Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 230 (2004) 581–590
Journal of Theoretical Biology
www.elsevier.com/locate/jtbi

# A proposal for using the ensemble approach to understand genetic regulatory networks

Stuart Kauffman[a,b,*]

[a] Cell Biology and Physiology, 1 University of New Mexico, MSC08-4750, Albuquerque, NM 8731-000, USA
[b] The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Received 8 December 2003; received in revised form 29 December 2003; accepted 31 December 2003

Available online 28 July 2004

**Abstract**

Understanding the genetic regulatory network comprising genes, RNA, proteins and the network connections and dynamical control rules among them, is a major task of contemporary systems biology. I focus here on the use of the ensemble approach to find one or more well-defined ensembles of model networks whose statistical features match those of real cells and organisms. Such ensembles should help explain and predict features of real cells and organisms. More precisely, an ensemble of model networks is defined by constraints on the "wiring diagram" of regulatory interactions, and the "rules" governing the dynamical behavior of regulated components of the network. The ensemble consists of all networks consistent with those constraints. Here I discuss ensembles of random Boolean networks, scale free Boolean networks, "medusa" Boolean networks, continuous variable networks, and others. For each ensemble, $M$ statistical features, such as the size distribution of avalanches in gene activity changes unleashed by transiently altering the activity of a single gene, the distribution in distances between gene activities on different cell types, and others, are measured. This creates an $M$-dimensional space, where each ensemble corresponds to a cluster of points or distributions. Using current and future experimental techniques, such as gene arrays, these $M$ properties are to be measured for real cells and organisms, again yielding a cluster of points or distributions in the $M$-dimensional space. The procedure then finds ensembles close to those of real cells and organisms, and hill climbs to attempt to match the observed $M$ features. Thus obtains one or more ensembles that should predict and explain many features of the regulatory networks in cells and organisms.


*Keywords:* Systems biology; Genetic regulatory networks; Ensemble approach; Statistical observables

## 1. Introduction

We have entered the post-genomic era. We know most of the genes, the coding regions, some of the *cis* regulatory sites and transcription factors, some of the protein components of cell signaling cascades that are driven by transcription and translation, and in turn feedback to regulate gene activities. Let me refer to this whole system as the genetic regulatory network. One of the outstanding problems of contemporary systems biology is to understand the structure, logic and dynamics of this network within and between cells.

In this new era, we will need new experimental, theoretical, experimental design, and data mining tools. We are, after all, attempting to understand systems with 30,000 or so genes and perhaps even more proteins, which interact in richly connected ways by differing rules. Indeed, with the new data suggesting that RNA transcripts from non-coding regions may play a regulatory role (Gibbs, 2003), the total number of "genes" may be far larger.

In this task, there are at least three theoretical approaches, each with advantages and disadvantages. The first builds detailed kinetic models of small isolated genetic circuits, (McAdams and Arkin, 1998, 1999). The virtue of this approach is that detailed comparison with experiments is possible. The disadvantage is that if one took 10 genes from a human genetic network and tried to build a detailed model of that sub-circuit, it would

*Corresponding author. Cell Biology and Physiology, 1 University of New Mexico MSC08-4750, Albuquerque, NM 8731-000, USA.
*E-mail address:* stu.kauffman@worldnet.att.net (S. Kauffman).

0022-5193/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2003.12.017

almost certainly be impacted by other genes from outside the sub-circuit, so detailed modeling would not capture all the dynamics.

The second approach I will call solving the "inverse problem". Using gene expression arrays, proteome arrays, and so forth, you show me the patterns of genes turning on and off, or proteins increasing and decreasing in abundance, and I deduce and tell you the circuitry and "logic" driving this dynamical behavior. The advantages of this approach include the fact that one is trying to deduce the actual circuitry and logic of the real network, or parts of it. The disadvantages include the fact that the data is noisy, that the inverse problem has, so far, met with limited success to my knowledge, (Hartemink et al., 2002) and will typically lead at best to a family of candidate networks. On the other hand, such a family provides hypotheses to be tested experimentally. The inverse approach will undoubtedly be an area of intense focus in the coming years.

The third approach will be termed the "ensemble approach". There is a fundamental ontological assumption underlying this approach, and it is not known if that assumption is true or false. Is it the case that the genetic network in an organism, or a species, or family of species, after 3.8 billion years of natural selection and evolution, is a highly crafted, "one off" design, brilliantly tuned by selection to achieve its functions? Or might it be the case that real genetic regulatory networks are more or less "typical" members of some class, or ensemble, of networks which selection has modified to some degree? In the latter case, we may be able to gain very considerable insight into the structure, logic, and dynamics of gene regulatory networks by examining the typical, or generic properties, of ensemble members. In addition, examination of the ensemble may help uncover any principles of organization that explain the behavior of the specific genetic networks in an organism, species, or a related set of species. In the last two cases, the diversity of real regulatory networks also forms some kind of ensemble with statistical features that can, ultimately, be studied.

At a minimum, the ensemble generic properties of ensembles that match cells can serve as useful null hypotheses about what we would expect to find, and direct further experimental work. Such experimental work is of use in itself, and also to provide more detailed information upon which to construct more refined ensembles that, iteratively, capture all we keep learning about genetic networks. That knowledge can be used to construct refined ensembles that utilize what has been learned as constraints and randomize over the remaining structural features and rules governing gene activities to create the improved ensembles.

The ensemble approach and the inverse problem can mutually inform one another. As noted, if the inverse approach suggests one or a family of networks, their statistical features can be used to construct refined ensembles reflecting those constraints. Conversely, if the ensemble approach successfully predicts many features of real cells and organisms, as described below, the ensemble(s) that do so have statistical features that can be used as constraints to narrow the search space for the inverse problem.

I will focus in this article on the ensemble approach. The first attempt, (Kauffman, 1969) was the introduction of random Boolean networks, where a gene is modeled as if it were an on–off device, and time is updated synchronously by a central clock (surely not characteristics of real cells). The first ensembles studied were networks with $N$ genes and $K$ regulatory inputs per gene, with $K = N$, and $K = 2$. This work, described momentarily, has been taken up and extended by many workers, (Kauffman, 1993; Derrida and Pommeau, 1986; Derrida and Weisbuch, 1986; Flyvbjerg, 1998; Bastolia and Parisi, 1996; Aldana et al., 2002). I will describe the behaviors of the first such networks, analyse of some of the predictions of the simplest model, and discuss a variety of currently measurable statistical features predicted by such networks. I will discuss the use of those, say, $M$, features, measured in $N$ different ensembles of networks, to create an $M$-dimensional feature space in which to locate the $N$ different ensembles. This leads to a discussion of the corresponding work to measure the same $M$ features in real cells and organisms and locate real cells and organisms as a cluster of points in the $M$-dimensional space. In turn, this leads to the use of the information about real cells to rule out some ensembles as inappropriate, find the best or better fitting ensembles which are closer to real cells in their predictions, and refine those ensembles to "hill climb" toward the real cell data. A resulting ensemble that fits the $M$ properties of real cells is a good candidate to capture many features of real genetic networks, and should offer a number of further independent predictions to be tested and the results used to refine further the ensemble in question. This iterative approach is, I hope, a step toward a productive use of the ensemble approach.

At least five initial ensembles of Boolean networks, distinguished by their "wiring" diagrams, are now available: (1) Classical random Boolean nets where each of $N$ genes receives the same number, $K$, inputs, randomly chosen among the $N$. (2) Boolean nets in which $K$ is not fixed, but exponentially distributed and each gene is randomly assigned its $K$ inputs. (3) Scale free networks in which there is a power law distribution of inputs or outputs or both from the genes, (Oosawa and Savageau, 2002; Aldana, 2003). (4) Small world networks. (5) "Medusa networks" in which a small "regulatory head" contains a network among transcription factors that regulate one another, and an acyclic directed graph hanging off that head which contains

genes that are regulated but not regulating. The sparse data available tend to support either a scale free network, (Oosawa and Savageau, 2002) in *E. coli*, or a medusa network wiring diagram in yeast, (Lee et al., 2002). In addition to these structural classes, different biases in the classes of Boolean functions can be introduced, particularly including canalysing and high $P$ functions, and certain post classes, (Shmulevich et al., 2003). Canalysing and high $P$ functions are explained below. Beyond Boolean nets, one can consider:

(1) Discrete $S$ state networks that remain synchronous (Sole et al., 2000)
(2) Boolean or $S$ state networks that are randomly asynchronous,
(3) Probabilistic Boolean nets where each gene is governed by a small set of Boolean functions given by best fits to data, and which rule governs each gene is chosen randomly from the small set at each moment (Shmulevich et al., 2002).
(4) Boolean nets in which genes have a distribution of time scales.
(5) Networks of piecewise linear equations (Glass and Hill, 1998).
(6) Networks with continuous Hill functions (Kauffman, 1970).
(7) Networks with more detailed and realistic equations representing both RNA and protein synthesis (Goutsias and Kim, 2004).

Clearly, the ensemble approach can include a wide variety of model genetic networks.

In the next section, I describe the basic features of classical Boolean networks. In the third section I describe an initial set of $M$ features that are currently largely measurable in cells or organisms, and are a start towards defining an $M$-dimensional space in which to locate real cells and organisms, as well as the initial set of $N$ ensembles mentioned above.

## 2. Classical random Boolean networks

A classical random Boolean network, as noted, has $N$ genes, each receiving $K$ inputs per gene. The $K$ inputs are chosen at random from among the $N$. Each gene is assigned at random one of the possible Boolean functions on $K$ inputs. This random construction is then fixed and the network is said to be "quenched", (Derrida and Pommeau, 1986). The point of random construction is this: There is an enormous ensemble, or class, of all possible networks with $N$ genes and $K$ inputs per gene, even for modest $N$. In order to study the typical, or generic, properties of the ensemble, the proper procedure is to sample the ensemble many times at random, examine each network sampled for a number, $M$, of properties, and thereby build up a statistical understanding of the distribution of those properties in the ensemble in question. I hasten to add that, while I have followed this approach for years, no one, myself included, thinks that real gene networks are random after 3.8 billion years of selection. On the other hand, the early hope was, and remains, that some ensembles inherently display properties that are biologically plausible and fit features of real cells and organisms. This hope seems reasonably justified, (Kauffman, 1993). If so, there is a deep implication, for those generic "self-organized" properties may then account for some of the order seen in organisms. Now it is not required by Darwinian theory, but most biologists believe that virtually all the order in organisms is due to natural selection. If the ensemble approach succeeds, we will have to confront the possibility that some of the order in organisms is self-organized, and therefore that there are at least two sources of order in biology, selection and self-organization. In turn we will be forced to analyse whatever sources of self-organization may exist, and how that self-organization mixes with and marries to natural selection.

Consider the dynamics of a random Boolean network. Each combination of gene activities, 1 or 0, across the $N$ genes constitutes a state of the network. Hence there are 2 raised to power $N$th states. Ponder this for a moment. A network with 30,000 genes would have about 10 raised to power 10,000th states. The known universe has something like 10 raised to power 80 particles. So 10 raised to power 10,000 is a truly vast number. A human is said to have on the order of 265 cell types, (Alberts et al., 1983). Evidently, not all possible states constitute cell types.

Start the network in an initial state, a combination of the on or off activities of the $N$ genes. Each gene examines the activities of its $K$ inputs, consults its Boolean function, and assumes the proper next state of activity at the next clocked moment. Thus the network progresses from a state to a state at each clocked moment. Over a succession of moments, the system traces out a trajectory in its state space. Since the number of states is finite, eventually, the system re-enters a state previously encountered on the trajectory. Since the system is deterministic, it thereafter cycles repeatedly around a loop of states called a state cycle attractor. It is a trivial property of such networks that they have at least one state cycle. The rest of the properties are highly non-trivial. Among the first of these properties are the number of states on state cycles, which might range from 1, a steady state, to 2 raised to power $N$. If the network is released from a different initial state, it might flow to the first state cycle along a trajectory, or flow to some other state cycle. Each state cycle attractor, plus the transients that flow into it, corresponds to a basin of attraction. The basins partition the state space, each state flows to a single

attractor. Hence a second property of interest is the number of state cycle attractors in the network. I will describe a variety of network properties below.

Work since the introduction of random Boolean networks has revealed that such networks behave in two broad regimes, ordered and chaotic, with a phase transition, sometimes dubbed the "edge of chaos", separating the two regimes. It is useful to visualize an hypothetical movie to characterize the two regimes. Start a network in an initial state. If a gene is turning on and off, color it green. It is "going". If a gene stops changing and is locked in the on or off state, color it red, it is "frozen". In the ordered regime, as the movie starts, all genes are green. Soon, more and more turn red, until a connected red sub-network of the genes spans, or percolates across the network, leaving behind green islands whose genes twinkle on and off. The size of the percolating red structure scales with the size of the system, $N$. In unpublished results on $K = 2$ networks achieved with Colin Hill, we found that in these networks, which lie on the phase transition hence are critical, the size distribution of green islands is a power law, whose mean increases logarithmically with $N$.

In the chaotic regime (here high dimensional chaos in a discrete deterministic system with state cycle attractors, hence not to be confused with low-dimensional chaos with positive Lyapunov exponents on chaotic attractors in continuous systems), the results of the movie are just the opposite. A green twinkling sea spans or percolates across the system leaving behind red frozen islands.

As parameters, such as $K$, or biases on the Boolean functions such as canalysing and $P$, described below, are tuned, the networks can pass through the phase transition. That transition occurs just as the green sea is breaking into green islands, (Kauffman, 1993). Thus, it is now known that $K = 2$ networks are exactly critical, lying just on the border between order and chaos, $K = 1$ networks lie in the ordered regime, while $K = 3$ or higher networks lie in the chaotic regime (Derrida and Pommeau, 1986).

## 3. New biological observables predicted by ensembles

### 3.1. The size distribution of avalanches of changing gene activities following perturbation

One of the features of both the ordered and chaotic regimes concerns the size distribution of avalanches of "damage" that spread through the network by transient perturbation of the activity of a single gene by transiently reversing its activity. Define a gene as damaged if its dynamical behavior in the perturbed system is ever different from the unperturbed system. Hence, even if the gene returns to normal behavior, once damaged, the gene is called damaged. A characteristic of the chaotic regime is that most avalanches are vast, altering the activities of 30–50%, approximately, of the $N$ genes. For a biological network with 30,000 genes that would predict that about 10,000–15,000 genes or so would commonly alter their behavior following a random perturbation to the activity of a single gene. This has never been seen to my knowledge, and is biologically implausible. In the chaotic regime, there is also a power law distribution of small avalanches. In the ordered regime, by contrast, avalanches of damage do not propagate through the red frozen sea, hence are confined to the green islands and perhaps the red genes that constitute the edges of the green islands. The size distribution of avalanches is a power law with a finite cut off that appears to scale as a square root function of $N$, (Harris et al., 2003). Thus, in humans, avalanches should typically affect one to several hundred genes. This is biologically plausible.

The size distribution of avalanches of damage constitutes the first of the $M$ properties that I wish to mention. Note that by using exogenously introduced promoters, RNAi, or other techniques, it is now perfectly possible to transiently activate or inactivate a given gene, then use gene arrays to study the size of the damage avalanche as it spreads from the initial perturbed gene. Such work faces technical issues, such as non-specific effects of introduced RNAi which may be controlled by subtractive experiments introducing nonsense RNA, the high signal to noise ratio for RNA expressed at low levels, which may be able to be handled by limiting attention to avalanches concerning genes whose levels of expression are above a defined higher signal to noise ratio, the sampling rates required to observe changes in gene activities along such avalanches, and general difficulties quantitating gene array data. Thus, we can, with effort, determine the size distribution of such avalanches for real cells, modulo the experimental issues, noted above, which also includes asynchrony of cells around the cell cycle, possible differences in gene expression even in synchronized cells, and sometimes, mixtures of more than one cell type. Use of synchronized unicellular organisms such as yeast, or synchronized cell lines for such experiments may be a preferred approach.

It is important to emphasize for this first of the $M$ properties, and the rest I will mention, that they are new biological observables that are currently largely measurable, but not the typical aspects of genetic networks upon which biologists have focused. For example, the size distribution of avalanches has not been measured to my knowledge.

### 3.2. Cell types as dynamical attractors

I now make a single biological interpretation. I noted above that humans have about 265 cell types. By

contrast, the state space of a gene network of 30,000 genes is huge. Most states of the genome cannot be cell types. In 1963 Jacob and Monod set the stage for modern thinking about why cell types are different (Jacob and Monod, 1963). They proposed a little circuit in which gene A represses B, while B represses A. This circuit has two steady state attractors, A on, B off, and A off, B on. Hence, this little circuit supports two cell types with the same set of genes. Of course it is clear that cell types are different patterns of gene activities. It is not clear that cell types, at the molecular dynamic level, are attractors. But, following Jacob and Monod, I shall propose that some or all of the alternative state cycle attractors of random Boolean nets, or alternative attractors of other dynamical models of the gene network, constitute the cell types of the organism. Aldana et al. (2002) have raised an important criticism of this concept for Boolean nets, namely that closure of a state cycle orbit is unstable to noise. This is correct. In commenting on this point, it may be useful to mention here that piecewise linear models of genetic nets also have alternative attractors. These attractors live in continuous state spaces. For molecular noise with finite variance, which seems plausible for real molecular noise in cells, the attractors may be stable to most or all such noise. This requires investigation. A possible manifestation of random molecular noise in real genetic networks may be the phenomenon of metaplasia, (Kauffman, 1993).

The hypothesis that attractors are cell types raises a number of important questions that can be addressed experimentally and theoretically. Experimentally, we can take synchronized cell line populations, perturb the current activity of randomly chosen genes, and ask, via gene arrays, whether the perturbed system returns to its former steady state or oscillatory pattern of gene activities. That is, is it typically the case that cell types are stable attractors to most small perturbations?

Note that the same question can be tested with cells asynchronously distributed around the cell cycle. One averages gene activities around an oscillatory attractor and asks for the attractor, whether the system returns to the average patterns of activity after perturbation of single model gene activities. One then asks, for asynchronous cells, if they return to their former pattern activities after transient perturbation of single gene activities. I emphasize that this is now testable, and count this homeostatic feature as another of the $M$ features that will constitute the $M$-dimensional space of properties for a hoped for productive use of the ensemble approach.

On a theoretical level, at least three further major questions are raised by the hypothesis that attractors correspond to cell types. First, is there a scaling law for the number of state cycle attractors as a function of $N$ that would predict correctly the scaling relation between the number of cell types of an organism and the number of its genes, (perhaps including the possibility of "junk" RNA playing regulatory roles (Gibbs, 2003)? In my own earlier work, I found numerical evidence that, for $K = 2$ network, the number of attractors scaled as a square root function of the number, $N$, of genes, (Kauffman, 1969). This square root scaling result has recently been shown to be wrong. I under-sampled very large state spaces and missed small basins of attraction. For small networks there is numerical evidence that the scaling law is linear for $K = 2$ networks (Aldana et al., 2002), which, as noted, lie on the phase transition between order and chaos. More recently, Socolar and I (Socolar and Kauffman, 2003) found evidence that the scaling for $K = 2$ networks is faster than linear, but is slower deeper in the ordered regime. This fact, if it holds up, means that we cannot predict the scaling law for the number of cell types as a function of the number of genes without knowing where, if at all, cells lie in the ordered regime, and also knowing the number of genes. On the other hand, the right ensemble, random Boolean nets, scale free nets, small world nets, medusa nets, or others, should be able to fit the observed scaling law, if position in the ordered regime is measured, as described next, and the number of genes is known. I count this scaling property as another of the $M$ properties.

Another feature of interest, if cell types are attractors, is to measure the overlap in gene expression patterns between all the attractors. There are several ways to define this overlap. For example, if attractors are state cycles longer than 1, then the average number of states that a gene is active can be calculated, and a real vector of activities represents that attractor. A real vector also represents the activities of genes on steady state attractors. Then both for steady state attractors and state cycles longer than 1, the Euclidian distance between each pair of attractors can be calculated. This gives a distribution of distances between attractors. What does the distribution look like? Small samples of $K = 2$ networks suggests a hierarchical pattern, with clusters of nearby attractors, linked to more distant clusters. These sparse results seem biologically plausible. However, the exact properties of such clusters are as yet unknown even for $K = 2$ random nets, let alone scale free, small world, medusa, or other ensembles, but clearly open to investigation theoretically. Meanwhile, in principle, the exact same distribution of distances is even now experimentally accessible using gene expression arrays on the cell types of one organism. Technical difficulties here include the usual problems concerning gene expression arrays, synchronous versus asynchronous populations of cells, (where asynchronous populations may be preferable to synchronous cells to average expression levels around an oscillatory pattern), and also obtaining populations of pure cell types which may be approached, if need be, by laser dissection. I count

this too as one of the $M$ features we should use. The correct ensemble should correctly predict the statistics of this distribution for realistic values of $N$ across organisms.

### 3.3. The Derrida curve: measuring position in the ordered or chaotic regime

Derrida and Pommeau (1986), showed analytically that there is a phase transition in what they termed the annealed model of random Boolean nets, between an ordered and chaotic regime. Indeed, they showed that $K = 2$ nets are critical, and $K > 2$ nets are chaotic. A Derrida plot is created as follows: Consider initial pairs of Boolean network states. Count the fraction of genes in the two states that are in different states of activity. This normalized Hamming distance is called the initial distance between the pair of states, $Dt$. Now let each state evolve to its successor state under the rules of the network. Measure the distance between the pair of successor states, call that normalized distance $Dt + 1$. In a Cartesian coordinate system, let the $X$-axis give all values of $Dt$, from 0.0 to 1.0, and the $Y$-axis give all the values of $Dt + 1$ from 0.0 to 1.0. Now plot, for any initial pair of states at a distance $Dt$, and their successor states at distance $Dt + 1$, a single point in the coordinate system corresponding to both $Dt$ on the $X$-axis, and $Dt + 1$ on the $Y$-axis. The main diagonal, $Dt = Dt + 1$ corresponds to a line where successor states are the same distance apart as the initial states. This does not mean that nothing has changed in going from initial to successor states, merely that the distance has not changed.

The main diagonal separates the ordered from the chaotic regimes. In the ordered regime, the Derrida curve, for all $Dt$, is everywhere below the main diagonal, implying that, on average, pairs of states in the ordered regime lie on trajectories which converge in state space. This convergence is the heart of homeostasis. In the chaotic regime, small $Dt$ states diverge, so that the curve passes above the main diagonal for small $Dt$, and falls below it for large enough $Dt$. Thus, nearby states in the chaotic regime, on average, lie on trajectories that diverge from one another in state space. This divergence is the first step in the avalanches of damage described above.

I count the Derrida curve as another feature among the $M$. And I emphasize that the Derrida curve, certainly for small $Dt$, is experimentally testable now by perturbing the activities of one or a modest number of genes and asking whether convergence or divergence in gene expression patterns is found in gene array data. Testing this would seem to require synchronized cell populations. Again, note that, while currently testable, the Derrida response of cells is a new biological observable never before probed, and in turn, that the curve itself probes the global dynamics of genetic regulatory networks.

Note also that, by measuring the Derrida curve for real cells, we can attempt to assess whether and where cells lie in the ordered or chaotic regime, hence we should be able to deduce the scaling law for the number of cell types in an organism as a function of the number of its genes. The missing data are a count of the number of "genes" in different organisms given the uncertainty of what counts as a gene playing a role in the network. As noted above, this scaling law is one of the $M$ properties that we can use. It should be properly predicted by the right ensemble of networks given the number of genes in organisms.

### 3.4. Pathways of differentiation

In the case of $K = 2$ random Boolean nets, attractors are stable to most small perturbations transiently reversing the activities of one gene at a time on all states of the attractor, showing homeostasis, but undergo transitions to a few other state cycles for some perturbations. I note that if cell types are attractors, then transitions between attractors model pathways of differentiation. I also note that the fact that each attractor can only undergo transition to a few other model cell types, and they to a few others, and so on, implies that differentiation from the zygote must follow branching pathways of differentiation. This property is seen in all multi-celled organisms developing from a zygote. Indeed, this fact raises the question of the role of self-organization. Do we think that natural selection has struggled for all these years to achieve branching pathways of differentiation, or is it a property so deeply embedded in the ensemble of genetic networks explored by evolution that it "shines through" and selection makes use of it? If the latter is true, then, as noted, there are two sources of order in biology, self-organization and selection.

The branching pathways of differentiation between attractors in a Boolean network in the ensembles described above, or other networks, create a directed graph showing which attractors can be perturbed to reach which attractors. The statistical features of these directed graphs can be compared to known branching pathways of differentiation in different organisms. I include the features of this directed graph among the $M$ features of interest.

A number of features of pathways of differentiation in ensembles and cells can be measured. As noted, perturbation of one attractor may yield one or several perturbations of genes on states that then lie on trajectories, or transients, that flow to another attractor. For such transients, which are models of differentiation pathways, we can measure the number of states on the transient, the gene activity differences between adjacent

states along the transient, the distribution in the number of times genes alter activities along the transients, and the distribution of the average time a gene is in a different state of activity along a transient compared to its activity in the state on the initial attractor state which was perturbed.

There are difficulties testing some of these predictions by examining the patterns of gene activities at sampled time points along pathways of differentiation in real cells. The clocked successive time moments in a Boolean net are clearly defined. But real cells have genes with continuous, somewhat noisy, levels of gene expression that wax and wane on different time scales. It is not clear how to map time moments along a Boolean net transient to time moments for real cell differentiation trajectories. This may make measuring the "length" of a transient difficult in real cells, and the differences in gene activities between adjacent states on a transient difficult.

At least the next to last and the last properties can probably be compared between ensembles and real cells. The distribution of the number of times genes change activities along a real differentiation transient is probably currently measurable by inducing differentiation of a synchronized cell line and, modulo accuracy of gene array data, obtaining a time series of arrays then counting how many times different genes change activity along the pathway, that is, how many change once, twice, and so forth. In more detail, make a fine-grained time series. For each gene, define a threshold above which it will be considered "on" and below which it will be considered "off". Using these thresholds, classify the time series moments of each gene into on and off activity levels. Count, along the time series, the number of times that gene has changed activities. Do so across the thousands of genes monitored along the differentiation transient and create the corresponding histogram of the number of genes changing activity 0, 1, 2, or more times. Compare the histogram to the histograms predicted by the Boolean net model, or those of other ensembles.

There are, with respect to Boolean nets, at least three remaining technical problems. How are the thresholds set per gene? How fine grained must the time series be to capture the changes in gene expression along a pathway of differentiation, and is that rate of sampling technically feasible? And for synchronous Boolean nets, transients of each length have well-defined histograms that, in initial results, change as the length of the transient changes. But we do not know how to map the corresponding length of the Boolean net transient to the real transient. One possible approach is to ask if the real cell histogram fails to match any of the histograms predicted by the ensemble. If so, the ensemble would appear to be a poor match to cells. Conversely, if the real cell histogram matches at least one theoretical histogram, the data are at least consistent with that ensemble.

The last property, the average time a gene is in a different state than the state it was in on the perturbed state of the attractor, should be directly measurable, and does not require mapping Boolean net time to real time. These experiments would appear to require synchronized cell populations.

I count these last two distributions as probably measurable now, hence count them among the $M$.

Two other features of transient pathways of differentiation should also be measurable currently. In $K = 2$ random Boolean nets, it is typical that there are multiple perturbations to one state cycle attractor that lead, via differentiation pathways, to the same target state cycle. Now these multiple transients may all be distinct, or some pairs or higher ordered sets may join in final common state transition pathways to the destination cell type. The distribution of numbers of transients that remain distinct versus joining to create final common pathways can be measured by perturbing synchronized cell lines in various ways and using time series arrays to measure the gene pathways engendered. In addition, there will be a distribution of fractions of pathway lengths where pathways join. This distribution can be compared to real differentiation pathways. I do count these among the $M$ properties of interest.

Finally, I would note that most perturbations of randomly chosen genes leave standard Boolean networks on states that return via some transient to the same attractor, thereby exhibiting homeostasis. All the questions and properties noted above for transients between state cycle attractors can also be asked of transients that return to the perturbed attractor. These features are also among the $M$.

### 3.5. *Functionally isolated "twinkling islands"*

The green islands of the ordered regime are deeply interesting if real cells have both a frozen red percolating component and such green twinkling islands. Such islands are the paragraph structure of the genome. Each island is functionally isolated from the other islands because no signal of gene activity changes can propagate between islands through the frozen red sea. Thus each island has its own attractors. Say there are three islands, one with two attractors, one with three attractors, and one with four attractors. Then the whole network has $2 \times 3 \times 4 = 24$ total state cycles attractors, each comprised, Chinese menu-like, by one choice of two for the first island, one of three for the second island and one of four for the third island. This suggests a combinatorial epigenetic code, and that the islands may be the major developmental decision taking sub-circuits of the network as a whole.

There are at least two ways currently available to test for the existence, and size distribution of such islands. The first uses mutual information between two genes at

either the same or different times, (Harris et al., 2003). The basic idea is that if genes are twinkling in the same island they are likely to do so in a correlated way, while if the genes are in different islands their twinkling should not be correlated. This is measured by the mutual information between two genes, which is the entropy of the first gene plus the entropy of the second gene minus their joint entropy. Note again that, using gene array data from time series of synchronized cell lines, this should be currently testable. Hopefully, we can find genes with high mutual information, hence presumably in the same island, and even evidence for the number and distribution of island sizes. This is another of the $M$ properties we might use. A second approach to find islands is to start avalanches. Such avalanches should not propagate through the red frozen structure, but should be limited to green twinkling islands and perhaps the first ring or so of red genes surrounding an island. Thus, the avalanches from two genes in the same island are likely to share downstream genes in the same island, but not affect genes in other islands. This provides evidence for which genes are in which islands, as well as evidence of causal connections between genes in the avalanche that should, in principle, be open to experimental test. Note again, that such avalanches are currently testable by perturbing the activity of single genes, and using gene arrays to measure avalanches and gene membership in each avalanche, hence overlap of gene membership in different avalanches. The existence, number and size distribution of such islands are among the $M$ features we should use. Experimentally, measuring mutual information in time series data across thousands of genes should be feasible, although it remains difficult to know how fine grained such a time series should be and whether that rate of sampling is feasible. Starting avalanches to look for overlapping downstream gene activity alterations may be feasible using pairs of genes known to be near neighbors in the regulatory network. Both mutual information and avalanches would seem to require synchronized cell lines.

### 3.6. Changes of gene expression following deletion mutations

Another property of interest is the size distribution alterations in gene activities by deletion of single genes. Serra et al. (2004), have used $K = 2$ canalysing networks, carried out such an analysis, and find that they can fit the distribution observed for several hundred deletion mutants in yeast. Experimentally, the comparisons were made between normal and knock-out asynchronous yeast cell populations using gene arrays. Theoretically, Serra et al. set randomly chosen genes to 0, mimicking a knockout, on one attractor, waited for the network to settle to a new attractor in the mutant network, and measured changes in gene expression averaged around the old and new attractors. This averaging is the natural match to asynchronous yeast cells. This is the first case I know in which a distribution predicted by an ensemble was tested and confirmed by data from cells. This case is the exemplary instance of the entire research program discussed here. Obviously, this property is among the $M$ that we should use.

### 3.7. Attractor time scales

I have not mentioned what was the first surprising feature of the ordered versus chaotic regime. The number of states on a state cycle in the chaotic regime is an exponential function of $N$, the number of genes, so soon becomes vast as $N$ increases. For even small networks and realistic times for gene activities to change, it would take billions of times the history of the universe to traverse the cycle, not a very attractive model of a cell type. Rather stunningly, for $K = 2$ networks, the median cycle length is a square root function of the number of genes, $N$, (Kauffman, 1969, 1993). So if there are 30,000 genes, the typical cycle is about 173 states long. If it takes 1–10 min to turn a gene on or off it would take 173–1730 min to traverse the cycle, a perfectly plausible biological time scale. I want to stress how stunning I still find this result, which has been confirmed analytically (Bastolia and Parisi, 1996). As noted, 30,000 genes have a state space of 2 raised to power 30,000, or 10 raised to power 10,000. Nevertheless, the system settles into tiny "black holes" of attractors in state space. If we looked at a typical network, its wiring diagram would be a mad scramble of connections, and each gene is governed by a randomly chosen Boolean function. Yet this order arises spontaneously in $K = 2$ nets, and presumably in other nets in the critical or ordered regime. Our intuitions about the requirements for order have simply been wrong.

Results comparing this square root scaling law with the obvious periodic behavior of cells, the lengths of cell cycles in organisms as a function of their DNA per cell, are interesting (Kauffman, 1969). The median cell cycle time is about a square root function of total DNA. The caveat of course, is that much of the DNA is junk DNA. So the scaling law with respect to coding regions and *cis* sites is steeper, but probably nothing like exponential. This hints that cells are, indeed, in the ordered regime. It has, as noted, recently been suggested (Gibbs, 2003) that some of the junk DNA may code for RNA that plays a regulatory role, so the total number of "genes" may be far larger than the number of structural genes. The true scaling law of cell cycle times as a function of the total number of genes may be far slower than if only structural genes are used. A proper ensemble should predict the true scaling law, and the cell cycle distribution around the median as well. So I tentatively count

this among the $M$ features. This feature is not quite testable now, for we do not know the total number of "genes" in cells.

### 3.8. Perturbing a fraction of input gene activities and the histogram of perturbations of target genes

Kim et al. (2003), have carried out an early and interesting comparison between random and scale free Boolean nets. They picked nets of different $N$, and different mean numbers of inputs, $K$. Then they picked target genes, and transiently perturbed the activities of fractions of the genes and obtained histograms for how many times the target gene was perturbed in its activity. The histograms differ between the two ensembles, random and scale free. They then trained a neural net on a subset of the data, and used it on the remaining data to predict from the histograms whether each histogram came from a scale free or a random Boolean net. As they point out, the fact that they can typically predict correctly may be related to the fact that the neural network only classified networks as random or scale free, with no finer details. In any case, the histograms they describe are useful among the $M$ properties.

There are other features of different ensembles to include among the $M$. For example, the wiring diagram of regulatory interactions is a directed graph. The statistics of indegrees and outdegrees of genes, taken as nodes of the graph, differ dramatically in random fixed or Poisson distributed $K$ networks, scale free networks with a power law distribution of indegree or outdegree or both, small world, and medusa networks. The data of Lee et al. (2002), on yeast suggest a "medusa" network, defined as a smallish head of genes (about 150 transcription factors) that mutually regulate one another, and an acyclic graph of thousands of genes that are regulated, but not regulating. The right ensemble should predict structural features such as the in and out degree distribution of the genes, the distribution of feedback loop lengths, the radius distribution, the descent distribution, and other graph features. Count these too among the $M$. Current techniques, such as those used by Lee et al. (2002), suggest that these structural properties can now be investigated.

This completes a starting list of $M$ testable features of real cells, and the distributions of those features in different ensembles, to create an $M$-dimensional space, with clusters of points representing the distribution for real cells and organisms, as well as for different ensembles. The strategy would be to use neural nets or other algorithms to measure which ensembles are closest on all $M$ features simultaneously to real cells, under the unproven assumption that real networks from different organisms belong to the same ensemble. Given that, then the idea is to refine the good ensembles, using any data that come to light as constraints the ensemble must respect, plus pure guessing and intuition, to create refined ensembles and hill climb towards the data of real cells and organisms. I stress again that these $M$ properties are not the current typical observables sought by molecular biologists, but they are valid, important, and reveal global features of the structure and dynamics of integrated genetic networks and are now largely measurable.

I also stress that ensembles that match the $M$ properties measured in real cells and organisms will be of use in a variety of ways. At a minimum, they are null hypotheses concerning real networks, which selection may have further modified. The statistical features of such ensembles, beyond the $M$, are hypotheses to be tested, both to learn more about real cell networks, and to further refine the ensemble. Further, the ensemble approach, if successful, aids the attempt to solve the inverse problem, for the ensemble statistical features are constraints that reduce the search space for solving the inverse problem. Finally, the ensemble should provide insight, analytic and otherwise, into the organizing principles that explain the generic behaviors of ensemble members, hence, hopefully, of cells.

## 4. Summary

I have suggested a new experimental approach to study implications of ensembles of Boolean or other regulatory networks. I propose that microarray technologies make possible the measurement of a considerable number of sophisticated dynamical features of real genetic networks, features that can be compared to the expectations from ensembles of theoretical networks constructed according to varying design criteria. Those ensembles that best conform to observed features could then be refined to give better fits to the data. If a network design can be found that fits reasonably well the experimental data set, then it may be that the genetic networks within real cells and organisms are members of that particular ensemble. It is an hypothesis that can and should be tested. Finally, the statistical features of successful ensembles can narrow the search space in the attempts to solve the inverse problem deducing network features from microarray and other data.

This article commemorates my dear friend, Art Winfree, with whom I shared my early career at the University of Chicago. Art was a brilliant scientist who devoted most of his work to the study of oscillatory phenomena, such as the eclosion rhythm in Drosophila melanogaster, and scroll waves in the BZ reaction, as well as cardiac phenomena. I admired him very much. He will be missed.

## Acknowledgements

This work was partially supported by NSF PHY-02-44957. I am glad to thank Ilya Shmulevich, Wei Zhang, Josh Socolar, Roberto Serra, Max Aldana, John Grefenstette, and Carsten Peterson for useful discussions.

## References

Alberts, A., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J.D., 1983. Molecular Biology of the Cell. Garland, New York.

Aldana, M., 2003. Dynamics of Boolean networks with scale-free topology. Physica D, in press.

Aldana, M., Coppersmith, S., Kadanoff, L.P., 2002. http://arXiv.org/abs/nlin.AO/020406.

Bastolia, U., Parisi, G., 1996. Closing probabilities in the Kauffman model: an annealed computation. Physica D 98, 1–25.

Derrida, B., Pommeau, Y., 1986. Random networks of automata: a simplified annealed approximation. Euophys. Lett. 1, 45–49.

Derrida, B., Weisbuch, G., 1986. Evolution of overlaps between configurations in random Boolean networks. J. Phys. 47, 1297–1303.

Flyvbjerg, H., 1998. An order parameter for networks of automata. J. Phys. A 21, L955–L960.

Gibbs, W.W., 2003. Unseen genome: gems among the junk. Scientific American 289 (5), 46–53.

Glass, L., Hill, C., 1998. Ordered and disordered dynamics in random networks. Europhys. Lett. 41, 599–604.

Goutsias, J., Kim, S., 2004. A nonlinear discrete dynamical model for transcriptional regulation: construction and properties. Biophys. J. 86, 1922–1945.

Harris, E.E., Sawhill, B., Wuensche, A., Kauffman, S., 2003. A model of transcriptional regulatory networks based on biases in the observed regulation rules. Complexity 7 (4), 23–40.

Hartemink, A., Gifford, D.K., Jaakkola, T.S., Young, R.A., 2002. Bayesian methods for elucidating genetic regulatory networks. IEEE Intell. System 17, 37–43.

Jacob, F., Monod, J., 1963. Genetic repression, allosteric inhibition and cellular differentiation. In: Locke, M. (Ed.), Cytodifferentiation and Macromolecular Synthesis. Academic Press, New York.

Kauffman, S.A., 1969. Metabolic stability and epigenesist in randomly connected nets. J. Theoret. Biol. 22, 437–1467.

Kauffman, S.A., 1970. Behavior of randomly constructed genetic nets: continuous element nets. In: Waddington, C.H. (Ed.), Towards a Theoretical Biology, Vol. 3. IUBS Symposium, Edinburgh University Press, pp. 38–46.

Kauffman, S.A., 1993. Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, New York.

Kim, S., Weinstein, J.N., Grefenstette, J.J., 2003. Inference of large-scale topology of gene regulation networks by neural nets. In: IEEE International Conference of Systems, Man, and Cybernetics, Washington, DC, USA, IEEE, New York, pp. 3969–3975.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J-B., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A., 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298, 799–804.

McAdams, H.H., Arkin, A., 1998. Simulation of prokaryotic genetic circuits. Annu. Rev. Biophys. Biomol. Struct. 27, 199–224.

McAdams, H.H., Arkin, A., 1999. Genetic regulation at the nanomolar scale. Trends Genet. 15 (2), 65–69.

Oosawa, C., Savageau, M.A., 2002. Effects of alternative connectivity on behavior of randomly constructed genetic networks. Physica D 170, 143–161.

Serra, R., Villani, M., Semeria, A., 2004. Geneticd network models and statistical properties of gene expression data in knock-out experiments. J. Theor. Biol. 227 (1), 149–157.

Shumuevich, I., Dougherty, E.R., Kim, S., Zhang, W., 2002. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 18 (2), 261–274.

Shmulevich, I., Lahdesmaki, H., Dougherty, E.R., Astola, J., Zhang, W., 2003. The role of certain Post classes in Boolean network models of genetic networks. Proc. Natl Acad. Sci. USA 100 (19), 10734–10739.

Socolar, J.E.S., Kauffman, S.A., 2003. Scaling in ordered and critical random Boolean networks. Phys. Rev. Lett. 90, 068702.

Sole, R., Luque, B., Kauffman, S.A., 2000. Phase transitions in random networks with multiple states, Santa Fe Institute Working Paper No. 00-02-011.