Modeling Pathways of Cell Differentiation in Genetic Regulatory Networks With Random Boolean Networks

by

Sheldon Ray Dealy

B.S., Oregon State University, 1995

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Computer Science

The University of New Mexico

Albuquerque, New Mexico

May, 2005

Dedication

To my wife Sandy, for all the support and understanding.

Acknowledgments

I would like to thank my advisor and committee member, Dr. Stuart Kauffman, for teaching me the definition of scientific research and for setting the highest of academic standards.

I would like to thank my committee chair Dr. Christopher Moore for giving me his honest insights and observations.

I would also like to thank committee member Dr. Robert Veroff for sharing with me his enthusiasm for Computer Science.

This research has been partially supported by grants from the National Science Foundation (PHY-0417660) and the National Institutes of Health (GM070600-01).

Modeling Pathways of Cell Differentiation in Genetic Regulatory Networks With Random Boolean Networks

by

Sheldon Ray Dealy

ABSTRACT OF THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Computer Science

The University of New Mexico

Albuquerque, New Mexico

May, 2005

Modeling Pathways of Cell Differentiation in Genetic Regulatory Networks With Random Boolean Networks

by

Sheldon Ray Dealy

B.S., Oregon State University, 1995

M.S., Computer Science, University of New Mexico, 2005

Abstract

It has long been known that some cells differentiate into other cell types. Under the assumption that living cells can be modeled using non-linear dynamical systems, then cell types are attractors. If a cell type is an attractor, then a series of steps leading from one attractor to another is a pathway of differentiation. As non-linear dynamical systems, random Boolean networks were analyzed for their suitability in a first attempt to model pathways of cell differentiation as model genetic regulatory networks. We hypothesize that perturbations of Boolean networks share some key behaviors of genetic regulatory networks in the process of differentiation. Furthermore, several solutions to the problem of mapping the discrete time nature of Boolean networks to the continuous nature of genetic regulatory networks are presented. The

findings suggest that Boolean networks are able to support some predictions about the pathways of differentiation in living cells. These predictions should be testable using gene array techniques.

Contents

List of Figures

List of Tables

Glossary

Canalizing Functions	Boolean functions where one value of at least one of the inputs controls the output regardless of the other input values.
Cell Surface Markers	Cell surface markers are proteins that present themselves on the surface of the cell giving an indication of what is going on inside the cell.
Deoxyribonucleic acid (DNA)	Molecules of genetic material inside the cell, segments of which have been identified as genes.
Dynamical system	A system that changes over time according to a fixed set of rules.
Gene Arrays	A Gene Array is a slide or wafer made from glass or silicon upon which 100's to 1000's of DNA fragments or genes are deposited in a regular ordering. This ordering allows the spots to be made available to probes which can detect and analyze the levels of expression in the genes over time.
Genetic Regulatory Network	A genetic regulatory network is a collection of genes in a cell (presented on strands of DNA) that interact

	with each other to determine the rate at which the genes are transcribed into RNA.
Hamming Distance	The number of bits which differ between two binary strings.
Histology	The scientific study of plant or animal tissues.
Medusa Networks	A connected network in which relatively few nodes, called the "Medusa head", control the actions of a majority of nodes in the network.
Morphology	The form and structure of an organism.
Nonlinear Dynamical System	A dynamical system is nonlinear when its output is not proportional to its input, i.e., when adding the inputs together and then operating on them produces a result that is different from operating on the inputs separately and then adding them together.
Power Law	The power law refers to the relationship between two quantities x and y such that their relationship can be written in the form $x = cy^n$ for some exponent n and a constant c.
RBN	See Random Boolean Network.
Random Boolean Network	A random Boolean network is a network containing N nodes. Each node has K input edges. For each node, the output value for each of the possible 2^K input permutations is determined by a Boolean function. The Boolean function takes as inputs the values of true or false from each of the K inputs

	and has an output, for each of the possible input permutations. The output values for the Boolean function are chosen at random during initialization and fixed for the life of the network. The initial state of the network is determined by choosing the value for each node, either true or false, at random. Subsequent network states are generated by feeding the present state of the network to the Boolean functions.
Regulatory Gene	A gene which produces a substance which inhibits the production of a protein.
Scale Free Networks	Networks which have a power law distribution for the number of connections on each node.
Stem Cells	Stem cells are undifferentiated cells which can both divide and differentiate into other cells.
Structural Gene	A gene which controls the production of a protein or peptide.
Transcription Factors	A transcription factor is a protein that binds DNA at a specific promoter site, where it regulates transcription. Transcription factors are regulated by other proteins.

Chapter 1

Introduction

1.1 Background

Since the early 1960s when Jacob and Monod postulated that cell differentiation is accomplished by the activation of specific groups of genes [8, 9], it has been accepted that genes can regulate one another's activity. Jacob and Monod proposed a model genetic circuit to explain the mechanics of cell differentiation [9]. For example, given two regulator genes RG_1 and RG_2, protein E_1 expressed from RG_1 represses the protein E_2 which is expressed from RG_2. The protein E_2 expressed from RG_2 represses the expression of protein E_1 from RG_1. This mutual repression produces one of two steady states; either RG_1 *on* and RG_2 *off* or RG_2 *on* and RG_1 *off* (see fig. 1.1). Such steady states are considered to be attractors of the genetic regulatory network dynamics. Jacob and Monod proposed that that these two steady state attractors correspond to two stable patterns of gene expression which might correspond to two cell types.

The hypothesis that cell types are attractors is credible [7, 10, 12] and of central importance. If we presume that a gene's activity is binary, to be restricted as either

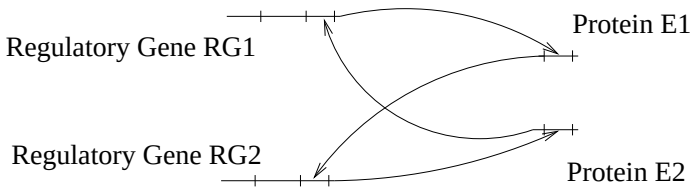

Figure 1.1: A model genetic circuit.

being on or off, then a genetic regulatory network with N genes has 2^N number of possible states[1]. Now the human genome has an estimated $25,000$ genes. There are $2^{25,000}$ possible configurations of gene activities in the human genome which is approximately 10^{7500} states where there is an estimated 10^{80} particles in the universe. Given that there are some 265 cell types in the human body, we can infer from the vast state space that only some configurations of genes can represent cell types. Therefore, genetic regulatory networks must restrict themselves to small subsets of the state space where each subset corresponds to a cell type. Thus it is natural to postulate that cell types correspond to attractors. Such an attractor should be able to approximate time scales of a cell orbiting the attractor, which to be biologically plausible would take from minutes to days. Random Boolean Networks (RBNs) in the ordered and critical regimes have these and other interesting properties [10, 11, 12].

While we understand how to induce cells to differentiate and why some cells differentiate naturally, little is understood about cell behavior along the pathway of differentiation as one cell type changes to another. For example, HL60 leukemia cells can be induced to differentiate to a benign state after the application of dimethylsulphoxide [2], but it is not known what happens during the process of differentiation. It is thought that with the proper computational model, we may understand the biology better. Through better understanding, predictions about the biology model can be made. Random Boolean networks may be able to fill in some of the pieces.

[1]A state in the network is defined to be a snapshot of all the node values in the network, each node being either on or off.

The findings presented here describe the first examination of pathways of cell differentiation in model genetic regulatory networks where cell types are assumed to be attractors. Random Boolean networks were found to be a good choice for modeling limited kinds of behavior in the differentiation of living cells. The results obtained offer for consideration what we might expect to find if we were to look at pathways of differentiation in living cells. Conclusions we have reached from these results:

- The large percentage of homeostatic transients for model cells in the ordered and critical regimes appeals to our intuition that cell types are stable attractors. After being perturbed, most homeostatic transients return to the attractor in one time-step. This behavior can be predicted in living cells.
- Pathways of differentiation often fuse and flow together in the Boolean network model which suggests these fusing pathways may be found in cell differentiation.
- For pairs of transients which terminate at the same attractor, the Hamming distance gradually decreases with proximity to the attractor. This behavior suggests that differentiating cells which share the same terminal cell type become more alike as they approach the attractor, even if they join the attractor at different states on the state cycle.
- The Hamming distance between successive states on model transients are found to decrease monotonically as an attractor is approached, substantiating recent results found in cells [7].
- The ratio of fused homeostatic transients to unfused homeostatic transients is greater for model cells in the ordered and critical regimes than it is for cells in the chaotic regime. Where this ratio reverses may be a marker for the critical regime.
- A large percentage of genes in chaotic models never change state on pathways of differentiation, where the reverse is true for the ordered and critical regimes.

If cells are in the ordered or critical regimes, then we would expect to find a wide distribution of the percentage of time which genes change state on the transient.

- The amount of gene variation over time in the Boolean network model suggests a linear relation between transient lengths in RBNs and the number of times a gene alters activity. This idea may be key to the problem of mapping discrete steps in a synchronous Boolean network to the continuous levels of gene expression within a cell.

Most of these conclusions are readily testable with living cells using gene array techniques.

1.2 Overview

The structure of this thesis is as follows. Chapter 2 describes how genetic regulatory networks operate, gives a short overview of how and why cells differentiate or change to another cell type, and concludes with a description of how cell differentiation is currently measured. In the Chapter 3, random Boolean networks are discussed and reasons why Boolean networks make a good model for genetic regulatory networks are shown. Chapter 4 discusses how the experiments were conducted and what was measured. In the Chapter 5, we address the outcome from the experiments and what kinds of behavior was observed. In Chapter 6 the meaning of the results is discussed and what might be taken away from modeling pathways of differentiation using Boolean networks. In Chapter 7 the experiments are summarized and the conclusions on the suitability of the classic Random Boolean Network as a model for the pathway of cell differentiation are presented. Finally in Chapter 8, we discuss possible ways to extend the experiments and the possibilities for improving the model.

Chapter 2

The Biology

2.1 Genetic Regulatory Networks

Cells contain molecules of deoxyribonucleic acid (DNA), the chemical building block for genetic information. Each structural gene is a segment of DNA which is transcribed into ribonucleic acid (RNA), which is translated into a protein. The DNA molecules in a human cell contain about $25,000$ genes. These genes interact with each other and other substances within the cell. The interactions induce the genes to express proteins which in turn can modify or change their own or another gene's behavior. This web of interactions reveals the connections of the genetic regulatory network. These interactions control the rate at which genes are transcribed into messenger RNA *(mRNA)*. Transcription is the vehicle which defines the cell type and enables cells to differentiate into other cell types. The genes collectively, with other genetic material make up the DNA in the cell. These genes are in a continuous level of activation, from being fully expressed (turned on), to being not expressed (turned off).

Each of the different cell types in the body contain an identical set of genes. There-

fore, each cell in the body contains a copy of the same genetic regulatory network. A cell type is defined by the particular subset of genes that are expressed [14]. The initial cell type in the human body, the fertilized egg or zygote, divides multiple times and differentiates into 265 [1] different cell types.

2.2 Cell Differentiation

Cell differentiation is the process by which a cell of one type becomes a cell of another type. The process of differentiation is initiated by the expression of specific genes within the cell. Small perturbations to the genetic regulatory network, such as the deviation of a gene's activity by a hormone, can force a cell to differentiate from one cell type to another. Differentiation of cell types can occur as a natural process, for example, embryonic stem cells differentiate into a variety of cell types. Cells may also be induced to differentiate into different, sometimes abnormal, cell types through the use of external stimuli such as chemicals. Likewise, abnormal cell types may be similarly effected to differentiate into normal cell types [2].

2.2.1 Measurement of Cell Differentiation

Historically, cell differentiation has been assessed histologically, that is, by staining the cell with certain stains or by observing the morphology. Research into the actual process of cell differentiation has been recently facilitated by using molecular probes, such as gene arrays. Gene arrays, or micro arrays can be thought of as a grid of DNA spots on a slide. The spots are treated with RNA derived from a cell population which determines the relative abundance of the RNA from different genes. The pathway of differentiation can then be measured by taking a survey of the spots at specific time intervals during cell differentiation.

Chapter 3

Model

3.1 Boolean Networks

Boolean networks, like living cells, are non-linear dynamical systems capable of exhibiting complex behavior [12]. A Boolean network can be thought of as a directed graph where each node in the graph evaluates to a binary value of one or zero. Random Boolean networks (RBNs) are networks of N nodes and K input edges for each node[1], where each of the connecting edges are generated randomly. To determine how state transitions in Boolean networks are facilitated, each node is equipped with a Boolean function (fig. 3.1). Inputs for the Boolean function are values of the nodes sending an edge to a given node. The output value of each Boolean function depends upon the function definition and the Boolean values of the input nodes, either a one or a zero. The wiring of the graph and the output of the Boolean functions remain static for the lifetime of the network.

The state of a Boolean network is a snapshot of the node values taken as a whole. Generating subsequent states in the Boolean network is determined by feeding the

[1]Random Boolean networks are also known as N-K networks.

present state of the network through a set of Boolean functions which generate the next network state deterministically. Boolean networks are synchronous; their activity is analogous to a mechanism being governed by a central clock. Each time the clock ticks, all genes examine their inputs and the Boolean functions concurrently update to produce the next state of the network. The network can be visualized as an interconnection of light bulbs, turning each other on and off with every change of state.

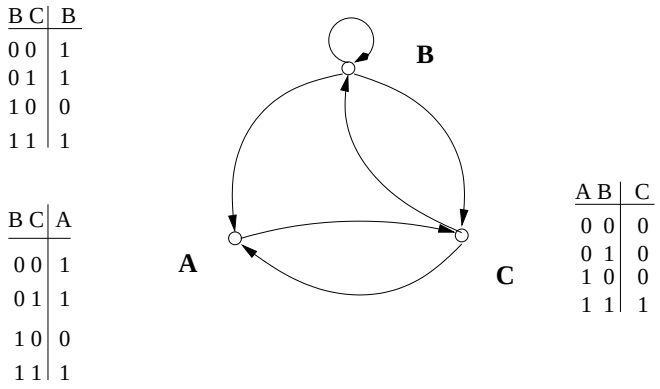

Figure 3.1: A simple Boolean network with $K = 2$ inputs.

Because the state space of a Boolean network is finite, a sequence of states generated by the Boolean functions will eventually repeat itself. A sequence of states which repeats itself is referred to as a state-cycle. State cycles in Boolean networks can be regarded as attractors. State cycles containing a single state are defined as point attractors. A state cycle which contains more than one state is referred to as being cyclic or periodic. The collection of states in a dynamical system which leads to a specific attractor plus the attractor is called the basin of attraction. Depending upon the configuration and implementation, a single Boolean network may have many such basins of attraction.

Our working hypothesis is that cell types are attractors and therefore can be modeled by the attractors exhibited in random Boolean networks. If cell types are attractors,

then getting from one attractor to another becomes the model for a pathway of cell differentiation. In biology, such a transition can occur, for example, from exogenous stimulation which perturbs a gene outside of its normal state. In random Boolean networks, pathways of differentiation are initiated by perturbing or toggling node values. Hence, a perturbation of a node can move a Boolean network into another basin of attraction.

Any pathway leading to a state in an attractor is called a transient. Following a perturbation of an attractor, a path or transient which flows back to the same attractor is called homeostatic. Conversely, a transient which terminates at an attractor different from the source attractor is non-homeostatic. The length of a transient is the number of state transitions between the source of the transient and the destination attractor.

The strength of the Boolean network model is its simplicity. From the simple implementation of an RBN, complex behavior can be observed. After perturbing a node on a state of a state cycle, the Boolean functions generate zero or more transient states which the network will pass through before reaching another state cycle. The transient states then represent a model pathway for cell differentiation as a cell changes from one cell type to another.

Random Boolean networks are one of a variety of models which could have been implemented to represent genetic regulatory networks. For example, Scale-free or Medusa networks[2] could also have been used to model genetic regulatory networks. The random Boolean network model was used because RBNs are a simple way to implement collections or ensembles of networks. Generation of Boolean networks by randomly constructing the wiring and the functions allow us to obtain random samplings of ensembles of networks for different K values.

[2]Medusa networks are described in Chapter 8.

3.2 Approximating the Biology

One of the goals of these experiments was to achieve a credible mapping between Boolean networks and genetic regulatory networks. Such a mapping would allow a state from a transient in a Boolean network to correspond to the state of a cell at some point in the process of differentiation. We hypothesized that it might be possible to map the discrete steps of a Boolean network to the continuous nature of a genetic regulatory network by allowing the total number of gene[3] activity changes in a transient to be a stand-in for transient length.

A possible way to achieving this mapping is by tracking the percentage of time each gene is in a state different from the unperturbed state. Suppose that a given gene in state cycle A is on and we flip it to off. Then the gene shows up in state cycle B. Calculate the fraction of states the gene is off. If the transient between the state cycles is of length ten and the number of states for which the gene is off is six, then the percentage of difference is 60%.

Another potential way to reach a mapping is by tracking the *incidence* of gene change or gene flips over a transient. This is computed by summing the product of the genes that change state x times over the length of the transient multiplied by x; which is the number of genes which flip once, plus the number of genes which flip twice times two, plus the number of genes which flip thrice times three, ... e.g.

$$M = \sum_{1}^{T_{len}} NumOfGenesWhichChange[i] * i \tag{3.1}$$

As will be shown in the results, this method yields a linear relationship between the number of gene variations and the transient length.

[3]For purposes of the Boolean network model, we refer to nodes and genes interchangeably as having equivalent meaning.

Chapter 4

Methods

4.1 Network Construction

The model starts with the implementation of a Boolean network of size N nodes, a value K which determines the number of inputs per node, and a seed to generate a series of pseudo-random numbers to determine the configuration of the network. The K input connections to each node are chosen at random from all nodes in the network. Network connections remain static for the life of the network. To determine the output value, each node in the network requires a Boolean function. The output value for each Boolean function is chosen at random to pair with each of the 2^K possible input configurations. Given K inputs, there are 2^{2^K} possible Boolean functions.

The network is initialized by setting the value of each node at random. Subsequent states in the network are generated using the Boolean functions. Network states are generated synchronously, meaning that all nodes in the network are updated simultaneously. Each state of the network is recorded and compared with previously generated states. When a duplicate state is found, a state cycle or attractor has been

reached. The newly discovered attractor is compared against previously recorded state cycles to determine if the attractor has been previously logged. If a previously undiscovered state cycle is exposed, a representative state is recorded to compare against the discovery of subsequent state cycles. Each gene of each state in the new state cycle is flipped or perturbed, one at a time to obtain all single unit perturbations from the attractor and subsequent states are generated using the Boolean functions. The functions generate a transient or pathway leading either back to the original attractor or to another state cycle.

New attractors discovered from the transients are also recorded and perturbed. When there are no more perturbations to perform, a new initial network state is generated at random and again the same Boolean functions use the network to generate additional states. This process of creating initial network states, discovery of state cycles, and perturbation continues until a user configurable limit of initial random states is reached at which time the experiment is considered complete. When the experiment is completed, all remaining data are written out to file. If additional investigations are required, the seed value is incremented, a new network is generated and the process repeats with a different set of Boolean functions and a new initial random configuration.

Because each random Boolean network has 2^N states, only a small fraction of states may be stored in memory for large N. Keeping data structure size to a minimum was the primary consideration. Wherever possible, network states were represented as bit vectors and a minimum number of states were kept in memory. Transient statistical data was written in compressed form to file and purged from memory as soon as possible. As K increases the average number of states in state cycles grows exponentially, hence the average amount of time required to process each state cycle grows exponentially with K as well.

4.2 What was Measured

The transient states between attractors in RBNs represent a cell in the process of differentiation and the destination attractor represents the terminal cell type. In order to make predictions about behaviors of living cells in the differentiation process, it is important to know how transients behave between attractors. It is intuitive to think that as a transient approaches an attractor, it becomes more like the attractor. If a transient does become more similar to the attractor, then it follows that pairs of transients approaching an attractor also become more similar to each other. The Hamming distance between subsequent states on a transient and the Hamming distance between pairs of transients as an attractor is approached were measured to prove or disprove these intuitions. The Hamming distance measures the size of the difference between two states in a given network. For example, the distance between two example states of a 4-gene network $\{1, 0, 1, 1\}$ and $\{1, 1, 0, 1\}$ is 2. The more similar two states in a network are, the smaller the Hamming distance.

At the termination of each run, a summary of statistics was produced. The distribution of transient lengths was calculated. The mean and standard deviation of gene activity changes over each transient length was recorded and written to file. Subsequent to a number of runs, a software utility was used to perform an aggregation of statistics. During this post-processing phase, Hamming distances between pairs of transients leading to the same attractors were recorded. The phenomena where transients join or fuse and flow together before reaching the first state on an attractor was also measured to provide predictions about the behavior of cell differentiation.

4.3 Testing the Network Dynamics

For a non-linear dynamical system to be complex suggests that it is neither completely regular nor irregular, but having qualities of both [4]. Chaos is the term used to describe irregular behavior in non-linear dynamical systems. Random Boolean networks exhibit increasingly irregular behavior for larger values of K. It appeals to our intuition that the model exhibits complex, as opposed to chaotic or entirely ordered behavior to represent cells as attractors. Chaotic behavior in cells suggests that cell types are unstable. On the other hand, behavior that is too ordered may not allow for the flexibility that differentiation requires.

A conspicuous feature of RBNs is that these networks behave in three broad regimes; the chaotic, the critical, and the ordered regime. Network dynamics needed to be tested to ensure that the networks produced fall within established criterion for complexity in Boolean networks. The accepted way of measuring the level of complexity found in an RBN is to measure the Derrida curves for each of the K values under consideration and compare the outcome against previously published results. The results substantiated that for $K = 1$, $K = 2$, and for $K > 2$, networks demonstrated ordered, critical, and chaotic behavior respectively. A Derrida curve is made by generating and initializing an RBN. From the initial RBN state, a copy is made and a percentage of genes from the copy state are perturbed. The normalized Hamming distance between the two states is labeled as δt for time step zero. Using the Boolean functions, calculate the next time states for both the original state and the copy state. The normalized Hamming distance between the two successor states label as $\delta t + 1$, for the next time step. The process is repeated for different percentages of perturbed genes and different random starting states. The mean of the results is plotted as a line $(\delta t, \delta t + 1)$.

For networks which generate lines that are above the main diagonal for small values

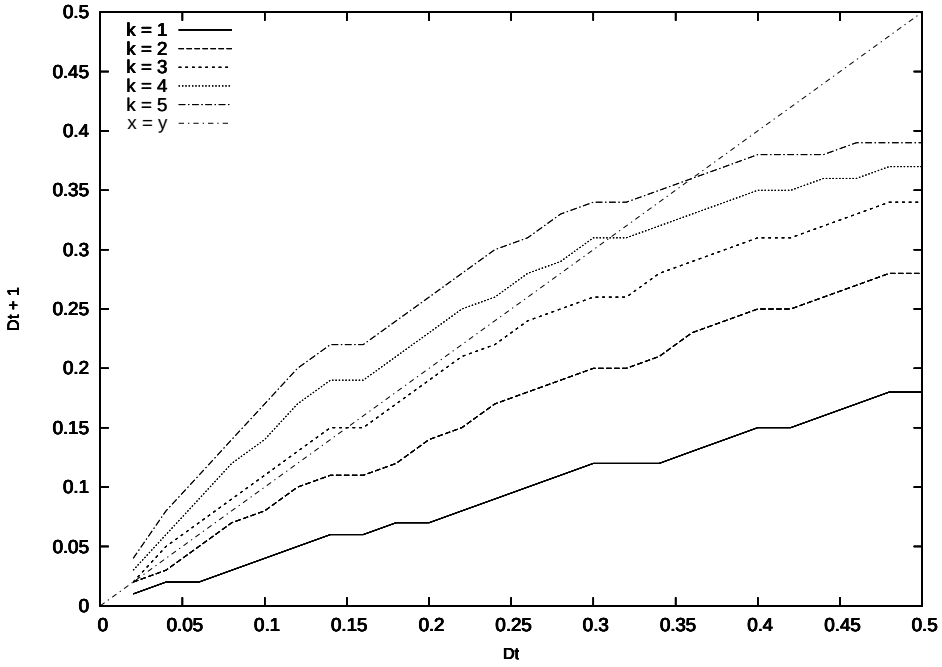

Figure 4.1: Derrida plot, $K = 1, 2, 3, 4, 5$.

of δ_t, it means that the trajectories are diverging from one another in state space and are therefore in the chaotic regime. For networks which generate lines that are below the main diagonal, it means that the trajectories are converging to each other in state space and are therefore in the ordered regime. Networks which generate lines near the main diagonal for small values of δ_t are neither diverging or converging in state space, this behavior is a marker of the critical regime [3].

Figure 4.1 affirms that for $K = 1$ the networks generated are in the stable or ordered regime, for $K = 2$ the networks are in the critical regime, and for $K = 3, 4, 5$, the networks are in the chaotic regime.

Chapter 5

Results

5.1 Homeostatic vs. Non-Homeostatic Behavior

The initial property measured was the number of modeled transients which return homeostatically to the perturbed attractor compared with transients which terminated non-homeostatically at another attractor. The percentage of homeostatic transients was found to be inversely proportional to the size of K, decreasing as K increases. Conversely, the percent of transients which returned homeostatically to the source attractor increased with N (see table 5.1). This is consistent with the idea that cell types are stable attractors. Networks where a single attractor was found were omitted from the calculations because in the case of networks containing a single attractor, non-homeostatic behavior is impossible.

Table 5.1: Percent of homeostatic transients as a function of N and K.

	$N = 10$	$N = 15$	$N = 20$	$N = 25$	$N = 30$	$N = 35$	$N = 40$
$K = 1$	71	82	83	86	88	89	93
$K = 2$	61	63	69	69	70	71	80
$K = 3$	56	58	59	61	63	64	67
$K = 4$	55	57	58	58	62	62	62

5.1.1 Transient Length versus. Percentage of Transients

To compare the relationship between homeostatic transients and non-homeostatic transients, transient lengths were compared with the number of transients occurring for $K = 1, 2, 3, \ \& \ 4$. Figures 5.1-5.4 show the distribution of transients of different lengths as K varies for homeostatic versus non-homeostatic transients.

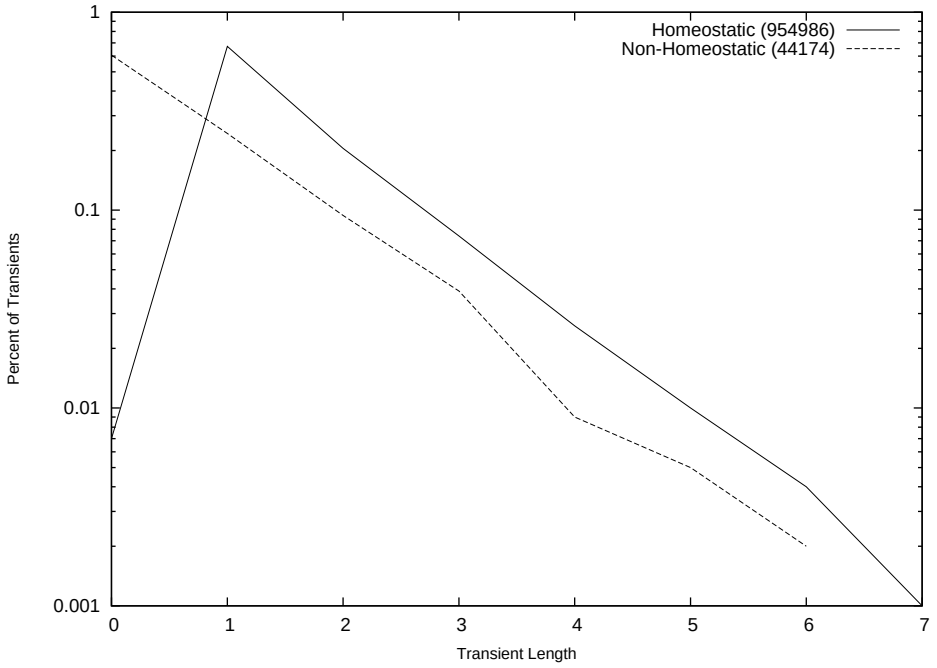

Figure 5.1: Distribution of transient lengths, homeostatic and non-homeostatic $K = 1$.

By observation, homeostatic but not non-homeostatic transients have a distinct characteristic; exhibiting a sharp peak for transients with a single state transition and quickly tapering off. The definitive peak indicates that most homeostatic transients quickly return to the original state cycle. If Boolean networks accurately model the biology then it can be predicted that for most exogenous perturbations cells will return to their original state cycle.

For $K = 2$, the number of homeostatic transients overwhelms the number of non-

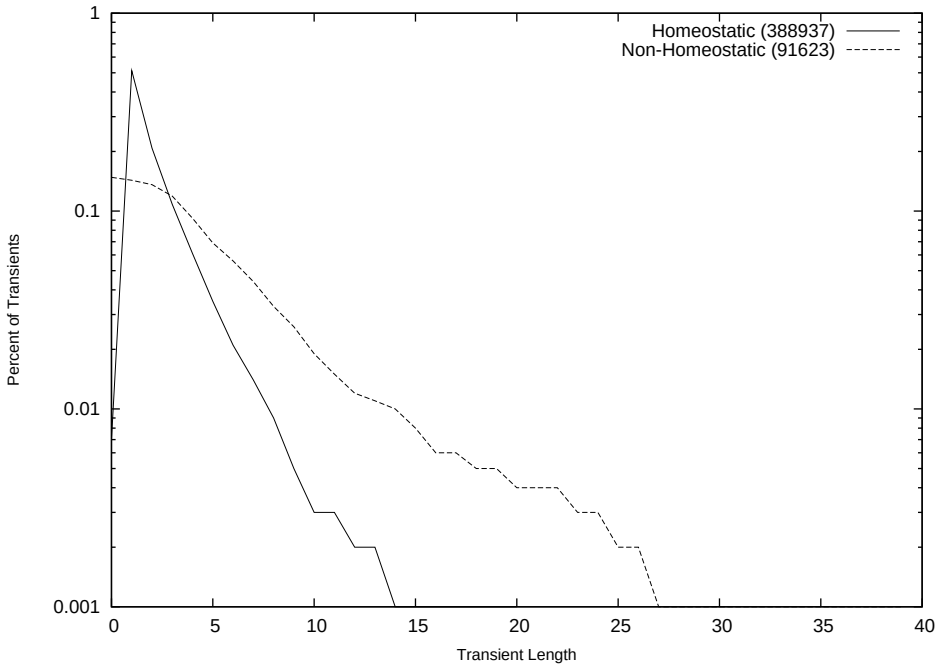

Figure 5.2: Distribution of transient lengths, homeostatic and non-homeostatic $K = 2$.

homeostatic transients by a ratio of approximately 4: 1. The tendency towards homeostasis lends stability to the model and exhibits behavior more typical of cell processes. As the K value is increased, the average transient length increases sharply and the concentration of transient lengths moderates rather than being concentrated close to zero.

It was expected that homeostatic transients of length one would return to their original state on the state cycle. This prediction was based on the assumption that shorter transients would return close to if not at their original state. Unexpectedly, it was confirmed that for all values of K, the majority of homeostatic transients length one terminated at a different state on the state cycle (see table 5.2). This result indicates that generally, length one homeostatic transients may return to any state in the attractor.

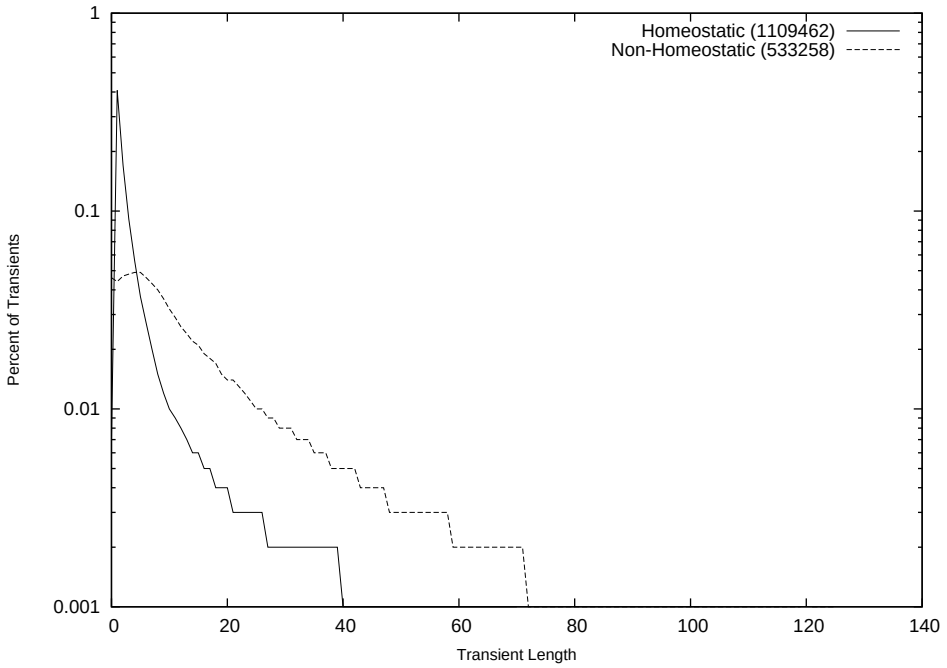

Figure 5.3: Distribution of transient Lengths, homeostatic and non-homeostatic $K = 3$.

5.2 Gene "Flips" Along Transients as a Function of K and Transient Length

As discussed in section 3.2 on page 10, the number of gene flips is the sum of the number of times each gene changes state on the transient. Figure 5.5 shows that for each value of $K = 2, 3, 4$, the total number of gene flips increases monotonically with

Table 5.2: For homeostatic transients of length one, does the transient return to the original state on the state cycle?

K-value	Runs	Same State	Different State
$K = 1$	1000	8170	11399
$K = 2$	1000	7587	88740
$K = 3$	1000	4943	194862
$K = 4$	100	189	42837

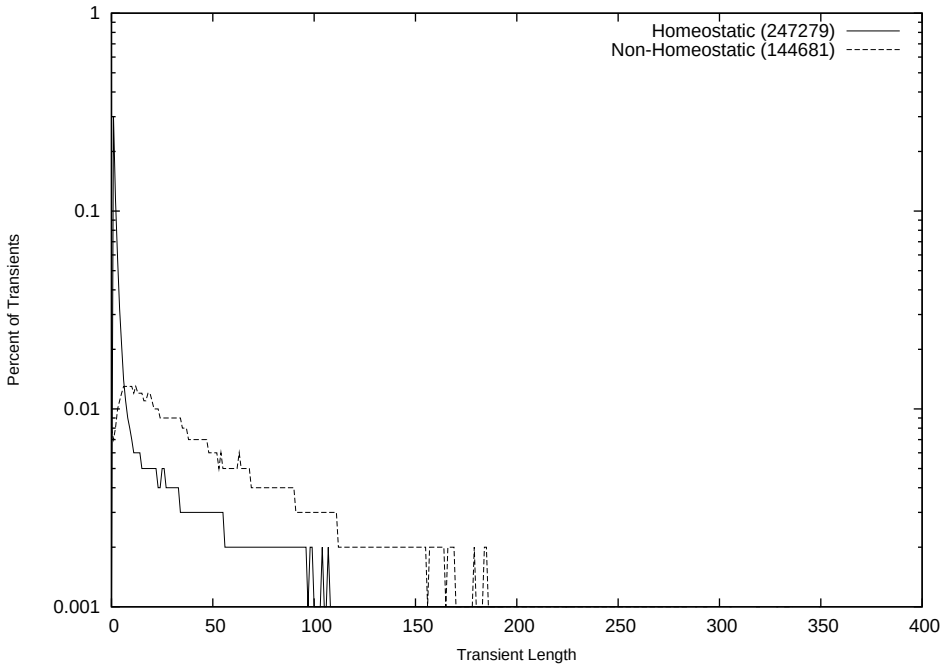

Figure 5.4: Distribution of transient lengths, homeostatic and non-homeostatic $K = 4$.

the length of the transient. As will be discussed later, this relationship is important because it provides a tool to bridge the synchronous nature of the Boolean network model to the asynchronous pathway of differentiation in real cells.

$K = 1$ did not follow the monotonic increase with the increase of transient length. The results for $K = 1$ suggest that these findings may not hold for cells deep in the ordered regime. It follows that if cells are in the ordered regime, we may be unable to use the number of gene flips as a mapping of transient length for cells in process of differentiation.

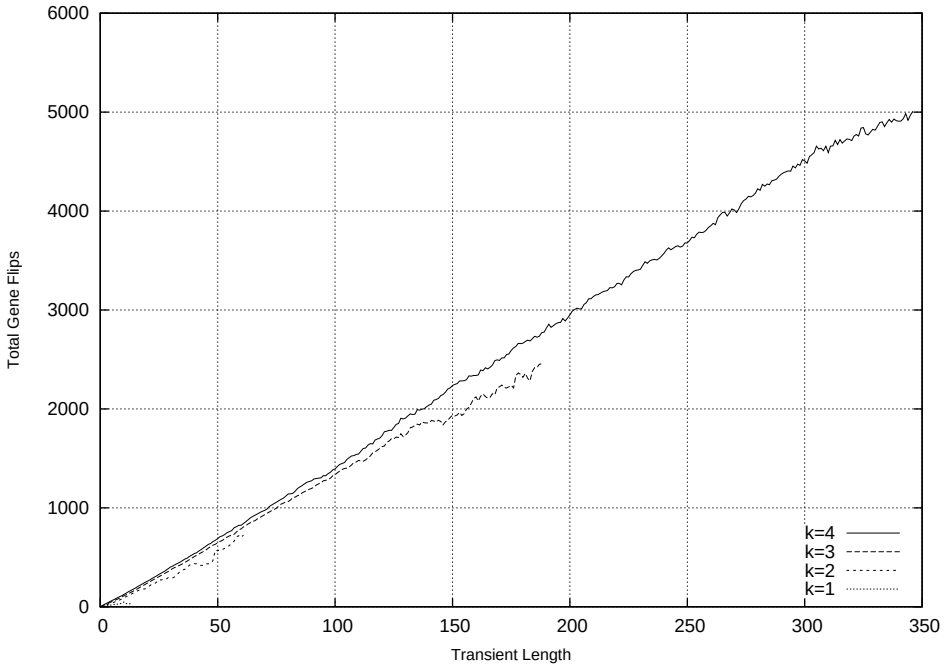

Figure 5.5: Gene Flips, $n = 40$, $K = 1, 2, 3, 4$.

5.3 Transient Fusion

Given a pair of transients flowing toward the same attractor, it is sometimes the case that the transients will join and flow together. Transient fusion is the measurement of common states between pairs of transients which join or fuse together. The length of fusion is measured by taking a pair of transients that share common states and counting backwards from the first state on the destination attractor to the point where the transient states differ. In general, this fusion between pairs of transients is more common in the ordered and critical regimes than in the chaotic regime [12].

5.3.1 Convergence With Transient Fusion

Given that a pair of transients fuse, there exists one or more pairs of states before the point of fusion in which the transients differ. The measure of difference between a pair of states is the Hamming distance. In biology, it has been recently discovered that the Hamming distance between differentiating cells which share the same destination cell type gradually diminishes as they approach the destination cell type [7].

It is not obvious that the Hamming distance between successive pairs of states on transients which fuse would diminish. Given a fully connected network, where $K = N$ and the Boolean functions are randomly assigned, for each state, the successor state is randomly chosen. On average, the distance between successive states is half of N or approximately $N/2$. So what you would expect from trajectories which join in a common pathway is that they stay about 50% apart and suddenly merge together. Hence, for networks where $K = N$, having transients in the same basin of attraction which fuse, there is no reason to believe their Hamming distance will shrink with proximity to the point of fusion. Thus, with respect to [7], it can be presumed that a mapping of a cell to an RBN would yield a network of a small K-value.

It was hypothesized that between pairs of transients which fuse, the Hamming distance measured between states along the trajectory would decrease or converge smoothly with proximity to the point of fusion similar to behavior observed in differentiating cells. We consider four classes of transients which converge:

- Purely Homeostatic: A pair of transients that originate from perturbations of the same state cycle and terminate after fusing on the same state cycle; i.e., $AA \rightarrow A$.

- Non-homeostatic A-type: A pair of transients that originate from perturbations of the same state cycle and terminate after fusing on a different state cycle;

i.e., $AA \rightarrow B$.

- Non-homeostatic B-type: A pair of transients that originate from perturbations of different state cycles and terminate after fusing on the source state cycle of one of the transients; i.e., $AB \rightarrow B$.

- Non-homeostatic C-type: A pair of transients that originate from perturbations of different state cycles and terminate on a third state cycle; i.e., $AB \rightarrow C$.

To measure the Hamming distance between the fusing pairs of transients, pairs of states from the transients were compared beginning with the states at the point of fusion. The Hamming distance was then measured pairwise at each step, by moving back towards the respective source state cycles up to the first state from the state cycle on transients of equal lengths.

Figure 5.6 is typical of the four fusion classes described above. Generally, the characteristics of the four classes are similar. The Hamming distance decreases as the attractor is approached and is generally greater for larger values of K. The results show that for stable, critical, and slightly chaotic K-values of transients which fuse, the Hamming distance decreases smoothly as the point of fusion is approached.

5.3.2 A Quantitative Comparison of Fusion Statistics

Figure 5.7 compares the number of transients with respect to fusion and homeostasis for $K = 1, \ldots, 4$. The histograms reveal that as K gets larger, the number of fused homeostatic transients grows smaller and the number of unfused non-homeostatic transients grows larger. This indicates a higher convergence in state space for networks in the stable and critical regimes where $K = 1$ and 2 respectively, than for networks in the chaotic regime, where $K > 2$. It is hypothesized that the ratio of

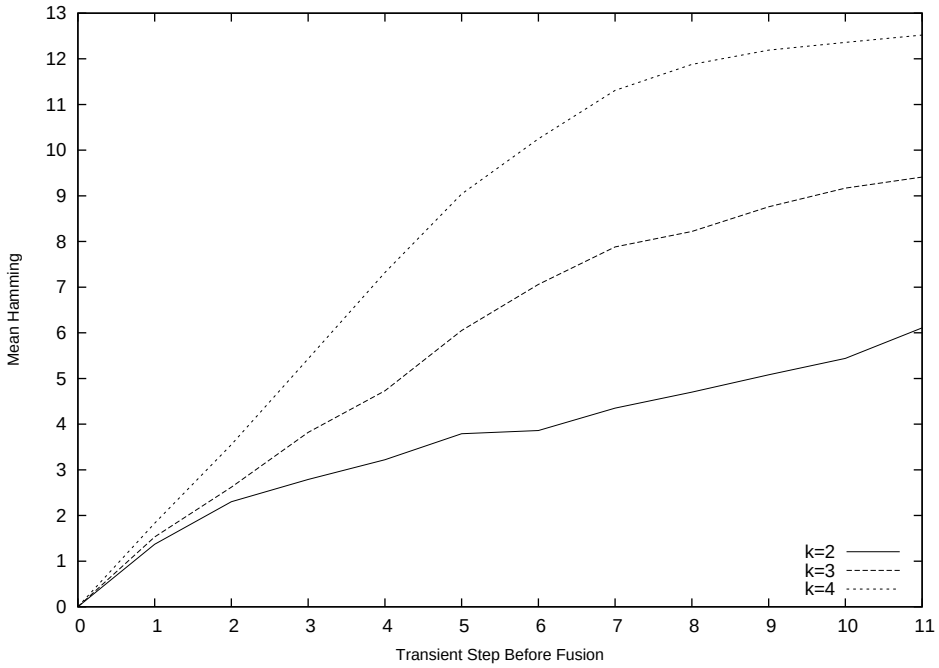

Figure 5.6: C-type fusion $AB- > C$, $n = 40$, $K = 2, 3, 4$

fused homeostatic transients to unfused non-homeostatic transients is a marker of the critical regime and is therefore measurable.

5.4 Trajectory Hamming Distance

It was hypothesized that the Hamming distance between successive states along the transient trajectory would be monotonically decreasing. The network values of successive states were recorded and the mean Hamming distance between them was calculated. All K input values exhibit similar behavior (see figures 5.8 and 5.9). The Hamming distance between successive states on the transient decreases until it reaches the attractor. Of course, at the attractor, the mean Hamming distance between states on the state cycle remains stable. These findings tend to support a gene array study undertaken at the Harvard Children's Hospital [7] showing similar

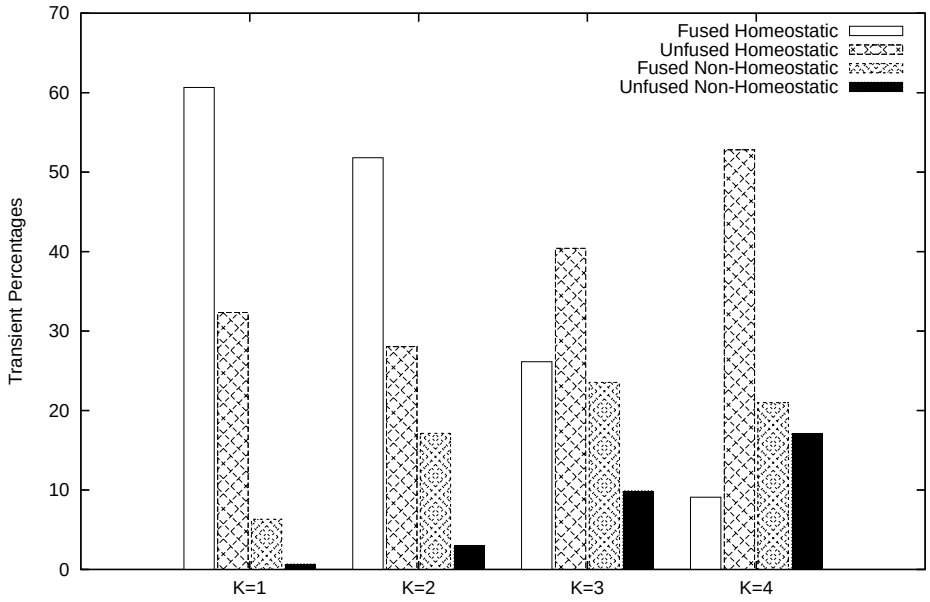

Figure 5.7: Fused and unfused transients, $K = 1$

behavior on pathways of differentiation in living cells. A caveat to these results is that the standard deviation measured for the Hamming distance between successive states is large which suggests that for any given trajectory, the gradual trend towards convergence would be difficult to notice in the simulated networks.

5.5 Distribution of Gene Flips as a Function of Transient Length

Figures 5.10-5.13 show the distribution of the fraction of times along a transient that a gene changes state, averaged over transients of all lengths in networks for $N = 40$, and $K = 1, 2, 3, 4$ graphed for homeostatic and non-homeostatic transients. The figures show a remarkable difference between networks in the ordered and critical

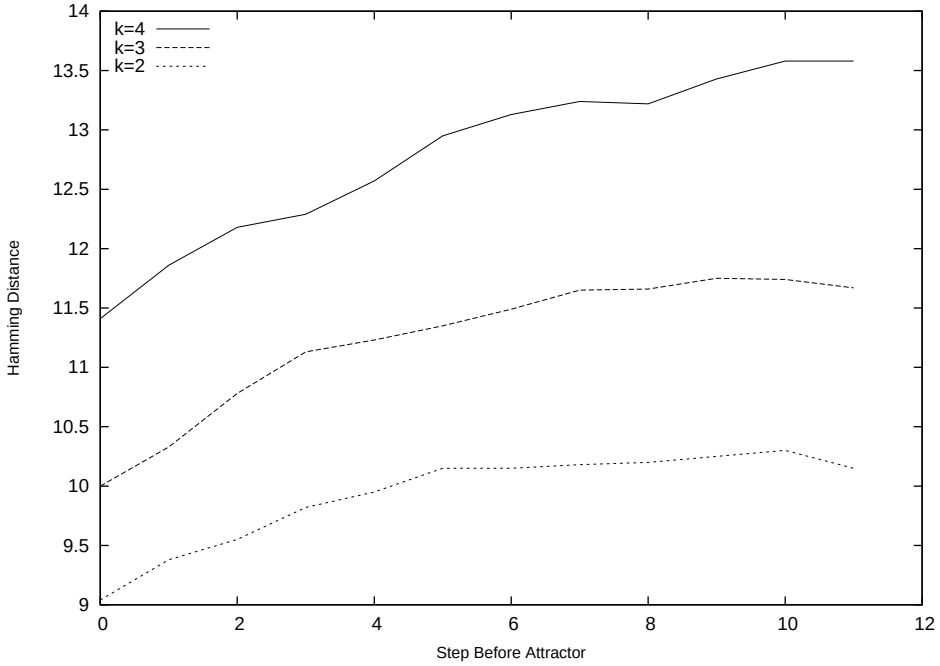

Figure 5.8: Non-homeostatic trajectory $K = 2, 3, 4$.

regimes, where $K = 1$ and $K = 2$ respectively, and networks in the chaotic regime for $K = 3$ and $K = 4$. Where $K = 1$ and $K = 2$, there is a wide distribution of fractions of times along transients that genes flip. Conversely, the chaotic networks for $K = 3$ and $K = 4$, show that a large fraction of genes never change state and a generally exponential drop off for larger fractions of times. The lengths of transients increase sharply in the chaotic regime. These features should be experimentally testable.

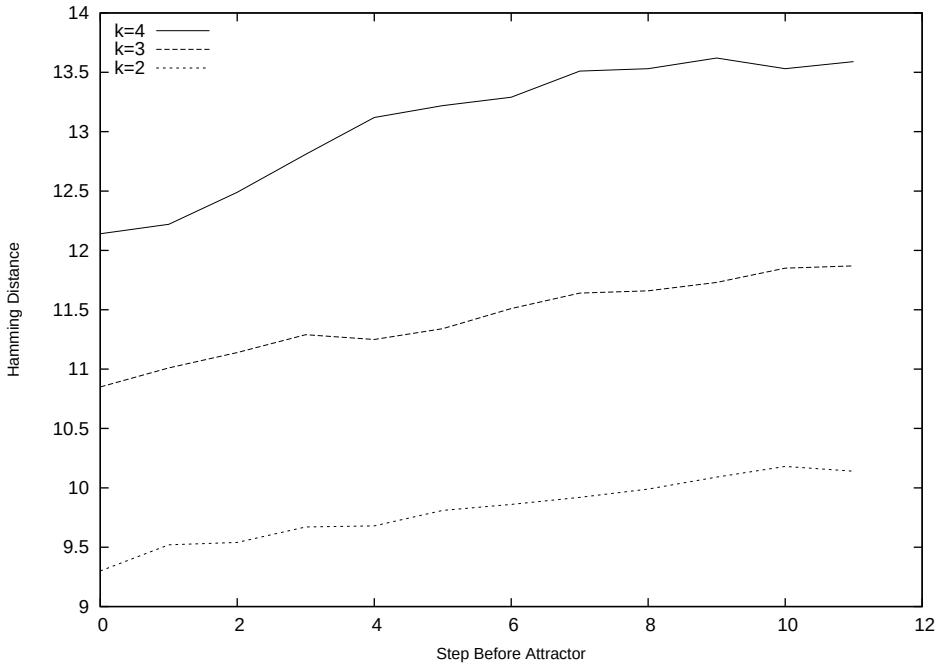

Figure 5.9: Homeostatic trajectory $K = 2, 3, 4$.

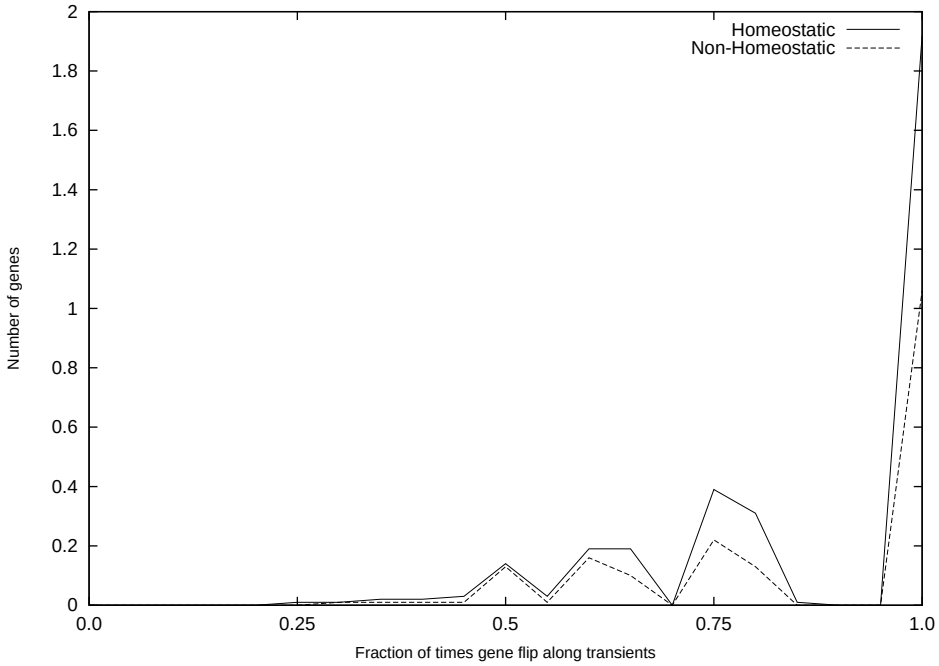

Figure 5.10: 10000 runs, $n = 40, K = 1$.

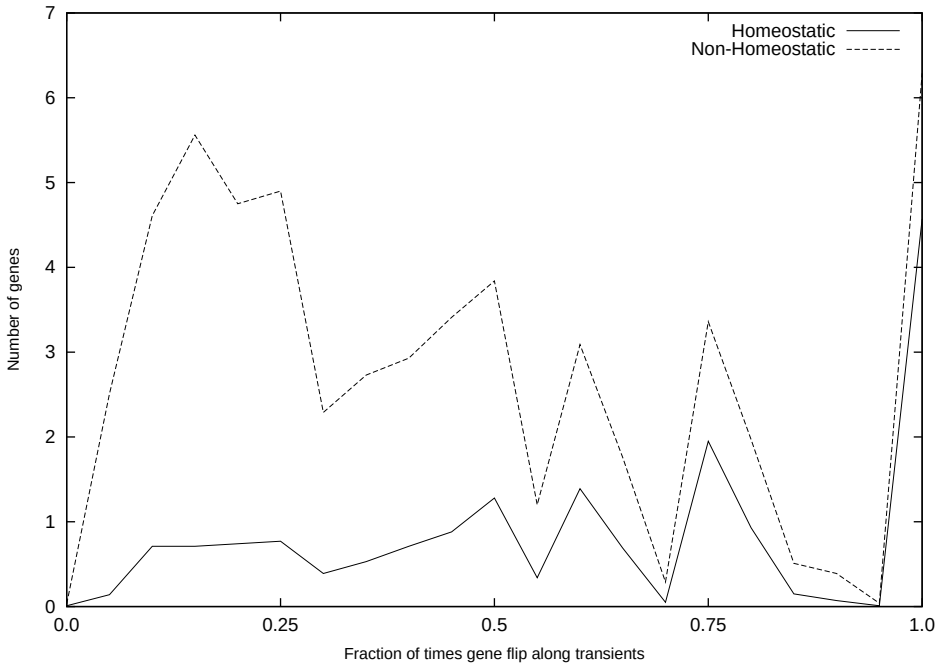

Figure 5.11: 1000 runs, $n = 40, K = 2$.

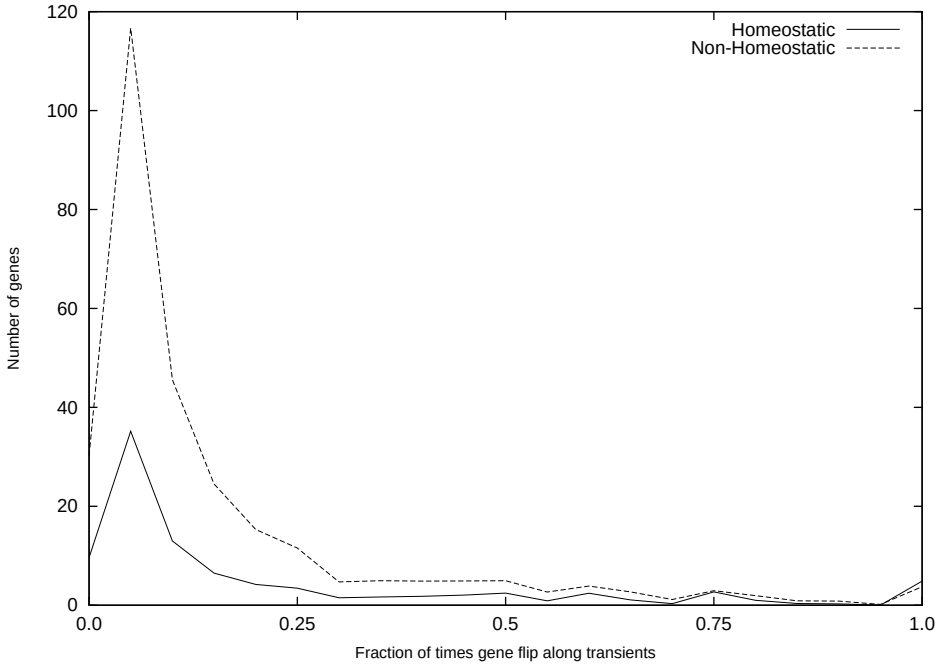

Figure 5.12: 1000 runs, $n = 40, K = 3$.

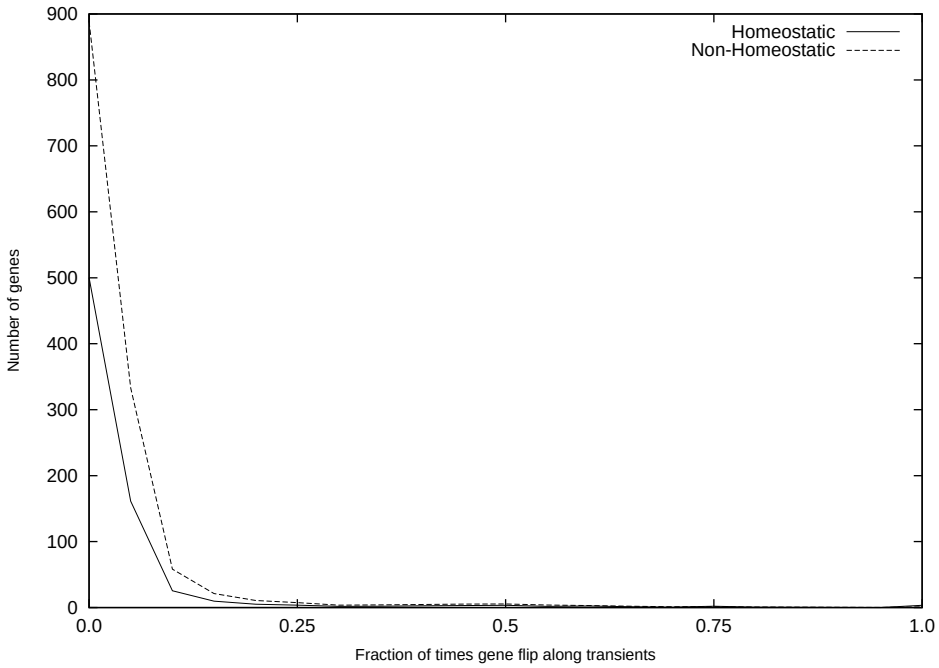

Figure 5.13: 100 runs, $n = 40, K = 4$.

Chapter 6

Discussion

This research is the first examination of pathways of differentiation under the hypothesis that cell types are attractors and pathways of differentiation correspond to perturbations of one attractor that place the network into another basin of attraction.

Using the model of deterministic random Boolean networks, a number of qualities of pathways of differentiation become open to inspection.

The number of fusing pairs represented by the four fusion classes on page 22, suggests that fusing pathways of differentiation can be found and the results indicate that the Hamming distance between pathways which fuse become closer together with proximity to the point of fusion. However, it should be noted that the profusion of fusion pathways is a general quality of Boolean networks and that the extent to which this mapping is valid for transient fusion in real cells still needs to be examined.

With respect to individual transients, the Hamming distances between successive states along transients terminating at the same attractor was observed to shrink. This behavior along with results obtained from the gene array study in [7] gives evidence that the use of random Boolean networks is appropriate. Because the

Hamming distances for large K was consistently greater than the Hamming distances for small K input values, we can predict that cells in the ordered regime are closer or more similar than cells in the chaotic regime.

6.1 Discrete vs. Continuous

It is important and necessary to establish a credible mapping from synchronous Boolean networks to the asynchronous real-time nature of living cells. A drawback of using Boolean networks is that it is difficult to model the continuous nature of genetic regulatory networks with discrete values. Boolean networks are synchronous in that each discrete step of the network is generated from the Boolean functions in a single step using the network state from the previous step. A synchronous Boolean network acts as if one central clock is directing activity. In contrast, genetic regulatory networks have no discrete clock. Cells which are differentiating are neither synchronized with other cells nor internally synchronized with expressed gene values. A genetic regulatory network is difficult to label as being precisely in one state or another. Even though gene activity on a molecular scale is made up of discrete events occurring in parallel [15], it is difficult to measure a differentiating cell as being in a specific state.

The mapping given by the results shown in figure 5.5 on page 21, suggests a linear relation between transient lengths in RBNs and the number of times which genes have altered their activity. If this is a valid mapping, it should be possible using gene arrays, to take a time series at closely timed intervals and binarize the data to yield a binary time series for each of the N examined genes. Using the previously mentioned mapping, we should be able to map transient lengths in terms of the total number of gene variations.

6.2 Determinism vs. Non-Determinism

If we remove the clock from the random Boolean network model and randomly update each of the nodes asynchronously at each time step, we introduce an element of non-determinism. The network can have more than one successor state. Rather than having state cycles in which a series of states is repeated, the dynamics of the network yields a closed set of possible states, or ergodic set, which can be accessed by the network. A basin of attraction then contains the states of the ergodic set plus all states which can reach the ergodic set. Perturbation would then initiate a pathway from one ergodic set. The pathway from the one ergodic set could have branches to more than one basin of attraction. The nature of these kinds of networks has not been well studied with respect to pathways of differentiation and is a topic for future study.

6.3 Proposed Gene Array Experiments

The following experiments could be implemented using gene arrays to provide validation of the predictions given by the RBN model.

1. Examine homeostatic transients: take a cell line, such as $HL60$ and first show that it is at a steady state using a gene array. Perturb the gene samples using retinoic acid [1] and perturb the gene's activity using a short exposure. Follow relaxation to $HL60$ with a time series of gene arrays. Retinoic acid is known to differentiate $HL60$ cells into white cells (polymorphoneutrophils). If the $HL60$ cells remain the same for small perturbations, then we know that the cells are stable to perturbations. This would provide proof we can get homeostatic perturbations from cells. There are about 15 known chemicals which cause

[1] Vitamin A.

$HL60$ cells to differentiate. It is possible to try small perturbations with all of them.

2. Examine pathways to other attractors: using the $HL60$ cell line, a long exposure to a high dose of retinoic acid can be done to see if the outcome is a polymorphoneutrophil (PMN). If a PMN results, that would show that a flow to another attractor could be achieved.

3. Examine trajectories within a basin of attraction: using different chemicals which are shown to differentiate $HL60$ cells into PMN's would show that there are different trajectories which lead to the same attractor. Different trajectories which lead to the same attractor would show that different trajectories in the same basin of attraction flow to the same attractor.

4. To test if there is a linear mapping between the number of gene variations and the transient length, we can count how many times a gene changes state during cell differentiation and binarize the data using some statistical clustering method. Ask how many times we change activity and binarize the data as a function of real time (30 minute samples). The results can be averaged over all the genes (about $20,000$). It is not clear this will work in the ordered regime if we recall that for $K = 1$, the graph was not monotonic.

5. To test for transient fusion, we can take the 10 chemicals known to differentiate $HL60$ cells into PNMs and initiate differentiation. Asking the question if any of these trajectories fuse before reaching the attractor. Using gene arrays, we can measure the concentrations of RNA and see if the last part of the pathways is the same between samples using pairs of chemicals.

6. The test to see whether the Hamming distance between successive states along a transient reduces with proximity to an attractor can be done with $HL60$ cells and differentiation initiated with retinoic acid using a gene array. The array

can be probed at closely timed intervals and the difference between subsequent samples can be measured. However, for our purposes, this has already been verified by the study done at Harvard Children's Hospital [7].

Chapter 7

Summary

The results represent the first analysis of pathways of differentiation under the theory that cell types are analogous to attractors and that pathways of differentiation correspond to perturbations of one attractor that place the network in the basin of attraction of another attractor.

A number of features of these pathways of differentiation have been measured. While it was not expected that random Boolean networks would yield a comprehensive model of cells, the results have exhibited some behaviors which should be measurable in living cells using current gene array techniques, such as:

- Measure the ratio of fused homeostatic transients to unfused non-homeostatic transients in cell types as a criterion for stability.
- Measure that the Hamming distance between successive states on a pathway of cell differentiation gets closer together (done [7]).
- Measure that the Hamming distance between pairs of differentiating cells gets closer together as they approach a terminal cell type.

- Confirm that the amount of gene activity in genetic regulatory networks can be mapped to transient length in Boolean networks via gene flips.

Our results show that the statistics of pathways of differentiation vary substantially in networks with different architectures, and hence suggest that measurements on real cells can yield important information.

Chapter 8

Directions for Future Work

The explorations into RBNs to model pathways of differentiation in genetic regulatory networks has prompted discussion on how to improve the model to more closely approximate the biology. Medusa networks and Asynchronous Random Boolean Networks may provide more clues into how the biology of cell differentiation works.

8.1 Medusa Networks

The Medusa network concept is the idea that a small group of genes called the "Medusa head" controls themselves and a larger group of genes or "Medusa tail" . What makes Medusa networks interesting is that they may more closely model the biology. Research in yeast cells suggests that the ratio of regulator to regulated genes in living cells is quite small, about 1:10 [13]. Intuitively we might assume that the stability of a genetic regulatory network is inversely proportional to the number of genes controlling the network, but we really don't know if this is true. The study of Medusa networks may enable the study of networks having a larger N value. Since the nodes in the Medusa tail do not control any other nodes, the network can be

represented using only those relevant nodes in the head whose value have an effect on changing the network state.

8.2 Asynchronous Boolean Networks

Another avenue of study in pathways of differentiation may be based on Asynchronous Random Boolean Networks (ARBNs). With ARBNs, nodes are chosen to update individually at random. The notable difference between classic random Boolean networks and asynchronous Boolean networks is the addition of nondeterminism [5]. ARBNs may be able to more closely mimic the continuous nature of cell differentiation [6].

References

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J.D. Watson. *Molecular Biology of the Cell*. Garland, New York, 1983.

[2] Steven J. Collins, Francis W. Ruscetti, Robert E. Gallagher, and Robert C. Gallo. Terminal differentiation of human promyelocytic leukemia cells induced by dimethyl sulfoxide and other polar compounds. *Proceedings of the National Academy of Sciences*, 75(5):2458–2462, May 1978.

[3] B. Derrida and Y. Pomeau. Random networks of automata: A simple annealed approximation. *Europhys. Lett.*, 1:45, 1986.

[4] Gary Flake. *The Computational Beauty of Nature*. MIT Press, Cambridge, MA, 1999.

[5] Carlos Gershenson. Updating schemes in random boolean networks: Do they really matter? *Artificial Life*, IX, 2004.

[6] I. Harvey and T. Bossomaier. Time out of joint:attractors in asynchronous random boolean networks. In P. Husbands and I. Harvey, editors, *Proceedings of the Fourth European Conference on Articial Life (ECAL97)*, pages 67–75. MIT Press, 1997.

[7] Sui Huang and Don Ingber. Cell fates as attractor states. Personal Communication. Harvard Children's Hospital., May 2004.

[8] F. Jacob and J. Monod. On the regulation of gene activity. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 26, Cold Spring Harbor, N.Y., 1961. Cold Spring Harbor Laboratory.

[9] F. Jacob and J. Monod. Genetic repression, allosteric inhibition and cellular differentiation. In M. Locke, editor, *Cytodifferentiation and Macromolecular Synthesis*. Academic Press, New York, 1963.

[10] S. A. Kauffman. Metabolic stability and epigenesis in randomly connected nets. *Journal of Theoretical Biology*, 22:437–467, 1969.

[11] S. A. Kauffman. Gene regulation networks: A theory for their global structure and behavior. *Current Topics in Developmental Biology*, 6:145, 1971.

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

[13] Denis Thieffry, Araceli M. Huerta, Ernesto Prerez-Rueda, and Julio Collado-Vides. From specific gene regulation to genomic networks: A global analysis of transcriptional regulation in escherichia coli. *BioEssays*, pages 20:433–440, 1998.

[14] A. Wuensche. Genomic regulation modeled as a network with basins of attraction. Technical report, Santa Fe Institute, Santa Fe, New Mexico, 1998.

[15] A. Wuensche. Basins of attraction in network dynamics: A conceptual framework for biomolecular networks. Technical report, Discrete Dynamics, Santa Fe, New Mexico, 2002.