

# Intrinsically Multivariate Predictive Genes

David C. Martins-Jr, Ulisses Braga-Neto, Ronaldo F. Hashimoto, Michael L. Bittner and  
Edward R. Dougherty

## Abstract

Canalizing genes possess such broad regulatory power, and their action sweeps across a such a wide swath of processes, that the full set of affected genes are not highly correlated under normal conditions. When not active, the controlling gene will not be predictable to any significant degree by its subject genes, either alone or in groups, since their behavior will be highly varied relative to the inactive controlling gene. When the controlling gene is active, its behavior is not well predicted by any one of its targets, but can be very well predicted by groups of genes under its control. To investigate this question, we introduce in this paper the concept of intrinsically multivariate predictive (IMP) genes for binary gene expression, and present a mathematical study of IMP with respect to the Coefficient of Determination (CoD), which measures the predictive power of a set of genes with respect to a target gene. A set of predictor genes is said to be IMP for a target gene if all properly contained subsets of the predictor set are bad predictors of the target but the full predictor set predicts the target with great accuracy. We show that logic of prediction, predictive power, covariance between predictors, and the entropy of the joint probability distribution of the predictors jointly affect the appearance of IMP genes. In particular, we show that high-predictive power, small covariance among predictors, a large entropy of the joint probability distribution of predictors, and certain logics, such as XOR in the 2-predictor case, are factors that favor the appearance of IMP. The IMP concept is applied to characterize the behavior of the gene DUSP1, which exhibits control over a central, process-integrating signaling pathway, thereby providing preliminary evidence that IMP can be used as a criterion for discovery of canalizing genes.

D. C. Martins-Jr is with the Department of Computer Science at University of São Paulo and with the Department of Electrical and Computer Engineering at Texas A&M University. Email: davidjr@vision.ime.usp.br

U. Braga-Neto and E. R. Dougherty are with the Department of Electrical and Computer Engineering at Texas A&M University

R. F. Hashimoto is with the Department of Computer Science at University of São Paulo

M. L. Bittner is with Translational Genomics Research Institute

This work was supported in part by the National Science Foundation (CCF-0514644), the National Cancer Institute (2R25CA090301), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - proc. 3696/06-5), FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo - proc. 04/03967-0), and the Translational Genomics Research Institute.

### Index Terms

Intrinsically Multivariate Prediction, Biological Networks, Canalization, Transcriptional Regulation, Microarray, Melanoma

## I. INTRODUCTION

The existence of genes that can constrain, or canalize, a biological system to particular choices was originally proposed by Conrad Waddington in 1942 [1]. The constraints he had in mind were those that caused the development of an organism to be quite reliable in its results even in the face of large applied perturbations during development. Here we are not speaking of sequential canalization, whereby a specific action of the master enforces a cascade of actions among a single highly correlated cohort of genes important in a single process, but rather where a gene has such broad regulatory power, and its action sweeps across such a wide swath of processes, that the full set of affected genes are not highly correlated under normal conditions. This type of canalization is required for any complex system to buffer the system from the effects of random alterations or operating errors.

Canalizing genes are frequently found in signaling pathways that deliver information from a variety of sources to the machinery that enacts central cellular functions such as cell-cycle, survival, apoptosis and metabolism. 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 [2], thus this gene is canalizing in its phosphorylated state, and DUSP1 is canalizing when it dephosphorylates MAPK1. As is often the case in such key pathways, there are multiple opportunities for canalizing behavior to be observed along the signal transducing pathway. Some of the earliest observations of canalization along the mitogenic pathway involved the RAS gene family, members of which were found to have frequent mutations in their 12th codon in cancers that produce uncontrolled proliferation [3]. Another locus of signal integration by a canalizing gene is in the area of stresses to the genome by the p53 (TP53) gene. The list of cellular stress responses where this gene exerts strong control continue to expand into new regulatory territories [4]. While canalizing genes can be extremely potent, their potency is nonetheless circumscribed by other features of the regulatory apparatus operating in the particular cell where control is attempted. A clear and very instructive example of this is the capability of a translocation

gene involving the BCR and ABL genes to induce a cancer phenotype in certain white blood cells by itself, a canalizing result of extraordinary scope. Recent work in mouse models strongly suggests that in order for this rearranged kinase to effectively produce a disease that strongly resembles human chronic myelogenous leukemia, the translocation gene must be delivered to a suitable hematopoietic precursor in the marrow [5].

A key characteristic of a canalizing gene is its ability to override other regulatory instructions. In terms of a model of transcriptional consequences, this kind of control would lead to a significant number of genes at various transcript abundance levels, and under the transcriptional control of other genes, to move to a single state when the broadly controlling gene is active – and thereby exercising overriding control. This can be envisaged as a reduction of the entropy of the controlled genes' states in those samples where the broadly controlling gene is active. This will result in a considerable change in the predictability of the controlling gene by those genes it controls. When not active, the controlling gene will not be predictable to any significant degree by its subject genes, either alone or in groups, since their behavior will be highly varied relative to the inactive controlling gene. When the controlling gene is active, its behavior is not well predicted by any one of its targets, but can be very well predicted by groups of genes under its control. This property of being *intrinsically multivariate predictive* (IMP) is investigated in this paper in terms of the nonlinear Coefficient of Determination (CoD). The notion of IMP genes for binary gene expression based on the CoD is introduced here for the first time. However, it has come to our attention that an information-theoretic approach has been proposed recently [6, 7], which also attempts to quantify the notion of intrinsic multivariate prediction. The present work is based instead on the CoD, which for a set of predictor genes predicting a target gene, in this case, the subject genes predicting the controlling gene, quantifies the relative increase in predictive power using the predictors over just using the statistics of the target itself [8]. The CoD was perhaps the first predictive paradigm utilized in the context of microarray data, the goal being to provide a measure of nonlinear interaction among genes [8–11]. We will demonstrate that the IMP model makes a number of very specific predictions about the distributions of controlled gene behaviors, along with the forms and accuracy of prediction logic where such control can be detected. Characterization of the behavior of a gene, DUSP1, which exhibits control over a central, process-integrating signaling pathway, shows that this gene exhibits the transcriptional

characteristics predicted by the model, thereby providing preliminary evidence that IMP can be used as a criterion for discovery of canalizing genes.

As might be expected when analyzing a complex, highly-integrated system, developing methods to identify canalizing genes via IMP and the CoD from a survey of the abundances of transcription products across a set of biological samples (set of expression profiles) is a very difficult exercise. There is considerable interest in the effectiveness of several techniques now being applied to this problem. One approach currently in use is to determine whether there is a shift in the transcript abundances of observed gene expression profiles consistent with a shift in the phenotypes of the samples examined. Under the usual assumptions of randomness and independence of the events being studied, statistical principles dictate that the certitude of the associations obtained in this type of study drop precipitously as the ratio of genes tested per sample observed increases. Because biological systems carry out their operations in highly non-random and highly integrated ways, the decay in certitude may be even greater than predicted. Recent publications in the area of genotypic associations with cardiovascular disease suggest that one should be very skeptical of approaches where many thousands of hypothesis tests are carried out on sample sizes of even a few hundred [12–14].

A possible way to deal with inference, given the complexity of biological systems, is to design analyses based on system models that predict system behavior via a simplified abstraction of the data, an approach developed by Norbert Wiener, who pioneered the field of nonlinear engineering. Wiener was responding to the analogous problem in electrical engineering, a shift from thinking about simple, linear systems to complex, nonlinear systems, and his methodology was biologically inspired. The formulation of the basic strategies he developed to solve these problems is grounded in his work with biologists studying feedback and control in biological systems. He dedicates his most notable book, *Cybernetics: Or Control and Communication in the Animal and the Machine* [15], to his longtime collaborator, Arturo Rosenblueth, a cardiologist and protégé of Walter Cannon, whose work established the importance of feedback control in the maintenance of homeostasis in organisms [16], some of the earliest work in Systems Biology.

To date, an extremely generalized model, co-transcription, has mainly been used to guide analysis of transcript abundance in biology. This model simply posits that genes whose transcription can be achieved by a particular set of factors will be jointly transcribed, and this can be detected by an analysis

of correlation. This approach remains essentially linear and can only identify genes participating in processes where the regulation is simple and linear. As regulation of transcription networks is studied in greater detail, the evidence for nonlinearity and complex regulation at the protein level by ever-finer specification of transcriptional initiation [17] and at the genomic level by the proliferation of promoter elements [18] indicates that the nonlinearity problem must be addressed if we are to develop transcription models that can be productively abstracted.

To uncover nonlinear regulation, we have developed inferential tools based on a model that characterizes the consequences of transcriptional regulation by incorporating known sources of nonlinearity in biological systems that influence genes controlled by one or more transcriptional regulators [19]. The inferences yielded by the model are based on using the behavior of the regulated genes (slaves) to identify genes (masters) that may be transcriptionally upstream of them and to estimate the possible extent of nonlinear behavior arising from common sources of nonlinearity. These sources include crosstalk, which refers to the regulatory effects of other genes when the master is not exhibiting control over the slave, and conditioning, which refers to the diminution of control by the master owing to different contextual conditions, these being determined by both intra-cellular and extra-cellular variables outside the model. In line with the scientific hypothetico-deductive method, the model constitutes a mathematical hypothesis from which observed facts are deduced [20], in this case, the distribution of the coefficients of determination for a master gene being predicted by its downstream slaves.

This paper is organized as follows. In Section II, the concept of IMP genes is defined formally. In Section III, different logics of prediction are analyzed and classified according to IMP. Section IV presents an analytical study of how predictive power impacts IMP, including the effects of correlation between predictors and different prediction logics. In Section V, preliminary evidence is given as to effectiveness of IMP in detecting canalization in real gene expression data of melanoma. Finally, Section VI presents concluding remarks.

## II. DEFINITION OF INTRINSICALLY MULTIVARIATE PREDICTION

To define the notion intrinsically multivariate predictiveness, we first define the coefficient of determination (here in the particular case of binary random variables, but this restriction is not necessary). Let

$\mathbf{X} = \{x_1, x_2, \dots, x_n\} \in \{0, 1\}^n$  be a set of predictor random variables and  $Y = y \in \{0, 1\}$  be the target random variable. The coefficient of determination (CoD) of  $\mathbf{X}$  with respect to  $Y$  [8] is given by

$$\text{CoD}_Y(\mathbf{X}) = \frac{\varepsilon_Y - \varepsilon_Y(\mathbf{X})}{\varepsilon_Y}, \quad (1)$$

where  $\varepsilon_Y$  is the error of the best predictor of  $Y$  in the absence of other observations and  $\varepsilon_Y(\mathbf{X})$  is the error of the best predictor of  $Y$  based on the observation of  $\mathbf{X}$ . By convention, we assume  $\frac{0}{0} = 1$  in the above definition. In the case of two predictors,  $\mathbf{X} = \{X_1, X_2\}$ ,  $\mathbf{X}$  is said to be intrinsically multivariate predictive (IMP) for  $Y$  with respect to  $\lambda$  and  $\delta$ , for  $0 \leq \lambda, \delta \leq 1$  and  $\lambda < \delta$ , if

$$\max_{X_1, X_2} \text{CoD}_Y(X_i) \leq \lambda \quad (2)$$

and  $\text{CoD}_Y(\mathbf{X}) \geq \delta$ . When there are more than two predictor variables, then the definition directly extends to

$$\max_{\mathbf{Z} \subseteq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) \leq \lambda \quad (3)$$

However, in the general case we can define different orders of IMP. In the case of  $n$  variables, we define  $\mathbf{X}$  to be IMP for  $Y$  of order  $k$  if

$$\max_{\{\mathbf{Z} \subseteq \mathbf{X}: |\mathbf{Z}| \leq n-k\}} \text{CoD}_Y(\mathbf{Z}) \leq \lambda \quad (4)$$

and  $\text{CoD}_Y(\mathbf{X}) \geq \delta$ . Note that order-1 IMP is just IMP. With this definition, in the case of three predictors, the  $\lambda$  requirement for order-2 IMP is  $\max_{X_1, X_2, X_3} \text{CoD}_Y(X_i) \leq \lambda$ .

For a fixed predictor-target pair  $(\mathbf{X}, Y)$ , the largest  $\delta$  for which the prediction is IMP is  $\delta = \text{CoD}_Y(\mathbf{X})$ . Hence, if we define the intrinsically multivariate predictiveness (IMP score) of the pair  $(\mathbf{X}, Y)$  by the maximum value of  $\delta - \lambda$ , then the IMP score is given by

$$\text{IMP}_Y(\mathbf{X}) = \text{CoD}_Y(\mathbf{X}) - \max_{\mathbf{Z} \subseteq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) \quad (5)$$

The definition is directly extended to different orders of IMP score.

In the next three sections we will consider three crucial factors for  $(\mathbf{X}, Y)$  to be classified as IMP: logic of prediction, predictive power, and the probability distribution of the predictor variables. Since our application of interest is gene regulation, henceforth we will refer to all random variables as genes.

### III. LOGIC OF PREDICTION

To focus on the logic of prediction, we let logic be the unique factor relative to IMP score. We assume that the predictive power of  $\mathbf{X}$  with respect to  $Y$  is the highest possible, i.e,  $P(Y = 1|\mathbf{x}) = 0$  or  $P(Y = 1|\mathbf{x}) = 1$ , for all  $\mathbf{x} \in \mathbf{X}$ ; equivalently,  $\varepsilon_Y(\mathbf{X}) = 0$  (deterministic prediction). We also assume that all instances of  $\mathbf{X}$  are equally probable, that is,  $P(\mathbf{X} = \mathbf{x}) = 1/2^n$  for all  $\{x_1, x_2, \dots, x_n\} \in \mathbf{X}$ . In this case, when we refer to a particular logic of prediction for the CoD, by definition this logic provides the best prediction among all possible predictors. As  $\varepsilon_Y(\mathbf{X}) = 0$ , we have that  $\text{CoD}_Y(\mathbf{X}) = 1$ . Under these conditions, we define three classes of logics with respect to IMP score:

- 1) **IMP logic:** a logic of prediction is IMP if  $\max_{\mathbf{Z} \subsetneq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) = 0$ ;
- 2) **anti-IMP logic:** a logic of prediction is anti-IMP if  $\max_{\mathbf{Z} \subsetneq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) = 1$ ;
- 3) **partly-IMP logic:** a logic of prediction is partly-IMP if  $0 < \max_{\mathbf{Z} \subsetneq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) < 1$ ;

For two binary predictors, the number of possible logics is  $2^4 = 16$  because there are 4 possible instances (predictor pairs),  $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$ , and for each instance, the target  $Y$  assumes a binary value. Each logic can be addressed by a string of four binary values in which the first character corresponds to the  $Y$  value when  $(0, 0)$  is observed, the second when  $(0, 1)$  is observed, the third when  $(1, 0)$  is observed, and the last when  $(1, 1)$  is observed. For example, the AND logic is expressed by the string 0001. For the constant logics 0000 and 1111, the target is insensitive to variability of the predictors. This corresponds to a degenerate case, where the error  $\varepsilon_Y$  in the absence of observations is zero, and so is the error  $\varepsilon_Y(\mathbf{Z})$  of any predictor set  $\mathbf{Z} \subsetneq \mathbf{X}$ . By definition, the CoDs in this degenerate case are all equal to 1, so that constant logics are anti-IMP.

From the 14 remaining logics, 4 present exactly one true value in their representative strings, or equivalently, one minterm in their sum-of-products expansion — we call these 1-minterm logics:  $Y = X_1 \wedge X_2$  (0001),  $Y = X_1 \wedge \bar{X}_2$  (0010),  $Y = \bar{X}_1 \wedge X_2$  (0100), and  $Y = \bar{X}_1 \wedge \bar{X}_2$  (1000). All four are

IMP logics. For instance, consider the AND logic,  $Y = X_1 \wedge X_2$  (0001). The optimal (Bayes) predictor function  $\psi_Y(X_1)$  of  $Y$  given  $X_1$  is given by

$$\psi_Y(X_1) = \begin{cases} 1, & \text{if } P(X_1 \wedge X_2 = 1|X_1) > 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Therefore,  $\psi_Y(0) = \psi_Y(1) = 0$ , since  $P(X_1 \wedge X_2 = 1|X_1 = 0) = 0$  and  $P(X_1 \wedge X_2 = 1|X_1 = 1) = 0.5$ . It follows that  $\varepsilon_Y(X_1) = P(\psi_Y(X_1) \neq Y) = P(X_1 = 1)P(X_1 \wedge X_2 = 1|X_1 = 1) = 0.25$ . We also have  $\varepsilon_Y = P(X_1 \wedge X_2 = 1) = 0.25$ . Therefore, the CoD for prediction of  $Y$  by  $X_1$  is given by  $\text{CoD}_Y(X_1) = (0.25 - 0.25)/0.25 = 0$ . Similarly, the CoD for prediction of  $Y$  by  $X_2$  is given by  $\text{CoD}_Y(X_2) = 0$ . The other 1-minterm logics are symmetric to the AND logic. Also, the 3-minterm logics,  $Y = X_1 \vee X_2$  (0111),  $Y = X_1 \vee \bar{X}_2$  (1011),  $Y = \bar{X}_1 \vee X_2$  (1101), and  $Y = \bar{X}_1 \vee \bar{X}_2$  (1110)), are symmetric to the 1-minterm logics, the difference being that  $P(Y = 1) = 0.75$  instead of  $P(Y = 1) = 0.25$ . Thus, all the 1-minterm and 3-minterm logics are IMP logics.

It remains to analyze the 2-minterm logics. There are 6 (4 choose 2) of those:  $Y = X_1$  (0011),  $Y = X_2$  (0101),  $Y = \bar{X}_1$  (1100),  $Y = \bar{X}_2$  (1010),  $Y = X_1 \oplus X_2$  (0110, XOR – exclusive OR),  $Y = X_1 \bar{\oplus} X_2$  (1001, negated XOR, also called EQV – equivalence). In the first four 2-minterm logics,  $Y$  is fully determined by one of the predictors, resulting in  $\text{CoD}_Y(X_1) = 1$  or  $\text{CoD}_Y(X_2) = 1$ , so that these four logics are anti-IMP. The last two logics, XOR and negated XOR, are IMP because in both cases,  $\text{CoD}_Y(X_1) = \text{CoD}_Y(X_2) = 0$ .

Thus, with two binary predictors, there are only two classes of logics: the anti-IMPs (constant logics and 2-minterm logics except XOR and negated XOR) and IMPs (1-minterm, 3-minterm, XOR and negated XOR logics).

The partly-IMP logics appear for the cases with three or more predictors. In fact, for the case with three predictors, from the 256 ( $2^8$ ) possible logics, 66 are IMP, 152 are partly-IMP, and the remaining 38 logics are anti-IMP. The IMP class is composed of all the 8 1-minterm logics (8 7-minterm logics by symmetry), 16 2-minterm logics (16 6-minterm logics), 8 3-minterm logics (8 5-minterm logics) and 2 4-minterm logics. The partly-IMP class contains 48 3-minterm logics (48 5-minterm logics by symmetry, all with  $\max_{\mathbf{Z} \subseteq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) = 0.67$ , and 56 4-minterm logics, all with  $\max_{\mathbf{Z} \subseteq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) = 0.5$ . Finally,

the anti-IMP class is composed of the two constant logics (00000000, 11111111), 12 2-minterm logics (12 6-minterm logics, by symmetry) and 12 4-minterm logics. More details about these logics can be seen in the supplementary material ([http://www.vision.ime.usp.br/~davidjr/imp\\_report/logics.html](http://www.vision.ime.usp.br/~davidjr/imp_report/logics.html)).

If one considers logic of prediction in a deterministic setting, the value of  $\text{CoD}_Y(\mathbf{X})$  does not matter (as  $\text{CoD}_Y(\mathbf{X}) = 1$  in all cases), and the logic of prediction is the only factor that produces IMP genes. The situation is different in the next section, which considers predictive power for stochastic logics.

#### IV. PREDICTIVE POWER

In this section we consider the case where prediction is not deterministic. The logics will correspond to stochastic truth tables, where all output values are of the form  $p$  or  $1 - p$ , for an indeterminacy factor  $0 \leq p \leq 1$ . The cases  $p = 0$  and  $p = 1$  produce standard deterministic logics, so the present framework is a stochastic generalization of the classic Boolean case. This model is reminiscent of, but distinct from, the noisy-OR model commonly used in Bayesian networks inference [21, 22]. As we will see, assuming  $p \geq 0.5$ , we have  $\varepsilon_Y(\mathbf{X}) = 1 - p$ , so that  $p$  measures the predictive power, i.e., the accuracy of the set of predictors in predicting the target behavior. We will analytically express the impact of  $p$  over the logics capable of producing IMPs in the two-predictor case. As seen in Section III, these are the 1-minterm logics, the 3-minterm logics, XOR, and negated XOR (we consider the case of three predictors in Section IV-B1). To analyze the impact of predictive power for two predictors, henceforth, without loss of generality, we will consider the AND and XOR logics, since other logics can be reduced to these two as far as IMP is concerned.

Throughout this section, for the sake of simplicity, we will use the following notation:  $\varepsilon_0 = \varepsilon_Y$ ,  $\varepsilon_1 = \varepsilon_Y(X_1)$ ,  $\varepsilon_2 = \varepsilon_Y(X_2)$ , and  $\varepsilon_{12} = \varepsilon_Y(X_1, X_2)$ . Similarly,  $\text{CoD}_1 = \text{CoD}_Y(X_1)$ ,  $\text{CoD}_2 = \text{CoD}_Y(X_2)$ , and  $\text{CoD}_{12} = \text{CoD}_Y(X_1, X_2)$ .

The probability model for  $X_1, X_2, Y \in \{0, 1\}$  can be described in terms of 8 parameters:

$$\begin{aligned} P(Y = 1 | X_1 = i, X_2 = j) &= \pi_{ij}, \quad i, j = 0, 1 \\ P(X_1 = i, X_2 = j) &= \alpha_{ij}, \quad i, j = 0, 1 \end{aligned}$$

The parameters  $\alpha_{ij}$  specify the joint probability distribution of  $X_1, X_2$  (only 3 of these parameters being

independent), while the parameters  $\pi_{ij}$  describe the stochastic truth table of the prediction logic. If all  $\pi_{ij}$  are either 0 or 1, then we are back to the case of a deterministic logic (Section III).

The error of the optimal predictor of  $Y$  in the absence of predictors is given by

$$\varepsilon_0 = \min\{P(Y = 1), P(Y = 0)\} = F[P(Y = 1)] = F\left(\sum_{ij} \alpha_{ij} \pi_{ij}\right)$$

where  $F : [0, 1] \rightarrow [0, 1]$  is given by  $F(x) = \min\{x, 1 - x\}$ . The function  $F$  introduces a nonlinearity in the dependence of the CoD on the model parameters. It has the following important symmetry property:  $F(p) = F(1 - p)$ , for all  $0 \leq p \leq 1$  (see Fig. 1).

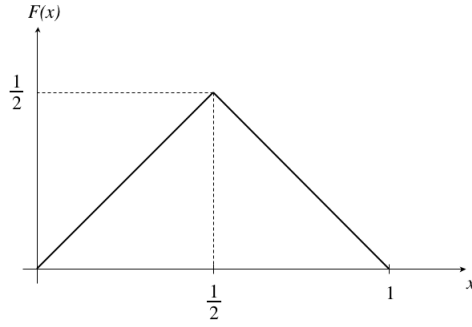


Fig. 1. Graph of the nonlinearity  $F(x)$ .

The optimal error for  $X_1$  predicting  $Y$  is

$$\varepsilon_1 = \sum_{i=0}^1 F[P(Y = 1|X_1 = i)]P(X_1 = i) = \sum_{i=0}^1 F\left(\frac{\pi_{i0}\alpha_{i0} + \pi_{i1}\alpha_{i1}}{\alpha_{i0} + \alpha_{i1}}\right)(\alpha_{i0} + \alpha_{i1}),$$

and, similarly, the optimal error for  $X_2$  predicting  $Y$  is

$$\varepsilon_2 = \sum_{j=0}^1 F[P(Y = 1|X_2 = j)]P(X_2 = j) = \sum_{j=0}^1 F\left(\frac{\pi_{0j}\alpha_{0j} + \pi_{1j}\alpha_{1j}}{\alpha_{0j} + \alpha_{1j}}\right)(\alpha_{0j} + \alpha_{1j}).$$

The optimal error based on  $X_1$  and  $X_2$  predicting  $Y$  is

$$\varepsilon_{12} = E\{F[P(Y = 1|X_1, X_2)]\} = \sum_{ij} F(\pi_{ij})\alpha_{ij} \quad (7)$$

We will consider truth tables parameterized by a variable  $p$ , with  $0 \leq p \leq 1$ , such that all entries are

$\pi_{ij} = p$  or  $\pi_{ij} = 1 - p$ . In this case, applying the symmetry property  $F(p) = F(1 - p)$ , it follows easily from (7) that  $\varepsilon_{12} = F(1 - p)$ . Assuming  $p \geq 0.5$ , we find that the optimal prediction error based on the pair  $(X_1, X_2)$  is just  $\varepsilon_{12} = 1 - p$ , that is,  $p$  is the *predictive power* of  $(X_1, X_2)$ .

The CoDs based on  $X_1$ ,  $X_2$ , and  $(X_1, X_2)$  are, respectively,

$$\begin{aligned} \text{CoD}_1 &= 1 - \frac{\varepsilon_1}{\varepsilon_0} = 1 - \frac{1}{F\left(\sum_{ij} \alpha_{ij} \pi_{ij}\right)} \sum_{i=0}^1 F\left(\frac{\pi_{i0} \alpha_{i0} + \pi_{i1} \alpha_{i1}}{\alpha_{i0} + \alpha_{i1}}\right) (\alpha_{i0} + \alpha_{i1}) \\ \text{CoD}_2 &= 1 - \frac{\varepsilon_2}{\varepsilon_0} = 1 - \frac{1}{F\left(\sum_{ij} \alpha_{ij} \pi_{ij}\right)} \sum_{j=0}^1 F\left(\frac{\pi_{0j} \alpha_{0j} + \pi_{1j} \alpha_{1j}}{\alpha_{0j} + \alpha_{1j}}\right) (\alpha_{0j} + \alpha_{1j}) \\ \text{CoD}_{12} &= 1 - \frac{\varepsilon_{12}}{\varepsilon_0} = 1 - \frac{1}{F\left(\sum_{ij} \alpha_{ij} \pi_{ij}\right)} \sum_{ij} F(\pi_{ij}) \alpha_{ij} = 1 - \frac{F(1 - p)}{F\left(\sum_{ij} \alpha_{ij} \pi_{ij}\right)} \end{aligned} \quad (8)$$

#### A. Uniform Predictors

Assuming that the distribution of  $(X_1, X_2)$  is uniform, i.e.,  $\alpha_{00} = \alpha_{01} = \alpha_{10} = \alpha_{11} = \frac{1}{4}$ , the preceding formulas reduce to

$$\begin{aligned} \text{CoD}_1 &= 1 - \frac{1}{2} \frac{F\left(\frac{\pi_{00} + \pi_{01}}{2}\right) + F\left(\frac{\pi_{10} + \pi_{11}}{2}\right)}{F\left(\frac{\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11}}{4}\right)} \\ \text{CoD}_2 &= 1 - \frac{1}{2} \frac{F\left(\frac{\pi_{00} + \pi_{10}}{2}\right) + F\left(\frac{\pi_{01} + \pi_{11}}{2}\right)}{F\left(\frac{\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11}}{4}\right)} \\ \text{CoD}_{12} &= 1 - \frac{1}{4} \frac{F(\pi_{00}) + F(\pi_{01}) + F(\pi_{10}) + F(\pi_{11})}{F\left(\frac{\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11}}{4}\right)} \end{aligned} \quad (9)$$

We will consider in the following two subsections the AND logic, which is representative of all 1- and 3-minterm logics, and the XOR logic, which is representative of the 2-minterm logics. The symmetry of 1- and 3-minterm logics to the AND logic and of the 2-minterm logics to the XOR logic was argued in Section III — see also the remarks in Section IV-B, in connection with Figure 6.

1) *AND Logic*: In the case of AND logic,  $\pi_{00} = \pi_{01} = \pi_{10} = 1 - p$  and  $\pi_{11} = p$ , with  $0.5 \leq p \leq 1$ . Substituting this into equation (9), and using the assumption  $p \geq 0.5$ , yields

$$\text{CoD}_1 = \text{CoD}_2 = 1 - \frac{1}{2} \frac{F(1 - p) + F\left(\frac{1}{2}\right)}{F\left(\frac{3 - 2p}{4}\right)} = 1 - \frac{1}{2} \frac{1 - p + \frac{1}{2}}{\frac{3 - 2p}{4}} = 0,$$

whereas the overall CoD is given by

$$\text{CoD}_{12} = 1 - \frac{1}{4} \frac{3F(1-p) + F(p)}{F\left(\frac{3-2p}{4}\right)} = 1 - \frac{4(1-p)}{3-2p} = \frac{p - \frac{1}{2}}{1 - \left(p - \frac{1}{2}\right)}. \quad (10)$$

The overall CoD is thus nonlinear as a function of the predictive power  $p$ , with  $\text{CoD}_{12} \rightarrow 1$  as  $p \rightarrow 1$ , and  $\text{CoD}_{12} \rightarrow 0$  for  $p \rightarrow \frac{1}{2}$  (see Fig. 2a). Since  $\text{CoD}_1 = \text{CoD}_2 = 0$ ,  $\mathbf{X}$  is IMP for  $Y$  with respect to  $\lambda = 0$  and  $\delta$  if and only if

$$\text{CoD}_{12} = \frac{p - \frac{1}{2}}{1 - \left(p - \frac{1}{2}\right)} \geq \delta \Rightarrow p \geq \frac{1}{2} \frac{1 + 3\delta}{1 + \delta}.$$

For example, with  $\delta = 0.8$ , the IMP condition is  $p \geq \frac{17}{18} = 0.944\dots$

2) *XOR logic*: In the case of XOR logic,  $\pi_{00} = \pi_{11} = 1 - p$  and  $\pi_{01} = \pi_{10} = p$ , with  $0.5 \leq p \leq 1$ . Substituting this into equation (9), and using the assumption  $p \geq 0.5$ , yields

$$\text{CoD}_1 = \text{CoD}_2 = 1 - \frac{1}{2} \frac{F\left(\frac{1}{2}\right) + F\left(\frac{1}{2}\right)}{F\left(\frac{1}{2}\right)} = 1 - \frac{1}{2} \frac{\frac{1}{2} + \frac{1}{2}}{\frac{1}{2}} = 0,$$

whereas the overall CoD is given by

$$\text{CoD}_{12} = 1 - \frac{1}{4} \frac{2F(p) + 2F(1-p)}{F\left(\frac{1}{2}\right)} = 1 - \frac{1}{4} \frac{4(1-p)}{\frac{1}{2}} = 2 \left(p - \frac{1}{2}\right). \quad (11)$$

The overall CoD in this case is a simple linear function of the predictive power  $p$ . Similarly to the AND case,  $\text{CoD}_{12} \rightarrow 1$  as  $p \rightarrow 1$ , and  $\text{CoD}_{12} \rightarrow 0$  for  $p \rightarrow \frac{1}{2}$  (see Fig. 2b). Again, the individual CoDs are null, so  $\mathbf{X}$  is IMP for  $Y$  with respect to  $\lambda = 0$  and  $\delta$  if and only if

$$\text{CoD}_{12} = 2 \left(p - \frac{1}{2}\right) \geq \delta \Rightarrow p \geq \frac{1}{2} + \frac{\delta}{2}. \quad (12)$$

For example, with  $\delta = 0.8$ , the IMP condition is  $p \geq 0.9$ . This is a less stringent condition than in the AND case with the same  $\delta$ . The plots of the critical value of  $p$  as a function of  $\delta$ , displayed in Fig. 3, show that the IMP condition is always more stringent for the AND logic than for the XOR logic.

### B. Non-Uniform Predictors

In this section we consider the situation in which the joint predictor distribution is not uniform and will see that the CoD can be strongly affected by this distribution. The only cases in which  $\text{CoD}_{12}$  is

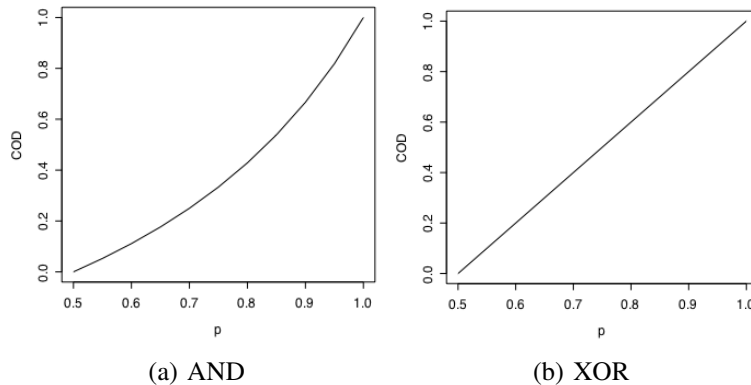


Fig. 2. Overall CoD for AND and XOR logics as a function of predictive power  $p$ , with uniform predictors.

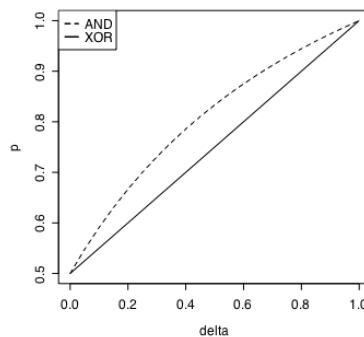


Fig. 3. Critical value of  $p$  for IMP as a function of  $\delta$  for the AND and XOR logics with uniform predictors, showing that the IMP condition is always more stringent for the AND logic than for the XOR logic.

not affected by the predictor distribution are when  $p = 1$  (maximum predictive power) and  $p = 0.5$  (no predictive power), with  $\text{CoD}_{12} = 1$  and  $\text{CoD}_{12} = 0$ , respectively, independently of the logic and the predictor distribution. For  $0.5 < p < 1$ , we will examine the AND and XOR logics to see what happens with  $\text{CoD}_1$ ,  $\text{CoD}_2$ , and  $\text{CoD}_{12}$ . First, we discuss the special case in which the predictors are independent for two- and three-predictor sets. Next, the general case in which predictor independence is not assumed will be discussed in depth only for two-predictor sets owing to the excess of parameters with three predictors.

1) *Two Independent Predictors:* In the case of two predictors, let  $P_1 = P(X_1 = 1)$  and  $P_2 = P(X_2 = 1)$ . Given predictive power  $p$  for  $(X_1, X_2)$  with regard to  $Y$ , the joint probability distribution (JPD) tables for AND and XOR logics are shown in Tables I and II respectively.

Note that  $\alpha_{00} = (1 - P_1)(1 - P_2)$ ,  $\alpha_{01} = (1 - P_1)P_2$ ,  $\alpha_{10} = P_1(1 - P_2)$ , and  $\alpha_{11} = P_1P_2$ . For

TABLE I  
JPD TABLE FOR THE AND LOGIC ASSUMING INDEPENDENCE BETWEEN  $X_1$  AND  $X_2$ .

$X_1$	$X_2$	$P(Y = 0, X_1 = i, X_2 = j)$	$P(Y = 1, X_1 = i, X_2 = j)$
0	0	$(1 - P_1)(1 - P_2)p$	$(1 - P_1)(1 - P_2)(1 - p)$
0	1	$(1 - P_1)P_2p$	$(1 - P_1)P_2(1 - p)$
1	0	$P_1(1 - P_2)p$	$P_1(1 - P_2)(1 - p)$
1	1	$P_1P_2(1 - p)$	$P_1P_2p$

TABLE II  
JPD TABLE FOR THE XOR LOGIC ASSUMING INDEPENDENCE BETWEEN  $X_1$  AND  $X_2$ .

$X_1$	$X_2$	$P(Y = 0, X_1 = i, X_2 = j)$	$P(Y = 1, X_1 = i, X_2 = j)$
0	0	$(1 - P_1)(1 - P_2)p$	$(1 - P_1)(1 - P_2)(1 - p)$
0	1	$(1 - P_1)P_2(1 - p)$	$(1 - P_1)P_2p$
1	0	$P_1(1 - P_2)(1 - p)$	$P_1(1 - P_2)p$
1	1	$P_1P_2p$	$P_1P_2(1 - p)$

the AND logic,  $\pi_{00} = \pi_{01} = \pi_{10} = 1 - p$  and  $\pi_{11} = p$ , with  $0.5 \leq p \leq 1$ . Replacing these values in Equation 8, and using the assumption  $p \geq 0.5$ , the AND logic CoDs are given by:

$$\begin{aligned}
 \text{CoD}_1 &= 1 - \frac{(1 - P_1)(1 - p) + F[P_2(1 - 2p) + p]P_1}{F[P_1P_2(1 - 2p) + p]} \\
 \text{CoD}_2 &= 1 - \frac{(1 - P_2)(1 - p) + F[P_1(1 - 2p) + p]P_2}{F[P_1P_2(1 - 2p) + p]} \\
 \text{CoD}_{12} &= 1 - \frac{1 - p}{F[P_1P_2(1 - 2p) + p]}
 \end{aligned} \tag{13}$$

For the XOR logic,  $\pi_{00} = \pi_{11} = 1 - p$  and  $\pi_{01} = \pi_{10} = p$ , with  $0.5 \leq p \leq 1$ . Thus, its CoDs are given by:

$$\begin{aligned}
\text{CoD}_1 &= 1 - \frac{F[P_2(1 - 2p) + p]}{F[P_1(1 - P_2) + P_2(1 - P_1) + (2P_1 - 1)(2P_2 - 1)p]} \\
\text{CoD}_2 &= 1 - \frac{F[P_1(1 - 2p) + p]}{F[P_1(1 - P_2) + P_2(1 - P_1) + (2P_1 - 1)(2P_2 - 1)p]} \\
\text{CoD}_{12} &= 1 - \frac{1 - p}{F[P_1(1 - P_2) + P_2(1 - P_1) + (2P_1 - 1)(2P_2 - 1)p]}
\end{aligned} \tag{14}$$

Note that, for  $p = \frac{1}{2}$ , the IMP score is zero for both the AND and XOR logics, regardless of the values of  $P_1$  and  $P_2$  (this corresponds to the case of no predictive power).

With these equations in mind, given  $p$  and the logic, we can plot graphs for  $\max\{\text{CoD}_1, \text{CoD}_2\}$ ,  $\text{CoD}_{12}$ , and the IMP score  $I = I_Y(\mathbf{X})$  with the values  $0 < P_1 < 1$  on the x-axis, the values  $0 < P_2 < 1$  on the y-axis, and associating with each point  $(P_1, P_2)$  an intensity (height) given by  $\max\{\text{CoD}_1, \text{CoD}_2\}$ ,  $\text{CoD}_{12}$ , or  $I$ . Figures 4 and 5 show these graphs for AND and XOR logics, respectively, with three values of  $p$ . Darker intensities correspond to smaller values. For AND logic, the most favorable regions are defined by  $P_1 \leq 0.5$  and  $P_2 \leq 0.5$  (left-bottom square), but lower predictive power  $p$  dramatically shrinks this region to  $P_1 \approx 0.5$  and  $P_2 \approx 0.5$ . For XOR logic, the most favorable regions are defined by  $P_1 \approx P_2$  and  $P_1 \approx 1 - P_2$  (diagonals) except for  $P_1 \approx 0$  and  $P_1 \approx 1$ .

The 1-minterm and 3-minterm logics are symmetric by rotation to the AND logic, as can be seen in Fig. 6, which shows four symmetrical IMP score surfaces and the logics associated to them (in binary string notation), for  $p = 0.99$ . For the XOR and NXOR logics, IMP score surfaces are exactly equal to one another for all values of  $p$ .

We make three observations based on the surfaces illustrated in Figs. 4, 5 and 6. First, for all situations, the central regions ( $P_1 \approx 0.5$  and  $P_2 \approx 0.5$ ) are the regions with highest IMP score intensity, indicating that uniformity in the predictor distributions facilitates IMP score. Second, the surfaces illustrate that the predictive power  $p$  exercises a dramatic effect on the  $\text{CoD}_{12}$  intensity. Even for  $p = 0.9$ , which is good prediction power, IMP score declines substantially. A last important observation is that the highest  $\text{CoD}_{12}$  values obtained for the XOR logic are higher than those obtained for the AND logic for the same  $p$  value. Therefore, if the predictive power is not so strong ( $p < 0.95$  or so), then XOR (or NXOR)

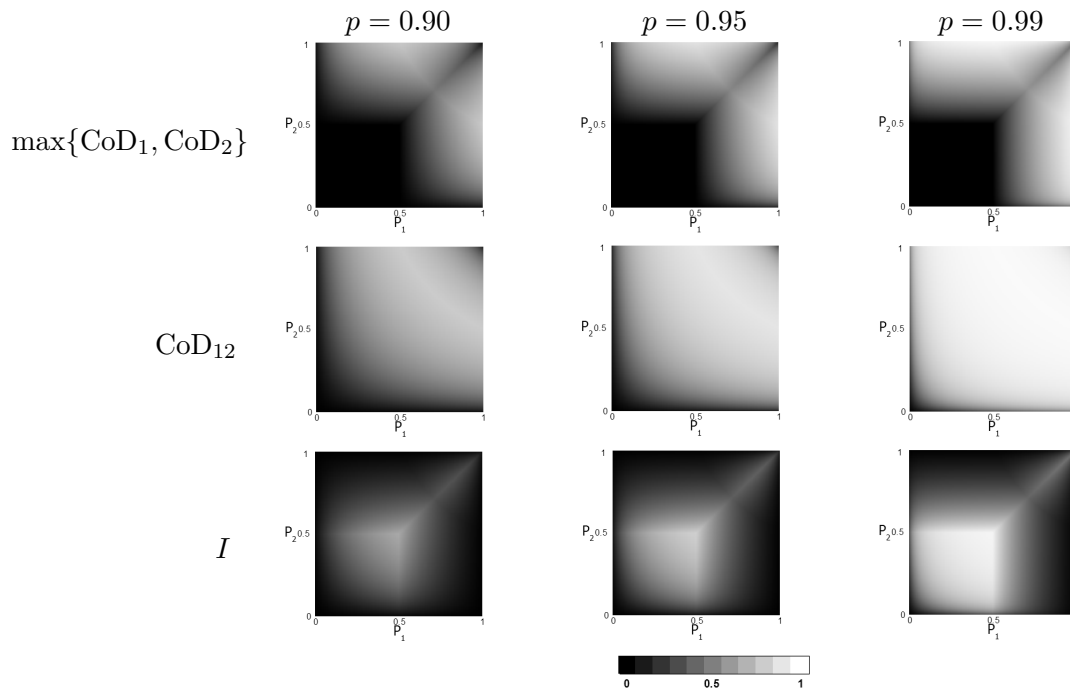


Fig. 4. Graphics representing  $\max\{\text{CoD}_1, \text{CoD}_2\}$  (first row),  $\text{CoD}_{12}$  (second row) and IMP score  $I$  values for the AND logic and three distinct values of  $p$  (0.90, 0.95 and 0.99) with the assumption that  $X_1$  and  $X_2$  are independent. The probabilities  $P_1$  and  $P_2$  are given by X-axis and Y-axis respectively. Darker color correspond to smaller values. Thus, the regions of interest to generate IMPs are composed by the darkest values in the first row and the brightest values in the second and third rows.

logic is required to obtain a substantial IMP score. On the other hand, if there is very strong predictive power ( $p \approx 1$ ), then IMP can be attained more easily with 1-minterm or 3-minterm logics because a considerable part of these surface regions possess high intensity.

2) *Three Independent Predictors*: Assuming independence between predictors, the same analysis done for sets with two predictors can be done for the three predictors except that now the analysis cannot be performed through visualization of surfaces, but only through statistics over the parameter space  $(P(X_1 = 1), P(X_2 = 1), P(X_3 = 1)) = (P_1, P_2, P_3)$ . As discussed in Section III, for three predictors sets there are  $2^8 = 256$  possible logics with 66 IMPs, 152 partly-IMP, and 38 anti-IMP, the latter not being of interest. Fortunately, the IMP and partly-IMP sets can be represented by few logics, with the remaining logics being obtained by symmetry. A key point regarding independence between two predictors is that uniformity in the predictor distribution offers the best conditions to generate IMPs. Thus, it is important to see if this uniformity also contributes to IMPs with three predictors. The Shannon entropy

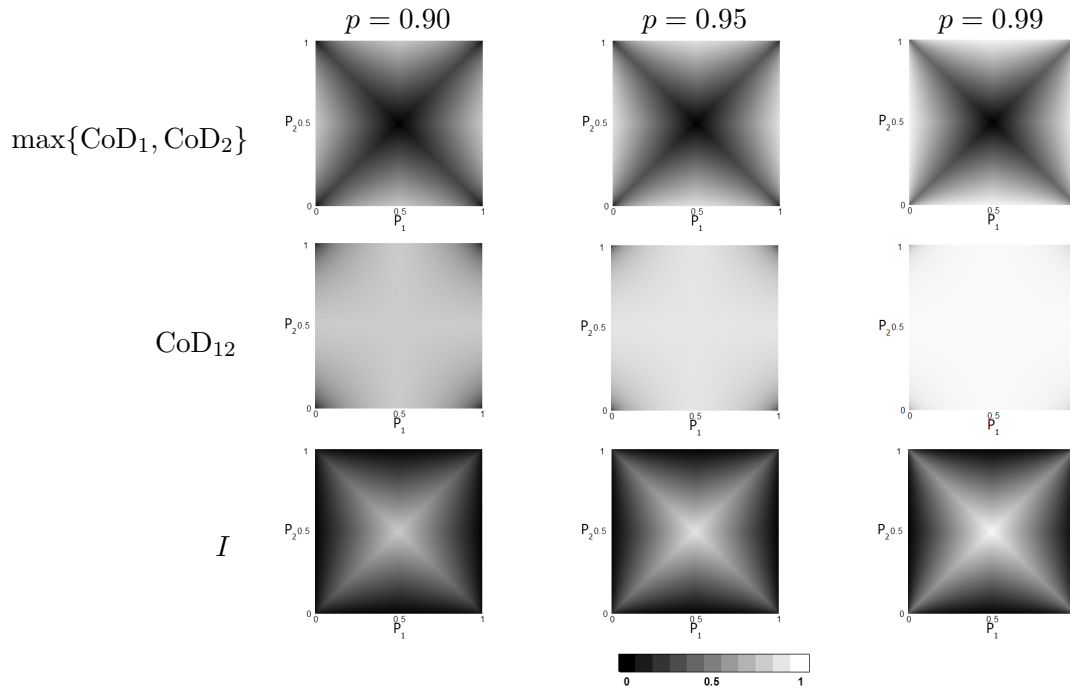


Fig. 5. Graphics representing  $\max\{\text{CoD}_1, \text{CoD}_2\}$  (first row),  $\text{CoD}_{12}$  (second row) and IMP score  $I$  values for the XOR logic and three distinct values of  $p$  (0.90, 0.95 and 0.99) with the assumption that  $X_1$  and  $X_2$  are independent. The probabilities  $P_1$  and  $P_2$  are given by X-axis and Y-axis respectively. Darker color correspond to smaller values. Thus, the regions of interest to generate IMPs are composed by the darkest values in the first row and the brightest values in the second and third rows.

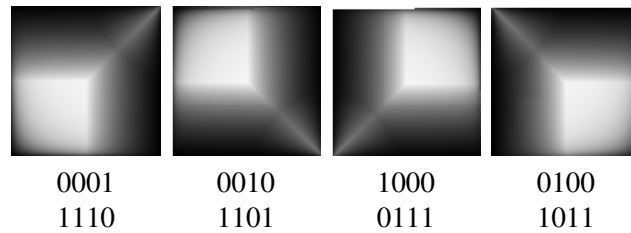


Fig. 6. 1,3-minterm logics and their respective symmetrical IMP score graphics for  $p = 0.99$  referenced by binary string notation.

[23] is adopted here to measure uniformity. It is defined by

$$H(\mathbf{X}) = - \sum_{\mathbf{x} \in \mathbf{X}} P(\mathbf{x}) \log P(\mathbf{x}), \quad (15)$$

where  $\mathbf{X}$  is the set of predictors and  $0 \cdot \log 0 = 0$ .

For a given predictive power  $p$ , we consider the correlation between entropy and IMP score. Consider the parameter triple  $(P_1, P_2, P_3)$  in the range  $(0.01 \leq P_1, P_2, P_3 \leq 0.99)$  with granularity 0.01, resulting a

total of 970,299 points in this parameter space. Each point has an entropy  $H$  for the predictor distribution and an IMP score  $I$  associated with it. We can generate the corresponding scatter plot and compute the correlation for each logic and  $p$ .

Beginning with the IMP logics, there are five subsets of logics with the same results in terms of correlation between  $H$  and  $I$ , one for 1-minterm and 7-minterm logics, two for 2-minterm and 6-minterm logics, one for 3-minterm and 5-minterm logics and one for 4-minterm logics. Figure 7 shows the scatterplot of entropy versus IMP score for  $p = 0.9, 0.95, 0.99$  and for the logic 00000001 representing 1-minterm and 7-minterm logics. The results in numeric table format corresponding to this subset and to the other four subsets of IMP logics can be viewed in the supplementary material ([http://www.vision.ime.usp.br/~davidjr/imp\\_report/imp\\_entropy\\_report.html](http://www.vision.ime.usp.br/~davidjr/imp_report/imp_entropy_report.html)). From these results we see that predictor entropy and IMP score are well correlated. Generally, the smaller  $p$  is, the greater the correlation is. If a set of predictors has a predictive power  $p$  which is not sufficiently high ( $p < 0.95$ ), then the entropy of the predictor distribution must be almost maximum in order to produce IMP. A last observation is that smaller  $p$  leads to a depopulation of the highest IMP score ( $0.8 \leq I \leq 1$ ) as  $p$  directly affects the joint CoD.

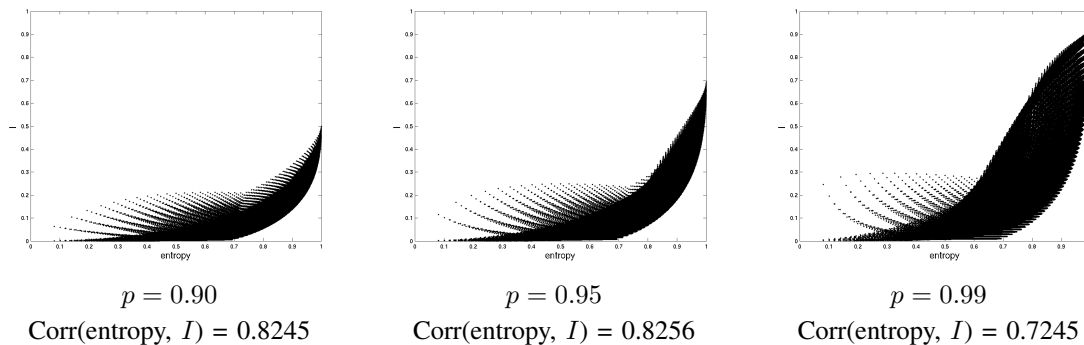


Fig. 7. Scatterplot of IMP score ( $I$ ) versus entropy with 970299 points of the parameter space ( $P_1, P_2, P_3$ ) for  $p = 0.90$ ,  $p = 0.95$  and  $p = 0.99$  and logic 00000001, (IMP logic). The respective correlations are below each plot.

The partly-IMP logics also have 5 subsets of logics in terms of the correlation between  $H$  and  $I$ , two for 3-minterm and 5-minterm logics and three for 4-minterm logics, and the same procedure is applied to these. Figure 8 shows the scatter plots for the logic 00011001. The results in numeric table format corresponding to its subset and to the other four subsets of partly IMP logics are in the supplementary material ([http://www.vision.ime.usp.br/~davidjr/imp\\_report/partly\\_imp\\_entropy\\_report.html](http://www.vision.ime.usp.br/~davidjr/imp_report/partly_imp_entropy_report.html)).

These tables show that the correlation of entropy with IMP score for partly-IMP logics is positive, but significantly smaller than those obtained from IMP logic results. Another difference is that the height of the scatter plot is smaller than  $I = 0.7$  even for  $p \approx 1$ , such as  $p = 0.99$ . In fact, this class does not present a well defined pattern for the height of the scatter plot since there are logics in which the height is always lower than  $I = 0.6$ , logics having some points higher than  $I = 0.6$  for  $p = 0.9$  but none for  $p = 0.99$ , and other logics having some points higher than  $I = 0.6$  for  $p \approx 1$ . The salient point is that the number of points higher than  $I = 0.6$  is dramatically reduced for all partly-IMP logics compared to IMP logics. The unique similarity of this class of logics with the IMP logics is the fact that smaller  $p$  generally increases the correlation between entropy and  $I$ .

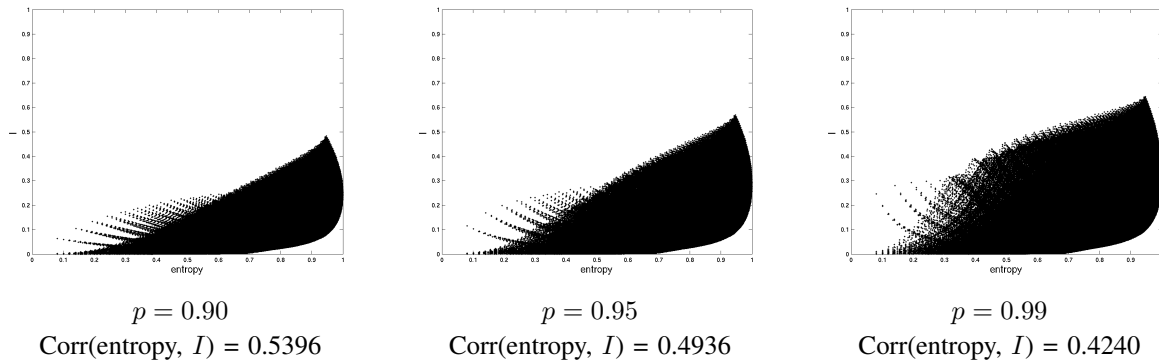


Fig. 8. Scatterplot of IMP score ( $I$ ) versus entropy with 970299 points of the parameter space  $(P_1, P_2, P_3)$  for  $p = 0.90$ ,  $p = 0.95$  and  $p = 0.99$  and logic 00011001, (partly-IMP logic). The respective correlations are below each plot.

Based on the foregoing results, we conclude that, at least for the case of independence between predictors, the logic associated with the predictive power and the entropy of the joint predictor distribution is crucial in the occurrence of IMP genes.

3) *Dependent Predictors*: We now drop the independence assumption. In this part, the analysis is done for two predictor sets only because the analysis for three predictors requires an excessive number of parameters. We utilize the covariance between the predictors as a measure of their dependence. Note that, in the binary case, the predictors are independent if and only if the covariance is null. The covariance  $\gamma$  between the predictors  $X_1$  and  $X_2$  is given by  $\gamma = E[X_1 X_2] - E[X_1]E[X_2]$ . Tables III and IV show the JPD tables adapted with the covariance  $\gamma$ .

### *Equiprobable Predictors*

TABLE III  
JPD TABLE FOR THE AND LOGIC ADAPTED WITH THE COVARIANCE  $\gamma$ .

$X_1$	$X_2$	$P(Y = 0, X_1 = i, X_2 = j)$	$P(Y = 1, X_1 = i, X_2 = j)$
0	0	$((1 - P_1)(1 - P_2) + \gamma)p$	$((1 - P_1)(1 - P_2) + \gamma)(1 - p)$
0	1	$((1 - P_1)P_2 - \gamma)p$	$((1 - P_1)P_2 - \gamma)(1 - p)$
1	0	$(P_1(1 - P_2) - \gamma)p$	$(P_1(1 - P_2) - \gamma)(1 - p)$
1	1	$(P_1P_2 + \gamma)(1 - p)$	$(P_1P_2 + \gamma)p$

TABLE IV  
JPD TABLE FOR THE XOR LOGIC ASSUMING INDEPENDENCE BETWEEN  $X_1$  AND  $X_2$ .

$X_1$	$X_2$	$P(Y = 0, X_1 = i, X_2 = j)$	$P(Y = 1, X_1 = i, X_2 = j)$
0	0	$((1 - P_1)(1 - P_2) + \gamma)p$	$((1 - P_1)(1 - P_2) + \gamma)(1 - p)$
0	1	$((1 - P_1)P_2 - \gamma)(1 - p)$	$((1 - P_1)P_2 - \gamma)p$
1	0	$(P_1(1 - P_2) - \gamma)(1 - p)$	$(P_1(1 - P_2) - \gamma)p$
1	1	$(P_1P_2 + \gamma)p$	$(P_1P_2 + \gamma)(1 - p)$

We first consider  $P_1 = P_2 = \frac{1}{2}$ . Based on Equation 8 and the JPD tables III and IV, we have  $\alpha_{00} = (\frac{1}{4} + \gamma)$ ,  $\alpha_{01} = (\frac{1}{4} - \gamma)$ ,  $\alpha_{10} = (\frac{1}{4} - \gamma)$ , and  $\alpha_{11} = (\frac{1}{4} + \gamma)$  for all logics.

The CoDs for the AND logic, assuming  $P_1 = P_2 = \frac{1}{2}$  and  $p \geq 0.5$ , are given by

$$\text{CoD}_1 = \text{CoD}_2 = 1 - \frac{1 - p + F\left(\frac{8p\gamma - 4\gamma + 1}{2}\right)}{2F\left(\frac{3 - 2p + 8p\gamma - 4\gamma}{4}\right)} \quad (16)$$

$$\text{CoD}_{12} = 1 - \frac{1 - p}{F\left(\frac{3 - 2p + 8p\gamma - 4\gamma}{4}\right)}$$

After some algebraic manipulation, the equations can be simplified by using the fact that  $-\frac{1}{4} \leq \gamma \leq \frac{1}{4}$  (since probabilities are non-negative):

$$\text{CoD}_1 = \text{CoD}_2 = \begin{cases} 0 & \text{if } -0.25 \leq \gamma \leq 0 \\ 1 - \frac{3 - 2p - 8p\gamma + 4\gamma}{3 - 2p + 8p\gamma - 4\gamma} & \text{if } 0 \leq \gamma \leq 0.25 \end{cases}$$

$$\text{CoD}_{12} = 1 - \frac{4(1 - p)}{3 - 2p + 8p\gamma - 4\gamma} \quad (17)$$

The CoDs for the XOR logic, assuming  $P_1 = P_2 = \frac{1}{2}$  and  $p \geq 0.5$ , are given by

$$\begin{aligned} \text{CoD}_1 = \text{CoD}_2 &= 1 - \frac{F\left(\frac{-8p\gamma+4\gamma+1}{2}\right)}{F\left(\frac{-8p\gamma+4\gamma+1}{2}\right)} = 0 \\ \text{CoD}_{12} &= 1 - \frac{1-p}{F\left(\frac{-8p\gamma+4\gamma+1}{2}\right)} \end{aligned} \quad (18)$$

The equation for  $\text{CoD}_{12}$  can be simplified by using the fact that  $-\frac{1}{4} \leq \gamma \leq \frac{1}{4}$ , and assuming  $p > \frac{1}{2}$ :

$$\text{CoD}_{12} = \begin{cases} 1 - \frac{2(1-p)}{8p\gamma-4\gamma+1} & \text{if } -0.25 \leq \gamma \leq 0 \\ 1 - \frac{2(1-p)}{-8p\gamma+4\gamma+1} & \text{if } 0 \leq \gamma \leq 0.25 \end{cases} \quad (19)$$

Note that, for  $p = \frac{1}{2}$ , the IMP score is zero for both the AND and XOR logics, regardless of the value of  $\gamma$  (this corresponds to the case of no predictive power).

Another important observation is that, for any given value of the predictive power  $p$ , the maximum IMP score is achieved by  $\gamma = 0$  for both logics (AND and XOR); i.e., the case of independent predictors. For the XOR logic, it is easy to see this, as the individual CoDs are zero, and the maximum IMP score is obtained by maximizing  $\text{CoD}_{12}$ , which corresponds to maximization of the denominator in the expression for  $\text{CoD}_{12}$  in Equation 18). But, as seen in Figure 1, the maximum of  $F$  is achieved at the point  $\frac{1}{2}$ ; the resulting equation leads to  $\gamma = 0$ . The AND logic case is a bit more complex, and the proof can be found in the Appendix.

Fig. 9 shows plots of  $\text{CoD}_{12}$ ,  $\max\{\text{CoD}_1, \text{CoD}_2\}$  and  $I = \text{CoD}_{12} - \max\{\text{CoD}_1, \text{CoD}_2\}$  as functions of  $p$  for some non-negative  $\gamma$  values. Fig. 10 presents the same plots for non-positive  $\gamma$  values. In all cases, the joint CoD, the individual CoDs (where they are non-zero), and the IMP score increase with increasing predictive power. Also in all cases, the IMP score increases with increasing correlation. The plots are exactly the same for the logics symmetric to the AND logic, namely, for the 1000 (NAND), 0111 (OR) and 1110 (NOR) logics. For the remaining 1,3-minterm logics (0100, 0010, 1011, 1101), the plots are identical to the plots for inverse signal of  $\gamma$  considering the AND logic. For the NXOR logic, the plots are the same as for the XOR logic.

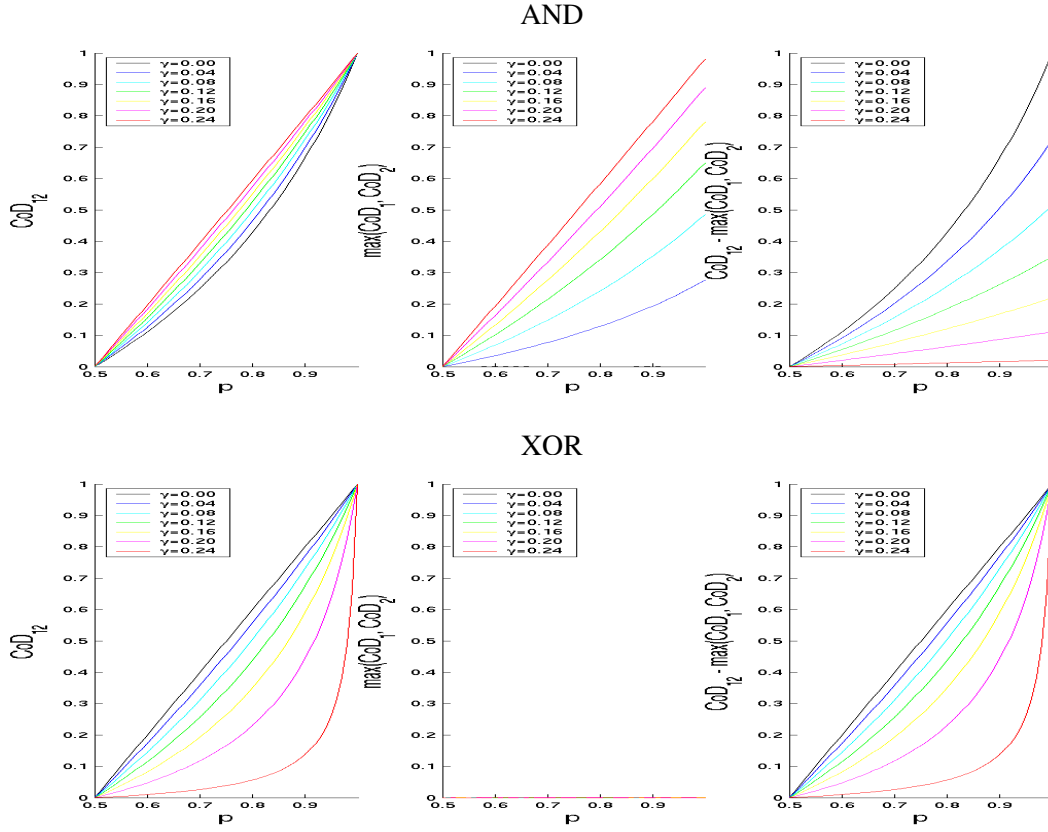


Fig. 9. Plots of CoDs and IMP score as a function of  $p$  for  $P_1 = P_2 = 0.5$  and non-negative covariance values ( $\gamma \geq 0$ ).

### General Case: Non-Equprobable Predictors

Now assuming that  $P_1$  and  $P_2$  can take any values between 0 and 1, based on Equation 8 and on the JPD tables III and IV, we have  $\alpha_{00} = (1 - P_1)(1 - P_2) + \gamma$ ,  $\alpha_{01} = (1 - P_1)P_2 - \gamma$ ,  $\alpha_{10} = P_1(1 - P_2) - \gamma$ , and  $\alpha_{11} = P_1P_2 + \gamma$  for all logics. For the AND logic (assuming  $p \geq 0.5$ ),

$$\begin{aligned}
 \text{CoD}_1 &= 1 - \frac{(1 - P_1)(1 - p) + F \left[ P_2 + \frac{\gamma - (2P_1P_2 + 2\gamma - P_1)p}{P_1} \right] P_1}{F [P_1P_2 + \gamma + (1 - 2P_1P_2 - 2\gamma)p]} \\
 \text{CoD}_2 &= 1 - \frac{(1 - P_2)(1 - p) + F \left[ P_1 + \frac{\gamma - (2P_1P_2 + 2\gamma - P_2)p}{P_2} \right] P_2}{F [P_1P_2 + \gamma + (1 - 2P_1P_2 - 2\gamma)p]} \\
 \text{CoD}_{12} &= 1 - \frac{1 - p}{F [P_1P_2 + \gamma + (1 - 2P_1P_2 - 2\gamma)p]}
 \end{aligned} \tag{20}$$

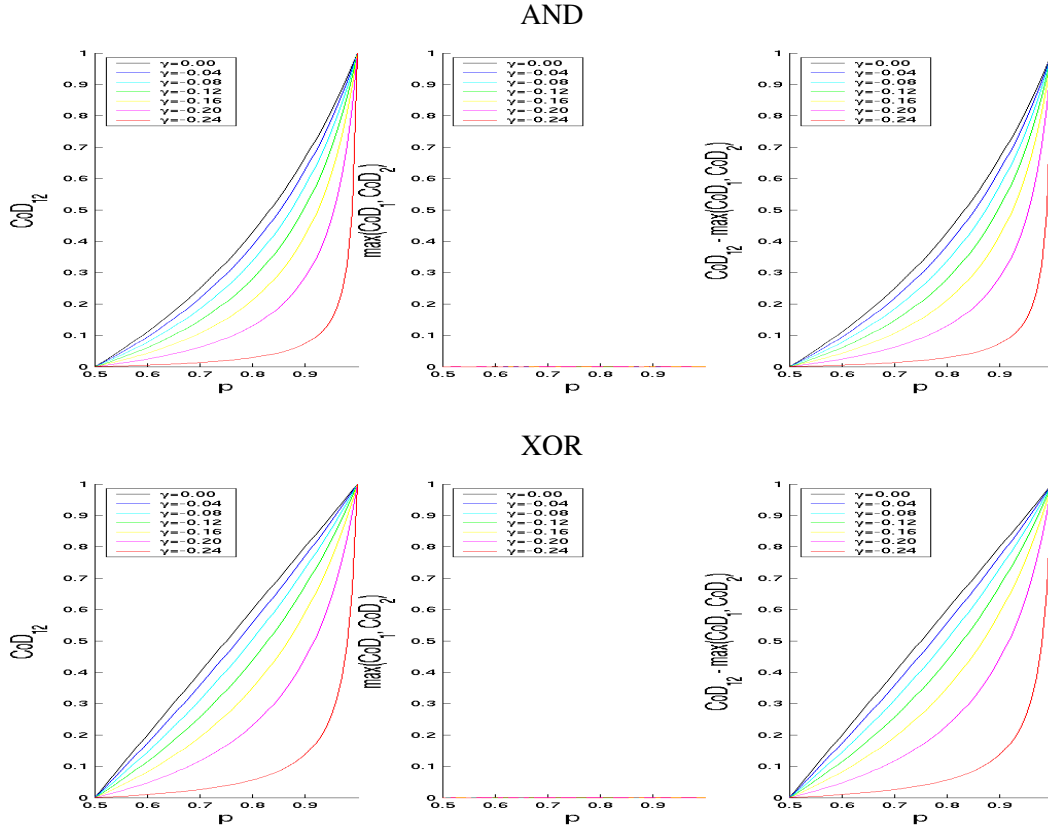


Fig. 10. Plots of CoDs as a function of  $p$  for  $P_1 = P_2 = 0.5$  and non-positive covariance values ( $\gamma \leq 0$ ).

For the XOR logic (assuming  $p \geq 0.5$ ),

$$\begin{aligned}
 \text{CoD}_1 &= 1 - \frac{F \left[ P_2 + \frac{(1-2P_2-P_1+2P_1P_2+2\gamma)p-\gamma}{1-P_1} \right] (1-P_1) + F \left[ P_2 + \frac{(P_1-2P_1P_2-2\gamma)p+\gamma}{P_1} \right] P_1}{F [P_1 + P_2 - 2P_1P_2 - 2\gamma + (1 - 2P_1 - 2P_2 + 4P_1P_2 + 4\gamma)p]} \\
 \text{CoD}_2 &= 1 - \frac{F \left[ P_1 + \frac{(1-2P_1-P_2+2P_1P_2+2\gamma)p-\gamma}{1-P_2} \right] (1-P_2) + F \left[ P_1 + \frac{(P_2-2P_1P_2-2\gamma)p+\gamma}{P_2} \right] P_2}{F [P_1 + P_2 - 2P_1P_2 - 2\gamma + (1 - 2P_1 - 2P_2 + 4P_1P_2 + 4\gamma)p]} \quad (21) \\
 \text{CoD}_{12} &= 1 - \frac{1-p}{F [P_1 + P_2 - 2P_1P_2 - 2\gamma + (1 - 2P_1 - 2P_2 + 4P_1P_2 + 4\gamma)p]}
 \end{aligned}$$

Note that, for  $p = \frac{1}{2}$ , the IMP score is zero for both the AND and XOR logics, regardless of the values of  $P_1$ ,  $P_2$  and  $\gamma$  (this corresponds to the case of no predictive power).

With these equations we plot graphs of IMP score for some  $p$  and  $\gamma$  values. Fig. 11 illustrates these

graphics for the AND and XOR logics. For the AND logic, we observe that positive covariance suppresses the region predisposed to generate IMPs (left-bottom square) while negative covariance expands the region around the square. For the XOR logic, both positive and negative covariance expand the central region with highest IMP score. Moreover, positive and negative covariance expand one of the diagonals. Therefore, a small covariance (positive or negative) has good impact on IMP score for the XOR logic, while for the AND logic, only small negative covariance is beneficial.

## V. CASE STUDY: DUSP1 IN MELANOMA

To test how well the model's predictions describe the behavior of broad controlling genes, an examination of a data set where samples containing active and inactive versions of a gene that acts as a general controller has been carried out. The dataset examined contains expression profiles of melanoma [24]. The gene exerting broad control is DUSP1, the prototypical member of a family of dual specificity (threonine/tyrosine) phosphatases that dephosphorylate MAP-kinase [25]. It is known that DUSP1 can serve as an effective antagonist to a variety of processes stimulated by activated MAPK, including both proliferation and apoptosis [26] and it has been recognized to be a functional antagonist to chronic replication driven by growth factors [27]. In its role as a guard against the high levels of proliferation frequently associated with cancer in mature, differentiated tissue, DUSP1 acts by dephosphorylating a mitogen-activated kinase, MAPK1, a protein that serves as an integration point for a diverse set of cellular processes in addition to proliferation. A significant fraction of the downstream transcriptional consequences of dephosphorylation of MAPK1 signaling derives from the inability of the dephosphorylated MAPK1 to activate transcription factors by phosphorylating them. Thus a DUSP1-induced change in MAPK1 phosphorylation status is expected to have a very significant effect on the abundance of the transcripts of many genes.

The dataset is composed of 31 samples with 587 gene expressions. Of the 31 samples, 19 are normal tissues and 12 are tissues with melanoma. Gene expressions are binarized to indicate change or no change relative to a reference expression level for each gene individually. A change can be under- or over-expression. Both cases are labeled as 1, whereas no significant change from the reference is labeled as 0.

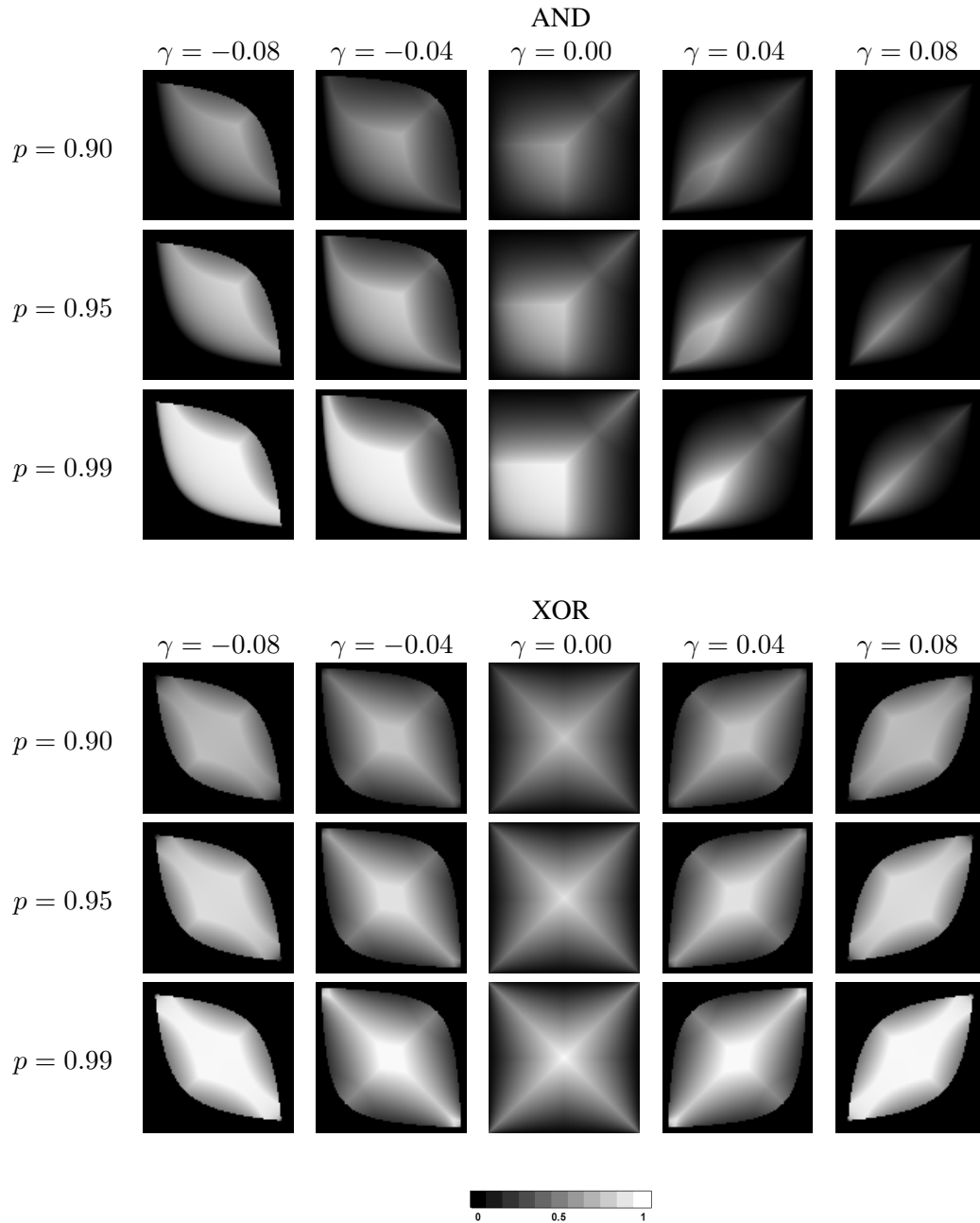


Fig. 11. Graphics representing IMP score values for AND and XOR logics and for three distinct values of  $p$  and five distinct values of  $\gamma$ . The probabilities  $P_1$  and  $P_2$  are given by X-axis and Y-axis respectively. Brighter gray-level intensities correspond to higher IMP score.

Treating each gene as a target, and given  $\lambda$  and  $\delta$ , the objective is to obtain a list of predictor pairs that are IMP with respect to  $\lambda$  and  $\delta$  for the gene, the desire being to analyze some factors affecting

IMP prediction: the probability of each predictor being 1 ( $P_1$  and  $P_2$ ), the covariance  $\gamma$  between the predictors, and the predictive power  $p$ . Whereas our model assumes uniform predictive power, in the sense that  $\pi_{ij} = p$  or  $\pi_{ij} = 1 - p$ , the purpose being to reduce the number of parameters, this assumption does not hold exactly for real data; nevertheless, for a specific logic of prediction, the assumption holds approximately, and the main points made in the theoretical analysis continue to hold.

We consider the following four  $(\lambda, \delta)$  pairs:  $(0.2, 0.7)$ ,  $(0.3, 0.7)$ ,  $(0.2, 0.8)$ ,  $(0.3, 0.8)$ . These are restrictive but still admit IMPs, whereas with  $\lambda = 0.1$  and  $\delta = 0.9$  very few IMP pairs are found. The full lists of the results can be found in the supplementary material (<http://www.vision.ime.usp.br/~davidjr/melanoma>); here, we confine our attention to  $\lambda = 0.2$  and  $\delta = 0.8$ .

From the 587 genes, we omit as targets 37 genes having less than 6 zeros or 6 ones. These show too little variability. Some of them yield a large quantity of IMPs, but given the small sample size and the minimal number of changes we have little confidence in the results. From the remaining 550 genes, there are 101 targets having at least one IMP predictor pair. The target gene possessing the largest number of IMP predictor pairs is DUSP1 (dual specificity phosphatase). It has 19 IMP pairs for  $\lambda = 0.2$  and  $\delta = 0.8$ . One of these 19 IMP pairs has the highest possible IMP score (individual CoDs equal to 0 and overall CoD equal to 1). The optimal predictors for all 19 pairs are determined by 1-minterm logic: 14 by AND logic (0001), 3 by  $Y = X_1 \wedge \bar{X}_2$  logic (0010), 1 by  $Y = \bar{X}_1 \wedge X_2$  logic (0100), and 1 by NAND logic (1000). Table V shows the list of IMP predictor pairs for DUSP1. Information about logic, number of occurrences of each instance,  $P_1$ ,  $P_2$ , and the covariance  $\gamma$  are presented. The corresponding IMP score plots with the locations  $(P_1, P_2)$  indicated as small black squares are shown in Fig. 12.

If we consider other less stringent values of  $\lambda$  and  $\delta$ , DUSP1 also lead to a considerable number of IMP sets. Considering the following pairs  $(\lambda, \delta) = \{(0.2, 0.7), (0.2, 0.8), (0.3, 0.7), (0.3, 0.8)\}$ , DUSP1 is the target gene with the largest number of IMP pairs for three of these pairs and with third largest number for  $(\lambda, \delta) = (0.3, 0.8)$ . Table VI shows the three target genes with largest number of IMPs for each  $(\lambda, \delta)$  pair.

The results also confirm the statement that if the predictive power  $p$  is not so strong, only the XOR and NXOR logic can produce IMPs, but if  $p$  is tight, the 1-minterm and 3-minterm logics are more likely to produce IMPs. In the melanoma data, assuming  $\lambda = 0.2$  and  $\delta = 0.8$ , the IMP sets assumed

TABLE V

LIST OF IMP SETS WHICH HAS DUSP1 AS TARGET. THE PREDICTIVE POWER IS INDICATED IN THE  $p$  COLUMN. THE COVARIANCE  $\gamma$  IS DISPLAYED IN THE LAST COLUMN. THE RESPECTIVE IMP SCORE GRAPHICS FOR EACH PREDICTOR SET ARE ILLUSTRATED IN FIGURE 12. THESE GRAPHICS WERE GENERATED BASED ON THE PREDICTION LOGIC,  $\gamma$ ,  $p$  AND THE PAIR  $(P_1, P_2)$  (THE LAST ONE INDICATED BY BLACK SQUARE IN THE GRAPHICS).

#	$X_1$	$X_2$	Logic	$CoD_1$	$CoD_2$	$CoD_{12}$	$p$	$P_1$	$P_2$	$\gamma$
1	RTN1	TEAD1	0001	0.0000	0.0000	1.0000	1.0000	0.5806	0.5161	-0.0763
2	CHN1	TOP1	0001	0.0000	0.1429	0.8571	0.9677	0.5484	0.4194	-0.0376
3	CASP3	STOM	0001	0.1429	0.0000	0.8571	0.9677	0.3548	0.4194	0.0462
4	EDG1	TEAD1	0001	0.0000	0.0000	0.8571	0.9677	0.3871	0.5161	-0.0065
5	MMP3	TEAD1	0001	0.0000	0.0000	0.8571	0.9677	0.4839	0.5161	0.0086
6	TGFBI	FOS	0001	0.0000	0.0000	0.8571	0.9677	0.4839	0.5161	-0.0581
7	UAP1	TOP1	0100	0.0000	0.1429	0.8571	0.9677	0.2581	0.4194	0.0548
8	TCF4	TEAD1	0001	0.0000	0.0000	0.8571	0.9677	0.5161	0.5161	-0.0753
9	TOP1	SERPINE1	0010	0.1429	0.0000	0.8571	0.9677	0.4194	0.4839	-0.0430
10	TOP1	TEAD1	0001	0.1429	0.0000	0.8571	0.9677	0.4194	0.5161	0.0430
11	TOP1	PLOD2	0010	0.1429	0.0000	0.8571	0.9677	0.4194	0.3548	0.0129
12	LAMA4	PCAF	0001	0.0000	0.0000	0.8571	0.9677	0.6452	0.4516	-0.0344
13	SERPINE1	PSFL	1000	0.0000	0.0000	0.8571	0.9677	0.4839	0.4516	-0.0258
14	IFIT1	TEAD1	0001	0.0000	0.0000	0.8571	0.9677	0.5484	0.5161	-0.0258
15	NR4A3	FOS	0001	0.0000	0.0000	0.8571	0.9677	0.6129	0.5161	-0.0602
16	CYP27A1	TEAD1	0001	0.0000	0.0000	0.8571	0.9677	0.4516	0.5161	0.0258
17	CYP27A1	ESTs	0010	0.0000	0.0000	0.8571	0.9677	0.4516	0.4839	0.0409
18	PCAF	FOS	0001	0.0000	0.0000	0.8571	0.9677	0.4516	0.5161	0.0258
19	ESTs	FOS	0001	0.0000	0.0000	0.8571	0.9677	0.6129	0.5161	-0.0602

TABLE VI

TARGET GENES WITH THE LARGEST NUMBER OF IMP SETS FOR EACH PAIR  
 $(\lambda, \delta) = \{(0.2, 0.7), (0.2, 0.8), (0.3, 0.7), (0.3, 0.8)\}$ .

	#	gene	#	gene
$\lambda = 0.2$	176	DUSP1	19	DUSP1
	103	SOCS1	16	ACY1
	87	MLANA	16	MLANA
$\lambda = 0.3$	350	DUSP1	49	MLANA
	225	MLANA	35	SOCS1
	215	SOCS1	33	DUSP1
	$\delta = 0.7$		$\delta = 0.8$	

one of the following four predictive power values: 1, 0.9677, 0.9355 and 0.9032. Table VII shows the distribution of each of these values for 1,3-minterm and 2-minterm logics IMP sets. Notice the correlation

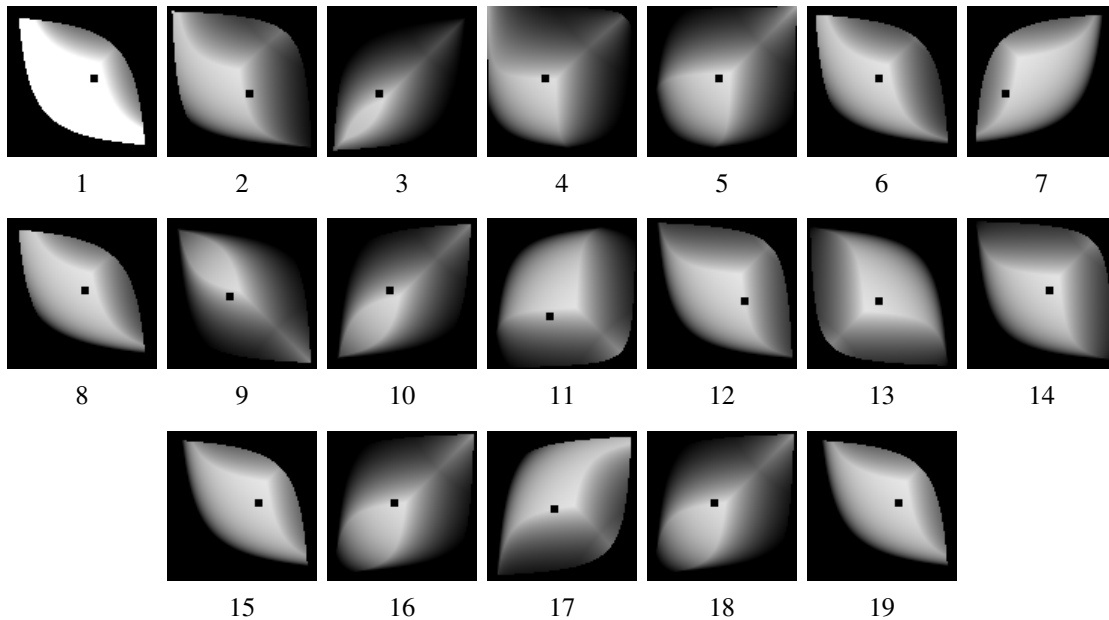


Fig. 12. IMP score plots relative to the DUSPI IMP sets given in Table V. Each graphic was generated taking as input the logic, the predictive power  $p$  and the covariance  $\gamma$  obtained. The small black squares point to the location  $(P_1, P_2)$  obtained.

between logics and  $p$  and the fact that none of the IMP sets with predictive power  $p = 0.9032$  is ruled by 1,3-minterm logic and none with predictive power  $p = 1$  is ruled by 2-minterm logic.

TABLE VII  
DISTRIBUTION OF THE IMP GENES CLASSIFIED AS 1,3-MINTERM LOGICS AND 2-MINTERM LOGICS (XOR AND NXOR)  
FOR EACH PREDICTIVE POWER VALUE APPEARING IN THE RESULTS WITH  $\lambda = 0.2$  AND  $\delta = 0.8$   
( $p = 0.9032, 0.9355, 0.9677, 1$ ).

	$p = 0.9032$	$p = 0.9355$	$p = 0.9677$	$p = 1$
1,3-minterm	0	92	102	3
2-minterm	7	39	3	0

## VI. CONCLUSION

In this paper we have studied the problem of binary prediction using the coefficient of determination (CoD). We have derived analytical formulas for the optimal, full-logic CoD as a function of the logic of prediction, predictive power, covariance between predictors, and marginal predictor probabilities. These have been used to determine conditions under which the phenomenon of intrinsic multivariate prediction (IMP) occurs, that is, the CoD of each proper subset of the predictors is low, but the joint

CoD of all predictors is large. In particular, this implies that no individual predictor allows determination of target behavior much better than chance, whereas combinations of these individually low-accuracy predictors determine the target almost deterministically. We have shown that IMP tends to occur for high predictive power of the joint set of predictors and for small correlation between predictors. We have also demonstrated that IMP is favored by specific prediction logics; for example, with two predictors, 2-minterm logics such as XOR and NXOR produce IMP at a lower threshold of predictive power than is required by 1- and 3-minterm logics such as AND, NAND and OR.

In our modeling of Boolean genomic regulation, the CoD measures how well the quantized transcriptional expression of a master gene is predicted by the quantized transcriptional expressions of its downstream slave genes. In this context, IMP corresponds to the case where the behavior of the master cannot be predicted reliably by any of its slave genes, but combinations of slave genes can predict it very reliably. The conditions for IMP derived in this paper mean that IMP occurs when a master gene exerts tight genomic regulation (i.e., there is high predictive power) of several normally uncorrelated pathways (i.e., there is low correlation between predictors). The IMP model therefore captures the idea of canalization of a biological system, proposed several decades ago by Conrad Waddington, regarding genes that can enforce broad corrective actions on cellular processes. By using actual gene expression data from a melanoma study, we have demonstrated that DUSP1, a gene hypothesized to be a canalizing gene in Waddington's sense, displays IMP features. Using these data, we have also demonstrated empirically the statement that 2-minterm logics of prediction produce IMP for lower predictive power than 1- and 3-minterm logics.

Based on these results, we putatively advance the IMP criterion as a practical tool for the identification of critical canalizing genes in quantized gene expression data. More studies, both analytical and experimental, are needed to achieve a better understanding of the IMP phenomenon, and confirm its usefulness as a gene discovery tool. One of the issues that will be examined in future studies concerns the use of the empirical CoD, estimated from microarray data, in assessing IMP. In the present study, we have considered the optimal prediction problem. The empirical CoD is asymptotically optimal. Therefore, as the number of available samples increases, its properties will approach the properties of the optimal CoD.

## APPENDIX

**Proposition:** For the AND logic and two equiprobable predictors with covariance  $\gamma$ , the IMP score  $I = \text{CoD}_{12} - \max(\text{CoD}_1, \text{CoD}_2)$  is maximum for  $\gamma = 0$ , i.e., for uncorrelated predictors.

**Proof:** We divide the proof in two parts. First, consider the case  $-0.25 \leq \gamma \leq 0$ . Then, according to Equation 17,  $\text{CoD}_1 = \text{CoD}_2 = 0$  and maximizing the IMP score corresponds to maximizing  $\text{CoD}_{12}$ , which is achieved by maximizing the term  $(3 - 2p + 8p\gamma - 4\gamma)$  in the expression for  $\text{CoD}_{12}$ . Since  $p \geq 0.5$ ,  $8p\gamma - 4\gamma = 4(2p - 1)\gamma$  is increasing with increasing  $\gamma$ . Therefore, the largest  $\gamma$  in the interval will result in the largest value of  $\text{CoD}_{12}$ , from which it follows that  $\gamma = 0$ . Now, consider the case  $0 \leq \gamma \leq 0.25$ . From Equation 17, and after simple algebraic manipulation, the IMP score is given by:

$$I = \frac{2p - 1 - (8p\gamma - 4\gamma)}{3 - 2p + 8p\gamma - 4\gamma}$$

Once again, since  $p \geq 0.5$ ,  $8p\gamma - 4\gamma = 4(2p - 1)\gamma$  is increasing with increasing  $\gamma$ . In the expression for the IMP score, the numerator is thus decreasing, while the denominator is increasing, with increasing  $\gamma$ . Therefore, the smallest  $\gamma$  in the interval will result in the largest value of  $\text{CoD}_{12}$ , from which it follows that  $\gamma = 0$ . Joining the two parts of the proof, the largest IMP score is obtained for  $\gamma = 0$ , as we would like to prove.  $\square$

## ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation (CCF-0514644), the National Cancer Institute (2 R25CA090301), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - proc. 3696/06-5), FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo - proc. 04/03967-0) and the Translational Genomics Research Institute.

## REFERENCES

- [1] C. H. Waddington. Canalization of development and the inheritance of acquired characters. *Nature*, pages 563–565, 1942.
- [2] L. Chang and M. Karin. Mammalian map kinase signalling cascades. *Nature*, 410(6824):37–40, 2001.
- [3] C.J. Tabin and R.A. Weinberg. Analysis of viral and somatic activations of the cHa-ras gene. *J Virol*, 53(1):260–265, 1985.

- [4] M. Gomez-Lazaro, F.J. Fernandez-Gomez, and J. Jordan. p53: twenty five years understanding the mechanism of genome protection. *J Physiol Biochem*, 60(4):287–307, 2004.
- [5] W.S. Pear, J.P. Miller, L. Xu, J.C. Pui, B. Soffer, R.C. Quackenbush, A.M. Pendergast, R. Bronson, J.C. Aster, M.L. Scott, and D. Baltimore. Efficient and rapid induction of a chronic myelogenous leukemia-like myeloproliferative disease in mice receiving P210 bcr/abl-transduced bone marrow. *Blood*, 92(10):3780–3792, 1998.
- [6] D. Anastassiou. Computational analysis of the synergy among multiple interacting genes. *Molecular Systems Biology*, 3:83, 2007.
- [7] V. Varadan, D.M. Miller III, and D. Anastassiou. Computational inference of the molecular logic for synaptic connectivity in *c. elegans*. *Bioinformatics*, 22(14):e497–e506, 2006.
- [8] E. R. Dougherty, S. Kim, and Y. Chen. Coefficient of determination in nonlinear signal processing. *EURASIP Journal on Signal Processing*, 80(10):2219–2235, October 2000.
- [9] S. Kim, E. R. Dougherty, M. L. Bittner, Y. Chen, K. Sivakumar, P. Meltzer, and J. M. Trent. A general framework for the analysis of multivariate gene interaction via expression arrays. *Biomedical Optics*, 5(4):411–424, October 2000.
- [10] S. Kim, E. R. Dougherty, Y. Chen, K. Sivakumar, P. Meltzer, J. M. Trent, and M. Bittner. Multivariate measurement of gene-expression relationships. *Genomics*, 67:201–209, July 2000.
- [11] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang. A rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18:261–274, February 2002.
- [12] D. J. Hunter and P. Kraft. Drinking from the fire hose—statistical issues in genomewide association studies. *N Engl J Med*, 357(5):436–439, 2007.
- [13] T. M. Morgan, H. M. Krumboltz, R. P. Lifton, and J. A. Spertus. Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. *JAMA*, 297(14):1551–1561, 2007.
- [14] N. J. Samani, J. Erdmann, A. S. Hall, C. Hengstenberg, M. Mangino, B. Mayer, R. J. Dixon, T. Meitinger, P. Braund, H. E. Wichmann, J. H. Barrett, I. R. Knig, S. E. Stevens, S. Szymczak, D. A. Tregouet, M. M. Iles, F. Pahlke, H. Pollard, W. Lieb, F. Cambien, M. Fischer, W. Ouwehand, S. Blankenberg, A. J. Balmforth, A. Baessler, S. G. Ball, T. M. Strom, I. Braenne, C. Gieger, P. Deloukas, M. D. Tobin, A. Ziegler, J. R. Thompson, and H. Schunkert. Genomewide association analysis of coronary artery disease. *N Engl J Med*, 357(5):443–453, 2007.
- [15] N. Wiener. *Cybernetics: Or Control and Communication in the Animal and the Machine*. Massachusetts Institute of Technology Press, Cambridge, 1948.
- [16] W. B. Cannon. *The wisdom of the body*. W. W. Norton & Company, New York, 1939.
- [17] W. Xu, L. H. Kasper, S. Lerach, T. Jeevan, and P. K. Brindle. Individual creb-target genes dictate usage of distinct camp-responsive coactivation mechanisms. *EMBO J*, 26(12):2890–2903, 2007.
- [18] E. Consortium. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.
- [19] E. R. Dougherty, M. Brun, J. M. Trent, and M. L. Bittner. A conditioning-based model of contextual genomic regulation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2007. (to appear).
- [20] H. Reichenbach. *The Rise of Scientific Philosophy*. University of California Press, Berkeley, 1951.

- [21] K. Murphy and S. Mian. Modelling gene expression data using dynamic bayesian networks. Technical report, University of California, Berkeley, CA, 1999.
- [22] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
- [23] C. E. Shannon and W. Weaver. *The mathematical theory of communication*. Univ. of Illinois Press, 1963.
- [24] M. Bittner, P. Meltzer, J. Khan, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, E. Gillanders, A. Leja, K. Dietrich, C. Beaudry, M. Berrens, D. Alberts, V. Sondak, N. Hayward, and J. M. Trent. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536–540, 2000.
- [25] D. R. Alessi, C. Smythe, and S. M. Keyse. The human c1100 gene encodes a tyr/thr-protein phosphatase which potently and specifically inactivates map kinase and suppresses its activation by oncogenic ras in xenopus oocyte extracts. *Oncogene*, 8(7):2015–2020, 1993.
- [26] K. S. Smalley. A pivotal role for erk in the oncogenic behaviour of malignant melanoma? *Int J Cancer*, 104(5):527–532, 2003.
- [27] T. Noguchi, R. Metz, L. Chen, M. G. Matti, D. Carrasco, and R. Bravo. Structure, mapping, and expression of erp, a growth factor-inducible gene encoding a nontransmembrane protein tyrosine phosphatase, and effect of erp on cell growth. *Mol Cell Biol*, 13(9):5195–5205, 1993.