SUPERPOLYNOMIAL GROWTH IN THE NUMBER OF ATTRACTORS IN KAUFFMAN NETWORKS*

Björn Samuelsson and Carl Troein

Complex Systems Division, Department of Theoretical Physics
Lund University, Sölvegatan 14A, 223 62 Lund, Sweden

The Kauffman model describes a particularly simple class of random Boolean networks. Despite the simplicity of the model, it exhibits complex behavior and has been suggested as a model for real world network problems. This work is based on an earlier paper where we introduced a novel approach to analyzing attractors in random Boolean networks. Applying this approach to Kauffman networks, we prove that the average number of attractors grows faster than any power law with system size.

PACS numbers: 89.75.Hc, 02.70.Uu

1. Introduction

We are increasingly often faced with the problem of modeling complex systems of interacting entities, such as social and economic networks, computers on the Internet, or protein interactions in living cells. Some properties of such systems can be modeled by Boolean networks. The appeal of these networks lies in the finite (and small) number of states each node can be in, and the ease with which we can handle the networks in a computer.

A deterministic Boolean network has a finite number of states. Each state maps to one state, possibly itself. Thus, every network has at least one cycle or fixed point, and every trajectory will lead to such an attractor. The behavior of attractors in Boolean networks has been investigated extensively, see *e.g.* [1–6]. For a recent review, see [7].

A general problem when dealing with a system is finding the set of attractors. For Boolean networks with more than a handful of nodes, state space is too vast to be searched exhaustively. In some cases, a majority of the attractor basins are small and very hard to find by random sampling.

One such case is the Kauffman model [8]. Based on experience with random samplings, it has been commonly believed that the number of attractors in that model grows like the square root of system size. Lately this has been brought into question [9–12].

Using an analytic approach, we are able to prove that the number of attractors grows faster than any power law [13]. The approach is based on general probabilistic reasoning that may also be applied to other systems than Boolean networks. The derivations in the analysis section are just one application of the approach. We have attempted to focus on the method and our results, while keeping the mathematical details at a minimum level needed for reproducibility using standard mathematical methods.

In 1969, Kauffman introduced a type of Boolean networks as a model for gene regulation [8]. These networks are known as N–K models, since each of the N nodes has a fixed number of inputs K. A Kauffman network is synchronously updated, and the state (0 or 1) of any node at time step t is some function of the state of its input nodes at the previous time step. An assignment of states to all nodes is referred to as a *configuration*. When a single network, a *realization*, is created, the choice of input nodes and update functions is random, although the update functions are not necessarily drawn from a flat distribution. This reflects a null hypothesis as good or bad as any, if we have no prior knowledge of the details of the networks we wish to model.

In this paper, we will first present our approach, and then apply it to Kauffman's original model, in which there are 2 inputs per node and the same probability for all of the 16 possible update rules. These 16 rules are the Boolean operators of two or fewer variables: AND, OR, TRUE, *etc.* This particular N–K model falls on the *critical line*, where the network dynamics is neither ordered nor chaotic [9, 14, 15].

2. Approach

Our basic idea is to focus on the problem of finding the number of cycles of a given length L in networks of size N. As we will see, the discreteness of time makes it convenient to handle cycles as higher-dimensional fixed point problems. Then it is possible to do the probabilistic averaging in a way which is suitable for an analytic treatment. This idea may also be expanded to applications with continuous time, rendering more complicated but also more powerful methods.

Our approach can be described as follows, given that we consider an ensemble of deterministic dynamical systems with discrete time and finite state space:

1. Write the average number of attractors as the sum of the average number of cycles of length $L = 1, 2, 3, \ldots$.

2. The problem of calculating the number of cycles of length L can be seen as an L-dimensional fixed point problem, which in turn is equivalent to calculating the partition function of a (more or less complicated) spin system.

We demonstrate our method by applying it to Kauffman networks, using four key assumptions: (i) the rules are chosen independently of each other and of N, (ii) the input nodes are independently and uniformly chosen from all N nodes, (iii) the dynamics is dominated by stable nodes, and (iv) the distribution of rules is invariant due to inversion of any set of inputs. (iv) means *e.g.* that the fraction of AND and NOR gates are the same whereas the fraction AND and NAND gates may differ. (iv) is presumably not necessary, but simplifies the calculations drastically. (iii) is expected to be valid for any non-chaotic network obeying (i) and (ii) [16]. Note that (i) does not mean that the number of inputs must be the same for every rule. We could write a general treatment of all models obeying (i)–(iv), but for simplicity we focus on the Kauffman model.

We will henceforth use $\langle C_L \rangle_N$ to denote the expectation value of the number of L-cycles over all networks of size N, with $L = 1$ referring to fixed points. The average number of fixed points, $\langle C_1 \rangle_N$, is particularly simple to calculate. For a random choice of rules, (i) and (iv) imply that the output state of the net is independent of the input state. Hence, the input and output states will on average coincide once on enumeration of all input states. This means that $\langle C_1 \rangle_N = 1$.

The problem of finding other L-cycles can be transformed to a fixed point problem. Assume that a Boolean network performs an L-cycle. Then each node performs one of 2^L possible time series of output values. Consider what a rule does when it is subjected to such time series on the inputs. It performs some Boolean operation, but it also delays the output, giving a one-step difference in phase for the output time series. If we view each time series as a state, we have a fixed point problem. $L \langle C_L \rangle_N$ is then the average number of input states (time series), for the whole network, such that the output is the same as the input.

To take advantage of assumption (iv), we introduce the notion of L-cycle patterns. An L-cycle pattern is s and s inverted, where s is a time series with period L. Let $\boldsymbol{Q}$ denote a choice of L-cycle patterns for the net, and let $P(\boldsymbol{Q})$ denote the probability that the output of the net is $\boldsymbol{Q}$. Using the same line of reasoning as for fixed points, we conclude that (i) and (iv) yield

$$\langle C_L \rangle_N = \frac{1}{L} \sum_{\boldsymbol{Q} \in \mathcal{Q}_L^N} P(\boldsymbol{Q}) \,, \tag{1}$$

where $\mathcal{Q}_L^N$ is the set of proper L-cycles of an N-node net. A proper L-cycle has no period shorter than L.

There are many ways to view Eq. (1) as a partition function. At its most detailed level, it is a sum over states where each state is a specific choice of $\boldsymbol{Q}$, couplings and rules. For the Kauffman model, it is convenient to work with the state space $\boldsymbol{Q} \in \mathcal{Q}_L^N$ and thus average out the choices of couplings and rules for each term in the partition function. This is equivalent to a system with 2^{L-1}-state spins where each spin interacts with all other spins.

The latter picture is used in order to evaluate Eq. (3) by simulated tempering [17,18], which is a method designed for simulation of thermodynamic systems. Simulated tempering is a Monte-Carlo scheme where the temperature is a dynamical parameter. This makes it easier to cross free-energy barriers and provides a convenient way to calculate ratios between values of the partition function at different temperatures. Since we know the total number of states, we know the partition function at infinite temperature and simulated tempering reveals the ratio between this number and the value of Eq. (3).

3. Analytic calculations

Assumption (ii) implies that $P(\boldsymbol{Q})$ is invariant under permutations of the nodes. Let $\boldsymbol{n} = (n_0, \ldots, n_{m-1})$ denote the number of nodes expressing each of the $m = 2^{L-1}$ patterns. For n_j, we refer to j as the pattern index. For convenience, let the constant pattern have index 0. Then

$$\langle C_L \rangle_N = \frac{1}{L} \sum_{\boldsymbol{n} \in \mathcal{P}_L^N} \binom{N}{\boldsymbol{n}} P(\boldsymbol{Q}) \,, \tag{2}$$

where $\binom{N}{\boldsymbol{n}}$ denotes the multinomial $N!/(n_0! \cdots n_{m-1}!)$ and $\mathcal{P}_L^N$ is the set of partitions $\boldsymbol{n}$ of N such that $\boldsymbol{Q} \in \mathcal{Q}_L^N$. That is, $\boldsymbol{n}$ represents a proper L-cycle.

What we have this far is merely a division of the probability of an L-cycle into probabilities of different flavors of L-cycles. Now we assume that each node has 2 inputs. Then, we get a simple expression for $P(\boldsymbol{Q})$ that inserted into Eq. (2) yields

$$\langle C_L \rangle_N = \frac{1}{L} \sum_{\boldsymbol{n} \in \mathcal{P}_L^N} \binom{N}{\boldsymbol{n}} \prod_{\substack{0 \le j < m \\ n_j \ne 0}} \left(\sum_{0 \le l_1, l_2 < m} \frac{n_{l_1} n_{l_2}}{N^2} (P_L)_{l_1 l_2}^j \right)^{n_j} \,, \tag{3}$$

where $(P_L)_{l_1 l_2}^j$ denotes the probability that the output pattern of a random 2-input rule has index j, given that the input patterns have the indices l_1 and l_2, respectively. Note that Eq. (3) is an exact expression for the average number of proper L-cycles in an N-node random Boolean network which satisfies the assumptions (i), (ii), (iv), and where each node has 2 inputs.

From now on, we only consider the Kauffman model, meaning that we also restrict the distribution of rules to be uniform. It is instructive to explore some properties of $(P_L)_{l_1 l_2}^j$; these will also be needed in the following calculations. We see that

$$(P_L)_{00}^0 = 1\,, \qquad (P_L)_{l_1 0}^0 = \tfrac{1}{2} \qquad \text{and} \qquad (P_L)_{l_1 l_2}^0 \geq \tfrac{1}{8}\,, \tag{4}$$

for $1 \leq l_1, l_2 < m$. Further, we note that for a given $j \neq 0$, $(P_L)_{l_1 0}^j$ has a non-zero value for exactly one $l_1 \in \{1, \ldots, m-1\}$. Let $\phi_L(j)$ denote that value of l_1. We can see ϕ_L as a function that rotates an L-cycle pattern one step backwards in time. With this in mind we define $\phi_L(0) = 0$. Now, we can write

$$(P_L)_{l_1 0}^j = \tfrac{1}{2}\delta_{l_1 \phi_L(j)}\,, \tag{5}$$

for $1 \leq j < m$. (δ is the Kronecker delta.)

We can view ϕ_L as a permutation on the set $\{0, \ldots, m-1\}$. Thus, we divide this index space into permutation cycles which are sets of the type $\{j, \phi_L(j), \phi_L \circ \phi_L(j), \ldots\}$. We refer to these permutation cycles as invariant sets of L-cycles. Let $\rho_L^0, \ldots, \rho_L^{H_L-1}$ denote the invariant sets of L-cycles, where H_L is the number of such sets. For convenience, let ρ_0 be the invariant set $\{0\}$. If two L-cycle patterns belong to the same invariant set, they can be seen as the same pattern except for a difference in phase.

We want to find the behavior of $\langle C_L \rangle_N$, for large N, by approximating Eq. (3) with an integral. To do this, we use Stirling's formula $n! \approx (n/e)^n \sqrt{2\pi n}$ while noting that the boundary points where $n_j = 0$ for some j can be ignored in the integral approximation. Let $x_j = n_j/N$ for $j = 0, \ldots, m-1$ and integrate over $\boldsymbol{x} = (x_1, \ldots, x_{m-1})$. x_0 is implicitly set to $x_0 = 1 - \sum_{j=1}^{m-1} x_j$. We get

$$\langle C_L \rangle_N \approx \frac{1}{L} \left(\frac{N}{2\pi} \right)^{(m-1)/2} \int\limits_{0 < x_0, \ldots, x_{m-1}} d\boldsymbol{x} \; \frac{e^{N f_L(\boldsymbol{x})}}{\prod_{j=0}^{m-1} \sqrt{x_j}}\,, \tag{6}$$

where

$$f_L(\boldsymbol{x}) = \sum_{j=0}^{m-1} x_j \ln \left(\frac{1}{x_j} \sum_{0 \leq l_1, l_2 < m} x_{l_1} x_{l_2} (P_L)_{l_1 l_2}^j \right)\,. \tag{7}$$

Eq. (7) can be seen as an average $\langle \ln X \rangle$, where X is the expression inside the parentheses. Hence, the concavity of $x \mapsto \ln x$ gives $f_L(\boldsymbol{x}) = \langle \ln X \rangle \leq \ln \langle X \rangle = 0$ with equality if and only if

$$x_j = \sum_{0 \leq l_1, l_2 < m} x_{l_1} x_{l_2} (P_L)^j_{l_1 l_2} \,, \tag{8}$$

for all $j = 0, \ldots, m - 1$.

Note that Eq. (8) can be interpreted as a mean-field equation of the model. Using Eq. (8) for $j = 0$ and Eq. (4) we see that $f_L(\boldsymbol{x})$ comes arbitrarily close to zero only in the vicinity of $\boldsymbol{x} = 0$, and for large N, the relevant contributions to the integral in Eq. (6) come from this region. Thus, the dynamics of the net is dominated by stable nodes, in agreement with [11, 16]. This means that assumption (iii) is satisfied by the Kauffman model. Using Eqs. (4) and (5), a Taylor-expansion of $f_L(\varepsilon \boldsymbol{x})$ yields

$$f_L(\varepsilon \boldsymbol{x}) = \varepsilon \sum_{j=1}^{m-1} x_j \ln \frac{x_{\phi(j)}}{x_j} + \varepsilon^2 \sum_{j=0}^{m-1} x_j \frac{\boldsymbol{x} \cdot A_L^j \boldsymbol{x}}{x_{\phi(j)}}$$
$$- \frac{\varepsilon^3}{2} \sum_{j=1}^{m-1} x_j \left(\frac{\boldsymbol{x} \cdot A_L^j \boldsymbol{x}}{x_{\phi(j)}} \right)^2 + O\left(\varepsilon^4\right) \,, \tag{9}$$

where $(A_L^j)_{l_1 l_2} = (P_L)^j_{l_1 l_2} - \frac{1}{2}(\delta_{l_1 \phi(j)} + \delta_{l_2 \phi(j)})$.

The first order term of Eq. (9) has 0 as its maximum and reaches this value if and only if $x_{\phi_L(j)} = x_j$ for all $j = 1, \ldots, m - 1$. The second order term is zero at these points, while the third order term is less than zero for all $\boldsymbol{x} \neq 0$. Hence, the first and third order terms are governing the behavior for large N. Using the saddle-point approximation, we reduce the integration space to the space where the first and second order terms are 0. Let $z_h = N^{1/3} \sum_{j \in \rho_L^h} x_j$ for $h = 1, \ldots, H_L - 1$ and let $(P_L')^h_{k_1 k_2}$ denote the probability that the output pattern of a random rule belongs to ρ_L^h, given that the input patterns are randomly chosen from $\rho_L^{k_1}$ and $\rho_L^{k_2}$, respectively. Thus, we approximate Eq. (6) for large N as

$$\langle C_L \rangle_N \approx \alpha_L \beta_L N^{\gamma_L} \,, \tag{10}$$

where

$$\alpha_L = \left(L \prod_{h=1}^{H_L-1} \left| \rho_L^h \right| \right)^{-1} \left(\frac{1}{2\pi} \right)^{(H_L-1)/2}, \tag{11}$$

$$\beta_L = \int\limits_{0 < z_1, \ldots, z_{H_L-1}} d\boldsymbol{z} \; \frac{\exp\left(-\frac{1}{2} \sum_{h=1}^{H_L-1} \frac{1}{z_h} \left(\boldsymbol{z} \cdot B_L^h \boldsymbol{z} \right)^2 \right)}{\prod_{h=1}^{H_L-1} \sqrt{z_h}}, \tag{12}$$

$$\gamma_L = \frac{H_L - 1}{3}, \tag{13}$$

and $(B_L^h)_{k_1 k_2} = (P_L')_{k_1 k_2}^h - \frac{1}{2}(\delta_{k_1 h} + \delta_{k_2 h})$. ($|\rho|$ denotes the number of elements of the set ρ.)

The error in Eq. (10) is presumably dominated by the error introduced by the continuous approximation in Eq. (6). Since

$$\sum_{h=1}^{H_L-1} z_h = N^{1/3} \sum_{j=1}^{m-1} x_j = N^{-2/3} \sum_{j=1}^{m-1} n_j, \tag{14}$$

we expect the length scale of the relevant region in Eq. (6) to scale like $N^{2/3}/m$. We can assume that the relative error is inversely proportional to this length scale and proportional to the number of dimensions. This yields $N \gg m^3 \sim 8^L$ as a rough estimate for the limit where Eq. (10) is useful as an approximation.

Though our calculations can not reveal an expression for the asymptotic growth of the total number of attractors, it is clear that the growth is superpolynomial. To show this, we observe that H_L grows rapidly with L. The number of elements in an invariant set of L-cycle patterns is a divisor of L. If an invariant set consists of only one pattern, it is either the constant pattern, or the pattern with alternating zeros and ones. The latter is only possible if L is even. Thus, $H_L - 1 \geq (2^{L-1} - 1)/L$, with equality if L is a prime number > 2. Applying this conclusion to Eqs. (13) and (10), we see that for any power law N^ξ, we can choose an L such that $\langle C_L \rangle_N$ grows faster than N^ξ.

H_L can be calculated using Burnside's formula in group theory. Let ϕ_L^i denote $\phi_L \circ \cdots \circ \phi_L$ with i factors and let $I_a = \{0, \ldots, a-1\}$ for non-negative integers a. Then, the set $\Phi_L = \{\phi_L^i | i \in I_L\}$ is a group under function composition $\circ$. Let Φ_L act on I_m with the group action $(\phi_L^i, j) \mapsto \phi_L^i(j)$ for

$i \in I_L, j \in I_m$. In this formulation, H_L is the number of invariant orbits in I_m under the group action of Φ_L. According to Burnside's formula, this is given by

$$H_L = \frac{1}{L} \sum_{i=0}^{L-1} \left| \{ j \in I_m | \phi_L^i(j) = j \} \right| . \tag{15}$$

$\phi_L^i(j) = j$ is equivalent to $\phi_L^{\gcd(L,i)}(j) = j$, which is true if and only if the L-cycle pattern with index j consists of repeated sequences of length $\gcd(L,i)$, possibly (if $L/\gcd(L,i)$ is even) with every second sequence inverted. ($\gcd(L,i)$ denotes the greatest common divisor of L and i.) This yields

$$H_L = \frac{1}{L} \sum_{i=0}^{L-1} 2^{\gcd(L,i) - \mathrm{odd}[L/\gcd(L,i)]} , \tag{16}$$

where $\mathrm{odd}(i) = 1$ if i is odd and $\mathrm{odd}(i) = 0$ otherwise. Applying $1 \leq \gcd(L,i) \leq L/2$ for $1 \leq i < L$ on Eq. (16) gives

$$\frac{2^{L-1} + L - 1}{L} \leq H_L < \frac{2^{L-1} + (L-1)2^{L/2}}{L} , \tag{17}$$

where the lower bound is the same result as above and holds with equality for primes > 2.

4. Numerical results

We have written a set of programs to compute the number of L-cycles in Kauffman networks, Eq. (3), both by complete enumeration and using Monte Carlo methods, and tested them against complete enumeration of the networks with $N \leq 4$. In order to push the limit for Monte Carlo evaluation of Eq. (3), we used the simulated tempering scheme [17, 18]. This makes it easier to cross the free-energy barrier between the bulk of all L-cycles and those that are significant in Eq. (3).

The results for $2 \leq L \leq 6$ are shown in Fig. 1, along with the corresponding asymptotes. The asymptotes were obtained by Monte Carlo integration of Eq. (12). For low N, $\langle C_L \rangle_N$ is dominated by the boundary points neglected in Eq. (6), and its qualitative behavior is not obvious in that region.

A straightforward way to count attractors in a network is to simulate trajectories from random configurations and count distinct target attractors. As has been pointed out in [11], this gives a biased estimate. Simulations in [8] with up to 200 trajectories per network indicated $\sqrt{N}$ scaling with system size, whereas [11] reported linear behavior with 1000 trajectories. That the true growth is faster than linear has now been firmly established [12, 13].
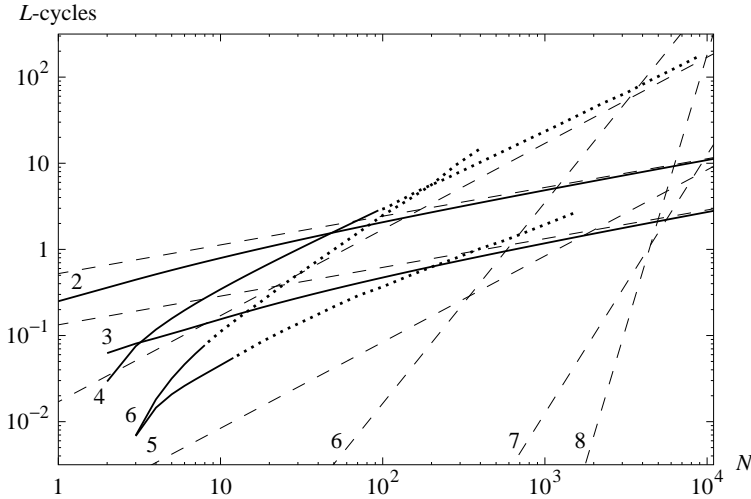
Fig. 1. The number of L-cycles as functions of the network size for $2 \le L \le 6$. The numbers in the figure indicate L. Dotted lines are used for values obtained by Monte Carlo summation, with errors comparable to the line width. The asymptotes for $N \le 8$ have been included as dashed lines. Their slopes are $\gamma_2 = \gamma_3 = \frac{1}{3}$, $\gamma_4 = \gamma_5 = 1$, $\gamma_6 = \frac{7}{3}$, $\gamma_7 = 3$, and $\gamma_8 = \frac{19}{3}$.

To closer examine the problem of biased undersampling, we have implemented the network reduction algorithm from [11], and gathered statistics on networks with $N \lesssim 10^4$. We repeated the simulations for different numbers of trajectories τ, with $100 \le \tau \le 10^5$. For each N and τ, 10^3 network realizations were examined, and we discarded configurations if no cycle was found within 2^{13} time steps. The results are summarized in Fig. 2(a).

For $\tau = 100$, the number of attractors follows $\sqrt{N}$ remarkably well, considering that $\tau = 10^3$ gives the quite different N behavior seen in [11]. If we extrapolate wildly from a log–loglog plot, $e^{0.3\sqrt{N}}$ fits the data rather well.

As another example of how severe the biased undersampling is, we have included a plot of the number of 2-cycles found in the simulations (Fig. 2(b)). The underlying distribution is less uniform than for the total number of attractors, so the errors are larger, but not large enough to obscure the qualitative behavior. The number of observed 2-cycles is close to $\langle C_2 \rangle_N$ for low N, but as N grows, a vast majority of them are overlooked. As expected, this problem sets in sooner for lower τ, although the difference is not as marked as it is for the total number of attractors.
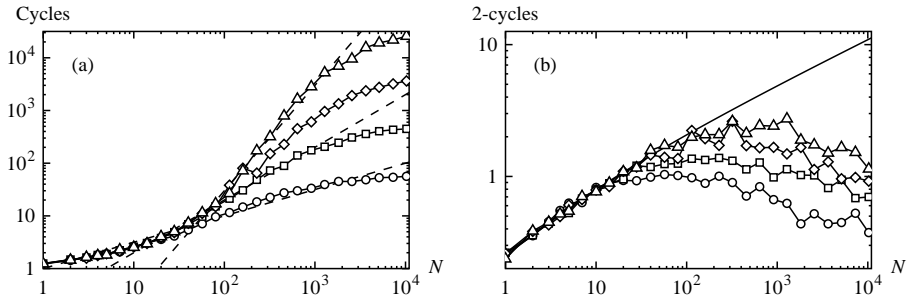
Fig. 2. The number of observed attractors (a) and 2-cycles (b) per network as functions of N for different numbers of trajectories τ: 100(circles), 10^3(squares), 10^4(diamonds), and 10^5(triangles). In (a), the dashed lines have slopes 0.5, 1, and 2. In (b) the solid line shows $\langle C_2 \rangle_N$, the true number of 2-cycles. In the high end of (b), 1σ corresponds to roughly 20%.

5. Summary

We have introduced a novel approach to analyzing attractors of random Boolean networks. By applying it to Kauffman networks, we have proven that the number of attractors in these grows faster than any power law with network size N. This is in sharp contrast with the previously cited $\sqrt{N}$ behavior, but in agreement with recent findings.

For the Kauffman model, we have derived an expression for the asymptotic growth of the number of L-cycles, $\langle C_L \rangle_N$. This expression is corroborated by statistics from network simulations. The simulations also demonstrate that biased undersampling of state space is a good explanation for the previously observed $\sqrt{N}$ behavior.

REFERENCES

[1] S. Bornholdt, K. Sneppen, *Phys. Rev. Lett.* **81**, 236 (1998).

[2] N. Lemke, J.C.M. Mombach, B.E.J. Bodmann, *Physica A* **301**, 589 (2001).

[3] A. Bhattacharjya, S. Liang, *Phys. Rev. Lett.* **77**, 1644 (1996).

[4] R.J. Bagley, L. Glass, *J. Theor. Biol.* **183**, 269 (1996).

[5] J.J. Fox, C.C. Hill, *Chaos* **11**, 809 (2001).

[6] C. Oosawa, M.A. Savageua, *Physica D* **170**, 143 (2002).

[7] M. Aldana-Gonzalez, S. Coppersmith, L.P. Kadanoff, *Boolean Dynamics with Random Couplings*, in *Perspectives and Problems in Nonlinear Science*, Eds. E. Kaplan, J.E. Marsden, K.R. Sreenivasan, Springer, 2003.

[8] S.A. Kauffman, *J. Theor. Biol.* **22**, 437 (1969).

[9] U. Bastolla, G. Parisi, *J. Theor. Biol.* **187**, 117 (1997).

[10] U. Bastolla, G. Parisi, *Physica D* **115**, 219 (1998).

[11] S. Bilke, F. Sjunnesson, *Phys. Rev.* **E65**, 016129 (2001).

[12] J.E.S. Socolar, S.A. Kauffman, *Phys. Rev. Lett.* **90**, 068702 (2003).

[13] B. Samuelsson, C. Troein, *Phys. Rev. Lett.* **90**, 098701 (2003).

[14] B. Derrida, Y. Pomeau, *Europhys. Lett.* **1**, 45 (1986).

[15] B. Derrida, D. Stauffer, *Europhys. Lett.* **2**, 739 (1986).

[16] H. Flyvbjerg, *J. Phys.* **A21**, L955 (1988).

[17] A.P. Lyubartsev, A.A. Martsinovski, S.V. Shevkunov, P.N. Vorontsov-Velyaminov, *J. Chem. Phys.* **96**, 1776 (1992).

[18] E. Marinari, G. Parisi, *Europhys. Lett.* **19**, 451 (1992).