DISCOVERING RELATIONSHIPS IN GENETIC REGULATORY NETWORKS

A Thesis

by

RANADIP PAL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

August 2004

Major Subject: Electrical Engineering

DISCOVERING RELATIONSHIPS IN GENETIC REGULATORY NETWORKS

A Thesis

by

RANADIP PAL

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Approved as to style and content by:

Aniruddha Datta (Chair of Committee)	
Edward Dougherty (Member)	Don R. Halverson (Member)
Thomas Vogel (Member)	Chanan Singh (Head of Department)

August 2004

Major Subject: Electrical Engineering

ABSTRACT

Discovering Relationships in Genetic Regulatory Networks. (August 2004)

Ranadip Pal, B. Tech., Indian Institute of Technology, Kharagpur

Chair of Advisory Committee: Dr. Aniruddha Datta

The development of cDNA microarray technology has made it possible to simultaneously monitor the expression status of thousands of genes. A natural use for this vast amount of information would be to try and figure out inter-gene relationships by studying the gene expression patterns across different experimental conditions and to build Gene Regulatory Networks from these data. In this thesis, we study some of the issues involved in Genetic Regulatory Networks. One of them is to discover and elucidate multivariate logical predictive relations among gene expressions and to demonstrate how these logical relations based on coarse quantization closely reflect corresponding relations in the continuous data. The other issue involves construction of synthetic Probabilistic Boolean Networks with particular attractor structures. These synthetic networks help in testing of various algorithms like Bayesian Connectivity based approach for design of Probabilistic Boolean Networks.

ACKNOWLEDGMENTS

I would like to especially thank my advisor, Dr. Aniruddha Datta, for his guidance and encouragement throughout my research. He always gave a patient hearing to my ideas and encouraged me to look for newer concepts.

I am also indebted to Dr. Dougherty for his invaluable suggestions and guidance throughout my research.

I am thankful to my parents and my brother for their immense support, love and inspiration throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Multicellular organisms, such as ourselves, are made up of billions of cells, each of which must behave in accordance with certain strict rules if the organism is to survive and carry on the basic functions of life. *Eucaryotic* cells, such as human cells, are characterized by the presence of an intra-cellular compartment called the *nucleus*. The nucleus contains the instructions that are necessary for the proper functioning of the cell. These instructions are written in the form of deoxyribonucleic acid (DNA) and must be replicated and handed down unchanged to its progeny when the cell divides. The DNA is a long polymeric molecule (chain-like molecule, comprising numerous individual units, called monomers, linked together in series) with the structure of a double helix [1]. Complementary base pairing is the fundamental idea by which the sequence of a DNA molecule is copied during replication of the double helix and is vital for expression of the biological information in a form utilizable by a cell. All genes undergo the first stage of gene expression which is called transcription. During transcription, the template strand of the gene directs synthesis of an RNA molecule. The second stage of the gene expression is translation for some genes while for others, the RNA transcript is the end product. Whatever be the case, once the information has been used to synthesize proteins, it cannot be transmitted back to the DNA. This is called the Central Dogma of Molecular Biology. We try to measure these gene expressions using cDNA microarrays. cDNA microarray technology has made it possible to simultaneously monitor the expression status of thousands of genes. A natural use for this vast amount of information would be to try and figure out inter-gene relationships by studying the gene expression patterns across different experimental conditions and to build

---

The journal model is *IEEE Transactions on Automatic Control.*

Gene Regulatory Networks from this data.

There are many approaches to modeling genetic regulatory networks. Some are based on Differential Equation models [2] while some are based on Boolean models [3]. The Boolean Model has received a lot of attention. In the Boolean model, a gene is modeled as ON or OFF, and its value at time $t + 1$ is a function of the value set at time t of a class of regulatory genes in the network. Certainly such a coarse quantization will result in the loss of information and render the model unsuitable for finer analysis, but it will permit much easier model identification from data and can also be useful for prediction in situations where the dominant discriminatory features are binary. Research indicates that many realistic biological questions may be addressed within the Boolean formalism, and Boolean networks, while structurally simple, are dynamically complex and have yielded insights into the overall behavior of genetic networks [4] [5] [6]. Recently, the Boolean model has been extended to probabilistic Boolean networks (PBN's), which are essentially a family of Boolean networks [7]. At any given time, the system occupies a state governed by the regulatory functions of one of the networks, and with small probability it has the possibility of switching to a different governing network at the next instant of time.

A. Previous Work

Correlation can identify pair-wise genetic co-regulative responses to a particular stimulus; however, correlation does not address the fundamental problem of determining sets of genes whose actions and interactions drive the cell's decision to set the transcriptional level of a particular gene. Transcriptional control is accomplished by a complex method that interprets a variety of inputs [8] [9]. Hence, it is necessary to apply analytical tools that detect multivariate influences on decision making present in complex genetic networks. This demand has motivated the use of the Coefficient of Determination (CoD) to measure

the strength of the relationship between a set of predictor genes and a target gene [10] [11] [12]. In these applications, working with cDNA microarray data, the continuous numerical expression data is usually reduced to ternary logical data via a method of internal standardization [13]. In essence, relative to a given target gene and a set of predictor genes, the CoD measures the relative increase in predictive capability using the predictor-gene expressions as opposed to predicting the target-gene expression based only on knowledge of the target gene's isolated behavior across the data set. Mathematically,

$$COD \;=\; \frac{\epsilon_0 - \epsilon_{opt}}{\epsilon_0}$$

where ϵ_0 is the error arising when using the best estimate of the target-gene expression level, given only the statistics relating to the target gene itself, without using any information concerning other genes, and ϵ_{opt} is the error arising using the best estimate of the target-gene expression level using the expression levels of a set of predictor genes. If a predictor set can perfectly predict a target, then ϵ_{opt} = 0 and CoD = 1; at the other extreme, if a predictor set provides no additional information about the target, then $\epsilon_{opt} = \epsilon_0$ and CoD = 0. In general, $0 \leq$ CoD ≤ 1. Reduction to logical data accomplishes the kind of extreme compression necessary to apply predictive analysis with small samples typical of microarray experiments and facilitates the understanding of predictive relations based essentially on an up-regulated/down-regulated paradigm.

There have been a number of approaches to modeling gene regulatory networks – in particular, Bayesian networks [14], Boolean networks [3], and Probabilistic Boolean Networks (PBNs)[15] [16], the latter providing an integrated view of genetic function and regulation. The original design strategy for PBNs in [15] is based on the Coefficient of Determination between the target and predictor genes. Although the Boolean framework leads

to computational simplification and elegant formulation of relations, the model extends directly to genes having more than two states. A salient area of investigation regarding PBNs concerns the prospect of designing intervention strategies based on their long-run dynamics [17]. In particular, the question arises as to whether the theory of automatic control can be applied in the context of PBN dynamics to prescribe optimal intervention (treatment) strategies - to the extent, of course, that a treatment strategy can be characterized effectively in an ON-OFF paradigm for some set of genes [18] [19]. We define a *Gene Regulatory Network* (*GRN*)[20] to consist of a set of n genes, $g_1, g_2, ..., g_n$, each taking values in a finite set V (containing d values), a family of regulatory sets, $R_1, R_2, ..., R_n$,where R_k contains the genes that determine the value of gene g_k, and a set of functions, $f_1, f_2, ..., f_n$, governing the state transitions of the genes. The value of gene g_k at time $t+1$ is given by $g_k(t+1) = f_k[g_{k1}(t), g_{k2}(t), ..., g_{kn}(t)]$, where $R_k = \{g_{k1}, g_{k2}, ..., g_{kn}\}$. For a Boolean Network, $V = \{0, 1\}$. A *Probabilistic Gene Regulatory Network* (*PGRN*) consists of a set of n genes, $g_1, g_2, ..., g_n$, each taking values in a finite set V (containing d values), and a set of vector-valued network functions, $\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_r$, governing the state transitions of the genes. Mathematically, there is a set of state vectors $S = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$, with $m = d^n$ and $\mathbf{x}_k = (x_{k1}, x_{k2}, ..., x_{kn})$, where x_{ki} is the value of gene g_i in state k. Each network function $\mathbf{f}_j$ is composed of n functions $\psi_{j1}, \psi_{j2}, ..., \psi_{jn}$, and the value of gene g_i at time $t+1$ is given by $g_i(t+1) = \psi_{ji}[g_1(t), g_2(t), ..., g_{i-1}(t), g_{i+1}(t), ..., g_n(t)]$. The choice of which network function $\mathbf{f}_j$ to apply is governed by a selection procedure. Specifically, at each time point a random decision is made as to whether to switch the network function for the next transition, with a probability q of a switch being a system parameter. If a decision is made to switch the network function, then a new function is chosen from among $\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_r$, with the probability of choosing $\mathbf{f}_j$ being the selection probability c_j. In other words, each network function $\mathbf{f}_j$ determines a GRN and the PGRN behaves as a fixed GRN until a random decision (with probability q) is made to change the network func-

tion according to the probabilities $c_1, c_2, ..., c_r$ from among $\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_r$. In effect a PGRN switches between the GRNs defined by the network functions according to the switching probability q. A final aspect of the system is that at each time point there is a probability p of any gene changing its value uniformly randomly among the other possible values in V. Since there are n genes, the probability of there being a random perturbation at any time point is $1 - (1 - p)^n$. The state space S of the network together with the set of network functions, in conjunction with transitions between the states and network functions, determine a Markov chain. The random perturbation makes the Markov chain ergodic, meaning that it has the possibility of reaching any state from another state and that it possesses a long-run (steady-state) distribution.

B. Motivation

To model the genetic Regulatory network we need to know the connectivity between genes (intergene-relationships). One of the issues in this thesis will be to figure out multivariate logical predictive relationships among gene expressions in a data set arising from radiation studies using the NCI 60 Anti-Cancer Drug Screen (ACDS) cell lines and to demonstrate how these logical relations based on coarse quantization closely reflect corresponding relations in the continuous data. Many of the current paradigms for modeling genetic regulatory networks are conditioned on the premise that genes interact with each other through Boolean logic. This thesis will try to show that not only do such relationships exist but they can also be unearthed via the coefficient of determination technique. The other issue which this thesis will address is the method to generate synthetic boolean networks with singleton attractor structures. The view many current biologists hold is that the state space of genes mostly have singleton attractor structures and large cycles are transient. The problem of generating these structures is quite computationally intensive and in this thesis a

moderately less intensive algorithm for designing this type of structures will be presented along with analytical and empirical studies on the frequency of such structures in a totally randomly generated network.

The first goal relates to our general desire to discover multivariate gene expressions that go beyond correlative relationships and to our interest in finding candidate genes from which to build genetic regulatory networks. This thesis also seeks to address the question as to how and to what extent do logical relations among the quantized expression levels reflect numerical relations among the analog data, the latter being more directly related to the actual mRNA concentrations governing transcription.

Going beyond the identification of multi-gene predictive relations, one would like to use these relations to model genetic regulatory networks. Whereas a fine model with many parameters may be able to capture detailed low-level phenomena, such as protein concentrations and kinetics of reactions, construction of such a model requires large amounts of data for inference. On the other hand, a coarse model with few parameters and low complexity is restricted to capturing high-level phenomena, such as whether a gene is ON or OFF, but requires far less data. The principle of Occam's razor dictates that model complexity should never exceed what is necessary to faithfully "explain the data."

If binary or ternary relations are sufficient to describe a predictive relation between predictor and target genes, then one might expect that the logical functions are discernable within the continuous, pre-quantized data. After all, the hypothesis is that somehow the ON-OFF model sufficiently characterizes the multivariate relations between mRNA concentrations, at least to the extent that those concentrations themselves characterize transcriptional control. It will be demonstrated using the NCI 60 cell lines that, in fact, strong predictive (high CoD) functions discovered in the ternary context have counterparts for the continuous data.

C. Organization

The next few sections are organized as follows. Chapter II deals with finding out the predictive relationships from gene expression data. In chapter III, the issue of generating Boolean Networks of particular attractor structures is discussed. Chapter IV concludes the thesis by summarizing the main conclusions and outlining the directions for future research.

CHAPTER II

PREDICTIVE RELATIONSHIPS AMONG GENES

A. Coefficient of Determination Analysis

In this section, we provide an intuitive discussion of the Coefficient of Determination (COD) analysis which is the main data analysis tool used in this analysis. As already mentioned, the Coefficient of Determination measures the degree to which the best estimate for the transcriptional activity of a target gene can be improved using the knowledge of the transcriptional activity of some other predictor genes, relative to the best estimate in the absence of any knowledge of the transcriptional activity of the predictors. Mathematically,

$$COD = \frac{\epsilon_0 - \epsilon_{opt}}{\epsilon_0}$$

where ϵ_0 is the error arising when using the best estimate of the target-gene expression level given only statistics relating to the target gene itself, without using any information concerning other genes, and ϵ_{opt} is the error arising using the best estimate of the target-gene expression level using the expression levels of a set of predictor genes. If a predictor set can perfectly predict a target, then ϵ_{opt} = 0 and CoD = 1; at the other extreme, if a predictor set provides no information about the target, then ϵ_{opt} = ϵ_0 and CoD = 0. In general, 0 $\leq$ CoD $\leq$ 1.

Let us now consider a concrete example to demonstrate the COD for quantized gene expression data measured across several cell lines. Suppose we are interested in two genes G1 and G2 and their ternary-quantized expression patterns across seven cell lines are given by G1: (1 -1 0 0 0 0 1) and G2: (-1 1 1 1 1 0 0). If we want to predict the expression pattern of G1, then a reasonable measure of prediction error would be the number of incorrectly

predicted values for G1 divided by the total number of cell lines. First suppose that we are attempting to predict G1 without observing the expression pattern of G2 or any other gene. Then we would probably assign the value 0 to gene G1 based on the fact that this value for G1 occurs in the largest number of cell lines. Then ϵ_0 = (2+1)/7. Now suppose that we would like to use the knowledge of the expression pattern of G2 to predict G1. For a -1 in G2 we will assign a 1 for G1 because there is only one value of -1 in G2 and the corresponding value for G1 is 1; for a 1 in G2,we will assign a 0 for G1 because the corresponding G1 values are three 0's and one -1; and for a 0 in G2, we will arbitrarily decide to assign a 1 for G1, since the corresponding values in G1 are one 1 and one 0. Using these assignment rules and the given expression pattern for G2, we would predict the expression pattern of G1 as (1,0,0,0,0,1,1). Comparing this with the actual expression pattern for G1, there are two mismatches,so that $epsilon_{opt}$ = 2/7. Hence the COD for G2 predicting G1 is CoD=(1/7)/(3/7) = 1/3.

The CoD technique has at least three advantages over standard correlation analysis. First, the CoD can be applied to multiple predictors, thereby giving it the ability to discern multivariate inter-gene relationships. Second, the CoD can discover both linear and non-linear relationships, whereas the correlation coefficient only addresses linear relationships. For instance, if gene G1 has the expression pattern (0, 0, 0, 1, 1, 0) across six cell lines and gene G2 has the corresponding expression pattern (1, 1, 1, 0, 0, -1), then there is the relation $G1 = G2^2 - 1$, which is picked up by the CoD, with CoD = 1 but not picked up by the correlation coefficient, with Corr = 0. A third advantage of the CoD is that, whereas the correlation coefficient is independent of the order, the CODs for G1 predicting G2 and G2 predicting G1 can be quite different. For instance, the example just given, the CoD of G2 predicting G1 is 1, whereas the CoD for G1 predicting G2 is only 2/3. It may be that G1 serves as a regulatory switch for G2.

B. Gene Expression

Heredity or the process by which characteristics are passed from parents to offsprings so that all organisms resemble their ancestors is controlled by a vast number of factors called genes. Genes are made of deoxyribonucleic acid (DNA) and carried by chromosomes which are made of protein and DNA. The DNA is a long polymeric molecule (chain-like molecule, comprising numerous individual units, called monomers, linked together in a series) with a structure of a double helix. Complementary base pairing is the fundamental idea by which the sequence of a DNA molecule is copied during replication of the double helix and is vital for expression of the biological information in a form utilizable by a cell.

All genes undergo the first stage of gene expression which is called transcription (Fig 1). During transcription the template strand of the gene directs synthesis of an RNA molecule.The second stage of the gene expression is translation for some genes while for others RNA transcript is the end product. We try to measure these gene expressions using DNA microarray.

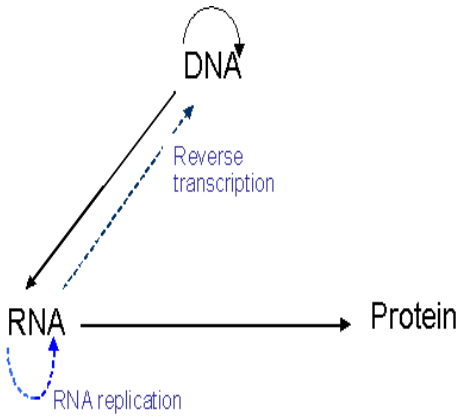

Fig. 1. Stages of gene expression

C. The Gene Expression Data and Its Processing

The data for the current study was obtained from radiation experiments conducted on cell lines from the National Cancer Institute(NCI) 60 AntiCancer Drug Screen(ACDS). The NCI 60 ACDS is a set of about 60 human cancer cell lines maintained at the National Cancer Institute. These cell lines have been derived from cancers of the colon, breast, ovary, lung, kidney, prostate, central nervous system, skin and bone marrow and serve as a screen for determining the efficacy of various compounds which are proposed as anti-cancer agents from time to time.

Sixty four cell lines from the NCI 60 ACDS were irradiated with high doses of ionizing radiation and harvested about 4 hours later. Microarrays were run with un-irradiated control and irradiated samples from the same cell line in each array. The genes which had responsiveness in at least 6 cell lines were selected. In this way, we identified about a thousand genes for further analysis. However, among this set, there were many genes with a large percentage of missing values (quality factor [13] less than .3) corresponding to different cell lines. If we attempted to use a missing value estimation algorithm for these genes, then the results would not have been very accurate. Accordingly, from the set of one thousand genes, we removed all the genes having poor quality data in more than 20 percent of the cell lines. Then we ran a missing value estimation algorithm (kNN impute [21]) on the remaining gene expression ratios and ternarized the estimated data using a threshold greater than 1.8 for induction and less than 0.5 for repression. The COD analysis [10] was then applied to the ternarized data to identify relationships between several genes responsive to ionizing radiation. Since the intergene relationships tend to be nonlinear and multivariate in nature, the COD technique is more appropriate than standard bivariate correlation analysis.

D. Missing Value Estimation Methods

Data from microarray experiments contains expression levels of genes under different experimental conditions. Due to the various reasons like noise and small sample spot sizes of the microarray (200 microns) there are frequently many missing values. Downstream analysis like clustering, classification and genetic network design is very much affected due to these missing values. Hence we need a method to estimate these absent data. There are many methods in the literature regarding missing value estimation but each has it's advantages and disadvantages. The kNN impute is one of the simplest of them and is quite robust. As explained in [21] this method selects genes with expression profiles similar to the gene of interest to impute missing values. If we consider gene A that has one missing value in experiment 1, this method would find K other genes, which have a value present in experiment 1, with expression most similar to A in experiments $2...N$ (where N is the total number of experiments). A weighted average of values in experiment 1 from the K closest genes is then used as an estimate for the missing value in gene A. In the weighted average, the contribution of each gene is weighted by similarity of its expression to that of gene A. Here we use the Euclidean Metric as the measure of gene similarity. Testing on other metrics like correlation coefficient, angle between gene expression vectors gives us better results sometimes but overall performance is better for Euclidean Metric. Only other method which gives slightly better results is Missing-value estimation using non-linear regression with Bayesian gene selection [22]. But this method is computationally intensive and hence we settle for the kNN method.

E. Robustness of Coefficient of Determination to Threshold

Since we are trying to figure out functions between genes from ternarized data, we would like to have the ternarized relationship hold for small changes in the threshold. In other

terms, based on the given threshold we find the optimal predictor, $z = f(x, y; t_0)$ where t_0 is the threshold. The error of this best predictor determines the CoD, say, C(x, y; z; t_0). Now, if we change the threshold, two things might happen. First, the error might change, thereby changing the CoD. Secondly, a different predictor may be optimal, thereby changing not only the CoD but the predictor function. If we focus on the first then upon finding f based on t_0 , f is fixed. Now let us change the threshold, so we are now considering the function $Q(t)$ defined by

$$Q(t) \quad = \quad \frac{\epsilon_0(t) - \epsilon_f(t)}{\epsilon_0(t)}$$

Where $\epsilon_0(t)$ is the error for the best predictor of z at threshold t given no observations and $\epsilon_f(t)$ is the error of predicting z using x and y at threshold t. The relationship is Robust relative to threshold when Q(t) is stable for small changes in t . Along with the graphs for expression values of genes we also plot Q(t) around the neighborhood of t_0 to find out the relevance of the logic function relative to the continuous data.

F. Instances of Boolean Relationships

We examined all genes that have a low COD when predicting a target individually but a high COD when they predict in conjunction with other genes. Specifically, we require that gene G1 predicting gene G3 and gene G2 predicting gene G3 have COD values at least 0.25 lower than the COD of genes G1 and G2 together predicting gene G3. This helps us to identify genes that in combination can significantly more strongly predict a particular target gene than individually. In the following sections we give some particular examples from such analysis while many other instances of strong relationship is maintained at the

website http://gsp.tamu.edu/Publications/supplement.htm .

1. Examples of “OR” Logic

If we consider the gene *mannose receptor, C type 1*(MRC1) as a target and the genes *visinin-like 1*(VSNL1) and *5-hydroxytryptamine (serotonin) receptor 2C*(HTR2C) as predictors for MRC1, then the individual CODs for predicting APC by the other two genes is 0.417 for each; however, used together to predict MRC1, the COD is 0.75.

The boolean relationship is as shown in Table I which defines the OR relation. Symbolically, MRC1 = VSNL1 $\vee$ HTR2C

Table I. Truth Table Showing OR Relationship

VSNL1	HTR2C	MRC1
0	0	0
0	1	1
1	0	1
1	1	1

To further investigate the relationship between the three genes, we produced the expression plots in Fig. 2 , where the blue bars represent the expression levels for MRC1, the green bars represent the expression levels for HTR2C, and the red bars represent the expression levels for VSNL1. The black Horizontal lines denotes the threshold for ternarizing which is 1.8 for +1 and .5 for -1. The plot demonstrates the OR relation between the target and the predictors in most of the cell lines.

This suggests if either of the predictors is high then the target gene expression level is also high, and that while the individual prediction by a single predictor may not be very

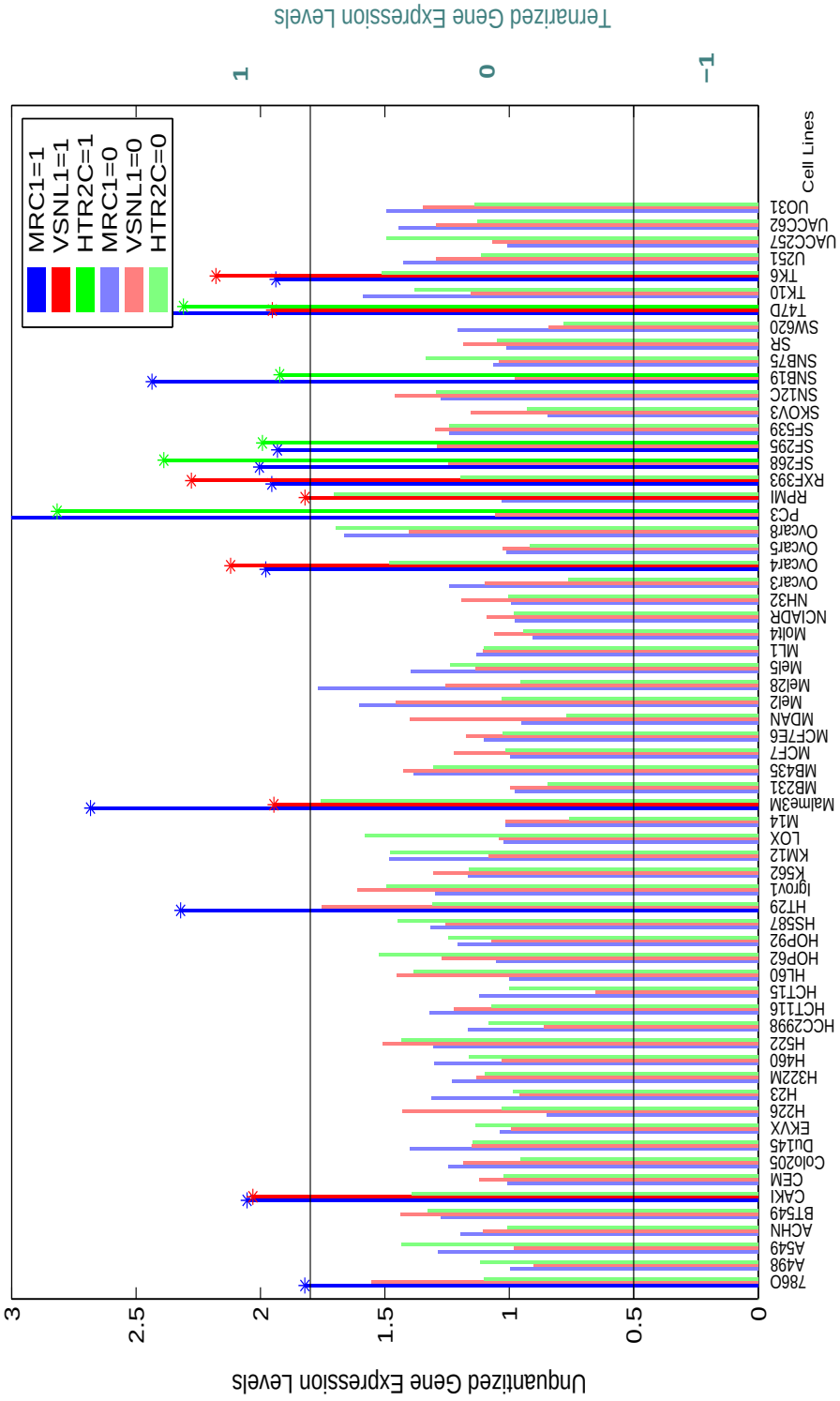

Fig. 2. OR relationship

reliable, a combined sum-type of prediction is quite accurate. This relationship is quite robust to changes in threshold as shown in Figure 3. Here we have plotted around the threshold for induction only as there are no repressed cell lines for these genes.

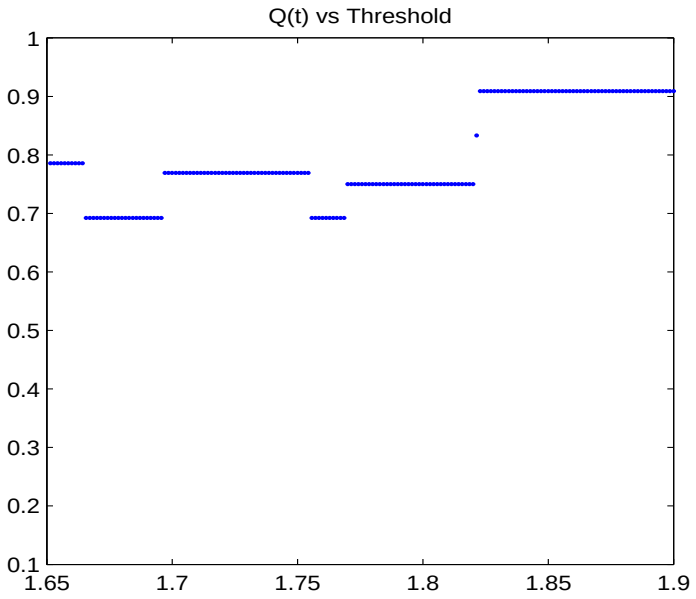

Fig. 3. Q(t) around 1.8 for MRC1

Another apparent instance of OR logic appear if we consider the gene *small inducible cytokine A7 (monocyte chemotactic protein 3)*(SCYA7) as the target and the genes *prosaposin (variant Gaucher disease and variant metachromatic leukodystrophy)*(PSAP) and *ribosomal protein L3*(RPL3) as the predictors.The COD for the predictor combination is 0.875 while the individual CODs are less than 0.65. Figure 4 shows the 'OR' relationship on the data. Here the blue bars represent the target SCYA7, the red bars represent PSAP, and the green bars represent RPL3.

The graph of Q(t) vs t is shown in Figure 5. The curve suggests that the relationship is robust to changes in threshold.

Some other instance of OR logic is visible when we consider the gene *adenomatosis polyposis coli*(APC) as a target and the genes *integrin, alpha L antigen CD11A p180*(ITGAL)

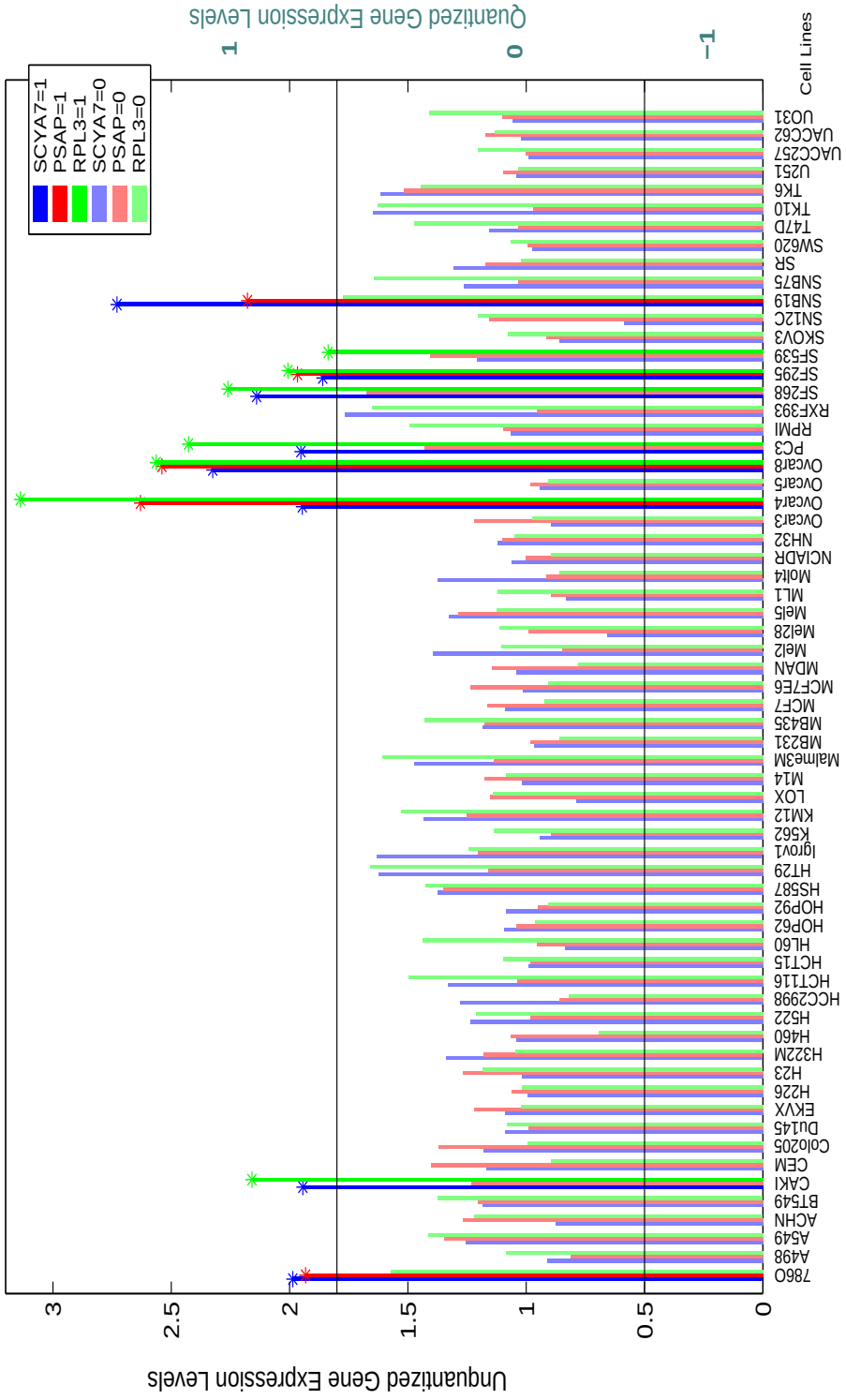

Fig. 4. Another OR relationship

and *Homo sapiens mRNA for TL132* as predictors for APC,then the individual CODs for predicting APC by the other two genes is 0.4 for each; however, used together to predict APC, the COD is 0.8. The individual correlation coefficients are 0.63 and 0.51. Owing to the OR relation above, we have evaluated the correlation coefficient of the sum of the gene expression profiles of the two predictor genes and the target gene, that is the correlation coefficient between ITGAL + mRNA and APC. This correlation coefficient is equal to 0.79. These numbers confirm that even without ternarizing the data, the two genes together seem to display a stronger relationship with the target than either of them considered individually.

2. Example of "AND" Logic

A case of AND logic is exhibited when we consider the gene *small inducible cytokine A7 -monocyte chemotactic protein 3*(SCYA7) as the target and the genes *mucin 5, subtypes A and C, tracheobronchial/gastric*(MUC5AC) and *calcium-sensing receptor (hypocalciuric hypercalcemia 1, severe neonatal hyperparathyroidism)*(CASR) as the predictors. The COD for combined prediction is 0.75 while the COD for each of the individual predictor genes is 0. The gene expression levels for these three genes are plotted in Figure 6, where the blue bars represent the gene expression level for the SCYA7, and the green and red bars represent the expression levels of the genes CASR and MUC5AC, respectively. The target clearly resembles an 'AND' function of the two predictors: when both predictors are high, then and only then is the target high. The boolean relationship is SCYA7 = CASR $\wedge$ MUC5AC and is shown in Table II. In terms of correlation coefficients the correlation coefficient between SCYA7 and MUC5AC is 0.61 and that between CASR and SCYA7 is 0.75, whereas the correlation coefficient between the product of the predictors and the target is 0.81. The robustness of this Boolean relationship is not that strong as depicted in Figure 7. The reason might be the optimum threshold for individual genes may be different.

Table II. Truth Table Showing AND Relationship

MUC5AC	CASR	SCYA7
0	0	0
0	1	0
1	0	0
1	1	1

3. Example of “EXOR” Logic

When the genes *mannose receptor, C type 1* (MRC1) and *interleukin 18 (interferon-gamma-inducing factor)* jointly predict the target gene *enoyl-Coenzyme A, hydratase/3-hydroxyacyl Coenzyme A dehydrogenase*(EHHADH), the output behaves as an XOR (exclusive OR) function of the inputs. This is clear from the plot shown in Figure 9, where the blue , red and green bars represent gene expression levels for the target gene EHHADH, the gene interleukin and the gene MRC1, respectively.When both predictors are upregulated or both are 0, then the target is also 0. Moreover it appears that interleukin can be a suppressor for EHHADH: whenever MRC1 is upregulated. The boolean relation, EHHADH = XOR(MRC1, Interleukin18), is shown in Table III. The plot for robustness is shown in Figure 8.

4. Boolean Relationships Among Four Genes

Thus far we have considered two genes predicting a target; now we treat a situation in which there are three predictor genes.

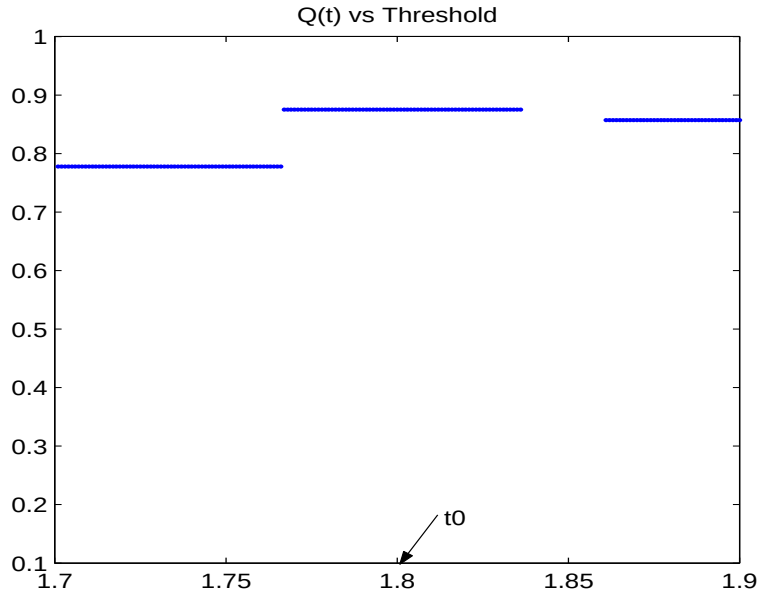

Fig. 5. Q(t) around 1.8 for SCYA7

Table III. Truth Table Showing EXOR Relationship

MRC1	InL	EHHADH
0	0	0
0	1	1
1	0	1
1	1	0

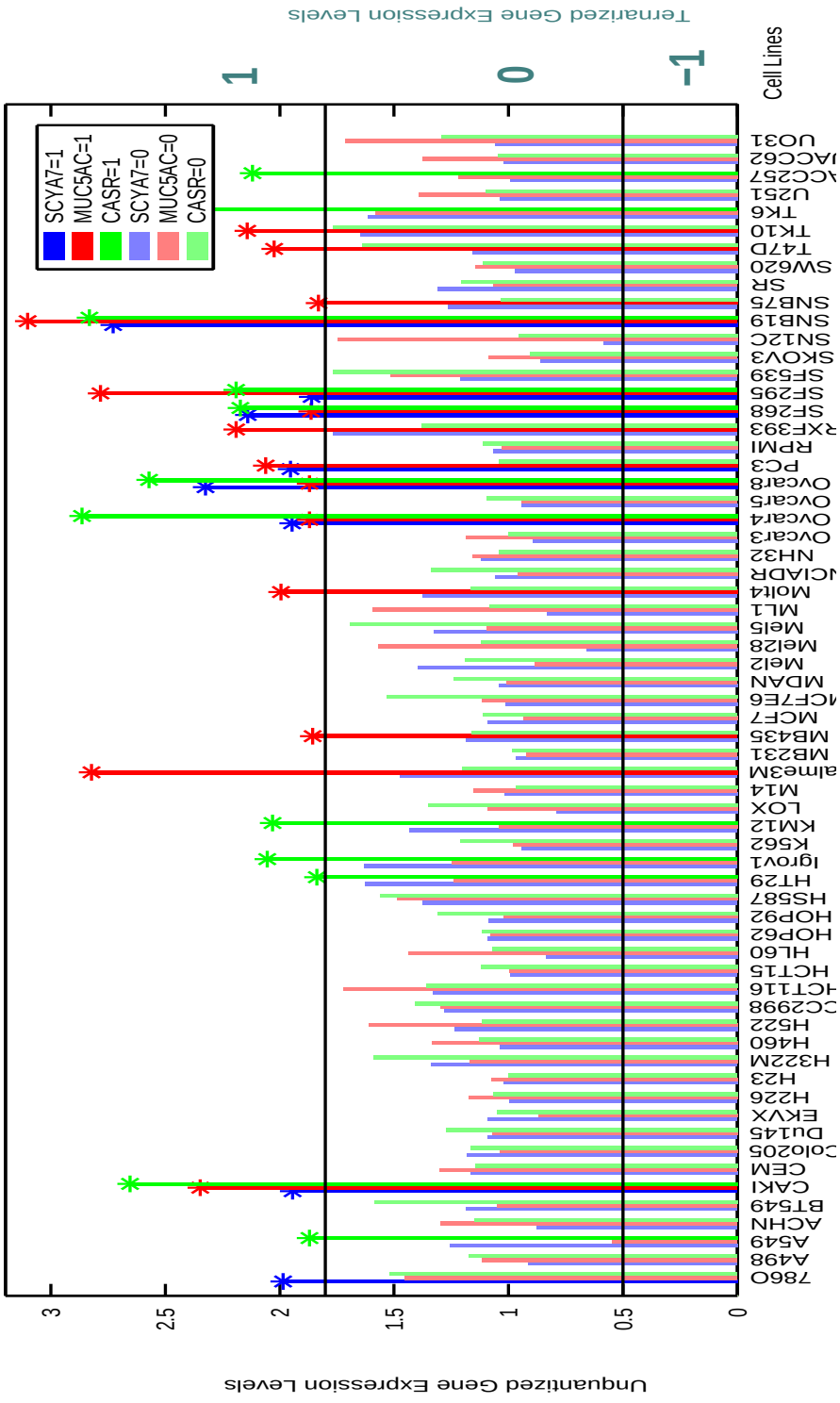

Fig. 6. AND relationship

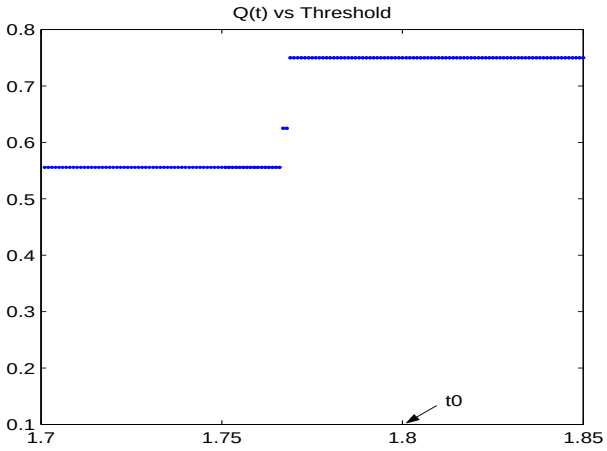

Fig. 7. Q(t) for AND function

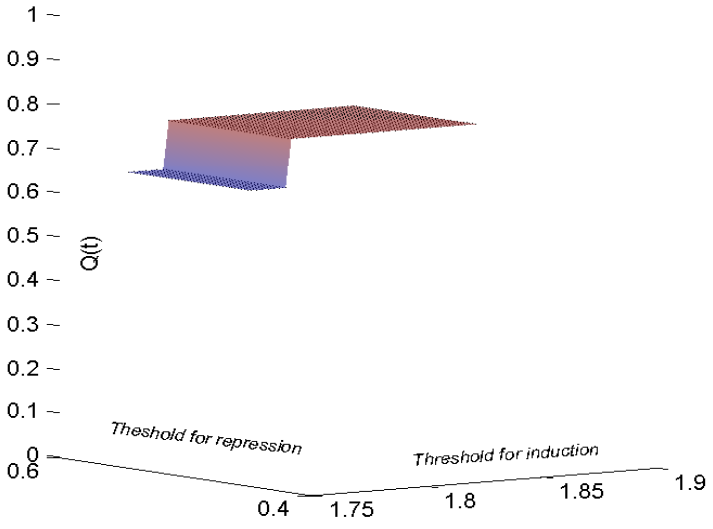

Fig. 8. Q(t1,t2) for EXOR function

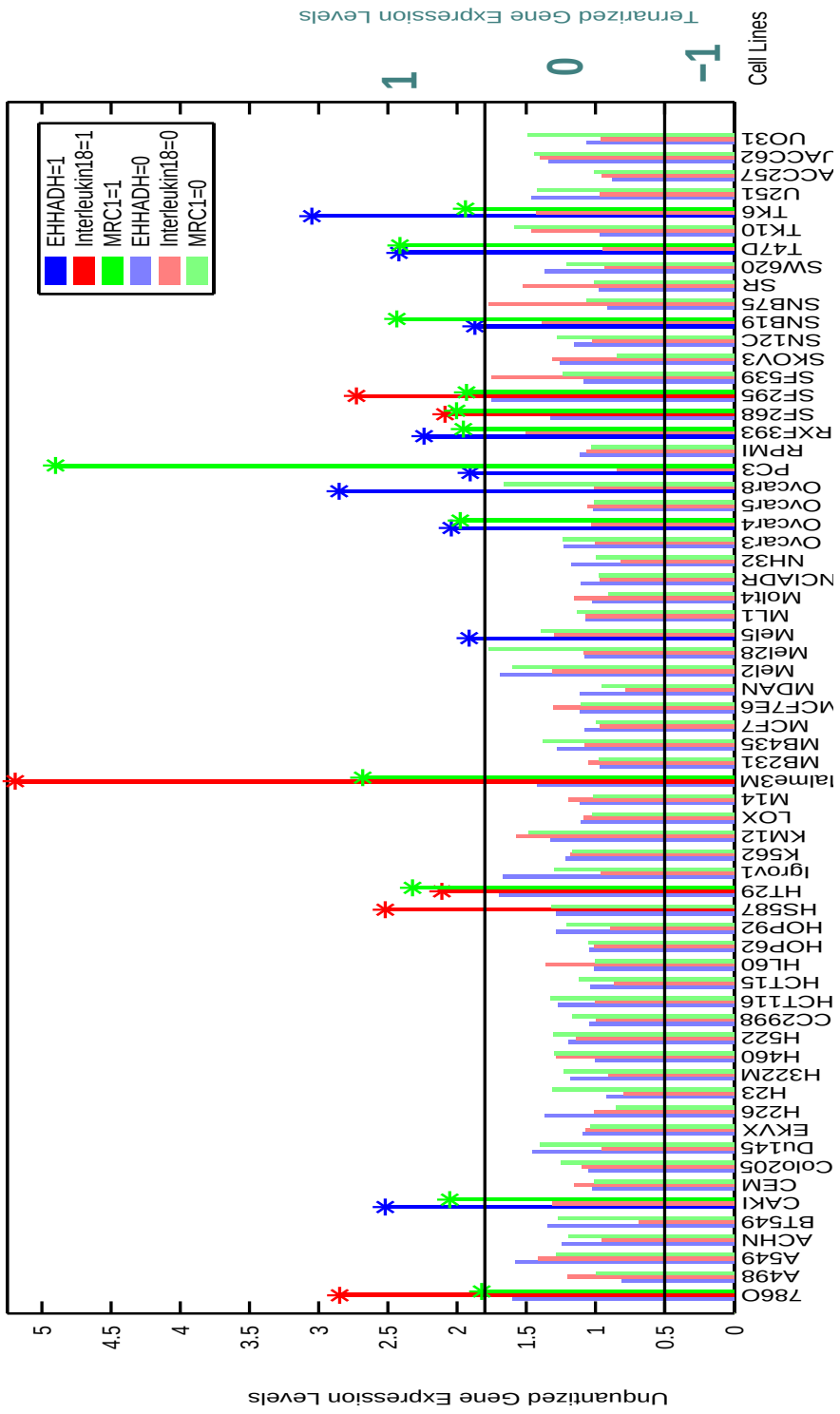

Fig. 9. Exor relationship

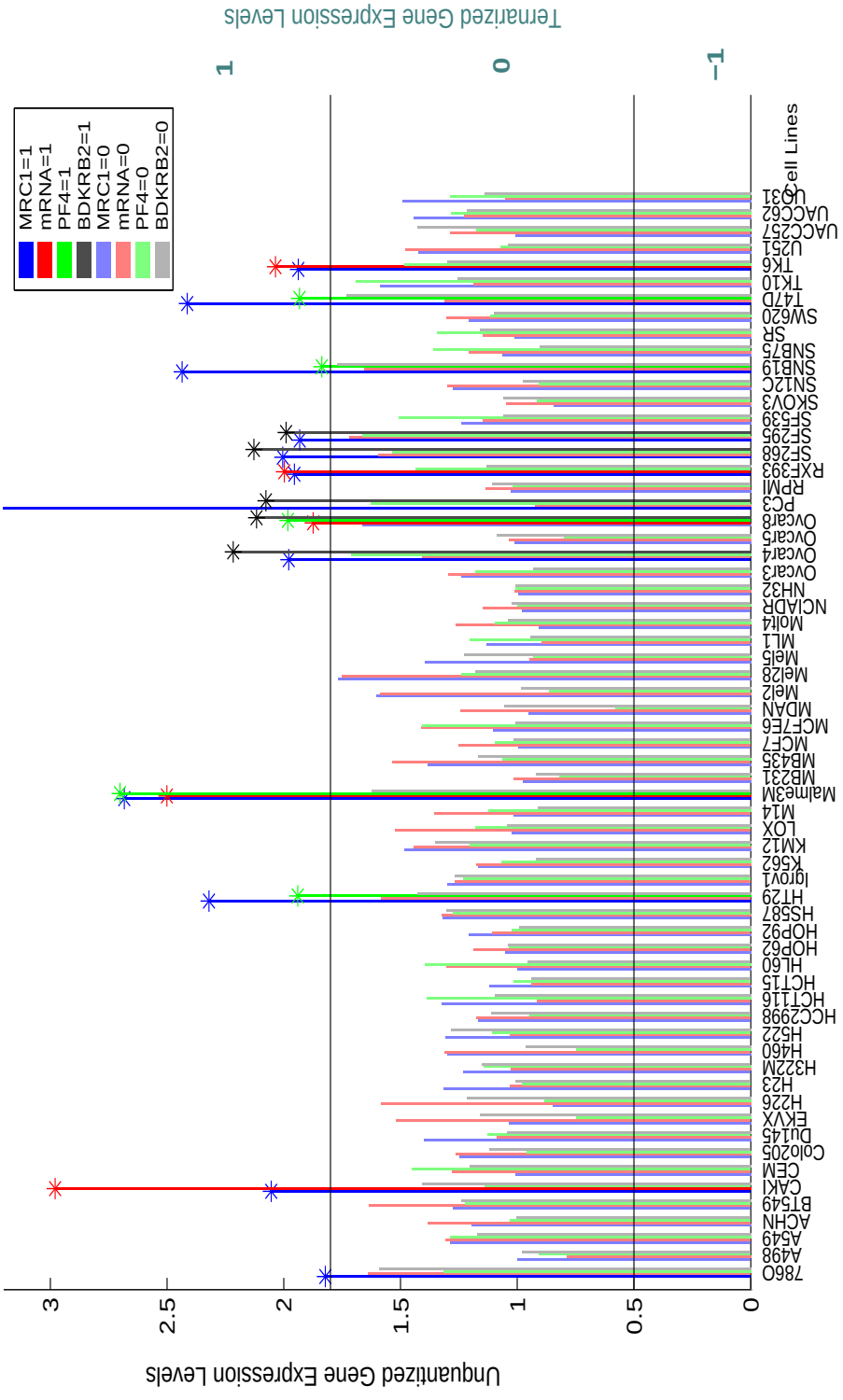

Fig. 10. Relationship among four genes

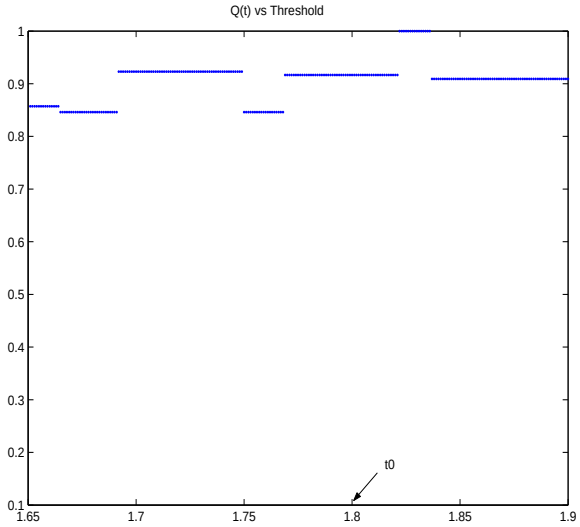

Fig. 11. Q(t) for four gene relationship

Figure 10 shows the relationship among three predictors and a target gene. The red, green, gray and blue bars represent the genes *Homo sapiens clone TCCCTA00151 mRNA sequence* (mRNA), *platelet factor 4* (PF4),*bradykinin receptor B2* (BDKRB2), and the target, *mannose receptor, C type 1* (MRC1). The relationship shows that the predictors up-regulate the target gene when one or two of them are upregulated but when all the predictor genes are upregulated then the target is not induced. The approximate boolean relationship,

MRC1 = (XOR(mRNA,PF4) $\vee$ XOR(mRNA,BDKRB2) $\vee$ XOR(PF4,BDKRB2))

is shown in Table IV. To verify that the same function holds for small changes in the induction threshold, we plot the COD vs Induction Threshold for threshold ranging from 1.65 to 1.9. Figure 11 suggests that the relationship as depicted by cod analysis is quite stable to small changes in induction threshold. We didn't consider the repression threshold for this case as almost all of the values were either ones or zeros.

By examining other CoD values, we find that the genes *interleukin 1, alpha*(IL1A) and *distal-less homeobox 4*(DLX4) seems to be repressors for *adenomatosis polyposis coli*(APC), whereas the gene *zinc finger protein, X-linked*(ZFX) appears to induce APC.

Table IV. Truth Table for Relationship Among 4 Genes

mRNA	PF4	BDKRB2	MRC1
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	0

For another instance of a Boolean relationship between four genes, we consider the genes *albumin*(ALB), EHHADH and *thyroid hormone receptor-associated protein* (TRAP 150) predicting the gene APC. From the data, they predict APC with a CoD of 1. The data also suggest that when gene TRAP150 is repressed, APC cannot be induced. APC is induced when either ALB or EHHADH is induced and Trap150 is not repressed.

Another strong relationship among genes is depicted in Fig. 12, where the red, green, grey and blue bars represent *islet cell autoantigen 1 (69kD)*(ICA1),*syndecan 2 (heparan sulfate proteoglycan 1,cell surface-associated, fibroglycan)*(SDC2), *heterogeneous nuclear ribonucleoprotein L*(HNRPL) and *mannose receptor, C type 1*(MRC1) (target), respectively. This relationship is quite robust with changes in threshold for both repression and induction as shown in Fig. 13. The X axis shows the threshold for induction and the Y axis the threshold for repression while Z axis shows the Q(t) which always stays above .75 for small changes in threshold values.

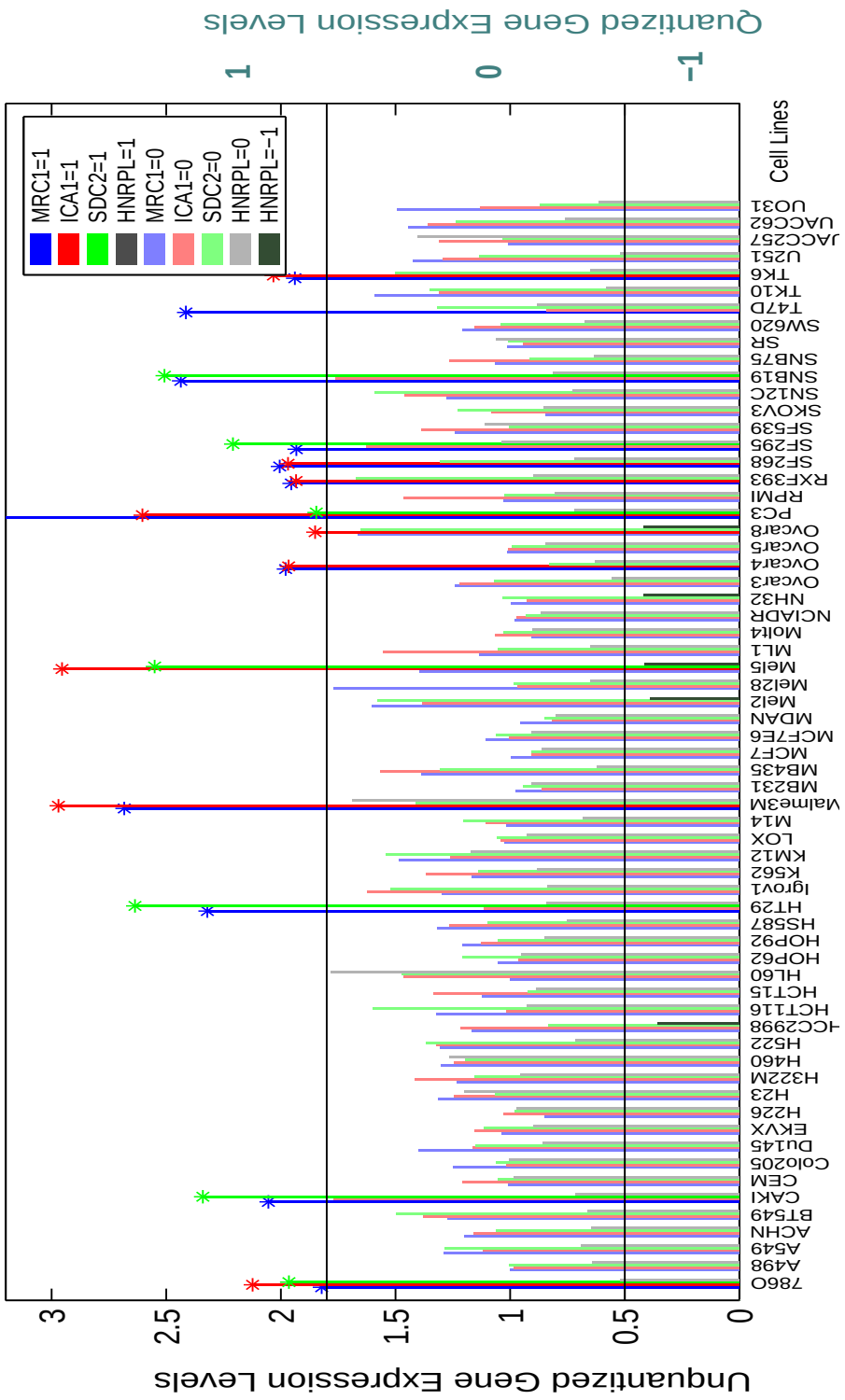

Fig. 12. Another relationship among four genes

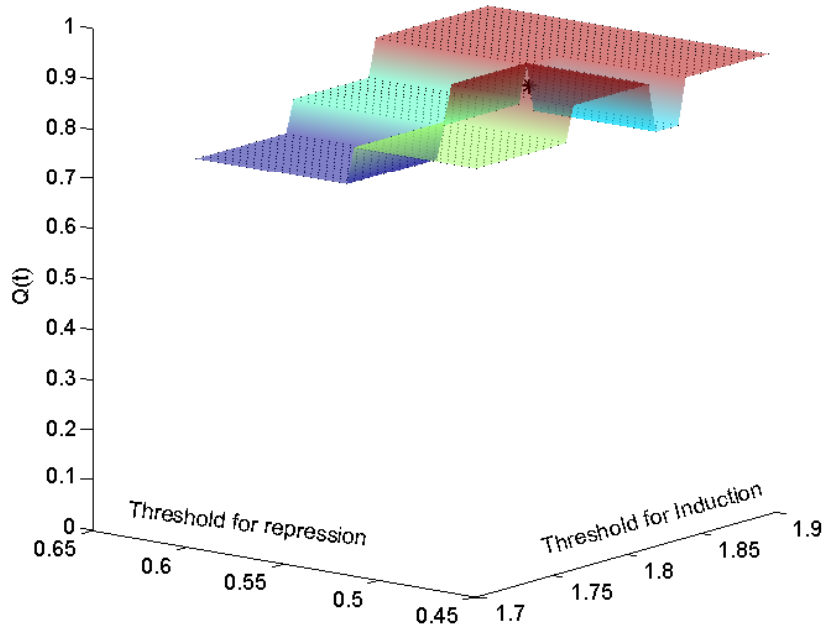

Fig. 13. Q(t1,t2) for second four gene relationship

G. Role of p53 Status in Determining Inter-Gene Relationships

In our data, there are 20 cell lines with functional p53 and 44 cell lines with mutant p53. To study the role of p53 status in determining inter-gene relationships, we have divided the expression data into two sets: one in which p53 is functional and the other in which it is mutant, and we have analyzed both sets separately. We have concluded that there are stronger inter-gene predictive relationships when p53 is functional that when p53 is mutant. Figure 14 shows the probability distribution function of the CoD values. The red and blue curves give the probability distribution functions of the CoD values for the wild type p53 and the mutant p53 cell lines, respectively. We observe that for wild type p53 the average and maximum cod values exceed the corresponding values for mutant p53 cell lines.

We next present a few specific instances of genes whose behavior seems to depend on

the p53 status.

Genes *TATA element modulatory factor 1*(TMF1),*cyclin T2*(CCNT2),*polymerase (DNA directed), delta 3*(POLD3) and *guanylate binding protein 1, interferon-inducible, 67kD* (GBP1) are induced when p53 is not active but they stay dormant when p53 is active. These genes might play an important role when the tumor suppressor (p53) is inactive. On the other hand, some genes show little responsiveness when p53 is mutant and display considerable variability when p53 is wild type. Instances of such genes are *serum-inducible kinase*(SNK), *tumor necrosis factor receptor superfamily*(TNFRSF10C), *UDP glycosyltransferase 2 family, polypeptide B10*(UGT2B10), *properdin P factor, complement*(PFC), ST14, PHLDA3,*damage-specific DNA binding protein 2*(DDB2), XPC and Killer/DR5.

Gene RAD52 homolog is predicted with a COD of 1 by an EST (Moderately similar to *LCP2 Human Lymphocyte Cytosolic protein 2*) when p53 is active but the COD falls to zero when p53 is mutant. Similar relationships exist in the un-quantized gene expression data.

Gene PPM1D is mostly zero when p53 is mutant; however when p53 is active it has large variations. Other genes can predict those variations quite well. The CoD for genes *centromere protein A*(CENPA) and *4-hydroxyphenylpyruvate dioxygenase*(HPD) predicting PPM1D is 0.85 when p53 is active. The strength of this relationship is borne out by the

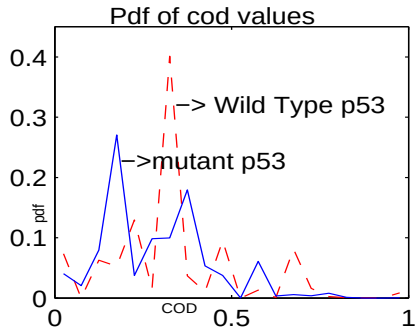

Fig. 14. Pdf of COD values

plot in Figure 15, where red and green curves represent PPM1D and CENPA, respectively. Clearly, there appears to be an inverse relationship between these two genes.

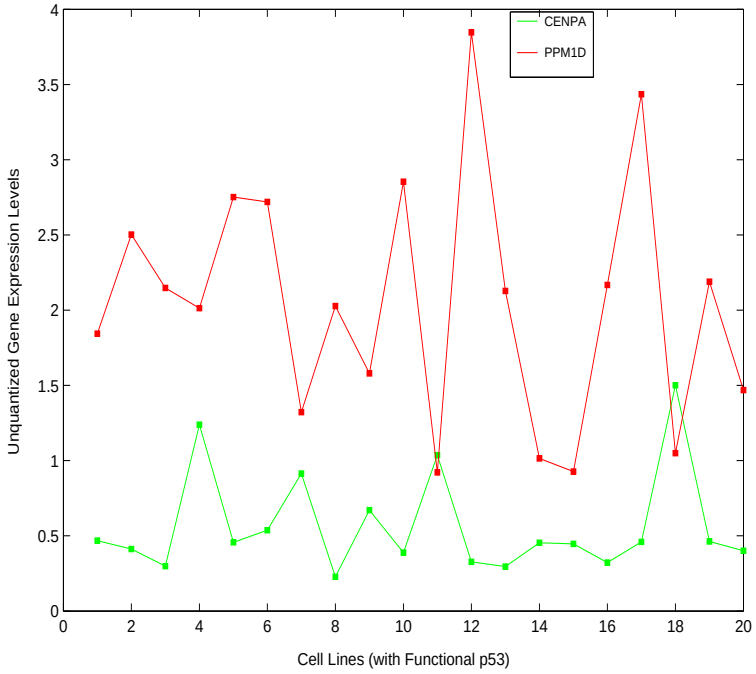

Fig. 15. Cenpa and PPM1D

Our data furnishes many other examples where p53 status is crucial for good prediction. For the gene CENPA, there are many variations when p53 is mutant but these are not predicted strongly by other genes. The best 2-gene predictor gives a CoD of 0.545 when p53 is mutant but when p53 is active the best 2-gene predictor gives a COD of 0.83. Similar relationships hold for gene PPP2R5A where for p53 mutant, the best value of COD is 0.35 whereas the best COD increases to 0.75 when p53 is functional.

H. Conclusion

In this thesis, we have used experimental data to show that Boolean relationships between genes do exist. Many of the current paradigms for modeling genetic regulatory networks

are conditioned on the premise that genes interact with each other through Boolean logic. The results of the current study using cancer cell line data shows that not only do such relationships exist but they can also be unearthed via the recently developed coefficient of determination technique. Another important observation that follows from the data is that several of the relationships unearthed between the different genes seem to be considerably stronger when p53 is functional as compared to when it is not. This is consistent with earlier findings in the literature.

CHAPTER III

CONSTRUCTING BOOLEAN NETWORKS

Random synthetic PBNs can possess a variety of attractor structures.The assumption that biological systems being modeled does not possess cyclic attractors requires us to find out synthetic networks with only singleton attractors.

A. Search Problem

Thus, we are presented with the following search problem: given a set of states earmarked to be singleton attractors and an upper bound on the number of predictors for each gene, find a Boolean network satisfying these conditions that contains no other singleton attractors and no cyclic attractors. There is an implicit consistency requirement for this search: if f predicts the value of gene g and has essential variables $x_1, x_2, ..., x_m$ (which without loss of generality we take to be first m variables) out of the full cohort $x_1, x_2, ..., x_n$, $m < n$, and if $f(a_1, a_2, ..., a_n) = 1$ for some particular set of values of $x_1, x_2, ..., x_n$, then $f(a_1, a_2, ..., a_m, x_{m+1}, ..., x_n) = 1$ for any values of $x_{m+1}, ..., x_n$. An analogous statement holds for $f(a_1, a_2, ..., a_n) = 0$. Here m denotes the number of predictors for gene g. This is a computationally intensive search.

1. Method 1 to Approach This Problem

Consider the task of randomly generating a Boolean Network (BN) of n genes, where each of the genes can not have more than m predictor genes. Furthermore, we assume that the BN to be generated has exactly k singleton attractors, i.e. there are exactly k states in the state space of the BN such that, each one of those states transitions into itself. The search problem is equivalent to finding a consistent BN in the space of BNs satisfying the above

conditions. Here consistency is understood in the sense of finding a BN that has a state transition diagram which is compatible with a given predictor set of the BN. The Predictor set for the BN is the collection of all of the predictor sets for each individual gene in the BN. If we assume that a gene has exactly p predictors, then for each gene there are P= $\binom{n-1}{p}$ possible predictor sets and a total of P^n possible predictors sets for the BN. Here we assume a gene cannot predict itself. For the case where the number of predictors can vary between 1 and p, P = $\sum_{i=1}^{p} \binom{n-1}{i}$. As the gene expressions are assumed to be binary, i.e. 0 or 1, the number of non-attractor states are M=$2^n - k$. We can think of these non-attractor states to be situated in levels where the j-th level signifies that the state reaches one of the singleton attractors in exactly j transitions. Let $N = 2^n$ be the total number of states. The possible number of such structures are $\sum \frac{N!}{k!m(1)!m(2)!...m(l)!}\ m(l-1)^{m(l)}\ m(l-2)^{m(l-1)}$ ....$k^{m(1)}$ where the summation is over all the different choices of $m(1)...m(l)$ satisfying $\sum_{i=1}^{l} m(i)$ = M. This is because if we consider the level l, then it has $m(l)$ states in it and the level just above has $m(l-1)$ states. Each state from level l has to go to one of the states of level $l-1$ and there are $m(l-1)^{m(l)}$ different ways to do that. Similarly for other levels. Once a sequence of $m(1)...m(l)$ has been found satisfying the summation property, then, there are $\frac{N!}{k!m(1)!m(2)!...m(l)!}$ ways to fill the levels. The search space of such structures with this method, even for small n is pretty huge. If we consider $n = 4$, $k = 4$ and number of levels $l = 4$, we have the search space in the range of 10^{15}. For $l = 4$ and for every possible number of attractors, the search space is around 24*10^{16}. In the following sections, we will show that for all possible number of levels, the search space is $N + 1^{N-1}$ which is 17^{15}=$28 * 10^{17}$ for $n = 4$. All these has to be multiplied by P^n to get the actual search space. For all possible levels, search space is equal to the number of graphs of N nodes with no cycle of length more than one and each node having a single directed edge going out from it.

2. Method 2 to Solve the Problem

This method modifies the truth table of the predictor functions and then checks whether the generated BN has only singleton attractors and a reasonable level structure. First, let us select a set of predictors genes P_{1i},P_{2i}.....P_{pi} for each target gene T_i. Then we try to fill up the function matrix (matrix containing the Boolean functions $f(P_{1i}, P_{2i}, ..., P_{pi})$). If we are aiming for a fixed singleton attractor set, then some of the entries of the function matrix is filled beforehand.The rest are filled randomly. Example: Let us consider a case where we have three genes(G_1,G_2,G_3) and two predictors for each target gene. As we have only three genes, the predictors are the other two genes. Thus if we want 001 and 100 to be the attractors then the starting point of the function matrix is as shown in Table V.

Table V. Function Matrix

Gene Values	f1	f2	f3
0 0	1		1
0 1	0	0	
1 0		0	0
1 1			

As the attractors are given to be 001 and 100 , the first gene G_1 will be 0 when the other two genes G_2 and G_3 are 0 and 1 respectively. Similarly from the second attractor 100, we get that gene G_1 is 1 when the other genes are 0. Hence the entry in the first row (corresponding to 00 for the other two genes) and the first column (corresponding to gene G_1) is 1 while the entry in the second row (corresponding to 01) and first column is 0. Similarly for the other two genes.

While filling the truth tables, some cases may arise which are not feasible. For exam-

ple if we consider the attractor set to be 000 and 100. Then for the first attractor 000, we have the function $f(0,0) = 0$, while from the second attractor, we have $f(0,0) = 1$ which is a contradiction. This arises as we have constrained the connectivity. For full connectivity, any structure is feasible. If we had full connectivity then the equations would have been $f(0,0,0) = 0$ and $f(1,0,0) = 1$. We randomly pick another set of attractors if the current set is not feasible. Once a feasible set of attractors is selected, the corresponding entries in the truth table is filled up. The remaining entries are filled randomly with equal probability of picking 1 or 0. After filling up the truth table, we find out the state transitions. The state transitions is an array NS of $N = 2^n$ elements where each entry contains the next transition state. This array is used to find out whether any cycles exist in the graph or not and what is the maximum level of the graph. To find out if any cycle of length more than one exist, we need to start from the first state and traverse to the next states till we reach a state already traversed. If that state is not a singleton attractor, then we have a cycle. Otherwise we move on to the next state which hasn't been traversed yet and start a new path. If the state to which the new path returns is a singleton attractor, then we go to the next non traversed state otherwise we have a cycle. Once a cycle is found the loop is broken and a new random function table is tried. As an example if we have $n = 2$ and there are four states 0,1,2 and 3. Let NS be NS(0)=1, NS(1)=3, NS(2)=1 and NS(3)=3. Then by our algorithm we will start from 0 and go to 1 and from one to 3 and as 3 is a single attractor (NS(3)=3) we move on to the next non traversed state 2 and start a new path. The state 2 transition to state 1 and 1 goes to 3 which is a singleton attractor. Hence this transition set is a possible structure containing only singleton attractors.

A network generated by this method is shown in Fig. 16. This network contains 6 Genes and hence $2^6 = 64$ states. The number of predictors for each gene is 2 and the maximum level for the network is 4.

In this method the search space is much smaller if the connectivity (i.e. the number

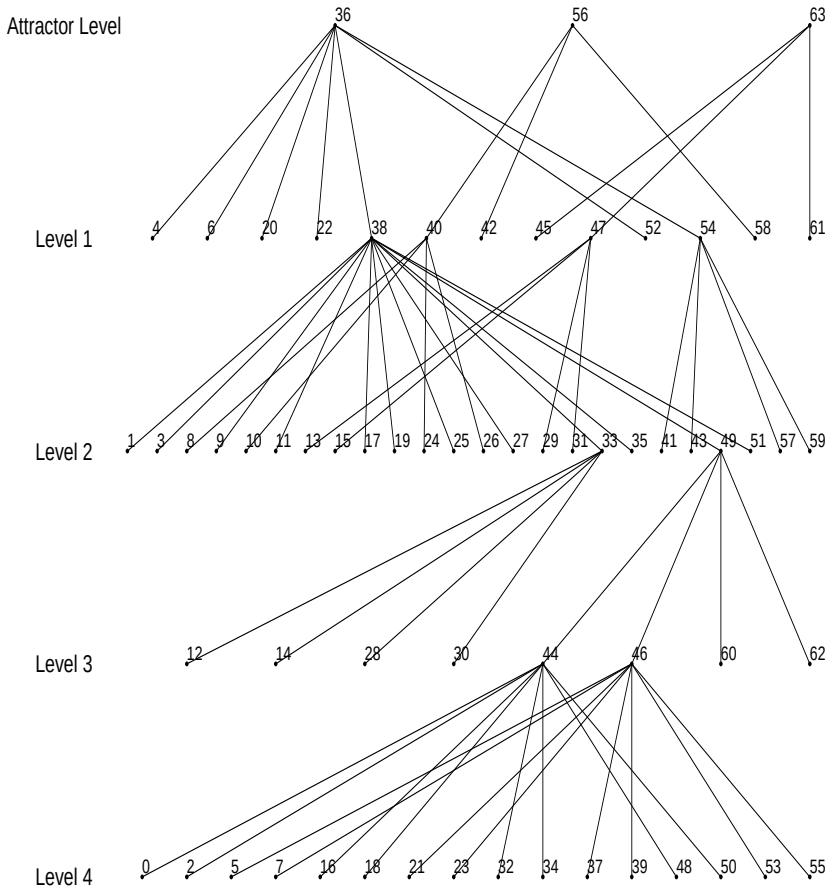

Fig. 16. Synthetic BN with only singleton attractors

of predictors for each gene) is small. In the previous method even if the connectivity was low, the search space will remain almost the same and many non feasible structures will be searched. As an example, if the first gene G_1 is not present in any of the predictor sets of the other genes then the states $0, x_2, ..., x_n$ and $1, x_2, ..., x_n$ should transition to the same state and hence should be at the same level. Here it is assumed that the other genes are equal for the two cases. In the first method, even the cases where they belong to different levels may come up and increase the number of computations unnecessarily. In the second method, this scenario won't come into picture as we generate the transition diagram from feasible truth tables.

The search space in the second method is $P^n 2^{n*2^p}$. The function table has n columns for the n genes and 2^p rows for the different predictor combinations $000...0$ $00....1$ .... $11...1$ as shown in Table VI. This $n * 2^p$ entries (marked x in Table VI) can be filled by 0 or 1. Hence the possible combinations are 2^{n*2^p}.

Table VI. Function Table

Gene Values	f1	f2	.....	fn
0 0..0	x	x	x	x
0 0..1	x	x	x	x
...	..	..	..	..
1 1...1	x	x	x	x

If p=2 and n=4 then $P^n\ 2^{n*2^p}$ is around $5.3 * 10^6$ which is much less than the previous method.

B. Number of Graphs with Only Singleton Attractors

To study analytically the percentage of networks with only singleton attractors and no other cycles, we consider a network build from n Genes (G_1,G_2,..... G_n). The total number of states in such networks is $N = 2^n$. Let us assume that we have full connectivity i.e. the predictors for a Gene G_i are all the available genes (G_1,G_2,..... G_n) including itself. Therefore a transition can occur from any state to any other state. If we reduce the connectivity then all the transitions may not be feasible. Total number of such graphs possible is N^N as a state has N choices and there are N states.

1. Unique Numbering of Graphs with Singleton Attractors

Let us consider the graph in Figure 17 which has only one singleton attractor. We can generate a unique sequence from this graph if we enumerate it as given in [23]. Remove the pendant vertex (and the edge incident on it) having the smallest label, say V_1. Let the state to which V_1 was incident be I_1. Now we do the same thing with the remaining vertices , remove the vertex with the smallest label V_2 and let V_2 be incident on I_2. This process is continued for $N - 1$ steps till we are left with only the singleton attractor. For the given figure the sequence of I_1,I_2,..... I_{N-1} is 15155.

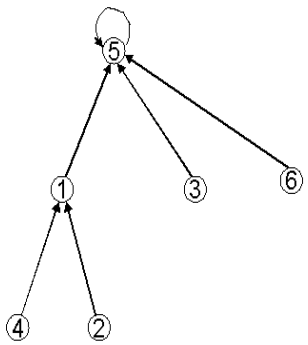

Fig. 17. Example of graph with no cycles

To construct back from a sequence of I_1,I_2,..... I_{N-1} we consider the first number in the sequence 1,2,..... N that does not appear in the I sequence. That number is V_1. Thus we get an edge from V_1 to I_1. Then we remove V_1 from the V sequence and I_1 from the I sequence. Now we do the same thing with the remaining sequences I and V. For the example considered we will get the edges $2->1$, $3->5$, $4->1$, $1->5$,$6->5$ and the last remaining one 5 as the singleton attractor.

The number of different such sequences possible is N^{N-1} as we have N choices for $N-1$ places. Now we consider the cases where there are two singleton attractors. Then if we enumerate them as done previously, we will get $N-2$ numbers in the sequence instead of $N-1$. For a particular choice of singleton atrractors A_1 and A_2 we have other $N-2-1$ places to be filled. So different ways in which we can select a graph with only 2 singleton attractors and no cycles is $\binom{N}{2}$ * 2 * N^{N-3}. The first number is the number of different ways to select two singleton atrractors from N states. The second 2 is due to the number of singleton attractors. The third number is the number of different ways the sequences can be formed with either one of the singleton attractors at the end. It is N^{N-3} and not N^{N-2} because if we fix a singleton attractor as the last number in the sequence then there are $N-3$ positions left to be filled. A point to notice is that the last number of the sequence has to be a singleton attractor.

For three singleton attractors the number becomes $\binom{N}{3}$ * 3 * N^{N-4}.

So the total number of graphs with no cycles of length more than one is

$\sum_{r=1}^{N} \binom{N}{r} r\ N^{N-r-1}$.

As $\binom{N}{r} r\ N^{N-r-1}$ = N^{N-1} $\binom{N-1}{r-1}$ $N^{-(r-1)}$

The sum becomes $N^{N-1} \sum_{r=0}^{N-1} \binom{N-1}{r} N^{-r}$ = N^{N-1} $(1+\frac{1}{N})^{N-1}$ which is equal to

$$(N+1)^{N-1} \tag{1}$$

Therefore the ratio of non-cyclic graphs to the overall number of graphs is

$$\frac{(N+1)^{N-1}}{N^N} \tag{2}$$

which is asymtotically equal to $\frac{e}{N}$.

The point to observe is that with the increase in number of genes, the percentage of graphs without any cyclic attractor set goes on decreasing exponentially on the number of genes n.

Table VII. Full Connectivity Table

Number of Genes	Number of States	Analytical ratio	Simulation Ratio
1	2	.75	.75
2	4	.48	.5
3	8	.285	.28
4	16	.15	.14
5	32	.08	.097
6	64	.041	.035
7	128	.02	.023
8	256	.0105	.013
9	512	.00529	.0051
10	1024	.00265	.0023

Table VII contains the analytic ratio of these kind of structures obtained from the stated formula along with the ratio obtained from simulations. The simulation were done by randomly picking 1000 state transitions and finding out how many of them contain only singleton attractors and no cycles. The ratios obtained from the analytic formula and the

simulations are nearly the same.

Table VIII contains the results of simulations done with different number of genes n and different number of Predictors p. The ratio of structures containing only singleton attractors among all feasible random networks increases when connectivity decreases. It's very difficult to find the exact number of such structures and there is no mention of any such analytic ratio in the literature.

Table VIII.: Low Connectivity Table

n	p	Simulation Ratio	Actual count	Number of Simulations
2	1	0.753800	11307	15000
3	1	0.778867	11683	15000
3	2	0.433933	6509	15000
4	1	0.802800	12042	15000
4	2	0.439733	6596	15000
4	3	0.222333	3335	15000
5	1	0.800200	12003	15000
5	2	0.425400	6381	15000
5	3	0.192200	2883	15000
5	4	0.105800	1587	15000
6	1	0.803600	12054	15000
6	2	0.410400	6156	15000
6	3	0.161867	2428	15000
6	4	0.072067	1081	15000
6	5	0.050400	756	15000
7	1	0.811533	12173	15000

Table VIII Continued

n	p	Simulation Ratio	Actual count	Number of Simulations
7	2	0.388000	5820	15000
7	3	0.132467	1987	15000
7	4	0.051867	778	15000
7	5	0.026400	396	15000
7	6	0.023133	347	15000
8	1	0.814067	12211	15000
8	2	0.382067	5731	15000
8	3	0.117400	1761	15000
8	4	0.040267	604	15000
8	5	0.018067	271	15000
8	6	0.012133	182	15000
8	7	0.009267	139	15000
9	1	0.812067	12181	15000
9	2	0.362600	5439	15000
9	3	0.107800	1617	15000
9	4	0.028133	422	15000
9	5	0.011600	174	15000
9	6	0.007267	109	15000
9	7	0.006000	90	15000
9	8	0.005733	86	15000
10	1	0.809400	12141	15000
10	2	0.361800	5427	15000
10	3	0.092067	1381	15000

Table VIII Continued

n	p	Simulation Ratio	Actual count	Number of Simulations
10	4	0.018867	283	15000
10	5	0.009067	136	15000
10	6	0.003400	51	15000
10	7	0.003133	47	15000

C. Conclusion

The singleton attractor structures are rare when connectivity is high as shown in (2) but they become common when the connectivity is decreased. The analytic ratio of these kind of structures for low connectivity is yet to be determined and is a future topic for research. The second method of finding these structures has a search space much lower than the first method and can be used to generate singleton attractor structures with reasonable complexity for less than 20 genes. But the function table method cannot give us full control on the maximum and minimum levels of the generated network as the first method. A combination of these two methods which can work efficiently for high number of genes and have more control on the level structure is a topic for further research.

CHAPTER IV

CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

We have used experimental cancer cell lines data in this thesis to show that Boolean relationships between genes do exist and can be discovered by using Coefficient of Determination Technique. Many of the current models of Genetic regulatory Networks are based on the idea of Genes interacting between themselves through Boolean Logic. The results reported in this thesis have laid the groundwork for further studies using the NCI 60 ACDS. A promising research direction seems to be to use the gene expression data to construct a Probabilistic Boolean Network which could then be used to design and evaluate possible intervention strategies for cancer treatment.

The other issue which this thesis covers is the generation of synthetic Boolean Networks with singleton attractors. The less intensive method gives us a way to construct such singleton attractor structures and use them in testing of algorithms for finding out connectivity of networks. The analytical and emperical studies on the frequency of such networks shows the scarcity of such networks for high connectivity. The analytical results for low connectivity is yet to be discovered and is a future research topic. The other area which needs to be researched is the efficient generation of such networks with large number of genes.

REFERENCES

[1] T. A. Brown, *Genetics: A Molecular Approach.* London, UK: Chappman & Hall, chap. 1-8, pp. 3-124, 1998.

[2] P. Smolen, D. Baxter, and J. Byrne, "Mathematical modeling of gene networks," *Neuron*, vol. 26, pp. 567-580, 2000.

[3] Kauffman, S. A., *The Origins of Order: Self-organization and Selection in Evolution*, New York: Oxford University Press, 1993.

[4] R. Somogyi and C. A. Sniegoski, "Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation," *Complexity*, vol. 1, pp. 45-63, 1996.

[5] Z. Szallasi and S. Liang, "Modeling the Normal and Neoplastic Cell Cycle With Realistic Boolean Genetic Networks: Their Application for Understanding Carcinogenesis and Assessing Therapeutic Strategies," *Pacific Symposium on Biocomputing*, vol. 3, pp. 66-76, 1998.

[6] S. M. Thomas,P. Soriano, and A Imamoto, "Specific and redundant roles of Src and Fyn in organizing the cytoskeleton," *Nature*, vol. 376, pp. 267-271, 1995.

[7] Shmulevich, I., E.R. Dougherty, and W. Zhang, "From Boolean to probabilistic Boolean networks as models of genetic regulatory networks," *Proceedings of the IEEE*, vol. 90, no. 11, 1778-1792, 2002.

[8] McAdams, H. H. and L. Shapiro, "Circuit simulation of genetic networks," *Science*, vol. 269, pp. 650-656, 1995.

[9] Yuh C-H, H. Bolouri, and E. H. Davidson, “Genomic cis-regulatory logic: Experimental and computational analysis of a sea urchin gene,” *Science*, vol. 279, pp. 1896-1902, 1998.

[10] Dougherty, E. R., S. Kim, and Y. Chen, “Coefficient of determination in nonlinear signal processing,” *Signal Processing*, vol. 80, no. 10, 2219-2235, 2000.

[11] Kim, S., 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*, vol. 4, no. 4, 411-424, 2000.

[12] Kim, S., E. R. Dougherty, Y. Chen, K. Sivakumar, P. Meltzer, J.M. Trent, and M. L. Bittner, “Multivariate measurement of gene-expression relationships,” *Genomics*, vol. 67, 201-209, 2000.

[13] Chen, Y., V. Kamat, E. R. Dougherty, M. L. Bittner, P. S. Meltzer, and J. M. Trent, “Ratio statistics of gene expression levels and applications to microarray data analysis,” *Bioinformatics*, vol. 18, 1207-15, 2002.

[14] N. Friedman, M. Linial, I. Nachman, and D. Pe’er, “Using Bayesian network to analyze expression data,” *Journal of Computational Biology*, vol. 7, pp. 601-620,2000.

[15] Shmulevich, I., E. R. Dougherty, S. Kim, and W. Zhang, “Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks,” *Bioinformatics*, vol. 18, 261-274, 2002.

[16] Shmulevich, I., E. R. Dougherty, and W. Zhang, “Gene perturbation and intervention in probabilistic Boolean networks,” *Bioinformatics*, vol. 18, 1319-1331, 2002.

[17] Shmulevich, I., E. R. Dougherty, and W. Zhang, “Control of Stationary Behavior in probabilistic Boolean networks by Means of Structural Intervention,” *Biological*

Systems, vol. 10, no. 4, 431-446, 2002.

[18] Datta, A., A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in Markovian genetic regulatory networks," *Machine Learning*, vol. 52, 169-191 ,2003.

[19] Datta, A., A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in Markovian genetic regulatory networks: The imperfect information case," *Bioinformatics*, vol. 20, no. 6, 924-930, 2004.

[20] Zhou, X., X. Wang, R. Pal, I. Ivanov, M. Bittner, and E. R. Dougherty, "A Bayesian connectivity-based approach to constructing probabilistic gene regulatory networks," *Bioinformatics*,in press. 2004.

[21] Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, 6, 520-525, 2001.

[22] Zhou, X., X. Wang, and E. R. Dougherty, "Missing-value estimation using linear and non-linear regression with Bayesian gene selection," *Bioinformatics*, 19, 2302-2307, 2003.

[23] Deo, N., *Graph Theory with Applications to Engineering and Computer Science*.Englewood Cliffs, NJ: Prentice-Hall, 1974.

VITA

Ranadip Pal was born on June 7, 1980 in West Bengal, India. He received his Bachelor of Technology in electronics and electrical communication engineering from Indian Institute of Technology (IIT), Kharagpur in June 2002. In September 2002, he started his M.S. in electrical engineering at Texas A&M University. His main area of research is genomic signal processing. His address is Department of Electrical Engineering, Texas A&M University, College Station, Texas, 77843.