PI CONTROL OF GENE EXPRESSION IN TUMOROUS CELL LINES

A Thesis

by

ROUELLA JOAN MENDONCA

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

May 2009

Major Subject: Electrical Engineering

PI CONTROL OF GENE EXPRESSION IN TUMOROUS CELL LINES

A Thesis

by

ROUELLA JOAN MENDONCA

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

Approved by:

Chair of Committee,	Aniruddha Datta
Committee Members,	Edward R. Dougherty
	Deepa Kundur
	Guy Battle
Head of Department,	Costas N. Georghiades

May 2009

Major Subject: Electrical Engineering

ABSTRACT

PI Control of Gene Expression in Tumorous Cell Lines. (May 2009)

Rouella Joan Mendonca, B.S., Texas A&M University

Chair of Advisory Committee: Dr. Aniruddha Datta

Recent experiments are bringing to the fore more and more information about the effects of different treatments on the gene expression of different genes. The results obtained from these experiments show that some definite trends are observed in different genes in the Human Embryonic Kidney and Human Colon Adenocarcinoma Grade II cell lines. The difference in the gene expressions of the two cell lines motivates the problem in this thesis. The thesis provided intervention methods to make the colon cancer cell line genes behave more like their Human Embryonic Kidney cell line counterparts. Two methods of intervention were introduced. The first method was the simpler on-off control intervention while the second method used a more advanced proportional integral control to meet the goal. A comparison of these two intervention methods showed the clear implementational advantages of proportional integral control over on-off control.

To my parents for their love, support, and encouragement

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Datta for his guidance, support, and assistance throughout the course of my study. I would also like to thank my committee members, Dr. Dougherty, Dr. Kundur, and Dr. Battle, for their guidance and support throughout the course of this research.

I also want to extend my gratitude to Dr. Ranadip Pal of Texas Tech University for his assistance, as well as the Translational Genomics Research Institute for providing me with the data that I used in my research. Thanks also go to my friends and colleagues and the departmental faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to my family for their encouragement and continuous support.

NOMENCLATURE

cDNA	Complementary Deoxyribonucleic Acid
CoD	Coefficient of Determination
DNA	Deoxyribonucleic Acid
FBS	Fetal Bovine Serum
GAP	Gene Activity Profile
GFP	Green Fluorescent Protein
HEK	Human Embryonic Kidney
HT29	Human Colon Adenocarcinoma Grade II
mRNA	Messenger Ribonucleic Acid
PBN	Probabilistic Boolean Network
PI	Proportional Integral
RNA	Ribonucleic Acid

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. INTRODUCTION

1.1 Problem Statement and Overview

Data obtained from recent experiments involving the addition of fetal bovine serum (FBS) to cells and the starvation of cells shows that there is a large difference in the gene expression of the Human Embryonic Kidney (HEK) cell line and the Human Colon Adenocarcinoma Grade II (HT29) or colon cancer cell line. The purpose of this research is to make the colon cancer cell line behave more like the near normal HEK cell line using intervention by addition and starvation of FBS to the genes.

To date, optimal intervention has been studied in the context of probabilistic Boolean networks (PBNs). A PBN is a collection of Boolean networks in which one constituent network governs gene activity for a random period of time before another randomly chosen constituent network takes over, possibly in response to some random event, such as an external stimulus or genes not included in the model network [1].

Major efforts have focused on manipulating external (control) variables to desirably affect dynamical evolution over a finite time horizon. These short-term policies, however, are not necessarily effective in changing long-run network behavior. To address this, stochastic control has been employed via dynamic programming algorithms to find stationary control policies that affect the steady state distributions of

---

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

PBNs [2]. Infinite horizon intervention strategies and mean first passage time intervention policies are the most recent algorithms. The important role of PBNs in current genomic research motivates its use in the selection of states for treatment modeling in this thesis.

The data that is used in this study is obtained through a lengthy experimental procedure that makes use of Green Fluorescent Protein (GFP) based promoter-reporter technology. Each type of cell (HEK and HT29 cell lines) is divided into three groups: the first is exposed to 10% FBS solution, the second group undergoes starvation after the 10% FBS exposure, and the third group is exposed to 20% FBS after the starvation. The response of the three genes – JUN, MYC, and EGR1 in the first group in the HEK cell line is used as a reference that we want the corresponding controlled colon cancer cell line genes to follow.

In order to apply intervention, modeling of the treatments is the first step. The addition of 20% FBS to the HT29 genes and starvation of the same genes constitute the two treatments to be used as inputs to the system. Modeling is done using parallels to probabilistic Boolean networks and by studying the behavior of HT29 genes under the different treatments. Once modeling is complete, intervention can be applied using the FBS addition and starvation models.

The first intervention method studied is that of on-off control. In this method of control, a control band is defined. When the output exceeds the top of the control band, the starvation treatment is applied. When the output moves below the bottom of the control band, the FBS addition treatment is used.

The second intervention method studied is that of proportional integral control. As there are only two treatments that can be applied, continuous variation of FBS is not possible. However, dividing time into small divisions simulates continuous variation of time. PI control is then used to determine when the system should be given the starvation treatment, and when the system should be given the FBS addition treatment.

A comparison of the two intervention methods shows the advantages of one over the other. The next section gives an outline of the thesis.

1.2 Outline of Thesis

The thesis begins with an introduction to tools that are helpful in explaining the study. These tools include a short description of PBNs followed by a description of the GFP based promoter-reporter technology. The document then proceeds to describe in detail the experimental procedure used to obtain the data that is used in the study. This is then followed by a description of treatment modeling. Finally, after describing the intervention techniques of on-off control and PI control, the two control methods are compared to show the advantages of one over the other.

2. BACKGROUND

This section briefly describes genetic regulatory networks. Such a description facilitates better understanding of treatment modeling which will be taken up in the next section. This is followed by a detailed description of the GFP based promoter-reporter technology, the experimental procedures used to obtain the data, and a description of the previously proposed on-off control, which is unfortunately not implementable.

2.1 Genetic Regulatory Networks

A central focus of genomic research concerns understanding the manner in which cells execute and control the enormous number of operations required for normal function and the ways in which cellular systems fail in disease. Modeling and analysis of gene regulation can substantially help to unravel the mechanisms underlying gene regulation and to understand gene function. This, in turn, can have a profound effect on developing techniques for drug testing and therapeutic intervention for effective treatment of disease [3] – [6]. Two salient aspects of a genetic regulatory system must be modeled and analyzed. One is the topology and the other is the set of interactions between the elements [7].

In order to explain Probabilistic Boolean Networks (PBNs), we first introduce its precursor, the so-called Boolean Networks. The regulatory network that has perhaps received the most attention is the Boolean network model [8] – [11]. In the Boolean

model, gene expression is quantized to two levels: ON and OFF. The expression level or state of each gene is functionally related to the expression states of other genes using logical rules. Other terminology sometimes used instead of 'on' and 'off' are -'up-regulated' and 'down-regulated', or 'responsive' and 'non-responsive' respectively [7]. This can be extended to a ternary network by adding the category of invariant.

A Boolean network is defined by a set of nodes, $V = \{x_1, x_2, ..., x_n\}$ and a list of Boolean functions, $F = \{f_1, f_2, ..., f_n\}$. Each x_k represents the state (expression) of a gene g_k, where $x_k = 1$ or $x_k = 0$, depending on whether the gene is expressed or not expressed. The Boolean functions represent the rules of regulatory interaction between genes. The value of gene g_k at time $t+1$ is determined by

$$x_k(t+1) = f_k(x_{k,1}, x_{k,1}, ..., x_{k,m(k)}) \quad (1)$$

where the nodes in the argument of f_k form the regulatory set for x_k(gene g_k) [7].

At time t, the state vector $x(t) = (x_1(t), x_2(t), ..., x_n(t))$ is called the gene activity profile (GAP). The functions together with the regulatory sets determine the network wiring. Attractors play a key role in Boolean networks. Given a starting state, within a finite number of steps, the network will transition into a cycle of states called an attractor, after which absent perturbation will continue to cycle thereafter. Each attractor is a subset of a basin composed of those states that lead to the attractor if chosen as starting states [7]. The attractors of a Boolean network characterize the long run behavior of the network and are indicative of the cell type and phenotypic behavior of the cell [11].

To establish multivariate relationships among genes, the coefficient of determination (CoD) quantifies how the estimate for the expression status of a particular target gene can be improved by knowledge of the status of some other predictor genes [12]. The CoD is defined by

$$CoD = \frac{\varepsilon_0 - \varepsilon_{opt}}{\varepsilon_0} \tag{2}$$

where,

ε_0 is the error of the best numerical predictor of the target gene in the absence of observation.

ε_{opt} is the error of the optimal predictor of the target gene based on predictor genes.

The CoD 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 predictor genes, relative to the best estimate in the absence of any knowledge of the transcriptional activity of the predictors. The CoD is a number between 0 and 1, a higher value indicating a tighter relationship [7].

Given a target gene, several predictor sets may provide equally good estimates of its transcriptional activity, as measured by the CoD. Moreover, one may rank several predictor sets via their CoDs. For a particular target gene, a good approach is to consider multiple predictor sets with high CoDs. One can compute the probability that the target gene will be transcriptionally active at time $t+1$ based on the gene activity profile at time t. The time evolution of the GAP then defines a stochastic dynamical system. In a PBN, the transcriptional activity of each gene at a given time point is a Boolean function

of the transcriptional activity of the elements of its predictor sets at the previous time point [13] – [14]. The choice of Boolean function and predictor set can vary randomly from one time point to another [7].

A PBN is a collection of Boolean networks in which one constituent network governs gene activity for a random period of time before another randomly chosen constituent network takes over, possibly in response to some random event, such as an external stimulus or genes not included in the model network. A PBN is composed of a set of n genes, $x_1, x_2, ..., x_n$ and a set of vector-valued network functions $\mathbf{f}_1$, $\mathbf{f}_2$, … ,$\mathbf{f}_r$, governing the state transitions of the genes. To every node x_i, there corresponds a set

$$F_i = \{f_j^{(i)}\}_{j=1,...,l(i)} \tag{3}$$

where each $f_j^{(i)}$ is a possible function, called a predictor, determining the value of gene x_i and $l(i)$ is the number of functions assigned to gene x_i [7].

Each network function is of the form $f_k = (f_{k1}^{(1)}, f_{k2}^{(2)}, ..., f_{kn}^{(n)})$ for k=1, …, r, $1 \le k_i \le l(i)$ and where $f_{ki}^{(i)} \in F_i (i = 1,2,...,n)$. Each vector function $\mathrm{f}_k : \{0,1\}^n \to \{0,1\}^n$ acts as a transition function (mapping) representing a possible realization of the entire PBN. Thus, given the value of all genes $(x_1, ..., x_n)$, $\mathrm{f}_k(x_1, x_2, ..., x_n) = (x_1', x_2', ..., x_n')$ gives us the state of the genes after one step of the network given by the realization $\mathbf{f}_k$. At each time point, a random decision is made whether to switch the network function for the next transition with a probability *q* of a change. If a decision is made to change 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 $\mathbf{f}_k$ being the selection probability c_k [7].

PBNs are commonly used in genomic research today to model cell behavior and interactions. It is because of this, that the treatments in this thesis are modeled using two states, the first state representing the ‘on’ state and the second state representing the ‘off’ state. More details on this will be presented in this thesis. The next section describes green fluorescent protein based promoter-reporter technology, which is used in obtaining the data used in this thesis.

2.2 GFP Based Promoter-Reporter Technology

Cellular control results from multivariate activity among cohorts of genes and their products. Since all three levels in the central dogma – Deoxyribonucleic Acid (DNA), Ribonucleic Acid (RNA), and protein – interact, it is not possible to fully separate them, and ultimately information from all realms must be combined for a full understanding; nevertheless, the high level of interactivity between levels insures that a significant amount of the system information is available in each of the levels, so that focused studies provide useful insights. Past efforts have focused at the RNA level owing to measurement considerations. A central aspect of RNA-based genomic analysis is measurement of the transcriptome, which is the collection of messenger Ribonucleic Acid molecules (mRNAs) in a cell at a given moment. High-throughput technologies make it possible to simultaneously measure the RNA abundance of tens of thousands of mRNAs [7].

In particular, expression microarrays result from a complex biochemical-optical system incorporating robotic spotting and computer image formation. These arrays are grids of thousands of different single-stranded DNA molecules attached to a surface to serve as probes. Two major kinds are those using synthesized oligonucleotides and those using spotted complementary Deoxyribonucleic Acid molecules (cDNAs). The basic procedure is to extract RNA from cells, convert the RNA to single-stranded cDNA, attach fluorescent labels to the different cDNAs, allow the single stranded cDNAs to hybridize to their complementary probes on the microarray, and then detect the resulting fluor-tagged hybrids via excitation of the attached fluors and image formation using a scanning confocal microscope [7]. Relative RNA abundance is measured via measurement of signal intensity from the attached fluors. This intensity is obtained by image processing and statistical analysis [15].

Expression microarrays are extensively used in bioinformatics and genomic signal processing. Due to focus on cancer patients, past efforts have concentrated on steady-state microarray data. Despite their usefulness for genomewide screenings, measuring gene activity using microarrays has a number of limitations. Firstly, microarray technology is destructive in the sense that the relevant cells have to be ruptured in order to extract the mRNA and, therefore, the cells are no longer available for subsequent measurements that would be required if one were interested in generating time-course data on the same cohort of cells. Secondly, the microarray technology measures the average expression level across the tissue from which the mRNA is extracted and, therefore, does not permit the measurement of gene expression levels of

individual cells. To overcome these limitations, the Translational Genomics Research Institute (TGen) is adapting a fluorescent protein based promoter-reporter technology that can be applied to living cells for monitoring gene expression patterns for a set of genes. The primary focus is not on genome-wide gene discovery but on monitoring the expressions over time of a set of preselected genes believed to be integral to a biological process of interest.

Green Fluorescent Protein (GFP) is a protein originally discovered in jellyfish and having the property that it emits green fluorescence. The key idea behind the technology is the following. For a gene to express, it is necessary that an RNA polymerase bind to the promoter region of the gene, which itself is a DNA sequence. Using genetic engineering methods, it is possible to insert the gene that codes for the GFP under the control of the same promoter that controls the transcription of the gene of interest [16]. In such a genetically engineered cell, whenever the gene of interest is transcribed, the GFP will also be transcribed (and translated), so that we can detect the expression level of the gene by measuring the intensity of fluorescence of the green fluorescent protein. TGen has utilized this approach in concert with a lentiviral delivery system for the reporters and a microscope system (InCell 3000, General Electric) that allows rapid and repetitive fluorescent data capture from cells in multi well plates, thereby making it possible to follow changes in transcription activity for a moderate number of reporters in parallel.

The above GFP based promoter-reporter technology is used in the experimental procedure next described. This provides the time-course data that is used in this thesis.

2.3 Experimental Procedure

The behavior of the genes: JUN, MYC, and EGR1 in the Human Embryonic Kidney (HEK) and the Human Colon Adenocarcinoma Grade II (HT29) cell lines is studied. Intensities of Green Fluorescent Protein (GFP) are measured at 20-minute intervals according to the following experimental design.

The cells are divided into three separate groups on which distinct operations are conducted. The flowchart in Figure 1 and plate layout table in Table 1 below depict the experimental design. A description of the flowchart is as follows: a plate with all the cells in a particular group is first placed in a tissue culture (TC) incubator before being moved to a Scottsdale facility. Here after being kept overnight in the TC incubator, the cells are moved to the InCell 3000 microscope in order that images may be taken. All the cells are now treated with a 10% FBS solution. Images are taken for the next two hours before the first and second groups of cells are starved of the serum. Images of the starved cells and not starved group 3 cells are taken for the following 6 hours. The cells of group 1 are then treated with a 20% FBS solution. Images of the three groups – group 1 treated with 10% FBS, starved, and treated with 20% FBS; group 2 treated with 10% FBS and then starved; and group 3 treated only with 10% FBS are then taken for the next 24 hours.

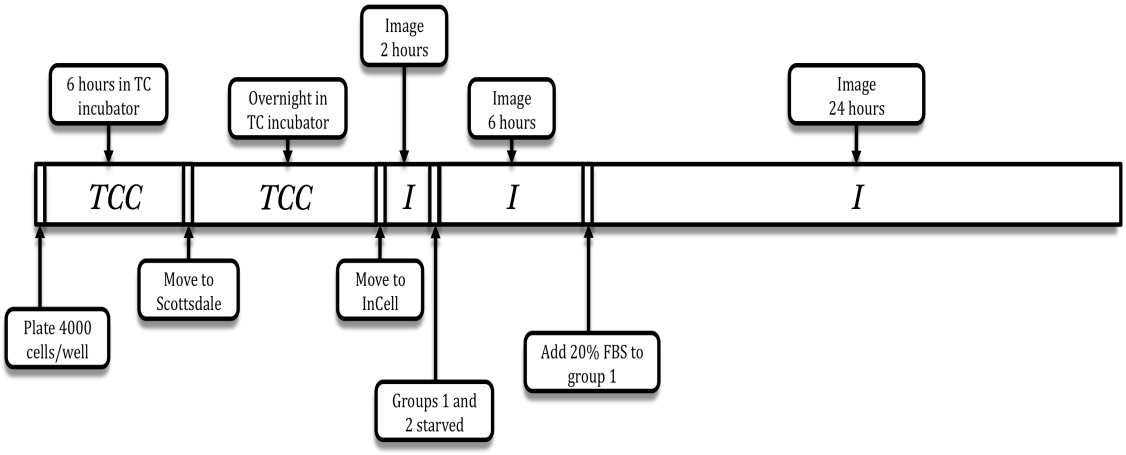


Figure 1: Experimental procedure flowchart.

TABLE 1

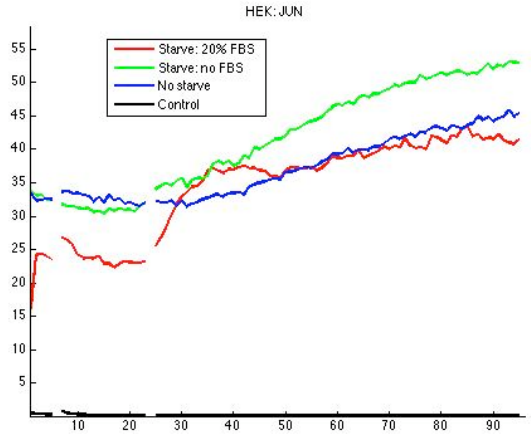
TABLE DEPICTING PLATE LAYOUT

	HEK						HT29 (Colon Cancer)						2 - 8 hours	8 - 32 hours
	1	2	3	4	5	6	7	8	9	10	11	12	Treatment 1	Treatment 2
A	JUN	JUN	JUN	JUN	Con	Con	JUN	JUN	JUN	JUN	Con	Con	Starve	Add 20% FBS Starved/Fed
B	MYC	MYC	MYC	MYC	Con	Con	MYC	MYC	MYC	MYC	Con	Con		
C	EGR1	EGR1	EGR1	EGR1	Con	Con	EGR1	EGR1	EGR1	EGR1	Con	Con		
D	JUN	JUN	JUN	JUN	Con	Con	JUN	JUN	JUN	JUN	Con	Con		No FBS Starved
E	MYC	MYC	MYC	MYC	Con	Con	MYC	MYC	MYC	MYC	Con	Con		
F	EGR1	EGR1	EGR1	EGR1	Con	Con	EGR1	EGR1	EGR1	EGR1	Con	Con		
G	JUN	MYC	EGR1		Con	Con	JUN	MYC	EGR1		Con	Con	Don't starve	Don't starve
H	JUN	MYC	EGR1		Con	Con	JUN	MYC	EGR1		Con	Con		

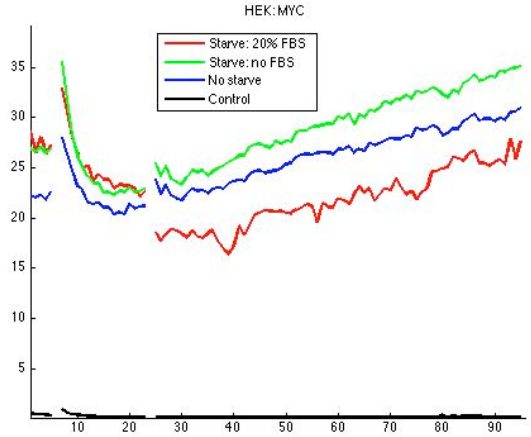
The experimental data is as follows:

- The first six data points correspond to cells cultured at 10% FBS for all three groups.
- The next 18 data points correspond to the starvation of serum for groups 1 and 2 and no treatment for group 3.
- The next 72 data points correspond to 20% FBS addition for group 1, continuation of starvation for group 2 and no treatment on group 3.

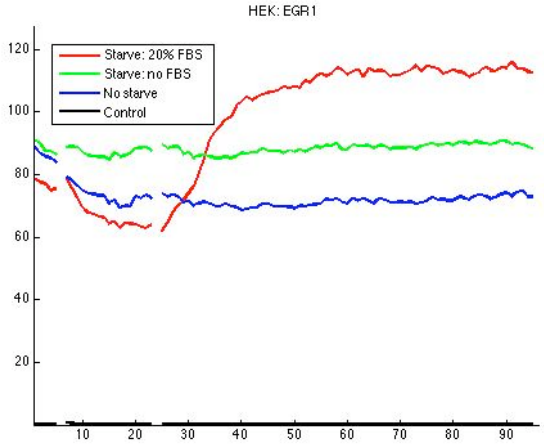
As the first six data points correspond to the same treatment for all three groups, they are used for normalization purposes. Figures 2 and 3 depict the original data for the three genes in the HEK cell line and HT29 colon cancer cell lines respectively. Figures 4 and 5 show the normalized (with respect to the first six data points) expression data for the three genes in HEK and HT29 respectively.



a) HEK: JUN before normalization.

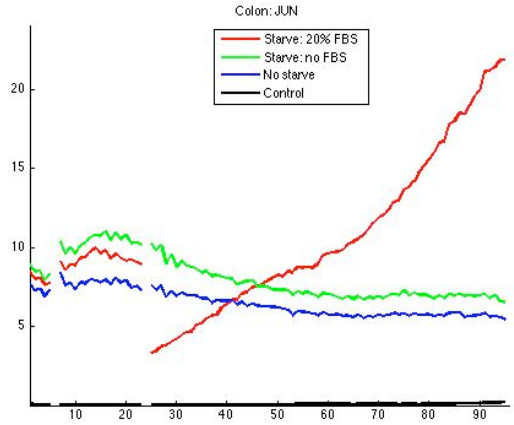


b) HEK: MYC before normalization.

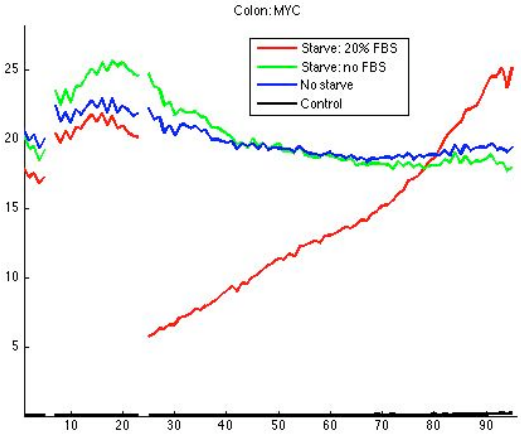


c) HEK: EGR1 before normalization.

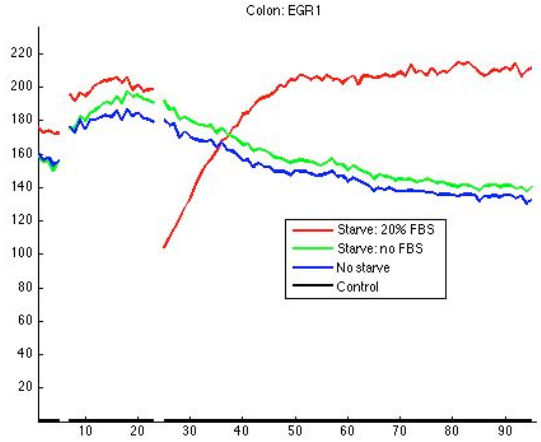
Figure 2: HEK cell line data for starved and then fed 20%FBS group (Group 1), starved with no FBS group (Group 2), and no starvation group (Group 3) for the genes JUN, MYC, and EGR1.



a) HT29: JUN before normalization.

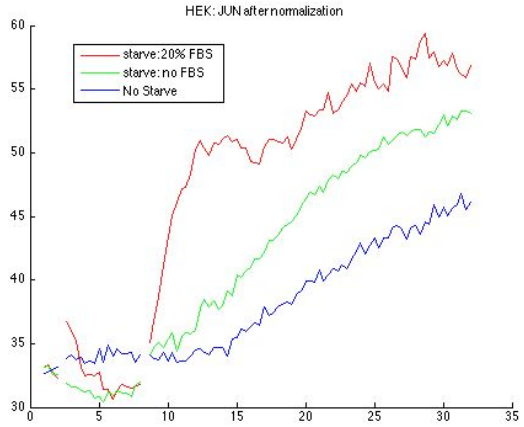


b) HT29: MYC before normalization.

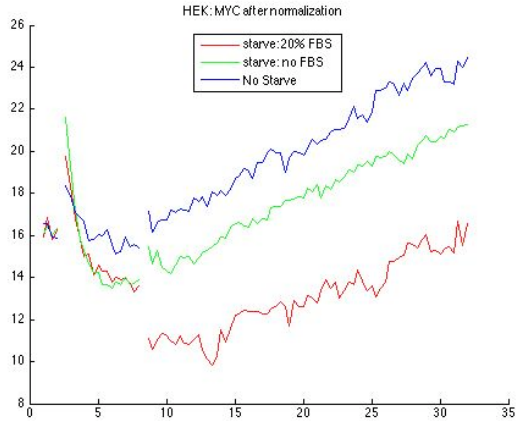


c) HT29: EGR1 before normalization.

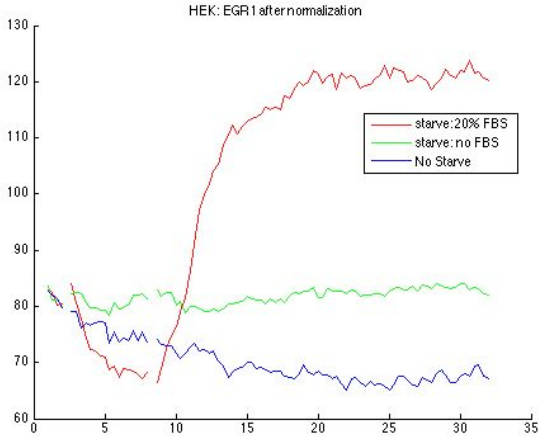
Figure 3: HT29 cell line data for starved and then fed 20%FBS group (Group 1), starved with no FBS group (Group 2), and no starvation group (Group 3) for the genes JUN, MYC, and EGR1.



a) HEK: JUN after normalization.

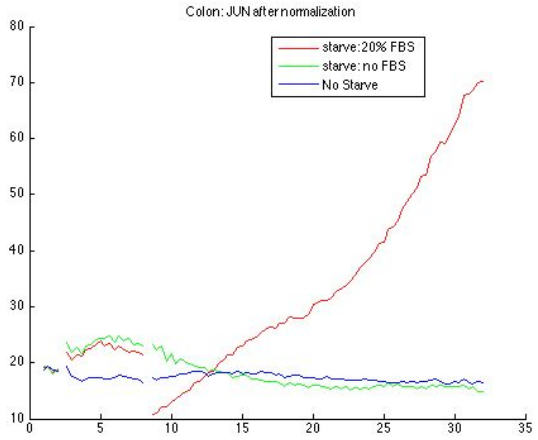


b) HEK: MYC after normalization.

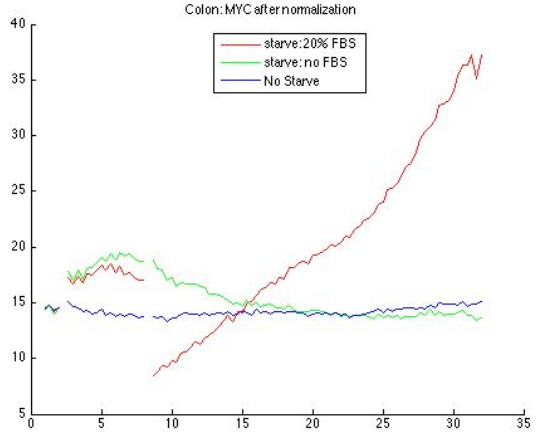


c) HEK: EGR1 after normalization.

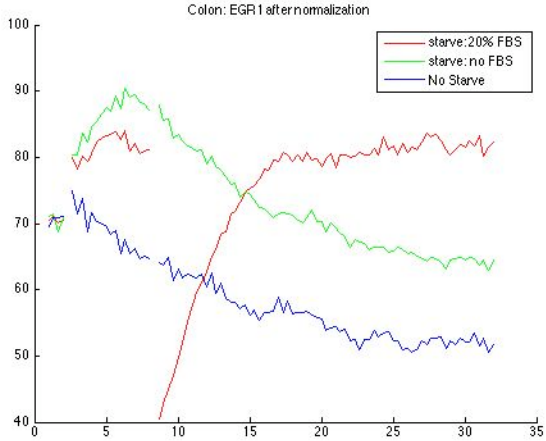
Figure 4: Normalized HEK cell line data for starved and then fed 20%FBS group (Group 1), starved with no FBS group (Group 2), and no starvation group (Group 3) for the genes JUN, MYC, and EGR1.



a) HT29: JUN after normalization.



b) HT29: MYC after normalization.



c) HT29: EGR1 after normalization.

Figure 5: Normalized HT29 cell line data for starved and then fed 20%FBS group (Group 1), starved with no FBS group (Group 2), and no starvation group (Group 3) for the genes JUN, MYC, and EGR1.

3. MODELING AND ON-OFF CONTROL

3.1 Treatment Modeling

The normalized data shows that different treatments have different effects on the GFP intensities of the three genes (JUN, MYC, and EGR1) in the colon cancer cell line. It can be observed that FBS addition (Group 1) increases the GFP intensity while continuous starvation (Group 2) decreases the GFP intensity. This property is used to model the two treatments of FBS addition and starvation using an exponential model.

A state space model with two states x_1 and x_2 is considered where x_1 signifies the closeness to the minimum GFP intensity – M_{ine}, and x_2 signifies the closeness to maximum GFP intensity - M_{axe}. While x_1 is equivalent to the 'off' state in a PBN or the low state in a cell, x_2 parallels the 'on' state in the PBN or the high state in a cell. The close parallel to the states of a PBN is why a second order model is initially selected.

$$x_1(t) = \frac{M_{axe} - y(t)}{M_{axe} - M_{ine}} \tag{4}$$

$$x_2(t) = \frac{y(t) - M_{ine}}{M_{axe} - M_{ine}} \tag{5}$$

It is observed from the above two equations ((4) and (5)) that

$$x_1(t) + x_2(t) = 1 \tag{6}$$

The observed output is then:

$$y(t) = x_1(t) * M_{ine} + x_2(t) * M_{axe} \tag{7}$$

We can assume that the GFP intensity is expressed as molecule counts that evolve following two Poisson processes of production and degradation. Let λ_1 and λ_2 be the inverse of the average switching times from uninduced to induced and vice versa. As an exponential second order model is considered, the model for the study is provided by:

$$\frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} -\lambda_1 & \lambda_2 \\ \lambda_1 & -\lambda_2 \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{8}$$

It is observed from equation (6) that $x_1(t)$ and $x_2(t)$ are dependant on each other. The second order model presented above can thus be reduced to a first order model. Let the state be represented by x such that

$$x(t) = x_2(t) = \frac{y(t) - M_{ine}}{M_{axe} - M_{ine}} \tag{9}$$

The first order model is now given by

$$\frac{d}{dt}x(t) = (-\lambda_1 - \lambda_2) * x(t) + \lambda_1 \tag{10}$$

The output is obtained by substituting equation (6) into equation (7)

$$y(t) = (M_{axe} - M_{ine}) * x(t) + M_{ine} \tag{11}$$

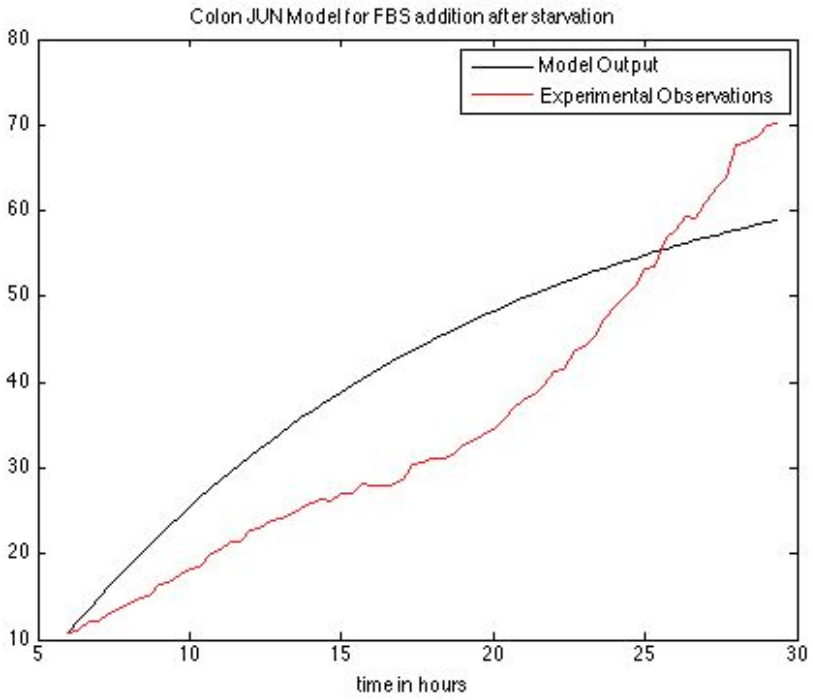
To obtain the values of λ_1 and λ_2, the equilibrium or steady state points are considered. The steady state point is given by

$$(-\lambda_1 - \lambda_2) * x_e + \lambda_1 = 0 \tag{12}$$

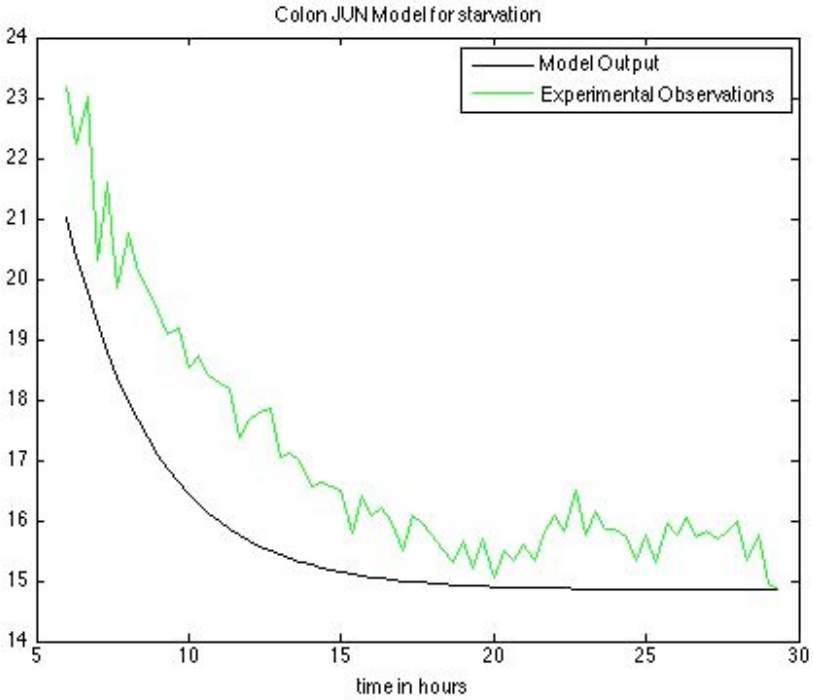
In the above equation (12), x_e is the steady state value at equilibrium, which can be estimated from the experimental data value at 32 hours. The switching time from uninduced to induced is taken to be the time where the intensity in the curve increases beyond a certain threshold. The inverse of this switching time is used to estimate λ_1.

The above model can be used to model both the treatments of starvation as well as FBS addition by using their respective steady state values, and maximum and minimum intensity values.

We next present model validation results for the HT29 colon cancer cell line. Figures 6 a) and b) validate the FBS addition and starvation models respectively with respect to the response of the JUN gene. Figures 7 a) and b) validate the two models with respect to the MYC gene response and Figures 8 a) and b) validate the two models with respect to the EGR1 gene responses.

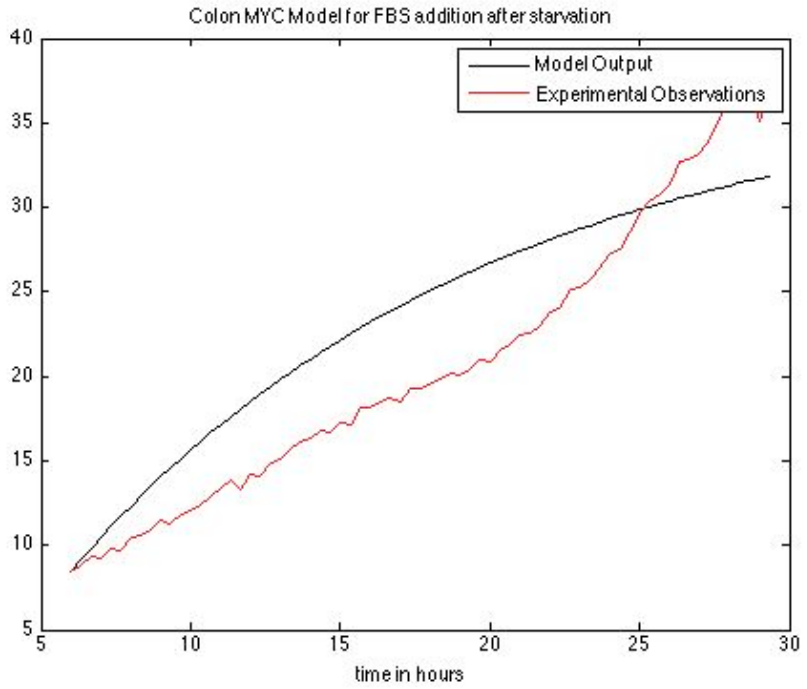


a) HT29 JUN model for FBS addition.

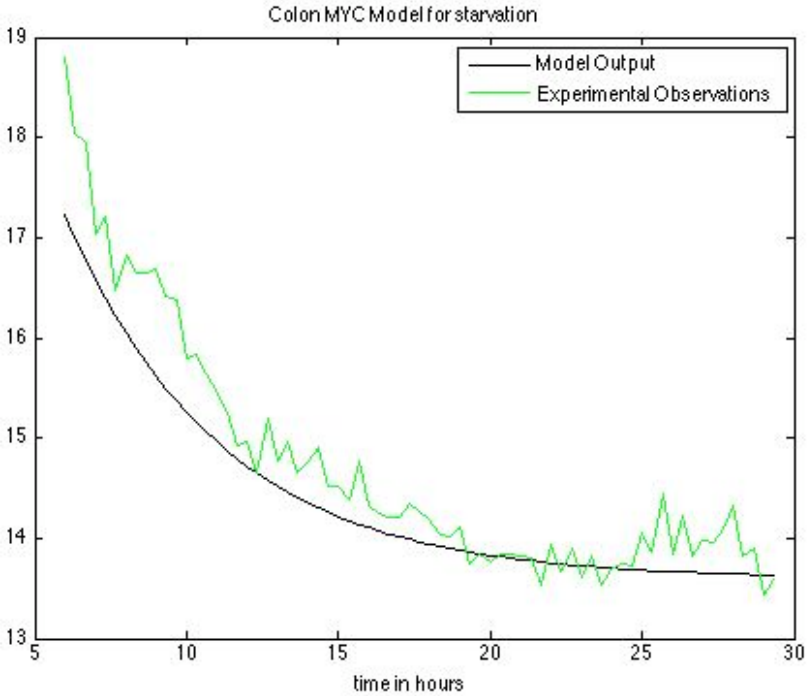


b) HT29 JUN model for starvation.

Figure 6: HT29 colon cancer JUN models for FBS addition and starvation respectively.

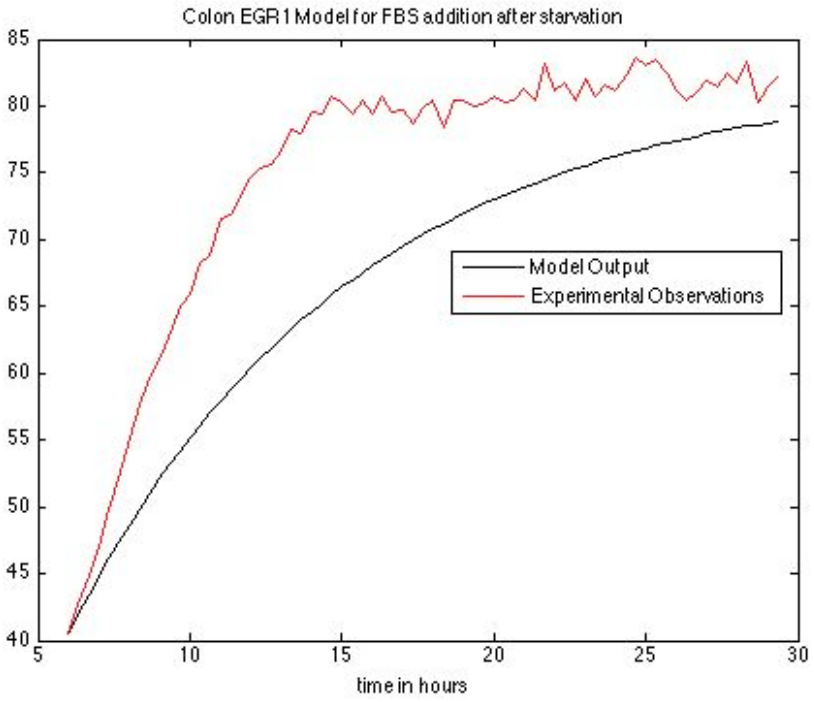


a) HT29 MYC model for FBS addition.

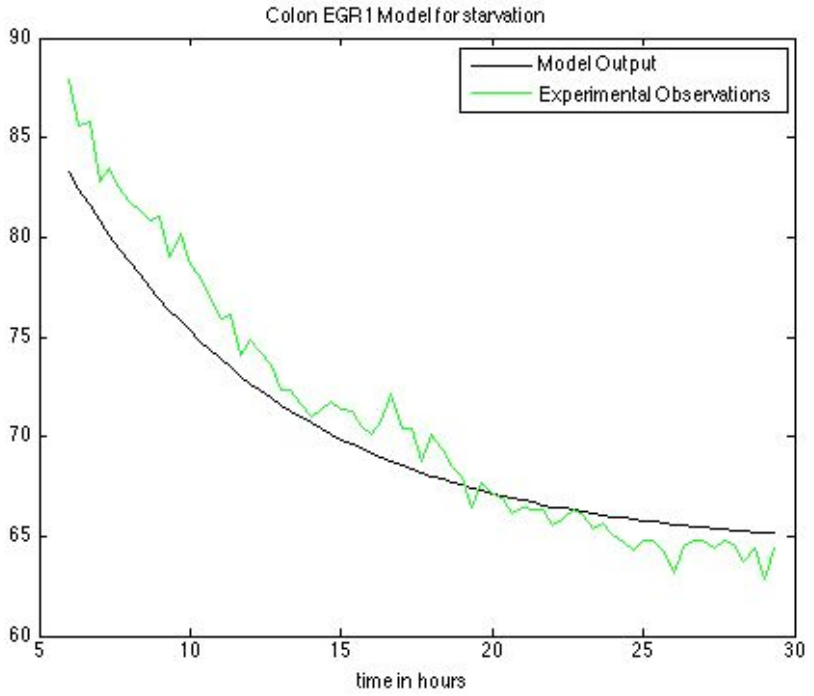


b) HT29 MYC model for starvation.

Figure 7: HT29 colon cancer MYC models for FBS addition and starvation respectively.



a) HT29 EGR1 model for FBS addition.



b) HT29 EGR1 model for starvation.

Figure 8: HT29 colon cancer EGR1 models for FBS addition and starvation respectively.

3.2 Motivation for Intervention

From the plots in Figures 4 and 5, it is clear that there exists a difference in the GFP intensities of the gene responses in the HEK and HT29 cell lines. Since the HEK cell line is considered to be near normal, a reasonable goal for intervention is to make the HT29 colon cancer cell line behave more like the HEK Group 3 cell line (the group in which no treatment was applied). This can be achieved by applying intervention in the colon cancer cell line to achieve gene expression response similar to that of the HEK cell line. Although there are many different intervention strategies that one could think of, in this thesis we will focus on only two possible intervention strategies:

- On-Off control.
- Proportional Integral (PI) control.

Following a description of these two types of intervention strategies, we will present a comparison between them.

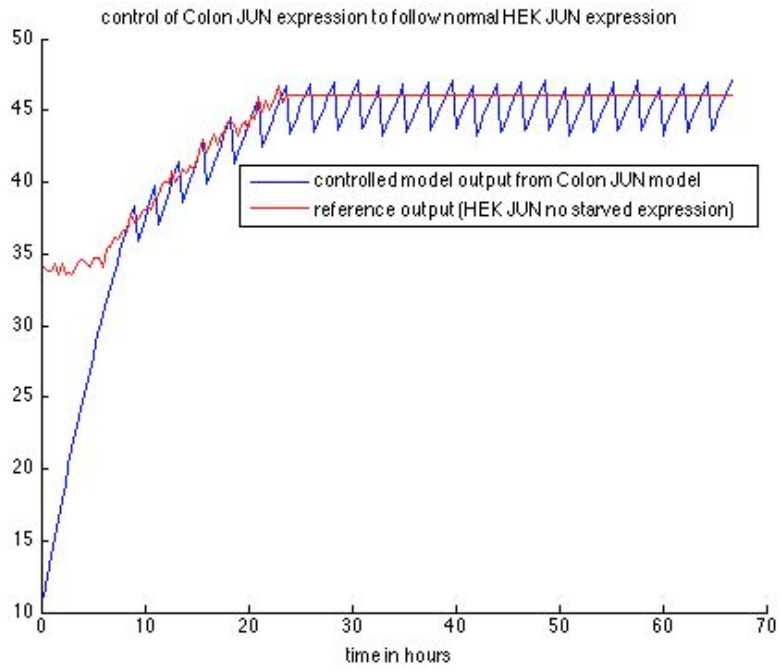
3.3 On-Off Control

The two treatments of FBS addition and starvation have been modeled in section 3.1 using the fact that FBS addition increases the gene expression level while starvation decreases the gene expression level. The starvation treatment model corresponds to the 'Off' control in the on-off intervention strategy, while the FBS addition model corresponds to the 'On' control in the on-off intervention strategy.

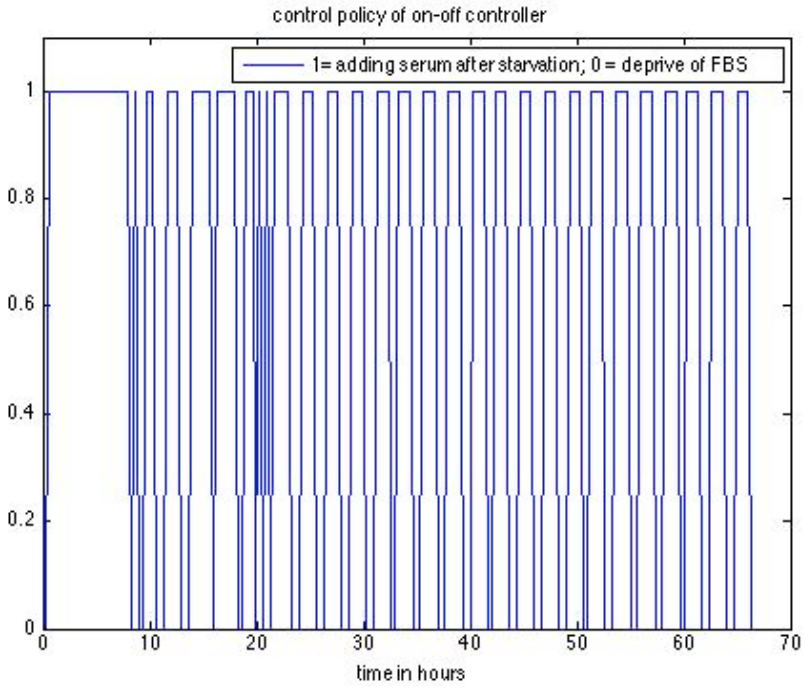
The reference signal is taken to be the HEK group 3 GFP intensity. A control band between $reference\ signal - \Delta ref$ and $reference\ signal + \Delta ref$ is defined. The on-off controller thus applies the FBS addition model when the GFP intensity of the HT29 colon cancer cell line is below the bottom of the control band and applies the starvation model when the GFP intensity of the HT29 colon cancer cell line exceeds the top of the control band.

This procedure is applied for each of the three genes to obtain the desired GFP intensity for the HT29 colon cancer genes. Figure 9 a) shows the controlled and reference gene expression for the JUN gene. Figure 9 b) shows the control policy applied where zero denotes starvation and one denotes FBS addition. Figure 10 a) and b) show similar plots for the MYC gene and figures 11 a) and b) show similar plots for the EGR1 gene.

As can be seen from the graphs below, the control policies are characterized by a lot of switching between FBS addition and starvation. Such control policies are not implementable because the effects of FBS addition remain in the system for a certain period of time and thus the switching between FBS addition and starvation should be necessarily infrequent. To remedy the situation, we introduce PI control in the next section. The integral of the error that will be used in PI control smoothens out the control actions (integration being a smoothing operation [17] – [18]) which leads to less frequent switching.

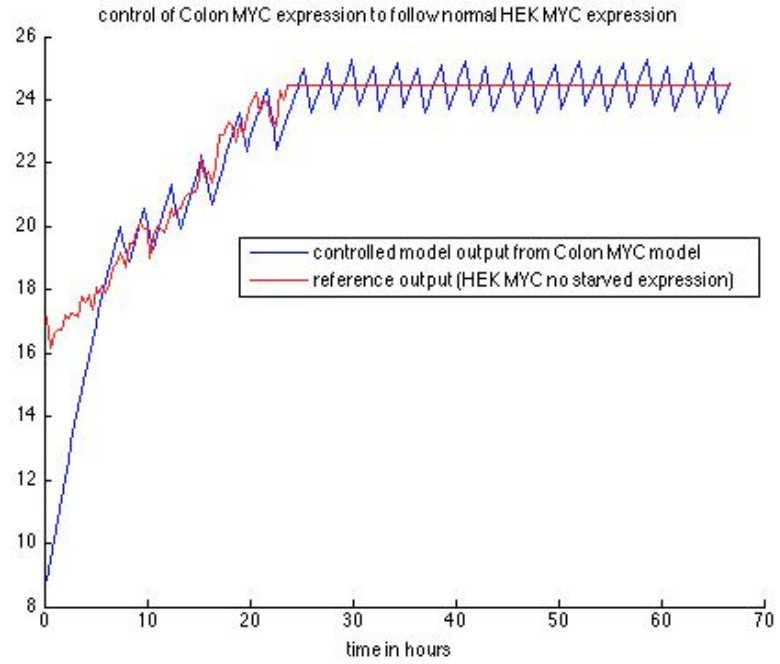


a) Controlled HT29 JUN expression following HEK JUN reference.

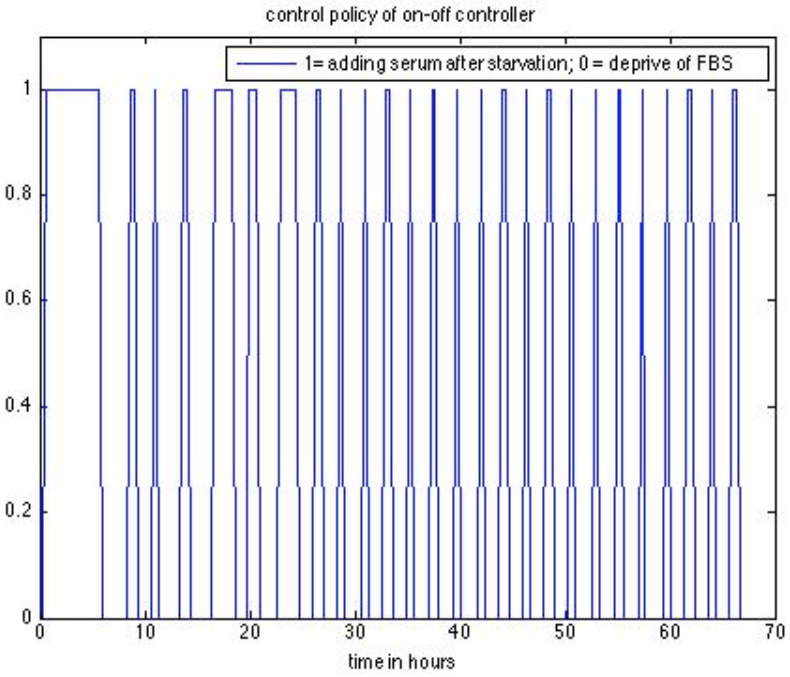


b) Control policy with zero denoting starvation and one denoting FBS addition.

Figure 9: Controlled HT29 colon cancer JUN gene level expression following HEK JUN reference and on-off control policy used to achieve desired results.

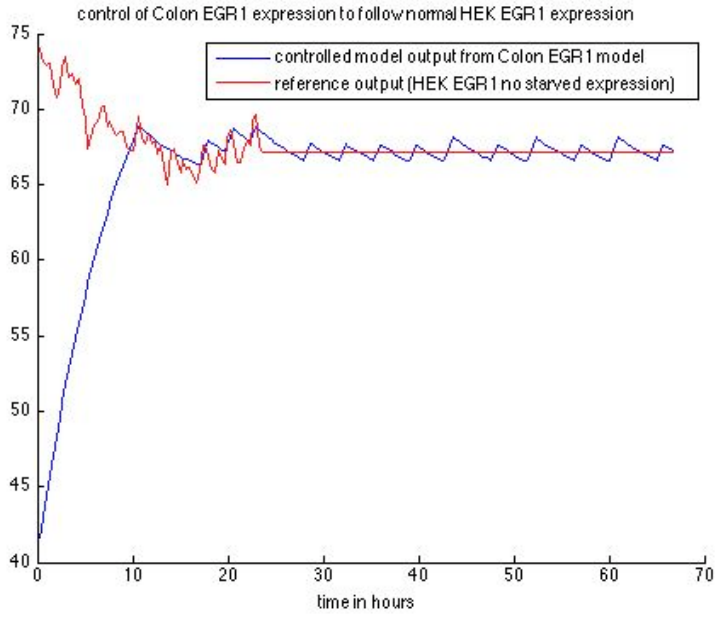


a) Controlled HT29 MYC expression following HEK MYC reference.

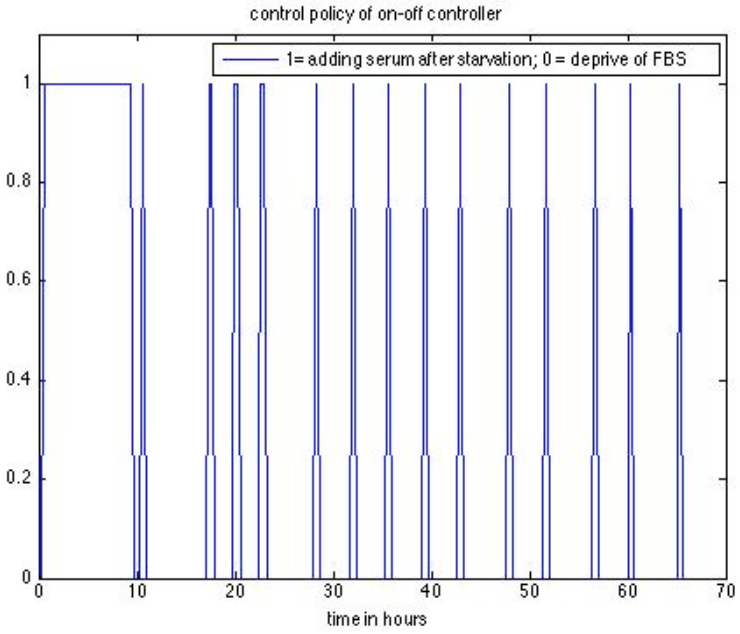


b) Control policy with zero denoting starvation and one denoting FBS addition.

Figure 10: Controlled HT29 colon cancer MYC gene level expression following HEK MYC reference and on-off control policy used to achieve desired results.



a) Controlled HT29 EGR1 expression following HEK EGR1 reference.



b) Control policy with zero denoting starvation and one denoting FBS addition.

Figure 11: Controlled HT29 colon cancer EGR1 gene level expression following HEK EGR1 reference and on-off control policy used to achieve desired results.

4. PI CONTROL

4.1 Proportional Integral Intervention

As already mentioned, the intervention goal is to make the HT29 colon cancer cell line behave like the HEK cell line with respect to the JUN, MYC, and EGR1 responses. On-off intervention to achieve this goal was discussed in the previous section. This section introduces PI control to achieve the same goal.

Ideally for PI control the control input should be a continuous variable. In this case, however, the set of controls is discrete. In order to simulate continuous variation in the control, we assume time to be divided into a series of short segments, each of length Δt [19] – [20]. Then, the desired continuous control signal can be approximated by leaving the system off (applying the starvation model) for a time $(1-u)\Delta t$ and then turning it on (applying the FBS addition model) for time $u\Delta t$. Here, Δt is selected such that sufficient accuracy in terms of tracking is achieved. It could range from between two to four time units (each unit being 20 minutes which is the time at which images were taken).

The next step involves finding u. In order to find u we consider the output signal f from the PI controller.

$$f(t) = P\varepsilon + I\int_0^t \varepsilon dt \tag{13}$$

where,

P is the proportional gain.

I is the integral gain.

ε is the error equal to the difference of the reference signal and output.

The contribution of each of the terms in the above expression is as follows. The proportional control term in the above expression boosts the output by an amount directly proportional to the extent to which the output falls below the desired value. The integral control term has a long-range effect and insures that there is no long-range permanent offset from the desired reference.

The value of u is decided as follows:

- If $f(t)$ lies in between zero and one, then the value of $u(t)$ is equal to $f(t)$.
- If $f(t)$ is less than zero, then the value of $u(t)=0$.
- If $f(t)$ is greater than one, then the value of $u(t)=1$.

Control is then applied in the following manner. The starvation model input is applied for $(1-u)\Delta t$ time, and for the remaining $u\Delta t$ time, the FBS addition model input is applied.

The values of P and I are left to be decided. The values of P and I are approximated as follows. In PI control, it is assumed that the error is equal to zero at steady state. This means that the proportional term is equal to zero. Assuming that steady state has been reached in 32 hours, P can be taken equal to zero and the only contribution to u comes from the integral term. The integral term (the sum of the errors) is approximated from the on-off intervention strategy. We consider an example to understand how the value of I is selected. Considering the value of u to be 0.5, we

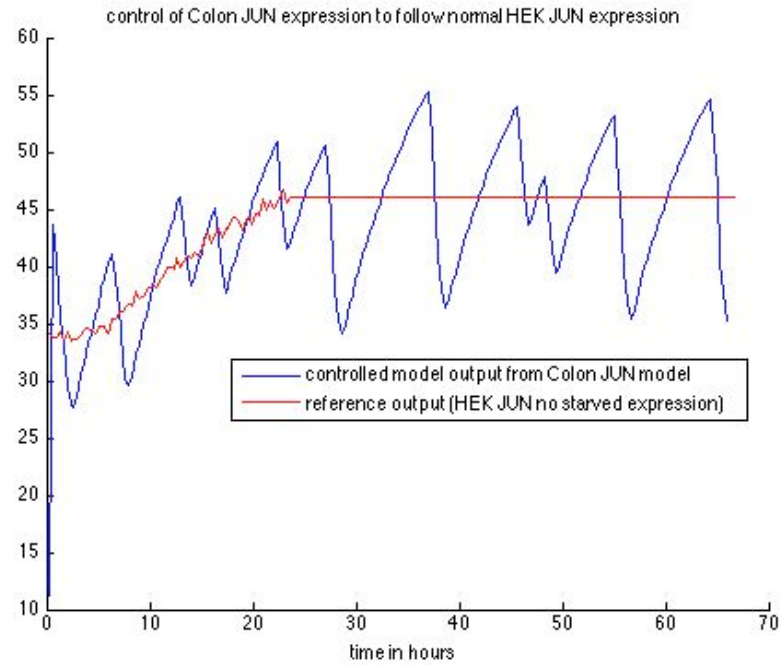
calculate the sum of the errors obtained at each time point through the on-off control intervention strategy. In this example, we consider the total error to be around 50. As the only contribution to u is from the integral term, we see that the estimate of I is 0.5/50, which is equal to 0.01.

The above procedure is used to get an estimate of *I*. Now, assuming steady state but *P* no longer equal to zero and I equal to zero, suppose that the control input causes the output to fall by a value of two below the reference point and we wish the value of *u* to rise by 0.2, from 0.5 to 0.7. An appropriate value of *P* then appears to be approximately 0.3. Using the above-described procedure, appropriate estimates of *P* and *I* can be found for controlling the three genes JUN, MYC, and EGR1.

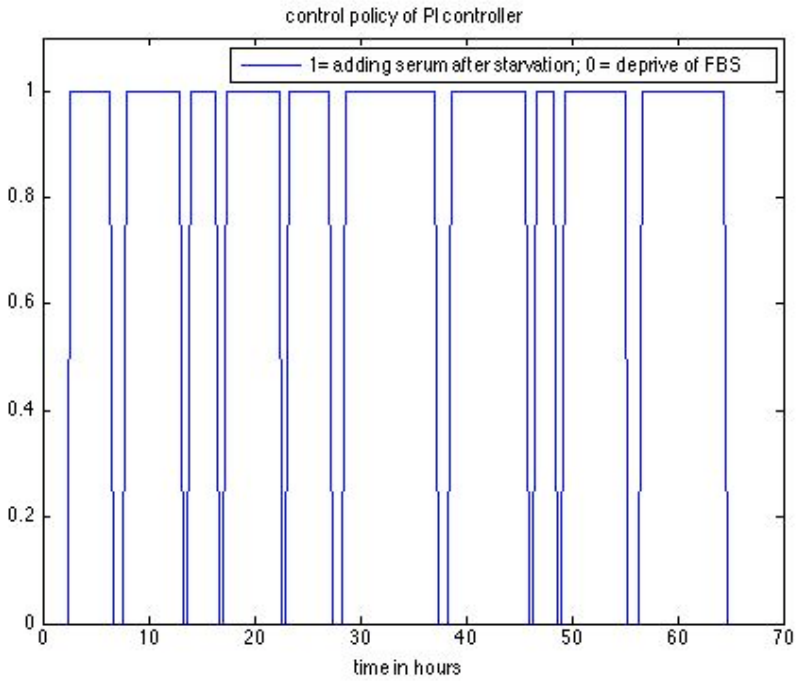
Now, knowing estimates for *P* and *I* and the value of *u*, the HT29 colon cancer cell line can be made to follow the reference HEK cell line using PI control.

4.2 Intervention Results

Simulations of PI control for the three genes JUN, MYC, and EGR1 are shown in the following three figures. Figure 12 a) shows the controlled and reference GFP intensities for the JUN gene. Figure 12 b) shows the control policy applied where zero denotes starvation and one denotes FBS addition. Figures 13 a) and b) show similar plots for the MYC gene and figures 14 a) and b) show the plots for the EGR1 gene.

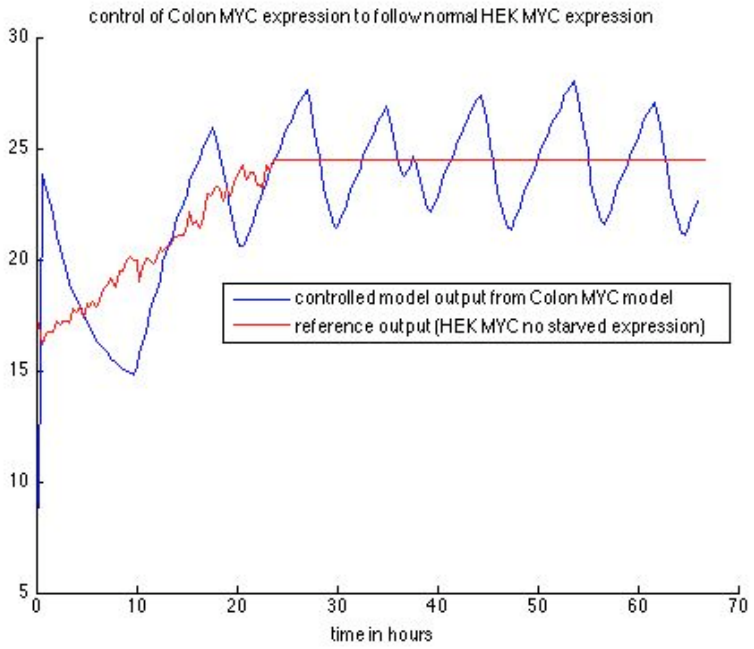


a) Controlled HT29 JUN expression following HEK JUN reference.

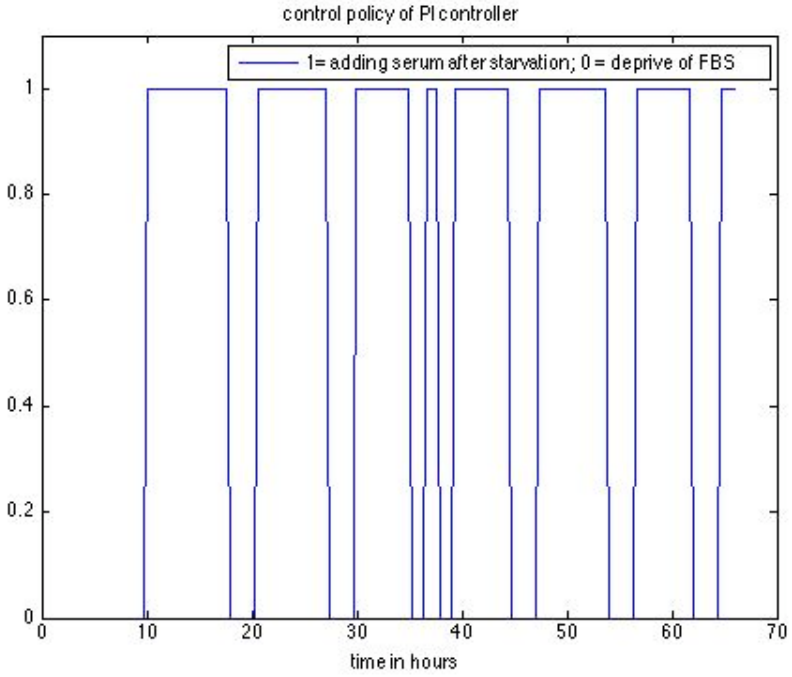


b) Control policy with zero denoting starvation and one denoting FBS addition.

Figure 12: Controlled HT29 colon cancer JUN gene level expression following HEK JUN reference and PI control policy used to achieve desired results.

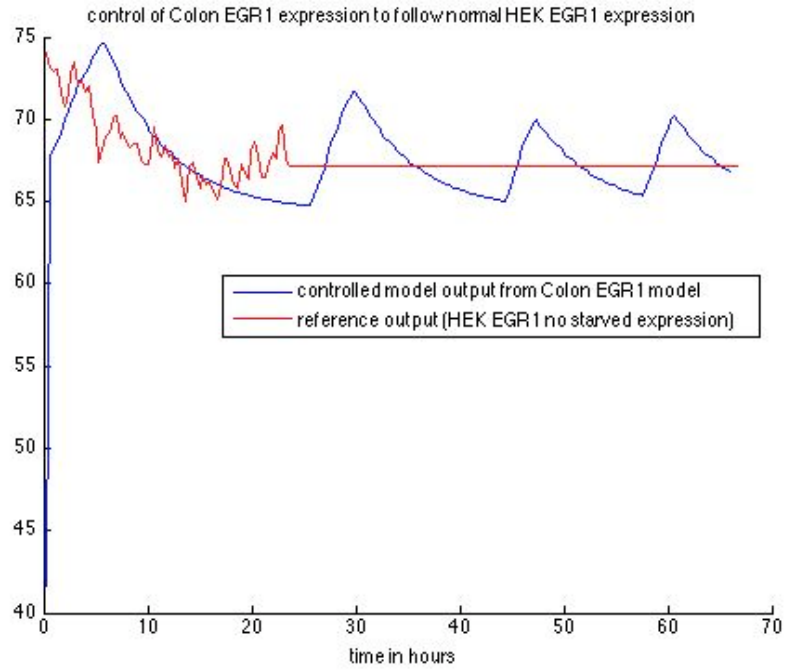


a) Controlled HT29 MYC expression following HEK MYC reference.

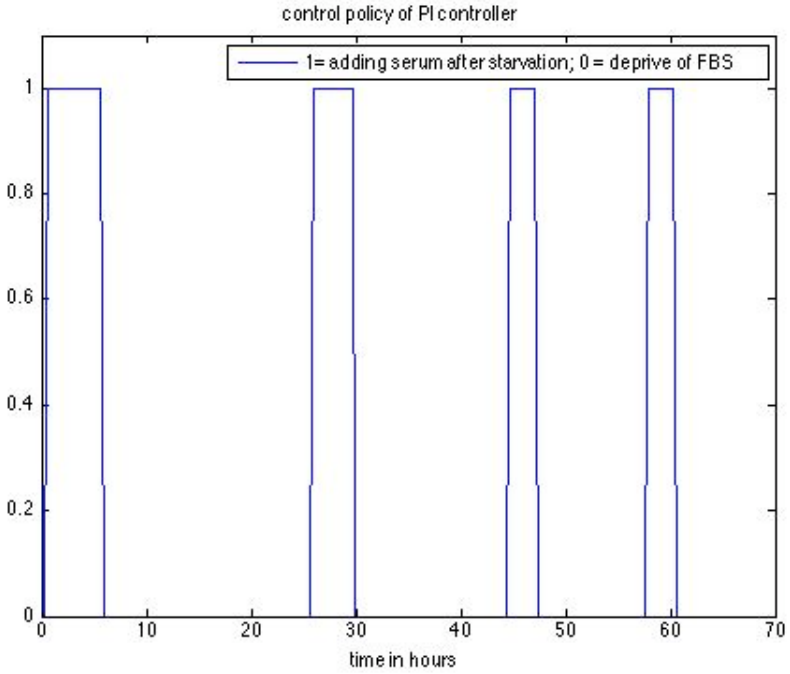


b) Control policy with zero denoting starvation and one denoting FBS addition.

Figure 13: Controlled HT29 colon cancer MYC gene level expression following HEK MYC reference and PI control policy used to achieve desired results.



a) Controlled HT29 EGR1 expression following HEK EGR1 reference.



b) Control policy with zero denoting starvation and one denoting FBS addition.

Figure 14: Controlled HT29 colon cancer EGR1 gene level expression following HEK EGR1 reference and PI control policy used to achieve desired results.

4.3 Comparison of PI Control with On-Off Control

The above simulations of on-off control (Figures 9, 10, and 11) and PI control (Figures 12, 13, and 14) are now compared to study the advantage of one type of control over the other. In experimental implementation of control, the effects of the 20% FBS once fed remain in the system for a period of time and so frequent switching between FBS addition and starvation is not implementable. As can be seen from the control policy of on-off control, switching between the FBS addition and starvation inputs is very frequent making it difficult to implement experimentally. The second approach of PI control however requires much less switching between the two inputs thereby making it a more favorable choice with regard to implementation. It is for this reason that PI control has an advantage over on-off control in this particular application of making the HT29 colon cancer cell lines behave like the HEK cell lines.

5. CONCLUSION

We have used experimental HEK and HT29 cancer cell line data in this research and controlled the HT29 cell line to make it behave more like the HEK cell line. Three of the genes contained in each of the cell lines – JUN, MYC, and EGR1 – are considered. The treatments of addition of 20% FBS to the system and its removal or starvation of the cells are modeled using exponential models. Each of these treatments behaves as inputs to the dynamic model.

The simplest form of on-off control is first simulated followed by simulations of PI control in order to make the HT29 cell line genes follow the reference HEK cell line genes. From the simulations, it is observed that PI control provides a better alternative to on-off control. This is because on-off control is not implementable in actual experiments. The serum (FBS) once added to the cells triggers a set of reactions, and thus cannot be taken out of the system immediately. The frequency of switching between FBS addition and starvation should thus be small, which is why PI control which has a far lower frequency of switching than on-off control is preferable.

Further research can be carried out where the modeling can be enhanced by looking at individual genes and estimating the Poisson parameter for protein molecule generation. As the GFP intensity of a cell is proportional to the number of protein molecules generated, a Poisson process can estimate it because time is continuous and the states are discrete. Also, the PI controller can be extended to control all the three gene expressions simultaneously instead of controlling one gene at a time.

REFERENCES

[1] A. Datta, R. Pal, A. Choudhary, and E.R. Dougherty, "Control approaches for probabilistic gene regulatory networks," *IEEE Signal Processing Magazine*, vol. 24, pp 54-63, 2007.

[2] G. Vahedi, B. Faryabi, J.F. Chamberland, A.Datta, and E.R. Dougherty, "Intervention in gene regulatory networks via a stationary mean-first-passage-time control policy," *IEEE Transactions of Biomedical Engineering*, vol. 55, pp 2319-2331, 2008.

[3] D. Endy and R. Brent, "Modelling cellular behavior," *Nature*, vol. 409, pp 391-395, 2001.

[4] T. Ideker, T. Galitski, and L. Hood, "A new approach to decoding life: systems biology," *Annual Review of Genomics and Human Genetics*, vol. 2, pp 343-372, 2001.

[5] J. Hasty, D. McMillen, F. Isaacs, and J.J. Collins, "Computational studies of gene regulatory networks: in numero molecular biology," *Nature Reviews Genetics*, vol. 2, pp 268-279, 2001.

[6] M. Karamouzis, V. Gorgoulis, and A. Papavassilou, "Transcription factors and neoplasia: vistas in novel drug design," *Clinical Cancer Research*, vol. 8, pp 949-961, 2002.

[7] A. Datta and E.R. Dougherty, *Introduction to Genomic Signal Processing with Control*, Boca Raton, FL : CRC Press, 2007.

[8] S. Kauffman, “Metabolic stability and epigenesis in randomly constructed genetic nets,” *Theor. Biol.*, vol. 22, pp 437-467, 1969.

[9] S. Kauffman, “Homeostasis and differentiation in random genetic control networks,” *Nature*, vol. 224, pp 177-178, 1969.

[10] L. Glass and S. Kauffman, “The logical analysis of continuous, nonlinear biochemical control networks,” *Journal of Theoretical Biology*, vol. 39, pp 103-129, 1973.

[11] S. Kauffman, *The Origins of Order: Self Organization and Selection in Evolution*, New York : Oxford Univ. Press, 1993.

[12] E.R. Dougherty, S. Kim, and Y. Chen, “Coefficient of determination in nonlinear signal processing,” *Signal Process.*, vol. 80, pp 2219-2235, 2000.

[13] I. Schmulevich, E.R. Dougherty, S. Kim, and W. Zhang, “Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks,” *Bioinformatics*, vol. 18, pp 261-274, 2002.

[14] I. Schmulevich, E.R. Dougherty, and W. Zhang, “From boolean to probabilistic boolean networks as models of genetic regulatory networks,” *Proceedings of IEEE*, vol. 90, pp 1778-1792, 2002.

[15] J. Hua, Y. Balagurunathan, Y. Chen, J. Lowey, M. Bittner, Z. Xiong, E. Suh, and E.R. Dougherty, “Normalization benefits microarray-based classification,” *EURASIP Journal on Bioinformatics and Systems Biology*, 2006.

[16] M. Chalfie, Y. Tu, G. Euskirchen, W.W. Ward, D.C. Prasher, “Green fluorescent protein as a marker for gene expression,” *Science*, vol. 263, pp 802-805, 1994.

[17] G.F. Franklin, J.D. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems*, 5th ed. Upper Saddle River, NJ : Pearson Prentice Hall, 2006.

[18] S.C. Goyal and U.A. Bakshi, *Principles of Control Systems*, 7th ed. Shaniwar Peth, Pune, India : Technical Publications, 2002.

[19] J.O. Wilkes and P. LaValle, "Air Temperature Control with Recycle," *umich.edu*, Jan. 3, 2006. [Online]. Available: http://www.engin.umich.edu/class/che360/coursepack/ch12-tempcontrol.doc. [Accessed: Aug. 8, 2008]

[20] D.R. Coughanowr, *Process Systems Analysis and Control*, 3rd ed. New York : McGraw-Hill, 2008.

VITA

Name:	Rouella Joan Mendonca
Address:	Department of Electrical and Computer Engineering, Texas A&M University, 214 Zachry Engineering Center, TAMU 3128 College Station, Texas 77843-3128
Email Address:	rouella@neo.tamu.edu
Education:	B.S., *Summa Cum Laude* Computer Engineering, Texas A&M University, 2007