Statistical Detection of Boolean Regulatory Relationships

Ting Chen, Student Member, IEEE, and Ulisses M. Braga-Neto, Senior Member, IEEE

Abstract—A statistic tool for the detection of multivariate Boolean relationships is presented, with applications in the inference of gene regulatory mechanisms. A statistical test is developed for the detection of a nonzero discrete Coefficient of Determination (CoD) between predictor and target variables. This is done by framing the problem in the context of a stochastic logic model that naturally allows the inclusion of prior knowledge if available. The rejection region, p-value, statistical power, and confidence interval are derived and analyzed. Furthermore, the issue of multiplicity of tests due to presence of numerous candidate genes and logic relationships is addressed via FWER- and FDR-controlling approaches. The methodology is demonstrated by experiments using synthetic data and real data from a study on ionizing radiation (IR) responsive genes. The results indicate that the proposed methodology is a promising tool for detection of gene regulatory relationships from gene-expression data. Software that implements the COD test is available online as an R package.

Index Terms—Coefficient of Determination; Stochastic Logic Model; Multiple Testing Procedure; Family-wise Error Rate; False Discovery Rate

I. INTRODUCTION

DNA regulatory circuits can be often described by networks of Boolean logical gates updated and observed at discrete time intervals [1]–[5]. In a stochastic setting, the degree of association between Boolean predictors and targets can be quantified by means of the discrete Coefficient of Determination (CoD) [6], which measures the relative decrease in prediction error when using predictor variables to estimate the target variable, as opposed to using no predictor variables.

The CoD is a function only of the joint distribution between predictors and target, and is thus a measure of the stochastic nonlinear interaction among genes. A high CoD value indicates tight gene regulation, whereas a low CoD indicates loose regulation. The CoD has found significant applications in Genomics [6]–[11]. The CoD is often used in the inference of gene regulatory networks from gene-expression data [8], [9], [11].

However, applications of the CoD so far have been based on user-selected thresholds to decide on the presence of gene regulation between the given predictor and target genes. The present paper develops a statistically rigorous tool for this inference problem, by providing a statistical test, and associated confidence interval, for a nonzero CoD between given Boolean predictors and a Boolean target. Rejection of the null hypothesis of zero CoD gives evidence for the presence of statistically-significant regulation. Even though the user still needs to choose the significance level, substituting this choice for the choice of an arbitrary CoD threshold has nevertheless advantages, beyond the fact that “standard” significance levels are available, such as $\alpha = 0.05$. The significance level can be interpreted as an upper bound on the false positive rate, whereas no such statistical interpretation can be attached to a user-selected CoD threshold.

This is accomplished by framing the problem in the context of a stochastic logic model that naturally allows the inclusion of prior knowledge if available; e.g., knowledge about the logic gate governing the relationship sought for. For example, knowledge about a canalizing relationship [12], i.e. a logic relationship in a class of AND or OR gates (with possibly negated inputs), can be easily added. Then an Intersection-Union Test (IUT) [13] based on likelihood-ratio tests for the individual model parameters is developed by deriving its rejection region, power, p-value, and associated confidence interval.

To be useful as an inferential tool, the proposed methodology must be able to deal with the multiple testing issue created by modern gene-expression experiments that monitor thousands of genes simultaneously. We address this by describing the application of two multiple testing procedures to control the overall Type I error rate, namely the single-step Bonferroni correction and the step-up Benjamini-Hochberg procedure, for controlling the family-wise error rate (FWER) and the false discovery rate (FDR), respectively [14], [15].

The properties of the proposed statistical test and multiple testing correction procedures are assessed by both theoretical analysis and Monte-Carlo experiments, in order to analyze how FWER, FDR, average power, and the confidence interval estimates behave under FWER- and FDR-controlling procedures, for varying sample size and number of multiple tests. To further evaluate the proposed methodology, we apply it to a real data set from a study on ionizing radiation (IR) responsive genes in [9]. This analysis reveals, for instance, that $p53$ possesses the largest number of significant predictive relationship, which is in accordance with the known fact that $p53$ is a significantly active gene involved in various pathways associated with stress responses. Additional relationships are detected that are consistent with current biological
knowledge of such pathways, among many other potentially useful regulatory relationships uncovered in the experiment. These preliminary results indicate that the proposed methodology could serve as a promising tool for discovery of gene regulatory relationships from gene-expression data. To our knowledge, this is the first work to define a rigorous hypothesis test for a multivariate Boolean association among binary random variables; in particular, a direct comparison with alternative methods was not possible.

The paper is organized as follows. In Section II, we present the model for stochastic Boolean relationships used by the proposed method, and the coefficient of determination based on the model. In Section III, we derive and analyze the rejection region, power, p-value, and confidence interval for the proposed CoD statistical test. In Section IV, we describe the FWER- and FDR-controlling multiple testing procedures for simultaneous detection of regulatory relationships. Section IV-C evaluates how FWER, FDR and average power are controlled under FWER- and FDR-controlling approaches, for varying sample size and number of multiple tests. In Section V, we apply the proposed methodology to expression data of genes responsive to ionization stresses. Finally, Section VI gives concluding remarks.

II. MATHEMATICAL PRELIMINARIES

A. Stochastic Boolean Model

After continuous measurements of gene expression have been binarized, a step that is not discussed here — for optimal methods to do this, see for example [16], [17] — the sample data consist of a binary target random variable \( Y \) and a vector of binary predictor random variables \( X = (X_1, \ldots, X_d) \in \{0,1\}^d \). Due to uncertainty, noise affects the Boolean relationship between the predictors and the target, which is addressed here by a simple Boolean “additive-noise” model, which has been discussed in our recent work [18], [19]:

\[
Y = f(X) \oplus N, \tag{1}
\]

where \( f: \{0,1\}^d \rightarrow \{0,1\} \) is a Boolean logic function, the symbol “\( \oplus \)” indicates modulo-2 addition, and the noise \( N \) is a Bernoulli random variable that is independent of \( X \) and \( Y \), such that \( P(N = 1) = 1 - p \), for \( 1/2 \leq p \leq 1 \). Here, \( 1 - p \) measures the amplitude of the noise. When \( N = 1 \), the value of \( Y \) is flipped because of the noise, otherwise it remains unperturbed. It follows that \( P(Y = f(X)) = p \), which indicates that \( p \) measures how strongly \( X \) predicts \( Y \), i.e., \( p \) is the predictiveness of the model. Notice that the case \( p < 1/2 \) need not be considered, since it corresponds to a negated logic gate with predictive power \( 1 - p \). When \( p = 1 \), the model is deterministic without noise, whereas, when \( p = 1/2 \), there is maximum confusion, i.e., no association between target and predictor.

We remark that the conditional distribution of the target given the predictor can be written entirely as a function of the logic function \( f \) and the parameter \( p \):

\[
P(Y = 1 \mid X = x) = I(f(x) = 1)p + I(f(x) = 0)(1 - p), \tag{2}
\]

where \( I(A) \) is 1 when \( A \) is true, and 0, otherwise.

B. Coefficient of Determination

The Coefficient of Determination (CoD) of \( X \) with respect to \( Y \) is defined to be [6]:

\[
\text{CoD} = \frac{\varepsilon_Y - \varepsilon_{X,Y}}{\varepsilon_Y} \tag{3}
\]

where \( \varepsilon_Y = \min\{P(Y = 0), P(Y = 1)\} \) is the optimal error of predicting \( Y \) in the absence of observations and \( \varepsilon_{X,Y} \) is the optimal error upon observation of \( X \). By convention, one assumes \( 0/0 = 1 \) in the above definition.

Let \( \xi = P(f(X) = 1) \); as we shall see, this distributional quantity plays a fundamental role in the sequel. In the context of model (1), it can be shown easily, by using (2), that the CoD is given by

\[
\text{CoD} = 1 - \frac{1 - p}{F[\xi_p + (1 - \xi)(1 - p)]}, \tag{4}
\]

where \( F[u] = \min\{u, 1 - u\} \), for \( 0 \leq u \leq 1 \). The CoD is therefore a function of the distributional parameters \( p \geq 1/2 \) and \( 0 \leq \xi \leq 1 \). Note that deterministic prediction is a function of \( p \) only: \( \text{CoD} = 1 \Leftrightarrow \varepsilon_{X,Y} = 0 \Leftrightarrow p = 1 \). The case \( \text{CoD} = 0 \) (i.e., no regulation) depends on both \( p \) and \( \xi \), and is stated in the next proposition.

**Proposition 1.** In the context of model (1), the following statements are equivalent:

(i) \( \text{CoD} = 0 \).

(ii) \( p = 1/2 \) or \( \xi \in \{0,1\} \).

**Proof.** The result follows from equating the numerator and denominator in the ratio appearing in (4). Q.E.D.

For \( 0 < \xi < 1 \), Proposition 1 assures us that \( \text{CoD} = 0 \Leftrightarrow p = 1/2 \), i.e., maximum noise. This would be the case, regardless of logic, if \( P(X = x) > 0 \) for all \( x \in \{0,1\} \). Without distributional knowledge, one cannot however ignore the boundary condition \( \xi \in \{0,1\} \) when testing for null CoD.

As a concrete example, consider the case of \( d = 2 \) predictors, \( X = (X_1, X_2) \). In this case, there are a total of \( 2^{2^2} = 16 \) possible prediction logics. Among those, six are either constant or depend only on one of the predictors, namely, \( 0, 1, X_1, X_2, X_1 \), and \( X_2 \). The remaining 10 logics are “true” 2-input logics, namely \( X_1 \land X_2 \) (AND), \( X_1 \lor X_2 \) (OR), \( X_1 \lor X_2 \) (XOR), \( X_1 \land X_2 \) and their negations. Logics can be represented by a bit string corresponding to the output column in its truth table; for example, 0001 (AND), 0111 (OR), 0110 (XOR), 0100 \((X_1 \land X_2)\), and 0010 \((X_1 \lor X_2)\). The bit string representation is particularly convenient when checking the distributional constraint \( \xi \in \{0,1\} \) in condition (ii) of Proposition 1. Now, note that if logic \( f \) is the negation of logic \( f \), then \( \xi = 1 - \xi \), so that the constraint
ξ ∈ {0, 1}, and in fact the expression for the CoD in (4), are the same for f and 1−f, as can be easily checked. Among the 10 2-input logics, there are therefore a total of five cases to consider, which are listed in Table I.

Similarly, for the case of CoD = θ ∈ [0, 1], we can prove that
\[
\text{CoD} = \theta \iff p = \frac{\theta + 2 \min\{\xi, 1 - \xi\}}{\delta + 2 \min\{\xi, 1 - \xi\}}.
\]
(5)

where \(\delta = \frac{\theta}{1-\theta}\). A small value of \(\theta\) implies a loose regulation between a target and its predictors, whereas a large value implies a tight regulation.

III. CoD Hypothesis Test

The CoD is a function of the distribution parameters \(p\) and \(\xi\) of \((X, Y)\), c.f. (4), and therefore statements about it can be statistically tested based on an i.i.d. sample \(S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}\) [13]. In particular, we are interested in the following hypothesis testing problem:

\[
H_0 : \text{CoD} = 0 \quad (p = 1/2 \text{ or } \xi \in \{0, 1\})
\]
\[
H_1 : \text{CoD} > 0 \quad (p > 1/2 \text{ and } 0 < \xi < 1).
\]
(6)

The null hypothesis \(H_0\) indicates the absence of useful prediction in \(X\) concerning the target \(Y\), whereas the alternative hypothesis \(H_1\) states that there is a degree of association between them.

This is a composite, multivariate hypothesis testing problem. As the null parameter space is a union of two subsets \([p = 1/2]\) and \([\xi \in \{0, 1\}]\), the appropriate strategy to employ here is the intersection-union test (IUT) method: the individual tests for \(p = 1/2\) and \(\xi \in \{0, 1\}\) are level-\(\alpha\) likelihood-ratio tests (LRTs), leading to an overall level-\(\alpha\) IUT test [20], [21]. This is summarized in the following result (details are found in the Appendix in the supplementary material).

Proposition 2. For given 0 ≤ \(\alpha\) ≤ 1, the test with rejection region

\[\mathcal{R} = \{s_n \mid \sum_{i=1}^{n} I(f(x_i) = y_i) \geq k \quad \text{and} \quad \exists 1 \leq i, j \leq n \text{ s.t. } f(x_i) \neq f(x_j)\}\],
(7)

where \(k\) is the 100(1 - \(\alpha\))\% percentile of a Binomial(\(n, 1/2\)) distribution, i.e., \(k\) is the smallest integer such that

\[\sum_{l \geq k} \binom{n}{l} \left(\frac{1}{2}\right)^n \leq \alpha\],
(8)

is a level-\(\alpha\) test for (6).

Proof. See Appendix in the supplementary material.

The following statements follow from Proposition 2.

1) Rejection region. Notice that \(\mathcal{R} = \mathcal{R}_1 \cap \mathcal{R}_2\), where \(\mathcal{R}_1 = \{s_n \mid \sum_{i=1}^{n} I(f(x_i) = y_i) \geq k\}\) is the rejection region for the \([p = 1/2]\) LRT, and expresses how tightly the data follows the proposed model, while \(\mathcal{R}_2 = \{s_n \mid \exists 1 \leq i, j \leq n \text{ s.t. } f(x_i) \neq f(x_j)\}\) is the rejection region for the \([\xi \in \{0, 1\}]\) LRT, and indicates that the null hypothesis cannot be rejected if \(f(x_i)\) is constant, for \(i = 1, \ldots, n\). Notice that

\[P_{\mathcal{R}_2} = P(S_n \in \mathcal{R}_2) = 1 - P([f(X_i) = 1, \forall i = 1, \ldots, n] \cup [f(X_i) = 0, \forall i = 1, \ldots, n]) = 1 - \xi^n - (1 - \xi)^n.
\]
(9)

It follows that, unless \(\xi \in \{0, 1\}\), in which case \(S_n \notin \mathcal{R}_2\) with probability 1, we have \(P_{\mathcal{R}_2} \rightarrow 1\) as sample size increases to infinity. Therefore, the criterion for rejecting the null hypothesis will be, with probability approaching 1, whether or not \(S_n \in \mathcal{R}_1\), and the proposed test approaches an LRT for \(p = 1/2\).

2) p-value. The rejection regions for varying significance level \(\alpha\) are nested, that is, \(\mathcal{R}(\alpha_1) \subseteq \mathcal{R}(\alpha_2)\), whenever \(\alpha_1 \leq \alpha_2\). This allows us to define a p-value for the proposed test as

\[\pi(s_n) = \begin{cases} \sum_{l \geq k} \binom{n}{l} \left(\frac{1}{2}\right)^n, & \text{if } s_n \in \mathcal{R}_2 \\ 1, & \text{otherwise.} \end{cases}
\]
(10)

It is clear that \(\pi(s_n)\) is a valid p-value [13], i.e., under the null hypothesis, \(P(\pi(s_n) \leq u) \leq u\), for all \(0 \leq u \leq 1\).

3) Statistical power. The power function [13] of the proposed test can be shown to be

\[\beta(p, \xi) = P(S_n \in \mathcal{R}) = \left(\sum_{l \geq k} \binom{n}{l} p^l (1-p)^{n-l}\right) \times (1 - \xi^n - (1 - \xi)^n),
\]
(11)

for \(p \geq 1/2\) and \(0 \leq \xi \leq 1\), where \(k\) is given by (8). Note that, under the null hypothesis, either \(\beta(p, \xi) = 0\), if \(\xi \in \{0, 1\}\), or \(\beta(p, \xi) \leq \alpha\), if \(p = 1/2\) and \(0 < \xi < 1\) (by virtue of eq. 8). Therefore, \(\sup \beta(p, \xi) \leq \alpha\) under the null hypothesis, so that this is indeed an \(\alpha\)-level test. Under the alternative hypothesis, \(\beta(p, \xi)\) gives the statistical power of the test. Notice from (11) that \(\beta(p, \xi)\) not only on the distributional parameters \(p\) and \(\xi\), but also on the level \(\alpha\) sample size \(n\), and logic function \(f\) (through \(\xi\)). Therefore, a power analysis has to take into account all of these factors. We consider below two important special cases for statistical power, where the analysis is facilitated.

Uniform (independent) predictors. Consider a uniform predictor distribution, \(P(X = x) = 1/2^d\), for \(x \in \{0, 1\}^d\). It is easy to see that this implies that the individual predictors \(X_1, \ldots, X_d\) are independent. Clearly, \(\xi = m/2^d\), where \(m\) is the number of minterms of logic \(f\), i.e., the number of 1’s in its bit string representation (c.f. Section II-B). The cases \(m = 0\) and \(m = 2^d\) are uninteresting, since they correspond to the constant logic \(f \equiv 0\) and \(f \equiv 1\), respectively. In addition, \(m\) and \(2^d - m\) lead to the same value for the CoD (c.f. equation 4), and hence for \(p\) and the power \(\beta(p, \xi)\).
Consider two predictors $X_1$ and $X_2$, such that $P(X_1) = P(X_2) = 1/2$; these are referred to as “unbiased” predictors in [7], [19]. Let
\[
\gamma = \text{Cov}(X_1, X_2) = E[X_1X_2] - E[X_1]E[X_2] = P(X_1 = 1, X_2 = 1) - \frac{1}{4}.
\]
From the constraint $P(X_1) = P(X_2) = 1/2$ it follows that $-\frac{1}{4} \leq \gamma \leq \frac{1}{4}$. When $\gamma = 0$ one obtains the case of uniform independent predictors previously considered, for $d = 2$.

With $d = 2$, there are only two families of useful logics to consider, according to number of minterms: the case $m = 1, 3$, represented here by the AND logic, and $m = 2$, represented here by the XOR logic. These cases correspond to the minimum (canalizing) and maximum (non-canalizing) number of minterms possible, respectively. For the AND logic, it is easy to see that $\xi = 1/4 + \gamma$. In addition, it can be shown that:
\[
\text{CoD} = \frac{(2p - 1)(1 + 4\gamma)}{4(1 - p) + (2p - 1)(1 + 4\gamma)} \Rightarrow
\]
\[
p = \frac{1}{2} \left( 1 + 4\gamma \right) + \left( 3 - 4\gamma \right) \text{CoD}.
\]
For the XOR logic, on the other hand, we have $\xi = 1/2 - 2\gamma$. Furthermore,

$$\text{CoD} = \frac{(2p - 1)(1 - 4|\gamma|)}{2(1 - p) + (2p - 1)(1 - 4|\gamma|)} \Rightarrow$$

$$p = \frac{1}{2} \frac{(1 - 4|\gamma|) + (1 + 4|\gamma|)\text{CoD}}{(1 - 4|\gamma|) + 4|\gamma|\text{CoD}}. \tag{15}$$

Substituting the expressions for $p$ and $\xi$ in each case above into (11) allows us to compute the power function in terms of the CoD effect size and the covariance parameter $\gamma$, which is displayed in Figure 3, for the AND and XOR logic cases. A few values of the covariance parameter are selected from the allowed interval $-1/4 < \gamma \leq 1/4$, but the case of perfectly negatively correlated predictors, $\gamma = -0.25$, is excluded, as it corresponds to the null hypothesis CoD = 0, in both AND and XOR cases. In addition, power is a function of $|\gamma|$ in the XOR case, so that only curves for $\gamma \geq 0$ are plotted (each of which give the cases of both positive $\gamma$ and negative $-\gamma$ correlation, of course). As in the previous example of uncorrelated predictors, we can see that for large sample size, power increases to 1 very rapidly with effect size. For $n = 500$, power decreases monotonically with increasing predictor correlation in the AND case, while it monotonically increases with increasing magnitude of predictor correlation, in the XOR case. However, as before, the behavior for small sample sizes is complex. It can be said that in the AND case, power generally is larger for negatively correlated predictors if the effect size is small, while positively correlated predictors lead to more powerful tests at large effect sizes. For the XOR logic, highly correlated predictors (regardless of sign) lead to more powerful tests for small effect size, while weakly correlated predictors produce more power at large effect sizes.

Figure 2(b-c) displays the minimum sample size necessary to achieve a standard power value of 80%, for the AND and XOR logic cases, respectively, and a few values of the covariance parameter in the allowed interval $-1/4 \leq \gamma \leq 1/4$. As in the previous example of uncorrelated predictors, we can see that the sample size requirement is monotonically decreasing with increasing CoD effect size. For small CoD effect size, the sample requirement is much larger for large values of covariance $\gamma$ (in the case of XOR, large values in magnitude). For large CoD size, the situation is reversed, dramatically so in the case of predictors with large negative correlation in the AND case, and uncorrelated predictors in the XOR case.

We remark than an extension of these results to $d \geq 3$ predictors is possible using an appropriate parametrization for the covariance structure of the predictor vector; such a parametrization is given in [19].

The results of the power analysis for the proposed test, displayed in Figures 1 and 3, may be summarized as follows. If a small CoD effect size is expected, then sample sizes in the neighborhood of $n = 100$ or larger are required for effective statistical power; in this case, small number of minterms (canalizing logics) lead to larger statistical power, while uncorrelated predictors lead to smaller power. If large CoD values, i.e., a tightly regulated target, is expected, then smaller sample sizes may be employed, as long as the logic of prediction contains a sufficiently large number of minterms and the predictors are weakly correlated, or, if the logic is closer to a canalizing type, the predictors are sufficiently positively correlated.

(4) **Confidence Interval.** A confidence interval for the CoD can be derived by considering a test of

$$H_0 : \text{CoD} = \theta \text{ vs. } H_1 : \text{CoD} \neq \theta, \tag{16}$$

where $\theta \in (0, 1)$. The likelihood ratio test statistic is given by ($\{i_1, \ldots, i_d\} \subseteq \{0, 1\}^d$)

$$\lambda(s_n; \theta) = \sup_{\text{CoD} = \theta} P(S_n = s_n) = \sup \frac{\text{CoD} = \theta}{P(S_n = s_n)} \frac{\text{CoD} \neq \theta}{P(S_n = s_n)} \sup (1 - \theta)^{n_f} \prod_{i \in \{i_1, \ldots, i_d\}} P(X = \{i_1, \ldots, i_d\})^{n_{i_1 \cdots i_d}}$$

$$= (n_f/n)^{n_f} (1 - n_f/n)^{n - n_f (n_{i_1 \cdots i_d}/n)^{n_{i_1 \cdots i_d}}}.$$  

where $n_f = \sum_{i=1}^{n} I(f(X_i) = Y_i)$, $n_{i_1 \cdots i_d} = \sum_{i=1}^{n} I(X_i = \{i_1, \ldots, i_d\})$, and $p$ is expressed by eq. (5). Note that the optimization problem of the numerator in eq. (17) can be solved by the method of gradient descent when there are multiple parameters [22].
under regularity conditions, the LRT statistic follows an asymptotic distribution, that is, under the \( H_0 \), as \( n \to \infty \), 
\[-2 \log \lambda (S_n; \theta) \to \chi^2_1[13]. \]
Hence, given some \( \theta \), the rejection region of such an asymptotic size \( \alpha \) test is formed by

\[ R = \left\{ s_n \mid -2 \log \lambda (S_n; \theta) \geq \chi^2_1(\alpha) \right\}, \tag{18} \]

where \( \lambda (S_n) \) is shown in eq. (17).

By inverting the LRT [13], the approximate \( 1 - \alpha \) confidence interval of the CoD, the set with plausible values of \( \theta \), is given by

\[ C(s_n) = \left\{ \theta \mid -2 \log \lambda (S_n; \theta) \leq \chi^2_1(\alpha) \right\}, \tag{19} \]

which can be numerically solved by the bisection method [23].

In the following, we consider again two important special cases (i.e., uniform and correlated predictors) for estimation of the confidence interval.

Uniform (independent) predictors. In the uniform predictor case, CoD is a function of only \( p \) (c.f. eq. 12). Since \( n_f = \sum_{i=1}^n I(f(x_i) = y_i) \) Binomial \((n, p)\), the Clopper-Pearson interval is employed to calculate the \( 1 - \alpha \) binomial confidence interval [24]. By substituting this confidence interval for \( p \) into eq. (12), we can obtain the confidence interval for the CoD, that is,

\[
\left[ \frac{m(2p_L - 1)}{mp_L + (2^d - m)(1 - p_L)} \right] \quad \left[ \frac{m(2p_U - 1)}{mp_L + (2^d - m)(1 - p_U)} \right]
\]

where \( p_L = \text{Beta}(\alpha/2; n_f, n - n_f + 1) \) and \( p_U = \text{Beta}(1 - \alpha/2; n_f + 1, n - n_f) \). Note that \( \text{Beta}(t; a, b) \)
is the \( t \)-th quantile from a beta distribution with parameters \( a \) and \( b \).

Correlated predictors, \( d = 2 \). In the correlated predictor case, the confidence interval is approximated by the asymptotic distribution, that is, \( \chi^2_1 \) distribution, as discussed in the general case. Table II shows the confidence interval estimate of the CoD based on random sample with \( n = 100 \) generated by a 2-input AND logic model in the general, uniform, and correlated predictor cases, respectively. We observe that the true values of \( \theta \) lie in the corresponding confidence intervals in all cases. Note that the approximation works better for a larger sample size.

### IV. Multiple Testing Procedure

For a given target \( Y \), the proposed test for multivariate Boolean relationships presupposes the model (1), which in turn depends on the choice of logic function \( f \) and predictor vector \( X \). Assuming dimensionality \( d \) and a number of genes \( G \) in the original gene-expression

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>CI (General)</th>
<th>CI (Uniform)</th>
<th>CI (Correlated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>[0.0000, 0.2153]</td>
<td>[-0.1229, 0.0668]</td>
<td>[0.0000, 0.1303]</td>
</tr>
<tr>
<td>0.1</td>
<td>[0.0000, 0.2369]</td>
<td>[0.0178, 0.2743]</td>
<td>[0.0096, 0.2506]</td>
</tr>
<tr>
<td>0.2</td>
<td>[0.0925, 0.4488]</td>
<td>[0.1113, 0.4037]</td>
<td>[0.0854, 0.3026]</td>
</tr>
<tr>
<td>0.3</td>
<td>[0.1023, 0.5545]</td>
<td>[0.2286, 0.5545]</td>
<td>[0.2917, 0.5201]</td>
</tr>
<tr>
<td>0.4</td>
<td>[0.1468, 0.4927]</td>
<td>[0.1813, 0.4953]</td>
<td>[0.3192, 0.4166]</td>
</tr>
</tbody>
</table>

For a given target \( Y \), the proposed test for multivariate Boolean relationships presupposes the model (1), which in turn depends on the choice of logic function \( f \) and predictor vector \( X \). Assuming dimensionality \( d \) and a number of genes \( G \) in the original gene-expression.
dataset, the total number of possible logic functions is $2^d$ and the number of distinct predictors is $\binom{G}{d}$. This creates a multiple testing issue; the total number of tests to be carried out would be, in this case, $M = 2^d \times \binom{G}{d}$. In typical gene-expression microarray or RNA-seq studies, $G$ tends to be very large (in the order of thousands or more) so that, even if $d$ is kept small, the number of tests may be very large indeed. In this section, we address the multiple testing problem in the context of the proposed detection method. We also comment on how to reduce the number of tests by use of prior knowledge.

A. Type-I Error Rates and Power

In a multiple testing procedure (MTP), there is a total of $M$ null hypotheses to be simultaneously tested, $\{H_0(m) \mid m = 1, \ldots, M\}$. While there is no ambiguity in defining a type-I error for a single test, in the case of MTPs the situation is less clear [25]. Let $0 \leq R \leq M$ be the number of hypotheses rejected by the test, and let $0 \leq V \leq R$ be the number of hypotheses falsely rejected (i.e., “false positives”). We consider in this paper two specific definitions of type-I error rates for MTPs:

- The family-wise error rate [26] is defined as $\text{FWER} = P(V \geq 1)$.
- The false discovery rate [14] is defined as

$$\text{FDR} = E \left[ \frac{V}{R} \mid R > 0 \right] = E \left[ \frac{V}{R} \right] P(R > 0). \quad (21)$$

The FWER gives the probability of at least one false positive, whereas the FDR essentially gives the average, or expected, proportion of false positives in the list of rejected hypotheses (with the proviso that, if no hypotheses are rejected, i.e., $R = 0$, then FDR = 0). It can be shown quite easily that the FDR is always smaller or equal than the FWER, with strict equality holding in the case where all the null hypotheses are true [14].

In the multiple testing procedures that control the Type-I error rate at a given level $\alpha$, one also expects to maximize power. We consider here the definition of the power for MTPs as formed by:

$$\text{PWR} = \frac{E[S]}{h_1}, \quad (22)$$

where $S$ is the true positives and $h_1$ is the number of false null hypotheses [15]. Obviously, The power gives the expected value of the proportion of true positives among the false null hypotheses. Note that the power estimate is mathematically equal to the true positive rate, that is, $S/h_1$.

B. Control of the Type-I Error Rate

For a given $0 < \alpha < 1$, an MTP is said to control the FWER at level $\alpha$ if FWER $\leq \alpha$. Similarly, an MTP is said to control the FDR at level $\alpha$ if FDR $\leq \alpha$. Notice that, since FDR $\leq$ FWER, any FWER-controlling procedure is also FDR-controlling, but the converse is not true in general, unless all null hypotheses are true, in which case FDR $=$ FWER, as mentioned previously.

Suppose that individual tests of the hypotheses $\{H_0(m) \mid m = 1, \ldots, M\}$ are performed, producing a set of (valid) unadjusted $p$-values $\{\pi_1, \ldots, \pi_M\}$. Let $\pi_m' = \min\{M \pi_m, 1\}, \quad m = 1, \ldots, M \quad (23)$

be the set of adjusted $p$-values. Then it can be shown, by an application of Boole’s inequality, that rejection of $H_0(m)$ if $\pi_m' \leq \alpha$, for $m = 1, \ldots, m$, is an MTP that controls the FWER at level $\alpha$ [25]. This is the well-known Bonferroni Correction method [26].

Similarly, let $\{\pi_1, \ldots, \pi_M\}$ be the list of unadjusted $p$-values sorted in increasing order, and define the set of adjusted $p$-values by

$$\pi_m'' = \min_{h=m,\ldots,M} \left\{ \min \left\{ \frac{M}{h} \pi^*(h), 1 \right\} \right\}, \quad (24)$$

Then it can be shown that rejection of $H_0(m)$ if $\pi_m'' \leq \alpha$, for $m = 1, \ldots, M$, is an MTP that controls the FDR at level $\alpha$, under the assumption of independence of the $p$-values for the true null hypotheses [14, Thm. 1] or for certain dependence structures among the $p$-values [27, Thm. 1.2]. If the $p$-values have an arbitrary dependence structure, the previous procedure will only control the FDR approximately. Here we utilize this FDR-controlling procedure, and assess its efficacy by means of simulation (see the next subsection).

As pointed out in [14], the power of the FWER- and FDR-controlling procedures described previously decreases as the number of tests $M$ increases. In practice, to have a useful MTP with reasonable power, the number of tests has to be reduced by using prior knowledge. In our case, let the true predictor set belong to a set $L$, and assume that it is related to the target via a logic function $f$ in a set $K$. The total number of tests is thus $M = |L| \times |K|$. Provided that $|L| \ll \binom{G}{d}$ and $|K| \ll 2^d$, which are the prior knowledge constraints of the problem, then the number of tests $M$ may be kept reasonably small.

From the previous considerations, we arrive at the following MTP.

Coefficient of Determination MTP.

(1) Set the significance level $\alpha$, and model sets $L$ and $K$. The total number of tests is $M = |L| \times |K|$. (2) For the given data set $S_a = s_a$, compute the unadjusted $p$-values $\{\pi_1(s_a), \ldots, \pi_M(s_a)\}$ for the tests $H_0(m) : \text{CoD} = 0$ vs. $H_1(m) : \text{CoD} > 0$, for $m = 1, \ldots, M$, using Eq. (10).

(3-a) FWER-controlling step. Compute the adjusted $p$-values $\{\pi_1''(s_a), \ldots, \pi_M''(s_a)\}$ according to Eq. (23). Reject those hypotheses $H_0(m)$ such that $\pi_m' \leq \alpha$, for $m = 1, \ldots, M$.

(3-b) FDR-controlling step. Compute the adjusted $p$-values $\{\pi_1''(s_a), \ldots, \pi_M''(s_a)\}$ according to Eq. (24).
Reject those hypotheses $H_0(m)$ such that $\pi''_m \leq \alpha$, for $m = 1, \ldots, M$.

It can be shown that the FDR-controlling step can be equivalently implemented by the following more efficient procedure [14]:

(3-b) **FDR-controlling step.** Find the list of increasing unadjusted p-values $\{\pi_{1}(s_{n}), \ldots, \pi_{M}(s_{n})\}$ and let $H^*_0(m)$ be the null hypothesis corresponding to $\pi^*_m(s_n)$, for $m = 1, \ldots, M$. Let $m^*$ be the largest $m$ such that $\pi^*_m(s_n) \leq \frac{m}{M} \alpha$. Reject all $H^*_0(m)$ for $m = 1, \ldots, m^*$. If $\pi^*_m(s_n) > \frac{m}{M} \alpha$ for all $m = 1, \ldots, M$, then reject none of the hypotheses.

C. Performance of Multiple Testing Procedures

In this section, we assess the effectiveness of the previous CoD MTP by means of simulation experiments. For the first experiment, we assume that each target $Y$ is regulated by predictors $X_1$ and $X_2$ among a set of possible predictors $X_1, \ldots, X_G$, such that $Y = X_1 \text{XOR} X_2 \oplus N$, where $N \sim \text{Bernoulli}(1 - p)$, for $1/2 \leq p \leq 1$, as before. Furthermore, we assume that the distribution of the random vector $(X_1, \ldots, X_G)$ is uniform. This specifies the stochastic model. Notice that here $L = \binom{G}{2}$. Provided that $G$ is not too large, this does not create a serious multiplicity issue; in our simulation, $G$ ranges from 4 to 24. In addition, we consider a number of targets $D$ varying from 1 to 8. As for the logic model set $K$, we consider three scenarios: (1) the prediction logic is known, $K_1 = \{\text{XOR}\}$; (2) $K_2 = \{\text{AND, XOR}\}$; and (3) $K_3 = \{\text{AND, XOR}, \bar{X}_1 X_2, \bar{X}_1 + X_2\}$. The total number of tests is given by $M_i = D \times \binom{G}{2} \times K_i$, under each of the prior-knowledge scenarios $i = 1, 2, 3$ described previously. Hence, the MTP increases in difficulty as the number of predictors and targets increase, and as less prior knowledge is available. We draw 5000 samples of varying size $n$ and form averages of FWER, FDR, and power estimates under FWER-controlling and FDR-controlling procedures.

We have selected to run the previous two simulation experiments with $n = 40$, small sample settings, due to limited availability of sample gene-expression data in practice. To investigate the appropriateness of this choice, we have re-run these simulations with $n = 20$ and $n = 60$ —results are shown in Figs. 1-4 in the supplementary material. We observed that the general conclusions from the $n = 40$ case were still valid. With a smaller sample size $n = 20$, the FDR-controlling procedure has a very clear superiority over the FWER-controlling one, as was already observed with $n = 40$. With $n = 60$, the performance of the FWER- and FDR-controlling procedures become very close due to the fact that larger sample size leads to stronger power.

The overall conclusion on the comparison between FWER- and FDR-controlling procedures is that in application with multiple targets, the FDR-controlling proce-
Fig. 4. Average FWER and FDR estimates (top row) and power estimates (bottom row) as a function of sample size under FWER- and FDR-controlling procedures, for three logic model sets, \( K_1 = \{\text{XOR}\}; \ K_2 = \{\text{AND, XOR}\}; \) and \( K_3 = \{\text{AND, XOR, } X_1X_2, X_1 + X_2\}, \) and predictive power \( p = 0.85. \) There is a single target to be predicted by two among \( G = 24 \) genes.

Fig. 5. Average FWER and FDR estimates (top row) and power estimates (bottom row) as a function of initial number of genes under FWER- and FDR-controlling procedures, for three logic model sets, \( K_1 = \{\text{XOR}\}; \ K_2 = \{\text{AND, XOR}\}; \) and \( K_3 = \{\text{AND, XOR, } X_1X_2, X_1 + X_2\}, \) and predictive power \( p = 0.85. \) There is a single target to be predicted by two among a varying number of initial genes. Sample size is fixed at \( n = 40. \)

dere is to be preferred due its superior power, whereas the FWER-controlling procedure is to be preferred in applications with very small number of targets since there is no appreciable difference in power, while the FWER and FDR rates are smaller.

V. CASE STUDY: GENOTOXIC STRESS RESPONSIVE GENES

In this section, we illustrate the application of the proposed multivariate Boolean detection methodology based on the CoD to real gene expression data, from a study on ionizing radiation (IR) responsive genes in [9].
This data set consists of 12 genes under 3 conditions (i.e., IR, MMS, UV) in 30 cell lines of both p53 proficient and p53 deficient cells. The data is ternary, indicating up-regulated (+1), down-regulated (-1), or no-change (0) status. Here we map this to binary expression using the following code: change (1), for either up-regulated or down-regulated genes, and no-change (0), as before. Additionally, we consider the three binary conditions (IR, MMS, and UV) as possible predictive factors, for a total of 15 Boolean variables in the data set.

### A. Detection of Significant Regulatory Relationships

In the first group of experiments, we use the proposed approach to find significant regulatory relationships between two predictors and a target. We assume no prior knowledge, and thus make no constraints on the allowed regulatory relationships, other than a gene does not predict itself. Hence, all \( \binom{12}{2} \) two-predictor sets and 10 possible “true” 2-predictor logic candidates are considered for each target, for a total of \( \frac{12 \times 10}{2} = 90 \) possible models; note that each gene can appear in multiple models, both as a member of different pairs and under different logic relationships. In addition, we consider each of the 12 genes in the data set as a possible target, so that the number of multiple tests performed is \( M = \binom{12}{2} \times 10 \times 12 = 10,920 \). We apply both the FWER- and the FDR-controlling procedures outlined in the previous section with a significance level \( \alpha = 0.05 \).

Figure 7 displays the gene targets possessing significant predictors and the number of significant predictive relationships (out of the maximum of 910) detected, under each of the two approaches.
Interestingly, $p53$ turns out to possess the largest number of significant predictive relationships, under both approaches. This is in accordance with the known fact that $p53$ is a significantly active gene involved in various pathways associated with stress responses. Notice that the FWER-controlling approach is more conservative than the FDR-controlling approach, for each of the targets. Table III provides examples of detected regulatory relationships that are consistent with well-known biological groundtruth. All of these relationships are detected under the FDR-controlling approach. As is known in the biological literature, $p53$ is found to be expressed when at least one of $p21$ and MDM2 is expressed, while $p21$ is found to be regulated in two ways: is is expressed when MDM2 is expressed or ATF3 is expressed, or when MDM2 is expressed or ATF3 is not expressed — the adjusted p-value for the former result is smaller than that for the latter, which may be evidence that the OR logic can provide a better model for this regulatory relationship. Table IV lists top 20 significant regulatory relationships under FDR- and FWER-controlling approaches. These results could serve as candidate regulatory relationships for further experimental verification.

Notice that the adjusted p-values in Table IV are identical for FWER- and FDR-controlling procedures, respectively. This is due to the discrete nature of the problem. For instance, considering the FWER-controlling procedure, all the 20 detections with their predicted logics share the same $k = 28$ in eq. (8) in form of the rejection region, which naturally leads to the same adjusted p-values according to eqs. (10) and (23).

### B. Detection of Synthetic Target Genes

Following [10], we further examine the properties of the proposed methodology by generating 8 synthetic target genes, SYN1, SYN2, . . . , SYN8, which are assumed to be predicted by two of 12 genes in the IR-response stress gene-expression data of [9]. Hence, each new data set consists of 23 genes (with 3 conditions included). The synthetic relationships are shown in Table V, where the noise $N \sim \text{Bernoulli}(1-p)$. A total of $M=100$ realizations are generated for the eight synthetic genes, based on the relationships in Table V. As for the logic model set $K$, we consider three cases: (1) the logic is known, $K_1 = \{\text{XOR}\}$; (2) $K_2 = \{\text{XOR}, \text{AND}, \text{NAND}\}$; and $K_3 = \{10 \text{ 2-predictor logics}\}$. We assume here that a gene cannot predict itself. Hence, with the addition of the 8 synthetic target genes, the total number of multiple tests is $M_1 = \left(\begin{array}{c}23 \cr 8 \end{array}\right) \times 1 \times 8 = 1848$ for the set $K_1$, $M_2 = \left(\begin{array}{c}23 \cr 2 \end{array}\right) \times 3 \times 8 = 5,544$ for the set $K_2$, and $M_3 = \left(\begin{array}{c}23 \cr 2 \end{array}\right) \times 10 \times 8 = 18,480$ for the set $K_3$. We apply the FWER- and FDR-controlling procedures with a significance level $\alpha = 0.05$. Figure 8 shows the power estimates as a function of the predictive power under each of the two procedures. It is observed that the FDR-controlling approach achieves larger power than the FWER-controlling one as expected. As the predictive power increases, the power increases to 1 for both
logic models, the power tends to be smaller. It is expected that this methodology will be a useful practical tool for the inference of gene regulatory relationships. Multiple-testing procedures are also described, which make the methodology applicable to large data sets. Furthermore, software that implements the COD test is made available to the scientific community as an R package through our website (http://gspt.cmu.edu/Publications-supplementary/ting13a). It is expected that this methodology will be a useful practical tool for the inference of gene regulatory relationships and networks from gene-expression data.

VI. CONCLUSION

We have described in this paper a rigorous statistical testing framework to investigate regulatory relationships among genes, by using the discrete Coefficient of Determination (CoD). This marks a significant change in the application of the CoD to such problems, since thus far its use depended on user-selected thresholds to characterize the presence of significant relationships. Multiple-testing procedures are also described, which make the methodology applicable to large data sets. Furthermore, software that implements the COD test is made available to the scientific community as an R package through our website (http://gspt.cmu.edu/Publications-supplementary/ting13a). It is expected that this methodology will be a useful practical tool for the inference of gene regulatory relationships and networks from gene-expression data.

VII. ACKNOWLEDGEMENTS

The authors acknowledge the support of the National Science Foundation, through NSF award CCF-0845407.

**TABLE V**

<table>
<thead>
<tr>
<th>Target</th>
<th>Synthetic Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SYN1 = PC1 XOR MDM2 ⊕ N</td>
</tr>
<tr>
<td>2</td>
<td>SYN2 = IAP1 XOR SSAT ⊕ N</td>
</tr>
<tr>
<td>3</td>
<td>SYN3 = PC1 XOR MMS ⊕ N</td>
</tr>
<tr>
<td>4</td>
<td>SYN4 = ATF3 XOR p53 ⊕ N</td>
</tr>
<tr>
<td>5</td>
<td>SYN5 = RCH1 XOR FRA1 ⊕ N</td>
</tr>
<tr>
<td>6</td>
<td>SYN6 = RELB XOR MMS ⊕ N</td>
</tr>
<tr>
<td>7</td>
<td>SYN7 = p53 XOR IR ⊕ N</td>
</tr>
<tr>
<td>8</td>
<td>SYN8 = BCL3 XOR IAP1 ⊕ N</td>
</tr>
</tbody>
</table>

Fig. 8. Power estimates as a function of predictive power for 8 synthetic targets using both FWER- and FDR-controlling procedures, for three logic candidate sets, $K_1 = \{\text{XOR}\}$, $K_2 = \{\text{XOR, AND, NAND}\}$ and $K_3 = \{10 \text{ meaningful logics}\}$.

References


**Ting Chen** received B.S. and M.S. degrees in electrical engineering from Fudan University, China, in 2006 and 2009. She is currently a Ph.D. student in Genomic Signal Processing Laboratory at the Department of Electrical and Computer Engineering of Texas A&M University, College Station, TX. Her current research interests include the study of discrete prediction and estimation methods and their applications in Genomics.

**Ulisses M. Braga-Neto** received the Ph.D. degree in electrical and computer engineering from The Johns Hopkins University, Baltimore, MD, in 2002. He held a post-doctoral fellowship in the section of clinical cancer genetics at the University of Texas M.D. Anderson Cancer Center, Houston, from 2002 to 2004. In 2004, he joined the Virology and Experimental Therapy Laboratory at the Oswaldo Cruz Foundation (FIOCRUZ), in Recife, Brazil. Since January 2007, he has been an assistant professor and member of the Genomic Signal Processing Laboratory at the Department of Electrical and Computer Engineering of Texas A&M University, College Station, TX. He received an NSF CAREER Award in 2008. His research interests include small-sample error estimation, statistical pattern recognition, and genomic signal processing, with applications in the study of cancer and infectious diseases.