

Eukaryotic cells are dynamically ordered or critical but not chaotic

Ilya Shmulevich^{*†}, Stuart A. Kauffman[‡], and Maximino Aldana^{§¶}

^{*}Institute for Systems Biology, Seattle, WA 98103; [†]Institute for Biocomplexity and Informatics, University of Calgary, Calgary, AB, Canada T2N 1N4; [§]Center of Physical Sciences, National Autonomous University of Mexico, Cuernavaca, Morelos, 62210, México; and [¶]Consortium of the Americas for Interdisciplinary Science, University of New Mexico, Albuquerque, NM 87131

Communicated by Leo P. Kadanoff, University of Chicago, Chicago, IL, August 9, 2005 (received for review December 13, 2004)

Two important theoretical approaches have been developed to generically characterize the relationship between the structure and function of large genetic networks: The continuous approach, based on reaction-kinetics differential equations, and the Boolean approach, based on difference equations and discrete logical rules. These two approaches do not always coincide in their predictions for the same system. Nonetheless, both of them predict that the highly nonlinear dynamics exhibited by genetic regulatory systems can be characterized into two broad regimes, to wit, an ordered regime where the system is robust against perturbations, and a chaotic regime where the system is extremely sensitive to perturbations. It has been a plausible and long-standing hypothesis that genomic regulatory networks of real cells operate in the ordered regime or at the border between order and chaos. This hypothesis is indirectly supported by the robustness and stability observed in the phenotypic traits of living organisms under genetic perturbations. However, there has been no systematic study to determine whether the gene-expression patterns of real cells are compatible with the dynamically ordered regimes predicted by theoretical models. Using the Boolean approach, here we show what we believe to be the first direct evidence that the underlying genetic network of HeLa cells appears to operate either in the ordered regime or at the border between order and chaos but does not appear to be chaotic.

Boolean network | genetic network | Lempel–Ziv complexity

It is an essential property of living organisms to be robust against perturbations. As noted by de Visser *et al.* (1), “robustness is the invariance of phenotypes in the face of perturbation.” This robustness is observed at various levels of biological organization, ranging from gene-expression and cell-cycle regulation to the adaptation of the whole organism to new environments. Two important questions arise in the study of robustness: What are the mechanisms that produce robust behavior? How can robustness be detected and measured? The answer to these questions will depend on the level of organization under consideration. At the genetic level, several mechanisms yielding robust behavior have been proposed. In most of them, such as gene redundancy and buffering epistasis, genetic robustness is associated with a small set of genes responsible for the expression of a specific phenotypic trait (1, 2). However, there is evidence suggesting that genetic robustness might also be associated with the structural and dynamical properties of the entire genetic network and not only with a particular set of genes (3–5). In this kind of “distributed robustness,” one would have to know all of the regulatory interactions among all of the genes to actually determine the dynamical properties of the whole network. Despite significant recent progress, a complete description of a genetic regulatory network, even for well characterized organisms, remains elusive.

A different strategy to determine and characterize the genetic robustness in living organisms is to compare the general dynamical properties of model genetic regulatory networks with those present in real systems. Over the last 35 years a great variety of

model genetic networks have been developed, ranging from continuous models based on differential equations for the temporal evolution of gene product concentrations, to Boolean models where genes can be in only two states of activity (“on” or “off”) and the dynamics evolve in discrete time steps. Continuous models have been considered as more realistic and complete for the description of intracellular processes than Boolean models. However, although the Boolean approach might seem to be an oversimplification of intracellular processes, recent work has shown that this approach actually captures the essential aspects of the gene-regulation dynamics, accurately reproducing the experimental gene-expression profiles of some real organisms (3, 4, 6–9). In this work, we will adopt the Boolean approach for the modeling of genetic networks.

In this context, the genetic network of an organism is represented by a set of N Boolean variables (genes), $\sigma_1, \sigma_2, \dots, \sigma_N$, each acquiring the values 1 or 0 corresponding to the two possible states of gene activity (either the gene is expressed or it is not). To each gene σ_n we assign a set of k_n randomly chosen genes, $\sigma_{n_1}, \sigma_{n_2}, \dots, \sigma_{n_{k_n}}$, which will control the value of σ_n through the equation $\sigma_n(t+1) = f_n(\sigma_{n_1}(t), \dots, \sigma_{n_{k_n}}(t))$. In this equation, the Boolean function f_n of k_n arguments is randomly chosen from the ensemble of all possible Boolean functions such that, for each configuration of its k_n arguments, $f_n = 1$ with probability p . [For a detailed description of this model, usually known as a random Boolean network (RBN), see refs. 10–12.] One of the main characteristics of RBNs is the occurrence of dynamical attractors. Starting out from a given initial configuration, after a transient time each particular realization of a RBN will fall into a pattern of cyclic activity called, as noted, an attractor. A network can have many attractors, but at least one must exist. It has been hypothesized that the dynamical attractors in a Boolean network correspond to the different cell types or cell fates in an organism (11). In other words, the phenotypic traits of the organism are encoded in the dynamical attractors of its underlying genetic regulatory network. This hypothesis has partially been corroborated in recent works (3, 4, 6–9, 13).

Another important aspect of RBNs is the existence of two dynamical regimes, ordered and chaotic, and a phase transition between the two (12, 14). Networks operating in the ordered regime are intrinsically robust. This robustness is reflected in the dynamical stability of the attractors both under structural perturbations (mutations to the wiring or function rules) and transient perturbations (changes to the activity or state of a gene). Contrary to this, in the chaotic regime the dynamical attractors are extremely sensitive to such small perturbations. The phase transition between the ordered and chaotic regimes is governed by the value of the so-called expected network sensitivity, defined as $S = 2Kp(1-p)$, where K is the average network connectivity and p is the probability of gene expression (15). For $S < 1$ the network is in the ordered regime, whereas

Abbreviations: KL, Kullback–Leibler; LZ, Lempel–Ziv; RBN, random Boolean network.

[†]To whom correspondence should be addressed. E-mail: is@iee.org.

© 2005 by The National Academy of Sciences of the USA

the chaotic regime is attained for $S > 1$. The phase transition occurs at $S = 1$. It is important to mention that the existence of the ordered and chaotic regimes mentioned above is not particularly associated with the Boolean approach but rather with the highly nonlinear behavior exhibited by genetic regulatory systems. Ordered and chaotic dynamics are also known to occur in continuous and hybrid models of genetic regulatory systems (16).

The robustness and stability observed in living organisms at the genetic level seem to point to the possibility that their underlying genetic networks operate in the ordered regime. For instance, cellular states are very robust to a large variety of perturbations, and evidence of the existence of robust attractors representing cell types and cell functions has been reported (3, 4). Moreover, cellular oscillations, such as the cell division cycle or the respiratory cycle (17, 18), have periods astonishingly small compared with the number of genes in the genome, which is a typical characteristic of networks in the ordered regime. All of these observables can be interpreted as indirect evidence favoring the idea that gene-regulation networks operate in the ordered regime. Nevertheless, such qualitative indirect evidence is not enough to assert whether the underlying machinery of intracellular processes is robust in part because it operates in the ordered phase of some appropriate dynamical space. For example, numerical simulations show that even RBNs in the chaotic regime can have relatively short-period attractors (and vice versa). For such networks, one has to probe the entire configuration space to detect chaotic behavior. Therefore, to determine whether the dynamics of the genetic networks of real organisms are ordered or chaotic, or even more, if order and chaos have any meaningful manifestation in such systems, we need a quantitative measure of robustness that does not depend on particular attractor structures (such as the attractor length). This requirement leads us to the second important question about robustness, namely, how to measure it, especially in the absence of knowledge about the specific network structure.

Methods and Results

The gene-expression patterns of real cells can be generated by using high-throughput transcriptional profiling technologies such as cDNA microarrays. From such data we would like to determine whether the underlying genetic network is ordered, critical, or chaotic. One possible approach is to first infer the underlying architecture and logical rules of the genetic network from the time-course gene-expression data and then analyze or simulate this network to determine the regime in which it operates. Such an approach, however, can be highly problematic, in part because it requires one to solve a much more difficult problem to answer a relatively simpler question. Indeed, using today's technology, genetic network inference is highly challenging and involves issues such as small-sample error estimation, variable selection, and, in the context of our central question, a thorough understanding of the way in which the inherent network estimation uncertainty affects our estimate of the degree of "chaoticity." It would thus be more prudent to attempt to answer our question directly by using the observed dynamical network behavior without having to first infer the network structure.

To this end, we compare the complexity of time series microarray data of real cells with that of mock data generated by RBNs operating in the ordered, critical, and chaotic regimes. There are different ways to measure the "complexity" of a time sequence, most of them aiming to measure the amount of information stored in the sequence. Here, we use the Lempel–Ziv (LZ) measure of complexity (19, 20), which applies to a sequence over a finite alphabet and counts the number of substrings (words) as the sequence evolves from left to right. Although there are numerous variants of this approach, the algorithm essentially parses the sequence into shortest words that have not occurred previously and the complexity is defined

as the number of such unique words, with the possible exception of the last word, which may not be unique. Because it is known that the LZ complexity of a random binary sequence is asymptotically Gaussian with a defined mean and standard deviation being a function of the sequence length, it can also be used for statistical tests of randomness (21).

Rather than give the mathematical formulation of the LZ complexity, which can be found in ref. 19, let us illustrate it by means of an example. Consider the sequence 01100101101100100110. The first digit, 0, is a new word because we have not seen it before. So is the second digit, 1. However, the third digit, also a 1, is one that we have seen before, so we increase the length of the word by one, resulting in the new word 10. Next, starting with the fifth digit, we see that the smallest new word is 010. If we repeat this process, the sequence gets parsed as follows: 0·1·10·010·1101·100100·110, where the dots delimit the new words. Thus, the LZ complexity of this word is 7. Note that all words, with the exception of the last one, are unique and that in this definition of LZ complexity, our search for previous occurrences of a word can span across previously seen word boundaries. Repetition results in lower LZ complexity. For example, the complexity of the sequence 010101010101010101 is 3. In general, time series with repetitive or simple patterns have a low LZ complexity, whereas series with a rich pattern structure exhibit high LZ complexities. Our simulations show that RBNs operating in the ordered or critical regimes exhibit lower LZ complexities of the sequences generated by each gene due to their pattern-like behavior over time, as compared with networks in the chaotic regime, which give rise to more random gene behavior with high LZ complexities.

We measured the LZ complexities of the time series obtained from the gene-expression patterns in eukaryotic cells. In particular, the cell data we used are in the public domain (22) and are derived from synchronized (in early S phase by double thymidine block) HeLa cells, on which cDNA microarray analyses were performed for 48 time points, representing two to three synchronous cell cycles. A total of 43,198 probes were used in these experiments representing an estimated 29,621 distinct genes. After filtering out genes that were flagged as unreliable by the experimenter, the data consist of a time series of length 48, at 1-h intervals except for the first two intervals, which were 30 min, for each of 42,159 gene probes. We used no additional data filters. To analyze these data, we used the absolute intensities from synchronized HeLa cells, normalizing each array by its median expression level across all genes.

Because the mock data generated by the RBNs are binary, the HeLa gene-expression data had to be binarized. We used the well known k -means algorithm (23) with two groups (or clusters), corresponding to the two binary values.^{||} Briefly, the general method consists of clustering n data points x_1, x_2, \dots, x_n into k disjoint clusters U_1, U_2, \dots, U_k . In our case, the number of clusters is $k = 2$. The clustering is carried out in such a way as to minimize the sum

$$\sum_{m=1}^k \sum_{\{x_n \in U_m\}} (x_n - \mu_m)^2,$$

where μ_m is the centroid of the cluster U_m . We iterate the following two steps: (i) Assign each data point to the cluster that has the closest centroid (e.g., mean). (ii) Recalculate the positions of the centroids. The algorithm minimizes the sum of

^{||}There is no "correct" procedure to binarize continuous-value time series data into a binary sequence. We also tried another method reported in ref. 24, but it did not exhibit the correct behavior on random data in the sense that it produced very different LZ complexities on binarized Gaussian and Bernoulli random sequences.

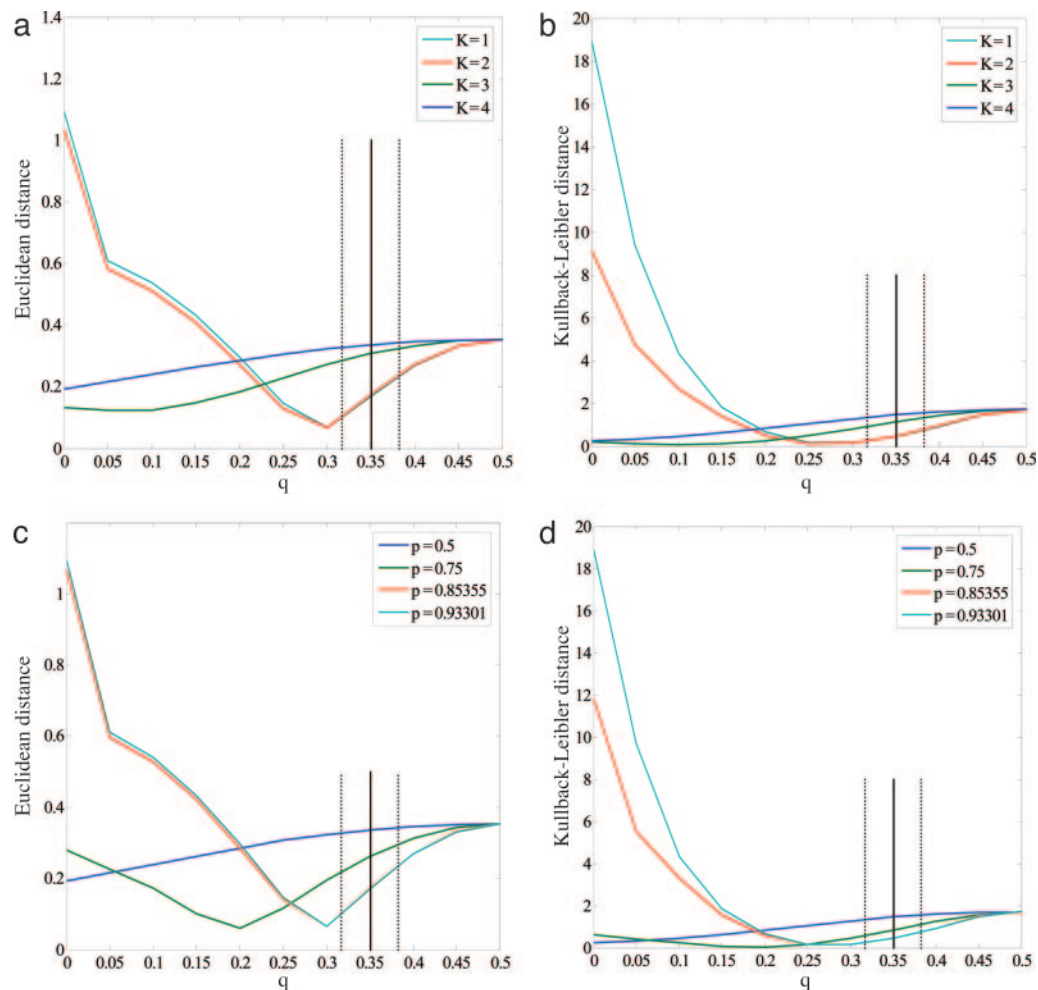


Fig. 2. Euclidean (a and c) and KL (b and d) distances between the LZ complexity distributions corresponding to the HeLa data and Boolean networks operating in different regimes, plotted against the noise probability, q . The four plots in each panel correspond either to the cases $K = 1, 2, 3,$ and 4 with $P = 0.5$ (a and b) or to the cases $P = 0.93301, 0.85355, 0.75,$ and 0.5 with $K = 4$ (c and d). The estimated noise probability $\hat{q} = 0.35$, along with its confidence interval, are indicated by vertical lines.

butions of LZ complexities rather than the LZ complexities themselves. Because there is a finite number of possible LZ complexities for a 48-element binary sequence, the resulting LZ distributions are discrete, meaning that they are essentially normalized histograms (see Fig. 1 for an example). If we represent as $P = [p_1, \dots, p_m]$ and $Q = [q_1, \dots, q_m]$ the LZ distributions corresponding to the HeLa data and mock data, respectively, it is clear that the actual form of the Q distribution will change according to whether the Boolean network is operating in the ordered, critical, or chaotic regime. (The distribution P corresponding to experimental data is given and cannot be changed.) There are different ways to compute the “distance” between two probability distributions. Here, we use the Euclidean distance, defined as $E(P, Q) = (\sum_{i=1}^m (p_i - q_i)^2)^{1/2}$, and the Kullback–Leibler (KL) distance, also called the relative entropy, defined as $D(P, Q) = \sum_{i=1}^m p_i \log(p_i/q_i)$. Note that if the distributions P and Q are identical, then $E(P, Q) = D(P, Q) = 0$. The more different the distributions P and Q , the larger the values of the Euclidean and KL distances. The purpose of using two different distances is to show that the results do not depend on the particular measure used.

Fig. 2 shows the results of comparing the experimental distribution Q with the distributions P corresponding to RBNs operating in different regimes. In each plot, either the Euclidean

distance (Fig. 2 a and c) or the KL distance (Fig. 2 b and d) is plotted against the noise probability q . The four plots in each figure correspond either to the cases $K = 1, 2, 3,$ and 4 with $p = 0.5$ (Fig. 2 a and b) or to the cases $p = 0.93301, 0.85355, 0.75,$ and 0.5 with $K = 4$ (Fig. 2 c and d), and each is an average taken over the 75 RBNs in each ensemble and for each value of q . In each figure, the estimated noise probability and its confidence interval are indicated by vertical lines.

It is evident from Fig. 2 a–d that in all cases, within the 95% confidence interval around the estimated noise level, the lowest Euclidean or KL distance, hence highest similarity between model and HeLa LZ distributions, corresponds to the ordered regime characterized by a sensitivity $S = 1/2$ ($K = 1$ with $p = 0.5$ or $p = 0.93301$ with $K = 4$), although these distances are extremely close to the ones corresponding to the critical phase $S = 1$ ($K = 2$ with $p = 0.5$ or $p = 0.85355$ with $K = 4$, respectively). Additionally, we reran all network simulations with the estimated probability of noise $\hat{q} = 0.35096$ and compared the distances (e.g., $K = 1$ vs. $K = 2$) using a t test and the Mann–Whitney test on samples generated from 75 random networks. In all cases but one, the distance corresponding to the ordered regime was smaller (with a P value $\ll 0.001$) than the distance corresponding to the critical regime, although their distributions exhibited considerable overlap:

