regulatory sequence-dependent way, with provisions for signal integration. The conceptual output of the model system was a means of explaining on a large scale the spatial expression of cohorts of genes. This model was based on the minimal knowledge then available about populations of nucleic acid species, and a much deeper store of descriptive knowledge of the events of development, but mainly just on pure logic given the requirement that a genomically encoded control system must exist. (Britten and Davidson, 1969)

- *The book (Davidson, 1986) proposes that causality in embryonic development must lie (i) in the multiple functions of the encoded gene regulatory system; (ii) in the informational characteristics of mRNA populations; (iii) in how the hugely complex genome works to make development happen. All these follow from what we call today* biological control system theory.
- *The basic precept of systems biology, that a defined living process can only be solved if all the moving parts of the system are known and their interactions are discovered, is what animates the major effort of our last 15 years, which has been to solve and to understand the functions of developmental GRNs. The distinguishing features of this now increasingly successful effort with respect to its predecessors over the decades before are a finally adequate knowledge base, the finally adequate technology potency of the methodologies available, and the development of a solid working theory of GRN structure/function. Our current trajectory is not due to recent advent of novel ideas about systems biology, as systems biology has been my own scientific purview for what is now over half a century.* (Davidson, 2015, 2017).
- Eric Davidson's work is very aligned with the present day theory of causality. In fact he anticipated all the three levels of causality in his visionary and pioneering work on gene networks in 1969. He made a most powerful case for systematic perturbations as the only way to extract causal explanation and understanding in developmental biology (Level 2) and Level 3, Counterfactuals, is essential ‘‘but mainly just on pure logic given the requirement that a *genomically encoded control system* must exist.’’ (Britten and Davidson, 1969).

Eric Davidson's causal experimental systems developmental biology principles present as the fundamental problem that needs to be solved, the unique focus, is to infer the system-wide ‘‘explanation’’ of mechanism for the GRN using the complete set of parts (regulatory genes) and complete linkages between them, each linkage being causality-direct (i.e., without unknown intermediates as in indirect causality).

Eric Davidson provided a strong critique to the claims of causality based on associations only, that is, Level 1. He forcefully advocated the essential need to go to Levels 2 and 3 for obtaining fully causal explanation and understanding:

> *The self-described field of ‘‘systems biology’’ has given protective cover, and beyond that, in government and institutional circles a pseudo rationale, to an enterprise, the object of which is to obtain very large, solely observational databases 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. The epistemological issue that arises is not attenuated no matter how elegant the instrumentation, how clever the mathematics, and how massive the datasets. This is whether scientific causality can ever be established without perturbations of the behavior of the system, without experimental tests of logical predictions of the results of perturbations or change of conditions,* i.e., *without ‘‘experimental hypothesis testing.’’…An inbuilt scientific agnosticism and tolerance of ignorance of prior scientific knowledge 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 prior knowledge* (Davidson, 2017)

7.3. Logic, computation, and the code

> *The bottom line is that the* cis-*regulatory code specifies Boolean or discrete as well as continuous operations, all of which are directly implied in the cis-regulatory DNA sequence. It is probably true that cis-regulatory modules always execute a mix of Boolean logic operations and processing of continuous driver inputs, and their total information processing capacities can be considered the product of the unit functions mediated by the interactions at their individual target sites. Most of the individual operations the regulatory code specifies will probably turn out to be mediated by diverse transcription factors, and there will clearly be no simple one-to-one correspondence between a given functional operation and a given target site recognizing a given species of factor.* (Istrail and Davidson, 2005)

In our article ''The regulatory genome and the computer'' (Istrail et al., 2007) we presented a side-by-side comparison of the major information processing components of the electronic computer and of the regulatory genome. Our article was written as a homage, at the 50th year anniversary of the publication of John von Neumann's book ''The computer and the brain'' (von Neumann, 1958). Our article's side-by-side comparison format was the same as von Neumann's book format. John von Neumann, the founding mathematician of computer science, designed the first electronic computer architecture and logical structure and led its hardware construction; this pioneering work was inspired by his work on mathematical models for neural networks modeling the activity of the brain. In his book, von Neumann described his research program to build a general *theory of automata* (von Neumann, 1951) for artificial computers, such as the electronic computer and the robots, and at the same time for natural computers such as the brain and other biological systems such as ''genes'' as well, as they all do information processing. von Neumann defined ''the computer'' as analogous to a brain, with an input ''organ'' (sensory neurons), a memory, an arithmetical and a logical ''organ'' (associative neurons), and an output ''organ'' (motor neurons). He was convinced that both the computer and the brain do sophisticated information processing so a general new automata theory that unifies the principles of both is needed. Although he did not finish developing his new theory in his much too short life of only 54 years, he made great progress toward that goal (von Neumann, 1952). Von Neumann's lifetime work's enormous impact could be seen in every electronic computer of today, every laptop and iPhone have the so called von Neumann architecture hardware. The computer and the brain are similar—at least in the von Neumann enormously influential mathematical brain model.

The quite remarkable fact is that *the regulatory genome computes in an entirely different way* than through the von Neumann architecture hardware, although there are some areas of shared conceptual structure, such as Boolean logic foundation, and the mixture of digital and analog information processing. von Neumann viewed the computer, artificial or natural automata, either digital or analog, as a set of devices that performed arithmetical operations on numerical data in an order determined by logical control. The decisive conclusion, however, is that all three, the electronic computer, the brain, and the regulatory genome, are digital computers that compute! The regulatory genome confirms von Neumann prophetic reflection:

> *The logical approach and structure in natural automata may be expected to differ widely from those in artificial automata.* (von Neumann, 1958).

In our article about analyzing side-by-side our natural automaton, the regulatory genome, with the artificial automata, the electronic computer, we proceeded with the comparison in the following categories: fundamental of information processing versus the properties of the genomic computational system, diffusion versus wires, time versus synchrony, multiplicity of processors, memory, robustness, hardware and software, and evolvability of the genomic computer.

> *Everybody who has worked in formal logic will confirm that it is one of the technically most refractory parts of mathematics. The reason for this is that it deals with rigid, all-or-none concepts, and has very little contact with the continuous concept of the real or of complex number, that is, with mathematical analysis. Yet analysis is the technically most successful and best-elaborated part of mathematics. Thus formal logic is, by the nature of its approach, cut off from the best cultivated portions of mathematics, and forced onto the most difficult part of mathematical terrain, into combinatorics.*—John von Neumann

Von Neumann raised a series of fundamental computer science questions, with then and now, unknown answers, characterizing the brain information processing, but that capture deep features of the regulatory genome as well: building reliable ''organs'' from unreliable components, the sequential versus parallel computation and mixed paradigms, the inavailability of ''large numbers'' as it seems unable to handle precision of encoding arithmetic to handle arithmetic overflow, and the logic gates compositions but only of short (shallow) depth.

Von Neumann reflected that the brain must be using a language, arithmetic, and logic and ''statistical encoding of information'' that were radically different from those invented by humans ''the primary language of the nervous system, not the language of math.'' He proposed a research program toward a new logical and computational theory for the biological cell, envisioned as a unifying theory for the discrete mathematics and combinatorics (most refractory) and the continuous mathematics (best elaborated) through a concept of thermodynamic error. The regulatory genome distinguishes itself in the biological cell, and has the von Neumannesque program axiomatics of mixed analog and digital computing differing widely from the artificial automata.

7.4. Mathematical truth and mathematical proof: ''gene regulatory networks solved'' and the completeness theorem

The Peter-Davidson Boolean Model of GRN Equations shows that the Endomesoderm GRN is ''*GRN solved*'' (Peter et al., 2012; Faure et al., 2013; Peter and Davidson, 2015) analogous with ''exactly solved'' models of statistical mechanics and other physical systems. The complexity of the approach involved, from experimental biology to computational biology, through effective causality completeness and experimental perturbation specific of direct causality, to Boolean logic and to a new programming logic language for GRN, is truly unparalleled. To reach the ''exactly solvable'' status, as in physics models, a mathematical framework is needed where a ''completeness theorem'' is to be proved. The extraordinary results of the Peter-Davidson model provided a compelling ''proof'' of the system completeness, in the sense of the Davidsonian causal systems biology axioms framework. The intuition of this proof is easy to articulate, at least informally. The computational engine of the endomesoderm GRN is a composition of mathematical functions, each associated with one of the 50 genes of the network. It is an appropriate approximation to the truth, which the network architecture of each gene in that network required a full PhD thesis work in experimental molecular biology for causal understanding of its structure, performed in a major molecular biology laboratory; indeed, they were done mostly with Eric Davidson as thesis advisor. Each such gene expression function has a complex mathematically defined function—defined by the Peter-Davidson Boolean Network Equations Model—capturing the gene behavior in vivo. As the GRN is the composition of 50 such complex mathematical functions, composed in a complex way as well, one can reason as follow that ''near completeness'' is ''provable.'' If such a composition, regarded as a computational engine, is able to predict almost perfectly the experimental output of experiments, of extraordinary stature from the last decade of experimental work, then clearly, it has all the genes in it, and their mathematical functions as well.

The mathematical intuition of such a mathematical proof of completeness is to prove the theorem by contradiction. If we assume that, say, one of the gene is missing, then it would be very unlikely that the computational engine would computationally predict with such a precision the experimental output of those landmark very complex experiments. Proceeding this way, one can establish a contradiction and, therefore, it would complete the proof of completeness. This would establish the compelling ''Truth'' of a Logical Completeness Theorem for the Peter-Davidson Model. We can call the *Peter-Davidson methodological equation* the following:

$$\boxed{\text{Experimental Causality System Completeness} \equiv \text{Computational Exact Predictability Completeness}}$$

In words, this truly revolutionary methodological equivalence equates the ''proving'' of the biological experimental causality-based system completeness, with the ''proving'' of the exact computational predictions of the computational engine that the biological system executes in its information processing protocol. Like von Neumann reminded us, ''Truth'' is much too complicated to allow anything but approximations, and so, the Peter-Davidson GRN-solved solution is the approximate realization of the Peter-Davidson methodological equation.

However, a formal mathematical proof, in a formal mathematical framework, is needed. In such a framework, the mathematical truth of the theorem could be achieved. Fortunately, the Peter-Davidson model also provided a programming logic language for expressing the mathematical and computational formulation of each gene's behavior. And, therefore, we have the mathematical functions associated for each of the 50 genes in the network. We will need next to analyze this class of mathematical functions and their compositions in computational engines such as the endomesoderm GRN and to study their mathematical properties. The mentioned informal argument needs to be made formal with such a study, and through further research the mathematical truth of a completeness theorem is to be established. This work is in progress (Peter and Istrail, in progress) that would bring new and interesting insights into the computational properties of individual genes mathematical functions and deeper mathematical theorems for computational engines of endomesoderm GRN type.

In my essay ''Eric Davidson: Master of the Universe'' (Istrail, 2016) published in the special issue of ''Developmental Biology'' dedicated to the memory of Eric Davidson, I ended my essay with what I called the *Eric Davidson's Axioms*, eight of them, lessons learned from him, to be shared, from our 15 years collaboration, mentorship, and friendship. I would like to add now one more axiom to the list:

Axiom 9. $Biology = Causality$

8. ACKNOWLEDGMENTS

This study was supported by the National Science Foundation grants to S.I.: Grant Nos. NSF 0645955, NSF 0714609, and NSF 1048831. This article is based on the talk I gave at the Eric Davidson's Memorial Symposium ''Genomic regulation: Experiments, computational modeling and philosophy'' held at the Jacques Loeb Centre for the History and Philosophy of the Life Sciences, Ben-Gurion University, Israel, December 4–5, 2017. I would like to thank Ute Decimann, the Symposium Chair, and Center Director for her superb organization of the Symposium and exceptional hospitality. Her influential symposia, which she organized over the years at the Jacques Loeb Centre, were favorites of Eric Davidson who enthusiastically attended several of them. In fact, his thinking on the causal regulatory genome and its epistemological and philosophical themes was blossoming in the exquisite intellectual atmosphere of those symposia with its participants of the highest scientific stature. I would also like to thank Ute for her very useful comments and feedback on this article.

I would like to thank Dr. Ryan Tarpine, my former PhD student, for his really exceptional work on his PhD thesis, whose brilliance and talent are witnessed in full in the results presented in great detail in this article. It was a delight to mentor Ryan together with Eric Davidson. As Ryan moved on to another area, we will still hope to publish the Part II of the *cis*GRN-Lexicon Meta-Analysis in the near future. I thank Isabelle Peter for the excitement and inspiration that our collaborative research brings to me. Her feedback on this article was very valuable. It is due to Isabelle's extraordinary work that we have to treasure much more of Eric Davidson's science inspiration and joint publications. In the past years (at least witnessed by me since the 2012 Davidson Lab Retreat), Isabelle's work acted as a force of nature in the Davidson Lab. A force aligned well, and mutually reinforcing each other, with the bigger than life research goals, and energy and extraordinary interdisciplinary technical potency of our beloved Eric Davidson. In addition to leading most of the major GRN projects in the laboratory, she coauthored the seminal Peter-Davidson Boolean GRN Equations Model, also coauthored their seminal book; she wrote several other fundamental articles, and coadvised PhD students, and co-designed and co-led and co-taught with Eric Davidson their annual MBL ''Gene Network'' Course. I am delighted that we could continue, hopefully for many years, just in two now, the collaboration that we started, the three of us, together with our Prof. Eric Davidson.

I thank very much Ellen Rothenberg for her inspiring articles both technical and memorial about Eric Davidson. Reading her articles means a lot to me ''time found'' with Eric Davidson. I thank Dr. Derek Aguiar, my former PhD student, for his wonderful collaboration on our Davidson project, especially for his outstanding solution to the *Transcription Factors Translation Table problem* that impressed Eric very much, a problem of great interest to Eric Davidson. I also thank Derek for his comprehensive edits and feedback on this article, which were very valuable. I thank the members of the Davidson Lab from whom I learned a lot in my 15 years of visits at the Davidson Lab: Jane Rigg, Andy Cameron, Andy Ransick, Cathy Yuh, Smadar Ben-Tabou De-Leon, Jongmin Nam, and Deanna Thomas. Thanks go to my former members of the Celera Genomics group who constructed the first genome assembly of the sea urchin at 3 × coverage, and collaborated on topics related to literature extraction of experimental articles: Granger, Sutton, Jason Miller, Russell Turner, Liliana Florea, and Hagit Shatkay. Last, and most importantly: *Dear Eric, thank you for the inspiration and for everything. We will carry on. Best as always. Yours ever, Sorin.*

AUTHOR DISCLOSURE STATEMENT

The author declares that no competing financial interests exist.

REFERENCES

Balmer, J.E., and Blomhoff, R. 2009. Evolution of transcription factor binding sites in mammalian gene regulatory regions: Handling counterintuitive results. *J. Mol. Evol.* 68, 654–664.

Barolo, S., and Posakony, J.W. 2002. Three habits of highly effective signaling pathways: Principles of transcriptional control by developmental cell signaling. *Genes Dev.* 16, 1167–1181.

Britten, R., and Davidson, E. 1969. Gene regulation for higher cells: A theory. *Science* 349–357.

Cameron, R.A., Chow, S.H., Berney, K., et al. 2005. An evolutionary constraint: Strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. *Proc. Natl Acad. Sci. U. S. A.* 102, 11769–11774.

Courey, A.J., and Jia, S. 2001. Transcriptional repression: The long and the short of it. *Genes Dev.* 15, 2786–2796.

Davidson, E. 1986. *Gene Activity in Early Development,* 3rd ed. Elsevier.

Davidson, E. 2015. Genomics, ''discovery science,'' systems biology, and causal explanation: What really works? *Perspect. Biol. Med.* 58, 165–181.

Davidson, E. 2017. Systems biology, choices arising, 69–78. *In Philosophy of Systems Biology: Perspectives from Scientists and Philosophers*. Academic Press.

Davidson, E., Longabaugh, B., and Bolouri, H. 2005. Computational representation of developmental genetic regulatory networks. *Dev. Biol.* 283, 1–16.

Davidson, E., Tarpine, R., Aguiar, D., et al. (in preparation).

Davidson, E.H. 2006. *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*, 1st ed. Academic Press.

Euskirchen, G.M., Rozowsky, J.S., Wei, C-L., et al. 2007. Mapping of transcription factor binding regions in mammalian cells by ChIP: Comparison of array- and sequencing-based technologies. *Genome Res.* 17, 898–909.

Faure, E., Peter, I., and Davidson, E. 2013. A new software package for predictive gene regulatory network modeling and redesign. *J. Comput. Biol.* 419–423.

Gallo, S.M., Gerrard, D.T., Miner, D., et al. 2010. REDfly v3.0: Toward a comprehensive database of transcriptional regulatory elements in drosophila. *Nucleic Acids Res.* 39(Database), D118–D123.

Glymour, C., Sprites, P., and Scheines, R. 2000. *Causation, Prediction and Search*. MIT Press.

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

Griffith, O.L., Montgomery, S.B., Bernier, B., et al. 2007. ORegAnno: An open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.* 36(Database issue), D107–D113.

Istrail, S. 2016. Eric Davidson: Master of the universe. *Dev. Biol.* 412, S47–S54.

Istrail, S., and Davidson, E.H. 2005. Logic functions of the genomic cis-regulatory code. *Proc. Natl Acad. Sci. U. S. A.* 102, 4954–4959.

Istrail, S, De-Leon, S.B., and Davidson, E.H. 2007. The regulatory genome and the computer. *Dev. Biol.* 310, 187–195.

Istrail, S., Tarpine, R., Schutter, K., et al. 2010. Practical computational methods for regulatory genomics: A cisGRN-lexicon and cisGRN-browser for gene regulatory networks, 369–399. *In* Ladunga, I., ed. *Computational Biology of Transcription Factor Binding*, volume 674 of *Methods in Molecular Biology*. Springer Science+Business Media, LLC, Humana Press.

Jiang, C., Xuan, Z., Zhao, F., et al. 2007. TRED: A transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res.* 35(Database issue), D137–D140.

Johnson, D.S., Li, W., Gordon, D.B., et al. 2008. Systematic evaluation of variability in chip-chip experiments using predefined DNA targets. *Genome Res.* 18, 393–403.

Kamachi, Y., Uchikawa, M., Tanouchi, A., et al. 2001. Pax6 and SOX2 form a co-DNA-binding partner complex that regulates initiation of lens development. *Genes Dev.* 15, 1272–1286.

Kirkpatrick, H., Johnson, K., and Laughon, A. 2001. Repression of dpp targets by binding of brinker to mad sites. *J. Biol. Chem.* 276, 18216–18222.

Lam, N., Chesney, M.A., and Kimble, J. 2006. Wnt signaling and CEH-22/tinman/Nkx2.5 specify a stem cell niche in *C. elegans*. *Curr. Biol.* 16, 287–295.

Latchman D.S. 2008. *Eukaryotic Transcription Factors*. Academic Press.

Lin, L., He, S., Sun, J.-M., et al. 2004. Gene regulation by sp1 and sp3. *Biochem. Cell Biol.* 82, 460–471.

Longabaugh, B. 2012. BioTapestry: A tool to visualize the dynamic properties of gene regulatory networks. *Methods Mol. Biol.* 359–394.

Matys, V., Kel-Margoulis, O.V., Fricke, E., et al. 2006. TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 34(Database issue), D108–D110.

Nakao, T., and Ishizawa, A. 1994. Development of the spinal nerves in the mouse with special reference to innervation of the axial musculature. *Anat. Embryol. (Berl).* 189, 115–138.

Ostlund, G., Schmitt, T., Forslund, K., et al. 2010. InParanoid 7: New algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38(Database issue), D196–D203.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Peter, I., and Davidson, E. 2015. *Genomic Control Process: Development and Evolution*. Academic Press/Elsevier, Oxford.

Peter, I., and Istrail, S. in progress.

Peter, I., Faure, E., and Davidson, E. 2012. Feature article: Predictive computation of genomic logic processing functions in embryonic development. *Proc. Natl. Acad. Sci. U. S. A.* 109, 16434–16442.

Sayers, E.W., Barrett, T., Benson, D.A., et al. 2012. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 40(Database issue), D13–D25.

Sea Urchin Genome Sequencing Consortium, Sodergren, E., Weinstock, G.M., et al. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* 314, 941–952.

Shipley, B. 2016. *Cause and Correlation in Biology*. Cambridge University Press.

Simon, H. 1957. *Models of Man*. John Wiley and Sons, Inc.

Smith, J., and Davidson, E.H. 2008. A new method, using cis-regulatory control, for blocking embryonic gene expression. *Dev. Biol.* 318, 360–365.

Tarpine, R. 2012. A database of causality-inferred structure-function for genomic *cis*-regulatory architecture. [Ph.D. thesis]. Brown University, 76 pp.

Tarpine, R., and Istrail, S. 2009. On the concept of cis-regulatory information: From sequence motifs to logic functions, 731–742. *In* Condon, A., Harel D., Kok, J.N., Salomaa, A., and Winfree, E., eds. *Algorithmic Bioprocesses*, Natural Computing Series. Springer Berlin Heidelberg.

von Neumann, J. 1958. *The Computer and the Brain*. Yale University Press.

West, A.G., Gaszner, M., and Felsenfeld, G. 2002. Insulators: Many functions, many mechanisms. *Genes Dev.* 16, 271–288.

Williamson, J. 2005. *Bayesian Nets and Causality*. Oxford University Press.

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

Yuh, C.H., Bolouri, H., and Davidson, E.H. 2001. Cis-regulatory logic in the endo16 gene: Switching from a specification to a differentiation mode of control. *Development* 128, 617–629.

Zeitlinger, J., Zinzen, R.P., Stark, A., et al. 2007. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev.* 21, 385–390.

Zeller, R.W., Griffith, J.D., Moore, J.G., et al. 1995. A multimerizing transcription factor of sea urchin embryos capable of looping DNA. *Proc. Natl Acad. Sci. U. S. A.* 92, 2989–2993.

Zinzen, R.P., Girardot, C., Gagneur, J., et al. 2009. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* 462, 65–70.

Address correspondence to:
Dr. Sorin Istrail
Department of Computer Science
Center for Computational Molecular Biology
Brown University
115 Waterman Street, 5th Floor, Box 1910
Providence, RI 02912

E-mail: sorin_istrail@brown.edu