Dynamics of Content-Based Networks

Duygu Balcan¹ and Ayşe Erzan^{1,2}

¹ Department of Physics, Faculty of Sciences and Letters
Istanbul Technical University,
Maslak 34469, Istanbul, Turkey

balcand@itu.edu.tr, erzan@itu.edu.tr

² Gürsey Institute, P.O.B. 6, Çengelköy 34680, Istanbul, Turkey
erzan@gursey.gov.tr

Abstract. Content-based networks are introduced and their topological properties are outlined. A content-based model with Random Boolean dynamics, designed to mimic the gene regulation network, exhibits an increase in the number and complexity of attractors for increasing number of nodes. However, contrary to expectations based on Mean Field calculations for random scale-free networks, the attractors are not chaotic, even for average connectivities in excess of 2. Thus, the present model offers a promising tool for understanding complex biological networks.

1 Introduction

Complex networks [1, 2, 3, 4] have become an essential tool for understanding coarse grained properties of discrete, distributed systems with often highly specific interactions as, for example, in genomic or proteomic networks [5, 6], which rely on recognition and binding mechanisms. Independently of the nature of the interaction, we submit that the specificity of the network connections can be modelled by requiring a certain amount of information to be shared between the nodes. We have coined the term “content-based” to denote networks where a certain information content coded in a common alphabet characterises each node. In two previous papers [7, 8] a null-version of such a model with random strings of letters, was formulated and simulations as well as analytical results regarding its topological properties were provided. (An earlier, more ambitious version of such a model was proposed by Reil [9].) In the present paper, we extend our model to incorporate certain features of the gene regulation network in a more realistic fashion [10, 11], and also include a dynamics, provided by randomly assigned Boolean functions at the nodes, in the same spirit as the Random Boolean Network (RBN) models introduced by Kauffman [12, 13].

In the following section we outline our content-based model and summarise earlier results. In Section 3 the details of the extended model and the dynamics will be introduced. In Section 4, preliminary results on the dynamics, and in Section 5 a short discussion will be presented.

2 A Content-Based Null Model

In our original model a linear code of length L , which can be thought of as an “artificial chromosome,” has elements which are randomly assigned from an alphabet containing $r + 1$ letters with the probability distribution $P(x)$,

$$P(x) = p\delta(x - r) + \frac{1 - p}{r} \sum_{\rho=0}^{r-1} \delta(x - \rho) \quad . \quad (1)$$

Each “gene”, G_i , is defined as the string between two successive occurrences of the letter “ r ” on the chromosome, $G_i \equiv (x_{i,1}, \dots, x_{i,j}, \dots, x_{i,l_i})$ where $x_{i,j} \neq r$ and l_i is the length of the i th gene. This results in an exponential distribution of string lengths, $P(l) \propto (1 - p)^l$. The model chromosome thus has only three parameters, the length L , besides p and r appearing in the above distribution.

We now associate a network with this model chromosome, by assigning a node to each “gene,” and postulating an interaction matrix w_{ij} based on the string matching condition, thus:

$$w_{ij} = \begin{cases} 1 & \text{if } G_i \subset G_j \\ 0 & \text{otherwise} \end{cases} \quad . \quad (2)$$

It should be noted that the linear codes associated with the “genes” do not necessarily represent a nucleotide sequence but rather should be considered as the information content of that node. The probability of finding an edge between two nodes with string-lengths l and k , directed from the first to the second, is $p(l, k) = 1 - (1 - r^{-l})^{k-l+1}$.

The resulting network is small-world type, having a large clustering coefficient and very small average shortest path length.[7] It is very robust to random removal of nodes, and we conjecture that the percolation threshold is 0, as for scale-free networks with exponents $\gamma \leq 3$ [3, 14]. Moreover, the network exhibits “Highly Optimised Tolerance” as defined by Carlson and Doyle [15].

The in- and out-degree distributions are displayed in Fig.1. The out-degree distribution has a continuous regime, decaying as a power law with an exponent γ_1 , followed by an oscillatory part whose envelope with an exponent γ_2 where [8]

$$\gamma_2 = \frac{1 \ln r + \ln(1 - p)}{2 \ln r - \ln(1 - p)} \quad , \quad \text{and} \quad \gamma_1 = \frac{1}{2} + \gamma_2 \quad . \quad (3)$$

The length distribution of the sequences identified with the nodes changes the topology of the network. Taking a Gaussian distribution for the lengths yields in- and out-degree distributions both similar to that shown in Fig. 1a, but with a sharper cut-off than for the case of the exponential length distribution.

3 A Content-Based Approach to Dynamics of Gene Regulation

The dynamics of gene regulation networks has been modelled by Random Boolean Networks (RBN) [12, 13], where the genes at the nodes take Boolean values

Fig. 1. The out-degree and in-degree distributions (for $L = 1.5 \times 10^4$ and $L = 4 \times 10^4$ averaged over 2×10^4 random realisations of the random string). Numerical results are indicated by circles, the solid lines are the theoretical expression. See Ref. [8].

$\sigma_i(t)$, $i = 1, \dots, N$, indicating their state of activation (“on” or “off”), determined by randomly assigned Boolean functions with inputs from other genes. The networks may be wired randomly, or with fixed in-degree. Here we present a model designed to mimic the dynamics of gene regulation, and which extends the content-based model of the preceding section.

In our extended model (see Fig.2) each gene, corresponding to a node of the regulatory network, is represented by a linear code having *two* distinct regions: *i*) a regulatory sequence (RS) through which the activation of this gene is regulated by the binding of transcription factors (TFs), and *ii*) a coding region which gets translated either into a TF or a structural protein (also see [9,11]). The sub-sequences occurring in a RS constitute the possible “binding sites” for the TFs.

Let us denote the regulatory and coding regions of the i th gene by ρ_i and π_i , $i = 1, \dots, N$, where the lengths of the two regions, l_{ρ_i} and l_{π_i} may be chosen from different distributions. In this study we have taken the minimum length of the codes associated with the TFs, and the corresponding “binding sites,” to be $l_{\min} = 1$. Larger values may be used to tune the cut-off of the degree distribution at large degrees.

Each character in the strings ρ_i or π_i is chosen from an alphabet of r letters with uniform probability $1/r$. We will denote the m th subsequence of the string ρ_i by $\rho_i^{(m)}$, where $m \equiv (\nu, l)$ with ν specifying the first letter, and l , the length of the sub-sequence. Note $\nu = 1, \dots, l_{\rho_i}$ and $l \leq l_{\rho_i} - \nu + 1$.

An interaction between a pair of genes is now defined via a sequence-matching condition (corresponding to Eq.(2)) between the TF of the i th gene and sub-sequences in the RS of the j th gene. Defining

$$w_{ij}^{(m)} = \begin{cases} 1 & \text{if } \pi_i = \rho_j^{(m)} \\ 0 & \text{otherwise} \end{cases}, \quad (4)$$

the element of adjacency matrix between the i th and j th nodes becomes $w_{ij} = 1 - \prod_m [1 - w_{ij}^{(m)}]$, being nonzero only if at least one of the $w_{ij}^{(m)} = 1$. Note that $w_{ij} \neq w_{ji}$, in general.

The activation state of the system at time t , $\Sigma(t) \equiv (\sigma_1(t), \dots, \sigma_N(t))$, evolves under a set of Boolean operators associated with each node of the network. The input to the Boolean operators B_j are the “binding states” $b_j^{(m)}(t) = w_{ij}^{(m)} \sigma_i(t)$ of the sub-sequences $\rho_j^{(m)}$. If there exists a protein matching the sub-sequence $\rho_j^{(m)}$ in the regulatory sequence of the j th gene, *and* if the gene coding this protein is “on” at time t ($\sigma_i(t) = 1$), then it is assumed that this protein will bind the said sub-sequence, in which case $b_j^{(m)}(t) = 1$; otherwise it is zero. For those sub-sequences which have no matches, the “binding states” are set identically equal to zero.

The truth table defining the operator $B_j(\{b_j^{(m)}(t)\})$ is constructed by assigning to each set $\{b_j^{(m)}\}$ of binding states, a value 1 or 0, with probability q or $1 - q$ as illustrated in Fig.2. Note that nodes with identical RSs will have identical Boolean operators associated with them. Given an initial state of the system, $\Sigma(0)$, the dynamics of the system is now given by $\Sigma(t + 1) = (B_1(t), \dots, B_i(t), \dots, B_N(t))$.

	RS	TF
A	abc	a
B	bb	bc
C	ca	c
D	caa	bc

(a)

Truth Table for RS abc						
a	b	c	ab	bc	abc	B_A
0	0	0	0	0	0	1
1	0	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	0	1	0	1
1	0	1	0	0	0	1
1	0	0	0	1	0	0
0	0	1	0	1	0	1
1	0	1	0	1	0	1
...	...	...	...	...	...	...

(b)

Fig. 2. A small network with four nodes (A,B,C,D) and their associated RS and TF sequences are shown in panel (a). The directed edges satisfy the matching conditions between the TFs and sub-sequences of the RSs. The truth table for the Boolean function associated with the RS of node A is shown on the left hand side of panel (b). The first six columns show the binding states of the “binding sites” a,b,...,abc. The last column gives the value B_A which has been randomly assigned to the set of binding states in that row (not all the sets are shown).

4 Simulation Results

We have simulated the dynamics on the content-based random Boolean networks defined above, with identical distributions for the string lengths associated with the RS and TF regions of each node i , choosing either (truncated) exponential

Fig. 3. Average values of the number, lengths, basin sizes and transient times for the attractors of the dynamics for Gaussian (●) and exponential (○) string-length distributions. We have taken $r = 2$, $p = 0.5$. The distributions are truncated and normalised over the interval $[1, 25]$. The “Gaussian” distribution has $\langle l \rangle = 13$, $\sigma^2 = 50$, whereas the “exponential” distribution has a characteristic length of $\xi \simeq 10$.

or Gaussian distributions. The choice of an exponential length distribution was motivated by the fact that it lends itself to analytical treatment [8]. However, a careful analysis [11] of the length distribution of the binding sites in the regulatory regions of yeast [18, 19] reveals that it is a somewhat skewed, unimodal distribution with a mean around 15 and a variance of about 4, when expressed in binary code. We adjusted the parameters of the Gaussian distributions we used so as to give networks that are not very sparse. Work is in progress to choose different appropriate length distributions for the TFs and the short binding sites that they recognize, and the regulatory regions (RSs) in which the binding sites are embedded. [20]

For systems having a finite number N of nodes, the volume of the phase space $\Omega = 2^N$, is also finite. If we start from an initial configuration and follow its trajectory we will revisit some already visited states and fall into a cyclic orbit known as an *attractor* of the system. The mean values of n_a and l_a , the number and length of the attractors, give us some clue about the system’s stability and versatility. We find (see Fig.3) that $\langle n_a \rangle$, $\langle l_a \rangle$ as well as the average transient time $\langle \tau \rangle$ increase linearly with system size N , for both the Gaussian and exponential string length distributions. However in all cases, the exponential has higher growth rates with N for the above quantities, whereas the Gaussian length distribution gives more stable results.

It is of interest to compare our finding with those of Aldana [16] who has simulated Random Boolean dynamics on random scale-free networks, especially with regard to whether the behaviour is ordered (where differences in initial

conditions are wiped out), critical, or “chaotic,” by which we would like to understand that the lengths of the orbits grow exponentially with N . The results we get for the distribution of the number of precursors of phase points, and the distribution of the number and length of the attractors, are very similar to those found by Aldana [16] for scale-free networks with $\gamma_c = 2.5$ for $p = 1/2$, exhibiting “critical” or borderline behaviour. The scaling exponents for our content-based network in the large N limit are decidedly smaller than this value, and would rather be expected to show chaotic behaviour within Aldana’s approach. This unexpected stability of the dynamics on content-based networks also holds for the mean connectivities: For the relatively small $N \leq 16$ realisations which we have simulated, the average undirected connectivities are smaller than 2 for the truncated Gaussian, and larger than 2 for $N \geq 9$, for the exponential length distribution.

We have also simulated the average overlap function $\langle x(t) \rangle$,

$$\langle x(t) \rangle \equiv 1 - \frac{1}{N} \sum_{i=1}^N \langle |\Sigma(t) - \Sigma^{(i)}(t)| \rangle \quad , \tag{5}$$

where $\Sigma^{(i)}(0)$ differs from $\Sigma(0)$ at the i th node. The average is performed over 10^4 realisations of the network, and all $\Sigma(0)$. Following the trajectory of this quantity under successive steps of the dynamics, we find, for each value of N considered, that it converges to a set of points in a rather small but finite interval lying below unity, as shown in Fig. 4, which becomes shifted to smaller values for larger N , as both $\langle n_a \rangle$ and $\langle l_a \rangle$ grow. Even for small N , $n_a > 1$ with small but finite probability, and the phase points to which trajectories originating in different basins of attraction converge are separated by finite distances. In the presence of periodic orbits of lengths $1 \leq l_i$, one obtains a set of $\binom{M}{2}$ such finite

Fig. 4. The long-time trajectories of the overlap function for different network sizes $5 \leq N \leq 13$, for the exponential length distribution. Averages have been taken over 10^4 network realisations and all pairs of configurations having initial overlap $x(0) = 1 - 1/N$.

distances between pairs phase points, where $M = \sum_i^{n_a} l_i$. Thus, the persistence of distances between randomly chosen points in phase space does not automatically signal “chaotic” behaviour, but the existence of multiple and/or periodic attractors. The evolution of the $x \neq 1$ fixed point with N and its consequences for the phase diagram, [17] is under further investigation.

5 Discussion

In conclusion, let us note that the content-based network which we have proposed offers ample possibilities for the modelling of biological networks. Due to the interactions that can arise between sub-sequences of sufficiently long codes, we see that networks with topological properties similar to those observed today [6] can arise spontaneously, without the need to be assembled from scratch by evolutionary processes.

The Boolean Dynamics on our content-based network yields an increase in the number and complexity of the attractors with the number of nodes, and seems to be near-critical, although, on the basis of the Mean-Field arguments in [16] one would expect “chaotic” behaviour. We note that the expression patterns of yeast have recently been analyzed by Živković et al. [21], and they find both periodic and stochastic behaviour, considering the differential expression levels of the genes. The discrete nature of the present model does not allow us to make detailed comparisons with their results. Nevertheless, we also find patterns where subsets of genes are going through cycles of different lengths, while some nodes are frozen in one or the other state, and that certain perturbations to the state of expression may switch the expression pattern to that of a neighbouring basin of attraction.

Acknowledgements. It is a pleasure to thank M. Mungan and A. Kabakçıoğlu, with whom we have collaborated on some of the results reviewed here, for many useful discussions. AE gratefully acknowledges partial support from the Turkish Academy of Sciences.

References

1. Bollobas, B.: *Modern Graph Theory*. Springer-Verlag, New York (1998)
2. Pastor-Satorras, R., Vespignani, A.: *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, London (2004)
3. Dorogovtsev, S.N., Mendes, J.F.F.: *Evolution of Networks*. *Adv. Phys.* **51**, (2002) 1079–1187
4. Albert, R., Barabási, A.-L.: *Statistical Mechanics of Complex Networks*. *Rev. Mod. Phys.* **74** (2002) 47–97
5. Sole, R.V., Pastor-Satorras, R.: *Complex Networks in Genomics and Proteomics*. In: Bornholdt, S., Schuster, H.G. (eds.): *Handbook of Graphs and Networks*. Wiley-VCH Verlag, Berlin (2002)
6. Barabási, A.-L., Oltvai, Z.N.: *Network Biology: Understanding the Cell’s Functional Organization*. *Nat. Rev. Genet.* **5** (2004) 101–113

7. Balcan, D., Erzan, A.: Random model for RNA interference yields scale free network. *Eur. Phys. J. B* **38** (2004) 253–260
8. Mungan, M., Kabakçioğlu, A., Balcan, D., Erzan, A.: Analytical solution of a stochastic content-based network model. *J. Phys A: Math Gen.* **38** (2005) 9599–9620
9. Reil, T.: Dynamics of gene expression in an artificial genome - implications for biological and artificial ontogeny. In: Floreano, D., Nicoud, J.D., Mondada, F., (eds.) *Advances in Artificial Life. Lecture Notes in Artificial Intelligence*, Berlin Heidelberg, Springer-Verlag (1999) 457–466
10. Alberts, B. et al.: *Molecular Biology of the Cell*. Chapter 9. Garland Science, N.Y. (2002)
11. Kabakçioğlu, A., Mungan, M., Balcan, D., Erzan, A.: in preparation
12. Kauffman, S.A.: Metabolic stability and epigenesis in randomly connected nets. *J. Theor. Biol.* **22** (1969) 437–467
13. Kauffman, S.A.: *The Origins of Order: Self-organization and Selection in Evolution*. Oxford University Press, N.Y. (1993)
14. Cohen, R., Erez, K., Ben-Avraham, D., Havlin, S.: Resilience of the Internet to Random Breakdowns. *Phys. Rev. Lett.* **85** (2000) 4625–4628
15. Carlson, J. M. and Doyle, J.: Highly Optimised Tolerance: A mechanism for power laws in designed systems. *Phys. Rev. E* **60**, (1999) 1412–1427
16. Aldana, M.: Boolean dynamics of networks with scale-free topology. *Physica D* **185** (2003) 45–66
17. Derrida B., Eckmann, J.-P., Erzan, A.: Renormalisation groups with periodic and aperiodic orbits. *J. Phys. A: Math. Gen.* **16** (1983) 893–906
18. Lee, T.I. et al. : Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. *Science* **298** (2002) 799–804
19. Harbison, C.T. et al. : Transcriptional regulatory code of a eukaryotic genome. *Nature* **431** (2004) 99–104
20. Oikonomou, Th. and Provata, A.: Non-extensive trends in the size distribution of Coding and Non-coding DNA sequences in the Human Genome. *Eur. Phys. J. B*, in press. The length distributions for the coding (non-coding) regions of the human genome are found to display different power law tails, here interpreted as indicative of short (long) range correlations, depending on the exponents. Note, however, that some of the so called intergenic non-coding regions actually code the highly conserved binding sites for the transcription factors. See Refs. [10, 19].
21. Zivković, J., Tadić, B., Wick, N., and Thurner, S. : Statistical Indicators of Collective Behaviour and Functional Clusters in Gene Networks of Yeast. *Eur. Phys. J. B*, in press.