Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

NJIT New Jersey's Science & Technology University

The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

A DYNAMICAL MODEL OF THE DISTRIBUTED INTERACTION OF INTRACELLULAR SIGNALS

by
Adrienne C. N. James

A major goal of modern cell biology is to understand the regulation of cell behavior in the reductive terms of all the molecular interactions. This aim is made explicit by the assertion that understanding a cell's response to stimuli requires a full inventory of details. Currently, no satisfactory explanation exists to explain why cells exhibit only a relatively small number of different behavioral modes.

In this thesis, a discrete dynamical model is developed to study interactions between certain types of signaling proteins. The model is generic and "connectionist" in nature and incorporates important concepts from the biology. The emphasis is on examining dynamic properties that occur on short term time scales and are independent of gene expression. A number of modeling assumptions are made. However, the framework is flexible enough to be extended in future studies.

The dynamical states of the system are explored both computationally and analytically. Monte Carlo methods are used to study the state space of simulated networks over selected parameter regimes. Networks show a tendency to settle into fixed points or oscillations over a wide range of initial conditions. A genetic algorithm (GA) is also designed to explore properties of networks. It evolves a "population" of modeled cells, selecting and ranking them according to a fitness function which is designed to mimic features of real biological evolution. An analogue of protein domain shuffling is used as the crossover operator and cells are reproduced asexually. The effects of changing the parameters of the GA are explored. A clustering algorithm is developed to test the effectiveness of the GA search at generating cells which display a limited number of different behavioral modes. Stability properties of equilibrium

states in small networks are analyzed. The ability to generalize these techniques to larger networks is discussed. Topological properties of networks generated by the GA are examined. Structural properties of networks are used to provide insight into their dynamic properties.

The dynamic attractors exhibited by such signaling networks may provide a framework for understanding why cells persist in only a small number of stable behavioral modes.

A DYNAMICAL MODEL OF THE DISTRIBUTED INTERACTION OF INTRACELLULAR SIGNALS

by
Adrienne C. N. James

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Mathematical Sciences

Department of Mathematical Sciences
Department of Mathematics and Computer Science, Rutgers-Newark

May 2001

Copyright © 2001 by Adrienne C. N. James
ALL RIGHTS RESERVED

APPROVAL PAGE

A DYNAMICAL MODEL OF THE DISTRIBUTED INTERACTION OF INTRACELLULAR SIGNALS

Adrienne C. N. James

Dr. Michael Recce, Dissertation Advisor — Date
Associate Professor of Computer and Information Science, and Mathematical Sciences, and Biology; Director, Center for Computational Biology and Bioengineering, NJIT

Dr. Denis Blackmore, Committee Member — Date
Professor of Mathematics, NJIT

Dr. Victoria Booth, Committee Member — Date
Assistant Professor of Mathematics, NJIT

Dr. James A. M. McHugh, Committee Member — Date
Professor of Computer and Information Science, Associate Chairperson of the Department, NJIT

Dr. Kay Swann, Committee Member — Date
Reader, Department of Anatomy and Developmental Biology, University College London, London, UK

BIOGRAPHICAL SKETCH

Author: Adrienne C. N. James

Degree: Doctor of Philosophy

Date: May 2001

Undergraduate and Graduate Education:

- Doctor of Philosophy in Applied Mathematics,
  New Jersey Institute of Technology, Newark, NJ, 2001
- Master of Arts in Mathematics,
  University of Cambridge, Cambridge, U.K., 1996
- Master of Science in Information Technology,
  University College London, London, U.K., 1994
- Bachelor of Arts in Mathematics,
  University of Cambridge, Cambridge, U.K., 1993

Major: Applied Mathematics

Presentations and Publications:

Adrienne James,
"Cell Behaviour as a Dynamic Attractor in the Intracellular Signalling System," *Invited Lecture*, Center for BioDynamics, Boston University, 2000.

Adrienne James,
"Cell Behaviour as a Dynamic Attractor in the Intracellular Signalling System," *Invited Lecture*, Institute of Biophysics and Biomedicine, University of Helsinki, (Helsinki, Finland), 2000.

A. James, K. Swann and M. Recce,
"Cell Behaviour as a Dynamic Attractor in the Intracellular Signalling System," *Journal of Theoretical Biology*, vol. 196, pp 269-288, 1999.

Adrienne James,
"Cell Behaviour as a Dynamic Attractor in the Intracellular Signalling System," *Invited Lecture as part of a Workshop on Cytokines and Cell Signalling*, Centre for Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, (London, U.K.), 1998.

Adrienne James,

"Cell Behaviour as a Dynamic Attractor in the Intracellular Signalling System," *Invited Lecture*, Centre for Non-Linear Dynamics and Its Applications, University College London, (London, U.K.), 1996

A. James, K. Swann and M. Recce,

"A dynamical model of the distributed interaction of intracellular signals," *International Journal of Neural Systems*, vol. 7, no. 4, pp 333-341, 1996.

To my parents, without whom this would not have been possible, and who have always believed in me and reminded me that hard work and perseverance will pay off in the end

To my mum:

LEISURE

What is this life if, full of care,
We have no time to stand and stare.

No time to stand beneath the boughs
And stare as long as sheep or cows.

No time to see, when woods we pass,
Where squirrels hide their nuts in grass.

No time to see, in broad daylight,
Streams full of stars, like skies at night.

No time to turn at Beauty's glance,
And watch her feet, how they can dance.

No time to wait till her mouth can
Enrich that smile her eyes began.

A poor life this, if full of care,
We have no time to stand and stare.

W.H. Davies (1871-1940)

To my dad:

YOU'LL NEVER WALK ALONE

When you walk through a storm
Hold your head up high, and don't be afraid of the dark.
At the end of the storm is a golden sky
And the sweet silver song of a lark.

Walk on through the wind, walk on through the rain,
Though your dreams be tossed and blown.
Walk on, walk on with hope in your heart
And you'll never walk alone.

The Liverpool Football Song.

ACKNOWLEDGMENT

There are many people whose contribution to this Ph.D. dissertation deserve acknowledgement. First and foremost, I would like to thank my advisor, Dr. Michael Recce, for his patience, support and encouragement throughout my time as a graduate student. I would also like to thank him for giving me the opportunity to come to the United States of America to finish my doctoral studies, continuing with him as my Ph.D. thesis advisor.

My heartfelt thanks to my parents, Thelma and David James, for their financial support, their understanding of my need to pursue further academic study and also for their continuing encouragement. I would also like to extend my thanks to the Pasquale family who have made the world of difference to my stay here in New Jersey. Thanks also to Steve Kunec, Hoa Tran, Lyudmyla Barannyk, Said Kas-Danouche and Dr. Shailesh Naire for their help in mathematical discussions and also for their friendship.

I am grateful to my former colleagues from the United Kingdom—Hajime Hirase, David Jack, Giorgio Grasso, Ken Harris and John Taylor—for innumerable discussions when I started my degree at University College London. I also wish to extend special thanks to Dr. Karl Swann for the original idea that sparked this study and for his help in continuing this as my Ph.D. project. This work was initially supported with the aid of a grant from the Medical Research Council (Studentship Reference Number G78/4557).

Thanks to Professor Bechtold and Professor Garcia-Reimbert who formed an important part of my decision to transfer to the Applied Mathematics Department at NJIT. I would also like to thank the department for providing financial support for this research and to the faculty, Irene Giouvanos (the system administrator) and administrative staff who have helped to make my stay here a memorable one. Finally, I wish to thank my remaining committee members for taking time out of their busy

schedules to provide help and constructive criticism. This work has been supported in its latter stages by the National Science Foundation (Grant Reference Number DMS-9973230).

Finally, I would like to express my gratitude to all the people who have proofread this document for errors in its various stages.

TABLE OF CONTENTS

Chapter		Page
1	INTRODUCTION	1
	1.1 A Look Ahead	4
	1.2 Progress in Cell Biology	5
	1.3 Cell Signaling Pathways as Linear Cascades	8
	1.4 Cross-talk Between Linear Pathways: A Signaling Network	19
	1.5 Further Levels of Complexity	22
	1.6 Proteins as the Computational Elements of Cells	27
	1.7 Desirable Properties of Cellular Networks	31
	1.8 Discrete Dynamical System Models	33
	1.8.1 Cellular Automata	36
	1.8.2 Random Boolean Networks	37
	1.8.3 "Connectionist" Protein Signaling Networks	40
	1.9 Cellular Observables	43
	1.10 Research Purpose and Design	45
	1.11 Thesis Outline	49
2	DEVELOPING A CONNECTIONIST MODEL	50
	2.1 Basic Assumptions	50
	2.2 Model Development	51
	2.2.1 Simulating the Dynamics of a Single Cell Type	55
	2.2.2 Derivation of Dynamical Update Rule	56
	2.3 Summary	61
3	MONTE CARLO SIMULATIONS	63
	3.1 Introduction	63
	3.2 Method	64
	3.3 Results	65

TABLE OF CONTENTS
(Continued)

Chapter		Page
	3.4 ATP Consumption in Modeled Networks	72
	3.5 Summary	73
4	A GENETIC ALGORITHM	75
	4.1 Introduction	75
	4.2 GA Development	76
	4.2.1 The Construction of a Fitness Function	77
	4.2.2 The Crossover and Mutation Operators	80
	4.3 Comparison of Genetic Search and Random Search	83
	4.4 A Parameter Study	84
	4.5 Summary	91
5	DETECTING DYNAMIC ATTRACTORS	93
	5.1 Introduction	93
	5.2 A Clustering Technique	93
	5.3 Performance of the Clustering Algorithm	98
	5.4 Numerical Evidence for Dynamic Attractors	104
	5.5 Summary	113
6	SMALL NETWORK ANALYSIS	114
	6.1 Introduction	114
	6.2 Examples of Fixed Point Stability Analysis.	116
	6.2.1 A Network with Two Protein Types: 2PKs	117
	6.2.2 A Network with Four Protein Types: 2PK, 2PP	120
	6.2.3 Analysis of the Unstable Fixed Point for the 2PK-2PP Network: A "Tanh" Approximation	124
	6.3 Stability Analysis of the Limit Cycle of the 2PK-2PP network	127
	6.3.1 Qualitative Explanation for the Periodic Orbit	127
	6.3.2 Proof of Local Asymptotic Stability of the Orbit	128

TABLE OF CONTENTS
(Continued)

Chapter **Page**

6.3.3 Parameter Variation 139

6.4 Comparison of Truncated and Double Precision Numerics 144

6.5 Summary 144

7 TOPOLOGICAL PROPERTIES OF NETWORKS 147

7.1 Introduction 147

7.2 Amplification Properties of Cascades 147

7.3 The Effect of Structural Properties on Oscillations 149

7.4 Structural Features of a Network Generated by the GA 154

7.5 Summary 155

8 DISCUSSION 157

8.1 Relaxing Model Assumptions 161

8.2 Genetic Algorithm Improvement 163

8.3 Small Network Analysis 164

8.4 Graph Theory Analysis 165

8.5 A Model for Long-Term Potentiation 167

8.6 Modeling Complex Networks 169

8.7 Concluding Remarks 173

APPENDIX A CELL SIGNALING SITES & PROTEIN DATABASES . . . 174

APPENDIX B A MODELED CASCADE SYSTEM 175

APPENDIX C SMALL NETWORK LOCAL STABILITY ANALYSIS 178

APPENDIX D GRAPH THEORY DEFINITIONS 179

APPENDIX E DYNAMICAL SYSTEMS TERMINOLOGY 181

BIBLIOGRAPHY 183

LIST OF TABLES

Table		Page
1.1	Major differences between neural networks and cell signaling networks	28
1.2	Similarities between neural networks and cell signaling networks	30
1.3	Cellular properties in the language of dynamical systems	32
1.4	Qualitative properties of cellular automaton models	37
5.1	Data from the analysis of the clustering algorithm	101
6.1	The rule for updating each node depends on the sign of the input $b_j^{(t)}$, $j = 1, 2$	121
6.2	Fixed points for the 2PK-2PP network when one of the input parameters takes either its maximum or minimum values	124
6.3	A period 14 limit cycle. Comparison between different schemes: (a) Numerics using f, (b) Numerics using $\hat{f}$, (c) Numerics using f_{tanh} with $\kappa = 100$	132
6.4	Eigenvalues of the iterated map f^{14}	140
6.5	Comparison of truncated and double precision numerical results for the 2PK-2PP network with $(\mu_1, \mu_2) = (5000, 5000)$	143
C.1	Summary of fixed points for some simple networks using the rule, f, and where inputs to each node always retain the same sign	178

LIST OF FIGURES

Figure		Page
1.1	Cells typically exhibit only a small, finite number of different behavioral modes	3
1.2	Phosphorylation (via protein kinases) and dephosphorylation (via protein phosphatases) reactions provide a dynamic and reversible mechanism for the modification of proteins	9
1.3	Phosphorylation of proteins involves the transfer of a phosphate group from an ATP molecule to a hydroxl (-OH) group on a Ser/Tyr/Thr residue of an amino acid	10
1.4	The intracellular signaling pathway thought to control glycogenolysis	12
1.5	Typical events in a cell signaling pathway	12
1.6	The MAP-kinase signaling cascade	13
1.7	A mono-cyclic cascade system	15
1.8	Representing protein activity in phase space	17
1.9	Interactions between signaling pathways can be extensive	19
1.10	The growth in protein phosphorylation research	26
1.11	Simplified diagram of the Chiva & Tarroux model	41
2.1	Schematic diagram of regulatory and catalytic domain structure and phosphorylation	51
2.2	Schematic diagram of interactions between signaling proteins	52
2.3	A single node in the network represents a pool of molecules	54
2.4	The updating function $f(o^{(t)}, b^{(t)})$	60
3.1	The Monte Carlo simulation method	64
3.2	Sample dynamics from a single cell simulation from two different starting states	66
3.3	Protein activities and range of fluctuations are shown for typical networks with $p_h = p_k =$ (A,B) 0.01, (C,D) 0.05	67
3.4	Protein activities and range of fluctuations are shown for a typical network with $p_h = p_k = 0.10$	68

LIST OF FIGURES
(Continued)

Figure **Page**

3.5 Average properties of Type II networks as $p_h = p_k$ is varied 70

3.6 A frequency plot of number of proteins against the difference in their maximum and minimum activity values over a series of Monte Carlo simulations 71

3.7 Estimate of ATP consumption for sample oscillating networks with $p_h = p_k$ = (A) 0.04, (B) 0.06, (C) 0.10 72

4.1 Genetic algorithm construction 80

4.2 Crossover in cells during the genetic algorithm 81

4.3 Mutation in cells during the genetic algorithm 82

4.4 Monte Carlo simulations on 10,000 randomly generated cell types 83

4.5 The effect of varying the mutation and crossover rates on the performance of the GA 85

4.6 The effect of varying the number of offspring per parent on the progression of the GA 89

5.1 Examples of potential difficulties with classifying clusters within data . . 94

5.2 One hundred random starting activity states for the fittest GA network. 99

5.3 Activity states for the same cell after $T = 500$ time steps 99

5.4 The clustering error decreases as the algorithm progresses 100

5.5 Distance matrix comparison of end states, D_{ij}, $1 \leq i, j, \leq S$, prior to clustering 103

5.6 Distance matrix after clustering 103

5.7 A cell settles into the same stable steady state configuration from two different starting states 105

5.8 The same cell (as in Figure 5.7) settles into a similar characteristic stable oscillation from two different random starting states 106

5.9 A switch between attracting states can be induced by an internal change in the network. 108

5.10 A cell can have many different oscillatory states 109

5.11 Phase space portrait of proteins in a network selected by the GA 111

LIST OF FIGURES
(Continued)

Figure **Page**

5.12 Plot of ATP consumption against time for the cases shown in Figure 5.11 112

6.1 A two node graph 117

6.2 A small network of four proteins 120

6.3 Limit cycle oscillation in the 2PK-2PP network for two selected pairs of input parameters: (μ_1, μ_2) = (A) (5000, 5000), (B) (2500, 7500) 122

6.4 Analyzing the iterative map in each of the four quadrants of the (o_1, o_2) phase plane 123

6.5 Surface plot of the updating rule with the tanh function (f_{tanh}) 125

6.6 The limit cycle has four main phases, corresponding to the four quadrants of Figure 6.4 127

6.7 Pre-images of the discontinuity (critical) lines $o_1 = \mu_1$, $o_2 = \mu_2$ 131

6.8 Mapping the lines LI-LIV 133

6.9 Evolution of a small neighborhood close to the limit cycle 135

6.10 An example of how parameter adjustment leads to a low period orbit which appears to be strongly attracting. 141

6.11 Variation of periodicity of the limit cycle for the rule $\hat{f}$ as the parameters μ_1 and μ_2 are varied 142

6.12 Scatter plot comparing data from truncated & double precision numerics 145

7.1 Positive amplification of a kinase cascade 148

7.2 Oscillating nodes in an acyclic graph 150

7.3 Oscillating nodes need not necessarily lie on a cycle 151

7.4 Illustration of conditions under which nodes may or may not oscillate . . 151

7.5 Nodes that are part of a strong component need not necessarily oscillate 153

7.6 Structural changes can introduce new dynamics 153

7.7 Graphical structure of the fittest network generated by the GA simulation of §5.4 154

8.1 A modern representation of signaling pathways that involve MAP-kinase and the EGF receptor. 170

LIST OF SYMBOLS

Summary of symbols used in model development

Symbol	*Description*
N	The number of different types of proteins in a modeled cell (or number of nodes in a given graph representing a single cell type)
T	The number of time steps that the dynamics is simulated for a single network (or modeled cell type)
S	The number of different random initial conditions from which dynamics are simulated for a given modeled cell type
M	An $N \times N$ connectivity matrix representing the interactions between modeled protein types
f_k	The fraction of protein types in a modeled cell that are protein kinases
f_p	The fraction of protein types in a modeled cell that are protein phosphatases ($f_p = 1 - f_k$)
p_k	The probability of interaction of a given kinase type with other protein types within a modeled cell
p_h	The probability of interaction of a given type of phosphatase with other protein types within a modeled cell
$b_i^{(t)}$	The number of attempts to attach or remove a phosphate group to molecules of protein type i at time step t
$a_i^{(t)}$	The average number of molecules of protein type i that are active at time t
$o_i^{(t)}$	The average number of molecules of protein type i that are phosphorylated at time t
o_{max}	The total number of molecules of each protein type (can be thought of as the concentration level of each protein type)
f	The sigmoidal shaped updating rule which relates the occupancy of each protein type at the next time step to a) the occupancy of the same protein type and b) the number of attempts to attach/remove phosphate groups at the current time step
g	The function relating the status of the catalytic domain of a protein type to the phosphorylation state of its regulatory domain/s

LIST OF SYMBOLS
(Continued)

Summary of symbols used in GA development

Symbol	*Description*
G	The number of generations for which the GA is simulated
C	The number of different cell types in the population at each generation of the GA
R	The number of randomly generated cell types sometimes included in early generations of a GA simulation
F	The number of cells per generation of the GA which reproduce
p_c	The crossover rate in the GA. Computed by dividing the total number of columns/rows actually swapped by the total number of protein types in the cell
p_m	The mutation rate in the GA. Computed by dividing the total number of elements flipped by the total number of non-zero connections in the matrix representing that cell type
F^c	The fitness of a cell type c in the population of the GA
$\bar{a}_i$	For a given protein type i, the average activity of that protein type at the final time step, T, calculated over S different starting simulations for a given cell type
A_i	The mean absolute deviation in activity of a given protein type i from the average $\bar{a}_i$
$\{p^c\}$	The set of proteins in a specified cell type passing the fitness criteria of the GA
p	The dimension of the set $\{p^c\}$ i.e., the number of proteins in a specified cell type passing the fitness criteria of the GA
$\{p_o^c\}$	The set of proteins in a specified cell type with no output interactions (corresponding row of M_{ij} has only zero elements)

Summary of symbols used in clustering algorithm

Symbol	*Description*
L	The number of cluster centers used in the clustering algorithm
E	The total error measure used for the clustering algorithm
E_j	The error for cluster j
c_j	The number of points classified to be part of cluster j
$\mathbf{V}_j$	A vector representing the cluster center of cluster j
$\tilde{a}$	The $S \times p$ matrix representing information on whether the activity of protein type j was fluctuating or stable during simulation i for a specified cell type
$\hat{a}$	The permuted version of $\tilde{a}$. Obtained from the completed clustering algorithm.
$d[\mathbf{x}, \mathbf{y}]$	Modified discrete (Hamming) distance between two vectors $\mathbf{x}$ and $\mathbf{y}$
w	The window length used for observing fluctuations in protein activity
Δ_{ij}	The maximum fluctuation in activity of protein j during the window w of simulation i for a specified cell type
D	The $S \times S$ matrix containing the distance between the modified activity vector of simulation i and the modified activity vector of simulation j for a specified cell type
$\hat{D}$	A matrix containing information on the average distance between points in cluster i from cluster j, calculated as the distance from cluster center j.

GLOSSARY OF BIOLOGICAL TERMS

ADP Adenosine 5'-diphosphate. Nucleotide that is produced by hydrolysis of the terminal phosphate of ATP. It regenerates ATP when phosphorylated by an energy-generating process such as oxidative phosphorylation.

amino acid Organic molecule containing both an amine group and a carboxyl group. Those that serve as the building blocks of proteins are alpha amino acids, having both the amino and carboxyl groups linked to the same carbon atom.

apoptosis Programmed cell death.

ATP Adenosine 5'-triphosphate. Composed of adenine, ribose and three phosphate groups. ATP is the principal carrier of chemical energy in cells. Synthesis in cells from ADP is driven by energy-yielding processes; The terminal phosphate groups are highly reactive in the sense that their hydrolysis, or transfer to another molecule, takes place with release of a large amount of free energy.

autophosphorylation The process whereby a protein kinase phosphorylates itself.

calmodulin Ubiquitous calcium binding protein whose binding to other proteins is governed by changes in intracellular Ca^{2+} concentration. Its binding modifies the activity of many target proteins.

CAMK-II CaM-kinase II or Ca^{2+}/calmodulin-dependent protein kinase II. A multi-functional protein kinase found in all animals but is especially enriched in the nervous system. Part of a family of Ca^{2+}/calmodulin-dependent protein kinases which phosphorylate serines or threonines in proteins. Thought to function as a molecular memory device since it can switch to an active state when exposed to Ca^{2+}/calmodulin and remain active even when the Ca^{2+} is withdrawn. This is because the kinase undergoes autophosphorylation.

cell cycle Reproductive cycle of the cell. The orderly sequence of events by which the cell duplicates its contents and divides in two.

cell signaling pathway A relay series of chemical reactions whereby a cell converts an extracellular signal into a response. The latter stages of the pathway often involve changes in gene expression.

GLOSSARY OF BIOLOGICAL TERMS (Continued)

checkpoint	Point in the eukaryotic cell-division cycle where progress through the cycle can be halted until conditions are suitable for the cell to proceed to the next stage.
chemotaxis	Motile response of a cell or organism that carries it toward or away from a diffusible chemical.
conformational state	Spatial location of the atoms of a molecule. For example, the precise shape of a protein or other macromolecule in three dimensions.
cross-talk	Communication between two separate linear signal transduction pathways whereby one molecule in a signaling pathway regulates the activity of a component considered to be in a separate pathway.
cyclic AMP (cAMP)	Nucleotide that is generated from ATP in response to hormonal stimulation of cell-surface receptors. Cyclic AMP acts as a signaling molecule by activating PKA. It is often referred to as a "second messenger" molecule.
cyclic AMP-dependent kinase	Also known as PKA or A-kinase. Enzyme that phosphorylates target proteins in response to a rise in intracellular cyclic AMP. First identified in skeletal muscle as part of the pathway of regulation of glycogen breakdown in response to adrenaline.
cytoplasm	Contents of a cell that are contained within its plasma membrane but, in the case of eukaryotic cells, outside the nucleus.
dephosphorylation	Removal of a phosphate group from an enzyme (reversing the effect of phosphorylation) by a protein phosphatase.
differentiation	Process by which a cell undergoes changes to an overtly specialized cell type.
domain	Portion of a protein that has a tertiary structure of its own. In larger proteins each domain is connected to other domains by short flexible regions of polypeptide.
domain shuffling	The process by which a segment of DNA is moved to another segment on the same chromosome. This causes a rearrangement of the DNA sequence which can result in the creation of a new gene. Many genes have common domains, hence the term "domain shuffling".

GLOSSARY OF BIOLOGICAL TERMS (Continued)

DNA	Deoxyribonucleic acid. The genetic material of all cells and many viruses. Polynucleotide formed from covalently linked deoxyribonucleotide units.
EGF receptor	Epidermal growth factor receptor. The first receptor protein discovered to be a tyrosine specific protein kinase. Binding of EGF activates the intracellular domain of the receptor and stimulates the proliferation of epidermal cells (and a variety of other cell types).
***Escherichia coli* (*E. coli*)**	A rod-like bacterium normally found in the colon of humans and other mammals and widely used in biomedical research.
enzyme	A protein that catalyzes a specific chemical reaction.
eukaryote	Living organism composed of one or more cells with a distinct nucleus and cytoplasm. Includes all forms of life except bacteria (prokaryotes) and viruses.
fibroblast	A cell type found in connective tissue.
gene knockouts	Also referred to as "gene targeting". A fundamental technology in molecular biology for testing the functional role of a particular gene in mammals. Gene targeting is most well studied in the mammalian mouse. Genes are inactivated or modified in the mouse embryo and the resulting defects are assessed. It is possible to target a gene in a specific cell type in a particular region of the brain (e.g., the CA1 region of the hippocampus) in order to study the effects of the "knocked out" gene on neurological development or memory.
genome	Total genetic information carried by a cell or organism.
glycolytic pathway	The sequence of chemical reactions resulting in the breakdown of glucose (glycolysis).
in vitro	Term used to refer to cells growing in culture or processes occurring in an isolated cell-free extract. (Latin for "in glass".)
in vivo	Term used to refer to processes occurring in an intact cell or organism. (Latin for "in life".)
lipid bilayer	Thin bimolecular sheet of mainly phospholipid molecules that forms the structural basis for all cell membranes. The two layers of lipid molecules are packed with their hydrophobic ("water-

lipid bilayer (ctd) hating") tails pointing inward and their hydrophilic ("water-loving") heads outward.

LTD Long-term depression. Phenomenon observed in some types of neurons whereby a reduction in synaptic strength can occur over a long time period.

LTP Long-term potentiation. A specific type of facilitation (increase in the effectiveness of specific synaptic connections) seen originally in hippocampal neurons [1] that can last for days or even weeks. LTP properties are not the same at all synapses. In the CA1 region of the hippocampus, LTP has three interesting properties: (*i*) cooperativity, (*ii*) associativity, (*iii*) specificity. For further detail see [2, 3]

MAP kinase Mitogen-activated protein kinase. A protein kinase that performs a crucial step in relaying signals from the plasma membrane to the nucleus. Turned on by a wide range of proliferation- or differentiation-inducing signals.

meiosis Special type of cell division whereby eggs and sperm cells are produced, involving a reduction in the amount of genetic material. Comprises two successive nuclear divisions with only one round of DNA replication, which produces four haploid daughter cells from an initial diploid cell.

membrane Double layer of lipid molecules and associated proteins that encloses all cells and, in eukaryotic cells, many organelles.

metabolism The sum total of the chemical processes that occur in living cells.

micron (μm) Unit of measurement often applied to cells/organelles (10^{-6}m).

mitosis Division of the nucleus of a eukaryotic cell, involving condensation of he DNA into visible chromosomes. (From the Greek *mitos* meaning "a thread", referring to the thread-like appearance of the condensed chromosomes.

neuron (nerve cell) Cells specialized to receive, conduct and transmit signals in the nervous system.

neurotransmitter Small signaling molecule secreted by the presynaptic nerve cell (neuron) at a chemical synapse to relay the signal to the postsynaptic cell. Examples include glutamate and acetylcholine.

GLOSSARY OF BIOLOGICAL TERMS (Continued)

NMDA receptor	N-methyl-D-aspartate receptor. A highly Ca^{2+}-permeable, ligand gated ion channel in neurons.
nucleus	Predominant membrane-bounded organelle in a eukaryotic cell, containing DNA organized into chromosomes.
oocyte	Developing egg; usually a large and immobile cell.
phospholipid	The major category of lipid molecules used to construct biological membranes. Generally composed of two fatty acids linked through glycerol phosphate to one of a variety of polar groups.
phosphorylase	The enzyme which breaks down glycogen. Also known as glycogen phosphorylase.
phosphorylase kinase	One of the first Ca^{2+}/calmodulin-dependent protein kinases to be discovered in 1956. Responsible for activating glycogen breakdown by activating the enzyme glycogen phosphorylase.
phosphorylation	Reaction in which a phosphate group becomes covalently coupled to another molecule.
phosphotransferase	An enzyme possessing the ability to transfer a phosphate group from a donor to an acceptor (e.g., protein kinases). Such enzymes are very important in metabolism.
plasma membrane	Membrane that surrounds a living cell.
PP-1	Phosphoprotein phosphatase-1; known to be involved in glycogen metabolism.
prokaryote	Organism made of simple cells that lack a well-defined membrane-enclosed nucleus (e.g., a bacterium).
protein	The major macromolecular constituent of cells. A linear polymer of amino acids linked together by peptide bonds in a specific sequence.
protein kinase	Enzyme that transfers the terminal phosphate group of ATP to a specific amino acid of a target protein (e.g., serine, threonine or tyrosine).
protein phosphatase	Enzyme that removes a phosphate group from a protein by hydrolysis.

GLOSSARY OF BIOLOGICAL TERMS (Continued)

PKC — Also known as C-kinase or Ca^{2+}-dependent protein kinase. A protein kinase that, when activated by DAG (diacylglycerol) and an increase in the concentration of Ca^{2+}, phosphorylates target proteins on specific serine and threonine residues.

postsynaptic density — A characteristic feature of a chemical synapse. It is an accumulation of opaque material on the cytoplasmic face of the postsynaptic membrane. It represents the aggregation of neurotransmitter receptors and signaling proteins essential for synaptic transmission.

pseudosubstrate — Polypeptide that binds in the active site of an enzyme, blocking and preventing its activation. Some protein kinases remain inactive for this reason. Activation is achieved by removing the pseudosubstrate from the active site allowing access for the protein's substrate. The pseudosubstrate interaction mimics some aspects of the substrate's structure.

receptor — Protein that binds a specific extracellular signaling molecule (ligand) and initiates a response in the cell. Cell surface receptors, such as the acetylcholine receptor and the insulin receptor, are located in the plasma membrane, with their ligand-binding site exposed to the external medium. Intracellular receptors, such as steroid hormone receptors, bind ligands that diffuse into the cell across the plasma membrane.

RNA — Ribonucleic acid. Polymer formed from covalently linked ribonucleotide monomers. All RNA species are synthesized by transcription of DNA sequences.

second messenger — A small molecule (or ion) that is formed or released into the cytosol in response to an extracellular signal and helps to relay the signal to the interior of the cell. Examples include cAMP, IP_3 and Ca^{2+}.

serine/threonine/ tyrosine kinase — Protein kinase that phosphorylates serine or threonines or tyrosines on its target protein.

signaling molecule — An extracellular or intracellular molecule that cues the response of a cell to the behavior of other cells or objects in the environment.

GLOSSARY OF BIOLOGICAL TERMS (Continued)

signal transduction	Relaying of a signal by conversion from one physical or chemical form to another. In cell biology, the process by which a cell converts an extracellular signal into a response.
substrate	Molecule on which an enzyme acts.
synapse	Communicating cell-cell junction that allows signals to pass from a nerve cell to another nerve cell. In a chemical synapse the signal is carried by a diffusible neurotransmitter; in an electrical synapse a direct connection is made between the cytoplasm of the two cells via gap junctions.
Tar	Transmembrane receptor involved in signal transduction during bacterial chemotaxis in *E. coli.*
targeting subunit	The part of a protein kinase or protein phosphatase which directs the catalytic subunit to the target locus. The target may be an organelle or a membrane component of the cell. This subunit may serve to position the catalytic subunit close to a particular substrate or to sequester it from other substrates and ligands.
Xenopus	African toad. Two species of *Xenopus* are commonly used for studies in developmental biology and genetics. These are *Xenopus laevis* and *Xenopus tropicalis.*
x-ray crystallography	Technique for determining the three-dimensional arrangement of atoms in a molecule based on the diffraction pattern of x-rays passing through a crystal of the molecule.

CHAPTER 1

INTRODUCTION

"The whole is more than the sum of the parts." —Aristotle (384-322 BC)

Mathematics pervades our everyday lives—"There is no branch of mathematics, however abstract, which may not some day be applied to phenomena of the real world."† The "golden age of mathematics" is often referred to as the 19th century post-Newtonian‡ era where the foundations of modern branches of mathematics were laid. One might argue that 20th century mathematics has been noted more for consolidation of existing discoveries rather than for profound innovation. Nevertheless, mathematics has greatly enhanced revolution in physics and is being increasingly applied to the biological sciences.

The first application of mathematics to a biological problem was by a British physician and experimentalist, William Harvey, in 1616 [5]. He used a quantitative proof to show that the blood circulates continuously around our bodies, an idea which had been hinted at earlier by Leonardo da Vinci. By 1850, two markedly different mathematical modeling approaches had emerged. The mathematical laws of physics had been laid down in differential equation form which enabled small, well-defined problems to be studied accurately and in detail. Statistical methods had been developed to study averaged quantities in order to coarsely analyze complex systems. Nowadays, techniques in non-linear dynamical systems are important in providing valuable insight into biological problems. The modern theory of dynamical systems has its foundations in the work of Henri Poincaré (1854-1912). The growing area of mathematical biology has extended to areas as diverse as developmental biology, population ecology, evolution, genetics and neuroscience.

---

†N. Lobatchevsky (1792–1856), from Ref. [4], pp. 90.

‡Newton (1642-1727) and Leibnitz (1646-1716) led the way with their development of the calculus.

Advances in the biological sciences progress at an ever expanding pace, and it would not be remiss to suggest that mathematicians might be in danger of missing out on some of the most exciting and interesting scientific problems of the century if they do not take an interest in some of the unresolved questions posed by biologists. This kind of research presents its own difficulties, not the least of which requires establishing a platform of understanding from which a biologist and a mathematician can communicate and exchange ideas. This work is also further challenged by the fast pace at which the vast and often contradictory experimental literature expands.

In the research field of cell biology, molecular techniques have propelled a truly reductionist approach, propounding the widespread view that "to understand better one need only go smaller" [6]. The announcement earlier last year of the completion of the first draft of the sequence of the human genome has increased interest in a variety of problems in genetics. It is currently estimated that there are probably about 30,000 genes encoded in the human genome. The general problem of predicting the structure of a protein from its amino acid remains unresolved, and the number of proteins encoded by the genome is still unknown. Only a small percentage of the proteins that have been identified has been studied in enough depth to reveal their structure and mechanisms of catalysis. However, there are wider issues still facing the community of cell biologists.

Researchers now begin to question whether the sheer complexity of a single cell may make a reductionist explanation impossible [6]. Could a mathematical model, which incorporated all this detail provide any useful information at all? How do all the components inside a cell cooperate to give rise to a complex and organized unit? Why are there only a relatively small number of different behavioral modes of a cell (see Figure 1.1)? A cell is, after all, the most fundamental unit of biological organisms which can be described as being alive. It adapts, self-regulates, reproduces and contains and organizes all the chemical reactions required for life.

Despite the amount of detailed information available, even now, in the year 2001, the most fundamental processes that occur in cells are not fully understood. Never was Aristotle's observation that the whole is more than the sum of the parts more evident. The field is ripe for mathematical insight.

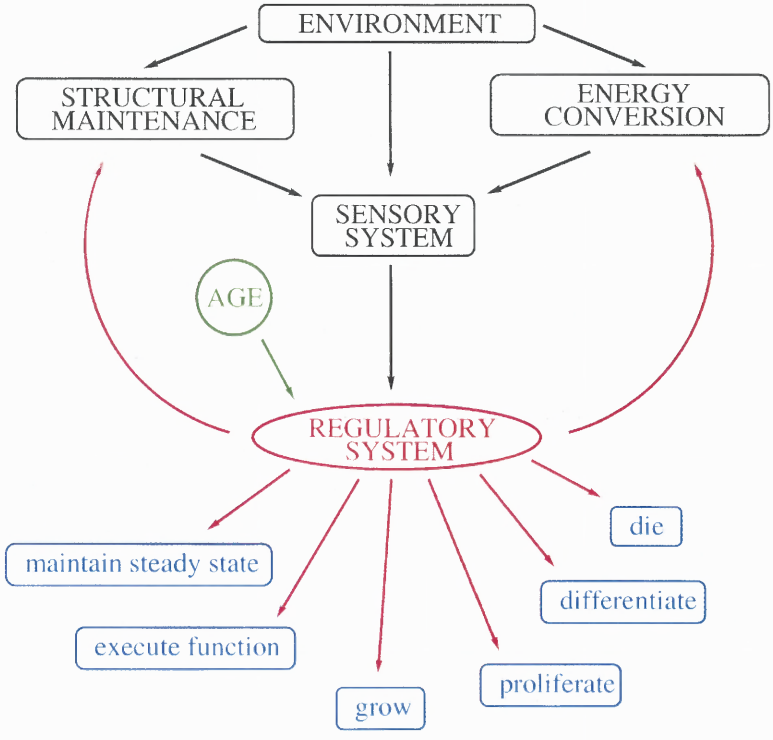


Figure 1.1: Cells typically exhibit only a small, finite number of different behavioral modes. For example, growth, proliferation, differentiation, death (adapted from Ref. [7]). There is currently no satisfactory explanation as to why an individual cell can only choose between a limited set of possible outcomes.

A sound mathematical model first requires a good understanding of the biological problem under investigation and often the mathematics will be dictated by what is known in the biology. The governing equations that are developed must provide as realistic a representation of the biological phenomenon as possible. They must also yield results which can be interpreted by an experimentalist in terms of predictions or insight which were not possible or practicable otherwise. Biological systems are complex and highly non-linear and thus realistic models will also tend to

be non-linear and may require computational methods for solution. Mathematical modeling has therefore also been aided by the steady progress in the speed and accessibility of high performance computers.

1.1 A Look Ahead

This chapter takes the reader through some of the biological terminology and literature necessary to appreciate the challenges involved in developing models of signal transduction. A glossary of terms is provided at the start of this document. The traditional manner in which cell signaling pathways are studied is discussed and the mathematical modeling approach for these "cascade" models is presented. The drawbacks of these models are then highlighted. The increasing awareness that pathways interact at many levels requires a shift in modeling approach from individual cascades to networks. The most well studied signal transduction network is that of the bacterium *E. Coli.* A continuum model of this signaling network is briefly discussed. The difficulties of applying detailed kinetic modeling methods to cellular systems where the data is less widely available are examined.

An alternative "connectionist" strategy is contrasted with prior approaches. Three different discrete parallel processing models are selected and discussed in turn: (a) Neural Networks, (b) Boolean Networks and (c) Cellular Automata. These discrete models are contrasted with the continuum approach. In particular, one noteworthy computational study of biological regulation networks is reviewed in this context. Desirable properties of cell signaling networks are translated into the mathematical terminology of dynamical systems. The predictive powers of the discrete versus the continuum modeling approach are also addressed. The specific modeling approach taken in this thesis is then outlined in light of these data.

1.2 Progress in Cell Biology

"In most sciences one generation tears down what another has built, and what one has established another undoes. In mathematics alone each generation builds a new story to the old structure." —Hermann Hankel (1839–1873)[†]

Progress in the biological sciences is very different from that of mathematics. Many fundamental mathematical techniques had been well formulated by the mid-nineteenth century. These results still today stand the test of time made possible by rigorous mathematical proof. Developments in cell biology, however, are often led by advances in experimental techniques and equipment: the invention of the centrifuge enabled molecules to be separated according to weight; the development of chromatography was used to aid the identification of the amino acids and hence the amino acid content of proteins; X-ray diffraction helped the study of the arrangement of atoms within proteins and other crystalline substances; the employment of chemical "tracers" such as radioactive isotopes enabled experimentalists to study the chemistry of living cells and hence reactions involved with metabolism; the improvement of microscopy enabled the fine structure of cells to be studied; and the discovery of dyes to stain parts of the cell meant that aspects of cell behavior could be studied for the first time (e.g., cell division). Ingenious and carefully controlled experiments are the crux of good research in molecular biology, biochemistry, biophysics and cell physiology. Each of these developments (and many more) has played an important role, and in one way or another contributed to the current understanding of cell structure and function.

The term "cell" was first used by Robert Hooke in 1665 to describe the small compartments he observed in cork cells under a light microscope [9]. In order to give some indication of scale, a light microscope gives enough resolution to discern shapes on the order of half a micron in length. This is roughly the length of a

[†]Quotation from Ref. [8], pp. 181

bacterial cell. A typical animal cell is somewhat larger, roughly of the order of 10–20 microns in diameter. By the early 19th century, light microscopes were useful enough to establish that all plant and animal tissues were comprised of cells, and that a single cell was an independent unit of life. In 1839, Theodore Schwann and Mattheis Schleiden recognized that cells are the elementary particles of all organisms [10]. Their observations and comparisons in plant and animal tissue indicated that there are basic features common to all cells (e.g., a cell membrane, nucleus and cell body). They also noticed that some organisms are unicellular and some are multicellular. This work was the beginning of "cell theory" and the basis for Virchow's "Cellular Pathology". By around 1860, Rudolph Virchow had asserted that all cells were derived from other cells[†] [11]. It was becoming clear that all living organisms, no matter how large or small, began life as a single cell.

The commercial introduction of the electron microscope in the late 1930s enabled researchers to increase their viewing resolution to scales in the nanometer range. The physical structure of cells could then be studied in far more detail. Membrane structure and organelles could be discerned. The age old view of the cell as a primitive sack of protoplasm was changing; cells were beginning to be viewed as structures containing a complex, ordered set of molecules.

Researchers then began to study cellular kinetics and were trying to identify the major pathways by which chemical changes occur in the cell. Part of this analysis was concerned with the understanding of the mechanisms of individual reactions. Evidence was emerging that cellular activity is controlled by chemical catalysts. These chemicals are called "enzymes". The smooth functioning of these enzymes enable reactions to proceed at high speed. Nowadays, it is recognized that most reactions require enzymes to catalyze them. As part of the study of the chains of these reactions

[†] "Where a cell arises, there a cell must have previously existed, (*omnis cellula e cellula*), just as an animal can spring only from an animal, a plant only from a plant." Ref. [11], pp. 27.

or metabolic pathways, researchers were looking at the synthesis and breakdown of compounds within the cell. They then had the problem of trying to understand how it is possible for the cell to strike an energy balance in these reactions. By the 1940s, it was discovered that the cell runs an energy bank which can trap and store energy that can be regulated on demand.

All enzymes are proteins, thus much of life revolves around the activities of proteins[†]. Proteins are important in biological systems as they have structural and functional roles. The first detailed molecular structure and chemical sequence of a protein were found for the hormone insulin in 1953 by a Cambridge researcher, Frederick Sanger [12], for which he was later awarded the Nobel Prize in chemistry. Insulin had the virtue of being a rather small protein yet was very important to bodily function. Sanger broke the protein up into smaller peptide fragments, each containing two or three amino acids. Each of the smaller fragments could be analyzed via chromatographic methods. Once all the peptide fragments had been identified, the problem was then how to piece the puzzle back together. To complete the sequence of the insulin protein took several years, but it enabled the complete structure of an important molecule to be identified. Nowadays, these procedures are all fully automated and can be achieved in a few days.

Chemical structures of other important intracellular molecules were also determined in the latter part of the 20th century, including the structure of DNA, RNA, proteins, phospholipids and carbohydrates. Small molecules were discovered which convey information through the cytoplasm of cells from the cell surface to a target enzyme inside the cell. As a result of this research, many individual reaction pathways have been identified and are now understood. The study of two of the first cell signaling pathways to be discovered will be discussed briefly in the next section.

[†]A protein is a compound made of amino acids.

1.3 Cell Signaling Pathways as Linear Cascades

Proteins, together with other molecules, integrate the cell's response to its environment in a way that affects whether the cell divides, differentiates, changes shape or metabolic rate. It is the proteins which perform the most wide ranging functions inside cells, and hence these will be the intracellular components that will be important to consider when developing a mathematical or computational model of intracellular signaling.

Modeling in cell biology has been dictated to by the manner in which cell biologists study cell signaling pathways. In order to understand the role of proteins in these pathways, the way in which proteins can affect one another must be understood. Signaling events, controlled by proteins, are highly dynamic processes. One specific, dynamic and reversible, mechanism via which proteins can be modified will be highlighted. This will be relevant later, as this is considered to be a salient feature of cell signaling that should be incorporated in any mathematical model. Historically, kinetic studies of individual signaling pathways have been commonplace. The general methodology of these studies will be briefly introduced. The problems of the traditional linear cascade models of intracellular signaling events will be highlighted in the light of more recent developments in the field.

This section will thus be concerned with the following questions:

- How are cell signaling pathways typically studied?
- What is the role of protein modification in these pathways (specifically, protein phosphorylation)?
- How are mathematical models for cascades typically developed?
- What are the problems with the cascade theory of signaling?

Every protein has a specific three-dimensional shape or "conformational state". This shape will determine how it interacts with other intracellular molecules. Moreover, the shape of the protein inside the cell can change dynamically during

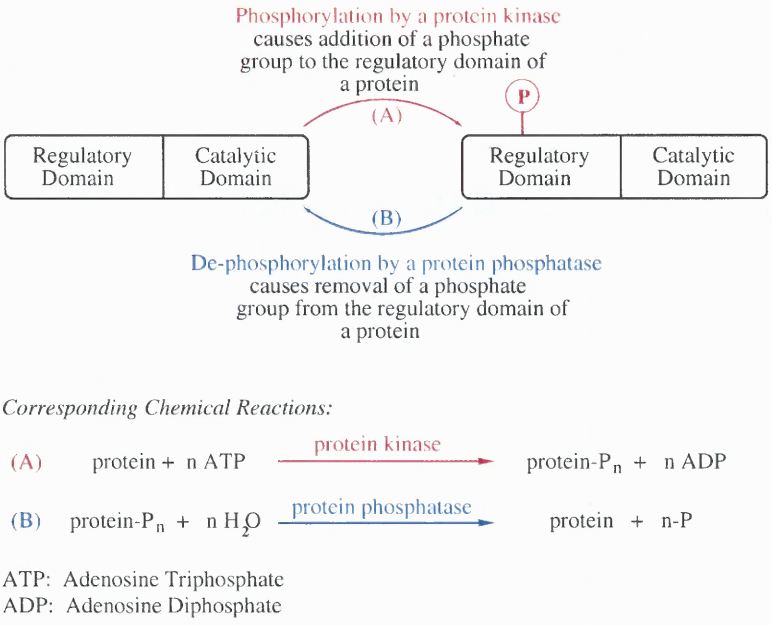


Figure 1.2: Phosphorylation (via protein kinases) and dephosphorylation (via protein phosphatases) reactions provide a dynamic and reversible mechanism for the modification of proteins.

its lifetime. There are several mechanisms of protein modification of which "phosphorylation" is known to be the most important. This is the process of adding a phosphate molecule to an amino acid side chain of the protein molecule. The phosphate molecule is taken from a molecule of ATP (see Figure 1.2).

In particular, the behavior of a cell is controlled by special classes of signaling proteins. Kinases and phosphatases form the majority of these signaling molecules. These signaling proteins affect the way in which the cell responds to hormones and other extracellular cues. Protein kinases are also controlled by small chemicals called "second messengers". Kinases phosphorylate other molecules. Phosphatases dephosphorylate other molecules.

The covalent attachment of a phosphate group to an amino acid side chain of a protein causes a conformational (structural) change in the protein. This is because the transfer of the phosphate group from an ATP molecule to a hydroxl (-OH) group on a serine (Ser), tyrosine (Tyr) or threonine (Thr) amino acid (see Figure 1.3) introduces

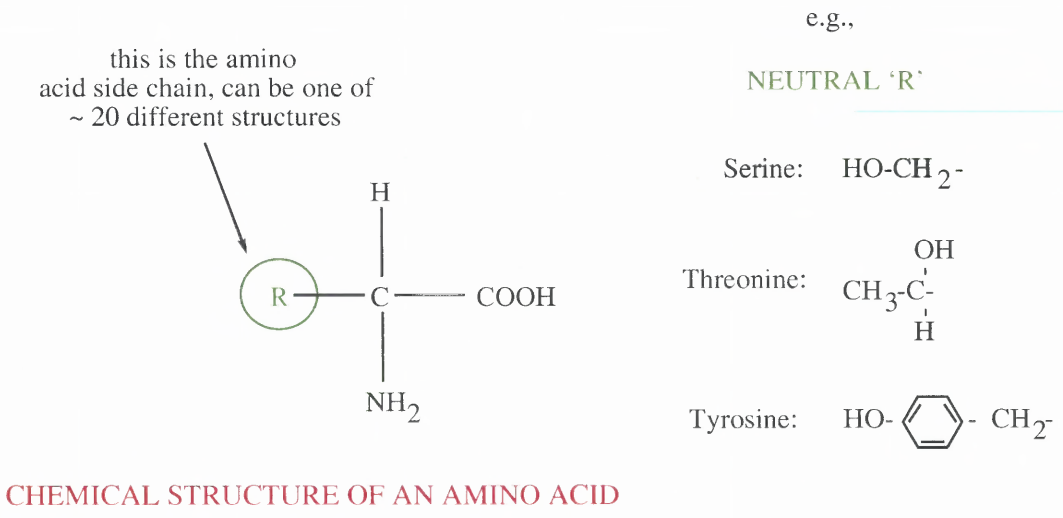


Figure 1.3: Phosphorylation of proteins involves the transfer of a phosphate group from an ATP molecule to a hydroxl (-OH) group on a Ser/Tyr/Thr residue of an amino acid. All proteins are made up of chains of amino acids. The amino acids are made up of C, H, O and N. Each amino acid contains a basic group ($-NH_2$) and an acidic group ($-COOH$)

two negative charges which can attract nearby postively charged side chains [13]. The roles of these residues (Ser/Tyr/Thr) during phosphorylation can be investigated by "site-directed mutagenesis" experiments [14, 15]. These procedures enable specific mutations of a protein to be performed. For example, a tyrosine (Tyr) amino acid residue can be replaced by glutamate (Glu). In this case, the negative charge of the glutamate has the same effect as phosphorylation, so that the protein acts as if it is phosphorylated. Site directed mutagenesis experiments are therefore useful to explore the relationship between protein structure and function, and particularly to explore the role of phosphorylation sites in the activation of enzymes [16, 17]. The conformational changes resulting from phosphorylation alter the biological properties of an enzyme, for example changing its affinity for substrates, activators or inhibitors.

Signals are typically received at the cell surface by receptors and are translated into changes in the activity of intracellular molecules such as the protein kinases and phosphatases. Signal transduction pathways are studied in one of two ways. The first approach is a "downstream" one, where the first step is to identify the receptor

substrate and then work down the pathway. The second is an "upstream" one where from the cell effect in question, one works backwards towards the receptor in order to identify the components in the pathway.

In the early 1940s work by Cori and Green [18] on an enzyme, "phosphorylase", found in rabbit skeletal muscle, indicated that the molecule could exist in two different forms, the active form, "a", and the inactive form, "b". Phosphorylase was known to play a role in glycogen synthesis and breakdown. The fact that the enzyme could exist in inter-convertible states led researchers to believe that it played a significant regulatory role within the cell. Another enzyme was then found to convert the inactive form of phosphorylase to the active form [18, 19]. The later discovery that ATP was required for this conversion led to the hypothesis that the enzyme involved in the inter-conversion was a protein kinase. This hypothesis was correct and the kinase was later named "phosphorylase kinase". This was the first time it was realized that a phosphorylation–dephosphorylation mechanism could play an important role in cellular processes such as metabolism. Moreover, the realization that these reversible reactions could occur dynamically inside cells was also important. Other small molecules [calcium and cyclic-AMP (cAMP), now known as second messengers] were also shown to be important in affecting the activity of phosphorylase kinase. In fact calcium was known to play a role in muscle contraction, coupling the breakdown of glycogen by phosphorylase (an energy-yielding process) with muscle contraction (an energy-utilizing process).

Later work indicated that phosphorylase kinase itself was also subject to phosphorylation. The protein kinase found to modify phosphorylase kinase was named "cAMP-dependent protein kinase" because its role in cellular function was far more wide ranging. This led to the description of the first cascade involving protein kinases in cell signaling (see Figure 1.4) via an upstream or bottom-up approach.

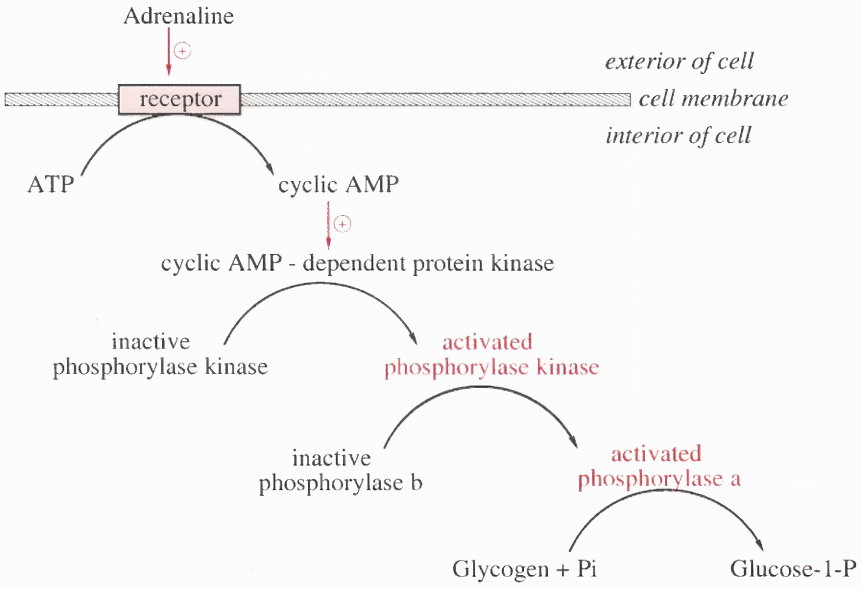


Figure 1.4: The intracellular signaling pathway thought to control glycogenolysis (adapted from Ref. [20]).

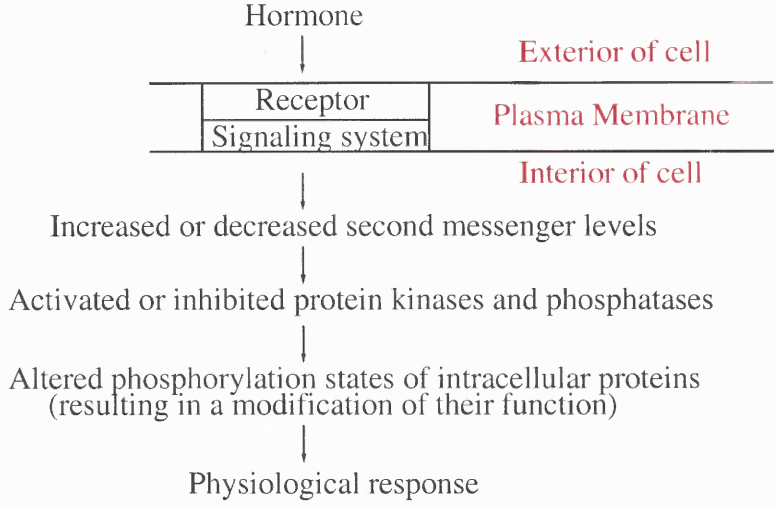


Figure 1.5: Typical events in a cell signaling pathway (adapted from Ref. [21]).

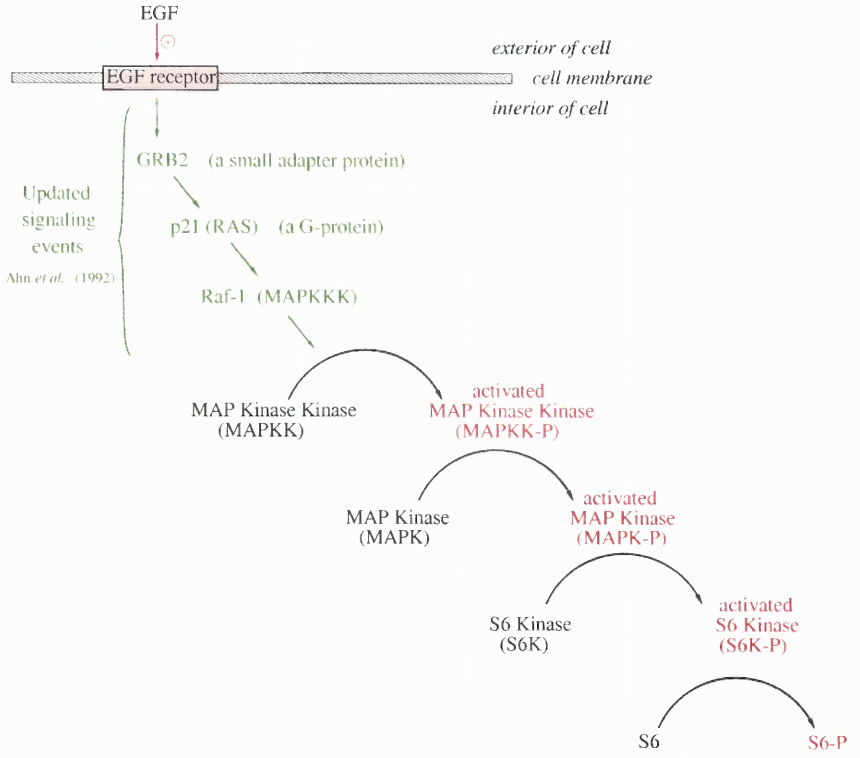


Figure 1.6: The MAP-kinase signaling cascade (adapted from Ref. [20]). More recently discovered signaling events are shown in green [22].

A signal arriving at the cell surface is translated into changes within the cell via a sequence of events, which are often classically described by the diagram in Figure 1.5. In this case, a hormone at the cell surface causes alterations in receptor molecules at the cell surface. This leads to changes in second messenger molecules which in turn affect the activity of protein kinases and phosphatases, ultimately resulting in changes in cell function. In the case of the glycolytic pathway, the downstream events involved one protein kinase phosphorylating and activating another protein kinase, resulting in a response which is the breakdown of the compound glycogen.

The second protein kinase cascade that was discovered turned out to play a role in insulin action [20]. This process demonstrated that not only do protein kinases and phosphatases occur intracellularly, but they are also present in the cell membrane as receptors. The "Epidermal Growth Factor receptor" (EGF receptor) and the insulin receptor are special types of protein kinase (protein tyrosine kinase). In the

case of insulin action, the approach was also an upstream one. Insulin was known to activate a specific downstream protein kinase, "S6-kinase", and this kinase was, itself, thought to be modified by phosphorylation. The S6-kinase was found to be a substrate of another kinase, "MAP-kinase" (MAP-K), which itself could be modified by phosphorylation. This led to the discovery of "MAP-kinase kinase" (MAP-KK) which also existed in phosphorylated and dephosphorylated forms (see Figure 1.6). By 1993, the search was on to discover the regulator of MAP-KK. A MAP-kinase kinase kinase was later discovered.

Although the discovery of another kinase signaling cascade was exciting, there were still no significant studies into questions such as why a cell needs such a complicated mechanism to regulate hormone action. The upstream and downstream approaches also overlook the possibility of multiple targets for each kinase in the scheme [20]. The possibility that branching could occur in the cascades was not considered in the initial identification of the glycolytic and MAP-K pathways.

The reductionist approach that has been so successful in the biological sciences has led to the extensive cataloging of biochemical pathways. Cellular processes have gradually been broken down into their most fundamental components. Advances in computer technology now enable this information to be provided in databases freely available over the World Wide Web (see Appendix A).

Early mathematical and computational studies tended to concentrate on theoretical analyses of specific signaling pathways with a small number of components. In the late 1970s, Chock and Stadtman studied the properties of mono-cyclic and multi-cyclic cascades of enzymes in metabolic regulation [23, 25]. One possible role for cascades was thought to be signal amplification. For example, in the case of glycogen metabolism, a very small concentration of an extracellular compound can then make use of a reserve pool of glycogen extremely effectively and quickly. The analytical results were later tested via an *in vitro* cyclic cascade system [24]. This study

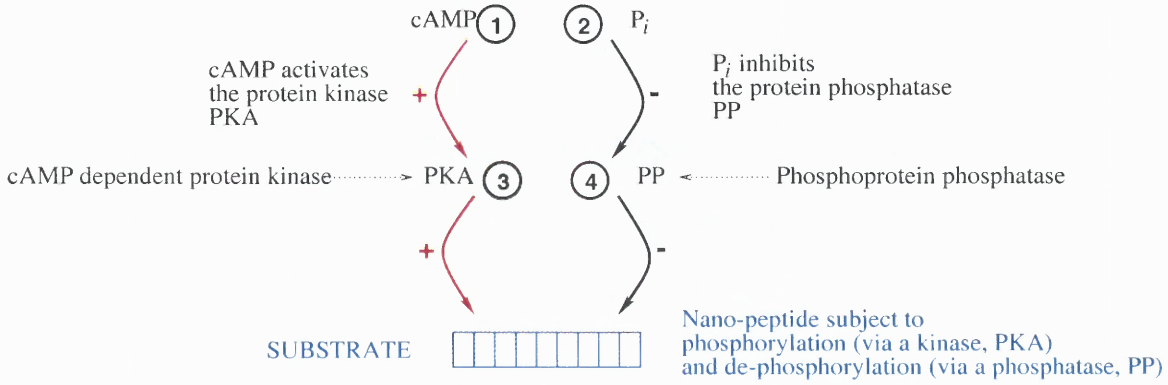


Figure 1.7: A mono-cyclic cascade system (studied by Chock & Stadtman [23, 24]). The higher the levels of cAMP, the larger the fractional number of substrate molecules in the phosphorylated state. The system reaches a steady state which is independent of the initial number of molecules of the nano-peptide that were phosphorylated. Even for this simple cascade system, there are a number of distinct reaction steps that need to be modeled. The system demonstrates increased flexibility because the activity of the substrate can be altered by changing the activity of the kinases and phosphatases either directly or via their effector molecules. An increased level of P_i will inhibit the protein phosphatase therefore enabling smaller concentrations of cAMP to activate the kinase and give an equivalent steady state level of phosphorylation. Fixing the concentration of cAMP and increasing levels of P_i causes an increase in the fraction of phosphorylated molecules of substrate. The change in response becomes more noticeable at higher levels of cAMP. In fact, increasing levels of cAMP and increasing levels of P_i have synergistic effects as both tend to favor increased levels of phosphorylation of the substrate.

examined the cyclic inter-conversion of a single enzyme between its phosphorylated and dephosphorylated forms.

The system consisted of a protein kinase (cAMP-dependent protein kinase) and a protein phosphatase (phosphoprotein phosphatase) together with regulatory molecules—cAMP as an activator for the protein kinase, and P_i (phosphate) as an inhibitor for the phosphatase (see Figure 1.7). The system constructed is referred to as a "mono-cyclic" cascade. The steady state of the system depends on the relative concentrations of each of these four components. The substrate molecule for the kinase and the phosphatase was an artificially constructed nano-peptide consisting of a chain of nine amino acids. By keeping the concentration of ATP sufficiently large for it to remain essentially constant, it was possible to study the effects of

altering the concentration of the regulatory or effector molecules cAMP and P_i. The cascade showed signal amplification and positive cooperativity. The latter feature of a signaling pathway means that the sensitivity of levels of phosphorylation of substrate to changes in levels of concentration of effector molecules was much greater than the sensitivity of the kinase to changing levels of cAMP. Even for this simple cascade system, there are a number of distinct reaction steps that need to be modeled (see Appendix B). The classical methods of enzyme kinetics used to study these reactions [26] involve the construction of a system of differential equations (B.1)–(B.6) which measure the rate of change of the concentration of each of the components of the system.

Given the importance of protein phosphorylation reactions inside cells, and as part of intracellular signaling pathways, it is natural to want to make predictions regarding the functionality and reliability of these pathways within cells. Cascades of protein kinases, such as the MAP-K cascade mentioned in §1.3, are postulated to demonstrate cooperativity, converting graded inputs into switch-like outputs [27, 28]. This has been investigated in the context of maturation of *Xenopus* oocytes [28]. The MAP-kinase cascade is thought to cause different cellular responses depending on the context [27].

Thus, researchers' current understanding of signaling proteins characterizes them as discrete units in series. Each protein acts as a relay station in a specific cascade of message amplification [13]. Advances in genetics have now made it possible to alter one specific type of molecule inside a cell at a time. The function of a specific protein can be investigated by disrupting its structure to prevent it from functioning normally. If signaling in the cell really occurs via a series of separate parallel pathways, then mutating individual proteins critical to these pathways should change the expected cellular effect. However, this is not always the case. The relative lack of effect in such "gene knockout" experiments of signaling proteins suggests the

system has multiple alternative pathways [29]. Modeling of single pathways may never give insight or provide satisfactory explanations as to why this is the case. There are also other inconsistencies with this viewpoint.

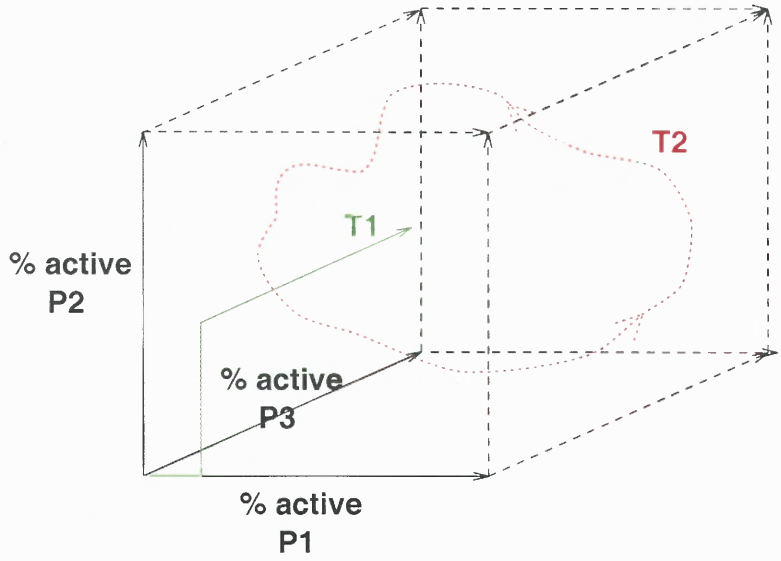


Figure 1.8: Representing protein activity in phase space. Trajectory T1 shows the effect that a small level of active molecules of protein type P1 can have on a cascade of three protein types P1 $\rightarrow$ P2 $\rightarrow$ P3 (where protein type P1 activates molecules of P2 which in turn activate molecules of P3). Trajectory T2 shows an example of periodic oscillations in activity of the three different protein types (not connected in a linear cascade). T2 therefore represents a closed curve in phase space.

Consider, for example, a cascade of three different types of proteins. As discussed above, a small change in the input layer can cause an avalanche or amplification of activity of molecules at lower levels of the cascade. Suppose there are 10,000 molecules of P1, P2 and P3 freely diffusing in part of the intracellular medium. Assume that each molecule of protein P1 can activate 100 molecules of P2 during a specified time interval, t, and that each molecule of P2 can activate 100 molecules of P3 in the same time interval. Then, if an input signal activates ten of the molecules of P1, these in turn, within a time t, can activate 1000 molecules of P2 and hence 100,000 molecules of P3 can be activated very easily. This change in percentage activity levels of each protein can be illustrated on a phase space diagram, shown in Figure 1.8. In such a case, the system will eventually be driven towards the

corners of the phase space, with a tendency to generate an all or none response (the latter could occur in the case of negative regulation). The changes that occur can be monitored by following a trajectory (T1) in the phase space.

If, however, a cascade of negative and positive regulation occurs, the system may not generate an all or none response. Instead, intermediate levels of activity may be possible. The system may then lie anywhere in its state space of possibilities with some probability. If the state of the system is interpreted as an indication of the total number of possible modes of behavior, there could be a vast number of possible states. How can this be reconciled with the knowledge that there are really only a small number of cellular fates (see Figure 1.1)?

In some cases, oscillations in intracellular molecules may occur. This could be represented by a closed curve (T2) in phase space. However, in such a case, the molecular species (P1,P2,P3) involved would not interact in series. Cycling patterns of activity are known to occur in intracellular signaling networks and these oscillations play functional roles [30]. The "cell cycle" is a prime example of a complex phosphorylation cycle involving a set series of activation and inactivation of proteins [31, 32]. At certain critical phases of the cell cycle, called "check-points", the signaling network is particularly responsive to external signals.

Oscillations of intracellular second messengers such as calcium are also a common feature of cells [33, 34]. Spontaneous oscillations are observed to occur during *in vitro* maturation of mouse oocytes [35]. Calcium oscillations can vary in length from a fraction of a second in muscle cells, to several minutes when they play a role after fertilization, to hours for the calcium cycles that control cell division. Calcium is a ubiquitous second messenger implicated in almost all aspects of intracellular signaling [36]. Calcium signals are known to trigger life at fertilization, to control differentiation and cellular development, and also to control cell growth and death. The duration and amplitude of the calcium signals appear to be extremely important.

It remains a mystery as to how a calcium signal can be utilized to control vastly opposing responses: on the one hand as a signal for life and on the other as a signal for death [37]. The context in which the cell receives its environmental cues therefore appears to be extremely important [38, 39]. Oscillations have also been suggested to have certain thermodynamic advantages [40].

Some limitations to the cascade perspective of signaling have been outlined, but the story is not yet complete. It is just beginning. Further complexities will be highlighted in the following sections (§1.4, §1.5). There are properties of cells that cannot be explained with a cascade-type model. The complexities of cell signaling deepen and as they do, so models must evolve to incorporate the ever changing biological data.

1.4 Cross-talk Between Linear Pathways: A Signaling Network

Researchers continue to study cell composition and to investigate the chemical interactions that occur inside cells. As data and information become available, it is becoming clearer that there is considerable interaction between intracellular signaling pathways (see Figure 1.9).

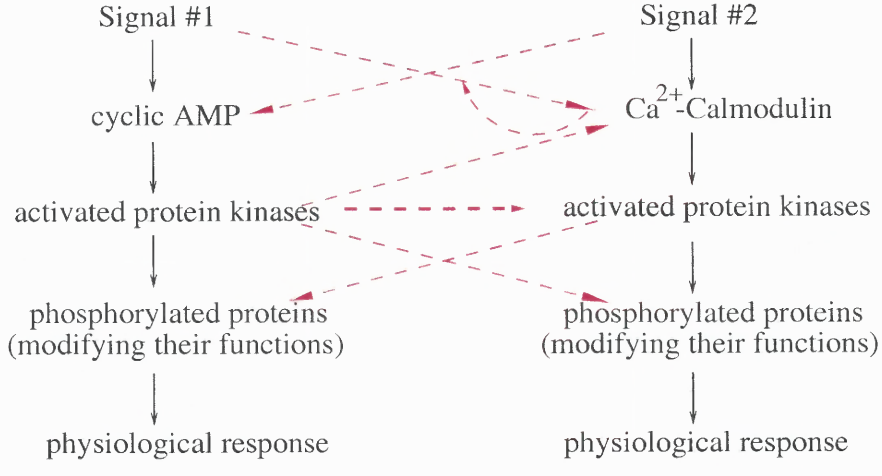


Figure 1.9: Interactions between pathways can be extensive. For example, cyclic AMP and calcium signaling pathways are interlinked at every level (adapted from Ref. [21]).

External signals received at the cell membrane by cell surface receptors cause changes in concentration levels of the second messengers calcium and cAMP. Pathways activated by calcium and cAMP are often studied independently, but they are in fact closely interlinked [21]. The simplistic view of cell signal transduction as many linear cascades is thus gradually being superseded by the realization that there are many diverse interactions between signaling pathways. In light of this, in this section the following questions will be addressed:

- How can mathematical models incorporate cross-talk?
- Are there any drawbacks with a detailed kinetic modeling approach?

The concept of the cell as a complex signaling network of molecules is therefore becoming more widely accepted [41]. A network model for intracellular signaling would seem more appropriate than a linear cascade model. The most detailed and extensive kinetic modeling of a small intracellular signaling system existing to date is that of a bacterial signaling network. The bacterium *Escherichia coli* is an example of a cell for which all the intracellular proteins involved in the chemotactic response have been identified. The components have been purified and sequenced and their kinetic data is available. The effect of deleting specific components of the signaling system have also been extensively documented.

Bray *et al.* [42] have explicitly modeled a proportion of this chemotactic signaling system using an extensive knowledge of these available data. A bacterium has the ability to adjust its swimming behavior in response to changes in the extracellular concentration of attractants and or repellent chemicals. Binding of these extracellular molecules to cell surface receptor molecules cause changes in the phosphorylation states of various intracellular protein kinases which in turn affect the rotation of the flagellar motor of the bacterium, allowing it to change its swimming behavior. This short term adjustment of the swimming pattern of the bacterium is achieved in about two tenths of a second.

The goal of the study by Bray and co-workers [42] was to combine all existing data into a single model and to predict the behavior of the bacterium in novel environments. The aim was also to have the potential to predict the effect of, as yet unperformed, mutation experiments. Standard results from enzyme kinetics were used to model the binding of the attractant to the cell surface receptor (named Tar). Protein auto-phosphorylation steps were modeled using Michaelis-Menten kinetics [26]. In fact the flagellar motor is a complicated structure being comprised of about forty proteins. Its phosphorylation state appears to determine whether it pauses or rotates counter-clockwise or clockwise. The resulting model describes the interactions between about half a dozen key proteins and has twenty-one different functions associated with it, each representing a single step in the signaling cascade. Computer simulations were used to study cell performance under a wide variety of different conditions. Numerical methods were used to solve precise differential equations arising from the detailed enzyme kinetics.

Despite the existing data available for this model, many aspects of the reaction pathways are still not clear and so Bray and his colleagues had to estimate the values of some of the unknown rate constants and average others from different experimental sources. Some of the reaction steps were purely hypothetical. Their model was therefore necessarily not correct in every detail. Results of their simulations managed to reproduce and explain some short term responses of the bacterium quite well but could not explain others. This is an example of an attempt to model a single, specific phosphorylation pathway which is part of a larger cell signaling network. It also contains only about half a dozen components.

It is thus possible to model specific cellular components based on kinetic parameters obtained via *in vitro* laboratory analysis. Once the data are known for several components, differential equations can be developed and predictions can be made which can then be tested experimentally. However, this approach is painstaking

and relies on the accuracy of the experimental data. In some cases, the data obtained in this manner may also not be representative of the phenomena which actually occur *in vivo* in the living cell.

It is still, however, a major goal of cell biology to understand the regulations of cell behavior in the reductive terms of molecular interactions of the intracellular signaling proteins. This aim is made explicit by the assertion that understanding a cell's responses to stimuli requires a full inventory of the details of all the molecular interactions involved [43]. The inventory of the kinetics and binding constants of all proteins and other molecules may then be used to construct a precise model of a cell's behavior. The possible predictive power of an explicit cell signaling model is potentially very extensive. If a model could be designed which could accurately predict the response of the cell in a given environment, this would reduce the need for extensive experimental testing of cells in a multitude of different extracellular environments. However, constructing a complete model of all the relevant details of molecular interactions in eukaryotic cells is likely to be a daunting task. This is evident by our growing knowledge of just a fraction of the proteins involved in cell signaling. More dauntingly, however, there are also somewhere between 10,000 and 100,000 different proteins inside mammalian cells and yet only a fraction of these have been studied in detail to the extent that their three-dimensional structures and mechanisms of catalysis are known. The lack of available data makes it hard to perform detailed simulations of this kind.

1.5 Further Levels of Complexity

Although many specific reactions inside cells are now well understood, it is the extensive interconnections between pathways that still make it difficult to predict the way in which the cell will respond to its environment. There are a multitude of proteins involved in cell signaling that can interact in a variety of ways. The control

of the activity of these proteins involves reversible binding and reversible covalent modification [31, 44, 45, 46]. There are many types of second messengers that bind to proteins and modulate their function and there are also several forms of protein modification.

This section will therefore be primarily concerned with:

- introducing further levels of complexity of protein phosphorylation,
- emphasizing the importance of phosphorylation in the control of cell behavior.

Phosphorylation is the most common and widespread form of modification that changes the activity of a protein [13, 45]. Protein phosphorylation by specific protein kinases (PKs), and dephosphorylation by specific protein phosphatases (PPs) may regulate as many as one third of all the proteins inside a cell [47, 48]. There are also many more protein kinases and protein phosphatases than other types of signaling proteins. It is estimated that approximately two percent of the genome of budding yeast and approximately five percent of the mammalian genome contains sequences encoding for protein kinases [47, 49]. This suggests that around 2,000 different types of protein kinases may exist in the mammalian genome [31, 50, 51]. There may also be about half as many different types of protein phosphatases as protein kinases [31, 52], although more recently Tony Hunter has estimated that the ratio of PKs:PPs may be closer to a one to one [53]. While not all of these kinases, phosphatases and other second messenger systems will be present in every different cell type, conservative estimates suggest that at least one hundred protein kinases, and more likely between 100 and 1,000 different protein kinases and protein phosphatases will be found in any one mammalian cell type [48, 54].

There are a variety of ways in which protein kinases and protein phosphatases interact, but some central themes emerge. Protein kinases generally possess one catalytic domain that is responsible for substrate recognition, ATP binding and phosphotransferase activity. There is then either one, or a series of regulatory

domains that may themselves be regulated by second messengers, phosphorylated by protein kinases or dephosphorylated by other protein phosphatases. Protein phosphatases also have a similar functional modular structure [44, 45]. For some protein phosphatases and protein kinases there are targeting subunits that perform the regulatory role [55]. These may also be important in directing the protein to specific intracellular locations. The substrates that any given protein kinase or protein phosphatase affects are mostly determined by the catalytic domain (for PKs) or targeting subunit (for PPs). Protein kinases and protein phosphatases can often phosphorylate substrates on more than one site with different affinities and this process is known as multi-site phosphorylation [44]. These kinds of interactions provide a considerable base level of complexity.

Another level of complexity is provided by the interconnections between protein kinases and protein phosphatases. Many of the protein kinases, phosphatases and other signaling proteins are thought to form part of pathways from the plasma to the nucleus or other cellular compartments [13] (e.g., the glycolytic and MAP-K pathways that were discussed in §1.3). Just as there is evidence for the involvement of protein kinases in signaling cascades, there is also evidence for protein phosphatase cascade systems [21]. Protein phosphatases are known to be activated by other protein kinases. Some protein kinases and protein phosphatases can be found on the plasma membrane as receptor molecules, others are located in the cytoplasm, the nucleus, or in other intracellular compartments [31, 44, 52]. Some of the most common substrates of protein kinases and protein phosphatases in these pathways are other kinases and phosphatases [31].

There is now growing evidence that these pathways "cross-talk" since protein kinases and protein phosphatases in one pathway interact with components in another [31, 41]. Many signaling pathways are now known to be conserved between prokaryotes and eukaryotes. The MAP-kinase pathway is the most notable of

these [56]. Many different types of contemporary proteins are built from domains that have been shuffled and mutated during evolution [57, 58]. This suggests that there is an elaborate network of interactions between protein kinases and protein phosphatases inside the cell. The connection patterns for such phosphorylation networks are likely to be different for each cell type. They are also dynamic since phosphorylation and dephosphorylation reactions are liable to occur continuously. It is estimated that the continuous turnover of phosphates on proteins consumes 20% of the cell's ATP [59]. There is also evidence that in red blood cells, approximately 80% of the cell's ATP is consumed in maintaining the cell's structural integrity [60]. This process involves phosphorylation reactions.

Since the discovery of a role for protein phosphorylation in glycogen metabolism, phosphorylation has now been shown to be part of every aspect of the life of a cell. Diseases such as cancer are now known to be linked to deficiencies in the cell signaling system and particularly to mutations in protein kinase or phosphatase genes[†]. Kinnunen [61] has indicated that cell transformation might be represented as a transition in the cell system, whereby a change in one sub-cellular system might induce a transition of other subsystems changing the state of the cell. Protein phosphorylation is now known to play a role in proliferation, metabolism, differentiation and movement [62]. In fact, some of the same kinases involved in signaling proliferation may also play a role in cell death. It has also been suggested that the cell may have several different options for cell death that can be activated by different perturbations of the protein phosphorylation signaling system [63].

Protein phosphorylation is now an area of vigorous scientific activity and numerous reviews on the role of protein phosphorylation have appeared in recent years [52, 53, 65, 66, 67, 68, 69, 70]. The number of enzymes found to be controlled by protein phosphorylation appears to be growing in a dramatic manner

[†]Volume 6, Issue 4, of *Seminars in Cancer Biology* in 1995 was dedicated entirely to a discussion of the role of Protein Serine/Threonine Phosphatases in Growth Control of Cells.

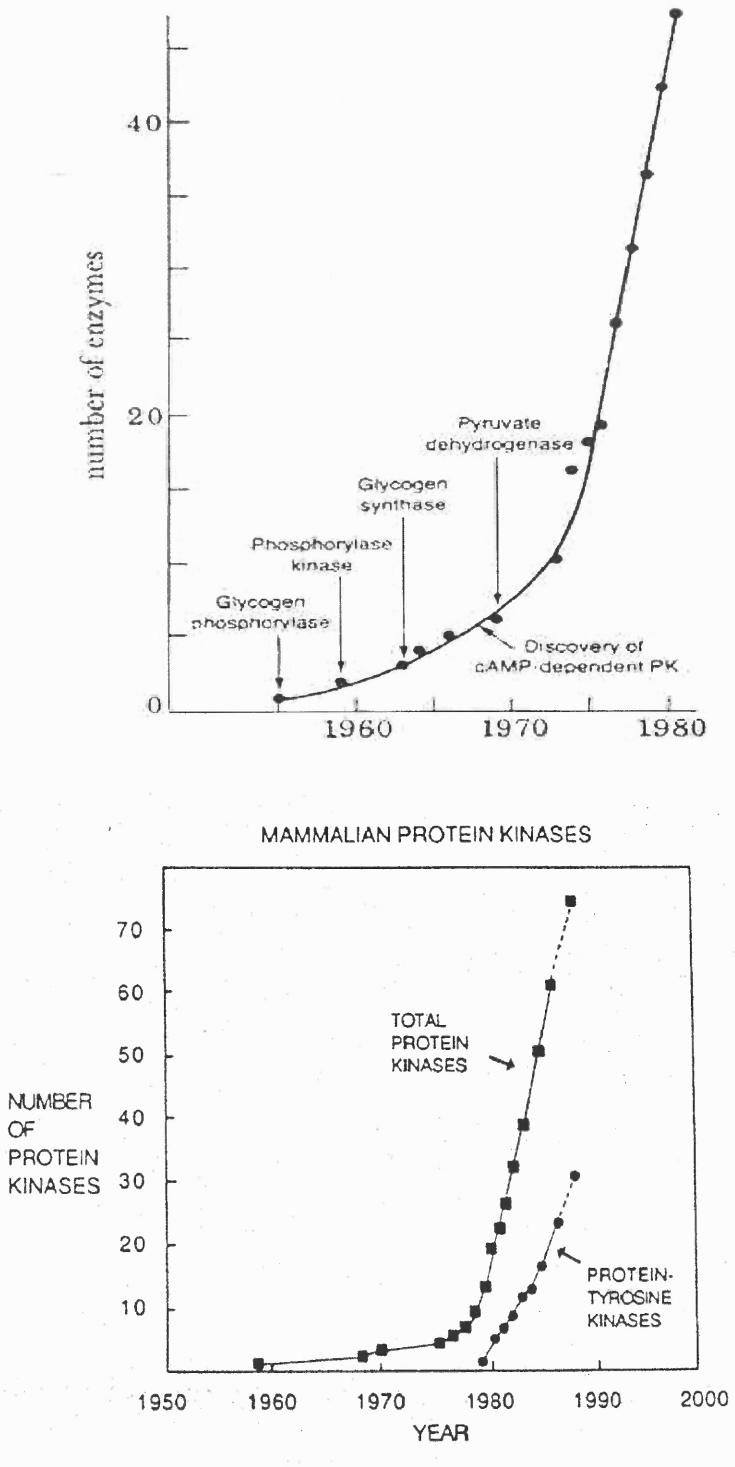


Figure 1.10: The growth in protein phosphorylation research. [Sources: [64] (top) and [49] (bottom).]

(see Figure 1.10). Even so, there are very few proteins whose three-dimensional structures are known for the phosphorylated and dephosphorylated states [53]. The importance of protein phosphorylation must also be indicated by the increasing number of databases containing information related specifically to this area of cell signaling (see Appendix A).

1.6 Proteins as the Computational Elements of Cells

Cellular networks therefore have a very complicated topology. Given the complexity of the protein kinases and protein phosphatases alone, it is not surprising that there have been no reports of an explicit model of a complete cell signaling system. Such a model would have to be different for each cell type. Furthermore, since these models might be as complex as the cell itself, they may not provide much insight into cell function [71]. Kauffman [72] has already argued that for any biochemical network to evolve it must be relatively insensitive to slight differences in enzymatic details. It may be possible to design a model of intracellular signaling before the precise details of all the molecular interactions are known.

Over the last decade researchers have begun to think of cells as computational units, with the protein molecules within them relaying information [39, 48]. The structure of such an intracellular network depends on molecular encounters that are limited by diffusion. Changes in the activity of a protein typically alters many target molecules and may even feed-back on themselves. Bray [41] has indicated that a protein can be thought of as an input-output device, with a non-linear function (typically hyperbolic or sigmoidal) which relates the input (substrate concentration) to the output (protein activity). Timing will be important in intracellular regulatory circuits and there is acknowledgement that many aspects of cell behavior require a capacity for information processing that is independent of the genome. Such behavior is essentially controlled by the activity of many thousands of proteins.

Bray [41] was the first to suggest that protein networks may have similarities to neural networks. Specifically, his paper discussed how one might try to model the response of a liver cell to the hormone glucagon. He suggested that an interconnected network could be used to model the parallel processes that occur in living cells.

Table 1.1: Major differences between neural networks and cell signaling networks.

	Salient features of networks	*Neural Networks*	*Cell Signaling Networks*
(i)	nodes	all nodes typically identical	nodes are not all equivalent in performance
(ii)	structure	typically layered with feed-forward connections only	regulatory signals often give rise to feedback connections resulting in cycles
(iii)	connectivity	highly connected	more sparsely connected
(iv)	learning rule	connectivity altered to perform a single function e.g., via a back-propagation algorithm	cells must be able to respond to multiple stimuli effectively;changes typically occur via evolutionary processes

The classical model for a neural network is a multi-layer perceptron network which is comprised of several layers, each node in a layer being connected to every element in the next layer. Multilayer perceptron networks usually consist of a single type of unit (node), often referred to as a neuron. The feed-forward connections between the nodes have weights associated with them. The network typically learns to perform a pattern recognition task based on a training set of inputs and associated outputs. During this procedure the weights in the network are adjusted to minimize the error between the actual output of the network and the expected output of the network. This is generally known as a supervised learning rule. A test set of inputs

is then used to determine whether the network has learned to perform the desired task. In such a network the nodes are generally binary decision units which are "ON" if the weighted sum of inputs into that node is greater than some pre-determined threshold. Bray compared features of neural networks to cell signaling networks and acknowledged that there are several fundamental differences that must be considered when modeling cell signaling networks (see Table 1.1).

A multi-layer perceptron model may not therefore provide a particularly effective way to study signaling networks with a parallel distributed architecture. In terms of a more realistically structured neural network model, a Hopfield network would be a more likely candidate [73]. However, there are important concepts to be taken from Bray's paper. There are also aspects of a cellular network that can be modeled in a similar fashion to that of a neural network (see Table 1.2). Bray discussed how models of cell signaling networks might help to explain how cells may respond to combinations of simultaneous extracellular signals, how cells robustly respond to their environment, and also how new signaling pathways may have emerged through evolutionary changes.

The cell can thus be considered as an information processing system [39, 48] with the intracellular proteins as the individual computational units of the cell. This kind of approach does not specifically involve modeling individual kinetic parameters, but rather concentrates on the resulting properties of the system as a whole. It is not currently possible to simulate all the detailed interactions inside cells, primarily because they are not fully understood. This leads us to the idea that to model a cell it is necessary to extract some essential features and base the simulation on general properties of interaction of these signaling molecules. This kind of a model is generally referred to as a "connectionist" model in the literature.

Barabasi *et al.* [74] have commented that techniques in science to describe properties of large connected networks of non-identical components are extremely

Table 1.2: Similarities between neural networks and cell signaling networks.

	Structural Features	*Neural Networks*	*Cell Signaling Networks*
(i)	node	a neuron	a molecular species
(ii)	dependency of nodes on input	a weighted sum of inputs from other neurons is computed	the concentration of the molecular species is a function of regulating molecules and their binding affinities
(iii)	output of the node	binary 0/1, (neuron fires or does not fire), depending on whether weighted sum of inputs exceeds a threshold	graded enzymatic activity of the molecular species is a complex (often sigmoidal) function of its inputs
(iv)	weights of connections	represent the strength of synaptic connections between neurons	represent cytoplasmic factors that affect enzymatic activity e.g., temperature, pH, and ATP concentration
(v)	modification of connectivity	occurs via a training rule which modifies the weights	occurs via evolutionary changes e.g., via random mutation events, fittest networks survive

inadequate. However, they also suggest that such networks often exhibit a high degree of self-organization. A complex, parallel network of intracellular signaling molecules may possess these emergent properties [41]. These are properties which cannot easily be predicted by looking at the individual elements themselves. Bray was the first biologist to suggest that cells might exhibit such features. He suggested that a parallel processing network, described by a connection matrix which identifies the interactions between the molecules in the system, might be an appropriate approach to model a cellular network. In fact, he also suggested that modeling the cell as a discrete dynamical system would be one way to approach the problem. His seminal work has opened the door for a different kind of modeling in cell biology.

1.7 Desirable Properties of Cellular Networks

Biochemical networks at the cellular level are extremely complex. It is natural to ask whether such complicated networks can function in a stable manner. In fact, it is important to ask questions such as whether these networks are relatively stable to small perturbations. For example, are specific reaction rate constants and enzyme concentrations finely tuned in order for optimal performance of the cell, or is the behavior of the network as a whole relatively insensitive to the precise details of the individual components? It has been argued that in order for cells to function properly, key properties of the system must be fairly robust [75]. Many questions remain unanswered as to why cell signaling networks are so complicated. Are there particular advantages of a large distributed system that cannot be obtained by small networks with only a small number of different molecular components?

Moreover, is the system as a whole stable to structural perturbations? Does the system fail when a single component is corrupted, as can be achieved nowadays by gene knockout experiments [39]? The system may well have a degree of redundancy so that elimination of individual units in a distributed network will not have a catastrophic result. To survive in an extracellular environment, the cell must be stable, but not so much so that it is unable to adapt to changes in the external world. However, it must also not be so fragile that small internal changes causes the network to collapse. In other words, the cell should display a degree of homeostasis. Table 1.3 illustrates how some of the properties of cells discussed in this section may be translated into the terminology of dynamical systems. This framework will be useful when reading subsequent chapters where other dynamical models are discussed.

Bray has suggested that the following properties of cellular networks would be desirable [48]. Cells should be able to recognize and react to environmental cues; they should also be robust and resistant to damage. He also suggests that cells may have

Table 1.3: Cellular properties in the language of dynamical systems. It is suggested in this thesis that a stable cell behavior may be thought of as a stable attractor state in the intracellular signaling system.

Cellular Properties	*Dynamical Systems Description*
Relatively small number of persistent cell behaviors	Relatively small number of stable attractor states
Homeostasis (stability to small perturbations)	Sizeable basins of attraction
Adaptable (to environment)	Ability to switch between different attractors via changes in input

the capability of encoding features of the environment. In fact, during evolution, cells should be amenable to adaptation via mutation and natural selection.

A feature of cell signaling pathways, and thus of cell signaling networks of which these pathways form a part, is that they are known to produce signal amplification. The cell can translate a small change in concentration of a chemical in its surrounding environment into a specific response such as the breakdown of glycogen. Cells need to be able to respond to relative changes in the extracellular environment. Often only a limited degree of amplification is required. The cell also needs mechanisms to shut off these signals. One may speculate that this is why cells have developed dynamic positive and negative regulatory mechanisms such as protein phosphorylation and dephosphorylation.

The level of protein phosphorylation in the steady state will also have considerable impact on levels of ATP consumption. In fact, it has been suggested that the cell signaling system would conserve energy if molecules remained primarily in their phosphorylated or dephosphorylated states [59]. In the literature, it has been noted that sometimes the formation and hydrolysis of ATP occurs continuously

and yet does not seem to perform any obvious functional role. These apparently futile cycles of ATP may play a role in maintaining cell stability by controlling phosphorylation states of important intracellular proteins.

One may also speculate that the cell signaling system may have evolved in such a way as to minimize the metabolic cost of information. In a study of the metabolic cost of neural information Laughlin and co-workers [76] have suggested that in noise-limited signaling systems, "a weak pathway of low capacity transmits information more economically, which promotes the distribution of information among multiple pathways". They estimated that a chemical synapse uses approximately 10^4 ATP molecules to transmit a single bit of information. Information processing in cells is therefore likely to have significant impact on its utilization of energy.

1.8 Discrete Dynamical System Models

Suppose a researcher were to think of constructing a discrete dynamical systems model for a single cell as Bray has suggested. A neural network model has been shown to be inadequate for such a task. So, how might one set about actually doing it? This section will be concerned with information that will help in answering this question, by first posing some more fundamental ones:

- How can the desirable properties of cells be translated into the language of dynamical systems?
- Can simple discrete, dynamical models generate complex behavior?
- How complicated would the dynamics of an n-dimensional discrete model be?
- What types of discrete models have typically been studied in the biological field?

When studying dynamical systems, there are general properties of the system in question that one would like to investigate. For example, one goal might be to predict the large-time behavior of the system and to determine all the possible states that the system can occupy. Given a set of final states of the system, it would also be

desirable to know the initial conditions that led to these final states. If the system is restricted in its phase space somehow, one would like to be able to analyze why and whether these sets of states are stable with respect to small perturbations.

The desirable properties of cellular networks discussed in the previous section can be described within the mathematical framework of dynamical systems. Homeostasis is a stability to small perturbations, both structural and temporal. In view of the fact that cells exhibit only a small number of different behaviors, it is possible that the states that a cell can occupy at any given time are somehow restricted. Stable states can be thought of as stable dynamic attractors, either fixed points or limit cycles. Each stable state will have a basin of attraction associated with it which is related to the homeostatic property of cellular networks discussed in the previous section. The response of any dynamical system to a given stimulus (initial condition) will be related to its current state and therefore the context in which a signal is received will be important. Changes in some nodes may cause a transition from one attractor to another. The ability to switch from one attractor to another has been demonstrated for a number of connectionist networks [72, 77]. These network models will be discussed in some detail in §1.8.2 and §1.8.3. If certain stable states can exist independently of extracellular signals then these patterns may determine how the cell senses the extracellular cues. In this case it may be inappropriate to view cell signaling as a problem of understanding how a series of proteins respond to an extracellular stimulus. It may be more useful to view extracellular stimuli as cues that shift the cell from one intrinsic dynamic attractor to another.

To return to the mathematics for a moment, and to answer the second question posed above, a non-linear model will be discussed here. Linear systems are more straightforward to study, but biological phenomena often exhibit non-linear characteristics. Mathematically speaking, even a simple one-dimensional, non-linear map can give rise to extremely complicated behavior. The logistic map, described

by Equation 1.1, is known to display chaotic behavior in certain parameter regimes. In fact the behavior of the system is critically dependent on a single parameter, ρ:

$$x_{n+1} = \rho x_n(1 - x_n). \tag{1.1}$$

A map in a single dimension, controlled by a single parameter, is also easy to visualize graphically. Cob-web diagrams [78, 79] and bifurcation diagrams can be drawn to explore the properties of the mapping.

In the case of a one-dimensional map, the asymptotic stability of fixed points may be determined by studying the first derivative of the map. Consider the mapping, $x_{n+1} = f(x_n)$, $x_n \in [0,1]\ \forall n$, where f is differentiable at all points of the domain. A fixed point, x_0, satisfies the relation $f(x_0) = x_0$. If $|f'(x_0)| < 1$ then the fixed point is said to be asymptotically stable. If $|f'(x_0)| > 1$ then the fixed point is said to be asymptotically unstable. Similarly, suppose that x_0 were a k-periodic point of f, then $f^k(x_0) = x_0$, where $k \in \mathbb{N}$ and $f^n(x_0) \neq x_0$ for $n = 1, 2, \ldots, k-1$. The asymptotic stability of the k-periodic orbit can be studied by examining the derivative of the new map $F(x) = f^k(x)$. One can also construct and study one-dimensional maps where the function f is non-differentiable. For example, the tent map is differentiable everywhere except at a single point.

$$x_{n+1} = f(x_n) = \begin{cases} 2\mu x, & 0 \leq x \leq 1/2, \\ 2\mu(1-x), & 1/2 \leq x \leq 1. \end{cases} \tag{1.2}$$

The parameter μ lies in the range $(0,1]$. Here the function f is a map $f : [0,1] \to [0,1]$. This map exhibits attracting sets for certain values of the parameter μ (see Appendix E for the definition of an attracting set).

To return to the question of whether simple models can generate complex behavior, it is thus not necessary to look any further than a simple one-dimensional, non-linear, discrete dynamical system. What happens when such a system is studied in multiple dimensions? Does it become so complicated that we can make no

sense of it? These questions will be addressed in the next three sections in the light of three different types of discrete dynamical models. Each of these mathematical models has been extensively studied and has its own applications in the field of biology.

1.8.1 Cellular Automata

Complexity is everywhere in nature, even in a tiny object like a single bacterial cell. One might imagine that biological systems may be impossible to characterize and describe mathematically. What is the nature of the complexity? How simple could a model be and still generate the complexity that mimics the real world of biology?

Von Neumann was fascinated by these very questions. He was the first person to study a certain class of mathematical models known as "cellular automata" [80]. He proved that a system requires a certain minimum level of complexity in order to replicate. In 1970, Conway developed what is now one of the most famous cellular automaton models of all—"The Game of Life" [81, 82]. This automaton has since been shown to be equivalent to a "Turing machine". Turing had developed this idealized model which he showed had the potential for universal computation [83]. Turing's work essentially laid the foundations for the modern theory of computation and computability, but he too found his inspiration from the complexity of the natural world around him. Cellular automaton models have since been used to study biological pattern formation, bacterial growth, the immune system and have even been used in the field of population biology. (For a discussion, see [84].)

A cellular automaton model is a discrete non-linear dynamical system comprised of a lattice of finite-state automata at each site. Each site takes input from a local region of the lattice (from the k nearest neighbors). The state of any given site is a function of the state of its neighbors. Cellular automata models demonstrate that via local interactions, complex dynamical properties can emerge and the system can self-organize from random initial conditions. The dynamics of cellular automata can be

Table 1.4: Qualitative properties of cellular automaton models. There are four classes of cellular automaton behavior. Conway's "Game of Life" is a famous example of a cellular automaton model.

Category	*Cellular Automaton Description*	*Dynamical Systems Description*
Class I	evolution to a spatially homogeneous fixed state	evolution to fixed points
Class II	evolution to a sequence of simple or periodic structures	evolution to limit cycles
Class III	chaotic, aperiodic behavior	chaotic behavior associated with strange attractors
Class IV	complex localized structures, some propagating	no direct analogy with theory of dynamical systems; long transients observed

observed qualitatively from computer simulations and can be classified into four broad categories [85] (see Table 1.4). Their dynamics can also be studied quantitatively [86].

1.8.2 Random Boolean Networks

The first connectionist study of a network of components inside a cell was performed by Kauffman [87]. He studied the behavior of large, randomly connected networks of binary "genes" called "Boolean networks". His results suggested that even randomly connected networks can exhibit a degree of stability.

A random Boolean network is a complex, parallel network. Genes regulate one another inside this network according to rules of interconnection described by random Boolean functions. The emergent global behavior of the network is controlled by the local features of gene interaction. Each gene is modeled as a binary device in that it can only take one of two possible values. The output of a gene at time $t + 1$ is

determined by applying a Boolean function to its inputs and therefore depends only on its activity at the previous time step, since the random Boolean function is fixed once it has been chosen. The state of the network is described by a vector which lists the current values (0 or 1) of each gene in the network. The current state of the network determines the next state and so forth; time evolves in discrete steps. A genetic network is constructed by selecting the number of genes in the network (N) and the number of inputs to each gene (K). Each node (gene) is then assigned a random Boolean function. From this moment on, the network evolves state by state according to a deterministic rule. There are a finite number of possible states that the network can enter (2^N in total). As a direct consequence of this, the network must eventually re-enter a previous state. The networks modeled by Kauffman do not use external inputs.

The system is investigated by randomly assigning Boolean functions and initial random values for each gene in the network, varying the network parameters N and K, and observing the resulting behavior. In particular, Kauffman calculated average network properties and examined the effect of small perturbations on the system. The latter fall into two categories—structural and minimal perturbations. The emergent network features vary from chaos to stability depending on the choice of K.

Kauffman's 1969 paper on "Homeostasis and Differentiation in Random Genetic Control Networks" suggested that his randomly constructed networks were in fact dynamically stable and exhibited the kinds of behavior expected from a complex, parallel network of interacting genes. His simulations used networks of 1000 elements with randomly assigned binary inputs and an update rule where the present state of the network determined the next state according to a Boolean logic rule. The many biological parameters that he needed for his networks were not known at that time. He overcame this problem by using statistical parameters and investigating the biological plausibility of the resulting networks.

In particular, when $K = 2$ the networks exhibit spontaneous cyclic behavior through small numbers of states and order emerges from random initial conditions. This is a direct consequence of the fact that networks have only a finite number of states. The systems also possess stable dynamic attractors. These are states which are stable to small perturbations. In other words, networks with different starting conditions sometimes reach the same cycle. These cycles were interpreted as temporal modes of behavior of the network.

A network satisfying $K = N$ displays maximally disordered, chaotic behavior with a high degree of sensitivity to changes in initial conditions. There are a small number of cycles ($\approx N/e$) with an average cycle length of approximately $2^{N/2}$. When $K = 2$, networks exhibit spontaneous collective order with minimal sensitivity. There are roughly $\sqrt{N}$ cycles with an average cycle length of $\sqrt{N}$. When $K = 1$, networks break into isolated loops. Thus, a fully connected network with $K = N = 200$ would have 2^{200} different states and an average cycle length of $2^{100} \approx 10^{30}$. At one microsecond per transition, it would take longer than the known age of the universe to traverse all the states of the attractor. However, a network of 100,000 nodes with only two inputs per component ($K = 2, N = 100,000$), has $2^{100,000}$ different states but only possesses around 370 state cycles. In other words, the actual number of states (or stable attractors) that the system occupies is much smaller than the total number of possible states.

Kauffman then asked the question of how realistic and relevant the simulations were. Typically, in the genetic networks of living things there are perhaps no more than about ten interactions per gene. Gene interactions can be described using canalizing functions which are a subclass of Boolean functions. A canalizing function is one where a change in at least one input can guarantee a monotonic change in the output function. He compared an *attractor* with a *cell type* and the number of nodes in the network with the number of genes in an organism. Since he could predict the

number of attractors in his networks, it meant that he could estimate the number of cell types in an organism. It also meant that he could predict cell cycle lengths as the length of the state cycle attractors. The cycle lengths of bacterial cells and other organisms appear to fit his predictions. In a similar way, the stability of cells can be related to the stability of the attractors in the networks. Cell differentiation can be explained as a process which carries a cell into the basin of attraction of another cell type. Kauffman also assumed that the number of genes is proportional to the amount of DNA in a cell, enabling him to estimate the number of genes and cell types in humans. These predictions are also in the right range. Mainstream biology has, however, for the most part ignored his results.

1.8.3 "Connectionist" Protein Signaling Networks

Very few connectionist models of cell signaling networks have been developed. One approach to studying biological regulation networks is to consider a model which has a sufficient level of abstraction to make it biologically plausible and yet computationally tractable. Chiva and Tarroux [77] have utilized some of the ideas suggested by Bray and have developed a connectionist model of a biological regulation network. A genome is used to encode the architecture of the network and the individual nodes of the network represent proteins with a continuous level of activity. Their model is deterministic, recurrent and synchronous. The activity of each node in the network describes the concentration of each protein. The activation of each node is updated by computing the weighted sum of inputs into that node. This assumes that protein interactions are additive and independent. The signaling network is represented by a matrix of weights W_{ij}, the activity of the network, at a given time step t, is represented by a vector $v_i^{(t)}$ and the state of the network is described by another vector $x_i^{(t)} \in [0, 1]$. The network is updated with the following rule:

$$v_i^{(t+1)} = W_{ij}\, x_j^{(t)}, \quad \text{where } 1 \leq i, j \leq N \text{ and } |W_{ij}| < 1, \tag{1.3}$$

and

$$x_i^{(t)} = F\left(v_i^{(t)}\right) = \frac{1}{1 + e^{-v_i^{(t)}/T}}, \tag{1.4}$$

where T is the mean interactivity level between units.

Each node i has a genome vector, G_i, associated with it. The genome for each protein determines its conformational structure which, in turn, determines the interactions of that protein (see Figure 1.11). Each weight element W_{ij} is determined by how well the conformational structures of the two proteins i and j match. This is achieved by performing correlations between the part of the protein j that can be regulated and the part of the protein i that tries to regulate it.

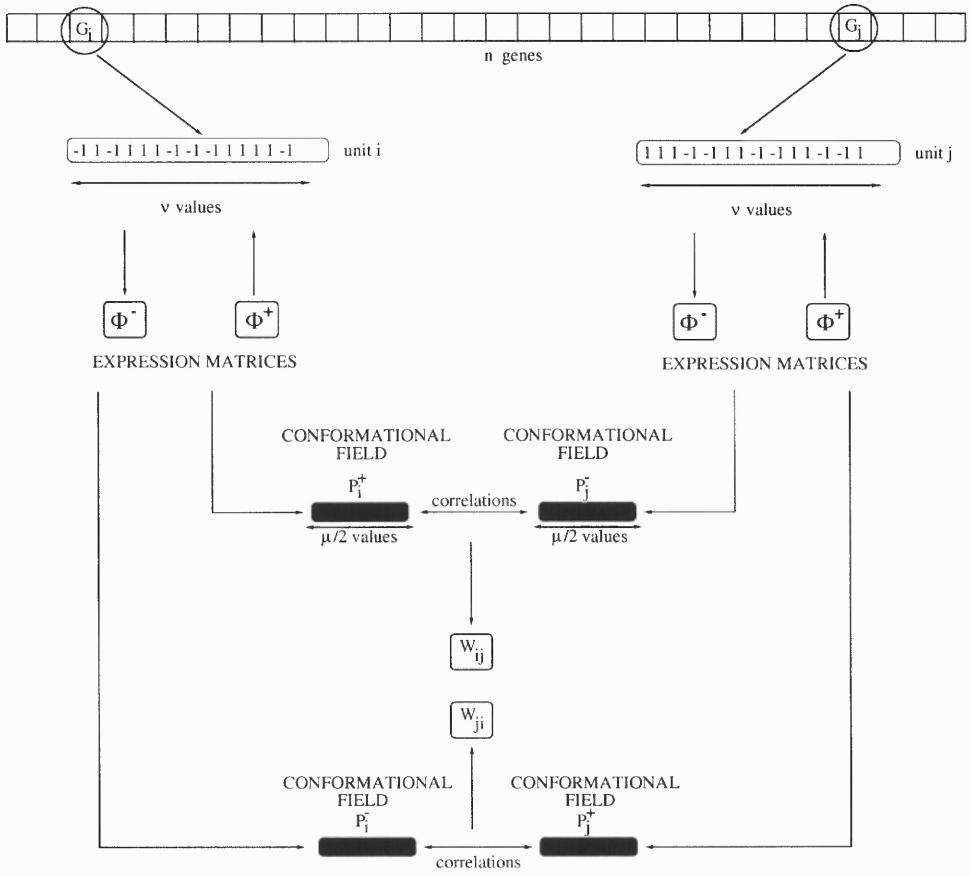


Figure 1.11: Simplified diagram of the Chiva & Tarroux model (adapted from Ref. [77]).

Networks are evolved using a "Genetic Algorithm", (GA), which is used to mimic evolution. The genetic algorithm allows the network to change over time. The goal was to develop an algorithm to allow the regulation networks to adapt to environmental cues. The resulting efficiency of the system as well as the resulting structural changes in the architecture were then discussed.

A population of cells is altered by selecting a number of "parents" at random and allowing crossover between the chromosomes of the parents. Offspring only replace parents when they have a higher fitness than the worst cell in the population. Mutation events can also occur randomly during each generation of the GA. The fitness function is designed to select networks that are robust to short perturbations and adapt to long lasting perturbations. In other words, fitness is a measure of adaptability to different environmental contexts. Two mechanisms are also allowed to increase the genetic material from time to time. This is achieved by firstly adding a new gene at random and secondly allowing duplication of previously existing genetic material.

The resulting simulations illustrate several interesting properties. The first observation is that the weight matrix develops a hierarchical structure. Units emerge where the size of the weights on the input connections of some of the nodes are much lower than the weights of their output connections. These are described as canalizing units. They also identified a phase transition parameter, $T = T_c$ below which the networks could not sustain limit cycle behavior, where T is the mean interactivity parameter and controls the steepness of the sigmoidal fitness function. At low interactivity, random networks exhibit a high degree of homeostasis with a small number of attractors. The networks always evolve to a small number of stable states, irrespective of initial conditions. At higher interactivity parameters, networks exhibit a larger number of attractors and are sensitive to perturbations in the extracellular environment. Networks with a smaller number of nodes exhibited

less varied behavioral modes. Also implementing the GA with only one of the two operators, mutation and crossover, limited the fitness of the population.

The Chiva and Tarroux model cleverly uses some important aspects of the biology. Specifically, they have used a genetic algorithm approach which mimics properties of evolution and have based their network model on the fact that the manner in which proteins interact depends on their conformational structure. The protein structure, in turn, is controlled by a modeled genome. However, there are also several drawbacks of their approach. The values at the nodes are used to represent the concentration of proteins which occur in discrete quantities, but yet the representation used is continuous. The weight matrix for their model signaling networks is fully connected, whereas real signaling networks will tend to be more sparsely connected. Networks of only 20 to 30 nodes were examined whereas in a real cell there may be somewhere between 100 and 1000 components or more.

1.9 Cellular Observables

There are thus several different types of dynamical models and approaches which one might try to use to model cellular processes. There is the highly detailed approach which would involve the development of systems of differential equations based on rate constants and known chemical reactions. In contrast, there is the more abstract connectionist approach which relies on extracting some fundamental features of the biology, which hopefully captures the most essential properties of the system under study. In either case, the model should provide some new insight into intracellular events that have already been observed but not yet satisfactorily explained, or alternatively should suggest new experimental paradigms or make testable predictions. In this section the following questions will be explored:

- What types of features of cells can currently be studied by biologists in real-time?
- What kinds of predictions can be made by connectionist models of intracellular signaling?
- Are there any important problems in biology to which such a model could be applied?

Freely diffusible second messengers, such as calcium, can already be studied in real-time. Features of calcium signaling were discussed briefly in §1.3. Techniques are likely to become available in the near future to study fluctuations in levels of ATP inside cells. The gradual development of cellular techniques (for example, the development of fluorescent markers) which enable monitoring of the interactions of proteins in real-time, and in intact living cells, will eventually enable researchers to monitor the dynamics of intracellular signaling both spatially and temporally. This will give crucial insight into signal transduction. In fact, review by Hunter [53] indicates that it is becoming possible to monitor, in real-time, the movement of protein kinases inside living cells.

A generic or connectionist model of intracellular signaling would never hope to reproduce the complete range of intracellular protein interactions nor would it expect to predict how any particular cell performs specified tasks. The questions addressed concern predictions about observables that are the averaged and conserved features of cells. For example, is there a constant and optimal ratio of kinases to phosphatases in a cell? What is the typical degree of divergence in a signaling pathway? On average how many target substrates does a kinase or a phosphatase have? What is the probability of disrupting a cellular function by a gene knockout experiment? How many distinct behaviors can one cell type display? How many of the cell's signaling proteins show oscillations in activity and what is the range of periodicities? It could thus provide a different kind of insight into properties of cell signaling networks to those of more explicit models.

Protein phosphorylation plays a fundamental role in processes in the central nervous system. In particular, it is thought to be important in changes in communication between neurons in the brain. Long term potentiation (LTP) is the long lasting increase in strength of synaptic transmission following certain types of high frequency stimulation of neurons. In the hippocampus this appears to involve certain receptor types (NMDA receptors) and certain protein kinases (e.g., CAMKII, PKC). LTP is a specific type of synaptic plasticity thought to underlie all constructs of learning and memory, but the cellular mechanisms responsible for these changes are not well understood. The mechanisms may be presynaptic, postsynaptic or involve changes in both neurons. There are several phases of LTP, but the early phase of LTP (1 hour) does not involve changes in gene expression. A generic model may be able to be used to predict the types of intracellular interactions that may be important in such processes.

1.10 Research Purpose and Design

Cells can undergo a variety of changes in activity during their lifetime. These activities may include aerobic or anaerobic metabolism, a mitotic or meiotic cell cycle, crawling movements, differentiation or apoptosis [13]. Although not precisely defined, these types of activity are generally regarded as examples of different modes of cell behavior. The switching between these kinds of behaviors can be triggered by external stimuli, such as hormones in multicellular organisms, or can occur autonomously, as in some developing embryos.

A cell is an inherently parallel structure like the brain. The brain can be thought of as a complex interwoven structure made up of components known as neurons. The cell can be viewed as a network of signaling molecules in much the same way. Proteins control the cell's response to its environment, resulting in the management of cell reproduction, cell differentiation, cell metabolism and cell stability. In the attempt to

understand how cells work, it is therefore natural to turn to a model based solely on these signaling molecules. Although many different kinds of proteins exist inside the cell, kinases and phosphatases can be identified as playing a particularly important role in the cell's control mechanism. Each protein has a highly complex structure and this structure dictates how it interacts with other molecules. Each protein regulates many others forming a complicated network inside the cell.

The rapidly growing data on cell signaling pathways make it difficult to envisage how a model, which includes all this detail, could provide any insight into cell function. It has, however, been assumed that a complete knowledge of all the molecular details of interactions are a prerequisite to constructing any model of how all these different cascades interact [43]. In any case, constructing such a highly detailed model would be daunting and may also provide no additional insight into how such a complex system works [71]. It may be many years before this information is complete. Other approaches are needed to complement existing techniques if researchers are to gain insight into how cells function and behave.

Some features of cell signaling may not require a full list of details. It has been suggested that the more traditional, reductionist approach of studying individual components and cascades in detail may not be adequate [39]. A study of individual components of a cell may miss essential features of the system as a whole. Rather than trying to model a cell in all its detail, it may be more expedient to try to model the system at a greater level of abstraction.

The goal of this project has been to develop a generic discrete dynamical system model of the interactions between protein kinases and protein phosphatases. In particular, the emphasis has been on examining dynamic properties on short term time scales that can occur independently of gene expression. This is a reasonable assumption since these proteins are known to be some of the most important intracellular signaling components of cells. The model is general and holistic in that it

attempts to represent cell behavior in terms of the overall activity of all the signaling proteins in the cell. In contrast to simulating a few proteins in detail, in our model there as many as one hundred highly idealized protein kinase and protein phosphatase components. The model is designed with a view to building in more detail in later versions.

This type of modeling is similar to that of recurrent neural networks designed to provide insight into the functioning of the brain [73], or of Boolean networks designed to understand homeostasis in gene regulation networks [72, 87]. The model concentrates on the rapid time scale ($\sim$ 10 mins) epigenetic information processing ability of cells that is to some extent independent of the genome and predominantly controlled by protein phosphorylation networks [48].

A discrete dynamical system model is developed to model the intracellular dynamics. Monte Carlo methods and a Genetic Algorithm are used to explore the state space of simulated networks. These simulations are intended to investigate whether simplified random networks of signaling components can display limited numbers of dynamic attractor states (in a manner analogous to the studies of Kauffman on genetic regulatory circuits). The purpose has not been to represent any particular cell type but to examine how the general features of signaling proteins may generate the limited range of behaviors that are seen in living cells.

The dynamical states of the system are explored both computationally and analytically. The hypothesis is that the attractor states of the intracellular signaling system may be the biophysical correlate of the behavioral mode of the cell. The activity of the signaling proteins in the cytoplasm of cells may be organized around a limited number of high dimensional dynamic attractors. It is suggested here that the stability of the attractors may provide a framework for understanding why a cell may persist in forms of cell behavior such as crawling. Moreover, since cell behavior is a rather poorly defined concept, ascribing the different attractors in the cell signaling

network to different modes of cellular behavior may actually help introduce it as a more precise concept. A shift in cell behavior may correspond to a change in the attractor dynamics (i.e., a switch from one attractor to another). This may provide a framework to help explain why:

- there are only a finite, limited number of stable cell behaviors,
- cells display sensitivity to some external signals but not to others,
- gene knockout experiments do not always abolish the desired behavior.

The ultimate goal of such a project is to make general statistical predictions about the range of network parameters that is possible for building any functional cell signaling network. Once the complete genome of a number of eukaryotic organisms, and the expression pattern in specific cell types are known, it may be possible to verify the ratio and specificity of PPs to PKs in a real cell. From these data we could judge the extent to which evolution has moved the range of different types of cells in nature away from the average features of all possible cell types. Statistical predictions might also be made to see how often the elimination of a component protein type disrupts the dynamics of an attractor. This could offer an insight into why some gene knockout experiments that eliminate signaling proteins appear to have very little effect upon cell function [29]. The overall frequency and potency of disruptions of cell behavior by such gene knockouts may represent the kind of observable in cell biology that is only predictable by a generic model of intracellular signaling. It may also be possible that a network with a large number of signaling elements with a relatively low level of connectivity can provide cellular networks with the stability that they need.

1.11 Thesis Outline

In this section, an outline of the thesis layout is presented. Chapter 2 develops a discrete dynamical systems model based on salient features of signal transduction networks. The model concentrates on control of proteins by phosphorylation. The assumptions and limitations of this modeling approach are presented.

The numerical and analytical techniques that are used to study the model are outlined in subsequent chapters. In Chapter 3 a Monte Carlo simulation method is described and utilized to study dynamics of networks over large parameter regimes.

In Chapter 4 a Genetic Algorithm (GA) is designed and developed to study networks in a more localized parameter range. The GA is designed specifically to incorporate certain important evolutionary concepts drawn from the biology. The effects of systematically altering parameters of the algorithm are discussed. This chapter also compares the Monte Carlo approach with the genetic algorithm.

The dynamics of networks evolved by the GA are analyzed with the use of a clustering algorithm presented in Chapter 5. The hypothesis that cell signaling networks may possess a limited number of attractor states is discussed in light of these data.

Chapter 6 examines the dynamic properties of small networks with the aid of various mathematical and numerical tools. Specifically, proofs of local stability of fixed points and periodic limit cycles are given in certain specialized parameter cases. The relevance of analyzing approximations to the fully discretized model is argued.

Chapter 7 presents a short collection of insights that can be gained from studying topological networks via graph theoretic methods. The signal amplification properties observed in cascade models will be discussed in the context of this model. The structure of a network generated by the GA is also examined.

In Chapter 8, the results of these data are summarized. The flexibility of this model is discussed and ideas for future work are presented.

CHAPTER 2

DEVELOPING A CONNECTIONIST MODEL

2.1 Basic Assumptions

This chapter will be concerned with developing the discrete dynamical model. To construct the model a number of idealizations have been made concerning the interactions between signaling proteins. These are summarized below:

- Consider only interactions between PKs and PPs. Inputs to the system are not explicitly represented.
- The model of intracellular signaling is designed to concentrate on aspects of intracellular signaling that occur epigenetically (i.e., independently of the genome).
- As shown in Figure 1.2, a phosphorylation reaction requires a molecule of ATP. It is therefore assumed here that there is an unlimited supply of ATP available inside the cell. In fact, in an average cell it is estimated that approximately 10^9 molecules are available for use [13].
- The time scales of diffusion are taken to be small when compared to protein phosphorylation. This is a reasonable assumption for small cells (e.g., platelets) and for smaller cellular compartments or synaptic cleft regions.
- All regulatory processes are taken to occur on the same time scales, leading to a synchronous updating rule.
- No compartmentalization is included. The cell is essentially considered as a homogeneous stirred medium.
- The initial model considers only one regulatory domain per protein type.
- It is assumed that addition of phosphate groups activates proteins.
- It is also assumed that all proteins have equal affinity for substrate molecules and that all attempts to phosphorylate and dephosphorylate are independent.
- The concentration levels of all protein types are the same. This is controlled by the parameter o_{max}.

2.2 Model Development

The main focus of this study is on the control of protein activity by phosphorylation. This simplification seems reasonable since kinases and phosphatases make up the bulk of the different types of elements in the intracellular signaling network and are known to control many features of cell behavior [32, 45]. Some cell surface receptors also appear to be directly linked to kinases and phosphatases [52].

The functional structure of a protein kinase usually consists of separate units. The first is a unit with catalytic activity; the second unit (or more commonly a series of units) is referred to as a regulatory unit. The latter are so named because they can be regulated by other molecules. For example, they can be phosphorylated by other protein kinases. Protein kinases work by using the free ATP in the cytoplasm to attach a phosphate molecule to the regulatory site of another specified target protein (see Figure 2.1).

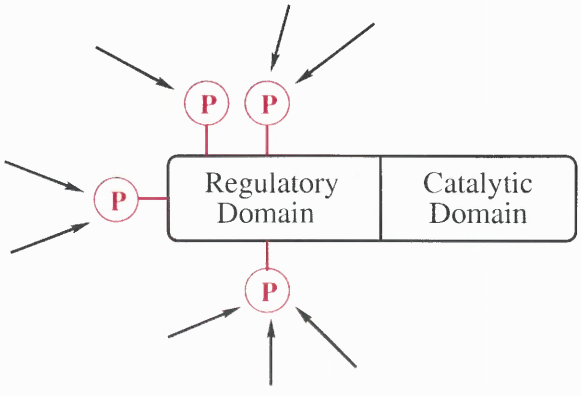


Figure 2.1: Schematic diagram of regulatory and catalytic domain structure and phosphorylation. In the initial model only one phosphorylation site per regulatory domain is considered.

Protein phosphatases have a similar functional organization with a single catalytic unit and separate regulatory units [44, 45]. The phosphatases bind to specific phosphorylated proteins removing the phosphate molecule from the regulatory site. The attachment of phosphate groups by protein kinases is assumed to be stable unless removed by specific protein phosphatases. It is the phosphorylation status of the one

or more regulatory units that determines the activity of the catalytic unit of each kinase or phosphatase. The specificity of the output for a kinase (i.e., the target it will phosphorylate) is generally determined by a region in the catalytic unit that recognizes a specific part of the regulatory units of other proteins. For some kinases and phosphatases there are targeting subunits that perform the same role [55].

The simplified generic network, therefore, consists of kinases and phosphatases that are cross-coupled to interact with one another (Figure 2.2). A particular collection of kinases and phosphatases and their interactions can be thought of as a particular cell type. The initial emphasis is on exploring the intrinsic dynamics of the phosphorylation system and not so much the specific tasks it performs. Inputs to and outputs from the system have not been explicitly represented.

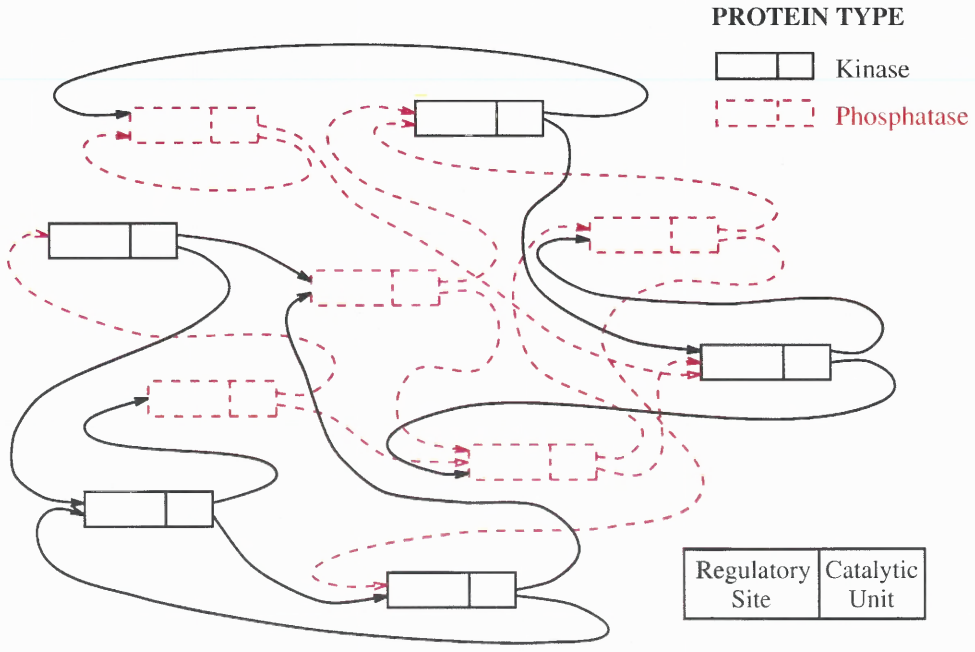


Figure 2.2: Schematic diagram of interactions between signaling proteins. Each box represents a particular type of signaling protein with kinases represented by solid boxes and phosphatases by dashed red line boxes. Each protein type consists of a regulatory site and a single catalytic subunit. The interactions between proteins are represented by lines emanating from the catalytic unit linking to target regulatory sites. Connections are randomly assigned with predefined probabilities. Each protein type can regulate a number of other types (kinase or phosphatase). Some proteins may be self-regulating.

The time for diffusion of proteins around the cell is imagined to be fast compared with the time for phosphorylation reactions. This is a reasonable assumption for small cells (e.g., platelets) or intracellular compartments. The diffusion coefficient of an average sized protein is estimated to be between 10^{-7} and 10^{-8} cm^2/s [88, 89, 90]. Using a diffusion constant, D, of 5×10^{-8} cm^2/s, the diffusion equation can be used to show that this corresponds to the protein molecule being found within a root-mean-square distance, $\bar{x}$, of 0.1μm from its initial position† in roughly $t_{mix} = 1$ ms ($\bar{x}^2 = 2Dt_{mix}$). The turnover rate (t_{turn}) for most enzymes operates in the interval of 10^{-3} to 10^{-2} seconds [88], but these can vary [91]. Turnover rates for protein kinases are more likely to be in the range 0.1 to 1 second. Therefore, in small cells or intracellular compartments reaction rates are determined by the turnover rate, rather than the diffusion rate ($t_{turn} \gg t_{mix}$) [88].

The exact manner in which the phosphorylation of each of the regulatory units affects the catalytic activity is not generally known. In this initial presentation complex interactions between regulatory units have not been investigated. All protein types have a single regulatory site. The activity of each type of kinase or phosphatase is simply determined by the occupancy of available phosphate sites.

The model of signaling proteins consists of a set of interacting nodes where each node represents a "type" of kinase or phosphatase. In the description of the dynamics of the model, the "activity" of a protein refers to the mean activity of all the molecules of that specific protein type across the cell rather than the behavior of a single molecule (see Figure 2.3). A connection matrix is used to represent the specific set of interactions between kinases and phosphatases in any given network.

Each of the modeled signaling proteins has a single regulatory unit and thus all of the interactions between proteins are represented in a single connection matrix M_{ij}. Non-zero elements in the position (i, j) of the matrix indicate that the catalytic site of

†The radius of an average eukaryotic cell varies from between 5 μm and 50 μm.

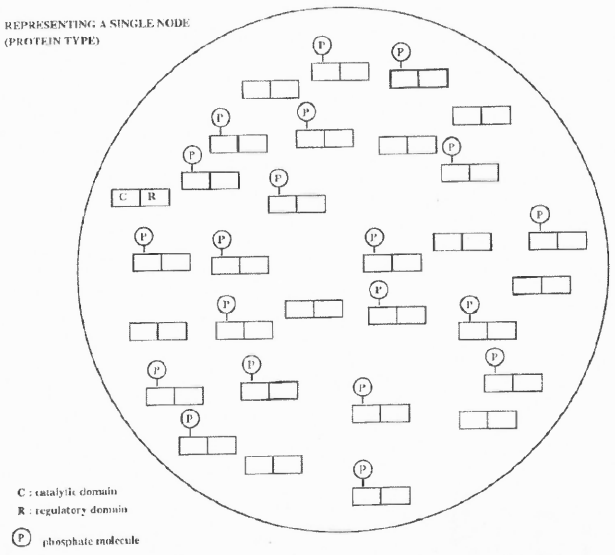


Figure 2.3: A single node in the network represents a pool of molecules. Not all of the molecules will be phosphorylated at any given time. The node illustrated in this figure is shown with thirty molecules, and is represented schematically by a *single* box in Figure 2.2. Here, 20 of the 30 molecules are phosphorylated (i.e., have a phosphate group attached). These are considered to be active molecules.

node (protein type) i interacts with the regulatory unit of node j. If i is a kinase, then the non-zero elements are $+1$, and if i is a phosphatase, the non-zero elements are -1. The nodes in the network are labelled in such a way that the upper rows of the matrix are always different protein kinases and lower rows are distinct phosphatases. Each instance of a network, with a specified set of nodes and a connection matrix determining how they interact, corresponds to an individual cell type. Different cell types (M_{ij}) can be generated using the following control parameters:

- total number of different types of signaling proteins: N
- the fraction of protein types that are kinases or phosphatases: f_k; $f_p = 1 - f_k$
- the probabilities of interaction of each type of kinase and phosphatase: p_k, p_h.

The set of interactions for each protein of each cell type is randomly selected using the probabilities p_k and p_h. Some proteins in the networks may have no regulatory input and others may be self-regulating (Figure 2.2). In any given network

every kinase has the same p_k value (similarly for phosphatases). Later simulations are presented where protein kinases and protein phosphatases do not have the same output specificity.

2.2.1 Simulating the Dynamics of a Single Cell Type

A deterministic rule, called the update rule, is implemented. This changes the activity of the proteins within any network over time as proteins influence one another. The activity of the catalytic unit and level of phosphorylation (phosphate occupancy) of the regulatory unit of any node i (at a specified instant in time t) are denoted by $a_i^{(t)}$ and $o_i^{(t)}$ respectively, where $1 \leq i \leq N$. The activity of each node represents the number of active molecules of protein type i. A molecule is activated when it is phosphorylated and inactivated when it has no phosphate group attached (i.e., it is in the dephosphorylated state). Since the activity and occupancy of each protein type represent information about the properties of a limited number of molecules, they are represented by integer values in the range $[0, o_{max}]$, where $o_{max} = 10,000$. A maximum activity of ten thousand was chosen as a plausible estimate of the maximum number of molecules of any protein type in a real cell. The occupancy of any protein type is the average number of molecules of that type whose regulatory units have become phosphorylated. See Figure 2.3. For example, if $o_i^{(t)} = 9,000$, then 90% of the molecules of type i are phosphorylated.

At each time step t, the change in occupancy of each node depends on its current occupancy, which other nodes interact with it, and their level of activity. If a node is currently highly occupied (i.e., many molecules are already phosphorylated) it is more difficult for interacting nodes to increase this level than if the level of phosphorylation were low. The level of activation of each node represents the instantaneous, average activity level of all the molecules of a particular type of kinase or phosphatase.

The dynamic behavior of each network is evaluated by a simulation based on a non-linear, synchronous, recursive update rule which determines the activities of the proteins at a series of discrete time steps. All allowed phosphorylation and dephosphorylation reactions occur between consecutive time steps. The iteration is repeated for a specified number of time steps, T. The time for diffusion of proteins within the cell is assumed to be negligible compared with the time for phosphorylation reactions. The occupancy level of each node at time $t + 1$ depends on its current occupancy, which other nodes interact with it, and their level of activity at time t.

The model, therefore, synchronizes all of the changes in binding and makes a simplification by having all of the regulatory processes occur on the same time scale. The update rule also has all signaling proteins with the same dynamic range of concentration and all attempted phosphorylations and dephosphorylations occur with equal probability (i.e., the substrate affinity of every kinase and phosphatase is the same).

2.2.2 Derivation of Dynamical Update Rule

Given an interaction matrix and the activity state of the network, it is possible to estimate whether on average there are more phosphorylation attempts on any node than dephosphorylation attempts. This is calculated through Equation 2.1. It is assumed that all phosphorylation or dephosphorylation attempts by kinases or phosphatases are independent, so that the number of attempts to bind a new phosphate to any node j, b_j, is calculated from the weighted sum of inputs into that node:

$$b_j = \sum_{i=1}^{N} a_i M_{ij}, \quad \text{where} \quad 1 \leq i, j \leq N. \tag{2.1}$$

It is also assumed that affinity for substrates is the same for each protein kinase and protein phosphatase (i.e., that all attempts to phosphorylate and dephosphorylate

have equal weighting). If b_j is greater than zero, then on average there are more kinase molecules influencing that node than there are phosphatase molecules trying to dephosphorylate molecules of type j. If, however, b_j is less than zero, then on average there are more phosphatases trying to dephosphorylate node j than there are kinases trying to attach phosphates. If the number of phosphorylation and dephosphorylation attempts exactly balance, then b_j will be zero. Therefore, given the current occupancy of node j and the proteins that interact with it,[†] the number of attempts to phosphorylate protein molecules of this type can be calculated. It is then possible to calculate the new occupancy of node j resulting from these interactions. From the computation of b_j, basic probability theory can be used to estimate how many of the binding attempts were successful, given the fact that a proportion of the protein molecules of type j were already phosphorylated. Not every attempt to bind a phosphate group will be successful. Suppose, then, that o_j molecules of node j are phosphorylated (out of a possible $o_{max} = 10,000$ molecules).

The probability of a phosphate group being added to a "free" molecule which is currently in a non-phosphorylated state is simply

$$\text{P[phosphorylating one "free" molecule]} = p_f = \frac{(o_{max} - o_j)}{o_{max}}.$$

Similarly, the probability of removing a phosphate group from a molecule which is currently phosphorylated is

$$\text{P[dephosphorylating a phosphorylated molecule]} = p_e = \frac{o_j}{o_{max}} = 1 - p_f.$$

Then the number of attempts needed, on average, to bind one new phosphate group is $1/p_f$ and the number of attempts needed to remove one phosphate group from the pool of molecules is $1/p_e$. The underlying assumption here is that phosphate

[†]These can be found by looking at the non-zero elements of column j of the interaction matrix M_{ij}.

groups are added or removed at random. In other words, it is assumed that all molecules exist in a homogeneous stirred medium and are free to diffuse at random. To bind one new phosphate group to a molecule in the pool requires, on average, $o_{max}/(o_{max} - o_j)$ attempts. After this many attempts, $o_j + 1$ molecules will now be occupied/phosphorylated. Therefore, the number of phosphorylation attempts required to bind two new phosphate groups, N_f, assuming that each attempt to bind is independent, is written as follows:

$$N_f = \frac{o_{max}}{o_{max} - o_j} + \frac{o_{max}}{o_{max} - (o_j + 1)}.$$

This formula is generalized to obtain the average number of attempts required to bind n new phosphate groups (i.e., to go from a total of o_j to $o_j + n$ phosphorylated molecules). This is given by Equation 2.2.

$$\sum_{k=0}^{n-1} \frac{o_{max}}{o_{max} - (o_j + k)}, \quad \text{where} \quad 0 \leq o_j < o_{max}, \quad 0 \leq n < o_{max} - o_j. \tag{2.2}$$

Similarly, the number of attempts required to dephosphorylate a single phosphorylated site is equal to o_{max}/o_j. Hence, the number of attempts required to dephosphorylate two molecules, N_e, is given by:

$$N_e = \frac{o_{max}}{o_j} + \frac{o_{max}}{o_j - 1}.$$

Thus, to remove n phosphate groups (i.e., to go from o_j to $o_j - n$ sites phosphorylated) requires a certain number of attempts (see Equation 2.3).

$$\sum_{k=0}^{n-1} \frac{o_{max}}{o_j - k}, \quad \text{where} \quad 0 \leq o_j < o_{max}, \quad 0 \leq n < o_j \tag{2.3}$$

Equations 2.2 and 2.3 are Riemann sums for the following integrals, where $o'_j = o_j \pm n$, implying $b'_j > 0$ when the overall effect of binding attempts is to attach new phosphate

groups and $b'_j < 0$ where the overall effect is to remove phosphate groups†, respectively (see Equation 2.4).

$$\begin{aligned} b'_j &= \sum_{k=0}^{n-1} \frac{o_{max}}{o_{max} - (o_j + k)}, \approx \int_{o_j}^{o'_j} \frac{o_{max}}{o_{max} - x}\, dx, \quad \text{where} \quad o'_j > o_j, \\ b'_j &= -\sum_{k=0}^{n-1} \frac{o_{max}}{o_j - k}, \approx -\int_{o'_j}^{o_j} \frac{o_{max}}{x}\, dx, \quad \text{where} \quad o'_j < o_j. \end{aligned} \tag{2.4}$$

On average, the number of attempts needed to bind a new phosphate molecule increases as o_j increases since it is harder to find an unphosphorylated molecule. Similarly, the number of attempts required to remove a phosphate increases as o_j decreases since there are more chances to bump into molecules which are already dephosphorylated.

The problem is as follows. If there are b_j attempts to attach or remove phosphates (as calculated from Equation 2.1) then how many molecules will be become phosphorylated as a result? The known quantities o_j, b_j are used to calculate the unknown quantity o'_j. Evaluating the integrals in Equation 2.4 gives (respectively):

$$b_j = \begin{cases} \displaystyle\int_{o_j}^{o'_j} \frac{o_{max}}{o_{max} - x}\, dx = o_{max} \log\left[\frac{o_{max} - o_j}{o_{max} - o'_j}\right], & \text{where } o'_j > o_j, \\ \displaystyle -\int_{o'_j}^{o_j} \frac{o_{max}}{x}\, dx = -o_{max} \log\left[\frac{o_j}{o'_j}\right], & \text{where } o'_j < o_j. \end{cases} \tag{2.5}$$

Each summation can be approximated by an integral because o_{max} is large. These relations can be equated with the expression for b_j given in Equation 2.1. Since b_j and o_j are known, o'_j can be calculated by integrating and rearranging the equations in (2.5). There are three cases to consider: (i) $b_j < 0$ (more attempts to dephosphorylate), (ii) $b_j > 0$ (more attempts to phosphorylate) and (iii) $b_j = 0$.

†In the latter case, the binding variable, b_j, is defined to be a negative quantity in order to distinguish between overall phosphorylation and dephosphorylation attempts which are computed from Equation A.

$$o_j^{(t+1)} = \lfloor f\left(o_j^{(t)}, b_j^{(t)}\right) \rfloor^{\dagger}$$

$$= \begin{cases} \lfloor o_{max} - \left(o_{max} - o_j^{(t)}\right) \exp\left(-\dfrac{b_j^{(t)}}{o_{max}}\right) \rfloor, & \text{if } b_j^{(t)} \geq 0, \\ \lfloor o_j^{(t)} \exp\left(\dfrac{b_j^{(t)}}{o_{max}}\right) \rfloor, & \text{if } b_j^{(t)} \leq 0. \end{cases} \tag{2.6}$$

The number of new sites filled at the next time step, $o_j^{(t+1)}$, is dependent on the number of new attempts to bind and the current number of phosphorylated molecules of that protein type. The saturation function f places upper and lower bounds on the occupancy level (or number of phosphorylated molecules) for each protein type since the occupancy level must be non-negative and the upper bound is a result of the finite concentration of any protein type in the cell (see Figure 2.4). Currently each node has the same o_{max} value.

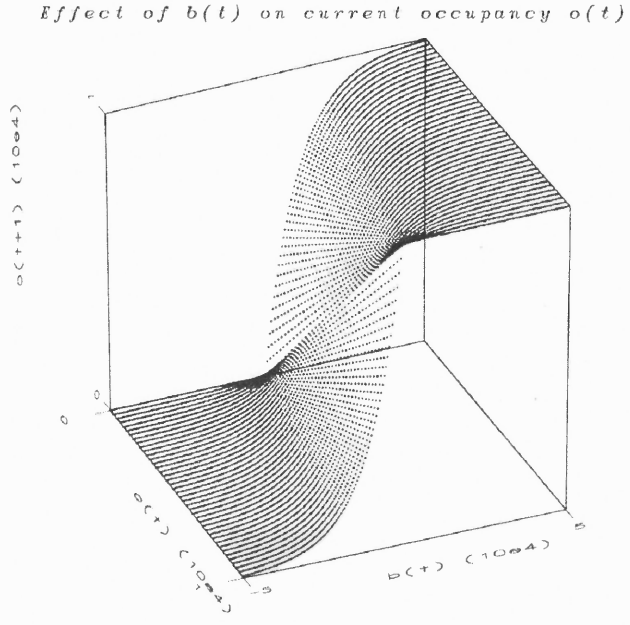


Figure 2.4: The updating function $f(o^{(t)}, b^{(t)})$.

[†]where $\lfloor x \rfloor$ denotes the "flooring" function (the greatest integer smaller than x) in Eqn. 2.6.

Note that the update rule implies that if a node has zero input, its level of phosphorylation remains constant. Also, the function f is continuous at $b_j^{(t)} = 0$ and the value of $o_j^{(t+1)}$ never goes below zero or above o_{max}. If $o_j^{(t)} = 0$ and $b_j^{(t)} < 0$, then $o_j^{(t+1)} = 0$, similarly if $o_j^{(t)} = o_{max}$ and $b_j^{(t)} > 0$ then $o_j^{(t+1)} = o_{max}$.

A function g determines the relationship between the activity level of a protein at time $t + 1$ and the occupancy of its regulatory site at the same time step. As an initial simplification, g is the identity function (Equation 2.7). At time t, therefore, the state of the network (cell) is described by an N-dimensional vector $\mathbf{o}^{(t)}$ from which the activity vector $\mathbf{a}^{(t)}$ is calculated:

$$a_j^{(t+1)} = g(o_j^{(t+1)}) = o_j^{(t+1)}, \quad 1 \leq j \leq N. \tag{2.7}$$

This iteration is repeated for a specified number of time steps, T. The symbols used in the model development are included in the "List of Symbols" section for reference purposes.

2.3 Summary

A number of modeling assumptions were listed at the start of this chapter. Despite these apparent restrictions, the model is flexible enough to be extended in a manner which allows several of these assumptions to be relaxed. This will be briefly touched upon here and discussed in more detail in the final chapter.

The model also incorporates several concepts that were introduced in the previous chapter. While intracellular signaling in biological cells utilizes many different types of molecules, the proteins are considered to be some of the most important computational elements of cells. A network of protein phosphorylation reactions is considered here. This simplification seems reasonable since kinases and phosphatases make up the bulk of the different types of elements in the intracellular signaling network and are known to control many features of cell behavior

[32, 45]. Some cell surface receptors also appear to be directly linked to kinases and phosphatases [52]. This was discussed earlier in §1.5. The effect of changing model parameters can be used to investigate the effect of altering the ratio of kinases to phosphatases. The current assumption of equal numbers of each type may not be unreasonable [53]. The model also incorporates the idea of cross-talk. This is captured by the connectivity parameters and the fact that the each modeled cell type is represented by a matrix of interactions.

An important feature of the model is that the new occupancy level depends both on the activity and occupancy values at prior time steps so the network shows a hysteresis effect. The model synchronizes all of the changes in binding, and implicitly assumes that all of the regulatory processes have the same time scales. The possibility of incorporating different time scales into the model will be discussed in Chapter 8.

The model also currently assumes that there is only one regulatory domain per protein. Using several sites would involve considerations of how protein domains interact. Biologically speaking, this relates back to the concept of multi-site phosphorylation which was discussed in the introductory chapter.

Another limitation of the current model is that all protein types currently exist at the same concentration levels. This assumption can be relaxed by careful consideration of the parameter o_{max}. This will be discussed further in the final chapter.

Second messenger molecules could be included by feeding additional nodes into the network to create signals that feed into the phosphorylation network. The model, as it stands, can be used to examine effects of changes in inputs to the network, if nodes with no inputs are considered to be external influences. Hence, changes in initial conditions of such a network mimic the effect of external signal perturbation. This will be seen in Chapter 5.

CHAPTER 3

MONTE CARLO SIMULATIONS

3.1 Introduction

In the previous chapter, a discrete dynamical model for interactions between signaling proteins is developed. The model is designed from a knowledge of the biology. The primary goal of the model is to extract some generic features from the biological system, and it is also intended to capture essential features of the intracellular signaling system. The basic model is considered, where there is no distinction between activity and occupancy of signaling proteins. The model remains sufficiently general to be extended and improved upon in later versions.

The Monte Carlo simulation method is outlined in §3.2. The motivation for this approach is that it is an effective way to sample a broad spectrum of different parameter cases. The simulations randomly sample the dynamics of modeled networks over the parameter space. It is important to understand whether even the simplest version of the model can generate rich dynamic behavior or whether the modeled networks simply exhibit highly unordered, chaotic features. Previous work on connectionist models by Kauffman [87], Hopfield [73], Wolfram [85] and Chiva and Tarroux [77] indicate that the latter properties would be unlikely. Nonetheless, the model properties must be thoroughly explored.

Data are presented in §3.3 from a systematic study of certain parameter regimes of the model via the Monte Carlo simulation approach. Specifically, it examines the effect of increasing the connectivity within simulated networks. Parameters are always maintained within a biologically plausible range. Resulting networks are classified into two different types depending on their dynamic properties. Some of the drawbacks with this classification will be discussed. A biological indicator of dynamic behavior is also suggested in §3.4.

3.2 Method

The Monte Carlo simulation method is outlined in this section. A set of C "modeled" cell types are constructed by generating a set of randomly constructed connection matrices constrained by the parameter values defined in §2.2. Fixed values for $N = 100$ and $f_k = f_h = 0.5$ are used. Connectivity parameter values are chosen in a biologically realistic range, varying $p_k = p_h$ from 0.01 to 0.10, incrementing in steps of 0.01. To examine the variety of behaviors of the modeled networks, each network is investigated from a series of S random initial starting states (see Figure 3.1).

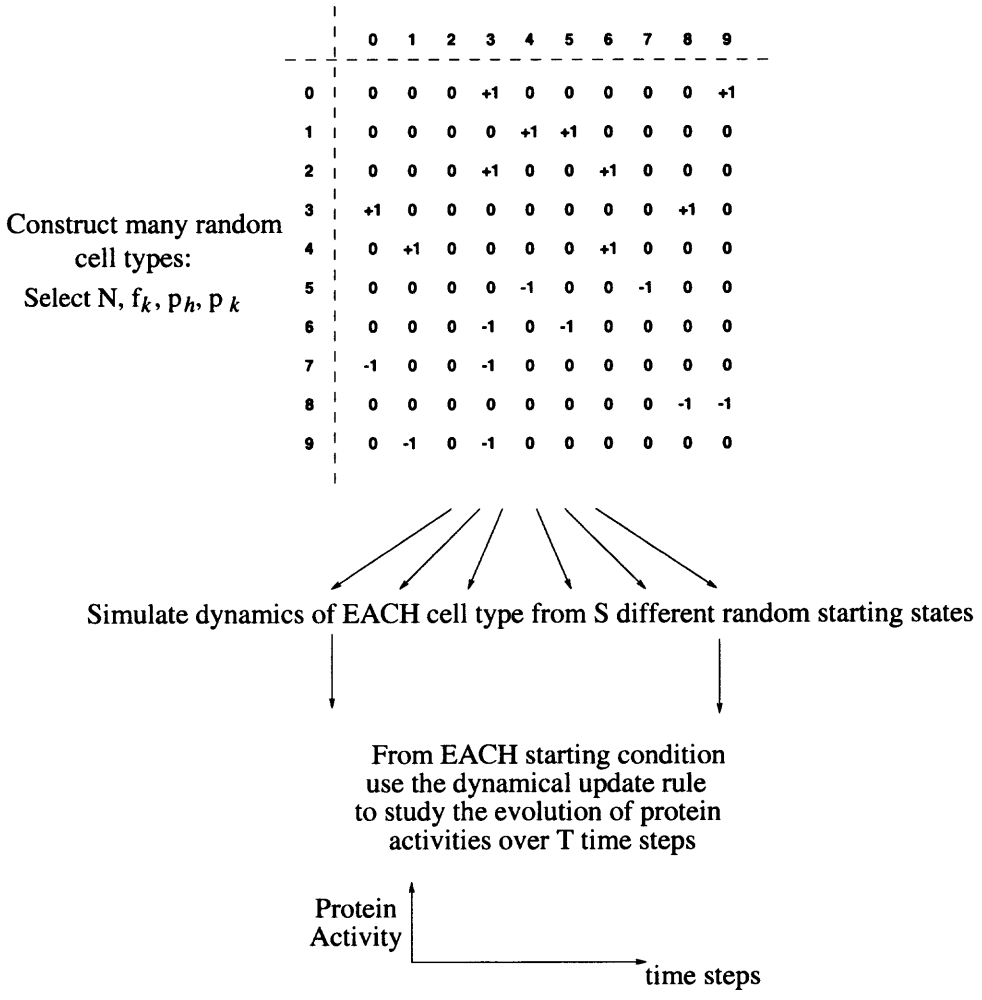


Figure 3.1: The Monte Carlo simulation method. The random cell type shown here has the following parameters: $N = 10$, $f_k = 0.5$, $p_h = p_k = 0.2$.

Typically, between one hundred and one thousand random networks and starting conditions are simulated. For low connectivity values, each network is simulated for T ranging from 100 to 500 time steps. In a model where the function g is the identity function, an occupancy vector $\mathbf{o}$ is sufficient to characterize the starting state of the network. The phosphorylation level of each protein type is set independently as a random integer, with $0 \leq o_j \leq o_{max}$, where $1 \leq j \leq N$. The dynamics of the proteins are then simulated over a series of T time steps using the update rule defined in §2.2.2. A summary of symbols used in the Monte Carlo simulation method can be found in the "List of Symbols".

3.3 Results

For each set of parameters, $[p_h, p_k, f_h, N]$, the behavior of one hundred randomly selected cell types are investigated, each from one hundred random initial conditions. A striking feature of these networks is that they settle down to states where many nodes do not change at all or change by less than 10%. This is seen at all values of $p_h = p_k$, but particularly at low probabilities $(0.01, 0.02, 0.03)$. Other observed features are that protein types within networks are seen to oscillate in and out of phase with one another at the same frequencies. This will be discussed in Chapter 6 in the light of more rigorous analysis of smaller networks.

Figure 3.2 broadly illustrates some of the general features of the simulations. Typical dynamic behaviors are stable fixed points and stable limit cycles. The network shown in this figure settles into two different stable configurations within 150 time steps from two different starting states. One configuration is a fixed point (A) and the other is an oscillation with a period of twenty-eight time steps (C).

Figures 3.3 and 3.4 show examples of protein activities from some sample networks with $p_h = p_k = 0.01$, 0.05 and 0.10. It can be seen that oscillations occur in the activity of both kinases and phosphatases even at low interaction

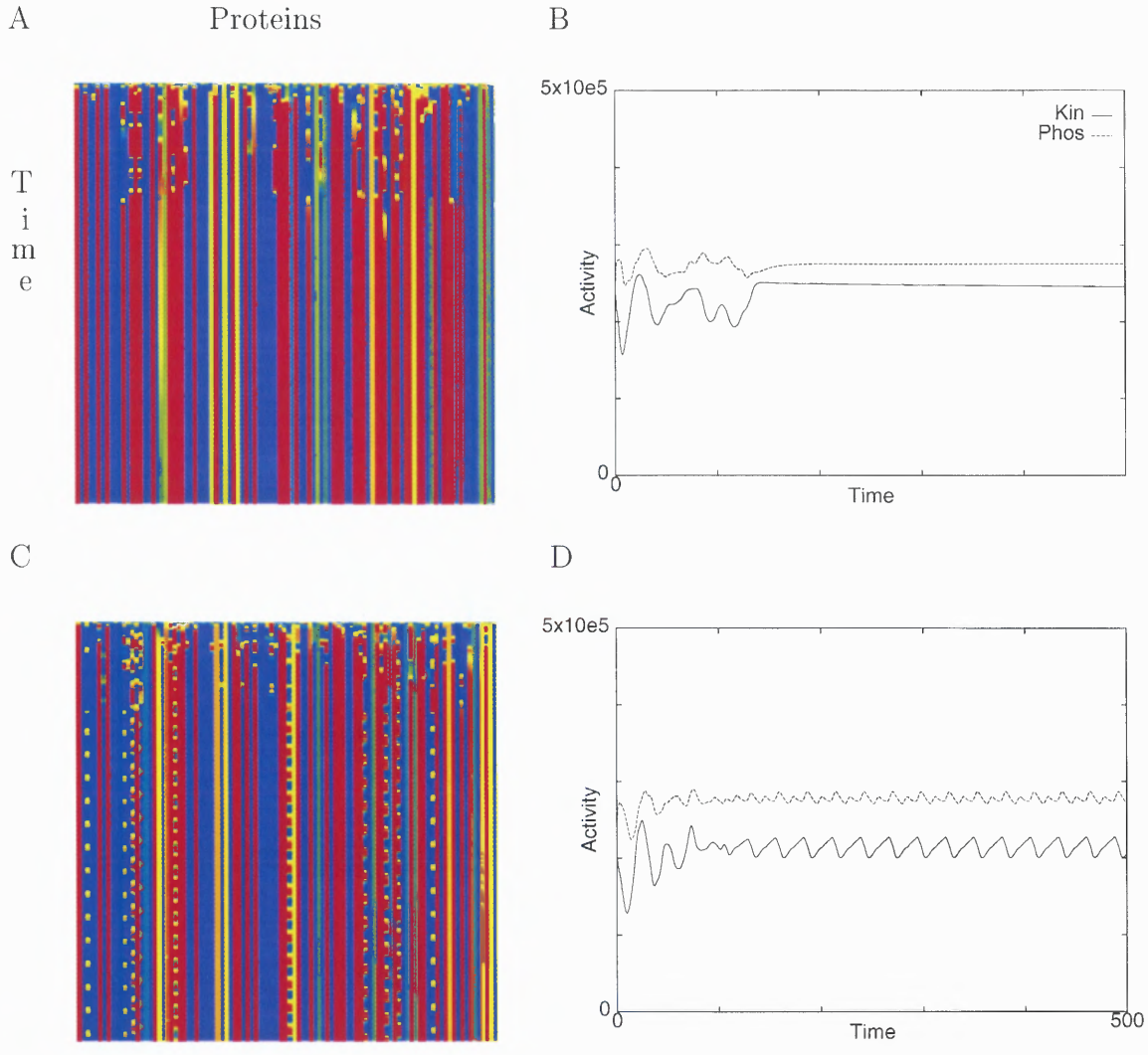


Figure 3.2: Sample dynamics from a single cell simulation from two different starting states. The cell parameters are $[p_h, p_k, f_h, T, N, S] = [0.03, 0.03, 0.5, 500, 100, 100]$. The left hand plots (A, C) illustrate the change in the activity state of the proteins in the network over time. Each row represents the activity of the one hundred different protein types in the network (one pixel represents one protein type), and each column represents the change in activity of the protein over time. Proteins with low activity values are shown in blue and a graded color scale is used with highly active proteins shown in red. The corresponding right hand plots show the same example, but illustrate how the sum of the activity of all the kinases (bold line) and phosphatases (dashed line) varies during the respective simulations. Plot A (and B) shows an example where the protein activities settle into a fixed point configuration. Plot C (and D) illustrate the same cell, but from a different starting configuration. The network settles into a limit cycle oscillation. In (A) and (B) every protein settles to a fixed activity value whereas in (C) and (D) some of the proteins are cycling through a regular pattern of activity values.

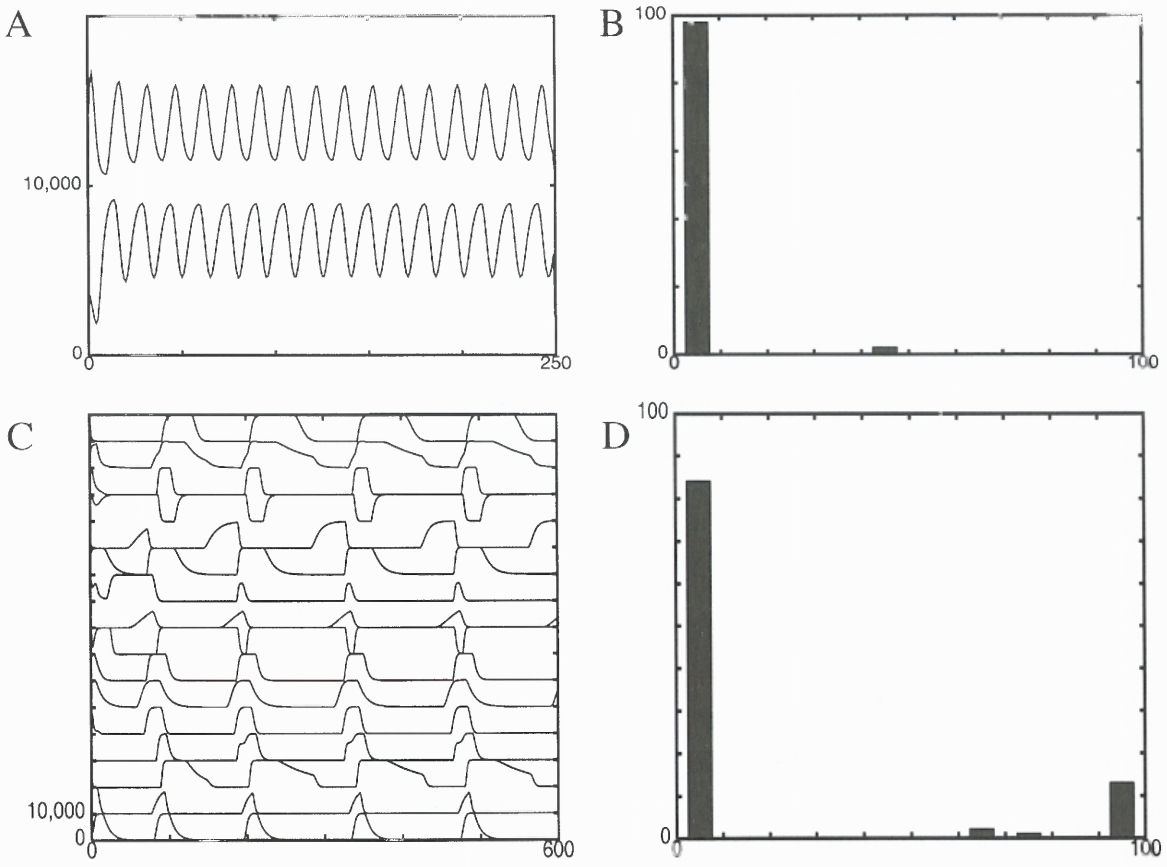


Figure 3.3: Protein activities and range of fluctuations are shown for typical networks with $p_h = p_k =$ (A,B) 0.01, (C,D) 0.05. The left hand diagrams (A,C) show the variation in activity, over time, of those protein types which show fluctuations in activity values. Protein types which stabilized to fixed values are not shown. The protein activity plots in (A,C) are stacked on top of one another, kinases first. The ordinate axis is drawn so that the full range of activity of each fluctuating protein is shown. The corresponding plots (B,D) illustrate, in each case, the percentage of the total number of proteins that are found to fluctuate against their corresponding fluctuation range. A window of $w = 50$ time steps and a run-in of 200 time steps are used. (See text for description of method.) In the $p_h = p_k = 0.01$ network (A,B), only two oscillating proteins are found, one kinase and one phosphatase. The period of the oscillation is 15 time steps and the network has a settle time $T_s < 50$ time steps. In (C,D) there are nine kinases and seven phosphatases oscillating with a period of 140 time steps and $T_s \approx 100$ time steps.

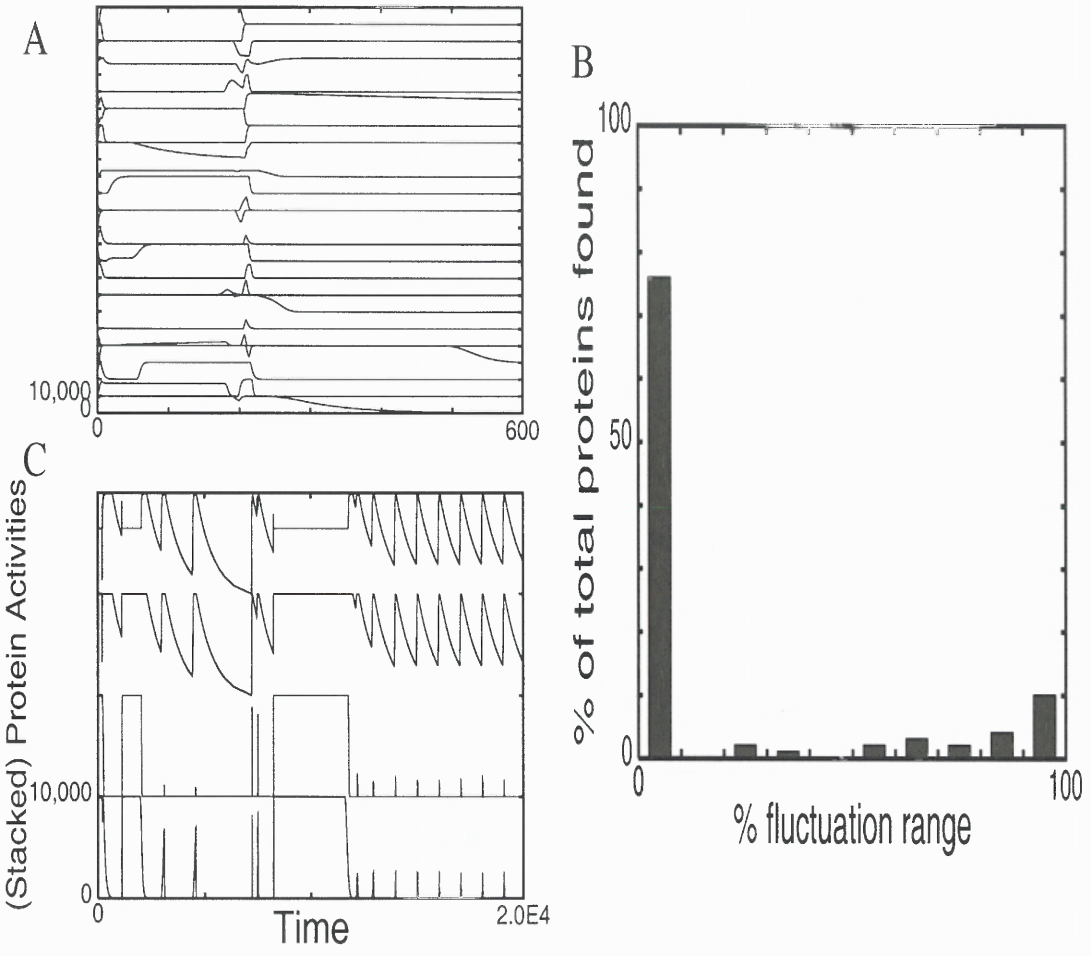


Figure 3.4: Protein activities and range of fluctuations are shown for a typical network with $p_h = p_k = 0.10$. This figure follows the same format as Figure 3.3. The plots (A,B) indicate that over the first 600 time steps, more than 20% of the proteins are found to be fluctuating by more than 60% (twelve kinases and twelve phosphatases), this is because the network is still in a transitory period. The network finally settles after about 15,000 time steps into an oscillation of length 1,000 time steps (C). At this time four proteins are found to be oscillating, three are kinases and one is a phosphatase.

probabilities. The settle time, T_s, or the number of time steps required for networks to stabilize, of more densely connected networks (where $p_h = p_k = 0.10$) can increase dramatically from $T_s < 100$ to $T_s > 1,000$. However, instances at higher connectivity are found with $T_s < 100$. Steady states (fixed points) are still apparent at these higher connectivities. However, settle times are typically much longer, with networks undergoing rapid fluctuations in activities for early times. This kind of behavior is not unexpected, as increasing the probability of interaction between modeled kinases and phosphatases is likely to increase the complexity of interactions within the cell types, thereby increasing the variety and complexity of the network evolution. Figure 3.4C with $p_h = p_k = 0.10$ shows an example where the network settles into an oscillation with a period of 1,000 time steps after an initial settle time of $T_s \approx 15,000$.

Networks which display only fixed point behavior are unlikely to be representative of real biological signaling networks. Oscillations are readily observed features of real cells, as discussed in Chapter 1. In order to eliminate those networks which show stability over all simulated starting conditions, a classification criterion is introduced.

The difference, Δ_w, between the maximum and minimum activity of each protein in any given network is measured after a suitable "run-in" time and over a specified number of time steps called a "window". Any network with no proteins satisfying $\Delta_w / o_{max} > 0.6$ (over all of the simulated starting conditions) is said to be a "Type I" network[†]. Otherwise, the constructed network is said to be a "Type II" network. The Type II networks shown in Figure 3.5 are either networks which contain oscillations in protein activity levels, or networks that were still undergoing transient fluctuations in protein activity levels before finally stabilizing. This classification, while being somewhat crude, is enough to enable some preliminary observations on network stability to be made. Figure 3.5A, illustrates that the dominant property of

[†]Recall that in this model, activity and occupancy are the same, and hence o_{max} also represents the maximum activity level of any protein type.

networks is that many nodes do not change, indicating a degree of stability. At higher connectivity, fixed point behavior is still common. For $p_h = 0.1$, approximately 10% of the simulated networks are classified as Type II.

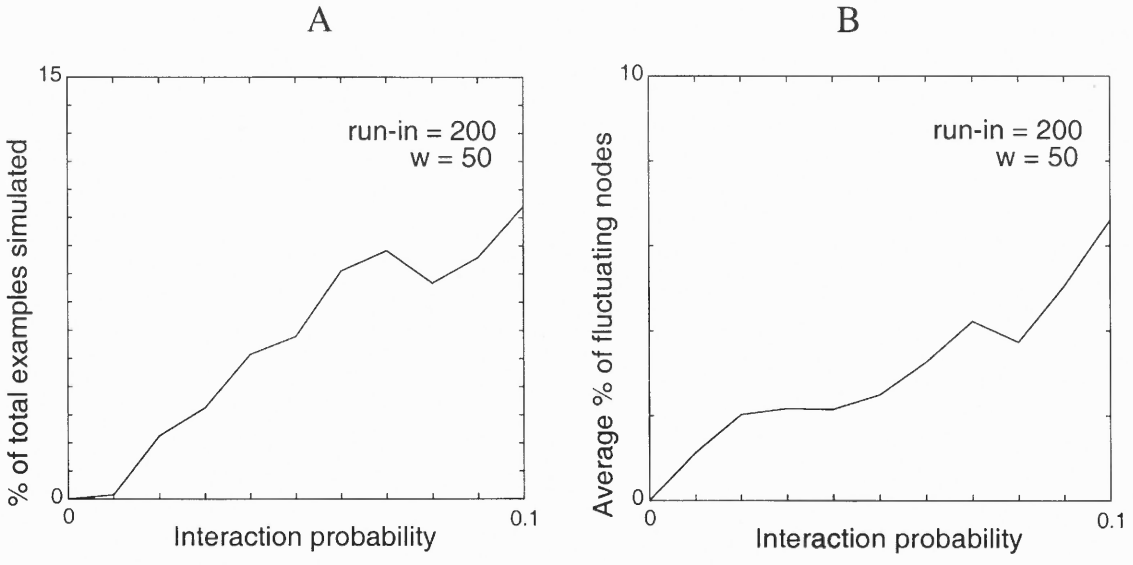


Figure 3.5: Average properties of Type II networks as $p_h = p_k$ is varied. For each of the values of p_h tested, one hundred different random networks are simulated, each from one hundred random starting conditions. (A) The percentage of Type II networks found (of the 100 different networks simulated) is plotted as a function of the interaction probability, p_h. The number of Type II networks increases as the probability of interconnection increases. This is reasonable since the interactions within the matrix will be more complex. (B) For each Type II network, at a given p_h value, the number of protein types satisfying $\Delta_w/o_{max} > 0.6$ is computed and divided by the total number of protein types in the network. This number is then averaged over the 100 different simulations for each Type II network and then averaged again over all Type II networks. It is this "Average percentage of fluctuating nodes" found in Type II networks that is plotted against the interaction probability. The average number of fluctuating proteins per network found also increases as p_h increases.

Figure 3.5B examines the number of fluctuating nodes found in Type II networks. At low connectivities, typically only one or two proteins are found to fluctuate within any given network. However, as $p_h = p_k$ approaches 0.10 nearly 10% of the proteins are found to be changing by more than 60% in any typical network. The cut-off of $\Delta_w/o_{max} = 0.6$ is chosen from a frequency plot of the

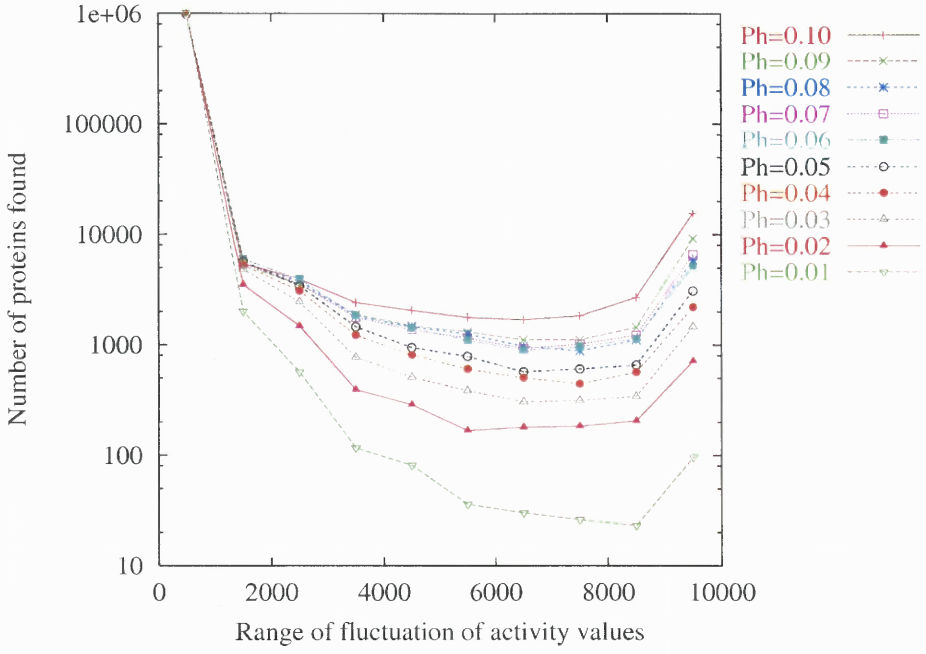


Figure 3.6: A frequency plot of number of proteins against the difference in their maximum and minimum activity values over a series of Monte Carlo simulations. Networks have p_h set between 1% and 10% . Data are grouped into activity range bins of size 1000. A total of 10,000 random networks were simulated, each with 100 protein types, giving a total sampling of 10^6 protein types for each connectivity parameter tested.

range of fluctuations in protein activities over all simulated networks (see Figure 3.6). This distribution is bimodal with more than 90% of all nodes fluctuating less than 10%, with the remainder, especially at higher connectivities, fluctuating through 90%. The minimum in this distribution occurs at around the cut-off value.

There are several limitations of this classification strategy that are worth mentioning here. The classification will obviously not capture the richness of the dynamics of the random networks. It may also misclassify networks if the run-in time is too narrow or the window too small. In the latter case, networks with large settle times will be classified as Type II if the window is too short, and networks with large periodicities will not be captured if the window is too small. Also, the network shown in Figure 3.3A might be classified as Type I network if a cut off value of 0.6 is used. However, it is likely that it would be classified as Type II once a

large number of simulations are performed from different random initial conditions. This said, the purpose of this classification is not to illustrate all the dynamic features of networks. Instead, it is used as a powerful tool to illustrate very clearly that stable states dominate at low connection probabilities.

3.4 ATP Consumption in Modeled Networks

Given the fact that as yet no simple technique exists for looking at real-time oscillations in kinase activity (or phosphatase activity) in real cells, one might question the relevance of examining activities of individual protein types in simulated networks. This point will be addressed later in §5.4.

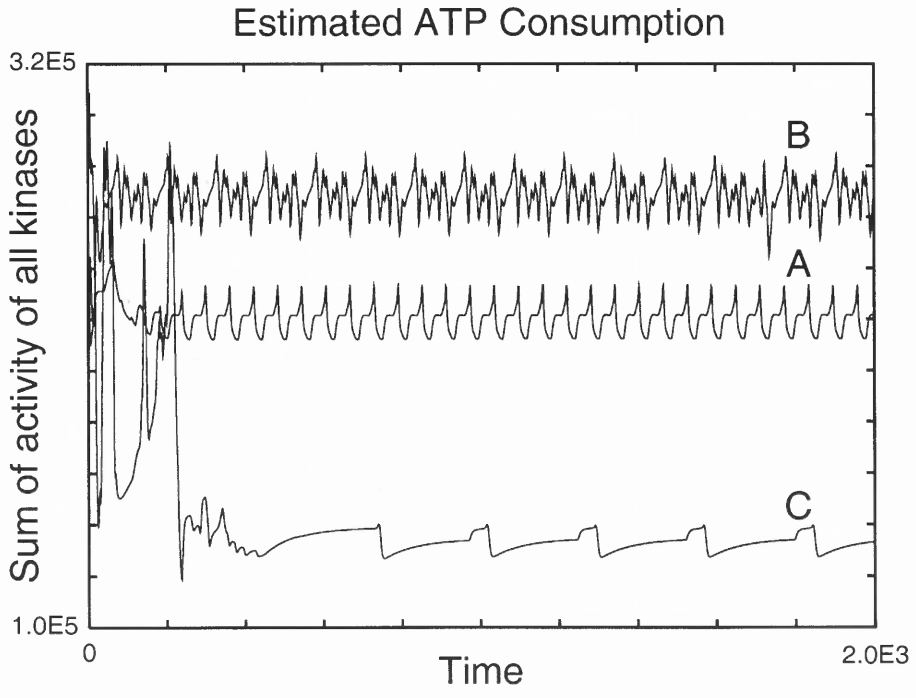


Figure 3.7: Estimate of ATP consumption for sample oscillating networks with $p_h = p_k =$ (A) 0.04, (B) 0.06, (C) 0.10. Three oscillatory networks are shown here. The estimate of ATP consumption is based on the fact that one molecule of ATP is used for each phosphorylation event. (See Figure 1.2.) Hence, the ordinate axis represents a summation of the total kinase activity in each network and gives a measure of the ATP requirements of the modeled cell. More complex ATP oscillations are seen in network with higher connectivities and this is usually coupled with an increase in periodicity.

It was noted in Chapter 1 that a single phosphorylation reaction utilizes one molecule of ATP. Summing the total kinase activity at any given time in a simulated network, therefore, gives a measure of the ATP requirements of a simulated cell at that instance in time. ATP is a small, freely diffusible intracellular molecule, and it is likely that techniques in cell biology will enable the monitoring of ATP levels inside cells in the very near future. Thus, it is prudent to examine the ATP consumption patterns of modeled cell types as part of this study. Figure 3.7 shows some sample plots of ATP consumption in networks which show spontaneous oscillations in protein activity. More complex oscillations are seen in networks with higher connectivities and this is usually coupled with an increase in periodicity.

3.5 Summary

Given any fixed p_h and p_k value within a biologically plausible range, any individual matrix produces different types of behaviors (e.g., rapid settling to a fixed point or oscillations). These behaviors are classified into two types (Type I and Type II). More variety occurs with increasing p_h, p_k, although no obvious phase transition phenomenon has been identified. Phase transitions have, however, been observed in other connectionist models [77, 92].

Some networks show a tendency to settle into fixed points over a wide range of initial conditions. Other networks typically show a core set of oscillating proteins whose amplitude and periodicity shift when the initial conditions are varied. At higher probabilities, a change in the initial conditions can alter the core set of oscillating proteins. This is often accompanied by a change in the settle time and nature of the network periodicity.

As yet, no simple technique exists for looking at real-time oscillations in kinase activity in cells. A novel suggestion here is the idea that ATP consumption may reveal information about the intracellular dynamics. This indicator might, in the

near future, be a measurable quantity in the laboratory. (Currently, the signal to noise ratio for ATP measurements is poor.) It might then be possible to carry out a laboratory "test-tube" experiment in which ATP levels of a mixture of kinase and phosphatase molecules could be measured.

CHAPTER 4

A GENETIC ALGORITHM

4.1 Introduction

In the previous chapter, the Monte Carlo simulation method was discussed. The main purpose of this chapter is to introduce a computational technique for large search problems is a "Genetic Algorithm" (GA). GAs are used in engineering and operations research, image processing and pattern recognition, and have also been applied to biological problems [93].

The high dimensionality of the modeled signaling networks dictates that a random selection samples only a minute fraction of the parameter space. A more robust and effective search technique, a GA, is implemented to complement the Monte Carlo simulation method. The GA is used in order to search more effectively for phosphorylation networks displaying a wide range of dynamic properties from a given set of random starting conditions. Since responsive and dynamic activity in the intracellular protein phosphorylation system is a characteristic feature of real cells, the genetic algorithm search is designed to select networks which contain proteins that exhibit different patterns of activity in response to different input patterns (i.e., where a protein can switch on or off or oscillate in a variety of ways according to the state of its neighbors). The genetic algorithm developed here is also designed to mimic some of the processes which occur naturally during the evolution of real cells.

In this chapter, a comparison is also made between the Monte Carlo simulation method and the genetic algorithm approach. The data show that the latter approach provides a more efficient search strategy with which to explore network properties. Moreover, not only does the GA provide a more effective way to search for networks with certain specified properties, it is also designed to be more biologically motivated.

A study of the effect of changing certain parameters of the genetic algorithm is presented in §4.4. A specific network generated by this algorithm will be discussed in some depth in §5.4. Results of the clustering algorithm are shown for this network. In light of these data, the hypothesis that networks may exhibit limited numbers of attractor states will be discussed.

4.2 GA Development

A standard genetic algorithm uses a combination of operations, called reproduction, crossover and mutation [93]. In such algorithms the search problem is generally encoded as a string of data, typically a string of binary digits [94, 95]. A population of strings is then evolved according to these three operators. Survival to the next generation is determined by a "fitness" function. In this implementation, networks are selected based on properties of the dynamics of the protein interactions within them. It is the model cells themselves that are reproduced. Modeled proteins are composed of domains which are defined as regions of the protein that perform a particular function. Since duplication and shuffling of these functional units, or domains, is thought to have occurred during evolution [57, 58], crossover and mutations are carried out upon the analogue of protein domains *within* the cells. Each cell represents a single "species" in the genetic algorithm simulation. The general structure of the algorithm is described below.

A starting population of cells is generated at random, as described in §3.2. The cellular dynamics are then investigated from a series of random starting conditions and for a series of T time steps, as before. The population is then ranked according to fitness (see §4.2.1). The best cells are selected for reproduction and the new population is subjected to crossover and mutation (see §4.2.2). The process is then repeated for a specified number of generations, G.

4.2.1 The Construction of a Fitness Function

The fitness function is designed to select networks which demonstrate a variety of distinct dynamic behaviors. Most networks at low connectivity settle into stable patterns of activity within fifty time steps. The level of activity of each protein in the network at the final time step T is compared over the set of S random starting conditions. If the protein activity always remains stable at the same fixed value over a variety of starting conditions, the protein does not appear to have the potential to switch its activity according to a different signal. If a protein has a wide range of stable fixed state values at time T, it may exhibit the richness of behavior desired. If, however, this occurs because the protein is not a substrate of any of the other proteins in the network, then this is less interesting. The final activity state of every protein i in any given cell of the population can be recorded at time T for each of the S simulations. These states are stored in a matrix $a_{ij}^{(T)}$ where $1 \leq i \leq S$ (S is the total number of simulations for cell C) and where $1 \leq j \leq N$ and N is the total number of protein types in cell C. The superscript T indicates that all data is taken from the final time step of each simulation. For each protein in any given cell the mean of these final activity values, over all starts, is calculated (Equation 4.1).

$$\bar{a}_j = \frac{1}{S}\sum_{i=1}^{S} a_{ij}^{(T)}, \quad 1 \leq j \leq N \tag{4.1}$$

The mean absolute deviation in activity from this value, $\bar{a}_j$, for each protein j in a specified cell is then calculated using Equation 4.2. The proteins with a low A_j value will be those which settle into nearly the same activity state for most of the simulations.

$$A_j = \frac{1}{S}\sum_{i=1}^{S} \left| a_{ij}^{(T)} - \bar{a}_j \right|, \quad 1 \leq j \leq N \tag{4.2}$$

A second measure is used to indicate the variation in the activity of individual proteins over each of the S simulations. This value, denoted by S_{ij}, is determined by evaluating the absolute change in activity of a protein between consecutive time steps, and then summing over the T time steps of the simulation. The maximum of this value for protein type j is calculated over all S simulations and is denoted S_j. This value will indicate the maximum variation in activity of that node for all simulations of any chosen cell in the population. If a protein never changes its activity, then either it has no inputs or the effects of the other nodes upon it always exactly balance. In general, those proteins with a low S_j value can be proteins with either no inputs or proteins whose activity value typically changes only slightly over time. The rows of the matrix $a_{ij}^{(t)}$ denote the activity vector of simulation i at time step t.

$$S_{ij} = \sum_{t=0}^{T-1} \left| a_{ij}^{(t+1)} - a_{ij}^{(t)} \right| \quad 1 \leq i \leq S, \;\; 1 \leq j \leq N \tag{4.3}$$

$$S_j = \max_{1 \leq i \leq S} S_{ij} \tag{4.4}$$

A threshold is then set for both of the parameters A_j and S_j. The number of proteins in any specified cell of the population passing both thresholds (i) and (ii) of Relation 4.5 is evaluated.

$$\left.\begin{array}{ll} (i) & A_j > \frac{1}{4}\, o_{max} \\ (ii) & S_j > 5\, o_{max} \end{array}\right\} \quad 1 \leq j \leq N \tag{4.5}$$

These thresholds are set based on a typical simulation of length $T = 500$. This simple test is designed to eliminate those proteins with high A_j yet low S_j values (and vice versa). The set of proteins, for a given cell c in the population ($1 \leq c \leq C$), passing both thresholds in Relation 4.5 is denoted by $\{p^c\}$ and is used as a measure of fitness, F^c, by which to rank the cells. A fit cell has a high percentage of proteins which pass the thresholding criteria.

During the course of the search, cells are penalized if high fitness results from a network structure consisting of a few control nodes, with high output connectivity, whose substrate nodes themselves have no substrates. If the controlling nodes of such networks show a wide range of different activity values, then the nodes they influence may also demonstrate the same properties which causes an artificially high fitness value. Protein kinases that are active and do not influence other elements of the cellular network can be thought of as molecules which use ATP but perform no functional role. Usage of valuable ATP resources in the cell in this manner would not be of selective advantage.

Cells which have large numbers of proteins with no inputs will perform poorly when tested with the threshold criteria of Relation 4.5. The additional condition mentioned in the previous paragraph penalizes cells where many proteins have no outputs resulting in a reduction of their fitness. The fitness value is then calculated by counting the number of proteins in the intersection of the two sets $\{p^c\}$ and $\{p_o^c\}$. The latter is the set of proteins in cell c $(1 \leq c \leq C)$ which satisfy $M_{oj} = 0$, where $1 \leq j \leq N$.

At each generation the population of cells is ranked according to fitness. By ranking cells in order of fitness, rather than weighting cells by fitness, highly fit cell species are prevented from dominating the population at the early stages of the genetic algorithm. The best F cells from the current generation, G, are compared to a record of the F fittest cells found in all prior generations up to and including generation $G-1$. The fittest cells from the combined list are allowed to reproduce, but the number of cells introduced from the previous generations is restricted. This selection process allows some of the parent cells to survive in the population, but the remaining cell strains die off and do not survive to the next generation. The overall population size, C, remains constant, and each parent produces the same number of offspring. Variation in the offspring is achieved through crossover and mutation (see §4.2.2).

In some simulations, random cell strains, R, are introduced into the population at specified generations. The overall structure of the GA is shown in Figure 4.1.

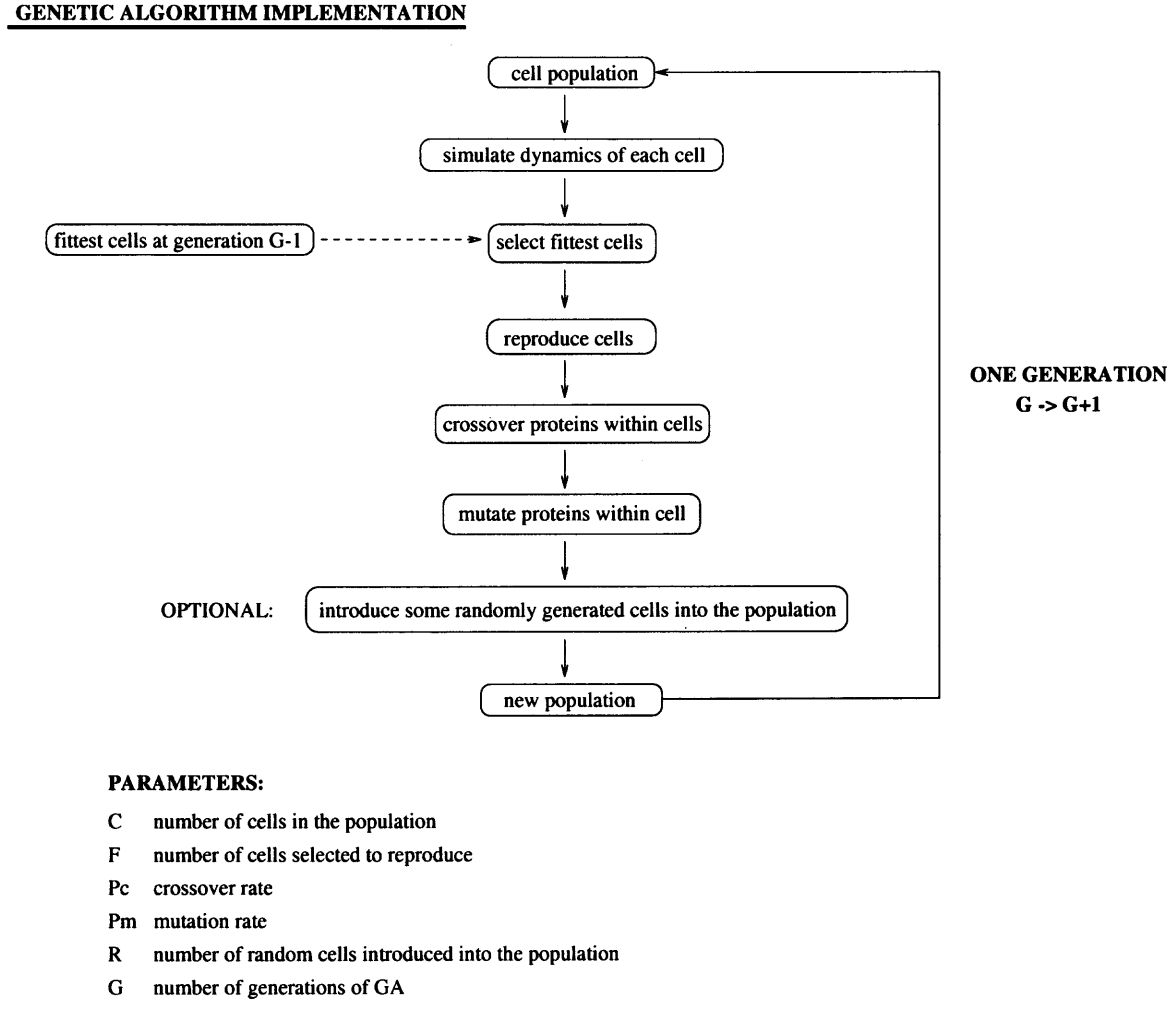


Figure 4.1: Genetic algorithm construction.

4.2.2 The Crossover and Mutation Operators

Protein domain shuffling is mimicked by swapping rows or columns within the daughter cell connection matrix. This corresponds to shuffling catalytic or regulatory domains between protein types, respectively. The crossover rate, p_c, determines the number of rows or columns that are shuffled within each daughter cell. The crossover rate is obtained by dividing the number of domains to be swapped by the total

number of protein types in the cell. The total number of connections in each cell remains constant, although the number of outputs from and inputs to each protein varies. Exchanging regulatory domains alters the number of regulatory inputs of the proteins involved. Swapping the catalytic domains of two proteins alters their substrate specificity. This allows greater flexibility in the connection matrices (cell types) because the parameters p_h, p_k, described in §2.2, may vary. A protein will have Np_k output connections if it is a kinase and Np_h output connections if it is a phosphatase. Therefore, some time after the genetic algorithm is initiated, the number of output connections per protein will no longer be identical. The crossover operation is illustrated in Figure 4.2.

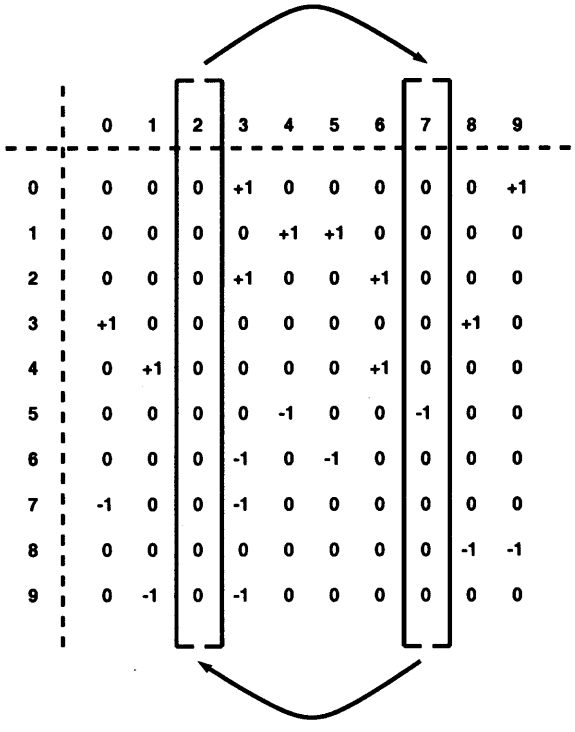


Figure 4.2: Crossover in cells during the genetic algorithm.

Mutations are generated by choosing a cell at random and selecting an individual connection (matrix element) to mutate. A zero connection in M_{ij} is mutated to $+1$ if protein i is a kinase, or -1 if protein i is a phosphatase. If mutated, a non-zero connection is set to zero in order to keep the total number of connections in the matrix

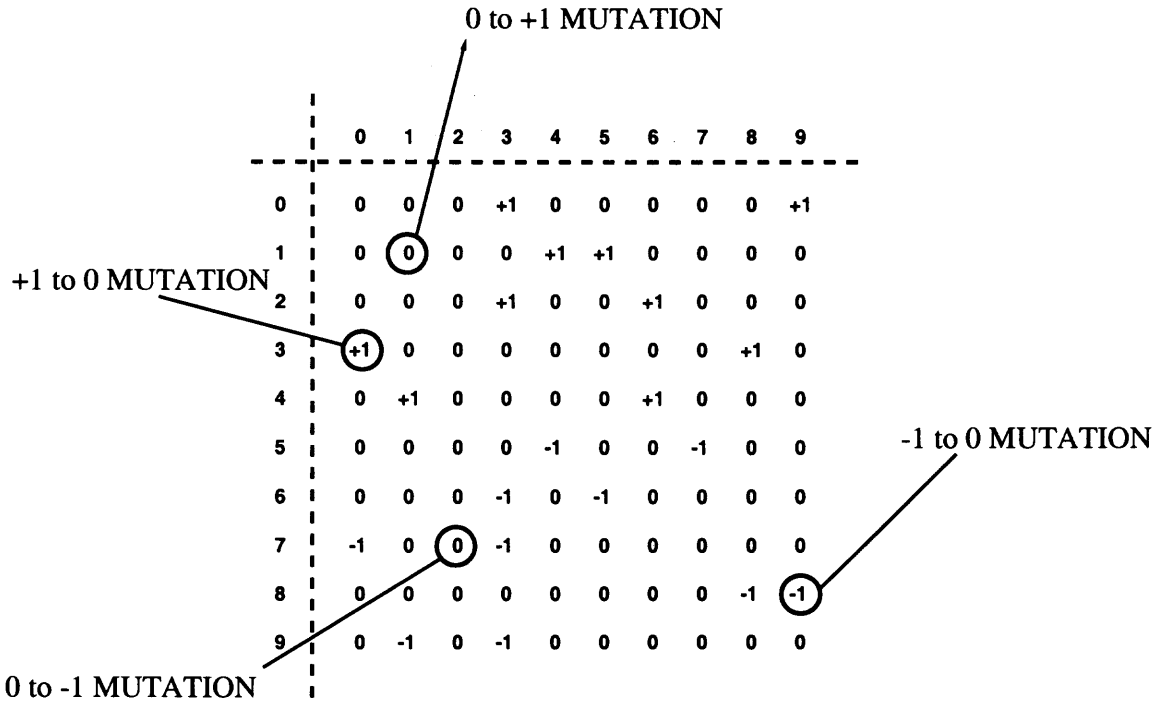


Figure 4.3: Mutation in cells during the genetic algorithm.

constant. When a zero to non-zero mutation is performed, a balancing non-zero to zero mutation is also made. This means that a row i of a connection matrix may, after some number of generations, have no positive or negative connections. If a zero connection in such a row is selected for mutation, a decision to introduce a $0 \rightarrow +1$ or $0 \rightarrow -1$ mutation is made at random. Hence, this algorithm allows variation in the ratio of kinases to phosphatases. Some proteins may change so that they evolve to a state where they have either no input interactions (the protein is essentially unregulated) or no output connections (no substrates are present). The mutation rate, p_m, is defined as the number of mutated connections divided by the total number of non-zero connections of the cell's connection matrix. Some of the different types of possible mutations are illustrated in Figure 4.3.

4.3 Comparison of Genetic Search and Random Search

Simulating the dynamics of cellular networks of one hundred components requires an investigation of a vast state space, not to mention the effects of altering individual network parameters. The Monte Carlo simulation method is highly useful for a preliminary investigation into properties of random networks over large connectivity ranges. Data are presented here that compare the designed GA to an equivalent number of Monte Carlo simulations.

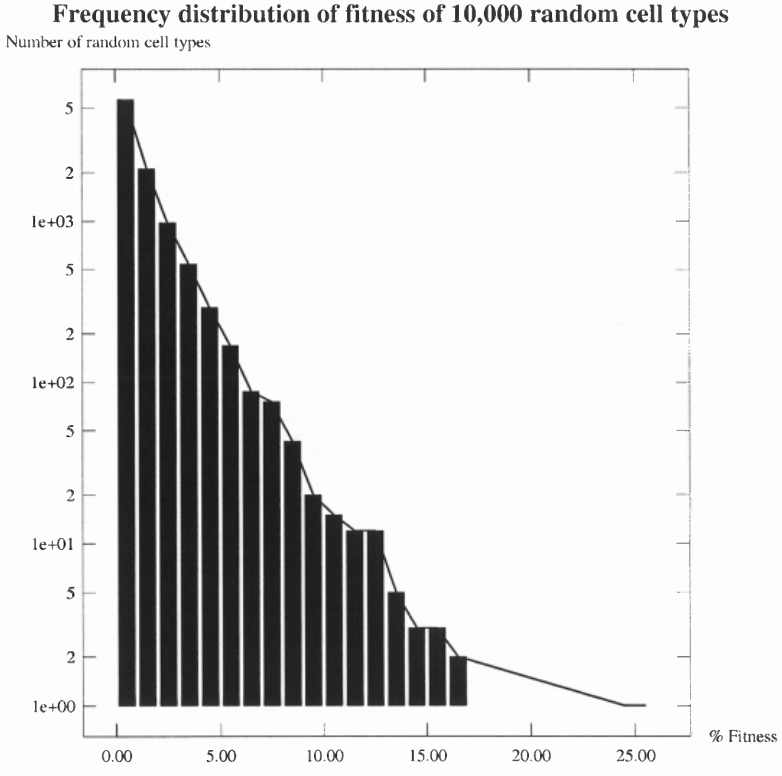


Figure 4.4: Monte Carlo simulations on 10,000 randomly generated cell types. The networks have the following parameter restrictions: $[p_h, p_k, f_h, T, N, S] = [0.03, 0.03, 0.5, 500, 100, 100]$. The majority of the cells have low fitness values. Note the logarithmic scale used on the vertical axis. The graph illustrates the distribution of fitness of the random cells simulated without the use of a GA. Of the 10,000 cells simulated, 5,628 cells have a fitness level of one percent or below. There are 7,742 cells of fitness less than 2%. The best cell has a fitness of 26%.

The cell types in the Monte Carlo simulations are generated at random according to the following set of parameters: $[p_h, p_k, f_h, T, N, S] = [0.03, 0.03, 0.5, 500, 100, 100]$. In order to compare the results of this data with the GA, 10,000 networks are simulated using the Monte Carlo method. (See Figure 4.4.)

The GA is run for 100 generations with a fixed population size of one hundred cells, $C = 100$. Of the resulting 10,000 cells generated during this time, the best one hundred cells have the following properties. The best cell has a fitness of 26% (see Figure 4.4) and the mean and median fitness of these one hundred cells are 11% and 10% respectively. The worst of these cells has a fitness of 8%. These data are comparable to a population of the GA at generation 100, provided that the population at each generation contains 100 cells. Even with a poor choice of parameters for the GA, the search still outperforms random search in the sense that the fitness of the best one hundred cells at a comparative point is higher when the GA is used. Implementation of the GA provides a more efficient way to search for cells with varied dynamic properties. From this point forward, all the data presented are from networks that have been evolved via the GA.

4.4 A Parameter Study

The GA has several parameters associated with it: the mutation and crossover rates, the population size and the number of offspring per parent. This section will methodically examine how each of these factors affects the quality of the local and global properties of this search algorithm.

Figure 4.5 illustrates data from simulations where the mutation and crossover rate have been systematically varied. The fitness of each cell at each generation is evaluated from its dynamics, as described in §4.2.1. The cellular dynamics are investigated by simulating each cell from a set of one hundred random initial activity vectors for $T = 500$ time steps.

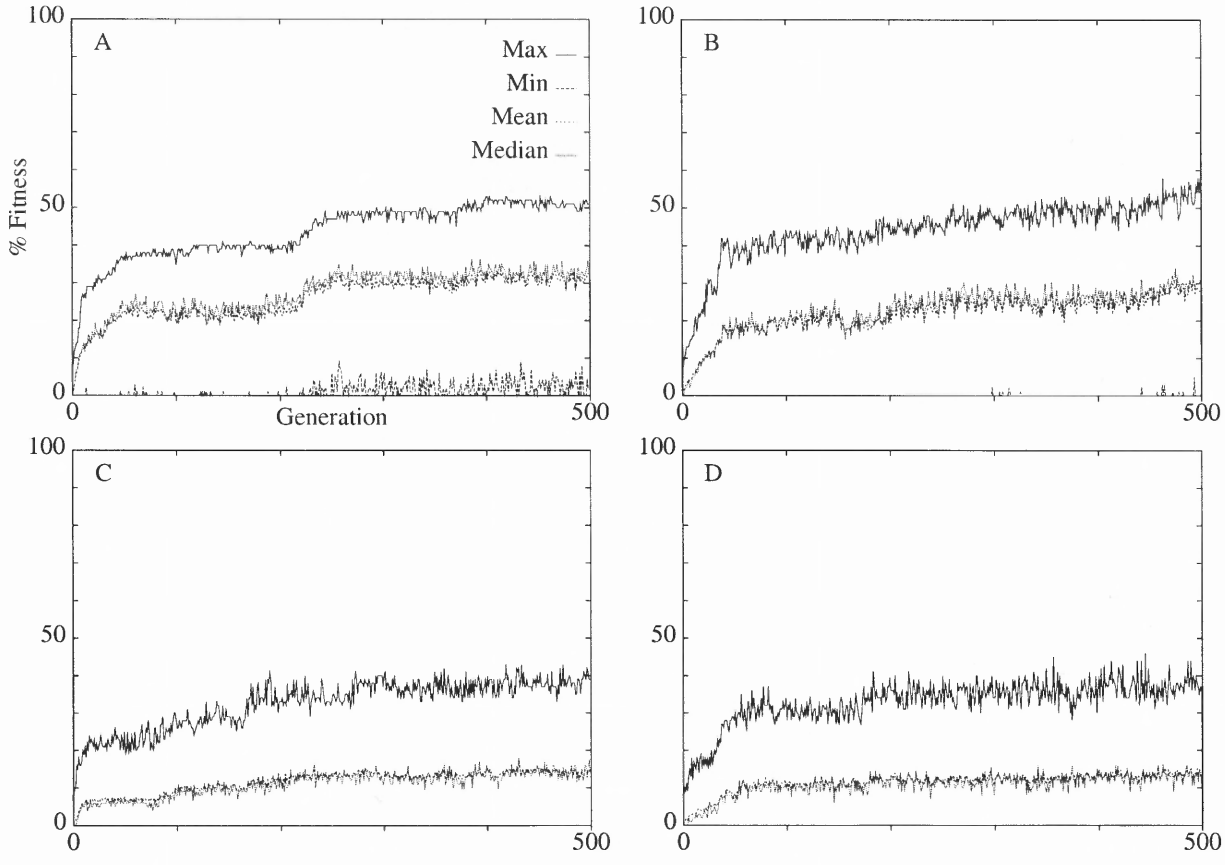


Figure 4.5: The effect of varying the mutation and crossover rates on the performance of the GA. Each simulation uses the same random initial population of cells whose parameters are $[p_h, p_k, f_h, T, N, S] = [0.03, 0.03, 0.5, 500, 100, 100]$. The parameters of the algorithm are set at $[G, p_m, p_c, C, F]$: (A) $[500, 0.0067, 0.02, 100, 20]$, (B) $[500, 0.033, 0.02, 100, 20]$, (C) $[500, 0.0067, 0.10, 100, 20]$, (D) $[500, 0.033, 0.10, 100, 20]$. For the first five generations, $R = 20$ randomly generated cells are introduced into the population. A maximum of three cells are re-introduced into the population at each generation G. Each plot shows the maximum, minimum, mean and median fitness in the population at each generation.

The data of Figure 4.5 show cells whose parameters are set as follows: $[p_h, p_k, f_h, T, N, S] = [0.03, 0.03, 0.5, 500, 100, 100]$. At every generation, the twenty fittest cells are selected for reproduction. Together, these selected cells then give rise to the population at the next generation. These twenty cells each generate four offspring for the first five generations and during this time twenty randomly generated cells are also introduced into the population. The reason for this is to prevent the algorithm from locking into one or two cells early in the simulated evolutionary process. Thereafter, each cell generates five daughters and no more random cells are introduced into the population. After this point, the average fitness of the population is usually much higher than the fitness of a randomly generated cell, so continuing to introduce these cells into the population has no effect. The size of the population remains fixed as the algorithm progresses. Protein domains are swapped between proteins within any given cell at a specified rate, p_c. Every daughter cell in the population undergoes crossover. Every cell has an even number of mutations determined by the mutation rate, p_m. This ensures that the cell connectivity does not increase steadily through the simulation (see §4.2.2). No more than three of the fittest cells from generation $G-1$ are reintroduced into the population at generation G. All the data shown in Figure 4.5 illustrate the change in fitness of a population of one hundred cells over $G = 500$ generations.

In Figure 4.5, the fitness of the best and worst cell at each generation are shown together with the mean and median percentage fitness of the entire population at each generation. In panel A, the crossover and mutation rates are set at $[p_c, p_m] = [0.02, 0.0067]$. At generations $G > 5$, the fittest twenty parents each generate five offspring which, after mutation and crossover, become the new population at the next generation. Every daughter cell possesses one hundred different protein types of which one pair of proteins, chosen at random, exchange regulatory domains. The mutation rate allows an average of two mutations per cell. The population fitness

increases rapidly in the first fifty generations. By generation 25 all cells are descended from one single parent from the initial population. During certain periods of evolution, the fitness appears to increase in small bursts $G \in [0, 50]$ and $G \in [200, 250]$ and then plateau. After five hundred generations, the best cell has a fitness $\approx 52\%$. The mean and median of the cell population are similar, at 34% and 35%, respectively. The median is slightly higher, indicating that the distribution is slightly biased towards cells of lower fitness. The minimum fitness of the population always remains near zero indicating that even small perturbations of high fitness cells can drastically reduce cell fitness.

In Figure 4.5B, $[p_c, p_m] = [0.02, 0.033]$. The increase in the mutation rate to an average of ten mutations per cell causes a sharp increase in fitness in the early stages of the algorithm. A cell of 40% fitness is reached in fewer than fifty generations, although the average population fitness takes over five hundred generations to reach 30% fitness, as compared to 250 generations in panel A. The population fitness shows a steadier gradual increase, as compared to panel A, although the fitness level fluctuated much more widely between consecutive generations. The maximum, mean, median and minimum fitness after five hundred generations are $[58, 29, 30, 0]$ respectively.

In Figure 4.5C, $p_c = 0.10, p_m = 0.0067$. In this case, five pairs of domains per cell are swapped at random. The increase in the crossover rate, as compared to panel A, causes a substantial reduction in the average population fitness and increases the noise component in the evolution of cell fitness. There are sharp rises and falls in the maximum population fitness, but there is a smoother increase in the average population fitness. A cell of 40% fitness is not generated until after nearly 200 generations. After five hundred generations the maximum, mean, median and minimum fitness of the population are $[39, 14, 14, 0]$. The fitness progression is more rugged; a larger perturbation from the peaks in the landscape appears to have a detrimental effect on the mutated offspring. In Figure 4.5D, $p_c = 0.10, p_m = 0.033$.

Here, the effect of higher crossover and mutation rates is marked, although the population shows greater improvement after fifty generations, as compared to panel C. The population graph is similar to Figure 4.5C but the magnitude of noise component is larger. At the final generation the maximum, mean, median and minimum fitness are $[38, 14, 12, 0]$ respectively.

In each case there is a general improvement in the fitness of the population as the algorithm progresses. The fluctuations or noise present in the graphs increases noticeably from panels A to D in Figure 4.5. This corresponds to an increase in both the mutation and crossover rates. Lower values for p_c and p_m allow more exhaustive search in the local neighborhood of the best cells which leads (via small steps) towards a few local peaks in the search space. Higher mutation and crossover cause larger leaps across the search space. If this search space consists of many sharp peaks or ridges, a few large steps may cause a drastic decrease in fitness (into local minima) from which it is harder to recover (Figure 4.5D). Comparing Figures 4.5A and 4.5D after one hundred generations, the difference in fitness of the best cell at the final generation is 14 percent.

Each graph shows a two to three-fold rise in fitness within the first twenty generations. All the graphs show the greatest improvement in fitness in the early generations $G \in [0, 50]$. This may be due to the nature of the fitness function and the choice of the number of offspring per parent. The fitness of the overall population will also depend on the quality of the starting population. From an analysis of the evolutionary history of the cells, it is evident that the entire population is descended from one single parent from the starting population.

Figure 4.6 illustrates the effect of varying the number of offspring per parent and the number of parent cells allowed to reproduce to the next generation. In the instance where the population size is kept fixed, an increase in the number of offspring per parent will automatically reduce the the number of parents that can be selected

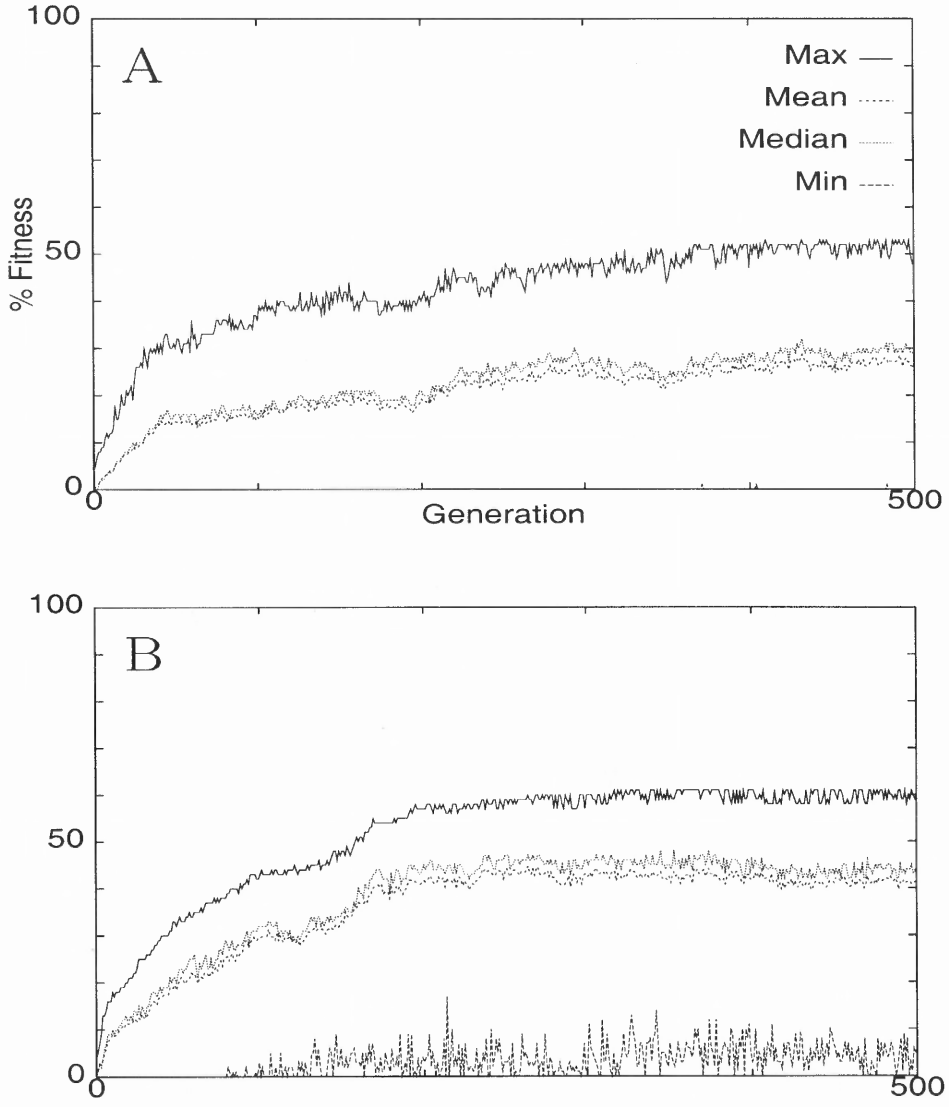


Figure 4.6: The effect of varying the number of offspring per parent on the progression of the GA. The population size is different from Figure 4.5. Each simulation is performed using the same random initial population of two hundred cells whose parameters are set as follows $[p_h, p_k, f_h, T, N, S] = [0.03, 0.03, 0.5, 500, 100, 100]$. The parameters of the algorithm are $[G, p_m, p_c, C, F]$: (A) $[500, 0.0067, 0.02, 200, 100,$ (B) $[500, 0.0067, 0.02, 200, 20$. No randomly generated cells are introduced into the population at any stage. A maximum of three cells are re-introduced into the population at each generation G from the previous generation. The maximum, minimum, mean and median fitness of the population are shown in each plot.

for reproduction at each generation. The GA parameters $[G, p_m, p_c, C, F]$ are set at (A) $[500, 0.0067, 0.02, 200, 100]$ and (B) $[500, 0.0067, 0.02, 200, 20]$. In this case, no random cells are introduced into the population at any stage. The population size is increased to two hundred cells. In panel A, one hundred parents are selected for reproduction and each parent is allowed to generate two offspring which are then subjected to mutation and crossover. In panel B, twenty parents are selected and each of these cells then gives rise to ten offspring. Each daughter cell has on average two mutations, and a pair of regulatory domains are swapped between one pair of proteins per cell. This is designed to detect whether the search can be improved by altering the parent to offspring ratio.

The fitness graph of Figure 4.6A is broadly similar to that of Figure 4.5A, although the curves are more rugged. The initial phase of the population fitness growth during generations 0 to 50 appears to be qualitatively similar to the same phase of Figure 4.5B. Increasing the number of parents, but reducing the number of offspring per parent, allows the algorithm to maintain variety in the population for longer. It takes approximately fifty generations until the entire population is ancestrally related to a single parent. This is longer than in all the simulations shown in Figure 4.5. The population fitness, however, does not outperform that shown in Figure 4.5A. There is clearly a penalty in reducing the number of offspring per parent. In Figure 4.6B, a larger number of offspring per parent appears to improve the search. In both panels A and B, a cell of 40% fitness is not generated for approximately one hundred generations (which is comparable to Figure 4.5A). In Figure 4.6A, at the final generation the maximum, mean, median and minimum fitness are $[49, 26, 30, 0]$.

In Figure 4.6B, the cell fitness properties began to plateau after about 250 generations. The mean and median even decreased slightly during the last one hundred generations. After about 15 generations, every cell in the population is descended from a single parent. The minimum cell fitness in each generation is

higher and, at $G = 500$, the maximum, mean, median and minimum fitness are $[60, 42, 41, 0]$. A larger set of parents allows exploration of more areas of the fitness landscape (potentially more peaks) but a smaller set of offspring per parent allows less effective search in the local region around these peaks.

4.5 Summary

This section has been concerned with the design and development of a genetic algorithm. The GA is used to study networks in a more localized parameter range than the Monte Carlo simulations. The three main aspects of the algorithm are the fitness function, the crossover operator and the mutation operator. A population of cells are evolved according to these operators. Symbols and parameters of the GA are summarized in the "List of Symbols" pages at the start of the thesis.

The fitness function is designed to select networks which demonstrate a variety of distinct dynamic behaviors. Cells are ranked according to fitness and reproduce asexually to the next generation. The crossover operator results in exchange of protein domains. This is designed to mimic the evolutionary mechanism of protein domain shuffling. The mutation operator is a common feature of most genetic algorithms. This is also motivated by the biology as genetic mutations occur at a low rate in all cells.

The GA was compared to the performance of the Monte Carlo method and found to perform better with respect to the designated fitness measure. The parameter study of §4.4 illustrated that if the mutation and crossover rates are set too high, then this can have a detrimental effect on the cells' offspring. The fitness landscape is highly rugged. Large perturbations in cell connection structure have a large impact on fitness. Lower values for p_c and p_m allow more exhaustive search in the local neighborhood of the best cells. Increasing the population size and allowing more offspring per parent also appears to improve the search. With a fixed population

size, increasing the number of parent cells that are allowed to reproduce restricts the number of offspring per parent. A larger set of parents allows exploration of more areas of the fitness landscape (potentially more peaks). In all cases tested the GA showed the greatest improvement in early generations and the population always contained a significant number of cells of low fitness. This point will be addressed again in the final chapter.

CHAPTER 5

DETECTING DYNAMIC ATTRACTORS

5.1 Introduction

In this chapter, a numerical method is presented which is designed to test the hypothesis that modeled networks exhibit only a relatively small number of different attracting states. The algorithm that is developed is an adapted form of a standard clustering technique. In order to cluster data, a distance metric must be defined. A generalized form of the Hamming distance metric (a discrete metric) is applied to modified activity state vectors. The difficulties in classifying data with clustering methods are discussed. An error measure for the clustering algorithm is defined.

The clustering method is used to detect similarities in the stable states of a network selected from the final generation of a genetic algorithm simulation. It has been noted in previous chapters that networks typically exhibit fixed point behavior and also oscillatory states. The clustering algorithm is designed to deal with behaviors of these types. The results of the clustering technique are presented in §5.3.

The cell that is chosen has the highest fitness in the final generation of the GA. The cluster grouping of the network states is presented in §5.4. ATP consumption properties are examined for several of the cluster groups. These data not only provide evidence of the effectiveness of the clustering algorithm, but also illustrate its limitations.

5.2 A Clustering Technique

A clustering algorithm is developed to verify that the fitness criteria selects networks which display a relatively small number of dynamic behaviors. It is an adapted form of "k-means" clustering [96]. The algorithm classifies similarities in activity states at the end of each of S simulations for a given cell C. A distance measure is required to quantify the similarities in these states. The distance of vectors within a cluster (the

intra-cluster distance) from the cluster center should be small, whereas the distance of states in other clusters to this cluster center (the inter-cluster distance) should be large. Typical difficulties that can arise when clustering data are illustrated in Figure 5.1.

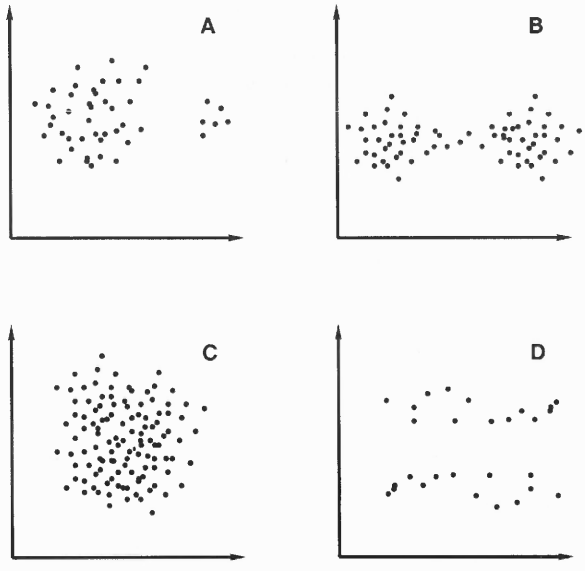


Figure 5.1: Examples of potential difficulties with classifying clusters within data. A good clustering procedure should find natural groupings in the data. If the data are normally distributed, a covariance matrix would constitute a compact representation of the data. It would indicate the amount of scatter in different directions from a mean representing the center of gravity of the cloud of data points. However, in the case where the structure of the data is non-normal, this representation would not capture all of the features of the data, see (C) and (D). Another problem can arise during partitioning of data. Occasionally a partitioning may split a cluster incorrectly because the mean square error is slightly reduced (the goal being a partitioning of the data with the smallest error). This can occur when there are outliers in the data, as in (A). In (B) one would need to decide whether the data lie in one large cluster or in two clusters that lie close together.

The fittest cell from the final generation G of a genetic algorithm simulation is studied, although in principle, any cell from any generation could be chosen. The set of proteins $\{p^c\}$ passing the thresholding criteria of Relation 4.5 are examined. Those proteins with no inputs have already been eliminated (i.e., those whose state is purely determined by their initial random starting condition). They can be thought of as input nodes to the network which can be altered by the action of an external agent.

These can be thought of as input parameters to the network. The proteins which always tend to settle into the same final activity state, regardless of input, have also been discarded. The reason for this is that a working hypothesis for a desirable property of a protein in a signaling network is that it should be able to shift its activity readily according to the states of the proteins interacting with it. Similarly, proteins with no output connections have also been rejected (see §4.2.1). The goal is to study the possible number of dynamic behaviors of cell types of high fitness by using an algorithm which can quantify in some manner the distance between the states of the network. For example, if the modeled cell type shows a small number of stable attractor states, one would expect to see a small number of clustering centers with the clustering algorithm.

Given a set of proteins, $\{p^c\}$, the clustering algorithm first detects whether the protein type shows:

(i) a high/low fixed activity level, or

(ii) oscillations in activity.

To detect proteins with oscillatory activity, proteins with significant fluctuations in activity were chosen. This simplifying assumption (i.e., that fluctuation implies oscillation) is based on the observation that network behaviors generally appear to be either fixed equilibrium points or limit cycle oscillations. The initial transient period $[0, T-w]$ is ignored, so that the network settles into its stable behavioral mode, where the observation window has length w. Potential errors of measurement caused by an early transient period are eliminated. The fluctuation in activity of a protein j, where $1 \leq j \leq N$, is measured over the window $t \in [T-w+1, T]$ for each of the simulations i, where $1 \leq i \leq S$. This variable is denoted by Δ_{ij}. Denote the dimension of the set $\{p^c\}$ as p. Equation 5.1 is used to calculate the size of the maximum fluctuation in activity of the protein over the specified windowing period.

$$\Delta_{ij} = \left| \max_{t \in [T-w+1,T]} a_{ij}^{(t)} - \min_{t \in [T-w+1,T]} a_{ij}^{(t)} \right|, \quad 1 \leq i \leq S, \;\; 1 \leq j \leq p \tag{5.1}$$

Preliminary simulations indicate that proteins generally fluctuate through the entire range of activity or remain stable. A protein j is considered to be fluctuating during simulation i if its corresponding Δ_{ij} value is found to be more than half the maximum activity range. Proteins which are thought to be oscillating are denoted using the symbol "X". The final activity state of the proteins from one of the S simulations of the chosen cell has then been reduced to a vector with fewer than N components. Each component takes one of three possible values (0,1,X). The new matrix $\tilde{a}_{ij}$ contains this modified data representation (see Equation 5.2).

$$\tilde{a}_{ij} = \begin{cases} \text{X}, & \text{if } \Delta_{ij} > \frac{1}{2} o_{max} \\ 0, & \text{if } \Delta_{ij} < \frac{1}{2} o_{max} \quad \text{and} \quad a_{ij}^T < \frac{1}{2} o_{max} \\ 1, & \text{if } \Delta_{ij} < \frac{1}{2} o_{max} \quad \text{and} \quad a_{ij}^T > \frac{1}{2} o_{max} \end{cases} \qquad 1 \le i \le S,\ 1 \le j \le p \tag{5.2}$$

The distance between components of a vector is then calculated using a distance measure that is a generalization of the Hamming distance or discrete metric. This metric satisfies all the standard properties of a distance metric. The distance between two activity vectors is calculated by summing the difference between each corresponding pair of components. A symmetric distance matrix, D, is generated whose elements represent the distances between pairs of final activity vectors (Equation 5.3):

$$D_{ij} = \sum_{k=1}^{p} d\left[\tilde{a}_{ik}, \tilde{a}_{jk}\right], \quad 1 \le i, j \le S, \tag{5.3}$$

$$d\left[\tilde{a}_{ik}, \tilde{a}_{jk}\right] = \begin{cases} 0, & \text{if } \tilde{a}_{ik} = \tilde{a}_{jk}, \\ 1, & \text{otherwise.} \end{cases} \tag{5.4}$$

The clustering algorithm is initialized with a preliminary set of L cluster centers. These are chosen either by:

(i) randomly generating binary vectors of length p where each component of the vector was chosen at random from $\{0, 1\}$, or

(*ii*) taking a set of cluster centers randomly selected from the data set (i.e., a set of rows from the matrix $\tilde{a}$).

Clustering of high dimensional data is a difficult problem if the number of attractors and shape of data clustering are not known, which is the case here. The algorithm must therefore be tested with different numbers of initial cluster centers.

Once the initial cluster centers are set, the distance measure is used to calculate the closest cluster center to each data point. Each data point is defined by a simplified activity vector, denoted by $\tilde{\mathbf{a}}_i$, where $1 \leq i \leq S$ (see Equation 5.2). This data point is then added to the cluster nearest to it. When all the data points have been assigned to a cluster (call it cluster l, say) the corresponding cluster center vector, $\mathbf{V}_l$, where $1 \leq l \leq L$, is adjusted to an estimate of the average of the data points within that cluster. For example, if the majority of points in the cluster have a component equal to zero, then the corresponding component of the cluster vector is set to zero. If equal numbers of points have the value 0, 1 or X then the value of the corresponding component of the cluster center is selected at random from the set {0,1,X}.

For each clustering, an error measure is defined which indicates how well the clustering algorithm is performing. The lower the error, the better the clustering. The error of a vector in a cluster is given by its distance from the cluster center. The error for cluster l, E_l, is defined as the sum of the distances of each data point in the cluster to the cluster center vector $\mathbf{V}_l$. The size of cluster l is denoted c_l. The total error E of the clustering algorithm is defined by Equation 5.5. A new permuted list of activity vectors is generated from the clustering algorithm. The new matrix $\hat{a}$ contains the vectors regrouped by cluster. It is essentially a matrix obtained by permuting the rows of matrix $\tilde{a}$.

$$E = \sum_{l=1}^{L} E_l, \quad \text{where} \quad E_l = \sum_{k=1}^{c_k} \sum_{m=1}^{p} d\left[V_{lm}, \hat{a}_{km}\right] \tag{5.5}$$

A data point (representing a modified final activity vector) is chosen at random and the effect of moving it into a new cluster is estimated. The data point is moved into a new cluster only if the total error measure decreases. The cluster selected is the one which results in the largest decrease in the value E. This process continues until the error E reaches an asymptotic value.

The data points are thus regrouped by cluster. The distance matrix of Equation 5.3 is recalculated. The clustering algorithm results in a permutation of the original distance matrix D. Other information on inter and intra cluster point distances are also computed. Although the algorithm is designed to find the best clustering using a fixed number of cluster centers, by systematically varying the number of cluster centers it is possible to estimate the value of L which provides the best clustering of the data.

The final clustering provides information on the similarities in the data set. In particular, the smaller the number of cluster centers, the greater the likelihood of the data falling into a small number of highly similar activity states. This provides evidence of a restriction in the possible state space that the network can occupy, and thus an indication of the possibility that the network may settle into a relatively small number of stable attractor states.

5.3 Performance of the Clustering Algorithm

In this section a specific cell type is examined in some detail and the results of the clustering algorithm are presented. The GA is run for five hundred generations with its parameters set as follows: $[G, p_m, p_c, C, F] = [500, 0.1, 0.1, 200, 50]$. Every daughter cell undergoes crossover (one pair of proteins, chosen at random, exchange regulatory domains) and mutation (on average two mutations per generation). Each cell generates three offspring for the first five generations, and during this time fifty randomly generated cells are also introduced into the population. After this time,

every parent cell generates four offspring. A maximum of three cells are re-introduced at each generation. Individual cell simulations are performed with the following parameter set: $[p_h, p_k, f_h, T, N, S] = [0.10, 0.10, 0.5, 500, 20, 100]$. Each cell contains twenty protein types where each protein has output interactions with two others.

The dynamics of the best cell selected from the final generation of the GA are analyzed from one hundred starting states. The initial random starting states for each of the S (=100) different simulations of this cell are shown in Figure 5.2. Each row represents an activity state vector (e.g., row 1 represents the initial random activity state of the network for simulation 1). Protein types that are inactive are displayed in blue, and protein types that are active are shown in red. The graded color scale is shown alongside this plot. The corresponding final activity states are shown in Figure 5.3.

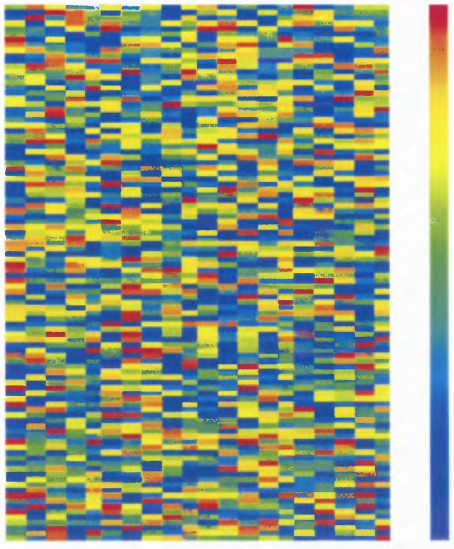


Figure 5.2: One hundred random starting activity states for the fittest GA network. Each state is shown horizontally.

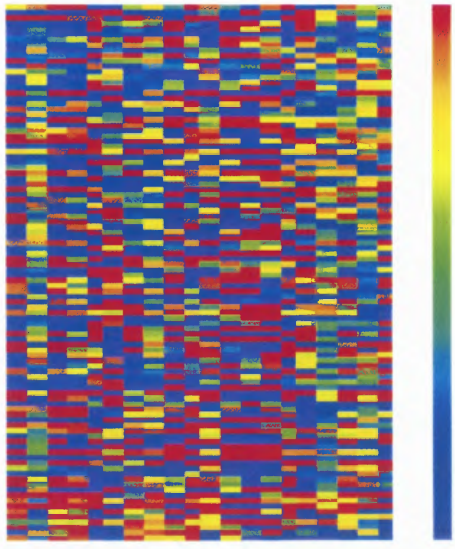


Figure 5.3: Activity states for the same cell after $T = 500$ time steps.

The selected cell has a fitness of 85%. The clustering analysis is therefore performed using seventeen out of the twenty proteins. The proteins which pass the fitness criteria defined in §4.2.1 are used in the clustering method. From the new matrix of states defined from these proteins, $\tilde{a}$, the data are clustered according to the algorithm defined in §5.2.

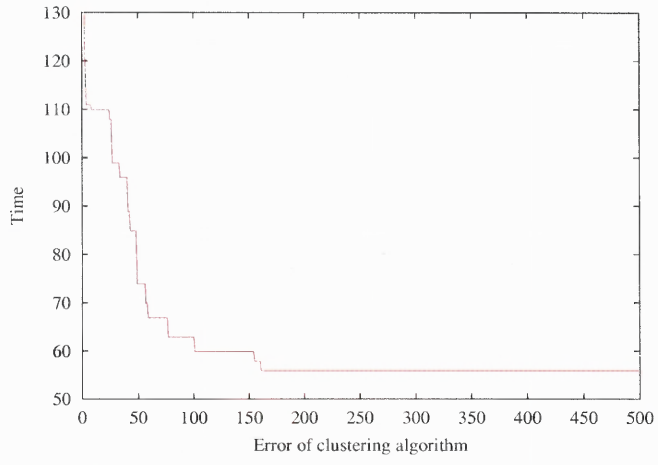


Figure 5.4: The clustering error decreases as the algorithm progresses

The change in error of the clustering algorithm is shown in Figure 5.4. A good cluster separation is obtained with five cluster centers (see Table 5.1). The number of data points grouped into each cluster is indicated by a number beside the corresponding row of the distance matrix. The number seven beside row one (denoted [1: 7]) indicates that seven data points are grouped into cluster one. Each data vector has dimension seventeen (i.e., seventeen proteins, of a possible twenty, from the cell selected pass the fitness criteria).

Table 5.1A is a distance matrix comparison, $\hat{D}_{ij}$, of the average distance between points in cluster i from cluster j, calculated as the distance from the cluster center j. Distances are rounded to the nearest integer and the maximum distance between end states is seventeen. The data show that the distance between data points in the same cluster is low (diagonal elements). The average distance between data points in cluster i to cluster j, $j \neq i$, is higher (off-diagonal elements). Note that

Table 5.1: Data from the analysis of the clustering algorithm (§5.2). One hundred final activity states are compared from a selected cell taken from the final generation of a GA simulation. The simulation has the following cell parameters: $[p_h, p_k, f_h, T, N, S] = [0.10, 0.10, 0.5, 500, 20, 100]$. The GA parameters are: $[G, p_m, p_c, C, F] = [500, 0.1, 0.1, 200, 50]$. Good separation is obtained with five cluster centers. The data in (A) are evidence for distinct attractor states in the state space. Table A is a distance matrix comparison, $\hat{D}_{ij}$, of the average distance between points in cluster i to cluster center j. Distances are rounded to the nearest integer and the maximum distance between end states is seventeen. The corresponding cluster centers are listed in (B), together with the average distance of points in the cluster from the cluster center. An "x" denotes that the protein is fluctuating, a zero denotes that the protein is "OFF", and a one denotes that the protein is "ON".

(A) Intra- and Inter-Cluster Distances:

		(1)	(2)	(3)	(4)	(5)
Cluster 1	(1: 7)	1	5	5	15	14
Cluster 2	(2: 68)	5	0	5	17	17
Cluster 3	(3: 8)	5	4	1	15	15
Cluster 4	(4: 8)	15	17	15	1	17
Cluster 5	(5: 9)	14	17	15	16	0

(B) Symbolic Representation of Cluster Centers:

		Cluster Center Vector	Avg. Dist. to Cl. Center
Cluster Center	(1)	`xxxx0xxxx0x0xx00x`	0.71
Cluster Center	(2)	`xxxxxxxxxxxxxxxx`	0.47
Cluster Center	(3)	`xxxx1xxxx1x1xxx1x`	1.00
Cluster Center	(4)	`11110110110001111`	1.38
Cluster Center	(5)	`00001001001110000`	0.00

some of the clusters are relatively close (see Table 5.1B). For example, cluster centers 2 and 3 are four units of distance apart, but clusters 2, 4 and 5 are maximally distant from one another.

The corresponding cluster centers are listed in Table 5.1B, together with the average distance of points in the cluster from the cluster center. An "**x**" denotes that the protein is fluctuating, a zero denotes that the protein is inactive ("OFF") and a one denotes that the protein is considered to be active ("ON"). Clusters 4 and 5 contain stable points, with all proteins either "ON" or "OFF". In fact, all proteins have precisely opposite characteristics in these two clusters. A protein which is "ON" in cluster four is "OFF" in cluster five (and vice versa). Although the components of these two cluster centers are in opposite states, this is not generally the case. In cluster 2 all the proteins selected are oscillating. Clusters 1 and 3 are similar to cluster 2, but several of the proteins have stable values.

The clustering process can only identify similarities in composition of dynamics at the level of "ON" or "OFF" or "oscillating" (i.e., it cannot detect differences in phase or frequency between oscillating proteins). Low diagonal elements and high off-diagonal elements indicate that points in cluster i are generally not close to points in the other clusters. These are precisely the properties that one would expect for a reasonable clustering of the data.

Figure 5.5 illustrates the symmetric distance matrix, D, prior to clustering. In this example, the maximum distance between any two state vectors is 17. A graded color scale is shown to the right. Vector pairs that are maximally distant are shown with a red pixel, and those that are highly similar are shown in blue. After the clustering algorithm has completed, (see Figure 5.6), the data points have been regrouped into the clusters shown in Table 5.1. The data indicate that the algorithm is extremely effective.

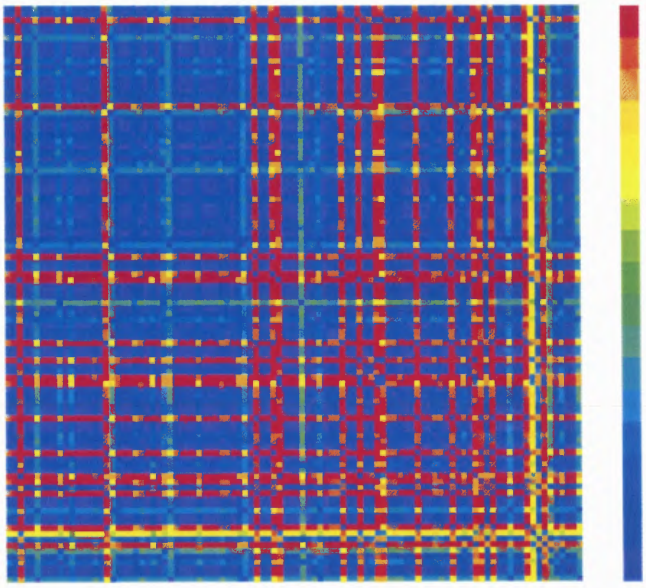


Figure 5.5: Distance matrix comparison of end states, D_{ij}, $1 \leq i, j, \leq S$, prior to clustering.

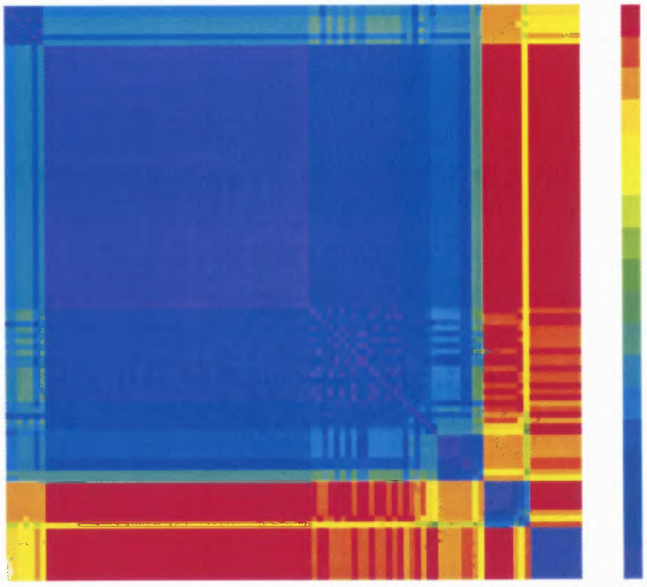


Figure 5.6: Distance matrix after clustering.

5.4 Numerical Evidence for Dynamic Attractors

The evidence that a single cell can settle into a variety of patterns of stable activity according to the initial signal provided to the cell was seen earlier in §3.3. Table 5.1 illustrates data for a particular cell generated by the genetic algorithm and highlighted the similarities between cellular states. In this section, data points from two of the clusters described in the previous section will be shown in more detail. Figure 5.7 shows the dynamics from two different starting states which are grouped into cluster five. Cluster center five is denoted, for the purposes of the cluster analysis, as "00001001001110000". Only 17 of the proteins are used for the clustering process. If the proteins are labeled in order, 0 to 19, the proteins not used in the clustering are numbered 1,16,18 The latter are marked with arrows in Figure 5.7. The actual activity vector of the network is represented by "0*00010010011100*0*0".

The left hand plots of Figure 5.7 denote the change in activity of the proteins over time during simulations from two different randomly generated starting states. In each case, the network settles into the same stable steady state. Protein activity is shown horizontally and time is plotted downwards. A horizontal snapshot of the network therefore shows the instantaneous activity of the proteins in the network. The right hand plots indicate the change in summed activity of all kinases (solid line) and all phosphatases (dashed line). The format of the data is the same as in Figure 3.2. The initial transient phase is different in each case, but the network settles into the same state after approximately two hundred time steps. In the lower figure, the network seems to start with an oscillation which eventually diminishes to the steady state.

Figure 5.8 shows data from two different starting states which settle into the same pattern of oscillation. The data shown here are from the same cell as in Figure 5.7. The dynamics are different because the network settles into a limit cycle oscillation in a different part of the state space. Again, the initial transient

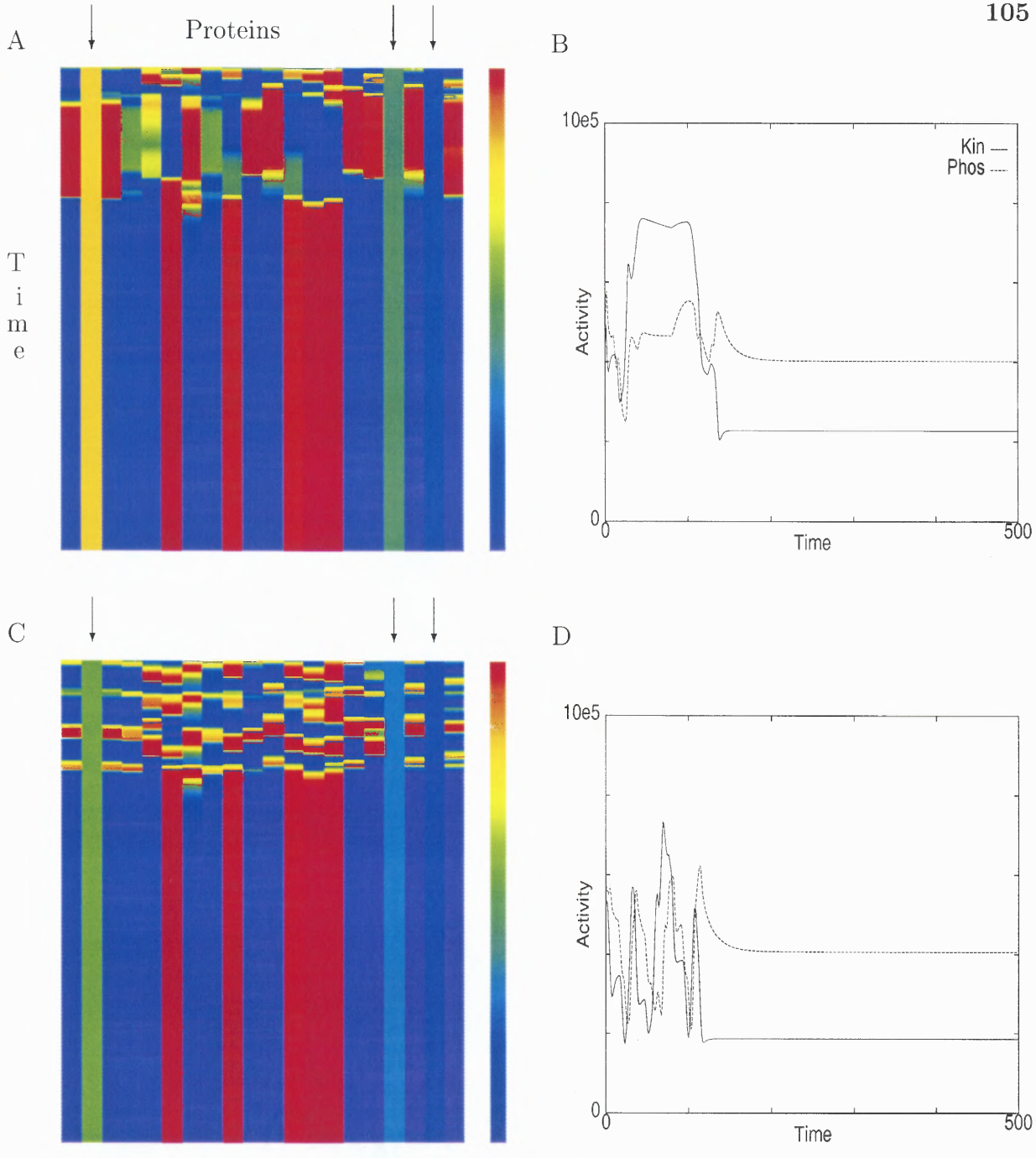


Figure 5.7: A cell settles into the same stable steady state configuration from two different starting states. Diagrams (A) and (C) illustrate the change in activity of the network of proteins over a simulation of $T = 500$ time steps. Protein activities are illustrated horizontally and the time axis is shown vertically. Time increases in the downward direction. Proteins with high activity values are shown in red and inactive proteins are shown in blue. The color scale is graded in-between. Plots (B) and (D) show the total summed activity of the protein kinases (solid line) and the protein phosphatases (dashed line) in the cell against time. The network falls into the same fixed point attractor, despite the difference in initial conditions. Both these states are found in cluster group five shown in Table 5.1.

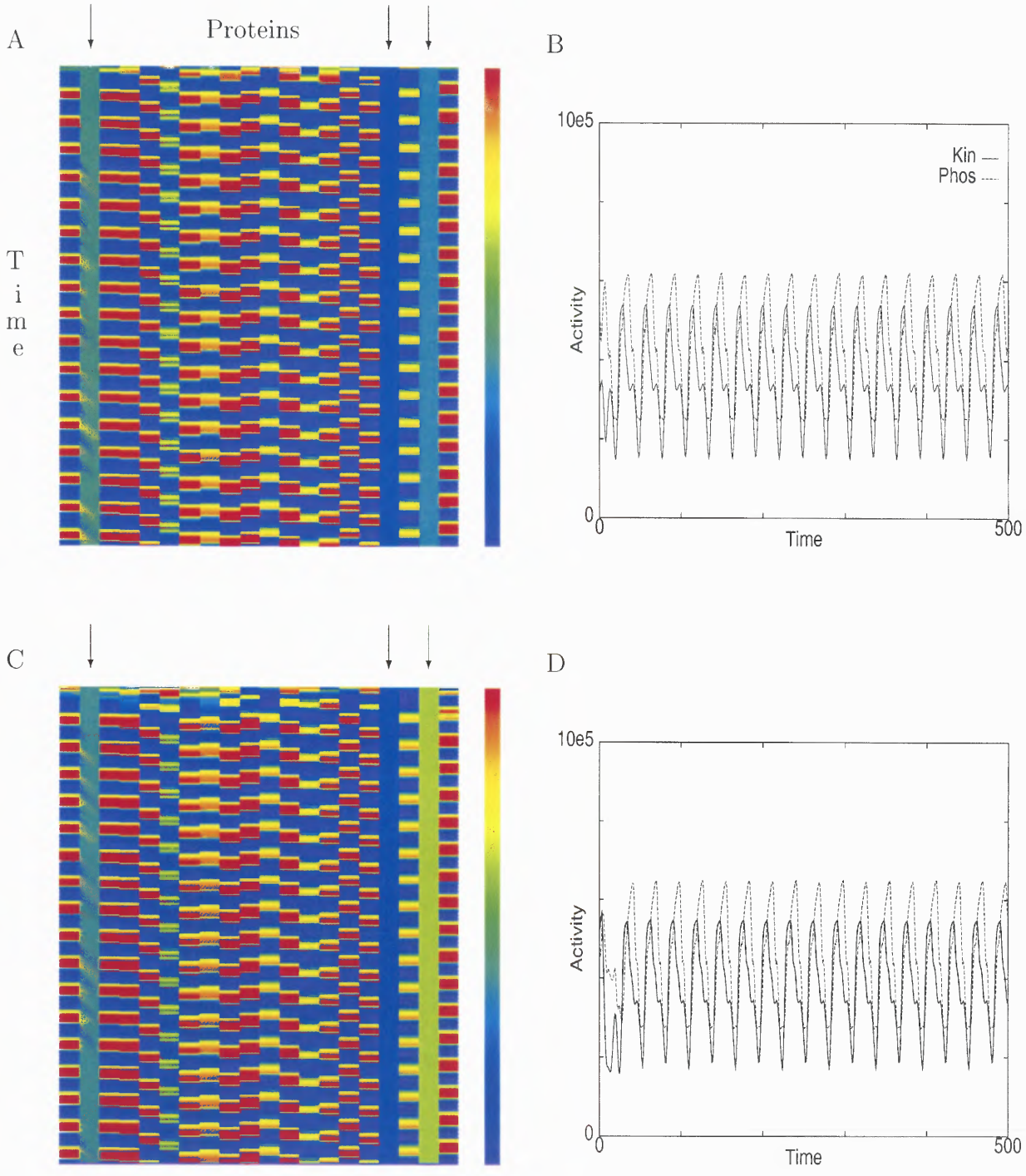


Figure 5.8: The same cell (as in Figure 5.7) settles into a similar characteristic stable oscillation from two different random starting states. The figure has the same format as Figure 5.7. Some proteins in the network have a stable activity level, but the remainder show a periodic oscillation in activity. Both these states fall into cluster group two shown in Table 5.1. The oscillation repeats every 28 time steps.

is different and is also short-lived as the network settles into the oscillation after less then 50 time steps. Note that the summed activity of the kinases and phosphatases are very similar. The period of the oscillation is 28 time steps. These data points are selected from cluster two. The proteins not included in the clustering process are also marked with arrows. The actual activity vector of the network is represented by "x*xxxxxxxxxxxxxx*x*x".

In fact, it is possible to flip between these states by forcing an internal change in the network (see Figure 5.9). Activation of a single node inside the network forces a change in the state of the entire network. Many nodes that were inactive become activated, and vice versa. This is because this node is highly connected. This point will be discussed in Chapter 7 in light of structural analysis of this network using graph theoretic methods. At time step t=500 an oscillation is forced in node 19. This causes a switch to a state where all the nodes are oscillating (excluding the three input nodes labelled as 1, 16 and 18) which is similar to that shown in Figure 5.8. Interestingly, a subsequent return of node 19 to an inactive state does not return the network to its original state. The network has undergone a permanent shift of state arising from a transient change in one of the internal nodes.

Figure 5.10 shows data from three different starting states which settle into the same pattern of oscillation. The data here are from the same cell as Figures 5.7 and 5.8. These data points are also grouped into cluster two. This is a good illustration of the fact that the clustering algorithm only captures some of the features of the data. The right hand plots show the similarity in the oscillation pattern in the three cases. Although the periodicity is not identical in each case, the network shows a period of relative inactivity where all proteins appear to maintain a constant level of activity. This is followed by a period of change until the network returns to the previously stable phase. The oscillation period is approximately 180 time steps in the upper and lower plots, but the central data plot appears to have a slightly longer

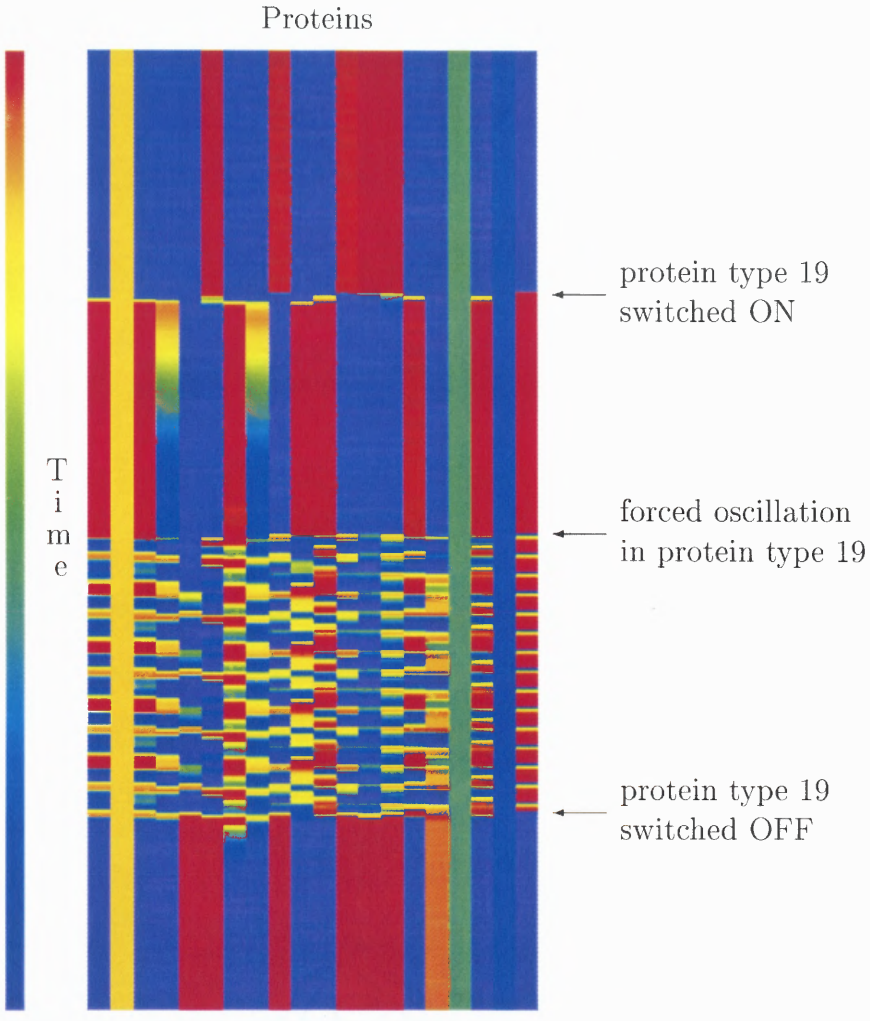


Figure 5.9: A switch between attracting states can be induced by an internal change in the network. The color map is shown to the left. At time step 250, node 19 is switched ON. At a later time (t=500), an oscillation in node 19 can force a shift in the entire network to a different state. Once this node is shut off (t=790), the network returns to a steady state. Interestingly, the network does not return to its initial state after this transient internal change.

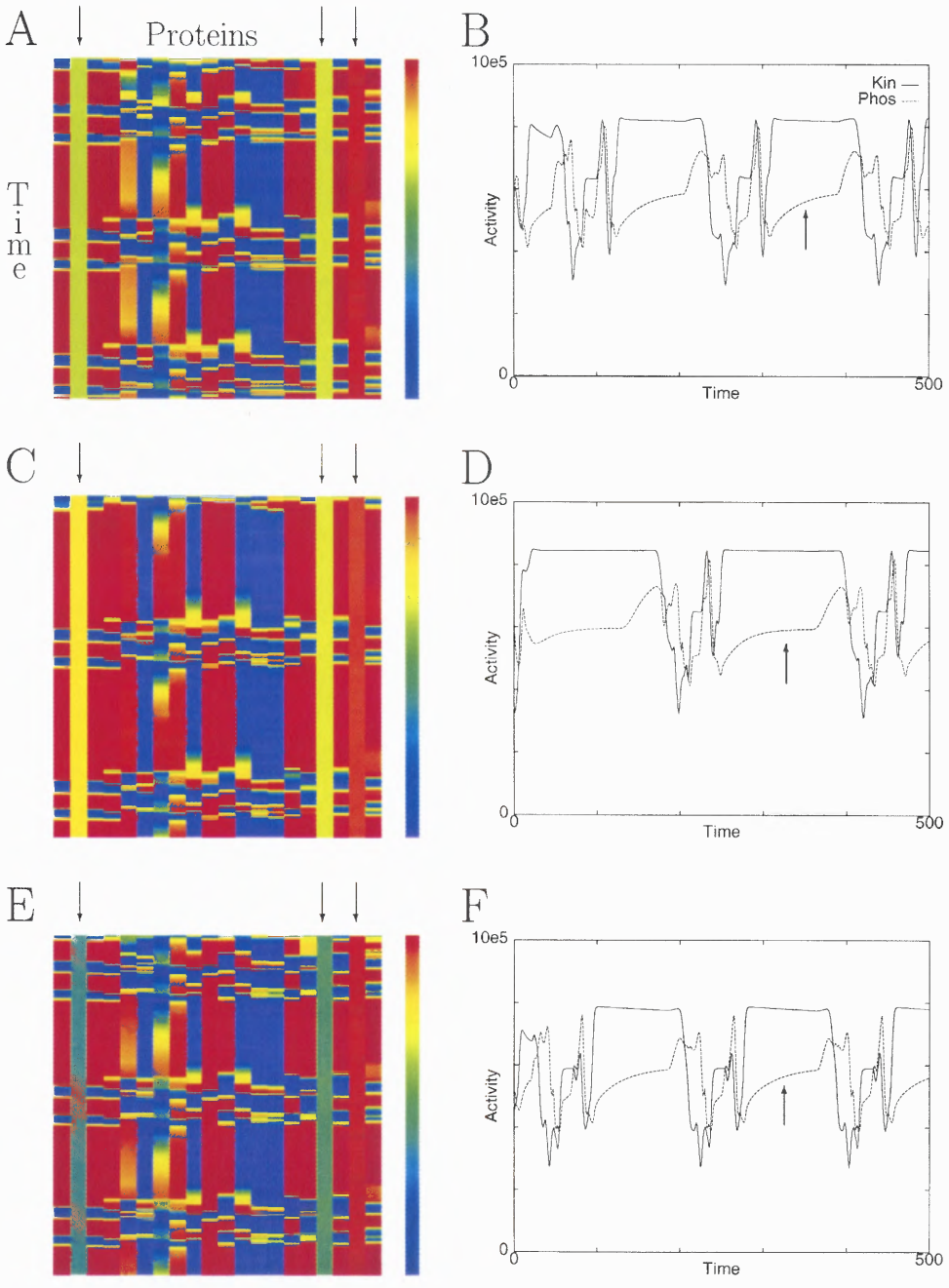


Figure 5.10: A cell can have many different oscillatory states. The same cell settles into an oscillation of different frequency to that shown in Figure 5.8. The cell is shown starting from three different random initial conditions settling into a new pattern of oscillation. These data points are also classified into cluster group two shown in Table 5.1. The center diagrams (C,D) show a slightly longer period of oscillation, ≈ 220 time steps (compared to ≈ 180 time steps in the upper and lower plots). The period of the oscillation appears to be governed by the time for the phosphatase activity to reach a critical value, denoted by arrows on the right hand plots.

period of around 220 time steps. There is a period of time (denoted by an arrow on the right hand plots) which determines when the stable phase of the oscillation ends. The phosphatase activity reaches a critical value, triggering the change, allowing the oscillation to complete.

Figures 5.7 to 5.10 demonstrate that a single network can settle into the same steady state configuration from a different starting state. The cell can also settle into different patterns of oscillatory behavior from different points in the state space. The dynamics show a degree of stability, in that certain changes in initial conditions do not affect the network behavior, yet other changes do alter the nature of the stability. For example, the cell can flip from a steady state configuration to a limit cycle, or change from one limit cycle oscillation to another. Figure 5.11 illustrates the dynamics of three proteins selected from the cell whose dynamics are shown in Figures 5.7 to 5.10. The relative activities of these proteins (labeled as 2,5,17) are shown using a three-dimensional phase space portrait. The phase portraits in Figure 5.11A,B are from the two simulations of Figure 5.8. The phase space plots in Figure 5.11C,D are from the two simulations of Figure 5.7. The proteins settle to a steady state value, which is indicated with an arrow. The plots of Figure 5.11E,F,G are from the three simulations of Figure 5.10. Data from the same attractor are therefore shown horizontally, and data from different attractors are shown vertically. There is a striking similarity in the horizontal plots and a distinct difference between the vertical plots in Figure 5.11A,C,E. These phase space portraits provide a useful way of distinguishing between different states in the cellular dynamics. As mentioned in §3.4, the ATP plots for each of the cases shown in Figure 5.11 are also considered as a useful way of distinguishing the intracellular states (see Figure 5.12).

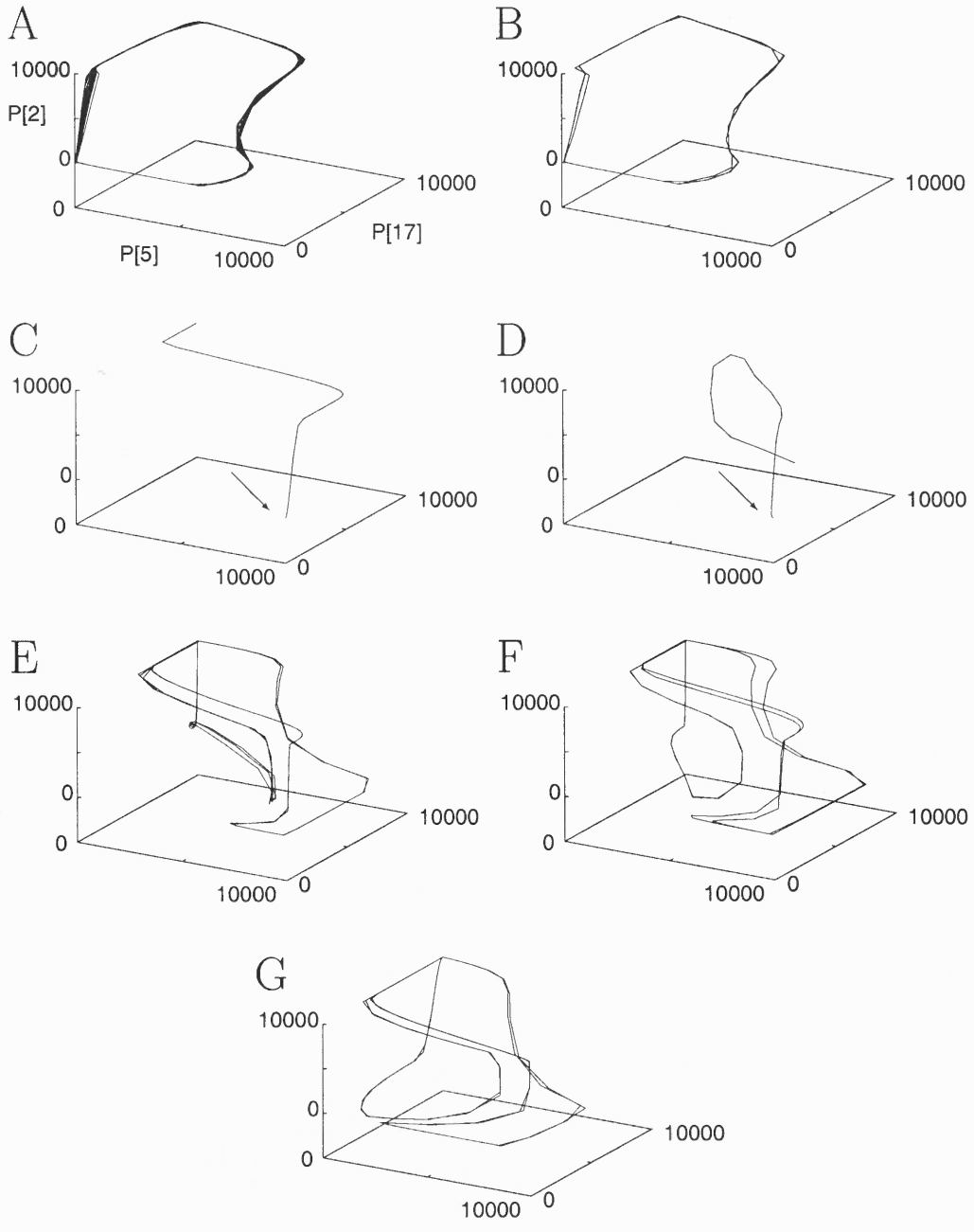


Figure 5.11: Phase space portrait of proteins in a network selected by the GA. The relative activities of three proteins (labeled as 2, 5, 17) are shown for the seven cases that are illustrated in Figures 5.7 to 5.10. (A), (B) are the protein activities corresponding to the two simulations of Figure 5.8. (C), (D) are from the two simulations of Figure 5.7. The proteins settle to a steady state value which is indicated with an arrow. (E),(F),(G) show data from the three simulations of Figure 5.10. This figure illustrates how relative protein activities differ when their dynamics are compared between clusters (vertically) and within the same cluster (horizontally).

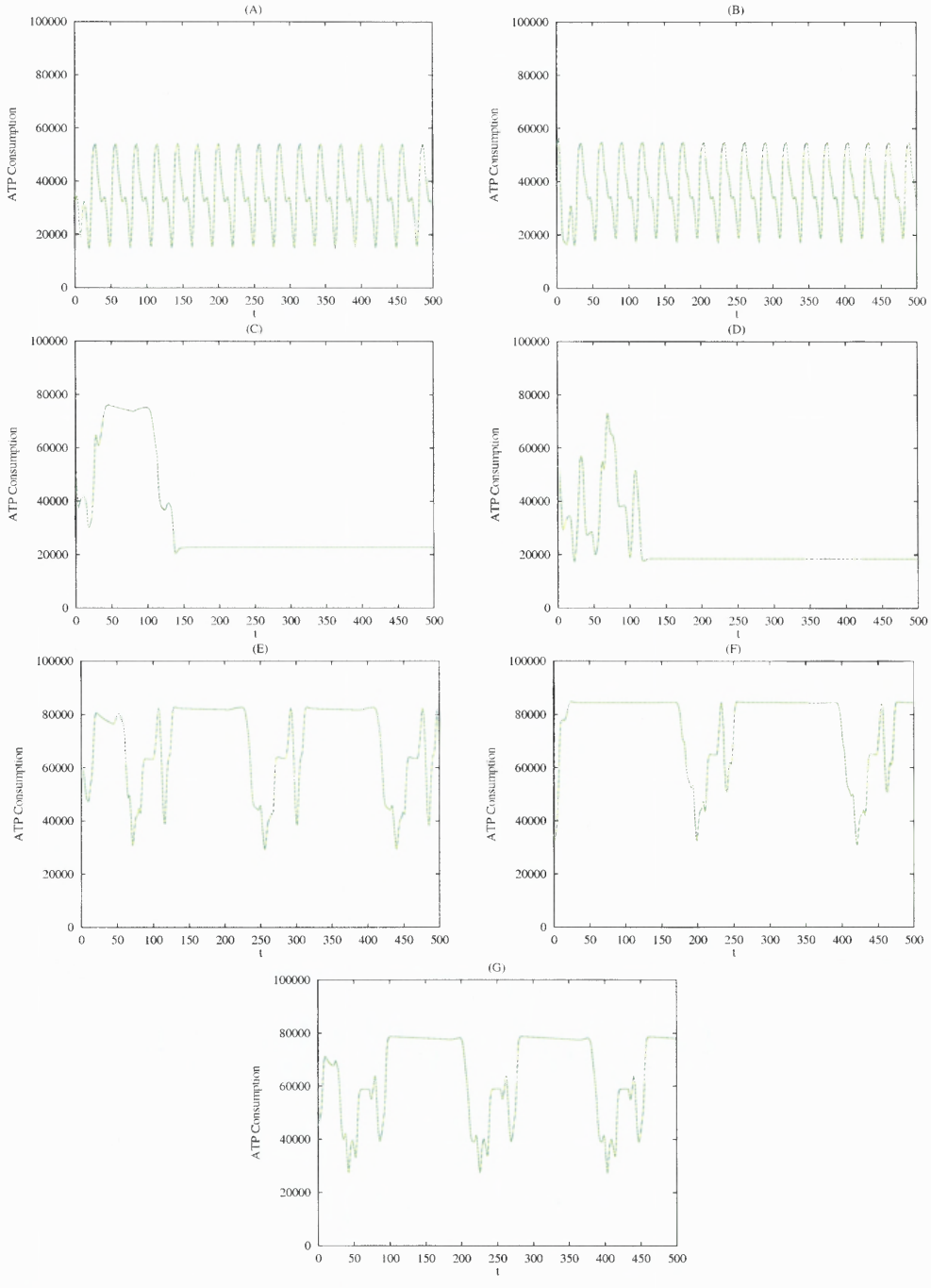


Figure 5.12: Plot of ATP consumption against time for the cases shown in Figure 5.11. The plots are organized in the same respective order as in Figure 5.11.

5.5 Summary

In this section a clustering algorithm is developed to analyze the dynamic states of simulated protein networks. The algorithm modifies the activity states of proteins and groups them into three different categories: (*i*) "ON", (*ii*) "OFF", or (*iii*) "OSCILLATING". Thus, activity states are represented as vectors whose components take the symbolic values "0", "1" and "X", respectively. A distance measure is introduced (modified Hamming distance/discrete metric) in order to evaluate similarities between data points. The error of the clustering algorithm is also defined in order to measure its performance. The notational symbols used in the clustering algorithm are summarized in the "List of Symbols" pages at the start of this thesis.

The clustering algorithm is used to evaluate the performance of the GA. Numerical work shows that the genetic algorithm has successfully selected for networks which display a relatively small number of distinct attractor basins. Despite widely different initial conditions, the network settles into a small number of different patterns of protein activities. The properties of GA networks will depend heavily on the nature and design of the fitness function.

Figure 5.9 indicates that a transient internal change in a network can have a significant impact on dynamics. In fact, in this example, a transient oscillatory change in a single internal node shifts the network from one equilibrium state to another. In fact, the new state of the network appears to be different from the states detected by the clustering algorithm. Both these points will be discussed further in the final chapter.

The stability of fixed point and limit cycle behaviors in simulated networks will be examined analytically in the next chapter. The topological structure of the network shown in this chapter will be discussed in Chapter 7. A combination of insights from graph theory techniques and small network analysis are important in order to gain a deeper insight into the dynamic properties of simulated networks.

CHAPTER 6

SMALL NETWORK ANALYSIS

6.1 Introduction

In the model developed in Chapter 2, the activity state of any node represents the instantaneous activity of a discrete number of molecules of any given protein type. Since a molecule that is 90% bound is still not fully bound, a decision was made to truncate the occupancy variable at each step (see Equation 2.6, §2.2.2). The numerical results presented in Chapters 3 through 5 are, therefore, based on an updating rule which involves discretization at each time step.

The updating rule for a node j in a simulated network, $1 \leq j \leq N$, is summarized below for reference:

$$\begin{aligned} b_j^{(t)} &= \sum_{i=1}^{N} a_i^{(t)} M_{ij}, \\ o_j^{(t+1)} &= \lfloor f(o_j^{(t)}, b_j^{(t)}) \rfloor, \\ &= \begin{cases} \lfloor o_{max}\{1 - e^{-b_j^{(t)}/o_{max}}\} + o_j^{(t)} e^{-b_j^{(t)}/o_{max}} \rfloor, & \text{if } b_j^{(t)} \geq 0, \\ \lfloor o_j^{(t)} e^{b_j^{(t)}/o_{max}} \rfloor, & \text{if } b_j^{(t)} \leq 0, \end{cases} \\ a_j^{(t)} &= g\{o_j^{(t)}\} = o_j^{(t+1)}. \end{aligned} \tag{6.1}$$

In this chapter, this truncated updating rule will be denoted by

$$\mathbf{o}^{(t+1)} = \hat{\mathbf{f}}\left(\mathbf{o}^{(t)}, \mathbf{b}^{(t)}\right) = \left(\hat{f}(o_1^{(t)}, b_1^{(t)}), \ldots, \hat{f}(o_N^{(t)}, b_N^{(t)})\right),$$

where $\hat{f}(o_j^{(t)}, b_j^{(t)}) = \lfloor f(o_j^{(t)}, b_j^{(t)}) \rfloor$, $\forall j$. Therefore, at the start of each of the numerical simulations presented in the aforementioned chapters, the variable $o_j^{(t)}$ is set to a random integer value in the range $0 \leq o_j^{(t)} \leq o_{max}$, where o_{max} is currently set to $10,000$. The function f is applied to each node and the resulting real value obtained is then truncated to the nearest integer. The numerical function $\hat{f}$ obtained is, therefore, a function acting on a set of integers: $\hat{f} : ([0, o_{max}] \cap \mathbb{Z}) \times \mathbb{Z} \to [0, o_{max}] \cap \mathbb{Z}$.

The properties of the non-discretized updating rule, f, are examined in this chapter

$$\mathbf{o}^{(t+1)} = \mathbf{f}\left(\mathbf{o}^{(t)}, \mathbf{b}^{(t)}\right) = \left(f(o_1^{(t)}, b_1^{(t)}), \ldots, f(o_N^{(t)}, b_N^{(t)})\right),$$
$$\text{where} \qquad f : [0, o_{max}] \times \mathbb{R} \to [0, o_{max}].$$

The reason for studying properties of the non-discretized rule f is because the discretized rule is not easily amenable to analysis with dynamical systems methods.

Numerical results are presented in this chapter which illustrate the strong similarities in the dynamic properties of both these rules. It is argued in light of these data that analytical results of the non-discretized mapping, f, are useful to provide insight into the discretized numerics (using $\hat{f}$). The remainder of this chapter is concerned with proving results regarding the stability of persistent dynamic states of some small networks (when the non-discretized updating rule f is used).

When the input to any given node in the network is only positive or negative, networks can be examined easily for fixed points. Their local stability is established with the use of linear stability analysis (see §6.2.1). However, the cases where the inputs for each node can vary between positive and negative values over time require more care. In these instances, it is difficult to study the stability of fixed points with f written in the form shown in Equation 6.1. In the basic model studied in this thesis, where g is the identity function, the non-discretized rule f can be re-written in a simplified form:

$$f(o_j^{(t)}, b_j^{(t)}) = H(b_j^{(t)}) \left\{ o_{max} \left[1 - e^{-b_j^{(t)}/o_{max}} \right] + o_j^{(t)} \left[e^{-b_j^{(t)}/o_{max}} - e^{b_j^{(t)}/o_{max}} \right] \right\} + o_j^{(t)} e^{b_j^{(t)}/o_{max}}, \tag{6.2}$$

where $H(x)$ represents the Heaviside function:

$$H(x) = \begin{cases} 0, & x \leq 0, \\ 1, & x > 0. \end{cases}$$

Another approximation to f (and referred to in this chapter as f_{tanh}) is introduced which smoothes out the discontinuity in its derivative:

$$f_{tanh}(o_j^{(t)}, b_j^{(t)}) = 0.5\left\{1+\tanh(\kappa b_j^{(t)})\right\}\left[o_{max}(1-e^{-b_j^{(t)}/o_{max}})+o_j^{(t)}e^{-b_j^{(t)}/o_{max}}\right]+ 0.5\left\{1-\tanh(\kappa b_j^{(t)})\right\}o_j^{(t)}e^{b_j^{(t)}/o_{max}}. \quad (6.3)$$

The parameter κ controls the steepness of the tanh function. In the limit $\kappa \to \infty$ the function $\frac{1}{2}[1+\tanh(\kappa x)]$ approaches $H(x)$. Results will show that this enables the resolution of these difficulties. Numerical data for the three different versions of the dynamical rule (f, $\hat{f}$, and f_{tanh}) show good agreement (see §6.4).

A four node network which exhibits simple periodic behavior under certain parameter regimes is discussed in some detail. The usefulness of analysis in the phase plane for interpreting features of the dynamics will be evident. Analysis is used to verify the numerical results for a specific parameter case. The locally attracting property of the limit cycle will be proved with the aid of the Banach fixed point theorem [97]. Global stability of the limit cycle is conjectured. The variation in periodicity of this limit cycle is also examined for a number of different parameter cases. Data relating to this four protein network will be discussed again in Chapter 7 in the context of a larger, more realistic protein network generated by the GA.

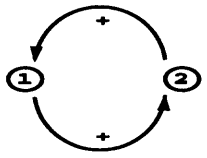
6.2 Examples of Fixed Point Stability Analysis.

In this section, a linear stability analysis is used to examine the local stability of fixed points of some simple networks. The non-discretized updating rule, f, is used for this analysis. Analysis is shown in §6.2.1 for a network with two protein types. A network with four protein types is considered in §6.2.2.

6.2.1 A Network with Two Protein Types: 2PKs

The network examined in this section contains two nodes which represent modeled protein kinases. The connection matrix for this network is also shown (see Figure 6.1).

Graphical Representation



Connection Matrix Representation

$$M_{ij} = \begin{pmatrix} 0 & +1 \\ +1 & 0 \end{pmatrix}$$

Figure 6.1: A two node graph.

In this example, the non-discretized updating rule, f, takes the form

$$\left.\begin{aligned} o_1^{(t+1)} &= o_{max}\left(1 - e^{-b_1^{(t)}/o_{max}}\right) + o_1^{(t)} e^{-b_1^{(t)}/o_{max}}, \\ o_2^{(t+1)} &= o_{max}\left(1 - e^{-b_2^{(t)}/o_{max}}\right) + o_2^{(t)} e^{-b_2^{(t)}/o_{max}}. \end{aligned}\right\} \tag{6.4}$$

The b_j $(j = 1, 2)$ terms are found by computing the weighted sum of inputs for each node (see §6.1). The fact that g represents the identity function in the simple model, means that activity and occupancy are the same (i.e., $a_j^{(t)} = o_j^{(t)}$, $\forall t$). Hence, $b_1^{(t)} = a_2^{(t)} = o_2^{(t)}$ and $b_2^{(t)} = a_1^{(t)} = o_1^{(t)}$. Here, the updating rule also simplifies because $b_1^{(t)} \geq 0$, $\forall t$ and $b_2^{(t)} \geq 0$, $\forall t$.

Fixed points of the system satisfy $o_1^{(t+1)} = o_1^{(t)} = \hat{o}_1$ and $o_2^{(t+1)} = o_2^{(t)} = \hat{o}_2$. The following system of equations must then be solved:

$$\left.\begin{aligned} (\hat{o}_1 - o_{max})(1 - e^{-\hat{o}_2/o_{max}}) &= 0, \\ (\hat{o}_2 - o_{max})(1 - e^{-\hat{o}_1/o_{max}}) &= 0. \end{aligned}\right\} \tag{6.5}$$

The two fixed points lie at $(\hat{o}_1, \hat{o}_2) = (0, 0)$ and $(\hat{o}_1, \hat{o}_2) = (o_{max}, o_{max})$.

The local stability of these points is considered next. This is examined by making a small perturbation, away from the fixed point $(\hat{o}_1, \hat{o}_2)$:

$$\left.\begin{aligned} o_1^{(t)} &= \hat{o}_1 + \epsilon x^{(t)}, \\ o_2^{(t)} &= \hat{o}_2 + \epsilon y^{(t)}. \end{aligned}\right\} \tag{6.6}$$

Each of the fixed points is examined in turn by substituting the expressions from Equation 6.6 into Equation 6.4. The eigenvalues of the resulting linearized system, in the new variables $(x^{(t)}, y^{(t)})$, are then calculated.

Analysis of fixed point at $(0, 0)$**:**

$$\left.\begin{aligned} \epsilon x^{(t+1)} &= o_{max} - (o_{max} - \epsilon x^{(t)}) e^{-\epsilon y^{(t)}/o_{max}} \\ \epsilon y^{(t+1)} &= o_{max} - (o_{max} - \epsilon y^{(t)}) e^{-\epsilon x^{(t)}/o_{max}} \end{aligned}\right\} \tag{6.7}$$

Expanding the exponential terms in a Taylor series gives

$$\left.\begin{aligned} \epsilon x^{(t+1)} &= o_{max} - (o_{max} - \epsilon x^{(t)}) \left[1 - \frac{\epsilon y^{(t)}}{o_{max}} + \ldots\right], \\ \epsilon y^{(t+1)} &= o_{max} - (o_{max} - \epsilon y^{(t)}) \left[1 - \frac{\epsilon x^{(t)}}{o_{max}} + \ldots\right]. \end{aligned}\right\} \tag{6.8}$$

After grouping terms of similar orders, only the $O(\epsilon)$ terms are retained. The $O(1)$ terms balance and terms of higher order are neglected. The eigenvalues of the resulting linear system in Equation 6.9 (denoted λ_i), determine the local stability of the fixed point:

$$\begin{pmatrix} x^{(t+1)} \\ y^{(t+1)} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x^{(t)} \\ y^{(t)} \end{pmatrix}. \tag{6.9}$$

In this example, the eigenvalues are found to be $\lambda_1 = 0$ and $\lambda_2 = 2$, indicating that the fixed point at $(0, 0)$ is *unstable* to small perturbations.

Analysis of fixed point at (o_{max}, o_{max}):

$$\left.\begin{aligned} o_{max} - \epsilon x^{(t+1)} &= o_{max} - \left\{o_{max} - (o_{max} - \epsilon x^{(t)})\right\} e^{-(o_{max} - \epsilon y^{(t)})/o_{max}} \\ o_{max} - \epsilon y^{(t+1)} &= o_{max} - \left\{o_{max} - (o_{max} - \epsilon y^{(t)})\right\} e^{-(o_{max} - \epsilon x^{(t)})/o_{max}} \end{aligned}\right\} \tag{6.10}$$

As before, expanding the exponential terms in a Taylor series yields

$$\left.\begin{aligned} x^{(t+1)} &= x^{(t)} e^{-1} \left[1 + \frac{\epsilon y^{(t)}}{o_{max}} + \ldots\right], \\ y^{(t+1)} &= y^{(t)} e^{-1} \left[1 + \frac{\epsilon x^{(t)}}{o_{max}} + \ldots\right]. \end{aligned}\right\} \tag{6.11}$$

The resulting linear system (Equation 6.12) follows from equating the coefficients of the order ϵ terms. Writing $\alpha = e^{-1}$,

$$\begin{pmatrix} x^{(t+1)} \\ y^{(t+1)} \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix} \begin{pmatrix} x^{(t)} \\ y^{(t)} \end{pmatrix}. \tag{6.12}$$

This time the matrix is diagonal, so the eigenvalues can be read directly from the diagonal elements of the matrix.

$$\lambda_1 = \lambda_2 = \alpha \tag{6.13}$$

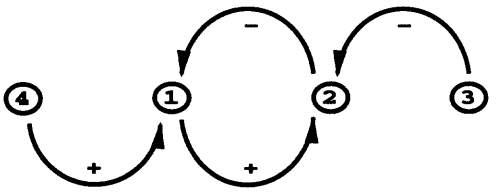
In this example, the eigenvalues are equal and $|\lambda_i| < 1$, $i = 1, 2$. Therefore, the fixed point at (o_{max}, o_{max}) is *stable* to small perturbations.

To summarize, this two node network (where both nodes represent protein kinases) has two equilibrium states. The trivial state, where neither node is active, is locally unstable. The state where both nodes are maximally active is locally stable. Numerics indicate that the fixed point at $(0, 0)$ is globally unstable and the fixed point at (o_{max}, o_{max}) is globally stable. Similar stability analysis results for other small networks with monotonic node input can be found in Appendix C.

6.2.2 A Network with Four Protein Types: 2PK, 2PP

In the case where the input for any given node may vary between positive and negative values over time, the analysis becomes more involved. In this section a simple four protein network is considered (see Figure 6.2).

Graphical Representation



Connection Matrix Representation

$$M_{ij} = \begin{pmatrix} 0 & +1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ +1 & 0 & 0 & 0 \end{pmatrix}$$

Figure 6.2: A small network of four proteins.

The matrix for this network is written alongside, with the labeling of proteins as indicated in the figure. Since the activities of nodes 3 and 4 are fixed by the given initial condition, the values of these nodes will be denoted by the parameters μ_1 and μ_2, respectively. The b_j $(j = 1, 2)$ terms can be found, as before, from the weighted sum of inputs for each node:

$$\left.\begin{aligned} b_1^{(t)} &= \mu_2 - o_2^{(t)}, \\ b_2^{(t)} &= o_1^{(t)} - \mu_1. \end{aligned}\right\} \tag{6.14}$$

Substituting the expressions for $b_1^{(t)}$ and $b_2^{(t)}$ into the non-discretized updating rule f (see §6.1) gives

$$\left.\begin{aligned} o_1^{(t+1)} &= o_1^{(t)} e^{(\mu_2 - o_2^{(t)})/o_{max}} + H(\mu_2 - o_2^{(t)}) \left[o_{max} \left(1 - e^{-(\mu_2 - o_2^{(t)})/o_{max}} \right) \right] \\ &+ H(\mu_2 - o_2^{(t)}) \left[o_1^{(t)} \left(e^{-(\mu_2 - o_2^{(t)})/o_{max}} - e^{(\mu_2 - o_2^{(t)})/o_{max}} \right) \right], \\ o_2^{(t+1)} &= o_2^{(t)} e^{(o_1^{(t)} - \mu_1)/o_{max}} + H(o_1^{(t)} - \mu_1) \left[o_{max} \left(1 - e^{-(o_1^{(t)} - \mu_1)/o_{max}} \right) \right] \\ &+ H(o_1^{(t)} - \mu_1) \left[o_2^{(t)} \left(e^{-(o_1^{(t)} - \mu_1)/o_{max}} - e^{(o_1^{(t)} - \mu_1)/o_{max}} \right) \right]. \end{aligned}\right\} \tag{6.15}$$

By inspection it is possible to find two fixed points. One occurs when all nodes are inactive (i.e., when $\mu_1 = \mu_2 = 0$ and $o_1^{(t)} = o_2^{(t)} = 0$). The second case occurs when the inputs for nodes 1 and 2 exactly balance. In such a case, the fixed point will occur where $o_1^{(t)} = \mu_1$ and $o_2^{(t)} = \mu_2$.

Numerical work suggests that in the cases where $0 < \mu_1, \mu_2 < o_{max}$ the network possesses a single unstable fixed point at $(o_1^{(t)}, o_2^{(t)}) = (\mu_1, \mu_2)$ around which there appears to be a single globally attracting limit cycle orbit (see Figure 6.3). A proof of the instability of this fixed point is outlined in the next section. The limit cycle properties will be discussed further in §6.3.

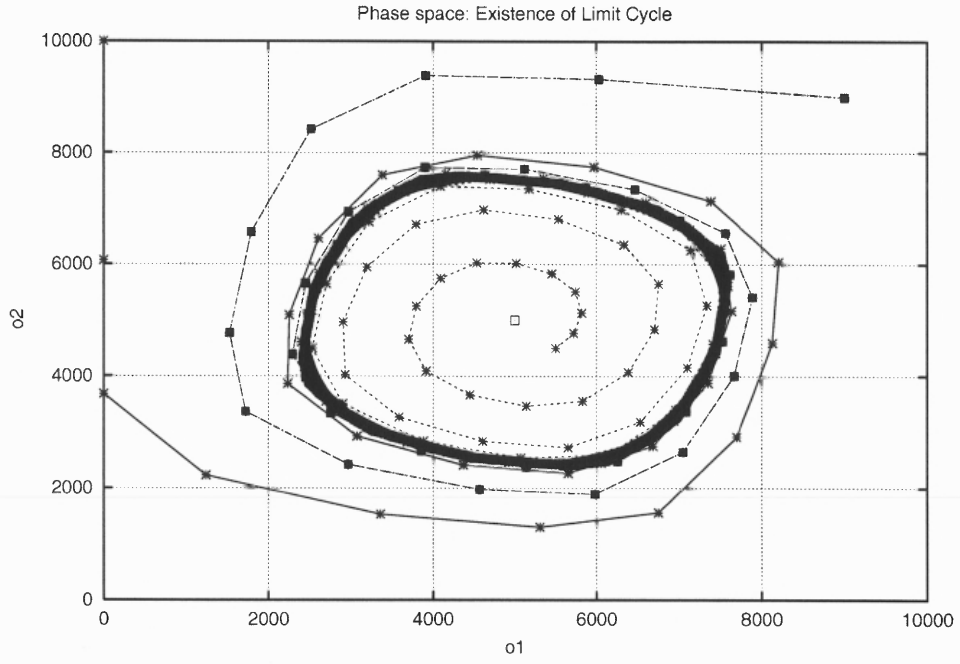
It is not easy to prove analytically the instability of the fixed point at $(o_1^{(t)}, o_2^{(t)}) = (\mu_1, \mu_2)$. The direction of the perturbation away from the fixed point affects how each node is updated at the next time step (see Figure 6.4). The different values of $b_j^{(t)}$, $j = 1, 2$, in these quadrants are shown in Table 6.1.

Table 6.1: The rule for updating each node depends on the sign of the input $b_j^{(t)}$, $j = 1, 2$. The motion in the $(o_1^{(t)}, o_2^{(t)})$ phase plane is different in each of the quadrants.

Region	
QI	$b_1^{(t)} > 0,\ b_2^{(t)} > 0\ [o_1^{(t)} > \mu_1,\ o_2^{(t)} < \mu_2]$
QII	$b_1^{(t)} < 0,\ b_2^{(t)} > 0\ [o_1^{(t)} > \mu_1,\ o_2^{(t)} > \mu_2]$
QIII	$b_1^{(t)} < 0,\ b_2^{(t)} < 0\ [o_1^{(t)} < \mu_1,\ o_2^{(t)} > \mu_2]$
QIV	$b_1^{(t)} > 0,\ b_2^{(t)} < 0\ [o_1^{(t)} < \mu_1,\ o_2^{(t)} < \mu_2]$

For the cases where either μ_1 or μ_2 take their minimum or maximum values, the equilibrium points are listed in Table 6.2. Note also that in the cases that $\mu_1 = 0$ or $\mu_2 = 0$ the network is reduced to a three node network. The results of this analysis agree with the data presented in Table C.1 (Appendix C).

A



B

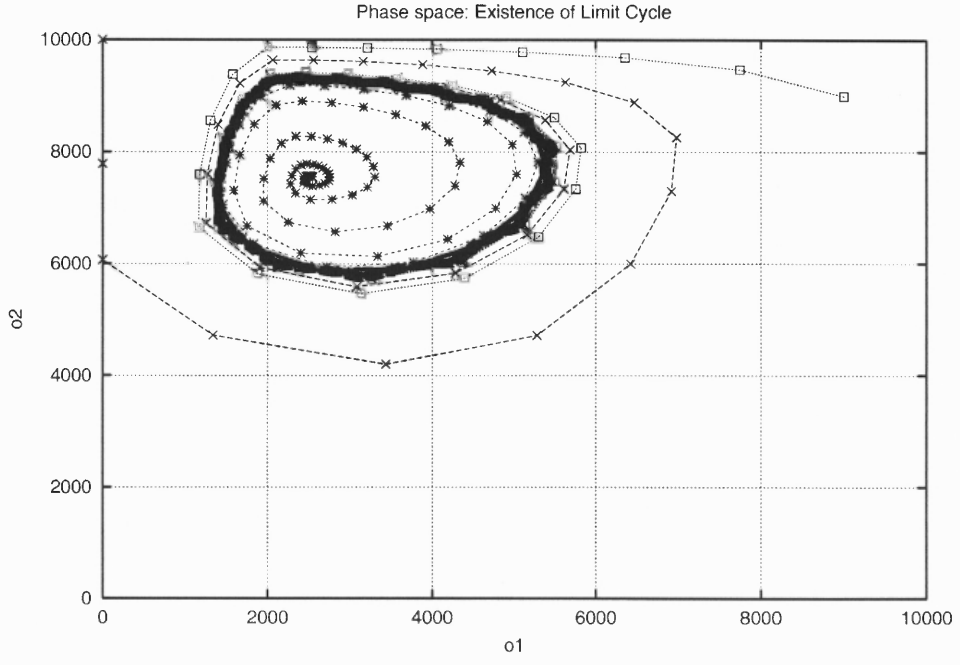


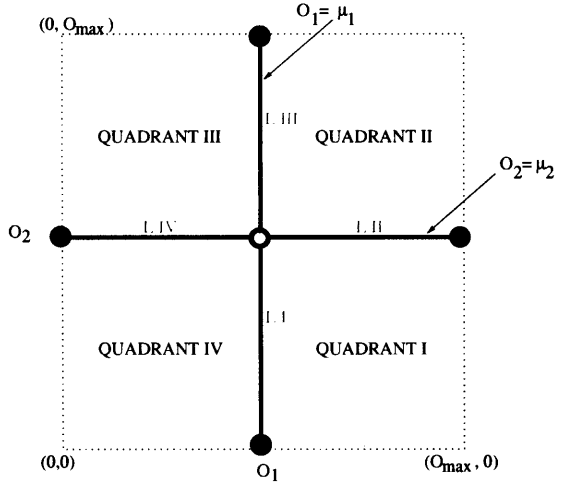
Figure 6.3: Limit cycle oscillation in the 2PK-2PP network for two selected pairs of input parameters: (μ_1, μ_2) = (A) $(5000, 5000)$, (B) $(2500, 7500)$. The network is tested from several different initial conditions in each case. The limit cycle seems to be globally attracting, with the exception of the fixed point at $(\hat{o}_1, \hat{o}_2) = (\mu_1, \mu_2)$. The input parameters to the network tune the amplitude and frequency of the oscillation. Global stability remains to be proven analytically.

QIII: $o_1^{(t)} < \mu_1,\ o_2^{(t)} > \mu_2.$

$$o_1^{(t+1)} = o_1^{(t)} e^{(\mu_2 - o_2^{(t)})/o_{max}},$$
$$o_2^{(t+1)} = o_2^{(t)} e^{(o_1^{(t)} - \mu_1)/o_{max}}.$$

QII: $o_1^{(t)} > \mu_1,\ o_2^{(t)} > \mu_2.$

$$o_1^{(t+1)} = o_1^{(t)} e^{(\mu_2 - o_2^{(t)})/o_{max}},$$
$$o_2^{(t+1)} = o_{max} - (o_{max} - o_2^{(t)}) e^{(\mu_1 - o_1^{(t)})/o_{max}}.$$



QIV: $o_1^{(t)} < \mu_1,\ o_2^{(t)} < \mu_2.$

$$o_1^{(t+1)} = o_{max} - (o_{max} - o_1^{(t)}) e^{(o_2^{(t)} - \mu_2)/o_{max}},$$
$$o_2^{(t+1)} = o_2^{(t)} e^{(o_1^{(t)} - \mu_1)/o_{max}}.$$

QI: $o_1^{(t)} > \mu_1,\ o_2^{(t)} < \mu_2.$

$$o_1^{(t+1)} = o_{max} - (o_{max} - o_1^{(t)}) e^{(o_2^{(t)} - \mu_2)/o_{max}},$$
$$o_2^{(t+1)} = o_{max} - (o_{max} - o_2^{(t)}) e^{(\mu_1 - o_1^{(t)})/o_{max}}.$$

Figure 6.4: Analyzing the iterative map in each of the four quadrants of the (o_1, o_2) phase plane. Quadrants QI-QIV and lines LI-LIV are identified because the map is different in each of these regions and on each of these lines. The point $(\hat{o}_1, \hat{o}_2) = (\mu_1, \mu_2)$ is a fixed point for the map.

Table 6.2: Fixed points for the 2PK-2PP network when one of the input parameters takes either its maximum or minimum values.

(μ_1, μ_2)	$(\hat{o}_1, \hat{o}_2)$
$(0, \mu_2)$, $\mu_2 \in [0, o_{max})$	$(0, \xi)$, $\xi \in [\mu_2, o_{max}]$
(o_{max}, μ_2), $\mu_2 \in (0, o_{max}]$	(o_{max}, η), $\eta \in [0, \mu_2]$
$(\mu_1, 0)$, $\mu_2 \in (0, o_{max}]$	$(\zeta, 0)$, $\zeta \in (0, o_{max}]$
(μ_1, o_{max}), $\mu_2 \in [0, o_{max})$	$(\rho, 0)$, $\rho \in [\mu_1, o_{max})$

6.2.3 Analysis of the Unstable Fixed Point for the 2PK-2PP Network: A "Tanh" Approximation

In simple cases, therefore, when the input to any given node in the network is only positive or negative, small networks can be examined easily for fixed points and for their stability (as discussed in §6.2.1). The following method has been considered as a way to examine the local stability of fixed points when the inputs for each node can take on positive and negative values. This involves using an approximation to the updating function, f.

In this section, a fixed point of the 2PK-2PP network is considered. The analysis is performed for the single binding site model where activity and occupancy are the same. A single closed form expression for the updating rule for a node j (which is independent of the sign of the inputs to the node, $b_j^{(t)}$) makes the stability analysis more tractable.

The resulting surface (Figure 6.5) can be compared to that illustrated in Chapter 2, Figure 2.4. The use of this rule is vindicated by numerical work. (See §6.4, Tables 6.3 & 6.5.) In practice, $\kappa = O(100)$ gives good numerical results.

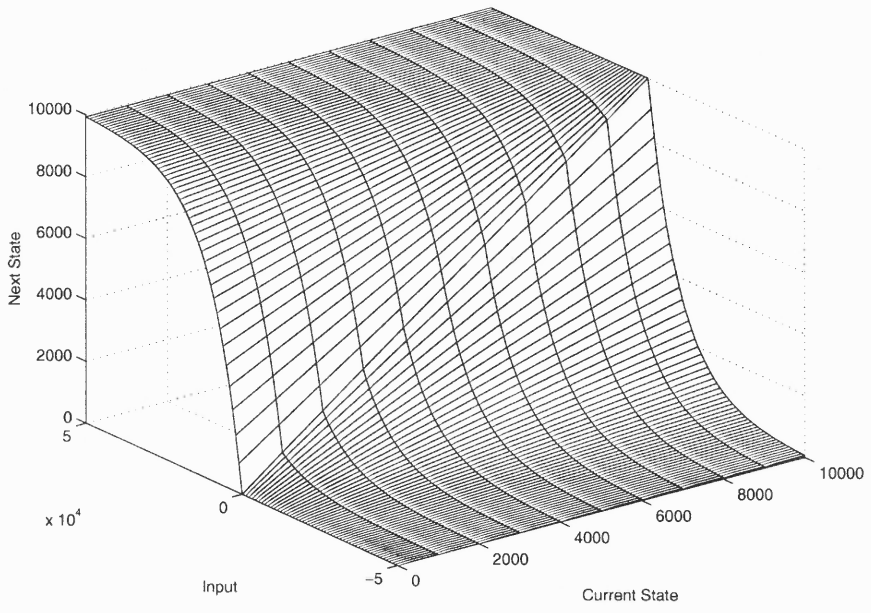


Figure 6.5: Surface plot of the updating rule with the tanh function (f_{tanh}).

A C^∞ approximation of the original updating rule, f, is considered here. It is referred to as f_{tanh}:

$$\left.\begin{aligned} o_j^{(t+1)} &= f_{tanh}(o_j^{(t)}, b_j^{(t)}) \\ &= \frac{1}{2}\left\{1 - \tanh(\kappa b_j^{(t)})\right\} o_j^{(t)} e^{b_j^{(t)}/o_{max}} \\ &\quad + \frac{1}{2}\left\{1 + \tanh(\kappa b_j^{(t)})\right\}\left[o_{max}(1 - e^{-b_j^{(t)}/o_{max}}) + o_j^{(t)} e^{-b_j^{(t)}/o_{max}}\right], \end{aligned}\right\} \quad (6.16)$$

where the parameter κ controls the steepness of the tanh function. In the limit $\kappa \to \infty$ the function $\frac{1}{2}[1 + \tanh(\kappa x)]$ approaches $H(x)$.

When the continuously differentiable function f_{tanh} is used, the occupancy values for nodes 1 and 2 of the four protein network shown in Figure 6.2 are

$$\begin{aligned} o_1^{(t+1)} &= 0.5\left\{1 - \tanh[\kappa(\mu_2 - o_2^{(t)})]\right\} o_1^{(t)} e^{(\mu_2 - o_2^{(t)})/o_{max}} \\ &\quad + 0.5\left\{1 + \tanh[\kappa(\mu_2 - o_2^{(t)})]\right\}\left\{o_{max}[1 - e^{-(\mu_2 - o_2^{(t)})/o_{max}}] + o_1^{(t)} e^{-(\mu_2 - o_2^{(t)})/o_{max}}\right\}, \\ o_2^{(t+1)} &= 0.5\left\{1 - \tanh[\kappa(o_1^{(t)} - \mu_1)]\right\} o_2^{(t)} e^{(o_1^{(t)} - \mu_1)/o_{max}} \\ &\quad + 0.5\left\{1 + \tanh[\kappa(o_1^{(t)} - \mu_1)]\right\}\left\{o_{max}[1 - e^{-(o_1^{(t)} - \mu_1)/o_{max}}] + o_2^{(t)} e^{-(o_1^{(t)} - \mu_1)/o_{max}}\right\}. \end{aligned}$$

The fixed point at $(o_1^{(t)}, o_2^{(t)}) = (\mu_1, \mu_2)$ can be found by inspection. A linear stability analysis is performed by making a small perturbation from the fixed point

$$\left.\begin{aligned} o_1^{(t)} &= \mu_1 + \epsilon x^{(t)}, \\ o_2^{(t)} &= \mu_2 + \epsilon y^{(t)}, \end{aligned}\right\} \tag{6.17}$$

where ϵ is a small parameter and $x^{(t)}, y^{(t)}$ are $O(1)$ quantities. If the tanh and exponential terms are expanded in a Taylor series, then the following is true:

$$\begin{aligned}
\mu_1 + \epsilon x^{(t+1)} &= \frac{1}{2}(\mu_1 + \epsilon x^{(t)})[1 + \epsilon\kappa y^{(t)} + \cdots]\left[1 - \frac{\epsilon y^{(t)}}{o_{max}} + \cdots\right] \\
&\quad + \frac{1}{2}(1 - \epsilon\kappa y^{(t)} + \cdots)\left\{o_{max}\left[-\frac{\epsilon y^{(t)}}{o_{max}} + \cdots\right] + (\mu_1 + \epsilon x^{(t)})\left[1 + \frac{\epsilon y^{(t)}}{o_{max}} + \cdots\right]\right\}, \\
\mu_2 + \epsilon y^{(t+1)} &= \frac{1}{2}(\mu_2 + \epsilon y^{(t)})[1 - \epsilon\kappa x^{(t)} + \cdots]\left[1 + \frac{\epsilon x^{(t)}}{o_{max}} + \cdots\right] \\
&\quad + \frac{1}{2}(1 + \epsilon\kappa y^{(t)} + \cdots)\left\{o_{max}\left[\frac{\epsilon x^{(t)}}{o_{max}} + \cdots\right] + (\mu_2 + \epsilon y^{(t)})\left[1 - \frac{\epsilon x^{(t)}}{o_{max}} + \cdots\right]\right\}.
\end{aligned}$$

After further manipulation and grouping of terms of similar orders, the new linearized system is

$$\begin{pmatrix} x^{(t+1)} \\ y^{(t+1)} \end{pmatrix} = \begin{pmatrix} 1 & -\frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} x^{(t)} \\ y^{(t)} \end{pmatrix}. \tag{6.18}$$

The eigenvalues of this new system are found to be $\lambda_j = 1 \pm i/2$, which satisfy $|\lambda_j| > 1$, for $j = 1, 2$. The fixed point at (μ_1, μ_2) is, therefore, locally *unstable*. Note, this analysis applies to the updating rule f_{tanh} which is a C^∞ approximation to the function f. The non-discretized updating rule, f, is not differentiable on the lines $o_1 = \mu_1$ and $o_2 = \mu_2$. The fixed point at (μ_1, μ_2) lies precisely where these lines intersect.

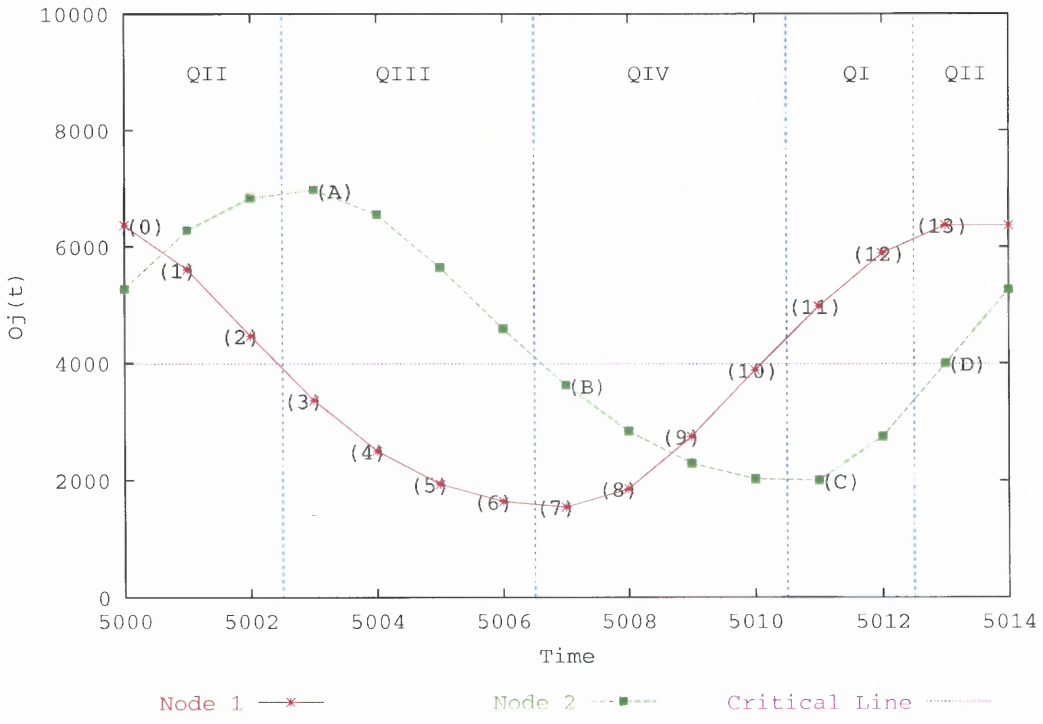


Figure 6.6: The limit cycle has four main phases, corresponding to the four quadrants of Figure 6.4.

6.3 Stability Analysis of the Limit Cycle of the 2PK-2PP network

In this section, the periodic orbit of the 2PK-2PP network is examined. Firstly, a qualitative explanation for the existence of the oscillation is presented. Secondly, a specific parameter case is examined, $(\mu_1, \mu_2) = (4000, 4000)$, for which the network exhibits a periodic orbit of length 14 time steps. The local asymptotic stability of the orbit is examined. Finally, the effect of varying the input parameters to the network is considered.

6.3.1 Qualitative Explanation for the Periodic Orbit

The periodic orbit can be divided into four distinct phases. This is discussed with reference to Figure 6.6. In the first phase, which corresponds to points $(o_1^{(t)}, o_2^{(t)})$ which lie in QII, the value of node 2 is increasing because its input is positive. This can be seen by noticing that the occupancy value of node 1 lies above the critical line,

$o_1^{(t)} = \mu_1$. Recall that the input for node 1 at time step t is given by $b_1^{(t)} = \mu_2 - o_2^{(t)}$, and the input for node two is $b_2^{(t)} = o_1^{(t)} - \mu_1$. At the same time, the occupancy of node 1 is decreasing, because its input is negative. The occupancy of node 2 will continue to increase until the occupancy of node 1 decreases below the critical line $o_1^{(t)} = \mu_1$. This occurs at time step 3 (see Figure 6.6). At this juncture, the data point has moved into quadrant three and the second phase of the oscillation starts. Now the input for node 2 is negative because the value of node 1 has fallen below its critical value (μ_1). In this phase, both occupancy values start to decrease. Once the value of node 2 has crossed its critical line, $o_2^{(t)} = \mu_2$, the input for node 1 becomes positive. This occurs at time step 7. The point $(o_1^{(t)}, o_2^{(t)})$ has now moved into QIV and phase three starts. The value of node 1 starts to increase (now $o_2^{(t)} < \mu_2$). This phase continues until node 1 crosses its critical value again. At time step 11 the input for node 2 becomes positive and now its occupancy value starts to rise. The point enters QI at this time. The fourth and final phase of the oscillation corresponds to an increase in the occupancy values of both nodes until finally node 2 crosses its critical value at time step 13. The data point returns to quadrant II and at time step 14 returns to its initial starting position. The phases of this periodic orbit illustrate why, in the cases when either μ_1 or μ_2 are set at the upper or lower limits, no oscillations are observed.

6.3.2 Proof of Local Asymptotic Stability of the Orbit

The goal of this section is to show that the 2PK-2PP network has a locally asymptotically stable periodic orbit of length 14 time steps when $(\mu_1, \mu_2) = (4000, 4000)$. Numerical evidence is presented which shows that a small neighborhood close to the limit cycle is mapped back completely within itself for the first time after 14 time steps. The eigenvalues of the Jacobian of the iterated map, $Df^{(14)}$, are shown to have magnitude less than 1. The Banach fixed point theorem is then applied to

demonstrate that the orbit is locally attracting with a period of 14 time steps. In order to apply the theorem, it must first be shown that the limit cycle points never lie on the lines where f has a discontinuity in its derivative. These are the lines $(o_1^{(t)}, o_2^{(t)}) = (\mu_1, \mu_2) = (4000, 4000)$ (LI-LIV).

The expressions for updating the occupancy of nodes 1 and 2 in this network are as follows:

$$\left.\begin{aligned} o_1^{(t+1)} &= H(\mu_2 - o_2^{(t)}) \left[o_{max} \left(1 - e^{-(\mu_2 - o_2^{(t)})/o_{max}} \right) \right] \\ &+ H(\mu_2 - o_2^{(t)}) \left[o_1^{(t)} \left(e^{-(\mu_2 - o_2^{(t)})/o_{max}} - e^{(\mu_2 - o_2^{(t)})/o_{max}} \right) \right] + o_1^{(t)} e^{(\mu_2 - o_2^{(t)})/o_{max}}, \\ o_2^{(t+1)} &= H(o_1^{(t)} - \mu_1) \left[o_{max} \left(1 - e^{-(o_1^{(t)} - \mu_1)/o_{max}} \right) \right] \\ &+ H(o_1^{(t)} - \mu_1) \left[o_2^{(t)} \left(e^{-(o_1^{(t)} - \mu_1)/o_{max}} - e^{(o_1^{(t)} - \mu_1)/o_{max}} \right) \right] + o_2^{(t)} e^{(o_1^{(t)} - \mu_1)/o_{max}}. \end{aligned}\right\}$$

These expressions simplify in each of the four quadrants (QI-QIV) (see Equations 6.19–6.22 and Figure 6.4):

QI: $o_1^{(t)} > \mu_1$, $o_2^{(t)} < \mu_2$,

$$\left.\begin{aligned} o_1^{(t+1)} &= o_{max} - (o_{max} - o_1^{(t)}) e^{(o_2^{(t)} - \mu_2)/o_{max}}, \\ o_2^{(t+1)} &= o_{max} - (o_{max} - o_2^{(t)}) e^{(\mu_1 - o_1^{(t)})/o_{max}}, \end{aligned}\right\} \tag{6.19}$$

QII: $o_1^{(t)} > \mu_1$, $o_2^{(t)} > \mu_2$,

$$\left.\begin{aligned} o_1^{(t+1)} &= o_1^{(t)} e^{(\mu_2 - o_2^{(t)})/o_{max}}, \\ o_2^{(t+1)} &= o_{max} - (o_{max} - o_2^{(t)}) e^{(\mu_1 - o_1^{(t)})/o_{max}}, \end{aligned}\right\} \tag{6.20}$$

QIII: $o_1^{(t)} < \mu_1$, $o_2^{(t)} > \mu_2$,

$$\left.\begin{aligned} o_1^{(t+1)} &= o_1^{(t)} e^{(\mu_2 - o_2^{(t)})/o_{max}}, \\ o_2^{(t+1)} &= o_2^{(t)} e^{(o_1^{(t)} - \mu_1)/o_{max}}, \end{aligned}\right\} \tag{6.21}$$

QIV: $o_1^{(t)} < \mu_1$, $o_2^{(t)} < \mu_2$,

$$\left.\begin{aligned} o_1^{(t+1)} &= o_{max} - (o_{max} - o_1^{(t)}) e^{(o_2^{(t)} - \mu_2)/o_{max}}, \\ o_2^{(t+1)} &= o_2^{(t)} e^{(o_1^{(t)} - \mu_1)/o_{max}}. \end{aligned}\right\} \tag{6.22}$$

The lines LI-LIV (see Figure 6.4) are also considered separately, because the map has a discontinuity in its derivative there. The map is continuous everywhere else and derivatives of all orders exist away from these lines. The point $(o_1^{(t)}, o_2^{(t)}) = (\mu_1, \mu_2)$ is a fixed point. The mappings of these lines are shown in Equations 6.23 to 6.26.

LI: $o_1^{(t)} = \mu_1,\ 0 \leq o_2^{(t)} < \mu_2,$

$$\left.\begin{aligned} o_1^{(t+1)} &= o_{max} - (o_{max} - \mu_1)e^{(o_2^{(t)} - \mu_2)/o_{max}} \\ o_2^{(t+1)} &= o_2^{(t)} \end{aligned}\right\}. \tag{6.23}$$

LII: $\mu_1 < o_1^{(t)} \leq o_{max},\ o_2^{(t)} = \mu_2,$

$$\left.\begin{aligned} o_1^{(t+1)} &= o_1^{(t)} \\ o_2^{(t+1)} &= o_{max} - (o_{max} - \mu_2)e^{(\mu_1 - o_1^{(t)})/o_{max}} \end{aligned}\right\}. \tag{6.24}$$

LIII: $o_1^{(t)} = \mu_1,\ \mu_2 \leq o_2^{(t)} \leq o_{max},$

$$\left.\begin{aligned} o_1^{(t+1)} &= \mu_1^{(t)} e^{(\mu_2 - o_2^{(t)})/o_{max}} \\ o_2^{(t+1)} &= o_2^{(t)} \end{aligned}\right\}. \tag{6.25}$$

LIV: $0 \leq o_1^{(t)} < \mu_1,\ o_2^{(t)} = \mu_2,$

$$\left.\begin{aligned} o_1^{(t+1)} &= o_1^{(t)} \\ o_2^{(t+1)} &= \mu_2 e^{(o_1^{(t)} - \mu_1)/o_{max}} \end{aligned}\right\}. \tag{6.26}$$

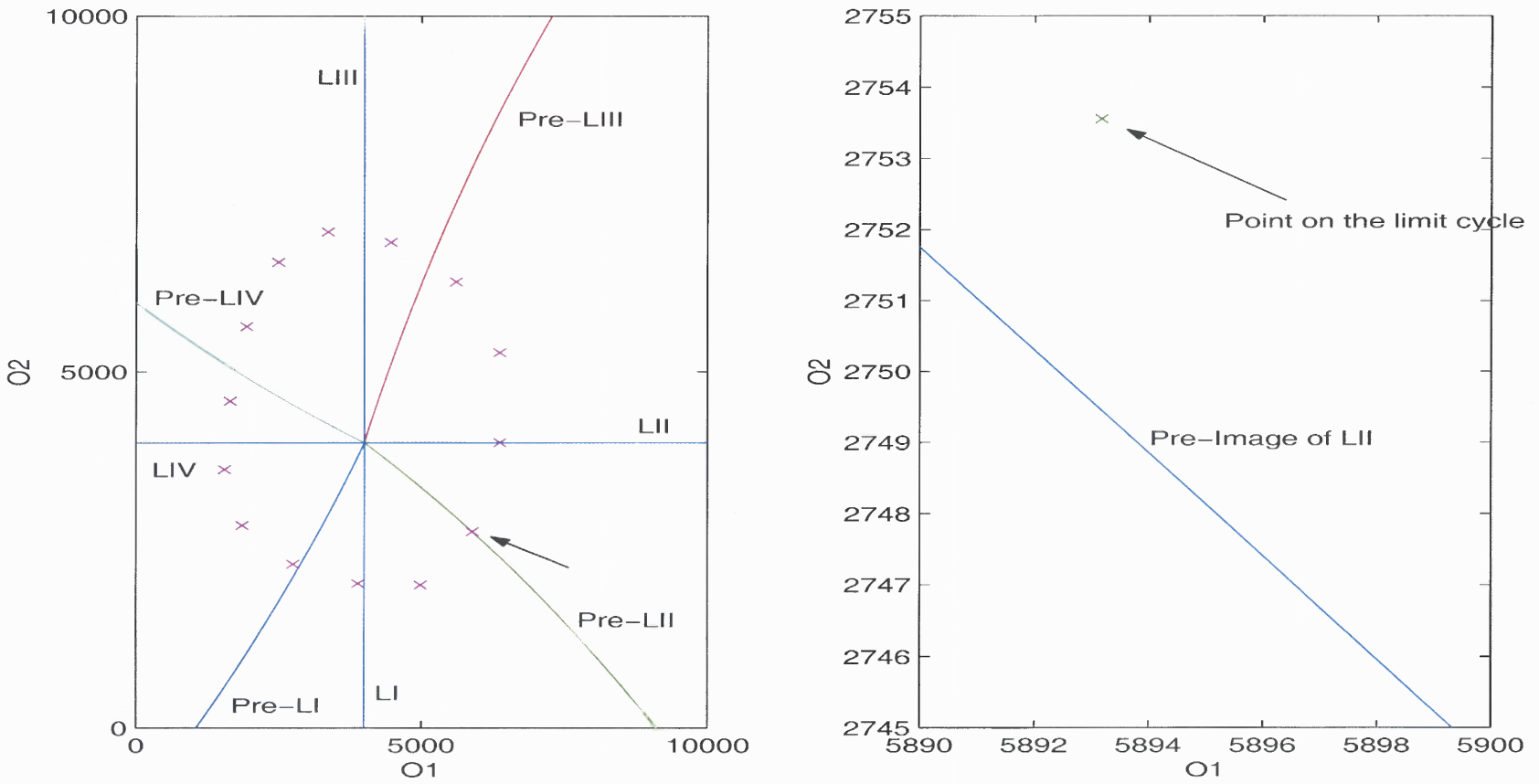


Figure 6.7: Pre-images of the discontinuity (critical) lines $o_1 = \mu_1$, $o_2 = \mu_2$. The left plot shows each of the pre-images of LI-LIV together with the limit cycle points. The right hand plot shows a blow up of the point which looks close to the pre-image of LII. The data show that no limit cycle point ever maps onto the discontinuity lines.

Table 6.3: A period 14 limit cycle. Comparison between different schemes: (a) Numerics using f, (b) Numerics using $\hat{f}$, (c) Numerics using f_{tanh} with $\kappa = 100$. Double precision numerical data is given to 8 d.p., the periodic repeat is evident when comparing data at time steps 5000 and 5014.

P_i	Time step	(a) Using f	(b) Using $\hat{f}$	(c) Using f_{tanh}
	5000	(6372.28057609, 5270.83217239)	(6375,5271)	(6372.28057608, 5270.83217245)
P_4	5001	(5611.81505542, 6269.57863244)	(5614, 6270)	(5611.81505537, 6269.57863248)
P_5	5002	(4472.36049082, 6824.89821878)	(4473, 6825)	(4472.36049077, 6824.89821880)
P_6	5003	(3371.73187339, 6971.39038937)	(3372, 6971)	(3371.73187334, 6971.39038937)
P_7	5004	(2504.99686350, 6546.87524636)	(2505, 6546)	(2504.99686346, 6546.87524633)
P_8	5005	(1941.77005738, 5637.76415528)	(1941, 5637)	(1941.77005736, 5637.76415523)
P_9	5006	(1648.43036676, 4589.01115527)	(1647, 4588)	(1648.43036675, 4589.01115522)
P_{10}	5007	(1554.14013796, 3627.36903939)	(1552, 3626)	(1554.14013796, 3627.36903934)
P_{11}	5008	(1863.06748499, 2840.33390303)	(1862, 2838)	(1863.06748502, 2840.33390300)
P_{12}	5009	(2754.02126035, 2293.84251438)	(2754, 2291)	(2754.02126040, 2293.8425143)
P_{13}	5010	(3890.58569704, 2025.12310497)	(3892, 2022)	(3890.58569710, 2025.12310496)
P_0	5011	(4985.45233807, 2003.08613948)	(4988, 2000)	(4985.45233812, 2003.08613948)
P_1	5012	(5893.16838644, 2753.55889692)	(5896, 2752)	(5893.16838648, 2753.55889697)
P_2	5013	(6374.44375307, 4003.39409082)	(6377, 4003)	(6374.44375309, 4003.39409088)
P_3	5014	(6372.28057609, 5270.83217239)	(6375, 5271)	(6372.28057608, 5270.83217245)

Although from a first glance at Figures 6.6 and 6.10 it seems as though the 13th iterate lies on the line LII, in actuality it does not ($o_2 \neq 4000$). See Table 6.3, data at time step 5013. This is true for all versions of the updating rule (f, $\hat{f}$ and f_{tanh}). Exact expressions for the pre-images of the lines LI-LIV can be found. For example, consider the line LI. Only points in QIV can map to LI. A point $(o_1^{(t+1)}, o_2^{(t+1)})$ which lies on the line LI satisfies the following conditions: (i) $o_1^{(t+1)} = \mu_1$, and (ii) $0 \leq o_2^{(t+1)} < \mu_2$. A relationship between $(o_1^{(t)}, o_2^{(t)})$ can be determined from this information using Equation 6.22. Pre-images of LII-LIV are computed similarly and are shown in Figure 6.7 in relation to the points of the periodic orbit.

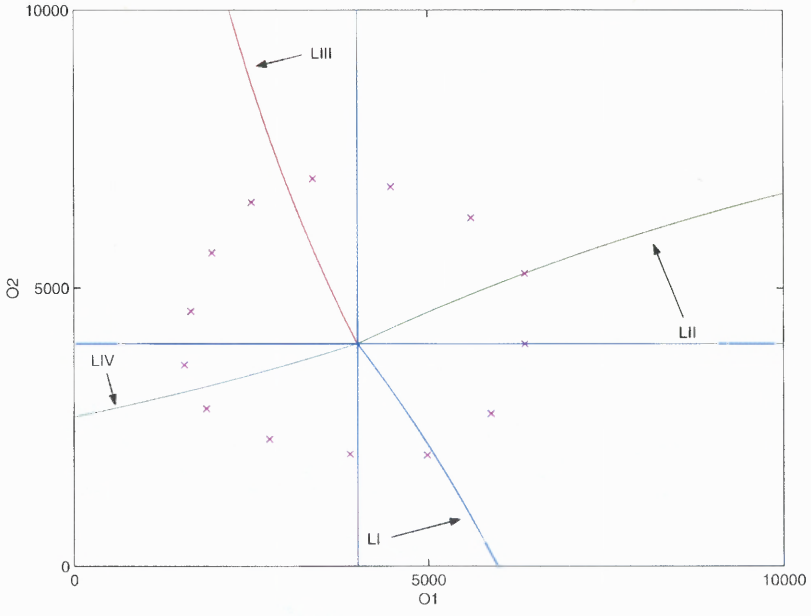


Figure 6.8: Mapping the lines LI-LIV.

The mapping of each of these lines is illustrated in Figure 6.8. Note that LI maps to QI, LII maps to QII and so forth. At this juncture, it is noted that in this four node network the rule can be represented in the form shown in Equation 6.27. The movement of any point is a combination of a translation and a scaling:

$$\left.\begin{aligned} o_1^{(t+1)} &= o_1^{(t)} F_1(o_2^{(t)}, \mu_2) + G_1(o_2^{(t)}, \mu_2), \\ o_2^{(t+1)} &= o_2^{(t)} F_2(o_1^{(t)}, \mu_1) + G_2(o_1^{(t)}, \mu_1). \end{aligned}\right\} \tag{6.27}$$

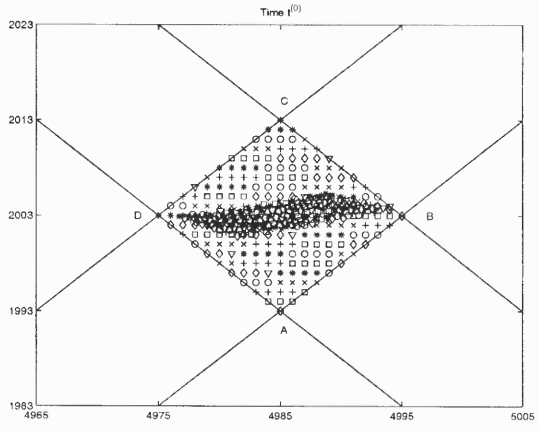
This can be seen clearly when the mapping of a small region is followed through the quadrants (see Figure 6.9). A four-sided polygonal region is considered in Quadrant I. When considered in the original phase plane (o_1, o_2), then the neighborhood is a rhombus of side length $o_{max}\epsilon$ where $\epsilon = 0.001$. Its vertices are labeled A–D in a counterclockwise direction, starting from the lower-most vertex.

Successive iterates of this region are shown in Figure 6.9. The data points within the polygon are shown. At each successive iteration of the map, the edges of the original polygon AB,BC,CD and DA are approximated by straight lines joining the mapped vertex points. This provides a good approximation to the region as is seen in Figure 6.9. When the mapped region crosses the discontinuity line at the time step $t = 2$ (Plot C of Figure 6.9) the polygonal approximation is improved by adding in two new vertices E and F which correspond to the points where the sides AD and CD crossed the discontinuity line LII at that time step. The new six sided polygon is shown in plot D which is obtained by mapping the 6 vertices from the previous time step. Straight lines are used to approximate the mapped region. After 14 time steps the four sided region ABCD (which has become AEBCFD) maps back within itself for the first time. This is shown in illustration A of Figure 6.9.

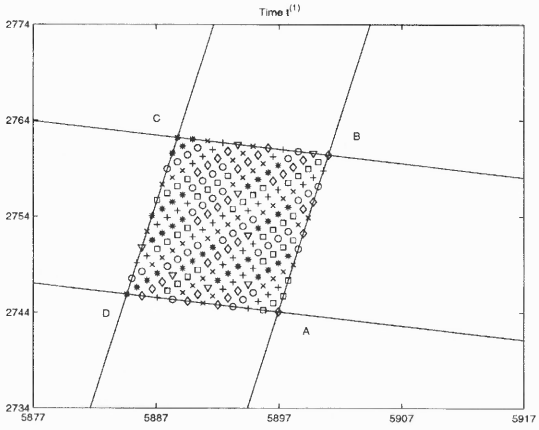
The nth-iterate of a point, P, under the map, f, is written $f^n(P)$. If there exists a cycle of k distinct points, $P_j = f^j(P_0)$, where $j = 0, \ldots, k-1$ and $f^k(P_0) = P_0$, then f has a periodic orbit of length k. A map $F\colon (Y, d) \to (Y, d)$ of a metric space into itself, where d is a given metric for Y, is called d-contractive if $\exists\ \alpha < 1$ such that $d(Fx, Fy) \leq \alpha d(x, y)$, $\forall\ (x, y) \in Y \times Y$. If S is a complete metric space and F is d-contractive, then F has a *unique* fixed point. This is a statement of the Banach fixed point theorem (BFPT) [97].

Consider the 14th iterate of the mapping $\mathbf{f}$, where $\mathbf{o}^{(t+1)} = \mathbf{f}(\mathbf{o}^{(t)})$. Let $\mathbf{P_0} \in Y$. For a point $\mathbf{P}$ lying in $\mathrm{B}(\mathbf{P_0}; \epsilon) \subset Y$, then in regions where f is C^2:

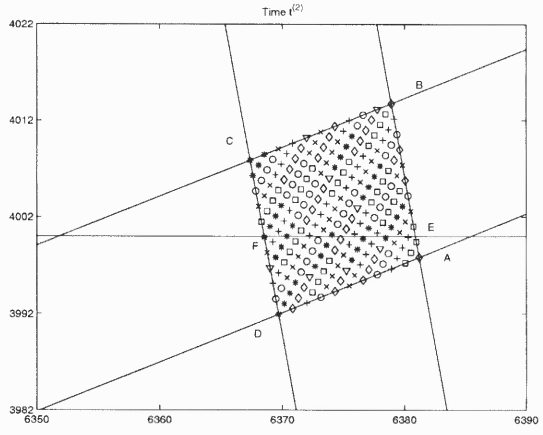
$$\mathbf{f}^{(14)}(\mathbf{P}) \;=\; \mathbf{f}^{(14)}(\mathbf{P_0}) + \mathbf{Df}^{(14)}(\mathbf{P_0})\mathbf{\Delta P} + \mathrm{O}(\|\mathbf{\Delta P}\|^2), \quad \text{where } \mathbf{\Delta P} = \mathbf{P} - \mathbf{P_0}.$$



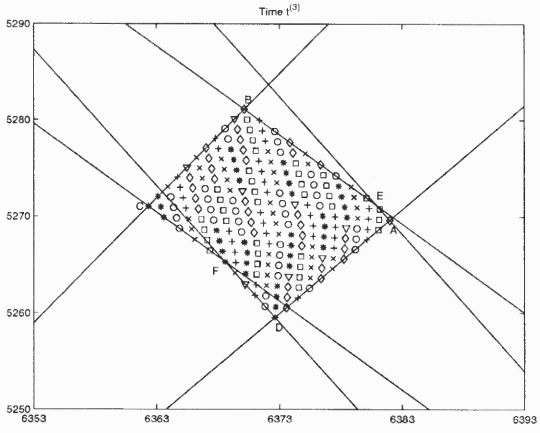
(A) t=0 and t=14



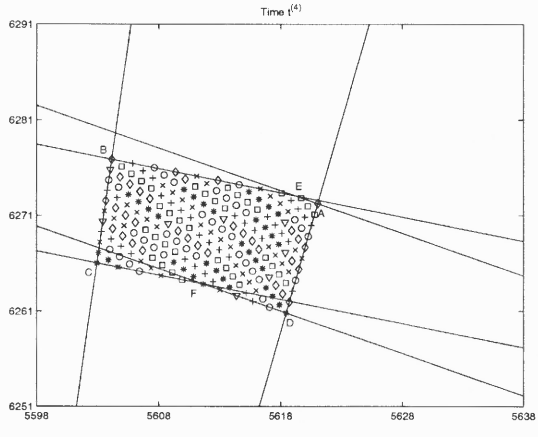
(B) t=1



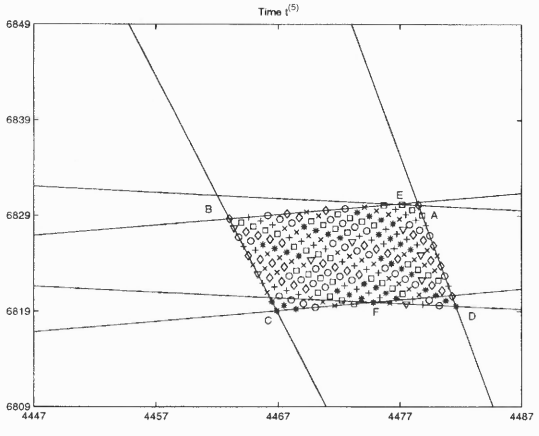
(C) t=2



(D) t=3

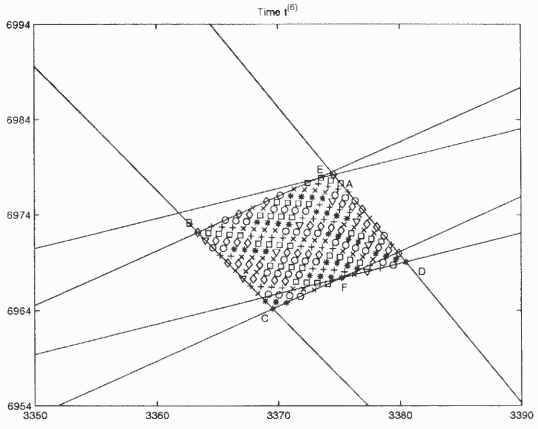


(E) t=4

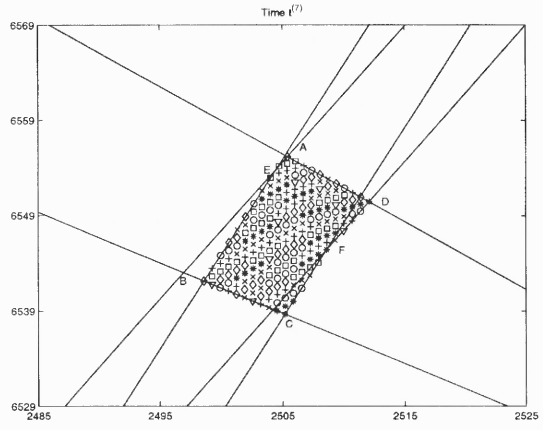


(F) t=5

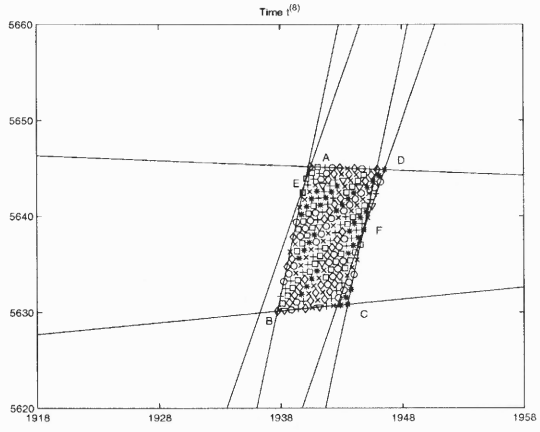
Figure 6.9: Evolution of a small neighborhood close to the limit cycle. Note that in panel (A) the mapped region after 14 time steps is also shown. This illustrates that the starting region ABCD maps back completely within itself for the first time after 14 time steps.



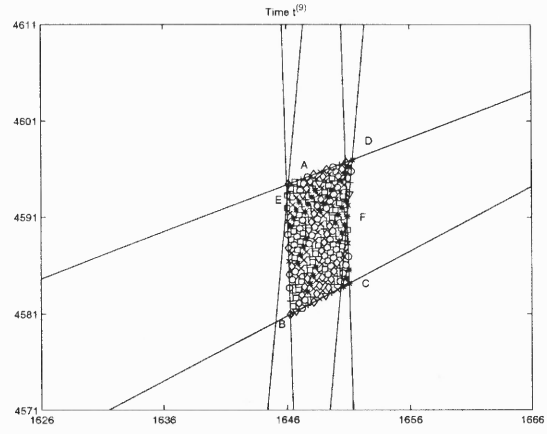
(G) t=6



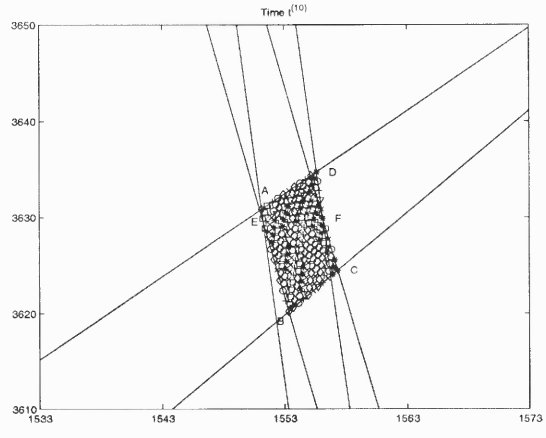
(H) t=7



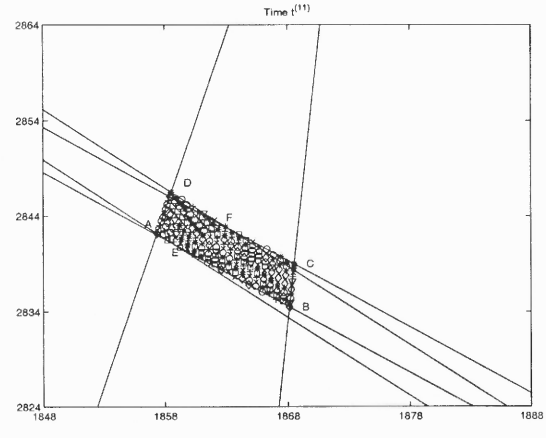
(I) t=8



(J) t=9

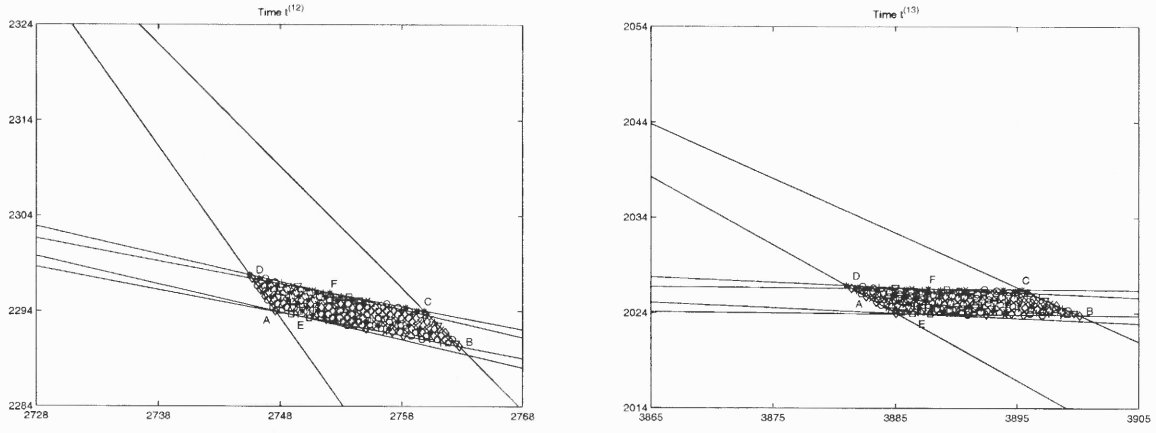


(K) t=10



(L) t=11

Figure 6.9 (ctd)



(M) t=12 (N) t=13

Figure 6.9 (ctd)

This can be seen by generalizing the following from a scalar function f of a single variable

$$\begin{aligned}
f(P) &= f(P_0) + f'(P_0)\Delta P + a_0(\Delta P)^2, \\
f(f(P)) &= f(f(P_0) + f'(P_0)\Delta P + a_0(\Delta P)^2), \\
&= f(f(P_0)) + f'(f(P_0))\left[f'(P_0)\Delta P + a_0(\Delta P)^2\right] + a_1(f'(P_0)\Delta P + a_0(\Delta P)^2)^2, \\
&= f^2(P_0) + f^{2'}(P_0) + \mathrm{O}(\Delta P)^2, \\
&\vdots \\
f^{14}(P) &= f^{14}(P_0) + f^{14'}(P_0) + \mathrm{O}(\Delta P)^2,
\end{aligned}$$

to a vector valued function of two variables:

$$\mathbf{f}^{(14)}(\mathbf{P}) = \mathbf{f}^{(14)}(\mathbf{P_0}) + \mathbf{Df}^{(14)}(\mathbf{P_0})\mathbf{\Delta P} + O(\|\Delta \mathbf{P}\|^2), \quad \text{where } \mathbf{\Delta P} = \mathbf{P} - \mathbf{P_0}.$$

Then, if $\|\mathbf{Df}^{(14)}(\mathbf{P_0})\| \leq c < 1$, then it follows that

$$\|\mathbf{f}^{(14)}(\mathbf{P}) - \mathbf{f}^{(14)}(\mathbf{P_0})\| \leq \tilde{c}\|\mathbf{P} - \mathbf{P_0}\|, \tag{6.28}$$

where $0 < \|\mathbf{Df}^{(14)}(\mathbf{P_0})\| + \delta \leq \tilde{c} < 1$, in a sufficiently small ball, $B(P_0; R)$ which does not intersect the lines LI-LIV. An estimate for the radius of this ball can be found.

The region of interest is now scaled to the unit interval in $\mathbb{R}^2$, denoted here as I, $I = [0,1] \times [0,1]$. The new variables $u^{(t)} = o_1^{(t)}/o_{max}$, $v^{(t)} = o_2^{(t)}/o_{max}$ are introduced. The parameters μ_1 and μ_2 are also rescaled so that $\hat{\mu}_1 = \mu_1/o_{max}$, $\hat{\mu}_2 = \mu_2/o_{max}$. The metric subspace (I, d) is complete because I is a compact subset of $\mathbb{R}^2$, which is complete. The local asymptotic stability of the orbit is determined by the eigenvalues of the linearized map $Df^k(P_0)$. By the chain rule, this is equivalent to

$$\begin{aligned} Df^k(P_0) &= Df(f^{k-1}(P_0)).Df(f^{k-2}(P_0)) \ldots Df(f(P_0)).Df(P_0), \\ &= Df(P_{k-1}).Df(P_{k-2}) \ldots Df(P_1)Df(P_0). \end{aligned} \tag{6.29}$$

The exact form of f depends on the sign of the input $b_j^{(t)}$ into each node j each of the four quadrants (see Equations 6.19–6.22 and Figure 6.4). The Jacobian matrix Df is computed for each of these cases. They are denoted Df_{QI} to Df_{QIV}. To simplify the Jacobian matrices, let

$$\eta = e^{v-\hat{\mu}_2}, \tag{6.30}$$

$$\zeta = e^{\hat{\mu}_1-u}. \tag{6.31}$$

QI:

$$Df_{QI} = \begin{bmatrix} \eta & (u-1)\eta \\ (1-v)\zeta & \zeta \end{bmatrix} \tag{6.32}$$

QII:

$$Df_{QII} = \begin{bmatrix} \eta^{-1} & -\eta^{-1}u \\ (1-v)\zeta & \zeta \end{bmatrix} \tag{6.33}$$

QIII:

$$Df_{QIII} = \begin{bmatrix} \eta^{-1} & -\eta^{-1}u \\ \zeta^{-1}v & \zeta^{-1} \end{bmatrix} \tag{6.34}$$

QIV:

$$Df_{QIV} = \begin{bmatrix} \eta & (u-1)\eta \\ \zeta^{-1}v & \zeta^{-1} \end{bmatrix} \tag{6.35}$$

The eigenvalues of the composition of these matrices are computed numerically since the points, P_i, $i = 0, \ldots, 13$ are known (see Table 6.3). The original data points have been scaled to the region I. At each mapping stage (i.e., at each point P_j, $0 \leq j \leq 13$) in Table 6.4) the eigenvalues of the mapping $Df(P_j).Df(P_{j-1})\ldots Df(P_1).Df(P_0)$ are shown, together with their corresponding eigenvectors. The magnitude of the largest eigenvalue is shown in the rightmost column. Thus, the eigenvalues of Df^{14} can be read off from the bottom row of the table: $\lambda_1 = 0.9320, \lambda_2 = 0.1996$. Both of these eigenvalues are smaller than one, showing that the map f^{14} is contractive.

This completes the proof of the existence and local asymptotic stability of the periodic orbit. Global stability of this orbit is conjectured, but not proven. Numerical work suggests that the periodic orbit is strongly attracting (see Figure 6.10).

6.3.3 Parameter Variation

A systematic test of the (μ_1, μ_2) parameter space indicates that the periodicity of the limit cycle does not vary smoothly with changes in these two parameters. Data are shown here for the truncated rule $\hat{f}$ (see Figure 6.11).

Table 6.4: Eigenvalues of the iterated map f^{14}. At each mapping stage the eigenvalues of the composed map are computed, together with its eigenvectors. For example, at point P_j, $0 \leq j \leq 13$, the eigenvalues of the mapping $Df(P_j).Df(P_{j-1})\ldots Df(P_1).Df(P_0)$ are shown, together with their corresponding eigenvectors. The magnitude of the largest eigenvalue is shown in the rightmost column. The eigenvalues of Df^{14} can be read off from the bottom row of the table: $\lambda_1 = 0.9320, \lambda_2 = 0.1996$. Both of these eigenvalues are smaller than one, showing that the map f^{14} is a contraction mapping.

P_i	Quad.	Eigenvalues of Composed Maps (λ_1, λ_2)	Eigenvectors of Composed Maps (v_1, v_2)	Magnitude of Largest Eigenvalue $\lvert\lambda\rvert_{max}$
P_0: (0.49854523, 0.20030861)	QI	$0.8626 \pm 0.5438i$	$(-0.5995 - 0.0481i,\ 0.7989i)$ + c.c.	1.0197
P_1: (0.58931684, 0.27535589)	QI	$0.4819 \pm 0.8680i$	$(-0.6226 - 0.0155i, 0.7824i)$ + c.c.	0.9928
P_2: (0.63744438, 0.40033940)	QII	$-0.0823 \pm 1.0331i$	$(-0.6884 - 0.1017i, 0.7182i)$ + c.c.	1.0364
P_3: (0.63722806, 0.52708322)	QII	$-0.5669 \pm 0.8060i$	$(-0.7093 - 0.2155i, 0.6711i)$ + c.c.	0.9854
P_4: (0.56118151, 0.62695786)	QII	$-0.7788 \pm 0.4359i$	$(-0.7008 - 0.3364i, 0.6291i)$ + c.c.	0.8925
P_5: (0.44723605, 0.68248982)	QII	$-0.8005 \pm 0.1154i$	$(-0.7353 - 0.4865i, 0.4719i)$ + c.c.	0.8088
P_6: (0.33717319, 0.69713904)	QIII	$-0.6438, -0.8754$	$(0.3772, -0.9261)$, $(0.0517, 0.9987)$	0.8754
P_7: (0.25049969, 0.65468752)	QIII	$-0.5794 \pm 0.3198i$	$(0.3605 - 0.1781i, 0.9156i)$ + c.c.	0.6618
P_8: (0.19417701, 0.56377642)	QIII	$-0.3761 \pm 0.4408i$	$(0.4528 - 0.1564i, 0.8778i)$ + c.c.	0.5794
P_9: (0.16484304, 0.45890112)	QIII	$-0.2018 \pm 0.4780i$	$(0.5291 - 0.1367i, 0.8375i)$ + c.c.	0.5188
P_{10}: (0.15541401, 0.36273690)	QIV	$0.1885 \pm 0.4793i$	$(0.5605 - 0.4113i, 0.7188i)$ + c.c.	0.5150
P_{11}: (0.18630675, 0.28403339)	QIV	$0.4633 \pm 0.1421i$	$(0.2326 - 0.7579i, 0.6095i)$ + c.c.	0.4846
P_{12}: (0.27540213, 0.22938425)	QIV	$0.9132, 0.2232$	$(0.9842, -0.1769)$, $(-0.5603, 0.8283)$	0.9132
P_{13}: (0.38905857, 0.20251231)	QIV	$0.9320, 0.1996$	$(0.9960, 0.0889)$, $(-0.4698, 0.8828)$	0.9320

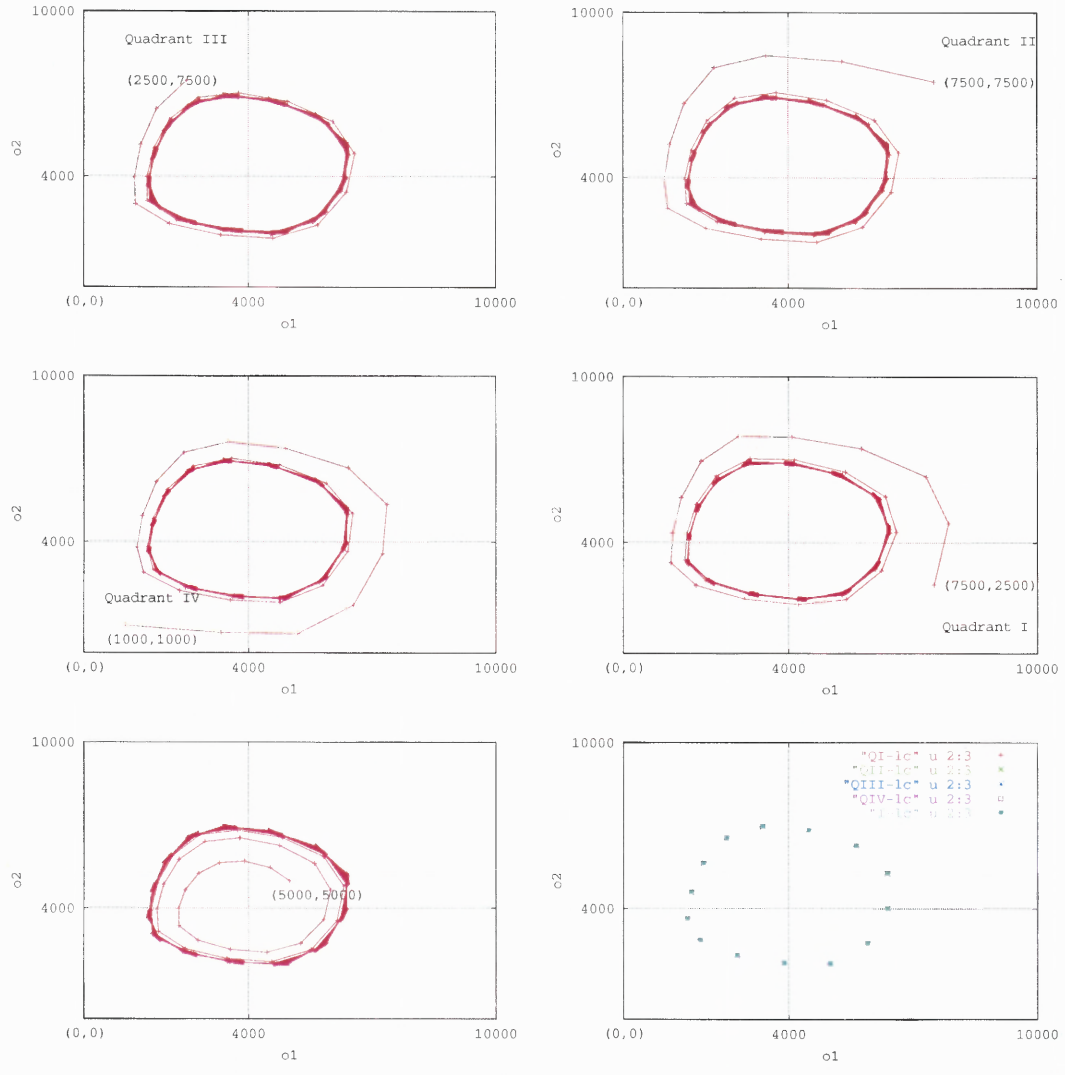
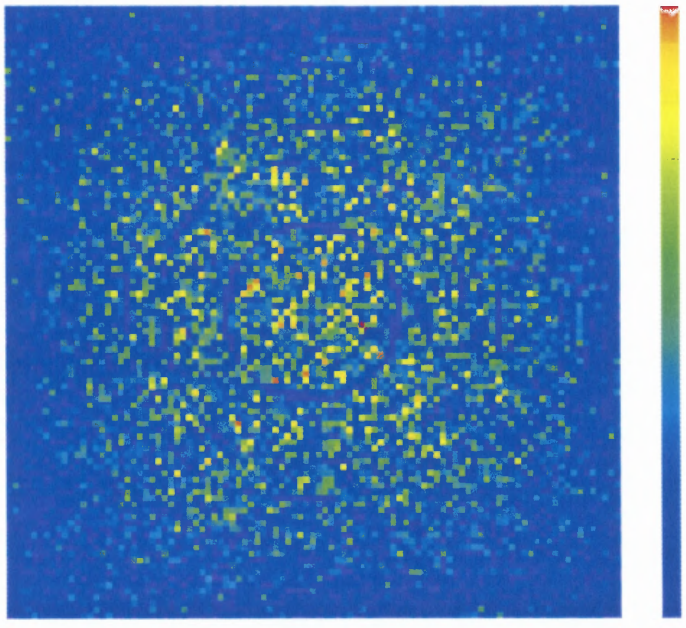


Figure 6.10: An example of how parameter adjustment leads to a low period orbit which appears to be strongly attracting. The non-discretized rule, f, is used to iterate several initial starting points in each of the four quadrants and also a point in the interior of the limit cycle. The plot in the bottom right hand corner shows that all points are attracted to the same period 14 orbit. The critical lines $o_1 = \mu_1$, $o_2 = \mu_2$ are shown on each plot.

A



B

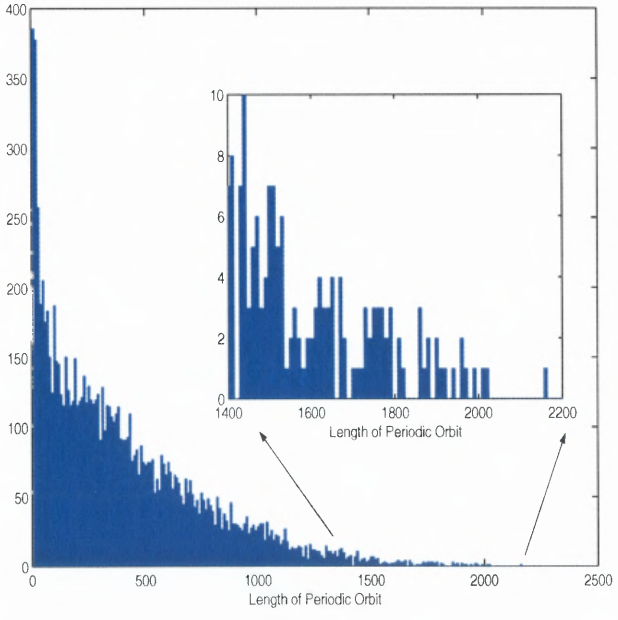


Figure 6.11: Variation of periodicity of the limit cycle for the rule $\hat{f}$ as the parameters μ_1 and μ_2 are varied. (A) A parameter range for (μ_1, μ_2) was taken to be $100 \leq \mu_1 \leq 9900$, incrementing μ_1 in jumps of 100 (similarly for μ_2). The largest periodicity value detected was 2,169 time steps. In this plot $(\mu_1, \mu_2) = (100, 100)$ is the data point at the top left, and $(\mu_1, \mu_2) = (9900, 9900)$ is the point at the bottom right corner. Limit cycles of short periodicity appear in blue. The graded color map is shown to the right. (B) Histogram plot of periodicities as μ_1 and μ_2 are varied. Data are grouped into bins of size 10.

Table 6.5: Comparison of truncated and double precision numerical results for the 2PK-2PP network with $(\mu_1, \mu_2) = (5000, 5000)$. The starting state used for the network is $(1000, 9000, 5000, 5000)$. The truncated rule (column 2) indicates an exact repetition every $t = 435$ time steps. The non-discretized rule, f, gives a repeat time of $t = 435$ (5 d.p. precision) after 10,000 time steps. The fourth column illustrates data from the application of the updating rule f_{tanh}, where $\kappa = 200$.

Time	Data for $\hat{f}$	Data for f	Data for f_{tanh}
383	(2511, 5175)	(2510.57085, 5174.76318)	(2510.57085163, 5174.76318027)
384	(2467, 4034)	(2467.07648, 4034.37204)	(2467.07648470, 4034.37203856)
818	(2511, 5175)	(2510.56838, 5174.73113)	(2510.56837656, 5174.73113067)
819	(2467, 4034)	(2467.08196, 4034.34605)	(2467.08195939, 4034.34605338)
1253	(2511, 5175)	(2510.56601, 5174.70049)	(2510.56600962, 5174.70048649)
1254	(2467, 4034)	(2467.08719, 4034.32121)	(2467.08195939, 4034.34605338)
1688	(2511, 5175)	(2510.56375, 5174.67120)	(2510.56374746, 5174.67120365)
1689	(2467, 4034)	(2467.09219, 4034.29747)	(2467.09219498, 4034.29746530)
2123	(2511, 5175)	(2510.56159, 5174.64324)	(2510.56158670, 5174.64323771)
2124	(2467, 4034)	(2467.09697, 4034.27479)	(2467.09697109, 4034.27479067)
2558	(2511, 5175)	(2510.55952, 5174.61654)	(2510.55952393, 5174.61654405)
2559	(2467, 4034)	(2467.10153, 4034.25315)	(2467.10152962, 4034.25314749)
2993	(2511, 5175)	(2510.55756, 5174.59108)	(2510.55755575, 5174.59107810)
2994	(2467, 4034)	(2467.10588, 4034.23250)	(2467.10587823, 4034.23249962)
3428	(2511, 5175)	(2510.55568, 5174.56680)	(2510.55567878, 5174.56679548)
3429	(2467, 4034)	(2467.11002, 4034.21281)	(2467.11002452, 4034.21281112)
3863	(2511, 5175)	(2510.55389, 5174.54365)	(2510.55388965, 5174.54365222)
3864	(2467, 4034)	(2467.11398, 4034.19405)	(2467.11397605, 4034.19404632)
4298	(2511, 5175)	(2510.55219, 5174.52160)	(2510.55218502, 5174.52160485)
4299	(2467, 4034)	(2467.11774, 4034.17617)	(2467.11774026, 4034.17617000)
4733	(2511, 5175)	(2510.55056, 5174.50061)	(2510.55056162, 5174.50061052)
4734	(2467, 4034)	(2467.12132, 4034.15915)	(2467.12132450, 4034.15914744)
5168	(2511, 5175)	(2510.54902, 5174.48063)	(2510.54901623, 5174.48062716)
5169	(2467, 4034)	(2467.12474, 4034.14294)	(2467.12473597, 4034.14294452)
89993	(2511, 5175)	(2510.52072, 5174.11509)	(2510.52071822, 5174.11509118)
89994	(2467, 4034)	(2467.18711, 4033.84655)	(2467.18711035, 4033.84654935)
99998	(2511, 5175)	(2510.52072, 5174.11509)	(2510.52071789, 5174.11508686)
99999	(2467, 4034)	(2467.18711, 4033.84655)	(2467.18711109, 4033.84654584)

6.4 Comparison of Truncated and Double Precision Numerics

In the previous sections in this chapter, f has been considered as the mapping function. Here data are presented for the truncated numerics, which utilize the function $\hat{f}$. A comparison between all three of the rules used is presented. Table 6.5 illustrates data between the three rules (Col 2: $\hat{f}$, Col 3: f, Col 4: f_{tanh} with $\kappa = 200$). The parameter case selected here is $(\mu_1, \mu_2) = (5000, 5000)$). The four protein network of §6.2.2 is considered and the initial state of the network is set to $(1000, 9000, 5000, 5000)$. The strong agreement between the simulations is evident.

Even after $50,000$ time steps, the numerical values remain "close". When the distance between respective components in the vectors $(o_1^{(t)}, o_2^{(t)})$ is compared, each is never more than $d = 8.14030$ apart over $t = 50,000$ time steps, using the following metric (L_0 norm):

$$d(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{i=1,2.}} |\mathbf{x_i} - \mathbf{y_i}|. \tag{6.36}$$

Recalling that the 2-D space is restricted to the region $0 \leq o_1^{(t)}, o_2^{(t)} \leq o_{max}$, this is approximately 8 parts in 10000. This is demonstrated in the scatter plot of Figure 6.12

6.5 Summary

The suggestion in this thesis has been that a stable cell behavior might be thought of as a stable dynamic attractor in the intracellular signaling system. It is therefore important to investigate stability properties of equilibrium states of the model. The simple example shown in §6.3 illustrated an attracting limit cycle generated by a network with only four different protein types. The network stability was investigated numerically and also analytically. Local stability properties of fixed points was also established in other simple networks. It would be useful to be able to enumerate the total number of different attractors in the intracellular dynamics of any given network.

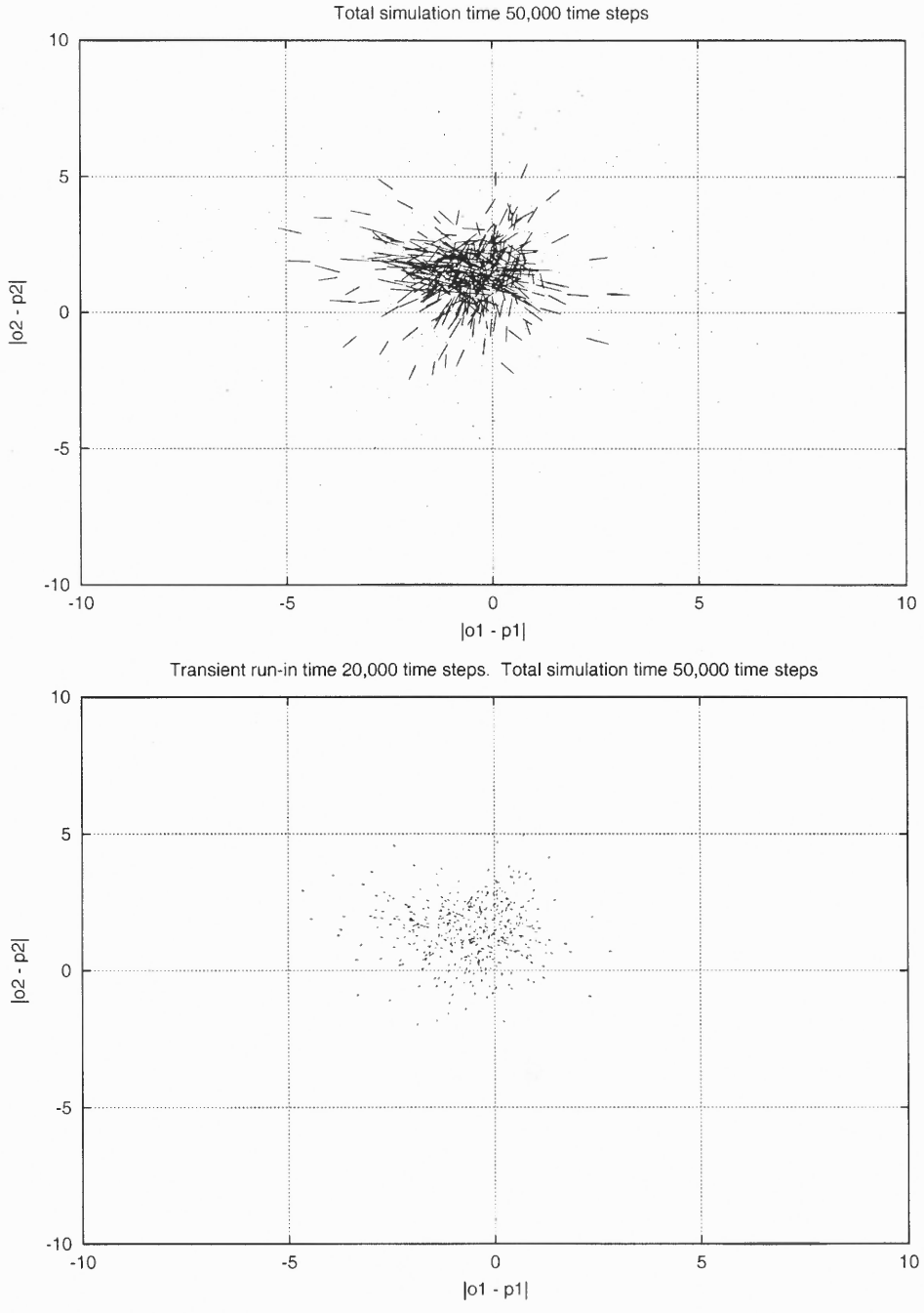


Figure 6.12: Scatter plot comparing data from truncated & double precision numerics. The data are generated by comparing the distance between respective nodes (1,2) in the network. In the case where the numerics is truncated, $\mathbf{p}^{(t+1)} = \hat{f}(\mathbf{p}^{(t)})$, the variable p is used to distinguish the truncated occupancy. The absolute distance $|o_2^{(t)} - p_2^{(t)}|$ is plotted against $|o_1^{(t)} - p_1^{(t)}|$. The limit cycle is strongly attracting, so the data remain close (i.e., within a 10×10 grid). The initial starting data for the cycle is (1000,9000,5000,5000), and the simulation is run for 50,000 time steps (upper). Data are also collected after a transient time of 20,000 time steps (lower). Note that in the right hand plot, the data are less scattered.

However, this proves to be a difficult prospect for a system of such large dimension. Nonetheless, analysis for some simple cases sheds light on possible procedures for predicting properties of larger networks.

CHAPTER 7

TOPOLOGICAL PROPERTIES OF NETWORKS

7.1 Introduction

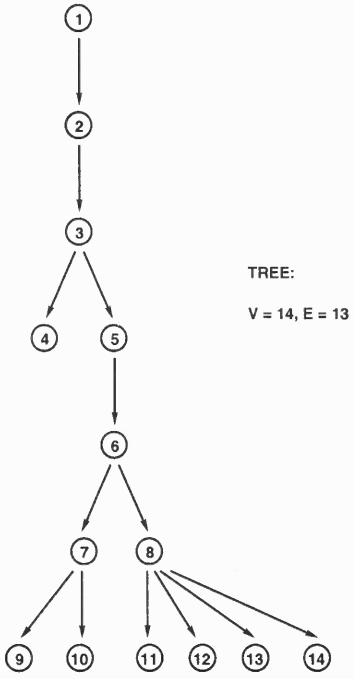
Previous chapters have been devoted to studying dynamic properties of networks both numerically and analytically. In this chapter, topological properties of networks will be explored, particularly the manner in which network structure can affect dynamics. Some terminology from graph theory is necessary, but any concepts used in this chapter are summarized for reference purposes in Appendix D. Section 7.2 explores the properties of a signaling cascade of PKs, and examines whether the model developed in this thesis can exhibit the qualitative features of biological signaling cascades. The following sections explore the relationship between structural properties and propensity of networks to generate oscillations.

7.2 Amplification Properties of Cascades

A graph can be used to describe the structure of a network. If the network structure is a linear cascade with no feedback, the graph is described as a "tree". A tree is an acyclic, connected graph satisfying the property that $E = V - 1$, where E is the number of edges in the graph and V is the number of vertices. Every node of the graph has an "in-degree" of one, which means that there is one input edge for each node of the graph (see Figure 7.1). With the synchronous updating rule, it takes one time step for a signal to propagate down one level of the tree. Since each node has a single input, the activity of each node can only increase or decrease monotonically. Eventually the activities of the nodes within the tree will converge to a steady state.

In the cascade shown in Figure 7.1 all nodes represent types of kinase, the activity values of the nodes will increase monotonically. Assume the null state and make a small perturbation in the value of the input node (Node 1), set $o_1^{(0)} = 100$. This 1% activation of the input node causes a cascade of changes which filters

A



B

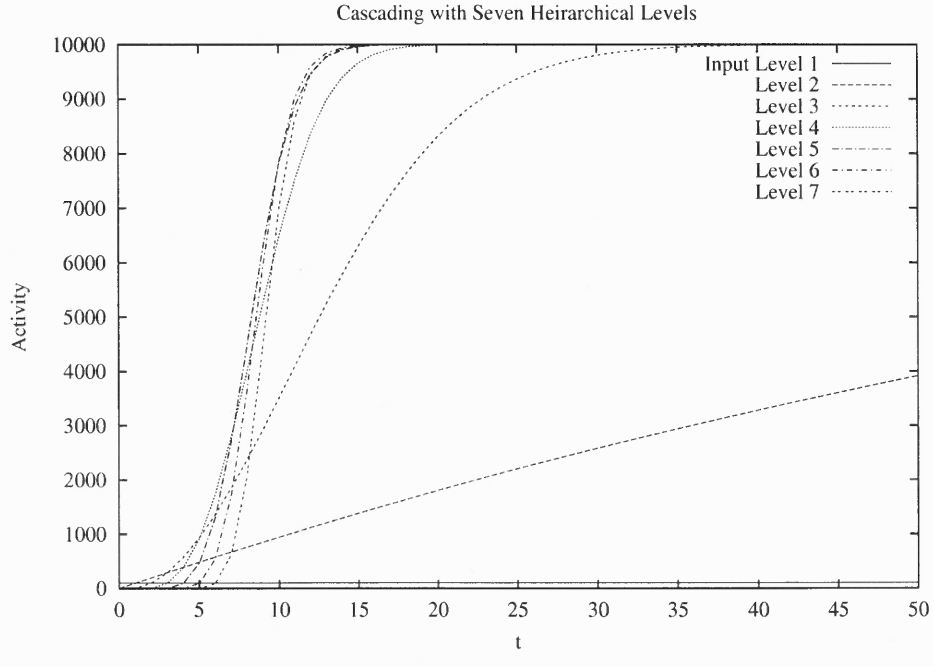


Figure 7.1: (A) A tree structure, (B) Positive amplification in a cascade of seven levels of protein kinases. A small change (1%) in the input level can cause a rapid increase of phosphorylation levels in protein types at lower levels. There is little change in the sigmoidal response beyond five levels of the pathway.

down the tree. The changes in activity of nodes 1, 2, 3, 5, 6, 7 and 9 are shown in Figure 7.1. A steeper sigmoidal response is observed in nodes at lower levels of the tree. However, there does not appear to be a significant change in response benefit beyond five levels of hierarchy.

Lemma 7.1 (Monotonicity principle) *(i) If the input to node j, $b_j^{(t)} \geq 0$, then its occupancy increases monotonically. The occupancy $o_j^{(t)}$ is bounded above by o_{max}. (ii) If the input to node j, $b_j^{(t)} \leq 0$, then its occupancy decreases monotonically. The occupancy $o_j^{(t)}$ is bounded below by zero.*

(i) $b_j^{(t)} \geq 0, \ \forall \ t$:

$$o_j^{(t+1)} = f(b_j^{(t)}, o_j^{(t)}) = o_{max}\left(1 - e^{-b_j^{(t)}/o_{max}}\right) + o_j^{(t)} e^{-b_j^{(t)}/o_{max}},$$

$$\Longrightarrow \quad \begin{cases} \dfrac{\partial f}{\partial o_j^{(t)}} = e^{-b_j^{(t)}/o_{max}} > 0, \\ \dfrac{\partial f}{\partial b_j^{(t)}} = \left(\dfrac{o_{max} - o_j^{(t)}}{o_{max}}\right) e^{-b_j^{(t)}/o_{max}} \geq 0, \end{cases}$$

$$\Longrightarrow \quad o_j^{(t+1)} \geq o_j^{(t)}.$$

(ii) $b_j^{(t)} \leq 0, \ \forall \ t$:

$$o_j^{(t+1)} = o_j^{(t)} e^{b_j^{(t)}/o_{max}},$$

$$\Longrightarrow \quad o_j^{(t+1)} \leq o_j^{(t)}, \qquad \textit{since} \ \ o_{max} > 0.$$

7.3 The Effect of Structural Properties on Oscillations

By the monotonicity principle, nodes with a "tree" (cascade) structure (with no feed-forward or feed-back loops) cannot oscillate in activity. This cannot occur even with a transient or sustained oscillatory input to the tree. A question of more general interest is whether a general acyclic graph can exhibit oscillations in activity. A simple

example illustrates that an acyclic graph can, in fact, oscillate in the special case when its input is sustained and oscillatory (see Figure 7.2). In general, an acyclic graph must equilibrate if the input signal is maintained at a constant level.

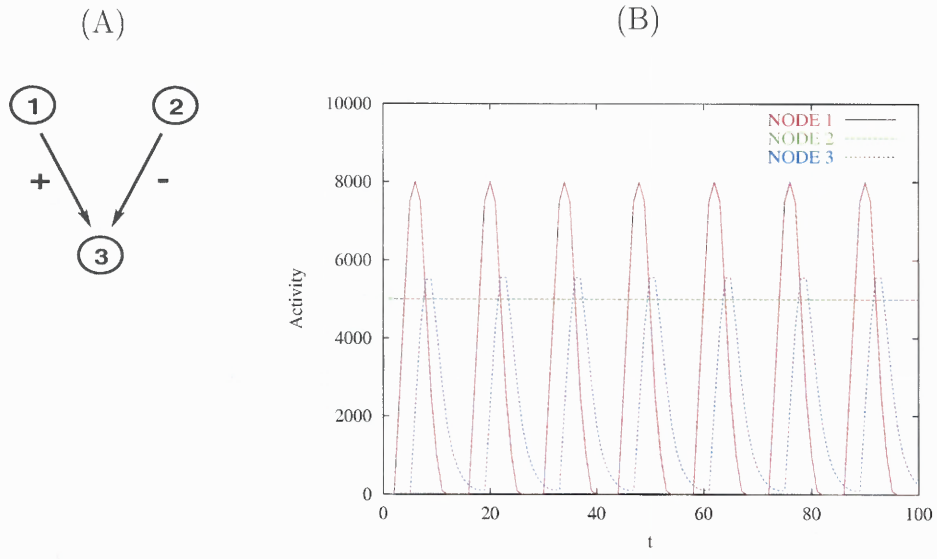


Figure 7.2: Oscillating nodes in an acyclic graph. (A) A three node acyclic graph. (B) An acyclic graph can have oscillating nodes if one of the inputs is oscillating.

If an acyclic graph can oscillate only if its input nodes oscillate, what kind of network structure can sustain an oscillation? An example of a small four node network which exhibits stable oscillations under certain conditions was demonstrated in §6.2.2. If a node oscillates, must it lie on a cycle? A simple counterexample suffices to show that the statement that an oscillating node must lie on a cycle is false. This is achieved by adding a single node and two extra edges to the four protein graph of Chapter 6 (see Figure 6.2). The new graph (Figure 7.3) illustrates that, under certain conditions, the activity of the additional node will oscillate. Must there, however, be at least one node that lies on a cycle in order to generate an oscillation?

Suppose that the inputs to the graph are sustained at a constant value. The following observations are noted. A node which receives input from an oscillating node does not necessarily oscillate itself. By the monotonicity principle, a node receiving a single input from an oscillating node will eventually stabilize. If a node oscillates,

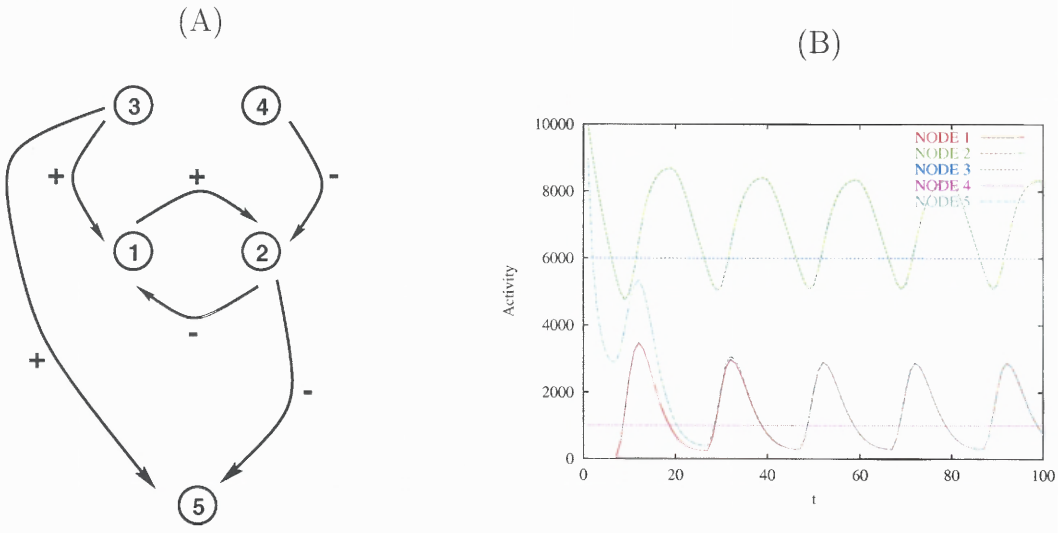


Figure 7.3: Oscillating nodes need not necessarily lie on a cycle. (A) A five node graph constructed by adding a single node and two edges to the graph studied in Chapter §6.2.2. (B) Under certain conditions, node 5 can oscillate, but it does not lie on a cycle.

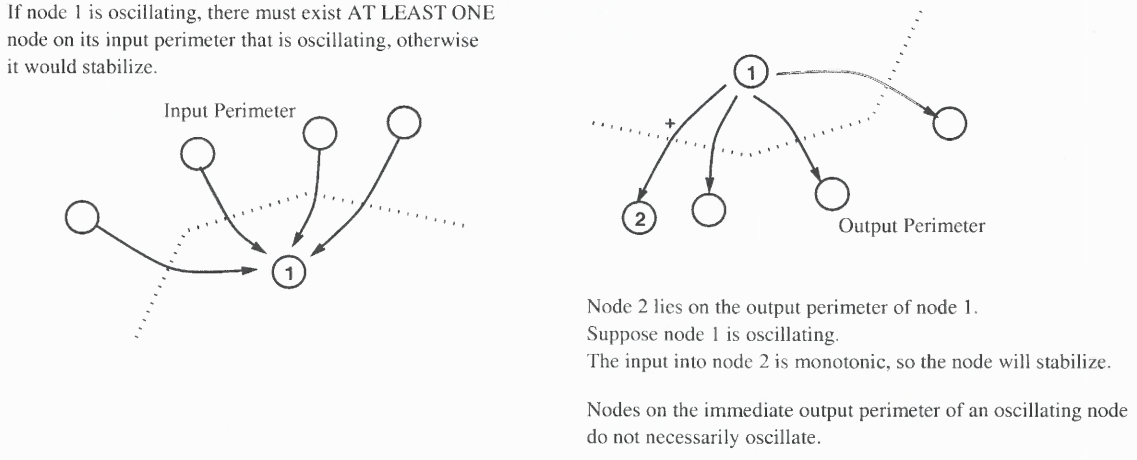


Figure 7.4: Illustration of conditions under which nodes may or may not oscillate. If a node is oscillating, then at least one of the nodes on its input perimeter must also be oscillating (left). This is not true for nodes on its output perimeter (right).

then there must exist an input to this node which is oscillating, since otherwise the node would equilibrate. If a graph exhibits oscillations, but the inputs to the graph are constant, then there must exist at least one cycle in the network. It is therefore important to consider the input and output perimeters of any given node (see Figure 7.4).

Common techniques used to traverse and study the structure of graphs are depth-first search (DFS) and breadth-first search (BFS) algorithms [98]. Depth-first search is particularly useful when one is interested in studying structural properties of graphs. DFS algorithms exist to search for cycles in graphs, to detect the strong components of a directed graph or even to search for the block/articulation point structure of a graph. A graph, G, is said to be strongly connected if there exists a path between every pair of vertices in G. An articulation point is defined as a vertex whose removal splits the graph into disjoint components. For example, node 8 in the graph of Figure 7.5 is an articulation point. Such a node could play a critical role in controlling dynamic patterns of modeled networks. One might imagine that information on the strong component structure of a graph would be a reasonable indicator of a tendency of a network to have large groups of oscillating nodes. Once one node starts to oscillate, the signal propagates throughout the component. However, as noted before, the signs of the inputs into any given node will be important in determining whether or not that node can oscillate.

Once strong components are embedded inside larger graphs it becomes harder to predict the dynamics of the network as a whole. In Figure 7.5 nodes 2 and 6 form a strong component and they are both oscillating. However, nodes 3, 7 and 8 form another strong component,, yet they are not oscillating. This is primarily because once the size of the positive input of node 8 (from node 3) rises above the amplitude of the oscillation of node 6, the input into node 8 saturates. This shuts off node 7, which prevents the nodes in the strong component from oscillating.

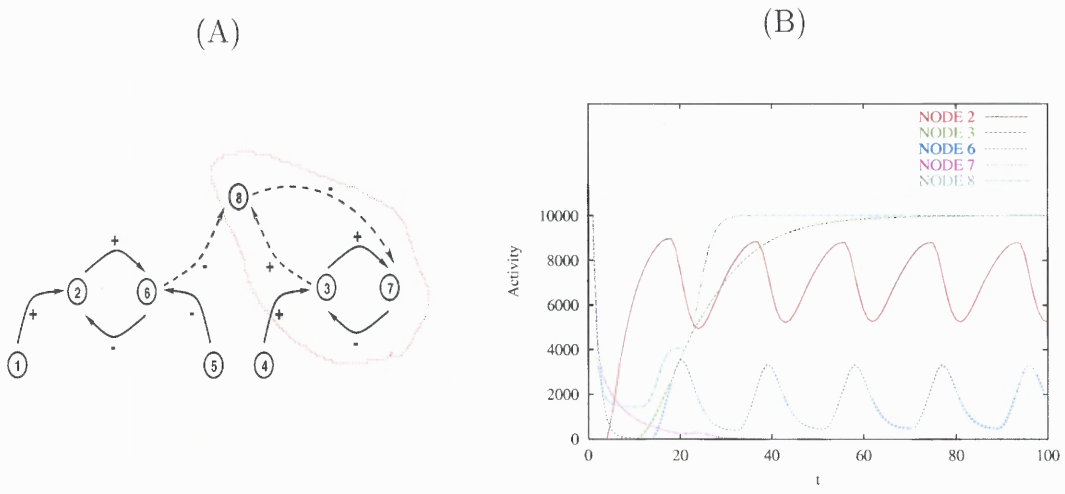


Figure 7.5: Nodes that are part of a strong component need not necessarily oscillate. (A) A network constructed to illustrate that, under certain conditions, nodes that are part of a strong component need not necessarily oscillate. (B) Plot of the activity of selected nodes against time. The values of nodes 1, 4, 5 were set to 2000, 1000, 8000, respectively. Nodes 3, 7, 8 form a strong component, but none of these nodes is oscillating.

Changes in structural properties can cause changes in the allowable range of dynamic behaviors of networks. For example, the three node graph shown in Figure 7.6, whose fixed points are listed in Table C.1, can never oscillate. The addition of a new node can introduce new behaviors under certain parameter regimes. The oscillatory behavior of the four node network in Figure 7.6 was examined in some detail in Chapter 6. Attention will be given in the next section to the structural properties of a network generated by the GA.

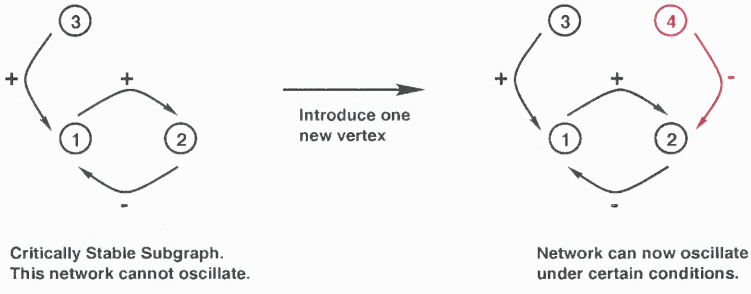


Figure 7.6: Structural changes can introduce new dynamics. The left graph can never show oscillations in activity. However, the introduction of a single extra vertex (node 4) can introduce new behaviors in certain parameter ranges.

7.4 Structural Features of a Network Generated by the GA

Tests on networks from the final generation of the GA search shown in §4.3 show that they often posses one large strong component, with some nodes with an "in-degree" of zero, which can essentially be regarded as input nodes. Tests on the block structure of these networks also indicate a small number of articulation points and hence a small number of large blocks. More sophisticated algorithms are needed to shed light on how the network structure can affect its dynamical properties. This is discussed further in the final chapter.

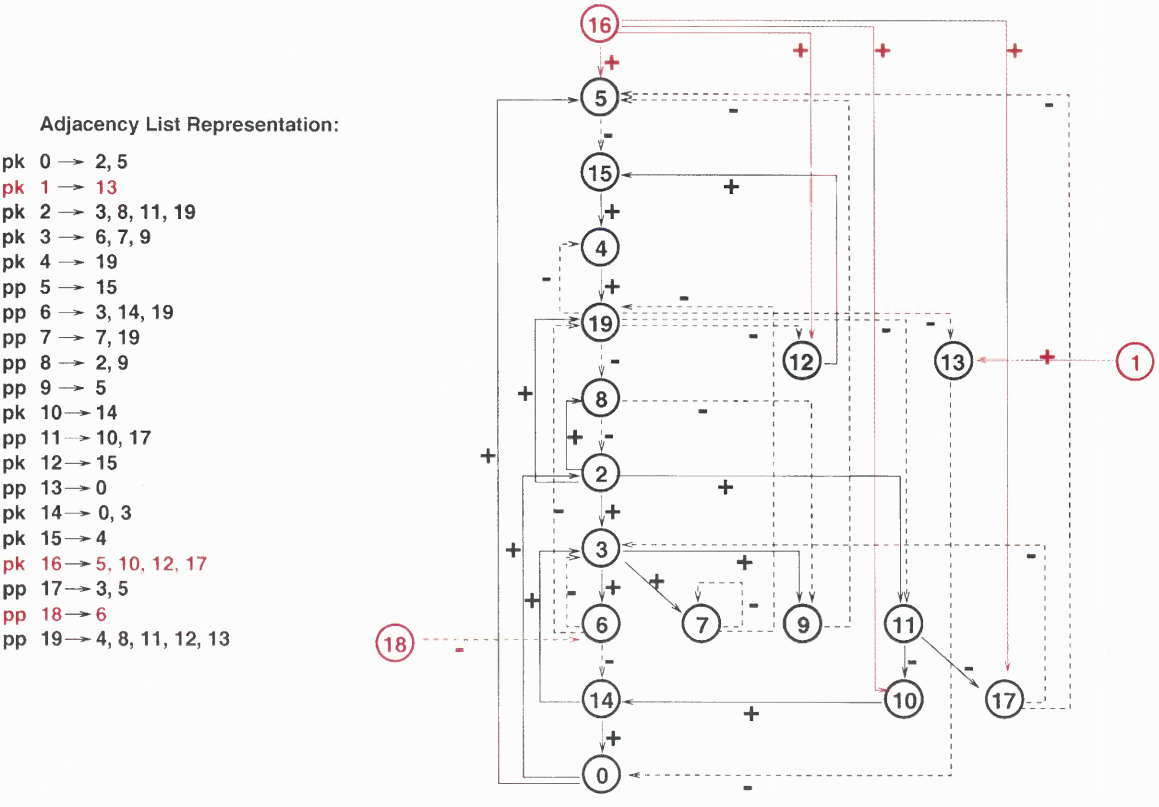


Figure 7.7: Graphical structure of the fittest network generated by the GA simulation of §5.4. Input nodes are shown in red. Dashed lines denote negative interactions. This graph was generated using a DFS algorithm. All nodes (excluding nodes 1, 16, 18) form one large strong component. Note the genetic algorithm has altered the ratio of types of protein kinases (PKs) to protein phosphatases (PPs) in the network. The fundamental cycles of the graph were found to be: (2,3,6,14,0,2); (5,15,4,19,8,2,3,6,14,0,5); (3,6,14,3); (19,8,2,3,6,19); (19,8,2,3,7,19); (5,15,4,19,8,2,3,9,5); (5,15,4,19,8,2,11,17,5); (19,8,2,19); (15,4,19,12,15). The starting vertex for the DFS algorithm was node 16.

An algorithm [98] was implemented to detect the fundamental cycles of a directed graph. The cycles detected are noted in the caption of Figure 7.7. Oscillations in activity of many nodes were clearly evident under certain input conditions (see Chapter 5). The detection of closed structural loops is not sophisticated enough to elucidate a structural explanation for oscillations, as the sign of the inputs into each node is important. This is where the mathematical analysis of small networks becomes important.

The four node substructure discussed in Chapter 6 is embedded in this larger network:

$$(15) \xrightarrow{+} (4) \underset{-}{\overset{+}{\rightleftharpoons}} (19) \xleftarrow{-} ((6)-(2))$$

$$(0) \xrightarrow{+} (2) \underset{-}{\overset{+}{\rightleftharpoons}} (8) \xleftarrow{-} ((19)+(7))$$

$$((2)-(17)+(14)) \xrightarrow{+} (3) \underset{-}{\overset{+}{\rightleftharpoons}} (6) \xleftarrow{-} (18)$$

Nodes 4 and 9 will oscillate provided that neither of their inputs hits the upper or lower activity limits, and provided that the negative input from node 6 to node 19 is greater than the positive input of node 2. Similarly, nodes 2 and 8 will oscillate provided that neither the activity of node 0 nor the combined activity of nodes 19 and 7 hits the upper or lower limits. Nodes 3 and 6 will oscillate under the condition that the positive inputs of nodes 2 and 14 to node 3 outweigh the negative input of node 17, and that neither of the inputs to nodes 3 and 6 hits the maximum or minimum levels. This preliminary analysis helps to shed light on why oscillations might be such a prevalent feature of this network. However, more detailed analysis is required.

7.5 Summary

In this chapter, some attention was given to trying to understand how structural properties of graphs affect their dynamics. However, structural properties alone are not sufficient to elucidate all the complexities. Fundamental cycles were examined and the strong component structure of graphs was studied in certain special examples. For

example, the nodes in the GA network, excluding the three input vertices (1, 16, 18), form one large strong component. This may explain the tendency for all nodes in the graph to oscillate. In Chapter 5 a transient change in node 19 was shown to have significant impact on the dynamic properties of the network. Structural analysis of this network indicates that node 19 has the greatest number of output connections. This may explain why this node is so critical to the state of the network. The sensitivity of the network state to changes in individual nodes could be interesting to explore in any future work.

A "tree" structure was studied and shown to possess the typical signal amplification properties seen in real signaling cascades. The benefit of studying several small networks in depth becomes clear when properties of larger networks are considered. The four node network configuration studied in Chapter 6 is embedded inside the network generated by the GA studied in this chapter. It then becomes more obvious why the network dynamics typically show large numbers of oscillating nodes.

It is of interest to examine whether there are certain underlying structural principles governing the design of intracellular signaling networks. For example, studies of "small-world" networks [99] indicate that graphs with a high degree of local clustering but with a short characteristic path length between vertices may have advantages over completely random or completely regular networks. This is discussed further in the final chapter.

Since gene knockouts experiments are also possible in real signaling networks, it would be interesting to study the effect of deleting nodes in simulated networks. This would be a fruitful area of study for future work. In this thesis, a decision was made to concentrate on mathematical analysis of several specific small example networks in order to try to understand the mathematical basis for the generation of oscillations.

CHAPTER 8

DISCUSSION

A correspondence between the nodes in the simulated networks and the signaling proteins in a cell has been described. When these networks are compared with real cells, any specific matrix of interconnections is analogous to the fixed set of kinases and phosphatases found in a particular cell type (i.e., one connection matrix represents one generic cell type). Since the number of model cell types that can be generated is enormous, the initial Monte Carlo search technique was replaced with a more efficient search algorithm known as a genetic algorithm [77]. The advantage of using a genetic algorithm is that a population of cells can be compared and evolved using a biologically motivated fitness function.

One clear feature of simulations of these signaling protein networks is that the activities of protein kinases and protein phosphatases settle rapidly into limit cycle attractors or fixed point attractors at low levels of connectivity. Even randomly selected networks seem to show a reduced set of distinct patterns of behavior. Protein types appear to show oscillations in activity that vary through the entire range, while stability appears to occur at the upper and lower saturation limits. As the average network connectivity is increased, the occurrence of periodic oscillations in activity also increases. Networks with a fixed set of interconnections are capable of maintaining themselves in distinct steady states or cyclical configurations. It is possible that these features may be observed in real cells. Futile cycles of ATP [59] may play a role in maintaining cell stability by controlling phosphorylation states of important intracellular proteins. These kinds of phenomena may be readily achieved in phosphorylation networks with a reasonable degree of interconnection.

These phenomena occurred with simulations of networks containing twenty protein types and one hundred protein types. Data from one hundred protein networks also showed evidence of attractor basins, although the dynamics were

more complex. Regardless of their exact nature, the settling into a dynamic attractor indicates that there is a considerable restriction in the number of possible states that the complete cell signaling system can occupy. This suggests a degree of spontaneous order in network behavior [71, 72].

These initial observations are similar to those seen in other models of complex biological systems (e.g., immune system and neural network models [100, 101]). The strong tendency to settle into attractors is also a feature of a previous generic network model of intracellular protein-protein interactions [77]. This work can be seen as modeling an epigenetic system that operates in the cytoplasm of cells on a faster time scale. The genomic regulatory system and cell signaling networks may be viewed as slow and fast lattice automata respectively [102]. In fact, the time steps in the model can be interpreted in terms of real time scales. One time step is probably readily interpretable as approximately one second of real time. This interpretation is based on turnover rates of protein kinases (as discussed in Chapter 2).

The model of Chiva *et al.* differs from that presented here in the sense that their protein-protein interaction is even more generic. In this study, a more specific analogue of the most common form of protein regulation, namely protein phosphorylation, has been modeled. In their model, simulated networks are also fully connected.

The dominance by attractors is also seen in one of the original presentations of a connectionist model applied to cells [87]. Kauffman has shown that Boolean networks that are constructed with biologically plausible constraints placed upon the number and type of connections can display a limited number of dynamic attractors [72]. The Boolean rules in the networks are considered to represent different types of regulation of gene expression in the nucleus of cells. The attractors in such genomic networks are suggested to correspond to distinct differentiated cell types [103]. In contrast to gene regulation networks, the attractors in the networks presented here may represent

a form of epigenetic cell stability that exists principally in the cytoplasm of cells. It probably operates on a faster time scale than that which involves changes in gene expression.

The protein kinase and protein phosphatase based signaling networks simulated here show evidence of a finite, relatively small number of distinct attractor states. Since the connection pattern in a particular network is fixed, this shows that any one modeled cell can display a variety of different and stable dynamical patterns. Each attractor pattern may correspond to a mode of cell behavior. This suggests that the rules of signaling proteins may readily give rise to a small variety of distinct modes of cell behavior. These stable patterns exist as a consequence of simulating protein-protein interactions and are not dependent upon changes in gene expression.

Real cells can show stable states, or even stable changes in behavior, in the absence of expression of new protein. This is illustrated by a process such as long-term potentiation in hippocampal neurons where changes in synaptic transmission can occur for up to about four hours without new gene expression [104]. For a review see [3].

The short-term effects of insulin on a cell amounts to a variety of kinases and phosphatases switching on, or off, in a manner that defies easy explanation as a single cell signaling pathway, or as a simple gene switch response [105]. However, it is readily interpretable as a shift of the cell from one signaling attractor to another. Interestingly, there is also evidence that some uncharacterized persistent epigenetic state accounts for the differences in the transformation potential of cultured fibroblasts [106]. Epigenetic and cytoplasmic states have also been suggested to underlie various modes of cell movement [107].

Nevertheless, in most biological instances, any network of stability in the cytoplasm will be interwoven with constant changes in the concentration of proteins involved in signaling. These fluctuations in protein concentrations could be included

in future models of protein kinases and protein phosphatases of this type. The simulated cell may then resemble the structure of modified cellular automata lattices where sets of rules interact on a fast time scale, but the rules themselves change on a slower time scale [102].

In practice it may be difficult to observe directly a range of different attractor patterns that this study predicts to exist in a real "resting" or stimulated cell. It may require the recording of the activities of many proteins in a single cell on a time scale of a few seconds. However, some simulations have suggested that observing the dynamics of the most variable three proteins in real-time may reveal signs of the distinct underlying attractor states. In addition, a novel suggestion presented by this work suggests that the pattern ATP usage may reflect the state of the cell and the underlying dynamics of the intracellular signaling network. This would be similar to the manner in which EEG recording reflects the complex dynamics of the activity of billions of neurons in the brain. Measuring the rate of use, or else the precise concentrations, of ATP in a living cell may be possible in the near future. Currently, the signal to noise ratio in experimental ATP measurements is poor.

It is suggested here that the finite number and character of different attractors in a given network could be an indication of the range of allowable distinct modes of cellular behaviors that any one cell type can display. A different perturbation of the initial state may shift the network into a new state, either a different stable state or limit cycle attractor. Altering the activity of some nodes may not alter the activities of the other proteins, but altering other nodes may have a more dramatic effect.

It will be difficult to test some of these ideas since there is really no precise way of defining a cell's behavior. In fact, as a converse, a particular attractor of the intracellular signaling system may be a new way of defining a behavioral state of a cell.

8.1 Relaxing Model Assumptions

The simplifying assumptions that were used in developing the initial model were listed in Chapter 2. The model structure is flexible enough to enable several of these assumptions to be relaxed. Most importantly, though, it incorporates certain important aspects from the biology:

- *Dynamic Regulation of Proteins*: A network of protein phosphorylation reactions has been considered in this thesis. This seems reasonable as kinases and phosphatases make up the bulk of the different types of elements in the intracellular signaling network and are known to control many features of cell behavior [32, 45].
- *Crosstalk*: This is captured by the connectivity parameters and the fact that the each modeled cell type is represented by a matrix of interactions.
- *Parameter Encodings*: Parameters allow the ratio of protein kinases and phosphatases to be varied, as well as the degree to which they interact with other protein types. Networks of different sizes can also be simulated. The current assumption of equal numbers of each type may not be unreasonable [53].

The flexibility of the framework allows future investigation into:

- *Multiple regulatory domains per protein type*: For example, Boolean logic rules could be used to determine the interactions between regulatory sites. This would translate into a more complicated function relating the activity of protein types to their occupancy. Biologically speaking, this relates back to the concept of multi-site phosphorylation that was discussed in the introductory chapter.
- *Asynchronous updating*: Currently the updating rule acts in a synchronous fashion. It would be interesting to investigate the effect of adding stochastic noise to asynchronize nodal updating. This would remove the assumption that all regulatory processes occur on the same time scales. For example, the shortest time-scale could be established first and updating performed relative to this. This is the method used in Morton-Firth's stochastic simulation program "StochSim" [108]. The role of noise in

signaling systems is another current topic of interest [109, 110]. However, stochastic models of signaling networks are still uncommon [108, 111].

• *Variation in concentration levels of proteins*: A change in the maximal concentration of each protein type involves a change in o_{max}. This number could be varied over time or made to be different for each node. This would then involve the addition of a new state of the network representing concentration levels of all the nodes within it.

• *Addition of genetic complexity*: Additional levels of complexity, such as gene expression, could be linked to changes in concentration levels of nodes (as mentioned in the previous point) to mimic the feedback of genetic information to the protein signaling network.

• *Addition of second messengers*: Second messenger molecules could be included by introducing additional nodes into the network to create signals that feed into the phosphorylation network. The model, as it stands, can be used to examine effects of changes in inputs to the network, if nodes with no inputs are considered to be external influences. Hence, changes in initial conditions of such a network mimic the effect of external signal perturbation. The phosphorylation network could be considered as a subnetwork of intracellular signaling processes.

• *Connection to an extracellular environment*: The simulation results presented in this thesis have concentrated on the internal or intrinsic dynamics of cell signaling networks. There are no explicit connections to an extracellular environment. This is not unreasonable as there are examples of biological systems which can change their state independently of "external" receptor inputs (e.g., developing embryos and some protozoa). However, such connections could be achieved by designating particular nodes in the networks as "receptors" whose activity state would be determined by some parameter outside the network. It would then also be possible to select networks via a genetic algorithm designed to have the cells perform specific computational tasks [94, 112].

8.2 Genetic Algorithm Improvement

A genetic algorithm was developed in Chapter 4 to select networks demonstrating a variety of distinct dynamic behaviors. The GA provided an effective "hill-climbing strategy" as it allowed a simultaneous, parallel search of a rugged landscape by evolving a population of cells against a "fitness" criteria. Cells were ranked according to fitness and reproduced asexually to the next generation. The crossover operator was designed to mimic the evolutionary mechanism of protein domain shuffling. The mutation operator is a common feature of most genetic algorithms. This is also motivated by the biology, as genetic mutations occur at a low rate in all cells.

The mutation and crossover rates, together with the population size and number of offspring per parent, were seen to have an impact on the performance of the algorithm. Increasing the population size by allowing more offspring per parent appeared to improve the search. The GA generally showed the greatest improvement in early generations and then fitness levels would often plateau. Future improvements to this algorithm might include a mutation or crossover rate that could be varied over time.

The GA fitness function was designed to search for networks with varied dynamical properties after fixing the levels of connectivity. (The search was restricted after initial investigations via Monte Carlo simulations.) A more biologically motivated fitness function might be possible using techniques from information theory. The information processing ability of cells is important and addressed in the literature [48, 113]. The concept of mutual information has also been used to study properties of cellular automaton models [86]. One possible approach might be also be to start the GA with a population of small networks, but allow addition of protein types to "grow" networks.

8.3 Small Network Analysis

The main reason to study a dynamical rule whose nodal representation takes discrete values is that the interpretation of the activity is related to the number of protein molecules, which is a discrete quantity. Another reason for studying a discretized mapping was to take an approach that was different from the study performed by Chiva & Tarroux [77] where their nodal description of protein activity was based on continuous variables. However, the form of the dynamical rule used to study protein interactions has been difficult to analyze mathematically,

In Chapter 6 the stability of fixed points in several small networks was examined by analyzing properties of the the non-discretized map, f. This chapter was also concerned with understanding how oscillations arise in small networks. A proof of the existence and local stability of a periodic orbit was presented for a simple network containing only four protein types. The premise for these studies was that there was good agreement between numerical studies for the truncated map, $\hat{f}$, and the non-discretized map, f. Plausible arguments were given for useful methods to study stability of fixed points in these networks. This said, however, the correlation between these data has not been proven conclusively and needs to be studied further. The truncation alters the space on which the function is acting from the reals to the integers, and so some further justification of the correspondence should be provided.

When large numbers of molecules are concerned one might argue that it would be reasonable to move to a continuum model and consider the percentage of activated protein molecules as a continuous quantity. In this case, the analysis provided in this chapter could be used. In the case of small numbers of molecules, both the rule and the mathematical simplifications necessary to analyze it fall short of the requirements set by the biology.

The analysis in the phase plane of §6.3 illustrated in a certain special case, how the mapping takes different forms depending on inputs to each node. This made the

local stability analysis of the fixed points more involved. Analysis of these points was achieved by considering an approximation of the map (the rule f_{tanh}). This analysis may extend the realms of possibilities for predictive methods for stability of fixed points. However, the fixed points lie precisely where f has a discontinuity in its derivative. Therefore, care must be exercised when interpreting these results and using them to gain insight into properties of the truncated map $\hat{f}$, or the non-discretized map, f. Global stability of fixed points and periodic orbits also remains a further area for analysis.

In Chapter 6 it was noted that the periodicity of the limit cycle of the four protein network was sensitive to changes in input parameters. This was based on an analysis of an exact repeat time using the updating rule, $\hat{f}$. The biological interpretation of periodicity might be better represented by the dominant frequency response.

8.4 Graph Theory Analysis

Chapter 7 was concerned with investigating how network structure affects function. The strong component and block structure of a network generated by the GA was investigated. These data indicated that most networks generated by the GA consisted of one single strong component with a small number of articulation points. Data visualization (i.e., the presentation and layout of complex graphs) is also of some interest. A depth-first search algorithm was used in this study to investigate the topology of a GA network.

Topological structure of signaling networks will impact their robustness to nodal deletion (analogous to "gene knockout"). It would be interesting to study systematically the effect of knocking out individual nodes on the dynamic properties of evolved networks. The sensitivity of the network state to internal changes in individual nodes appears to be important. This was seen in Chapter 5, where changes

in a single, highly connected internal node shifted the network from one attracting state to another. Transient changes in internal nodes may also cause a permanent shift of state. These remain areas for future study.

Watts and Strogatz [99, 114] have suggested that signaling networks might have a selective advantage were they to have a "small-world" structure. Small-world networks are defined to have a high degree of clustering and a short characteristic path length. The characteristic path length is a measure of the typical separation distance between vertices in the network and can be considered to be a global property of the network. The characteristic path length is a measure of the "cliqueishness" of a typical neighborhood of any given vertex in the network and, hence, is a local property of the network. Small-world networks are conjectured to have a structure which provides enhanced signal propagation speed, computational ability. It is also suggested that in such networks nodes can synchronize their activity more easily than if the network possessed an regular or random structure. GA networks could be tested for such properties.

Jeong *et al.* have shown that metabolic networks taken from 43 different organisms possess the same topological scaling properties. "Scale-free" properties of metabolic networks were first suggested by Barabasi and co-workers [74]. The probability that a substrate participates in k reactions follows a "power-law" scaling: $P(k) \approx k^{-\gamma}$. Similarly, the probability that any chosen substrate is produced by k reactions, not only possesses the same scale-free property, but a log-log plot of $P(k)$ against k is a straight line with the same gradient. Furthermore, the constant γ also appears to be independent of organism. GA networks could also be tested for similar "scale-free" features.

It might also be interesting to construct a generic network which possesses structural properties analogous to the bacterial chemotactic signaling network to see if qualitative properties of the model presented here and the Bray model are comparable.

8.5 A Model for Long-Term Potentiation

The process by which the nervous system changes and adapts over time is described as "neuronal plasticity". The cellular mechanisms involved are not yet clearly understood. There may be changes at the morphological level (e.g., narrowing of the synaptic cleft or the enlargement of functionally active synaptic areas), at the electro-physiological level (e.g., at the membrane level), or at the biochemical level (e.g., involving regulatory changes at the molecular level or changes in transmitter release or metabolism) [104]. Long-term potentiation (LTP) is the process by which the response of a postsynaptic neuron is enhanced and sustained following certain types of high frequency stimulation. It is a type of activity-dependent modification of synapses and it is currently thought that this may underlie the formation of certain types of memory in the brain. LTP is generally seen when the presynaptic and postsynaptic cells are both active simultaneously. This is often explained as an NMDA channel effect. It is generally accepted that there are three phases of LTP which are on different time scales. There are short term, intermediate and long term changes. Short term changes may involve post-translational mechanisms such as conformational changes of molecules; intermediate changes may be changes in the structure of the synapse, and long term changes may involve gene expression [104].

Different types of molecules may be involved in LTP at different synapses. The morphology of synapses may also be important, as well as the age of the tissue sample. There is also no general consensus in the literature regarding whether the mechanism of LTP is presynaptic, postsynaptic, or whether retrograde messengers (such as nitrous oxide (NO) or arachadonic acid (AA)) are involved in signaling. There is, however, evidence for an explicit role of signaling proteins in LTP and also in long-term depression (LTD). Specifically, certain types of kinases and phosphatases play an important role in LTP in some neurons (e.g., CAMK-II, PKC, PP-1).

CAMK-II has been identified as a major protein in the postsynaptic density of synapses and is therefore ideally situated to respond to calcium influx through the NMDA receptors. CAMK-II has also been shown to be controlled by auto-phosphorylation *in vitro*. This kinase can auto-phosphorylate and maintain activity in the absence of calcium/calmodulin which triggers its activity. Lisman [115, 116] has suggested that CAMK-II can become a bistable molecular switch which may be responsible for the maintenance of LTP. The activity of CAMK-II has been shown to be increased for long periods, up to one hour subsequent to the induction of LTP. It has been suggested by Lisman that the number of molecules of phosphorylated CAMK-II might correspond to the strength of the synapse. Genetically modified mice with a specific form of CAMK-II "knocked out" have been shown to have a spatial memory deficit [117, 118].

One idea for future uses of a generic model of the type presented in this thesis is to attack problems like LTP where the data are predominantly unknown. The intermediate form of LTP (which is independent of gene expression) could be considered and a genetic algorithm developed to search for networks which display some of the dynamics features of LTP [119]. This kind of modeling might therefore provide insight into the kinds of intracellular interactions that may underlie changes in cellular behavior such as synaptic efficacy. Different topological structures of networks may underlie LTP in distinct cell types. A shift from normal synaptic transmission to enhanced synaptic transmission could be interpreted as a shift from one attractor to another as a result of changes in input conditions (stimulation paradigms). Time scales of interaction would then become important as synaptic transmission occurs on a much faster time scale than post-translational changes such as protein phosphorylation.

8.6 Modeling Complex Networks

The structure and dynamics of complex networks are of general interest to the scientific community. Electronic communications networks, metabolic networks, neural networks, and social networks impact upon our lives in one way or another. Intracellular signaling systems illustrate some of the difficulties that arise when studying the properties of such ensembles. Not only do these systems possess inherent non-linearity, but complexity is also added due to the large number of different types of interactive elements involved. Furthermore, the structural properties of the network are in a constant state of flux, changing over time as the intracellular signals alter gene expression. Cells are also structurally compartmentalized, and so some signaling events will occur exclusively within membrane bound organelles (e.g., the nucleus).

Advancing techniques in cellular and molecular biology continue to propel the reductionist approach. The ultimate underlying and understated assumption is that once the details are known, everything will be understood. However, one only has to look into the intracellular world (see Figure 8.1) to wonder how all the databases of detailed and sometimes conflicting information can be integrated into a framework that sheds light on how and why cells behave in the way that they do.

Mathematical modeling in the area of cell biology presents many challenges. A researcher must decide what kind of detail to incorporate into a model. Mathematical techniques to analyze the dynamics of networks with massive structural (and hence dimensional) complexity are minimal. The predictive goals will most likely dictate the choice of the model. If the decision is to develop a continuum model, how will the model differ from previous work? What are its primary aims? What predictions might it suggest to experimentalists? How are the parameters set for the enzymatic equations if the data are not known? How sensitive is the model to changes in its parameters? What happens when there are only a small number of molecules of a particular molecular species present? Does the continuum model break down in

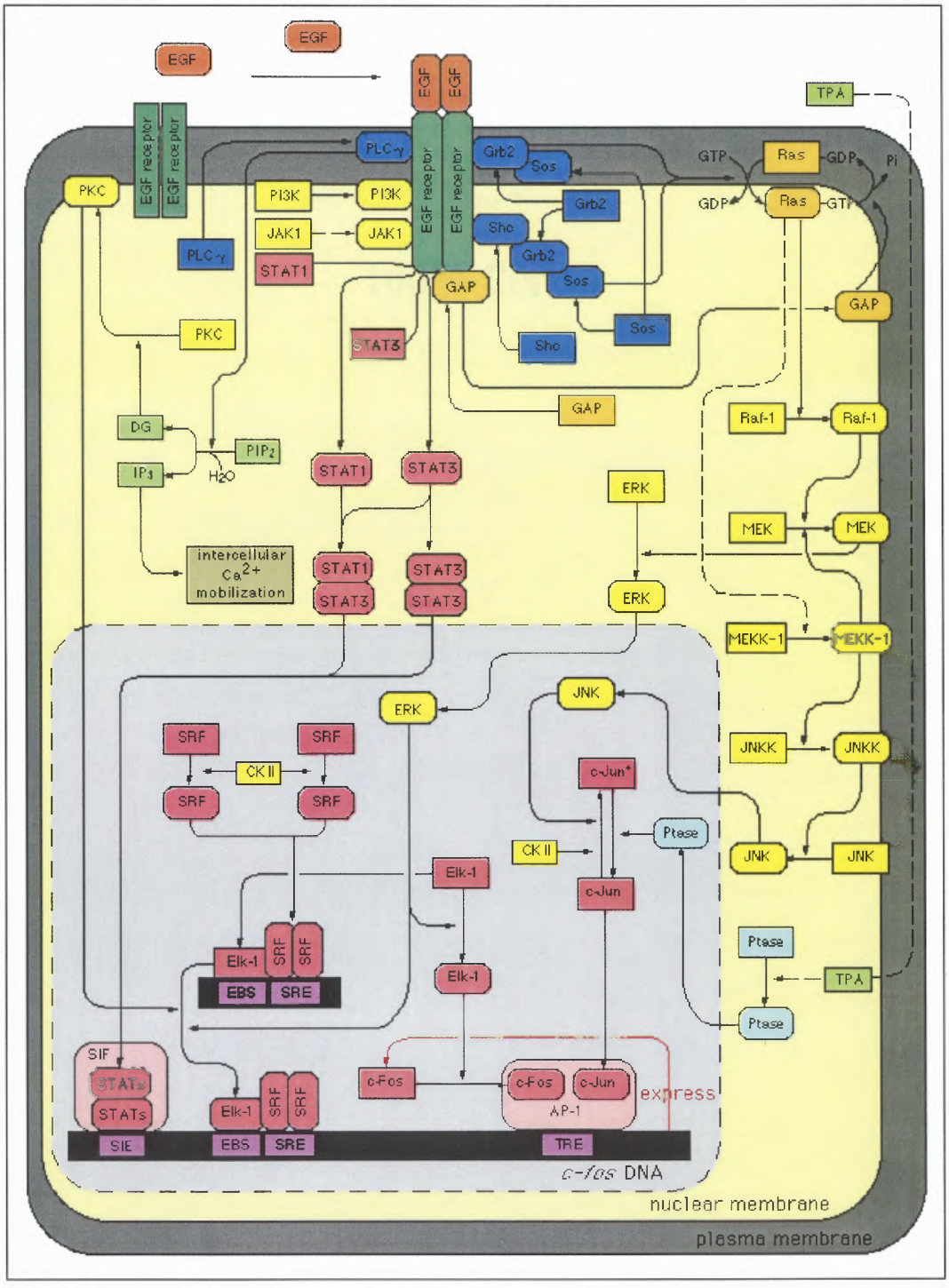


Figure 8.1: A modern representation of signaling pathways that involve MAP-kinase and the EGF receptor. Taken from Ref. 2, Appendix A. Models must take into account the cross-talk between pathways. It is interesting to compare this modern view of EGF signaling to the signaling pathway presented in Chapter 1, Figure 1.6.

these circumstances? The model should also be accessible, understandable and usable by biologists. This has motivated the discussion and development of novel software platforms for modeling metabolic networks [108, 120, 111, 121, 122]. Some of these are now freely available on the Web.

Continuum models of biochemical systems therefore exist in many forms. Specific intracellular components have been particularly well studied. Calcium dynamics have been of special interest because of the fact that real-time techniques already exist to study the spatio-temporal properties of calcium signals [36, 123, 124]. Limited numbers of studies address the issues of whether there are oscillations in protein phosphorylation levels in cells [125, 126]. This is because the biochemical techniques to monitor real-time changes in protein phosphorylation simply do not exist. Stage-specific changes in phosphorylation levels (e.g., during development) require painstaking gel electrophoresis work [127].

The intracellular signaling network of *E. Coli* remains the most well studied in the context of bacterial chemotaxis. It is the first system where some behavioral events have been explained at the molecular level [128]. This is because of the small number of intracellular components involved, which enables a comprehensive listing of features. The enzymatic components and data are thus fairly complete. This not only allows testing of models for predictive ability, but also permits investigation of the structural properties of this network and its sensitivity to parametric perturbations. In fact, it has been suggested that some functional properties of bacterial chemotaxis are robust to perturbations and some are not [129]. This suggests the question of how structure affects function. Are there any underlying design principles common to all intracellular networks, regardless of organism? Conservation and transmission of signals is evident across species by the MAP-K pathway [130], but it is not currently known whether signal transduction mechanisms are more evolved in mammalian cells [131].

Topological properties of complex networks are also a topic of current interest in the literature. Watts & Strogatz have suggested that a "small-world" structure may have important dynamical consequences for complex networks [99, 114]. Metabolic networks also seem to possess "scale-free" features seen in other large networks [74, 132]. Continuous addition of new nodes is also a common feature of real world networks. The addition or deletion of vertices are often not distributed in a uniformly random pattern. Such networks, therefore, exhibit some form of preferential connectivity [74]. Cell signaling networks are no exception. Jeong *et al.* have proposed that the most "connected" intracellular substrates may be preserved between species [132]. However, as in other complex systems, detailed topological data are currently unavailable for many intracellular signaling networks.

The more one becomes embroiled in the data the more one wonders whether there isn't a place for generic models of signaling networks which will necessarily not include all the detail. These types of models would not be used to predict what a specific cell will do when a particular chemical is applied in certain concentrations. However, what they can do are develop frameworks to address more statistical questions that are are not well addressed by more detailed continuum models. Such models might offer insight into why some gene knockout experiments appear to have very little effect upon cell function [29]. The overall frequency and potency of disruptions of cell behavior by such gene knockouts may represent the kind of observable in cell biology that is only predictable by a generic model of intracellular signaling. It may also be possible that a network containing a large number of signaling elements with a relatively low level of connectivity can provide cellular networks with the stability that they need. These models may follow in the footsteps of other connectionist models [87, 101, 133, 134]. Whatever modeling approach researchers decide to take will depend on personal bias and beliefs. However, they will still need to think about the question: How much of the detail is really important?

8.7 Concluding Remarks

With all the recent excitement in the human genome project [135], endeavors of bioinformatics, proteomics, and genomics may well be the emphasis of biology in the 21st century. It must be remembered, however, that beyond the genome, there is also the combinatorial complexity of the protein machinery. Many different proteins can be encoded in a single gene, and proteins themselves have their own epigenetic regulatory mechanisms. The genome may once have been thought to be the blueprint for life, but the subtleties of control mechanisms inside cells still remain largely elusive. Cells are the most fundamental units at which "life" properties emerge. Living systems are not chaotic, nor are they completely random or completely deterministic. It has also been proposed that living systems may naturally evolve towards a region showing complex, or "edge of chaos" characteristics [72, 86].

Despite the rapid accumulation of genetic data, the most well studied signaling network is still that of the bacterium *E. Coli.* There is no satisfactory explanation as to why any given cell type exhibits only a limited number of behavioral modes. Very few mathematical or computational models currently exist to address this question. For these reasons, the dynamics of intracellular signal transduction networks will remain a fruitful area of study for the future. Perhaps a combination of different modeling approaches will finally shed light on these mysteries. In any case, the road ahead will be, no doubt, an exciting one.

"Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference."

Robert Frost (1874–1963)[†]

[†]An excerpt from "The Road Not Taken" by Robert Frost (1916), pp. 70, Ref. [136].

APPENDIX A

CELL SIGNALING SITES & PROTEIN DATABASES

	Web Site (Cited March 25, 2001)	*Description*
1.	http://vlib.org/Science/Cell_Biology/signal_transduction.shtml	The WWW Virtual Library: Cell Biology—with information on other signal transduction sites of interest
2.	http://www.grt.kyushu-u.ac.jp/spad/	Signaling Pathway Database—contains diagrams of cell signaling pathways
3.	http://geo.nihs.go.jp/csndb/	Cell Signaling Networks Database—a signal transduction database [137]
4.	http://www.sdsc.edu/kinases/	The Protein Kinase Resource—data available on the enzymology, genetics, molecular and structural properties of protein kinases [138]
5.	http://www.expasy.ch/sprot/	SWISS PROT Database—contains protein sequences with functional and structural information [139]
6.	http://www.expasy.ch/prosite/	Prosite Pattern Database—contains information on protein families and protein domain structure [140]
7.	http://www.cbs.dtu.dk/databases/PhosphoBase	PhosphoBase—a database of phosphorylation sites in proteins [141]
8.	http://www-lmmb.ncifcrf.gov/phosphoDB	Phosphoprotein Database—site dedicated to protein phosphorylation
9.	http://www.rcsb.org/pdb	PBD Brookhaven Crystallographic Database—a protein data bank containing 3-d structural X-ray crystallographic data [142]
10.	http://www-nbrf.georgetown.edu/	Protein Information Resource—maintains a protein sequence database, the PIR-International Protein Sequence Database [143]
11.	http://www.ncbi.nlm.nih.gov/	National Center for Biotechnology Information
12.	http://www.ncgr.org/software/pathdb	PATHDB: Metabolic Pathways Database—contains information on pathways relating to metabolism in plants.

APPENDIX B

A MODELED CASCADE SYSTEM

The mono-cyclic cascade system [23, 24] consisted of a protein kinase (cAMP-dependent protein kinase) and a protein phosphatase (phosphoprotein phosphatase) together with regulatory molecules—cAMP as an activator for the protein kinase, and P_i (phosphate) as an inhibitor for the phosphatase (see Figure 1.7). The steady state of the system depends on the relative concentrations of each of these four components. The substrate molecule for the kinase and the phosphatase was an artificially constructed nano-peptide consisting of a chain of nine amino acids. By keeping the concentration of ATP sufficiently large for it to remain essentially constant, it was possible to study the effects of altering the concentration of the regulatory or effector molecules cAMP and P_i. The cascade showed signal amplification and *positive cooperativity.* The latter feature of a signaling pathway means that the sensitivity of levels of phosphorylation of substrate to changes in levels of concentration of effector molecules was much greater than the sensitivity of the kinase to changing levels of cAMP.

Even for this simple cascade system, there are a number of distinct reaction steps that need to be modeled. The classical methods of enzyme kinetics used to study these reactions [26] involve the construction of a system of differential equations (B.1)–(B.6) which measure the rate of change of the concentration of each of the components of the system. The equilibrium state of the system is studied, whereby the rate of the phosphorylation reaction (i.e., rate of formation of S-P, the phosphorylated form of the substrate molecule) in (B.1) exactly balances the reverse reaction (i.e., the rate of regeneration of S) in (B.2). The square brackets around the terms in the equations denotes concentration.

The phosphorylation of the substrate, S, by the activated catalytic subunit of the kinase, PKA:

$$\mathrm{S} + \mathrm{C} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{S \cdot C} \overset{k_f}{\rightarrow} \mathrm{S\text{-}P} + \mathrm{C}. \tag{B.1}$$

The dephosphorylation of the substrate, S, by the phosphatase, PP:

$$\mathrm{S\text{-}P} + \mathrm{PP} \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} \{\mathrm{S\text{-}P}\} \cdot \mathrm{PP} \overset{k_r}{\rightarrow} \mathrm{S} + \mathrm{PP}, \tag{B.2}$$

with associated rate equations:

$$\frac{d[\mathrm{S\text{-}P}]}{dt} = k_f[\mathrm{S \cdot C}] - k_3[\mathrm{S\text{-}P}][\mathrm{PP}] + k_{-3}[\{\mathrm{S\text{-}P}\} \cdot \mathrm{PP}], \tag{B.3}$$

$$\frac{d[\mathrm{S}]}{dt} = k_r[\{\mathrm{S\text{-}P}\} \cdot \mathrm{PP}] + k_{-1}[\mathrm{S \cdot C}] - k_1[\mathrm{S}][\mathrm{C}], \tag{B.4}$$

$$\frac{d[\mathrm{S \cdot C}]}{dt} = k_1[\mathrm{S}][\mathrm{C}] - (k_{-1} + k_f)[\mathrm{S \cdot C}], \tag{B.5}$$

$$\frac{d[\{\mathrm{S\text{-}P}\} \cdot \mathrm{PP}]}{dt} = k_3[\mathrm{S\text{-}P}][\mathrm{PP}] - (k_{-3} + k_r)[\{\mathrm{S\text{-}P}\} \cdot \mathrm{PP}]. \tag{B.6}$$

By mass conservation arguments, the total concentration of the kinase, phosphatase and substrate can be found. The steady state is defined as the equilibrium state where the rate of the phosphorylation reaction (i.e. rate of formation of S-P) in (B.1) exactly balances the reverse reaction (i.e. the rate of regeneration of S) in (B.2).

The system is further complicated by the manner in which cAMP-dependent protein kinase phosphorylates its substrates. In fact the inactive enzyme is composed of four subunits—two regulatory units and two catalytic units, and is denoted ($\mathrm{R_2C_2}$). Binding of cAMP to its regulatory domains causes dissociation of the subunits and activation of the catalytic domains upon separation.

$$\mathrm{R_2C_2} + 4\ \mathrm{cAMP} \overset{K_a}{\rightarrow} \mathrm{R_2.cAMP} + 2\mathrm{C} \tag{B.7}$$

Symbol List:

Symbol	*Description*
S	substrate
S-P	phosphorylated form of the substrate, S
PKA	cAMP dependent protein kinase
PP	protein phosphatase
$S \cdot C$	substrate and cAMP-dependent protein kinase complex
$\{S\text{-}P\} \cdot PP$	phosphorylated substrate and protein phosphatase complex
k_1	rate constant for the formation of the substrate-PKA complex
k_{-1}	rate constant for the reaction involving the dissociation of the substrate-PKA complex into substrate and PKA
k_f	rate constant for the reaction involving phosphorylation of the substrate by PKA
k_3	rate constant for the formation of the phosphorylated substrate-PP complex
k_{-3}	rate constant for the reaction involving the dissociation of the phosphorylated substrate-PP complex
k_r	rate constant for the reaction involving dephosphorylation of the phosphorylated substrate by PP
cAMP	cyclic-AMP
R_2C_2	inactive PKA is composed of 4 subunits (R: regulatory subunit, C: catalytic subunit)
K_a	rate constant for the reaction whereby cAMP causes dissociation of 2 of the catalytic subunits of PKA, forming a complex with the remaining 2 regulatory subunits

APPENDIX C

SMALL NETWORK LOCAL STABILITY ANALYSIS

A summary of local stability analysis of fixed points for some simple networks is shown in Table C.1. All examples shown here are networks where the nodes have inputs such that $b_j^{(t)} \geq 0\ \forall t$, or $b_j^{(t)} \leq 0\ \forall t$, where $1 \leq j \leq N$.

Table C.1: Summary of fixed points for simple networks using the rule, f, and where inputs to each node always retain the same sign. Local stability of the fixed points is examined via linear stability analysis.

Network Description	*Critical Points*	*E'values*	*Local Stability*
Self-Excitation	0	2	Unstable
	N	e^{-1}	Stable
Self-Inhibition	0	1	Critically Stable
(1) $\overset{+}{\underset{-}{\rightleftharpoons}}$ (2)	$(0, y)$ $0 \leq y \leq N$	$e^{-y/N}$, 1	Critically Stable
(1) $\overset{+}{\underset{+}{\rightleftharpoons}}$ (2)	$(0,0)$	0, 2	Unstable
	(N, N)	e^{-1}, e^{-1}	Stable
(1) $\overset{-}{\underset{-}{\rightleftharpoons}}$ (2)	$(x, 0)$ $0 \leq x \leq N$	1, $e^{-x/N}$	Critically Stable
	$(0, y)$ $0 \leq y \leq N$	$e^{-y/N}$, 1	Critically Stable
(1) $\overset{+}{\underset{-}{\rightleftharpoons}}$ (2) $\overset{+}{\leftarrow}$ (3)	$(0, N, \mu)$ $0 < \mu \leq N$	e^{-1}, $e^{-\mu/N}$	Stable
(3) $\overset{-}{\rightarrow}$ (1) $\overset{+}{\underset{-}{\rightleftharpoons}}$ (2)	$(0, y, \mu)$ $0 \leq y \leq N, 0 < \mu \leq N$	$e^{-(y+\mu)/N}$, 1	Critically Stable
(1) $\overset{+}{\underset{+}{\rightleftharpoons}}$ (2) $\overset{+}{\leftarrow}$ (3)	(N, N, μ) $0 < \mu \leq N$	e^{-1}, $e^{-(\mu+N)/N}$	Stable

APPENDIX D

GRAPH THEORY DEFINITIONS

Definition D.1 (Graph) *A graph, $G(V, E)$, consists of a set V of elements called vertices and a set E of unordered pairs of members of V called edges. The order of a graph is the number of vertices in the graph. A directed graph is a graph where a direction is imposed upon the edges of the graph. The edges are then interpreted as ordered pairs of vertices.*

Definition D.2 (Subgraph) *A subgraph of $G(V, E)$ is a graph $S(V', E')$ such that $V' \in V$ and $E' \in E$ and the endpoints of any edge in E' are also in V'. Suppose an edge in the graph is denoted as the unordered pair (u, v), then the endpoints of this edge are defined to be the vertices u and v.*

Definition D.3 (Induced Subgraph) *S is an induced subgraph of G if whenever $u \in V'$ and $v \in V'$ and $(u, v) \in E$, then $(u, v) \in E'$.*

Definition D.4 (Connected Graph) *A graph G is connected if and only if every pair of vertices in G are connected. Two vertices u and v are said to be connected if there exists a path from u to v. A graph that is not connected is said to be disconnected.*

Definition D.5 (Path) *A path from a vertex $u \in V$ to a vertex $v \in V$ in a graph $G(V, E)$ is an alternating sequence of vertices and edges, $v_1, e_1, v_2, e_2, \ldots, e_{k-1}, v_k$ where all the vertices and edges in the sequence are distinct and successive vertices v_i, v_{i+1} are the endpoints of the intermediate edge e_i. The length of the path is the number of edges in the sequence.*

Definition D.6 (Degree of a Vertex) *The degree of a vertex V is denoted $deg(V)$ and is the number of edges incident with V.*

Definition D.7 (Strong Component) *A strong component of a directed graph $G(V, E)$ is a connected, induced subgraph of G of maximal order, (i.e., a strong component is a maximally internally connected subgraph of a directed graph G). A directed graph $G(V, E)$ is said to be strongly connected if there exists a path between every pair of vertices in G.*

Definition D.8 (Cycle) *If the first and last vertices of a path coincide, the resulting closed path is called a cycle.*

Definition D.9 (Non-Separable Graph) *A non-separable graph is a non-trivial graph with no articulation points.*

Definition D.10 (Articulation Point) *An articulation point or cut-vertex of a graph $G(V, E)$ is a vertex whose removal increases the number of components of the graph G.*

Definition D.11 (Block) *A block is a maximal, non-separable subgraph of a graph.*

APPENDIX E

DYNAMICAL SYSTEMS TERMINOLOGY

In this appendix, some basic definitions of dynamical systems are given. These mainly relate to stability and large-time properties of a discrete dynamical system.

Definition E.1 (Positive Semi-Orbits) *Consider the mapping* $f : \mathbb{R}^n \to \mathbb{R}^n$ *and the point* $x_0 \in \mathbb{R}^n$. *The positive semi-orbit of* x_0 *is the set of points* $\{f^k(x_0)\}_{k=0}^{\infty}$.

Definition E.2 (Fixed Points) *The mapping* $f : \mathbb{R}^n \to \mathbb{R}^n$ *is said to have a fixed point,* x_0, *if* $f(x_0) = x_0$.

Definition E.3 (Asymptotically stable fixed points) *Suppose* x_0 *is a fixed point of* f. *If there exists a neighborhood,* U, *of* x_0 *such that* $\lim_{n\to\infty} f^n(x) = x_0$ $\forall x \in U$, *then the point* x_0 *is said to be an asymptotically stable fixed point of* f.

Definition E.4 (Domain of attraction of an asymp. stable fixed point.) *Suppose* x_0 *is an asymptotically stable fixed point of* f. *The set of points which converge to* x_0 *via iteration of the mapping* f *is said to be the domain (or basin) of attraction of the point* x_0.

Definition E.5 (Periodic Points) *The mapping* $f : \mathbb{R}^n \to \mathbb{R}^n$ *has a k-periodic point,* x_0, *associated with it if* $f^k(x_0) = x_0$ *where* $k \in \mathbb{N}$ *and* $f^n(x_0) \neq x_0$ *for* $n = 1, 2, \ldots, k-1$.

Definition E.6 (Asymptotically stable periodic points) *Suppose* x_0 *is a k-periodic point of* f. *If there exists a neighborhood,* U, *of* x_0 *such that* $\lim_{n\to\infty} F^n(x) = x_0$ $\forall x \in U$, *and where* $F = f^k$, *then the point* x_0 *is said to be an asymptotically stable k-periodic point of* f *(or an asymptotically stable fixed point of* F*).*

Definition E.7 (Domain of attraction of an asymp. stable periodic pt.) *Suppose* x_0 *is an asymptotically stable k-periodic point of* f. *The set of points which converge to* x_0 *via iteration of the mapping* f^k *is said to be the domain (or basin) of attraction of the point* x_0.

Definition E.8 (Aperiodic points) *A point x_0 is called an* aperiodic *point of the mapping f if the orbit of x_0 is bounded and there is* no *value of $k \in \mathbb{N}$ such that $\lim_{n\to\infty} f^{nk}(x_0)$ exists.*

Definition E.9 (Attracting Sets) *A critical or fixed point, x_0, of the map $x_{n+1} = f(x_n)$ in $\mathbb{R}^n$ is called a positive attractor if there exists a neighborhood U of $x = x_0$ such that if $x_t \in U$ for some $t > 0$ and $t \in \mathbb{N}$, then $\lim_{n\to\infty} x_{t+n} = x_0$.*

For non-degenerate fixed points, a critical point, which after linearization is a positive attractor, is also asymptotically stable. However, there are examples of systems with critical points which are attracting but are not asymptotically stable (see pp. 62 of [79]).

Definition E.10 (Invariant Set) *A set $M \subseteq \mathbb{R}^n$ is (positively) invariant under $f : \mathbb{R}^n \to \mathbb{R}^n$ if $x \in M \Rightarrow f(x) \in M$, $\forall x \in M$.*

Definition E.11 (Attractor) *An attractor is a closed invariant set, A, such that there is an open set $U \supset A$ such that if $x \in U \Rightarrow f^n(x) \to A$ as $n \to \infty$,*

An attractor is also sometimes referred to as an attracting set that contains a dense orbit [144].

Definition E.12 (Bifurcation point in 1-d discrete dynamical systems) *Suppose we define a one-dimensional dynamical system*

$$x_{n+1} = f(\mu, x_n) \tag{E.1}$$

where μ is a regarded as a parameter. Then, if the stability number or type and stability of fixed points of the mapping changes as we pass through a specific parameter value $\mu = \mu_0$, then, μ_0 is said to be a bifurcation point.

BIBLIOGRAPHY

[1] T. Bliss and T. Lomo, "Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path," *Journal of Physiology*, vol. 232, pp. 331–356, 1973.

[2] E. Kandel, J. Schwartz, and T. Jessell, *Principles of Neural Science.* New York: Elsevier Science Publishing Co., Inc., 1991.

[3] T. Bliss and G. Collingridge, "A synaptic model of memory: long-term potentiation in the hippocampus," *Nature*, vol. 361, pp. 31–39, 1993.

[4] N. Rose, *Mathematical Maxims and Minims.* Rayleigh, North Carolina: Rome Press Inc., 1988.

[5] W. Harvey, *On the Motion of the Heart and Blood in Animals.* New York: "Great Minds Series," Prometheus Books, a translation by Richard Willis. Originally presented as a lecture series in 1616. ed., 1993.

[6] A. I. Tauber and S. Sarkar, "The human genome project: has blind reductionism gone too far?," *Perspectives in Biology and Medicine*, vol. 35, no. 2, pp. 220–235, 1992.

[7] P. K. Kinnunen, "On the principles of functional ordering in biological membranes," *Chemistry and Physics of Lipids*, vol. 57, pp. 375–399, 1991.

[8] L. Motz and J. H. Weaver, *The Story of Mathematics.* New York: Avon Books, 1993.

[9] R. Hooke, *Micrographia: or some physiological descriptions of minute bodies made by magnifying glasses with observations and inquiries thereupon.* London, U.K.: John Martyn and James Allestry, printers to the Royal Society of London, 1665.

[10] T. Schwann, *Microscopical researches into the accordance in the structure and growth of animals and plants.* London: The Sydenham Society, translated by Henry Smith from the original: "Mikroskopische Untersuchungen über die Übereinstimmung in der Strucktur und dem Wachstume der Tiere and Pflanzen" (1839) ed., 1847.

[11] R. Virchow, *Cellular pathology as based upon physiological and pathological histology.* London: John Churchill, edited by Frank Chance, with notes and and numerous emendations. ed., 1978.

[12] H. Brown, F. Sanger, and R. Kitai, "The structure of pig and sheep insulin," *Biochemical Journal*, vol. 60, pp. 556–565, 1955.

[13] B. Alberts *et al.*, *Molecular Biology of the Cell.* New York: Galand, third ed., 1994.

[14] M. Smith, "In vitro mutagenesis," *Annual Review of Genetics*, vol. 19, pp. 423–462, 1985.

[15] M. Smith, "Site-Directed Mutagenesis," *Philosophical Transactions of the Royal Society of London Series A*, vol. 317, pp. 295–304, 1986.

[16] J. Zhang, F. Zhang, D. Ebert, M. Cobb, and E. Goldsmith, "The structure of the MAP kinase ERK2 is controlled by a flexible surface loop," *Structure*, vol. 3, pp. 299–307, 1995.

[17] Y. J. Zhou *et al.*, "Distinct tyrosine phosphorylation sites in JAK3 kinase domain positively and negatively regulate its enzymatic activity," *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, vol. 94, no. 25, pp. 13850–13855, 1997.

[18] G. Cori and A. A. Green, "Crystalline muscle phosphorylase. II. Prosthetic group," *Journal of Biological Chemistry*, vol. 151, pp. 31–38, 1943.

[19] G. Cori and C. Cori, "The enzymatic conversion of phosphorylase *a* to *b*," *Journal of Biological Chemistry*, vol. 158, pp. 321–332, 1945.

[20] E. Krebs, "Protein phosphorylation and cellular regulation I," *Bioscience Reports*, vol. 13, no. 3, pp. 127–142, 1993.

[21] P. Cohen, "Protein phosphorylation and hormone action," *Proceedings of the Royal Society of London Series B - Biological Sciences*, vol. 234, pp. 115–144, July 1988. Review lecture.

[22] N. G. Ahn, R. Seger, and E. Krebs, "The mitogen-activated protein kinase activator," *Current Biology*, vol. 4, pp. 992–999, 1992.

[23] E. Stadtman and P. Chock, "Superiority of interconvertible enzyme cascades in metabolic regulation: analysis of monocyclic systems," *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, vol. 74, pp. 2761–2765, 1977.

[24] E. Shacter, P. Chock, and E. Stadtman, "Regulation through Phosphorylation / Dephosphorylation cascade systems," *Journal of Biological Chemistry*, vol. 259, pp. 12252–12259, 1984.

[25] P. Chock and E. Stadtman, "Superiority of interconvertible enzyme cascades in metabolic regulation: analysis of multicyclic systems," *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, vol. 74, pp. 2766–2770, 1977.

[26] A. Fersht, *Enzyme Structure and Mechanism.* WH Freeman and Co., 2nd ed., 1985.

[27] J. E. Ferrell Jr, "Tripping the switch fantastic: how a protein kinase cascade can convert graded inputs into switch-like outputs," *TIBS*, vol. 21, pp. 460–466, 1996.

[28] J. E. Ferrell Jr. and E. M. Machleder, "The biochemical basis of an all-or-none fate switch in Xenopus oocytes," *Science*, vol. 280, pp. 895–898, 1998.

[29] G. E. Lienhard, "Insulin – life without the IRS," *Nature*, vol. 372, pp. 128–129, 1994.

[30] D. Lloyd and L. Rossi, Ernest, "Biological rhythmns as organisation and information," *Biological Reviews of the Cambridge Philosophical Society*, vol. 68, pp. 563–577, 1993.

[31] T. Hunter, "Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling," *Cell*, vol. 80, no. 2, pp. 225–236, 1995.

[32] A. Murray and T. Hunt, *The Cell Cycle: An Introduction.* New York: W.H. Freeman, 1993.

[33] M. Berridge and G. Dupont, "Spatial and temporal signaling by calcium," *Current Opinion in Cell Biology*, vol. 6, no. 2, pp. 267–274, 1994.

[34] J. Meldolesi, "Oscillation, activation, expression," *Nature*, vol. 392, pp. 863–866, 1998.

[35] J. Carroll and K. Swann, "Spontaneous cytosolic calcium oscillations driven by inositol trisphosphate occur during in vitro maturation of mouse oocytes," *Journal of Biological Chemistry*, vol. 267, no. 16, pp. 11196–11201, 1992.

[36] M. Berridge, "Inositol trisphosphate and calcium signalling," *Nature*, vol. 361, pp. 315–325, 1993.

[37] M. Berridge, M. Bootman, and P. Lipp, "Calcium – a life and death signal," *Nature*, vol. 395, pp. 645–648, 1998.

[38] A. Lloyd *et al.*, "Cooperating oncogenes converge to regulate cyclin/cdk complexes," *Genes and Development*, vol. 11, pp. 663–677, 1997.

[39] P. Nurse, "Reductionism and explanation in cell biology," in *The limits of reductionism in biology*, vol. Novatis Foundation Symposium 213, pp. 93–105, Chichester, U.K.: Wiley, 1998.

[40] Y. Termonia and J. Ross, "Oscillations and control features in glycolysis in numerical analysis of a comprehensive model," *Proceedings of the National Academy of Sciences of the United States of America–Biological Sciences*, vol. 78, pp. 2952–2956, 1981.

[41] D. Bray, "Intracellular signalling as a parallel distributed process," *Journal of Theoretical Biology*, vol. 143, no. 2, pp. 215–231, 1990.

[42] D. Bray *et al.*, "Computer simulation of the phosphorylation cascade controlling bacterial chemotaxis," *Molecular Biology of the Cell*, vol. 4, pp. 469–482, 1993.

[43] B. Alberts and R. Miake-Lye, "Unscrambling the puzzle of biological machines: the importance of the details," *Cell*, vol. 68, pp. 415–420, 1992.

[44] P. Roach, "Multisite and hierarchical protein phosphorylation," *Journal of Biological Chemistry*, vol. 266, no. 22, pp. 14139–14142, 1991.

[45] G. Barrit, *Communication within animal cells.* Oxford, UK: Oxford University Press, 1992.

[46] P. Chock and E. Stadtman, "Cyclic cascades in cellular regulation," in *Chemistry of the Living Cell* (E. Bittar, ed.), vol. 3B, pp. 391–411, Conneticut: JAI Press Inc, 1992.

[47] P. Cohen, "Dissection of the protein phosphorylation cascades involved in insulin and growth factor action," *Chemical Society Transactions*, vol. 21, pp. 555–567, 1993. Twenty–fourth Ciba Medal Lecture.

[48] D. Bray, "Protein molecules as computational elements in living cells," *Nature*, vol. 376, pp. 307–312, 1995.

[49] T. Hunter and G. D. Plowman, "The protein kinases of budding yeast: six score and more," *TIBS*, vol. 22, pp. 18–22, 1997.

[50] T. Hunter, "1001 protein kinases redux: towards 2000," *Seminars in Cell Biology*, vol. 5, pp. 367–376, 1994.

[51] T. Hunter, "Tyrosine phosphorylation: past, present and future," *Biochemical Society Transactions*, vol. 24, no. 2, pp. 307–327, 1996.

[52] H. Charbonneau and N. Tonks, "1002 protein phosphatases?," *Annual Review of Cell Biology*, vol. 8, pp. 463–493, 1992.

[53] T. Hunter, "Signaling – 2000 and beyond," *Cell*, pp. 113–127, 2000.

[54] T. Hunter, "A thousand and one protein kinases," *Cell*, vol. 50, pp. 823–829, 1987.

[55] M. Hubbard and P. Cohen, "On target with a new mechanism for the regulation of protein phosphorylation," *TIBS*, vol. 18, no. 5, pp. 172–177, 1993.

[56] E. Nishida and Y. Gotoh, "The MAP kinase cascade is essential for diverse signal transduction pathways," *TIBS*, vol. 18, pp. 128–131, 1993.

[57] W. Li and D. Graur, *Fundamentals of molecular evolution.* Sunderland, Mass.: Sinauer Associates, 1991.

[58] R. Doolittle, "The multiplicity of domains in proteins," *Annual Review of Biochemistry*, vol. 64, pp. 287–314, 1995.

[59] D. Koshland, "Switches, thresholds and ultrasensitivity," *TIBS*, vol. 12, pp. 225–229, 1987.

[60] P. Kinnunen, "Personal communication." May 2, 2000.

[61] P. K. Kinnunen, "On the mechanisms of the lamellar to hexagonal h_{II} phase transition and the biological significance of the h_{II} propensity," in *Handbook of non-medical applications of liposomes: Theory and Basic Sciences* (D. D. Lasic and Y. Barenholz, eds.), vol. 1, ch. 6, pp. 153–171, New York: CRC Press, 1996.

[62] J. Woodgett, *Protein Kinases.* New York: IRL Press at Oxford University Press, 1995.

[63] B. Gjertsen and S. Doskeland, "Protein phosphorylation in apoptosis," *Biochimica et Biophysica Acta*, vol. 1269, pp. 187–199, 1995.

[64] E. Krebs, "Historical perspectives on protein phosphorylation and a classification system for protein kinases," *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences*, vol. 302, pp. 3–11, 1983.

[65] M. Mumby and G. Walker, "Protein serine/threonine phosphatases: structure, regulation, and functions in cell growth," *Physiological Reviews*, vol. 73, no. 4, pp. 673–699, 1993.

[66] H. Sun and N. Tonks, "The coordinated action of protein tyrosine phosphatases and kinases in cell signaling," *TIBS*, vol. 19, pp. 480–485, 1994.

[67] T. Woodford *et al.*, "The biological functions of protein phosphorylation-dephosphorylation," in *Fundamentals of Medical Cell Biology*, vol. 3B, pp. 453–507, JAI Press Inc, 1992.

[68] T. Hunter, "Protein–tyrosine phosphatases: the other side of the coin," *Cell*, vol. 58, pp. 1013–1016, 1989.

[69] D. Alexander, "The role of phosphatases in signal transduction," *The New Biologist*, vol. 2, no. 12, pp. 1049–1062, 1990.

[70] A. Nairn *et al.*, "Protein kinases in the brain," *Annual Review of Biochemistry*, vol. 54, pp. 931–976, 1985.

[71] A. Gelfand and C. Walker, *Ensemble modeling: inference from small–scale properties to large–scale systems.* New York: Marcel Dekker Inc., 1984.

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

[73] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences of the United States of America–Biological Sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[74] A. Barabàsi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[75] N. Barkal and S. Leibler, "Robustness in simple biochemical networks," *Nature*, vol. 387, pp. 913–917, 1997.

[76] S. B. Laughlin, R. R. de Ruyter van Steveninck, and J. C. Anderson, "The metabolic cost of neural information," *Nature Neuroscience*, vol. 1, no. 1, pp. 36–41, 1998.

[77] E. Chiva and P. Tarroux, "Evolution of biological regulation networks under complex environmental constraints," *Biological Cybernetics*, vol. 73, no. 4, pp. 323–333, 1995.

[78] J. Murray, *Mathematical Biology.* New York: Springer-Verlag, 1991.

[79] F. Verhulst, *Nonlinear Differential Equations and Dynamical Systems.* New York: Springer Verlag, 1991.

[80] J. Von Neumann, *Theory of Self-Reproducing Automata.* Urbana, Illinois: University of Illinois Press, completed and edited by Arthur W. Burks ed., 1966.

[81] M. Gardner, "The fantastic combinations of John Conway's new solitaire game 'life'," *Scientific American*, vol. 223, pp. 120–124, 1970.

[82] M. Gardner, "On cellular automata, self-reproduction, the Garden of Eden and the game 'life'," *Scientific American*, vol. 224, pp. 112–117, 1971.

[83] A. M. Turing, "On computable numbers, with an application to the Entscheidungsproblem," *Proceedings, London Mathematical Society*, vol. 42, no. 2, pp. 230–265, 1936.

[84] G. B. Ermentrout and L. Edelstein-Keshet, "Cellular automata approaches to biological modelling," *Journal of Theoretical Biology*, vol. 160, pp. 97–133, 1993.

[85] S. Wolfram, "Cellular automata as models of complexity," *Nature*, vol. 311, pp. 419–424, 1984.

[86] C. G. Langton, "Life at the Edge of Chaos," in *Artifical Life II, SFI Studies in the Sciences of Complexity* (C. Langton, C. Taylor, J. Farmer, and R. S., eds.), vol. X, pp. 41–91, Addison-Wesley, 1991.

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

[88] B. Hess and A. Mikhailov, "Microscopic self-organization in living cells: a study of time matching," *Journal of Theoretical Biology*, vol. 176, pp. 181–184, 1995.

[89] A. Mikhailov and B. Hess, "Fluctuations in living cells and intracellular traffic," *Journal of Theoretical Biology*, vol. 176, pp. 185–192, 1995.

[90] G. Brown and B. Kholodenko, "Spatial gradients of cellular phospho-proteins," *FEBS Letters*, vol. 457, pp. 452–454, 1999.

[91] L. Stryer, *Biochemistry.* New York: Freeman, 3rd ed., 1980.

[92] C. Langton, "Computation at the edge of chaos: phase transitions and emergent computation," *Physica D*, vol. 42, pp. 12–37, 1990.

[93] D. Goldberg, *Genetic algorithms in search, optimization and machine learning.* USA: Addison–Wesley, 1989.

[94] J. Holland, *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence.* Cambridge, Mass.: MIT Press, 2nd ed., 1992. First printed in 1975 by University of Michigan Press, Ann Arbour, MI.

[95] S. Forrest, "Genetic algorithms: principles of natural selection applied to computation," *Science*, vol. 261, pp. 872–878, 1993.

[96] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* USA: John Wiley and Sons, Inc., 1973.

[97] J. Dugundji, *Topology.* Boston: Allyn and Bacon, Inc., 1966.

[98] E. Reingold, J. Nievergelt, and N. Deo, *Combinatorial Algorithms: Theory and Practice.* NJ, USA: Prentice Hall, 1977.

[99] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[100] J. Farmer, "A rosetta stone for connectionism," *Physica D*, vol. 42, pp. 153–187, 1990.

[101] J. Stewart and F. Varela, "Exploring the meaning of connectivity in the immune network," *Immunological Reviews*, vol. 110, pp. 37–61, 1989.

[102] S. Wolfram, "Approaches to complexity engineering," *Physica D*, vol. 22, pp. 385–399, 1986.

[103] S. A. Kauffman, "Antichaos and adaptation," *Scientific American*, vol. 265, no. 2, pp. 78–84, 1991.

[104] C. Marsan and H. Mattheis, *Plasticity in the Nervous System — An Approach to Memory Research*, pp. 1–15. NY: Raven Press, 1982.

[105] A. Saltiel, "The paradoxical regulation of protein phosphorylation in insulin action," *The FASEB Journal*, vol. 8, no. 13, pp. 1034–1040, 1994.

[106] S. Cuthill, "Cellular epigenetics and the origin of cancer," *Bioessays*, vol. 16, pp. 393–394, June 1994.

[107] G. Albrecht-Beuhler, "In defense of 'nonmolecular' cell biology," *International Review of Cytology*, vol. 120, pp. 191–241, 1990.

[108] C. J. Morton-Firth and D. Bray, "Predicting temporal fluctuations in an intracellular signalling pathway," *Journal of Theoretical Biology*, vol. 192, pp. 117–128, 1998.

[109] J. Liu, J. Crawford, and R. Viola, "The consequences of interactive noise for understanding the dynamics of complex biochemical systems," *Dynamics and Stability of Systems*, 1996.

[110] J. White, J. Rubinstein, and A. Kay, "Channel noise in neurons," *TINS*, 2000.

[111] M. Kraus and B. Wolf, *Structured Biological Modelling: A New Approach to Biophysical Cell Biology.* CRC Press, 1995.

[112] D. Bray and S. Lay, "Computer simulated evolution of a network of cell–signaling molecules," *Biophysical Journal*, vol. 66, no. 4, pp. 972–977, 1994.

[113] W. Bialek and R. Fred, "Reliability and information transmission in spiking neurons," *TINS*, vol. 15, no. 11, pp. 428–434, 1992.

[114] S. Strogatz, "Exploring complex networks," *Nature*, 2001.

[115] J. E. Lisman, "A mechanism for memory storage insensitive to molecular turnover: a bistable autophosphorylating kinase," *Proceedings of the National Academy of Sciences of the United States of America–Biological Sciences*, vol. 82, pp. 3055–3057, 1985.

[116] J. Lisman, "The CaM kinase II hypothesis for the storage of synaptic memory," *TINS*, vol. 17, no. 10, pp. 404–412, 1994.

[117] A. Rotenberg, M. Mayford, R. D. Hawkins, E. R. Kandel, and R. U. Muller, "Mice expressing activated CaMKII lack low frequency LTP and do not form stable place cells in the CA1 region of the hippocampus," *Cell*, vol. 87, pp. 1351–1361, 1996.

[118] M. Mayford, M. E. Bach, Y.-Y. Huang, L. Wang, R. D. Hawkins, and E. R. Kandel, "Control of memory formation through regulated expression of a CaMKII transgene," *Science*, vol. 274, pp. 1678–1683, 1996.

[119] H. Markram and M. Tsodyks, "Redistribution of synaptic efficacy between neocortical pyramidal neurons," *Nature*, vol. 382, pp. 807–810, 1996.

[120] http://www.physiome.com, "In Silico Cell." Cited: April 20, 2001.

[121] S. Ball, V. Mah, and P. Miller, "SENEX: a computer-based representation of cellular signal transduction processes in the central nervous system," *Computer Applications in the Biosciences*, 1991.

[122] H. Sauro, "SCAMP: a general purpose simulator and metabolic control analysis program," *Computer Applications in the Biosciences*, 1993.

[123] A. Goldbeter, *Biochemical Oscillations and Cellular Rhythms: The molecular bases of periodic and chaotic behaviour.* England: Cambridge University Press, 1996.

[124] A. Goldbeter, D. G., and M. Berridge, "Minimal model for signal-induced Ca^{2+} oscillations and for their frequency encoding through protein phosphorylation," *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, vol. 87, pp. 1461–1465, 1990.

[125] Y. Fukushima and Y. Tonomura, "Oscillations of the amount of phosphorylated protein and the ATPase activity of microsomes," *Journal of Biological Chemistry*, vol. 72, pp. 623–634, 1972.

[126] B. Kholodenko, "Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades," *European Journal of Biochemistry*, vol. 267, pp. 1583–1588, 2000.

[127] A. Lopo and P. Calarco, "Stage–specific changes in protein phosphorylation during preimplantation development in the mouse," *Gamete Research*, vol. 5, pp. 283–290, 1982.

[128] B. Taylor and M. Johnson, "Universal themes of signal transduction in bacteria," in *Signal Transduction: Prokaryote and Simple Eukaryote Systems* (J. Kurjan and B. Taylor, eds.), ch. 1, pp. 3–15, Academic Press, Inc., 1993.

[129] U. Alon, M. Surette, N. Barkai, and S. Leibler, "Robustness in bacterial chemotaxis," *Nature*, vol. 397, pp. 168–171, 1999.

[130] A. Neiman, "Conservation and reiteration of a kinase cascade," *TIG*, vol. 9, no. 11, pp. 390–394, 1993.

[131] S. Noselli and N. Perrimon, "Are there close encounters between signaling pathways?," *Science*, 2000.

[132] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A. Barabàsi, "The large-scale organization of metabolic networks," *Nature*, 2000.

[133] S. A. Kauffman, "Autocatalytic sets of proteins," *Journal of Theoretical Biology*, vol. 119, pp. 1–24, 1986.

[134] R. Bagley, J. Farmer, S. Kauffman, N. Packard, A. Perelson, and I. Stadynk, "Modeling adaptive biological systems," *BioSystems*, vol. 23, pp. 113–138, 1989.

[135] "The human genome," *Science*, vol. 291, no. 5507, 2001.

[136] R. Ellmann and R. O'Clair, eds., *Modern Poems: An Introduction to Poetry.* New York: W. W. Norton & Company, 1976.

[137] T. Takai-Igarashi, "Guide to the Cell Signaling Networks Database," *Trends in Glycoscience and Glycotechnology*, vol. 11, no. 60, pp. 201–210, 1999.

[138] C. Smith, I. Shindyalov, S. Veretnik, M. Gribskov, S. S. Taylor, L. Ten Eyck, and P. Bourne, "The Protein Kinase Resource," *TIBS*, vol. 22, no. 11, pp. 444–446, 1997.

[139] A. Bairoch and R. Apweiler, "The SWISS–PROT protein sequence database and its supplement TrEMBL in 2000," *Nucleic Acids Research*, vol. 28, pp. 45–48, 2000.

[140] K. Hofmann, B. P., L. Falquet, and A. Bairoch, "The PROSITE database, its status in 1999," *Nucleic Acids Research*, vol. 27, pp. 215–219, 1999.

[141] A. Kreegipuu, N. Blom, and S. Brunak, "PhosphoBase, a database of phosphorylation sites: release 2.0," *Nucleic Acids Research*, vol. 27, pp. 237–239, 1999.

[142] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[143] W. Barker, J. Garavelli, Z. Hou, H. Huang, R. Ledley, P. McGarvey, H. Mewes, B. Orcutt, F. Pfeiffer, A. Tsugita, C. Vinayaka, C. Xiao, L. L. Yeh, and C. Wu, "Protein Information Resource: a community resource for expert annotation of protein data," *Nucleic Acids Research*, vol. 29, pp. 29–32, 2001.

[144] L. Perko, *Differential Equations and Dynamical Systems.* New York: Springer Verlag, 1996.