Nonlinear Dynamics of Transcription Regulatory Networks – Application to E.coli.

Kun Qu

Advisor: Prof. Peter J. Ortoleva

Center for Cell and Virus Theory, Department of Chemistry, Indiana University,

Bloomington, Indiana 47405

Contents

1. Objectives

The purpose of this work is to construct a general methodology for the nonlinear analysis of transcription regulatory network and thereby overcome the impediment to progress in understanding cellular complicity.

2. Introduction

Genetics and genomics start with Gregor Mendel's discovery of the law of heredity [1] and their rediscovery in the early days of the twentieth century. Scientists then realized that DNA is the main hereditary material [2] and began to determine its structure. 50 years ago, Watson and Crick's discovery of the double-helical structure of DNA [3], is a landmark event, after which the main focus of life science changed to elucidating the genetic code [4], developing recombinant DNA technologies [5], and establishing increasingly automatable methods for DNA sequencing [6-7], and eventually make the ambitious Human Genome Project (HGP) finished as least two years ahead of expectation. The next phase of genomics is to catalogue characterize and comprehend the entire set of functional elements encoded in the human and other genomes, which is a more challenge for its complicity. For instance, gene and gene products do not function independently, but participate in complex, interconnected pathways, networks of molecular system that taken together give rise to the workings of cells, tissues, organs and organisms [8]. Qualitatively as well as quantitatively defining these systems and determining their properties and interactions are crucial to understanding how biological systems function [9]. Yet these systems are far more

complex than any problem that molecular biology, genetics has yet approached [8]. Image how many genes in human and how many interactions! Because of its complicity, genomics attracts a lot people from different fields, and has become a central and productive area of research. Biologists and chemists start with some simple motifs to collect the information of the products of genes, and transcriptional regulatory relationship, and make the basic bricks for the whole building. As data accumulate, physicist and mathematician begin to analyze the properties of the gene-transcription factor (TF) regulatory networks, and post translation reaction networks etc. and construct models to represent the structure of these networks in expect to understand the relationship between DNA sequence information and nonlinear cellular responses [10].

In this report, I present an automatic approach for nonlinear analysis of gene regulatory networks based on our transcription and translation reactions model and our understanding of a given regulatory network. We use programs to automatically read this information and generate ordinary differential equations that are readable by AUTO [11] – a bifurcation analyzer, and use AUTO to give a nonlinear analysis to a known regulatory network. We keep the whole work flow general and automatic and then we apply this approach to E.coli. We obtained a branch of results that are not only interesting but also of great importance in drug discovery and caner therapy.

3. Work flow

Our overall work flow is shown in Figure 1, we first define a mathematical model

describe mRNA and TF behavior, concerning with transcription, translation, and post translation reactions. Meanwhile we obtain the regulatory network, which we call B matrix here, from websites and literatures [12]. The B matrix tells which genes are up/down regulated by which transcription factors, also we obtained a C matrix, which contains the information of which genes make which transcription factors. With this information, we can import microarray data into our KAGAN [13] program to calibrate those parameters that will be useful in our model, we write these parameters into different data files and used as the input of the next step. On the other hand, I decomposed and filtered B and C matrix to be independent central B and C matrix. Then we combine all the information and write equations into Auto, I use AUTO to get the bifurcation result and draw the graphs. So, in a word, we make an automatic work flow with the input of gene regulatory network and have the cell bifurcation graph output. More detailed algorithm method is described in the next chapter.

It is worth to pointing out that this project is interdisciplinary. It evolves bioinformatics, chemical kinetics, numerical analysis – applied mathematics and biochemistry. In this report, I will clarify which element of the project is specifically done by me and my colleagues.

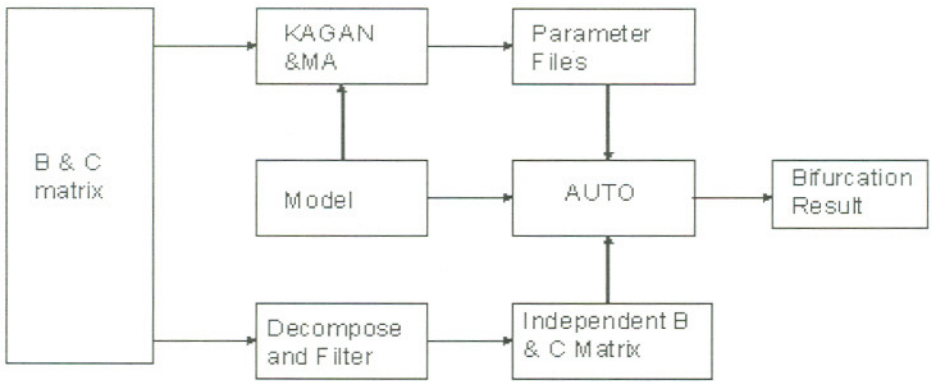

Figure 1 Work Flow

4. Methods

Our approach is to develop a very general cell process network and then take the point of view that as we add more and more detailed processes, cell differentiation and other phenomenon should emerge. Thus we propose to add TF/gene control interactions, post translational reactions and other factors as gathered from the literature. This is distinct from the more traditional approach whereby one restricts the analysis to a small network and negates the majority of its interaction with the rest of the network even though the latter certainly is strongly coupled to it. While a great difficult of the latter is that this approach built a small network that is designed to arrive the answer we seek and therefore is not objective, rather, we believe that the answer should naturally follows to as much information as we have about the whole network.

We process the whole work flow in answering 5 questions: 1. How to construct

our B and C matrix. 2. How to construct a reasonable cell model to describe mRNA and TF behavior. 3. What is KAGAN and mirocarray data. 4. How to decompose and filter the original B and C matrix. 5. What is AUTO.

4.1 How to construct our B and C matrix.

We have a script that can automatically search from literature and database for Gene – TF regulatory information to arrive at a putative, albeit incomplete and likely error – prone regulatory networks, which we call B and C matrix here. The B matrix is a Gene * TF dimensional matrix and the element of the matrix are 0, 1, -1, which means this Gene is non-regulate, up-regulate, down-regulate by that TF respectively. Some of the proteins will dimerize with another one or two proteins before it can work as a TF, in these cases we use 2 or 3 to indicate their dimerization and higher order complexing. We labeled the Gene and TFs from the first to last, and now we have a 984 Genes and 144 TFs B matrix for E.coli K12, which consists 1/3 of the whole regulatory network. See Figure 2 as an example of B matrix. The C matrix contains three columns, the first column is the TF index number, the second column is the index of the gene that makes this TF, and the third column is the name of that gene. See Figure 3 as an example. In our 144 TFs database, 116 of which are made by those Genes in our known regulatory network, and 28 of which are made by other genes that we have no idea. The regulatory information is obtained from Ecocyc and Regulon DR [12], by my group mates Kranthi, Frank, Leshin, Al, Tim and organized by Lisa.

	NtrC-Phosphorylated			Fur	CpxR	IHF	ModE-Molybdate	
fur	0	-1	0	0	0	0	1	0
cbl	1	0	0	0	0	0	0	0
cspA	0	0	0	0	0	0	0	0
glcC	0	0	0	0	0	0	0	0
gadX	0	0	0	0	0	0	-1	0
treR	0	0	0	0	0	0	0	0
argR	0	0	0	0	0	0	0	0
fdp	0	0	0	0	0	0	0	0
uxuR	0	0	0	0	0	0	1	0
exuR	0	0	0	0	0	0	0	0
uidR	0	0	0	0	0	0	0	0
araC	0	0	0	0	0	0	1	0
fis	0	0	0	1	0	-1	0	0
fnr	0	0	0	0	0	0	0	0
galS	0	0	0	0	0	0	1	0
glnG	2	0	0	0	0	0	2	0

Figure 2 An Example of B matrix

1	16	glnG
2	1	fur
6	13	fis
7	31	crp
8	65	mlc
9	14	fnr
10	37	hns
11	63	nac
12	19	oxyR
13	47	soxS
14	20	phoB
15	26	cysB
16	58	arcA
19	42	lrp
20	45	marA

Figure 3 An example of C matrix

4.2 How to construct a reasonable cell model to describe mRNA and TF behavior

Gene regulation is considered here to be a Markov process involving the attachment/detachment of transcription factors (TFs) to sites on each gene while the dynamics of high-population species (e.g. TFs) are treated via chemical kinetics. Let

P_{ij} be the probability that site j on gene i is occupied. It is assumed that site ij can only be occupied by a unique TF labeled n_{ij}. Each site is considered to be independent so that the Markov dynamics take the form

$$\frac{dP_{ij}}{dt} = k_{ij}^{+} T_{n_{ij}} (1 - P_{ij}) - k_{ij}^{-} P_{ij} \tag{1}$$

where $T_{n_{ij}}$ is the concentration of TF n_{ij}. Let b_{ij} indicate the nature of the regulation of gene i by TF n_{ij} due to site j:

$$b_{ij} = \begin{cases} +1, & up \quad regulation \\ -1, & down \quad regrlation \\ 0, & no \quad regulation \end{cases} \tag{2}$$

Introduce a function $\Psi(P,b)$ such that

$$\Psi = \begin{cases} P, b = +1 \\ 1 - P, b = -1 \\ 1, b = 0. \end{cases} \tag{3}$$

Assuming that a gene is most conducive for transcription if its up-regulating sites are occupied and its down regulating ones are not, the probability that gene i is conducive, Θ_i, is given by

$$\Theta_i = \prod_{j=1}^{N_{(i)}} \Psi(P_{ij}, b_{ij}). \tag{4}$$

where $N_{(i)}$ is the number of sites on gene i. With this it is assumed that the dynamics of the cellular RNA content R_i for the single RNA type assumed to be associated with gene i (e.g. splicing is ignored) is given by

$$\frac{dR_i}{dt} = k_i^{\max} [RP] \{\Theta_i + \zeta_i\} / (1 + \zeta_i) - \lambda_i R_i \tag{5}$$

where $k_i^{\max}$ is the maximum rate, ζ_i is a small parameter that allows a minimal rate of transcription even if gene i is not optimally conducive and $[RP]$ is the concentration

of intra-nuclear or intra-bacterial RNA polymerase, while λ_i is a rate coefficient for RNA degradation. To complete the model we assume that N_{TF} transcription factors each arise either out of the translation of a specific gene or the complexing of other TFs and factors. Let TF n be translated via gene I_n, or from a dimerization of other TFs and factors, the latter having net rate W_n for the n-th TF. Then the following simple model is adopted

$$\frac{dT_n}{dt} = \alpha_n R_{I_n} - \beta_n T_n + W_n(I, \underline{c}) - U_n \tag{6}$$

for rate factors α_n, β_n, and set $\underline{c}$ of concentration of other factors. When a TF is a single translated protein, $W_n = 0$; when it arises out of dimerization or other complexing then $\alpha_n = \beta_n = 0$; this allows for a rather general structure to the model. The contribution U_n arises from the complexing of TFs to the N_g gene:

$$U_n = \sum_{i=1}^{N_g} \sum_{j=1}^{N_{(1)}} \delta_{nn_{ij}} \left\{ k_{ij}^{+} T_{n_{ij}} (1 - P_{ij}) - k_{ij}^{-} P_{ij} \right\} \tag{7}$$

where $\delta_{nn_{ij}}$ is 1 or 0 depending on whether $n = n_{ij}$ or $n \neq n_{ij}$, respectively.

In the above model a number of assumptions were made including

- T_n, as it effects the occupation probability Θ_i is assumed to be a concentration and not a thermodynamic activity; and
- For eukaryotic cells the factors $k_i^{\max}$ and α_n depend on intra-nuclear nucleotide and cytoplasmic amino acid levels respectively, whose time variation can be ignored, the analogous situation for bacterial cell; and similarly for [RP].

The above model yields rapidly responsive control if the rate constants k_{ij}^{+} and k_{ij}^{-}

are large. Let $Q_{ij} = k_{ij}^{+} / k_{ij}^{-}$. Then in the limit k_{ij}^{-} large

$$P_{ij} = \frac{Q_{ij} T_{n_{ij}}}{1 + Q_{ij} T_{n_{ij}}} \tag{8}$$

This is the limiting case considered by Sayyed-Ahmad et al. (2004) in their microarray analysis approach.

To implement the above model for analysis via the AUTO software, we wrote a program that automatically writes the above differential equation as a file. In addition the parameters $k_i^{\max}, [RP], \lambda_i, k_{ij}^{+}, k_{ij}^{-}, \alpha_n, \beta_n$ must be provided as must the matrix $\underline{\underline{b}}$ and the vector $\underline{I}$ that is our B and C matrix.

The model supports a great richness of multiple steady-states, periodic and other nonlinear dynamical system behaviors. Objective is to develop a methodology for automating the discovery of these phenomena. The straightforward approach is to write a program， which automatically write a file for input to AUTO [11] which is a bifurcation analyzer The input is the set of model parameters and the matrices $\underline{\underline{b}}$ and $\underline{I}$ determining the structure of the regulatory control system. Before applying the approach to *E.coli*, consider the nature of the steady states problem and the potential of the system to support a myriad of distinct states.

The structure of the model allows for gene simplification when the system is at steady state. Under that condition

$$P_{ij} = \frac{Q_{ij} T_{n_{ij}}}{1 + Q_{ij} T_{n_{ij}}} \tag{9}$$

As seen from (1). This implies that U_n in (6) vanishes and hence

$$\begin{cases} W_n = 0 \quad if \quad \alpha_n = \beta_n = 0 \\ T_n = \alpha_n R_{I_n} / \beta_n, \quad \alpha_n \quad and \quad \beta_n \neq 0 \end{cases} \tag{10}$$

This implies that either T_n is given by the equilibrium relation ($W_n = 0$), or is obtained from the steady state balance of translation and degradation.

In the simplest case where all TFs are monomers (i.e. one gene and without dimerization) that we obtain:

$$T_n = \Gamma_n[\Theta_{I_n} + \zeta_{I_n}] \tag{11}$$

$$\Gamma_n = \alpha_n k_{I_n}^{\max}[RP]/\beta_n \lambda_{I_n} \tag{12}$$

Thus the problem reduces to solving N_{TF} equations for N_{TF} $\underline{T}$ and involves the N_{TF} parameters $\underline{\Gamma}$. This greatly simplifies the analysis and will be referred to as the reduced problem henceforth.

The reduced problem can be decomposed further in terms of the simple motifs, self-regulating gene can be solved independently, i.e. for a single site, self-upregulating gene one has

$$T = \Gamma\left[\frac{QT}{1+QT} + \zeta\right] \tag{13}$$

which can be solved exactly. Similarly for a two site self-upregulating case

$$T = \Gamma\left[\left(\frac{Q_1 T}{1+Q_1 T}\right)\left(\frac{Q_2 T}{1+Q_2 T}\right) + \zeta\right] \tag{14}$$

This also can be solved analytically as can mix two sites cases (i.e. one up and the other down). For the single site, or the symmetrical, the obvious fundamental variables are $T' = QT$ and $\Gamma' = Q\Gamma$, i.e. $T' = \Gamma'\left[\left(\frac{T'}{1+T'}\right)^m + \zeta\right]$ for the m symmetric site case.

Two gene motifs also lend themselves to analytical solution in the reduced problem. For gene 1 and 2, such that T_1 up-regulate G_2 and T_2 down regulate G_1 we have

$$T_2 = \Gamma_2\left[\frac{Q_1 T_1}{1+Q_1 T_1} + \zeta_1\right] \tag{15}$$

$$T_1 = \Gamma_1\left[\frac{1}{1+Q_2 T_2} + \zeta_1\right] \tag{16}$$

which can also be solved analytically, and similarly for other two and three gene motifs. Finally "follower" genes (controlled by genes in the motifs) can be obtained from TFs involved in the motifs. More complex cases involving the common dimmer TFs can also be solved analytically.

If there are M^* basic motifs, which support multiple (i.e. 2 stable) steady states, that the system can support 2^{M^*} distinct states. This underlies the richness of the cellular problem.

4.3 What is KAGAN and microarray data

KAGAN (KAryote Genome ANalyzer) is a software package that receives raw time series microarray data, the list of factors that regulate each gene, and yields the timecourse of thermodynamic activities within the nucleus or prokaryotic cell. Software packages that could be used to provide input to KAGAN include PAINT (developed at Thomas Jefferson University), which gives gene/transcription factor interactions, but does not give the sense, up/down, of the regulation because PAINT is simply a sequence analysis package. The results of KAGAN are provided graphically in terms of the predicted timecourse of transcription factor activities. The latter provides information that can be used to detect errors in the gene control network. For example, if a gene is upregulated by one transcription factor and unaffected by all

others then a change in transcription factor activity should be well correlated with a microarray response for that gene; if this is not the case, the network requires correction.

KAGAN is built on an entropy maximization principle. Entropy is a quantitative measure of uncertainty. It is used in our methodology to provide a framework for data/cell model integration. The information theory approach of Sayyed-Ahmad et al. (2003) and a transcription kinetic model similar to that in Weitzke and Ortoleva (2003) [14] are used. More details of the methodology and technician could be found in Ref [15].

DNA microarrays are a new technology that allows the whole genome to be monitored on a single chip so that a better picture of the interactions among thousands of genes can be observed simultaneously [16]. Recently, this technique has been widely used in many fields of life science. It has already yielded many discoveries in the field of gene discovery, disease diagnosis, drug discovery, and toxicology research [17]. KAGAN combined cDNA microarray with a transcription kinetic modeling through information theory, and get more information about the gene regulatory networks than obtained preciously [15]. for example, as we said before, our gene regulatory network is incomplete, using the cDNA microarray with information theory, we can not only get most probable gene-TF relationship for those we were not sure, but also expend our previous regulatory network as well, further more, KAGAN yield rate and equilibrium constants for the transcription factors and RNA degradation reactions.

So we input our raw cDNA microarray and gene regulatory network, run in

KAGAN and get a more convinced gene regulatory network as well as calibrated parameters that are required by our model.

4.4 How to decompose and filter the original B and C matrix

The original B and C matrix contains as much information about the regulatory network as we know. According to the model we showed above, it is possible that we can construct a whole set of equations to describe the behavior of a single mRNA or TF as well as the entire mRNAs and TFs, however the network we have consist of several sub networks which are apparently independent and therefore should be more efficiently analyzed separately, thus I wrote a program to decompose and filter the original network. I eliminated those irrelevant genes and TFs, and decomposed the original whole matrix to be a smaller sub-network that is independent of others. An Algorithm to do this is to consider the whole B matrix as a 'tree' in data structure. Those non-zero elements shows that gene and TF somewhat related, and those zero ones indicate that they are not related. I start from a gene and find out those non-zero elements in this particular row, which means we are trying to find out those TFs that relate to this gene, and then set all those TFs as been visited, the next step is to start from one of those visited TFs, and begin to search for non-zero elements in that column, so on and on until we visit all the genes and TFs that are direct or indirect related. I can manage this simply with the information of B matrix and use the algorithm of traversal of the 'tree'. A visualization of the structure is shown on line at http://fan.gotdns.org/cgi-bin/gene.pl and we just need to upload the B matrix file and

will get the visualized structure we need. (this work is done by my group mate Jianmiao Fan) The result shows that the E.coli gene regulatory network have a major sub network and some smaller independent networks, which is corresponded to another network structure analysis of E.coli regulatory network published in Bioinformatics that shows the whole network could be decomposed to a primary network and some small sub-networks [18]. Compare the result we got to the result from literature, as shown in Figure 4, we can see, that the topology of the two graphs are almost the same, although they come from different database.

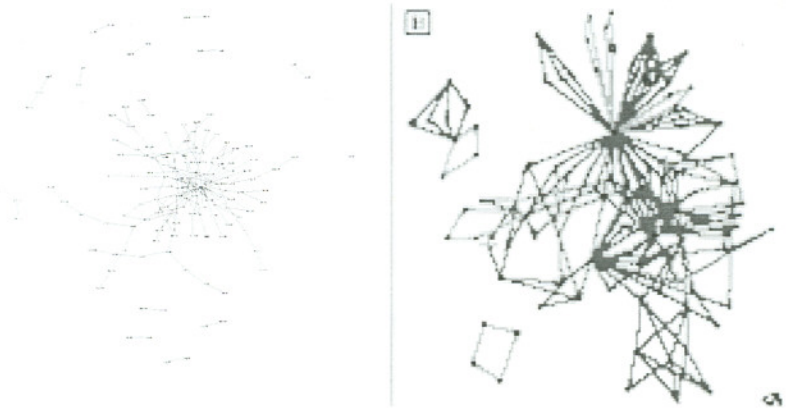

Figure 4, topology of regulatory network, the left is obtained from our database, while the right is from literature [18].

Actually, the primary B matrix, in out case, we have 889 genes and 116 TFs, which is still large, however the fact is although those genes and TFs are somewhat related, only those genes that make TFs will get a feedback and will contribute the nonlinear behavior to the whole system. See our model in detail, if a gene does not

make any TFs, its dynamics is determined by those TFs that regulate it, without giving any feed back, and of course they will not affect the whole system, so those genes could be eliminated, and will not change the nonlinear behavior of the whole system. I have written a program to decompose and filter those networks, the input files are the original B and C matrixes, and the output files are the decomposed and filerred B and C matrixes, whose structure could be seen in Fig 5. We also rearranged the index of the genes and TFs. After decomposing and filtering the network, we get a 67 genes * 71 TFs central network. We know that 60 of 71 TFs are made by those genes that are listed in a total 67 genes, and for those we do not know which gene makes them, I put them as constants.

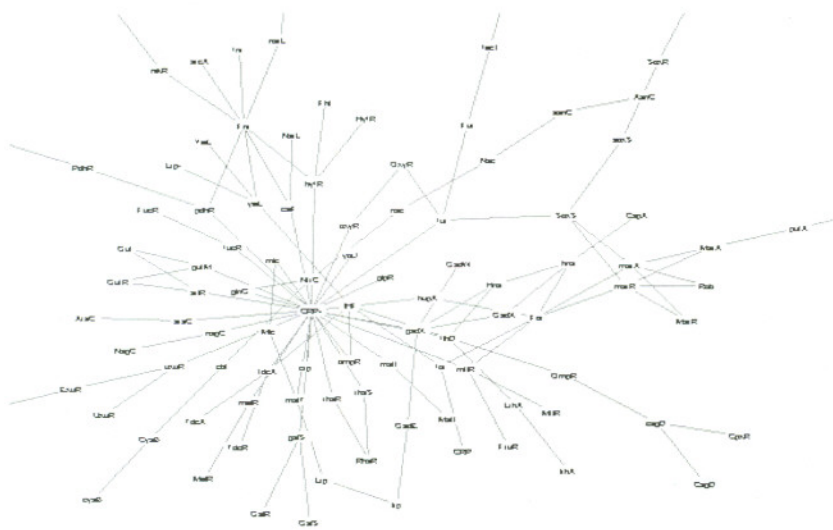

Figure 5, the central independent regulatory network.

4.5 What is AUTO

AUTO is a continuation and bifurcation software for ordinary differential equations [11]. It was developed by Eusebius J. Doedel etc. It is quite powerful to solve Algebraic Systems, Ordinary Differential Equations, Parabolic PDEs and show the

steady states, bifurcation points, and Hopf bifurcations, which is well suited for nonlinear system analysis of cell. We use AUTO to search for nonlinear behavior such as multiple steady states, and bifurcations of limit cycles, in our gene regulatory network. The AUTO software requires two input files: the constant file "r.xxx" and equation file "xxx.f", ("xxx" stands for a user-select name of the file, for example, "xxx" could be gene, and those two files could be "r.gene", and "gene.f" to indicate different problems) the former defines the computation conditions for a particular problem, such as the dimension of the system, the boundary conditions, the free control parameter, the region of the free control parameter that is required to compute, the number of bifurcation points we want to find, the tolerance and continuation step size etc. The later file defines the mathematical form of the problem and contains the Fortran subroutines FUNC, STPNT, BCND etc. which specifies the mathematical equations, the starting point variable values and parameter values, and the boundary condition respectively etc. AUTO requires the starting point a steady state of the system, we call it an initial steady state, thus it can trails out along the steady state bifurcation graph. In order to make this point, Al wrote a script that could read our B and C matrix, and automatically generate equations that describe gene and TFs dynamics according our model, and then create "gene.f" file, "r.gene" file, and another Fortran file "steady.f90" as well, the last file is used to calculate the initial steady state using Monte Carlo Method, that is to say, I randomly generate a set of values indicate the concentrations of those RNAs and TFs, and then let it run for a long time until the system goes to steady state, I think it is the initial steady state, and

write this set of values in a file which is readable by AUTO, thus everything is prepared, and the whole work flow is automatic.

In answering the above 5 questions, it is obvious that I integrate several elements to arrive at an approach to the analysis of complex gene regulatory networks:

- automated literature and database searches to arrive at a putative, albeit incomplete and likely error-prone, regulatory networks;
- a microarray analysis methodology (KAGAN) that allows one to use time series data to correct errors in, fill in gaps of and expand the putative regulatory network;
- a Markov/chemical kinetic model of the gene transciption regulatory networks;
- scripts to automatically transform the above information into a format acceptable to nonlinear phenomenon discovery package AUTO; and
- analysis of the network by AUTO to identify multiple steady states or dynamical attractors in the cellular network.

With the above information and analysis, we are able to generate an automatic methodology to identify multiple steady-states, periodic and other nonlinear dynamical system behaviors, which may be related to cell differentiation or other key cellular phenomena. Since we have a relative complete regulatory network for E.coli, we apply this methodology to E.coli.

5. Results and discussion

5.1 Test the methodology

I first tested the approach to some simple and known results motifs, in order to make sure that the whole work flow works.

Motifs are small, overrepresented topologically distinct regulatory interaction patterns (sub graphs), which have been considered as the basic bricks of the whole regulatory network [18]. Researches show that bacterium like E.coli exists different types of motifs, such as the auto-regulation, feed back loop, and feed forward loop etc. [19], among which auto-regulation is the most common case, and studies shows that it consists half of the regulatory motifs for E.coli [20]. An auto-regulation means a gene makes a TF (gene is transcribed to be mRNA, which is then translated into protein, this protein is a transcription factor, we call this process as gene makes a TF), which regulate – either up or down regulate the transcription of this gene. As for our central regulatory network, we found that 60 of total 71 TFs could be found that are made by those genes in the B matrix, and 53 of which have the behavior of auto-regulation, this number consists 71% of the total genes we have in the central network. For example, gene phoB makes TF ArcA-Phosophorylated that up regulate the transcription of gene phoB, gene fur makes TF Fur that down regulate the transcription of gene fur. We take gene phoB and TF ArcA-Phosophorylated and get a small B and C matrix, to test the work flow.

The dynamic of this one gene one TF system could be described by these two equations:

$$\frac{dR}{dt} = K_{\max} * (\frac{Q*T}{1+Q*T} + \zeta) - \lambda * R$$

$$\frac{dT}{dt} = \alpha * R - \beta * T$$

We set value of $K_{\max}, Q, \lambda, \alpha, \beta, \zeta$ as 1.0, 1.0, 2.0, 2.0, and 0.02 respectively. We solve the above equations in Mathematic and get the normalized RNA concentration ~ Kmax graph below (see Figure 6 left). Also, we get a solution from AUTO, and show the result in Figure 6 right. Compare two graphs, although we got from different ways, (the former is analytical, and the later is AUTO) they look the same, which means the work flow that we explained above works well.

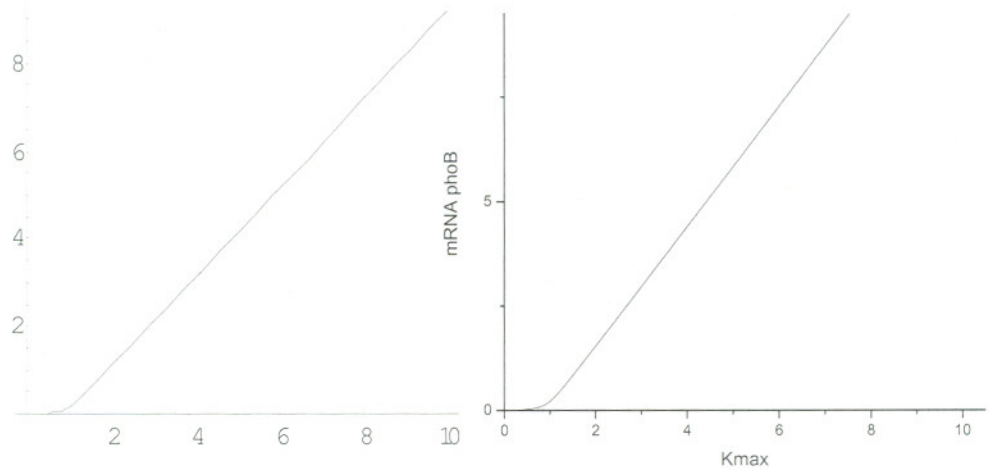

Figure 6,Test of the work flow, the left is obtained from mathematics, and the right is obtained from AUTO

5.2 Application to E.coli

As we said before, we have a relatively complete regulatory network of E.coli, which contains 984 genes and 144 TFs, and we have 100 mRNA microarray data, so, we take these microarray data and run it in KAGAN, we get the equilibrium constants -- Q values for 100 genes and 16 TFs, I then put the whole matrix into my decomposition and filtering program and get a central network of 67 genes and 71

TFs. Unfortunately, we do not have the microarray data for those mRNAs that correspond to the genes in our central network, so, for those gene-TF equilibrium constants – Q values, I take the average value of those we know. I set $K_{\max}, \lambda, \alpha, \beta$ as free control parameters, and write the equations into AUTO according our model.

The result of AUTO is rich and exciting, here I show some of the most important results.

(1). I recorded the steady state of all single mRNA and TF versus the changing of free parameters, as for example, the $K_{\max}$ value. It is possible for us to see the behavior of all the single mRNAs and TFs, which implies that if we are interested in a certain mRNA or TF, we can directly see the nonlinear behavior: does it have multiple steady states, limit cycles, bifurcations etc.? I take gene rcp for example. Gene crp has been studied from 1980s [21], and is considered a very important gene in bacteria, because A) the product of gene crp is TF CRP-cAMP, which regulates over 260 gene transcription in E.coli, among those genes, many of them are response to glucose levels. Research shows that high levels of glucose reduce the levels of cylic AMP (cAMP) within the cell, and conversely, glucose starvation leads to an increase in cAMP levels allowing a molecule of cAMP to bind to CRP [22]. B)The complex of the regulation of crp gene expression by CRP-cAMP attracts a lot interests [21, 23, 24]. The crp gene is regulated autogenously, which means the product of gene crp up regulated the expression of itself. Protein CRP and RNA polymerase bind to the crp regulatory region simultaneously, which suggests a different mechanism for transcriptional repression of the crp gene by CRP-cAMP from that of a typical operator-repressor

model [21]. Figure 7 shows the bifurcation diagram of mRNA crp, and TF CRP-cAMP, both of them shows a clear S-shape, which is not surprising, because TF CRP-cAMP is the product of mRNA crp, so its dynamical behavior should follow that of the mRNA crp, however the S-shape does means a lot, as is shown in the picture, the increase of RNA polymerase activity causes a slightly increase of mRNA crp level and TF CRP-cAMP level, but when the activity reaches point a, both the mRNA level and TF level will suddenly jump up to point b, which is a quite different steady state, after that, when the RNA polymerase activity decrease, the mRNA level and TF level will not drop down to point a, but will decrease slowly until it reaches point c. This kind of hysteretic behavior is typical and considered important for life systems, because it makes the whole system robust to the fluctuation of the surroundings, which means the system is willing to keep what it was until the disturbance of the circumstance reaches to a certain limit point like a and c here.

Well, although RNA rcp and TF RCP-cAMP show a beautiful S-shape graph, it is not always the case that all the mRNAs or TFs should be an S-shape, in fact different mRNAs or TFs show quite different behaviors. Fig 11 shows several kinds of mRNAs and TFs, where mRNA cspA shows a straight line, mRNA gadX shows a curve, mRNA uxuR shows a big S-shape, and mRNA cbl shows a complex graph. The bifurcation diagrams show the number of steady states one can hold for a given physical condition, for instance, RNA cspA shows a single steady state as the change of RNA polymerase activity, while RNA uxuR could be one or three steady states.

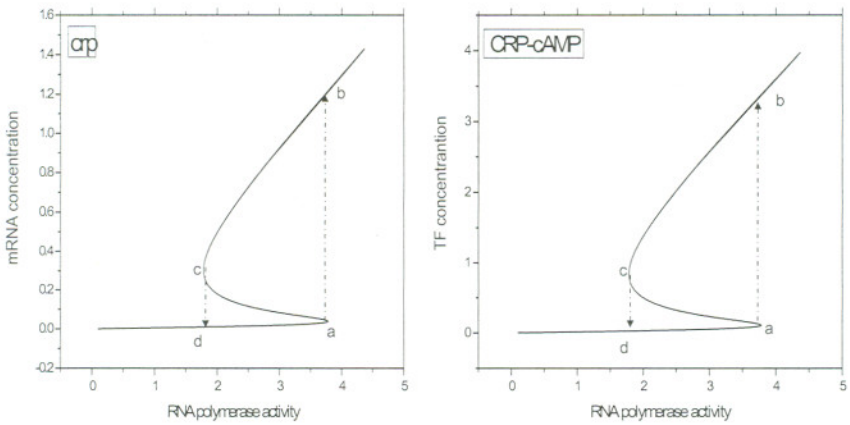

Figure 7 the bifurcation diagram -- S-shape of mRNA crp and TF CRP-cAMP versus RNA polymerase activity

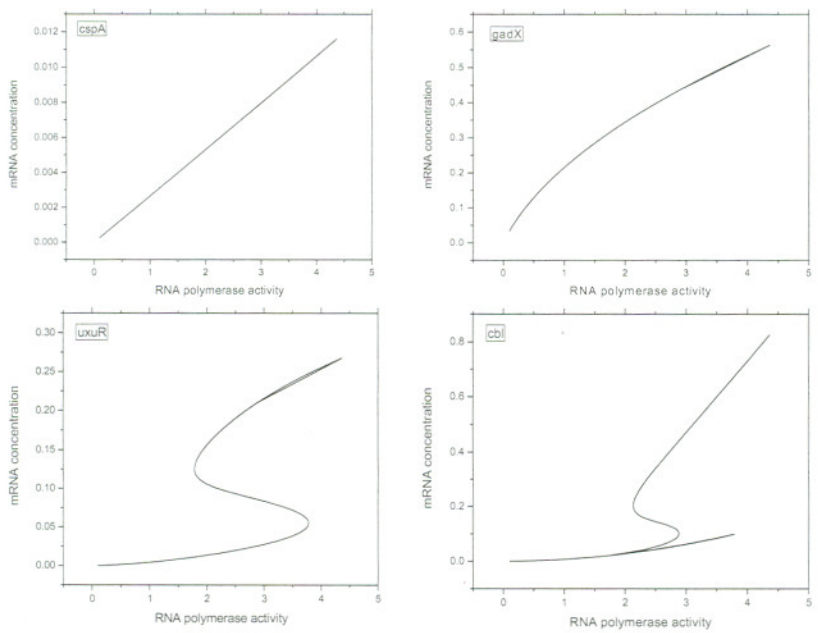

Figure 8, bifurcation diagrams for mRNA cspA(left above), gadX(right above), uxuR(left below) and cbl(right below).

(2). I recorded the steady state of normalized total RNA content. I use L_2

normalization, which is defined as the square root of the sum of the square of all RNA levels. See:

$L_2 = \sqrt{\sum_{i=1}^{N} (mRNA(i))^2}$ where N is the number of mRNA.

To show the normalized concentration of the total RNA contents is meaningful, because the total RNA level represents the cell states. Those RNAs, together with TFs, determines the protein level, glucose level etc. and eventually determines the cell behavior. The result of this normalized content is shown in Figure 9, which is much more complicated than a single mRNA or TF. Actually, the diagram contains several bifurcation points, which comes to be the fact that E.coli cell do exist multiple steady states as the different values of RNA polymerase activity, and I have a quite fine structure of this phenomena.

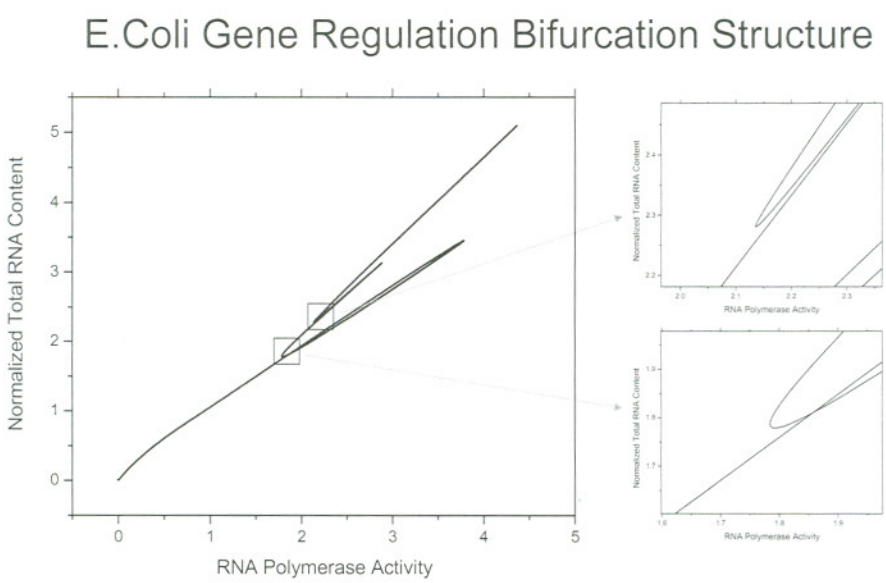

Figure 9, Transcription/translation regulatory network of E.coli shows bifurcation fine-structure as a function of RNA polymerase activity

(3). I changed the free parameter to RNA degradation rate λ, TF creation rate α and degradation rate β to see how the bifurcation diagram changes with the increase of those values, which means the nonlinear behavior versus λ, α and β.

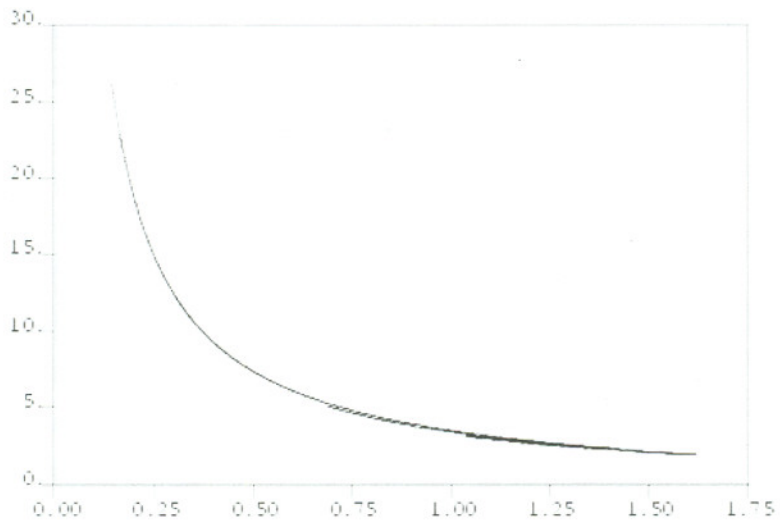

Figure 10 E.coli Gene Regulation Bifurcation Structure, the X-axis is the RNA degradation rate λ, and the Y-axis is the normalized mRNA total contents. Actually, this graph contains a lot loops and bifurcation points when the λ value is between 0.5-1.75.

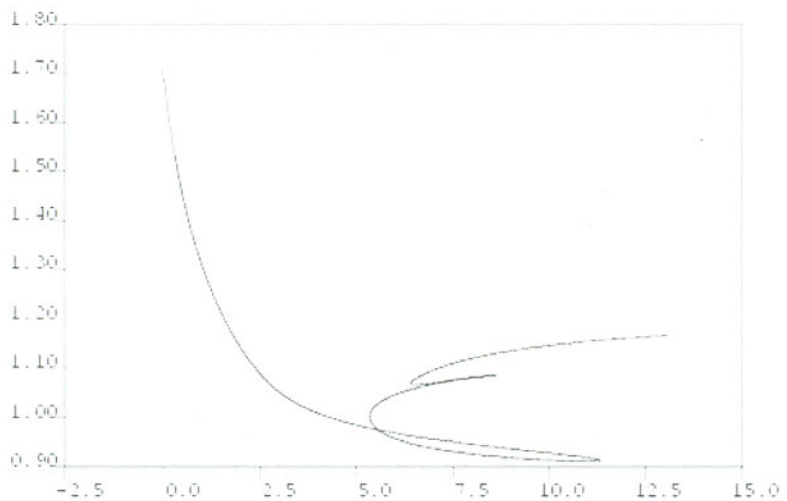

Figure 11 E.coli Gene Regulation Bifurcation Structure, the X-axis is the TF creation rate α,

and the Y-axis is the normalized mRNA total contents.

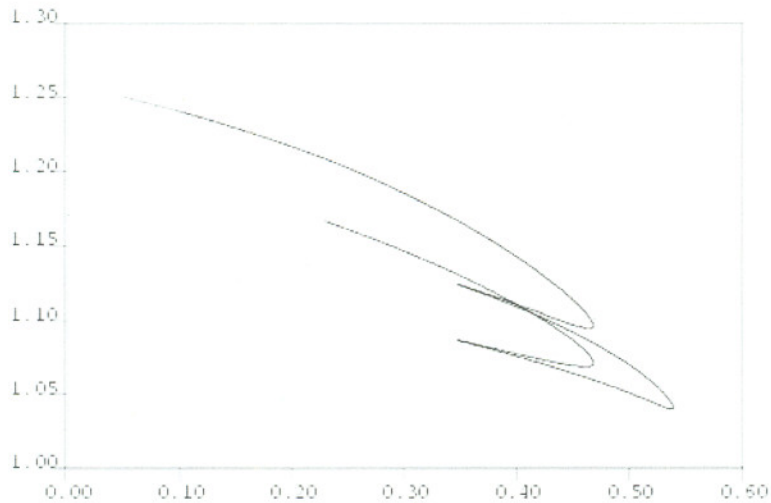

Figure 12 E.coli Gene Regulation Bifurcation Structure, the X-axis is the TF degradation rate β, and the Y-axis is the normalized mRNA total contents.

Meanwhile, I get the data for the bifurcation diagram of single mRNA and TF as I did for RNA polymerase activity $K_{\max}$ however I am not going to show these graphs here in order to save space.

In summary, we use our methodology and get some of the most important results concerning the relationship between the gene regulatory network and nonlinear cell behavior. My data shows how the steady state as characterized by total mRNA contents as well as single mRNA and TF behave as the change of RNA polymerase activity $K_{\max}$, RNA degradation rate λ, TF creation rate α, and TF degradation rate β, so a true case is that we get a 4 dimensional space for steady state diagram. These results at lease contain two quite valuable applications: A) theoretical study of the nonlinear behavior of gene transcription and RNA translation dynamical reaction network. B) the bifurcation diagram of the total mRNA contents suggest an approach

to characterizing a cell steady state, and may provide a method for developing strategy to avoid transitions from a normal cell state to abnormality. The latter is especially important to cancer therapy, which is trying to discover new drug target to block some of the mRNAs or proteins' concentration exceed some certain limit value.

Since we have a method to make the network in more detail, we believe that we will be able to explain how these bifurcation and S-shape come from, as far as we get adequate information about the regulatory network. Also I did some work in trying to explain, for instance, I tried to change the model to consider those dimerized TF as normal TF, and did not find any of those nonlinear behaviors, which suggest dimerization or high complexing may play an important role in nonlinear cell behavior, though we need more evidence to support this argument.

5.3 Stochastic and uniqueness of regulatory networks

As I said above, our gene regulatory network is incomplete so how could we convinced others that our result is accurate enough, in other words, how accurate our results are? In order to answer this question, I take a small test. Suppose we have a model $Y = B * X$, where B is a 1000*200 matrix, and the elements of B is 0, 1 or -1, X is a 200*1 column, and Y is the product of these two matrixes, which is a 1000*1 column. If the true value of column Y, X, and B are Y_star, X_star and B_star respectively, and then we set X unchanged and randomly change part of the B matrix, for each changed B matrix, we have a Y column value, which is different from Y_star. I define an error E as

$E = \sqrt{\sum_{i=1}^{1000} (Y(i) - Y\_star(i))^2}$ (the L_2 norm) which indicates the difference

between column Y and its true value Y_star.

Then I randomly change the B matrix 15,000 times and get 15,000 different E, the maximum E, is called Emax, and the minimum E is called Emin. I then divided 10 average segments from Emin to Emax, and counted the number of tries that have an error between each section. We got a histogram shown in Figure 13, where we changed 1% of the matrix that is 2000 elements. The same kind of histogram could be obtained when we change 0.01% to 10% of the matrix, but both Emax and Emin would increase as the increase of the changing percentage, as shown in Figure 14. These results could be reduced to two points: 1) both the absolute error and relative error are inclined to stay at the average with a statistical bell shape to both sides instead of going to the extremes of minimum or maximum error. 2) the more uncertainty the matrix has the larger error will be. We believe the same kind of behavior would happen in our gene regulatory network analysis, so the answer to how much information we have to know or when is a critical point that we can say that our result is accurate enough depend on how accurate is the experimental measurements. The more accurate experimental measurements requires a more comprehensive understanding of the network, otherwise, a roughly construct network could give us enough confidence to our results because the experimental measurements take up the majority of the error.

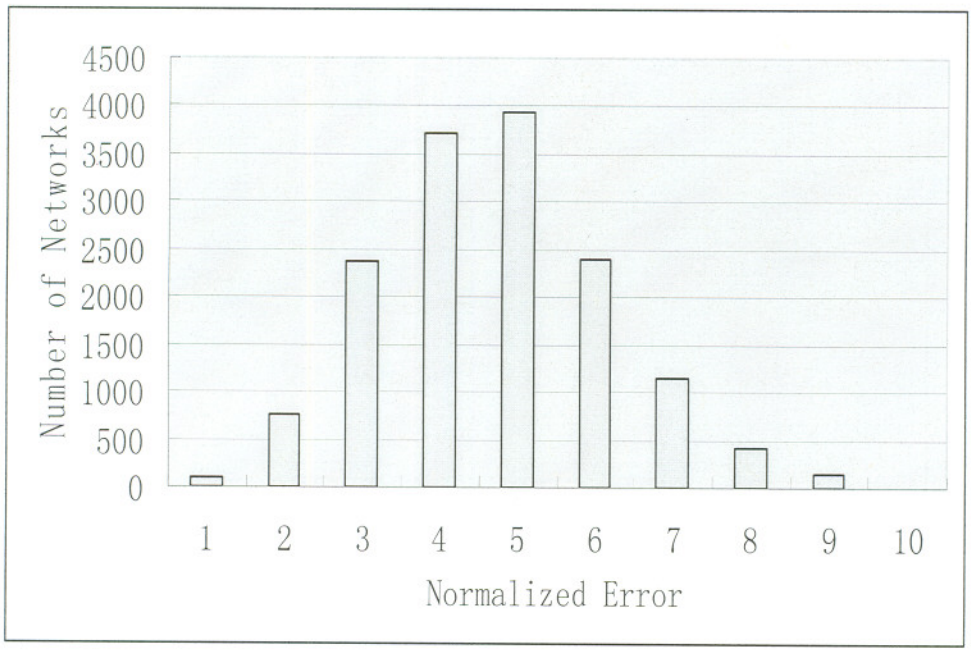

Figure 13, the histogram of number of random networks versus normalized error (from Emin to Emax)

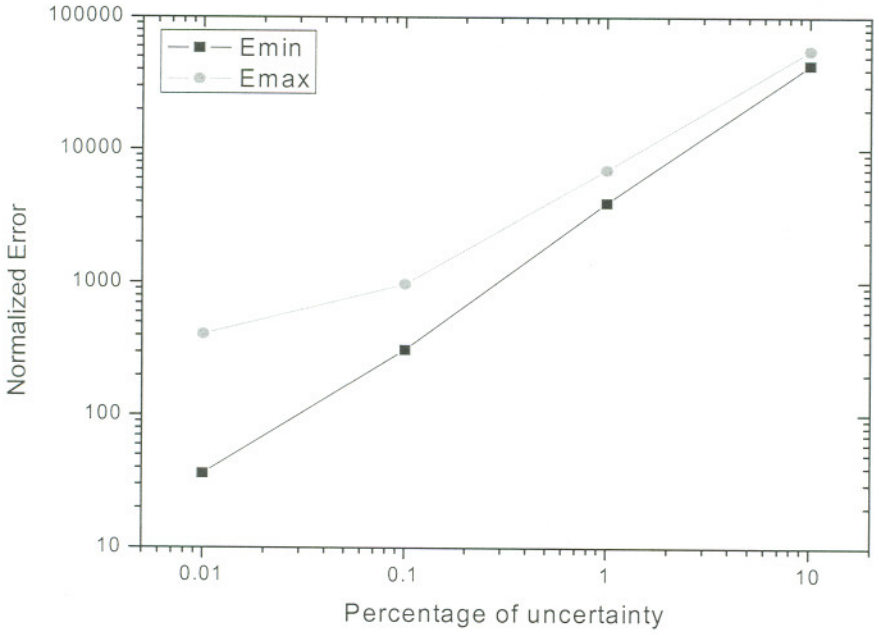

Figure 14, the normalized error change with the percentage of uncertainty, we use log scale for both the axis.

Conclusion

In conclusion, we developed a general methodology for nonlinear analysis of gene transcriptional regulatory network. We proposed a reasonable model to describe the transcription and translation reaction dynamics and constructed an automatic work flow for this nonlinear analysis. I applied this methodology to E.coli and got some quite interesting results, which not only prove that E.coli regulatory network do exist nonlinear behavior but also got a fine structure of these behaviors as well. The results contain the bifurcation information of the total RNA contents as well as that of every single mRNA and TF. I also selected different control parameter to see how a cell behaves in changing with different free parameter. At last, I discussed the critical point problem, to see how convincible our results are. We believe our methodology is of great importance to understand the relationship between the DNA sequence and the nonlinear cellular behavior, when we applied this general methodology into a real human cell system. Consider, even for a system as simple as E.coli, our method discovered multiple distinct steady states followed from the E.coli genome, and therefore, one can reasonably expect that when we turn to the human genome with its 25,000 genes we should expect of a multiplicity of steady states which reflect the many distinct human cell types that follow from the same genome, further more, the discover of RNA and protein bifurcations that are controlled by some physical conditions will be quite useful in drug discovery and cancer therapy.

Acknowledgements

Tuncay for his patience in helping me to understand our dynamic model for gene regulatory network. I thank Professor *Michael S. Jolly* for his kind help with AUTO program. I thank *Alaa Elie Abi Haidar* for his help to write scripts that could generate equations according to our model. I thank Jianmiao Fan for his help in visualizing the networks. I thank *A. Sayyed-Ahmad* and *K. Tuncay* for developing the KAGAN program that calibrates rate constants and equilibrium constants that are required in our model. I thank *Kranthi Varala, Lisa Ensman, Timothy Burger,* and *Le-Shin Wu* for their hard working on the database that supports my work. I thank *Kyle Hubbard*, and *Frank Stanley* for their help of my computer problems. I thank all group members at Center for Cell and Virus Theory, Indiana University for giving me help during my research work of this project.

Reference:

1. Mendel,G.Versuche über Pflanzen-Hybriden.*Verhandlungen des naturforschenden Vereines,Abhandlungen, Brünn* **4**, 3–47 (1866).
2. Avery,O. T.,MacLeod,C.M.& McCarty, M. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. *J. Exp.Med.* **79**, 137–158 (1944).
3. Watson, J.D. & Crick, F. H. C.Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171**, 737 (1953).
4. Nirenberg, M.W. The genetic code: II. *Sci.Am.* **208**, 80–94 (1963).

5. Jackson,D. A., Symons, R. H. & Berg, P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. *Proc.Natl Acad. Sci. USA* **69**, 2904–2909 (1972).

6. The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

7. Sanger, F.& Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J.Mol. Biol.* **94**, 441–448 (1975).

8. A vision for the future of genomics research Nature Vol **422** 24 (2003)

9. Gene regulation at the single-cell level, Science Vol **307** 25, 2005

10. Dynamical analysis of gene network requires both mRNA and protein expression information. Kelvin H. Lee, Metabolic Engineering **1**, 1999

11 AUTO 97: continuation and bifurcation software for ordinary differential equations. Eusebus J. Doedel, Alan R. Champneys, etc.

12. http://www.uni-giessen.de/~gx1052/ECDC/ecdc.htm

http://www.ncbi.nlm.nih.gov/Entrez/

13. http://biodynamics.indiana.edu

14. Weitzke, E. and P. Ortoleva. 2003. Simulating cellular dynamics through a coupled transcription, translation, metabolic model, *Computational Biology and Chemistry*, Vol **27**(4-5), 469-481.

Sayyed-Ahmad, A., K. Tuncay and P. Ortoleva. 2003. Toward Automated Cell Model Development through Information Theory, *Journal of Physical Chemistry*

A **107**, 10554-10565.

15. Karyote Genome Analyzer

16. A.Brazma, A.Robinson, G.cameron, M.Ashburner. One-stop for microarray data. Nature **403**:699-700, 2000

17. A review of DNA microarrray data analysis.

18. Aggregation of topological motifs in E.coli transcriptional regulatory network, R. Dobrin etc. Bioinformatics 2004, **5**:10

19.Network motifs in the transcriptional regulation network of *Escherichia coli.* Shai S. Shen-Orr1, Ron Milo, Shmoolik Mangan & Uri Alon, Nature Genetics April 22 2002.

20. Network Motifs: Simple Building Blocks of Complex Networks. R.ilo,1 S. Shen-Orr,1 S. Itzkovitz,1 N. Kashtan,1 D. Chklovskii,U. Alon1*, Science **298**: 25, 2002

21. Autoregulation of the Escherichia coli crp gene: CRP is a transcription repressor for its own gene. Cell 1983 Jan; **32**(1): 141-9

22. www.biochemistry.bham.ac.uk/sjwb/crprot.html

23. Study of the regulation of crp gene expression in Escherichia coli K12

24 http://www.uni-giessen.de/cgi-bin/cgiwrap/gx1052/ecgetrf.pl?cG00253