Evolving populations of random boolean networks

Ney Lemke[1], José C. M. Mombach[1], and Bardo E. J. Bodmann[1]

[1] Centro de Ciências Exatas e Tecnológicas,
Universidade do Vale do Rio dos Sinos,
93022-000 São Leopoldo, RS, Brazil
{lemke,mombach,bardo}@exatas.unisinos.br

Abstract. We investigate the adaptation of Random Boolean Networks that are a model for regulatory gene networks. The model considers a general genetic algorithm and a fitness function that takes into account the full network dynamical behavior.Our simulations shows a fast decrease on algorithm perfomance when we consider larger networks or networks with more complex dynamical behavior. Finally, we discuss a scenario that describes the adaptation on the proposed fitness landscape.

1. Introduction

The cell genome stores all information required for the construction and function of an organism. Its basic units, the genes, interact with each other to perform these tasks in an orchestrated way. Kauffman proposed a cellular automata model for the functioning genome where the dynamics are due to mutuous activations and inactivations of regulatory genes represented by a network of boolean variables. A Kauffman NK network is a set of N boolean variables each one connected randomly to K other variables in the set. The state of each variable is determined from a random logical function of the K inputs.

The underlying dynamics are setup as follows: The state of a variable, S_i at instant $(t+1)$ is determined from a logical function (B_i) evaluating the states of the K input variables connected to it at instant t,

$$S_i(t+1) = B_i(S_{j_1}(t), S_{j_2}(t), S_{j_3}(t), ..., S_{j_K}(t)). \quad (1)$$

Since the phase space of the networks is discrete and finite, the attractors are cycles with period length between 1 and 2^N (the total number of states of a network of size N). The networks are known to possess distinct dynamical behaviors dependent on K including a dynamical transition that separates an ordered phase at $K = 2$ from a disordered phase for $K > 2$. For $K = 2$ the average period length and the number of cycles scale with $\sim \sqrt{N}$ while for $K > 2$ they scale with $\sim e^N$ and $\sim N$, respectively [1, 2, 3, 4, 5, 6].

An important application of Random Boolean Networks (RBN) is the study of evolution where we can investigate the relation between genotype represented by a RBN and its phenotype defined by a fitness function. Despite of its relevance, the adaptation of RBN has received little attention from the scientific community. The only study on the subject was conducted by Kauffman [1]. The issue has both biological and technological appeal. In biology, boolean networks are a powerful model to understand the behavior and evolution of gene networks. In technology research, they provide a method to understand and develop, through trial and error, real intensive parallel computer programming [7, 8, 9].

Kauffman, based on small scale simulations, suggests that $K = 2$ RBN adapt on a "*well correlated, good landscapes*" [1]. In other words the fitness landscape for $K = 2$ is equivalent to a NK-surface of low K. This property and the dynamical behavior of these networks inspired Kauffman to propose the hypothesis that "*living systems exist in the solid regime near the edge of chaos ...*", since these networks can have reasonable complex behavior and are also highly evolvable. Given the importance of these conclusions, and the fact that they are based on small scale simulations we understand that further work on this area is essential to answer some open questions:

1. Are these results general, or do they depend on the way the fitness function is defined or in the specific details of the adaptation process?
2. Does the phase transition at $K = 2$ change qualitatively the adaptation process?
3. Why does the complexity catastrophe set in?

2. The model

We investigate the capacity of the networks to be "programmed" to reach an arbitrary target cycle. This arbitrary dynamical state represents the goal of the selective process. For example, it may represent the cellular cascade of differentiations during ontogeny, the mutations of tumor cells, the selection of clone cells of the immune system [1], etc.

We start by defining a target net and its initial state, S_0. The target net evolves from S_0 and eventually reaches a cycle in which it remains. The goal is to find boolean networks with cycles similar to the target's. To this end we employ a genetic algorithm (GA) [10] that mutates the connections and boolean functions of the networks and also performs crossovers. This is achieved by representing the boolean networks as a bit string and defining a fitness function (see below) to control the differential reproduction rate in the population, favoring networks with cycles closest to the target's.

The following criteria yield a guide to construct the fitness function: The highest possible fit (f=1) will be obtained only if the asymptotic dynamical behavior of a given net is exactly the same of the target net. If both networks have cycles with the same length (period), we calculate the Hamming distance (the number of different bits) between the states belonging to each cycle. If the two cycles have different periods, the fitness function depends on the difference between the two periods and the similarity of both cycles. The fitness decreases exponentially with the difference between the two periods simulating the effect of deleterious mutations.

To calculate the fitness of a given net η, we start both η and the target net on S_0 and let them evolve until they reach cycles. Let P_η and P_τ be the cycle length of η and of the target net, respectively. We choose the longest one and call it P_m. In the next step both networks are evolved from their first state after the transient (labeled t_η and t_τ). The fitness f is calculated according to

$$f(\eta) = \frac{e^{-A(P_\eta - P_\tau)^2}}{P_m N} \sum_{i=1}^{P_m} d_H(\vec{S_\eta}(t_\eta + i), \vec{S_\tau}(t_\tau + i)) \ , \tag{2}$$

where A is a constant, which without restrictions is defined as 1 and d_H is the Hamming distance between any two states. The exponential factor was included to favor networks with cycle lengths closer to the target's.

Once the fitness function is defined we use a genetic algorithm to simulate the adaptation process. We evolve the population using a steady state GA [10].

The simulation code was implemented in C++ and uses the Galib package developed at Massachussets Institute of Technology
(see `http://lancet.mit.edu/ga/`).

3. Results

We have performed simulations using the following parameters:

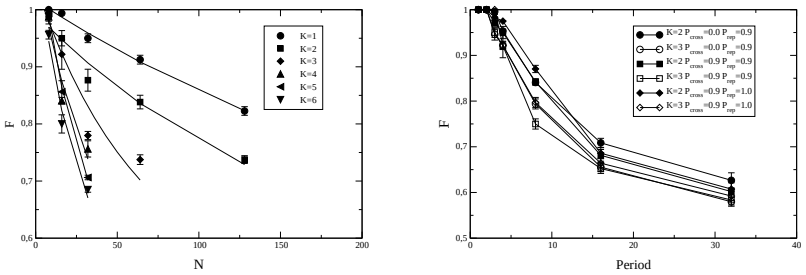
- Population size: 30
- Mutation probability: 0.1
- Crossover (P_{cross}): 0 and 0.9
- Target cycle length: 1,2,3,4,8,16 and 32
- Replacement probability (P_{rep}): 1 and 0.9

The results are averages calculated over five different runs for a given set of parameters.

Our concern is to investigate the GA performance to find solutions for the proposed problem. Our analysis was based on parameter F, the best fitness found in a population. In order to characterize the performance we investigate F time dependence and its value attained after 50,000 generations (this value was chosen to maintain our computational time manageable - the longer simulations took 2 hours on a 400 Mhz Pentium III computer).

In Figure 1(a) we present the F dependence on N for target period 4 for different K values. The graph shows clearly that as K and N grows the GA performance dwindles. Since fitness is a central quantity reflecting dynamical properties of the complex system, we can synthesize the set of graphs in Figure 1(a) with different K described by single parameter a:

$$F(N\,, K) = \frac{1 + e^{-aNK}}{2}. \tag{3}$$



(a) (b)

Figure 1: (a) The F dependence on N and NK for $K = 1, 2, 3, 4, 5$ and 6. Full lines correspond to the function $(1 + e^{-0.0042NK})/2$. The results illustrate the complexity catastrophe, as we get more complex networks, F decreases. (b) The fitness dependence on the target period for crossover probability $P_{cross} = 0$ and 0.9, replacement probability $P_{rep} = 1$ and 0.9 and K=2 and 3. The GA performance is slightly better for $K = 2$. The system size is 16.

From parametric inference using the Maximum Likelihood [12] $a = 0.0042 \pm 0.0004$, where the uncertainty is based on a 95% confidence level (CL). The fairly good agreement of the multiple fit with the simulation data at the given CL may be interpreted as a confirmation of the proposed parametrization of the fitness function, as well as its significance.

In Figure 1(b) we present the dependence of F on the target net cycle period length for $K = 2$ and 3, and $N = 16$. We also compare the performance by changing P_{cross} and P_{rep}. The GA performance is slightly better for $K = 2$, but remains effectively unchanged as we vary other parameters. This is an indication that the fitness landscape presents no long range correlations [1].

From figures 1(a) and 1(b) we observe that performance decreases as we get more complex landscapes or longer periods, respectively. This is the manifestation of a complexity catastrophe caused by the complexity of the imposed task differently from the catastrophe caused by a structural property of the network. This result appears to be universal, being robust against changes on P_{cross} and P_{rep}, as well as modifications in the crossover schema used.

The $1 - F$ time dependence is well approximated by a function of the type $A \ln t + B$ for F close to 1, suggesting a very slow relaxation towards equilibrium, where A measures the rapidity that the GA takes to find an optimum solution. Thus the parameter A dependence on N and K represents the network structure influence on the GA performance. As N and K grows A tends to 0 which implies a divergent relaxation time.

In order to get a better insight in the complexity catastrophe phenomenon, a deeper analysis was performed. In the traditional view the population sticks to local maxima,

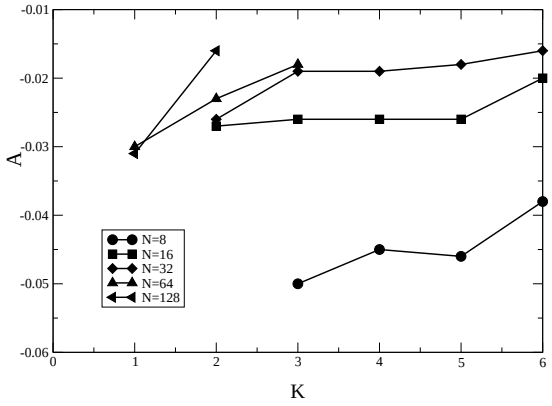


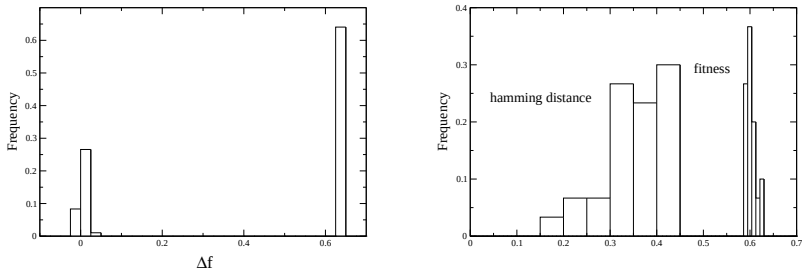
Figure 2: Values of parameter A from the fitting function: $1 - F(t) = A\ln(t) + B$ for different values of N and K for $P = 4$. The data illustrate the complexity catastrophe, as N and K grow $|A|$ gets smaller, implying a divergent relaxation time.

there the variability decreases and the system has to cross a fitness barrier to reach another maximum. The complexity catastrophe in such scenario occurs because the number of local maxima grows exponentially with system size and so does the time spent at a given maximum.

To verify if this picture is adequate to explain our system behavior, we quantified some features of the late stage evolving population. After performing a series of mutations on the best individual on a typical population we verified that it was not located on a local maxima. As shown in Figure 3(a), approximately 40% of point mutations are adaptive or neutral. Figure 3(b) shows for a given population the fitness and the hamming distance histogram. The graph suggests that while fitness values are strongly concentrated around the mean, the hamming distances have a spread histogram, indicating that the population is not concentrated around a given maximum. These results contradict the hypothesis of "population climbing" to fitness maxima, since we should obtain a large frequency of small (~ 0) hamming distances, corresponding to mutants of the best fit network.

These data suggest a new picture for the fitness landscape, although the landscape has a large number of local maxima there always exist some directions where the change in fitness is small or even zero. That is a direct consequence of the fitness function definition. Following the classical argument of Eigen and Schuster [13] we conclude that the *error catastrophe* sets in. Instead of being strongly located at the maxima, the population diffuses through the phase space. The image we have in mind is that the population diffuses on plateaus with small variation in fitness.

As the population reaches higher plateaus with small variations in fitness values, diffusion slows down. This happens since the plateaus are irregular structures, possibly fractals. For higher fitness values they get more labirinthyne increasing substantially the



(a) (b)

Figure 3: Characterization of the late stage population. (a) The best individual is selected and all possible mutations are executed. The figure shows histogram for the fitness variations Δf between the best individual and its mutants ($N = 16$, $K = 2$, and $P = 4$). (b) We consider the distribution of fitness values for a given population and the hamming distance between the best individual and other individuals of the population. The figure shows that while the fitness distribution is concentrated around the maximum value, the hamming distance presents a broad distribution.

relaxation times. The situation is equivalent to the diffusion occurring on percolating clusters over the hypercube as described on [14, 15].

4. Discussion

In the present work we investigated the adaptation of random boolean networks with genetic algorithms (GA's). We find that the convergence is very slow (logarithmic) with time. Suggesting the system could reach arbitrarily high fitness values ($F \sim 1$), but the time involved grows exponentially. The maximum fitness found in a population after a fixed number of generations is an exponentially decreasing function of N, K and P. We have not found any numerical evidence that the behavior for $K = 2$ networks are qualitatively different from other K values. We also found that the GA parameters do not influence significantly in the population evolution.

The situation depicted here is similar to a model studied by Mitchell and coworkers [11], that considers a mutation only GA evolving on a simple fitness landscape called the Royal Road Genetic Algorithm. In their study they find that the major contribution to the collapse of adaptation is the fast increase with GA parameters of the phase space for a finite population to adapt. In this model the complexity catastrophe is the result of an increasing phase space volume and not due to the ruggedness of the fitness landscape. They also find that crossover operators are inefficient in such a scenario as we also verify in our model (see Figure 1(b)).

We believe that the main cause of the complexity catastrophe observed here is a consequence of fast increase of the phase space volume, the ruggedness plays only a marginal role.

Acknowledgements

Work partially supported by FAPERGS and CNPq.

References

References

[1] S. A. Kauffman, *The Origins of order: self-organization and selection in evolution* (Oxford University Press, New York, 1993).

[2] A. Bhattacharjya and S. Liang, Physica D **95**, 29 (1996).

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

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

[5] B. Derrida and G. Weisbuch, J. Physique **47**, 1297 (1987).

[6] R. Albert and A. L. Barabási, Phys. Rev. Lett. **84**, 5660 (2000).

[7] J. P. Crutchfield and M. Mitchell, Proc. Natl. Acad. Sci. USA **92**, 10742 (1995).

[8] S. Wolfram, *Cellular Automata and Complexity: Collected Papers*, 1st ed. (Addison Wesley, Reading, Massachussets, 1994).

[9] M. Mitchell, J. P. Crutchfield, and P. T. Hraber, Physica D **75**, 361 (1994).

[10] D. E. Goldberg, *Genetic algoritms in search, optimization and learning* (Addison-Wesley, USA, 1989).

[11] E.Van Nimwegen, J. P. Crutchfield, and M. Mitchell, Theoretical Computer Science **229**, 41 (1999).

[12] M. Tanner, *Tools for Statistical Inference* (Springer Verlag, New York, 1996).

[13] M. Eigen and J. McCaskill, J. Chem. Phys. **92**, 6881 (1988).

[14] I. A. Campbell, J. M. Flesseles, J. R. Jullien, and R. Botet, J. Phys. C **20**, L47 (1987).

[15] N. Lemke and I. A. Campbell, Physica A **230**, 554 (1996).