# Identification of essential regulators in biological networks

E. Oh[1,2], J.S. Lee[1], Y.S. Yoo[2] and B. Kahng[1]

[1] *CTP & FPRD, Department of Physics and Astronomy,*
*Seoul National University, NS50, Seoul 151-747, Korea*

[2] *Bioanalysis and Biotransformation Research Center,*
*Korea Institute of Science and Technology, Seoul 136-791, Korea*

(Dated: June 16, 2008)

## Abstract

The Boolean dynamics model is a simple tool to understand the generic dynamic features in biological regulatory networks. From this model, one can construct a network in state space (SS), in which the nodes are sets of Boolean states of each protein and the links connect two consecutive sets. Here we study how the topological features of the SS network are related to the protein functions in the original network. We find that essential regulators, responsible for coordinating the dynamics in the original network, such as checkpoint proteins in the yeast cell cycle, remain in the same state in the SS nodes with large degrees and basin sizes. Using this finding, we then introduce a mathematical tool that can identify the essential regulators in the original network, and successfully select them. We apply this method to another system, *Arabidopsis thaliana*, and confirm its validity.

PACS numbers: 89.75.Hc, 05.45.Df, 64.60.Ak

Recent developments in high-throughput experimental technology and network analysis have provided a global perspective on cellular networks [1]. It has been discovered that various biological properties such as lethality and functionality were related to the structural properties of the protein interaction network [1, 2]. For example, lethal proteins are likely to be located at the hub of the protein interaction network [2].

One of the most important cellular functions is reproduction, which is achieved through the cell cycle [3]. In the cell-cycle network (CCN), there are a few proteins that are responsible for the coordination of regulatory processes and determine if the system moves on to further steps or which dynamic pathway the system would take. Such proteins are known as checkpoint proteins. Here we refer to such proteins as essential regulators.

The dynamics of a biological regulatory network can be studied using the Boolean model. For the CCN, the Boolean dynamics model was introduced with selected proteins [6, 7]. From this model, one can construct a network in state space (SS), in which a node is a set of the Boolean states of each protein and a directed link connects two consecutive sets. In this Letter, we show that in the SS network, the Boolean states of the essential regulators remain the same at the nodes with large degrees and basin sizes. From this dynamic property, we detect the essential regulators through a mathematical tool that we introduce here. We also apply this method to other dynamic systems such as the cell-fate determination regulatory network in floral organs, and successfully identify essential regulators.

We first recall a previous Boolean model for the cell-cycle network [6], in which 14 selected proteins and 33 links compose a regulatory network. The network topology is shown in Fig. 1, which is the same as that shown in Fig. 1A in Ref. [6], but for a few links that have been newly added to Cdc14. Fourteen proteins are connected with two types of directed links, positive (green solid) and negative (red solid) interactions. Further, some proteins contain self-degradation loops (blue links). The dashed green arrow between Cdc20 and Cdc14 is newly added here, signifying that Cdc20 can activate Cdc14 directly as well as via Pds1 [8]. The self-degradation link on Cdc14 is also added here, based on biological evidence presented in Ref. [8]. The checkpoint proteins are denoted as red nodes (octagons) in Fig. 1. These proteins are Cln3, Mcm1/SFF, Pds1, and Cdc20.

Boolean dynamics operates as follows: Each node $i$ is in one of two states, $S_i = 1$ or 0, representing the active and the inactive states, respectively. Following the threshold rule

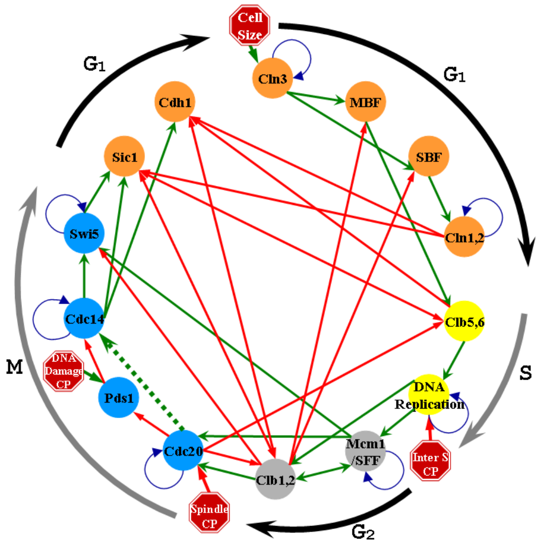

FIG. 1: (Color online) CCN with 14 key proteins. Circles represent proteins, solid green arrows represent activation, and solid red arrows represent inactivation. The dashed green arrow from Cdc20 to Cdc14 also represents activation, and is newly added here.

introduced in Ref. [6], $S_i(t+1)$ at time $t+1$ is updated by the rule below.

$$S_i(t+1) = \begin{cases} 1 & \text{if} \quad \sum_j a_{ij} S_j(t) > 0, \\ 0 & \text{if} \quad \sum_j a_{ij} S_j(t) < 0, \\ S_i(t) & \text{if} \quad \sum_j a_{ij} S_j(t) = 0, \end{cases} \tag{1}$$

where $a_{ij} = 1$ $(-1)$ for an activation (deactivation) link from $j$ to $i$, and 0, otherwise. In addition, the proteins with self-degradation links reset their states as $S_i(t+1) = 0$ when $S_i(t) = 1$, provided that the total input $\sum_j a_{ij} S_j = 0$. Following this rule, the state of each node is subsequently updated.

We obtain $2^{14} = 16,384$ sets generated by the binary states of 14 proteins $(S_1, \cdots, S_{14})$. Each set is indexed with a Greek letter $\alpha$ and a node in the SS network. A cluster comprises connected nodes. From the abovementioned Boolean dynamics rule, we find that there are 14 clusters in the SS network. Each cluster has one attractor, and thus, each cluster size is the basin size of its attractor. Due to the deterministic process of Boolean dynamics, the SS network is a directed tree. Fig. 2 shows a giant cluster of the SS network, containing 14,835 nodes. Due to the large number of nodes, the peripheral nodes are not shown, however, the nodes that are connected from those peripheral nodes are drawn larger in size, proportional to the number of hidden peripheral nodes. The dynamic trajectory under normal conditions

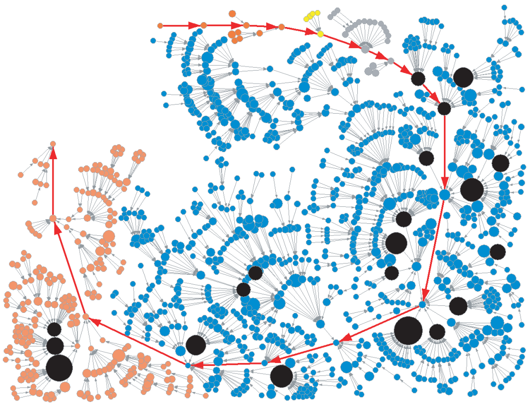

FIG. 2: (Color online) Dynamic trajectory of the cell-cycle sequence in SS. Only the largest cluster is shown. The trajectory under normal conditions is indicated with red arrows. First 20 nodes with large degrees are indicated by black filled circles, which are located at scattered positions near or on the red arrows in the SS network. Colors of nodes represent phases in the cell-cycle sequence. G1, S, G2 and M phases are indicated by peach, yellow, gray, and royalblue, respectively.

is indicated by the red solid arrows.

The degree distribution of the giant cluster in the SS network agrees with the formula $P_d(k) \sim (k+k_0)^{-\gamma}$ with $\gamma \approx 2.8$ and $k_0 \approx 8.7$. Here the degree is the sum of the in-degree and out-degree, however, since the dynamics is deterministic, the out-degree is 1, and the in-degree is the degree minus one for each node in the SS network. In the SS network, the nodes with large degrees, represented by black dots in Fig. 2, are located in a decentralized manner in the SS network. The decentralization of hubs is closely related to the fact that the four checkpoint proteins operate in respective phases in the CCN.

To identify the checkpoint proteins using the network topology, we recall the role of checkpoint proteins. A checkpoint is a coordinator that checks if the cell-cycle process has taken the right step. Otherwise, it triggers a recovery system to repair the damage. Thus, the checkpoint proteins are in hibernation when the cell cycle is in normal operation, but become active otherwise. In Fig. 2, the trajectory represented by the red arrows in the SS network follows the trail of the cell-cycle sequence in normal operation. The other trajectories represent the dynamic trails under abnormal conditions. It is the role of checkpoints to recover from such abnormal processes to the normal trail. Thus, we could say that if the checkpoint proteins were located at the bottlenecks in the SS network under abnormal conditions, then they could play their roles most efficiently. We show that the bottlenecks are the nodes with large degree and basin sizes in the SS network.

We select the top 10 nodes in order of degree in the SS network and tabulate their Boolean states at the individual protein level in Table I. The checkpoint proteins such as Cln3, Cdc20, Pds1, and Mcm1/SFF remain in the same state at the first six nodes. Moreover, other proteins such as SBF, MBF, Sic1, Cdh1, and Swi5 remain in the same state as well. This result yields a large difference between the quantities $D_i^{\rm on}$ and $D_i^{\rm off}$ defined below.

We introduce two mathematical quantities. The first is a degree-related quantity. Since a SS node $\alpha$ represents a set of Boolean states of each protein, one can determine the state of the $i$-th protein at node $\alpha$, i.e., either $S_i^\alpha = 1$ or 0. Using this, we calculate $D_i^{\rm on} = \sum_\alpha' k_\alpha$, where $k_\alpha$ is the degree of SS node $\alpha$ and the primed summation $\sum'$ is subject to the condition that $S_i^\alpha = 1$ for all $\alpha$. $D_i^{\rm off} = \sum_\alpha'' k_\alpha$ is similarly defined, but the double primed summation $\sum''$ is subject to $S_i^\alpha = 0$ for all $\alpha$. The normalized difference between these two quantities for each protein $i$ is defined as

$$\Delta D_i = \frac{D_i^{\rm on} - D_i^{\rm off}}{D_i^{\rm on} + D_i^{\rm off}}. \tag{2}$$

The second is a basin-size-related quantity. The basin size $b_\alpha$ of SS node $\alpha$ is defined as the number of SS nodes which dynamics can start and reaches the node $\alpha$. Similar to the degree-related quantity, we introduce $B_i^{\rm on} = \sum_\alpha' b_\alpha$, and $B_i^{\rm off} = \sum_\alpha'' b_\alpha$. Further, we define

$$\Delta B_i = \frac{B_i^{\rm on} - B_i^{\rm off}}{B_i^{\rm on} + B_i^{\rm off}}. \tag{3}$$

The average of $\Delta D_i$ and $\Delta B_i$, viz.

$$(\Delta D_i + \Delta B_i)/2, \tag{4}$$

is meaningful: For the checkpoint proteins Cdc20 and Mcm1/SFF, this quantity becomes the largest and the second largest positive value, respectively, while for the check proteins Cln3 and Pds1, it becomes the largest and the second largest negative value, respectively, as shown in Fig. 3 (wide red bars). This result implies that the checkpoint proteins operate at the nodes with large degrees and basin sizes in the SS network. However, for proteins with large degrees in the original network such as Clb1, -2 and Sic1, quantity (4) decreases. Thus, the checkpoint proteins can be identified as the ones having extreme values of $(\Delta D_i + \Delta B_i)/2$.

To support this result, we select the region where the basin size is large, which is surrounded by solid blue curve in Fig. 4. This region encloses the trajectory of the cell-cycle sequence under normal condition. For the nodes with large degree in this region, $\Delta D + \Delta B$

TABLE I: Protein states for top 10 nodes in order of degree in the SS network. Here, the numbers in parenthesis in the first column are the degree of respective node in the SS network. The second column "step" is the location of the attractor on the trajectory under the normal condition.

| Rank | Step | Cln3 | SBF | MBF | Cln1,2 | Clb5,6 | Sic1 | Cdh1 | Clb1,2 | DNA Rep. | Mcm1/SFF | Swi5 | Cdc20 | Cdc14 | Pds1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1(226) | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| 2(201) | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 3(177) | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 4(176) | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| 5(159) | 11 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| 6(144) | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| 7(139) | 9 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |
| 8(133) | 11 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| 9(132) | 15 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 10(119) | 11 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |

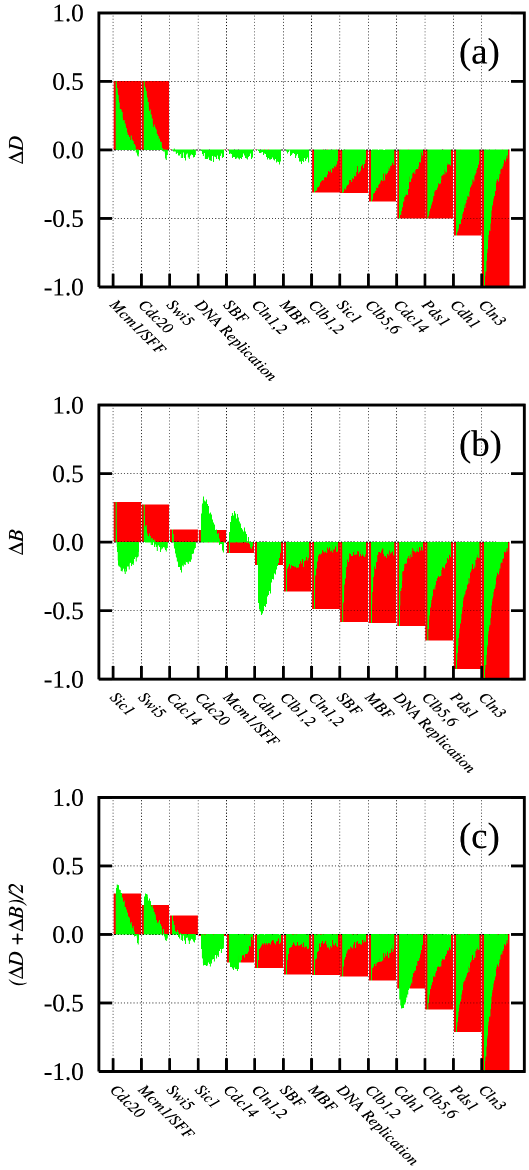

FIG. 3: (Color online) Plot of $\Delta D_i$ versus protein $i$ (a), $\Delta B_i$ versus protein $i$ (b) and $(\Delta D_i+\Delta B_i)/2$ versus protein $i$ (c) for the original network (wide red bars). The same plots for each case but for perturbed networks (thin green bars).

is also large. Further, by representing green the nodes at which Cdc20 is in the active state, as shown in Fig. 4(a). Then, one can see that Cdc20 always remains in the active state at all the nodes within the region. However, non-checkpoint proteins Clb1 and 2 can be inactive at the nodes within the region, as shown in Fig. 4(b). Thus, we confirm that the checkpoint proteins are in the active state at the nodes with large degrees and basin sizes in the SS network.

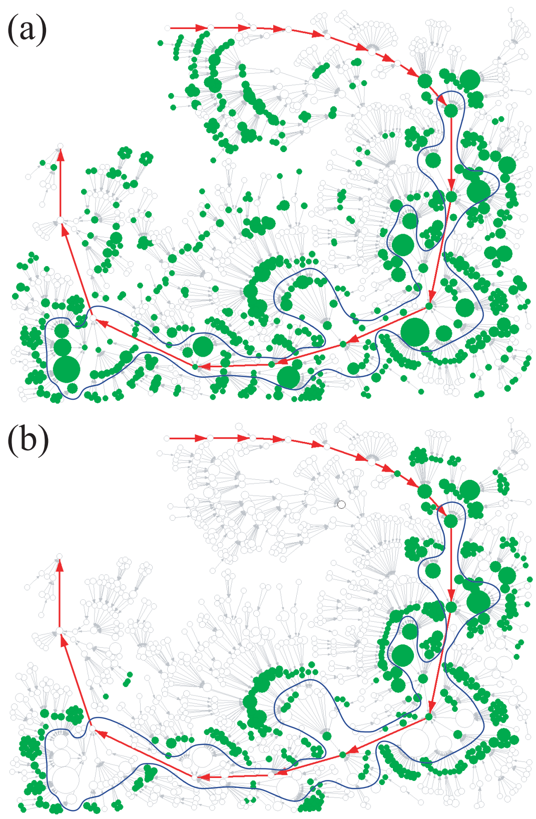

FIG. 4: (Color online) Active (inactive) state of a given protein is indicated in green (open circle) for each node in SS. The given proteins are Cdc20 (checkpoint protein) in (a) and Clb1 and 2 (non-checkpoint proteins) in (b). Nodes with large degrees and basin sizes are distinguished by the blue solid curve.

To determine the robustness of our results to network perturbation, we remove $n$ randomly selected links and reattach them between previously unconnected pairs of nodes, where $n$ runs from 1 to 40. For perturbed networks, we perform the same analysis and measure $(\Delta D + \Delta B)/2$. For each $n$, we construct 1,000 different perturbed networks, and calculate the mean $\langle(\Delta D + \Delta B)/2\rangle$ over the ensemble. The measured value for each case is indicated by the thin green bar in Fig. 3. We find that indeed, as $n$ increases, the quantity $\langle(\Delta D + \Delta B)/2\rangle$ decreases to zero for all proteins. This implies that the original network is well designed and the checkpoint proteins are located at the proper positions. Moreover, the quantity $(\Delta D + \Delta B)/2$ can be successfully used to detect the checkpoint proteins in the original network.

We apply our method to another Boolean model for cell-date determination during *Ara-*

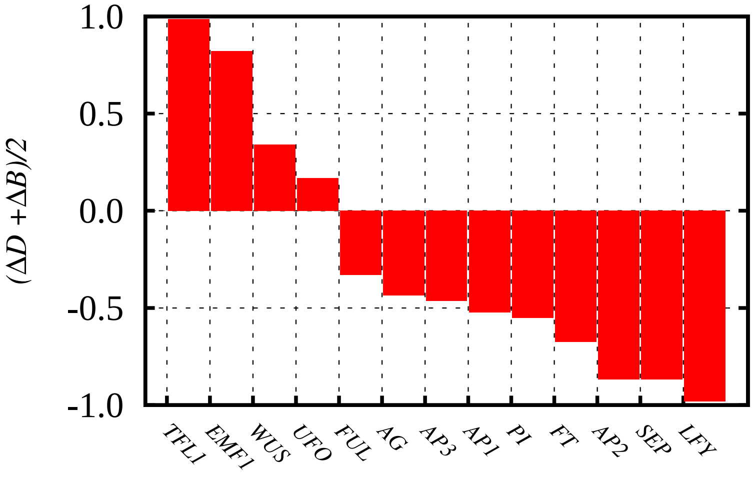

FIG. 5: (Color online) Plot of the difference between two $(\Delta D_i + \Delta B_i)/2$ values measured in the inflorescence and primordial basins versus protein $i$.

*bidopsis thaliana* floral organ development [9]. The Boolean network model consists of 15 proteins, among which two proteins are irrelevant. The SS network for this system consists of 10 disconnected clusters. Their sizes are comparable, which is in contrast to the ones in the CCN. The role of the essential regulators is to determine the cell-fate in the differentiation process. There are five key steps for this. First is to determine whether to differentiate into inflorescence or primordial cells, the essential regulators for this step are known as LFY and TFL. According to the Boolean state of the essential regulator, either ON or OFF, the dynamics flows to different basins, one phenotype "inflorescence" and the other phenotype "primordial." The two basins signify two disconnected clusters. Thus, we measure $(\Delta D + \Delta B)/2$ for each cluster and calculate the difference between them, divided by the normalization factor 2. The result obtained is shown in Fig. 5. Indeed, TFL1 and LFY proteins are located at the extremes, indicating that the quantity we used can identify the essential regulators correctly. Similar analysis can be carried out to select essential regulators in other differential processes for sepal, petal, stamen, and carpel primordial cells.

In summary, we have introduced a mathematical method to identify essential regulators such as checkpoint proteins through the structural features of the SS network. We found that the checkpoint proteins robustly remain in either the on- or off-state at the SS nodes with both large degree and basin size. On the basis of this finding, we introduced the mathematical quantity, $(\Delta D_i + \Delta B_i)/2$ for each protein, which link the structural features of the SS network and the protein functions in the original network. We showed that our tool can also be useful in other systems, for example, the cell-fate determination process during floral organ development.

The authors thank the anonymous referee for introducing the flower problem. This work was supported by the 21C Frontier Microbial Genomics and Application Center Program, MOST (Grant MG05-0203-1-0) and Acceleration Research (CNRC) of MOST/KOSEF in SNU and the KRF Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2007-355-C00030) in KIST.

[1] A.-L. Barabási and Z.N. Oltvai, Nature Reviews Genetics **5**, 101 (2004).

[2] H. Jeong, S.P. Mason, A.-L. Barabási, and Z. N. Oltvai, Nature **411**, 41 (2001).

[3] K. Nasmyth, Cell **107**, 689 (2001).

[4] P.T. Spellman, *et al.*, Mol. Biol. Cell **9**, 3273 (1998).

[5] T.I. Lee, *et al.*, Science **298**, 799 (2002).

[6] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang, Proc. Natl. Acad. Sci. USA **101**, 4781 (2004).

[7] K.-Y. Lau, S. Ganguli, and C. Tang, Phys. Rev. E **75,** 051907 (2007).

[8] K.C. Chen, *et al.*, Mol. Biol. Cell **15**, 3841 (2004).

[9] C. Espinosa-Soto, P. Paddilla-Longoria and E.R. Alvarez-Buylla, The Plant Cell **16,** 2923 (2004).