MAPPING DYNAMICAL STATES TO STRUCTURAL CLASSES FOR BOOLEAN NETWORKS USING A CLASSIFICATION ALGORITHM

Septimia Sarbu[1], *Ilya Shmulevich*[2], *Olli Yli-Harja*[1], *Matti Nykter*[3], *Juha Kesseli*[3]

[1] Department of Signal Processing, Tampere University of Technology,
PO Box 527 FI-33101, Tampere, Finland, septimia.sarbu@tut.fi
[2] Institute for Systems Biology, Seattle, WA, USA
[3] Computational Biology, University of Tampere, Tampere, Finland

ABSTRACT

Complex systems have received growing interest recently, due to their universal presence in all areas of science and engineering. Complex networks represent a simplified description of the interactions present in such systems. Boolean networks were introduced as models of gene regulatory networks. Simple enough to be computationally tractable, they capture the rich dynamical behaviour of complex networks. Structure-dynamics relationships in Boolean networks have been investigated by inferring a particular structure of a network from the time sequence of its dynamical states. However, general properties of network structures, which can be obtained from their dynamics, are lacking. We create a mapping of dynamical states to structural classes, using time-delayed normalized mutual information, in an ensemble approach. The high accuracy of our classification algorithm proves that structural information is embedded in network dynamics and that we can extract it with information-theoretic methods.

Index Terms— Boolean networks, structural classes, information theory, classification, feature extraction

1. INTRODUCTION

The study of complex systems is an extremely active area of research, as they are part of nature and of all fields of human activity, such as physics, biology, ecology, social sciences, economy, engineering [1]. Computationally tractable algorithms to analyse, modify and synthesize complex systems are lacking. The overall behaviour of such systems can be better understood if we analyse the interactions between the elements of the system, and not take into consideration the particular details of each element. This level of simplification gives rise to complex networks [2], [3].

One fundamental aspect in the study of complex networks is the relationship between the structure and the dynamics of the network [4]. The general questions that are being addressed are how to define classes of structure and classes of dynamics and how different classes of structures and classes of dynamics affect one another. Significant research effort has been devoted to inferring the structure of particular complex networks from the time sequence of their dynamical states. The author of [5] defined classes of structures of complex networks and classified several real-world networks. Structure-dynamics relationships were studied in [6] for a specific example of complex networks.

The principles underlying the mutual influence of structure and dynamics of models of complex networks have also been studied with information-theoretic methods. The authors of [7] inferred the structure of several complex networks from time series measurements of their dynamical behaviour and assessed the reconstruction accuracy. In [8], the authors inferred a Boolean network model of the gene regulatory networks of five biological systems and explored their dynamical behaviour. Using Kolmogorov complexity, the macrophage biological system was proven to have critical dynamics [9]. However, the universal principles that govern these structure-dynamics relationships have generally not been investigated in detail. The following studies have initiated the discovery of these general laws of interaction between the structure and the dynamics. Within the Kolmogorov complexity framework and using an ensemble approach, the authors of [10] illustrated that critical random Boolean networks had the greatest variety of dynamical behaviour. Here and in [11], dynamical properties were investigated for certain classes of random Boolean networks.

Our novel contribution is the characterization of structural properties of Boolean networks, from the dynamics, without inferring the topology. We identify information about the class of structure of the network. We are able to accurately separate classes of structures from the observed dynamics, without computing dynamical properties, in an ensemble approach. With our time-delayed normalized mutual information method, we prove that structural information is embedded in the dynamics and that it can be revealed with such methods from information theory. Taking the dynamical states from two structural classes of Boolean networks, we can separate the two systems with high accuracy, without explicitely knowing the structure of the system or its dynamical

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

properties. Thus, our study of ensembles of Boolean networks shows that we can create a mapping from general dynamical states to structural classes.

The article is organized as follows: in section 2 we present background information about Boolean networks, the normalized mutual information and the clustering coefficient. In the next section, we describe in detail the classification method. In section 4, we illustrate the accuracy results of the classification process and, in the last section, we present the conclusions of this study.

2. BACKGROUND

2.1. Boolean networks

Boolean networks were introduced in [12], as models of gene regulatory networks. They are graphs of interconnected elements, with the following properties: the state of each node can be either 0 or 1 and each node has an n-dimensional Boolean function associated to it. For each node, the dimension of the Boolean function is given by the number of inputs to that node. An n-dimensional Boolean function is defined as $f : \{0,1\}^n \rightarrow \{0,1\}$ and, for each combination of input values, the function is either 0 or 1. The total number of Boolean functions with n inputs is 2^{2^n}. For a network, each node can have any number of input nodes, the in-degree, and any number of output nodes, the out-degree, as long as the mean in-degree of the network is equal to its mean out-degree. The network topologies investigated in this paper, i.e. the structural classes, are described in section 3.1.

In terms of dynamical behaviour, the network starts in an initial state, where the state of each node is either 0 or 1. Each node updates its state at time point $t+1$ according to the Boolean function associated with it. Each node has one Boolean function associated with it. The state of the node at time point $t+1$ comes from its Boolean function, when the input to the function is the collection of the states of the input nodes to that node, at time point t. When the structural classes are different, the Boolean functions are selected at random from the uniform distribution on all Boolean functions. When the structural class is the same for both categories of networks that we want to classify, the Boolean functions come from two different sets. They are selected at random from the uniform distribution on all the functions from that set. For example, one category of networks has only canalizing Boolean functions, which are taken at random from the uniform distribution on all canalizing Boolean functions. More details about this selection can be found in section 4.

2.2. Time-delayed normalized mutual information

The mutual information, MI, is defined as [13]

$$MI(X,Y) = \sum_x \sum_y p(x,y) \cdot \log_2 \frac{p(x,y)}{p(x) \cdot p(y)}. \quad (1)$$

In our study, the MI has a wide range of variation when applied to the dynamical behaviour from different types of Boolean networks. Thus, it is impossible to conduct a comparative analysis of the dynamical behaviour of such networks, to obtain meaningful results. To solve this issue, we performed the analyses with the normalized version of the mutual information, which has values between 0 and 1. We used the following definition of the normalized mutual information, nMI, between two random variables X, Y [14]:

$$nMI(X,Y) = \frac{MI(X,Y)}{H(X,Y)} = \frac{H(X)+H(Y)}{H(X,Y)} - 1. \quad (2)$$

If $X = Y$, then $nMI(X,Y) = 1$, and if X, Y are independent, then $nMI(X,Y) = 0$.

2.3. Clustering coefficient

The clustering coefficient for undirected and unweighted networks was introduced in [15] and extended for directed and weighted networks in [16]. In this study, we use the directed version of the clustering coefficient, which will be referred to as clustering coefficient throughout the paper. Let CM be the connectivity matrix of a directed graph of N nodes. Let N_{O_i} be the number of neighbours of node O_i, $\forall i = 1 : N$. The number of all possible connections between these neighbours is equal to $N_{O_i} \cdot (N_{O_i} - 1)$. The local clustering coefficient for a node O_i, CC_i, can be defined as the ratio between the actual connections between the neighbours of this node and all the possible connections between them, as

$$CC_i = \frac{\sum_{j=1}^{N_{O_i}} \sum_{k=j+1}^{N_{O_i}} [CM(O_j,O_k) + CM(O_k,O_j)]}{N_{O_i} \cdot (N_{O_i} - 1)}. \quad (3)$$

The average clustering coefficient of a network, CC, is defined as the mean of the local clustering coefficients of all the nodes of the network: $CC = \frac{1}{N} \cdot \sum_{j=1}^{N} CC_i$.

3. DESCRIPTION OF THE METHOD

3.1. Boolean network design

We simulated and classified the dynamics for the following structural classes of Boolean networks: **fixed $\mathbf{K} = 3$ Boolean network**, where each node has a fixed in-degree $K = 3$ and a fixed out-degree $K = 3$, **fixed $\mathbf{K} = 2$ Boolean network**, where each node has a fixed in-degree $K = 2$ and a fixed out-degree $K = 2$, **Poisson Boolean network**, where the in-degrees of the nodes are drawn from a Poisson distribution with mean $\overline{K} = 3$ and the out-degrees from a Poisson distribution with mean $\overline{K} = 3$, **scale-free Boolean network**, where the in-degrees of the nodes are drawn from a scale-free distribution with mean $\overline{K} = 3$ and the out-degrees from

a scale-free distribution with mean $\overline{K} = 3$ and **modular Boolean network**, which comprises three modules that are pairwise connected by one random connection (the first module is a **fixed $\mathbf{K = 3}$ Boolean network**, the second module is a **fixed $\mathbf{K = 4}$ Boolean network** and the third module is a **fixed $\mathbf{K = 5}$ Boolean network**). In all structural classes, the nodes are pairwise connected at random.

3.2. Estimation of the nMI matrix

As the nodes of the network are interconnected, they influence each other from one time step to another. A time-delayed version of the nMI captures more accurately the correlation between the dynamical states of the nodes of the network. We define the normalized mutual information matrix as $nMI(i,j) = nMI(O_i(t), O_j(t+1))$, where $O_i(t)$ represents the state of node O_i, at time point t, and $O_j(t+1)$ represents the state of node O_j, at time point $t+1$. We use a plug-in method to estimate the nMI matrix. We first estimate the joint probability mass function, pmf, of $O_i(t)$ and $O_j(t+1), \forall i,j = 1:N$.

After the structure has been created and the Boolean functions have been assigned to each node, we run the network forward in time starting from an initial state. It comprises the inital states of all the nodes, randomly selected from the uniform distribution on 0 and 1. We run the network forward in time for $N_s = 100$ time steps and record the dynamical state of the entire network, for each time point. We repeat this process $N_n = 100$ times, as follows: we have the same connectivity matrix and the same Boolean functions, but, we restart the network from a random initial state each time. As a result, we have $N_n \cdot (N_s + 1) = 10100$ time points for the estimation of the time-delayed normalized mutual information, including the initial state. We collect the states of node O_i from $t = 1 : N_s - 1$ and the states of node O_j from $t = 2 : N_s$. We estimate the joint pmf of node O_i and O_j by using a frequency of appearance estimator. We count the frequency of each of the states: $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ and then we divide by the total number of states.

3.3. Thresholding the nMI matrix

The nMI matrix contains real values in the $[0\ 1]$ interval. They represent the degree of correlation of the dynamics of the nodes between which they were computed. A high nMI value between two nodes is a good indicator of the presence of a link between them. However, even if two nodes are linked, their nMI may be lower than the nMI for other pair of connected nodes. In addition, a relatively high nMI value may come from indirect connections between two nodes. As a result, it is impossible to define a universal value of nMI, above which all nMI values indicate a link between two nodes.

To overcome this problem, we develop a thresholding scheme to obtain several approximations of the connectivity pattern of the nodes. For a given threshold level, the nMI values that are greater than this level are replaced with 1 and the rest are replaced with 0. For each nMI matrix, the threshold levels are the 2.5^{th}, 25^{th}, 50^{th}, 75^{th} and 97.5^{th} percentiles of the set of all the nMI values from the matrix. For example, the 25^{th} percentile of a set of numbers is the value for which 25% of all the elements in this group are smaller or equal to this value. For each level, we create a thresholded nMI matrix of 0 and 1.

3.4. SVM classification

We compute the clustering coefficient of the thresholded nMI matrix, which represents one feature for classification. We associate 5 classification features for each nMI matrix, corresponding to the 5 thresholding levels. Two structural classes of Boolean networks are separated with a support vector machine (SVM) algorithm [17], with a Gaussian kernel and a 10-fold cross-validation scheme. The Gaussian kernel parameters are optimized using grid search, having the missclassification rate as an optimization criterion.

4. SIMULATION RESULTS

The accuracy results presented in table 1 and in table 2 are averaged over 100 experiments. For each experiment, we repeat the procedure of creating the structure, assigning the functions and running the network, $N_e = 1000$ times. One accuracy result is computed as follows: we have $N_e = 1000$ samples for classification, each sample containing 5 features. We train the SVM classifier on 100 samples and we test the algorithm on 200 samples. The training samples are selected at random from the uniform distribution on all samples and the test samples are selected at random from the uniform distribution on the remaining samples after training. The number of nodes of the network is $N = 100$.

In table 1, the Boolean networks have a variety of classes of structures. The Boolean functions are randomly chosen from the uniform distribution on all Boolean functions with a given number of inputs. In table 2, we perform the experiments for one structural class, a fixed $K = 3$ Boolean network. The types of Boolean functions are different for each class. In the first test case, one group of networks contains functions randomly chosen from the uniform distribution on canalizing Boolean functions [18] and the other group contains functions randomly chosen from the uniform distribution on noncanalizing Boolean functions. In the second test case, one class contains functions randomly chosen from the uniform distribution on canalizing Boolean functions and the other class contains functions selected at random from the uniform distribution on all Boolean functions.

Table 1. *Classification of Boolean networks from their dynamics. The Boolean functions are selected at random from the uniform distribution on all Boolean functions.*

Class	In-degree distribution	Out-degree distribution	Accuracy
1	fixed $K = 3$	fixed $K = 3$	94%
2	Poisson $\overline{K} = 3$	Poisson $\overline{K} = 3$	
1	fixed $K = 3$	fixed $K = 3$	99.73%
2	modular $K_1 = 3$ $K_2 = 4$, $K_3 = 5$	modular $K_1 = 3$ $K_2 = 4$, $K_3 = 5$	
1	fixed $K = 3$	fixed $K = 3$	50.17%
2	fixed $K = 3$	fixed $K = 3$	
1	Poisson $\overline{K} = 3$	Poisson $\overline{K} = 3$	50.05%
2	Poisson $\overline{K} = 3$	Poisson $\overline{K} = 3$	
1	Poisson $\overline{K} = 3$	Poisson $\overline{K} = 3$	94.57%
2	Scale-free $\overline{K} = 3$	Scale-free $\overline{K} = 3$	
1	fixed $K = 2$	fixed $K = 2$	98.39%
2	fixed $K = 3$	fixed $K = 3$	

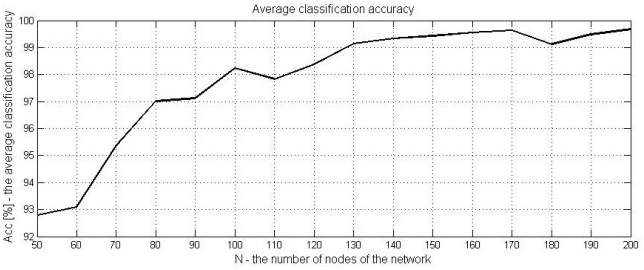

Fig. 1. *The classification accuracy of two Boolean networks with fixed degree K=2 and fixed degree K=3, shown as a function of the nodes of the network. The results are averaged over 100 experiments.*

5. CONCLUSIONS

Our algorithm produces extremely high classification accuracy results, which are averaged over 100 experiments. The range is 92% to 99.8%, for networks with $N = 100$ nodes. Figure 1 shows that the classification accuracy falls between 92.8% and 99.5%, for networks with the total number of nodes ranging from 50 to 200 nodes. The accuracy rises sharply for networks with 60 to 90 total number of nodes, continues with a more moderate increase in the range of $90 - 150$ nodes, leveling off at around 99%, for networks with the total number of nodes above 150.

Table 2. *Classification of Boolean networks from their dynamics. The structure is the same for both classes, a random Boolean network with fixed $K = 3$, but the types of Boolean functions are different for each class.*

Class	Boolean update functions	Accuracy
1	canalizing	99.27%
2	noncanalizing	
1	canalizing	92.85%
2	all functions	

We tested our classification algorithm on data from one class of Boolean networks to identify whether or not our algorithm yielded false results. We obtained a classification accuracy of around 50%, indicating that it is not possible to separate the two data sets. These results are correct, as the data sets belong to the same ensemble of Boolean networks and are characterized by the same features. In addition, this result proves that our method performs well on ensembles of Boolean networks, as it cannot classify individual networks coming from the same ensemble.

Information about the structure of real biological networks is available indirectly from their dynamics. For example, gene regulatory networks are identified indirectly through gene expression microarray experiments [19], [20]. Classifying ensembles of such networks into structural classes, using their dynamical states, will uncover the general laws that govern the bidirectional relationship between their structure and their dynamics. Such theoretical framework will facilitate the understanding of how complex networks, and biological networks in particular, are organized, how their constituent parts are interconnected and how they function together as a whole, giving rise to intricate dynamical behaviour. In addition, this methodology can be the first step in the inference process of biological networks. Once the structural class is identified, more targeted methods can be used to infer the structure of the biological network under study. The results of our experiments with Boolean networks, as models of complex networks, support these hypotheses. Our classification method, based on time-delayed normalized mutual information, discovers structural information from trajectories of dynamical states of Boolean networks. The results indicate that we have successfully created a mapping of dynamical states to structural classes of Boolean networks.

6. ACKNOWLEDGEMENTS

The work presented in this paper was partly funded by the Graduate School in Electronics, Telecommunication and Automation (GETA), Aalto University, Helsinki, and the Academy of Finland project number 310071.

REFERENCES

[1] M.E.J. Newman, "Complex systems: A survey," *American Journal of Physics*, vol. 79, no. 800, 2011.

[2] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, pp. 47–97, January 2002.

[3] M.E.J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 165–381, 2003.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex neworks: structure and dynamics," *Physics Reports*, vol. 424, pp. 175–308, 2006.

[5] E. Estrada, "Topological structural classes of complex networks," *Physical Review E*, vol. 75, pp. 016103–1–016103–12, 2007.

[6] J.M. Siqueiros-García, R. García-Herrera E. Hernández-Lemus, and A. Robina-Galatas, "Mapping the structure and dynamics of genomics-related MeSH terms complex networks," *PLoS One*, vol. 9(4):e92639, 2014.

[7] E. Hernández-Lemus and Jesus M. Siqueiros-García, "Information theoretical methods for complex network structure reconstruction," *Complex Adaptive Systems Modeling*, vol. 1:8, 2013.

[8] E. Balleza, E.R. Alvarez-Buylla, A. Chaos, S. Kauffman, I. Shmulevich, and M. Aldana, "Critical dynamics in genetic regulatory networks: examples from four kingdoms," *PLoS One*, vol. 3(6): e2456, 2008.

[9] M. Nykter, N.D. Price, A. Larjo, T. Aho, S.A. Kauffman, O. Yli-Harja, and I. Shmulevich, "Critical networks exhibit maximal information diversity in structure-dynamics relationships," *Physical Review Letters*, vol. 100, pp. 058702–1–058702–4, 2008.

[10] M. Nykter, N.D. Price, M. Aldana, S.A. Ramsey, S.A. Kauffman, L.E. Hood, O. Yli-Harja, and I. Shmulevich, "Gene expression dynamics in the macrophage exhibit criticality," *PNAS*, vol. 105(6), pp. 1897–1900, 2008.

[11] T. MakiMarttunen, J. Kesseli, and M. Nykter, "Balance between noise and information flow maximizes set complexity of networks dynamics," *PLoS One*, vol. 8(3):e56523, 2013.

[12] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, pp. 437–467, 1969.

[13] T.M. Cover and J.A. Thomas, *Elements of information theory*, John Wiley & Sons, Second edition, 2006.

[14] F.M. Malvestuto, "Statistical treatment of the information content of a database," *Information Systems*, vol. 11, no. 3, pp. 211–223, 1986.

[15] D.J. Watts and S.H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[16] G. Fagiolo, "Clustering in complex directed networks," *Physical Review E*, vol. 76, pp. 026107–1–8, 2007.

[17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[18] I. Shmulevich and S.A. Kauffman, "Activities and sensitivities in Boolean network models," *Physical Review Letters*, vol. 93, pp. 048701–1–4, 2004.

[19] Donna K. Slonim and Itai Yanai, "Getting started in gene expression microarray analysis," *PLoS Computational Biology*, vol. 5(10): e1000543, 2009.

[20] Guocai Chen, Michael J. Cairelli, Halil Kilicoglu, Dongwook Shin, and Thomas C. Rindflesch, "Augmenting microarray data with literature-based knowledge to enhance gene regulatory network inference," *PLoS Computational Biology*, vol. 10(6):e1003666, 2014.