Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 340 (2004) 733–740
PHYSICA A
www.elsevier.com/locate/physa

The ensemble approach to understand genetic regulatory networks

Stuart Kauffman

Cell Biology and Physiology, MSC08-4750, 1 University of New Mexico, Albuquerque, NM 8731-000, USA

Received 23 January 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 to explain and predict features of real cells and organisms.

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 feed back 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 task, there are at least three theoretical approaches, each with advantages and disadvantages. The first builds detailed kinetic models of small isolated genetic circuits [1,2]. 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 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 "in-

0378-4371/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.physa.2004.05.018

verse 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 [3], and will typically lead at best to a family of candidate networks. The inverse approach will undoubtedly be an area of intense focus in the coming years. The third approach will be termed the "ensemble approach".

I will focus in this article on the ensemble approach. The first attempt [4], was the introduction of random Boolean networks, where a gene is modeled as if it were an on-off device, the network has N genes, each with K inputs, and time is updated synchronously by a central clock (surely not characteristics of real cells). This work has been taken up and extended by many workers [5–10].

At least three 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) Scale-free networks in which there is a power-law distribution of inputs or outputs or both from the genes [11,12]. (3) "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 [11], in *E. coli*, or a medusa network wiring diagram in yeast [13]. In addition to these structural classes, different biases in the classes of Boolean functions can be introduced, particularly including canalyzing and high P functions [5,14], and certain Post classes [15]. Beyond Boolean nets, one can consider:

(1) Discrete S state networks that remain synchronous [16].
(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 [17].
(4) Boolean nets in which genes have a distribution of time scales.
(5) Networks of piecewise linear equations [18].
(6) Networks with continuous Hill functions [19].
(7) Networks with more detailed and realistic equations representing both RNA and protein synthesis [20].
(8) Networks governed by stochastic equations of motion.

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

I comment that medusa networks are of interest in that evolution can add new genes to the acyclic graph of genes that are regulated by the head, but play no regulatory role, without altering the attractors of the regulatory head. Since attractors emerge as model cell types, this may be of importance.

2. Classical random Boolean networks

A classical random Boolean network has N genes, each receiving K randomly chosen inputs per gene. Each gene is assigned at random one of the possible Boolean functions on K inputs. Random construction samples at random from the enormous ensemble of all NK Boolean networks, hence allows assessment of the distributions of properties of interest.

Each combination of gene activities, 1 or 0, across the N genes constitutes a state of the network. Hence there are 2 to the Nth power states. Start the network in an initial state. 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 reenters 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.

Work since the introduction of random Boolean networks has revealed that such networks behave in two broad regimes, ordered and chaotic, with a critical phase transition 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. In the chaotic regime 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 canalyzing and P [5,14], are tuned, the networks can pass through the phase transition. That transition occurs just as the green sea is breaking into green islands [5]. Thus, it is now known that $K = 2$ networks are exactly critical, $K = 1$ networks lie in the ordered regime, while $K = 3$ or higher networks lie in the chaotic regime [6].

3. New biological observables predicted by ensembles

It is important to emphasize that the properties I will describe are new biological observables that are currently largely measurable, that they probe the integrated behaviors of networks, but not the typical aspects of genetic networks upon which biologists have focused. Space does not permit discussion of experimental approaches to measuring these properties [21].

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. 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 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, for $K = 2$ networks, as a square root function of N [14]. Thus, in humans, avalanches should typically affect one to several 100 genes. This is biologically plausible.

3.2. Cell types as dynamical attractors

I now make a single biological interpretation. Humans have about 265 cell types [22]. By contrast, the state space of a gene network of 30,000 genes is huge. Most states of the genome cannot be cell types. Following Jacob and Monod [23], 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. [10], have raised an important criticism of this concept for Boolean nets, namely that closure of a state cycle orbit is unstable to noise. This may be correct. However, if noise is rare compared to convergence onto attractors, or if error-correcting mechanisms such as majority organs are present, attractors may be stable to noise. A possible manifestation of random molecular noise in real genetic networks may be the phenomenon of metaplasia [5].

On a theoretical level, at least two 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 [24]). Recently, Socolar and I [25], 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.

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 the average number of states on an attractor that a gene is active can be calculated, and a real vector of activities represents that attractor. The Euclidean

distance between each pair of attractors can be calculated. This gives a distribution of distances that can be compared to gene expression patterns in different cell types.

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

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. Plot the results for many pairs of initial states at different Dt values. The Derrida curve is curve connecting the mean for each Dt [5].

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 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.

Note that, by measuring the Derrida curve for real cells, by perturbing gene activities transiently and measuring divergence or convergence in expression patterns, 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.

3.4. Pathways of differentiation

If cell types are attractors, then transitions between attractors induced by perturbations model pathways of differentiation. For $K=2$ networks each attractor can only flow to a few other model cell types by single gene perturbations, and they to a few others, and so on. This 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. Statistical features of such pathways, as directed graphs, can be studied in ensembles and organisms.

A number of features of pathways of differentiation in ensembles and cells can be measured: 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.

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. 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 properties can be measured in cells.

3.5. Functionally isolated "twinkling islands": co-regulated genes

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 and an attractor of the whole network is a combination of choices on each 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. By measuring mutual information between genes in the same and different islands one achieves a cluster analysis of the similarities in gene expression patterns that can be compared between ensembles and real cells.

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. [26], have used $K=2$ canalyzing networks, carried out such an analysis, and find that they can fit the distribution observed for several hundred deletion mutants in yeast.

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 [4,5]. 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 [9].

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 or distributions representing ensembles or real cells. Many other features could be added. The strategy would be to use Kullbeck–Leibler distances or other means to measure which ensembles are closest on all M features simultaneously in real cells. Given that, then the idea is to refine the good ensembles, using any data that comes to light as constraints the ensemble must respect to create refined ensembles and hill climb towards the data of real cells and organisms.

An ensemble that matches the M properties measured in real cells and organisms will be of use in a variety of ways. At a minimum, such ensembles 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.

A word about my friend Per Bak. I found him persistently creative, brilliant, open, argumentative, and delightful. I admired him a great deal, shared his last weeks at a distance, and miss him.

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

[1] H.H. McAdams, A. Arkin, Ann. Rev. Biophys. Biomol. Struct. 27 (1988) 199–224.
[2] H.H. McAdams, A. Arkin, Trends Genet. 15 (2) (1999) 65–69.
[3] A. Hartemink, D.K. Gifford, T.S. Jaakkola, R.A. Young, IEEE Intell. Systems 17 (2002) 37–43.
[4] S.A. Kauffman, J. Theor. Biol. 22 (1969) 437–467.
[5] S.A. Kauffman, Origins of Order: Self-organization and Selection in Evolution, Oxford University Press, New York, 1993.
[6] B. Derrida, Y. Pommeau, Europhys. Lett. 1 (1986) 45–49.
[7] B. Derrida, G. Weisbuch, J. Phys. 47 1297–1303.
[8] H. Flyvbjerg, J. Phys. A 21 (1998) L955–L960.
[9] U. Bastolia, G. Parisi, Physica D 98 (1996) 1–25.
[10] M. Aldana, S. Coppersmith, L.P. Kadanoff, 2002, http://arXiv.org/abs/nlin.AO/020406.
[11] C. Oosawa, M.A. Savageau, Physica D 170 (2002) 143–161.
[12] M. Aldana, Physica D, 2003, in Press.

[13] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J-B. Tagne, T.L. Volkert, E. Fraenkel, D.K. Gifford, R.A. Young, Science 298 (2002) 799–804.
[14] E.E. Harris, B. Sawhill, A. Wuensche, S. Kauffman, Complexity 7 (4) (2003) 23–40.
[15] I. Shmulevich, H. Lahdesmaki, E.R. Dougherty, J. Astola, W. Zhang, Proc. Natl. Acad. Sci. USA 100 (19) (2003) 10734–10739.
[16] R. Sole', B. Luque, S.A. Kauffman, Santa Fe Institute Working Paper no. 00-02-011, 2000.
[17] I. Shumuevich, E.R. Dougherty, S. Kim, W. Zhang, Bioinformatics 18 (2) (2002) 261–274.
[18] L. Glass, C. Hill, Europhys. Lett. 41 (1998) 599–604.
[19] S.A. Kauffman, in: C.H. Waddington (Ed.), Towards a Theoretical Biology, Vol. 3, IUBS Symposium, Edinburgh University Press, Edinburgh, 1970, pp. 38–46.
[20] J. Goutsias, S. Kim, Biophys. J., 2004, in Press.
[21] S.A. Kauffman, J .Theor. Biol., 2004, in Press.
[22] A. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson, Molecular Biology of the Cell, Garland, New York, 1983.
[23] F. Jacob, J. Monod, in: M. Locke (Ed.), Cytodifferentiation and Macromolecular Synthesis, Academic Press, New York, 1963.
[24] W.W. Gibbs, Sci. Am. 289 (5) (2003) 46–53.
[25] J.E.S. Socolar, S.A. Kauffman, Phys. Rev. Lett. 90 (2003) 068702.
[26] R. Serra, M. Villani, A. Semeria, in: W. Banzhaf, T. Christaller, P. Dittrich, J.T. Kim, J. Zeiegler (Eds.), Advances in Artificial Life. Springer Lecture Notes in Artificial Intelligence, Springer, Berlin, 2003, pp. 706–715.