Logic motif of combinatorial control in transcriptional networks

Xuebing Wu, [1] Zhirong Sun [2,*], Rui Jiang [1]

[1]*MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST / Department of Automation*
[2]*MOE Key Laboratory of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, Department of Biological Science and Technology，Tsinghua University, Beijing 100084, China*
[*] *Corresponding author*

Combinatorial control is prevalent in transcriptional regulatory networks. However, whether there are specific logic patterns over- or under-represented in real networks remains uninvestigated. Using a theoretic model and *in-silico* simulations, we systematically study how the relative abundance of distinct regulatory logic patterns influences the network's global dynamics. We find that global dynamic characteristics are sensitive to several specific logic patterns regardless of the detailed network topology. We show it is possible to infer logic motifs based on the sensitivity profile and the biological interpretations of these global characteristics.

Introduction

Gene regulatory networks, or more specifically, transcriptional regulatory networks, are vital for many important biological processes, such as development and the response to environmental changes. The study of the static architecture and dynamic behaviors of gene regulatory networks has long been one of the central problems in genetics. Recent works have shown that transcriptional regulatory networks are composed of basic building blocks — network motifs [1, 2]. Network motifs are enriched small sub-network patterns that are more frequently observed than random. Some motifs, such as feed-back loops and feed-forward loops, are found to be highly enriched in real transcriptional networks and are supposed to be important for the dynamics and robustness of the gene network. Essentially, a network motif dissects the network at the static topology level, while the network dynamics depends more on a higher level: combinatorial control logic. In real biological networks, the behavior of genes are controlled by complex combinatorial regulations of multiple inputs, and it has been shown that many complex combinatorial control logics could be easily implemented even in simple organisms [3]. Similar to the concept of network motif, we would ask: are there any “motifs” of combinatorial logic patterns? i.e., are there specific logic patterns favored by nature? Are all theoretically available logic patterns uniformly distributed in real biological networks? To answer these questions, one needs to know the exact combinatorial control logics for relatively large numbers of genes in specific organism. However, currently we do not have sufficient data on real networks.

As a first step towards the logic motif problem, we turn to theoretic models and *in silico* simulations. We propose a framework targeting this problem in the reverse direction: we study how the relative abundance of distinct regulatory logic patterns influences the network's global dynamic characteristics, and by forcing these characteristics to resemble those of real organisms, we may infer whether a logic pattern is over-represented or under-represented in real networks.

Methods

NK model

We use Kauffman's *NK* model [4] to model the transcriptional network. *NK* model is the first attempt in deciphering the integrated dynamic behavior of the complex genetic network. Genes are taken as boolean variables (1-active, 0-inactive) and the state of which is determined by the state of the inputs through some Boolean function, or logic patterns here, such as AND, OR and XOR when K=2. In *NK* models, there are N genes, and each gene has K inputs, or regulated by K genes. *NK* model is much simplified, yet is quite powerful in exploring the dynamics of gene regulatory networks [5]. In this report we set K=2, i.e. each gene receives regulation from 2 genes. Previous research shows that networks with K=2 are ordered and critical, whose dynamics are always stable [4], which is an important characteristics for real regulatory networks. In addition, large scale experimental data in yeast [6] shows that the average number of regulators for each gene is 1.9. Therefore it is reasonable to study our problem with K=2 networks. Another benefit is, for K=2, there are only 16 ($=2^{2^K}$) potential logic patterns, which are feasible to be studied all at one time.

Global dynamic characteristics

Four global characteristics are defined and studied in the attractor state space ***S*** of each *NK* network: distinct attractor number (*DAN*), average attractor cycle length (*ACL*), gene expression rate (*GER*), and gene expression variance (*GEV*). A *state* is a string of '0' and '1', representing the expression level (active or inactive) of all genes in the network. The attractor state space ***S*** consists of all possible states for all attractors. It is well recognized that the number of distinct attractors can be interpreted as distinct cell types and the attractor cycle length represents the cell differentiation period [4]. The expression rate of a *state* is the proportion of genes activated (taking on the value 1), and the expression rate of the network (*GER*) is the average of expression rates of all states in the attractor space. The global characteristic *GER* can be interpreted as the average level of how the network is activated. For simple species (with relative small networks), this value should be high, for example 4664 out of about 6000 yeast genes are active under general conditions, with *GER*>0.75 [7]. The fourth characteristic gene expression variance (*GEV*) is defined by

$$GEV = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{P_i - 0.5}{0.5}\right)^2}$$

Where N is the number of genes in the network and P_i measures how likely gene i is active in a given states and equals to the ratio of the number of *states* where gene i is active to the total number of *states*. *GEV* describes how far away P is from 0.5, the random case. Considering the fact that housekeeping gene tends to express in all cell types (thus $P_i \rightarrow 1$) while tissue-specific genes only express in a few special cell types ($P_i \rightarrow 0$), GEV is supposed to be high in real organisms.

Multi-parameter sensitivity analysis (MPSA)

In this report, we want to find out if some of the defined global characteristics are sensitive to specific logic patterns, and if so, how the relative abundance of the logic patterns influence the global characteristics. In traditional sensitivity analysis, if one wants to find out whether the output is sensitive to specific parameters (or inputs), all other parameters are fixed, then one checks if the output changes significantly when the parameter understudy varies. However, the relative abundance of logic patterns is inherently correlated. If the relative abundance (proportion) for one logic pattern changes, the numbers of other logic patterns also changes, as their sum is fixed. At this point, we need a tool that simultaneously tests all parameters at one time.

The tool we use is multi-parameter sensitivity analysis (MPSA), a method recently introduced in finding the relative important factor of a complex system [8]. We first generate a random network with N genes and $2N$ edges. This is done by randomly assign K (=2) genes as inputs to each gene in the network. For each network, we randomly sampled 1000 N-by-1 vectors by Latin Hypercube Sampling method. Each number in the vector indicates the number of a certain pattern in the network. Then logic patterns are randomly assigned to each gene, so that the number of each logic pattern is equal to that in the specified vector. Finally, the *NK* model is run, attractors are identified, and the four global characteristics are calculated. The vector specifying the relative abundance of all logic patterns is taken as the multivariate input and the global characteristics as the multivariate output. Then a threshold is selected for each characteristic and each vector is assigned *acceptable* or *unacceptable* depending on whether the corresponding output passes the threshold or not. Finally the cumulative frequencies of *acceptable* and *unacceptable* vectors are calculated for each pattern, and we define the maximum vertical difference of the cumulative frequency curve as the *KS* value, which indicates the relative importance of the pattern to the global characteristic. Thus in each network, for each characteristic, each pattern gets a *KS* value. A higher *KS* value indicates that the characteristic is more sensitively affected by changes in the number of the corresponding pattern. We define the vector of the *KS* values for all patterns as the *sensitivity profile*. A sensitivity profile shows the sensitivity of all logic patterns with respect to a certain characteristic in a network with specified topology.

Results & Discussion

As a proof-of-concept analysis, we study network with $N = 20$ and $K = 2$. In total 10 networks with distinct topology are generated, and we find that the sensitivity profile shows high consistency across different networks. The sensitivity profiles and their consistency between different networks can be visualized by colored map (Fig. 1). We list all 16 logic patterns on the left side in Fig. 1. Note that there is symmetry in sensitivity profile between patterns and their inverse.

To determine whether the sensitive patterns positively or negatively influence the characteristics, we use the simulation data to show how the average value of each characteristic varies when the abundance of logic patterns increases. In Fig. 2, for a fixed point (x, y) on the curve, x is the proportion of a pattern in the network, and y is the characteristic values averaged on the output of 10000 abundance vectors where the pattern's proportion is x. The slope of the curve also indicates the sensitivity of the characteristic to the pattern and is consistent with results of Fig. 1.

Now we consider the problem of identifying over- or under-represented logic motifs in gene regulatory networks. If real organisms favor larger values of a global characteristic, then logic patterns positively correlated with this characteristic will be over-represented, while those patterns negatively correlated will be under-represented. For example, it can be seen from Fig. 1 that pattern 7 and pattern 10 are sensitive ones for GEV. From Fig. 2 we see they both negatively correlated with GEV. If GEV tends to be high in real organisms, then pattern 7 and pattern 10 should be under-represented. Interestingly, this conclusion is supported with some evidences. Pattern 7 and pattern 10 are both "exclusive or" (XOR) type, which are also the only two patterns that do not belong to the class of canalyzing functions. It has been shown that network constructed from canalyzing functions exhibit a tendency toward ordered behavior and are widely observed in real genetic networks [9]. Some explanations have been proposed for the rarity of XOR patterns, such as the difficulty in constructing such functions in reality, or its contribution to the instability of the network. Here we propose another explanation based on our sensitivity analysis. Fig.2 shows that both XOR type patterns have negative impact on the GEV characteristics. The more XOR patterns exist in the network, the lower GEV is. As long as housekeeping genes take a large part of the genome, the GEV should be high. Networks with low GEV would have lots of housekeeping genes inactivated; which means that some of the fundamental processes remain silent. Thus networks with such configuration are not favored by nature, therefore such patterns will eventually become under-represented.

The current knowledge on gene interactions of genome wide is not sufficient to test our conclusions. Nevertheless, our proof-of-concept analysis shows that it is beneficial to explore the logic motif problem by theoretical analysis and simulation studies. Further work is needed to discuss the situation with larger *N* and *K*.

Acknowledgements

We thank Dr. R Wang, Dr. J.S. Zheng, Dr. L Li, Mr. H Chen and Mr. Y Zhou for useful discussions. This work was supported by the National Key Foundation Research Grant in China (973-2003CB715903; 863-2002AA234041; NSFC-90303017).

Figures

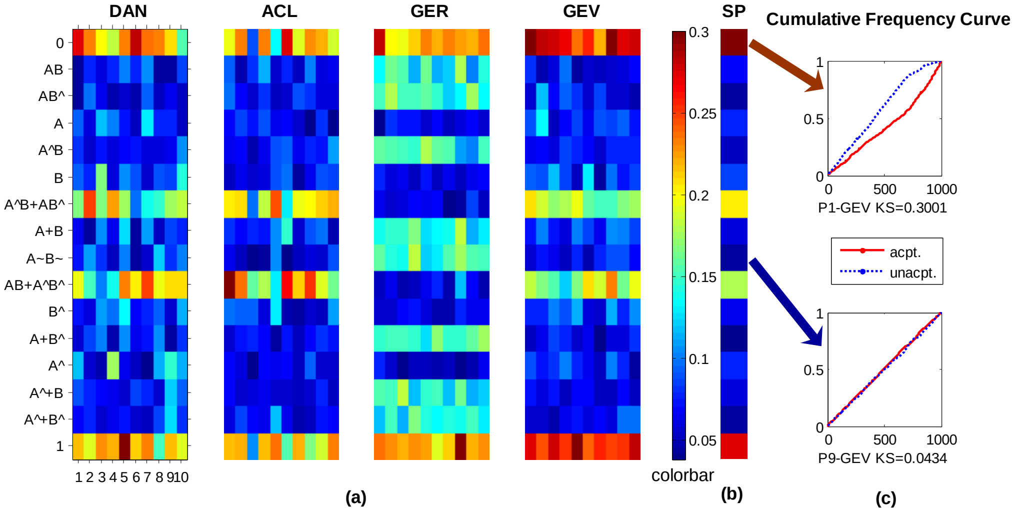

Fig. 1. Sensitivity Profiles. (**a**) the sensitivity profiles for all the four global characteristics studied. Each column represents the SP of one network, such as (**b**), the first column of GEV. Each row represents one logic pattern labeled on the left side, in which '^' means the inverse. The color encodes the KS value, red for higher KS value thus higher sensitivity, blue ones on the contrary. KS value is the maximum vertical distance between the cumulative frequency curves for acceptable and unacceptable abundance vectors, see illustration in (**c**).

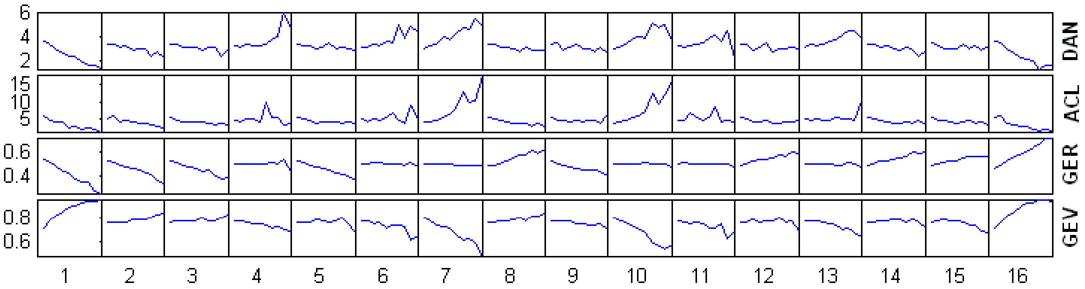

Fig. 2. Relationship between pattern abundance and global characteristics. Each row shows the curves of all 16 patterns for one characteristic. Each subplot in one row illustrates how the expectation of the characteristics varies with the abundance of a certain pattern (labeled by the number on the bottom). The horizontal axes all vary between [0, 0.5], representing the proportion of a pattern in a network.

References

1. Shen-Orr, S.S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* **31**, 64-68 (2002).
2. Milo, R. et al. Network motifs: Simple building blocks of complex networks. *Science* **298**, 824-827 (2002).
3. Buchler, N.E., Gerland, U. & Hwa, T. On schemes of combinatorial transcription logic. *P Natl Acad Sci USA* **100**, 5136-5141 (2003).
4. Kauffman, S.A. Metabolic stability and epigenesis in randomly constructed nets. *J Theor Biol* **22**, 437-467 (1969).

5. Kauffman, S., Peterson, C., Samuelsson, B. & Troein, C. Random Boolean network models and the yeast transcriptional network. *P Natl Acad Sci USA* **100**, 14796-14799 (2003).

6. Lee, T.I. et al. Transcriptional Regulatory Networks in Saccharomyces cerevisiae. *Science* **298**, 799-804 (2002).

7. Lewin, B. Gene VIII. (Pearson Education, Inc., 2004).

8. Zi, Z.K. et al. In silico identification of the key components and steps in IFN-gamma induced JAK-STAT signaling pathway. *Febs Lett* **579**, 1101-1108 (2005).

9. Kauffman, S., Peterson, C., Samuelsson, B. & Troein, C. Genetic networks with canalyzing Boolean rules are always stable. *P Natl Acad Sci USA* **101**, 17102-17107 (2004).