Bioinformatics, 35(4), 2019, 643–649
doi: 10.1093/bioinformatics/bty665
Advance Access Publication Date: 25 July 2018
Original Paper

OXFORD

Systems biology

Quantifying the notions of canalizing and master genes in a gene regulatory network—a Boolean network modeling perspective

Eunji Kim[1,*], Ivan Ivanov[2] and Edward R. Dougherty[1]

[1]Department of Electrical and Computer Engineering and [2]Department of Veterinary Physiology and Pharmacology, Texas A&M University, College Station, TX 77843, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on March 11, 2018; revised on July 18, 2018; editorial decision on July 23, 2018; accepted on July 24, 2018

Abstract

Motivation: Canalizing genes enforce broad corrective actions on cellular processes for the purpose of biological robustness maintaining a constant phenotype to remain unchanged in spite of genetic mutations or environmental perturbations. Despite their central role in biological systems, the observation/detection of canalizing genes is often impeded because the behavior of affected genes is highly varied relative to the inactive canalizer. Therefore, the activity of canalizing genes is difficult to predict to any significant degree by their subject genes under normal cell conditions.

Results: We investigate this question and present a quantitative framework that allows for the estimation of the power of canalizing genes in the context of Boolean Networks (BNs) with perturbation. This framework borrows tools from the Pattern Recognition theory and uses the coefficient of determination (CoD) to capture the capacity of the canalizing genes. The canalizing power (CP) of a gene is quantitatively characterized by two terms: regulation power (RP) and incapacitating power (IP). We base this assumption on the idea that canalizing power of a gene should be quantified by the extent of its regulation on the overall network and the extent of control that the gene takes over from other master genes when it is activated, which is equivalent to reduction of the control of other master genes upon its activation. Following this, the CP concept is illustrated with examples in which the goal is to provide preliminary evidence that CP can be used to characterize the ability of canalizing genes.

Availability and implementation: A library of functions written in MATLAB for computing CP is available at http://github.com/eunjikim-angie/CanalizingPower.

Contact: eunjikim@tamu.edu

1 Introduction

The concept of genes that can constrain, or canalize, a biological system to a specific behavior was first proposed by C. Waddington in 1942 (Waddington, 1942). Waddington proposed the existence of genes that can produce reliable developmental effects against genetic mutations or environmental changes during evolution (Waddington, 1942; Wagner, 2005). Lehner investigated Waddington's intuition and stated that canalizing genes are hub genes that present similar robustness when faced with environmental, stochastic and genetic perturbations (Ben Lehner, 2010). The term *canalizing gene* has been used by Martins *et al.* (2008) to refer to genes that possess broad regulatory power, and their action sweeps across a wide swath of processes for which the full set of affected genes are not highly correlated under normal conditions. Zhao *et al.* (2012) made a clear distinction between master genes and canalizing genes. Both master and canalizing genes exert a strong control over many downstream gene pathways; however, canalizing genes have an additional ability of taking over the control and overriding other regulatory

instructions. In this paper, canalizing genes refer to genes that are not highly active under normal conditions, but are capable of taking over the control of many pathways and exerting broad regulatory power upon such activation. Canalizing genes produce adaptive and optimal reactions to environmental, stochastic and genetic perturbations and they are essential in a complex system so it can achieve biological robustness and buffer itself from the effects of random alterations or operating errors. We also suggest that the currently adopted definitions of canalizing and master genes could be modified so that a particular gene does not have to be exclusively a master or a canalizing gene. It is important to emphasize that the notions of canalizing and master gene are relative. Any gene possesses some degree of canalizing power over its subnetwork. The notion of canalizing gene can only be defined relative to other genes and the notation C, M and S used in this paper for a canalizing, master and slave gene, respectively, is used with this understanding.

The principle of a canalizing gene is similar to the concept of an interrupt in computer architecture. In systems programming, an interrupt is a mechanism by which the hardware or software alerts the processor to a high-priority condition indicating an event that needs immediate attention and requests the processor to stop the normal processing or current code it is executing and perform a specific action (Linux Device Drivers, 2001). The processor responds by suspending its current activities and jumping to a separate piece of code to deal with the event. Similar to an interrupt handler or an interrupt service routine (ISR) that is invoked by a special instruction or by an exceptional condition and puts the program into a different execution context (Comp Arch And Org, 2E, 2010), the activation of a canalizing gene occurs in response to diverse stress signals or situations where special attention is needed and results in a regulatory mode switch. There are multiple opportunities for canalizing behavior to be observed along the signal-transducing pathway that governs central cellular functions such as cell-cycle, survival, apoptosis and metabolism (Martins *et al.*, 2008). Early observations of canalization along the mitogenic pathway involved dual specificity protein phosphatase 1 (DUSP1) and Ras (Tabin and Weinberg, 1985). DUSP1 antagonizes the activity of the p38 mitogen activated kinase, MAPK1 (ERK), which is a central component of the pathway by which extracellular signal-regulated kinases send mitogenic signals (Chang and Karin, 2001); thus, this gene is canalizing in its phosphorylated state, and DUSP1 is canalizing when it dephosphorylates MAPK1(Martins *et al.*, 2008). Another important instance of canalization involves the tumor protein 53 (*p53*) gene with regard to stresses to the genome (Gomez-Lazaro *et al.*, 2004). While canalizing genes can be extremely potent, their potency is often obscured by other features of the regulatory apparatus operating in the particular cell where control is attempted (Martins *et al.*, 2008).

Martins *et al.* (2008) proposed Intrinsically Multivariate Predictive (IMP) scores, which quantify the synergistic prediction effect of multiple genes, and provided evidence that IMP could potentially be used as a practical tool for discovery of canalizing genes. Chen and Braga-Neto (2015) developed a statistical tool for this inference problem based on the IMP score, by providing a test for a nonzero IMP score between a Boolean target and its respective Boolean predictors. Rejection of the null hypothesis of zero IMP score at a given level of statistical significance gives evidence for the presence of IMP properties. Zhao *et al.* defined canalizing power in a tree model in the context of Bayesian networks (Zhao *et al.*, 2012). The canalizing power of a gene in the paper by Zhao and co-authors measures the total increase in prediction power using pairs of predictors over the maximum prediction power of the respective single predictors, which is equivalently the sum of the IMP scores from all genes in the model. The paper concludes that target genes showing large IMP scores with multiple predictor sets tend to be canalizing. However, when single predictors provide perfect predictions for a canalizing gene, the sum of the IMP scores becomes zero, leading to a paradoxical result: the canalizing power is zero. Furthermore, a key characteristic of a canalizing gene is its ability to override other regulatory instructions and none of the previously mentioned papers considers terms associated with the regulation power of other controlling genes that lose control by the activation of the canalizing gene. Although Zhao *et al.* suggested a formula to measure the canalizing power of a gene, their definition fails to capture the incapacitating trait of canalizing genes. Therefore, we introduce a novel definition of the canalizing power that can quantitatively characterize the power of a canalizing gene based on two important characteristics: (i) It has to be sensitive to the strength of the influence of the canalizing gene on downstream genes and (ii) It should be able to detect how much the canalizing gene incapacitates other regulatory instructions upon its activation. The novelty of this paper lies in the introduction of the notion of incapacitating power and development of a mathematical formula for canalizing power in terms of regulation power and incapacitating power.

This paper is organized as follows. In Section 2, we present the Boolean Networks (BNs) with random gene perturbations as a model for gene regulatory networks and the concept of CoD. Then, we define the regulation power, incapacitating power and canalizing power. In Section 3, we apply the novel definition of canalizing power to both synthetic data and real gene expression data to evaluate effectiveness of the proposed measurements in quantitatively characterizing canalizing genes. Finally, Section 4 gives concluding remarks.

2 Materials and methods

We restrict ourselves to the binary case and note that the methodology presented here presupposes that gene expression has been preprocessed and quantized into binary values. There are several methods that accomplish this (Shmulevich and Zhang, 2002; Zhou *et al.*, 2003). We do not address these methods in this paper, but they are naturally central to the accuracy of the results.

2.1 Boolean networks with gene perturbations as a model for gene regulatory networks

A Boolean network $G(\boldsymbol{X}, \boldsymbol{F})$ is defined by a set of binary-valued nodes $\boldsymbol{X} = (X_1, \ldots, X_n)$ and a corresponding list of Boolean functions $\boldsymbol{F} = (f_1, \ldots, f_n)$. Each node X_i represents the state (expression) of gene i, where $X_i = 1$ means that gene i is expressed and $X_i = 0$ means it is not expressed. $\boldsymbol{F}$ represents the rules of regulatory interactions between genes. To every node X_i, a Boolean function $f_i : \{0,1\}^n \to \{0,1\}$ determining the value of gene X_i is assigned. It is known that genes may become either activated or inhibited due to external stimuli. Moreover, noise could also affect the Boolean relationships. To capture this uncertainty, we consider a BN with perturbation, which has been discussed in (Shmulevich *et al.*, 2002). A Boolean additive-noise model with a random *perturbation vector* $\boldsymbol{N} \in \{0, 1\}^n$ is given by

$$X_i = f_i(\boldsymbol{X}) \oplus N_i \tag{1}$$

where f_i is a Boolean logic function of gene X_i, N_i is the ith component of $\boldsymbol{N}$ and $\oplus$ is modulo-2 addition. $\boldsymbol{N}$ does not need to be

independent and identically distributed (i.i.d.) and we suppose that $P\{N_i = 1\} = p_i$. Then, Equation (1) states that when $N_i = 1$, the ith gene is flipped with probability p_i because of the noise, independently of other genes; otherwise it remains unperturbed. If $p_i = 0$ for all i, then the model is reduced to a deterministic Boolean Network and the standard network transition function $\boldsymbol{F}$ determines the evolution of the model. If $p_i > 0$, then with a probability $1 - \prod_{i=1}^{n}(1 - p_i)$, the current network state will change due to at least one random bit perturbation.

The randomness of this particular network model is encoded by the selection of the initial starting state of the network and also by the gene perturbation probabilities. In order to have a useful probabilistic description of this dynamical system, it is necessary to consider the joint probabilities of all of the genes over time. The dynamics of BNs can be modeled by Markov chains, consisting of 2^n states $D(\boldsymbol{X}), \boldsymbol{X} \in \{0,1\}^n$, with the $2^n \times 2^n$ state transition matrix A where $A(\boldsymbol{X}, \boldsymbol{X}')$ is the probability of transitioning from $\boldsymbol{X}$ to $\boldsymbol{X}'$ (Shmulevich *et al.*, 2002). We are interested in computing the joint probability distribution after one step of the network, which can be achieved iteratively by $D^{t+1} = D^t \cdot A = D^0 \cdot A^{t+1}$, where D^t and D^{t+1} are 1×2^n vectors containing the joint probability distribution at time t and $t+1$, respectively, and D^0 is the starting (prior) joint distribution.

2.2 Coefficient of determination

Let $Y \in \{0,1\}$ be a binary target random variable and $\boldsymbol{X} = (X_1, \ldots, X_r) \in \{0,1\}^r$ be a vector composed of r binary predictor random variables, which in our model represent the values of network nodes perturbed with a random noise vector $\boldsymbol{N}$. The CoD for $\boldsymbol{X}$ predicting Y is defined by

$$CoD_{\boldsymbol{X}}(Y) = \frac{\varepsilon_0(Y) - \varepsilon^{\circ}(\boldsymbol{X}, Y)}{\varepsilon_0(Y)} \tag{2}$$

where $\varepsilon_0(Y) = \min\{P(Y=0), P(Y=1)\}$ is the optimal error of predicting Y in the absence of observations and

$$\varepsilon^{\circ}(\boldsymbol{X}, Y) = \sum_{\boldsymbol{x} \in \{0,1\}^n} \min\{P(Y=0, \boldsymbol{X}=\boldsymbol{x}), P(Y=1, \boldsymbol{X}=\boldsymbol{x})\}$$

is the optimal error upon observation of $\boldsymbol{X}$ (Dougherty *et al.*, 2000). By convention, one assumes $0/0 = 1$ in the above definition because zero prediction error indicates strong interaction between discrete predictor and target variable. The CoD measures the relative decrease in the classification/prediction error when optimally predicting a random variable Y using random vector $\boldsymbol{X}$ as opposed to optimally predicting Y based only on its own statistics. The CoD measures the inherent strength of the nonlinear interaction between a target gene and its predictors and is therefore more appropriate to genomics than the correlation coefficient, which only measures linear interaction. If $CoD_{\boldsymbol{X}}(Y) = 0$, there is no association between $\boldsymbol{X}$ and Y, whereas if $CoD_{\boldsymbol{X}}(Y) = 1$, then $\boldsymbol{X}$ and Y are deterministically related. The CoD measures nonlinear association (increase in prediction power), not causality. The CoD is often used to measure the strength of downstream genes predicting upstream genes. The intuition behind this interpretation is that, if gene Y regulates genes X_1 and X_2, the observation of X_1 and X_2 should allow one to predict the behavior of Y. Moreover, the stronger the control by Y, the stronger is the prediction based on X_1 and X_2.

2.3 Regulation power

In Zhao *et al.*, (2012), the mean CoD value of a gene was defined to represent its regulatory importance in the model. Specifically, the mean CoD of a node Y using all single predictors $X = (X_1, \ldots, X_n) \in \{0,1\}^n$ is given by

$$\overline{CoD_{\boldsymbol{X},1}(Y)} = \frac{\sum_{i=1}^{n} CoD_{X_i}(Y)}{n} \tag{3}$$

Similarly, the mean CoD of a node Y using all sets of double predictors is given by

$$\overline{CoD_{\boldsymbol{X},2}(Y)} = \frac{\sum_{1 \le i < j \le n} CoD_{X_i, X_j}(Y)}{{}_nC_2} \tag{4}$$

A generalized definition for the mean CoD using d predictors is given by

$$\overline{CoD_{\boldsymbol{X},d}(Y)} = \frac{\sum_{i=1}^{{}_nC_d} CoD_{\boldsymbol{X}_d^{(i)}}(Y)}{{}_nC_d} = RP_{\boldsymbol{X},d}(Y) \tag{5}$$

where $\boldsymbol{X}_d^{(i)} \in \mathbb{R}^d$ is the ith d-dimensional vector composed of the elements of $\boldsymbol{X}$ when all possible combinations of size d from the array $\boldsymbol{X}$ are lexicographically ordered for $i = 1, \ldots, {}_nC_d$ (e.g., $\boldsymbol{X}_3^{(1)} = (X_1, X_2, X_3)$, $\boldsymbol{X}_3^{(2)} = (X_1, X_2, X_4), \ldots, \boldsymbol{X}_3^{(20)} = (X_4, X_5, X_6)$ when $\boldsymbol{X} = (X_1, \ldots, X_6)$ and $d = 3$). Equation (5) gives the average strength of predicting Y by using all possible combinations of size d formed by the genes in the network. This general mean CoD measures the influence of the gene Y on the overall network and therefore we call it the "d-regulation power" of the gene Y in the network when d predictors are used for measurements and denote it by $RP_{\boldsymbol{X},d}(Y)$.

2.4 Incapacitating and enhancing power

In this section, we assume that $\boldsymbol{S} = \{S_1, \ldots S_m\}$ is a set of regulated/slave genes. Furthermore, suppose that there is a master gene M which controls the slave genes in a regular network regime and there is a canalizing gene C that is capable of overriding the instructions from the master genes. Intuitively, the regulation power of M could experience significant changes depending on the activation of C. The conditional CoD for $\boldsymbol{S}$ predicting M given that C is on is defined by

$$CoD_{\boldsymbol{S}}(M|C=1) = \frac{\varepsilon_0(M|C=1) - \varepsilon^{\circ}(\boldsymbol{S}, M|C=1)}{\varepsilon_0(M|C=1)} \tag{6}$$

where $\varepsilon_0(M|C=1)$ is the error of the best predictor of M in the absence of observations under the condition that C is turned on and $\varepsilon^{\circ}(\boldsymbol{S}, M|C=1)$ is the error of the best predictor of M based on the observation of $\boldsymbol{S}$ when C is on. Change in control of $\boldsymbol{S}$ by M relative to the activity of C is defined by

$$\Delta CoD_{\boldsymbol{S}}(M|C) = CoD_{\boldsymbol{S}}(M|C=0) - CoD_{\boldsymbol{S}}(M|C=1). \tag{7}$$

A positive value of $\Delta CoD_{\boldsymbol{S}}(M|C)$ indicates that C incapapcitates M as C is turned on. We call this value the *incapacitating power* (IP) of C relative to the regulation of $\boldsymbol{S}$ by M. A negative value of $\Delta CoD_{\boldsymbol{S}}(M|C)$ means that there is an increase in control of M over $\boldsymbol{S}$ as C is turned on and the magnitude is referred to as the *enhancing power* (EP) of C with respect to M upon the activation of C. This can be written as

$$\left|\Delta CoD_{\boldsymbol{S}}(M|C)\right| = \begin{cases} IP_{\boldsymbol{S}}(M|C) & \text{if } \Delta CoD_{\boldsymbol{S}}(M|C) > 0 \\ EP_{\boldsymbol{S}}(M|C) & \text{if } \Delta CoD_{\boldsymbol{S}}(M|C) < 0 \end{cases} \tag{8}$$

Equation (7) can be generalized to equation (9) if one wants to consider all possible subsets of size $d \le m$ of predictors in $\boldsymbol{S}$ being used:

$$\Delta CoD_{\boldsymbol{S},d}(M|C) = \frac{\sum_{i=1}^{{}_mC_d} CoD_{\boldsymbol{S}_d^{(i)}}(M|C=0) - CoD_{\boldsymbol{S}_d^{(i)}}(M|C=1)}{{}_mC_d} \tag{9}$$

where $\boldsymbol{S}_d^{(i)} \in \mathbb{R}^d$ is ith d-dimensional vector consisting of the entries

of $\boldsymbol{S} = (S_1, \ldots S_m)$ when all possible combinations of size d from $\boldsymbol{S}$ are lexicographically ordered for $i = 1, \ldots, {}_mC_d$.

2.5 Canalizing power

In this section, we define the *canalizing power* of the gene C, $CP_{\boldsymbol{S}\cup\boldsymbol{M}}(C)$, as a quantitative measure of canalization potential of a gene C relative to the set of genes $\boldsymbol{S}\cup\boldsymbol{M} = \{S_1, \ldots, S_m, M_1, \ldots, M_p\}$, where $\boldsymbol{S} = \{S_1, \ldots, S_m\}$ and $\boldsymbol{M} = \{M_1, \ldots, M_p\}$ are sets of slave genes and master genes, respectively. The canalizing power of gene C is expressed in terms of the regulation power and incapacitating power of C. This follows from the intuition that canalizing power should be quantified by the control of a gene C on the overall network and the extent of control that the gene C takes over from master genes when C is activated, which is equivalent to reduction of the control of the master genes $\boldsymbol{M}$ due to gene C. Thus, the canalizing power of gene C is given by

$$\begin{aligned} CP_{\boldsymbol{S}\cup\boldsymbol{M}}(C) &= RP_{\boldsymbol{S}\cup\boldsymbol{M}}(C) + \sum_i IP_{\boldsymbol{S}}(M_i|C) \\ &= RP_{\boldsymbol{S}\cup\boldsymbol{M}}(C) + \sum_i \Delta CoD_{\boldsymbol{S}}(M_i|C) \\ &\quad \times 1[CoD_{\boldsymbol{S}}(M_i|C=0) - CoD_{\boldsymbol{S}}(M_i|C=1) > 0] \end{aligned} \tag{10}$$

where $1[\cdot]$ is an indicator function. $RP_{\boldsymbol{S}\cup\boldsymbol{M}}(C)$ and $IP_{\boldsymbol{S}}(M_i|C)$ can be obtained by replacing $\boldsymbol{X}$ with $\boldsymbol{S}\cup\boldsymbol{M}$ in (5) and M with M_i in (8). Note that the summation is over only those master genes that have been incapacitated by the activation of C.

2.6 Applications

Consider a network consisting of n genes and assume that it has a canalizing gene and one is interested in detecting it. One possible approach to do this is to sort out all of the controlling genes which could be either a master gene or a canalizing gene by computing the mean CoD because both master and canalizing genes should exhibit high regulation power. Hypothesis testing based on user-selectable thresholds or statistical tools presented in (Chen and Braga-Neto, 2015, 2013) can be also used for picking out controlling genes. Suppose that we constitute a vector of controlling genes $\boldsymbol{X} = (X_1, \ldots, X_p)$ and slave genes $\boldsymbol{S} = (S_1, \ldots, S_l)$, where $n = p + l$. Furthermore, let $\boldsymbol{X}_{-c} = \{X_1, \ldots, X_{c-1}, X_{c+1} \ldots, X_p\}$ be the vector $\boldsymbol{X}$ without the element X_c. The canalizing power of X_c is

$$\begin{aligned} CP_{\boldsymbol{X}_{-c}\cup\boldsymbol{S}}(X_c) &= RP_{\boldsymbol{X}_{-c}\cup\boldsymbol{S}}(X_c) + \sum_{k\neq c} IP_{\boldsymbol{S}}(X_k|X_c) \\ &= RP_{\boldsymbol{X}_{-c}\cup\boldsymbol{S}}(X_c) + \sum_{k\neq c} \Delta CoD_{\boldsymbol{S}}(X_k|X_c) \\ &\quad \times 1[CoD_{\boldsymbol{S}}(X_k|X_c=0) - CoD_{\boldsymbol{S}}(X_k|X_c=1) > 0]. \end{aligned}$$

By taking turns, compute the canalizing power for each of the gene in the vector of controlling genes $\boldsymbol{X}$. The gene X_{i^*} possessing the maximum canalizing power is the most likely candidate for the canalizing gene with respect to our model assumptions, where

$$\begin{aligned} i^* &= \underset{c\in 1,\ldots,p}{argmax} CP_{\boldsymbol{X}_{-c}\cup\boldsymbol{S}}(X_c) \\ &= \underset{c\in 1,\ldots,p}{argmax} \left[RP_{\boldsymbol{X}_{-c}\cup\boldsymbol{S}}(X_c) + \sum_{k\neq c} IP_{\boldsymbol{S}}(X_k|X_c) \right] \end{aligned} \tag{11}$$

Since the power of incapacitation is a key attribute of canalizing genes which can be used to distinguish canalizing genes from master genes, only the second term in (11) can be utilized for a fast approximate search. Thus,

$$\begin{aligned} i^* &\approx \underset{c\in 1,\ldots,p}{argmax} \sum_{k\neq c} IP_{\boldsymbol{S}}(X_k|X_c) \\ &= \underset{c\in 1,\ldots,p}{argmax} \sum_{k\neq c} \Delta CoD_{\boldsymbol{S}}(X_k|X_c) \\ &\quad \times 1[CoD_{\boldsymbol{S}}(X_k|X_c=0) - CoD_{\boldsymbol{S}}(X_k|X_c=1) > 0] \end{aligned}$$

3 Results

In this section, we illustrate the application of canalizing power in a number of experiments using both synthetic data and real data sets.

3.1 Synthetic data

We generate a synthetic BN with $n = 10$ genes as shown in Figure 1 which is composed of one canalizing gene C, two master genes M_1 and M_2 and three levels of slave genes $S_{11}, \ldots S_{32}$. Regulatory influences on downstream genes are transferred between master genes and the canalizing gene depending on the activity of C. Thus, Boolean functions that govern the activity of downstream genes are designed to differ according to the expression of the canalizing gene C and therefore, the canalizing gene is embedded in the network by these Boolean rules. When there is no noise, the system transitions in accordance with its structural rules as defined by the Boolean functions listed in Table 1. The regulation power, incapacitating power and canalizing power of controlling genes are measured at each time point along the network evolution under various settings of the

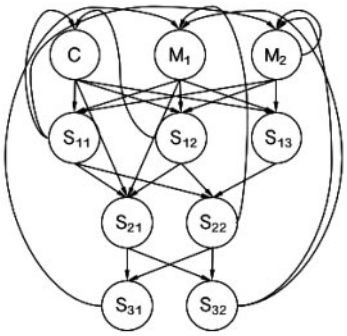

Fig. 1. A synthetic BN with $n = 10$ genes which is composed of one canalizing gene C, two master genes M_1 and M_2 and three levels of slave genes $S_{11}, \ldots S_{32}$. Upstream genes C, M_1 and M_2 regulate downstream genes and downstream genes provide feedback signals to the upregulators

Table 1. Boolean functions of genes in the synthetic BN

		Boolean expression	C inactivated	C activated
Controlling Genes	C	$S_{11} \oplus S_{12}$		
	M_1	$S_{22} \wedge (\overline{S_{31}} \oplus S_{32})$		
	M_2	$M_2 \wedge S_{11} \wedge S_{32}$		
Level 1	S_{11}	$C \vee M_1 \vee M_2$	$M_1 \vee M_2$	C
	S_{12}	$\overline{C} \wedge M_2$	M_2	$\overline{C}$
	S_{13}	$C \vee (M_1 \oplus M_2)$	$M_1 \oplus M_2$	C
Level 2	S_{21}	$S_{11} \wedge S_{12} \vee \overline{C} \wedge M_1$	$M_1 \vee M_2$	$\overline{C}$
	S_{22}	$S_{11} \vee S_{12} \wedge \overline{S_{13}}$	$M_1 \vee M_2$	C
Level 3	S_{31}	$S_{21} \wedge S_{22}$	$M_1 \vee M_2$	$\overline{C}$
	S_{32}	$S_{21} \oplus S_{22}$	0	C

Note: The symbols $\vee$, $\wedge$ and $\oplus$ denote the Boolean disjunction, conjunction and exclusive-OR, respectively.

model parameters. Each target is predicted by $d = 3$ predictors. Given the network, we consider four different simulation scenarios: (i) no gene is perturbed, (ii) only one specific gene is perturbed while other genes are noiseless, (iii) all genes are perturbed with equal probability and (iv) all of the genes are susceptible to noise where the perturbation probability for each gene is randomly generated from a beta distribution. Since the behavior of the network depends not only on the perturbation probabilities but also on the initial state distribution, we compute the average RP, IP and CP over ten thousand random generations of its initial joint probability distribution, D_0.

In the first case, no gene is perturbed and we plot the mean RP, IP and CP measured at each time step averaged over 10, 000 random starting joint probability distributions. Figure 2A shows that the mean regulation power of C is similar to or even less than that of M_1, whereas incapacitating power is exhibited only for the canalizing gene. This leads to higher CP of C, which indicates that the incapacitating power is a key attribute of canalizing genes that can be used to distinguish canalizing genes from other controlling genes.

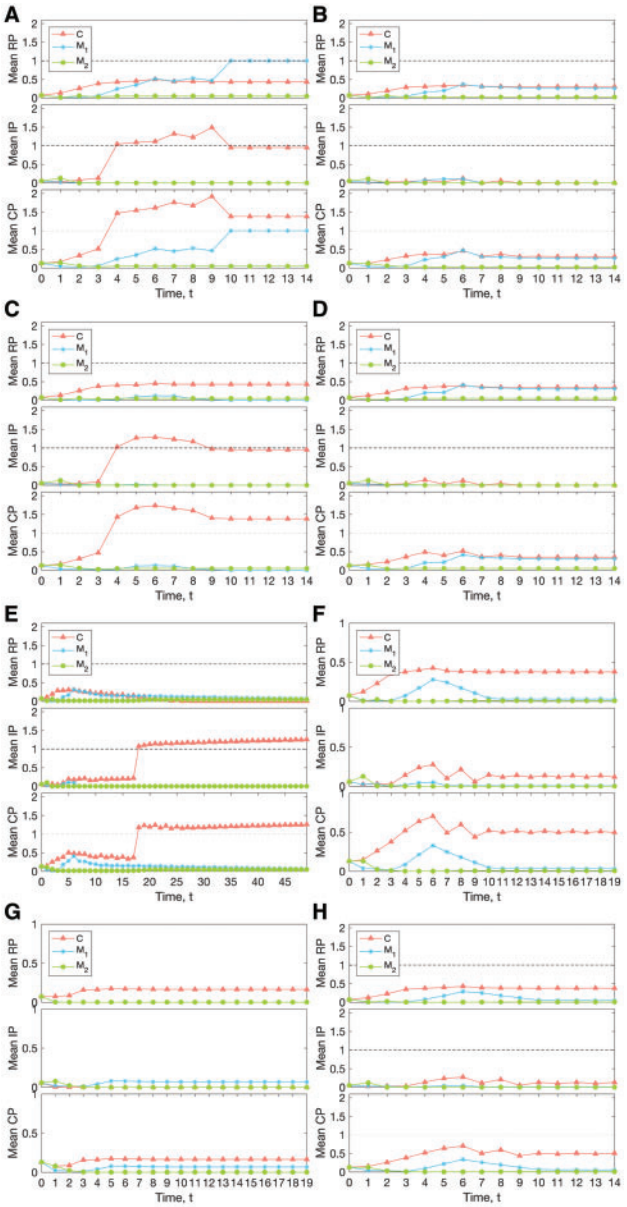

Fig. 2. Mean regulation power, incapacitating power and canalizing power over time **(A)** when no gene is perturbed. **(B)** A particular controlling gene is perturbed with $P_C = 0.1$ and **(C)** $P_{M_1} = 0.1$. **(D)** Effects of noise in the expression of downstream genes on mean RP, IP and CP when $P_{S_{12}} = 0.1$ and **(E)** $P_{S_{22}} = 0.1$. All genes are perturbed with the same probability **(F)** $P = 0.01$ and **(G)** $P = 0.1$. **(H)** All genes are perturbed with different probabilities which are randomly generated from *beta*(2, 200) distribution

The results for the second case where only one specific controlling gene is perturbed are presented in Figure 2B and C. In Figure 2B, the canalizing gene is perturbed with a probability $P_C = 0.1$ and other genes are not perturbed at all. Presence of noise in the canalizing gene corrupts its gene expression resulting in its lower IP, which negatively impacts its canalizing power. A case where only M_1 is perturbed with a probability $P_{M_1} = 0.1$ is shown in Figure 2C. In concordance with intuition, RP and IP of the canalizing gene is hardly impacted which does not compromise its predominance in CP. Effects of noise imposed on each of the downstream genes S_{12} and S_{22} with perturbation probabilities $P_{S_{12}} = 0.1$ and $P_{S_{22}} = 0.1$ are presented in Figure 2D and E, respectively. S_{12} is given as an input to C and therefore, IP of C deteriorates substantially which makes CP of C contiguous to that of M_1 across the timeline when $P_{S_{12}} = 0.1$ (Fig. 2D). S_{22} provides a feedback signal to the master gene M_1, thus, the RP of M_1 dwindles when $P_{S_{22}} = 0.1$ as illustrated in Figure 2E. While there is little noticeable distinction in regulation power between C and M_1, IP of the canalizing gene is remarkably higher in comparison to the rest of controlling genes, resulting in CP of C being greater than M_1 and M_2 (Fig. 2E).

For the next group of experiments, all of the genes are equally perturbed. Figure 2F shows that when the common perturbation probability is relatively small, $P = 0.01$, C experiences a decrease in its IP and CP. However, it still remains the gene with the highest canalizing potential in the network. When the perturbation probability is increased to $P = 0.1$, IP of C is virtually nonexistent and mean canalizing power of C has fallen below 0.18 as depicted in Figure 2G. This suggests that the amount of noise in the regulatory network could negatively affect the canalizing potential of certain genes.

For the final group of simulation experiments, all genes are perturbed with different probabilities. The beta distribution, which is defined on the interval $[0, 1]$, can represent all the possible values of a probability and it is widely used as a probability distribution of probabilities (Mun, 2015). The perturbation probability for each gene is randomly generated from a beta distribution with two parameters $\alpha = 2$, $\beta = 200$, which is a right-skewed distribution with mean 0.0099 to introduce a moderately small perturbation. The results are displayed in Figure 2H. When the entire network is exposed to such type of random noise, RP of M_1 decreases over time and while the canalizing gene C remains the most potent canalizer in the network despite its diminished IP. Boxplots of incapacitating power of controlling genes measured at $t = 14$ are shown in Figure 3 and the first boxplot represents a decrease in control of M_1 over downstream genes as C is turned on. The boxplots are based on the 10 000 samples which are generated from random starting joint probability distributions under the same experimental conditions as used for Figure 2H. The expected canalizing power of C is clearly higher than the CP of the rest of the controllers in the network: $E[CP_C] = 0.483$, $E[CP_{M_1}] = 0.053$, $E[CP_{M_2}] = 0.005$.

3.2 Real data

In this section, the proposed definition of the canalizing power is applied to a real data set from a study on ionizing radiaiton (IR) responsive genes in (Kim *et al.*, 2000) to assess the usefulness of our quantification in characterizing a canalizing gene. Note that our

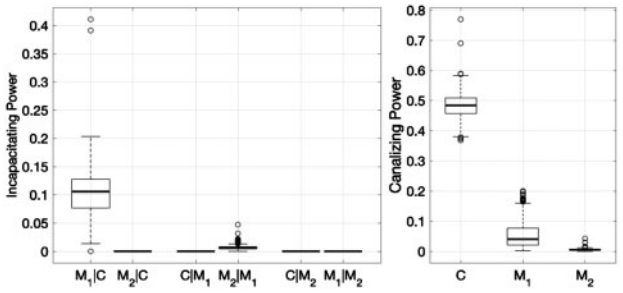

Fig. 3. Boxplots of IP and CP of upstream genes. The label $M_1|C$ on the horizontal axis of the left panel represents that the first boxplot indicates a decrease in control of M_1 over downstream genes as C is turned on. The boxplots are based on the data measured at $t = 14$ and generated from random starting joint probability distributions

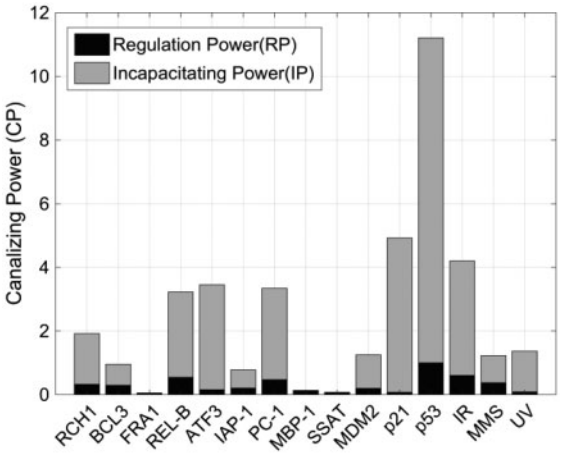

Fig. 4. A stacked bar chart of CP for each gene in the real dataset. The height of the black and gray bar segments represent contributions of RP and IP to CP, respectively

goal is not to discover new canalizing genes, but rather to illustrate the potential of our measurement on well-known canalizing genes. The data set consists of 12 genes under three conditions (i.e., IR, MMS, UV) in 30 cell lines of both *p53* proficient and *p53* deficient cells. The data are ternary, indicating up-regulated (+1), down-regulated (−1), or no-change (0) status. Here we map this to binary expressions using the following rules: change (1), for either up-regulated or down-regulated genes, and no-change (0). Additionally, we consider the three binary conditions (IR, MMS, and UV) as possible predictive factors, for a total of 15 Boolean variables in the BN model of this data set. We then apply the definitions of regulation power, incapacitating power and canalizing power outlined in the previous section. Figure 4 shows a bar chart with the canalizing power of each gene when triple predictors are used ($d = 3$). It is stacked to display the regulation power and incapacitating power of each gene. *p53* turns out to be the most powerful canalizing gene in the data set. This is in accordance with the known fact that *p53* is kept at a low level/dephosphorylated in unstressed cells and becomes significantly activated/phosphorylated in response to environmental stresses like UV, IR and oxidative stress, leading to a quick accumulation of *p53* in stressed cells (Collot-Teixeira *et al.*, 2004).

4 Discussion

It is a well-established notion in biology that canalizing genes possess broad regulatory power and can enforce broad corrective actions. Canalizing genes can be extremely potent not only because they produce optimal reactions to operating errors and external stimuli, but also because they don't act alone. Canalizing genes are more like master switches that set in motion a cascade of regulatory events that have huge impacts on downstream genes for the sake of driving the system to a desired condition. Discovering such potential drug targets that affect the disease trajectories is a strong step toward significant therapeutic benefits. From the perspective of optimal control, this is viewed as obtaining the best estimates of inputs which are most probable to elicit certain behavior of the network. However, the detection of these genes is circumscribed by their particular behavior. Under normal cell conditions, canalizing genes are not active and they are turned on only when cells encounter unfavorable situation. One of the most intensively studied tumor suppressor genes, p53 best describes this situation in which it is found at very low levels in normal cells while it is frequently observed in its phosphorylated state cancer-prone cells.

Although there have been several studies attempting to mathematically characterize canalizing genes and their power, they all missed the opportunity to characterize an important property of canalizing genes; that is, their incapacitating power. Therefore, we introduce a conditional CoD that characterizes predictive power of a set of genes with respect to a target gene under a specific condition of other genes. Our approach also suggests that the currently adopted definitions of canalizing and master genes could be modified so that a particular gene does not have to be exclusively a master or a canalizing gene. The newly introduced canalizing power resides in the continuous domain; therefore it presents a relative characterization of controlling genes. Although we have focused on BNs with perturbations to validate our ideas in a simplified environment, the same concept can be easily extended to Probabilistic Boolean Networks (PBNs), which offers more model flexibility.

Funding

This work was supported in part National Institutes of Health Grants [U01CA-162077, P30ES-023512].

Conflict of Interest: none declared.

References

Lehner,B. (2010) Genes confer similar robustness to environmental, stochastic, and genetic perturbations in yeast. *PLoS One*, **5**, e9035.

Chang,L. and Karin,M. (2001) Mammalian MAP kinase signalling cascades. *Nature*, **410**, 37–40.

Chen,T. and Braga-Neto,U.M. (2013) Statistical detection of boolean regulatory relationships. *IEEE/ACM Trans. Comput. Biol. Bioinfo.*, **10**, 1–1.

Chen,T. and Braga-Neto,U.M. (2015) Statistical detection of intrinsically multivariate predictive genes. *IEEE/ACM Trans. Comput. Biol. Bioinfo.*, **12**, 951–963.

Collot-Teixeira,S. *et al.* (2004) Human tumor suppressor p53 and DNA viruses. *Reviews Med. Virol.*, **14**, 301–319.

Comp Arch And Org, 2E (2010) *Comp Arch and Org*, **2E**.

Dougherty,E.R. *et al.* (2000) Coefficient of determination in nonlinear signal processing. *Signal Process.*, **80**, 2219–2235.

Gomez-Lazaro,M. *et al.* (2004) p53: twenty five years understanding the mechanism of genome protection. *J. Physiol. Biochem*, **60**, 287–307.

Kim,S. *et al.* (2000) General nonlinear framework for the analysis of gene interaction via multivariate expression arrays. *J. Biomed. Opt.*, **5**, 411–425.

Linux Device,D. (2001) Linux Device Drivers.

Martins,D.C. *et al.* (2008) Intrinsically Multivariate Predictive Genes. *IEEE J. Sel. Top. Signal Process*, **2**, 424–439.

Mun,J. (ed.) (2015) Understanding and choosing the right probability distributions. In: *Advanced Analytical Models, over 800 Models and 300 Applications from the Basel II Accord to Wall Street and Beyond*. Wiley, Hoboken, NJ, USA, pp. 899–917.

Shmulevich,I. and Zhang,W. (2002) Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics*, **18**, 555–565.

Shmulevich,I. *et al*. (2002) Gene perturbation and intervention in probabilistic Boolean networks. *Bioinformatics*, **18**, 1319–1331.

Shmulevich,I. *et al*. (2002) Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, **18**, 261–274.

Tabin,C.J., and Weinberg,R.A. (1985) Analysis of viral and somatic activations of the cHa-ras gene. *J. Virol*., **53**, 260–265.

Waddington,C.H. (1942) Canalization of development and the inheritance of acquired characters. *Nature*, **150**, 563–565.

Wagner,A (2005) *Robustness and Evolvability in Living Systems*. Princeton University Press, Princeton.

Zhao,C. *et al*. (2011) Pathway regulatory analysis in the context of bayesian networks using the coefficient of determination. *J. Biol. Syst*., **19**, 651–682.

Zhou,X. *et al*. (2003) Binarization of microarray data on the basis of a mixture model1. *Mol. Cancer Ther*., **2**, 679–684.