Utah State University

DigitalCommons@USU

All Graduate Theses and Dissertations

Graduate Studies

5-2015

Algorithmic Information Theory Applications in Bright Field Microscopy and Epithelial Pattern Formation

Hamid Mohamadlou

Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/etd

Part of the Computer Sciences Commons

Recommended Citation

Mohamadlou, Hamid, "Algorithmic Information Theory Applications in Bright Field Microscopy and Epithelial Pattern Formation" (2015). *All Graduate Theses and Dissertations*. 4539.
https://digitalcommons.usu.edu/etd/4539

This Dissertation is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Theses and Dissertations by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.

UtahStateUniversity
MERRILL-CAZIER LIBRARY

ALGORITHMIC INFORMATION THEORY APPLICATIONS IN BRIGHT FIELD MICROSCOPY AND EPITHELIAL PATTERN FORMATION

by

Hamid Mohamadlou

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Computer Science

Approved:

Dr. Nicholas Flann Major Professor	Dr. Gregory Podgorski Committee Member
Dr. Xiaojun Qi Committee Member	Dr. Minghui Jiang Committee Member
Dr. Haitao Wang Committee Member	Dr. Mark R. McLellan Vice President for Research and Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2015

Copyright © Hamid Mohamadlou 2015

All Rights Reserved

ABSTRACT

Algorithmic Information Theory applications in Bright Field Microscopy and Epithelial Pattern Formation

by

Hamid Mohamadlou, Doctor of Philosophy
Utah State University, 2015

Major Professor: Dr. Nicholas Flann
Department: Computer Science

Algorithmic Information Theory (AIT), also known as Kolmogorov complexity, is a quantitative approach to defining information. AIT is mainly used to measure the amount of information present in the observations of a given phenomenon. In this dissertation we explore the applications of AIT in two case studies. The first examines bright field cell image segmentation and the second examines the information complexity of multicellular patterns. In the first study we demonstrate that our proposed AIT-based algorithm provides an accurate and robust bright field cell segmentation. Cell segmentation is the process of detecting cells in microscopy images, which is usually a challenging task for bright field microscopy due to the low contrast of the images. In the second study, which is the primary contribution of this dissertation, we employ an AIT-based algorithm to quantify the complexity of information content that arises during the development of multicellular organisms. We simulate multicellular organism development by coupling the Gene Regulatory Networks (GRN) within an epithelial field. Our results show that the configuration of GRNs influences the information complexity in the resultant multicellular patterns.

(99 pages)

PUBLIC ABSTRACT

Algorithmic Information Theory applications in Bright Field Microscopy and Epithelial Pattern Formation

Hamid Mohamadlou

The incredible patterns of multicellular organisms emerge as a result of the operation of Gene Regulatory Networks (GRN) that work during development. Understanding how GRNs produce these complex multicellular patterns is a significant challenge in biology. The primary goal of this dissertation is to employ Algorithmic Information Theory (AIT), also known as Kolmogorov complexity, to unravel the information complexity of GRNs and the resultant multicellular patterns. To obtain a better understanding of Kolmogorov complexity performance, first we study an application in cell image segmentation.

There are an estimated 20,000-25,000 protein-coding genes in the human genome. The sheer size of the human genome, as well as the huge number of protein and other gene product networks, requires systems biologists to use simplified computational models to gain insight into the behavior of the system. The approach taken in this work was to use a simplified model of a genetic regulatory network called a Boolean network, in which each gene is represented as a network node that takes binary values. Boolean networks represent a qualitative description of gene states and their interactions.

In this work, a model of embryonic cells in an epithelium field was simulated. Each cell holds a Boolean network and each Boolean network is designed to connect to the neighboring cells through cell-cell signaling. The state of each cellular network is initialized randomly by setting the state of each gene to 0 or 1. The state of the system during simulation is run synchronously until steady or cyclic state is reached for all individual cells. The steady or cyclic state, which is also referred to as attractor, is used to construct the multicellular body patterns by treating cells with the same attractor as the same cell types. The states of all the genes during the simulation of gene network dynamics along with multicellular patterns were encoded to strings and recoded for further analysis of information content. Kolmogorov complexity-based algorithms were applied to understand how the complexity of GRN configuration relates to the complexity of the spatial patterns that emerge as a consequence of network operation.

CONTENTS

Page

ABSTRACT iii

PUBLIC ABSTRACT iv

LIST OF TABLES viii

LIST OF FIGURES ix

1 INTRODUCTION 1
  1.1 Research objectives 2
  1.2 Modeling approaches and methodology 3
  1.3 Project outlines 7
  1.4 Research impacts 9

2 MAXIMIZING KOLMOGOROV COMPLEXITY FOR ACCURATE AND ROBUST BRIGHT FIELD CELL SEGMENTATION 10
  2.1 Abstract 10
  2.2 Introduction 11
  2.3 Implementation 13
  2.4 Results 21
  2.5 Conclusions 27

3 KOLMOGOROV COMPLEXITY OF EPITHELIAL PATTERN FORMATION: THE ROLE OF REGULATORY NETWORK CONFIGURATION 29
  3.1 Abstract 29
  3.2 Introduction 30
  3.3 Methods 30
  3.4 Experimental study 39
  3.5 Results and discussion 40
  3.6 Summary 46

4 THE ROLE OF NETWORK MOTIFS IN EPITHELIAL PATTERN FORMATION: A KOLMOGOROV COMPLEXITY STUDY 49
  4.1 Abstract 49
  4.2 Introduction 49
  4.3 Modularity of gene regulatory networks 52
  4.4 Multicellular model and its implementation 55
  4.5 Methodology 58
  4.6 Results and discussion 60
  4.7 Summary 65

5 CONCLUSIONS 67

APPENDICES 81
CURRICULUM VITAE 85

LIST OF TABLES

Table Page

2.1 The *NCD* values for the four image frames given in Figure 2.2 19

2.2 Set complexity values for two different approaches 24

2.3 Segmentation results for three data sets for human embryonic kidney cells (HEK 293T) 26

4.1 Orthogonal signaling. Dynamics and pattern complexity change when non modular networks become highly modular networks. 61

4.2 Symmetric signaling. Dynamics and pattern complexity change when non modular networks become highly modular networks. 61

4.3 ANOVA analysis of dynamics complexity illustrates differences for various groups of motifs over networks with orthogonal and symmetric signaling . 64

4.4 ANOVA analysis of pattern complexity illustrates differences for various groups of motifs over networks with orthogonal and symmetric signaling. . 65

LIST OF FIGURES

Figure Page

1.1 A Drosophila embryo at the cellular blastoderm stage triple-labeled for three segmentation proteins. Courtesy of Stephen W. Paddock, Eric J. Hazen, and Peter J. DeVries, HHMI, University of Wisconsin, Madison, WI, USA. . . 1

1.2 This diagram is discussed in chapter 3. The Drosophila segment polarity network in Boolean framework introduced by Albert et al. [11]. The network shows interactions between segment polarity genes and gene products. Genes are shown with rectangles, mRNAs with ovals and proteins with hexagons. The influence of one gene on another is indicated by the directed edges, those terminating in an arrow are activating, those with a dot are inhibiting. 5

2.1 Relationship between frame entropy as the focus level changes in the z-stack is shown in (a) and (b). In (a) there is a monotonic increasing and then decreasing relationship between focus and entropy, with the in-focus frame containing minimum entropy. In (b) a nosier data set is employed and the relationship between focus and entropy is irregular. As can be seen in frame 6, banding and stripe noise introduced by the microscope unexpectedly increases entropy. (c) Illustrates four corresponding frames for data set analyzed in graph (b). 16

2.2 Strongly and weakly defocused selected frames from time step 1 in data set one. Top row is the raw image frames. The second row is the binary image following Otsu thresholding that is linearized and compressed. 19

2.3 *NCD* values shown as a heatmap for all pairs of image frames in the z-stack of three selected defocused image stacks from the same experiment. Color code blue specifies pairs of frames with lowest *NCD* values and red specifies highest *NCD* values. The lowest z frame is in the lower left, the highest z frame is in the upper right. Analysis illustrates that off-diagonal *NCD* values range from 0.6 (most similar images) to 1 (red, most dissimilar images). Along the diagonal *NCD* equals zero (blue). Note the diversity of similarity relationships among the frames of each z-stack. 22

2.4 A parametric plot of set complexity values for the four defocused frames selected by the two algorithms. The X axis indicates the complexity value of the frame set selected by *maximal-information* and the Y axis indicates complexity value for the frame set selected by *sephaCe*. Each data point represents one z-stack from the 192 z-stacks in the human embryonic kidney cells (HEK 293T) data set. 23

2.5 Example cell segmentation results for two z-stacks of human embryonic kidney cells (HEK 293T) overlaid on the in-focus frame. Segmentations produced by *maximal-information* are shown in red; segmentations produced by *sephaCe* are shown in blue. In (a) both algorithms select similar frames and produce similar and accurate results. In (b) *maximal-information* selects a alternative set of frames at different focus planes from the fixed strategy and produces significantly lower segmentation errors. Here the *sephaCe* method fails to accurately detect four cells along with over segmenting another. In (c) segmentation results are shown closeup. 25

3.1 Example of pattern formation in *Drosophila* embryos. 31

3.2 The *Drosophila* segment polarity network in Boolean framework introduced by Albert et al. [1]. Network shows interactions between segment polarity genes and gene products. Some interaction are inter-cellular and connect two cells (represented by the two tan boxes). Gene's are shown with rectangles, mRNAs with eclipses and proteins with hexagons. 32

3.3 Gene expression in *Drosophila* segment polarity genetic regulatory network. The vertical axis shows the genes (lower case) and proteins (upper case) in the network shown in Figure 3.2, and the horizontal axis is a linear sequence of individual cells. Each repeating unit in the embryo (a paragsegment) is four cells wide. Eight paragsegments are represented in the figure. (a) The initial gene expression values, with the exception of the *sloppy-paired (SLP)* gene are set randomly. (b) A steady-state gene expression pattern that emerges from the operation of the Boolean network. 33

3.4 Network model used in this work. Each box is a cell within the epithelium containing an intracellular Boolean network that is identical to all other cells within the epithelium. A genetic regulatory network is represented as a graph where nodes are Boolean functions (representing a gene regulatory function) and edges denote an interaction between the output of one function and the input of another (a different regulatory gene). In (a) there is no signaling between cells; (b) illustrates orthogonal communication where one gene regulates the expression of another gene in an adjacent cell. Red nodes represent communicating genes; white are intracellular genes. 35

3.5 Examples of patterns from result sets showing their Ψ_p value (the set complexity of patterns), the bandwidth (BW is number of communicating genes), the intercellular signaling configuration (orth; is orthogonal, sym; is symmetrical), the cell-cell communication configuration (comm: sym is symmetric, orth is orthogonal) and the complexity domain (ordered, critical, chaotic) of the intracellular network. Each cell in the pattern is colored according to its attractor (same attractor, same color). Patterns are ordered left to right by increasing compression size. 39

3.6 Distributions of network *NCD* values between network pairs as a function of the number of communicating genes from 0 to 8 along the horizontal axis. The vertical axis for each plot is *NCD* from 0.0 at the bottom to 1.0 at the top. High probability is red, low probability is blue. 42

3.7 Distributions of pattern *NCD* values between pattern pairs as a function of the number of communicating genes from 0 to 8 along the horizontal axis. The vertical axis for each plot is *NCD* from 0.0 at the bottom to 1.0 at the top. High probability is red, low probability is blue. (a)(c) refer to the equivalence classes discussed in Section 3.5.1. 43

3.8 Network vs. pattern complexity. The relationship between network complexity and the subsequent pattern complexity for the two signaling configurations. Each line shows the trajectory as the signaling bandwidth increases from zero to eight for ordered, critical and chaotic networks. 45

4.1 Ventral view of stage 16 *Drosophila melanogaster* embryo immunostained for tropomyosin (green; a protein expressed in muscle), Pax 3/7 (blue; a regulatory protein expressed in central nervous system nuclei and ectoderm), and HRP (red; neurons). All nuclei shown in gray (DAPI). Courtesy of Julieta Mara Acevedo and Lucas Leclere, Marine Biological Laboratory, Woods Hole, www.mbl.edu/ dev.biologists.org/ 50

4.2 Structural modularity. (a) A network with modular structure where intramodular connectivity is higher than inter-module connectivity. (b) A randomly connected network. 53

4.3 Functional modularity. (a) A positive feedback loop (double-negative loop with two positive autoregulatory loops [2]). (b) A positive feedback loop (double-positive loop with two positive autoregulatory loops). (c) A negative feedback loop [3] with two positive autoregulatory loops. (d) Coupled positive-positive feedback loops. (e) Coupled positive-negative feedback loops. (f) The type-1 coherent feed forward loop [4]. 54

4.4 Structural modularity implementation. Lattice of cells is illustrated in this figure. Cell-cell signaling is implemented through specific modules referred to as signaling modules (depicted in green). In this example, cell-cell signaling is orthogonal such that two adjacent cells signal directionally [north-south and east-west]. These directions can be thought of as corresponding to the anterior-posterior and dorsal-ventral embryonic axis. 56

4.5 Motif insertion. Example motif insertion into a random GRN. Dashed arrows represent random outgoing signals from the motif. Outgoing and incoming signals from and to the random GRN are randomly connected to genes within the random network. 57

4.6 Non modular vs. modular. Average dynamics and pattern complexity for 60 non-modular (arrow tails) and 60 highly-modular (arrow heads) networks. 61

4.7 Effect of insertion of a double-positive feedback loop on network dynamics and pattern complexity. Insertion of a double-positive loop significantly increases only pattern complexity in the case of orthogonal signaling, and increases both pattern and dynamics complexity in the case of symmetric signaling. 63

4.8 Effects on network dynamics and pattern complexity of inserting regulatory motifs into random GRNs. Average dynamics and pattern complexity for 60 random GRNs (arrow tails) and 60 GRNs with the indicated inserted motifs (arrow heads). 63

CHAPTER 1

INTRODUCTION

The patterns that emerge during development are the consequence of Genetic Regulatory Networks (GRNs) that operate within and between cells [5]. GRNs are dynamic systems made up of a set of interacting genes where combinations of genes control the expression of other genes. Gene expression within each cell is determined by signaling within and between cells, ultimately forming the body plan and subsequent morphology [6]. While it is known that multicellular patterns emerge as a result of GRN interactions, the detailed processes by which complex variety of cellular patterns develop remains a significant challenge in biology.

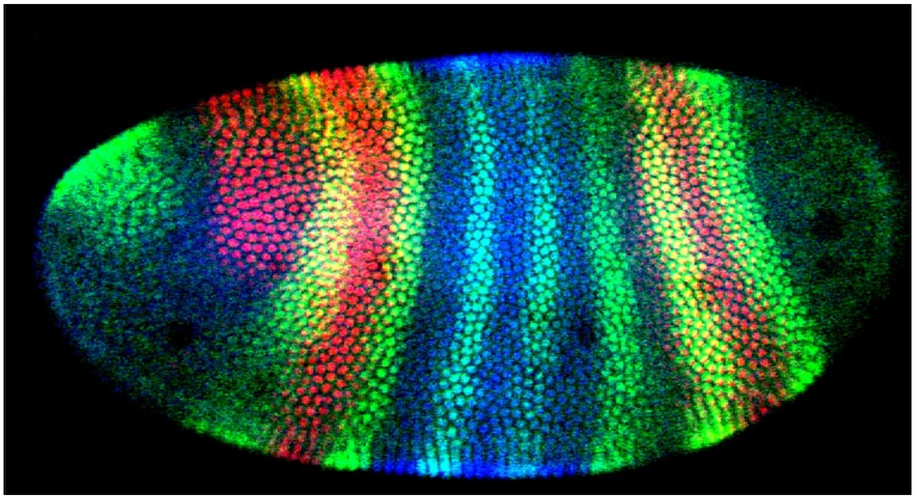


Figure 1.1. A Drosophila embryo at the cellular blastoderm stage triple-labeled for three segmentation proteins. Courtesy of Stephen W. Paddock, Eric J. Hazen, and Peter J. DeVries, HHMI, University of Wisconsin, Madison, WI, USA.

The primary objective of this dissertation is to answer this question: How is the information content in gene network dynamics and multicellular patterns influenced by functional and structural properties of the genetic regulatory network? The challenges in answering this question are twofold. First is quantifying the complexity of information that arises in GRN dynamics and resultant multicellular patterns during development. Second is determining the influential properties of a GRN in terms of the complexity of resultant multicellular patterns.

In order to quantify the information complexity of biological networks and the resultant spatial and temporal dynamics and patterns, we use the Algorithmic Information Theoretic (AIT) approach, also known as Kolmogorov complexity. First, in chapter two, we perform a study on Kolmogorov complexity, then in chapter three and four we explore the impact of some of the functional and structural properties of GRNs on developing complex network dynamics and multicellular patterns. The detailed objectives of each chapter are discussed in the next section.

1.1 Research objectives

To have a better understanding of Kolmogorov complexity performance, in chapter two we propose an application of cell image segmentation. The goal is to provide a preprocessing step toward bright field microscopy cell segmentation, or detecting cells in microscopy images by an image processing technique. Bright field microscopy is the simplest and most common method of cell imaging but it does not provide sufficient contrast needed by image processing methods for an accurate segmentation. Some studies have used images in stack of defocused microscopy frames (also known as Z-stack) to acquire more information for an accurate cell segmentation. In this study we propose a Kolmogorov complexity based algorithm called *maximal-information* to select the most informative images from Z-stack for an accurate and robust cell segmentation. *maximal-information* is compared with a recent approach that uses a fixed frame selection strategy over embryonic kidney cells (HEK 293T) image data from multiple experiments.

In chapter three we study pattern formation in a simulated two-dimensional lattice of cells containing identical GRNs, representing a simple model of an embryonic epithelium. We explore the role of complexity domains of intracellular GRNs and the nature of cell-to-cell signaling. We examine contact-mediated signaling, where cells can only send signals to neighboring cells, over a range that extends from no signaling to eight different signals. Kolmogorov complexity is employed to evaluate the information content of the genetic regulatory networks, the network dynamics, and the emergent cellular patterns. The objective of this chapter is to provide insight into the relationship between network dynamics and cellular patterns as a function of the types of cell-cell signaling and complexity domains of intracellular GRNs.

In chapter four we study the impact of GRN modularity on network dynamics and multicellular patterns. There are two kinds of modules: structural and functional. When GRNs are composed of tightly connected clusters of genes that are linked to other clusters by sparse connections, then these networks are said to exhibit structural modularity. Modules that occur frequently and consist of few interacting genes are referred to as functional modules or, more commonly, network motifs. Due to the importance of network structural modularity and the significance of biological motifs that naturally exist in a diversity of organisms, the objective of this chapter is to explore the influence of modularity on network dynamics and multicellular pattern complexity. The simulation model of chapter three is extended by adding following features to the model: representation of a GRN as a modular network and insertion of motifs into the complete GRN.

1.2 Modeling approaches and methodology

1.2.1 Boolean networks

To capture the behavior of gene regulatory system, scientists have developed mathematical and computational models for gene regulatory networks with the purpose of generating predictions to explain experimental observations. Among these modeling alternatives is a system of differential equations with quantitative values to capture the temporal and spatial expression levels of the genes. Despite their accuracy for small well known networks, differential equation suffer from the need for parameter values that are difficult to obtain. They are also computationally and conceptually too complex to model larger networks [7].

The massive scale of gene and protein networks requires systems biologists to simplify the computational models to gain insights into the behavior of these systems. A Boolean network is a simplified model of a genetic regulatory network, first introduced by Kauffman [8], [9], in which each gene is represented as a network node that takes binary values (1 for expressed and 0 for not expressed). The state of a gene (0 or 1) is determined by its Boolean function defined as an expression of AND, OR, NOT over the inputs from other genes represented as directed edges in the network graph. Boolean networks represent a qualitative description of gene states and their interactions.

An example of the use of Boolean networks is given for the inferred Drosophila segment polarity network [1] illustrated in Figure 1.2 (d). This is the network that determines the polarity (anterior-posterior axis) of each segment of the developing fruit fly. The large tan box shows the intracellular network within one cell, and the connections between boxes are the molecular signals that are products of intercellular signaling genes that link gene outputs of one cell with the inputs of regulatory functions in neighboring cells. In silico simulation of all known interactions among segment polarity genes in Drosophila has helped to determine whether the polarity network suffices to produce the organized spatial pattern in which cells only communicate with their adjacent neighbor [10].

1.2.2 Implementation of epithelial field of embryonic cells

To study pattern formation, a two dimensional lattice of cells containing identical Boolean GRNs is employed as a simple model of an embryonic epithelium [11]. This is an abstraction of many developmental systems, such as the cellularized Drosophila embryo [11], the sensory epithelia of the developing vertebrate retina [12] and the inner ear [13]. Signaling is implemented in the model as an edge connecting the state of one gene in a cell to an input of a Boolean function of one or more of its neighbors. Such genes are called communicating genes and these model ligand receptor interactions among contacting cells. The number of communicating genes is referred to as the signaling bandwidth.

Two kinds of signaling configurations are considered: (a) Symmetric, where each cell contains a gene (output of a Boolean function) that receives inputs from all four neighboring cells. This gene is activated if any of the gene inputs of neighboring cells are active (as

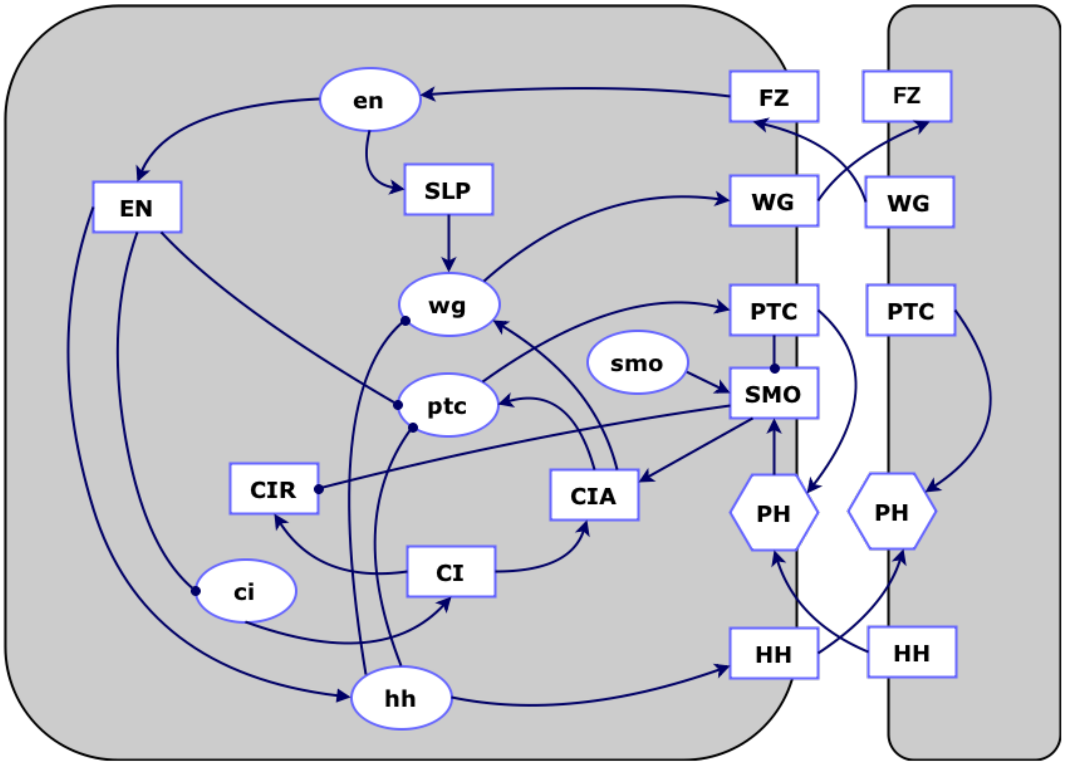


Figure 1.2. This diagram is discussed in chapter 3. The Drosophila segment polarity network in Boolean framework introduced by Albert et al. [11]. The network shows interactions between segment polarity genes and gene products. Genes are shown with rectangles, mRNAs with ovals and proteins with hexagons. The influence of one gene on another is indicated by the directed edges, those terminating in an arrow are activating, those with a dot are inhibiting.

in [14]); or (b) Orthogonal, where two adjacent cells signal directionally (North-South, East-West). Orthogonal signals correspond to intercellular communication along the anterior-posterior and dorsal-ventral embryonic axes.

Boolean network simulation will eventually generate a sequence of states that repeat, and so represent a fixed point. These fixed points within each intracellular network are referred to as attractors [15] and may be either a single state (a point attractor) or a cycle of states (a cyclic attractor). Some of these cyclic attractors may have an undetermined cycle length and are classified as chaotic. Stable attractors of genetic regulatory networks can be interpreted to represent terminally differentiated cell type [16] [17]. The process of cell differentiation where a cell transitions from a pluripotent cell to one with specific character, then is the sequence of network states that converges to an attractor. In this dissertation, if the state of two cells converge to the same attractor, even if they are cyclic attractors that are out of phase, they are considered as the same cell type.

1.2.3 Set Complexity

Set complexity was first introduced by Galas [18] to discover and reflect all computable similarities and information residing in a set of molecular sequences. Set complexity can be used to quantify the information content of a regulatory network, its temporal dynamics, and the spatial pattern produced. By measuring information content, set complexity can distinguish between critical systems that encode maximal information, and ordered and chaotic systems that encode less information. Set complexity (symbolized as Ψ) applies Kolmogorov's intrinsic complexity [15] to quantify contextual information in a set of objects by discounting the combination of pairs of objects that are randomly related or redundant.

Set complexity is independent of any specific application, so long as each object in the set (such as a GRN or multicellular pattern) can be encoded as a string. The Kolmogorov complexity of two strings is the length of the shortest algorithm that can transform one string to the other. Exact computation is undecidable [19], but minimum algorithm length can be approximated by the normalized compression distance (NCD) described in [20]. *NCD* is defined below, where s_i and s_j are strings, $s_i + s_j$ is the concatenation of s_i and s_j, and $C(s)$ is the compression size of string s:

$$0.0 \leq \mathit{NCD}(s_i, s_j) = \frac{C(s_i + s_j) - \min(C(s_i), C(s_j))}{\max(C(s_i), C(s_j))} \leq 1.0 \tag{1.1}$$

NCD is a measure of the similarity of the two strings. If the two strings compress smaller together than separately, then they are similar and *NCD* will be closer to 0.0.

Then set complexity of a set of n strings $S = \{s_1, \ldots, s_n\}$ is defined:

$$\Psi(S) = \frac{1}{n(n-1)} \sum_{s_i \in S} C(s_i) \sum_{s_j \neq s_i} \mathit{NCD}(s_i, s_j)(1 - \mathit{NCD}(s_i, s_j)) \tag{1.2}$$

According to [21] [11] set complexity is generally insensitive to the specific method of encoding objects (transforming objects to strings), as long as compression methods are lossless and effective. As we will see in the next section, we have employed a diversity of encoding methods selected based on the types of the data. For instance, network dynamics and cellular patterns each require different encoding methods.

Set complexity can distinguish between ordered, chaotic and critical systems. If the objects in a set are similar and $NCD(s_i, s_j) \simeq 0$, the set belongs to the ordered domain and $\Psi(S) \simeq 0$ indicating minimum information. A set belongs to the chaotic domain when the objects in the set are random. In this case $\Psi(S)$ takes low value because $NCD(s_i, s_j) \simeq 1.0$, however it will be greater than the ordered domain due to the effect of the $C(s_i)$ multiplicative term. $\Psi(S)$ is maximized when the set of objects describe an information dense system where the objects are all distinct from one another but share some similarity.

1.3 Project outlines

1.3.1 Project 1: An accurate and robust bright field cell segmentation: A Kolmogorov complexity study

Cell segmentation is the identification of cells and their observable properties from biological microscopy images. Florescent microscopy and bright field microscopy are two main methods of cell imaging. While bright field microscopic imaging is the most common method of cell microscopy, it presents a challenge to image processing techniques due to low image contrast and lack of nuclei reporters available with florescent microscopy. For some studies using bright field cell segmentation, researchers used images in stack of defocused microscopy frames (also known as a Z-stack) to acquire more information for an accurate segmentation. SephaCe is a recent method that uses images in the Z-stack for segmentation. SephaCe presents a series of algorithms to automatically segment images without the need for any florescent channel. The key to discriminating cells is to initialize a level-set algorithm with the difference of two strongly defocused images, chosen based on their entropy values, and then guide contour expansion using the difference of two weakly defocused images.

In an ideal case, entropy value increases monotonically as defocused distance in Z-stack increases, implying that there is no irregularity in the frames. Using this fixed frame selection strategy produces reasonable results in this case, but a fixed strategy cannot take into account random and systemic noises, variability in experimental configurations including microscope configurations, and multiple unknowns in the biological system under study.

In this project we present an optimization-based approach that searches the combinations of Z-stack frames to select the four frames that contain the most information. We propose a method called *maximal-information*, which applies Kolmogorov complexity measures to identify specific out-of-focus frames that encode the maximum information. *maximal-information* then searches the space of all possible combinations of two frames from above the in-focus frame and two frames from below the in-focus frame, evaluates each set, then picks the set that maximizes information content.

1.3.2 Project 2: Epithelial pattern formation: role of complexity domains and cell-cell signaling

In this project, we study the information complexity of simulated theoretical GRNs, their dynamics, and the resultant multicellular patterns. In the simulation we see the multicellular patterns emerge in a range from random to fairly organized patterns. This variation is due to different configurations for intra- and intercellular GRNs. We employ a Kolmogorov complexity-based approach to evaluate the information content of classes of GRNs with different configurations, their dynamics, and the emergent cellular patterns. For example, we examine contact-mediated signaling, where cells can only send signals to their neighboring cells, over a range that extended from no signaling to eight different signals. Finding a relationship between information complexity of GRNs and the information complexity of the emergent multicellular patterns has potential biological importance.

1.3.3 Project 3: Epithelial pattern formation: role of network modularity and motifs

Several biological studies have claimed that GRNs are modular and contain network motifs. In this project we explore the influence of modularity of GRNs on GRN dynamics and multicellular patterns. In the first part of this project the GRNs is designed to be structurally modular. In the second part of the project the most significant motifs are inserted into random GRNs. The simulation model is unchanged from the previous experiment. We employ a Kolmogorov complexity-based approach as an information theoretic measure to evaluate the information complexity of random GRNs, GRN with modular structure, and GRNs with inserted motifs.

1.4 Research impacts

In this dissertation we demonstrate that Kolmogorov complexity is a powerful tool to quantify the amount of information contained within a phenomenon. We apply Kolmogorov complexity-based algorithms to solve some challenging problems in developmental biology and in bright field image processing. We demonstrate that our Kolmogorov-based algorithm significantly improves the results for bright field cell segmentation. Also, we target a challenging problem in developmental biology. The challenge is to quantify the information complexity that arises as consequence of gene interactions and the information complexity of the resultant multicellular patterns. A tool that enables us to quantify information complexity in GRNs and emergent multicellular patterns will help us explore a potential relationship between the complexity of GRNs and the complexity of information in multicellular patterns. Such relationships can potentially have biological significance.

CHAPTER 2

MAXIMIZING KOLMOGOROV COMPLEXITY FOR ACCURATE AND ROBUST BRIGHT FIELD CELL SEGMENTATION

2.1 Abstract

Background. Analysis of cellular processes with microscopic bright field defocused imaging has the advantage of low phototoxicity and minimal sample preparation. However, bright field images lack the contrast and nuclei reporting available with florescent approaches and therefore present a challenge to methods that segment and track the live cells. Moreover, such methods must be robust to systemic and random noise, variability in experimental configuration, and the multiple unknowns in the biological system under study.

Results. A new method called *maximal-information* is introduced that applies a non-parametric information theoretic approach to segment bright field defocused images. The method utilizes a combinatorial optimization strategy to select specific defocused images from each image stack such that set complexity, a Kolmogorov complexity measure, is maximized. Differences among these selected images are then applied to initialize and guide a level-set based segmentation algorithm. The performance of the method is compared with a recent approach that uses a fixed defocused image selection strategy over an image data set of embryonic kidney cells (HEK 293T) from multiple experiments. Results demonstrate that the adaptive *maximal-information* approach significantly improves precision and recall of segmentation over the diversity of data sets.

Conclusions. Integrating combinatorial optimization with non-parametric Kolmogorov complexity has been shown to be effective in extracting information from microscopic bright field defocused images. The approach is application independent and has the potential to be effective in processing a diversity of noisy and redundant high throughput biological data.

2.2 Introduction

Cell segmentation is the identification of cell objects and their observable properties from biological images. Current cell segmentation methods perform most accurately when applied to high contrast and minimal noise images obtained from samples where the cells have fluorescently-labeled cell nuclei and stained membranes, and are distinct with minimal adherent membranes. However, these ideal conditions rarely exist.

Fluorescently tagging cells using green fluorescent protein (GFP) leads to robust identification of each cell during segmentation. While GFP tagging is widespread, there are disadvantages when applying the method repeatedly to the same sample since under repeated application of high-energy light the cells can suffer phototoxicity. Such light can disrupt the cell behavior through stress, shorten life and potentially confound the experimental results [22–24]. Significantly, a requirement for GFP labeling adds a step before a new cell line can be studied, thus making it difficult to apply this method in a clinical setting.

The alternative is to use bright field microscopy, the original and the simplest microscopy technique, wherein cells are illuminated with white light from below. However, using only bright field imaging of unstained cells presents a challenging cell detection problem because of lack of contrast and difficulty in locating both cell centers and borders, particularly when cells are tightly packed. Bright field imaging, while eliminating phototoxicity, leads to an excess of segmentation errors that significantly reduce biological and medical utility.

We seek to remedy the disadvantages and harness the experimental advantages of bright field microscopy of living cells by applying information-theoretic measures over defocused images to improve segmentation accuracy. The approach applies Kolmogorov complexity to identify the most informative subset of images within the focal stack that maximize information content while minimizing the effect of noise.

The paper first briefly reviews existing methods for segmentation of living cells, with a focus on recent approaches to defocused bright field images. Next, measures of Kolmogorov complexity are introduced and applied to image data. The new *maximal-information* method is then defined and evaluated by comparing its performance with a recent method *sephaCe* [24] over image sequence data sets from three separate experiments. An analysis and a discussion of the results follows.

2.2.1 Cell segmentation methods

Several cell segmentation approaches have been developed over time for detection of live cells in microscopy images [25–28]. Most of the approaches binarize an image with certain thresholding techniques, and then use a watershed or level-set based method on either intensity, gradient, shape, differences in individual defocused images (referred to as frames) [24, 29], or other measures. The algorithms then remove small artifacts with size filters, and apply merge and split operations to refine the segmentation [25–27].

Florescent microscopy cell segmentation

Most studies can primarily be categorized into a few key approaches. Wavelets are used for decomposing an image in both the frequency and spatial domain, and can be an effective tool since wavelets are robust to local noise and can discard low frequency objects in the background. Genovesio et al. [30] developed an algorithm to segment cells by combining coefficients at different decomposition levels. Wavelet approaches work well with whole cell segmentation, but have difficulty to segment internal cell structures. In Xiaobo et al. [31] a watershed algorithm was introduced for cell nuclei segmentation and phase identification. Using adaptive thresholding and feature extraction, Harder et al. [32] classified cells into four cell classes comprising of interphase cells, mitotic cells, apoptotic cells, and cells with clustered nuclei. In Solorzano et al. [33] the level-set method determines cell boundaries by expanding an active contour around each detected cell nuclei.

While these cell segmentation algorithms have been developed for fluorescence microscopy images, defocused bright field cell segmentation demands more complex and advanced level of image processing. Broken boundaries, poor contrast, partial halos, and overlapping cells are some of the shortcomings of available algorithms [24, 29] when applied to images lacking fluorescent reporters.

Defocused bright field microscopy approaches

Selinummi et al. [34] introduced z-projection based method to replace whole cell florescent microscopy with bright field microscopy. This method computes an intensity variation over a stack of defocused images (referred to as the z-stack) to obtain a contrast-enhanced image called a z-projection. Since variability of pixel intensity inside a cell is high compared to the background, the resulting z-projection image has high contrast and can substitute for an image obtained through whole cell florescent microscopy. The z-projection approach is straightforward and free from parameters setting. However, in order to distinguish between adherent cells, a second channel of nuclei florescent microscopy is required. As a final step *CellProfiler* [35] software is applied to both the z-projection and nuclei florescent channel to produce cell segmentation. While the z-projection approach avoids whole cell florescence, it still requires an additional nuclei channel of florescent microscopy and so does not eliminate potential problems with cell toxicity.

2.3 Implementation

A recent method that needs only bright-field defocused images has been introduced in *sephaCe* [24]. This system is capable of both the detection and segmentation of adherent cells and can be downloaded from (http://www.stanford.edu/rsali/sephace/seg.htm) as a free and open source image analysis package. In contrast to Selinummi et al. where all the frames of the z-stack are utilized, *sephaCe* selects only a subset of five frames as input to the image processing system. *sephaCe* selects this subset using a hard-coded strategy independent of each data set and each individual z-stack contained within that data set. Therefore, *sephaCe* does not adapt to the inevitable equipment and biological sample variation. While parameters of the image processing method can be tuned for specific data sets somewhat ameliorating the problem, a more general purpose non-parametric frame selection method is needed for high-throughput processing of diverse data sets. This work introduces a new adaptable frame selection method that applies an information theoretic measure to select frame subsets specific to the idiosyncracies of each z-stack. This method is referred to as *maximal-information.*

Following frame subset selection, the *maximal-information* method applies the same image processing and segmentation algorithm of *sephaCe.* Ali et al. [24,29] presents a series of algorithms that automatically segment each z-stack without the need for any florescent channel. The key to discriminating adherent cells is to initialize a level-set algorithm [36] with the difference between two strongly defocused frames and then guide contour expansion using the difference of two weakly defocused frames. As an initial step, the in-focused frame is detected by selecting that image from the z-stack in which the Shannon entropy [37] is minimized. Given an image histogram I, entropy is defined as:

$$E(I) = -\int_{y=1}^{n}\int_{x=1}^{m} p(I(x,y))\, log\, p(I(x,y)))dxdy \tag{2.1}$$

Where $p(I(x,y))$ is the probability of pixel intensity values. Entropy value is expected to be maximized for strongly out of focus images and minimized for the in-focus image. Let the in-focus image frame be I^0.

After detecting the in-focus image, four additional images from the z-stack are selected, two above the in-focus frame and two below. To initialize the level-set algorithm, a difference image is generated from two strongly defocused images selected at a fixed distance of $\pm 25\mu m$ from the in-focus frame, referred to as I^{++} and I^{--}. This image is binarized using the Otsu [38] thresholding method and then small artifacts are removed by labeling connected components and applying size filter.

To guide the level-set algorithm in expanding the initial cell boundaries, another difference image is generated between two slightly defocused images $\pm 10\mu m$ from the in-focus frame, referred to as I^{+} and I^{-}. Details on how this difference image is applied to compute local phase and local orientation images that direct the border expansion is given in [29] and [24].

2.3.1 Motivation for the *maximal-information* approach

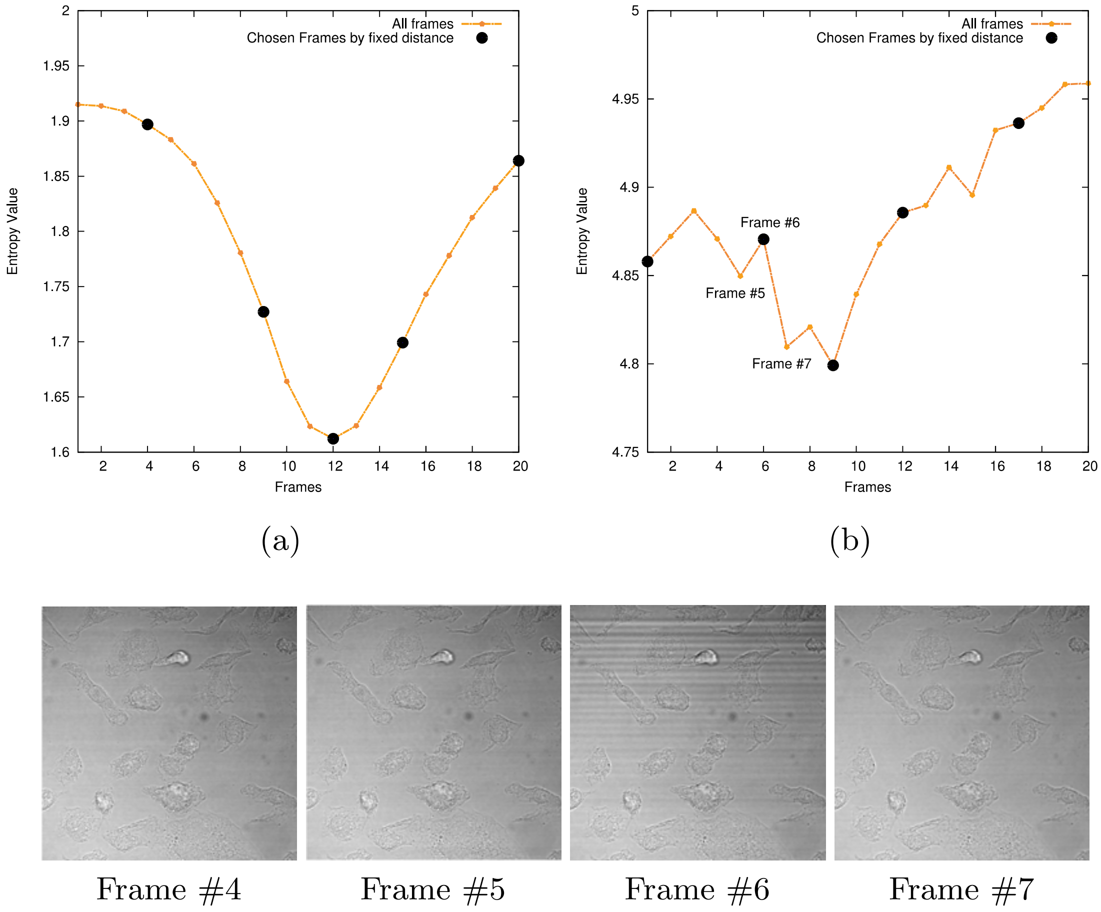
In the *sephaCe* package, the four defocused frames are chosen at fixed distances ($\pm 10\mu m, \pm 25\mu m$) from the in-focused frame to initialize and guide the level-set algorithm. Figure 2.1(a) illustrates an entropy analysis of a z-stack with 21 frames in which the image separation is $3\mu m$. The in-focus frame I^0 is determined as the 12'th frame, the 9'th and 15'th frames are the weakly defocused frames I^- and I^+ (in this case $\pm 9\mu m$ due to sampling resolution), the strongly defocused frames I^{--} and I^{++} are the 4'th and 20'th frames. In this z-stack image, as the frames become more blurred, their entropy increases monotonically implying that there are no irregularities within the frames. In this ideal case, the fixed strategy can produce reasonable results.

However, in experiments over a diversity of images (given in Section Results) this fixed selection of out-of-focus frames is demonstrated to produce poor segmentation. A fixed strategy cannot take into account random and systemic noise, variability in experimental configurations including microscope configurations, and multiple unknowns in the biological system under study. Some of these conditions are illustrated in selected frame images in Figure 2.1(c). Two possible reasons to account for the irregular entropy-focus plane relationship in Figure 2.1(b) are:

- Biological variability where cells do not adhere to the flat surface of the culture medium but vary in the z-dimension as they change morphology and form cell-cell adhesive bonds. That is, a focused frame for one cell could be a defocused frame for other cells. In Figure 2.1(c), the bright upper cell is positioned higher than the rest. Therefore a semi-random level of sharpness resides in the all defocused images.

- Systemic noise introduced by microscopy and imaging. For instance in Figure 2.1(c), frame 6 has strip noises introduced by the camera. Strip noise residing in the image increases the entropy value from the 5'th frame to 6'th frame while a decrease is expected.

Applying this fixed distance strategy to select strongly defocused frames can add unwanted initial active contours resulting in over-segmentation and also can miss initial active



(a) (b)

(c)

Figure 2.1. Relationship between frame entropy as the focus level changes in the z-stack is shown in (a) and (b). In (a) there is a monotonic increasing and then decreasing relationship between focus and entropy, with the in-focus frame containing minimum entropy. In (b) a nosier data set is employed and the relationship between focus and entropy is irregular. As can be seen in frame 6, banding and stripe noise introduced by the microscope unexpectedly increases entropy. (c) Illustrates four corresponding frames for data set analyzed in graph (b).

contours resulting in under-segmentation. Likewise, fixed selection of weakly defocused frames can add anomalies into the local phase and orientation images and thus misdirect the contour expansion to include or exclude cells, particularly when cells are tightly packed.

Overall, the fixed approach in selecting initial images in the *sephaCe* package is brittle and error-prone. The unavoidable variation requires an *adaptable* method rather than a fixed approach. The *maximal-information* method uses an optimization based approach that searches the combinations of z-stack frames to select the four frames that contain the highest information, evaluated using Kolmogorov information-theoretic measure [39]. This process is repeated for each individual z-stack and so adapts to the distinctiveness of each sample. Since the *maximal-information* method is adaptive, it can be applied to a diversity of data sets utilizing different microscopes, lighting conditions and biological samples.

2.3.2 Kolmogorov information set complexity

Set complexity [40], denoted Ψ, is applied to quantify the amount of information contained within each possible set of four image frames. The measure is general purpose and non-parametric in that it computes the information content of set of objects so long as they can be encoded as strings. Set complexity has been applied to understand the organization and information content of biological data sets including developmental pattern formation [5], genetic regulatory network dynamics [41], and gene interaction network structure [42]. The Kolmogorov complexity [39] of a string is the length of shortest algorithm that can be used to generate the string. Exact computation is undecidable, but it can be approximated by the compression size of a string. Bzip2 and zip compressor with block size of 900 Kbytes have been tested and shown robust for this purpose.

A related Kolmogorov complexity measure is the Normalized Compression Distance *NCD*) defined as the length of the shortest program that computes one given string from another. This measure provides a quantification of similarity between the strings since the more similar they are, the shorter the program needed. Again, this measure is undecidable but can be estimated using compression. Normalized Compression Distance described in [19] and [20] defined below, is such a measure of similarity between two objects that applies compression size $C(s)$ of string s:

$$NCD(s_i, s_j) = \frac{C(s_i + s_j) - \min(C(s_i), C(s_j))}{\max(C(s_i), C(s_j))} \tag{2.2}$$

where $s_i + s_j$ is the concatenation of s_i and s_j string. If the two strings compress smaller together than separately, then *NCD* will be closer to 0.0. As the two strings are more similar, the concatenated string is more compressed resulting in a lower *NCD* value. Random strings or dissimilar regular patterns are not as compressed and so *NCD* will be closer to 1 [43, 44].

1. $C(s^s_i + s^s_j) \simeq C(s^s_i) \simeq C(s^s_j)$ then $NCD(s^s_i, s^s_j) \simeq 0.0$
2. $C(s^r_i + s^r_j) \simeq C(s^r_i) + C(s^r_j)$ then $NCD(s^r_i, s^r_j) \simeq 1.0$
3. $C(s^r_i + s^s_j) \simeq C(s^r_i)$ and $C(s^s_j) \simeq 0.0$ then $NCD(s^r_i, s^s_j) \simeq 1.0$

where s^r is from the set of random strings and s^s are simple strings containing regular patterns. Set complexity [40] of a set of n strings $S = \{s_1, \ldots, s_n\}$ is defined:

$$\Psi(S) = \frac{1}{n(n-1)} \sum_{s_i \in S} C(s_i) \sum_{s_j \neq s_i} NCD(s_i, s_j)(1 - NCD(s_i, s_j)) \tag{2.3}$$

Set complexity captures the relationships among strings in the set, discounting when strings are very similar (*NCD* close to 0.0) and so contain the same information, or highly dissimilar so that they have nothing in common and appear random (*NCD* closer to 1.0). The value is maximized when each string is intrinsically complex (high $C(S_i)$) and the similarity between the strings lies between maximally dissimilar and maximally similar $NCD(s_i, s_j) \simeq 0.5$, which occurs when $C(s_i + s_j) \simeq C(s_i)/2 - C(s_j)$, assuming $C(s_i) > C(s_j)$.

Figure 2.2 gives an example of applying $\Psi(S)$ to defocused images. Along the top are the original frames and below them is their binary representation following an Otsu thresholding step. Each binary image is encoded as a string by concatenating each column scanning from left to right (more details are provided in Algorithm 2.3.3). For each image the compression size is given. *NCD* values between each pair of the images is provided in Table 2.1.

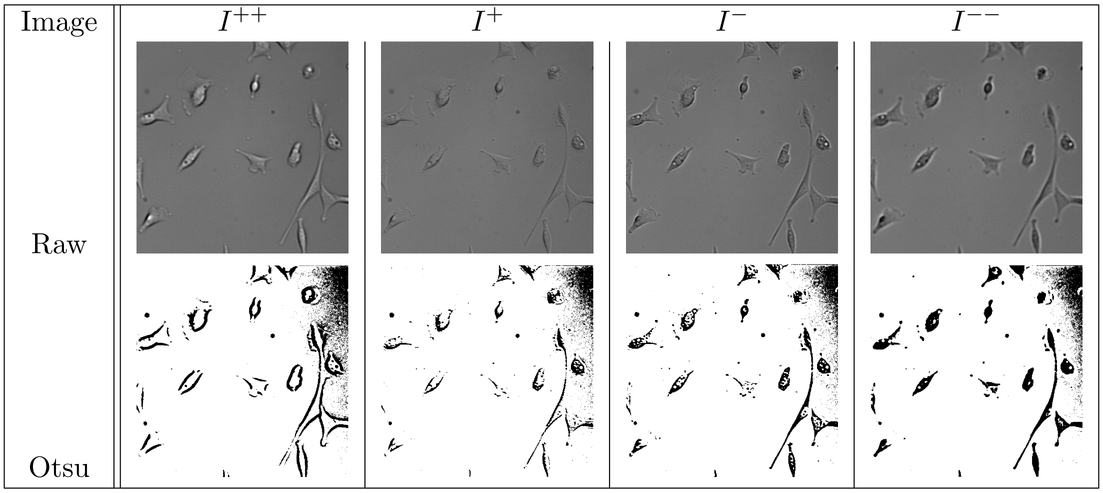


Figure 2.2. Strongly and weakly defocused selected frames from time step 1 in data set one. Top row is the raw image frames. The second row is the binary image following Otsu thresholding that is linearized and compressed.

Table 2.1. The *NCD* values for the four image frames given in Figure 2.2

NCD	I^{++}	I^{+}	I^{-}	I^{--}
I^{++}	0.0	0.1429	0.2154	0.1071
I^{+}	0.0	0.0	0.2615	0.1296
I^{-}	0.0	0.0	0.0	0.2000
I^{--}	0.0	0.0	0.0	0.0

2.3.3 The *maximal-information* segmentation method

To select the four most informative frames from a z-stack with n frames, the method searches the space of all possible combinations of two frames from above the in-focus frame (I^{++} and I^{+}) and two frames from below the in-focus frame (I^{-} and I^{--}), evaluates each set for Ψ, then picks the maximizing combination. The method is given in Algorithm 2.3.3.

Algorithm 1. The *maximal-information* algorithm to select the four z-stack frames needed to initialize the level-set method for segmentation. Let the input z-stack be $\boldsymbol{I}$ containing n frames. The algorithm returns the in-focus frame and four defocused frames. Note that all compression calculations are calculated once and cached.

```
maximal-information($\boldsymbol{I}$)
% binarize and linearize images
for $i = 1$ to $k$ do
    $\boldsymbol{Ip}[i] = Otsu(\boldsymbol{I}[i])$
end for
% compress individual and pairwise strings
for $i = 1$ to $k$ do
    $C[i] = C(\boldsymbol{Ip}[i])$
end for
for $i = 1$ to $k$ do
    for $j = i + 1$ to $k$ do
        $C[i, j] = C(\boldsymbol{Ip}[i] + \boldsymbol{Ip}[j])$
        $NCD[i, j] = (C[i, j] - \min(C[i], C[j]))/\max(C[i], C[j])$
    end for
end for
% find in-focus frame
$m \leftarrow E(\boldsymbol{I}[i])|1 \leq i \leq k$
$I^0 \leftarrow \boldsymbol{I}[m]$
% search for weakly and strongly out-of-focus frames
$\Psi_{\min} \leftarrow \infty$
for $i = 1$ to $m - 2$ do
    for $j = i + 1$ to $m - 1$ do
        for $k = m + 1$ to $n - 2$ do
            for $l = m + 2$ to $n - 1$ do
                $\Psi_0 \leftarrow \Psi(i, j, k, l, NCD, C)$
                if $\Psi_{\min} > \Psi_0$ then
                    $\Psi_{\min} \leftarrow \Psi_0$
                    $I^{++} \leftarrow \boldsymbol{I}[i]; I^{+} \leftarrow \boldsymbol{I}[j]; I^{-} \leftarrow \boldsymbol{I}[k]; I^{--} \leftarrow \boldsymbol{I}[l];$
                end if
            end for
        end for
    end for
end for
return $I^{++}, I^{+}, I^0, I^{-}, I^{--}$
```

First each image in the z-stack is binarized using the Otsu [38] thresholding method and then converted to a string (linearization) by concatenating each column of the image to the next column [45]. Many methods of linearization were explored in [45] and column concatenation was found to be effective because spatially located regularities are picked up by compression. Bzip2 is applied to compute the compression size of each individual string and also each pairwise concatenated string (for *NCD*, Equation 3.1). From these cached compression values, pairwise *NCD* values are determined.

The $O(n^2)$ compression step dominates the computation time since strings must be written to file before processing; the final Ψ calculation involves only matrix operations and is very fast, even though more combinations must be computed. For the three data sets studied in this work, the preprocessing and level-set algorithms of *sephaCe* take approximately 10 seconds per z-stack. The *maximal-information* frame selection method adds approximately 20 seconds per z-stack to the run time. Timings were on an Intel Pentium G640 Processor 2.8 GHz (3 MB cache).

2.4 Results

2.4.1 Set complexity analysis of image data

To understand how Kolmogorov Complexity measures could reveal information in z-stacks, an initial study was performed by computing the *NCD* between each pair of 21 frames for three data sets each containing 192 z-stacks. The data sets used for in this work are human embryonic kidney cells (HEK 293T) sampled at 5 minute intervals for 16 hours. Each z-stack sequence is from a distinct experiment. Data was obtained using a Leica DM6000 microscope with each z-stack containing 21 image frames each separated by 10μm, with resolution 1024×1024 12-bit grey-scale pixels. Since the z-stack was sampled at a 10μm resolution, the strongly defocused frames for *sephaCe* were set at $\pm 30\mu$m.

Figure 2.3 presents values of *NCD* in the form of a heatmap for each pair of frames along the z-stack sequence for a selection of three images. Frames tend to decrease in similarity as the focus distance increases so that blue areas (low *NCD*) are mostly around the diagonal, and red areas off the diagonal. However, each image displays significant individuality due to noise, microscope variability over time and changes in the biological sample as cells divide, die and move. This inconsistency among *NCD* matrices over time justifies the need for an adaptive frame selection strategy.

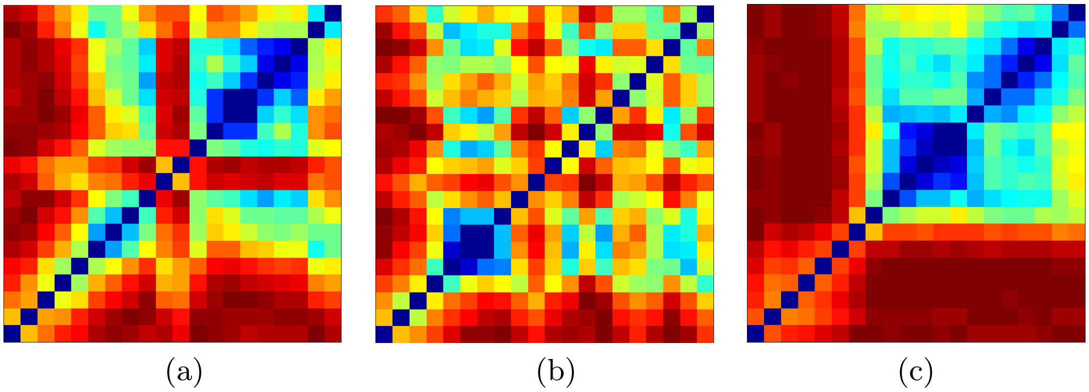

(a) (b) (c)

Figure 2.3. *NCD* values shown as a heatmap for all pairs of image frames in the z-stack of three selected defocused image stacks from the same experiment. Color code blue specifies pairs of frames with lowest *NCD* values and red specifies highest *NCD* values. The lowest z frame is in the lower left, the highest z frame is in the upper right. Analysis illustrates that off-diagonal *NCD* values range from 0.6 (most similar images) to 1 (red, most dissimilar images). Along the diagonal *NCD* equals zero (blue). Note the diversity of similarity relationships among the frames of each z-stack.

Four frames of the z-stack are chosen to start and guide the level-set algorithm. Figure 2.4 compares the computed Ψ of frames obtained by the *maximal-information* method with the Ψ of the frames identified using the fixed distance method of *sephaCe*, for all 192 z-stacks. In all cases the *maximal-information* frame set has a higher information content then the fixed *sephaCe* set. While this result is not surprising, it supports the need for adaptability as it demonstrates the inability of a fixed strategy to pick those images that have high intrinsic information. A mean difference hypothesis statistical analysis demonstrates that these differences are significant, see Table 2.2. According to the p-value in Table 2.2, that is much lower than 0.05, the mean difference hypothesis is rejected and so there is a significant difference between the mean values of the two groups. That is, selecting images using *maximal-information* guarantees sets with higher Ψ than the *sephaCe* method.

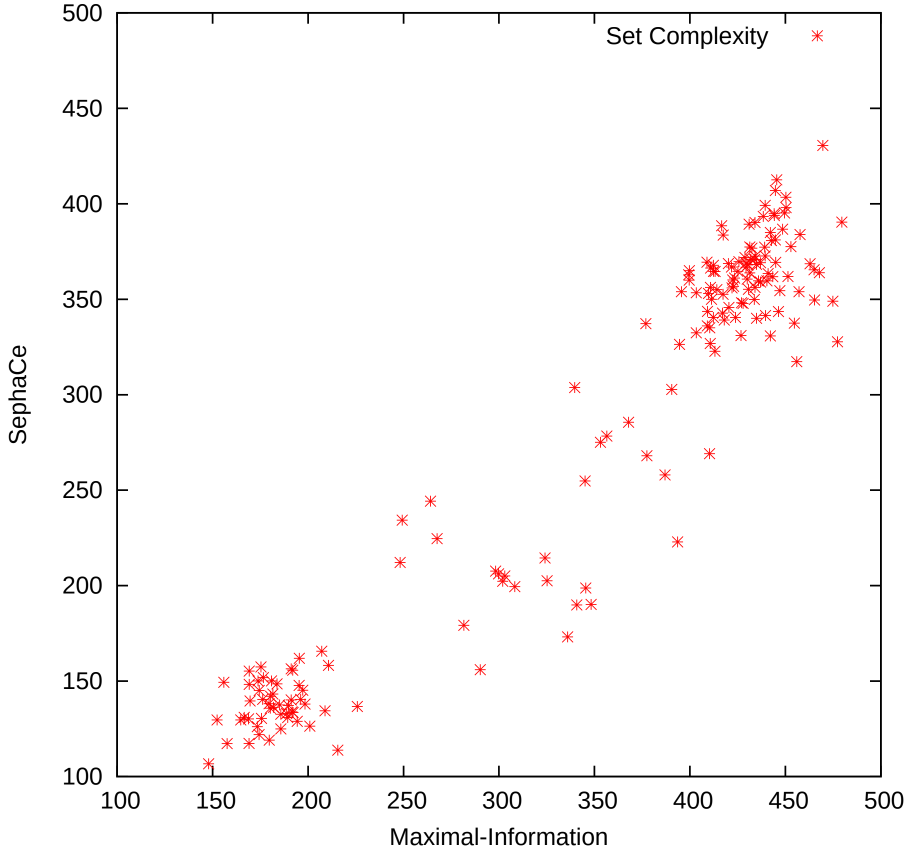


Figure 2.4. A parametric plot of set complexity values for the four defocused frames selected by the two algorithms. The X axis indicates the complexity value of the frame set selected by *maximal-information* and the Y axis indicates complexity value for the frame set selected by *sephaCe*. Each data point represents one z-stack from the 192 z-stacks in the human embryonic kidney cells (HEK 293T) data set.

2.4.2 Precision and recall analysis

Two examples of segmented bright field microscopy frames are shown in Figure 2.5. In (a) both algorithms select similar frames and produce similar and accurate results. In (b) *maximal-information* selects a alternative set of frames at different focus planes (com-

Table 2.2. Set complexity values for two different approaches

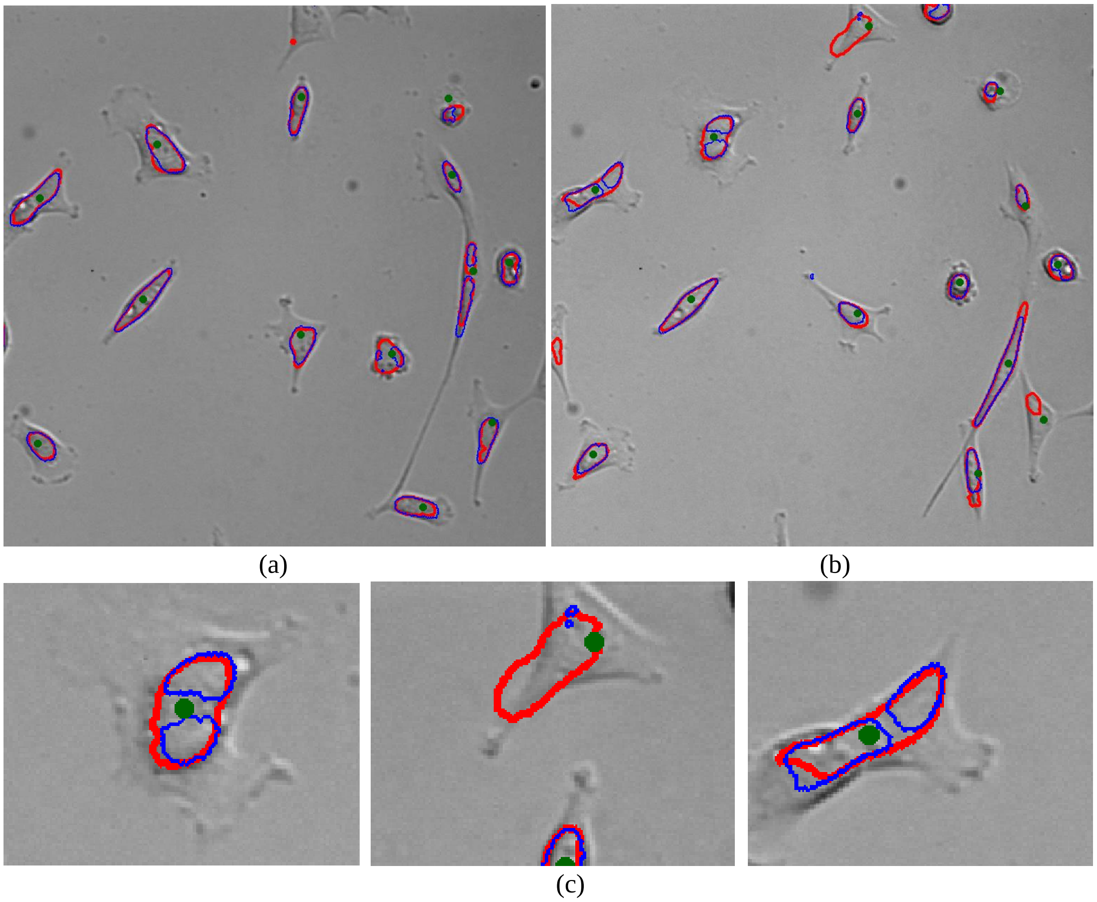
	Fixed defocused distance (*sephaCe*)	**Selected by *maximal-information***
Mean	278.5049	345.1289
Variance	10620.73	12336.47
Observations	192	192
Pearson Correlation	0.9603	
P(T¡=t) one-tail	1.19825E-67	
t Critical one-tail	1.6536	
P(T¡=t) two-tail	2.3965E-67	
t Critical two-tail	1.9736	

pared to the fixed strategy) and produces significantly lower segmentation errors. Here the *sephaCe* method fails to accurately detect four cells along with over-segmenting another.

In order to evaluate the segmentation results, the raw microscope z-stacks were provided to a human expert (Joseph C. Shope, Utah State University) who identified the cells using *Image-Pro Plus* (Media Cybernetics). Optimal z-frames were selected and cell centers determined by fitting a major and minor axis to produced excel files of cell center coordinates for each z-stack. No segmentation results were given to the expert during this initial cell identification. In parallel, the two methods were applied to the data sets to produce segmentation results for each z-stack, drawn as overlays with red (*maximal-information*) and blue (*sephaCe*) as in Figure 2.5. Next, the segmentation results were overlaid with the expert-determined cell centers and for both methods a count was made of the correctly identified cells (true positive), missing (false negative) and fragments of cells identified as one cell or spurious objects (false positive). To measure the quality and utility of the methods overall, the precision Pr and recall Re of *maximal-information* and *sephaCe* correction was determined, where:

$$Pr = \frac{tp}{tp + fp} Re = \frac{tp}{tp + fn}$$

with tp, fp, fn being the count of detected true positive, false positive, and false negative objects, respectively. In Table 2.3 the precision and recall of *maximal-information* are both


(a)

(b)

(c)

Figure 2.5. Example cell segmentation results for two z-stacks of human embryonic kidney cells (HEK 293T) overlaid on the in-focus frame. Segmentations produced by *maximal-information* are shown in red; segmentations produced by *sephaCe* are shown in blue. In (a) both algorithms select similar frames and produce similar and accurate results. In (b) *maximal-information* selects a alternative set of frames at different focus planes from the fixed strategy and produces significantly lower segmentation errors. Here the *sephaCe* method fails to accurately detect four cells along with over segmenting another. In (c) segmentation results are shown closeup.

significantly better than *sephaCe* for each of the three data sets.

In Table 2.3 the average correctly segmented cells for *maximal-Information* is higher than *sephaCe* method and demonstrates the advantage of extracting more informative frames in the z-stack. The average of both missing and unexpected cell segmentation

Table 2.3. Segmentation results for three data sets for human embryonic kidney cells (HEK 293T)

Data set one	***maximal-information***	***sephaCe***	***Correlation***	**t- stat**	$P(T \leq t)$ **one-tail**
Correct Segmentation *tp*	9.12	5.76	0.3970	9.4557	0.0
Unexpected areas fp	0.68	0.80	0.2355	-0.5492	0.2939
Missing cells fn	1.60	4.72	-0.0909	-9.0929	0.0
Precision *Pr*	93.20%	89.36%	0.3295	1.4461	0.0805
Recall *Re*	85.37%	54.34%	-0.2903	8.2830	0.0
Data set Two	***maximal-information***	***sephaCe***	**Correlation**	**t stat**	$P(T \leq t)$ **one-tail**
Correct Segmentation *tp*	13.35	12.60	0.4344	3.4701	0.0012
Unexpected areas fp	1.15	2.20	0.1633	-4.0977	0.0003
Missing cells fn	0.50	1.25	0.2939	-3.4701	0.0012
Precision *Pr*	92.30%	85.45%	0.1690	4.3714	0.0001
Recall *Re*	96.40 %	91.08%	0.2822	3.4407	0.0013
Data set three	***maximal-information***	***sephaCe***	**Correlation**	**t stat**	$P(T \leq t)$ **one-tail**
Correct Segmentation *tp*	15.56	11.86	0.4549	10.18	0.0
Unexpected areas fp	1.72	2.00	0.3642	-0.9434	0.1759
Missing cells fn	2.81	6.36	0.4926	-9.9501	0.0
Precision *Pr*	91.66%	86.23%	0.3887	2.6898	0.0
Recall *Re*	85.94%	65.21%	0.4256	10.12	0.0

tp is the average count of correctly identified cells, fp is unexpected segmentations and fn is cells that were missed. Recall and precision are given as percentages.

for *maximal-information* are lower than *sephaCe* method. All three of these measures of quality are shown to be significantly better for *maximal-information* than for the *sephaCe* using a paired one-tail T-test (values that are less than 10^{-8} are reported as 0.0 in the table).

In addition, Table 2.3 includes the inter-method correlation of *tp*, fp, fn over the z-stack data sets. High correlation implies that the performance of both methods is consistent in that they perform poorly on the same set of "difficult" images, and well on the same set of "easy" images. Results in Table 2.3 show that true positives are highly correlated implying that the cells correctly identified by *maximal-information* include some of the set of cells recognized by *sephaCe*.

2.5 Conclusions

This work has presented a method for identifying live cells in bright field defocused images. The method applies Kolmogorov complexity measures to identify specific out-of-focus frames that encode the maximum information. These frames are then used to initialize active contours and guide contour expansion for level-set segmentation algorithms as applied in the *sephaCe* method.

The new *maximal-information* approach is compared with a selection strategy employed in the original *sephaCe* that picks out-of-focus frames using fixed offsets from the estimated in-focus frame. An empirical study using a large data set of embryonic kidney cells (HEK 293T) z-stacks taken from different experimental runs has demonstrated that the adaptive method significantly improves the recall and precision of the segmentation.

Kolmogorov set complexity identifies the most informative frames by exploiting similarity measures between all pairs of frames contained within the *NCD* matrix. Each selected frame is sufficiently dissimilar (high *NCD*) to other frames in the set so as to provide unique and synergistic information about each cell in the z-stack. Recall that the dissimilarity is due to changes in cell appearance as the focal plane is moved through the cell profile. By selecting the best degree of dissimilarity, the differences between frames (used to initialize and guide the active contour of the level-set method) maximize sensitivity to the presence and shape of cells. Kolmogorov set complexity also tempers the effects of noise by discounting frames that have too higher dissimilarity since this is most likely due to noise.

The method introduced here is generally applicable because it relies on fundamental non-parametric information-theoretic properties and treats data as simple strings, ignoring the actual semantics. Robustness is achieved by viewing frame selection as combinatorial optimization problem with set complexity as the scoring function. The full potential of the method in dealing with noise, variability in experimental configurations, and multiple unknowns across a diversity of biological data will be explored in further studies.

Software and data availability

The software is written in Matlab and is available for download at https://sites.google.com/site/maximalinformation. Selected z-stack files are also available for download at https://sites.google.com/site/maximalinformation. For the full data set, please email nick.flann@gmail.com.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HM and NSF conceived the method and wrote the manuscript. HM wrote the code and performed all the computational experiments. JS assisted with the writing, analyzed all the raw images and evaluated performance. All the authors have read the paper and approve its contents.

Acknowledgements

This work was supported by the Luxembourg Centre for Systems Biomedicine, the University of Luxembourg and the Institute for Systems Biology, Seattle, USA. Research reported in this publication was partially supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number P50GM076547. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Thanks to Ilya Shmulevich and Pekka Ruusuvuori for helpful discussions, and to Adrian Ozinsky for image data.

CHAPTER 3

KOLMOGOROV COMPLEXITY OF EPITHELIAL PATTERN FORMATION: THE ROLE OF REGULATORY NETWORK CONFIGURATION

3.1 Abstract

The tissues of multicellular organisms are made of differentiated cells arranged in organized patterns. This organization emerges during development from the coupling of dynamic intra- and intercellular regulatory networks. This work applies the methods of information theory to understand how regulatory network structure both within and between cells relates to the complexity of spatial patterns that emerge as a consequence of network operation. A computational study was performed in which undifferentiated cells were arranged in a two dimensional lattice, with gene expression in each cell regulated by identical intracellular randomly generated Boolean networks. Cell-cell contact signaling between embryonic cells is modeled as coupling among intracellular networks so that gene expression in one cell can influence the expression of genes in adjacent cells. In this system, the initially identical cells differentiate and form patterns of different cell types. The complexity of network structure, temporal dynamics and spatial organization is quantified through the Kolmogorov-based measures of normalized compression distance and set complexity. Results over sets of random networks that operate in the ordered, critical and chaotic domains demonstrate that: (1) Ordered and critical networks tend to create the most information-rich patterns and networks; (2) signaling configurations in which cell-to-cell communication is non-directional mostly produce simple patterns irrespective of the internal network domain; and (3) directional signaling configurations, similar to those that function in planar cell polarity, produce the most complex patterns when the intracellular networks function in non-chaotic domains.

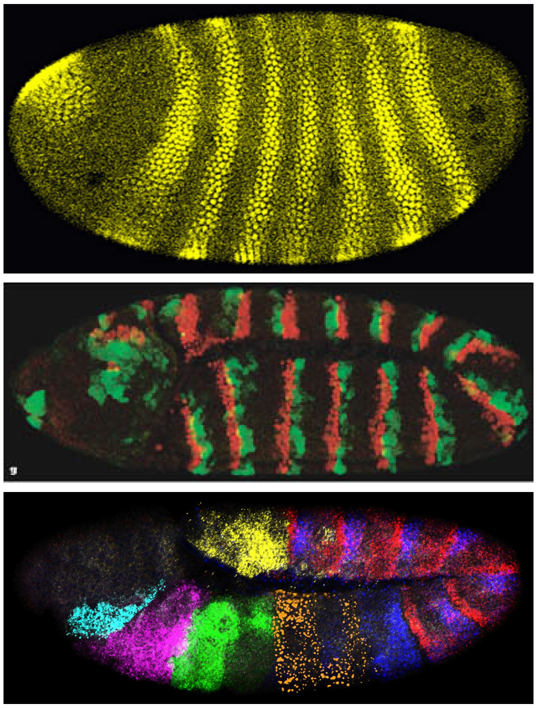
3.2 Introduction

Multicellular organisms exhibit an incredible variety of cellular patterns, for instance, those in the Drosophila embryo illustrated in Figure 3.1. These patterns arise during development and are a consequence of genetic regulatory networks (GRNs) that operate within cells and that respond to communication between cells [46, 47]. One interesting question to explore is the relationship between the structure of GRNs and the complexity of cellular patterns that can emerge from the operation of these networks. A related question is how GRNs and their evolution contributed to the transition from unicellularity to multicellularity. Although details are not known about the evolution of multicellularity in any lineage, this process almost certainly involved the co-option of GRNs and intercellular communication systems that existed in single-celled organisms (Knoll, 2011).While the actual paths of evolution to complex multicellularity may never be known, potential paths open to evolution can be explored and understood through computational studies. This is a long term goal of the investigations reported here.

Evidence suggests that living processes lie “on the edge of chaos,” and that biological systems experience selection to maximally retain information yet allow evolution [48–50]. Dynamic systems, including biological systems, operate in three complexity domains: ordered, critical and chaotic. Ordered systems are robust in that they dampen perturbations to retain information, but at the cost of limited potential for change. Chaotic systems magnify perturbations and lose information, rendering them unsuitable for homeostatic living systems; in fact, chaotic systems are implicated in diseases like cancer [51]. Critical systems, which operate on the cusp between order and chaos, are the most information dense in both network organization and dynamics [52]. This work focuses on how the information content of multicellular patterns is influenced by the complexity domain of intracellular GRNs and the nature of cell to cell signaling.

3.3 Methods

An empirical study was performed with a simulated embryonic epithelium consisting of a grid of undifferentiated cells, each containing identical Boolean networks to model a



(a) Expression of *hairy* (yellow) in the cellular blastoderm (Courtesy of Langeland, S. Paddock, and S. Carroll, HHMI)

(b) Expression of segment polarity genes, *wingless* (*wg*; green) and *engrailed* (*en*; red). Courtesy of C. Tomlin and J. D. Axelrod [53]

(c) Expression of seven *Hox* genes at the extended germ band stage. (Courtesy of Dave Kosman, UCSD)

Figure 3.1. Example of pattern formation in *Drosophila* embryos.

genetic regulatory network. Cell-cell communication was modeled by linking the output of a Boolean function to the input of the genetic regulatory network controlling one or more adjacent cells. The complexity domain of the network, its temporal dynamics and resultant pattern were quantified using the information theoretic measure called set complexity [54]. Empirical studies were performed over ensembles of randomly generated Boolean networks and the results compiled. Each step is defined in detail below.

3.3.1 Regulatory Network Models

Boolean networks [8] represent sets of expressed or non-expressed genes that are regulated by other genes using logic functions. They represent a qualitative description of gene states and there interactions. For instance, the inferred *Drosophila* segment polarity network is illustrated in Figure 3.2 (from [1]). Within the large tan box is the intracellular network of one cell, and the connections between boxes are the intercellular signaling genes

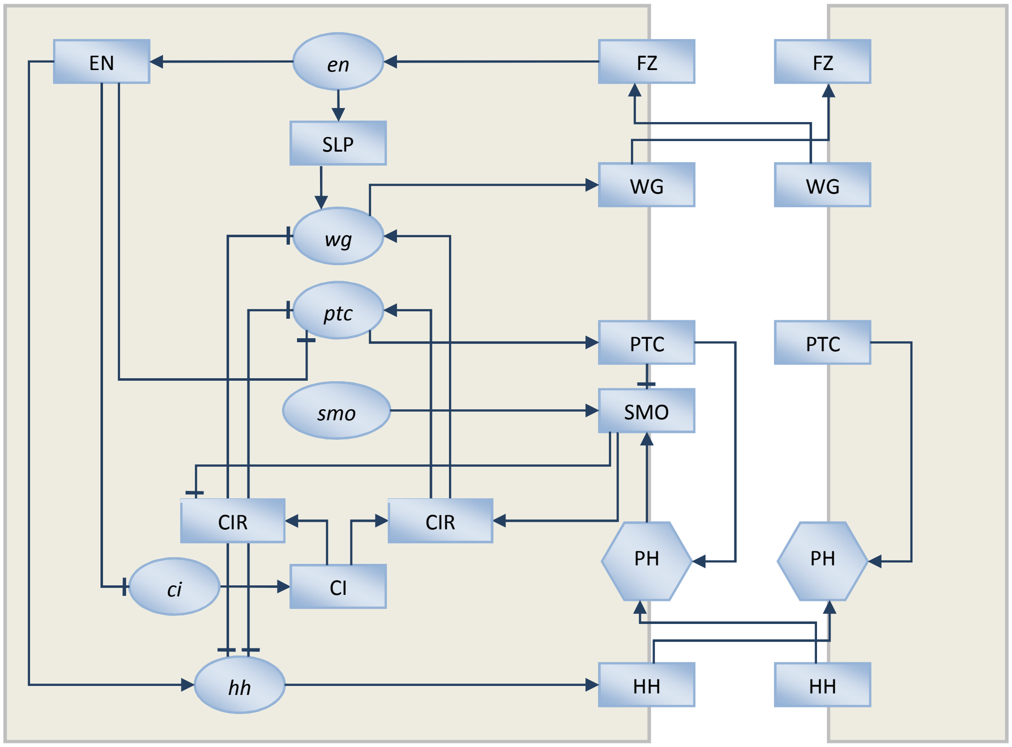


Figure 3.2. The *Drosophila* segment polarity network in Boolean framework introduced by Albert et al. [1]. Network shows interactions between segment polarity genes and gene products. Some interaction are inter-cellular and connect two cells (represented by the two tan boxes). Gene's are shown with rectangles, mRNAs with eclipses and proteins with hexagons.

that link the outputs with the inputs of regulatory functions that are activating (→) or inhibiting (⊤).

An assignment of true or false (representing expressed or not expressed) to each node in the network describes the state of the Boolean network. In this work, a node is a gene, mRNA, or protein and nodes are referred to generically as "genes." This is illustrated for the *Drosophila* segment polarity network in Figure 3.3. Each column is the gene expression values in a single cell, with rows corresponding to each gene and black showing expression and grey showing non-expression of a gene. Figure 3.3(a) shows the state of the network

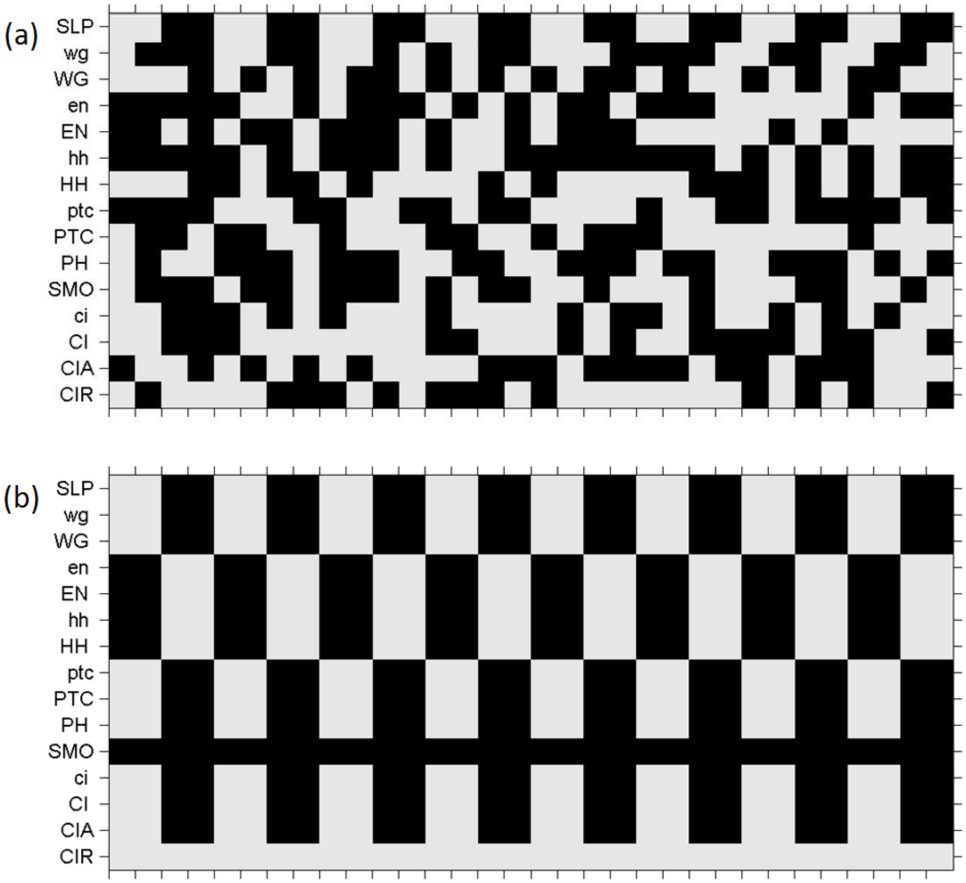


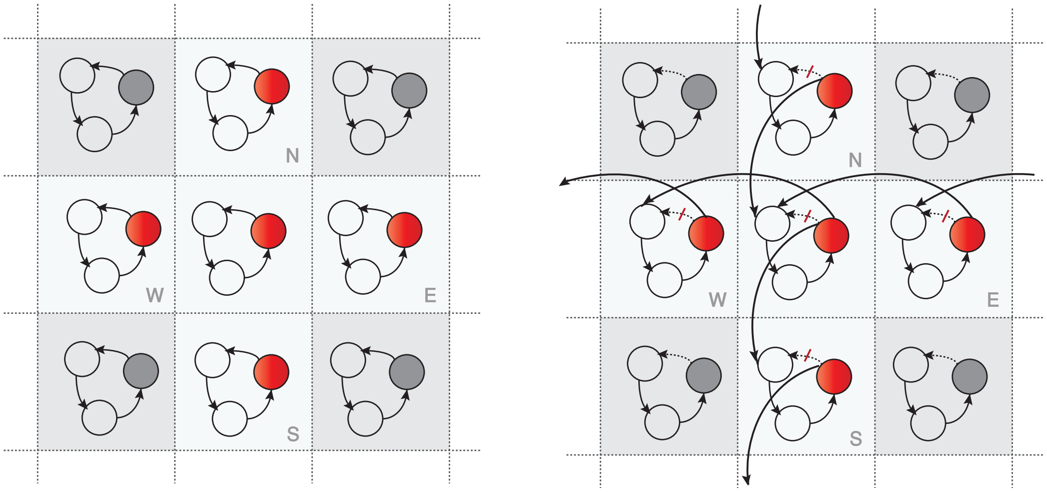
Figure 3.3. Gene expression in *Drosophila* segment polarity genetic regulatory network. The vertical axis shows the genes (lower case) and proteins (upper case) in the network shown in Figure 3.2, and the horizontal axis is a linear sequence of individual cells. Each repeating unit in the embryo (a paragsegment) is four cells wide. Eight paragsegments are represented in the figure. (a) The initial gene expression values, with the exception of the *sloppy-paired (SLP)* gene are set randomly. (b) A steady-state gene expression pattern that emerges from the operation of the Boolean network.

before execution, with all but the *sloppy-paired (SLP)* gene assigned random values. *SLP* is a member of the previously activated pair-rule gene network, and serves as a initiating gene for activation of the segment polarity network [1]. To simulate the dynamics of the network, the state of the system is clocked by applying each regulatory function to recompute output gene values, from all gene values of its inputs. Figure 3.3(b) shows the state of the network when it has reached a steady state. The pairwise patterning has emerged, with the values of genes *engrailed (en)* and *wingless (wg)* matching the biological embryo given in Figure 3.1(b).

In this work, sets of random intracellular Boolean networks were generated by randomly interconnecting a varying number of nodes within one cell then instantiating each regulatory node with a randomly generated logic function. To produce networks of different complexity domains, the number of inputs to each Boolean function (node in the network) is set according to $s = 2kp(1 - p)$ where s is the sensitivity of the network to perturbations in gene values, p is the probability of the output of each Boolean function being 1, and k is the count of inputs to each Boolean function [55]. When $s = 1$ a single bit change is on average propagated to one other node and the network is in the critical domain. In an ordered network, $s < 1$ and perturbations tend to die out, while in a chaotic network, $s > 1$ and perturbations tend to grow. In this work, p was fixed at 0.5 and k was changed to create networks of different domains: $k = 1$ for ordered, $k = 2$ for critical, and $k = 3$ for chaotic.

To study multicellular pattern formation, a two dimensional lattice of cells containing identical GRNs was employed as a simple model of an embryonic epithelium [56]. This is an abstraction of many developmental systems, such as the cellularized *Drosophila* embryo [57], and the sensory epithelium of the developing vertebrate retina [12] and inner ear [13]. Signaling is implemented in the model as an edge connecting the state of one gene in a cell to an input of a Boolean function of one or more of its neighbors (see Figure 3.4). Such genes are called communicating genes and these model ligand-receptor interactions among contacting cells. The number of communicating genes is referred to as the signaling bandwidth.

Two kinds of signaling configurations are considered: (a) Symmetric, where each cell contains a gene (output of a Boolean function) that receives inputs from all four neighbors. This gene is activated if any of the gene inputs of neighboring cells are active (as in [14]);



(a) (b)

Figure 3.4. Network model used in this work. Each box is a cell within the epithelium containing an intracellular Boolean network that is identical to all other cells within the epithelium. A genetic regulatory network is represented as a graph where nodes are Boolean functions (representing a gene regulatory function) and edges denote an interaction between the output of one function and the input of another (a different regulatory gene). In (a) there is no signaling between cells; (b) illustrates orthogonal communication where one gene regulates the expression of another gene in an adjacent cell. Red nodes represent communicating genes; white are intracellular genes.

or (b) Orthogonal, where two adjacent cells signal directionally (North-South, East-West). Orthogonal signals correspond to intercellular communication along the anterior-posterior and dorsal-ventral embryonic axes. A mechanism to autonomously generate intercellular directional signaling via a morphogen gradient has been elegantly demonstrated in [58]. This is implemented in the Boolean network by connecting an output function of the originating cell to the input function of the destination cell.

Since this work focusses on the self-organization of patterns, the state of each intracellular network is initialized randomly by setting the activation of each gene to on or off with equal probability. To simulate the emergence of patterns over the modeled epithelium, the state of the system is clocked synchronously until either a steady state or the maximum number of updates is reached.

Synchronously clocking the network is an abstraction of the actual process of biological regulation [8], in which the underlying molecular events are stochastic and execute at different temporal scales. To better represent this process, Boolean network dynamics can be asynchronously updated [59] by applying the regulatory rules in a random order, along with differential time delays to pre- and post-translational events [60]. The inherent molecular stochasticity can be likewise modeled by randomly switching the state of genes [61] during dynamics.

In this work synchronous updating is applied both for its simplicity and its ability to reproduce biological observations with sufficient fidelity; evidenced by the diversity of modeling systems that employ this approach including eukaryotic cell dynamics [62], yeast transcription networks [63], and *Drosophila* segment formation [1, 64]. In [65] a study was presented suggesting that synchronous methods can approximate those that employ asynchronous approaches. Furthermore, extensive studies of the *Drosophila* segment formation network [60] demonstrated that synchronous updating can converge to the same attractors as asynchronous updating. Finally, using the synchronous approach considerably eases the detection of identical intracellular states and subsequently the patterns that emerge in the simulated epithelium.

Network simulation will eventually generate a sequence of states that represent a fixed point, where states start to repeat. These fixed points within each intracellular network are referred to as attractors [8] and may be either a single state, or point attractor, or cyclic, where the state transitions return to a previous state. To detect whether a cell has reached a fixed point, the state of each intracellular network at each time point is compared to all its previous states. If a single match is found, an attractor has been reached since the updates are deterministic. If no cycle is detected within the maximum number of steps, the cell is considered to be in a chaotic state.

The assertion that attractors of genetic regulatory networks are terminally differentiated cell types is gaining acceptance in the scientific community [16, 17, 66]. The process of cell differentiation is then the sequence of network states that converges to an attractor. In this work, if the state of two cells converge to the same attractor, even if they are out of phase, then they are considered as the same cell type.

3.3.2 Information Complexity

Set complexity [54] can be used to measure the information content of a regulatory network, its temporal dynamics, and the spatial pattern produced. By measuring information content, set complexity can distinguish between critical systems that encode maximal information, and ordered and chaotic systems that encode low information. Set complexity (symbolized as Ψ) applies Kolmogorov's intrinsic complexity [67] to quantify contextual information in a set of objects by discounting pairs of objects that are randomly related or redundant. Set complexity is independent of any specific application, so long as each object in the set can be encoded as a string.

The Kolmogorov complexity of two strings is the length of the shortest algorithm that can transform one string to the other. Exact computation is undecidable, but minimum algorithm length can be approximated by the normalized compression distance (*NCD*) described in [19] and [20]. *NCD* is defined below, where s_i and s_j are strings, $s_i + s_j$ is the concatenation of s_i and s_j, and $C(s)$ is the compression size of string s:

$$0.0 \leq \mathit{NCD}(s_i, s_j) = \frac{C(s_i + s_j) - \min(C(s_i), C(s_j))}{\max(C(s_i), C(s_j))} \leq 1.0 \tag{3.1}$$

NCD is a measure of the similarity of the two strings [43,44]. If the two strings compress smaller together than separately, then *NCD* will be closer to 0.0. Consider the following cases, where s^r is from the set of random strings and s^s are simple strings containing regular patterns:

1. $\mathit{NCD}(s_i^s, s_j^s) \simeq 0.0$ since $C(s_i^s + s_j^s) \simeq C(s_i^s) \simeq C(s_j^s)$.
2. $\mathit{NCD}(s_i^r, s_j^r) \simeq 1.0$ since $C(s_i^r + s_j^r) \simeq C(s_i^r) + C(s_j^r)$
3. $\mathit{NCD}(s_i^r, s_j^s) \simeq 1.0$ since $C(s_i^r + s_j^s) \simeq C(s_i^r)$ and $C(s_j^s) \simeq 0.0$

Then set complexity of a set of n strings $S = \{s_1, \ldots, s_n\}$ is defined:

$$\Psi(S) = \frac{1}{n(n-1)} \sum_{s_i \in S} C(s_i) \sum_{s_j \neq s_i} d_{ij}(1 - d_{ij}) \tag{3.2}$$

where $d_{ij} = NCD(s_i, s_j)$. The distance d_{ij} is maximized when $NCD(s_i, s_j) = 0.5$, which occurs when $C(s_i + s_j) \simeq C(s_i)/2 - C(s_j)$, assuming $C(s_i) > C(s_j)$. In the case of strings in the set being similar, $\Psi(S) \simeq 0$ indicating the set belongs to the ordered domain and contains little information. Chaotic systems generate strings that appear random and so $\Psi(S)$ is minimized, but not zero because of the $C(s_i)$ multiplicative term. In [54] it is shown that $\Psi(S)$ is maximized when the set of strings describe an information dense critical system.

To ensure accurate measurement of compression length the block size of the compressor must be greater than the string length. Here we used the bzip2 compression algorithm with a block size of 900 Kbytes [68].

3.3.3 String encoding of networks, network dynamics, and patterns

To compute the set complexity of any set of objects, each must be encoded as a string by a one-to-one mapping so that no information is lost. The method by which each random network, temporal dynamics, and the spatial pattern produced are encoded as a string is described below. Studies in [21] suggest that *NCD* and Ψ are in general insensitive to the specific encoding methods employed so long as the compression methods are effective. Let n be the number of Boolean functions in each intracellular network, k be the number of input connections of each function and m^2 be the total number of cells in the pattern (for a square pattern of $m \times m$). The following mappings were employed:

Network: The method used is described in the supplementary materials of [69]. Here the complete intercellular network is represented as a directional connectivity matrix with side m^2nk where each Boolean variable is assigned a unique identifier. The matrix is then represented in row-order and encoded as a string. Each Boolean function is encoded by 2^k 1's or 0's, one for each row in the function table, along with the k identifiers of its input variables. The two strings are then concatenated.

Temporal dynamics: To simulate pattern formation, each network is executed for 300 time steps with a "burn in" period of 100 steps [21]. The burn in period is ignored in the analysis of the dynamics. The 2D space-time matrix of the network state trajectory with size $200m^2n$ is then encoded as a row-order string of 1's and 0's.

Spatial pattern: At the completion of the forward simulation of the network, the dynamics of each intracellular network is analyzed to identify cyclic attractors by searching for repeating states. Then each cell is assigned a cell type ID by performing $200m^2$ comparisons where matching attractors are assigned the same type (irrespective of phase). The string is then a row-order concatenation of each cell's type ID in the $m \times m$ simulated epithelium.

3.4 Experimental study

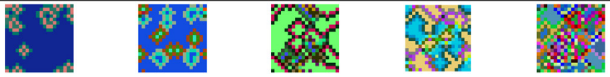
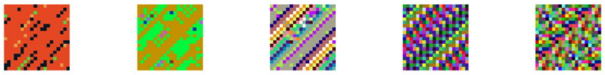
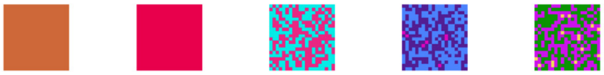
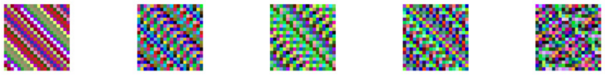
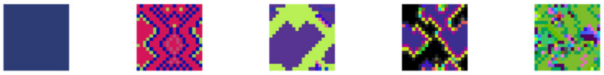
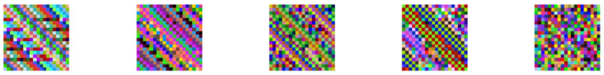
Ψ_p	BW	Comm.	Domain	Example Patterns
9.38	6	sym.	ordered	
18.85	5	orth.	ordered	
19.92	5	sym.	critical	
26.94	6	orth.	critical	
6.72	7	sym.	chaotic	
12.26	8	orth.	chaotic	

Figure 3.5. Examples of patterns from result sets showing their Ψ_p value (the set complexity of patterns), the bandwidth (BW is number of communicating genes), the intercellular signaling configuration (orth; is orthogonal, sym; is symmetrical), the cell-cell communication configuration (comm: sym is symmetric, orth is orthogonal) and the complexity domain (ordered, critical, chaotic) of the intracellular network. Each cell in the pattern is colored according to its attractor (same attractor, same color). Patterns are ordered left to right by increasing compression size.

In this study the number of Boolean functions in each intracellular network is fixed at eight and the pattern is fixed at a 20 by 20 square arrangement of cells. These values represent a balance between computational feasibility and realism. The entire intercellular network contains 3200 Boolean functions. To simulate the activity of the network, each gene in each cell is randomly assigned a value of true or false and stepped forward 300 iterations as described in Section 3.3.3.

With three complexity domains of intracellular networks (ordered, critical and chaotic), two communication configurations (symmetric and orthogonal), and nine bandwidths (zero through eight) there are 54 experimental conditions. For each condition, 100 random networks were constructed and each executed 10 times from a distinct random initial state. For each run, the specific network, its temporal dynamics and the resulting spatial pattern were encoded into strings as described in Section 3.3.3 and stored in separate folders. Given these parameters, the string size of the network is $3200k2^k$ characters; the string size of the dynamics is $64 \times 10^3 k$ characters; and the string size of the pattern is 400 characters, where 400 is the maximum number of unique attractors. Additionally, each spatial pattern was recorded as an image, examples of which are provided in Figure 3.5.

Results presented in Section 3.5 were computed for each experimental condition above using four hundred network repeats. For every execution of a network, its dynamics and pattern were encoded as strings and stored. For each of these string sets, 2000 *NCD* values were computed by randomly sampling string pairs. Not all pairs were considered because the total number of *NCD* values grows as the square of the string set cardinality (see Equation (3.1)). Next, Ψ was computed for the network, dynamics and pattern string sets for each of the 54 experimental conditions. Ψ was estimated from sampling by averaging 100 distinct set complexity computations, each determined from a random sampling of 10 *NCD* values. Sampling was used since the run time of set complexity grows as the square of *NCD* set cardinality (see Equation (4.1)).

3.5 Results and discussion

Three studies were conducted that are described below.

A sample of the results from the first study is given in Figure 3.5. This figure illustrates six pattern sets that emerged from running the 54 combinations of network complexity domains, signaling configurations, and bandwidths of communication described above. Each experimental condition produced a diversity of patterns depending on the topology of each randomly generated intracellular network and its Boolean function values. However, even in the face of these randomized conditions, common patterning themes are apparent within each experimental condition. For example, the first row of patterns in Figure 3.5 was generated by networks operating in the ordered domain and with cells communicating symmetrically. These patterns are all simple and composed of regular patches set on a uniform background, and have a low Ψ_p (pattern set complexity) of 9.38. In contrast, the fourth row of patterns that emerged from networks operating in the critical domain and cells linked by orthogonal communication shows complex diagonal repeating elements with varying periodicity and high Ψ_p of 26.94. In general, symmetric signaling tends to produce patterns that contain contiguous regions and maze-like interfaces that have low Ψ_p, while orthogonal signaling tends to produce repeating regular pattern elements that have high Ψ_p.

3.5.1 Distributions of *NCD* values

The second study investigated the distributions of *NCD* values within each set of networks Figure 3.6 and patterns Figure 3.7. Distributions of the dynamics were not included in the results because the emphasis is on the relationship between network complexity and pattern complexity. These *NCD* values were computed from ordered, critical and chaotic intra-cellular networks that communicate by symmetrical or orthogonal signaling. When many pairs of strings have an *NCD* value near 0.5, then Ψ is often high because the sum of mutual Kolmogorov information $\mathit{NCD}(s_i, s_j)(1 - \mathit{NCD}(s_i, s_j))$ will be high. If most pairs of strings are identical or random, then *NCD* will exhibit a bimodal distribution at 0.0 and 1.0, and the set typically has a low Ψ value. The specific value of Ψ for each string set will be dependent on both the *NCD* distribution and the compression sizes of the individual strings. Results of the Ψ experiments are discussed in Section 3.5.1.

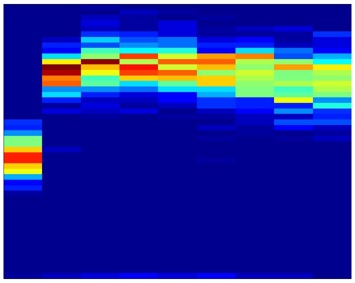
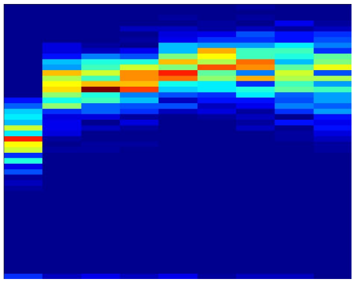
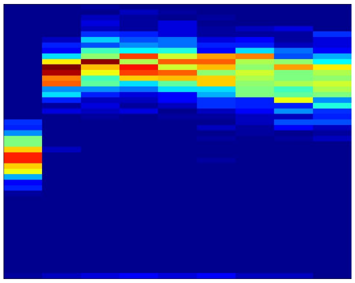
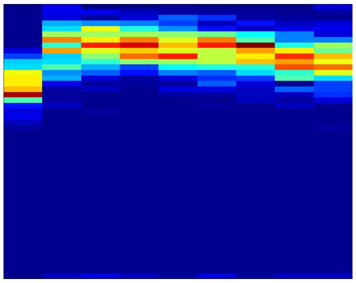
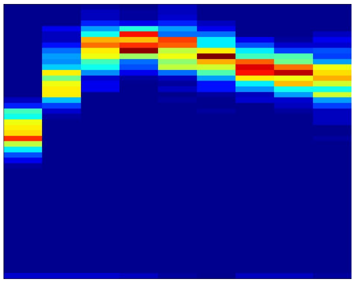
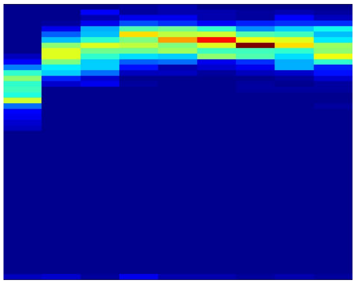
Network Domain	**Communication Type**	
	Symmetric	Orthogonal
Ordered	(a) 	(c) 
Critical	(a) 	(c) 
Chaotic	(b) 	(b) 

Figure 3.6. Distributions of network *NCD* values between network pairs as a function of the number of communicating genes from 0 to 8 along the horizontal axis. The vertical axis for each plot is *NCD* from 0.0 at the bottom to 1.0 at the top. High probability is red, low probability is blue.

Distribution of *NCD* network values

Figure 3.6 shows that *NCD* distributions are remarkably similar for all the network configurations. In addition, each experimental condition exhibits low variance, irrespective of the bandwidth of the networks. Higher *NCD* values indicate dissimilarity between network strings due to the random construction of the intracellular networks in each simulated epithelium. The signaling configuration plays a small role in shaping the distributions be-

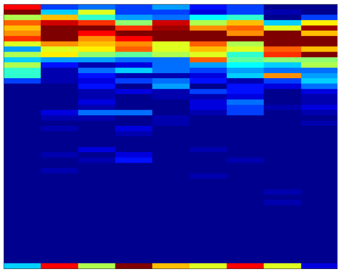
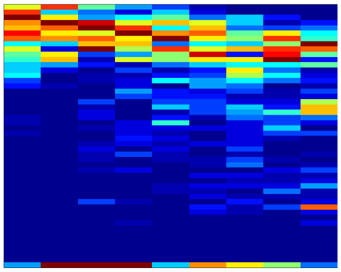
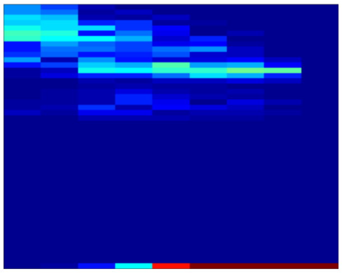
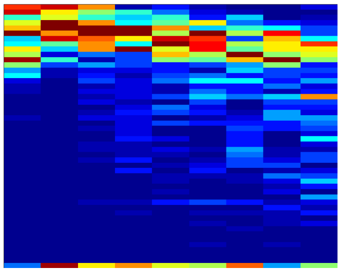
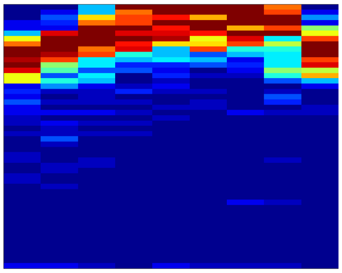
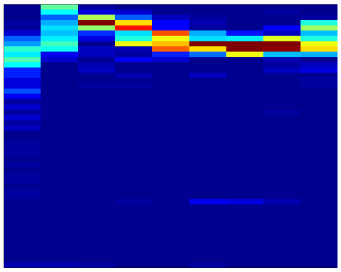
Pattern Domain	**Communication Type**			
	Symmetric		Orthogonal	
Ordered	(a)		(c)	
Critical	(a)		(c)	
Chaotic	(b)		(b)	

Figure 3.7. Distributions of pattern *NCD* values between pattern pairs as a function of the number of communicating genes from 0 to 8 along the horizontal axis. The vertical axis for each plot is *NCD* from 0.0 at the bottom to 1.0 at the top. High probability is red, low probability is blue. (a)(c) refer to the equivalence classes discussed in Section 3.5.1.

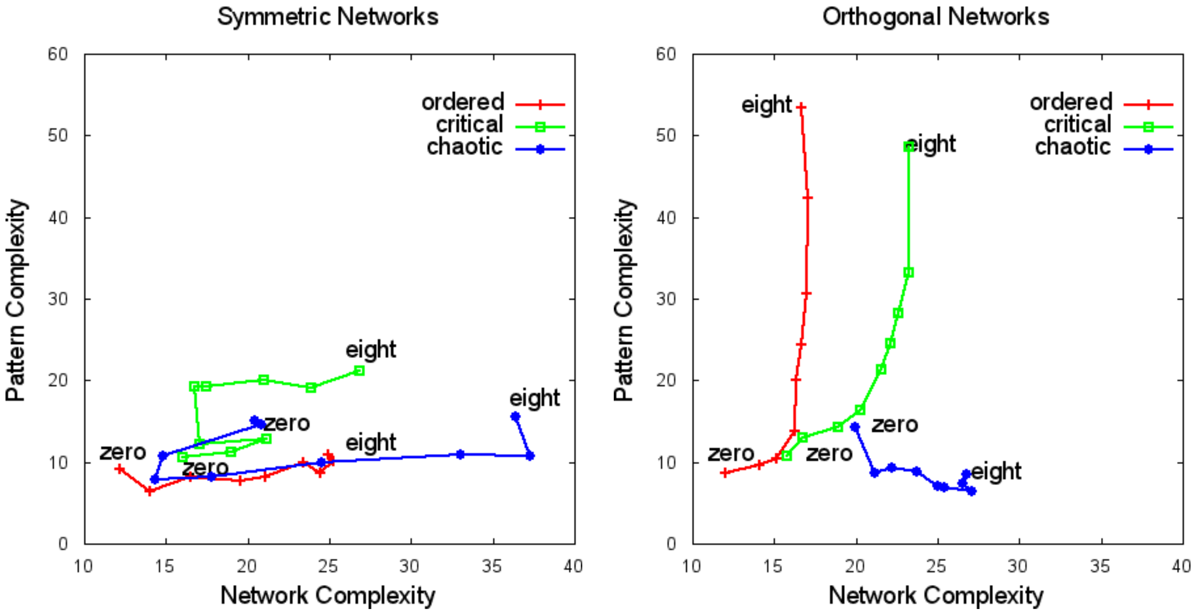
cause specific encodings of intercellular connectivity produces little variation in the strings. The minor role of variation in *NCD* values as bandwidth increases is predicted by the binomial distribution of signaling configurations. Given a network with bandwidth i, there are $\binom{n}{i}$ possible signaling configurations. At low and high bandwidths the number of possible signaling configurations is low, so *NCD* is low. At intermediate bandwidths, the number of signaling configuration is high and therefore *NCD* is high.

Distribution of *NCD* pattern values

Figure 3.7 shows that the six network configurations produce three different types of *NCD* distributions that we call equivalence classes. The first equivalence class (Figure 3.7(a)) is created by ordered or critical networks connected by symmetric signaling between cells. These network configurations produce bimodal *NCD* distributions with maxima at 0.0 and near 1.0. The 0.0 maximum is a consequence of the majority of cells differentiating to the same attractor. This occurs when the intracellular network only has a few attractors, or when the intercellular connectivity over-constrains the attainable attractors. The 1.0 maximum is a consequence of cells converging to many distinct and independent attractors. This results in patterns with little or no spatial organization. Symmetric connections limit information transfer among cells because signals from neighbors are combined using disjunction, and this leads to loss of directional information.

The second equivalence class (Figure 3.7(b)) is observed when the intracellular networks are chaotic, irrespective of the signaling configuration. Here, all *NCD* values are near 1.0 because the patterns are either disordered or complex but with many imperfections (as illustrated in Figure 3.7). These imperfections are cells whose intracellular network dynamics are in long or unlimited attractor cycles (a characteristic of chaotic networks (Kauffman, 1993) and are therefore classified as unique cell types. Significantly, the addition of information transfer between cells by orthogonal signaling prevents adjacent cells from converging to the same attractor.

The third equivalence class (Figure 3.7(c)) is the most complex and is observed with orthogonal signaling and intracellular networks that are either ordered or critical. Here, the distribution has a significant population around 0.5 *NCD* and each pattern has a high compression size. We also observe that these high information patterns increase with the number of signals, particularly when there are six or more signals sent through communicating genes. This highest complexity equivalence class appears only under orthogonal signaling, likely because this signaling configuration promotes long range information transfer between cells.



(a) Symmetric communication (b) Orthogonal communication

Figure 3.8. Network vs. pattern complexity. The relationship between network complexity and the subsequent pattern complexity for the two signaling configurations. Each line shows the trajectory as the signaling bandwidth increases from zero to eight for ordered, critical and chaotic networks.

Relationship among network and pattern complexity

The third study investigated how the complexity of intra-and intercellular networks impact the resulting epithelia patterns. Parametric plots that relate network and pattern Ψ are given in Figure 3.8 for symmetric and orthogonal signaling configurations. Each graph includes relationships for ordered, critical and chaotic networks as the signaling bandwidth grows.

Results show that symmetric communication is sufficient to generate low complexity patterns in the simulated epithelium (Figure 3.8(a)). The network complexity domain has a negligible effect on pattern complexity, with ordered, critical and chaotic domains producing a narrow range of low complexity patterns. Increases in signaling bandwidth produce modest increases in network complexity for ordered or critical networks. When intracellular networks are chaotic, increasing bandwidth leads to increases in network complexity, but without an increase in pattern complexity.

The introduction of directionality in signaling results in significant changes in pattern complexity (Figure 3.8(b)). For critical and ordered networks, as signaling bandwidth grows from 1 to near half the number of intracellular genes, the network complexity reaches a maximum (as discussed in Section 4.2.1). Increasing signaling bandwidth beyond this number of genes maintains network complexity but significantly increases pattern complexity. A maximum in pattern complexity is reached when every intracellular gene is communicating. In contrast to ordered and critical networks, chaotic networks that use orthogonal signaling do not develop complex patterns at any communication bandwidth.

3.6 Summary

This work has explored the potential of ordered, critical or chaotic genetic regulatory networks to create complex patterns in a simulated field of embryonic cells. The impact of the transition from autonomous cells to cells that communicate by contact-mediated signaling was examined as the number of signaling connections increase. An information theoretic measure was used to evaluate the information content of the originating networks, the network dynamics and the emergent cellular patterns. The most complex patterns emerge from ordered and critical networks that communicate directionally. When cells communicate with all neighbors isotropically, only simple, low information patterns emerge. Low information patterns also emerge from chaotic networks regardless of the signaling bandwidth or configuration.

In networks that operate in an isotropic environment (symmetric networks), critical networks generated the most complex patterns (see Figure 3.8(a)), but only when there were four or more communicating genes. This is consistent with previous reports that conclude that critical networks are centrally important in biology [69–71]. More complex patterns arise when there is directionality to intercellular signaling (orthogonal signaling; Figure 3.8(b)). A surprising result was that with directional signaling, ordered networks produce patterns as complex as critical networks and do so at lower levels of network complexity. With orthogonal signaling, there appeared to be a critical point as signaling bandwidth increased. Below this point, there was little effect on pattern complexity of increasing the number of communicating genes. Above this point, there was a sharp increase in pattern complexity as the number of communicating genes increased. This transition occurred between 2 and 3 communicating genes for ordered networks and 3 and 4 communicating genes for critical networks.

What is the biological significance of these results? The first point is that without directionality within a field of cells, critical networks are much more effective in generating simple patterns than either ordered or chaotic networks. Significantly, for these networks to create patterns effectively, there must be a minimum number of communicating genes. In a biological context, communicating genes correspond to independent signals sent and received by neighboring cells. Once primitive patterns are generated, they break symmetry and may then be used as a stepping stone to more complex patterns. The newly established asymmetry creates directionality within the field of cells. This directionality may been visioned as corresponding to one or more of the embryonic axes. Regardless of whether directionality in the embryo is created solely by interactions between adjacent cells or is imposed by a longer range morphogen gradient, once symmetry is broken, much more complex patterns can be generated.

Within an anisotropic environment, ordered networks appear to be at least as effective as critical networks in producing complex patterns. As for the initial symmetry breaking event, there appears to be a minimum number of communicating genes required for effective pattern generation within an anisotropic environment. Once this threshold is crossed, increasing the number of communicating genes produces a linear increase in pattern complexity.

A second biological implication of this work relates to the evolution of patterning mechanisms. A speculative interpretation of these findings is that if an ancestral unicellular organism possessed a relatively small number of genes that orchestrated the collective behavior of these cells, for example, in processes such a quorum sensing, then it is possible that if these cells formed aggregates, few if any additional genes would be needed to create patterns relevant to multicellular development. If these primitive multicellular aggregates presented a selective advantage, then the evolution of additional intercellular communication genes could produce a monotonic increase in pattern complexity. Complex patterns of differentiated cells could emerge in a two-step process in which critical networks operating in an anisotropic group of cells create one or more axes, followed by the operation of either ordered or critical networks to increase pattern complexity.

CHAPTER 4

THE ROLE OF NETWORK MOTIFS IN EPITHELIAL PATTERN FORMATION: A KOLMOGOROV COMPLEXITY STUDY

4.1 Abstract

Genetic regulatory networks consists of quasi-autonomous subnetworks referred to as modules. Such modular networks determine the cellular patterns in multicellular organisms during development. However, the role of modularity in this process is poorly understood. This study applies methods of information theory to explore how network modularity influences the complexity of multicellular patterns that emerge from the dynamics of the regulatory networks. A computational study was performed by creating Boolean intracellular networks of varying degrees of modularity within a simulated epithelial field of embryonic cells. Each cell contains the same network and communicates with adjacent cells using contact-mediated signaling. The study explored two types of modules: motifs, which are subnetworks with unique connectivity and regulatory functions, and clusters, which are densely connected sets of genes sparsely connected to other genes. Comparison of random networks to those with clusters and motifs demonstrated that: (1) Networks with clusters tend to produce more complex multicellular patterns without a significant increase in the gene expression dynamics. (2) Motifs with feedback loops increase information complexity of the multicellular patterns while simplifying the network dynamics. (3) Negative feedback loops effect the dynamics complexity more significantly than positive feedback loops.

4.2 Introduction

Understanding how multicellular patterns form during development is a significant challenge in biology (Figure 4.1). These multicellular patterns emerge as a result of genetic regulatory networks (GRNs) that operate within cells [5]. GRN's are networks of interacting genes that control biological processes. The gene expression profile for each cell is then determined by signaling within and among other cells, differentiation thus making the body plan and subsequence morphology [6].

To understand how GRNs regulate biological events, scientists have developed mathematical and computational models to generate predictions and explain experimental observations. Among these modeling approaches is a simplified modeling technique that considers GRNs as Boolean networks in which the activity of a gene is either on or off, with the activity of a particular gene controlled by a set of logical rules involving the set of regulatory inputs to that gene [8].

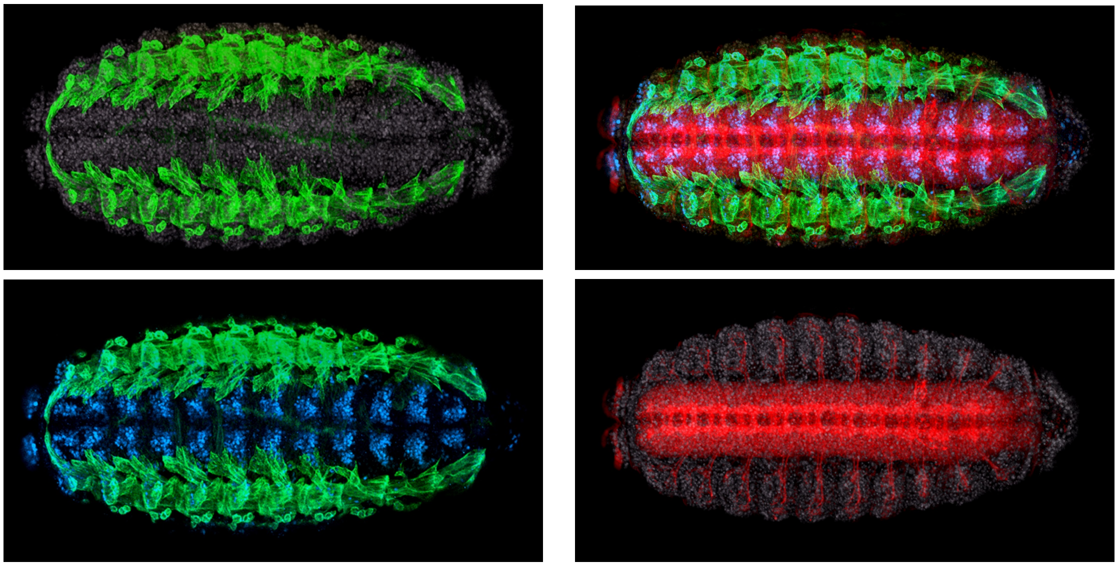


Figure 4.1. Ventral view of stage 16 *Drosophila melanogaster* embryo immunostained for tropomyosin (green; a protein expressed in muscle), Pax 3/7 (blue; a regulatory protein expressed in central nervous system nuclei and ectoderm), and HRP (red; neurons). All nuclei shown in gray (DAPI). Courtesy of Julieta Mara Acevedo and Lucas Leclere, Marine Biological Laboratory, Woods Hole, www.mbl.edu/ dev.biologists.org/

Boolean networks were employed in this study to investigate how Genetic regulatory networks operate in three complexity domains: ordered, critical and chaotic [48] [49] [50]. In the Order systems some events happens more frequently than others and they are more accurately predictable, but at the cost of limited potential for change. The parameter that defining the behavior of chaotic systems are random, magnify the perturbation and do not evolve with time. Chaotic systems are unsuitable for homeostatic living systems and in fact they are implicated in diseases like cancer [72]. Critical systems, which operate in between order and chaos, are the most information dense in both network organization and dynamics and support the efficient prorogation of this information in evolution. This work aims to explore how the information complexity of multicellular pattern and dynamics of GRN is impacted by one of the most influential network configurations: modularity.

GRNs consist of clusters of genes referred to as modules. Modules are the building blocks of the complete cellular network. There are two kinds of modules, structural and functional. If a module contains a set of genes that are densely connected to one another but sparsely connected to other genes within the network then the module is structural [73], [74]. Modules are functional when they are defined as small interconnected networks of genes that are not necessarily structurally distinguishable from other part of network, but by distinctive gene regulatory rules that have been found to be enriched over the population of extant networks. Functional modules also are referred to as a network motif [2].

Developmental biologists have proposed that modularity in organisms arises from modularity in the gene regulatory networks [75], [76]. However this question is difficult to answer since modular developmental networks are poorly understood. In this study we perform computational experiments that aim to answer these questions: How is the information content of multicellular patterns and the dynamics of GRNs influenced by structural modularity of the networks? How do network motifs impact multicellular development to produce information dense patterns? To begin answering these questions we use an information theoretic approach known as Kolomogrov complexity to measure the information content of GRN dynamics and multicellular pattern complexity.

To evaluate the influence of structural and functional modularity on network dynamics and pattens, we design GRNs that are embedded into cells arranged in a 2D grid, simulating an epithelium. Each cell contains an identical Boolean network, referred to as a complete network. In the first study, complete networks were created with multiple modules that are sparsely connected to each other. We investigated how the structural modularity of the network influences the dynamics of the network and complexity of the multicellular pattern formed by altering the sparsity of module connections.

In the second part of this work we explore the influence of the best understood motifs on network dynamics and multicellular patterns by inserting them into randomly generated Boolean networks. The goal was to understand the influence of these motifs on the behavior of the global regulatory system and the multicellular patterns that arise from the network dynamics.

4.3 Modularity of gene regulatory networks

The role of modularity in cellular function and organization has been extensively studied [77]. It is believed that modules perform relatively independent tasks in gene regulatory networks [78], [75]. The modular organization of biological structure is supported by experimental studies from pathogen structure, gene networks, and protein-protein interaction networks [79]. For example, Kim et al. [78] studied the connected subset of protein networks in protein-protein interaction data for budding yeast. Their analysis suggests that the yeast protein network is significantly modular. Networks are structurally modular if they contain highly connected clusters of genes that are linked by sparser connections than those within the modules. Figure 4.2 shows a small network with a modular structure and a randomly connected networks.

We refer to the type of modularity illustrated in Figure.4.2(a) as structural modularity where individual modules are densely connected networks without any specific function. Of course, structural modules may have a function, but functionality is not how they are recognized. In contrast to structural modules, functional modules are defined as a set of interconnected genes that produce a distinct function, regardless of whether they are structurally isolated within a network. Functional modules that occur frequently and consist of few interacting genes are referred to as regulatory motifs [2]. This work considers both structural and functional kinds of modularity.

Regulatory motifs were first noted in *Escherichia coli*, where they were detected at a higher frequency than would be expected in random networks. Since then multiple motifs

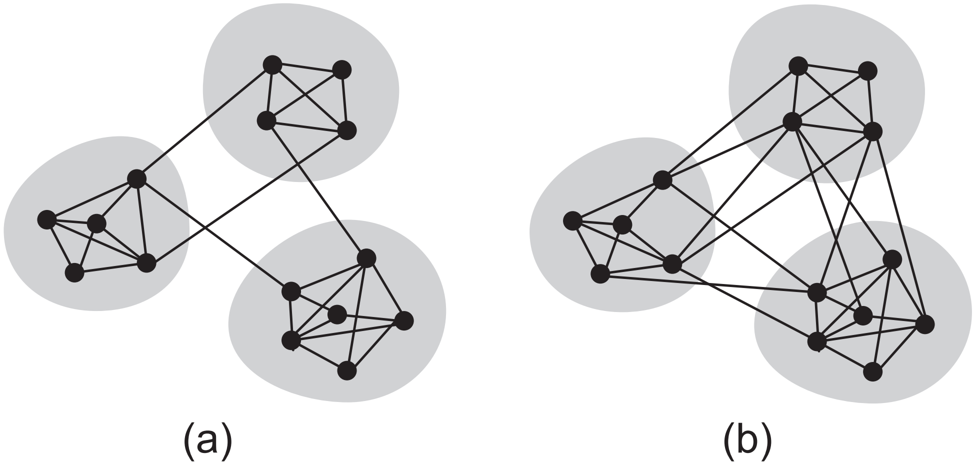


Figure 4.2. Structural modularity. (a) A network with modular structure where intra-modular connectivity is higher than inter-module connectivity. (b) A randomly connected network.

have been identified in bacteria and yeast [80]. This finding suggests that motifs are building blocks of transcription networks and that they may have evolved to achieve specific regulatory behaviors in cellular transcription networks [3]. Regulatory motifs may be found in two different regulatory networks: 1- Developmental networks that guide differentiation and cell fate determination by transducing signals into irreversible cell-fate decisions [81] [82] and 2- Sensory networks that respond to signals such as stresses and nutrients rapidly and make reversible decisions [83].

The motifs that are associated with developmental networks are commonly comprised of feedback loops. Positive feedback loops are most common and are made up of two transcription factors that regulate each other. There are two kinds of positive feedback loops, a double-positive loop (Figure 4.3(b)) and a double-negative loop (Figure 4.3(a)). The regulatory dynamics of these gene pairs coupled by positive feedback loops often results in two or more steady states and is referred to as multistability [3]. Positive feedback loops amplify signals and elongate the time required to reach to a steady state [80]. This slowed response can be helpful when a cell makes significant decisions such as irreversible cell

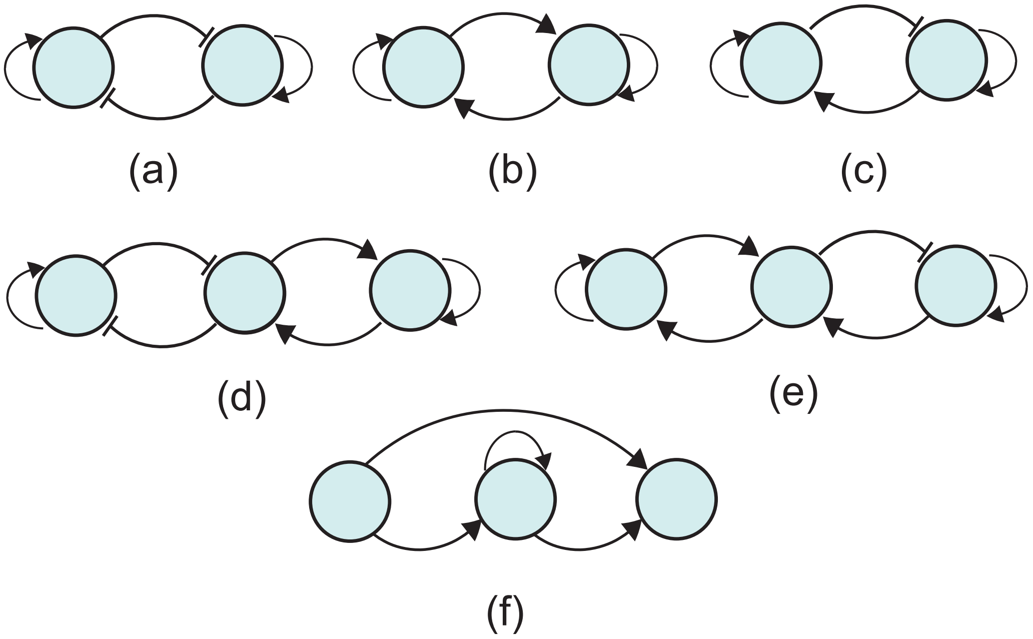


Figure 4.3. Functional modularity. (a) A positive feedback loop (double-negative loop with two positive autoregulatory loops [2]). (b) A positive feedback loop (double-positive loop with two positive autoregulatory loops). (c) A negative feedback loop [3] with two positive autoregulatory loops. (d) Coupled positive-positive feedback loops. (e) Coupled positive-negative feedback loops. (f) The type-1 coherent feed forward loop [4].

specification and apotosis. Unlike positive feedback loops, negative feedback loops (Figure 4.3(c)) often enhance attractor stability. They also function as noise filters and make cells more robust to signal noises. In addition, positive and negative feedback loops are coupled into structures containing two feedback loops, such as positive-positive, positive-negative and negative-negative feedback loops (Figure 4.3(d,e)). Coupled feedback loops perform functions that single feedback loops cannot. In particular, Kim et al. [3] found that a positive-positive feedback loop enhances signal amplification and bistability and a positive-negative feedback loop increases reliable decision-making by modulating signal responses and effectively dealing with noise.

Feed-Forward Loops (FFL) are another family of motifs that are associated with sensory networks. FFL are found in variety of organisms such as Saccharomyces cerevisiae, Bacillus subtilis, Caenorhabditis elegans and humans [4]. FFL consists of a three genes (Figure 4.3(f)). The first regulatory gene controls the second and the third genes. The third gene also is regulated by the second gene. Logical gates such “AND gate” or “OR gate” could be applied for the three regulatory interactions in the FFL. The best known FFL which occurs frequently in (*E. coli and yeast*), is the coherent type-1 FFL [84] with all “AND gates”.

4.4 Multicellular model and its implementation

The large scale of gene and protein networks drove the decision to use Boolean networks as the framework for this computational study. A Boolean network can be used as a simplified model of a genetic regulatory network. In this application, each gene is represented as a network node that takes binary values (1 for expressed and 0 for not expressed). The state of a gene (0 or 1) is determined by its Boolean function defined as the expressions of AND, OR, NOT on the inputs from other genes. These inputs are represented as directed edges in the network graph. Boolean networks provide a qualitative description of gene states and their interactions, first introduced by Kauffman [8], [9].

This work extends our previous study [5] of complexity of multicellular pattern formation by adding the following features to the model: 1- Representation of gene regulatory network as a structural modular network (Figure 4.4, Also see methodology section) 2- Insertion of motifs into the complete GRN (Figure 4.5).

The simulation model was unchanged, and considered a lattice of cells, with each cell holding a complete Boolean network. Cell-cell signaling was implemented in the model as an edge connecting the state of one gene in a cell to an input of a Boolean function of one or more of its neighbors. Such genes are called communicating genes (indicated at the tails of the larger arrows in Figure.4.4) and the modules containing these genes are referred to as signaling modules, shown with green background color. The number of communicating genes is referred to as the signaling bandwidth.

Signaling bandwidth is set to half of the total number of genes in a cell as our previous study showed that this configuration established effective cell-cell signaling. Also

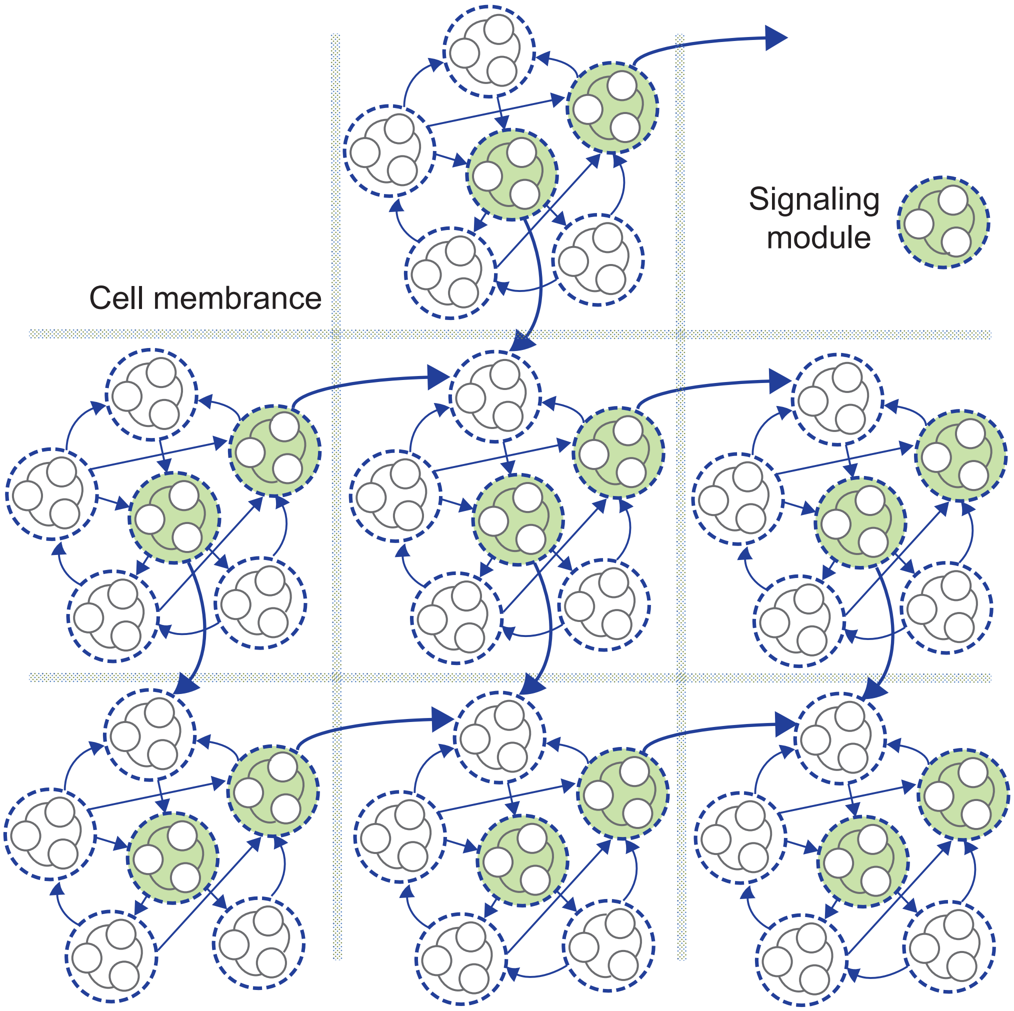


Figure 4.4. Structural modularity implementation. Lattice of cells is illustrated in this figure. Cell-cell signaling is implemented through specific modules referred to as signaling modules (depicted in green). In this example, cell-cell signaling is orthogonal such that two adjacent cells signal directionally [north-south and east-west]. These directions can be thought of as corresponding to the anterior-posterior and dorsal-ventral embryonic axis.

as previously shown, both symmetric and orthogonal signaling configuration is considered for cell-cell signaling [5]. When each cell signals to any other north, south, east and west neighboring cells that signaling is called symmetric signaling and when two adjacent cells signal directionally [north-south and east-west] to corresponded to anterior-posterior and dorsal-ventral embryonic axis, the signaling is called orthogonal signaling.

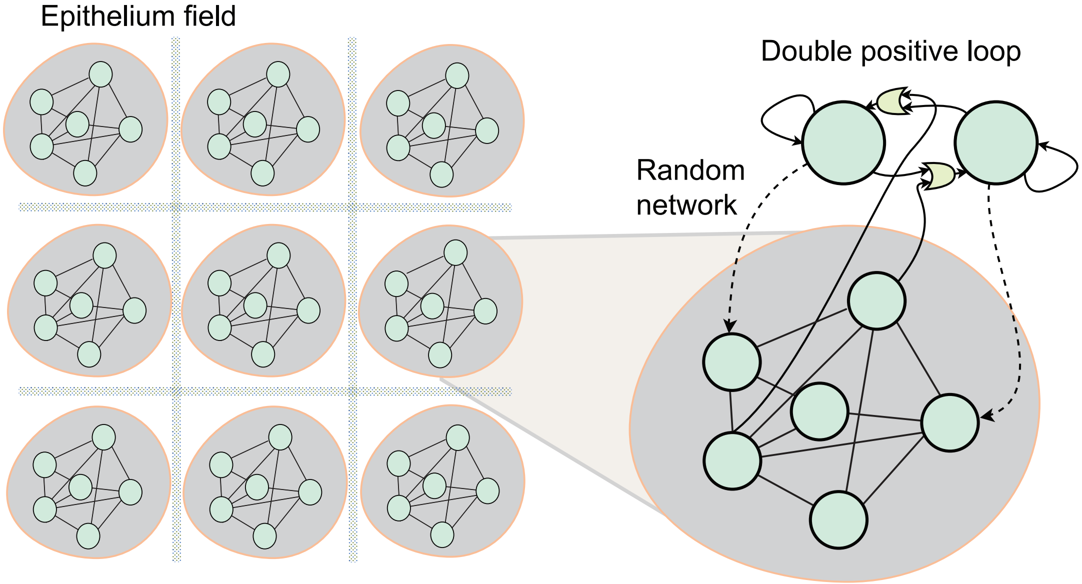


Figure 4.5. Motif insertion. Example motif insertion into a random GRN. Dashed arrows represent random outgoing signals from the motif. Outgoing and incoming signals from and to the random GRN are randomly connected to genes within the random network.

The state of each cellular GRN is initialized randomly by setting the state of each gene to 0 or 1. Randomly generated logic functions are assigned to networks as the transition rules used to determine the state of genes [5]. The state of the system during simulation is clocked synchronously until a steady or cyclic state (in up to 300 repeats) is reached for all individual cells. When the state of genes change in a repetitive cycle or reach to a fixed state then cell are in attractor state [5]. Attractor is used to construct multicellular patterns by treating cells with the same attractor as the same cell types. The state of all the genes as the networks are run along with multicellular patterns is recorded for analysis of information content.

After running the randomly-generated GRNs, single and coupled feedback and feed-forward loops are inserted into the randomly generated GRN (Figure 4.5). The network is run again with the inserted motifs to identify the attractors and visualize the multicellular patterns that are formed. The information complexity of both the gene network dynamics and multicellular patterns is analyzed by an information theoretic measure called Set

Complexity.

4.5 Methodology

4.5.1 Building networks with different modularity degree

To produce networks with different modularity scores, we first construct each individual module as a network where each gene has one or two inputs from other genes. With this configuration, networks operate in critical domain [5]. Critical domains are on the cusp between order and chaos and are the most information dense in both network organization and dynamics [52]. In our previous study [5] we shown that coupled Boolean networks in grid of cells, where each gene in a single network receives up to two incoming signals from other genes randomly, operate in critical domains. Interconnection of modules are implemented by adding random connections between modules. The modularity score of the complete network is decreased when more connections are added. For example if we consider 4 modules to build the complete network, each module contained 4 genes, then we alter the number of random incoming signals to each module from 1 to 6 and we see that by adding more random incoming connections the modularity score of the whole network is decreased.

4.5.2 Information Complexity

Set complexity is an information complexity metric that will be used to measure the information content of each regulatory network, its temporal dynamics and resulting multicellular patterns. Set complexity distinguishes between chaotic, critical and ordered set of objects and is based on Normalized Compression Distance (NCD) [19]. By employing NCD as a metric to evaluate similarity of pairs of objects in a set, set complexity discounts the influence of the pairs of objects that are randomly related or redundant. As long as any object can be encoded as a string, set complexity is able to compute the information content that resides in the set.

Set complexity of a set of n strings $S = \{s_1, \ldots, s_n\}$ is defined:

$$\Psi(S) = \frac{1}{n(n-1)} \sum_{s_i \in S} C(s_i) \sum_{s_j \neq s_i} \mathit{NCD}(s_i, s_j)(1 - \mathit{NCD}(s_i, s_j)) \tag{4.1}$$

where $C(s_i)$ is the compression size of string s_i. The term $\mathit{NCD}(s_i, s_j)(1 - \mathit{NCD}(s_i, s_j))$ is maximized when $\mathit{NCD}(s_i, s_j) = 0.5$, which occurs when $C(s_i + s_j) \simeq C(s_i)/2 - C(s_j)$, assuming $C(s_i) > C(s_j)$.

To encode an object to a string, a one-to-one mapping is required so that no information is lost. The method by which each random network, temporal dynamics, and the spatial pattern produced are encoded as a string is described in the next section.

4.5.3 Encoding objects to strings

Studies in [21] suggest that NCD and Set Complexity are in general insensitive to the specific encoding methods employed so long as the compression methods are effective. Let n be the number of Boolean functions in each intracellular network, k be the number of input connections of each function and m^2 be the total number of cells in the pattern (for a square pattern of $m \times m$). The following mappings were employed:

Temporal dynamics: To simulate pattern formation, each network is executed for 300 time steps with a burn in period of 100 steps [21]. The burn in period is ignored in the analysis of the dynamics. The 2D spacetime matrix of the network state trajectory with size $200 \times m^2 \times n$ is then encoded as a row-order string of 1's and 0's.

Spatial pattern: At the completion of the forward simulation of the network, the dynamics of each intracellular network is analyzed to identify cyclic attractors by searching for repeating states. Then each cell is assigned a cell type ID by performing $200 \times m^2$ comparisons where matching attractors are assigned the same type (irrespective of phase). The string is then a row-order concatenation of each cell's type ID in the $m \times m$ simulated epithelium.

4.5.4 Structural modularity score measurement

The most common method used in the literature to score structural modularity is a method by Newman [85]. Newman's method to compute the modularity score Q for a given network is as follows:

$$Q = \sum_{i=1}^{c} (e_{ii} - a_i^2) \tag{4.2}$$

where c is the total number of modules, e_{ii} is fraction of edges in module i, e_{ij} is the fraction of edges that connect module i to module j and a_i is the fraction of edges that connect module i to other modules and is as follows:

$a_i = \sum_j e_{ij}$

The module break-downs are known since we generate each modules with higher intra-connection than the random connections connecting the modules.

4.6 Results and discussion

4.6.1 Modularity of intracellular gene networks influences multicellular pattern and gene network dynamic complexity

In order to analyse how modularity of a network influences the dynamics and pattern complexity, we partition the population of the networks created by percentile of their modularity score distribution. Networks that lie in lower third are defined as non-modular, and networks that lie in upper third are defined as highly modular. Figure 4.6 illustrates the average dynamics complexity and pattern complexity for 60 non-modular and 60 highly-modular networks. Experiments are run for two orthogonal and symmetric signaling (Figure 4.6), each with two configuration for modules, critical and chaotic. Critical configuration is when genes in each module receive up to two incoming signals from other genes, make them to behave in critical domain when isolated. In chaotic configuration the genes receive up to three incoming signals [5]. Results in Figure 4.6(a) show how highly modular networks produce higher dynamics and pattern complexity when cell-cell signaling is orthogonal. A statistical analysis confirms the distribution of the two classes of networks are significantly different Table 4.1 and Table 4.2.

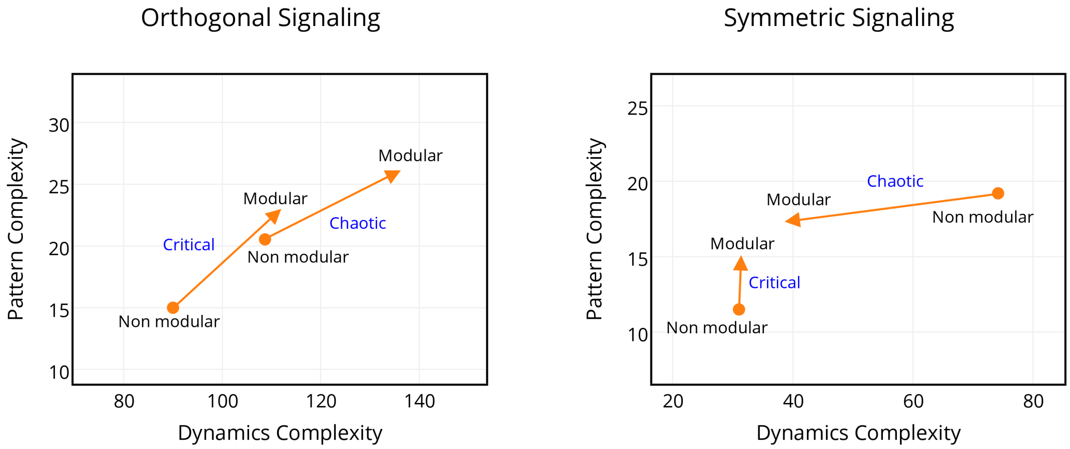


Figure 4.6. Non modular vs. modular. Average dynamics and pattern complexity for 60 non-modular (arrow tails) and 60 highly-modular (arrow heads) networks.

Table 4.1. Orthogonal signaling. Dynamics and pattern complexity change when non modular networks become highly modular networks.

	Critical modular				**Chaotic modular**			
	Pattern complexity		**Dynamics complexity**		**Pattern complexity**		**Dynamics complexity**	
	Non Modular	Highly Modular	Non Modular	Highly Modular	Non Modular	Highly Modular	Non Modular	Highly Modular
Mean	15.593	23.328	90.059	112.25	20.921	26.565	107.85	134.96
Variance	11.282	15.255	1315.4	1680.0	18.769	19.127	1212.8	2035.9
df	59		59		59		59	
t Stat	11.480		3.265		7.993		3.714	
P(T¡=t)	5.6E-17		9E-5		2.8E-11		2E-4	
t Critical	1.671		1.671		1.671		1.671	

Table 4.2. Symmetric signaling. Dynamics and pattern complexity change when non modular networks become highly modular networks.

	Critical modular				**Chaotic modular**			
	Pattern complexity		**Dynamics complexity**		**Pattern complexity**		**Dynamics complexity**	
	Non Modular	Highly Modular	Non Modular	Highly Modular	Non Modular	Highly Modular	Non Modular	Highly Modular
Mean	11.499	14.940	31.017	31.844	19.193	17.479	74.169	39.422
Variance	12.603	11.919	175.18	138.65	18.212	20.827	690.37	273.23
df	59		59		59		59	
t Stat	7.046		0.382		2.526		9.192	
P(T¡=t)	1.1E-09		0.3517		7E-04		2E-13	
t Critical	1.671		1.671		1.671		1.671	

The observation for symmetric signaling with critical configuration for each module Figure 4.6(b)) suggest that the networks with modular structure are able to produce multicellular patterns with higher information content without an increase in the network dynamics complexity.

Another interesting observation was that for symmetric signaling with chaotic configuration (Figure 4.6(b)) both network dynamics and pattern complexity decrease. As shown in the previous study [5] symmetric signaling produces only low information patterns and dynamics because information transfer among adjacent cells combined using disjunction, which loses directional information. Now when configuration of each module is chaotic, structural modularity of network helps to produce more simplified pattern and dynamics.

4.6.2 Insertion of motifs into the randomly generated global networks

In this section we explore the effect of insertion of the best known motifs into randomly generated coupled GRNs. Figure 4.7 shows the influence of insertion of a positive feedback loops (double-positive loop) into a random GRN. The results are represented for two orthogonal Figure 4.7(a) and symmetric Figure 4.7(b) of cell-cell signaling. Insertion of a double-positive feedback loop into random GRNs increases dynamics complexity while leaving pattern complexity unchanged in the case of orthogonal signaling (Figure 4.7(a)). In contrast, insertion of this same double-positive feedback loop into a GRNs operating under symmetric signaling decreases both dynamics and pattern complexity (Figure 4.7(b)).

Figure 4.8 illustrates the average dynamics and pattern complexity for 60 random GRNs (arrow tails) and 60 GRNs with the insertion of various types of regulatory motifs (arrow heads). Under conditions of orthogonal and symmetric signaling, insertion of a negative feedback loop, a double-negative feedback loop or a double-positive feedback loop all have the same qualitative effect of decreasing network dynamics complexity and increasing pattern complexity; however, a double-positive loop increases pattern complexity much more than either of the negative feedback loops (Figure 4.8(a)). Insertion of Feed-Forward loops decreases dynamics with almost no effect on the pattern complexity. ANOVA analysis presented in Table.4.3 and Table 4.4 shows that insertion of these motifs makes a significant change in the average network dynamics and patterns complexity.

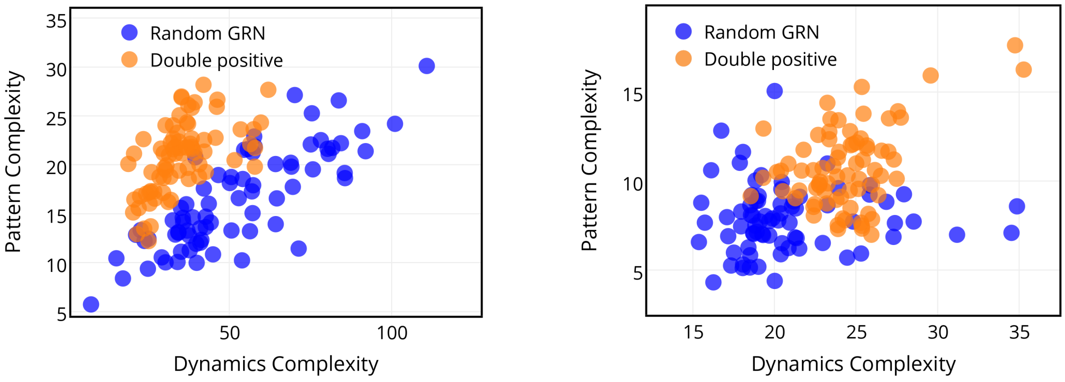
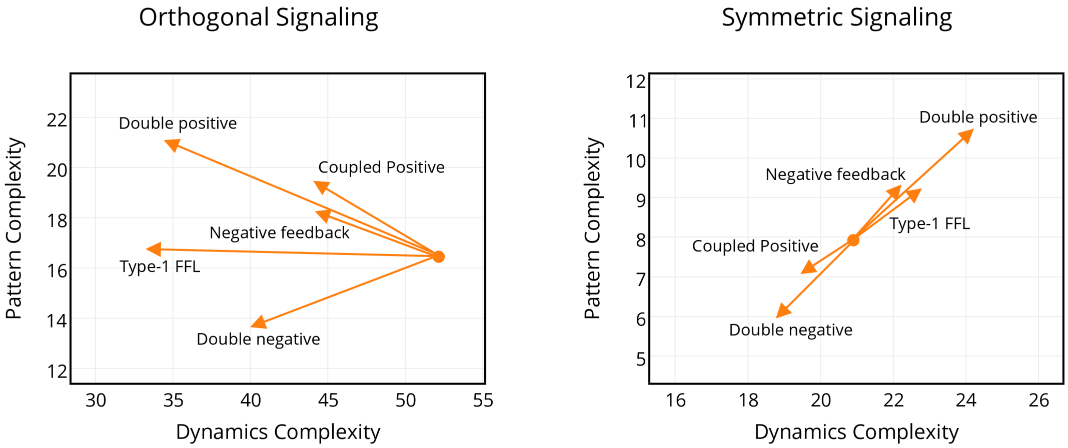


Figure 4.7. Effect of insertion of a double-positive feedback loop on network dynamics and pattern complexity. Insertion of a double-positive loop significantly increases only pattern complexity in the case of orthogonal signaling, and increases both pattern and dynamics complexity in the case of symmetric signaling.


Orthogonal Signaling

Symmetric Signaling

Figure 4.8. Effects on network dynamics and pattern complexity of inserting regulatory motifs into random GRNs. Average dynamics and pattern complexity for 60 random GRNs (arrow tails) and 60 GRNs with the indicated inserted motifs (arrow heads).

All the motifs with feedback loops affect the pattern complexity in orthogonal cell-cell signaling. The only motif in this study that has no effect on pattern complexity are Feed-Forward loops Type-1. This observation is consistent with the association of feedback loop motifs with developmental networks that mediate important cell fate decisions.

We hypothesize that variation in dynamics complexity originates from two different sources. 1- The time for GRN to reach to steady state. 2- The proportion of single state

Table 4.3. ANOVA analysis of dynamics complexity illustrates differences for various groups of motifs over networks with orthogonal and symmetric signaling

	Orthogonal signaling			**Symmetric signaling**		
Groups	Sum	Average	Variance	Sum	Average	Variance
Without Motifs	3911.45	52.15	474.35	1560.25	20.80	15.96
Double-negative	3017.06	40.22	284.05	1423.40	18.97	8.18
Double-positive	2633.55	35.11	101.26	1828.02	24.37	7.64
Negative-Feedback	3322.99	44.30	181.91	1658.00	22.10	8.15
Coupled P-P	3355.92	44.74	306.40	1466.11	19.54	4.17
Type-1 feed Forward	2560.45	34.13	249.67	1718.50	22.91	4.02
Orthogonal signaling ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	17099.6	5	3419.93	12.84	1.1E-11	2.234
Within Groups	118226.6	444	266.27			
Total	135326.3	449				
Symmetric signaling ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1594.73	5	318.94	39.74	9.5E-34	2.234
Within Groups	3563.42	444	8.02			
Total	5158.16	449				

versus cyclic attractors produced by the GRN. Since the motifs studied here act as multi-stable switches, they simplify the complex cyclic attractors to attractors with a few states. Genes in feedback loops reach a steady state expression quickly and reduce the length and complexity of cyclic attractors. We hypothesize that this is why in all the cases of orthogonal signaling, dynamics complexity decreases from that of the original random networks. The rate of dynamics complexity reduction associated with the addition of the motifs with negative feedback loops is significantly lower than for positive loops. Unlike positive feedback loops, negative feedback loops do not increase the time to reach to steady states [3]. Therefore they don't effect the dynamics complexity noticeably. As results shows negative feedback loop motifs (such as single negative feedback loop and coupled positive-negative loops) have the lowest reduction in their dynamics complexity.

With symmetric signaling all the motifs except for those containing double-negative loops increase the pattern and dynamics complexity. The pattern complexity variation are

Table 4.4. ANOVA analysis of pattern complexity illustrates differences for various groups of motifs over networks with orthogonal and symmetric signaling.

	Orthogonal signaling			**Symmetric signaling**		
Groups	Sum	Average	Variance	Sum	Average	Variance
Without Motifs	1233.65	16.44	25.15	592.51	7.90	3.55
Double-negative	1026.61	13.68	22.74	456.57	6.08	1.97
Double-positive	1559.75	20.79	13.10	805.59	10.74	5.08
Negative-Feedback	1457.84	19.43	19.91	687.03	9.16	2.66
Coupled P-P	1372.30	18.29	24.27	530.80	7.07	2.35
Type-1 feed Forward	1250.56	16.67	9.22	692.85	9.23	3.22
Orthogonal signaling ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	2366.77	5	473.35	24.82	4.4E-22	2.234
Within Groups	8466.32	444	19.06			
Total	10833.10	449				
Symmetric signaling ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1057.66	5	211.53	67.26	3.0E-52	2.234
Within Groups	1396.30	444	3.14			
Total	2453.97	449				

similar to networks with orthogonal signaling that confirms the behavior of feedback loops as a cell differentiation facilitator. Tendency of motifs to simplify the cyclic and random attractors emerges primarily in networks with symmetric signaling.

4.7 Summary

In the first part of this study we explored the role of compartmentalization of GRNs into modules on network dynamics and pattern complexity. The results show that networks with a modular structure tend to produce more complex multicellular patterns without a significant increase in gene expression dynamics. In the second part of this study we explored the role of common regulatory motifs on network dynamic complexity and pattern complexity. These motifs appear frequently in biological networks and often play critical roles in overall network function. Although the significance of these motifs have been shown in multiple studies, there is a lack of computational studies to explore how and to what degree biological network dynamics and the resulting multicellular patterns are influenced by network motifs. The results shows that network motifs that are associated with feedback loops increase the information complexity of the multicellular patterns regardless of whether cell-cell signaling occurs symmetrically or orthogonally. Another important observation was that negative feedback loops do not effect the dynamics complexity significantly as positive feedback loops do.

CHAPTER 5

CONCLUSIONS

In this dissertation we explored the application of Algorithmic Information Theory (AIT) for two case studies: bright field cell image segmentation and pattern formation in multicellular organisms. As the first study showed, AIT can be employed as an effective preprocessing step in cell image segmentation. We demonstrated that selecting frames with our proposed AIT-based algorithm will result in more accurate cell image segmentation because it discards noisy images. In the second study, which was the primary contribution of this dissertation, we employed an AIT-based algorithm to quantify the complexity of information content that arises during the development of multicellular organisms. We simulated multicellular organism development by coupling the Gene Regulatory Networks (GRN) within an epithelial field. Primary results showed that structure and function of GRNs impact the complexity of the information content in the resultant multicellular patterns. We demonstrated that some of the GRN classes, in terms of structure and function, produce more complex patterns than others. This finding has biological significance.

In chapter 2 we proposed an AIT-based algorithm called *maximal-information* to solve an image processing challenge in a biological context. Cell segmentation is the identification of cells and their observable properties from cell microscopy images. Bright field microscopy is a simple and common method of cell imaging. Bright field microscopy, however, presents challenges due to low image contrast. Some studies have used a defocused stack of images to acquire more information for an accurate cell segmentation. In this study, the performance of the *maximal-information* method was compared with a recent approach that uses a fixed frame selection strategy in image data of embryonic kidney cells (HEK 293T) from multiple experiments. Results demonstrated that the adaptive *maximal-information* approach significantly improves precision and recall of segmentation over the diversity of data sets.

In chapter 3 of this dissertation, we studied simulated coupled gene networks in an epithelium field of embryonic cells. We used a Kolmogorov complexity-based algorithm to evaluate the information complexity of given Genetic Regulatory Networks, the network dynamics, and the emergent cellular patterns. Our results demonstrated that the most complex dynamics and patterns emerge from networks that communicate directionally. When cells communicate with all neighbors isotropically, only simple, low information patterns emerge. Low information patterns also emerge from chaotic networks (networks in which each gene accepts 3 signals) regardless of the signaling bandwidth or configuration. In networks that operate in an isotropic signaling environment, critical networks (networks in which each gene accepts up to 2 signals) generate the most complex patterns, but only when there were four or more communicating genes. This is consistent with previous reports that conclude critical networks are centrally important in biology. Directional signaling among cells leads to more complex patterns. A surprising result was that directional signaling in ordered networks produces patterns as complex as critical networks, and do so at lower levels of network complexity.

In chapter 4 of this dissertation, we studied the concept of GRN modularity and motifs. It is believed that modules perform relatively independent tasks in cellular function. Due to the importance of network modularity and the significance of biological motifs that naturally exist in the diversity of organisms, in chapter 4 we explored the influence of modularity on network dynamics and patterns. The results demonstrated that networks with modular structure tend to produce more complex multicellular patterns without producing significantly high complexity in gene dynamics. Another important result was that the insertion of some of the well-recognized motifs had a significant effect on the patterns complexity. For example, network motifs associated with feedback loops increase the information complexity of the multicellular patterns regardless of the type of cell-cell signaling.

In this work we demonstrated that Kolmogorov complexity is a powerful measurement tool to quantify the amount of information contained within a phenomenon. We applied Kolmogorov complexity-based algorithms to successfully solve some challenging problems in developmental biology and in bright field image processing.

REFERENCES

[1] R. Albert and H. G. Othmer, “The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster.” *Journal of theoretical biology*, vol. 223, no. 1, pp. 1–18, Jul. 2003. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/12782112

[2] A. Ghaffarizadeh, N. Flann, and G. Podgorski, “Multistable switches and their role in cellular differentiation networks,” *BMC Bioinformatics*, vol. 15, no. Suppl 7, pp. S7+, 2014. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-15-s7-s7

[3] J.-R. R. Kim, Y. Yoon, and K.-H. H. Cho, “Coupled feedback loops form dynamic motifs of cellular networks.” *Biophysical journal*, vol. 94, no. 2, pp. 359–365, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1529/biophysj.107.105106

[4] S. Kalir, S. Mangan, and U. Alon, “A coherent feed-forward loop with a SUM input function prolongs flagella expression in Escherichia coli.” *Molecular systems biology*, vol. 1, no. 1, pp. msb4 100 010–E1–msb4 100 010–E6, Mar. 2005. [Online]. Available: http://dx.doi.org/10.1038/msb4100010

[5] N. S. Flann, H. Mohamadlou, and G. J. Podgorski, “Kolmogorov complexity of epithelial pattern formation: The role of regulatory network configuration,” *Biosystems*, vol. 112, no. 2, pp. 131–138, May 2013. [Online]. Available: http://dx.doi.org/10.1016/j.biosystems.2013.03.005

[6] E. H. Davidson, “Emerging properties of animal gene regulatory networks,” *Nature*, vol. 468, no. 7326, pp. 911–920, Dec. 2010. [Online]. Available: http://dx.doi.org/10.1038/nature09645

[7] B. Ristevski, “A survey of models for inference of gene regulatory networks,” *Nonlinear Analysis: Modelling and Control*, vol. 18, pp. 444–465+. [Online]. Available: http://www.researchgate.net/publication/258285498_A_survey_of_models_for_inference_of_gene_regulatory_networks

[8] S. A. Kauffman, “Metabolic stability and epigenesis in randomly constructed genetic nets.” *Journal of theoretical biology*, vol. 22, no. 3, pp. 437–467, Mar. 1969. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/5803332

[9] S. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press,USA. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1226010/

[10] G. von Dassow, E. Meir, E. M. Munro, and G. M. Odell, “The segment polarity network is a robust developmental module,” *Nature*, vol. 406, no. 6792, pp. 188–192, Jul. 2000. [Online]. Available: http://dx.doi.org/10.1038/35018085

[11] F. Emmert-Streib, “Exploratory analysis of spatiotemporal patterns of cellular automata by clustering compressibility.” *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 81, no. 2 Pt 2, Feb. 2010. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/20365627

[12] S. J. Eglen and D. J. Willshaw, “Influence of cell fate mechanisms upon retinal mosaic formation: a modelling study.” *Development (Cambridge, England)*, vol. 129, no. 23, pp. 5399–5408, Dec. 2002. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/12403711

[13] R. Goodyear and G. Richardson, “Pattern formation in the basilar papilla: evidence for cell rearrangement.” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 17, no. 16, pp. 6289–6301, Aug. 1997. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/9236239

[14] M. Villani, R. Serra, P. Ingrami, and S. A. Kauffman, “Coupled Random Boolean Network Forming an Artificial Tissue,” in *Cellular Automata*, ser. Lecture Notes in Computer Science, S. Yacoubi, B. Chopard, and S. Bandini, Eds. Springer Berlin Heidelberg, 2006, vol. 4173, ch. 63, pp. 548–556. [Online]. Available: http://dx.doi.org/10.1007/11861201_63

[15] A. N. Kolmogorov, “Three approaches to the quantitative definition of information.” *Prob. Inf. Transm.*, vol. 1, pp. 17+, 1965.

[16] S. Huang, G. Eichler, Y. Bar-Yam, and D. E. Ingber, “Cell fates as high-dimensional attractor states of a complex gene regulatory network.” *Physical review letters*, vol. 94, no. 12, Apr. 2005. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/15903968

[17] C. Furusawa and K. Kaneko, “A Dynamical-Systems View of Stem Cell Biology,” *Science*, vol. 338, no. 6104, pp. 215–217, Oct. 2012. [Online]. Available: http://dx.doi.org/10.1126/science.1224311

[18] D. J. Galas, M. Nykter, G. W. Carter, N. D. Price, and I. Shmulevich, “Set-based complexity and biological information,” Jan. 2008. [Online]. Available: http://arxiv.org/abs/0801.4024

[19] X. Chen, B. Francia, M. Li, B. McKinnon, and A. Seker, “Shared information and program plagiarism detection,” *Information Theory, IEEE Transactions on*, vol. 50, no. 7, pp. 1545–1551, Jul. 2004. [Online]. Available: http://dx.doi.org/10.1109/tit.2004.830793

[20] R. Cilibrasi and P. M. B. Vitanyi, “Clustering by compression,” *Information Theory, IEEE Transactions on*, vol. 51, no. 4, pp. 1523–1545, Apr. 2005. [Online]. Available: http://dx.doi.org/10.1109/tit.2005.844059

[21] M. Nykter, N. D. Price, A. Larjo, T. Aho, S. A. Kauffman, O. Yli-Harja, and I. Shmulevich, “Critical networks exhibit maximal information diversity in

structure-dynamics relationships." *Physical review letters*, vol. 100, no. 5, Feb. 2008. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/18352443

[22] M. Folkard, K. M. Prise, G. Grime, K. Kirkby, and B. Vojnovic, "The use of microbeams to investigate radiation damage in living cells," *Appl Radiat Isot*, vol. 67, no. 3, pp. 436–439, 2010.

[23] J. Selinummi, P. Ruusuvuori, I. Podolsky, A. Ozinsky, E. Gold, O. Yli-Harja, A. Aderem, and I. Shmulevich, "Bright Field Microscopy as an Alternative to Whole Cell Fluorescence in Automated Analysis of Macrophage Images," *PLoS ONE*, vol. 4, no. 10, pp. e7497+, Oct. 2009. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0007497

[24] R. Ali, M. Gooding, T. Szilágyi, B. Vojnovic, M. Christlieb, and M. Brady, "Automatic segmentation of adherent biological cell boundaries and nuclei from brightfield microscopy images," *Machine Vision and Applications*, pp. 1–15, May 2011. [Online]. Available: http://dx.doi.org/10.1007/s00138-011-0337-9

[25] P. S. U. Adiga and B. B. Chaudhuri, "An efficient method based on watershed and rule-based merging for segmentation of 3-D histo-pathological images," *Pattern Recognition 34*, vol. 34, pp. 1449–1458, 2001.

[26] G. Lin, M. K. Chawla, K. Olson, J. F. Guzowski, C. a. Barnes, and B. Roysam, "Hierarchical, model-based merging of multiple fragments for improved three-dimensional segmentation of nuclei," *The journal of the Cytometry*, vol. 63, no. 1, pp. 20–33, Jan. 2005.

[27] Xiaowei Chen, Xiaobo Zhou and S. T. C.Wong*, "Automated Segmentation, Classi?cation, and Tracking of.pdf," *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, 2006.

[28] N. Harder, B. Neumann, M. Held, U. Liebel, H. Erfle, J. Ellenberg, R. Eils, and K. Rohr, "Automated recognition of mitotic patterns in fluorescence microscopy images

of human cells," *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro*, pp. 1016–1019, 2006.

[29] R. Ali, M. Gooding, M. Christlieb, M. Brady, and F. R. S. Feng, "Advanced phase-based segmentation of multiple cells from brightfield microscopy images," *5th IEEE International Symposium on Biomedical Imaging*, pp. 181–184, 2008.

[30] A. Genovesio, T. Liedl, V. Emiliani, W. J. Parak, M. Coppey-Moisan, and J.-C. Olivo-Marin, "Multiple particle tracking in 3-D+t microscopy: method and application to the tracking of endocytosed quantum dots." *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 15, no. 5, pp. 1062–70, May 2006.

[31] Xiaobo Zhou, Fuhai Li, Jun Yan and S. T. C. Wong, "A Novel Cell Segmentation Method and Cell Phase Identification Using Markov Model," *IEEE Trans Inf Technol Biomed*, vol. 13, no. 2, pp. 152–157, 2010.

[32] V. Kovalev, N. Harder, B. Neumann, M. Held, U. Liebel, H. Erfle, J. Ellenberg, R. Eils, and K. Rohr, "Feature Selection for Evaluating Fluorescence Microscopy Images in Genome-Wide Cell Screens," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 276–283, 2006.

[33] C. O. De Solorzano, R. Malladi, S. a. Lelièvre, and S. J. Lockett, "Segmentation of nuclei and cells using membrane related protein markers." *Journal of microscopy*, vol. 201, no. Pt 3, pp. 404–15, Mar. 2001.

[34] J. Selinummi, P. Ruusuvuori, I. Podolsky, A. Ozinsky, E. Gold, O. Yli-Harja, A. Aderem, and I. Shmulevich, "Bright field microscopy as an alternative to whole cell fluorescence in automated analysis of macrophage images," *PloS one*, vol. 4, no. 10, p. e7497, Jan. 2009.

[35] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. a. Guertin, J. H. Chang, R. a. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini,

"CellProfiler: image analysis software for identifying and quantifying cell phenotypes." *Genome biology*, vol. 7, no. 10, p. R100, Jan. 2006.

[36] M. J. Gooding, S. Kennedy, and J. A. Noble, "Volume segmentation and reconstruction from freehand three-dimensional ultrasound data with application to ovarian follicle measurement," *Ultrasound in medicine & biology*, vol. 34, no. 2, pp. 183–95, Feb. 2008.

[37] C. E. Shannon, "A mathematical theory of communication," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, Jan. 2001. [Online]. Available: http://dx.doi.org/10.1145/584091.584093

[38] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE transactions on systems*, vol. C, no. 1, pp. 62–66, 1979.

[39] Andrey N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems in Information Transmission*, vol. 1, pp. 1–7, 1965.

[40] D. J. Galas, M. Nykter, G. W. Carter, N. D. Price, I. Shmulevich, and S. Member, "Biological Information as Set-Based Complexity," vol. 56, no. 2, pp. 667–677, 2010.

[41] T. Mäki-Marttunen, J. Kesseli, S. Kauffman, O. Yli-Harja, and M. Nykter, "Of the complexity of Boolean network state trajectories," in *Proceedings of the Eighth International Workshop on Computational Systems Biology, WCSB*, 2011, pp. 6–8.

[42] N. A. Sakhanenko and D. J. Galas, "Complexity of networks I: The set-complexity of binary graphs," *Complexity*, vol. 17, no. 2, pp. 51–64, Nov. 2011. [Online]. Available: http://dx.doi.org/10.1002/cplx.20382

[43] M. Li, X. Li, X. Li, B. Ma, and P. Vitanyi, "The similarity metric," in *IEEE Transactions on Information Theory*, 2003, pp. 863–872. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.58.1259

[44] M. C. Cebrián, M. Alfonseca, and A. Ortega, "Common pitfalls using normalized compression distance: what to watch out for in a compressor," *Communications*

in Information and Systems, vol. 5, pp. 367–384, 2005. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.104.9265

[45] J. Mortensen, J. Wu, J. Furst, J. Rogers, and D. Raicu, "Effect of Image Linearization on Normalized Compression Distance," in *Signal Processing, Image Processing and Pattern Recognition*, ser. Communications in Computer and Information Science, D. Ślezak, S. Pal, B.-H. Kang, J. Gu, H. Kuroda, and T.-h. Kim, Eds. Springer Berlin Heidelberg, 2009, vol. 61, pp. 106–116. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-10546-3_14

[46] A. D. Lander, "Morpheus unbound: reimagining the morphogen gradient." *Cell*, vol. 128, no. 2, pp. 245–256, Jan. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2007.01.004

[47] ——, "Pattern, Growth, and Control," *Cell*, vol. 144, no. 6, pp. 955–969, Mar. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2011.03.009

[48] M. Mitchell, P. T. Hraber, and J. P. Crutchfield, "Revisiting the Edge of Chaos: Evolving Cellular Automata to Perform Computations," in *Complex Systems*, vol. 7, 1993, pp. 89–130. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.55.6034

[49] M. G. Kitzbichler, M. L. Smith, S. R. Christensen, and E. Bullmore, "Broadband criticality of human brain network synchronization." *PLoS computational biology*, vol. 5, no. 3, pp. e1 000 314+, Mar. 2009. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1000314

[50] S. A. Kauffman and S. Johnsen, "Coevolution to the edge of chaos: coupled fitness landscapes, poised states, and coevolutionary avalanches." *Journal of theoretical biology*, vol. 149, no. 4, pp. 467–505, Apr. 1991. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/2062105

[51] E. D. Schwab and K. J. Pienta, “Cancer as a complex adaptive system,” *Medical Hypothesis*, vol. 47, no. 3, pp. 235–241, Sep. 1995. [Online]. Available: http://www.medical-hypotheses.com/article/S0306-9877(96)90086-9/abstract

[52] P. Bak, C. Tang, and K. Wiesenfeld, “Self-organized criticality: An explanation of the 1/ f noise,” *Physical Review Letters*, vol. 59, no. 4, pp. 381–384, Jul. 1987. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.59.381

[53] C. J. Tomlin and J. D. Axelrod, “Biology by numbers: mathematical modelling in developmental biology,” *Nature Reviews Genetics*, vol. 8, no. 5, pp. 331–340, May 2007. [Online]. Available: http://dx.doi.org/10.1038/nrg2098

[54] D. J. Galas, M. Nykter, G. W. Carter, N. D. Price, and I. Shmulevich, “Biological information as set-based complexity,” *IEEE Transactions on Information Theory*, vol. 56, no. 2, pp. 667–677, Feb. 2010. [Online]. Available: http://dx.doi.org/10.1109/TIT.2009.2037046

[55] B. Derrida and Y. Pomeau, “Random networks of automata: A simple annealed approximation,” *EPL (Europhysics Letters)*, vol. 1, no. 2, pp. 45–49, Jan. 1986. [Online]. Available: http://dx.doi.org/10.1209/0295-5075/1/2/001

[56] R. Serra, M. Villani, C. Damiani, A. Graudenzi, and A. Colacci, “The Diffusion of Perturbations in a Model of Coupled Random Boolean Networks,” in *Cellular Automata*, ser. Lecture Notes in Computer Science, H. Umeo, S. Morishita, K. Nishinari, T. Komatsuzaki, and S. Bandini, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, vol. 5191, ch. 40, pp. 315–322. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-79992-4_40

[57] A. Mazumdar and M. Mazumdar, “How one becomes many: blastoderm cellularization in Drosophila melanogaster.” *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 24, no. 11, pp. 1012–1022, Nov. 2002. [Online]. Available: http://dx.doi.org/10.1002/bies.10184

[58] J. F. Knabe, C. L. Nehaniv, and M. J. Schilstra, "Evolution and morphogenesis of differentiated multicellular organisms: autonomously generated diffusion gradients for positional information," in *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems.* MIT Press, 2008, pp. 321–328. [Online]. Available: http://www.alifexi.org/papers/ALIFExi_pp321-328.pdf

[59] R. Thomas, "Regulatory networks seen as asynchronous automata: A logical description," *Journal of Theoretical Biology*, vol. 153, no. 1, pp. 1–23, Nov. 1991. [Online]. Available: http://dx.doi.org/10.1016/S0022-5193(05)80350-9

[60] M. Chaves, R. Albert, and E. D. Sontag, "Robustness and fragility of Boolean models for genetic regulatory networks," *Journal of Theoretical Biology*, vol. 235, no. 3, pp. 431–449, Aug. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.jtbi.2005.01.023

[61] I. Shmulevich and E. Dougherty, *Probabilistic Boolean Networks*, 1st ed. SIAM Press, 2010.

[62] I. Shmulevich, S. A. Kauffman, and M. Aldana, "Eukaryotic cells are dynamically ordered or critical but not chaotic," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13 439–13 444, Sep. 2005. [Online]. Available: http://dx.doi.org/10.1073/pnas.0506771102

[63] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein, "Random Boolean network models and the yeast transcriptional network," *Proceedings of the National Academy of Sciences*, vol. 100, no. 25, pp. 14 796–14 799, Dec. 2003. [Online]. Available: http://dx.doi.org/10.1073/pnas.2036429100

[64] J. W. Bodnar, "Programming the *Drosophila* embryo," *Journal of Theoretical Biology*, vol. 188, no. 4, pp. 391–445, Oct. 1997. [Online]. Available: http://dx.doi.org/10.1006/jtbi.1996.0328

[65] C. Gershenson, "Introduction to Random Boolean Networks," in *Workshop and Tutorial Proceedings, Ninth International Conference on the Simulation and*

Synthesis of Living Systems (ALife IX), M. Bedau, P. Husbands, T. Hutton, S. Kumar, and H. Suzuki, Eds., Aug. 2004, pp. 160–173. [Online]. Available: http://arxiv.org/abs/nlin/0408006

[66] S. Huang, I. Ernberg, and S. Kauffman, "Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective." *Seminars in cell & developmental biology*, vol. 20, no. 7, pp. 869–876, Sep. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.semcdb.2009.07.003

[67] A. N. Kolmogorov, "Three approaches to the quantitative definition of information." *Problems in Information Transmission*, vol. 1, pp. 1–7, 1965.

[68] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," Systems Research Center, Tech. Rep., 1994. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.6177

[69] E. Balleza, E. R. Alvarez-Buylla, A. Chaos, S. Kauffman, I. Shmulevich, and M. Aldana, "Critical Dynamics in Genetic Regulatory Networks: Examples from Four Kingdoms," *PLoS ONE*, vol. 3, no. 6, pp. e2456+, Jun. 2008. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0002456

[70] M. Nykter, N. D. Price, M. Aldana, S. A. Ramsey, S. A. Kauffman, L. E. Hood, O. Yli-Harja, and I. Shmulevich, "Gene expression dynamics in the macrophage exhibit criticality," *Proceedings of the National Academy of Sciences*, vol. 105, no. 6, pp. 1897–1900, Feb. 2008. [Online]. Available: http://dx.doi.org/10.1073/pnas.0711525105

[71] T. Mora and W. Bialek, "Are biological systems poised at criticality?" *Journal of Statistical Physics*, vol. 144, no. 2, pp. 268–302, Dec. 2010. [Online]. Available: http://dx.doi.org/10.1007/s10955-011-0229-4

[72] E. D. Schwab and K. J. Pienta, "Cancer as a complex adaptive system," *Medical Hypotheses*, vol. 47, no. 3, pp. 235–241, Sep. 1996. [Online]. Available: http://dx.doi.org/10.1016/s0306-9877(96)90086-9

[73] Z. Wang and J. Zhang, "In search of the biological significance of modular structures in protein networks." *PLoS computational biology*, vol. 3, no. 6, pp. e107+, Jun. 2007. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.0030107

[74] A. Hintze and C. Adami, "Evolution of Complex Modular Biological Networks," *PLoS Comput Biol*, vol. 4, no. 2, pp. e23+, Feb. 2008. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.0040023

[75] J. Clune, J.-B. Mouret, and H. Lipson, "The evolutionary origins of modularity," *Proceedings of the Royal Society B: Biological Sciences*, vol. 280, no. 1755, p. 20122863, Mar. 2013. [Online]. Available: http://dx.doi.org/10.1098/rspb.2012.2863

[76] K. H. Ten Tusscher and P. Hogeweg, "Evolution of networks for body plan patterning; interplay of modularity, robustness and evolvability." *PLoS computational biology*, vol. 7, no. 10, Oct. 2011. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/21998573

[77] D. Papatsenko, "Stripe formation in the early fly embryo: principles, models, and networks." *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 31, no. 11, pp. 1172–1180, Nov. 2009. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/19795410

[78] M. S. Kim, J. R. Kim, D. Kim, A. Lander, and K. H. Cho, "Spatiotemporal network motif reveals the biological traits of developmental gene regulatory networks in Drosophila melanogaster," *BMC Systems Biology*, vol. 6, no. 1, pp. 31+, 2012. [Online]. Available: http://dx.doi.org/10.1186/1752-0509-6-31

[79] D. M. Lorenz, A. Jeng, and M. W. Deem, "The Emergence of Modularity in Biological Systems," Apr. 2012. [Online]. Available: http://arxiv.org/abs/1204.5999

[80] U. Alon, "Network motifs: theory and experimental approaches," *Nat Rev Genet*, vol. 8, no. 6, pp. 450–461, Jun. 2007. [Online]. Available: http://dx.doi.org/10.1038/nrg2102

[81] M. Levine and E. H. Davidson, “Gene regulatory networks for development,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 14, pp. 4936–4942, Apr. 2005. [Online]. Available: http://dx.doi.org/10.1073/pnas.0408031102

[82] G. Swiers, R. Patient, and M. Loose, “Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification,” *Developmental Biology*, vol. 294, no. 2, pp. 525–540, Jun. 2006. [Online]. Available: http://dx.doi.org/10.1016/j.ydbio.2006.02.051

[83] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, “Network motifs in the transcriptional regulation network of Escherichia coli,” *Nature Genetics*, vol. 31, no. 1, pp. 64–68, Apr. 2002. [Online]. Available: http://dx.doi.org/10.1038/ng881

[84] H.-W. Ma, B. Kumar, U. Ditges, F. Gunzer, J. Buer, and A.-P. Zeng, “An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs,” *Nucleic Acids Research*, vol. 32, no. 22, pp. 6643–6649, Jan. 2004. [Online]. Available: http://dx.doi.org/10.1093/nar/gkh1009

[85] M. E. J. Newman, “Finding community structure in networks using the eigenvectors of matrices. Phys,” in *Rev. E.* [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.261.8782

APPENDICES

BioSystems 112 (2013) 131–138

Contents lists available at SciVerse ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Kolmogorov complexity of epithelial pattern formation: The role of regulatory network configuration

Nicholas S. Flann[a,b,*], Hamid Mohamadlou[a], Gregory J. Podgorski[c,d]

[a] *Department of Computer Science, Utah State University, United States*
[b] *Institute for Systems Biology, Seattle, United States*
[c] *Department of Biology, Utah State University, United States*
[d] *Center for Integrated BioSystems, Utah State University, United States*

ARTICLE INFO

Keywords:
Kolmogorov complexity
Information theory
Pattern formation
Gene regulatory networks
Normalized compression distance
Boolean networks

ABSTRACT

The tissues of multicellular organisms are made of differentiated cells arranged in organized patterns. This organization emerges during development from the coupling of dynamic intra- and intercellular regulatory networks. This work applies the methods of information theory to understand how regulatory network structure both within and between cells relates to the complexity of spatial patterns that emerge as a consequence of network operation. A computational study was performed in which undifferentiated cells were arranged in a two dimensional lattice, with gene expression in each cell regulated by identical intracellular randomly generated Boolean networks. Cell–cell contact signalling between embryonic cells is modeled as coupling among intracellular networks so that gene expression in one cell can influence the expression of genes in adjacent cells. In this system, the initially identical cells differentiate and form patterns of different cell types. The complexity of network structure, temporal dynamics and spatial organization is quantified through the Kolmogorov-based measures of normalized compression distance and set complexity. Results over sets of random networks that operate in the ordered, critical and chaotic domains demonstrate that: (1) ordered and critical networks tend to create the most information-rich patterns; (2) signalling configurations in which cell-to-cell communication is non-directional mostly produce simple patterns irrespective of the internal network domain; and (3) directional signalling configurations, similar to those that function in planar cell polarity, produce the most complex patterns, but only when the intracellular networks function in non-chaotic domains.

© 2013 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Multicellular organisms exhibit an incredible variety of cellular patterns, for instance, those in the *Drosophila* embryo illustrated in Fig. 1. These patterns arise during development and are a consequence of genetic regulatory networks (GRNs) that operate within cells and that respond to communication between cells (Lander, 2007, 2011). One interesting question to explore is the relationship between the structure of GRNs and the complexity of cellular patterns that can emerge from the operation of these networks. A related question is how GRNs and their evolution contributed to the transition from unicellularity to multicellularity. Although details are not known about the evolution of multicellularity in any lineage, this process almost certainly involved the co-option of GRNs and intercellular communication systems that existed in single-celled organisms (Knoll, 2011). While the actual paths of evolution to complex multicellularity may never be known, potential paths open to evolution can be explored and understood through computational studies. This is a long term goal of the investigations reported here.

Evidence suggests that living processes lie “on the edge of chaos,” and that biological selection operates to maximally retain information yet allow evolution (Mitchell et al., 1993; Kitzbichler et al., 2009; Kauffman and Johnsen, 1991). Dynamic systems, including biological systems, operate in three complexity domains: ordered, critical and chaotic. Ordered systems are robust in that they dampen perturbations to retain information, but at the cost of limited potential for change. Chaotic systems magnify perturbations and lose information, rendering them unsuitable for homeostatic living systems. In fact, chaotic systems are implicated in diseases like cancer (Schwab and Pienta, 1995). Critical systems, which operate on the cusp between order and chaos, are the most information dense in both network organization and dynamics (Bak et al., 1987). This work focuses on how the information content of

---

* Corresponding author at: Department of Computer Science, Utah State University, United States.
E-mail addresses: nick.flann@usu.edu, nick.flann@gmail.com (N.S. Flann), hamidmohamadlou@yahoo.com (H. Mohamadlou), gregory.podgorski@usu.edu (G.J. Podgorski).

0303-2647/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.biosystems.2013.03.005

Mohamadlou *et al. BMC Bioinformatics* 2014, **15**:32
http://www.biomedcentral.com/1471-2105/15/32

BMC
Bioinformatics

SOFTWARE Open Access

Maximizing Kolmogorov Complexity for accurate and robust bright field cell segmentation

Hamid Mohamadlou[1], Joseph C Shope[4] and Nicholas S Flann[1,2,3*]

Abstract

Background: Analysis of cellular processes with microscopic bright field defocused imaging has the advantage of low phototoxicity and minimal sample preparation. However bright field images lack the contrast and nuclei reporting available with florescent approaches and therefore present a challenge to methods that segment and track the live cells. Moreover, such methods must be robust to systemic and random noise, variability in experimental configuration, and the multiple unknowns in the biological system under study.

Results: A new method called *maximal-information* is introduced that applies a non-parametric information theoretic approach to segment bright field defocused images. The method utilizes a combinatorial optimization strategy to select specific defocused images from each image stack such that set complexity, a Kolmogorov complexity measure, is maximized. Differences among these selected images are then applied to initialize and guide a level set based segmentation algorithm. The performance of the method is compared with a recent approach that uses a fixed defocused image selection strategy over an image data set of embryonic kidney cells (HEK 293T) from multiple experiments. Results demonstrate that the adaptive *maximal-information* approach significantly improves precision and recall of segmentation over the diversity of data sets.

Conclusions: Integrating combinatorial optimization with non-parametric Kolmogorov complexity has been shown to be effective in extracting information from microscopic bright field defocused images. The approach is application independent and has the potential to be effective in processing a diversity of noisy and redundant high throughput biological data.

Background

Cell segmentation is the identification of cell objects and their observable properties from biological images. Current cell segmentation methods perform most accurately when applied to high contrast and minimal noise images obtained from samples where the cells have fluorescently-labeled cell nuclei and stained membranes, and are distinct with minimal adherent membranes. However, these ideal conditions rarely exist.

Fluorescently tagging cells using green fluorescent protein (GFP) leads to robust identification of each cell during segmentation. While GFP tagging is widespread, there are disadvantages when applying the method repeatedly to the same sample since under repeated application of high-energy light the cells can suffer phototoxicity. Such light can disrupt the cell behavior through stress, shorten life and potentially confound the experimental results [1-3]. Significantly, a requirement for GFP labeling adds a step before a new cell line can be studied, thus making it difficult to apply this method in a clinical setting.

The alternative is to use bright field microscopy, the original and the simplest microscopy technique, wherein cells are illuminated with white light from below. However, using only bright field imaging of unstained cells presents a challenging cell detection problem because of lack of contrast and difficulty in locating both cell centers and borders, particularly when cells are tightly

---

*Correspondence: Nick.Flann@usu.edu
[1] Department of Computer Science, Utah State University, Logan, UT 84322, USA
[2] Institute for Systems Biology, Seattle, WA 98109, USA
Full list of author information is available at the end of the article

BioMed Central

© 2014 Mohamadlou et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The role of network motifs in epithelial pattern formation: A Kolmogorov complexity study

Hamid Mohamadlou, Gregory Podgorski, and Nicholas Flann [⋆]

Department of Computer Science,
Department of biology
{nicholas.flann,gregory.podgorski@usu.edu}
{hamid.mohamadlou@aggiemail.usu.edu}

Abstract. Genetic regulatory network consists of quasi-autonomous subnetworks referred to as modules. Such modular networks determine the organized patterns in multicellular organisms during development. However, the role of modularity in this process is poorly understood. This study applies methods of information theory to explore how network modularity influences the complexity of multicellular patterns that emerge from the dynamics of the regulatory networks. A computational study was performed by creating Boolean intracellular networks of varying modularity within a simulated epithelium field of embryonic cells. Each cell contains the same network and communicates with adjacent cells using contact-mediated signaling. The study explored two types of modules: motifs, which are subnetworks with unique connectivity and regulatory functions, and clusters, which are densely connected sets of genes sparsely connected to other genes. Results comparing random networks to those with cluster and motif modularity demonstrate that: (1) Networks with modular clusters tend to produce higher information-dense multicellular patterns without a significant increase in the gene expression dynamics. (2) Network motifs with feedback loops increase information complexity of the multicellular patterns while simplifying the network dynamics. (3) Positive feedback motifs don't effect the dynamics complexity as significantly as positive feedback loops do.

Keywords: Network motifs, Kolomogrov complexity, Pattern formation

1 Introduction

Understanding the process by which the complex variety of cellular patterns form during development of multicellular organisms is a significant challenge in biology (Fig.1). While many challenges remain, it is known that these multicellular patterns emerge as a result of genetic regulatory networks (GRNs) that operate within cells [1]. GRN's represent the interactions among genes where combinations of genes control the expression of other genes, forming feedback loops. The gene expression profile for each cell is then determined by signaling within and among other cells, differentiation thus making the body plan and subsequence morphology[2]. To capture the behavior of this regulatory system, scientists have developed mathematical and computational models for gene regulatory networks with the purpose of generating predictions to explain experimental observations. Among these modeling alternatives is a simplified modeling technique called Boolean networks that is the approach employed in this study [3].

It is hypothesized that biological systems operate to maximally retain information across evolutionary time by which they fall into three complexity domains: ordered, critical and chaotic [4] [5] [6]. In the Order systems some events happens more frequently

---

[⋆] Supported by Luxembourg Centre for Systems biomedicine and the University of Luxembourg.

CURRICULUM VITAE

Hamid Mohamadlou

EDUCATION

Ph.D., Computer Science. Utah State University, Logan, UT. 2015.

M.S., Systems Engineering. Tehran University, Tehran, Iran. 2009.

B.S., Mathematics. Zanjan University, Zanjan, Iran. 2007.

RESEARCH INTERESTS

Algorithms, Data mining, Machine learning, Computational biology.

PUBLICATIONS

Mohamadlou, H., Flann, N. (2015). The role of network motifs in epithelial pattern formation: A Kolmogorov complexity study. 10th International Conference on Information Processing in Cells and Tissues IPCAT 2015, San Diego, USA.

Mohamadlou, H., Flann, N. (2014). Maximizing Kolmogorov complexity for accurate and robust bright field cell segmentation. BMC Bioinformatics Journal.

Flann, N., Mohamadlou, H. (2013). Kolmogorov Complexity of Epithelial Pattern Formation: the role of Regulatory Network Configuration. BioSystems Journal.

Flann, N., Mohamadlou, H. (2012). Criticality of Spatiotemporal Dynamics in Contact Mediated Pattern Formation. Theoretical Computer Science and General Issues, Springer.

Arefan, D., Mohamadlou, H. (2012). Automated Abnormal Mass Detection in the Mammogram Images Using Chebyshev Moments. Research Journal of Applied Sciences, Engineering and Technology.

Azadeh, A., Mohamadlou, H. (2010). Modeling road traffic accident reporting system by discrete event simulation: A case study of Irans road. 8th International Conference of Modeling and Simulation, MOSIM, Hemmamet.

Mohamadlou, H. (2009). A method for mining association rules in quantitative and fuzzy data. Computer and industrial engineering conference, France.

TECHNICAL SKILLS

Expert level in Matlab and prototype programming (over 10 years of experience).

Coding skill in C++, Java, and Python.

Comfortable working in UNIX environment.

Familiar with technologies such as Map Reduce, Apache Spark.

Proficient with relational databases and SQL, HTML, CSS.

PROJECTS AND EXPERIENCE

Research Assistant. Flanns lab, Utah State University. 2010-2015.

- Published several technical articles in Algorithmic Information Theory, Image processing and data mining in adjudicated journals.
- Developed an accurate segmentation algorithm for bright field cell microscopy images.
- Developed an Algorithmic Information Theory based method to detect information-dense patterns that result from running coupled Gene Regulatory Network in an in silco epithelial field.
- Studied the role of gene network motifs in multicellular pattern formation.

Teaching Assistant Advanced Algorithms. Utah State University. 2014-2015.

Teaching Assistant C++. Utah State University. 2012-2014.

- Developed an auto-grader script in UNIX shell to grade C++ course assignments.

Teaching Assistant HTML. Utah State University. 2010-2011.

Web Programmer. Utah State University. 2013-2014.

Research Assistant. Tehran University. 2007-2009.

- Developed an algorithm for effective Association Rule mining using fuzzy clustering techniques.
- Developed a Monte Carlo simulation for business process. Case study of Irans road accident reporting system.

Network Operator and Controller engineer. Huawei Technology, Tehran, Iran. 2009-2010.

MATHEMATICS, STATISTICS AND ALGORITHMS SKILLS

Technically proficient in mathematics, statistics and computer science.

Translate unstructured problems into abstract mathematical frameworks.

Deliver actable and predictive models to answer business questions and improve strategy delivery.

Experience in data mining approaches such as association rules, clustering, classification, regression, Principle Component Analysis.

Able to develop algorithms to solve quantitative and qualitative problems.

Expert in advanced algorithms such as text mining, graph e.g., Dijkstra and network flow, data compression.

Familiarity with big data, scalable machine learning and data mining techniques.

Apply optimization techniques such as simplex programming, genetic algorithm and simulating annealing.

Conduct research and analysis over large volume of data.

Nurture idea from prototype to launch and write production quality code while implementing new ideas.

Publish data driven, scientific and technical articles.

Curious, out of box thinker and interdisciplinary data scientist.