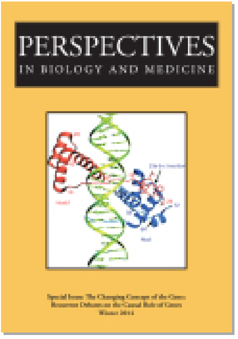
Brief Notes on the Meaning of a Genomic Control System for Animal Embryogenesis

Eric Davidson



Perspectives in Biology and Medicine, Volume 57, Number 1, Winter 2014, pp. 78-86 (Article)

Published by The Johns Hopkins University Press
DOI: 10.1353/pbm.2014.0009

Brief Notes on the Meaning of a Genomic Control System for Animal Embryogenesis

Eric Davidson

ABSTRACT This article presents some reflections on how the recently published Boolean gene regulatory network (GRN) model for sea urchin endomesoderm development affects the problem of what we can expect to know about a developmental process. The Boolean computation demonstrated that, on a system-wide level, a topological GRN model can contain sufficient regulatory information to predict in silico all the spatial and almost all the temporal processes of regulatory gene expression observed in this phase of embryonic development. Conclusions that can be drawn illuminate the general and fundamental characteristics of developmental regulatory systems, such as their innate hierarchy and their reliance on logic-processing functions. The automaton-like performance which the Boolean model displayed reflects the basic quality of genomically controlled developmental process. This quality is of course the underlying requirement for a genetically encoded developmental mechanism. The accessibility of system-wide mechanistic explanation is something new in developmental biology, and turns on their head old truisms that for a century have been implicit in science aimed at small parts of systems.

Division of Biology, Caltech, Pasadena, CA 91125.
E-mail: davidson@caltech.edu.

This work was supported by a grant from the National Institute for Child Health and Development. The ideas discussed were all developed with my colleague Isabelle S. Peter, coauthor of the automaton computation that is the subject of this commentary.

Perspectives in Biology and Medicine, volume 57, number 1 (winter 2014): 78–86.

In 2012, we published a computational automaton, based on the most comprehensive gene regulatory network (GRN) model yet available (Peter, Faure, and Davidson 2012). This model had been synthesized over the previous years from extensive experimental studies on specification mechanisms in the endomesodermal territories of the sea urchin embryo. The GRN model explicitly indicated the dynamically changing interactions occurring at the *cis*-regulatory control sequences of almost 50 genes, mostly encoding transcription factors (the proteins that specifically recognize *cis*-regulatory DNA sequence and cause expression or inactivity of the genes these sequences control). The GRN model encompasses all regulatory genes specifically expressed in four different spatial domains of diverse embryonic fate (Oliveri, Tu, and Davidson 2008; Peter and Davidson 2009, 2011; for an always-current version of the endomesoderm GRN, see http://sugp.caltech.edu/endomes). These four domains constitute about half the embryo and encompass about 30 hours of development, from early in cleavage to the onset of gastrulation.

Regulatory States and GRN Models

The GRN model thus putatively provided causal explanations for the generation of regulatory states all over the endodermal and mesodermal portions of the embryo. As used here, "regulatory state" specifically denotes the sum of regulatory gene products present in a given cell at a given time—in other words, nuclear transcription factors or the cytoplasmic mRNAs encoding them. The genomic *cis*-regulatory sequence controlling transcription of regulatory genes in time and space determines the conditions causing each such gene to be expressed or silenced, as it is this sequence that contains the DNA sequences recognized by those transcription factors controlling each gene. Thus, a gene is made to be expressed when it was previously inactive, or silenced when it was previously expressed, only when the positively or negatively acting transcription factors for which it contains target sites appear in the nucleus, and not otherwise.

Embryogenesis obviously depends exclusively on the program of gene expression in developmental space and time. But since expression of all genes in turn depends directly on regulatory state, understanding the causal origin of spatial regulatory states amounts to understanding why the embryonic process happens as it does. The map of all the putative interactions controlling regulatory gene expression (and non-expression) in the relevant spatial components of the pre-gastrular embryo can be considered a topological model, which captures the encoded regulatory logic by which spatial regulatory states are progressively elaborated in the embryo. The endomesoderm GRN implicitly proposes such a causal explanation of pre-gastrular development, in delineating the control functions that generate the regional regulatory states.

Rationale of the Automaton Computation

But could the GRN model in fact achieve such an end, which is a rather tall order? Possibly it could: we found that the multiple kinds of analyses on which the endomesoderm GRN model was based, criss-crossed and overlapped synergistically at many points of evidence; from the outset, we sought completeness by making deliberate and comprehensive efforts to include every regulatory gene predicted in the genomic sequence that is expressed in the endomesoderm during the period in question; and we also insisted experimentally that the structure of the GRN model derive primarily from system-wide perturbation analyses. Thus, the approach we took to defining the topological GRN model conforms to the basic epistemological rule of gene regulatory network construction, which is that causality can only be established by use of perturbations of the normal interrelationships among genes that reveal what these interrelations are. Network "solutions" derived solely from observations made in unperturbed conditions—that is, absent perturbation analysis—can never be taken as representations of causality in genomic biology. Examples are proposed networks based only on statistical clustering, or on other kinds of correlations, or on unperturbed kinetics. In addition to perturbation results, we also had available extensive *cis*-regulatory evidence to aid in distinguishing direct from indirect inputs. But none of the attributes of our endomesoderm GRN model, even if they succeed in establishing its causal relationships, were sufficient to demonstrate that this model indeed contains enough information to account for the overall observed progression of regulatory states.

To challenge this fundamental issue we built the computational automaton. Essentially, we asked how many black swans in the forest we might have missed, and whether what we did include in the GRN model actually predicts the overall regulatory process that transpires in life. To cut to the chase, the answer is that with the exception of a few now perfectly defined, isolated mysteries, the computation demonstrated that the GRN model indeed does suffice for prediction of progressive spatial regulatory state patterns that, overall, match observed reality. This is a new departure for system-level developmental biology, and its implications are also new. In the remainder of this article, I shall touch on some of these implications. First, however, a brief précis of the working structure of the automaton computation is in order.

Working Structure of the Automaton Computation

The automaton is a Boolean generator of hour-by-hour, on-versus-off spatial expression patterns. The term "automaton" is properly applied, in that, just as does the developmental regulatory system in vivo, the computation proceeds without external intervention, using its set rules to process the regulatory state at each step and calculate the output at the next step. It was specifically conceived to afford a direct comparison, for all the dozens of genes in the endomesoderm GRN model, between the calculated gene expression patterns and the observed, experimentally measured Boolean expression patterns. Mathematical analyses of the kinetics of transcription, of *cis*-regulatory occupancy, and of translation in sea urchin embryos shows that the limits of sensitivity of in situ hybridization closely approximates the limits of functionality of a typical transcription factor (~10 molecules of mRNA per cell) (Bolouri and Davidson 2003). Thus, if in situ hybridization reveals a regulatory gene mRNA, the transcription factor to which these mRNAs give rise is likely to have an effect on downstream genes, while if its mRNAs are too rare to be detected, it cannot provide enough transcription factor occupancy to affect downstream gene expression.

The automaton is powered by "vector equations," which for each gene state in machine-readable form the combinatorial logic that causes the gene to be transcriptionally active or inactive, according to the interactions that in the GRN model control its function. To provide the flavor of such functions, an example might be:

> Gene X =1 (i.e., is active) if At -3 Gene A=1 AND At -3 Gene B=1, AND At -3 Gene C=0, Else 0 (i.e., or else gene X is inactive)

This equation states that Gene X will be transcribed if, and only if, at three hours earlier than the time the assessment is being made, both genes A and B are being transcribed, where Genes A and B encode two activators that are specifically required to work together in the *cis*-regulatory enhancer of Gene X to activate it—but if and only if gene C, encoding a repressor, was not being transcribed at that time in those cells.

The three-hour interval is the "step time" for the sea urchin embryo at 15°, according to the kinetic analysis referred to above. The step time is defined as the interval in hours between the activation of an upstream regulatory gene encoding a transcription factor, and the time a target gene of this transcription factor begins to be transcribed. With remarkable accuracy, the computation assumed a uniform three-hour step time. Vector equations thus were built for every gene in the system, and for genes operated by multiple modules a different equation was installed for each module, according to its inputs. Every hour the spatial outputs of the whole system were computed in silico, according to the conditions set forth in these equations, and these outputs were used as inputs to compute anew the on/off state for every gene at the next hour. A key aspect of the automaton is that just as in life, the genomic regulatory sequence never changes during development, so the vector equations are always the same in every domain and at every time. In addition, a realistic algorithm was installed to represent inter-domain signaling. Thus, transcription of signal ligand genes in silico is determined as the output of vector equations for these ligand genes that were also erected according to evidence in the GRN model. The effects on the signaling target genes in the receiving tissue were made to depend on response to the signal on the part of the signal transducing regulatory factor, which acts as an activator if the cell receives the signal, or else as a default repressor. For signaling interrelations to work, the automaton was fed information encompassing the embryonic geometry through time, so that signaling will function only between embryonic cells within reach, just as in life. Initial conditions, i.e., the maternal mRNAs and other factors of regulatory significance were manually installed to get the computation going.

Over the 30-hour duration, there were over 2,770 hour-long space/time intervals in which all the genes in the model could be computed to be either on or off. In the event, except for a handful of these intervals, the automaton correctly predicted the Boolean activity states that had been observed experimentally for each of these same genes. There were only a few discrepancies, and all were temporal. All genes were computed to be on or off in the correct spatial domains, but in three cases an observed extinction of gene expression at a certain time was not predicted by the automaton, and in a handful of others genes were computed to turn on a few hours too early or too late. These prediction failures represent regulatory information missing from the GRN model used to design the vector equations. But by the same token, the sufficiency of the automaton overwhelmingly shows that the information resident in the GRN model accounts for almost the whole of the changing endomesodermal regulatory gene expression pattern in time and space, from early cleavage to gastrulation.

Fundamental Hierarchical Structure of the Embryonic Control System

This strong result specifically excludes all but a single interpretation of the informational structure of the genomically encoded regulatory system controlling (at least) this phase of embryogenesis. The information used to build the vector equations that power the automaton consists entirely of representations of interactions between transcription factors and *cis*-regulatory target sites, plus evidence of the logic transactions executed by the *cis*-regulatory modules (such as the AND logic in the above sample vector equation). But that information nonetheless suffices to generate an accurate prediction of almost the whole complex progression of regulatory states in time and space.

On the other hand, we know from many other studies that there are other levels at which gene expression can be affected, as by miRNAs, histone modifications, and DNA methylation. With respect to the basic Boolean control function of determining which genes are expressed and are not expressed in each spatial domain through time, the only way these last two statements are consistent is if all other levels of control beyond transcription factor/DNA interaction operate downstream of such interactions. There is very little room for primary control of gene expression in the developing sea urchin endomesoderm by any other mechanism. A few exceptions are allowed by the results of the automaton computation: for example, the missing inputs for the three cases where we encountered an unexplained silencing of genes could perhaps be mediated by miRNAs. But the general and strong conclusion is that genomic regulatory transactions at the DNA level directly control almost all regulatory gene expression. Furthermore, these interactions constitute the very top level of a hierarchical control apparatus.

Of Black Swans Somewhere

During the second half of the 20th century, experimental analyses of almost all processes of developmental biology lived in the shadow of Karl Popper's criticism of inductive scientific process, that all it takes is a single instance to the contrary to prove an inductive mechanistic idea wrong. In the experimental developmental biology of the last century, one could examine only a single little piece of a process within the focus of any given research project. Popper used the inductive assumption "all swans are white" as his paradigmatic example, of which the discovery of black swans in Australia provided a previously unexpected falsification.

In biological research on developmental gene regulation, almost always focused on a given gene or a given small set of genes, who could be sure where the next Australian black swan would turn up? But in our time, genomics has changed everything, particularly with respect to the fundamental problems of developmental control systems. Thus, the foundation principle of systems developmental biology, that all parts of a system must be included in mechanistic analysis, in principle offers a waterproof counter to the concern that it is extremely difficult or impossible to know if there are black swans somewhere else in the world. Control of developmental processes is mediated primarily in and by the regulatory genome, and what parts of the regulatory genome are engaged in any given such process can now be determined exactly, a priori. Thus, properly executed systems developmental biology turns the black swan either into a myth or into just another moving part. The automaton project illustrates this with respect to what is now a demonstration of the true locus of causal control. The automaton analysis is neither solely an induction nor a deduction, but both. Systems developmental biology is for this reason very different, in its epistemological quality, from that which came before.

The General Dimensions of Regulatory Complexity

The performance of the automaton indicates that the GRN model approximates completeness in terms of genes and causal linkages, in that the missing components are a small minority in number. There could be additional linkages, but they are unlikely to be required in the Boolean sense of necessity. Thus, we can now ask in numerical terms, what are the informational requirements for the process of endomesoderm specification in sea urchin embryos up to gastrulation?

A thumbnail list of the functions required for execution of this process is as follows: (1) it begins with interpretation of a few localized initial (maternally originating) inputs; (2) it mandates formation of four domains of distinct embryonic fate and regulatory state, namely, skeletogenic mesoderm, other mesoderm, anterior, and posterior endoderm, as well as the bordering ectoderm; (3) it establishes the diversification of these spatial regulatory state domains from common embryonic ancestors and sets up the boundaries between them; (4) within each domain it controls a progression from initiation of the transcriptional regulatory state in response to transient inputs, to lockdown of the regulatory domain state, to maturation and elaboration of the domain regulatory state; and (5) these functions include the accompanying expression and interpretation of signals from adjacent domains. If the control system is nearly encompassed in the automaton computation (and indirectly in the underlying GRN model), we should be able to infer the values of quantitative metrics of its complexity, in terms of genomic information.

One such metric is the number of regulatory and signaling genes required for all these functions, ~50 (many utilized in changing ways at multiple stages of the process). Another metric is the number of vector equations required, which state the conditional inputs into each *cis*-regulatory module necessary for function, ~80, roughly the equivalent of the number of *cis*-regulatory modules operating the whole system. The number of inputs figured in these equations, on the average, is about four per equation, or per *cis*-regulatory module. However, detailed investigations of given enhancers show that the number of factors bound may greatly exceed the number required just to obtain the requisite qualitative spatial outputs. The additional ones perform quantitative output level control through time, and interactions within the regulatory system, for example enhancer/promoter interactions and inter-module interactions (Peter and Davidson 2015; Yuh, Bolouri, and Davidson 2001). We can ignore these kinds of inputs in complexity considerations, however, as the same small set of general "workhorse" factors is usually bound repetitively and is likely used over again for the same purposes in many genes. These parameters are relevant to the sum amount of information processing that occurs at the *cis*-regulatory modules controlling the regulatory genes of the embryo (Istrail and Davidson 2005; Istrail, De-Leon, and Davidson 2007), here those that are engaged in the regulatory specification of the endomesoderm.

But information processing occurs at a second level of GRNs, as well as at the primary level of the *cis*-regulatory module. This is the level of the network subcircuit, consisting of from three to six genes wired together in unique ways. The term "information processing" refers as legitimately to this as to *cis*-regulatory modules, by the definition that genomically encoded subcircuits generate outputs that are distinct from any one of their multiple inputs, while these outputs are conditionally dependent both on the inputs and on the architectural structure/function characteristics of the subcircuit. The subcircuits do the regulatory jobs, such as signal interpretation, transformation of transient initial inputs into stable regulatory states, and setting and maintaining boundaries (Davidson, 2010; Peter and Davidson, 2009, 2015). As a generality, in each of the four domains of the endomesoderm GRN model there can be distinguished several such sub-circuits, to a total of at least a dozen of these little regulatory machines encoded structurally in the GRN.

We may ask how typical this GRN is, in terms of its informational complexity. A current survey of developmental GRNs that encompasses all phases of development, from embryogenesis to adult body part formation, and includes examples from flies, worms, fish, and mammals, as well as sea urchins, reveals the endomesoderm GRN to be of fairly typical complexity (Peter and Davidson 2015), that is, for the developmental process of setting up a new individual regulatory state domain**.** For instance, in our case, since we have essentially complete knowledge of what it takes at the regulatory level, we can ask what is the cost in regulatory transactions to create a new developmental domain such as that giving rise to the skeleton or endoderm of the sea urchin embryo. What we find is that the participation of perhaps 10 to 30 regulatory genes and several network subcircuits is typically required. Processes such as building an adult body part, with all of its subparts, simply add more such GRN episodes hierarchically and sequentially, so that the overall network becomes both deeper and broader as more subdomains are formed and their regulatory states are specified. But one's impression is that from flies to mice, the regulatory price, in terms of informational transactions, is about the same for a single episode of developmental specification.

Automatons as Models

The automaton computation that gives rise to these comments is unlike most computational GRN models (Peter and Davidson 2015). It is not a simulation, in which a more-or-less arbitrary mathematical form (with respect to actual mechanism) is used to generate an output that resembles a natural process. It is not designed to extract parameters or reveal interactions by statistical function fitting. Although it utilizes kinetics, it is not oriented toward rationalizing kinetic behavior. It has, instead, three other basic functions. First, it provides a direct test of the predictive completeness of the underlying experimentally determined GRN. Second, and unusually for GRN models, it deals directly in the Boolean spatial output of regulatory states such as underlie all development. Third, it has the strange quality that it runs by itself, iteratively utilizing its own output at each step to generate a new output at the next step.

The informative conclusions that devolve from the first two of these functions are briefly considered above, but the last is thought-provoking in another way. For though on a tiny stage relative to the developmental life of the whole complex animal, this computation shows how a static genomic regulatory code can be used to generate an automaton-like, unidirectional, progressive series of functions just as does the developmental process in life. In the computation, the code is represented as authentically as possible in the form of the *cis*-regulatory functionalities written in the genomic control systems for each gene, and encompassed in silico in the vector equations of the automaton. The requirement that makes it possible for the static code to serve as the animating force of the sequential computation is that the outputs of genes in the system provide the regulatory inputs of other genes in the system: that is, the requirement for use of the static code is nothing else but the existence of the gene regulatory network, by definition a set of interacting regulatory genes. The automaton behavior of the network, when recast in this form, confirms that indeed the gene regulatory network is per se the locus of the genomic regulatory code for development.

References

Bolouri, Hamid, and Eric H. Davidson. 2003. "Transcriptional Regulatory Cascades in Development: Initial Rates, Not Steady State, Determine Network Kinetics." *Proc Natl Acad Sci USA* 100 (16): 9371–76.

Davidson, Eric H. 2010. "Emerging Properties of Animal Gene Regulatory Networks." *Nature* 468 (7326): 911–20.

Istrail, Sorin, and Eric H. Davidson. 2005. "Logic Functions of the Genomic Cis-Regulatory Code." *Proc Natl Acad Sci USA* 102 (14): 4954–59.

Istrail, Sorin, Smadar Ben-Tabou De-Leon, and Eric H. Davidson. 2007. "The Regulatory Genome and the Computer." *Dev Biol* 310 (2): 187–95.

Oliveri, Paola, Qiang Tu, and Eric H. Davidson. 2008. "Global Regulatory Logic for Specification of an Embryonic Cell Lineage." *Proc Natl Acad Sci USA* 105 (16): 5955–62.

Peter, Isabelle S., and Eric H. Davidson. 2009. "Modularity and Design Principles in the Sea Urchin Embryo Gene Regulatory Network." *FEBS Lett* 583 (24): 3948–58.

Peter, Isabelle S., and Eric H. Davidson. 2011. "A Gene Regulatory Network Controlling the Embryonic Specification of Endoderm." *Nature* 474 (7353): 635–39.

Peter, Isabelle S., and Eric H. Davidson. 2015. *Genomic Control Process: Development and Evolution*. Amsterdam: Elsevier (forthcoming).

Peter, Isabelle S., Emmanuel Faure, and Eric H. Davidson. 2012. Feature Article: "Predictive Computation of Genomic Logic Processing Functions in Embryonic Development." *Proc Natl Acad Sci USA* 109 (41): 16434–42.

Yuh, Chiou-Hwa, Hamid Bolouri, and Eric H. Davidson. 2001. "Cis-regulatory Logic in the Endo16 Gene: Switching from a Specification to a Differentiation Mode of Control." *Development* 128 (5): 617–29.