Signal Processing in Biological Cells: Proteins, Networks, and Models

by

Maya Rida Said

S.B. Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 1998
S.B. Biology
Massachusetts Institute of Technology, 1998
S.M. Toxicology
Massachusetts Institute of Technology, 2000
M.Eng. Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2001

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2005

Author .....
Department of Electrical Engineering and Computer Science
January 24, 2005

Certified by .....
Alan V. Oppenheim
Ford Professor of Engineering
Thesis Supervisor

Certified by .....
Douglas A. Lauffenbuger
Whitaker Professor of Bioengineering
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Signal Processing in Biological Cells: Proteins, Networks, and Models

by

Maya Rida Said

Submitted to the Department of Electrical Engineering and Computer Science
on January 24, 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Science

Abstract

This thesis introduces systematic engineering principles to model, at different levels of abstraction the information processing in biological cells in order to understand the algorithms implemented by the signaling pathways that perform the processing. An example of how to emulate one of these algorithms in other signal processing contexts is also presented.

At a high modeling level, the focus is on the network topology rather than the dynamical properties of the components of the signaling network. In this regime, we examine and analyze the distribution and properties of the network graph. Specifically, we present a global network investigation of the genotype/phenotype data-set recently developed for the yeast *Saccharomyces cerevisiae* from exposure to DNA damaging agents, enabling explicit study of how protein-protein interaction network characteristics may be associated with phenotypic functional effects. The properties of several functional yeast networks are also compared and a simple method to combine gene expression data with network information is proposed to better predict pathophysiological behavior.

At a low level of modeling, the thesis introduces a new framework for modeling cellular signal processing based on interacting Markov chains. This framework provides a unified way to simultaneously capture the stochasticity of signaling networks in individual cells while computing a deterministic solution which provides average behavior. The use of this framework is demonstrated on two classical signaling networks: the mitogen activated protein kinase cascade and the bacterial chemotaxis pathway.

The prospects of using cell biology as a metaphor for signal processing are also considered in a preliminary way by presenting a surface mapping algorithm based on bacterial chemotaxis.

Thesis Supervisor: Alan V. Oppenheim
Title: Ford Professor of Engineering

Thesis Supervisor: Douglas A. Lauffenbuger
Title: Whitaker Professor of Bioengineering

Acknowledgments

I am extremely fortunate and blessed to have met, early on in my career, Prof. Alan Oppenheim. Who would have ever imagined that stepping into an office hour on a fall afternoon would change my life forever? Al was the main reason I decided to pursue a PhD at MIT. He has had an incredible influence on my development, nurturing me both professionally and personally. He has taught me by example how to always aspire for excellence in teaching and research, how to never give up, and that luck is the residue of hard work. I am and will be forever indebted to my advisor, mentor, and above all, close friend. Al, we have come a long way since the first 6.011 office hour, the squash challenge, Irish music, the water-start, randomized sampling, Itzhak Perlman, and BSP. Thank you for being such a great teacher and coach. Thank you for believing in me, thank you for watching out for me, and most importantly thank you for making me discover a whole new world through your passion, creativity, and perseverance. I am looking forward to closing this chapter of my life and starting a new one with us interacting as colleagues and partners. It's been a roller coaster ride...fasten your seat belt, it's about to get more intense!

I also had the incredible good fortune early on in my graduate career to meet and come under the guidance and mentorship of Prof. Douglas Lauffenburger. Doug listened to my thoughts and believed in my ideas before almost anyone would hear them. His great enthusiasm and support have been spectacular. Doug, thank you for being such a great advisor and mentor. I look forward to a lifelong collaboration and friendship.

My thesis committee members have been an essential element of my thesis work and development; I would like to thank Prof. Denny Freeman, Prof. Peter Sorger, and Prof. Jacob White for great interactions and discussions and for providing feedback on the thesis document. Thanks to Denny for being such a great teacher to work with; I learned so much teaching 6.021: it was fun, challenging, and exciting. Thanks to Peter for great insights and for advising me not to pursue the Masters in biology which indirectly introduced me to Doug. Thanks to Jacob for being the first to awaken my interest in signals and systems and through this indirectly introducing me to Al. I am also indebted to Prof. Leona Samson and Prof. Thomas Begley for sharing the genomic phenotyping data and for being such great and generous people to work with. Chapters 3, 4 and 5 are the result of this very special collaboration. Prof. George Verghese provided valuable feedback and discussions on the interacting Markov chains model and the comments of Dr. Charlie Rohrs on Chapter 7 have been very helpful. I am grateful to Dr. Ily Setubal for stimulating discussions on the technical aspects of the thesis and Dr. Amy Baer for proposing early on creative research directions and for sharing unpublished data. My appreciation goes to Dr. John Stegeman for developing my interest in biology and Prof. Steve Tannenbaum for his encouragement and advice. I am particularly thankful to Dean Philip Khoury for his advice, guidance, support, and wisdom beginning with his role as my freshman advisor and continuing throughout my student career. I would also like to thank Alaa Kharbouch for providing technical help and spending endless hours on the bacterial chemotaxis simulations and Sabrina Spencers for help with some simulations.

I have been fortunate to be part of two great research environments while at MIT, the Digital Signal Processing Group and the Lauffenburger Lab both of which provided the opportunity to interact with very special people. I would like to thank Giovanni Aliberti, Tom Baran, Richard Barron, Alecia Batson, Albert Chan, Darla Chupp, Stark Draper, Uri Erez, Angela Glass, Vivek Goyal, Joonsung Lee, Yonina Eldar, Zahi Karam, Nick Lanneman, Li Lee, Emin Martinian, Andrew Russell, Matt Secor, Charles Sestok, Melanie

Shames, Eric Strattman, Charles Swannack, Wade Torres, and Dianne Wheeler for making my years at MIT so much richer. Special thanks go to Petros Bounfonos for always having the right answer to my endless questions, Sourav Dey for great technical discussions, and John Albeck, Kevin Janes, Suzanne Gaudet, Karen Sachs, and Birgit Schoeberl for sharing data and great brainstorming sessions on the BIM project and beyond.

My friends have always been a constant support to everything I do. I'd like to particularly thank Nazem Atassi and Samaan Rafeq for making sure I didn't work too hard and for great Starbucks breaks, Jamil Sobh for always checking on me and keeping me e-company, Joe Saleh for discussions about life, the universe, and for closely following the thesis progress, Ibrahim Abou-Faycal and Fadi Karameh for great advice and friendship, and Hur Koser, Diego Syrowicz, Jenn Healey, and Lukasz Weber for making my life at MIT, since the very beginning, so memorable. I would also like to thank Phyllis, Justine, and Jason Oppenheim for many wonderful Thanksgivings at Woods Hole.

My deepest thanks and gratitude go to my uncle Wafic Said, his wife, Rosemary, and children Khaled and Racha for providing me with the opportunity of a lifetime. Coming to MIT would have never materialized without their generosity. Thank you for launching me on this great path.

My sister Dana Said and my brother Mohamad Said have been and will always be my best friends. I thank them for putting up with me and for being so loving and caring.

I am especially grateful to Ziad Sabah for his deep love and emotional support, for sharing the low and high times, and for being so patient particularly these last few months. Thank you Ziad for coming into my life and for making it more meaningful and fulfilling. I am truly lucky to have you by my side.

This thesis, like all accomplishments in my life, could not have been without the unlimited support and unconditional love of my parents Rida and Layla. They have always taught me to continuely aim higher and never settle for good when I can do better. Dad, Mom, I will always strive to be as good a parent to my children as you are to me. You have been and continue to be incredible role models. You are a treasure that I will cherish forever. I love you so much. My gratitude is beyond any words. This thesis is dedicated to you.

To my parents Rida and Layla,
for being such incredible role models
and an endless source of inspiration.

Contents

1 Introduction **23**

1.1 General Description 24
1.2 Summary of Contributions 25
1.3 Chapter Outline 25

2 Graph Theoretic Techniques Applied to Biology **27**

2.1 Random Graph Theory 28
2.1.1 Degree, loops, and degree distribution 28
2.1.2 Adjacency matrix, spectra of random graphs 29
2.1.3 Clustering coefficient 29
2.1.4 Small worlds 30
2.1.5 Betweenness 30
2.1.6 The giant connected component 30
2.1.7 Methods for generating random graphs 31
2.2 Random Graphs in Biology 31
2.2.1 Networks of metabolic reactions 31
2.2.2 Genetic regulatory networks 32
2.2.3 Protein interaction networks 33
2.3 The Yeast Interactome 34

3 Global Network Properties of Damage Recovery Proteins **37**

3.1 Introduction 38
3.2 Phenotypic Annotation of the Yeast Interactome 39
3.2.1 Genomic phenotyping 39
3.2.2 Network construction 39
3.3 Network and Significance Measures 40
3.3.1 Network metrics 40
3.3.2 Randomizations and p-values 41
3.4 Degree Distribution 43
3.5 Shortest Path Distribution, Characteristic Path Length, and Global Centrality Distribution 43
3.6 The Centrality of Essential and Damage-Recovery Proteins in the Yeast Interactome 46

4 Local Properties of Synthesized Phenotypic Networks **49**
4.1 Synthesis of Phenotypic Networks . . . 51
4.2 Network Metrics . . . 52
4.2.1 Connected component analysis . . . 52
4.2.2 Clustering coefficient . . . 52
4.3 Network Connectivity . . . 52
4.4 Local Protein Environments in the Phenotypic Sub-Networks . . . 52
4.5 Identification of Toxicologically Important Protein Complexes and Signalling Pathways . . . 55
4.6 Towards a Systems Level Understanding of Vital Cellular Processes . . . 63

5 Topological Organization of Functional Yeast Networks **65**
5.1 Functional Yeast Networks . . . 66
5.1.1 Metabolism . . . 66
5.1.2 Sporulation . . . 66
5.1.3 Environmental stress response (ESR) . . . 67
5.1.4 Damage-recovery transcription response (DRR) . . . 67
5.2 Network Properties of Functional Yeast Networks . . . 67
5.2.1 Regulatory versus energy generation networks . . . 67
5.2.2 Expression profiling versus phenotypic analysis . . . 69
5.3 Predicting Phenotypic Outcome Using Expression Profiling Data . . . 70
5.3.1 Protein overlap between expression profiling and phenotypic data . . . 70
5.3.2 Using network information to better predict phenotype from expression profiling data . . . 72
5.3.3 Up-regulated versus down-regulated genes . . . 73
5.3.4 Towards a better prediction of phenotype using expression profiling data . . . 74
5.4 Robustness of the Network Results . . . 75
5.5 Damage-Recovery Party and Date Hubs . . . 75
5.6 Beyond Pairwise Interactions: Understanding Network Integrity Through Constrained Graph Partitioning . . . 81
5.6.1 Interaction-centric partitioning . . . 81
5.6.2 Protein-centric partitioning . . . 85

6 Low-Level Modeling of Biological Signaling Networks **87**
6.1 Current Low-Level Models . . . 88
6.2 Deterministic Formulation . . . 88
6.3 Stochastic Master Equation Formulation . . . 89
6.4 Stochastic Petri Nets . . . 91

7 Stochastic Models for Cell Signaling **93**
7.1 Markov Modulated Markov Chains (3MC) . . . 94
7.1.1 Additive model . . . 94
7.1.2 Fading model . . . 95
7.2 Notation . . . 95
7.3 Evolution of the State Probabilities . . . 95
7.4 Application to Cellular Signaling . . . 96
7.5 Derivation of the Modulating Function $f()$ . . . 97

7.5.1 Reactive collisions . . . 97
7.5.2 Associative interactions . . . 98
7.5.3 Conformational activation . . . 98
7.5.4 Simplified notation . . . 99
7.6 Revisiting the Single Event Assumption . . . 99
7.7 Model Approximation: State Dependency versus Probabilistic Dependency 100
7.7.1 The a3MC model . . . 100
7.7.2 Relationship between the 3MC and the a3MC . . . 101
7.7.3 Advantages of the a3MC Model . . . 103
7.8 Reconciling the Stochastic and Deterministic Formulations . . . 103
7.8.1 Unimolecular reactions . . . 103
7.8.2 Bimolecular reactions . . . 105
7.9 Example . . . 106
7.9.1 State probabilities . . . 107
7.9.2 Stochastic realizations . . . 107
7.9.3 Distributions . . . 107
7.10 Related Models . . . 108
7.10.1 The Gillespie algorithm . . . 108
7.10.2 Stochastic cellular automata . . . 112
7.11 A Unified Framework for Modeling the Dynamics of Signaling Pathways . . 112

8 The Mitogen Activated Protein Kinase Cascade 113
8.1 Background . . . 114
8.2 Properties of the MAP Kinase Cascade . . . 115
8.3 Model Formulation . . . 116
8.3.1 Topology . . . 116
8.3.2 Interactions . . . 117
8.3.3 Parameters values . . . 118
8.4 Dynamics of the State Probabilities of the a3MC Model Implementation . . 120
8.4.1 Steady-state . . . 120
8.4.2 Dynamic behavior . . . 121
8.5 Stochastic Simulations . . . 122
8.5.1 Time realizations . . . 122
8.5.2 Distributions . . . 129
8.6 Summary . . . 129

9 Bacterial Chemotaxis 131
9.1 The Biology of Bacterial Chemotaxis . . . 132
9.1.1 Signaling cascade . . . 132
9.2 Current Models of Bacterial Chemotaxis . . . 133
9.2.1 Biophysical models . . . 133
9.2.2 Biochemical models . . . 134
9.2.3 Stochastic models . . . 135
9.3 A Markov Modulated Markov Chains Model of Bacterial Chemotaxis . . . 136
9.3.1 Ligand binding and receptor states . . . 136
9.3.2 Chains and states . . . 138
9.3.3 Interactions and parameters . . . 138
9.3.4 The flagellar motor . . . 139

9.4 Simulations . . . 143
9.4.1 The dynamic behavior of the state probabilities . . . 143
9.4.2 Adaptation time . . . 145
9.5 Stochastic Implementation of Enzyme Kinetics . . . 147
9.5.1 Interacting Markov chains model of enzyme kinetics . . . 148
9.5.2 Applying the Michaelis-Menten approximation . . . 149
9.6 Updated Two-State Receptor Model . . . 151
9.6.1 Stochastic enzyme kinetics implementation for methylation by CheR 151
9.6.2 Using the Michaelis-Menten approximation . . . 153
9.7 Six-State Receptor . . . 157
9.7.1 Dynamics of the state probabilities of the a3MC model . . . 159
9.7.2 Stochastic simulations using the a3MC model . . . 160
9.8 Stochastic Analysis of Single Cell Behavior . . . 165
9.8.1 Correlations underlying the binary time series of switching events . . 165
9.8.2 Distributions of run and tumble lengths . . . 169
9.9 Summary . . . 171

10 Using Biology as a Metaphor: A Surface Mapping Algorithm Based on Bacterial Chemotaxis 173
10.1 Nature as a Metaphor . . . 174
10.1.1 Simulated annealing . . . 174
10.1.2 Genetic algorithms . . . 175
10.1.3 Ant colony optimization . . . 175
10.1.4 Optimization based on bacterial chemotaxis . . . 175
10.2 An Interactive Markov Chains Algorithm for Surface Mapping Based on Bacterial Chemotaxis . . . 176
10.2.1 BASM: Bacterial Algorithm for Surface Mapping . . . 177
10.2.2 Two-dimensional BASM . . . 177
10.2.3 One-dimensional BASM . . . 180
10.3 One-Dimensional Simulations . . . 181
10.3.1 Unimodal test function . . . 181
10.3.2 Multimodal test function . . . 182
10.4 Two-Dimensional Simulations . . . 184
10.4.1 Unimodal test function . . . 184
10.4.2 Multimodal test function . . . 185
10.5 Bacterial Algorithm for Surface Flattening (BASF) . . . 186
10.6 Summary . . . 191

11 Conclusions and Contributions 193
11.1 Graph Theoretic Modeling . . . 194
11.2 Multiresolution Modeling . . . 194
11.3 Biologically Inspired Surface Mapping and Flattening Algorithm . . . 194

A Randomizations 197

List of Figures

2-1 The yeast interactome. 34

3-1 Yeast protein categories and global network measures. A. Proteins that prevent damaging agent induced cell death are shown in green, essential proteins are shown in black and proteins associated with no-phenotype are shown in red. B. Degree of a node in a graph. As an example, the degree of (a) is 10 while for protein (b) it is 3. C. Shortest-path length. D. Characteristic path length. E. Global centrality. 40

3-2 Degree distributions of selected proteins. Black squares, green diamonds, and red triangles represent essential, and no-phenotype proteins respectively. The solid vertical lines represent the average degree. Blue (red italic) font in the inset indicates an average greater (smaller) than the corresponding randomized average. 42

3-3 Shortest-path length distribution, characteristic path length (a), and global centrality distribution (b) of selected proteins. Black squares, green diamonds, and red triangles represent essential, and no-phenotype proteins respectively. The solid vertical lines give the characteristic path length. n is the number of proteins in each category. Blue (red italic) font indicates an average greater (smaller) than the corresponding randomized average. . . . 44

3-4 Shortest-path length and global centrality of selected proteins. Non-essential (black diamonds), no-phenotype (red triangles), MMS (green squares), 4NQO (violet stars), UV (blue crosses), and t-BuOOH (gold circles) proteins. . . . 45

4-1 Newly-defined networks and clustering coefficient analysis. A. Derivations of new networks: networks comprised of only essential proteins and connecting edges are shown in black, proteins that prevent agent induced cell death and connecting edges are shown in green and no-phenotype proteins and connecting edges are shown in red. The clustering coefficient (C), can be determined for each protein to identify the degree of connectivity between a given protein's neighbors. B. Clustering coefficient analysis. C. p values. . . 50

4-2 Newly-defined phenotypic sub-networks. 51

4-3 Phenotypic protein networks with $C > 0$. Thick blue lines represent previously reported protein complexes. E. Selected complexes identified using clustering coefficient analysis. From top to bottom: MMS (RNA polymerase II holoenzyme/mediator complex and vacuolar H-ATP assembly complex); 4NQO (SWI/SNF complex and Nuclear pore complex); UV (Nucleotide excision repair pathway and C-terminal domain kinase I complex); and t-BuOOH (putative vacuolar sorting sub-network and SAGA transcriptional regulatory complex). 56

4-4 MMS protein network with $C > 0$. 57

4-5 4NQO protein network with $C > 0$. 58

4-6 UV protein network with $C > 0$. 59

4-7 t-BuOOH protein network with $C > 0$. 60

5-1 Number of proteins important for phenotypic growth only (phe only), proteins showing differential expression only (expr only), and proteins both important for phenotypic growth and showing differential expression (both). . 71

5-2 Average degree of proteins in different categories. 71

5-3 Specific interactions with edges cut in the FYI partitioning. ORF names are followed by common names in italic. DR, Ess, and Nophe correspond to damage-recovery, essential, and no-phenotype classifications respectively. . . 83

5-4 Node partitioning results using the FYI interactome. 85

7-1 Graphical representation of the interacting Markov chain model. Two representative Markov chains are illustrated for nodes 4 and 5: a two-state and a three-state chain respectively. 95

7-2 Simplified notation for the Markov chains model. For simplicity, the self loops on each state are not drawn. 99

7-3 Markov chains model representation. 106

7-4 Time evolution dynamics. 107

7-5 Examples of time realizations. 108

7-6 3MC model histograms. 109

7-7 a3MC model histograms. 110

7-8 Examples of a time realization using the Gillespie algorithm. 110

7-9 Gillespie algorithm histograms. 111

8-1 Schematic of the MAPK cascade. Forward arrows correspond to activation (phosphorylation for MAP2K and MAPK) while back-arrows correspond to deactivation steps (phosphatase activity in the case of MAP2K and MAPK). 114

8-2 Experimental stimulus/response data for MAPK and MAP2K(MEK) activation in Xenopus oocytes reproduced with permission of the authors from [67]. malE-Mos is the relevant MAP3K in this system. For more details regarding the experimental setup, the reader is referred to [67]. 115

8-3 Time evolution data. (a) ERK2 (the relevant MAPK) adaptation in Chinese hamster ovaries obtained from [10] with permission of the authors. Insulin is an activating factor which operates upstream of the cascade. The reader is referred to [10] for experimental details. (b) Time course of JNK (MAPK) activation and inactivation in sorbitol-treated Xenopus oocytes obtained from [12] with permission of the authors. Sorbitol is an activating factor upstream of the cascade. The reader is referred to [12] for experimental details. . . . 116
8-4 Interactive Markov chains model of the mitogen activated protein kinase cascade with a single phosphorylation mechanism. 116
8-5 Interactive Markov chains model of the mitogen activated protein kinase cascade with a dual phosphorylation mechanism. 117
8-6 Steady-state state probabilities predicted by the a3MC model using the parameters given in Table 8.4. (a) The input stimulus is expressed in multiples of EC_{50} (the input concentration that produces a 50% maximal response. (b) Semi-log scale. Here, the input stimulus is expressed in absolute values. 120
8-7 Steady-state state probabilities predicted by the a3MC model using a concentration of MAP2K of $6\mu M$. (a) The input stimulus is expressed in multiples of EC_{50} (the input concentration that produces a 50% maximal response. (b) Semi-log scale on an absolute scale. 121
8-8 Steady-state state probabilities predicted by the a3MC model using K_m values of 60nM (as opposed to 300nM) for the reactions that convert MAP2K among its various phosphorylation states. (a) The input stimulus is expressed in multiples of EC_{50}.(b) Semi-log scale on an absolute scale. 122
8-9 Steady-state state probabilities predicted by the a3MC model using a single phosphorylation. (a) The input stimulus is expressed in multiples of EC_{50}. (b) Semi-log scale. 123
8-10 Steady-state state probabilities predicted by the a3MC model using negative feedback. (a) The input stimulus is expressed in multiples of EC_{50}. (b) Semi-log scale. 123
8-11 Steady-state state probabilities predicted by the a3MC model using positive feedback on a semi-log scale. 124
8-12 State probabilities time evolution for a step of input stimulus of different strengths. 124
8-13 Time realization using the a3MC model for a step input stimulus of varying strengths. 125
8-14 Time realization using the 3MC model for a step input stimulus of varying strengths. 126
8-15 Distributions at steady state of relevant protein states using the a3MC model for step input stimuli of different strengths. 127
8-16 Distributions at steady state of the relevant protein states using the 3MC model for step input stimuli of different strengths. 128

9-1 The core Tar complex (reproduced from Bray [28]). CheY, CheB, and CheZ are not shown. 133
9-2 The signal transduction pathway in bacterial chemotaxis (reproduced from Bray *et al.* [26]). Solid arrows represent phosphorylation reactions. Dashed arrows indicate regulatory interactions, and open arrows are methylation reactions. 134

9-3 Conditional Markov chains representing the receptor states. 137
9-4 Core interactive Markov chains model of the bacterial chemotaxis pathway. 139
9-5 Markov chains model of the flagellar motor. 142
9-6 Non-zero input ligand concentration stimulus. 143
9-7 State probabilities using default parameters. (a) ap, yp, and bp, (b) receptor states (c) methylated receptors, (d) motor bias. 144
9-8 State probabilities using a high rate constant for demethylation by bp. (a) ap, yp and bp, (b) receptor states (c) methylated receptors, (d) motor bias. 145
9-9 State probabilities using the case where all forms of CheB can demethylate the receptor. (a) ap, yp, and bp, (b) receptor states (c) methylated receptors, (d) Motor bias. 146
9-10 Average adaptation time and standard deviation using the original model parameters. The error bars correspond to the standard deviation. 147
9-11 Interacting Markov chains illustration of enzyme kinetics. 148
9-12 Simplified Markov chains model using the Michaelis-Menten approximation. 151
9-13 Receptor model updated to include the stochastic implementation of enzyme kinetics for the methylation of the receptor by CheR. 152
9-14 State probabilities for the two-state model with enzymatic kinetics implementation. (a) ap, yp and bp, (b) receptor states (c) methylated receptors, (d) motor bias. 154
9-15 State probabilities for different concentrations of CheB of the two-state enzymatic kinetics model. 155
9-16 State probabilities for different concentrations of CheR of the two-state enzymatic kinetics model. 155
9-17 State probabilities for the Michaelis-Menten two-state model. (a) ap, yp and bp, (b) receptor states (c) methylated receptors, (d) motor bias. 156
9-18 State probabilities for different concentrations of CheB for the Michaelis-Menten two-state model. 157
9-19 State probabilities for different concentrations of CheR for the Michaelis-Menten two-state model. 158
9-20 3MC implementation of the Michaelis-Menten two-state model. 158
9-21 3MC implementation of the Michaelis-Menten two-state model with CheA always attached to the receptor. 159
9-22 Average number of phosphorylated CheY for varying concentrations of CheR (a) and CheB (b) The average is based on simulations of 100 CheY molecules using the Michaelis-Menten two-state model where CheB only demethylates phosphorylated receptors. 160
9-23 Distribution of the phosphorylated CheY state for different concentrations of CheR using the Michaelis-Menten two-state model where CheB only demethylates phosphorylated receptors. 160
9-24 Distribution of the phosphorylated CheY state for different concentrations of CheB using the Michaelis-Menten two-state model where CheB only demethylates phosphorylated receptors. 161
9-25 Adaptation error for different concentrations of CheR (a) and CheB (b) using the Michaelis-Menten two-state model where CheB only demethylates phosphorylated receptors. Error bars correspond to standard deviations around the average. 161
9-26 Interacting Markov chains model of the six-state receptor 162

9-27 State probabilities for the six state receptor. 162
9-28 State probabilities for varying concentrations of CheB using the six-state receptor model. 162
9-29 State probabilities for varying concentrations of CheR using the six-state receptor model. 164
9-30 Distribution of the phosphorylated CheY state for different concentrations of CheR using the six-state receptor model. 164
9-31 Normalized adaptation time (a) and percent adaptation error (b) for different concentrations of CheR using the six-state receptor. 165
9-32 Power spectral density of the six-state receptor *a*3MC model. 166
9-33 Power spectral density of the two-state model with stochastic enzyme kinetics 3MC implementation. 167
9-34 Power spectral density of the Michaelis-Menten two-state 3MC implementation. 167
9-35 Average power spectral density of the two-state Michaelis-Menten 3MC model. This plot was generated using a 2^{15} point FFT. 168
9-36 Distribution of run lengths computed using the Michaelis-Menten two-state model. 169
9-37 Distribution of tumble lengths computed using the Michaelis-Menten two-state model. 169
9-38 Distribution of run and tumble lengths computed using the Michaelis-Menten two-state model. 170
9-39 Distribution of run lengths for various CheB and CheR concentrations computed using the Michaelis-Menten two-state model. 170
9-40 Distribution of tumble lengths for various CheB and CheR concentrations computed using the Michaelis-Menten two-state model. 171

10-1 State probabilities time evolution for a step input stimulus at 100s from 0 to 10^{-3}M. (a) [Ap] and [Yp], (b) [Bp] and methylated receptors, (c) receptor states, (d) Motor bias. 178
10-2 State probabilities time evolution for a step input stimulus at 100s from 10^{-3}M to 10^{-6}M. (a) [Ap] and [Yp], (b) [Bp] and methylated receptors, (c) receptor states, (d) Motor bias. 179
10-3 One-dimensional test function. 181
10-4 Simulations using the default parameters and a one motor configuration. (a) individual bacterium variables. (b) Average density plot using 10 simulations. 182
10-5 Simulations using the default parameters and a nine motor configuration. (a) individual bacterium variables. (b) Average density plot using 40 simulations. 182
10-6 Simulations using the rotation in only one direction. (a) individual bacterium variables. (b) Average density plot using 10 simulations. 183
10-7 Simulations using the rotation in only one direction, 9 motor chains, and an increase in the motor backward probabilities by a factor of two. (a) individual bacterium variables. (b) Average density plot using 40 simulations. 183
10-8 Simulations using the rotation in only one direction, and an increase in the motor backward probabilities by a factor of two. (a) individual bacterium variables. (b) Average density plot using 10 simulations. 184
10-9 One-dimensional multimodal test function. 184

10-10Multimodal simulations using the nine motor configuration. The test function (scaled) is shown in solid red. (a) individual bacterium variables. (b) Average density plot using 10 simulations. 185
10-11Average density plot using 40 multimodal simulations with nine motor configuration. $v = 0.6$ units/s and $w_r = \pi$ rad/s. 185
10-12Average density plot using 40 multimodal simulations with nine motor configuration. $v = 0.5$ units/s and $w_r = 1.25\pi$ rad/s. 186
10-13Two-dimensional test function. 186
10-14Two-dimensional bacterial run. 187
10-15Two-dimensional run using a uniformly dropped bacterium. 187
10-16Average density plot of 10 uniformly dropped bacteria. 188
10-172D multimodal test function. 188
10-18Average density of 10 runs using the multimodal test function. 189
10-19Surface flattening using one bacterium. 189
10-20Average of 20 independent runs using one bacterium. 190
10-21Average of two simultaneous bacteria. 190

A-1 Representative Normal quantile-quantile (q-q) plots for different measures of the randomized networks corresponding to the damage-recovery category. 199

List of Tables

3.1 Node degree and characteristic path length for proteins with different degrees of sensitivity. 47

4.1 Large connected component (LC) size in newly defined sub-networks. *n is the total number of proteins in a given category. **The large component (LC) size is the number of connected proteins. Blue (red italic) font indicates a positive (negative) deviation from the random average. 53

4.2 Significance of the clustering coefficient analysis using biased randomizations. Blue (red italic) font indicates a positive (negative) deviation from the random average. The metabolic network will be discussed in the next chapter. 54

4.3 Protein complexes and signaling pathways identified by clustering coefficient analysis of the MMS and t-BuOOH networks. 61

4.4 Protein complexes and signaling pathways identified by clustering coefficient analysis of the 4NQO and UV networks. 62

5.1 Network measures for a number of functional yeast networks. n is the total number of proteins in a given category. (N-E) corresponds to non-essential versions of the networks. The large connected component (LC) size is the number of proteins in the LC. Blue font indicates a positive deviation from the random average and red italic font indicates a negative deviation from the random average. 68

5.2 Network measures for the metabolic network as compared to the networks defined in the previous chapter. Blue (red italic) font indicates that the measured value is greater (smaller) than the one obtained using the randomized networks. 68

5.3 Comparison of the expression and phenotypic damage-recovery data. *expr* corresponds to proteins that are differentially expressed in response to a damaging agent, *phe* corresponds to proteins whose deletion leads to impaired growth upon exposure to a damaging agent and *both* indicates proteins that are both differentially expressed and important for proper growth in response to a damaging agent. 70

5.4 Ratios quantifying the importance of one measurement (expression (E), phenotype (P), high degree (H)) compared to the other two. 72

5.5 Difference between down and up regulated genes. 73

5.6 Ratios quantifying the importance of one measurement (expression (E), phenotype (P), high degree (H)) compared to the other two for down-regulated genes. . . . 74
5.7 Network measures based on the core yeast interactome. . . . 76
5.8 Network measures based on the FYI yeast interactome. . . . 77
5.9 Number of total proteins, date hubs, and party hubs in the different phenotypic categories. . . . 77
5.10 MMS date hubs. . . . 78
5.11 MMS party hubs. . . . 79
5.12 4NQO date hubs. . . . 79
5.13 4NQO party hubs. . . . 80
5.14 UV date and party hubs. . . . 80
5.15 t-BuOOH date and party hubs. . . . 81
5.16 Summary of edge partitioning results. . . . 82
5.17 Proteins identities (top 5) with the highest number of edges cut in the full and core networks. All proteins with edges cut in the FYI network. . . . 84

7.1 Example parameters. . . . 107
7.2 Gillespie algorithm parameters. . . . 108

8.1 Transition probabilities for the model in Figure 8-4. For bimolecular reactions, the relevant γ is obtained from the relevant k (in $\mathrm{M^{-1}s^{-1}}$) using the expression $\gamma = \frac{k\Delta t}{v_r A_V V}$. . . . 117
8.2 Transition probabilities for the model in Figure 8-5. For bimolecular reactions, the relevant γ is obtained from the relevant k (in $\mathrm{M^{-1}s^{-1}}$) using the expression $\gamma = \frac{k\Delta t}{v_r A_V V}$. . . . 118
8.3 Parameters values for the model in Figure 8-4. All K values are in $M^{-1}s^{-1}$ and all concentrations are in M. . . . 118
8.4 Parameter values for the model in Figure 8-5. All K values are in $M^{-1}s^{-1}$ and all concentrations are in M. . . . 119

9.1 Transition probabilities for the model shown in Figure 9-4. For bimolecular reactions, the relevant γ is obtained from the relevant k (in $\mathrm{M^{-1}s^{-1}}$) using the expression $\gamma = \frac{k\Delta t}{v_r A_V V}$. . . . 140
9.2 Model rate constants. Rate constants were obtained from Spiro *et al.* [118]. 140
9.3 Model parameters. N_X is the total number of molecules of species X and were obtained from the concentrations in [118]. . . . 141
9.4 Transition probabilities for the motor model. For bimolecular reactions, the relevant γ is obtained from the releval k (in $\mathrm{M^{-1}s^{-1}}$) using the expression $\gamma = \frac{k\Delta t}{v_r A_V V}$. . . . 141
9.5 Motor model rate constants obtained from [97]. . . . 141
9.6 Interactions of the updated two-state receptor model. . . . 151
9.7 Parameter values of the updated two-state receptor model. . . . 153
9.8 Transition probabilities for the six-state receptor model. For bimolecular reactions, the relevant γ is obtained from the relevant k (in $\mathrm{M^{-1}s^{-1}}$) using the expression $\gamma = \frac{k\Delta t}{v_r A_V V}$. . . . 163
9.9 Six state receptor model rate constants. . . . 163

10.1 Model rate constants used for the surface mapping algorithm. 176
10.2 Motor model rate constants used for the surface mapping algorithm. 177

A.1 Rank of each statistic in the tested network with respect to the values obtained in the 1,000 randomized sets. 198

Chapter 1

Introduction

Signal processing is an integral part of cell biology. The associated biological algorithms are implemented by signaling pathways that cell biologists are just beginning to understand and characterize. One of our long term objectives is to understand and model these biological algorithms. Another long term objective is to emulate these processes, i.e. to use them as metaphors in other engineered contexts such as *ad-hoc* wireless networks, distributed sensor networks, and general algorithm development with the possibility of developing new, efficient, robust signal processing systems. Toward these ends, the main focus of this thesis is on developing new frameworks for modeling cellular signal processing. An example of how the results obtained from the modeling can be exploited to develop a new generation of algorithms for engineered systems is presented in the last chapter.

1.1 General Description

Biological signaling takes on different forms ranging from electrical signals through nerve synapses, physical signals such as mechanical stress or pressure at the surface of cells, to chemical signals such as hormone concentrations in the bloodstream. While some of these signals, notably electrical and physical signals, have been historically easier to study and control than others, the emergence of high throughput technologies for molecular biology is making the study of biochemical signaling networks a possibility. In addition, as more signaling networks and elements are identified, their complexities and the intricate interactions among signaling molecules, or cross-talk, are quickly becoming intractable. In fact, although the molecular components comprising cells are being cataloged, at a continually accelerating rate, there is no effective knowledge of how these components work together as an integrated dynamical system to yield output cell functions (e.g. survival, proliferation, death, differentiation, migration, secretion, etc.) as responses to information (chemical, mechanical, electrical, ...) presented in the cell environment. Understanding how cells do signal processing therefore eventually requires models that define layers of abstraction in order to view signaling algorithms at different resolutions.

Biology also presents a potentially very fruitful metaphor for signal processing. In fact, there is very strong evidence of interesting sophisticated signal processing operations performed by living systems including ones reminiscent of frequency modulation coding [119], multi-resolution filterbanks, gradient search algorithms and many more. Historically, nature and biology on a non-cellular level have provided the inspiration for widely exploited signal processing methods such as neural networks, chirp signals, etc. The possibility of exploiting cell biology to develop new signal processing algorithms is another underlying motivation for the work presented in this thesis.

We limit our study to signaling networks *within* cells. The intracellular signaling networks we are interested in are composed of proteins and enzymes as well as other molecules such as DNA, phosphates, and ATP. Signaling usually occurs through a series of protein modifications such as phosphorylation (the addition of phosphate groups) and cleavage, translocation from the cytoplasm to the nucleus, as well as control of gene expression. Usually, signals propagate through cascades where one protein affects the activity of another downstream protein. However, cross-talk also plays an important role in the signaling mechanism where several proteins coming from different upstream signals converge onto one single downstream signal or one upstream signal diverges to affect different downstream signals. These protein signaling *networks* are the focus of this thesis. Specifically, this work presents a detailed study and exploration of protein networks from an engineering perspective and at different levels of resolution. The first part of this thesis starts by looking at

these networks at the highest level of abstraction by focusing on the interconnectivity of proteins rather than on their identities. We then add resolution to our model by examining protein dynamics and how they interact. This leads to the formulation of a new stochastic framework for modeling internal fluctuations in biological cells. Finally, using the tools developed to understand and model biological networks, we then use these models to explore a new generation of networks where nodes are no longer biological components but mathematical entities or engineered systems such as sensors and processors. This leads to the formulation of a new search algorithm.

■ 1.2 Summary of Contributions

The main contribution of this work is that it provides a framework for examining biological signaling networks at different levels of abstraction and therefore presents a methodology for addressing different biological questions. The work also includes an example of how such a study can lead to new algorithms for engineered systems. Specifically in this thesis, we have developed:

- A graph theoretic model of the whole-genome relationship between cell genotype (genomic content) and phenotype (pathophysiological behavior) in response to toxic agents (chemicals and radiation) in the environment, for yeast as the currently more genomically-complex available experimental system.
- A unified modeling framework for simultaneously capturing the stochasticity of signaling networks in individual cells and computing a deterministic solution which provides average behavior.
- A biologically inspired surface mapping algorithm that provides an example for exploiting biological models to develop signal processing algorithms.

■ 1.3 Chapter Outline

There are two main sections in this thesis. The first section presents the high-level modeling results. Specifically, Chapter 2 provides an introduction to graph theoretic techniques and how they have been applied to biology. The graph theoretic analysis of damage-recovery networks is given in Chapters 3 and 4: Chapter 3 investigates the global properties of the damage-recovery proteins in the context of the full yeast interactome while in Chapter 4 we define newly synthesized phenotypic networks and examine their local properties. Finally Chapter 5 extends this analysis to other functional yeast networks. The second section presents the low-level modeling results. Chapter 6 provides the necessary background and a survey of current low-level models for biological signaling. The general model description is given in Chapter 7. The model is then applied to two biological pathways: the mitogen-activated protein kinase cascade (Chapter 8) and bacterial chemotaxis (Chapter 9). Exploiting bacterial chemotaxis to map and flatten surfaces is explored in Chapter 10. Finally Chapter 11 concludes the thesis and outlines contributions.

Chapter 2

Graph Theoretic Techniques Applied to Biology

At the highest level, biological networks (or any other network) can be viewed as connected graphs of "black boxes". From this point of view, the focus is solely on the connectivity of the network rather than on the dynamical properties of the individual elements (proteins, DNA, molecules). Using this level of abstraction, one can analyze the properties conferred by the topology of the network independently of the dynamics of the individual components. Specifically, there exists a rich mathematical framework for analyzing graphs at this level, namely the field of random graph theory. In this chapter, we first give some background on random graph theory. We then overview the current work applying this theory to biology, and finally introduce the yeast protein-protein interactome.

■ 2.1 Random Graph Theory

Random graphs have been used to model diverse complex systems such as social, biological, and economic systems. While much is known about some random graph structures such as uncorrelated random graphs [14] [45] [59], much less is known about real-world graphs describing complex systems such as the Internet, metabolic pathways, networks of power stations, etc. In this section we define random graphs and introduce some basic notions associated with them. The reader is referred to [41] and [23] for more details.

A graph (network) is a set of vertices (nodes) connected via edges (links). If these connections are statistically distributed, then the corresponding network is a random graph. We restrict our attention to graphs with unit weight on all edges and that do not include unit loops. These graphs can have non-symmetric edges with directionality in which case they are termed *directed graphs*. On the other hand, *undirected graphs* have their nodes simply connected via undirectional edges. In addition, there can be equilibrium graphs and non-equilibrium graphs. *Equilibrium graphs* have a fixed total number of vertices and random connections between them while in *non-equilibrium* graphs new vertices are added to a network all the time and new edges emerge between the growing number of vertices. An example of an equilibrium graph is the classical random graph introduced by Erdos and Renyi [43] [44] where the total number of nodes are fixed and randomly chosen pairs of nodes are connected via undirected edges. The citation graph [107], on the other hand, represents an example of a non-equilibrium graph where at each time step a new node is added to the graph and it is connected with some old node via an undirected edge.

There are several measures and functions associated with random graphs that allows their characterization, we present some of the most important ones below.

■ 2.1.1 Degree, loops, and degree distribution

The total number of connections of a node is called its *degree* k. In a directed graph, the number of incoming edges of a node is the *in-degree* and denoted by k_i while the number of outgoing edges of a node is its *out-degree* k_o. Clearly $k = k_o + k_i$. In general, the number I of irreducible loops (i.e. loops that cannot be reduced to a combination of smaller ones) in an arbitrary undirected connected graph is related to the number of edges, L, and the number of nodes, N, as follows: $I = L + 1 - N$.

Since the degree of nodes in random graphs are statistically distributed, another property of the graph is its *degree distribution*. For an undirected graph we define $p(k, s, N)$ as the probability that the node s in the graph of size N has k connections (i.e. k nearest

neighbors). The total degree distribution of the graph is then given by:

$$p(k, N) = \frac{1}{N} \sum_{s=1}^{N} p(k, s, N) \tag{2.1}$$

If all nodes of a random graph are statistically equivalent each of them has the same degree distribution $p(k, N)$, in this case, the degree distribution completely determines the statistical properties of the graph. A similar set of distributions can be defined for the in-degree and out-degree of directed graphs (see [41] for more details). Common degree distributions are Poisson, exponential, and power-law. For example, a power-law graph is defined as a network where the number of vertices of degree d is proportional to $d^{-\delta}$ for some positive constant δ. δ is usually referred to as the exponent of the power-law graph. Such graphs have very few nodes of high degree (generally referred to as hubs) and many nodes of low degree. They exhibit scale-free behavior, i.e. their properties or behaviors are invariant across changes in scale. As a result, knowledge of the topology of a small part of a network provides a reliable means of estimating the topology of the complete network.

■ 2.1.2 Adjacency matrix, spectra of random graphs

A graph can be characterized by its *adjacency matrix*, A, which indicates which of the vertices are connected (adjacent). A is a square $N \times N$ matrix where N is the total number of vertices in the graph. For directed graphs, the entries of A are such that $a_{ij} = 1$ if there is a directed edge going from vertex i to vertex j. For undirected graphs, A is symmetric with $a_{ij} = a_{ji}$. Since unit loops are not allowed, the diagonal elements of A are equal to zero, i.e. $a_{ii} = 0$. Furthermore, since the graphs described here are random, the adjacency matrix completely describes only one particular realization of the graph. From the adjacency matrix, one can obtain several characteristics of an undirected graph. In particular, the degree of a node is given by $k_i = \sum_j a_{ij}$ while the total degree of the graph, $K = \sum_i k_i$, is double the total number of edges L, and therefore is given by $K = 2L = Trace\{A^2\}$. The total number of loops of length three in the graph is $N_3 = \frac{1}{6} Trace\{A^3\}$ and the total number of connected triples of vertices in the graph is $T = \frac{1}{2} \sum_i k_i(k_i - 1)$.

The *spectrum* of a random graph is the distribution of the eigenvalues of the adjacency matrix. The spectral densities of uncorrelated random graphs, for example, follow the semi-circle law [14] [45] [59] while it has been shown that the spectral density of the eigenvalues of the adjacency matrix of a power-law graph approaches a triangle-like shape as the nodes in the graph increase [45]. Further details can be found in [45].

■ 2.1.3 Clustering coefficient

The *clustering coefficient* characterizes the 'density' of connections in the environment close to a node and is only well defined for undirected graphs. It is the ratio between the total number y of edges connecting its nearest neighbors, and the total number of all possible edges between these nearest neighbors:

$$C = \frac{2y}{z(z-1)} \tag{2.2}$$

where z is the number of nearest neighbors. The average value $\bar{C}$ is the clustering coefficient of the network. Intuitively, $\bar{C}$ can be thought of as the probability that if a triple of vertices

of a network is connected together by at least two edges, then the third edge is also present: it shows the 'density' of small loops of length 3 in the network. For undirected graphs, the clustering coefficient of the graph can be computed from the adjacency matrix as

$$\bar{C} = \frac{1}{9}\frac{Trace\{A^3\}}{\sum_{i\neq j}(A^2)_{ij}}. \tag{2.3}$$

■ 2.1.4 Small worlds

Consider a network in which all edges are unit length. We define the *distance* between two vertices of a network as the length of the shortest path between them. The distance l between a pair of nodes of a random graph is distributed with some distribution function $p(l)$ which is the probability that the length of the shortest path between two randomly chosen nodes is equal to l. $p(l)$ describes one of the main structural characteristics of the network. For undirected graphs, the distance l of the shortest path between nodes i and j is equal to the minimal power of the adjacency matrix with non-zero ij element, i.e.

$$(A^{l-1})_{ij} = 0 \text{ and } (A^l)_{ij} \neq 0 \tag{2.4}$$

The average shortest path length for an uncorrelated network is small even for very large networks. Graphs that exhibit small average shortest path lengths while being highly clustered are termed "small-world" graphs [130] [131]. It has been further shown that the time required for the spreading of a perturbation in a small-world network is close to the theoretically possible minimum for any graph with the same number of nodes and edges [130].

■ 2.1.5 Betweenness

The *betweenness* $\sigma(m)$ (also called *load* or *betweenness centrality*) of a node m is the total number of shortest paths between all possible pairs of nodes that pass through this node. This quantity was first introduced in sociology [49] to characterize the "social" role of a node. It indicates whether or not a vertex is important in traffic on a network. $\sigma(m)$ is given by:

$$\sigma(m) = \sum_{i\neq j}\frac{B(i,m,j)}{B(i,j)} \tag{2.5}$$

where $B(i,j)$ is the total number of shortest paths between i and j and $B(i,m,j)$ is the proportion of these shortest paths that pass through node m. The sum is over all pairs of nodes for which at least one path exists, that is with $B(i,j) > 0$. The nodes with high betweenness actually control the net. A network can also be decomposed into a pair of separated subnets by successive deletion of the edges with maximum betweenness (note that each deletion changes the betweenness of other edges).

■ 2.1.6 The giant connected component

All the characteristics described so far are local properties of the graph; they do not give a picture of the global topology of the networks involved. To get a global picture, one needs a description of the percolation properties of the network. A distinct connected component of a network is defined as the set of mutually reachable vertices. The size of a connected component is the total number of nodes in it. When the relative sizes of all connected

components of a network tend to zero as the number of nodes in the network tends to infinity, the network is below the percolation threshold. If the relative size of the largest connected component approaches a finite (non-zero) value in the limit of large networks, the network is above the percolation threshold. In this case, the huge connected component plays the role of a percolating cluster and is called the *giant connected component (GCC)*. The *giant strong connected component (GSCC)* is the set of nodes which are mutually reachable by a directed path. The *giant out-component (GOUT)* is the set of nodes approachable from the GSCC by a directed path while the *giant in-component (GIN)* contains all nodes from which the GSCC is approachable. If the GCC is absent, an undirected graph is only a set of separate clusters.

■ 2.1.7 Methods for generating random graphs

We end this subsection with three methods for generating random graphs with different degree distributions.

We start with an equilibrium network with attachment of edges without preference: starting with a large number of edges, we connect their ends to randomly chosen nodes i.e. the attachment of edges occurs without preference. The probability that an edge becomes attached to a node is independent of the degree of the node. It can be proved that the degree distribution of the resulting graph is a *Poisson distribution.*

One obtains the *stationary exponential distribution* as the degree distribution in the case of a growing network with attachment of edges without preference. This network is generated by adding at each time step a node and connecting a pair of randomly chosen nodes.

Finally, in the growing network with preferential attachment, the probability that the end of a new edge becomes attached to a node of degree k is proportional to $k+b$ (linear type of preferential linking) where $b > 0$ is a constant (this is a generalization of the Barabasi-Albert model [14]). In this case, it can be shown that the degree distribution is a *power-law* with $p(k) \propto k^{-(2+\frac{b}{2})} = k^{-\gamma}$. γ changes from 2 to infinity as b grows from 0 to infinity. It is the combination of growth and *linear* preferential linking that naturally leads to power-law degree distributions.

■ 2.2 Random Graphs in Biology

In the past few years, scientists have conducted several empirical studies of very diverse real world networks including electrical power grids, world wide web, the Internet backbone, telephone call graphs, co-authorship and citation networks of scientists, etc. Within biology, several networks have been approached using random graph theory including neural networks, networks of metabolic reactions, genomic and protein networks, ecological and food webs, and world web of human language [41]. Since the thesis focus is on biochemical signaling networks, we will only give background on the approaches investigating such types of networks, namely models of metabolic networks, genomic regulation networks, and protein interaction networks.

■ 2.2.1 Networks of metabolic reactions

In metabolic reaction networks, the nodes are substrates, i.e. molecular compounds that are either educts (inputs) or products (outputs) in metabolic reactions. Directed edges connect the educts and products that participate in a metabolic reaction. Jeong *et al.* [74] carried

out a systematic comparative analysis of the metabolic networks of 43 organisms representing all three domains of life. The data they used was obtained from the WIT database [101] which is an integrated pathway-genome database that predicts the existence of a given metabolic pathway on the basis of the annotated genome of an organism combined with firmly established data from the biochemical literature. They showed that these networks have similar topological scaling properties, namely they exhibit a scale-free structure , i.e. the probability that a given substrate participates in k reactions follows a power-law distribution. In addition, the measured value of the average directed shortest-path length is very close for all 43 networks: $\bar{l} = 3.0 \sim 3.5$. This average is robust to random removal of nodes: the random removal of 10% of the total number of nodes in the metabolic network of the bacterium *E.Coli* does not produce any noticeable variation of the average shortest-path length. However, when the most connected nodes (substrates) are removed, the average shortest path increases rapidly, suggesting the special role of these metabolites in maintaining constant average shortest path of the metabolic network. They also found that practically the same substrates act as hubs in all organisms. Since only 4% of all substrates that are found in all 43 organisms are present in all species, these substrates represent the most highly connected substrates found in any individual organism. It implies the generic utilization of the same substrates by each species while species-specific substrates are less connected. As a result, these highly connected substrates may provide the connections between modules responsible for distinct metabolic functions.

Wagner and Fell [127] also investigated the metabolic network of the bacterium *E.Coli* and found that it is a small-world graph and that the connectivity of the metabolites follows a power-law. They also suggested that the small-world architecture may serve to minimize transition times between metabolic states, i.e. it may allow a metabolism to react rapidly to perturbations.

■ 2.2.2 Genetic regulatory networks

Genetic regulatory networks describe which genes regulate each other. They are usually inferred from gene expression data sets which contain measured responses of the studied genes to various stimuli. Specifically, a genetic network is a set of genes that interact through directed transcriptional regulation, i.e. a set of genes that encode transcription factors (regulating proteins) that can bind to specific DNA control regions of regulated genes to activate or inhibit their transcription. Since regulated genes can themselves act in a regulatory manner, they create the genetic regulatory network. As a result, the transcriptional regulatory network is represented by a graph where vertices are genes and directed edges denote activating or repressing effects on transcription. It should also be noted that the transcription regulatory network is naturally directed as opposed to the network of physical protein interactions which we will discuss in the next section and which in principle lacks directionality.

Guelzim *et al.* [62] created a graph of 909 genetically or biochemically established interactions among 491 yeast genes. They found that the number of regulating proteins per regulated gene (in-degree) has a narrow distribution with an exponential decay while the number of regulated genes per regulating protein (out-degree) has a broader distribution with a decay resembling a power-law.

Farkas *et al.* [46] used microarray data on 287 single gene deletion yeast *Saccharomyces cerevisiae* mutant strains to elucidate generic relationships among perturbed transcriptomes. They defined a similarity measure between two segments and constructed a graph

such that two nodes were connected if the similarity score computed for the two transcriptomes they represent exceeded a fixed threshold. Ultimately, they created a similarity network in which each node represents one of the 287 deleted genes and their corresponding transcriptional response programs.

Similarly, Rung *at al.* [110] built genome wide disruption networks for yeast. However, in their graph, nodes represent genes and arcs connect nodes if the disruption of the source gene significantly alters the expression of the target gene.

Finally, Maslov and Sneppen [89] studied the yeast genetic regulatory network formed by 1,289 directed positive or negative direct transcriptional regulations within a set of 682 proteins. They found that links between highly connected proteins are systematically suppressed whereas links between a highly connected protein and a lowly connected one are favored. They concluded that the resulting topology decreases the likelihood of cross-talk between different functional modules of the cell and as a result increases the overall robustness of a network by localizing effects of deleterious perturbations.

■ 2.2.3 Protein interaction networks

Protein interaction networks are networks whose nodes are proteins and whose edges represent pairwise direct physical protein-protein interactions. The edges of the network are generally undirected since the physical interactions among proteins lack directionality. These edges form biochemical and signalling pathways by which in turn the production and degradation of proteins is regulated.

Jeong *et al.* [74] studied the protein-protein interaction network of the yeast *Saccharomyces cerevisiae* based on data obtained mostly from two-hybrid analyses. Their network consisted of 1,870 nodes (proteins) and 2,240 edges which represent direct physical protein-protein interactions. They suggested that the probability that a given yeast protein interacts with k other yeast proteins follows a power-law. In addition they tested the network robustness to random errors by removing random nodes in the network and examining the effect on the topology. They found that the network is robust to these random errors. This was further checked with results from systematic mutagenesis experiments which show a striking capacity of yeast to tolerate the deletion of a substantial number of individual proteins from its proteome. They also ranked all interacting proteins based on the number of links they have and correlated this with the phenotypic effect of their individual removal from the yeast proteome. They found that the likelihood that removal of a protein will prove lethal correlates with the number of interactions it has. As a result, they concluded that highly connected proteins with a central role in the network's architecture are three times more likely to be essential than proteins with only a small number of links to other proteins.

Maslov and Sneppen [89] also analyzed the topological properties of interaction networks in the yeast *Saccharomyces cerevisiae.* The interaction network they studied consisted of 4,549 physical interactions between 3,278 yeast proteins measured by two-hybrid screens. They found that the network followed a power-law with exponent ranging from 2.2 to 2.8 and the number of neighbors ranging from 2 to 100. They further observed that direct links between hubs are suppressed and that hubs tend to share fewer of their neighbors with other hubs thereby extending their isolation to the level of next-nearest neighbor connections. This reinforces the picture of functional modules clustered around individual hubs.

■ 2.3 The Yeast Interactome

The yeast *Saccharomyces cerevisiae* (also abbreviated as *S. cerevisiae*) is currently one of the most extensively studied eukaryotic organisms. Published data include information about transcriptional regulation networks as well as protein interaction networks. As can be seen from the previous section, the yeast is therefore currently the organism of choice for applying network analysis and characterization techniques. In this thesis we also use the yeast for our network studies and focus on the yeast protein-protein interaction graph (also called interactome).

The yeast protein-protein interaction network used in this thesis is compiled from various online databases including the Bimolecular Interaction Network Database (BIND) [11] [20] and the Database of Interacting Proteins (DIP) [37]. This data is then processed in MATLAB to create a yeast protein-protein interaction graph. The protein-protein graph has 4,684 nodes (proteins) and 14,493 undirected edges (protein-protein interactions). The interactome is shown in Figure 2-1. This figure was generated using the program Cytoscape (http://www.cytoscape.org).

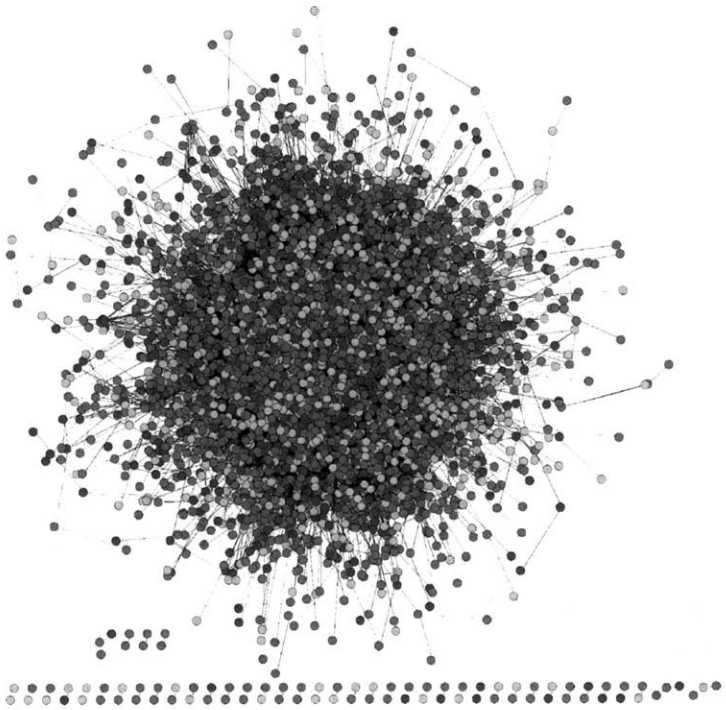


Figure 2-1. The yeast interactome.

Nodes in the graph correspond to proteins while edges are experimentally characterized physical protein-protein interactions. Some of these interactions are identified through small-scale experiments carried out in different laboratories while the majority of these interactions are identified through genome-wide experiments such as yeast two hybrid screens [123] and mass spectrometric analysis of protein complexes [65] [54]. These genome-wide screens carry with them sources of errors that one needs to be aware of. For example yeast

two hybrid experiments give rise to false positives of two kinds. In one case, the interaction between proteins is real but it never happens in the course of the normal life cycle of the cell, due to spatial or temporal separation of participating proteins. In another case, an indirect physical interaction is mediated by one or more unknown proteins localized in the yeast nucleus. In addition, there are false negatives where a binding may not be observed if the conformation of the bait or prey heterodimer block relevant interaction sites or if the corresponding heterodimer altogether fails to fold properly. It is therefore crucial to test for the potential effect of false positives and false negatives when analyzing the network in order to draw meaningful, robust conclusions.

In the next three chapters, we use the yeast interactome and some of the graph theoretic techniques presented here to investigate network properties of phenotypic effects in the yeast *S. cerevisiae*.

Chapter 3

Global Network Properties of Damage Recovery Proteins

Using genome-wide information to understand holistically how cells function is a major challenge of the post-genomic era. Recent efforts to understand molecular pathway operation from a global perspective have lacked experimental data on phenotypic context, so insights concerning biologically-relevant network characteristics of key genes or proteins have remained largely speculative. In this and the next chapters, we present a global network investigation of the genotype/phenotype data-set developed for the recovery of the yeast *S. cerevisiae* from exposure to DNA damaging agents, enabling explicit study of how protein-protein interaction network characteristics may be associated with phenotypic functional effects. We show that proteins important for damage-recovery have topological properties similar to essential proteins suggesting that cells initiate highly coordinated responses to damage similar to those needed for vital cellular functions.

■ 3.1 Introduction

Cells represent complex systems with thousands of proteins, carbohydrates, lipids, nucleic acids and small molecules interacting to maintain growth and homeostasis. Such maintenance requires that cells appropriately respond to both endogenous and exogenous environmental cues. The recent completion of several genome projects provides us with parts lists of genes and proteins that contribute to maintaining growth and homeostasis. Our current challenge is to use this information to understand holistically how cells function. This means understanding how global responses are orchestrated by communication between the components in the network that mediate responses to environmental stimuli. Toward this goal, thousands of protein-protein interactions and genetic interactions have been mapped into complex networks for several organisms, including the budding yeast *Saccharomyces cerevisiae* [65] [123] [120] [121] [50] [71], the human gastric pathogen *Helicobacter pylori* [106], the fruit fly *Drosophila melanogaster* [58] and the nematode worm *Caenorhabditis elegans* [85] [129]. While some global network analysis has been performed on these interacting networks as described in the previous chapter, they have rarely been directly connected to systematic global genotype/phenotype information. We have chosen to connect protein-protein interaction maps for *S. cerevisiae* with a genomic-scale dataset describing the phenotypic role of all non-essential yeast proteins in helping cells recover after exposure to a number of DNA damaging agents, typical of those encountered in our endogenous and exogenous environment. This choice was motivated in part by the relevance of damage-recovery processes to human diseases such as cancer, aging, and other degenerative states.

As described in Chapter 2, graph theoretic approaches are now being used to study global properties of biological networks. Network analyses of this kind, however, are for the most part carried out in the abstract without any functional information, and when functional information is present it is usually based on cataloged information assembled from an array of unrelated experiments. Tools that allow for global systematic network perturbations are crucial for establishing biologically meaningful network characteristics, and the *S. cerevisiae* single-gene deletion library of strains provides a robust tool for such analyses. In theory, each gene deletion strain represents an engineered cellular model in which one node, and its corresponding edges, has been removed from the yeast genetic and protein network. High throughput phenotypic studies that use gene deletion strains to identify the associated phenotypic effects under specific experimental conditions (a procedure termed genomic phenotyping [18] [19]) provide biologically relevant data sets for network characteristics studies.

■ 3.2 Phenotypic Annotation of the Yeast Interactome

For the network characteristics study presented in this chapter, we integrated the recently generated complete genomic phenotyping data-set that catalogs the sensitivity of thousands of mutant S. cerevisiae strains to DNA damaging agents [19], with published *S. cerevisiae* protein-protein interaction data.

■ 3.2.1 Genomic phenotyping

Genomic phenotyping was carried out by Begley *et al* [19] and is briefly described here. 4,733 haploid *S. cerevisiae* single gene deletion strains were used to identify deletions that affect growth (compared to wild type) upon exposure to the simple methylating agent methyl methanesulfonate (MMS), the bulky alkylating agent 4-nitroquinoline-N-oxide (4NQO), the oxidizing agent tert-butyl hydroperoxide (*t*-BuOOH), or 254 nm UV radiation. Only the 4,733 non-essential yeast genes could be examined since deletion of the essential genes in haploid strains is, a priori, lethal. A highly sensitive, reproducible, multi-replicate and multi-dose screen was developed to monitor individual strain growth after exposure to the four DNA damaging agents; the results of this screen have been described [19]. The term *damage-recovery protein* in this study corresponds to the product of the gene that was deleted in a strain that displays significantly more growth inhibition than the wild type strain following exposure to a DNA damaging agent. It should be stressed that damage-recovery is a simplified label and meant to include proteins that prevent damage in addition to proteins that contribute to the repair of macromolecular damage. Such genomic phenotyping data was integrated with public protein-protein interaction data as described in the next section.

■ 3.2.2 Network construction

14,493 protein-protein interactions for 4,686 *S. cerevisiae* proteins found in the Database of Interacting Proteins (DIP) [135] as of November 2002 were used to build the yeast protein interactome. Essential, damage-recovery and no-phenotype classifications were based on results obtained by the *S. cerevisiae* gene deletion consortium (essential) and the high throughput genomic phenotyping study [19]. Specific damaging agents used for genomic phenotyping included methyl methanesulfonate (MMS), the bulky alkylating agent 4-nitroquinoline-N-oxide (4NQO), the oxidizing agent tert-butyl hydroperoxide (*t*-BuOOH), and 254 nm UV radiation. MMS-, *t*-BuOOH-, 4NQO- and UV-recovery proteins correlate to gene deletion strains that exhibit impaired growth, compared to wild-type, after agent treatment, and the damage-recovery phenotype represents gene-deletion strain sensitivity to one of these classical DNA damaging agents as compared to wild-type. The no-phenotype classification indicates that gene deletion strains, correlating to specified proteins, had no growth defects after agent treatment as compared to wild-type. Based on this genomic phenotyping data, nodes in the full network are color-coded, with essential proteins being black, damage-recovery proteins green (for all four damaging agents) and no-phenotype proteins red (Figure 3-1A). The resulting network structure thus represents a phenotypically annotated interactome of essential, damage-recovery, and no-phenotype proteins. This structure was analyzed using graph theoretic techniques. Each protein category, as defined by phenotype, was analyzed first in the context of the underlying global protein interaction network (the full network) and then in the context of several newly defined protein networks, each composed exclusively of proteins and edges from a particular phenotypic category. This

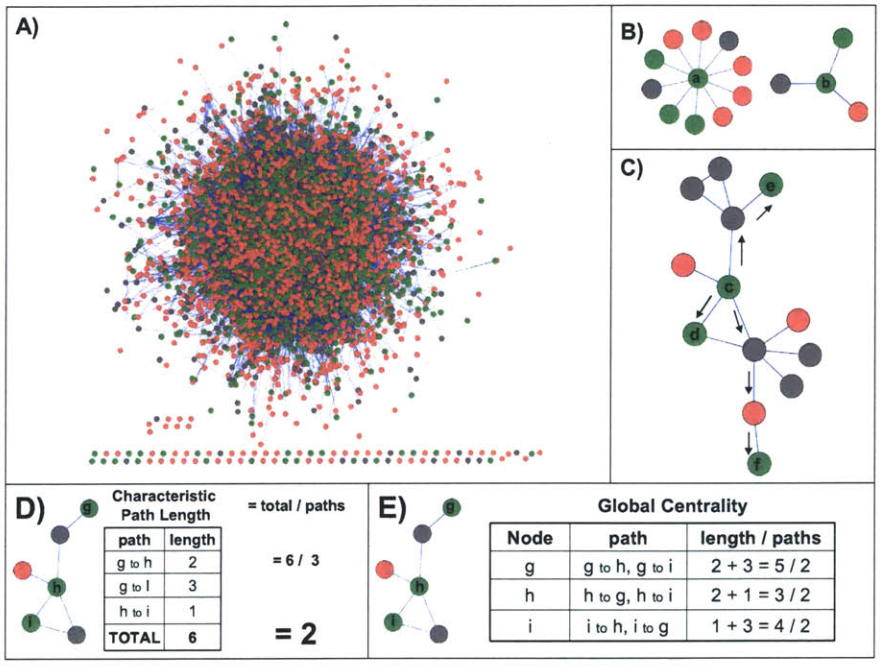


Figure 3-1. Yeast protein categories and global network measures. A. Proteins that prevent damaging agent induced cell death are shown in green, essential proteins are shown in black and proteins associated with no-phenotype are shown in red. B. Degree of a node in a graph. As an example, the degree of (a) is 10 while for protein (b) it is 3. C. Shortest-path length. D. Characteristic path length. E. Global centrality.

latter analysis is described in the next chapter.

■ 3.3 Network and Significance Measures

■ 3.3.1 Network metrics

We have exploited two fundamental network metrics to extract phenotype-dependent global network characteristics for essential, damage-recovery and no-phenotype proteins in the context of the full yeast network (Figure 3-1A). We determined the degree of each node, which details the number of interacting partners for each node in the network (Figure 3-1B), and the shortest-path distance between pairs of nodes, which details the shortest edge distance between similarly categorized pairs of proteins, but allowing transitions through proteins in other categories (Figure 3-1C). From the shortest-path distance, we computed two additional measures, the characteristic path length (Figure 3-1D) defined as the shortest-path distance averaged over all pairs of proteins, and a global centrality measure (Figure 3-1E) which, for a given protein, computes the average shortest-path distance to every other similarly categorized protein in the network. Each one of these measures reveals insight into the architecture of damage-recovery pathways and, as will be seen, all measures show that damage-recovery proteins are more similar to essential proteins than are no-phenotype proteins as well as proteins involved in other functional yeast networks.

In order to carry out the analysis, the proteins in the network were organized into seven different phenotypic categories: (1) essential; (2) damage-recovery (a composite of the next four categories); (3) MMS-recovery; (4) 4NQO-recovery; (5) UV-recovery; (6) t-BuOOH-

recovery; and finally (7) no-phenotype. The degree distribution, shortest-path distribution, characteristic path length and global centrality distribution were determined for each of the seven protein categories. We reasoned that the damage-recovery proteins, due to their importance in preventing cell death (the ultimate disturbance of cellular homeostasis), might have distinct network characteristics.

■ 3.3.2 Randomizations and p-values

Controlled randomizations based on the overall network structure were performed to determine the significance of analyzed characteristics. For the essential nodes, we compared the results dictated by the data to what would result from a corresponding number of nodes randomly selected from the full yeast protein network. For the non-essential proteins, i.e. the damage-recovery and no-phenotype proteins, the randomization consisted of a corresponding number of proteins randomly selected from the non-essential yeast network. This additional step was needed in order to un-bias the randomizations. Applying the randomization to the overall underlying structure of the yeast protein network by random selection of nodes is in contrast to other randomization network studies [130] [117] in which networks or sub-networks are constructed through random connections between nodes. It should be stressed that this randomization technique preserves the underlying distribution of the yeast interactome which is composed of mostly lowly connected proteins interfaced with a few highly connected hubs. As a result, the probability of selecting highly connected nodes is related to their relative abundance in the node set which is relatively low. Performing the randomization this way is crucial in order to reveal eventual bias in proteins with a phenotypic connection. Our results show that such bias does not always exist among proteins sharing a common function.

Specifically, eight sets of randomizations based on the eight phenotypic categories were generated. In each set, 1,000 randomized networks were obtained based on 1,000 independent experiments consisting of randomly selecting nodes from the full yeast network using a binomial distribution with mean equal to the ratio of the total number of nodes in that category over the total number of nodes in the full yeast network. Randomized networks were thus generated for the essential, non-essential, no-phenotype, damage-recovery, MMS-recovery, 4NQO-recovery, UV-recovery, and t-BuOOH-recovery categories. In order to factor out the bias introduced by essential genes, the randomizations corresponding to the no-phenotype, damage-recovery, MMS, 4NQO, UV, and t-BuOOH categories were generated by randomly selecting nodes from the non-essential yeast network; the essential proteins were excluded because we only have damage-recovery phenotype information for non-essential proteins. Based on these randomizations, p-values were computed using a two-sided hypothesis test with a Normal distribution assumption [39]. The Normal distribution assumption for each tested statistic was checked by generating Normal quantile-quantile (q-q) plots in MATLAB which graphically compare the distribution of the data set to the Normal distribution. Representative q-q plots are shown in Figure A-1 in Appendix A. Furthermore, to consolidate the reported p-values, we have computed the rank of the statistic in the tested network with respect to the values obtained in the 1,000 random sets. The rank is shown in Table A.1 in Appendix A.

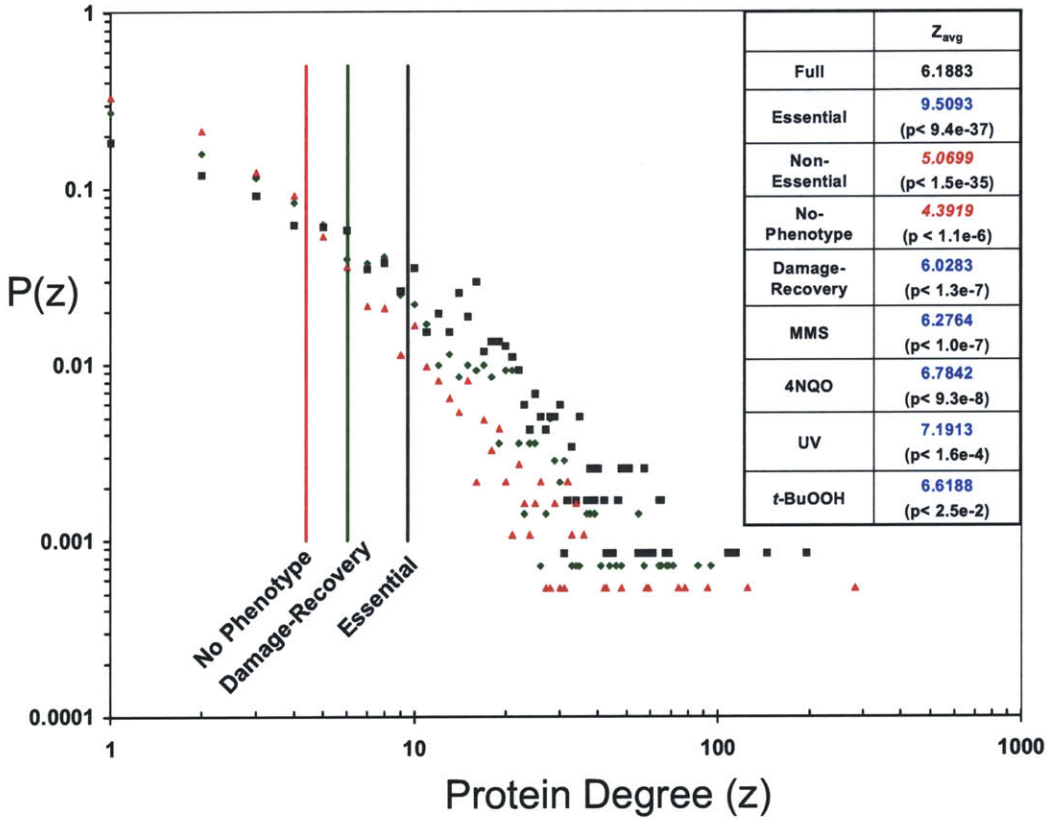


Figure 3-2. Degree distributions of selected proteins. Black squares, green diamonds, and red triangles represent essential, and no-phenotype proteins respectively. The solid vertical lines represent the average degree. Blue (red italic) font in the inset indicates an average greater (smaller) than the corresponding randomized average.

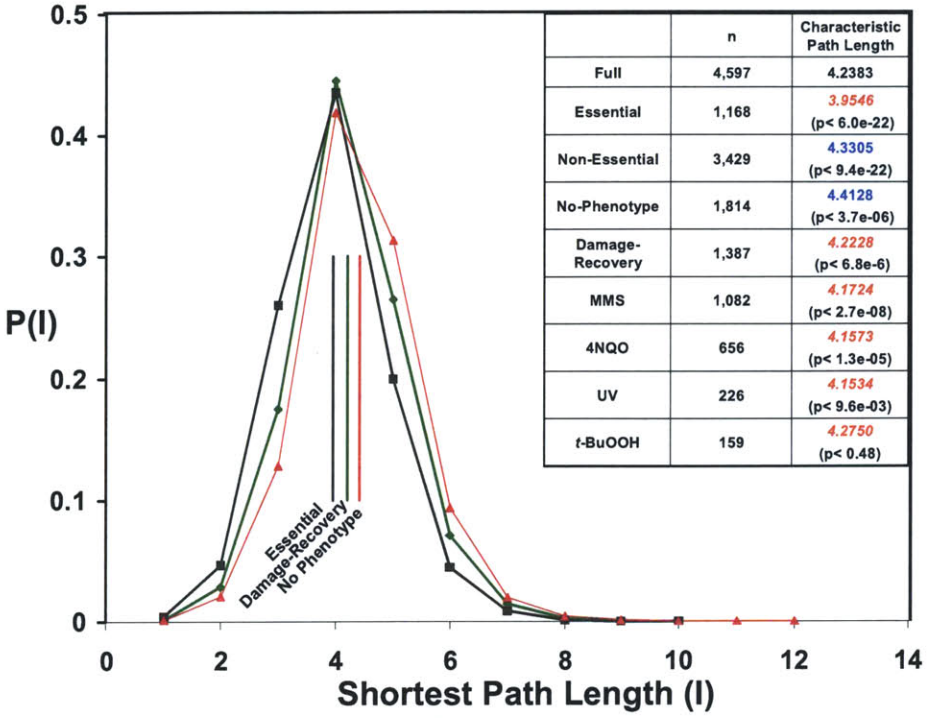
■ 3.4 Degree Distribution

We computed the degree distribution, as well as the average degree, for all of the nodes in the full yeast interactome, and for the nodes in each one of the seven categories defined above (Figure 3-2). These distributions have an inhomogeneous structure and are characterized by a number of highly connected proteins, or hubs, as previously observed by Jeong *et al.* and Maslov *et al.* for a much smaller yeast protein-protein interaction network [74] [89]. Furthermore, it has been suggested that an important consequence of this architecture is the network's simultaneous tolerance to random errors, and vulnerability to directed attacks on the most connected nodes [4]. When the full network is divided into essential and non-essential proteins, it is clear that essential proteins have a higher average degree. To assess the significance of this difference, the average degree value was compared to values obtained when the same number of proteins (1,180 in the essential category) was selected at random from the full network; we performed 1,000 such randomizations and the average degree for the 1,000 randomized networks was significantly smaller than that for the essential protein network ($p < 10^{-36}$); or, put another way, the average degree for essential proteins is significantly higher than that expected by chance. The results illustrated in Figure 3-2 show that not only is the essential proteins distribution skewed to the right (towards a higher degree) in agreement with previous results [74], but a similar trend is also true for damage-recovery proteins. Among the entire set of non-essential proteins, i.e. the set for which we have phenotypic data, it is clear that damage-recovery proteins have a higher average degree than the entire set of non-essential proteins as well as the no-phenotype proteins. On average each category of damage-recovery proteins, including the collective damage-recovery category, contained significantly more direct interactions than randomly selected proteins, (p values range from < 0.03 to $< 10^{-7}$) as well as non-essential and no-phenotype proteins. Moreover, the no-phenotype distribution shows the opposite characteristic, with significantly fewer direct interactions than the randomly selected proteins ($p < 10^{-6}$), fewer direct interactions than the non-essential category, and fewer direct interactions than for the full network. Note that there is an overlap of nodes among the specific agent networks and the fact that the collective damage-recovery category has a lower average degree than the individual agent specific categories indicates that most of the overlap is through high degree nodes.

The extent to which the essential and damage-recovery categories have a higher average degree can be further assessed by examining the probability of a yeast protein having a certain phenotype given it has a high degree where high is defined as having more than 15 direct interactions. Specifically, using the distribution shown in Figure 3-2, one can show that a protein with a high degree is two times more likely to be essential than a random protein in the yeast network. Furthermore, a non-essential protein with a high degree is one and a half times more likely to be important for damage recovery than a random non-essential yeast protein. This should be contrasted with the fact that a high degree protein is more than a third less likely to be involved in metabolism than a random protein in the yeast network (to be discussed further in Chapter 5).

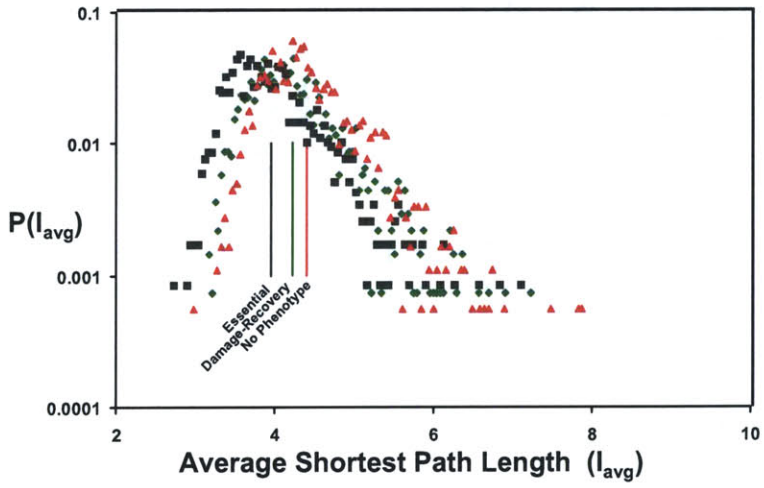
■ 3.5 Shortest Path Distribution, Characteristic Path Length, and Global Centrality Distribution

The existence of a few highly connected nodes (hubs) holding together a large number of lesser connected nodes serves to add shortcuts into a network and to create a smaller average



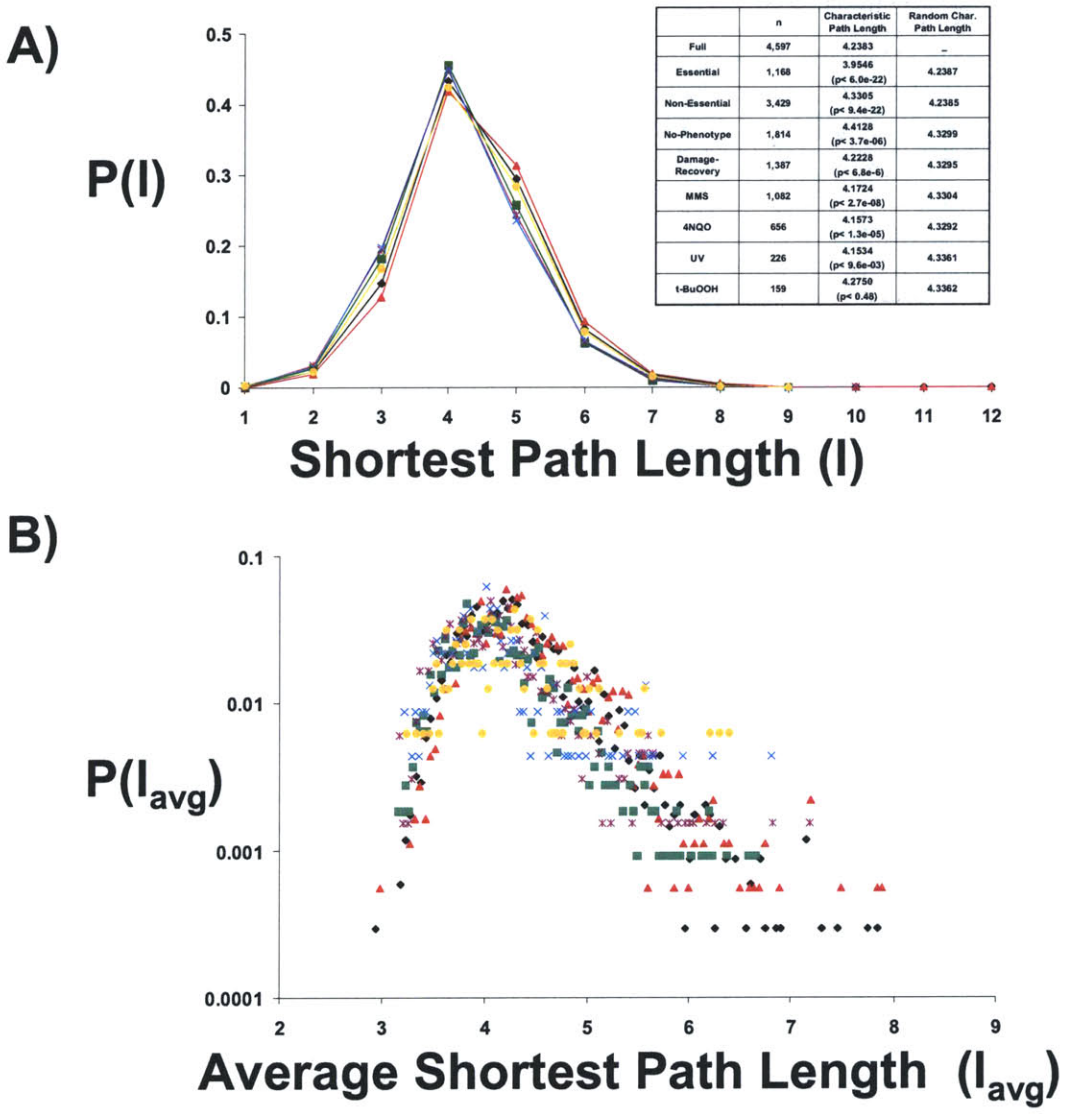
	n	Characteristic Path Length
Full	4,597	4.2383
Essential	1,168	*3.9546* (p< 6.0e-22)
Non-Essential	3,429	4.3305 (p< 9.4e-22)
No-Phenotype	1,814	4.4128 (p< 3.7e-06)
Damage-Recovery	1,387	*4.2228* (p< 6.8e-6)
MMS	1,082	*4.1724* (p< 2.7e-08)
4NQO	656	*4.1573* (p< 1.3e-05)
UV	226	*4.1534* (p< 9.6e-03)
t-BuOOH	159	*4.2750* (p< 0.48)

(a)



(b)

Figure 3-3. Shortest-path length distribution, characteristic path length (a), and global centrality distribution (b) of selected proteins. Black squares, green diamonds, and red triangles represent essential, and no-phenotype proteins respectively. The solid vertical lines give the characteristic path length. n is the number of proteins in each category. Blue (red italic) font indicates an average greater (smaller) than the corresponding randomized average.



	n	Characteristic Path Length	Random Char. Path Length
Full	4,597	4.2383	--
Essential	1,168	3.9546 (p< 6.0e-22)	4.2387
Non-Essential	3,429	4.3305 (p< 9.4e-22)	4.2385
No-Phenotype	1,814	4.4128 (p< 3.7e-06)	4.3299
Damage-Recovery	1,387	4.2228 (p< 6.8e-6)	4.3295
MMS	1,082	4.1724 (p< 2.7e-08)	4.3304
4NQO	656	4.1573 (p< 1.3e-05)	4.3292
UV	226	4.1534 (p< 9.6e-03)	4.3361
t-BuOOH	159	4.2750 (p< 0.48)	4.3362

Figure 3-4. Shortest-path length and global centrality of selected proteins. Non-essential (black diamonds), no-phenotype (red triangles), MMS (green squares), 4NQO (violet stars), UV (blue crosses), and *t*-BuOOH (gold circles) proteins.

shortest-path length between any two nodes in such a network. Because the essential and damage-recovery categories display high average degree relative to non-essential and no-phenotype proteins (respectively), it seems likely that their shortest path distribution will have a relatively short characteristic path length. We have computed the shortest-path length distribution, the characteristic path length, and the global centrality distribution for the full yeast protein network, as well as for each individual protein category, as specified by the seven phenotypic categories described above; note that in calculating the distance between two proteins in one particular category, the shortest-path can pass through proteins belonging to other categories (see Figure 3-1C). The results for three of the shortest-path length and global centrality distributions are shown in Figure 3-3, the rest are in Figure 3-4. The characteristic path length for all categories is given in the Figure 3-3 inset. At first sight, the shortest-path length distributions may not appear to be significantly different, but upon randomization of the networks by random protein sampling as described above, it becomes clear that the shortest path length distributions, and the characteristic path lengths, follow exactly the same highly significant trends as for the degree distributions and average degrees, respectively. In other words, among the proteins in the full network, essential proteins have a significantly shorter characteristic path length than non-essential proteins ($p < 10^{-21}$), and their shortest path length distribution is skewed away from the non-essential proteins, this time to the left (Figure 3-3 and Figure 3-4A). Among the non-essential proteins the damage-recovery proteins follow the same trend as the essential proteins (though less marked), in that they have a shorter characteristic path length than both the entire set of non-essential proteins and the no-phenotype proteins with p values ranging from < 0.01 to $< 10^{-7}$ (with the exception of the t-BuOOH category), and again their shortest-path length distributions are skewed to the left of no-phenotype proteins which have a larger characteristic path length ($p < 10^{-5}$) (Figure 3-3 and Figure 3-4A). The characteristic path length for t-BuOOH-recovery is not significantly different from randomized sampling, perhaps due in part to the small size of this category (159 proteins). The results show that essential proteins are visibly more central than proteins displaying no-phenotype, as well as being more central than non-essential proteins. Furthermore, damage-recovery proteins are also more central than non-essential and no-phenotype proteins. In addition, as will be shown in the next chapter, both essential and damage-recovery proteins are more central than proteins involved in metabolism. This kind of shortest-path analysis of the network may provide an idea of network navigability and of the efficiency with which a perturbation can spread throughout the network. However, in analyses of this type it is currently assumed that the connections between each node (i.e. the edges) are equivalent, and this seems very unlikely to be true in a biological system. This would be analogous to assuming that all edges are equal in a transportation network that contains highways and winding country roads as edges connecting bus stations as nodes. Ultimately, some metrics to describe the attributes of each edge in the protein-protein interaction network will be needed in order to be quantitative with respect to network navigability.

■ 3.6 The Centrality of Essential and Damage-Recovery Proteins in the Yeast Interactome

Degree distribution, shortest-path distribution, characteristic path length and global centrality distribution essentially provide different measures of centrality, and collectively these characteristics indicate that essential, damage-recovery, and no-phenotype proteins are quite distinguishable in the context of the yeast protein-protein interaction network. These pro-

	n	Average degree	Char. Path length
Damage-Recovery	**1,415**	**6.0283**	**4.2228**
D-R High	374	7.4893	4.0553
D-R Medium	692	6.7052	4.1821
D-R Low	757	5.0370	4.3131
MMS	**1,100**	**6.2764**	**4.1724**
MMS High	224	7.6741	4.0353
MMS Medium	397	7.3854	4.0985
MMS Low	479	4.7035	4.2897
4NQO	**672**	**6.7842**	**4.1573**
4NQO High	259	8.0270	3.9803
4NQO Medium	259	6.3900	4.2120
4NQO Low	154	5.3571	4.3520
UV	**230**	**7.1913**	**4.1534**
UV High	38	8.7895	3.7237
UV Medium	111	6.8919	4.2929
UV Low	81	6.8519	4.1334
***t*-BuOOH**	**160**	**6.6188**	**4.2750**
t-BuOOH High	10	8.1000	3.6222
t-BuOOH Medium	32	10.0313	3.9234
t-BuOOH Low	118	5.5678	4.4241

Table 3.1. Node degree and characteristic path length for proteins with different degrees of sensitivity.

tein categories are not only distinguishable from each other, but are also distinguishable from similarly sized groups of proteins obtained by random selection from the full network. Due to their phenotypic role in cell survival, it makes sense that damage-recovery proteins might represent a middle ground between essential and no-phenotype proteins. Essential proteins dictate cell viability under all conditions of life and their place in the network makes them the most centralized. The centrality of essential proteins may serve to provide facile communication between the processes vital for maintaining proper cellular function and homeostasis. Damage-recovery proteins are less centralized compared to essential proteins, perhaps because they are only required for cell viability some of the time (i.e., during stress). It may be that damage-recovery proteins are more centralized in the network than no-phenotype proteins, because under stressful conditions damage-recovery proteins need to rapidly coordinate a wide variety of cellular processes that will ultimately dictate cellular viability [19]. For MMS alone it has been postulated that extensive damage occurs to DNA, RNA, lipids, and proteins [18] [19] [73] [72] it therefore seems likely that a highly coordinated response to carry out the repair, removal and replacement of a multitude of damaged molecules is required for survival. Short path lengths via access to a number of highly connected nodes might serve to provide damage-recovery proteins with a means of optimizing cellular responses that together prevent damage induced cell death.

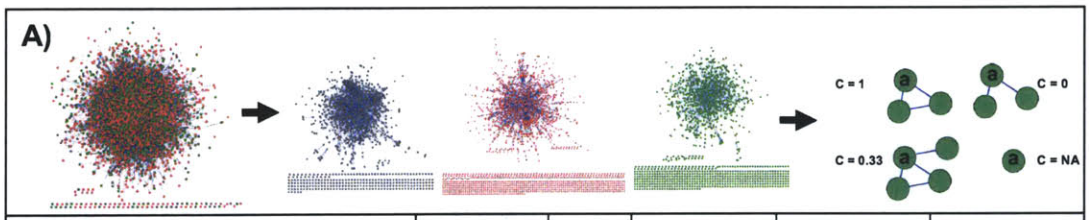
The fact that the importance of a protein may be reflected by its connectivity degree in the graph can be further investigated by quantitatively categorizing the damage-recovery proteins in high, medium, and low sensitivity categories and calculating the average degree

and characteristic path length of each category. Specifically, strains with high, medium, and low sensitivity were identified by using a range of exposure doses in our genomic phenotyping screen [19]. High, medium, and low sensitivity proteins were then compiled for each one of the four damaging agents and grouped into different categories. High, medium, and low damage-recovery categories were also obtained by combining the data of each of the four different agent-specific categories. The high damage-recovery category for example includes any protein that exhibits high sensitivity to any one of the four damaging agents. The global network measures were then computed for each category and the results are given in Table 3.1. The results indicate that highly sensitive mutants are distinct in their topological properties from their low-sensitivity counterparts. High sensitivity corresponds in most of the cases to higher connectivity degree and shorter characteristic path length, therefore further supporting the hypothesis that a protein with higher degree and greater centrality is on average more important for damage recovery than a lowly connected and less central one.

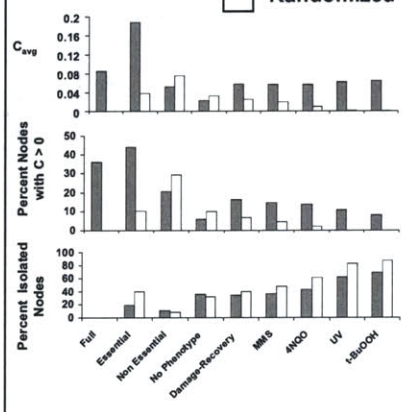
Finally, even though the degree distribution curves are significantly different from each other for several phenotypic categories, all seven phenotypic categories nevertheless embrace some number of highly connected proteins, or hubs, and some number of much lesser connected nodes. Thus, it is important to stress that it is the distributions and the average values that serve to distinguish between protein categories.

Chapter 4

Local Properties of Synthesized Phenotypic Networks



B)



C)

	n	p-value C_{avg}	p-value C> 0	p-value Isolated Nodes
Essential	1,180	2.0e-110	1.3e-105	1.5e-12
Non-Essential	3,504	2.5e-13	2.0e-22	1.5e-2
No-Phenotype	1,855	2.3e-02	3.6e-04	5.0e-2
Damage-Recovery	1,415	1.9e-11	1.1e-18	2.4e-2
MMS	1,100	2e-12	2.1e-19	3.1e-04
4NQO	672	1.1e-16	1.9e-31	1.8e-07
UV	230	8.5e-38	2.2e-61	1.2e-06
t-BuOOH	160	3.2e-53	1.9e-55	2.2e-05

Figure 4-1. Newly-defined networks and clustering coefficient analysis. A. Derivations of new networks: networks comprised of only essential proteins and connecting edges are shown in black, proteins that prevent agent induced cell death and connecting edges are shown in green and no-phenotype proteins and connecting edges are shown in red. The clustering coefficient (C), can be determined for each protein to identify the degree of connectivity between a given protein's neighbors. B. Clustering coefficient analysis. C. p values.

In the last chapter, we have analyzed the essential, damage-recovery, and no-phenotype proteins, in the context of the full yeast interactome, i.e. as selected groups of proteins within the overall network. We have looked at degree distributions in the overall network as well as shortest-path analysis, allowing proteins in each category to traverse non-category proteins to get to other proteins in the same category (see Figure 3-1). In order to gain further insight into the organization and local environment of proteins in each category, sub-networks were compiled that are composed solely of protein-protein interactions between proteins within a given phenotypic category. The local properties of these networks are the focus of this chapter.

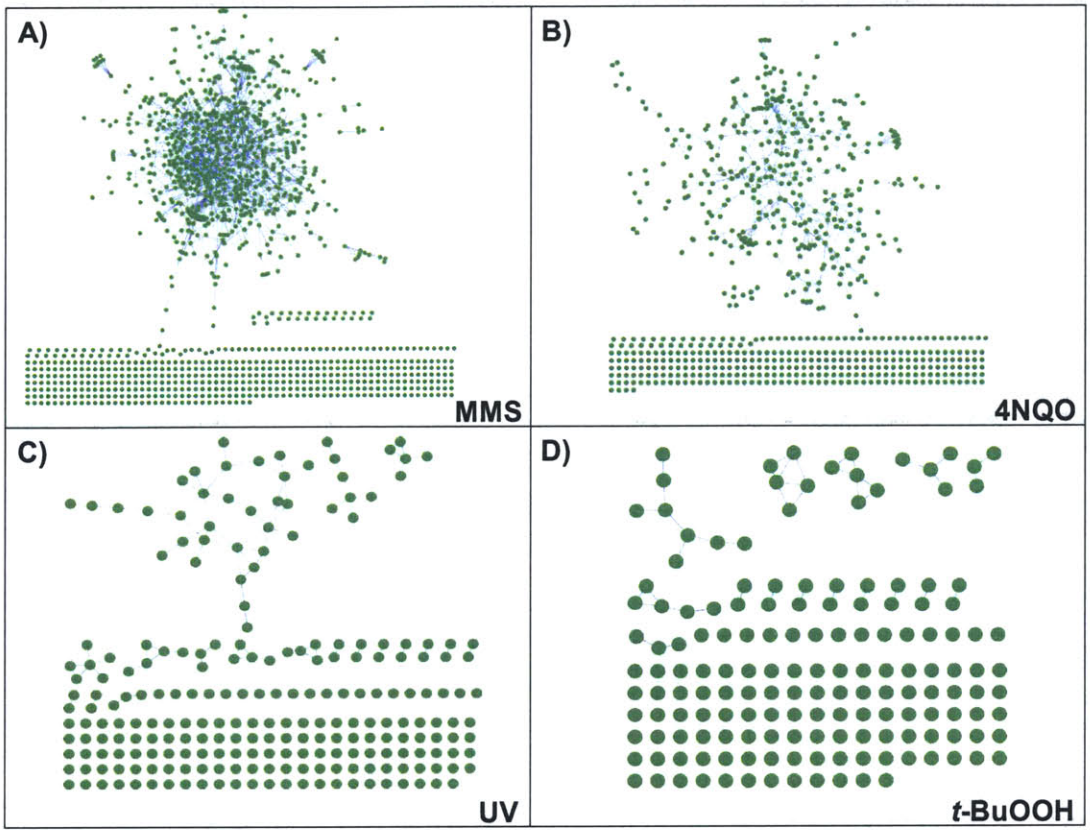


Figure 4-2. Newly-defined phenotypic sub-networks.

■ 4.1 Synthesis of Phenotypic Networks

The newly defined sub-networks have nodes corresponding to proteins exhibiting a given phenotype and edges representing experimentally characterized protein-protein interactions. We thus generated seven new sub-network structures, one for the essential proteins, five for the damage-recovery proteins, and one for the no-phenotype proteins.

Figure 4-1A illustrates the full network (as in Figure 3-1A), the essential sub-network, the no-phenotype sub-network, and the collective damage-recovery sub-network that includes MMS, 4NQO, UV and *t*-BuOOH recovery proteins. The individual MMS, 4NQO, UV and *t*-BuOOH sub-networks are shown in Figure 4-2. The overall network connectivity and the local properties of the nodes in these seven sub-networks were further investigated (Table 4.1 and Figure 4-1). The results were compared to the outcome from 1,000 randomized networks obtained by randomly selecting the same number of nodes in each one of the seven sub-networks from the full yeast interactome and assembling the corresponding network. For the non-essential sub-networks (no-phenotype and damage-recovery), the nodes in the corresponding randomized networks were picked from the non-essential yeast network in order to eliminate bias introduced by the essential proteins.

4.2 Network Metrics

4.2.1 Connected component analysis

We have computed the largest connected component size for each one of the eight phenotypic networks and compared it to the largest connected component size of the full yeast network as well as to the size of the largest connected component in the corresponding randomized networks.

4.2.2 Clustering coefficient

The clustering coefficient was compiled for each node in the essential, no-phenotype and each damage-recovery networks (Table 3.1). The average clustering coefficient per node, the percent of nodes with a positive clustering coefficient, as well as the percent of isolated nodes (i.e. that do not have any neighbors) was then computed for each network and compared to the corresponding randomized networks. For each phenotypic network, the set of nodes with a non-zero clustering coefficient was identified and visualized using Cytoscape [115].

4.3 Network Connectivity

The landscape of connected components in each newly defined network was explored to probe the connectivity of each structure. Formally, a connected component in a graph is defined by a set of nodes wherein every single node is reachable from every other node in the component, i.e. the set of nodes for which there exist paths to connect them all together. In each one of the seven networks (except the *t*-BuOOH-recovery network) one large connected component emerged (Table 4.1, Figure 4-1A, and Figure 4-2). For the essential, no-phenotype, collective damage-recovery, MMS-recovery, and 4NQO-recovery networks, the largest connected component included more than half of the nodes in the sub-network, i.e., using graph-theoretic terminology, a giant component emerged. It was further noted that the size of these components was significantly larger for the essential and each of the damage-recovery sub-networks (except that for *t*-BuOOH) than the size that would be expected if the nodes were selected at random from the full or non-essential yeast network (p values from < 0.009 to $< 10^{-8}$). In contrast, the no-phenotype sub-network had a smaller, though not statistically significant ($p < 0.074$), large connected component than the one expected if the nodes were randomly selected from the full non-essential yeast network. These observations indicate that the essential and damage-recovery proteins are relatively highly connected to each other, and we infer that this suggests that cohesive signalling pathways, protein complexes and biochemical pathways are at least partially represented in these novel sub-networks.

4.4 Local Protein Environments in the Phenotypic Sub-Networks

We further investigated the organization of the newly defined sub-networks using clustering coefficient analysis, which measures whether direct, first degree partners of a particular node interact with each other. This is a common and convenient measure used in graph theory to investigate the local neighborhood of nodes in the graph and it simply measures the amount of clustering within the local neighborhood of a node. The clustering coefficient (C) of a particular node is defined as the ratio of the number of edges that actually exist between the direct partners of that node, to the total number of edges that could possibly

	n*	LC size**
Full	4,684	4,597
Essential	1,180	914 (p< 1.4e-09)
Non-Essential	3,504	*2,967* (p< 1.8e-02)
No-Phenotype	1,855	*981* (p< 7.4e-02)
Damage-Recovery	1,415	851 (p< 8.9e-03)
MMS	1,100	639 (p< 1.4e-04)
4NQO	672	343 (p< 1.4e-05)
UV	230	31 (p< 3.8e-05)
t-BuOOH	160	8 (p< 0.3)

Table 4.1. Large connected component (LC) size in newly defined sub-networks. *n is the total number of proteins in a given category. **The large component (LC) size is the number of connected proteins. Blue (red italic) font indicates a positive (negative) deviation from the random average.

exist among these partners (Figure 4-1A). We considered that clustering coefficient analysis might help to identify protein complexes, signalling and biochemical pathways in the newly defined sub-networks. The clustering coefficient was computed for every node in each of the phenotypic sub-networks. The average clustering coefficient per node (C_{avg}), the percent of nodes with a non-zero clustering coefficient ($C > 0$) and the percent of isolated nodes (i.e. proteins that have no direct partners) was also computed for each sub-network and compared to the corresponding values for randomized sub-networks (Figure 4-1B).

It was previously observed that the average clustering coefficient of the full yeast network is significantly higher than that of a corresponding random network, i.e. a network containing the same number of nodes and edges but where edges connect nodes at random [134]. This observation indicates that protein interaction networks have an overall tendency to form clusters or groups; this translates biologically into biochemical pathways, signalling pathways and protein complexes [117]. The tendency to form protein clusters is significantly over-emphasized in the essential and damage-recovery networks (Figure 4-1B). The essential, MMS-recovery, and 4NQO-recovery networks are around five times more clustered than what would be expected from a random sampling of nodes from the full yeast network. Moreover, clustering in the UV-recovery network is more than one order of magnitude higher than expected by a random selection of nodes, while the *t*-BuOOH-recovery network is more than two orders of magnitude more clustered than the corresponding randomized network. The p values for the enrichment of protein clusters in these phenotypically derived sub-networks range from $< 10^{-10}$ to $< 10^{-109}$. In contrast, the no-phenotype network is actually less clustered than the corresponding randomized network ($p < 0.023$).

The average clustering coefficient (C_{avg}) results suggest that the phenotypic effects of the proteins in these sub-networks is governed by denser than normal interconnected biochemical pathways, signalling pathways and protein complexes. This notion is supported by a finer look at the distribution of the clustering coefficients (Figure 4-1B and 4-1C). The percentage of nodes having a non-zero clustering coefficient indicates that a certain degree of local clustering is taking place. This percentage is significantly higher for the essential and damage-recovery sub-networks, and significantly smaller for the no-phenotype sub-networks, compared to their randomized counterparts (p values range from $< 10^{-3}$ to

	C_{avg}	C> 0
Full	0.0846	36.3%
Essential	0.1879 (p<2.5e-60)	44.2% (p<3.9e-24)
Metabolic	0.0276 (p<0.0015)	3.5% (p<0.029)
Non-Essential	0.0529	20.7%
No-Phenotype	*0.0227* (p<2.5e-5)	*5.9%* (p<1.1e-6)
Metabolic (N-E)	0.0240 (p<7.0e-5)	3.3% (p<2.7e-4)
Damage-Recovery	0.0584 (p<1.9e-6)	16.3% (p<2.4e-6)
MMS	0.0569 (p<2.1e-6)	14.5% (p<1.2e-5)
4NQO	0.0572 (p<1.9e-7)	13.7% (p<1.5e-8)
UV	0.0625 (p<2.1e-14)	10.9% (p<2.7e-19)
t-BuOOH	0.0656 (p<5.1e-26)	8.1% (p<8.6e-22)

Table 4.2. Significance of the clustering coefficient analysis using biased randomizations. Blue (red italic) font indicates a positive (negative) deviation from the random average. The metabolic network will be discussed in the next chapter.

$< 10^{-104}$). The corollary of this is that the number of isolated nodes in the essential and damage-recovery sub-networks is significantly smaller than that expected from a random selection (p values range from < 0.0024 to 10^{-11}), while the no-phenotype sub-network actually has a larger proportion of isolated nodes ($p < 0.05$). In summary, these results indicate that sub-networks corresponding to essential and damage-recovery phenotypes are much more clustered and organized into groups than what would be expected based on a random selection of nodes from the full network, while the no-phenotype network exhibits the opposite trend.

We have further investigated the extent to which the high degree of clustering could be a simple consequence of the high average node degree observed in the essential and damage-recovery networks by carrying out a new set of biased randomizations that preserve the proportion of high degree nodes in each network. Specifically, we define a high degree node as a node having more than 10 direct interactions. For each network tested, we performed a set of 1,000 biased randomizations constrained to have the same proportion of high degree nodes as the network to be tested. Specifically, biased randomizations were carried out by dividing all the yeast proteins into two subsets: high degree nodes (nodes with a degree higher than 10) and low degree nodes (nodes with a degree between 0 and 10). The number of high and low degree nodes was then determined for each category and 1,000 biased randomized networks were obtained for each category based on 1,000 independent experiments consisting of randomly selecting nodes from the full yeast high

(low) degree nodes subset using a binomial distribution with mean equal to the ratio of the total number of high (low) degree nodes in that category over the total number of high (low) degree nodes in the full yeast network. The non-essential network was used for the non-essential categories as described above. The corresponding p-values were computed as described earlier. The results are shown in Table 4.2 and indicate that even when the randomized networks are constrained to have the same average degree as the tested network, the average clustering coefficient as well as the proportion of nodes with a non-zero clustering coefficient are significantly higher for the damage-recovery and essential networks than for the randomized networks, indicating that the degree of local clustering is not a consequence of the presence of higher degree nodes but a reflection of another underlying phenomenon related to protein complexes and dense signalling pathways.

■ 4.5 Identification of Toxicologically Important Protein Complexes and Signalling Pathways

Clustering coefficient analysis of the newly defined damage-recovery sub-networks (for MMS, 4NQO, UV, and *t*-BuOOH) appears to provide an analytical method to identify damage-recovery protein complexes, pathways and signalling information. It should be stressed that clustering coefficient analysis identified local neighbor interactions that we in turn have used to build higher order complexes *in silico* using the program Cytoscape (www.cytoscape.org) [115]. Figure 4-3(A-D) contains every protein in each of the damage-recovery sub-networks that has a non-zero clustering coefficient ($C > 0$); these proteins are displayed with their corresponding protein-protein interactions (edges) derived from each sub-network; in cases where the protein-protein interaction has previously been characterized in a biological context (i.e., not just as part of a high throughput screen [40]) the edge is shown in bold. All of the proteins in Figure 4-3(A-D) are shown with gene names in Figures 4-4, 4-5, 4-6, and 4-7. Tables 4.3 and 4.4 identify the protein complexes and signaling pathways derived by clustering analysis of the phenotypic networks.

From the bold edges in Figure 4-3 it is clear that previously recognized complexes, pathways and signalling modules are represented, even though many of these were not previously recognized as being important for cellular recovery after exposure to a DNA damaging agent. (Note that all of the protein nodes shown in Figure 4-3 play a role in damage-recovery, but now they are colored to represent their cellular function [40].) As expected, groups of proteins that participate in coordinated DNA damage responses (i.e. nucleotide excision repair, mismatch repair, DNA damage checkpoints) were identified among the clustered proteins in the damage recovery sub-networks (Figure 4-3E). However, groups of proteins involved in transcription regulation and chromatin remodeling are also well represented, and these include components of the SAGA regulatory complex (MMS^S, $4NQO^S$ and $t\text{-}BuOOH^S$), RNA polymerase II complex (MMS^S, $4NQO^S$, and UV^S) and SWI/SNF complex ($4NQO^S$). In addition signal transduction is represented for MMS and UV damage-recovery with the cyclin dependent kinase that phosphorylates the C-terminal domain of RNA polymerase II, and which contains CTK1, CTK2 and CTK3 as subunits [40]. Unexpectedly, protein complexes and pathways involving the nuclear pore complex (NUP proteins), RNA metabolism (MUD, STO1, PUB and TIF proteins), vacuolar function and targeting (VMA proteins) are also represented upon visualizing the proteins with non-zero clustering coefficients (i.e., $C > 0$) from the MMS-, 4NQO-, UV- and *t*-BuOOH-recovery sub-networks (Figure 4-3E). In addition, it is clear from Figure 4-3 that many other known and putative complexes await further investigation into their role in cellular recovery after exposure to the so-called

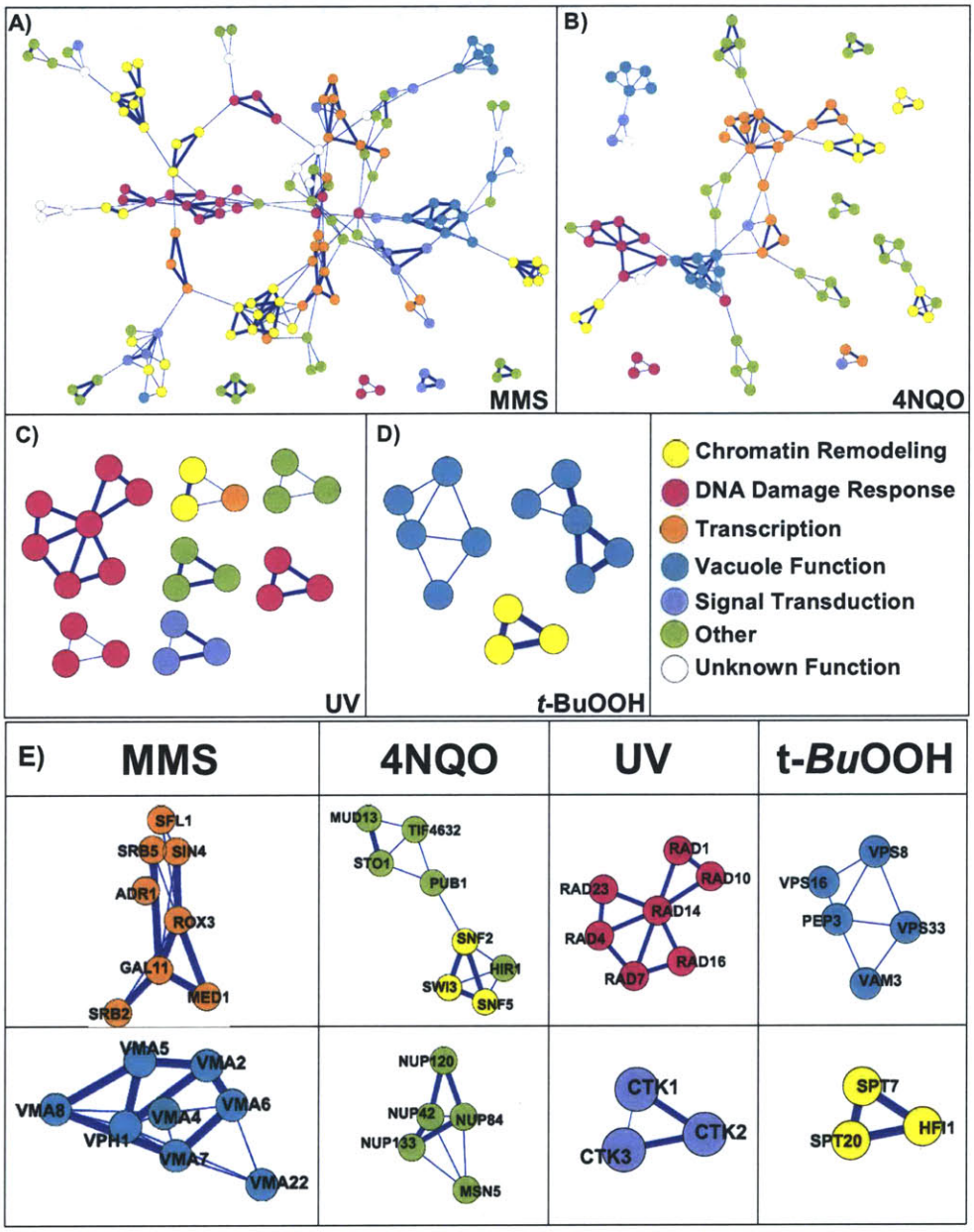


Figure 4-3. Phenotypic protein networks with $C > 0$. Thick blue lines represent previously reported protein complexes. E. Selected complexes identified using clustering coefficient analysis. From top to bottom: MMS (RNA polymerase II holoenzyme/mediator complex and vacuolar H-ATP assembly complex); 4NQO (SWI/SNF complex and Nuclear pore complex); UV (Nucleotide excision repair pathway and C-terminal domain kinase I complex); and *t*-BuOOH (putative vacuolar sorting sub-network and SAGA transcriptional regulatory complex).

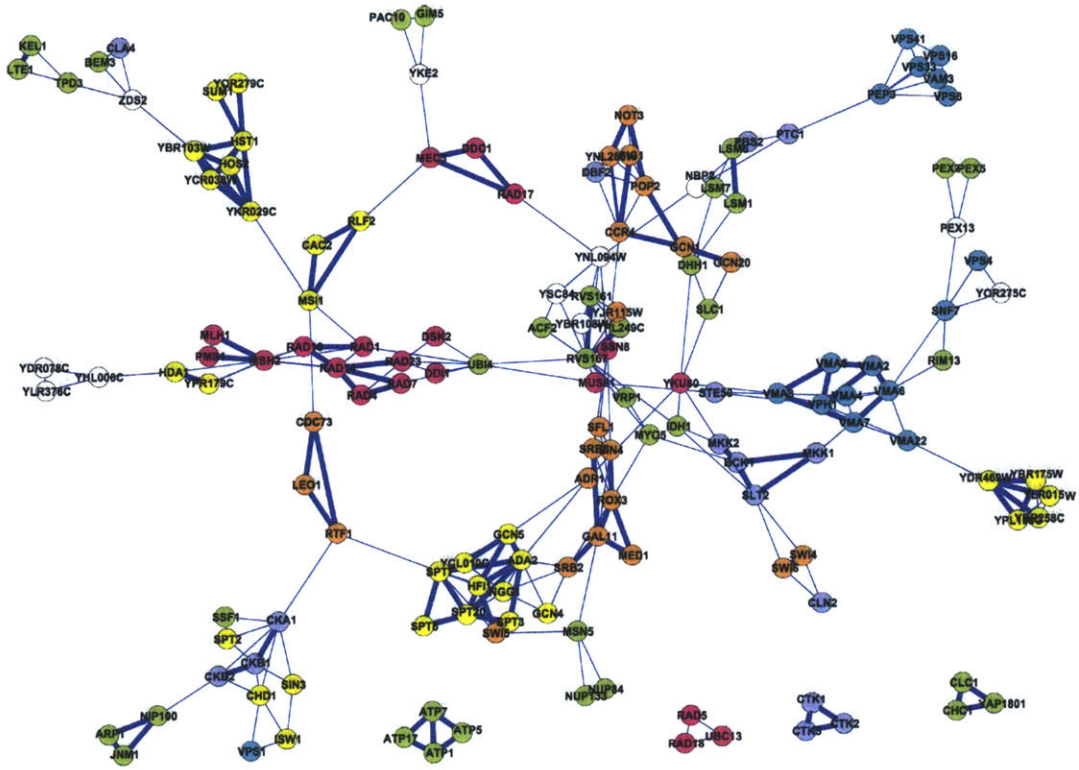


Figure 4-4. MMS protein network with $C > 0$.

DNA damaging agents, roles that may involve protein complexes, biochemical pathways or signalling modules.

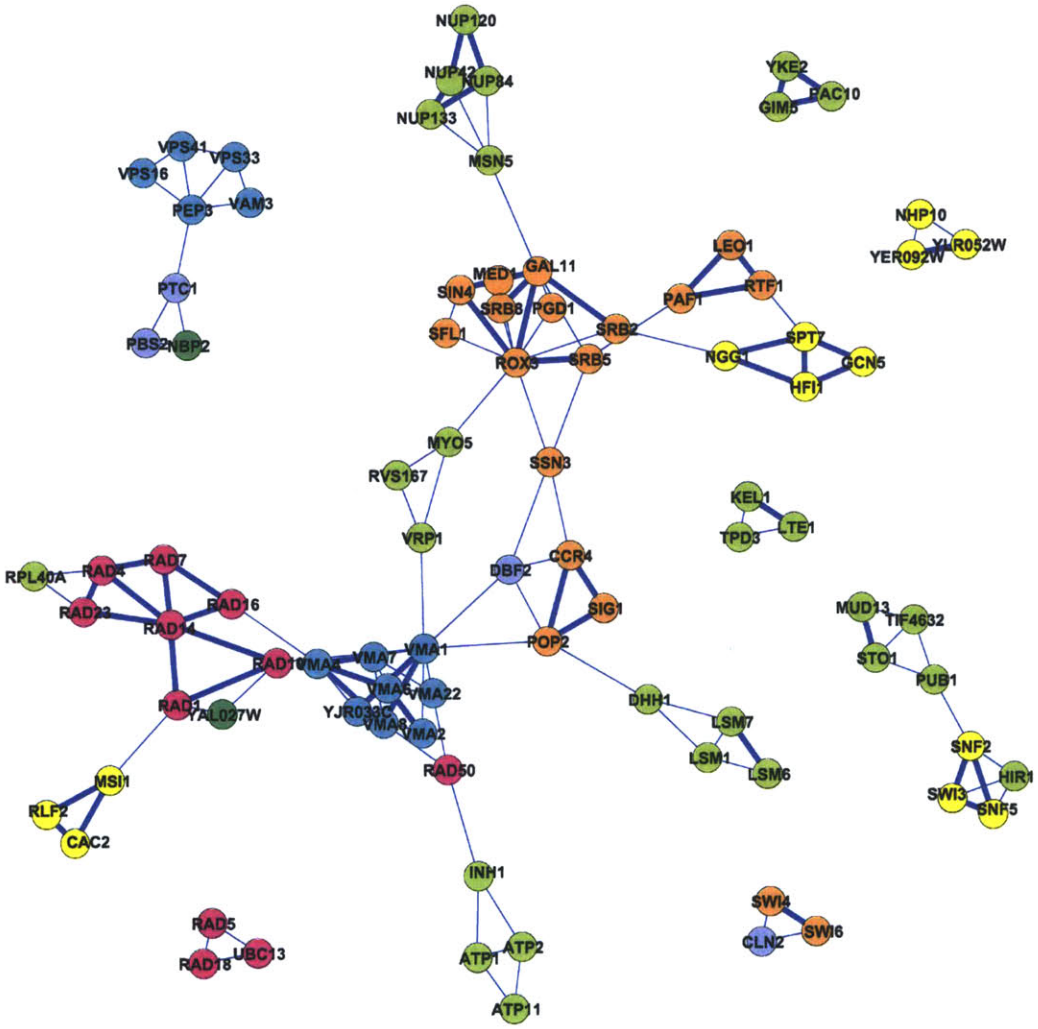


Figure 4-5. 4NQO protein network with $C > 0$.

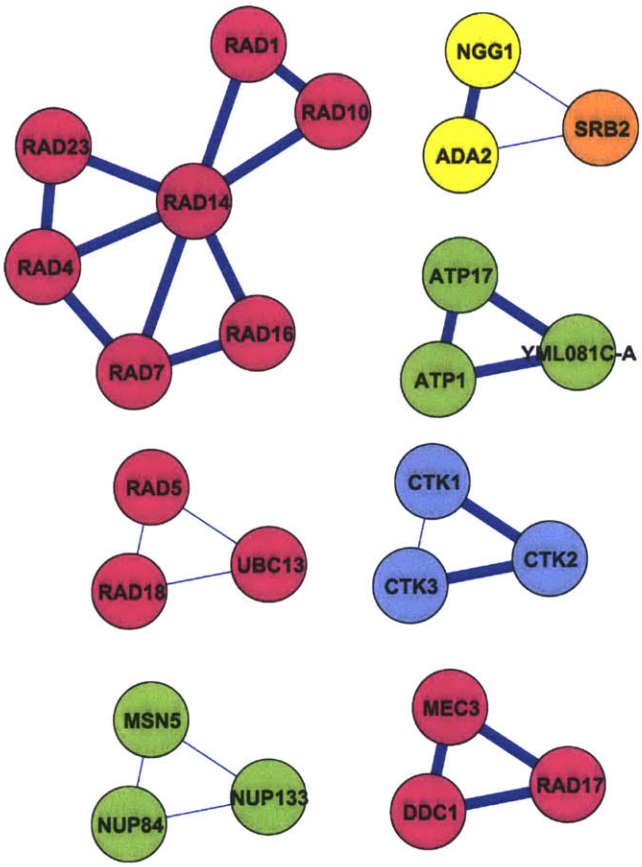


Figure 4-6. UV protein network with $C > 0$.

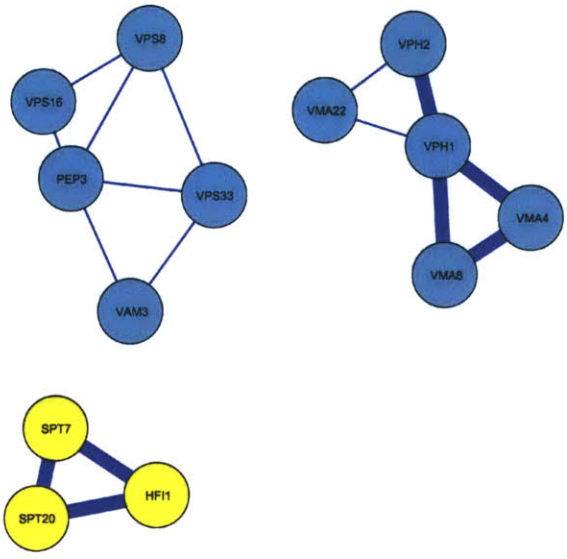


Figure 4-7. t-BuOOH protein network with $C > 0$.

Agent	Complex	Proteins
MMS	DNA clamp complex	Mec3, Ddc1, Rad17
MMS	Prefoldin complex	Gim5, Pac10, Yke2
MMS	SAGA transcriptional regulatory complex	Spt7, Spt8, Spt20, Spt3, Hfi1, Ycl010c, Ngg1, Ada2, Gcn5,
MMS	COMPASS complex	Ydr469w, Ybr75w, Ylro15w, Ybr258c, Ypl138c
MMS	Clatherin complex	Clc1, Crc1, Yap1801
MMS	C-terminal domain kinase I complex	Ctk1, Ctk2, Ctk3
MMS	Nuclelide excision repair	Rad1, Rad14, Rad4, Rad7, Rad23, Rad10,
MMS	Mismatch repair	Msh2, Pms1, Mlh1
MMS	Chromatin assembly complex	Msi1, Cac2, Rlf2
MMS	ATP synthase complex	Atp1, Atp5, Atp7, Atp17
MMS	dynactin complex	Nip100, Arp1, Jnm1
MMS	small nuclear ribonucleprotein complex	Lsm6, Lsm7
MMS	GTP binding complex	Gtr1, Gtr2,
MMS	Nucelar pore complex	Nup133, Nup84
MMS	BAR adaptor proteins	Rvs161, Rvs167, Ypl249c
MMS		Kel1, Lte1,
MMS	Histone deactylase complex	Hda1, Ypr179c
MMS		Gcn1, Gcn20
MMS	RNA polymerase II holoenzyme/mediator complex	Srb5, Sin4, Rox3, Gal11, Med1, Srb2,
MMS	Transcription elongation factor complex	Cdc73, Leo1, Rtf1
MMS	Transcription cofactor complex	Swi4, Swi6
MMS	Map kinase pathway	Bck1, Slt2, Mkk1, Mkk2
MMS	Histoine deacteylase complex	Hst1, Yor279c, Ybr103w, Hos2, Ycr033w, Ykr029c, Sum1
MMS	Protein kinase CK2	Cka1, Ckb1, Ckb2
MMS	CCR4-NOT transcription factor complex	Ccr4, Pop2, Not3, Sig1, Ynl288w
MMS	vacuolar H-ATP assembly complex	Vma2, Vma4, Vma5, Vma6, Vma7, Vma8, Vph1
t-BuOOH	SAGA transcriptional regulatory complex	Spt7, Spt20, Hfi1
t-BuOOH	vacuolar H-ATPasem complex	Vma4, Vma8, Vph1, Vph2

Table 4.3. Protein complexes and signaling pathways identified by clustering coefficient analysis of the MMS and *t*-BuOOH networks.

Agent	Complex	Proteins
UV	SAGA transcriptional regulatory complex	Ngg1, Ada2
UV	C-terminal domain kinase I complex	Ctk1, Ctk2, Ctk3
UV	Nucletide excision repair	Rad1, Rad14, Rad4, Rad7, Rad23, Rad10, Rad16
UV	ATP synthase complex	Atp1, Atp17, Yml081c-a
UV	Nucelar pore complex	Nup133, Nup84
4NQO	Prefoldin complex	Gim5, Pac10, Yke2
4NQO	SAGA transcriptional regulatory complex	Spt7, Hfi1, Ycl010c, Ngg1, Gcn5,
4NQO	SWI/SNF complex	Snf2, Snf5, Swi3
4NQO	Nucletide excision repair	Rad1, Rad14, Rad4, Rad7, Rad23, Rad10, Rad16
4NQO	Chromatin assembly complex	Msi1, Cac2, Rlf2
4NQO	GARP complex	Luv1, Yjl029c
4NQO	ATP synthase complex	Atp1, Atp2, Atp11, Inh1
4NQO	small nuclear ribonucleprotein complex	Lsm6, Lsm7
4NQO	nuclear cap-binding protein complex	Sto1, Mud13
4NQO	Nucelar pore complex	Nup120, Nup42, Nup133, Nup84
4NQO		Kel1, Lte1,
4NQO	RNA polymerase II holoenzyme/mediator complex	Srb2, Srb5, Srb8, Sin4, Rox3, Gal11, Med1
4NQO	Transcription elongation factor complex	Leo1, Rtf1, Paf1
4NQO	Transcription cofactor complex	Swi4, Swi6
4NQO	CCR4-NOT transcription factor complex	Ccr4, Sig1, Pop2
4NQO	INO80 chromatin remodeling complex	Ylr052w, Yer092w
4NQO	vacuolar H-ATPasem complex	Vma1, Vma2, Vma4, Vma5, Vma6, Vma7, Vma8, Vph1, Yjr033c

Table 4.4. Protein complexes and signaling pathways identified by clustering coefficient analysis of the 4NQO and UV networks.

■ 4.6 Towards a Systems Level Understanding of Vital Cellular Processes

In this and the previous chapter, we have presented a systematic investigation of global protein networks in a phenotypic context. Recovery from exposure to DNA damaging agents was chosen because of the wide range of cellular activities required to prevent cell death and because of the association of many damage-recovery pathways with human diseases such as cancer, aging, and other degenerative diseases. Our findings suggest that damage-recovery proteins have attributes somewhat similar to essential proteins. It is striking that all measures reported here (degree distribution, shortest-path distribution, characteristic path length, global centrality distribution, connected component analysis, and clustering coefficients) lead us to the same conclusion. Specifically, damage-recovery proteins have greater direct interactions, smaller shortest-path, are more connected, and are significantly more clustered than the average yeast protein, suggesting that there exists a higher-order organization for these damage-recovery networks. These results reflect two underlying phenomena: the damage-recovery proteins have more hubs allowing them to be more connected and to exhibit shorter path lengths, and the network composed of these proteins is more clustered indicating the existence of many protein complexes and dense signalling pathways. We were also able to identify, using global measures, targeted pathways and complexes essential for damage recovery. The integration of global genomic-scale computational techniques with targeted, experimentally established genome-wide phenotypic data, and the use of such integrated information to uncover biological network information, is one of the major contributions of this study.

The exact biological advantage associated with connectivity characteristics remains to be determined, but we would hypothesize that increased connectivity could allow for more efficient control of cellular processes. Efficiency could manifest itself in the form of redundant signalling events, faster response times and greater flexibility in responses; all of which could provide a biological advantage. Molecular interactions require conserved molecular interfaces. We suppose that proteins involved in vital cellular processes (i.e. essential and damage-recovery) may have "older" molecular interfaces than random proteins, due to the essentiality of their function. These older interface points represent the basic blue print for motifs employed by younger proteins and the high number of molecular interactions displayed by "older" proteins could be due to cross reactivity with new evolved partners. In fact, this view would be supported by recent efforts to model biological networks degree distributions using graph generation techniques. In the proposed network evolution models, high degree nodes are the ones that were added early in the history of the graph [126] [128].

Furthermore, damage-recovery pathways may be highly conserved across evolution (certainly this is true for DNA repair pathways) and as a result, it is expected that the pathway characteristics unraveled for *S. cerevisiae* will parallel those in higher organisms. Using homology mapping, siRNA-mediated genomic phenotyping, and global network information in higher organisms, we aim to explore the biological networks important for protecting mammalian cells against the cytotoxic effects of environmental damaging agents.

Chapter 5

Topological Organization of Functional Yeast Networks

So far, the main results have demonstrated that damage-recovery proteins have greater direct interactions, smaller shortest-path, are more connected, and are significantly more clustered than the average yeast protein. These properties are shared with the essential proteins while proteins not important for damage recovery show opposite trends. One possible explanation to these findings is that the no-phenotype network is composed of a wide selection of unrelated proteins that are part of different pathways and therefore it does not constitute a coherent functional biological network. One would then anticipate the topological network properties observed for the no-phenotype network and one would further expect that all functional cellular networks have the same kind of quantitative characteristics as the ones found for the damage-recovery and essential networks. In this chapter, we investigate the topological organization of a number of different functional yeast networks and compare them to the network properties of damage-recovery proteins described in the previous chapter. Based on the results, we propose new ways to combine expression profiling data with network information to better predict phenotypic outcome. We also expand the network analyses presented in the previous chapters by proposing additional measures for analyzing phenotypic data.

■ 5.1 Functional Yeast Networks

Functional networks are networks of proteins and/or genes that are involved in coordinating a common function such as damage-recovery. There has been a number of recent genome-wide studies that uncovered protein and gene networks involved in different functions. We restrict our study in this chapter to protein networks involved in metabolism, sporulation, the environmental stress response, and the damage recovery transcriptional profiling response.

■ 5.1.1 Metabolism

The metabolic network is composed of proteins, genes, and molecules involved in biochemical processes that sustain the living cell. These include the breaking down of complex substances into simple ones leading to the release of energy as well as the storing of energy by building up complex substances from simple ones. A genome-scale reconstruction of the metabolic network of *S. cerevisiae* was recently synthesized by Forster *et al* [48]. The network is based on currently available genomic, biochemical, and physiological information collected from the genome annotation, biochemical pathway databases, biochemistry textbooks, and recent publications. The network contains 708 structural open reading frames of which 508 are present in the yeast protein-protein interactome used in the previous chapter. Out of the 508 proteins, 395 are non-essential proteins.

■ 5.1.2 Sporulation

Yeast reproduction occurs mainly through vegetative growth, i.e. by mitotic division which is initiated when cells reach a critical cell size and are stimulated by exogenous or endogenous signals. However, when nutrient supplies are low, sexual reproduction is an alternative way for yeast to reproduce. Specifically, cells of opposite mating type (a and α) fuse to form a diploid a/α cell. The diploid cell can then be induced by external signals (starvation) to enter meiosis and undergo sporulation. Meiosis involves two successive nuclear divisions with only one round of DNA replication producing four haploid daughter cells (spores) from the initial diploid cell. It is a specialized process that requires the expression of many specialized

genes. A number of genetic screens have been carried out to identify the various contributors to the meiotic pathway in yeast. A functional screen has been recently described by Enyenihi and Saunders [42] where they have used diploid deletion mutants (constructed by mating independently mutated haploid strains) to screen for genes required for meiotic division and sporulation using differential interference contrast microscopic inspection for the formation of visible spores following growth in liquid population medium. Out of the 4,323 single-gene deletion mutants in non-essential genes, 334 were found to be sporulation-essential genes. This phenotypic analysis was further compared to the sporulation expression data of Chu *et al* [35]. The results show that there is little total overlap between the genes essential for sporulation and those showing the most increase in expression.

■ 5.1.3 Environmental stress response (ESR)

Gash *et al* [53] have recently studied genomic expression profiles in yeast responding to different environmental changes. Specifically, they used DNA microarrays to measure changes in transcription levels over time for every yeast gene as cells responded to temperature shocks, hydrogen peroxide, the superoxide-generating drug menadione, the sulfhydryl-oxidizing agent diamide, the disulfide-reducing agent dithirthreitol, hyper- and hypo-osmotic shock, amino acid starvation, nitrogen source depletion, and progression into stationary phase. Approximately 900 genes were found to respond to all stressful conditions tested. These genes define the Environmental Stress Response (ESR).

■ 5.1.4 Damage-recovery transcription response (DRR)

Damage-recovery transcriptional profiling was also performed by Samson and coworkers [72] where the transcriptional response of *S. cerevisiae* after exposure to methyl methanesulfonate (MMS), 4-nitroquinoline-N-oxide (4NQO) and *tert*-butyl hydroperoxide (*t*-BuOOH) was measured. 1,648 genes were found to be three-fold differentially expressed in response to at least one of these three damaging agents. Out of these 1,648 genes, 1,239 are present in the yeast protein-protein interactome used the previous chapter. 938 of the 1,239 genes are non-essential. We define this set of genes as the Damage-Recovery transcription Response (DRR). In addition, 229 genes were two-fold differentially expressed in response to all three agents. Out of these 229 genes, 167 are found in the protein interactome of which 146 are non-essential. We define this set of genes as the common Damage-Recovery Response (cDRR).

■ 5.2 Network Properties of Functional Yeast Networks

We have computed the network measures defined in the previous chapters: average degree, characteristic path length, largest component size, clustering coefficient average, proportion of non-zero clustering coefficients ($C > 0$), and the number of isolated nodes for all four functional yeast networks defined in the previous section. In addition, for the metabolic, ESR, DRR, and cDRR networks, we computed the measures for non-essential versions of these networks. The results are shown in Table 5.1.

■ 5.2.1 Regulatory versus energy generation networks

Table 5.2 shows all network measures computed for the phenotypic networks defined in the previous chapters as well as for the full metabolic network and for the metabolic network composed of only non-essential proteins. The results indicate that the metabolic network

	n	Average degree	Char. Path length	LC size	Cluster avg	C> 0	Isolated Nodes
Metabolic	508	*4.6024* (p<3.66e-4)	4.2888 (p<0.31)	*93* (p<0.98)	0.0276 (p<0.14)	3.5% (p<0.79)	64.8% (p<0.66)
Metabolic (N-E)	395	*4.6127* (p<0.28)	*4.3158* (p<0.81)	*21* (p<0.84)	0.0240 (p<4.99e-4)	3.3% (p<1.1e-3)	*70.4%* (p<0.39)
ESR	658	7.0365	4.0504	316	0.1054	22.6%	43.6%
ESR (N-E)	416	5.1779 (p<0.80)	*4.2785* (p<0.35)	74 (p<1.7e-2)	0.0155 (p<8.1e-2)	4.6% (p<1.35e-5)	*67.8%* (p<0.22)
DRR	1,239	6.3051 (p<0.64)	*4.1290* (p<2.01e-4)	668 (p<0.16)	0.0532 (p<0.027)	15.0% (p<0.002)	*38.3%* (p<0.42)
DRR (N-E)	938	5.1290 (p<0.79)	*4.2516* (p<5.6e-3)	*335* (p<0.15)	*0.0152* (p<0.41)	*4.4%* (p<0.83)	51.2% (p<0.38)
cDRR	167	*5.0120* (p<0.17)	4.3096 (p<0.5)	13 (p<0.12)	*0* (p<0.64)	*0* (p<0.63)	*81.4%* (p<0.67)
cDRR (N-E)	146	*4.5205* (p<0.46)	4.3650 (p<0.74)	*3* (p<0.53)	*0* (p<0.80)	*0* (p<0.80)	89.0% (p<0.72)
Sporulation (N-E)	375	7.1707 (p<5.82e-7)	*4.1239* (p<1.77e-4)	99 (p<1.13e-7)	0.0568 (p<6.92e-23)	10.4% (p<7.79e-34)	*51.2%* (p<4.23e-9)

Table 5.1. Network measures for a number of functional yeast networks. n is the total number of proteins in a given category. (N-E) corresponds to non-essential versions of the networks. The large connected component (LC) size is the number of proteins in the LC. Blue font indicates a positive deviation from the random average and red italic font indicates a negative deviation from the random average.

	n	Average degree	Char. Path length	LC size	C_{avg}	C> 0	Isolated Nodes
Full	4,684	6.1883	4.2383	4,597	0.0846	36.3%	0
Essential	1,180	9.5093 (p<9.4e-37)	*3.9546* (p<6.0e-22)	914 (p<1.4e-9)	0.1879 (p<2.0e-110)	44.2% (p<1.3e-105)	*19.9%* (p<1.5e-12)
Non-Essential	3,504	*5.0699* (p<1.5e-35)	4.3305 (p<9.4e-22)	*2,967* (p<1.8e-2)	*0.0529* (p<2.5e-13)	*20.7%* (p<2.0e-22)	11.0% (p<1.5e-2)
Metabolic	508	*4.6024* (p<4e-4)	4.2888 (p<0.3)	*93* (p<0.98)	0.0276 (p<0.15)	3.5% (p<0.8)	64.8% (p<0.68)
No-Phenotype	1,855	*4.3919* (p<1.1e-6)	4.4128 (p<3.7e-6)	*981* (p<7.4e-2)	*0.0227* (p<2.3e-2)	*5.9%* (p<3.6e-4)	36.4% (p<5.0e-2)
Damage-Recovery	1,415	6.0283 (p<1.3e-7)	*4.2228* (p<6.8e-6)	851 (p<8.9e-3)	0.0584 (p<1.9e-11)	16.3% (p<1.1e-18)	*34.5%* (p<2.4e-2)
Metabolic (non-essential)	395	*4.6127* (p<0.3)	*4.3158* (p<0.8)	*21* (p<0.85)	0.0240 (p<6e-4)	3.3% (p<1.2e-3)	*70.4%* (p<0.4)

Table 5.2. Network measures for the metabolic network as compared to the networks defined in the previous chapter. Blue (red italic) font indicates that the measured value is greater (smaller) than the one obtained using the randomized networks.

exhibits properties more similar to the randomized (and no-phenotype) networks than to the essential and damage-recovery networks. Since the metabolic network contains modules and complexes (as is confirmed by the more significant C_{avg} and $C > 0$ measures), this result indeed contradicts the supposition that all cellular networks must share similar quantitative features. For instance, as can be seen in Table 5.2, the metabolic network has a lower average degree than the corresponding randomized networks similar to the no-phenotype network. This contrasts with the essential and damage-recovery networks which have a higher average degree than the corresponding randomized networks.

We can speculate as to why the individual network metrics are so different for the metabolic network and how the damage-recovery network metrics can provide a biological advantage for achieving an effective damage-recovery response. One possibility is that since small diffusible metabolites play an important role in metabolism, they may be the crucial components for keeping the network coherent instead of protein-protein interactions that seem to be more important in signaling pathways. This contrast would be consistent with a view that regulatory networks may be more complex in organizational structure than those devoted to core functions such as metabolism and energy generation [81]. The results observed here would then suggest that essential and damage-recovery networks involve modular control elements tightly networked to each other allowing a fast and coordinated response. This may also be related to the fact that the phenotypic readout for both the essential and damage-recovery networks is the same, namely cell death since the damage-recovery proteins are themselves "essential" in the presence of a damaging agent. Preventing such a catastrophic outcome may need tight coordination and the rapid stimulation of multiple cellular processes that display, at the network level, more hubs, shorter path lengths, larger connected component size, and higher clustering coefficients. Consequently, phenotypic network studies of the kind presented here shed new light on the architecture of biological networks and are an important tool for further understanding their functionality.

■ 5.2.2 Expression profiling versus phenotypic analysis

In addition to metabolism, the results in Table 5.1 show that data obtained from expression profiling (ESR, DRR, and cDRR) have, for the most part, network properties similar to the randomized networks (low p values) which is in contrast to the phenotypic damage-recovery networks. These results are surprising, especially since the DRR network uses the same DNA damaging agents (MMS, 4NQO, t-BuOOH) as the ones used in the phenotypic study leading to the damage-recovery network. Proteins that are more than three-fold differentially expressed in response to DNA damaging agents have very different topological properties than proteins that are essential for normal growth after exposure to the same DNA damaging agents. In contrast, the sporulation network has highly significant network properties very similar to the ones observed for damage-recovery. Specifically, the results in Table 5.1 indicate that sporulation-essential proteins have a significantly higher degree, shorter characteristic path length, are more connected, and are more clustered than the average yeast protein. The sporulation network is derived from phenotypic analyses similar to the ones used to derive the damage-recovery networks. These observations show that data obtained from expression profiling do not seem to have any distinctive network properties whereas data obtained from phenotypic analyses have clear distinctive network properties. These results seem to suggest a fundamental difference between the topological properties of networks derived from expression profiling data and networks constructed using phenotypic studies. We explore this difference in the next section.

	n	Average degree	Char. Path length
DRR (expr)	938	5.1290	4.2516
DRR(phe)	1376	6.0560	4.2228
DRR (both)	**348**	**5.6695**	**4.1503**
MMS (expr)	420	5.8095	4.1758
MMS (phe)	1100	6.2764	4.1724
MMS (both)	**131**	**6.9771**	**4.0348**
4NQO(expr)	84	4.9167	4.1454
4NQO (phe)	672	6.7842	4.1573
4NQO (both)	**12**	**6.8333**	**4.3485**
t-BuOOH (expr)	723	5.0069	4.2748
t-BuOOH (phe)	160	6.6188	4.2750
***t*-BuOOH (both)**	**31**	**8.8710**	**4.1634**

Table 5.3. Comparison of the expression and phenotypic damage-recovery data. *expr* corresponds to proteins that are differentially expressed in response to a damaging agent, *phe* corresponds to proteins whose deletion leads to impaired growth upon exposure to a damaging agent and *both* indicates proteins that are both differentially expressed and important for proper growth in response to a damaging agent.

■ 5.3 Predicting Phenotypic Outcome Using Expression Profiling Data

In order to further investigate the difference between the expression profiling data and the phenotypic data, we computed the global network properties (degree and characteristic path length) for the individual MMS, 4NQO, and *t*-BuOOH expression networks and compared them to their phenotypic counterparts. Since we only have phenotypic data for non-essential genes, the expression networks only include non-essential genes. The protein overlap between the expression and phenotypic networks was also computed as well as the global network properties of the overlapping proteins. The results are shown in Table 5.3. Figure 5-1 displays, for each class, the number of proteins that are exclusively differentially expressed (expr only), i.e. are differentially expressed but do not, if deleted, affect growth upon exposure to a damaging agent. The number of proteins that are important for proper growth upon exposure to a damaging agent but are not differentially expressed (phe only) as well as the proteins that are both differentially expressed and important for proper growth upon exposure to a damaging agent (both) are also shown. Figure 5-2 graphs the average degree (computed in Table 5.3) for proteins in the different categories.

■ 5.3.1 Protein overlap between expression profiling and phenotypic data

The results in Figure 5-1 indicate that there is little overlap between differentially expressed damage recovery genes and the ones needed for proper growth. It should be noted that for this analysis the damage-recovery phenotypic category (DRR phe) is the union of the MMS, 4NQO and *t*-BuOOH phenotypic categories and excludes the UV data since we do not have expression profiling data on UV treated cells. Only 37% of the damage-recovery differentially expressed genes show a sensitive growth phenotype upon deletion. For MMS alone this fraction reduces to 31%, for 4NQO it is 14.2% while for *t*-BuOOH it is 4.3%. This is in agreement with previous analysis on a smaller set of genes tested for growth sensitivity [18]. Furthermore, as can be seen from Table 5.3, the individual mutagen network results for the expression profiling data follow the same characteristics as the ones for the collective damage recovery response in that their average degree and characteristic path length is not higher than what would be expected from a random selection of nodes showing

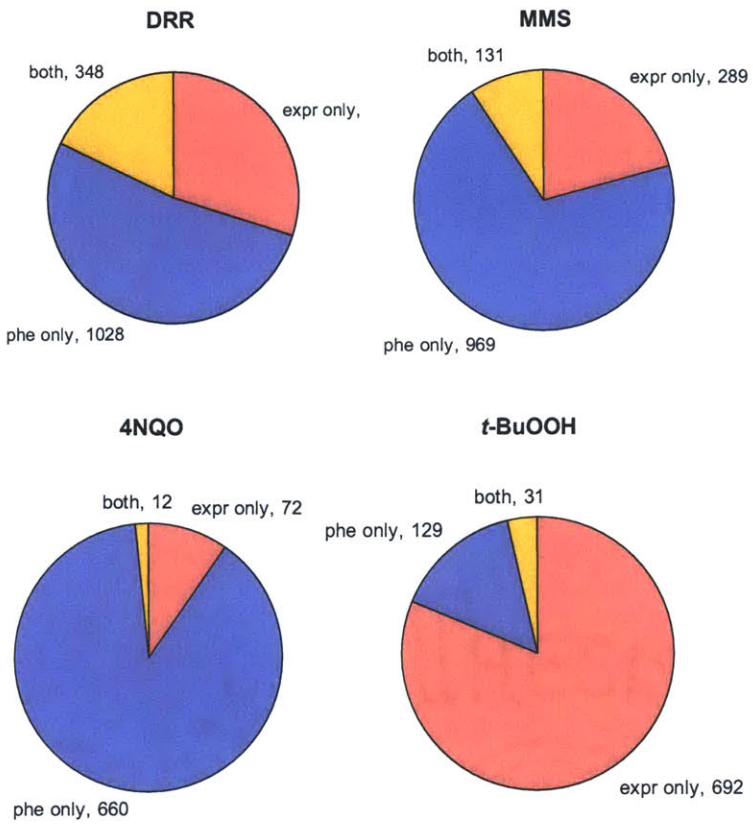


Figure 5-1. Number of proteins important for phenotypic growth only (phe only), proteins showing differential expression only (expr only), and proteins both important for phenotypic growth and showing differential expression (both).

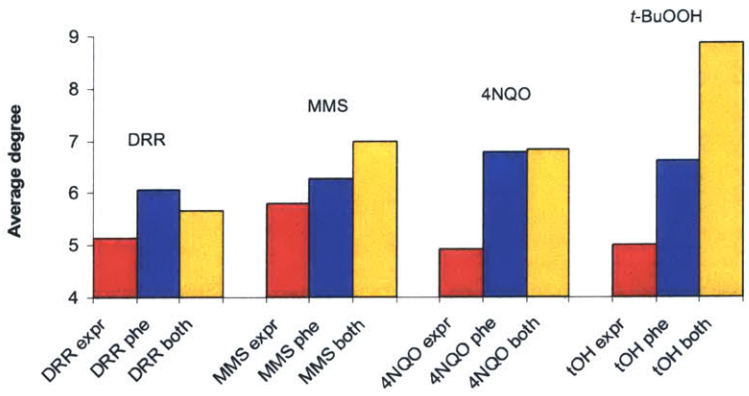


Figure 5-2. Average degree of proteins in different categories.

	DRR	MMS	4NQO	t-BuOOH
$p_x(P\|E,H)/p_x(P\|E)$	1.51	1.52	1.79	2.50
$p_x(P\|E)/p_x(P)$	0.92	0.99	0.75	0.95
$p_x(P\|H)/p_x(P)$	1.43	1.50	1.72	1.30
$p_x(E\|P,H)/p_x(E\|P)$	1.07	1.27	1.35	2.06
$p_x(E\|P)/p_x(E)$	0.92	0.99	0.75	0.95
$p_x(E\|H)/p_x(E)$	1.01	1.26	1.32	1.17

Table 5.4. Ratios quantifying the importance of one measurement (expression (E), phenotype (P), high degree (H)) compared to the other two.

very different topological behavior from their phenotypic counterparts. However, if the expression profiling data is intersected with the phenotypic data (the 'both' rows in the table) and network properties of the resulting set of overlapping nodes are investigated, significant topological trends arise. Specifically, the average degree of the differentially expressed genes that show phenotypic behavior is, for all networks, much higher than the one for all differentially expressed genes and this time the result is statistically significant. Furthermore, as can be seen from Figure 5-2, for the individual mutagen networks (MMS, 4NQO, and t-BuOOH), the average degree of the intersection of differentially expressed and phenotypic genes is higher than the one obtained only based on phenotypic data. These trends are also observed for the characteristic path length shown in Table 5.3 except for 4NQO possibly due to the very small number of overlapping genes.

■ 5.3.2 Using network information to better predict phenotype from expression profiling data

The observations discussed in the previous subsection suggest a way to combine expression profiling data with network information to better predict phenotypic outcome. In order to test the prospects of such a method, we have estimated the probability that a protein exhibits a growth phenotype given that it is differentially expressed and has a high degree. Specifically, let P denote the event that a protein x exhibits a growth phenotype, i.e. if x is deleted, the cell does not grow properly in response to exposure to a damaging agent. E is defined as the event that a protein x is differentially expressed and H the event that it has a high degree where high is defined as having more than 14 direct interactions. Then, $p_x^q(P|E,H)$ defines the probability that protein x belonging to category q (DRR, MMS, 4NQO, or t-BuOOH) exhibits a growth phenotype given that it is differentially expressed and has a high degree and is given by:

$$p_x^q(P|E,H) = \frac{p_x^q(P,E,H)}{p_x^q(E,H)} \approx \frac{N_q(P,E,H)}{N_q(E,H)} \tag{5.1}$$

where $N_q(P,E,H)$ and $N_q(E,H)$ are respectively the number of proteins in category q that are differentially expressed, exhibit a growth phenotype, and have a high degree, and the number of proteins in category q that are both differentially expressed and have a high degree. Table 5.4 shows the change in the probability of a protein x having a growth phenotype or being differentially expressed as one additional measurement (expression, phenotype, or high degree) is obtained. The results indicate that just knowing that a protein is differentially expressed does not increase its probability of showing a growth

	Down-regulation		Up-regulation	
	n	Average degree	n	Average degree
MMS (expr)	70	8.0857	350	5.3543
MMS (phe)	1100	6.2764	1100	6.2764
MMS (both)	**19**	**10.1053**	**112**	**6.4464**
4NQO(expr)	22	4.4091	62	5.0968
4NQO (phe)	672	6.7842	672	6.7842
4NQO (both)	**3**	**9.3333**	**9**	**6.0**
t-BuOOH (expr)	250	6.1120	473	4.4228
t-BuOOH (phe)	160	6.6188	160	6.6188
***t*-BuOOH (both)**	**11**	**13.4545**	**20**	**6.3500**

Table 5.5. Difference between down and up regulated genes.

phenotype and vice versa. However, a differentially expressed protein that has a high degree is more than one and a half times more likely to show a growth phenotype than a random differentially expressed protein. This trend is not as marked in the reverse direction, i.e. knowing that a protein exhibiting a growth phenotype is highly connected only shows a slight increase in the likelihood that it is differentially expressed compared to a random protein showing a growth phenotype. The results in Table 5.4 also confirm the earlier observation that a high degree node is as equally likely to be differentially expressed as a random protein however this is not true for a protein exhibiting phenotypic growth where a high degree node is more likely to be required for proper growth upon exposure to a damaging agent than a random protein.

■ 5.3.3 Up-regulated versus down-regulated genes

So far we have considered proteins whose corresponding genes are differentially expressed as one homogeneous category. This category includes proteins whose transcription is down-regulated in response to exposure to a damaging agent as well as proteins whose transcription is up-regulated. The proteins exhibiting a growth phenotype however correspond to proteins whose deletion yields to impaired growth upon exposure to a damaging agent. One may therefore suspect that proteins exhibiting a growth phenotype should correlate better with proteins whose expression is down-regulated in response to exposure damaging agents since gene deletion effectively corresponds to eliminating gene expression which is the goal of down-regulation. In order to investigate this hypothesis, we computed the number of proteins in each category that are down-regulated and up-regulated as well as the number of proteins that are both down-regulated or up-regulated and show a growth phenotype. For each category, we also computed the average degree of the proteins in that category. The results are shown in Table 5.5 and indicate that the extent of overlap between proteins that are down- or up-regulated and the ones important for growth is the same for both down-regulated and up-regulated proteins. However, the average degree results are very different for the down-regulated and up-regulated genes. The up-regulated proteins

	MMS	4NQO	t-BuOOH
$p_x(P\|E,H)/p_x(P\|E)$	1.33	7.34	2.27
$p_x(P\|E)/p_x(P)$	0.86	0.71	0.96
$p_x(P\|H)/p_x(P)$	1.50	1.72	1.30
$p_x(E\|P,H)/p_x(E\|P)$	1.94	2.73	2.91
$p_x(E\|P)/p_x(E)$	0.86	0.71	0.96
$p_x(E\|H)/p_x(E)$	2.19	0.64	1.67

Table 5.6. Ratios quantifying the importance of one measurement (expression (E), phenotype (P), high degree (H)) compared to the other two for down-regulated genes.

results are similar to the overall differentially expressed results in that the average degree for up-regulated proteins is not significantly higher than expected by chance while the overlap between the up-regulated proteins and the ones important for phenotypic growth has a significantly higher average degree. The down-regulated proteins however show a surprising trend. While the probability of a protein being important for growth given that it is down-regulated is still the same as the probability of a protein being important for growth given that it is differentially expressed, unlike generally differentially expressed and up-regulated proteins, proteins that are down-regulated have a much higher average degree than expected by chance. This is true for MMS responsive and t-BuOOH responsive proteins however the trend does not hold for 4NQO responsive proteins perhaps due to the small size of this category. Furthermore the overlap between down-regulated proteins and proteins important for proper growth has a much higher average degree than the overlap between differentially expressed proteins and proteins important for proper growth. To further quantify this effect, we computed the likelihood of having one additional measurement (expression, phenotype, or high degree) compared to the other for the down-regulated proteins. The results are shown in Table 5.6 and indicate that while knowing that a protein is down-regulated does not increase the probability of it exhibiting a growth phenotype and vice versa similar to the results obtained using all differentially expressed proteins, knowing that a protein is down-regulated and has a high degree greatly increases its probability of exhibiting a growth phenotype. The opposite also holds, i.e. knowing that a protein has a growth phenotype and has a high degree greatly increases the probability of it being down-regulated. This last observation does not hold for all differentially expressed proteins as was observed in Table 5.4.

■ 5.3.4 Towards a better prediction of phenotype using expression profiling data

The results presented in this section suggest a simple yet powerful way to combine network information with expression profiling data to better predict phenotypic outcome. Specifically, overlaying expression profiling data with network topology and screening for highly connected proteins provides a set of candidate proteins that are more likely to exhibit a growth phenotype than a random differentially expressed protein. Providing methods that better predict phenotype using expression profiling data is important because expression profiling data are readily available and can be easily obtained using non-invasive procedures. This is not the case for phenotypic data where the experiments are more complicated and sometimes impossible to do such as on humans for example. As a result, any method that improves our ability to predict phenotype using expression data is highly valuable since

ultimately one is interested in changes in observable behavior (phenotype) in response to some stimulus. We have presented here a very simple way to combine expression profiling data with network information in order to illustrate the power of the network models we provide in this thesis. Clearly, more sophisticated methods can and should be developed using the different measures presented in this thesis. This further analysis is proposed as a direction for future work.

■ 5.4 Robustness of the Network Results

All network measures so far relied on protein-protein interactions obtained from the Database of Interacting Proteins (DIP) [135] which includes high-throughput genome-wide data such as yeast two hybrid [123] [50] [106] and mass spectrometric analysis of protein complexes [65] [54] as well as interactions collected from small scale screens in hundreds of individual research papers. We performed computations on an exhaustive yeast interactome that includes all reported protein-protein interactions in order to evaluate the network characteristics of all our identified damage-recovery proteins. However, false-positive protein-protein interactions might affect our observed trends so it is important to assess the robustness of the results reported here. We have recomputed all network measures using an additional smaller yeast protein-protein interaction network: the core yeast interactome obtained from DIP as of October 2004 [135] which is a curated interactome using the method outlined by Deane *et al.* [38]. The network includes 6,337 high confidence interactions among 2,628 proteins. Results are shown in Table 5.7 and are in agreement with those obtained using the complete yeast interactome, indicating that the results are robust to false-positive protein-protein interactions.

We have also recomputed the network measures using the recently published Filtered Yeast Interactome (FYI) [63]. Table 5.8 shows the results. As can be seen by comparing Tables 5.7 and 5.8, the results for this network are different than for the core network as well as for the full yeast network. This is most probably due to the fact that this network is biased towards complexes and signaling pathways published by individual groups as opposed to genome-wide studies as can be confirmed from the clustering coefficient average results.

■ 5.5 Damage-Recovery Party and Date Hubs

In their recently published paper, Han *et al.* [63] analyzed the dynamic organization of the yeast protein-protein interaction network. In their study, they defined two types of hubs, i.e. proteins that interact with many partners, party hubs and date hubs. Party hubs are proteins that interact with most of their partners at the same time. In contrast, date hubs interact with their partners at different times or locations. The authors propose an organized modularity of the yeast proteome where modules are connected by the date hubs which represent regulators, mediators, or adaptors. Party hubs exist within modules coordinating the function mediated by these modules therefore functioning at a lower level of the organization of the proteome. The same proportion of party and date hubs was found to be essential.

We have identified the party and date hubs in the damage-recovery networks in order to extend the results of Han *et al.* and unravel the role of damage-recovery proteins in the context of the modular network organization of the yeast proteome. Table 5.9 summarizes the results and Tables 5.10-5.15 give the names and description of the party and date hubs for the MMS, 4NQO, UV, and *t*-BuOOH networks as well as the process they are involved

	n	Average degree	char. path length	Cavg	C>0	Isolated Nodes
FULL CORE	2628	4.8227	5.0052	0.2036	48.48%	1%
Essential	852	7.0352 (p<9.33e-36)	*4.6596* (p<2.13e-15)	0.2708 (p<2.82e-50)	55.63% (p<7.07e-65)	*10.92%* (p<1.57e-21)
Non-Essential	1767	*3.7731* (p<3.06e-35)	5.1519 (p<2.45e-13)	*0.1381* (p<1.6e-2)	*28.30%* (p<2.07e-6)	19.07% (p<1.63e-5)
Damage-Recovery	820	4.4549 (p<5.51e-9)	*5.0836* (p<5.84e-2)	0.1262 (p<5.53e-8)	22.07% (p<8.25e-9)	*32.80%* (p<3.02e-6)
No-Phenotype	842	*3.1758* (p<3.65e-8)	5.2093 (p<1.14e-1)	*0.0431* (p<4.47e-2)	*6.77%* (p<6.8e-3)	48.46% (p<2.75e-2)
MMS	655	4.5573 (p<4.03e-8)	*5.0957* (p<2.28e-1)	0.1245 (p<3.90e-10)	21.22% (p<1.74e-12)	*35.57%* (p<5.67e-8)
4NQO	410	4.9512 (p<6.99e-9)	*4.9523* (p<3.1e-3)	0.1071 (p<2.39e-11)	18.54% (p<1.38e-17)	*40.98%* (p<6.21e-10)
UV	147	4.8776 (p<2.00e-3)	5.2188 (p<4.95e-1)	0.0712 (p<1.77e-11)	11.56% (p<2.16e-20)	*59.86%* (p<1.39e-7)
t-BuOOH	101	5.4851 (p<1.16e-4)	*5.0469* (p<5.08e-1)	0.1271 (p<8.86e-55)	15.84% (p<3.61e-64)	*72.28%* (p<1.80e-3)
Metabolic	232	*2.6897* (p<2.77e-8)	5.4065 (p<3.11e-5)	0.0517 (p<1.35e-2)	6.47% (p<3.58e-2)	*61.64%* (p<4.82e-2)
Metabolic (non-essential)	176	*2.6648* (p<4.68e-4)	5.3968 (p<2.58e-2)	0.059 (p<3.26e-7)	7.39% (p<1.34e-7)	*70.45%* (p<2.73e-2)

Table 5.7. Network measures based on the core yeast interactome.

	n	Average degree	char. path. length	Cavg	C>0	Isolated Nodes
FULL FYI	1379	3.6157	9.4527	0.3336	48.15%	0
Essential	572	4.4231 (p<7.47e-9)	*9.3315* (p<3.95e-1)	0.3654 (p<1.26e-20)	48.78% (p<1.21e-25)	*12.94%* (p<5.42e-13)
Non-essential	799	*3.0513* (p<6.03e-9)	*9.3767* (p<4.72e-1)	*0.2087* (p<1.86e-1)	*27.16%* (p<7.95e-2)	*19.65%* (p<6.28e-1)
Damage-Recovery	403	3.268 (p<7.46e-2)	9.3701 (p<9.91e-1)	0.189 (p<2.72e-5)	21.84% (p<1.86e-4)	*36.97%* -5.00E-03
No-Phenotype	357	*2.9132* (p<3.38e-1)	*9.3003* (p<6.42e-1)	0.094 (p<9.47e-1)	13.73% (p<3.71e-1)	52.94% (p<3.90e-1)
MMS	322	3.2857 (p<1.36e-1)	*9.3086* (p<7.51e-1)	0.172 (p<8.93e-5)	19.88% (p<2.59e-4)	*38.20%* (p<4.65e-5)
4NQO	229	3.786 (p<3.15e-4)	9.6922 (p<2.42e-1)	0.1669 (p<2.32e-6)	18.78% (p<1.08e-5)	*37.99%* (p<2.44e-8)
UV	99	3.2323 (p<6.42e-1)	9.7096 (p<4.33e-1)	0.0606 (p<6.44e-2)	6.06% (p<1.14e-1)	*59.60%* (p<2.72e-4)
t-BuOOH	65	3.3077 (p<5.83e-1)	*8.8968* (p<3.78e-1)	0.1641 (p<4.05e-15)	20% (p<1.53e-18)	*64.62%* (p<1.69e-4)
Metabolic	83	*2.241* (p<2.0e-3)	10.1838 (p<1.24e-1)	0.2071 (p<1.27e-16)	22.89% (p<5.89e-17)	*28.92%* (p<8.92e-18)
Metabolic (non-essential)	60	*2.1833* (p<5.80e-2)	10.3788 (p<9.51e-2)	0.1944 (p<2.53e-20)	21.67% (p<4.79e-22)	*31.67%* (p<1.60e-20)

Table 5.8. Network measures based on the FYI yeast interactome.

Phenotype	total proteins	date hubs	party hubs
Essential	572	55	74
Damage-recovery	403	23	25
MMS	322	21	20
4NQO	229	16	19
UV	99	3	4
t-BuOOH	65	5	7

Table 5.9. Number of total proteins, date hubs, and party hubs in the different phenotypic categories.

MMS date hubs			
ORF		**Description**	**Process**
YBR200W	BEM1	Bud emergence mediator	Cell division
YAL040C	CLN3	Cyclin, G1/S-specific	Cell cycle
YPL031C	PHO85	Cyclin-dependent protein kinase	Cell cycle
YHR152W	SPO12	Sporulation protein	Cell cycle
YGL240W	DOC1	Component of the anaphase promoting complex	Cell cycle
YBL016W	FUS3	Mitogen-activated protein kinase (MAP kinase)	Cell signaling
YMR125W	STO1	Large subunit of the nuclear cap-binding protein complex CBC	mRNA metabolism
YPL254W	HFI1	Transcriptional coactivator	Sm
YDR378C	LSM6	Sm-like (Lsm) protein	Sm
YNL147W	LSM7	Sm-like (Lsm) protein	Sm
YBL093C	ROX3	Transcription factor	Transcription
YOL004W	SIN3	Transcription regulatory protein	Transcription
YDR392W	SPT3	General transcriptional adaptor or co-activator	Transcription
YBR081C	SPT7	Involved in alteration of transcription start site selection	Transcription
YNL236W	SIN4	Global regulator protein	Transcription
YOL148C	SPT20	Member of the TBP class of SPT proteins that alter transcription site selection	Transcription
YGR252W	GCN5	Histone acetyltransferase	Transcription
YBR289W	SNF5	Component of SWI/SNF transcription activator complex	Transcription
YLR055C	SPT8	Transcriptional adaptor or co-activator	Transcription
YDR448W	ADA2	General transcriptional adaptor or co-activator	Transcription
YDR176W	NGG1	General transcriptional adaptor or co-activator	Transcription

Table 5.10. MMS date hubs.

MMS party hubs			
ORF		**Description**	**Process**
YLR102C	APC9	Subunit of the Anaphase Promoting Complex	Cell cycle
YFR036W	CDC26	Subunit of anaphase-promoting complex (cyclosome)	Cell cycle
YDR388W	RVS167	Reduced viability upon starvation protein	Cell shape
YDR264C	AKR1	Ankyrin repeat-containing protein	Cell shape
YOL018C	TLG2	Member of the syntaxin family of t-SNAREs	Cell traffic
YMR167W	MLH1	DNA mismatch repair protein	DNA metabolism
YBL099W	ATP1	F1F0-ATPase complex, F1 alpha subunit	Mito. Atpase
YDR298C	ATP5	F1F0-ATPase complex, OSCP subunit	Mito. Atpase
YGL173C	KEM1	Multifunctional nuclease	mRNA metabolism
YGL070C	RPB9	DNA-directed RNA polymerase II, 14.2 KD subunit	Polymerase
YJL124C	LSM1	Sm-like (Lsm) protein	Sm
YEL051W	VMA8	H+-ATPsynthase V1 domain 32 KD subunit, vacuolar	Vacuolar atpase
YEL027W	CUP5	H+-ATPase V0 domain 17 KD subunit, vacuolar	Vacuolar atpase
YOR332W	VMA4	H+-ATPase V1 domain 27 KD subunit, vacuolar	Vacuolar atpase
YLR447C	VMA6	H+-ATPase V0 domain 36 KD subunit, vacuolar	Vacuolar atpase
YGR020C	VMA7	H+-ATPase V1 domain 14 kDa subunit, vacuolar	Vacuolar atpase
YBR127C	VMA2	H+-ATPase V1 domain 60 KD subunit, vacuolar	Vacuolar atpase
YPL234C	TFP3	H+-ATPase V0 domain 17 KD subunit, vacuolar	Vacuolar atpase
YOR270C	VPH1	H+-ATPase V0 domain 95K subunit, vacuolar	Vacuolar atpase
YHR026W	PPA1	H+-ATPase 23 KD subunit, vacuolar	Vacuolar atpase

Table 5.11. MMS party hubs.

4NQO date hubs			
ORF		**Description**	**Process**
YHR152W	SPO12	Sporulation protein	Cell cycle
YOR212W	STE4	GTP-binding protein beta subunit of the pheromone pathway	Cell signaling
YMR125W	STO1	Large subunit of the nuclear cap-binding protein complex CBC	mRNA metabolism
YDR378C	LSM6	Sm-like (Lsm) protein	Sm
YNL147W	LSM7	Sm-like (Lsm) protein	Sm
YBL093C	ROX3	Transcription factor	Transcription
YBR081C	SPT7	Involved in alteration of transcription start site selection	Transcription
YBR289W	SNF5	Component of SWI/SNF transcription activator complex	Transcription
YDR176W	NGG1	General transcriptional adaptor or co-activator	Transcription
YDR392W	SPT3	General transcriptional adaptor or co-activator	Transcription
YGL025C	PGD1	Mediator complex subunit	Transcription
YGR252W	GCN5	Histone acetyltransferase	Transcription
YLR055C	SPT8	Transcriptional adaptor or co-activator	Transcription
YNL236W	SIN4	Global regulator protein	Transcription
YOL004W	SIN3	Transcription regulatory protein	Transcription
YPL254W	HFI1	Transcriptional coactivator	Transcription

Table 5.12. 4NQO date hubs.

4NQO party hubs			
ORF		**Description**	**Process**
YFR036W	CDC26	Subunit of anaphase-promoting complex (cyclosome)	Cell cycle
YDR388W	RVS167	Reduced viability upon starvation protein	Cell shape
YLR268W	SEC22	Synaptobrevin (V-SNARE)	Cell traffic
YOL018C	TLG2	Member of the syntaxin family of t-SNAREs	Cell traffic
YJR121W	ATP2	F1F0-ATPase complex, F1 beta subunit	Mitocondrial ATPase
YBL099W	ATP1	F1F0-ATPase complex, F1 alpha subunit	Mitocondrial ATPase
YGL173C	KEM1	Multifunctional nuclease	mRNA metabolism
YGL070C	RPB9	DNA-directed RNA polymerase II, 14.2 KD subunit	Polymerase
YCR002C	CDC10	Cell division control protein	Septins
YJL124C	LSM1	Sm-like (Lsm) protein	Sm
YEL051W	VMA8	H+-ATPase V1 domain 32 KD subunit, vacuolar	Vacuolar ATPase
YHR026W	PPA1	H+-ATPase 23 KD subunit, vacuolar	Vacuolar ATPase
YGR020C	VMA7	H+-ATPase V1 domain 14 kDa subunit, vacuolar	Vacuolar ATPase
YOR332W	VMA4	H+-ATPase V1 domain 27 KD subunit, vacuolar	Vacuolar ATPase
YEL027W	CUP5	H+-ATPase V0 domain 17 KD subunit, vacuolar	Vacuolar ATPase
YDL185W	TFP1	H+-ATPase V1 domain 69 KD subunit, vacuolar	Vacuolar ATPase
YBR127C	VMA2	H+-ATPase V1 domain 60 KD subunit, vacuolar	Vacuolar ATPase
YLR447C	VMA6	H+-ATPase V0 domain 36 KD subunit, vacuolar	Vacuolar ATPase
YPL234C	TFP3	H+-ATPase V0 domain 17 KD subunit, vacuolar	Vacuolar ATPase

Table 5.13. 4NQO party hubs.

UV date hubs			
ORF		**Description**	**Process**
YPL031C	PHO85	Cyclin-dependent protein kinase	Cell cycle
YDR176W	NGG1	General transcriptional adaptor or co-activator	Transcription
YDR448W	ADA2	General transcriptional adaptor or co-activator	Transcription
UV party hubs			
ORF		**Description**	**Process**
YFR036W	CDC26	Subunit of anaphase-promoting complex (cyclosome)	Cell cycle
YBL099W	ATP1	F1F0-ATPase complex, F1 alpha subunit	Mitochondrial ATPase
YGL070C	RPB9	DNA-directed RNA polymerase II, 14.2 KD subunit	Polymerase
YGR020C	VMA7	H+-ATPase V1 domain 14 kDa subunit, vacuolar	Vacuolar ATPase

Table 5.14. UV date and party hubs.

t-BuOOH date hubs			
ORF		**Description**	**Process**
YPL031C	PHO85	Cyclin-dependent protein kinase	Cell cycle
YPL254W	HFI11	Transcriptional coactivator	Transcription
YBR081C	SPT7	Involved in alteration of transcription start site selection	Transcription
YOL148C	SPT20	Member of the TBP class of SPT proteins that alter transcription site selection	Transcription
YBR289W	SNF5	Component of SWI/SNF transcription activator complex	Transcription
t-BuOOH party hubs			
ORF		**Description**	**Process**
YDR388W	RVS167	Reduced viability upon starvation protein	Cell shape
YDR395W	SXM1	Putative beta-karyopherin	Cell traffic
YOR332W	VMA4	H+-ATPase V1 domain 27 KD subunit, vacuolar	Vacuolar ATPase
YOR270C	VPH1	H+-ATPase V0 domain 95K subunit, vacuolar	Vacuolar ATPase
YHR026W	PPA1	H+-ATPase 23 KD subunit, vacuolar	Vacuolar ATPase
YEL051W	VMA8	H+-ATPsynthase V1 domain 32 KD subunit, vacuolar	Vacuolar ATPase
YEL027W	CUP5	H+-ATPase V0 domain 17 KD subunit, vacuolar	Vacuolar ATPase

Table 5.15. *t*-BuOOH date and party hubs.

in. As can be seen from Table 5.9, although the majority of these hubs are essential, most of the remaining nonessential party and date hubs are involved in damage recovery. Tables 5.10-5.15 indicate that the date hubs mostly coordinate transcription and the cell cycle while party hubs are involved in complexes and signaling modules important for recovery from exposure to the mutagen such as the vacuolar ATPase, the mitochondrial ATPase, and cell trafficking. These results shed some light on the organization of damage-recovery proteins and emphasize the topological similarities between essential and damage-recovery proteins.

■ 5.6 Beyond Pairwise Interactions: Understanding Network Integrity Through Constrained Graph Partitioning

Biological signaling is ultimately about relaying signals and processing them. At relatively short time-scales, this process happens entirely at the level of protein-protein interactions. As a result, it would be expected that disabling protein-protein interactions such that entire sections of the network can no longer communicate with each other may have a profound effect on the ability of a cell to properly process signals and in some cases may lead to a catastrophic outcome. Alternatively, disconnection between sections of the network can potentially occur by disabling specific proteins whose collective action seems to hold the network together. We have investigated these two scenarios by applying different graph partitioning algorithms to the different versions of the yeast interactome.

■ 5.6.1 Interaction-centric partitioning

We first focused on specific interactions as opposed to full protein function by disconnecting edges in the graph rather than removing nodes. Specifically, the yeast interactome was partitioned into two equal partitions while minimizing the number of edges cut (i.e. the

	FULL		CORE		FYI	
Total number of proteins	4,597		2,435		778	
Essential	25.41%		33.68%		45.76%	
Damage-recovery (DR)	30.17%		31.17%		26.74%	
No-Phenotype	39.46%		30.84%		24.16%	
Total number of edges	14,448		6,227		1,798	
Number of edges cut	2,396 (16.58%)		557 (8.95%)		9 (0.5%)	
	part. 0	part. 1	part. 0	part. 1	part. 0	part. 1
Average degree	6.357	4.130	4.601	4.713	3.424	5.774
Essential	28.2%	22.6%	36.9%	30.5%	35.7%	55.8%
Damage-Recovery (DR)	30.1%	30.2%	30.9%	31.4%	29%	24.4%
No-Phenotype	36.6%	42.4%	28.3%	33.4%	31.4%	17.0%
Nodes in partition with edges cut	47.7%	39.41%	26.62%	24.38%	2.31%	2.06%
Essential edges cut/ Total edges cut	31.5%	29.7%	38.9%	41.4%	77.8%	62.5%
Essential edges cut/ Total essential	53.2%	51.8%	28.1%	33.2%	5%	2.3%
DR edges cut/ Total edges cut	32.2%	29.8%	32.4%	30.6%	22.2%	25.0%
DR edges cut/ Total DR	51.0%	38.8%	27.9%	23.8%	1.8%	2.1%
Nophe edges cut/ Total edges cut	30.9%	36.2%	24.1%	24.2%	0%	12.5%
Nophe edges cut/ Total nophe	40.4%	33.7%	22.7%	17.7%	0%	1.5%
Nodes with > 5 edges cut $\equiv N_5$	2.78%	3.83%	0.49%	1.15%	0%	0%
Essential with > 5 edges cut/ N_5	40.6%	45.5%	83.3%	78.6%	-	-
DR with > 5 edges cut/ N_5	35.9%	25.0%	0%	0%	-	-
Nophe with > 5 edges cut/ N_5	21.9%	26.1%	16.7%	21.4%	-	-

Table 5.16. Summary of edge partitioning results.

YBR133C *HSL7* ↔ YPL016W *SWI1*	DR ↔ Ess
YDL043C *PRP11* ↔ YKL074C *MUD2*	Ess ↔ Nophe
YIL129C *TAO3* ↔ YDR167W *TAF25*	Ess ↔ Ess
YOR250C *CLP1* ↔ YDR167W *TAF25*	Ess ↔ Ess
YIL144W *TID3* ↔ YOL069W *NUF2*	Ess ↔ Ess
YLR026C *SED5* ↔ YLR268W *SEC22*	Ess ↔ DR
YMR117C *SPC24* ↔ YPR010C *RPA135*	Ess ↔ Ess
YOL004W *SIN3* ↔ YDR207C *UME6*	DR ↔ DR
YOR326W *MYO2* ↔ YBR109C *CMD1*	Ess ↔ Ess

Figure 5-3. Specific interactions with edges cut in the FYI partitioning. ORF names are followed by common names in italic. DR, Ess, and Nophe correspond to damage-recovery, essential, and no-phenotype classifications respectively.

number of interactions disabled). This was done using the program METIS (http://www-users.cs.umn.edu/karypis/metis/metis/index.html) which is based on a multilevel graph partitioning algorithm described in [76]. The program provides high quality partitions and runs very fast. We applied the graph partition algorithm on the three versions of the yeast interactome described earlier: the full yeast interactome (full), the curated core yeast interactome (core), and the filtered yeast interactome (FYI). For each interactome, the minimum number of edges cut to partition the network into two equal partitions (part. 0 and part. 1) was computed and identified. The nature of these interactions was investigated as well as the phenotypic node distribution between the two partitions, i.e. the relative number of nodes with a given phenotype (essential, damage-recovery, and no-phenotype) in each partition. We also computed the average degree of each partition and identified the phenotype of nodes that had a large number of interactions disabled in order to partition the network. The results are summarized in Table 5.16 and indicate that the proportion of nodes with a given phenotype in each partition is similar to the proportion of nodes with a given phenotype in the non-partitioned network for all three versions of the network. However the relative number of proteins with a given phenotype that had their edges cut is different for each version of the network. While for the full and core networks the proportion of essential proteins that had their edges cut is slightly higher than the proportion of essential proteins in the partition, this proportion is much higher in the FYI network indicating that most of the edges cut in the network belong to essential proteins. Furthermore, the number of deleted edges needed to partition the network is higher for the full network (16.58%), than for the core network (8.95%) which is higher than for the FYI network (0.5%). This should not be surprising since the core and FYI networks are curated and include only high confidence interactions. As a result, these networks are sparser and therefore it would take fewer disconnected edges to partition them. The results also show that the proteins that had many (greater than five) interactions removed are mostly essential proteins emphasizing the special topological role of these proteins.

In order to gain further insight as to the nature of the interactions disabled, we identified the five proteins in each partition that had most of their edges cut for the full and core interactomes as well as all the proteins that had their edges cut for the FYI network. The results are shown in Table 5.17 and again show that the majority of these proteins are essential. The FYI network results are particularly interesting since cutting only nine edges disconnects the network. These nine interactions are shown in Figure 5-3 along with the

FULL					
Partition 0			Partition 1		
ORF	Edges cut	Class	ORF	Edges cut	Class
YML064C	38	Ess.	YJR091C	71	Nophe.
YBR055C	20	Ess.	YNL189W	70	Ess.
YMR308C	19	Ess.	YLR447C	39	DR
YDR034C	18	Nophe.	YMR047C	33	Ess.
YER133W	16	Ess.	YHR114W	31	-
CORE					
Partition 0			Partition 1		
ORF	Edges cut	Class	ORF	Edges cut	Class
YBR009C	10	Nophe	YBR160W	33	Ess.
YDR170C	8	Ess.	YNL189W	20	Ess.
YPL153C	7	Ess.	YDL188C	8	Nophe.
YPL204W	7	Ess.	YFL039C	8	Ess.
YPR041W	7	Ess.	YHR135C	8	Nophe.
FYI					
Partition 0			Partition 1		
ORF	Edges cut	Class	ORF	Edges cut	Class
YBR133C	1	DR	YDR167W	2	Ess.
YDL043C	1	Ess.	YBR109C	1	Ess.
YIL129C	1	Ess.	YDR207C	1	DR
YIL144W	1	Ess.	YKL074C	1	Nophe.
YLR026C	1	Ess.	YLR268W	1	DR
YMR117C	1	Ess.	YOL069W	1	Ess.
YOL004W	1	DR	YPL016W	1	Ess.
YOR250C	1	Ess.	YPR010C	1	Ess.
YOR326W	1	Ess.			

Table 5.17. Proteins identities (top 5) with the highest number of edges cut in the full and core networks. All proteins with edges cut in the FYI network.

common names of the proteins as well as their phenotype. The majority of these interactions involve at least one essential protein and most of them involve two essential proteins.

Proteins are classified as essential because their deletion is lethal. However when deleting a protein, all its interactions with other proteins are effectively disabled. Some of these interactions may be essential in the sense that without those functional interactions, the yeast would die, others however may not. Given the way experiments are carried out to identify essential proteins (gene deletion), it is impossible to determine which specific interactions or combinations of interactions are essential. In order to identify which interactions are essential, interactions have to be specifically inhibited through a specific antibody for example or an engineered protein that has an inactive binding site effectively disabling that interaction. These types of experiments can be very difficult to carry out and going through the combination of over 700 interactions (for the case of the FYI network) is practically infeasible. The results in this section suggest a method to identify potential combinations of interactions that are likely to be essential. Specifically, we hypothesize that inhibiting

interactions that lead to network disconnection will be catastrophic to the cell leading to cell death. As a result, we predict that if the interactions in Figure 5-3 were specifically inhibited, the yeast will die. Carrying out these types of experiments is suggested as a possible direction for future work. Being able to distinguish between essentiality based on full protein function and essentiality based on specific protein interactions is important for understanding fundamental network structure and function as well as for effective targeted drug design. The types of network analyses presented in this section present a first step towards building this kind of understanding.

■ 5.6.2 Protein-centric partitioning

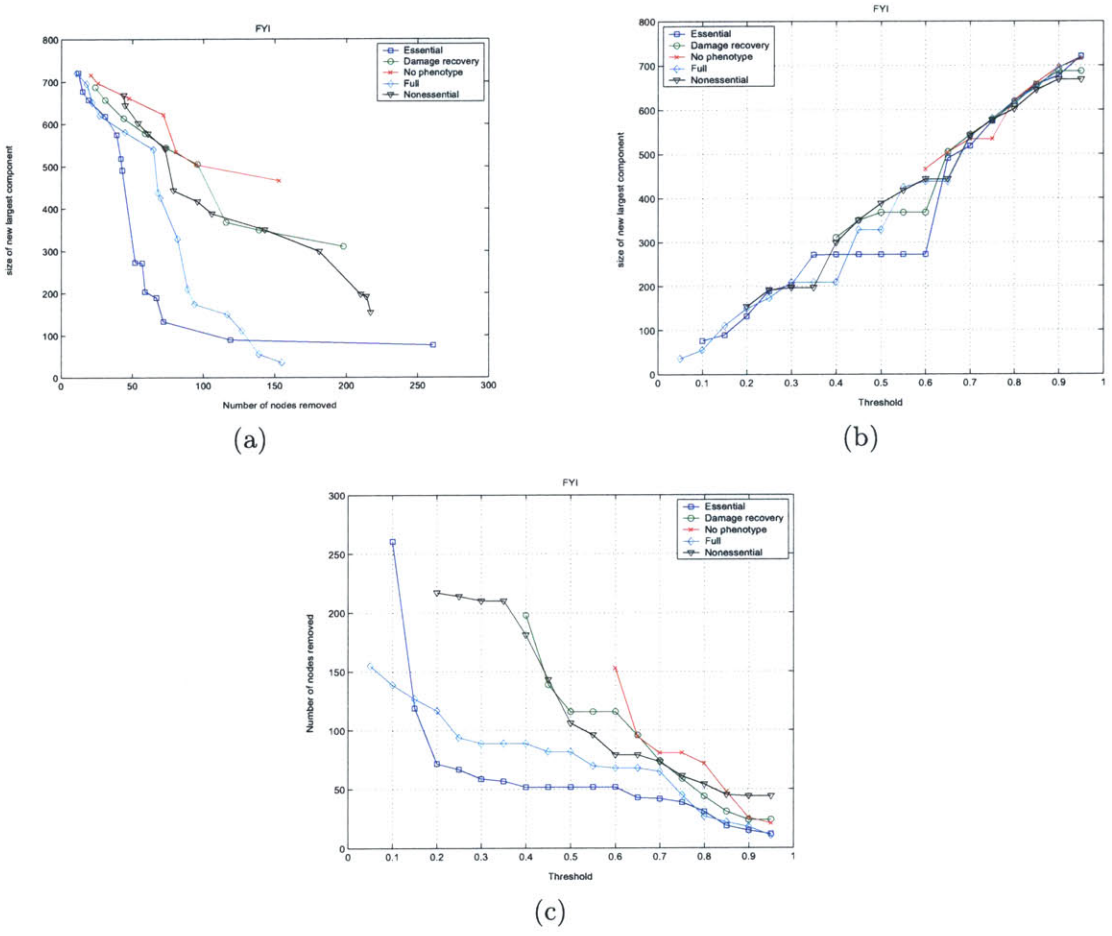


Figure 5-4. Node partitioning results using the FYI interactome.

We also investigated full protein function by examining the set of proteins that, if removed, disconnects the network. We call this approach protein-centric partitioning because it focuses on node (protein) removal as opposed to edge (interaction) cutting. Nodes can be removed and the network can be disconnected in a number of different ways. We present here one approach for removing nodes and disconnecting the network to illustrate the potential of this partitioning scheme. We use the FYI network and consider five classes of proteins (nodes): all proteins (full), the essential proteins (essential), the non-essential proteins (non-

essential), the damage-recovery proteins (damage-recovery), and the no-phenotype proteins (no-phenotype). Proteins in each class are ordered according to their degree in the full network and are serially removed starting from the highest degree protein until the ratio of the largest connected component size of the new network and the largest connect component size in the original network reaches a threshold which is varied in increments of 0.05 from 0 to 0.95. In addition, in order to avoid trivial solutions where single nodes are being disconnected, the size of the second largest connected component is constrained to be at least 10 nodes. Figure 5-4 shows the results.

Figure 5-4 shows that it takes fewer essential nodes to disconnect the network at a given threshold than non-essential nodes and similarly it takes fewer damage-recovery nodes to disconnect the network than no-phenotype nodes. These results further emphasize the topological similarities between essential and damage-recovery proteins. The results in Figure 5-4 also show a surprising trend: at almost all threshold values, it takes fewer proteins to disconnect the network if using the class of essential proteins than if using the class of all proteins. A similar trend, though less marked, holds for the damage-recovery proteins where for a few threshold values, fewer damage-recovery proteins need to be removed to disconnect the network than if the entire class of non-essential proteins was used. These results are not necessarily expected because the essential and damage-recovery proteins are subsets of the all and non-essential proteins respectively and therefore any essential (damage-recovery) protein is present in the class of all (non-essential) proteins. Furthermore, since we chose to remove nodes by classifying them according to their degree and by starting from the highest degree node, the average degree of the nodes removed to satisfy a given disconnection threshold is lower for the essential proteins than for the class of all proteins since the essential class is a subset of the class of all proteins and it takes fewer essential proteins to disconnect. The same observation holds for the damage-recovery proteins. These results indicate that essential and damage-recovery proteins have distinctive network properties beyond the number of direct physical interactions (degree). Specifically, they play an important role in maintaining network integrity.

We have chosen here to use protein degree as a classifying measure to systematically remove proteins in order to disconnect the network. This choice was motivated by the fact that node degree was extensively used as a distinctive network property in the other analyses presented in this chapter as well as the previous two chapters and continues to be used in many published articles studying network topology. The results presented here show that properties of phenotypic proteins extend well beyond first-order interactions (analyzed through node degree) emphasizing the importance of studying the properties of phenotypic proteins using the methods presented in this section. A number of other network measures can and should be used to classify proteins and systematically remove them in order to disconnect the network. This analysis is the focus of future work.

Chapter 6

Low-Level Modeling of Biological Signaling Networks

Looking at network topology and interconnectivity properties is important in order to understand the high-level properties of the network and for example the impact a node removal could have on the entire network. However, temporal information and dynamics cannot be inferred from such a high-level modeling technique. In other words, in order to get the dynamics of the interactions between different nodes in a graph, one has to know the identities and characteristics of the nodes. In this section, we add resolution to the models described in the previous section by examining the proteins that constitute the nodes in the graph and providing a model for their dynamics and interactions. We term this approach low-level modeling of signaling networks.

Most current low-level models for signaling pathways can be classified in one of two categories: biochemical models which define the system in terms of kinetic differential equations governed by the laws of mass action or Michaelis Menten kinetics and biophysical models which are concerned with the mechanical forces and spatial distributions necessary to generate a response. While these models usually have a direct physical basis, even the simplest of these includes hundreds of equations and many parameters that usually need to be estimated since they are not directly measurable. As a result, one can rarely gain intuition from such models. Also, the under-constrained parameter space makes such models somewhat arbitrary.

In this thesis, we propose a fundamentally different approach to modeling biological signaling at a low-level based on statistical control and signal processing theory. Our goal is to develop an intuition as well as some understanding of how cells perform signal processing based on approximate yet detailed models with a physical foundation inspired from our engineered signal processing networks. We start this section by reviewing some of the relevant current low-level models. We then present a new framework based on interacting Markov chains and discuss related models.

■ 6.1 Current Low-Level Models

Biological signaling networks can be represented as a series of biochemical reactions that can then be modeled using some mathematical formalism. If one further assumes that the volume V of the cell or compartment is fixed and contains a spatially uniform mixture of N species (proteins, chemicals, or other molecules) that can interact through M specified biochemical reactions, then given the number of molecules of each species present at some initial time, the modeling problem consists of finding the molecular population levels at any later time. Several approaches have been formulated to solve this problem, the oldest of which is the deterministic approach [125] [92]. More recently, a stochastic approach has been proposed [90] and successfully implemented using a direct algorithm [55] as well as a Petri net formalism [61]. We give below some background on these approaches.

■ 6.2 Deterministic Formulation

The traditional way of modeling biochemical reaction networks consists of translating each reaction involved in the network into a set of ordinary differential equations. This leads to the deterministic approach. The deterministic approach is based on two basic assumptions: the number of molecules of each species can be represented by a continuous, single-valued function and each of the M chemical reactions can be regarded as a continuous-rate process. Using these two assumptions one can easily construct a set of coupled, first-order, ordinary

differential equations of the form

$$\frac{dX_i}{dt} = f_i(X_1, ..., X_N) \tag{6.1}$$

where $i = 1, ..., N$ and N is the total number of species. The specific forms of the functions f_i are determined by the structures and rate constants of the M chemical reactions. These equations express the time-rate-of-change of the molecular concentration of one chemical species, X_i, as a function of the molecular concentrations of all the species, $X_1, ..., X_N$. They are termed the *reaction-rate equations* and their solution gives the time evolution of the network. In large volumes, the rate of the reaction is *proportional* to the concentration of the reactants. This is the Law of Mass Action which has been well established in large volumes based on experiments that date back to the mid-1800s [125]. Later, in the early 1900s, Michaelis and Menten [92] published a model for enzyme kinetics where an enzymatic reaction can be written as:

$$E + S \leftrightarrow ES \rightarrow P + E \tag{6.2}$$

The Law of Mass Action applied to this basic enzyme reaction leads to a set of coupled differential equations that can be approximated using perturbation theory or solved numerically. This approach is referred to as the *deterministic approach.*

While differential equations are a natural way to model chemical reactions in large containers, they do not necessarily represent the true state of the system in a cell. Specifically, the Law of Mass Action relies on two key assumptions: continuity and determinism, both become problematic at the level of the single cell. The low number of molecules, usually ranging from dozens to a few hundreds, may compromise the notion of continuity while fluctuations in the timing of cellular events may lead to non-deterministic outcomes. For example, two regulatory systems having the same initial conditions might ultimately end up in different states. The deterministic approach also ignores effects due to correlations. The stochastic approach, which we discuss next, tries to deal with some of these shortcomings.

■ 6.3 Stochastic Master Equation Formulation

The stochastic approach to chemical kinetics dates back to the mid 1950s [90] where the inability of the deterministic formulation to take into account fluctuations and correlations led scientists to explore a stochastic formulation. In this approach, the reaction constants are viewed not as reaction "rates" but as reaction "probabilities per unit time". As a result, the time evolution of the N species is analytically described not by a set of N coupled differential equations but rather by a single differential-difference equation for a grand probability function in which time and the N species populations all appear as independent variables, this equation is termed the *master equation* [55]. The solution to the master equation gives the probability of finding various molecular populations at each instant of time and it can be rigorously derived from a microphysical standpoint [57]. The fundamental hypothesis of the stochastic formulation, and the *only* one, is that there exists a reaction parameter c_μ which characterizes each reaction R_μ such that:

$c_\mu \delta t$ = average probability, to first order in δt, that a particular combination of R_μ reactant molecules will react accordingly in the next interval δt.

where c_μ depends on many physical properties of the system including volume, temperature, mass, and activation energy. One can thus see that the solution to the master equation can be thought of as a Markovian random walk in the space of reacting variables. The condition

above is satisfied when non-reactive molecular collisions occur much more frequently than reactive molecular collisions, i.e. the system is "well-mixed": the reactant molecules are always randomly distributed uniformly throughout the volume.

One can prove [80] that the stochastic formulation reduces to the deterministic formulation in the thermodynamic limit i.e. when the number of molecules of each species and the containing volume all approach infinity in such a way that the molecular concentrations approach finite values. In this limit, the deterministic rate constant k_μ can be expressed in terms of the reaction parameter c_μ as

$$k_\mu = V * c_\mu \tag{6.3}$$

where V is the volume.

Analytic solutions to the master equation are very difficult to obtain and earlier numerical methods encountered convergence issues. However, Gillespie in 1976 [55] presented a general method for numerically simulating the stochastic time evolution, and, as a result, the stochastic formulation gained more applicability in the modeling community. In the remainder of this section, we briefly present the Gillespie algorithm, for a more detailed discussion, the reader is referred to [55].

The Gillespie algorithm is based on defining the *reaction probability density function*, $P(\tau, \mu)$ where:

$P(\tau, \mu)d\tau$ = the probability at time t that the *next* reaction in V will occur in the differential time interval $(t + \tau, t + \tau + d\tau)$ *and* will be an R_μ reaction.
$P(\tau, \mu)$ is therefore a joint probability density function on the space of the continuous variable τ $(0 \leq \tau < \infty)$ and the discrete variable μ $(\mu = 1, 2, ..., M)$. In order to obtain an expression for $P(\tau, \mu)$, we define:

h_μ = the number of distinct molecular reactant combinations for reaction R_μ found to be present in V at time t.
Then,

$h_\mu c_\mu \delta t$ = probability, to first order in δt that an R_μ reaction will occur in V in the next time interval δt.
With these definitions, one can derive an expression for $P(\tau, \mu)$ as follows:

$$P(\tau, \mu) = h_\mu c_\mu e^{\sum_{\mu=1}^{M} h_\mu c_\mu \tau} \tag{6.4}$$

for τ positive real $(0 \leq \tau < \infty)$ and μ integer μ $(\mu = 1, 2, ..., M)$, for all other values of τ and μ, $P(\tau, \mu)$ is zero. Furthermore, one can write:

$$P(\tau, \mu) = P_1(\tau) P_2(\mu|\tau) \tag{6.5}$$

where

$$P_1(\tau) = a e^{-a\tau} \tag{6.6}$$

and

$$P_2(\mu|\tau) = \frac{a_\mu}{a} \tag{6.7}$$

where $a_\mu = h_\mu c_\mu$ and $a = \sum_{\mu=1}^{M} a_\mu$. With these equations, one can use a direct method where we first generate a random value τ according to $P_1(\tau)$ in equation (6.6) and then generate a random integer μ according to $P_2(\mu|\tau)$ in (6.7). The resulting random pair (τ, μ) will be distributed according to $P(\tau, \mu)$. The Gillespie algorithm uses this joint probability

to give an exact simulation to the corresponding master equation. Specifically, given a specified population of molecules at time 0, the Gillespie algorithm consists of the following three steps:

Step(1): Calculate the function $P(\tau, \mu)$ for the current population.

Step(2): Use Monte Carlo methods to generate a pair of random numbers (τ, μ) according to the density function computed in Step(1).

Step(3): The time variable is advanced by τ and the molecular population is adjusted to reflect the occurrence of one molecular reaction R_μ and return to Step(1).

Note that the algorithm generates its own, non-uniform, time samples: as the simulation proceeds, it generates time samples based on the probability density function in (6.6), i.e. simulation time steps are based on draws from an *exponential* distribution. Each reaction that occurs changes the quantity of at least one reactant. When this happens, h_μ changes and it is necessary to recalculate the a_μ values. If we did not need to recalculate the probabilities at every time step, the system would be a Markov process with a fixed transition matrix.

■ 6.4 Stochastic Petri Nets

Petri nets were first introduced by C.A. Petri in 1960 and have been widely used in computer science for the modeling and analysis of computer systems. They have a standard graphical representation that is easy to interpret and to use for defining models. Petri net models are composed of four basic elements: places P, transitions T (represented by bars or boxes), directed arcs, and tokens. If a directed arc connects a place to a transition, the place is called the input place to the transition while if a directed arc connects a transition to a place, the place is the output place to the transition. The directed arcs linking input and output places to transitions represent the input and output functions respectively. The arcs can have different weights depending on the input and output functions and are labeled with a natural number. By default unlabeled arcs have unit weight. Given these rules, multiple places and transitions can be connected to form a complex net to model the static view of a complex system.

In addition to the net structure that represents the static part of the system, Petri nets have a global marking M which represents the overall state of the structure. Each place contains tokens, the number of tokens in a place is its marking and represents the local state of the place. The token distribution among the places of a Petri net is called its global marking and represents the overall state of the system. The dynamic behavior of the system is modeled by the flow of tokens and the firing of transitions. Specifically, a transition is said to be enabled if each input place has at least as many tokens as the weight of the arc connecting them. Enabled transitions may be fired by removing from each input place the number of tokens equal to the weight of the arc. When the transition is fired, tokens, equal to the weight of the arc joining the transition to the output place, are added to the output places connecting the transition.

In conventional Petri nets, every enabled transition may fire. However this is not always true for other kinds of Petri nets. In particular, in the kind of Petri nets we are interested in, *stochastic Petri nets*, enabled transitions fire with an *exponentially* distributed time delay. The rate parameter for each transition t_j is given by a weight function w_j and may in general be a function of the global marking M. These stochastic Petri nets are equivalent to continuous-time Markov chains.

Since Petri nets are described as both a graphical tool and a mathematical tool, they

provide both a visual medium for a modeler to describe complex systems and present it to others and an analytic medium for formal analysis via linear algebraic equations. Specifically, one can perform structural analysis on the net to determine which states are transient, fixed and recurring. Numerical analysis can also be used to derive both steady-state and transient behavior for a small number of states and algorithms are available to simulate both steady-state and transient behavior as well as estimating the distributions of the results. However the major weaknesses of Petri nets is that as the system size and complexity evolve, the state-space of the Petri net grows exponentially which can become too difficult to manage both graphically and analytically.

The use of stochastic Petri nets to model biochemical reactions was first proposed by Goss and Peccoud [61]. In their model, each place represents a distinct molecular species and tokens represent the individual molecules. The initial marking M_0 is the number of molecules of each species in the system at time $t = 0$. Elementary chemical reactions are represented by transitions and input and output functions determine the stoichiometric coefficients of the molecular species involved in the reaction. The rate of the reaction is represented by the weight function, W. The input places are the reactants and the output places are the products. If the rate parameter w_j for a transition t_j of the stochastic Petri net is equal to the stochastic rate (as defined in the previous section) of the reaction the transition represents, then the stochastic Petri net represents molecular interactions *exactly*. In other words, the Kolmogorov equations for the corresponding stochastic Petri net are the same as the chemical master equation for the system of molecular interactions represented. In their work, Goss and Peccoud defined stochastic Petri nets models for plasmid ColE1 replication [61]. They used a computer package that integrates the graphical representation with the analysis of stochastic Petri nets to simulate the dynamics of the system.

One of the limitations of stochastic Petri nets in their application to biochemical systems modeling is the lack of spatial dimension. While compartmentalization by representing systems as separate but interacting with the movement of molecules between compartments represented by transitions is conceptually easy to perceive, the current definition of stochastic Petri nets does not allow the implementation of a true reaction-diffusion model. This limitation together with the exponential growth of the state-space creates the need for more manageable models that are still detailed enough to capture the dynamics of the system. In the next chapters, we present such a model based on interacting Markov chains.

Chapter 7

Stochastic Models for Cell Signaling

In this chapter, we propose a new framework for modeling the stochasticity of cellular signaling. Our goal is to develop an intuitive model that captures the stochastic behavior of cells while reconciling the results obtained from a deterministic modeling approach. The model is formulated and described in the following sections.

■ 7.1 Markov Modulated Markov Chains (3MC)

We define a Markov modulated Markov chain to be a doubly stochastic process composed of a discrete-time Markov chain whose transition probabilities vary according to a second Markov chain. The two chains are said to interact or more precisely, the states in the second chain modulate the state transition probabilities of the first chain.

The Markov Modulated Markov Chains (3MC) model is composed of a network of v interacting nodes. Each node is composed of a k-state Markov modulated Markov chain, i.e. the transition probabilities for node X_p at time n is given by:

$$P(X_p[n] = j | X_p[n-1] = i, \cdots, X_p[0] = m) = P(X_p[n] = j | X_p[n-1] = i) \equiv p_{ij}^{X_p}[n] \quad (7.1)$$

where $X_p[n]$ is the state of node X_p at time n. In the absence of any modulation, the transition probabilities of chain X_p are simply $p_{ij}^{X_p}[n] = q_{ij}^{X_p}[n]$ where $q_{ij}^{X_p}[n]$ is independent of all chains in the network. More generally, nodes can interact i.e. the Markov chain in one node is modulated by Markov chains in other nodes. These interactions between different nodes are defined by influences from states in a set of nodes onto transition probabilities in other nodes. The interaction may be either positive (activating) or negative (inhibiting). Specifically, if a node X_r affects the transition probability from state i to state j in node X_p then the transition probability from state i to state j in node X_p at time n due to the interaction with chain X_r, $\{p_{ij}^{X_p}[n]\}_{X_r}$, is given by:

$$\{p_{ij}^{X_p}[n]\}_{X_r} = f(X_r[n]) \quad (7.2)$$

where $X_r[n]$ is the state of node X_r at time n and $f()$ is a function that depends on the nature of the interaction and will be derived later in this chapter.

Furthermore, in addition to $q_{ij}^{X_p}[n]$, a given transition probability may be subject to interactions from several nodes, each one of which will have a form similar to equation (7.2). There are clearly many different ways in which the individual transition probabilities, or equivalently the interactions, may be combined with $q_{ij}^{X_p}[n]$ into one composite transition probability. Here we consider two models for combining the state influences: an additive model and a fading model.

■ 7.1.1 Additive model

As the name suggests, in the additive model, the interactions from different nodes are added together to form the total transition probability from state i to state j in node X_p at time n, $p_{ij}^{X_p}[n]$, as follows:

$$p_{ij}^{X_p}[n] = q_{ij}^{X_p}[n] + \sum_r \{p_{ij}^{X_p}[n]\}_{X_r} \quad (7.3)$$

This model is used whenever the individual interactions define mutually exclusive events and when the nature of the compound interaction is associative.

■ 7.1.2 Fading model

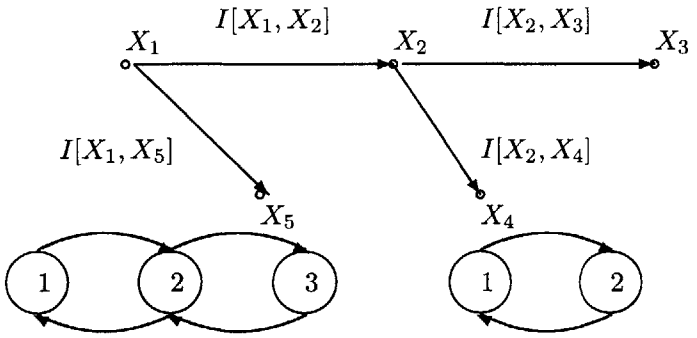
In the fading model, the individual interactions are combined through multiplication, i.e. the composite probability is given by:

$$p_{ij}^{X_p}[n] = q_{ij}^{X_p}[n] \prod_r \{p_{ij}^{X_p}[n]\}_{X_r} \tag{7.4}$$

This model is used when the nature of the compound interaction is cooperative.

■ 7.2 Notation

A graphic representation of the model is given in Figure 7-1 where directed arrows between nodes define influences of the Markov chain in the originating node onto specific transition probabilities in the receiving node. The interaction is labeled $I[X_r, X_p]$ where X_r is the originating node and X_p is the receiving node. Each $I[X_r, X_p]$ carries with it a set of relations of the form $f(X_r[n]) \xrightarrow{\epsilon} [p_{ij}, X_p]$ where p_{ij} is the transition probability in node X_p being affected by node X_r, $f(X_r[n])$ is the mapping of the influence from the state of node X_r at time n onto the transition probability, and ϵ describes the way the interactions is to be composed with other interactions. Specifically, $\epsilon = +$ if the additive model is to be used and $\epsilon = \times$ if the fading model is to be used. Examples of such interactions are given in Figure 7-1.



$$\begin{aligned}
I[X_1, X_2] &= f(X_1[n]) \xrightarrow{+} [p_{12}, X_2] \\
I[X_2, X_3] &= f(X_2[n]) \xrightarrow{+} [p_{21}, X_3] \\
I[X_1, X_5] &= f(X_1[n]) \xrightarrow{\times} [p_{23}, X_5] \\
I[X_2, X_4] &= f(X_2[n]) \xrightarrow{\times} [p_{11}, X_4] \\
&\quad\; f(X_2[n]) \xrightarrow{+} [p_{21}, X_4]
\end{aligned}$$

Figure 7-1. Graphical representation of the interacting Markov chain model. Two representative Markov chains are illustrated for nodes 4 and 5: a two-state and a three-state chain respectively.

■ 7.3 Evolution of the State Probabilities

Recursive formulae for the state occupancy probabilities of the individual Markov modulated Markov chains are similar to those of traditional Markov chains [52]. Specifically, let $\mathbf{p}_{X_p}[n]$ be a row vector of length k where k is the order of the Markov chain at node X_p and whose entries correspond to the state probabilities at time n, i.e. the l-th entry corresponds to the probability of chain X_p being in state l at time n, $P(X_p[n] = l)$, and the initial state probability distribution is given by $\mathbf{p}_{X_p}[0]$. Furthermore, let $\mathbf{A}_{X_p X_p}[n]$ be the transition

matrix of the Markov chain at node X_p at time n, i.e. $\mathbf{A}_{X_pX_p}[n]$ is written as follows:

$$\mathbf{A}_{X_pX_p}[n] = \begin{bmatrix} p_{11}^{X_p}[n] & \cdots & p_{1k}^{X_p}[n] \\ \vdots & \ddots & \vdots \\ p_{k1}^{X_p}[n] & \cdots & p_{kk}^{X_p}[n] \end{bmatrix} \tag{7.5}$$

Then, it follows from the Markovian property that:

$$\mathbf{p}_{X_p}[n+1] = \mathbf{p}_{X_p}[n]\mathbf{A}_{X_pX_p}[n] \tag{7.6}$$

and therefore:

$$\mathbf{p}_{X_p}[n+1] = \mathbf{p}_{X_p}[0] \prod_{i=0}^{i=n} \mathbf{A}_{X_pX_p}[i] \tag{7.7}$$

It is straightforward to show that if the individual Markov chains consist of a single recurrent class and if the graph representing the network is acyclic, then the state probabilities will *always* reach a steady-state which can be solved for by propagating the steady-state probabilities of the originating node (the node that is not subject to any influence) into the other nodes. If the graph contains cycles however, a steady-state may not exist. In order to have a full characterization of the network behavior such as the pattern of states that the network can be in at a given point in time, the joint probabilities of state occupancies in different chains is needed. This higher-order analysis is not considered here.

■ 7.4 Application to Cellular Signaling

In applying our model to cellular signaling, each type protein or enzyme corresponds to a node. The order of the associated Markov chain is given by the number of states the protein can have. For example, a protein can be in one of two states: active or inactive, alternatively it can have three possible states: unphosphorylated, singly phosphorylated, or dually phosphorylated. Non-zero transitions between states are given by our knowledge of the protein state transitions. For example if the protein cannot be dually phosphorylated without first being singly phosphorylated, the corresponding Markov chain will only allow for transitions between the unphosphorylated state and the singly phosphorylated state and between the singly phosphorylated state and the dually phosphorylated state, i.e. no direct transitions between the unphosphorylated state and the dually phosphorylated state are allowed.

Interactions between different nodes in the network are defined by interactions between the different proteins represented by those nodes including enzymatic reactions, binding events, and protein modifications. Furthermore, when more than one state affects a transition probability in a given node, whether the additive model or the fading model is used depends on the actual nature of the protein interaction. Specifically, if the protein interactions are known to be cooperative, i.e. the proteins either bind to the same site or all the proteins in question need to be present to get an activation or an inhibition effect, then the fading model is used. On the other hand, if the proteins bind at different sites or their actions are independent of each other, then the correct model to use is the additive model.

Multiple proteins or enzymes of the same type are represented by identical Markov chains independent of each other and evolving according to their transition probabilities. For simplicity, the identical Markov chains are never shown in the graphical representation

of the model. Only one Markov chain for each protein type is shown in the graphical representation. Since proteins of the same type are represented by identical Markov chains, they can be thought of as different realizations of the same random process. In addition, when proteins of different types interact, the specific pair of proteins interacting is selected at random from the set of proteins of the same type using a uniform probability distribution. For example, if a protein of type Y gets activated by interacting with a protein of type X and there is a total of 100 proteins of type X, a particular X is selected at random from the 100 proteins, using a uniform probability distribution, as the X protein the Y protein will interact with.

■ 7.5 Derivation of the Modulating Function $f()$

Biological signaling pathways are dominated by three main types of interactions: reactive collisions, such as enzymatic modifications, associative interactions such as protein-protein associations or ligand binding, and activations (or inhibitions) due to conformational changes such as proteins part of a receptor complex that get activated (or inhibited) in response to a conformational change of the receptor or proteins activated (or inhibited) through scaffolds. The form of the influence functions in our model will depend on the nature of the interaction. In this section, we derive an expression for $f()$ for these three types of interactions.

■ 7.5.1 Reactive collisions

Consider two types of molecules, X and Y, that can be in one of two states: an inactive state 0, and an active state 1. An inactive Y-type molecule gets activated by colliding and reacting with an active X-type molecule. In the 3MC model formulation, X and Y are represented by two nodes each comprising a two-state discrete-time Markov chain with states '0' (inactive) and '1' (active). Furthermore, the transition probability from state 0 to state 1 in the Y node is modulated by the states of node X.

We wish to derive an expression for the probability that an inactive Y-type molecule gets activated at time n, i.e. $P(Y[n] = 1|Y[n-1] = 0) = p_{01}^{Y}[n]$. This corresponds to the probability that an inactive Y collides into at least an X in the time interval $[n-1, n]$ *and* that the X is active *and* that the Y reacts with that X. Let C be the event that Y collides with an X in the time interval $[n-1, n]$, A the event that the X is active, and R the event that Y reacts with that X, we then have using Bayes rule:

$$\begin{aligned} P(Y[n] = 1|Y[n-1] = 0) &= P(C, A, R) \\ &= P(R|C, A)P(C, A) = P(R|C, A)P(A|C)P(C) \end{aligned} \tag{7.8}$$

The probability that X and inactive Y react given that they collided and that X is active, $P(R|C, A)$, is a constant, γ, that depends on the kinetic energy due to the relative motion of X and Y as well as temperature. The probability that X is active given that it collided with a Y, $P(A|C)$, is 1 if X is in the active state (state 1) and 0 if X is in the inactive state (state 0). We therefore have:

$$P(R|C, A) = \gamma \tag{7.9}$$

$$P(A|C, X) = \begin{cases} 0 & : \quad X = 0 \\ 1 & : \quad X = 1 \end{cases} \tag{7.10}$$

The probability that inactive Y collides with at least an X in the time interval $[n-1, n]$, $P(C)$, can be derived by considering the complement of C, $\bar{C}$: the event that inactive Y does not collide into any X in the time interval $[n-1, n]$. In order to get an expression for $P(\bar{C})$, we first define V to be the volume in which Y and X lie and N_X the total number of X species in V in the time interval $[n-1, n]$ (generally N_X will be a constant over all time). Let v be the volume swept by a given Y molecule in the time interval $[n-1, n]$. The relative volume swept by a Y molecule in the time interval $[n-1, n]$ is therefore given by $v_r \equiv \frac{v}{V}$.

In order to get an expression for the collision probability, we make two basic assumptions: (1) that X is uniformly distributed across V so that the probability that a given X is in the sweeping volume of Y is v_r and (2) that different X molecules are independent of each other. An X-Y collision occurs in the time interval $[n-1, n]$ if and only if an X is present in the sweeping volume of Y in that time interval. As a result, the probability that a given X will collide with Y is simply v_r. Using these assumptions, we get:

$$P(\bar{C}) = (1 - v_r)^{N_X} \tag{7.11}$$

and therefore:

$$P(C) = 1 - (1 - v_r)^{N_X} \tag{7.12}$$

Note that the above probability is the probability that a given Y collides with *at least* one X in $[n-1, n]$. We can guarantee that Y collides with *at most one* X in that interval by sampling at a high enough rate so that the sweeping volume is small enough such that the probability that more than one X exists in the sweeping volume is negligible. Constraints on v_r that guarantee an upper bound on this probability will be given in the next section. Adding this constraint, the above expression for $P(C)$ simplifies to the following:

$$P(C) = N_X v_r \tag{7.13}$$

Combining the above expressions, we get:

$$\boxed{P(Y[n] = 1 | Y[n-1] = 0, X[n-1]) = \begin{cases} 0 & : \quad X[n-1] = 0 \\ \gamma N_X v_r & : \quad X[n-1] = 1 \end{cases}} \tag{7.14}$$

■ 7.5.2 Associative interactions

Consider the case where X and Y in the previous subsection can only associate without reacting. Specifically, if Y can only bind to active X (i.e. X in state one), the transition probability would then just be the probability that Y collides with X and X is active. Therefore, we have in this case:

$$\boxed{P(Y[n] = 1 | Y[n-1] = 0, X[n-1]) = \begin{cases} 0 & : \quad X[n-1] = 0 \\ N_X v_r & : \quad X[n-1] = 1 \end{cases}} \tag{7.15}$$

■ 7.5.3 Conformational activation

Molecules associated as part of a complex that undergo activation as a result of a conformational change or modification of the complex react without colliding since they are already attached to the complex. As a result, $f()$ in this case is a much simpler function that is directly proportional to the probability that the complex is in the active state. Specifically,

if Y above was part of the complex X which can be in either the inactive (0) or active (1) state and if Y gets activated upon activation of the complex, we then have:

$$P(Y[n] = 1|Y[n-1] = 0) = P(A, R) = P(R|A)P(A) \tag{7.16}$$

i.e. the probability that Y gets activated is given by the probability that X is in the active state and that it reacts with X. As a result, following the same derivation as in the previous subsections, we get:

$$\boxed{P(Y[n] = 1|Y[n-1] = 0, X[n-1]) = \begin{cases} 0 & : \quad X[n-1] = 0 \\ \gamma & : \quad X[n-1] = 1 \end{cases}} \tag{7.17}$$

■ 7.5.4 Simplified notation

Using the definition of $f()$ described in this section, we can simplify the notation presented in Section 7.2 by drawing non-zero state interactions as dashed arrows from the originating state to the transition probability as shown in Figure 7-2 below where p_{yf} is non-zero only when X is in state X_1. For simplicity, the self loops on each state are not drawn but can be readily derived using the other transition probabilities.

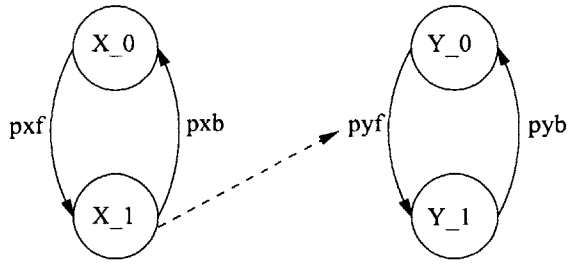


Figure 7-2. Simplified notation for the Markov chains model. For simplicity, the self loops on each state are not drawn.

■ 7.6 Revisiting the Single Event Assumption

In this section, we derive constraints on v_r such that the probability that Y collides with more than one X in the time interval $[n-1, n]$, $P(C > 1)$, is negligible. Formally, $P(C > 1)$ is given by:

$$\begin{aligned} P(C > 1) &= 1 - (1 - v_r)^{N_X} - \binom{N_X}{1} v_r (1 - v_r)^{N_X - 1} \\ &= 1 - (1 - v_r)^{N_X} - N_X v_r (1 - v_r)^{N_X - 1} \end{aligned} \tag{7.18}$$

We would like to choose v_r such that $P(C > 1) \approx 0$. Therefore:

$$1 - (1 - v_r)^{N_X} - N_X v_r (1 - v_r)^{N_X - 1} = 1 - (1 - v_r)^{N_X - 1}(1 + v_r(N_X - 1)) \equiv g(v_r) \approx 0$$

Or more precisely we would like to derive a condition on v_r such that $g(v_r) < \epsilon$ for some small value ϵ. Note that $v_r = 0$ is a root of $g(v_r)$. The Taylor series expansion of $g(v_r)$

around this root is:

$$\begin{aligned} g(v_r) &= g(0) + g'(0)v_r + \frac{g''(0)}{2!}v_r^2 + \frac{g^{(3)}(0)}{3!}v_r^3 ... \\ &= 0 + 0 + \frac{N_X(N_X-1)}{2!}v_r^2 + \frac{-2N(N-1)(N-2)}{3!}v_r^3 + ... \end{aligned}$$

Using the first three terms in the Taylor series expansion, we get the following condition:

$$\begin{aligned} \frac{N_X(N_X-1)}{2!}v_r^2 &< \epsilon \\ v_r < \sqrt{\frac{2\epsilon}{N_X(N_X-1)}} &\approx \frac{\sqrt{2\epsilon}}{N_X} \end{aligned}$$

The condition on v_r is therefore:

$$\boxed{v_r < \frac{\sqrt{2\epsilon}}{N_X}} \tag{7.19}$$

For example, for $\epsilon = 10^{-3}$, v_r has to be smaller than $\frac{0.044}{N_X}$.

■ 7.7 Model Approximation: State Dependency versus Probabilistic Dependency

In this section, we consider an approximation to the 3MC model that decouples the chains while maintaining a dependency between the distributions of the different states. We name this new model the *a priori* Markov Modulated Markov Chains (a3MC) model.

Consider the 3MC model described in Section 7.1 with $f()$ as described in Section 7.5. We have:

$$\boxed{P(Y[n]=1|Y[n-1]=0, X[n-1]) = \begin{cases} 0 & : \quad X[n-1]=0 \\ \gamma N_X v_r & : \quad X[n-1]=1 \end{cases} \equiv f(Y[n-1]=0, X[n-1])}$$

where Y and X are dependent random variables. One can obtain an expression for $P(Y[n] = 1|Y[n-1] = 0)$ by summing over the different values of $X[n-1]$, we have:

$$\begin{aligned} P(Y[n]=1|Y[n-1]=0) &= \sum_j P(Y[n]=1|Y[n-1]=0, X[n-1]=j) \\ &\quad \times P(X[n-1]=j|Y[n-1]=0) \\ &= \sum_j f(Y[n-1]=0, X[n-1]=j) P(X[n-1]=j|Y[n-1]=0) \\ &= \gamma N_X v_r P(X[n-1]=1|Y[n-1]=0) \end{aligned} \tag{7.20}$$

■ 7.7.1 The a3MC model

We define the *a priori* Markov Modulated Markov Chains (a3MC) model as a stochastic process composed of a discrete-time Markov chain whose transition probabilities are dependent on the *a priori* state probabilities of a second Markov chain. Specifically, using the example laid out in Section 7.5, in the a3MC model, the transition probability in chain Y from state 0 to state 1 is *independent* of the realizations of the random process X but is

related to its statistics as follows:

$$\begin{aligned}\hat{P}(Y[n]=1|Y[n-1]=0) &\equiv \sum_j f(Y[n-1]=0, X[n-1]=j)P(X[n-1]=j) \\ &= \gamma N_X v_r P(X[n-1]=1) \end{aligned} \tag{7.21}$$

Said differently, the transition probability from state 0 to state 1 in chain Y in the a3MC model is equal to the expected value of the function $f(Y[n-1]=0, X[n-1])$ with respect to $X[n-1]$ and is independent of chain X. Since the chains in the a3MC model are independent of each other, stochastic simulations of the chains can be performed independently, i.e. in order to obtain a stochastic simulation of a particular chain of interest, one needs to only simulate that chain using the transition probabilities that can be precomputed. This is in contrast to the 3MC model where the chains are dependent and therefore in order to stochastically simulate a chain, all the chains in the model have to be simulated. Recursive formulae for the state occupancy probabilities of the individual Markov chains can be obtained as described in Section 7.3.

■ 7.7.2 Relationship between the 3MC and the a3MC

The a3MC model is an approximation to the 3MC model where the chains are independent but the transition probabilities take on values derived from the state-dependency in the 3MC model. Specifically the transition probabilities are given by the average with respect to $X[n]$ of the function describing the state dependency in the 3MC model, $f(X[n])$ The relationship between the two models can be examined by computing the state probabilities. In the 3MC model, we have:

$$\begin{aligned} P(Y[n]=m) = \\ \sum_p \sum_j P(Y[n]=m|Y[n-1]=p, X[n-1]=j)P(Y[n-1]=p, X[n-1]=j) \\ = \sum_p \sum_j f(Y[n-1]=p, X[n-1]=j)P(X[n-1]=j|Y[n-1]=p)P(Y[n-1]=p) \\ = \sum_p \sum_j P(Y[n]=m, X[n-1]=j|Y[n-1]=p)P(Y[n-1]=p) \\ = \sum_p P(Y[n]=m|Y[n-1]=p)P(Y[n-1]=p) \end{aligned} \tag{7.22}$$

whereas in the a3MC model, we have:

$$\begin{aligned} \hat{P}(Y[n]=m) = \\ \sum_p \sum_j f(Y[n-1]=p, X[n-1]=j)P(Y[n-1]=p, X[n-1]=j) \\ = \sum_p \sum_j f(Y[n-1]=p, X[n-1]=j)P(X[n-1]=j)P(Y[n-1]=p) \end{aligned} \tag{7.23}$$

where $f(Y[n-1], X[n-1])$ is the function defining the state dependency in the 3MC model.

It should be clear from the above that the only difference between equations 7.22 and 7.23 is that the conditional probability $P(X[n-1]|Y[n-1])$ in the 3MC is approximated in the a3MC by the unconditional probability $P(X[n-1])$. Specifically, let $\mathbf{p_Y[n]}$ be the row vector containing the state probabilities of chain Y at time n and $\mathbf{A}_{YY}[n]$ be the transition matrix of the Markov chain in node Y at time n. We define the conditional transition matrix $\mathbf{A}_{YY|X}[n]$ as a block matrix composed of the vertical concatenation of the conditional transition matrices of chain Y given chain X is in state j for $j = 1, ..., p$ where p is the number of states in chain X. $\mathbf{A}_{YY|X}[n]$ has $m \times p$ rows and m columns where m is the number of states in chain Y and p is the number of blocks, i.e. we have:

$$\mathbf{A}_{YY|X}[n] \equiv \begin{bmatrix} \mathbf{A}_{YY|X=1}[n] \\ \mathbf{A}_{YY|X=2}[n] \\ \cdots \\ \mathbf{A}_{YY|X=j}[n] \\ \cdots \\ \mathbf{A}_{YY|X=p}[n] \end{bmatrix} \tag{7.24}$$

where

$$\mathbf{A}_{YY|X=j}[n] = \begin{bmatrix} p_{00}^{Y|X=j}[n] & \cdots & p_{0m}^{Y|X=j}[n] \\ p_{10}^{Y|X=j}[n] & \cdots & p_{1m}^{Y|X=j}[n] \\ \cdots & \cdots & \cdots \\ p_{m0}^{Y|X=j}[n] & \cdots & p_{mm}^{Y|X=j}[n] \end{bmatrix} \tag{7.25}$$

Let $\mathbf{B}_{X|Y}[n]$ be another block matrix composed of the horizontal concatenation of p diagonal matrices having m rows and m columns with entries equal to the conditional probability of X being in state p at time n given Y is in state m at time n, i.e.:

$$\mathbf{B}_{X|Y}[n] \equiv \begin{bmatrix} \mathbf{B}_{X=1|Y}[n] & \cdots & \mathbf{B}_{X=j|Y}[n] & \cdots & \mathbf{B}_{X=p|Y}[n] \end{bmatrix} \tag{7.26}$$

where

$$\mathbf{B}_{X=j|Y}[n] = \begin{bmatrix} P(X[n]=1|Y[n]=1) & 0 & .. & 0 \\ 0 & P(X[n]=1|Y[n]=2) & .. & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & ... & P(X[n]=1|Y[n]=m) \end{bmatrix} \tag{7.27}$$

Using these newly defined matrices, we have the following state evolution equation for the 3MC model:

$$\mathbf{p}_Y[n] = \mathbf{p}_Y[n-1]\mathbf{B}_{X|Y}[n]\mathbf{A}_{YY|X}[n] \tag{7.28}$$

while the state evolution equation for the a3MC model is given by:

$$\mathbf{\hat{p}_Y[n]} = \mathbf{\hat{p}_Y[n-1]}\mathbf{\hat{B}_{X|Y}[n]}\mathbf{\hat{A}_{YY|X}[n]} \tag{7.29}$$

where:

$$\mathbf{\hat{A}_{YY|X}[n]} \equiv \mathbf{A}_{YY|X}[n] \tag{7.30}$$

The a3MC model converges to the 3MC model when:

$$\mathbf{B}_{X|Y}[n] \approx \mathbf{\hat{B}_X[n]} \tag{7.31}$$

i.e. when:

$$P(X[n-1]=j, Y[n-1]=p) \approx P(X[n-1]=j)P(Y[n-1]=p) \tag{7.32}$$

This condition is satisfied when N_X is large enough since under such a regime, Y can collide into any X with a uniform probability and therefore the expected value of the state of the X molecule it collides into is $P(X[n]=1)$ and thus in the limit of large N_X, the 3MC model converges to the a3MC model.

■ 7.7.3 Advantages of the a3MC Model

As noted earlier in this section, the advantage of the a3MC model is that it is much easier to compute and therefore the algorithm is more efficient since the individual chains can be simulated separately. In a model with hundreds of nodes and interactions, the computational savings can be very significant. Furthermore, since the 3MC converges to the a3MC when the number of substrate molecules (N_X) is large enough, hybrid simulations can be performed on a large model by using the a3MC in subsets of the model where the number of substrate molecules is large enough and using the 3MC model in subsets where the number of molecules is small.

■ 7.8 Reconciling the Stochastic and Deterministic Formulations

The framework described in this chapter provides a stochastic formulation for modeling the behavior of biological networks. This framework is particularly valuable in situations where the number of molecules in the system is small enough that a deterministic formulation which relies on a continuum assumption is no longer valid. However, in the limit of large N (number of molecules), we would like the formalism to reduce to conventional mass action kinetics. Stated differently, we would like the expected value of the state probabilities in the model to equal the result obtained using the deterministic formalism of mass action kinetics.

In this section, we derive relationships between the microscopic reaction probability γ defined in our model and the macroscopic deterministic rate constant k such that the average in our model matches the deterministic result. In the following subsections, we address two types of reactions: unimolecular reactions, i.e. reactions that involve only one species that spontaneously gets transformed (in our model these reactions correspond to constant transition probabilities) and bimolecular reactions, that is reactions that involve two species.

■ 7.8.1 Unimolecular reactions

Consider the following reaction:

$$X_0 \xrightarrow{k_x} X_1 \tag{7.33}$$

where X_0 gets spontaneously transformed to X_1 at a rate of k_x measured in s^{-1}. Using mass action kinetics, we can write a differential equation that governs the dynamics of the reaction in 7.33:

$$\frac{d[X_1]}{dt} = k_x[X_0] \tag{7.34}$$

as well as a conservation equation:

$$[X_0] + [X_1] = [X_{tot}] \tag{7.35}$$

where brackets indicate concentrations in M and $[X_{tot}]$ is a constant and equal to the total concentration of species X. Using forward Euler, equation 7.34 can be approximated as follows:

$$\frac{\Delta[X_1]_t}{\Delta t} \approx k_x[X_0]_t \tag{7.36}$$

$$\frac{[X_1]_{t+\Delta t} - [X_1]_t}{[X_0]_t} \approx k_x \Delta t \tag{7.37}$$

The Markov modulated Markov chains model of the reaction in 7.33 leads to:

$$p_{X_1}[n+1] = p_{xf} \times p_{X_0}[n] + p_{X_1}[n] \tag{7.38}$$

where:

$$p_{X_0}[n] + p_{X_1}[n] = 1 \tag{7.39}$$

Multiplying the equation 7.38 by the total number of X molecules, N_X and taking the expectation with respect to X, we get:

$$N_{X_1}[n+1] = p_{xf} \times N_{X_0}[n] + N_{X_1}[n] \tag{7.40}$$

$$\frac{N_{X_1}[n+1] - N_{X_1}[n]}{N_{X_0}[n]} = p_{xf} \tag{7.41}$$

where $N_{X_0}[n]$ and $N_{X_1}[n]$ are the expected number of molecules in state X_0 and X_1 respectively. Furthermore, using the definition of concentration, we have:

$$\frac{N_{X_0}[n]}{A_v V} = [X_0]_n \tag{7.42}$$

$$\frac{N_{X_1}[n]}{A_v V} = [X_1]_n \tag{7.43}$$

where V is the total volume containing the molecules and A_v is Avogradro's number: the total number of molecules contained in one mole. Equating the two expressions above, we therefore get:

$$\frac{N_{X_1}[n+1] - N_{X_1}[n]}{N_{X_0}[n]} = \left.\frac{[X_1]_{t+\Delta t} - [X_1]_t}{[X_0]_t}\right|_{t=n} \tag{7.44}$$

$$\boxed{p_{xf} \approx k_x \Delta t} \tag{7.45}$$

The derivation above assumed that the reaction was unidirectional. It is straightforward to show that the expression in equation 7.45 holds even in the case of bidirectional reactions where a similar expression for the backward probability involving the backward reaction rate can also be derived.

■ 7.8.2 Bimolecular reactions

We now consider the following bimolecular reaction:

$$Y_0 + X_1 \xrightarrow{k_{yx}} Y_1 + X_1 \tag{7.46}$$

where Y_0 reacts with X_1 to create Y_1 without consuming X_1 at a rate k_{yx} measured in $M^{-1}s^{-1}$. Following the same methodology as in the previous subsection, we have, using mass action kinetics:

$$\frac{d[Y_1]}{dt} = k_{yx}[X_1][Y_0] \tag{7.47}$$

and

$$[Y_0] + [Y_1] = [Y_{tot}] \tag{7.48}$$

For a small period of time Δt, we have

$$\Delta[Y_1] \approx k_{yx}[X_1]_t[Y_0]_t\Delta t \tag{7.49}$$

$$\frac{\Delta[Y_1]}{[Y_0]_t} \approx k_{yx}[X_1]_t\Delta t \tag{7.50}$$

$$\frac{[Y_1]_{t+1} - [Y_1]_t}{[Y_0]_t} \approx k_f[X_1]_t\Delta t \tag{7.51}$$

Using the Markov modulated Markov chains model formulation, we get:

$$p_{Y_1}[n+1] = p_{yf}[n] \times p_{Y_0}[n] + p_{Y_1}[n] \tag{7.52}$$

$$N_{Y_1}[n+1] = p_{yf}[n] \times N_{Y_0}[n] + N_{Y_1}[n] \tag{7.53}$$

$$N_{Y_1}[n+1] - N_{Y_1}[n] = \gamma N_X v_r p_{X_1}[n] \times N_{Y_0}[n] \tag{7.54}$$

$$\frac{N_{Y_1}[n+1] - N_{Y_1}[n]}{N_{Y_0}[n]} = \gamma v_r N_{X_1} \tag{7.55}$$

where we have taken the expected value over both Y and X and $N_{X_1}[n]$, $N_{Y_0}[n]$, and $N_{Y_1}[n]$ are the expected numbers of molecules X_1, Y_0, and Y_1 at time n respectively. Using the definition of concentration and setting the two expressions above equal, we get:

$$\boxed{\gamma \approx \frac{k_y \Delta t}{v_r A_V V}} \tag{7.56}$$

Here again, it is straightforward to show that the expression in equation 7.56 holds even in the case of bidirectional reactions where a similar expression for the backward probability involving the backward reaction rate can also be derived.

■ 7.9 Example

To illustrate the models described in this chapter, consider the system composed of the following simple reactions:

$$X_0 \xrightarrow{k_{xf}} X_1 \tag{7.57}$$

$$X_1 \xrightarrow{k_{xb}} X_0 \tag{7.58}$$

$$Y_0 + X_1 \xrightarrow{k_{yf}} Y_1 + X_1 \tag{7.59}$$

$$Y_1 \xrightarrow{k_{yb}} Y_0 \tag{7.60}$$

where two species X and Y each exist in one of two states X_0, X_1, and Y_0, Y_1 respectively. X_0 gets converted to X_1 spontaneously at the rate k_{xf} and X_1 gets converted back to X_0 at the rate k_{xb}. Y_0, on the other hand, reacts with X_1 to create Y_1 at the rate k_{yf} without consuming X_1 and Y_1 spontaneously converts back to Y_0 at the rate k_{yb}. If we use a deterministic formulation (mass action kinetics) of the chemical dynamics, we get the following set of differential equations:

$$\frac{d[X_1]}{dt} = -\frac{d[X_0]}{dt} = k_{xf}[X_0] - k_{xb}[X_1] \tag{7.61}$$

$$\frac{d[Y_1]}{dt} = -\frac{d[Y_0]}{dt} = k_{yf}[X_1][Y_0] - k_{yb}[Y_1] \tag{7.62}$$

Figure 7-3 shows the Markov chains representation of the system where p_{xf}, p_{xb}, and p_{yb} are constant transition probabilities and p_{yf} is modulated by the X chain. For simplicity the self loops on the states are not shown. Specifically, we have $p_{xf} = k_{xf}\Delta t$, $p_{xb} = k_{xb}\Delta t$,

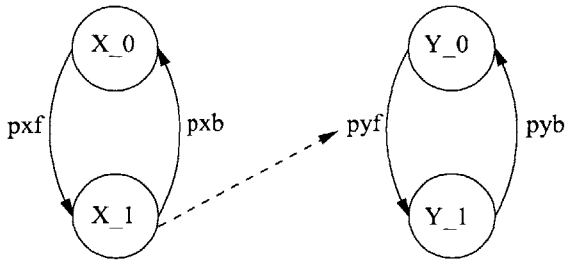


Figure 7-3. Markov chains model representation.

and $p_{yb} = k_{yb}\Delta t$ while

$$\boxed{p_{yf}[n] = \begin{cases} 0 & : \quad X[n] = X_0 \\ \gamma N_X v_r & : \quad X[n] = X_1 \end{cases}} \tag{7.63}$$

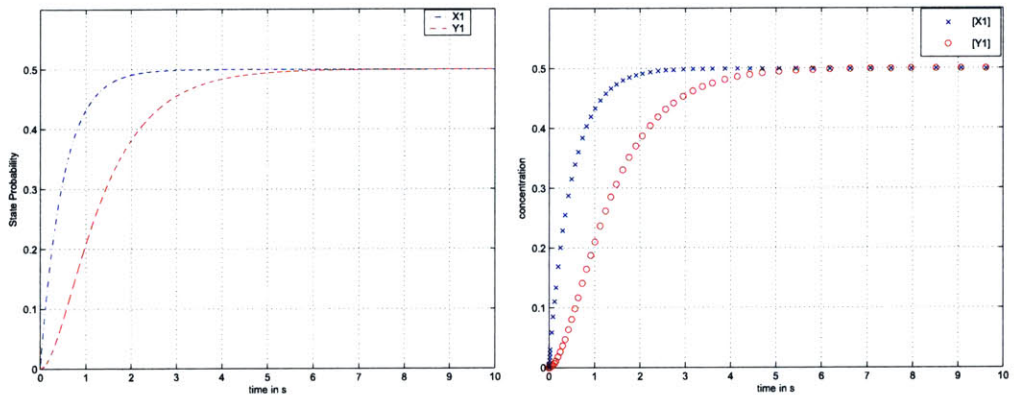
where $\gamma N_X v_r = k_{yf}\Delta t \frac{N_X}{A_V V} = k_{yf}\Delta t[X]$. $[X]$ is the total concentration of species X. We have simulated the behavior of this system under the different models using the parameters shown in Table 7.1.

k_{xf}	=	1 s^{-1}	$[X]$	=	1 M
k_{xb}	=	1 s^{-1}	N_X	=	100
k_{yf}	=	1 $M^{-1}s^{-1}$	N_Y	=	100
k_{yb}	=	0.5 s^{-1}	V	=	100
Δt	=	0.001 s			

Table 7.1. Example parameters.

■ 7.9.1 State probabilities

Figure 7-4-a shows the computed a3MC Markov chain model state probabilities as a function of time. The solution to the differential equations obtained using the deterministic



(a) a3MC Model (b) Deterministic Model

Figure 7-4. Time evolution dynamics.

formulation is shown in Figure 7-4-b. The results in Figure 7-4 show that the a3MC model state probabilities converge to the deterministic solution as expected by design.

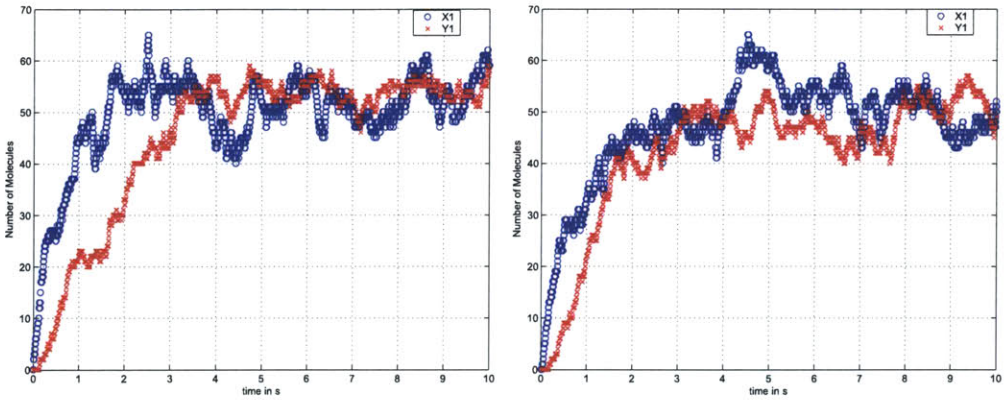
■ 7.9.2 Stochastic realizations

We performed Monte Carlo simulations of 100 molecules of each species using both the 3MC model and the a3MC model. Examples of time realizations are shown in Figure 7-5.

These simulations, as compared to the state probabilities and deterministic solution (Figure 7-4) illustrate the value of the stochastic simulations since they provide time-series that can be directly compared to biological experiments. Specifically, they provide a mean to quantify the expected intrinsic fluctuations in the biological system being studied and as a result, biological fluctuations can be separated from external fluctuations resulting from the experimental apparatus. The expected biological variance can be quantified by looking at the distributions of the time series.

■ 7.9.3 Distributions

Distributions were also computed using 1000 independent experiments consisting of simulations of 100 molecules at different time points. Figure 7-6 show the histograms for both X_1 and Y_1 at $0.5s$, $1s$, and $10s$ obtained using the 3MC model. The a3MC model histograms are shown in Figure 7-7.



(a) 3MC (b) a3MC

Figure 7-5. Examples of time realizations.

The figures illustrate how the variance as well as the shape of the distributions change depending on the molecules examined and the time points measured. Being able to predict the variance in the measurement as well as its variation with time can have a deep impact on the way biological experiments are performed and interpreted. Furthermore, the shape of the distributions for the 3MC and a3MC models are similar confirming that 100 molecules is large enough for the a3MC model approximation to hold.

■ 7.10 Related Models

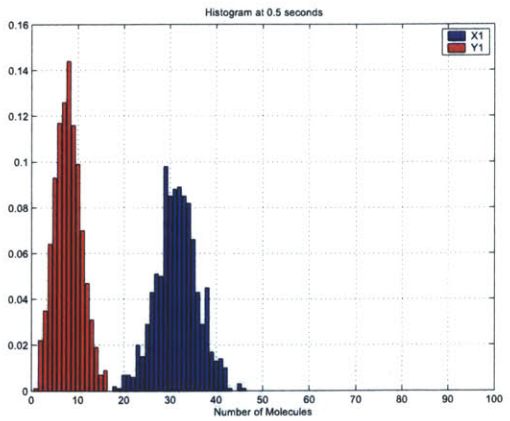
■ 7.10.1 The Gillespie algorithm

c_{xf}	=	1
c_{xb}	=	1
c_{yf}	=	$k_{yf}/V = 0.01$
c_{yb}	=	0.5
N_X	=	100
N_Y	=	100
V	=	100

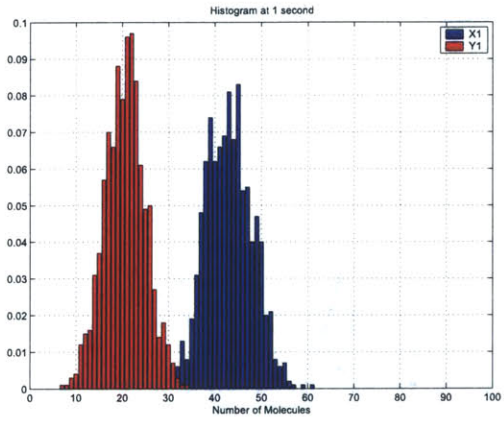
Table 7.2. Gillespie algorithm parameters.

Simulations of the example presented in the previous section were also performed using the Gillespie algorithm (discussed in detail in the previous chapter) with the parameters shown in Table 7.2. The results are shown in Figures 7-8 and 7-9.

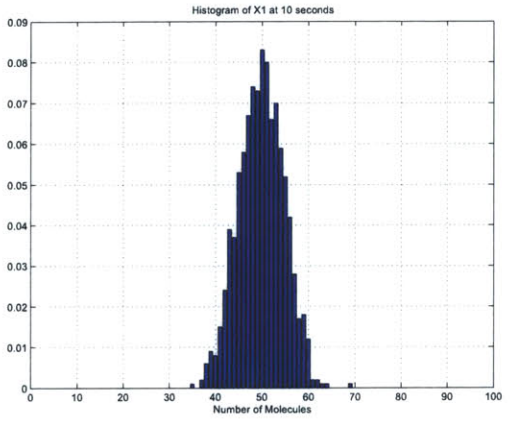
As can be seen from the figures the results are very similar to the results obtained using our models. However, the Markov model formulation presented in this thesis differs from the Gillespie algorithm in important ways. While the Gillespie algorithm adopts a reaction-centric approach, our model is molecule-centric. Specifically, in the Gillespie algorithm, it is assumed that only one reaction occurs in the volume at a given time. The time and identity of the reaction are then determined according to their probability distribution (obtained from the number of all reactant molecules in the volume and the propensity of



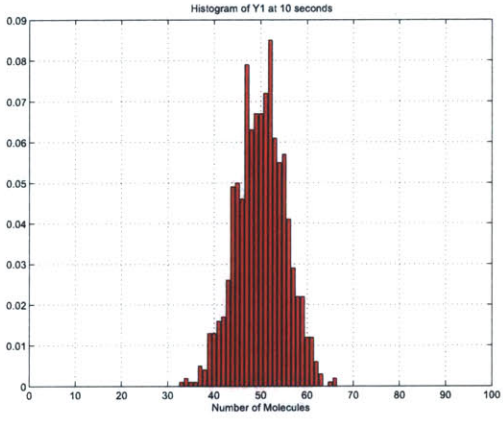
(a) $t = 0.5$ s



(b) $t = 1$ s



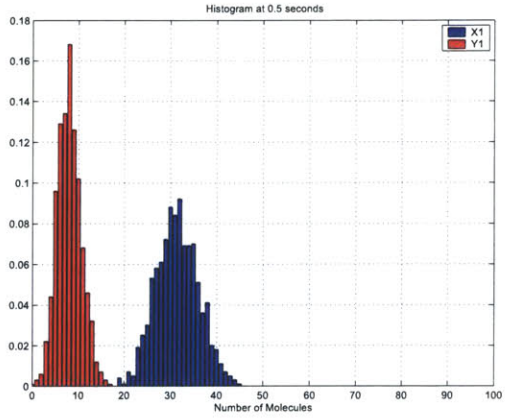
(c) X_1, $t = 10$ s



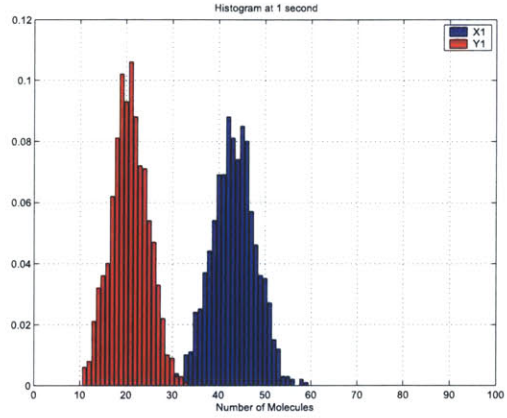
(c) Y_1, $t = 10$ s

Figure 7-6. 3MC model histograms.

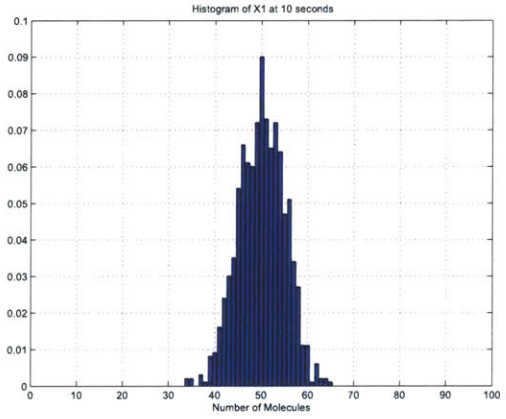
the reactions). Time is therefore a stochastic variable that needs to be computed. In our model, however, time is *discrete and sampled uniformly.* Reactions are viewed from the standpoint of individual molecules and the assumption is that at most one reaction occurs *per molecule* in a time interval $[n-1, n]$ however there are no restrictions on the number of simultaneous reactions that can occur if they involve different molecules. Since our method is molecule-centric, individual molecules in the volume can be tracked. This is not possible using the Gillespie algorithm since the algorithm simulates the occurrence of reactions not individual molecules involved in those reactions. Finally, the Markov chains model with its graphical notation is very intuitive and therefore formulating a model (defining the topology and interactions) can be done very easily using a non-mathematician intuitive understanding of the system under study, a property that neither the Gillespie algorithm nor the deterministic formulation have.



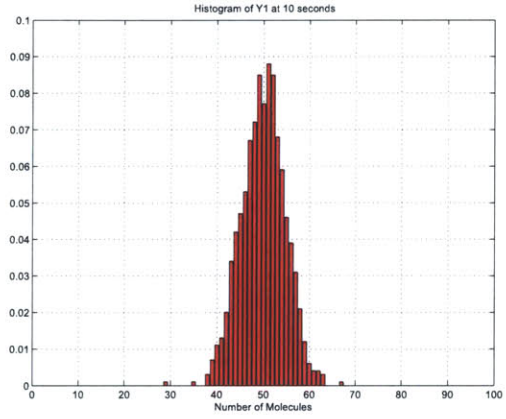
(a) $t = 0.5$ s



(b) $t = 1$ s



(c) X_1, $t = 10$ s



(d) Y_1, $t = 10$ s

Figure 7-7. *a*3MC model histograms.

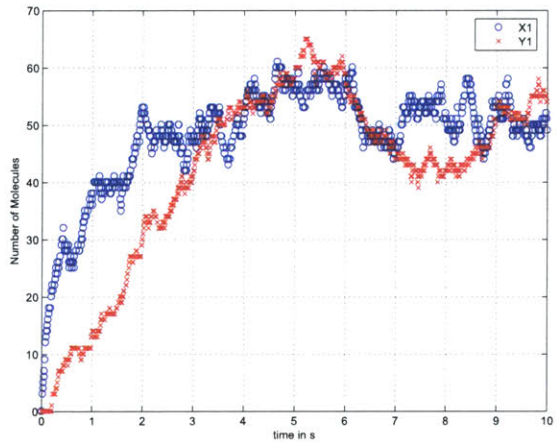
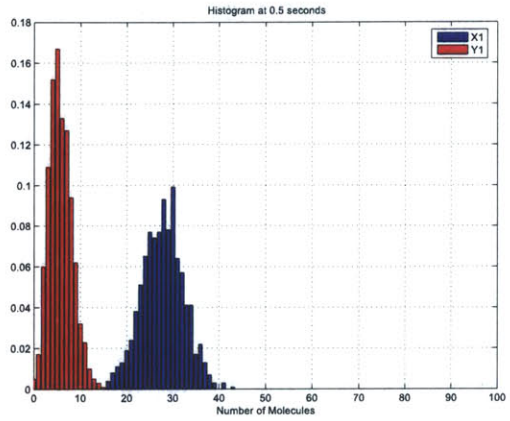
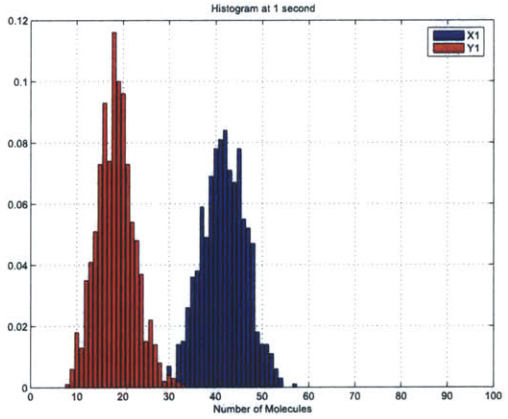


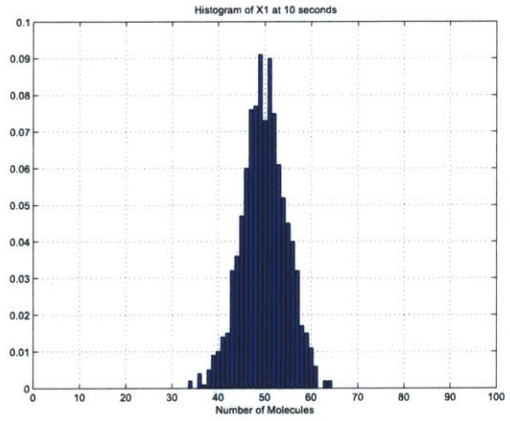
Figure 7-8. Examples of a time realization using the Gillespie algorithm.



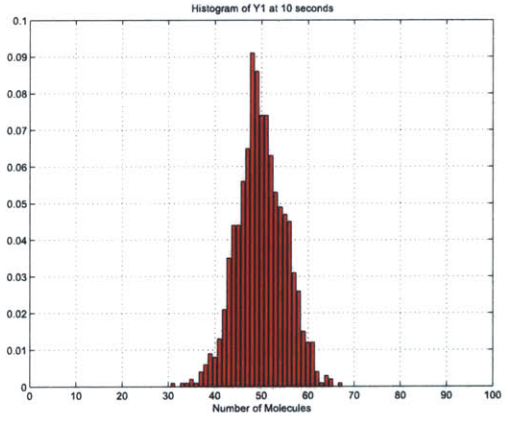
(a) $t = 0.5$ s



(b) $t = 1$ s



(c) X_1, $t = 10$ s



(d) Y_1, $t = 10$ s

Figure 7-9. Gillespie algorithm histograms.

■ 7.10.2 Stochastic cellular automata

A variety of interacting Markov chains or more generally stochastic cellular automata models have previously been formulated and described, most in the context of studying parallel systems, i.e. systems composed of interacting modules. The most recent of these models and probably the closest to our model are the influence model described in [9] and the stochastic automata network (SAN) in [104]. All three models, ours, the influence model and the SAN define a network of interacting Markov chains. However our model diverges from both of these models in very important ways. In the influence model [9], a node randomly selects a neighboring node according to an influence probability. Once the neighboring node is selected, it fixes the transition matrix to be used by the node being influenced. As a result, nodes interact with other nodes by affecting *all* transition probabilities in the receiving nodes. In contrast in our model, specific transitions in a given chain are influenced by *different* states in other nodes and at any given time different transitions in a node can be affected by different nodes in the model. This is not allowed in the influence model. Furthermore, in the influence model, the influence from neighboring nodes is constrained to take a multilinear form, while the interaction between Markov chains in our model can be non-linear. The difference between our model and the SAN is more subtle. In fact, the 3MC model presented here can be thought of as a special case of the SAN where interactions have been specialized to biological phenomena. However, the a3MC is very different from the SAN since the interaction is expressed through the unconditional probability of being in a certain state while in the SAN the influence is expressed through the conditional probability of being in a given state. As a result, while it can be shown that the SAN can always be expanded to a higher order Markov chain with much larger state-space, the a3MC model presented here is not always expandable to a higher order Markov chain. For a more detailed literature review of related models, the reader is referred to [9].

■ 7.11 A Unified Framework for Modeling the Dynamics of Signaling Pathways

In this chapter, we have formulated a stochastic framework for modeling biological signaling dynamics based on interacting Markov chains. We defined a Markov Modulated Markov Chains (3MC) model and explored an efficient way to approximate it by decoupling the Markov chains while maintaining the probabilistic influence. We call this new simplified model the *a priori* Markov Modulated Markov Chain (a3MC) model. These models can be viewed as a unified framework for simultaneously studying the fluctuations of signaling pathways (stochastic behavior) and computing their average behavior therefore allowing modeling at multiple resolutions within the same framework. Specifically, computing the state probabilities of the individual chains in the model is equivalent to computing solutions to deterministic models based on Mass Action kinetics while performing Monte Carlo simulations of the chains provide stochastic behavior. Furthermore, using the different versions of the model, one can perform hybrid simulations where in parts of the model, only average behavior is sought and therefore state probabilities are computed while in other parts, the a3MC approximation holds and therefore stochastic simulations are performed on the relevant chains only. Other sections of the model may not need to be simulated but should be graphically included for completeness, and finally, full stochastic behavior of some proteins in the model can be obtained by using the 3MC to simulate those chains. The intuitive graphical representation as well as ability to perform different levels of simulations make this modeling framework particularly attractive for studying biological signaling pathways.

Chapter 8

The Mitogen Activated Protein Kinase Cascade

In this chapter, we demonstrate the modeling framework proposed in the previous chapter using a well conserved and studied module: the mitogen activated protein kinase (MAPK) cascade. The goal here is to introduce and validate the model by applying it to a simple module and comparing the model predictions to established results rather than to use it to discover new phenomena which will be explored in the next chapter in the context of bacterial chemotaxis. We start by presenting some background and established properties of the cascade, we then formulate the model and perform simulations. We end this chapter by discussing the implications of the results.

8.1 Background

The mitogen-activated protein kinase (MAPK) cascade is an essential component of a wide variety of cell signaling pathways. It is a series of three conserved kinases (i.e. enzymes that add phosphate groups to other enzymes) organized in a hierarchy and found only in eukaryotes. This set of enzymes plays a role in relaying signals from receptors at the cell surface to regulatory elements inside the cell nucleus. They are involved in a variety of pathways ranging from growth, differentiation, and development to inflammation and programmed cell death. At the top of the hierarchy, activated MAPK kinase kinase (or MAP3K*) serially phosphorylates MAPK kinase (or MEK or MAP2K) at two serine residues. Activated MEK then serially phosphorylates MAPK at a threonine and a tyrosine residue which in turn proceeds to activate downstream signals. The cell also contains phosphatases that dephosphorylate activated kinases. The cascade can thus be perceived as a stand-alone module with the MAP3K activating enzyme as the input signal and the activated MAPK as the output signal. Figure 8-1 gives a schematic representation of the cascade.

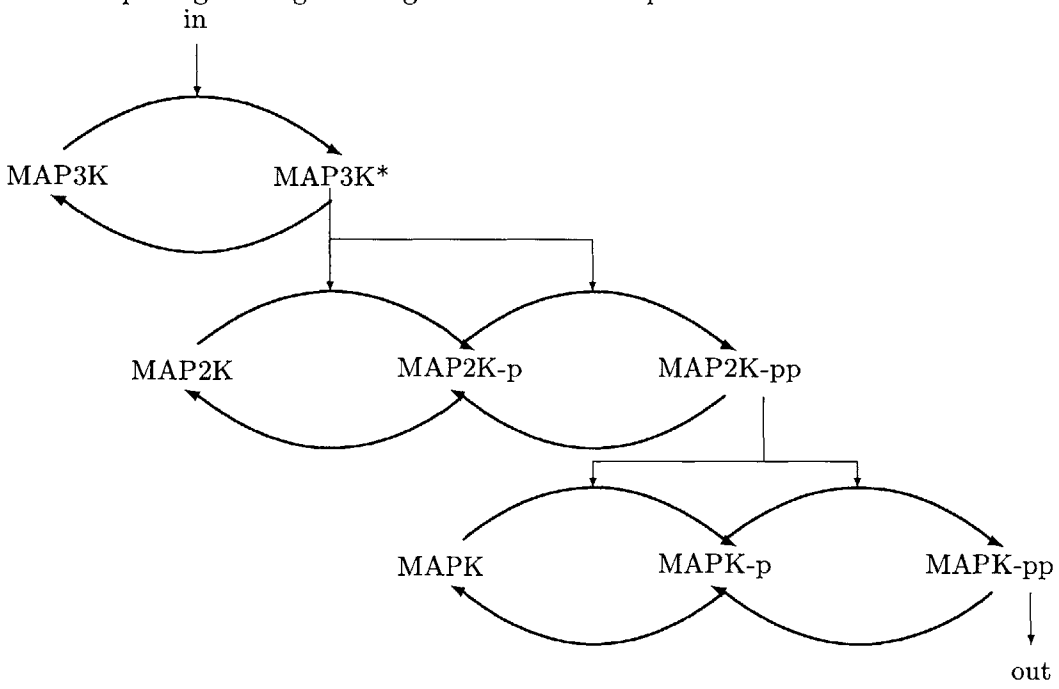


Figure 8-1. Schematic of the MAPK cascade. Forward arrows correspond to activation (phosphorylation for MAP2K and MAPK) while back-arrows correspond to deactivation steps (phosphatase activity in the case of MAP2K and MAPK).

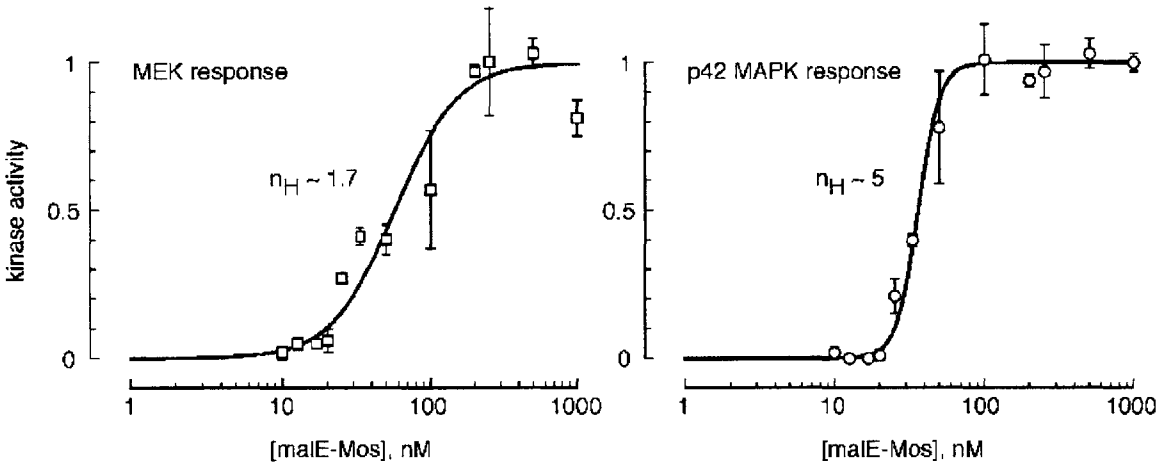
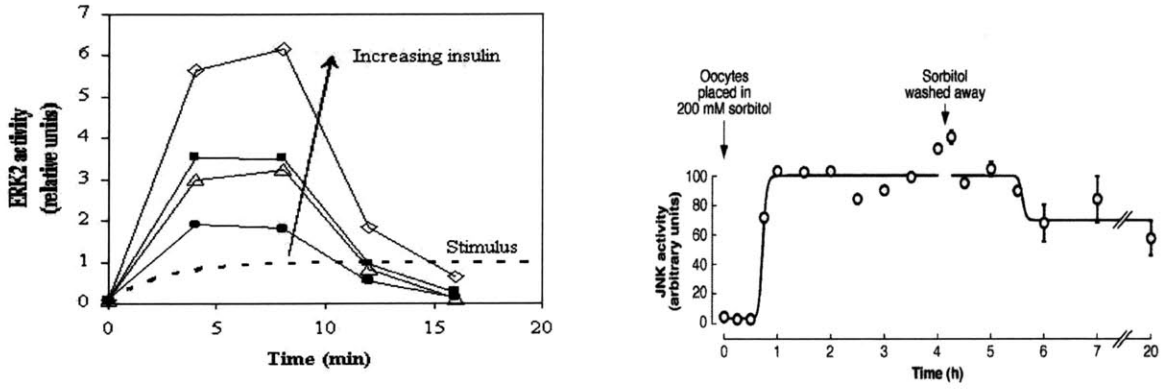


Figure 8-2. Experimental stimulus/response data for MAPK and MAP2K(MEK) activation in Xenopus oocytes reproduced with permission of the authors from [67]. malE-Mos is the relevant MAP3K in this system. For more details regarding the experimental setup, the reader is referred to [67].

■ 8.2 Properties of the MAP Kinase Cascade

Many MAP kinase cascade proteins have been studied and quantitative data can be found in the literature for both steady-state behavior [67] [12] and time behavior [10]. Three important properties implemented by this module have been suggested, namely ultrasensitivity, adaptation, and bistability. Probably the most important property is *ultrasensitivity*, i.e. the ability of the cascade to generate a highly switch-like response to a continuously variable stimulus. Specifically, Huang and Ferrell [67] studied the steady-state response of the cascade in Xenopus oocytes and were able to show that in response to a continuous stimulus, the response at the output of the cascade, i.e. the concentration of activated MAP kinase, was more switch-like than at intermediate stages of the cascade such as the concentration of activated MAP2K. The data is shown in Figure 8-2. In a parallel and complementary manner, Asthagiri and Lauffenburger studied the MAPK cascade in Chinese Hamster ovary cells [10]. They were able to show that there exists a negative feedback mechanism which leads to adaptation of the response, i.e. the time response to a step stimulus generates an output with a peak response followed by an adaptation of the output back to its original value, or close to that value, as if the stimulus was turned back off. The corresponding data is given in Figure 8-3(a). Furthermore, Bagowski and Ferrell [12] suggested that a positive feedback mechanism may occur within the cascade leading to bistability: they were able to measure a different steady-state response if the stimulus was stepped up from a low level than if the stimulus was stepped down from a high level as shown in Figure 8-3(b). In addition to the experimental findings, a kinetic model based on reaction-rate differential equations was first formulated by Huang and Ferrell in [67]. The simplest model they propose which does not include feedback has at steady-state a total of 25 equations and 17 parameters which need to be estimated. Asthagiri and Lauffenburger in [10] used a similar model with negative feedback.



(a) (b)

Figure 8-3. Time evolution data. (a) ERK2 (the relevant MAPK) adaptation in Chinese hamster ovaries obtained from [10] with permission of the authors. Insulin is an activating factor which operates upstream of the cascade. The reader is referred to [10] for experimental details. (b) Time course of JNK (MAPK) activation and inactivation in sorbitol-treated Xenopus oocytes obtained from [12] with permission of the authors. Sorbitol is an activating factor upstream of the cascade. The reader is referred to [12] for experimental details.

■ 8.3 Model Formulation

■ 8.3.1 Topology

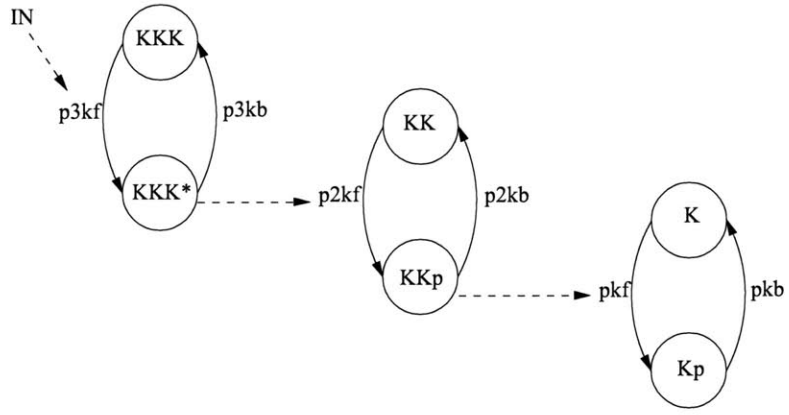


Figure 8-4. Interactive Markov chains model of the mitogen activated protein kinase cascade with a single phosphorylation mechanism.

Figures 8-4 and 8-5 show schematics of the interacting Markov chains model of the MAP kinase cascade using both a single and dual phosphorylation mechanism respectively. Both models have three chains which correspond to the three proteins in the cascade (MAP3K, MAP2K, and MAPK). In the single phosphorylation model, each protein is represented by two states: an inactive state (KKK, KK, and K) and an active state (KKK*, KKp, and Kp). The MAP2K and MAPK chains in the dual phosphorylation model are represented by three states corresponding to the unphosphorylated proteins (KK and K), singly phosphorylated

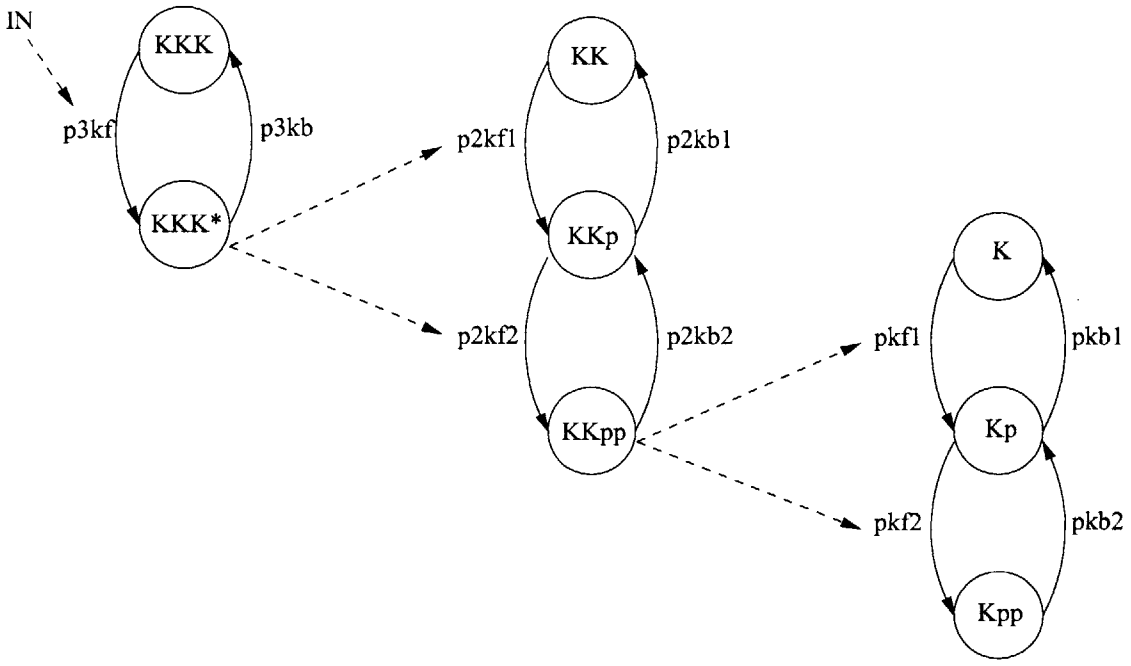


Figure 8-5. Interactive Markov chains model of the mitogen activated protein kinase cascade with a dual phosphorylation mechanism.

proteins (KKp and Kp), and dually phosphorylated proteins (KKpp and Kpp). Note that in this model, we do not differentiate between the different phosphorylation sites in the singly phosphorylated protein states, i.e. these states correspond to singly phosphorylated proteins whether the phosphorylation is on one site or the other.

■ 8.3.2 Interactions

$p_{3kf}[n] =$	$\gamma_{in} v_r N_{in} = k_{in} \Delta t[IN]$
$p_{2kf}[n] =$	$\gamma_{3k*1} v_r N_{3K} p_{3k*}[n-1] = k_{3k} \Delta t[3K] p_{3k*}[n-1]$
$p_{kf}[n] =$	$\gamma_{2kpp} v_r N_{KK} p_{2kpp}[n-1] = k_{2kpp} \Delta t[KK] p_{2kpp}[n-1]$
$p_{3kb}[n] =$	$k_{p3k} \Delta t[P3k]$
$p_{2kb}[n] =$	$k_{p2k} \Delta t[P2k]$
$p_{kb}[n] =$	$k_k \Delta t[Pk]$

Table 8.1. Transition probabilities for the model in Figure 8-4. For bimolecular reactions, the relevant γ is obtained from the relevant k (in $M^{-1}s^{-1}$) using the expression $\gamma = \frac{k \Delta t}{v_r A_V V}$.

The models interactions described by the full transition probabilities are shown in Table 8.1 for the single phosphorylation model and in Table 8.2 for the dual phosphorylation model. Note that for the dual phosphorylation, the interactions for the middle states of the three states Markov chains are modified to incorporate the competition between the relevant phosphatase and kinase.

$p_{3kf}[n] =$	$\gamma_{in} v_r N_{in} = k_{in} \Delta t[IN]$
$p_{2kf1}[n] =$	$\gamma_{3k*1} v_r N_{3K} p_{3k*}[n-1] = k_{3k*1} \Delta t[3K] p_{3k*}[n-1]$
$p_{2kf2}[n] =$	$k_{3k*2} \Delta t[3K] p_{3k*}[n-1] \times (1 - k_{p2k1} \Delta t[P2k])$
$p_{kf1}[n] =$	$\gamma_{2kpp1} v_r N_{KK} p_{2kpp}[n-1] = k_{2kpp1} \Delta t[KK] p_{2kpp}[n-1]$
$p_{kf2}[n] =$	$k_{2kpp2} \Delta t[KK] p_{2kpp}[n-1] \times (1 - k_{pk1} \Delta t[Pk])$
$p_{3kb}[n] =$	$k_{p3k} \Delta t[P3k]$
$p_{2kb1}[n] =$	$k_{p2k1} \Delta t[P2k1] \times (1 - k_{3k*2} \Delta t[3K] p_{3k*}[n-1])$
$p_{2kb2}[n] =$	$k_{p2k2} \Delta t[P2k2]$
$p_{kb1}[n] =$	$k_{pk1} \Delta t[Pk1] \times (1 - k_{2kpp2} \Delta t[KK] p_{2kpp}[n-1])$
$p_{kb2}[n] =$	$k_{k2} \Delta t[Pk2]$

Table 8.2. Transition probabilities for the model in Figure 8-5. For bimolecular reactions, the relevant γ is obtained from the relevant k (in $\mathrm{M^{-1}s^{-1}}$) using the expression $\gamma = \frac{k \Delta t}{v_r A_V V}$.

■ 8.3.3 Parameters values

The parameters values used in the initial set of simulations are shown in Tables 8.3 and 8.4. The parameters are based on the values used by Huang and Ferrell in their deterministic model [67]. Later, some of the parameters values will be modified. The new values will be introduced as the corresponding modifications are considered.

$\Delta t = 0.1\mathrm{s}$	Time step
$K_{in} = 8.3 \times 10^6$	Activation of MAP3K
$K_{astar} = 8.3 \times 10^6$	Phosphorylation of MAP2K
$K_{bstarstar} = 8.3 \times 10^6$	Phosphorylation of MAPK
$K_{pa} = 8.3 \times 10^6$	Deactivation of MAP3K
$K_{pb} = 8.3 \times 10^6$	Dephosphorylation of MAP2K
$K_{pc} = 8.3 \times 10^6$	Dephosphorylation of MAPK
$Atot = 3 \times 10^{-9}$	Total MAP3K concentration
$Btot = 1.2 \times 10^{-6}$	Total MAP2K concentration
$Ctot = 1.2 \times 10^{-6}$	Total MAPK concentration
$Pa = 3 \times 10^{-10}$	Total MAP3K deactivating enzyme concentration
$Pb = 3 \times 10^{-10}$	Total MAP2K phosphatase concentration
$Pc = 120 \times 10^{-9}$	Total MAPK phosphatase concentration

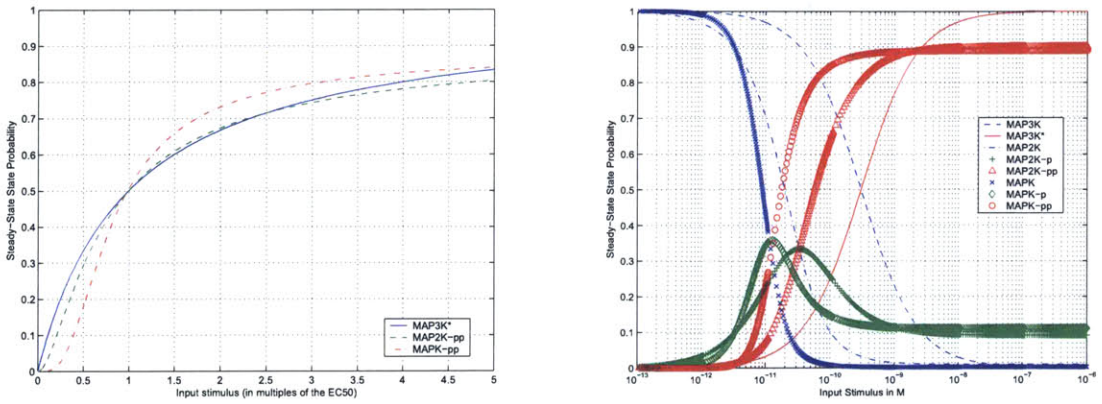
Table 8.3. Parameters values for the model in Figure 8-4. All K values are in $M^{-1}s^{-1}$ and all concentrations are in M.

$\Delta t = 0.1$s	Time step
$K_{in} = 8.3 \times 10^6$	Activation of MAP3K
$K_{astar1} = 8.3 \times 10^6$	First phosphorylation of MAP2K
$K_{astar2} = 8.3 \times 10^6$	Second phosphorylation of MAP2K
$K_{bstarstar1} = 8.3 \times 10^6$	First phosphorylation of MAPK
$K_{bstarstar2} = 8.3 \times 10^6$	Second phosphorylation of MAPK
$K_{pa} = 8.3 \times 10^6$	Deactivation of MAP3K
$K_{pb1} = 8.3 \times 10^6$	First dephosphorylation of MAP2K
$K_{pb2} = 8.3 \times 10^6$	Second dephosphorylation of MAP2K
$K_{pc1} = 8.3 \times 10^6$	First dephosphorylation of MAPK
$K_{pc2} = 8.3 \times 10^6$	Second dephosphorylation of MAPK
$Atot = 3 \times 10^{-9}$	Total MAP3K concentration
$Btot = 1.2 \times 10^{-6}$	Total MAP2K concentration
$Ctot = 1.2 \times 10^{-6}$	Total MAPK concentration
$Pa = 3 \times 10^{-10}$	Total MAP3K deactivating enzyme concentration
$Pb1 = 3 \times 10^{-10}$	Total MAP2K phosphatase 1 concentration
$Pb2 = 3 \times 10^{-10}$	Total MAP2K phosphatase 2 concentration
$Pc1 = 120 \times 10^{-9}$	Total MAPK phosphatase 1 concentration
$Pc2 = 120 \times 10^{-9}$	Total MAPK phosphatase 2 concentration

Table 8.4. Parameter values for the model in Figure 8-5. All K values are in $M^{-1}s^{-1}$ and all concentrations are in M.

■ 8.4 Dynamics of the State Probabilities of the a3MC Model Implementation

■ 8.4.1 Steady-state



(a) (b)

Figure 8-6. Steady-state state probabilities predicted by the a3MC model using the parameters given in Table 8.4. (a) The input stimulus is expressed in multiples of EC_{50} (the input concentration that produces a 50% maximal response. (b) Semi-log scale. Here, the input stimulus is expressed in absolute values.

The steady-state state probabilities as a function of the input stimulus (MAP3K activating enzyme) are shown in Figure 8-6 for the a3MC model implementation of the dual phosphorylation version of the model. These results are in agreement with the results obtained from the deterministic model presented by Huang and Ferrell [67] in that they predict the ultransensitivity of the cascade. The figures are directly comparable to dose-response experimental results such as the ones shown in Figure 8-2.

In their paper, Huang and Ferrell [67] investigate systematic changes to their model in order to study the effect of the model parameters on ultrasensitivity. We performed similar changes in our model and examined the corresponding response. As a start, we have examined the effect of increasing the concentration of MAP2K alone keeping the rest of the parameters the same. Figure 8-7 shows the steady-state state probabilities obtained when the concentration of MAP2K is assumed to be $6\mu M$ (5-fold greater than the concentration initially assumed and 20-fold above the measured K_m value for the phosphorylation of MAPK by active MAP2K (300nM) [67]). As can be seen from the figure, the activated MAPK response becomes more ultrasensitive in agreement with previous results [67] as well as more amplifying at high input concentrations. A similar effect, although not as pronounced (and without amplification), can be obtained by decreasing the K_m values from 300 nM to 60nM for the reactions that convert MAP2K among its various phosphorylation states (Figure 8-8). If the single phosphorylation model is used (Figure 8-4), the ultrasensitive response is lost (Figure 8-9) in agreement with previous results [67]. We have also examined the effect of negative and positive feedback on the steady-state response of the cascade. Feedback was implemented using the dual phosphorylation model in Figure 8-5. Specifically MAPK-pp was allowed to interact with inactive MAP3K as either an inhibitor (negative feedback) or an activator (positive feedback). We assumed that MAPK-pp inhibits the activation of MAP3K by preventing the MAP3K activating enzyme from binding inactive MAP3K. As a result, the fading version of the interactive Markov chains model was used for negative feedback and the transition probability from state KKK to state KKK$*$,

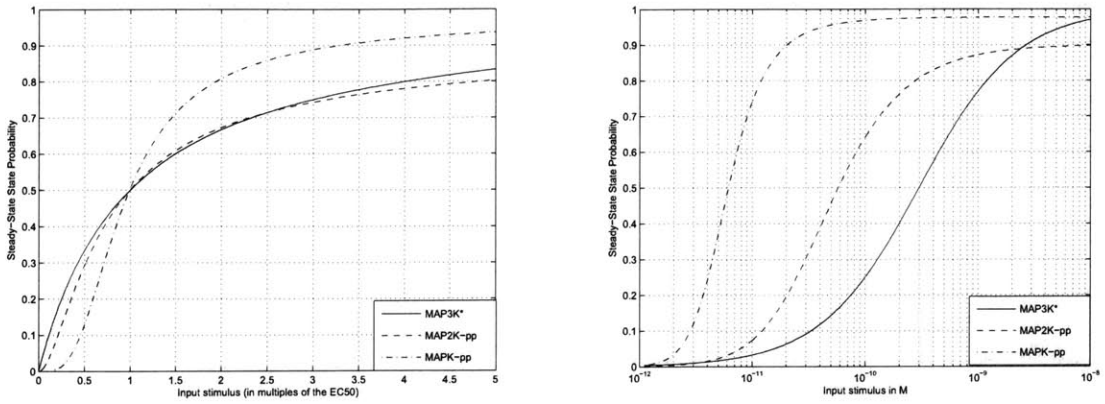


Figure 8-7. Steady-state state probabilities predicted by the $a3MC$ model using a concentration of MAP2K of $6\mu M$. (a) The input stimulus is expressed in multiples of EC_{50} (the input concentration that produces a 50% maximal response. (b) Semi-log scale on an absolute scale.

$p_{3kf}[n]$, was defined as:

$$p_{3kf}[n] = k_i n \Delta t[IN] \times k_{nfbk} \Delta t[K](1 - p_{kpp}[n-1]) \tag{8.1}$$

where

$$k_{nfbk} = 8.3 \times 10^6 M^{-1} s^{-1} \tag{8.2}$$

It was further assumed that the activation of MAP3K by MAPK-pp and the activation of MAP3K by MAP3K activating enzyme was associative and as a result the additive version of the interactive Markov chains model was used for positive feedback. $p_{3kf}[n]$ was therefore defined as:

$$p_{3kf}[n] = k_{in} \Delta t[IN] + k_{pfbk} \Delta t[K] p_{kpp}[n-1] \tag{8.3}$$

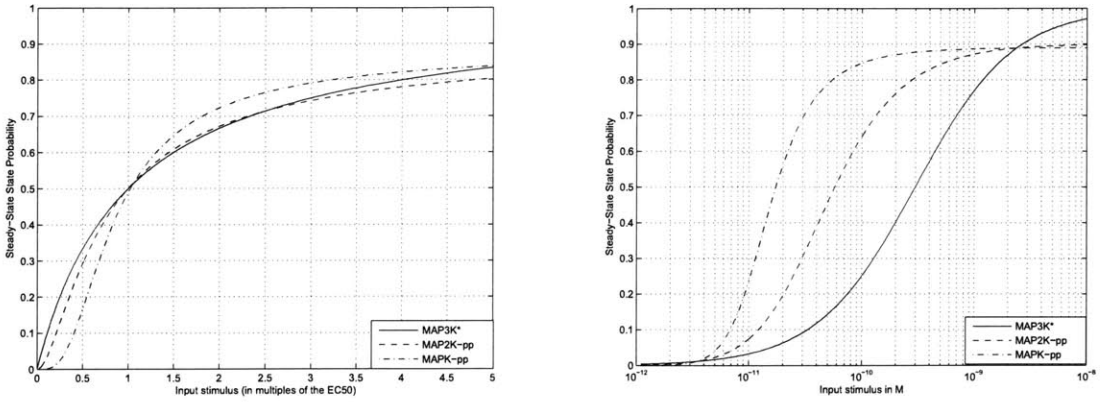
where again k_{pfbk} was chosen to be $8.3 \times 10^6 M^{-1} s^{-1}$. The negative feedback results are shown in Figure 8-10. As can be seen from the figure, the cascade still displays ultrasensitivity although it is slightly less marked. On the other hand, the response becomes almost a perfect switch (Figure 8-11) in the presence of positive feedback.

These results demonstrate the ability of the steady-state state probabilities of the $a3MC$ model to predict average behavior computed by the deterministic approach as well as to match experimental data obtained through dose response measurements.

■ 8.4.2 Dynamic behavior

We have also examined the time evolution of the state probabilities in response to different inputs. Figure 8-12 shows representative time plots for the state probabilities of the open loop dual phosphorylation model shown in Figure 8-5 in response to different step input stimuli of different strengths covering the entire dynamic range.

The results show that the active states (MAP3K*, MAP2K-pp, and MAPK-pp) reach their steady-state values monotonically. However the singly phosphorylated states do not exhibit monotonic behavior if the input is large enough. Specifically, the state probability of the singly phosphorylated state peaks a few minutes after the stimulus is applied and then adapts to a lower value at steady-state. Applying negative feedback leads to a partial



(a) (b)

Figure 8-8. Steady-state state probabilities predicted by the a3MC model using K_m values of 60nM (as opposed to 300nM) for the reactions that convert MAP2K among its various phosphorylation states. (a) The input stimulus is expressed in multiples of EC_{50}.(b) Semi-log scale on an absolute scale.

adaptation of the MAP3K* state and if the feedback is strong enough, it also leads to a partial adaptation of the MAP2K-pp and MAPK-pp states in agreement with previous results [10].

■ 8.5 Stochastic Simulations

The previous sections presented the dynamics of the state probabilities which lead to predictions of average behavior and are directly comparable to solutions of mass action deterministic models. These results are useful for predicting average behavior which can be obtained from repeating experiments numerous times and hold in the limit of a large number of molecules.

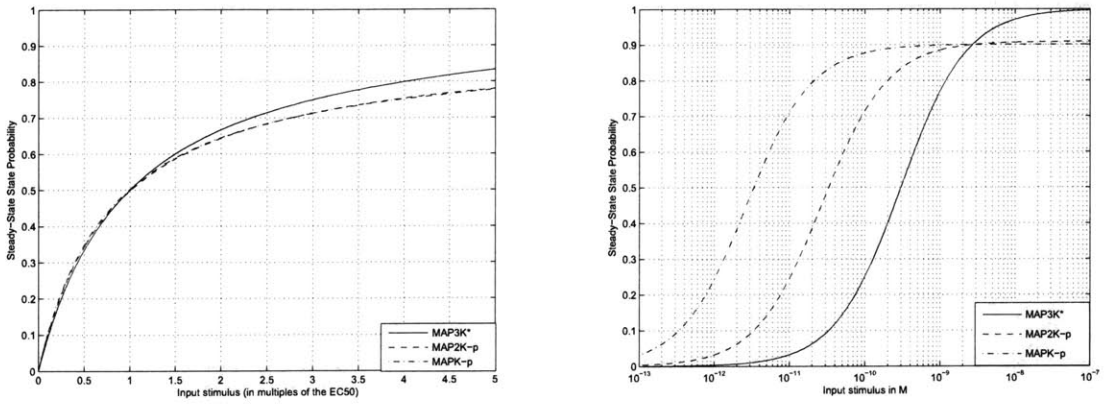
However, even more important is the ability of our model to generate meaningful predictions of the variance of single measurements and therefore proper interpretations of biological experiments. This is achieved by performing stochastic simulations and examining the individual time realizations as well as protein distributions at a given time. This section presents the results of such stochastic simulations.

Stochastic simulations were performed using the parameters in Table 8.4 and using both the 3MC and the a3MC model formulations. We simulated the stochastic behavior of 100 proteins for step inputs at time 0 of various concentrations and performed 1000 independent experiments. Both individual time realizations and distributions were investigated for the different proteins.

■ 8.5.1 Time realizations

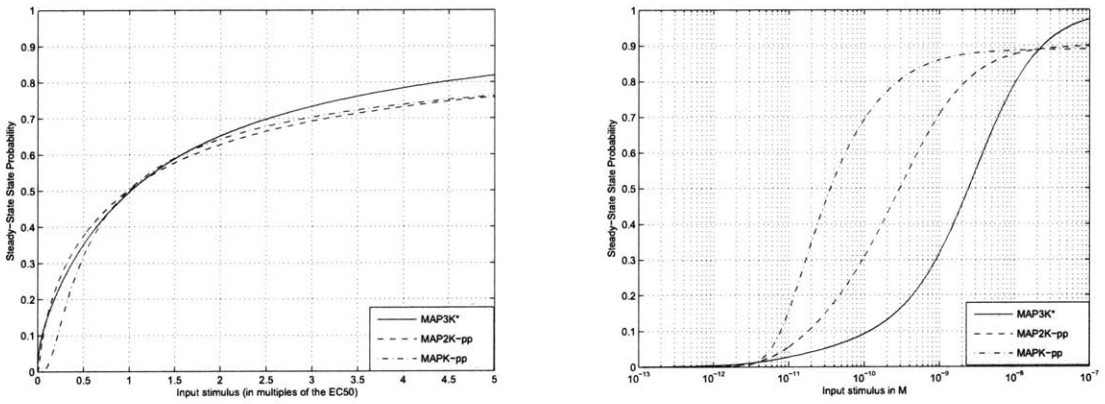
Figures 8-13 and 8-14 show sample paths of the 3MC and a3MC models simulation respectively for all molecules in the cascade in response to a step input stimulus of varying strength.

Note that the initial response is very different for both models. The difference is particularly striking for MAPK-p and MAPK-pp. This should not come as a surprise since the phosphorylation probabilities associated with MAPK are 400 times larger than those



(a) (b)

Figure 8-9. Steady-state state probabilities predicted by the a3MC model using a single phosphorylation. (a) The input stimulus is expressed in multiples of EC_{50}. (b) Semi-log scale.



(a) (b)

Figure 8-10. Steady-state state probabilities predicted by the a3MC model using negative feedback. (a) The input stimulus is expressed in multiples of EC_{50}. (b) Semi-log scale.

associated with MAP2K and MAPK because the concentration of MAP2K is 400 times larger than that of MAPK and since under the a3MC, the transition probabilities of a given chain are independent from the other chains in the model, we would expect the states of the MAPK chain in the model to reach their steady state values relatively fast even if the MAP2K chain is still in its initial state (i.e. the state probability of MAP2K-pp and MAP2K-p is zero), however this is not possible in the 3MC model implementation. In other words, one can clearly see from the figures that the joint distributions of states in different chains are different for both models. In the 3MC model, it is impossible to obtain activated MAPK molecules before observing activated MAP2K molecules whereas in the a3MC model that is possible since the relevant states are independent from each other.

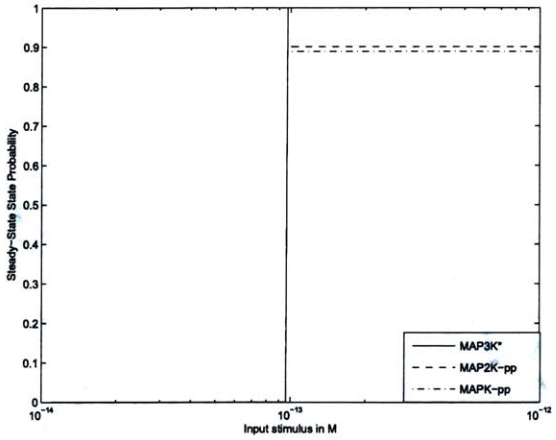
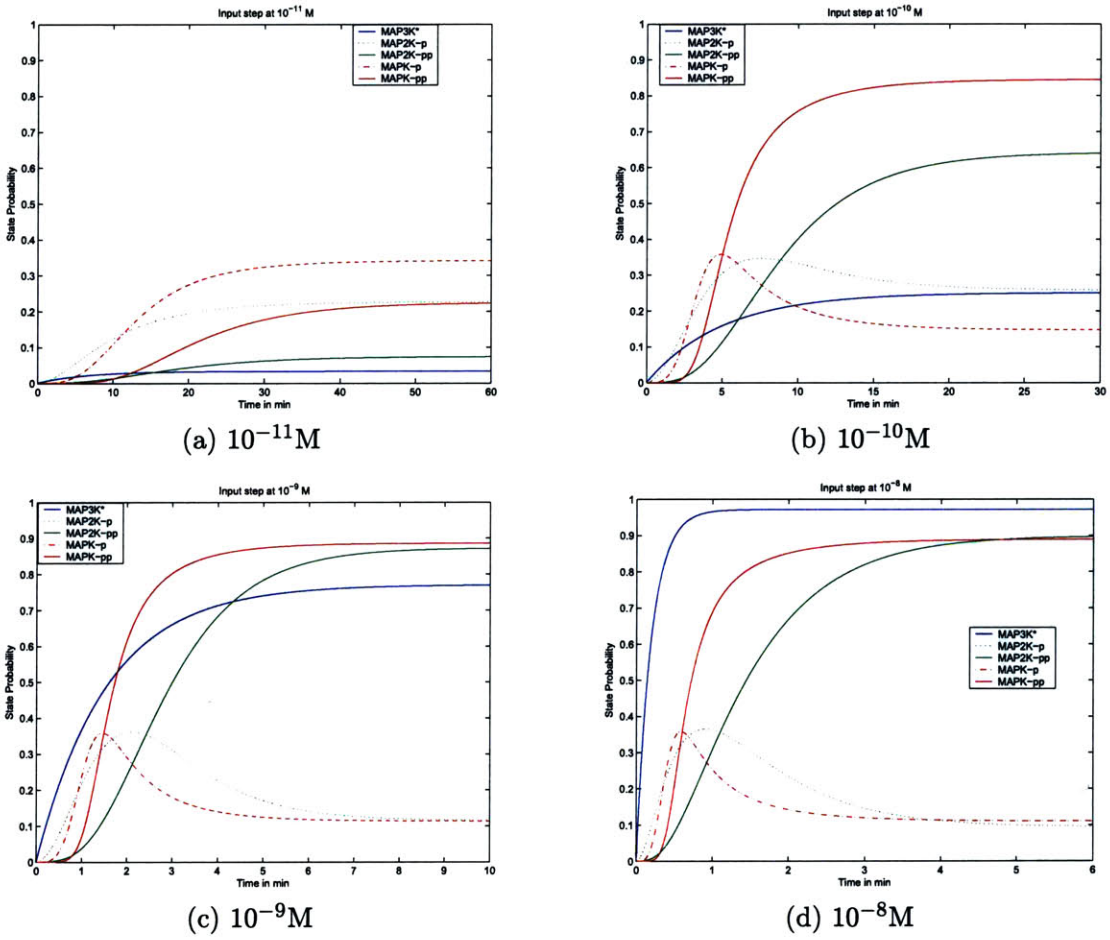


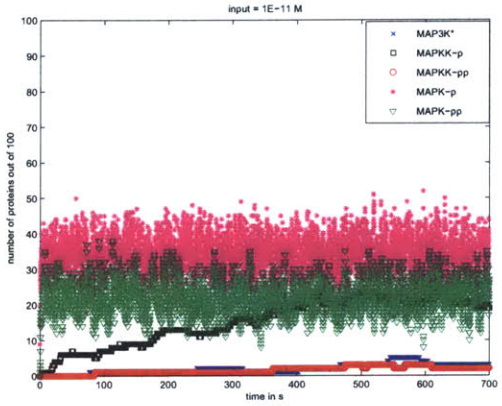
Figure 8-11. Steady-state state probabilities predicted by the a3MC model using positive feedback on a semi-log scale.



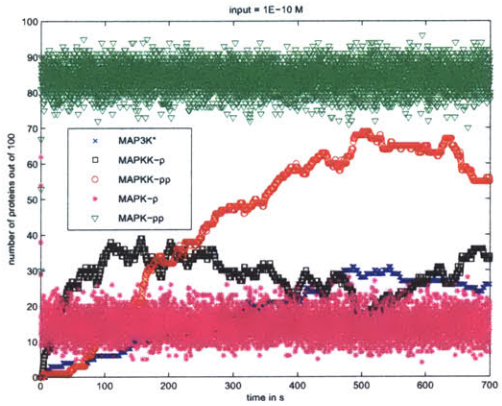
(a) 10^{-11}M (b) 10^{-10}M

(c) 10^{-9}M (d) 10^{-8}M

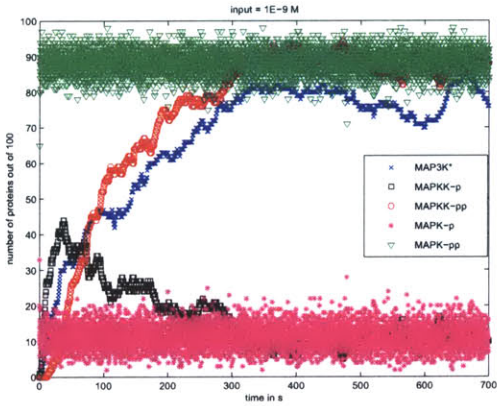
Figure 8-12. State probabilities time evolution for a step of input stimulus of different strengths.



(a) 10^{-11}M

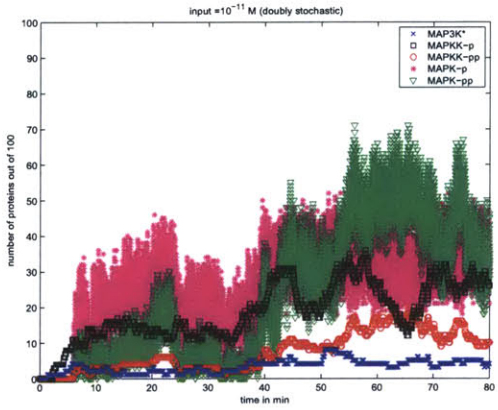


(b) 10^{-10}M

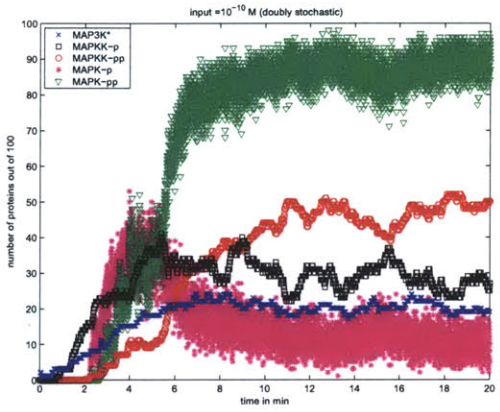


(b) 10^{-9}M

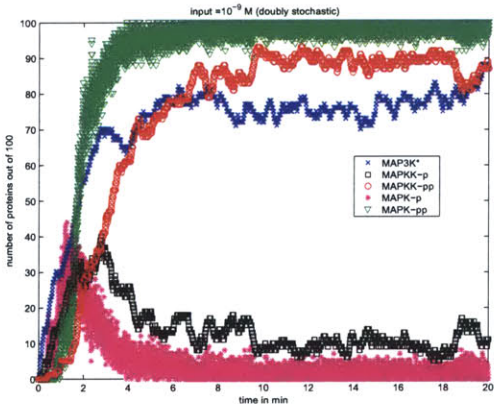
Figure 8-13. Time realization using the a3MC model for a step input stimulus of varying strengths.



(a) 10^{-11}M

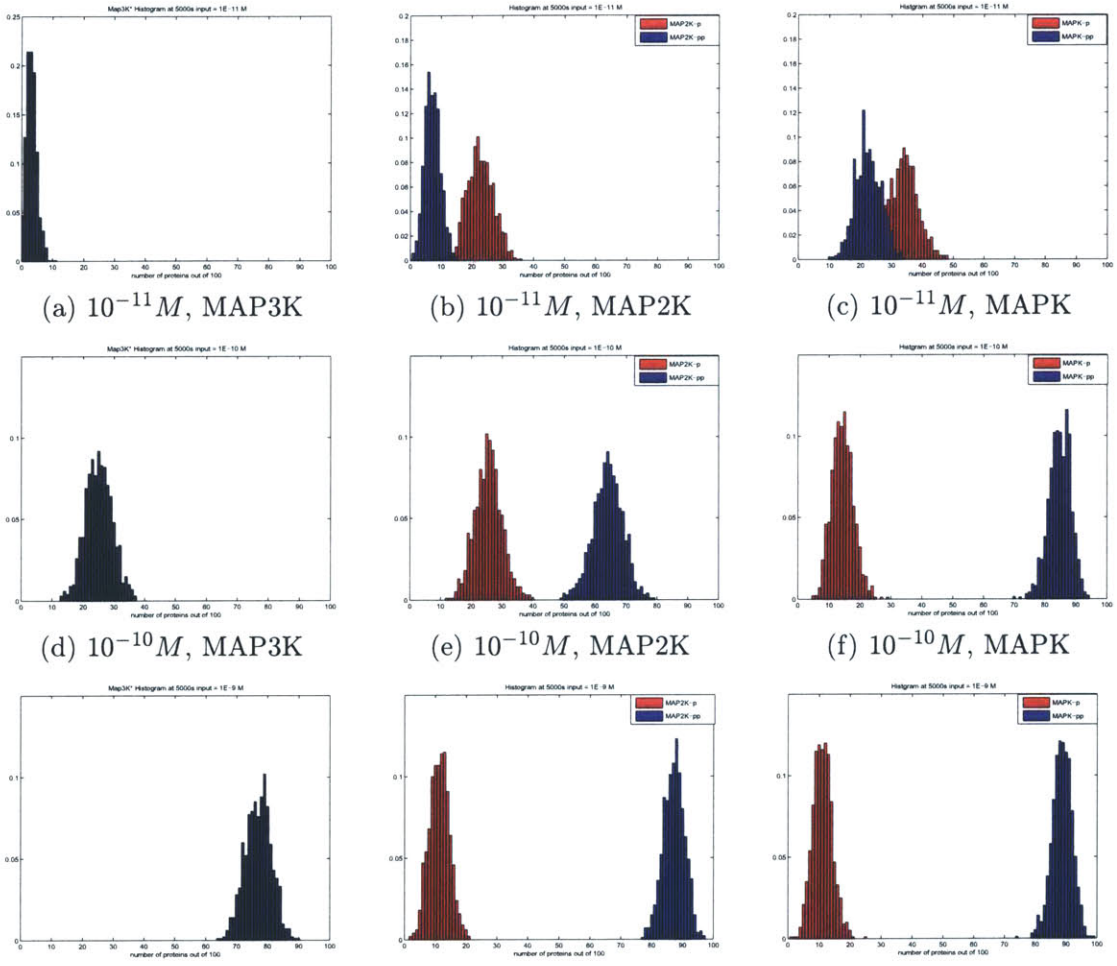


(b) 10^{-10}M



(c) 10^{-9}M

Figure 8-14. Time realization using the 3MC model for a step input stimulus of varying strengths.

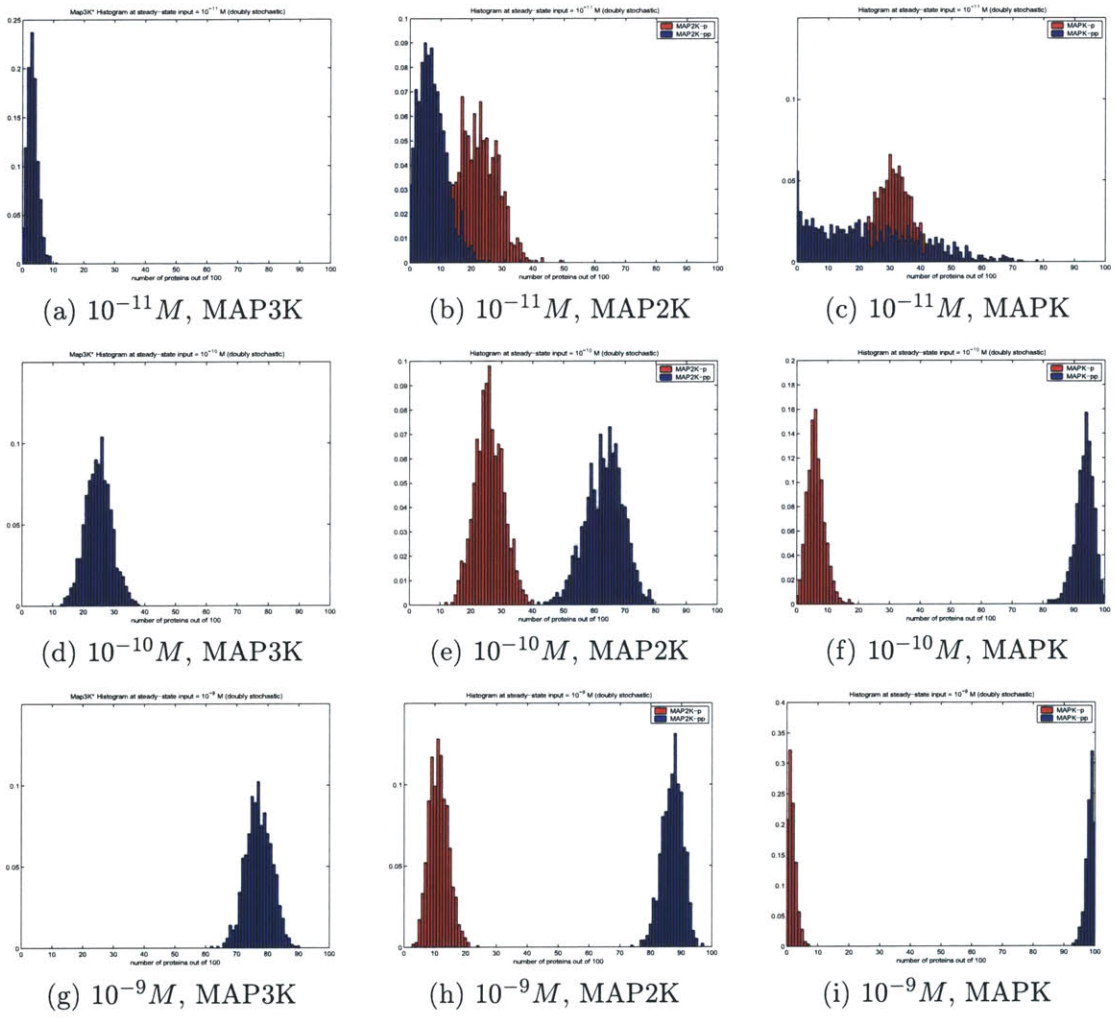


(a) $10^{-11}M$, MAP3K (b) $10^{-11}M$, MAP2K (c) $10^{-11}M$, MAPK

(d) $10^{-10}M$, MAP3K (e) $10^{-10}M$, MAP2K (f) $10^{-10}M$, MAPK

(g) $10^{-9}M$, MAP3K (h)$10^{-9}M$, MAP2K (i) $10^{-9}M$, MAPK

Figure 8-15. Distributions at steady state of relevant protein states using the a3MC model for step input stimuli of different strengths.



(a) $10^{-11}M$, MAP3K (b) $10^{-11}M$, MAP2K (c) $10^{-11}M$, MAPK

(d) $10^{-10}M$, MAP3K (e) $10^{-10}M$, MAP2K (f) $10^{-10}M$, MAPK

(g) $10^{-9}M$, MAP3K (h) $10^{-9}M$, MAP2K (i) $10^{-9}M$, MAPK

Figure 8-16. Distributions at steady state of the relevant protein states using the 3MC model for step input stimuli of different strengths.

■ 8.5.2 Distributions

While the individual time realizations presented in the previous subsection can correspond to the kind of data one would obtain from an individual experiment on a single cell. Investigating an isolated sample from a random process which is effectively what one experiment on a single cell would amount to, does not necessarily provide intuition and understanding as to the underlying nature of the process under study. A better approach to empirically analyzing random processes is to generate and study distributions which represent estimates of the probability density function of the underlying process. As a result, protein distributions were computed for all molecules in the model using 1000 independent experiments of 100 molecules.

The normalized distributions of the relevant protein states are shown in Figure 8-15 for the *a*3MC version of the model and in Figure 8-16 in the 3MC implementation of the model. The results indicate that the shape of the distributions are different depending on the protein state we are interested in as well as the strength of the input stimulus. It is clear from the figures that the average of the distribution which could be obtained using a more conventional modeling framework such as a mass-action deterministic model represents a very narrow view of the system's dynamics. Furthermore, the current state-of-the-art in single cell measurement techniques now allows the validation of such model predictions using flow cytometry therefore making the models presented here a potential formalism for interpreting single cell measurements.

■ 8.6 Summary

This chapter presented an example of how the interactive Markov chains model can be used to model biological pathways. Specifically, we applied the model to a classic biological signaling module: the mitogen-activated protein kinase cascade. The dynamics of the state probabilities obtained from the *a*3MC implementation of different versions of the model were validated by comparing the results to deterministic models based on mass-action kinetics as well as experimental data. We also generated results beyond those that could be obtained from deterministic models by performing stochastic simulations using both the *a*3MC and 3MC model implementations. Individual time realizations as well as distributions representing ensemble behavior illustrated the power of the modeling technique presented in this thesis. The model can potentially prove particularly useful as it generates predictions that are now readily testable using the current state of the art single cell measurements.

Chapter 9

Bacterial Chemotaxis

Bacterial chemotaxis is one of the most studied and well-understood signaling pathway in Biology. It involves a small, defined set of proteins and molecules that have been all purified and sequenced [26]. In addition, many of the enzymatic reactions they catalyze have been characterized and analyzed kinetically. In this chapter, we review some of the essential features of bacterial chemotaxis and the signaling cascade responsible for generating the chemotactic response. We then formulate an interactive Markov chains model for the chemotactic cascade and analyze it in the context of other models as well as experimental findings.

■ 9.1 The Biology of Bacterial Chemotaxis

Bacteria, such as *Escherichia coli* (*Ecoli*), cannot control their swimming direction, instead, they choose between swimming in a straight line or tumbling randomly. Some chemicals act as attractants while others are repellents. An attractant suppresses the tumbling behavior while a repellent induces it. However, if the swimming takes the bacteria away from the attractant, it will start tumbling randomly re-orienting itself before resuming its swimming behavior, possibly this time, in the direction of the attractant. *Ecoli*'s swimming mechanism consists of 6 to 10 flagella per cell [94]. The swimming behavior of the cell is determined by the direction of rotation of the flagella. If they are rotating in an anti-clockwise direction, they propel the cell in one direction. If, on the other hand, one or more flagella rotates in a clockwise direction, the cell tumbles.

■ 9.1.1 Signaling cascade

The underlying mechanisms governing bacterial chemotaxis revolve around a multifunctional, dimeric transmembrane receptor, the *Tar* complex. This receptor monitors the concentration of aspartate in the surrounding fluid, as well as the concentration of maltose (through a specific interaction with maltose binding protein), repellents such as nickel ions, ambient pH, and ambient temperature. On the cytoplasmic domain, Tar is associated with a cluster of "Che" (for chemotaxis) proteins and, together with these proteins, generates a signal that is sent to the flagellar motors. The magnitude of the signal depends on the rate of change of the various inputs to the Tar molecule [28]. The core Tar complex (the receptor with the associated signaling proteins) is shown in Figure 9-1.

Proteins associated with the Tar complex include CheA, CheB, CheW, CheR, CheZ and CheY. Binding of an attractant, or repellant, to the Tar receptor induces a conformational change that is propagated through the membrane to the cytoplasmic domain where it causes a change in the rate of autophosphorylation of CheA. The phosphorylated form of CheA (CheAp) transfers its phosphate group to a second protein, CheY, transforming it into CheYp. CheYp is the ultimate signaling molecule which diffuses to the flagellar motor where it interacts with it, increasing the clockwise rotation and, as a result, the bacteria starts tumbling. The default state of the motor (in the absence of CheYp) is therefore in counterclockwise rotation which causes the bacteria to swim [26]. Two additional molecules involved in this response are CheW, which is a transducing protein and mediates the effect of the receptor on the CheA phosphorylation, and, CheZ, a protein phosphatase, which counteracts the effect of CheYp by dephosphorylating it. In addition to this direct response, a slower adaptive response is orchestrated by CheR and CheB. CheR, a methyltransferase, deactivates receptor proteins by methylating them while CheB, a methylesterase, removes the methyl groups. The complete signal transduction pathway is shown in Figure 9-2.

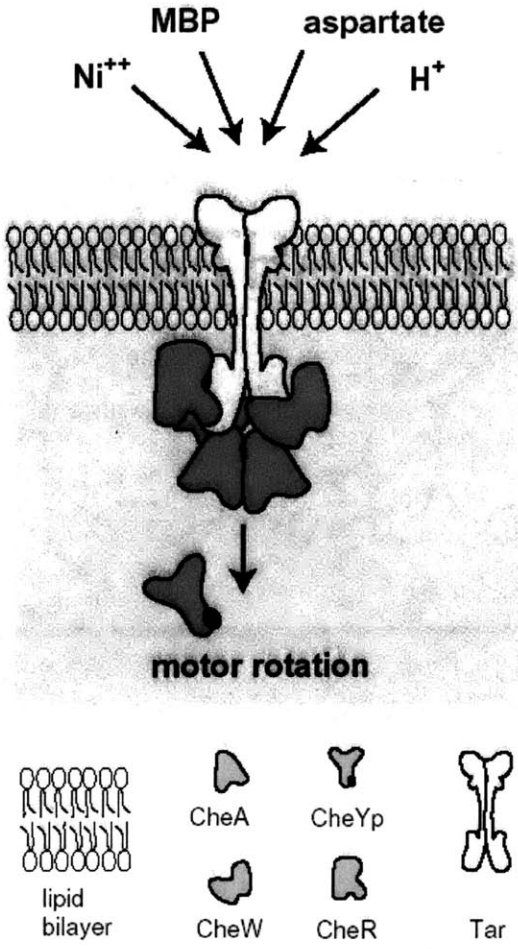


Figure 9-1. The core Tar complex (reproduced from Bray [28]). CheY, CheB, and CheZ are not shown.

The aspartate pathway is thus greatly simplified in bacteria. It consists of a single freely diffusing molecule (CheYp), a receptor complex concerned with the stimulus (the Tar complex), and a second protein complex concerned with the behavioral response, the flagellar motor [28].

■ 9.2 Current Models of Bacterial Chemotaxis

Current models of bacterial chemotaxis are of two types, biophysical and biochemical models. Biophysical approaches attempt to model the conformational change of the Tar receptor induced by signaling molecules, while biochemical approaches model the enzymatic reactions leading to the cell response.

■ 9.2.1 Biophysical models

Chervitz *et al.* [34] presented the first molecular description of a ligand-induced transmembrane conformational change using results from x-ray crystallography, solution ^{19}F NMR,

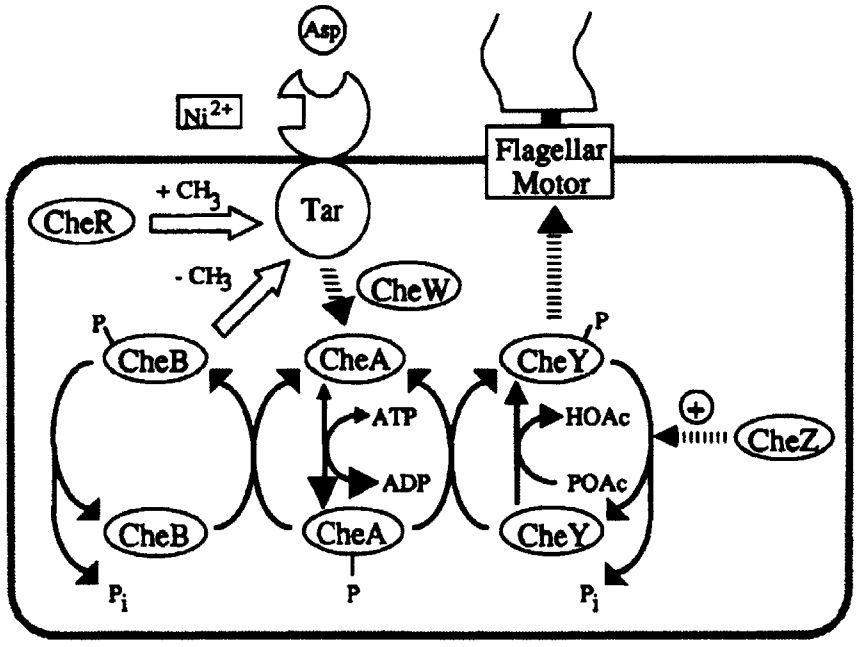


Figure 9-2. The signal transduction pathway in bacterial chemotaxis (reproduced from Bray *et al.* [26]). Solid arrows represent phosphorylation reactions. Dashed arrows indicate regulatory interactions, and open arrows are methylation reactions.

and disulfide-engineering studies. They concluded that the available evidence supports a swinging-piston mechanism for the transmembrane signal of the aspartate receptor. The signal is transmitted by a ligand-induced movement of a single transmembrane helix, located within the subunit providing most of the contacts to the bound aspartate molecule. They further noted that this model could be extended to any histidine kinase signaling pathway since the pathways are regulated by the same class of receptors possessing two transmembrane helices per subunit. They also appear to use the same mechanism of transmembrane signaling. Ottemann *et al.* [100] further explored the piston model for transmembrane signaling of the aspartate receptor by attaching nitroxide spin labels at strategic positions in the receptor and collecting the electron paramagnetic resonance spectra.

■ 9.2.2 Biochemical models

The piston model does not explain how the cytoplamic domain translates the conformational change into modulation of histidine kinase activity. A different approach is therefore needed to complement the results. An alternative consists of describing the signal transduction pathway as a biochemical network. A full characterization of the network could then be obtained by enumerating the Mass Action equations for the different biochemical reactions involved. Bray *et al.* [26] developed a computer simulation of the phosphorylation cascade controlling bacterial chemotaxis. In their model, each binding (where two molecular components associate or dissociate) and reaction (where a phosphate is added or removed by a protein) step is represented by an equation which is evaluated repeatedly at intervals of time Δt. The reactions rate constants and molecules concentrations are obtained from

previous published experiments and analyses. Their results show that the model provides an adequate explanation for most of the short-term chemotactic responses. The model was further generalized by Bray and Lay [27] to model the family of histidine kinase receptors. This simple signaling pathway is modeled by seven reactions, four of the reactions are treated as simple binding equilibria governed by the Law of Mass Action, and thus are characterized by a constant rate of association and a dissociation rate. The three other reactions are phosphorylation and dephosphorylation steps treated as first-order, irreversible, reactions, each characterized by a single rate. The slower response which leads to adaptation and is characterized by the methylation and demethylation of the receptor by respectively CheR and CheB is not considered in the original model. This bacterial chemotaxis model, named BCT, is available online at http://www.anat.cam.ac.uk/comp-cell/Chemotaxis.html and has been updated to include adaptation and conformational spread, i.e. the change in activity of a ligand-bound receptor can propagate to neighboring receptors in a cluster.

A number of additional models have been developed using a deterministic formulation of mass-action kinetics focusing on different aspects and properties of the signaling pathway. In particular, Spiro *et al.* [118] and Barkai and Leibler [15] proposed different models for adaptation in bacterial chemotaxis. In Spiro *et al.* [118], it is assumed that the receptor Tar, CheW, and CheA only exist as a complex (Tar-CheA-CheW) and never dissociate. The receptor has three methylation states (the three highest ones: two methyl groups attached, three methyl groups attached, and four methyl groups attached) since the unstimulated level is assumed to be at 1.5-2 methyl groups per receptor [25]. In the model, only phosphorylated CheB can demethylate the receptor and the phosphorylation state of CheA is independent of the ligand-binding state of the receptor and of the activities of CheR and CheBp, i.e. CheR and CheBp act in the same way on phosphorylated (active) and unphosphorylated receptors. Furthermore, the activity of CheZ is not modulated and the activation of CheY and CheB by CheAp is independent of the ligand binding and methylation states of the receptor. Finally, CheR methylates ligand-bound receptors faster than ligand-unbound receptors and the autophosphorylation of CheA is faster for higher methylation states of the receptor. Using these assumptions and experimentally based rate constants, Spiro *et al.* simulated the bacterial chemotaxis signaling pathway in response to a ramp, step and saturating levels of aspartate. They were also able to achieve perfect adaptation after fine-tuning the parameters using trial and error. In fact, it can be shown [124] that perfect adaptation is achieved using a very specific set of constants and small variations of the values will lead to a loss of perfect adaptation.

In contrast to Spiro *et al.*, Barkai and Leibler [15] proposed a robust model for perfect adaptation. The main assumption in this model is that CheB only demethylates active, i.e. phosphorylated, receptors and demethylation is independent of ligand binding. CheR, on the other hand, methylates both active and inactive receptors at the same rate. Under these assumption, Barkai and Leibler have shown that the system shows almost perfect adaptation over a wide range of parameter values. However while adaptation is a robust property of the system, they also show that adaptation time is not. Specifically, it is found that adaptation time is inversely proportional to receptor-complex activity.

■ 9.2.3 Stochastic models

There have been recently interest in modeling the stochastic aspects of the bacterial chemotaxis pathway. Specifically, Morton-Firth *et al.* [96] developed the first stochastic simulation of bacterial chemotaxis using the StochSim program. In this model, the activity of the re-

ceptor changes with both ligand-binding and methylation states. Furthermore, CheR binds exclusively to inactive receptor complexes while phosphorylated CheB binds exclusively to active receptor complexes. It is shown that the model achieves robust adaptation. The model was further extended by Shimizu *et al.* [116] to allow interactions between neighboring receptor complexes arranged in a regular lattice according to the Ising model.

■ 9.3 A Markov Modulated Markov Chains Model of Bacterial Chemotaxis

In this section, we use the interactive Markov chains formulation presented in Chapter 7 to model the bacterial chemotaxis pathway. The model includes a number of key assumptions which we discuss in the next subsections. Later in this chapter we will modify some of the assumptions and investigate their impact on the results.

■ 9.3.1 Ligand binding and receptor states

We assume in the model that ligand binding is independent of methylation state, i.e. the binding constant for the ligand and the receptor is the same whether or not the receptor is methylated, and that the receptor has only two methylation states: unmethylated and methylated. In addition, since ligand binding is much faster than any enzymatic transformation such as phosphorylation or methylation, we assume that the ligand and the receptor instantly equilibrate. As a result, we have:

$$\begin{aligned} k_D &= \frac{[T]_n[L]_n}{[TL]_n} \\ &= \frac{(N_{T_n}/A_V V)[L]_n}{N_{TL_n}/A_V V} \\ &= \frac{N_{T_n}}{N_{TL_n}}[L]_n \\ &= \frac{p_T[n]N_{Ttot}}{p_{TL}[n]N_{Ttot}}[L]_n \\ &= \frac{p_T[n]}{p_{TL}[n]}[L]_n \\ \frac{p_{TL}[n]}{p_T[n]} &= \frac{[L]_n}{k_D} \end{aligned} \tag{9.1}$$

where K_D is the dissociation constant of the receptor, A_V is Avogadro's number, V is the volume around the receptor, $[L]_n$ is the concentration of free ligand at time n, $[T]_n$ is the concentration of free receptor at time n, $[TL]_n$ is the concentration of ligand bound receptor at time n, N_{T_n} and N_{TL_n} are the numbers of free receptors and ligand bound receptors at time n, N_{tot} is the total number of receptors, and $p_T[n]$ and $p_{TL}[n]$ are the probabilities of the receptor being in the ligand unbound and bound states respectively. Similarly, we have:

$$\frac{p_{TmL}[n]}{p_{Tm}[n]} = \frac{[L]_n}{k_D} \tag{9.2}$$

where $p_{Tm}[n]$ and $p_{TmL}[n]$ are the probabilities of the methylated receptor being in the ligand unbound and bound states respectively.

The receptor is assumed to be in one of four states: unbound and unmethylated receptor (T), unbound and methylated receptor (Tm), ligand bound and unmethylated receptor

(TL), and ligand bound and methylated receptor (TmL). While it is known that Tar has four different methylation states (unmethylated, singly methylated, doubly methylated, and triply methylated), the first version of the model includes only two methylation states (unmethylated and methylated). We therefore have at all times:

$$p_T[n] + p_{Tm}[n] + p_{TL}[n] + p_{TmL}[n] = 1 \tag{9.3}$$

and using equations (9.1) and (9.2) above, we get:

$$p_T[n] + p_{Tm}[n] = \frac{1}{1 + \frac{[L]_n}{k_D}} = \frac{k_D}{k_D + [L]_n} \tag{9.4}$$

$$p_{TL}[n] + p_{TmL}[n] = \frac{\frac{[L]_n}{k_D}}{1 + \frac{[L]_n}{k_D}} = \frac{[L]_n}{k_D + [L]_n} \tag{9.5}$$

Since ligand binding is always assumed to be at equilibrium in the model, the receptor is represented as two *conditional* Markov chains: a ligand unbound chain and a ligand bound chain. Each chain is composed of two states corresponding to the unmethylated and methylated forms of the receptor. Specifically, the ligand unbound chain is composed of the states T' and Tm' corresponding to the unmethylated and methylated states of the receptor respectively *given* the receptor is not ligand bound. Similarly the ligand bound chain is composed of the states TL' and TmL' corresponding to the unmethylated and methylated states of the receptor respectively *given* the receptor is ligand bound. Figure 9-3 shows the receptor model with the conditional Markov chains.

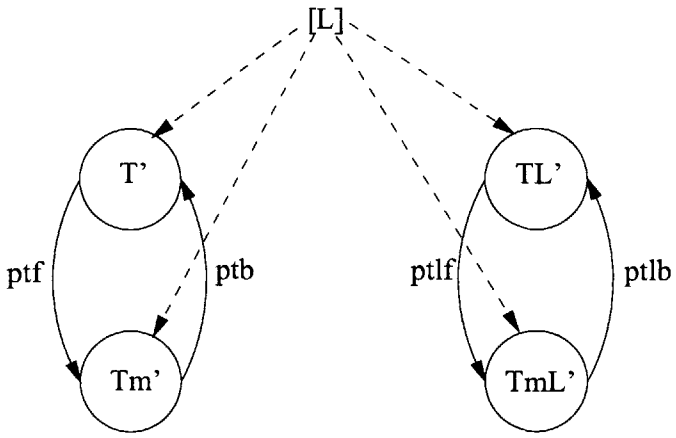


Figure 9-3. Conditional Markov chains representing the receptor states.

Each conditional chain evolves according to its transition matrix which is governed by the methylation and demethylation probabilities (p_{tf}, p_{tlf}, and p_{tb}, p_{tlb}). In order to get the unconditional probability of the receptor being in a given state, the conditional state probabilities are multiplied by the probability of being in the ligand bound or ligand unbound states. For example, the *a priori* probability of the receptor being in the unmethylated non-ligand bound state at time n, $p_T[n]$, is given by the product of the probability of the receptor being unmethylated at time n *given* it is not ligand bound, $p_{T'}[n]$, and the proba-

bility of the receptor being not ligand bound at time n which according to equation 9.4 is $\frac{1}{1+\frac{[L]_n}{k_D}}$, i.e.:

$$p_T[n] = p_{T'}[n] \times \frac{k_D}{k_D + [L]_n} \tag{9.6}$$

Similarly, we have:

$$p_{Tm}[n] = p_{Tm'}[n] \times \frac{k_D}{k_D + [L]_n} \tag{9.7}$$

$$p_{TL}[n] = p_{TL'}[n] \times \frac{[L]_n}{k_D + [L]_n} \tag{9.8}$$

$$p_{TmL}[n] = p_{TmL'}[n] \times \frac{[L]_n}{k_D + [L]_n} \tag{9.9}$$

■ 9.3.2 Chains and states

The core model (excluding the motor) includes six Markov chains representing the different molecules involved in the signaling pathway. The first chain represents the Tar receptor and has been discussed in the previous subsection. The second chain represents the CheA protein and has two states: unphosphorylated CheA (a) and phosphorylated CheA (ap). CheB and CheY are similarly represented with two-states Markov chains representing the phosphorylation states of the proteins: unphosphorylated (b and y) and phosphorylated (bp and yp). CheR and CheZ are represented by single states (R and Z respectively). The chains and states are shown in Figure 9-4 along with the interactions between chains (dashed arrows).

■ 9.3.3 Interactions and parameters

The interactions among the different Markov chains are obtained based on the known interactions between the different molecules in the signaling pathway and are depicted as dashed arrows in Figure 9-4. The expressions for the transition probabilities (using the a3MC model notation) are also given in Table 9.1 along with the rate constants values (Table 9.2). Table 9.3 tabulates the model parameters. Most rate constants and parameters were obtained from Spiro *et al.* [118] and are for the most part based on experimental measurements. (See [118] for further details.) In this version of the model, we have assumed that all enzymatic reactions occur in a regime of low substrate concentrations and therefore obey pseudo-first order kinetics where the effective rate constant is given by k_{cat}/K_m. Δt was chosen based on the fastest rate constant (largest) and v_r was chosen so that $v_r << \frac{1}{N}$ making sure that the one collision approximation is accurate. The transition probability describing the phosphorylation of CheA, $p_{af}[n]$ is simply expressed as a weighted sum of the activating states of the receptor since CheA is attached to the receptor complex and therefore does not need to collide with the receptor in order to get phosphorylated. In addition, each receptor state has an *a priori* probability of being active, i.e. being able to phosphorylate CheA. The activities of the different states are as follows:

$$\begin{aligned} A(T) &= 0.07 \\ A(Tm) &= 0.88 \\ A(TmL) &= 0.74 \\ A(TL) &= 0 \end{aligned}$$

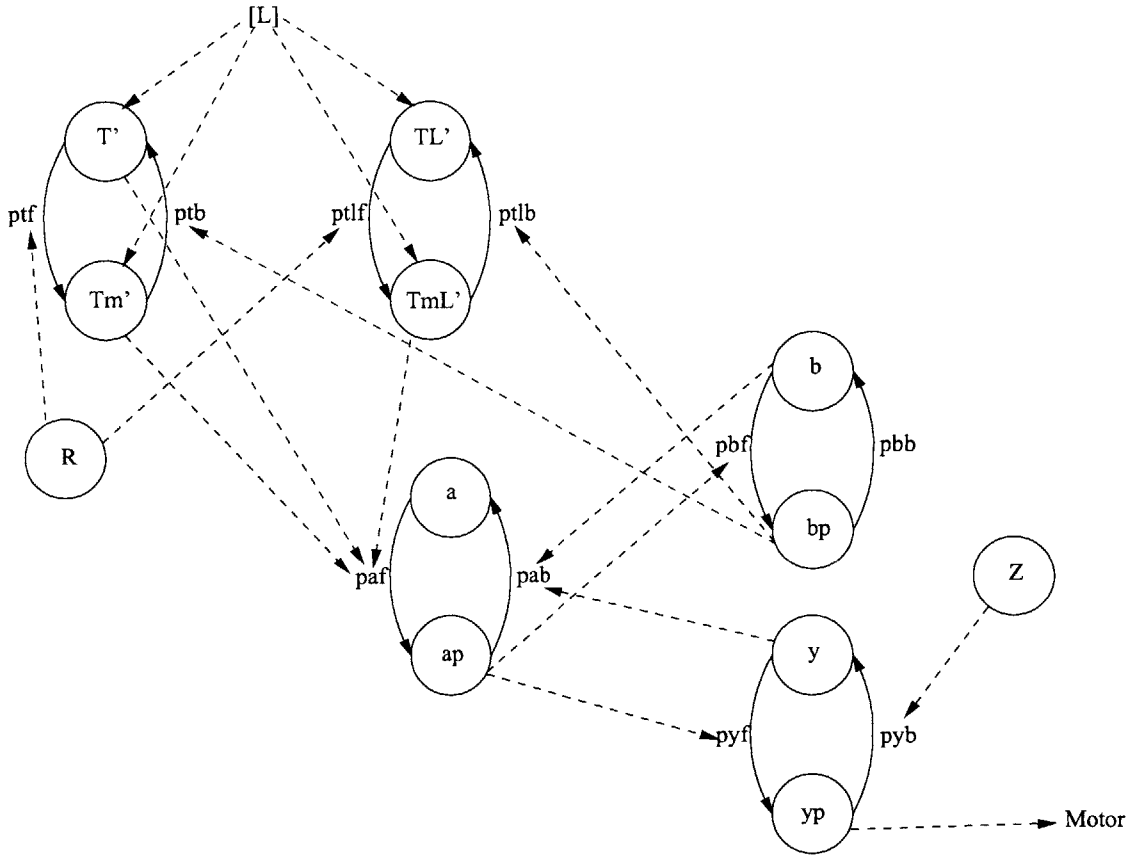


Figure 9-4. Core interactive Markov chains model of the bacterial chemotaxis pathway.

where $A(x)$ denotes the probability of state x being active (i.e. being able to phosphorylate CheA). These activities were obtained from Morton-Firth *et al.* [96] and are estimated from free energy changes. In [96] five different methylation states are considered. We map the two low methylation states (unmethylated, singly methylated) to our non-methylated state and use the average of the activities of those states for the activity of our state. Similarly, we map the two high methylation states (triply methylated and quadruply methylated) to the methylated state in our model where we use the average of the activities of those states as the activity of the methylated state in our model. We further assume that CheB demethylates only active receptors and therefore the activity probabilities are factored into the demethylation probabilities. As for the transition probability describing the dephosphorylation of CheA, $p_{ab}[n]$, both unphosphorylated CheB and unphosphorylated CheY contribute to the dephosphorylation of CheA in an associative manner. We therefore use the weighted sum model to combine the probabilities.

■ 9.3.4 The flagellar motor

The motor is modeled as an eight-state Markov Chain corresponding to the number of bound phosphorylated CheY molecules (ranging from zero CheYp bound to seven CheYp bound). Figure 9-5 shows the Markov chain model of the motor and Tables 9.4 and 9.5

$$
\begin{array}{lcl}
p_{tf}[n] & = & \gamma_{tf} v_r N_R \\
p_{tlf}[n] & = & \gamma_{tlf} v_r N_R \\
p_{tb}[n] & = & \gamma_{tb} v_r N_B p_{bp}[n-1] A(Tm) \\
p_{tlb}[n] & = & \gamma_{tlb} v_r N_B p_{bp}[n-1] A(TmL) \\
p_{af}[n] & = & k_{af} \Delta t (A(T) p_T[n-1] + A(Tm) p_{Tm}[n-1] + A(TmL) p_{TmL}[n-1] + A(TL) p_{TL}[n-1]) \\
p_{ab}[n] & = & \gamma_{ab_b} v_r N_B p_b[n-1] + \gamma_{ab_y} v_r N_Y p_y[n-1] \\
p_{bf}[n] & = & \gamma_{bf} v_r N_A p_{ap}[n-1] \\
p_{bb}[n] & = & k_{bb} \Delta t \\
p_{yf}[n] & = & \gamma_{yf} v_r N_A p_{ap}[n-1] \\
p_{yb}[n] & = & k_{yb} \Delta t
\end{array}
$$

Table 9.1. Transition probabilities for the model shown in Figure 9-4. For bimolecular reactions, the relevant γ is obtained from the relevant k (in $\mathrm{M}^{-1}\mathrm{s}^{-1}$) using the expression $\gamma = \frac{k \Delta t}{v_r A_V V}$.

$$
\begin{array}{lcl}
k_{tf} & = & 79992 M^{-1} s^{-1} \\
k_{tlf} & = & 79992 M^{-1} s^{-1} \\
k_{tb} & = & 79992 M^{-1} s^{-1} \\
k_{tlb} & = & 79992 M^{-1} s^{-1} \\
k_{af} & = & 45 s^{-1} \\
k_{ab_b} & = & 8 \times 10^5 M^{-1} s^{-1} \\
k_{ab_y} & = & 3 \times 10^7 M^{-1} s^{-1} \\
k_{bf} & = & k_{ab_b} \\
k_{bb} & = & 0.35 s^{-1} \\
k_{yf} & = & k_{ab_y} \\
k_{yb} & = & 20 s^{-1}
\end{array}
$$

Table 9.2. Model rate constants. Rate constants were obtained from Spiro *et al.* [118].

give the expressions for the transition probabilities and the values of the rate constants associated with the motor respectively.

The direction of rotation of the motor is governed by its state. States $m1$, $m2$, $m3$, $m4$, and $m5$ correspond to the motor rotating counter clockwise which propels the bacterium forward while states $m6$, $m7$, and $m8$ correspond to the motor rotating clockwise which makes the bacterium tumble. As a result, the probability of the motor rotating counter clockwise (run) at time n, $p_{ccw}[n]$, is given by:

$$p_{ccw}[n] = p_{m1}[n] + p_{m2}[n] + p_{m3}[n] + p_{m4}[n] + p_{m5}[n] \equiv p_{run}[n] \tag{9.10}$$

and the probability of the motor rotating clockwise (tumble) at time n, $p_{cw}[n]$, is given by:

$$p_{cw}[n] = p_{m6}[n] + p_{m7}[n] + p_{m8}[n] \equiv p_{tumble}[n] \tag{9.11}$$

The motor bias (or simply bias) is defined as the probability of a run, i.e. of rotating counterclockwise:

$$bias \equiv p_{ccw}[n] = p_{run}[n] \tag{9.12}$$

$k_D =$	10^{-6}M	Motor =	1
$[L] =$	x the input	$N_{Ttot} =$	6793
$N_R =$	255	$\Delta t =$	0.001 s
$N_B =$	1444	$v_r =$	10^{-6}
$N_A =$	6793	$A_V =$	6.022×10^{23}
$N_Y =$	17000	$V =$	1.41×10^{-15}

Table 9.3. Model parameters. N_X is the total number of molecules of species X and were obtained from the concentrations in [118].

$p_{mf1}[n]$	=	$\gamma_{mf1} v_r N_Y p_{yp}[n-1]$
$p_{mf2}[n]$	=	$\gamma_{mf2} v_r N_Y p_{yp}[n-1](1-k_{mb1}\Delta t)$
$p_{mf3}[n]$	=	$\gamma_{mf3} v_r N_Y p_{yp}[n-1](1-k_{mb2}\Delta t)$
$p_{mf4}[n]$	=	$\gamma_{mf4} v_r N_Y p_{yp}[n-1](1-k_{mb3}\Delta t)$
$p_{mf5}[n]$	=	$\gamma_{mf5} v_r N_Y p_{yp}[n-1](1-k_{mb4}\Delta t)$
$p_{mf6}[n]$	=	$\gamma_{mf6} v_r N_Y p_{yp}[n-1](1-k_{mb5}\Delta t)$
$p_{mf7}[n]$	=	$\gamma_{mf7} v_r N_Y p_{yp}[n-1](1-k_{mb6}\Delta t)$
$p_{mb1}[n]$	=	$k_{mb1}\Delta t(1-\gamma_{mf2} vr N_Y p_{yp}[n-1])$
$p_{mb2}[n]$	=	$k_{mb2}\Delta t(1-\gamma_{mf3} vr N_Y p_{yp}[n-1])$
$p_{mb3}[n]$	=	$k_{mb3}\Delta t(1-\gamma_{mf4} vr N_Y p_{yp}[n-1])$
$p_{mb4}[n]$	=	$k_{mb4}\Delta t(1-\gamma_{mf5} vr N_Y p_{yp}[n-1])$
$p_{mb5}[n]$	=	$k_{mb5}\Delta t(1-\gamma_{mf6} vr N_Y p_{yp}[n-1])$
$p_{mb6}[n]$	=	$k_{mb6}\Delta t(1-\gamma_{mf7} vr N_Y p_{yp}[n-1])$
$p_{mb7}[n]$	=	$k_{mb7}\Delta t$

Table 9.4. Transition probabilities for the motor model. For bimolecular reactions, the relevant γ is obtained from the releval k (in $\mathrm{M}^{-1}\mathrm{s}^{-1}$) using the expression $\gamma = \frac{k\Delta t}{v_r A_V V}$.

$k_{mf1} =$	$7 \times 10^6 M^{-1}s^{-1}$	$k_{mb1} =$	$1.43s^{-1}$
$k_{mf2} =$	$6 \times 10^6 M^{-1}s^{-1}$	$k_{mb2} =$	$2.86s^{-1}$
$k_{mf3} =$	$5 \times 10^6 M^{-1}s^{-1}$	$k_{mb3} =$	$4.29s^{-1}$
$k_{mf4} =$	$4 \times 10^6 M^{-1}s^{-1}$	$k_{mb4} =$	$5.72s^{-1}$
$k_{mf5} =$	$3 \times 10^6 M^{-1}s^{-1}$	$k_{mb5} =$	$7.15s^{-1}$
$k_{mf6} =$	$2 \times 10^6 M^{-1}s^{-1}$	$k_{mb6} =$	$8.58s^{-1}$
$k_{mf7} =$	$1 \times 10^6 M^{-1}s^{-1}$	$k_{mb7} =$	$10.01s^{-1}$

Table 9.5. Motor model rate constants obtained from [97].

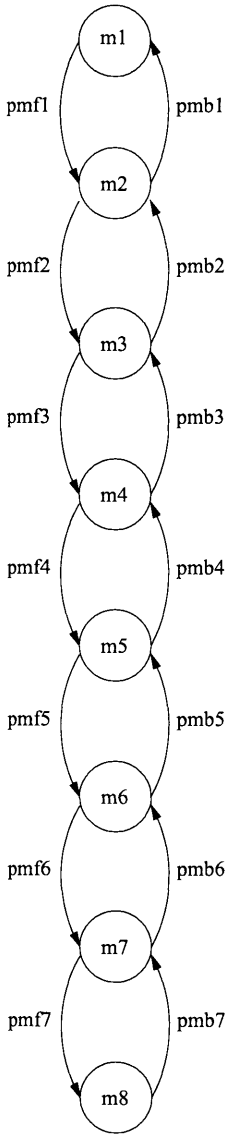


Figure 9-5. Markov chains model of the flagellar motor.

■ 9.4 Simulations

■ 9.4.1 The dynamic behavior of the state probabilities

A multistep input was applied to the model and the state probabilities were computed. Specifically, the input ligand concentration was stepped up from 0 to 10^{-6}M at 250 seconds, and from 10^{-6}M to 10^{-3}M at 400 seconds. The input concentration was then stepped down from 10^{-3}M to 10^{-6}M at 550 seconds, and from 10^{-6}M to zero at 700 seconds. Figure 9-6 plots the input ligand concentration between 250 and 700 seconds.

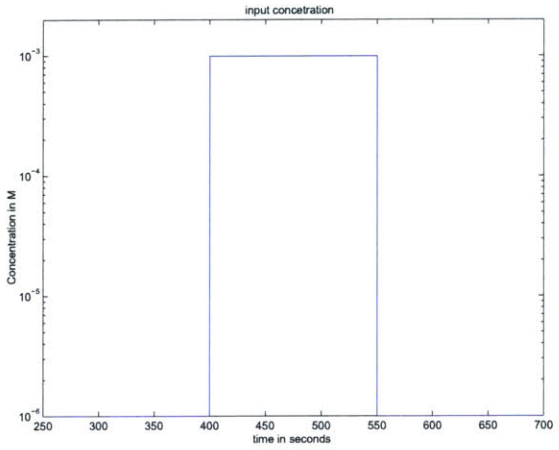


Figure 9-6. Non-zero input ligand concentration stimulus.

Simulations were carried out using the default parameters given in the previous sections and the state probabilities for the phosphorylated forms of CheA (ap), CheY (yp), and CheB (bp) were computed. The probabilities of the different receptor states (T, Tm, TL, TmL) were also computed as well as the probability of a receptor being methylated which is given by the sum of $p(Tm)$ and $p(TmL)$ and gives an indication of the level of receptor activity. Finally, the motor bias, defined as the probability of a run (i.e. the probability of rotating counterclockwise), was also computed. The results are shown in Figure 9-7. As expected, upon ligand addition, the receptor states shift from non-methylated non-ligand bound to predominantly ligand bound (both unmethylated and methylated). Since demethylation is a function of ligand activity and the methylated ligand bound receptor is less active than the methylated non-ligand bound receptor, demethylation is slower for the ligand bound state and as a result the proportion of methylated receptors continues to slowly increase after the initial step increase due to ligand addition and instantaneous equilibration. These dynamics lead to an immediate decrease in phosphorylated CheA upon ligand addition followed by a slow increase due to the methylation of the receptor. This effect is transmitted down to the motor as shown in the figure leading to a transient increase in motor bias upon ligand addition followed by a slower decrease. The amplitude of the motor bias response is larger upon ligand addition than upon ligand removal. This is largely due to the non-symmetric functional states of the motor. Specifically, as can be seen from equations 9.10 and 9.11, five out of the eight motor states correspond to a run whereas only three states correspond to a tumble. Upon ligand addition the transitions into the run states are increased and since the number of these states is larger than the number of the tumble states, the probability of being in a state corresponding to a run increases more than upon removal of the same

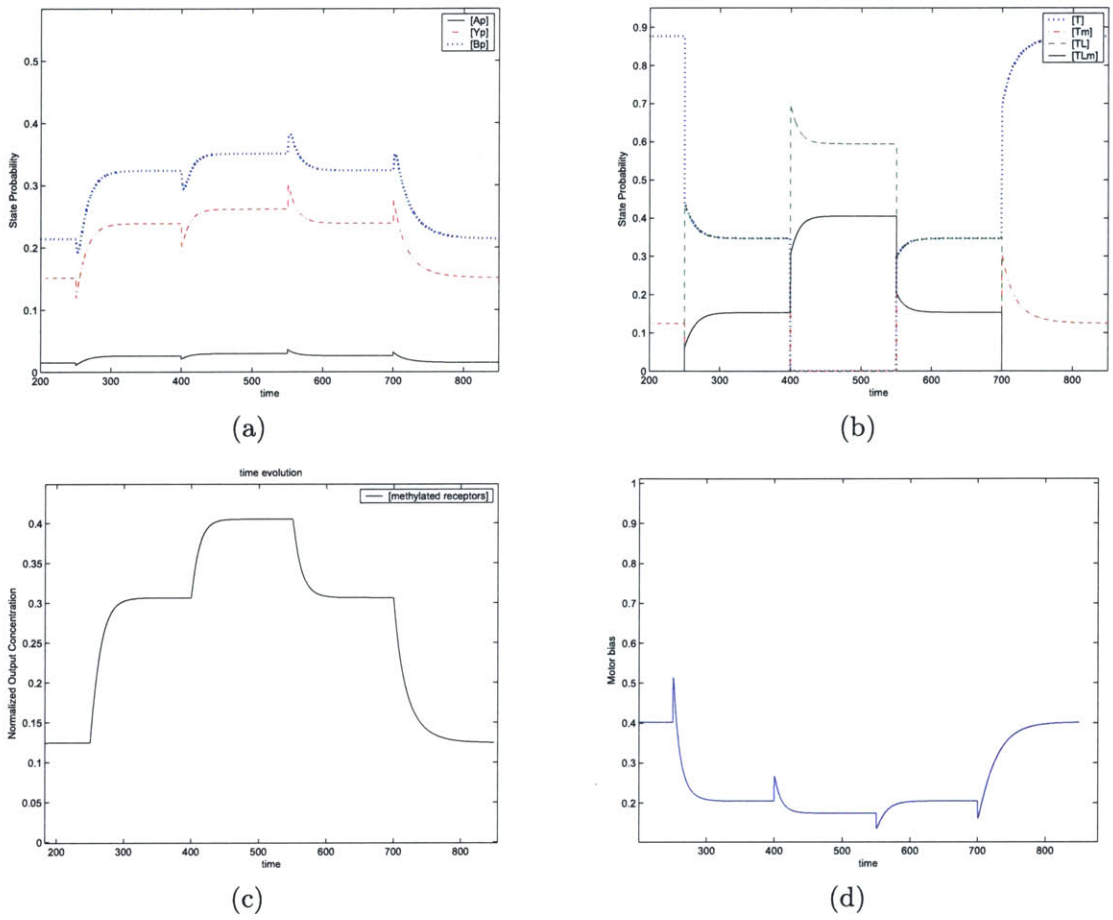


Figure 9-7. State probabilities using default parameters. (a) ap, yp, and bp, (b) receptor states (c) methylated receptors, (d) motor bias.

amount of ligand, since in this latter case, the transition probabilities into the tumbling states are higher but the number of corresponding states is smaller. This assymetry in the response upon ligand addition and removal is in agreement with previous results [15]. The response in Figure 9-7, however, fails to adapt or more precisely over-adapts, i.e. instead of settling back to its original value (perfect adaptation), the response overshoots its original value.

In order to address the failure of the model to adapt, we consider two modifications to the original model. First, the rate of receptor demethylation by phosphorylated CheB was increased by a factor of 3.75. The same factor was applied to the demethylation of the ligand bound and the ligand unbound receptor states. Increasing the demethylation rate reduces the activity of the receptor at steady state since unmethylated receptors are less active. As a result, the proportion of phosphorylated CheA is reduced which propagates to the motor increasing the bias at steady-state and therefore reducing over-adaptation. The value of the factor by which the demethylation rate was increased (3.75) was derived to achieve perfect adaptation in response to a saturating ligand stimulus (i.e. in response to a step in ligand concentration of 10^{-3}M). Figure 9-8 shows the results. As expected, in

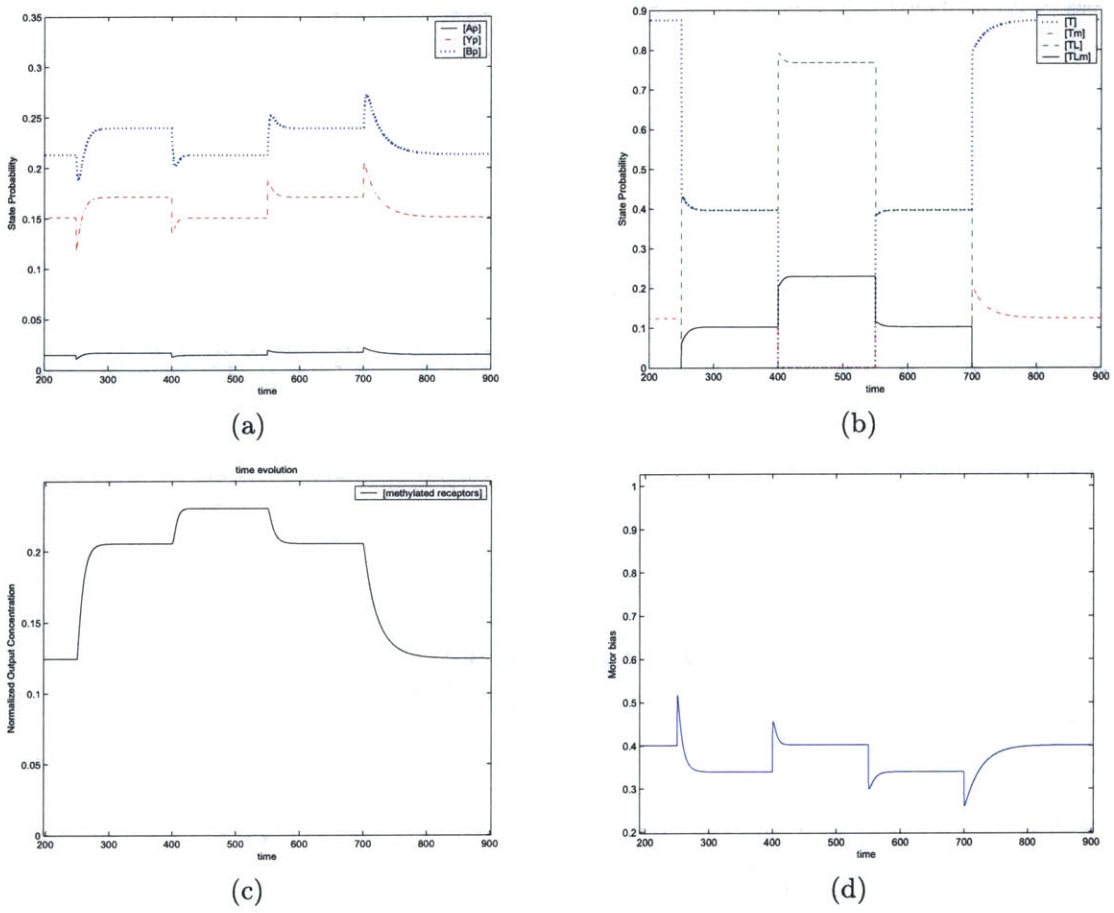


Figure 9-8. State probabilities using a high rate constant for demethylation by bp. (a) ap, yp and bp, (b) receptor states (c) methylated receptors, (d) motor bias.

response to a ligand concentration of 10^{-3}M (the input value between 400 and 550 seconds), the response perfectly adapts i.e. it returns to its pre-stimulus level. However, for lower concentrations (10^{-6}M in the figure, for the time points between 250 and 400 seconds as well as 550 and 700 seconds), the response overadapts however not as strongly as before. We also considered the case where the receptors can be demethylated by all forms of CheB (both phosphorylated and unphosphorylated). The results are shown in Figure 9-9. The amplitude of the response in this model is now larger than the amplitude of the response in both the original model and the model with higher demethylation rate. However, the system still overadapts as in the original model. Later in this chapter, additional modifications to the model leading to perfect adaptation will be considered and discussed.

■ 9.4.2 Adaptation time

The results obtained in the previous section (examining the dynamics of the state probabilities and their response to different modifications) are equivalent to the results one can obtain using models based on deterministic rate equations. However, while the model presented here can perform the kind of "average" analysis that the deterministic approach

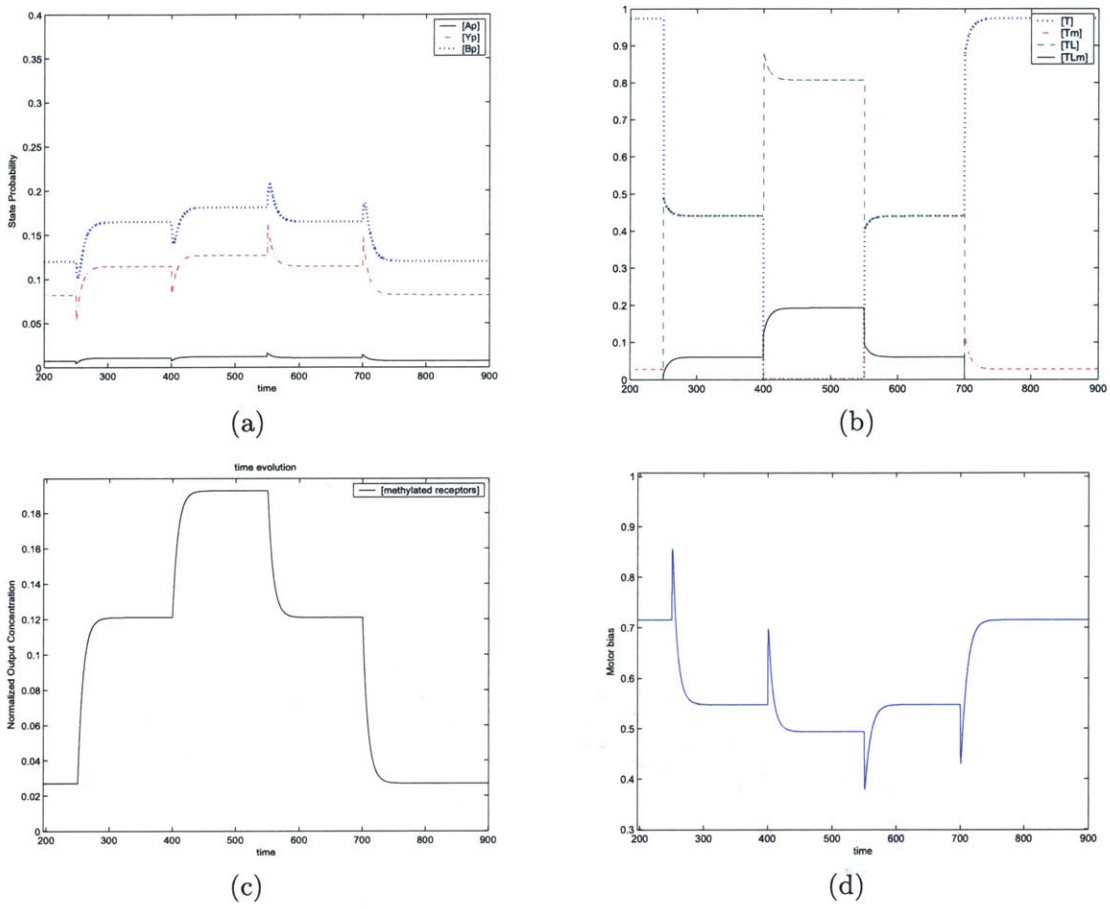


Figure 9-9. State probabilities using the case where all forms of CheB can demethylate the receptor. (a) ap, yp, and bp, (b) receptor states (c) methylated receptors, (d) Motor bias.

provides, it can also be used to examine the variations and stochasticity of the response which is not possible using a deterministic formalism. In order to illustrate this feature of the model, we investigate the effect of CheR concentration on the adaptation time (or more generally time to steady-state) of the motor in response to a step increase in ligand concentration. Stochastic simulations using the *a*3MC model implementation were performed for different CheR concentrations . For each CheR concentration, the average as well as the variance of the time to adaptation was computed using 100 simulations. The results are shown in Figure 9-10 and indicate that the adaptation time decreases with increased CheR concentration. This result should not come as a surprise since the higher the CheR concentration, the higher the transition probabilities out of the non-methylated states of the receptor which makes the chains reach steady-state faster. The results in Figure 9-10 also show that the standard deviation of the adaptation time decreases with increasing CheR. This result is not necessarily expected since CheR acts upstream in the signaling cascade while adaptation time is measured using the motor which is much further downstream. These results are in agreement with experimental observations. Specifically, Alon *et al.* [5] measured the average adaptation time in response to a step-like stimulus of 1mM L-aspartate using different bacterial strains including wild-type and a mutant lacking CheR.

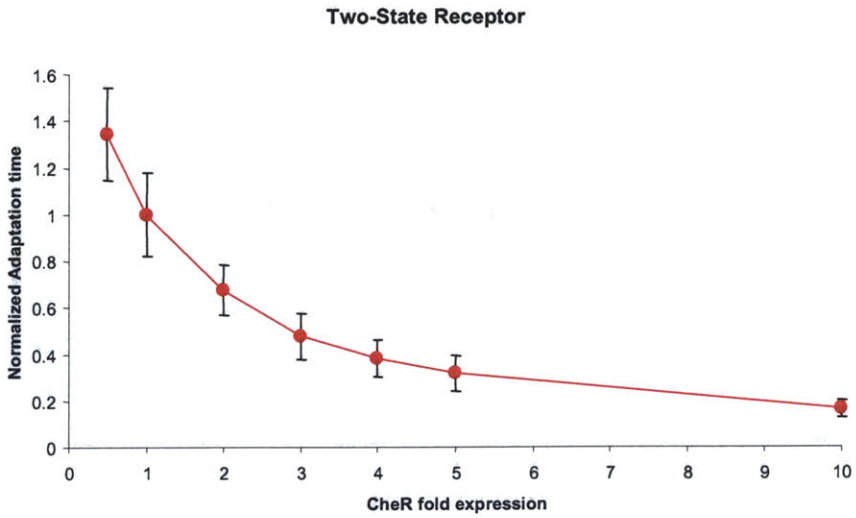


Figure 9-10. Average adaptation time and standard deviation using the original model parameters. The error bars correspond to the standard deviation.

The level of CheR expression in the mutant was controlled by inserting a plasmid containing a CheR gene that can be controlled by IPTG induction. The average adaptation time for different concentrations of CheR was measured and it was observed that the average adaptation time decreased with increased CheR concentration. In addition to the average, Alon *et al.* computed error bars corresponding to the standard deviation of triplicate experiments (see Figure 2b in [5]). The standard deviation results follow the same trend as the results shown in Figure 9-10 indicating that the change in standard variation is an intrinsic property of the signaling pathway as opposed to a consequence of external variations such as experimental noise. These results imply that more experimental replicates need to be performed for low CheR concentrations than for high CheR concentrations in order to get a good estimate of the average adaptation time since higher biological variance is expected at low CheR concentrations.

■ 9.5 Stochastic Implementation of Enzyme Kinetics

The macromolecular rates used to derive the transition probabilities in the model presented in the previous sections were all based on first order mass action kinetics. In particular, all enzymatic reactions were assumed to operate in a regime of low substrate concentration relative to the K_m of the enzyme. While for bacterial chemotaxis, this assumption holds for most enzymatic reactions, it does not always hold for the reactions governing the methylation of the receptor by CheR. Specifically, when the probabilities of the receptor being in the unmethylated non-ligand bound or in the unmethylated ligand bound states are greater than 0.2, using the first order approximation leads to a much larger transition probability than the one based on enzymatic kinetics. To address this issue, the original model was modified to include a stochastic implementation of enzymatic dynamics. In this implementation, we assume the existence of an intermediary state between the substrate and the product corresponding to the enzyme bound substrate. Specifically, the reaction

can be written as follows:

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} SE \overset{k_2}{\rightarrow} P + E \tag{9.13}$$

where S is the substrate, E is the enzyme, ES is the enzyme-substrate complex and P is the product. The total enzyme concentration, $[E_t]$ is fixed and is given by:

$$[E]_t = [E] + [ES] \tag{9.14}$$

and the total substrate-product concentration, $[S, P]_t$, is also fixed (in the absence of additional reactions involving these species) and is given by:

$$[S, P]_t = [S] + [ES] + [P] \tag{9.15}$$

where brackets indicate concentration.

■ 9.5.1 Interacting Markov chains model of enzyme kinetics

The reaction given in 9.13 has a natural interacting Markov chains representation based on two Markov chains: the enzyme chain and the substrate-product chain. The enzyme chain has two states corresponding to the free and bound states while the substrate-product chain has three states corresponding to the substrate, substrate bound to enzyme, and product states. Furthermore, since the last reaction in 9.13 is irreversible, there is no transition from the product to the enzyme-bound substrate state in the substrate-product chain. The interacting Markov chains representation of enzyme kinetics is shown in Figure 9-11. In

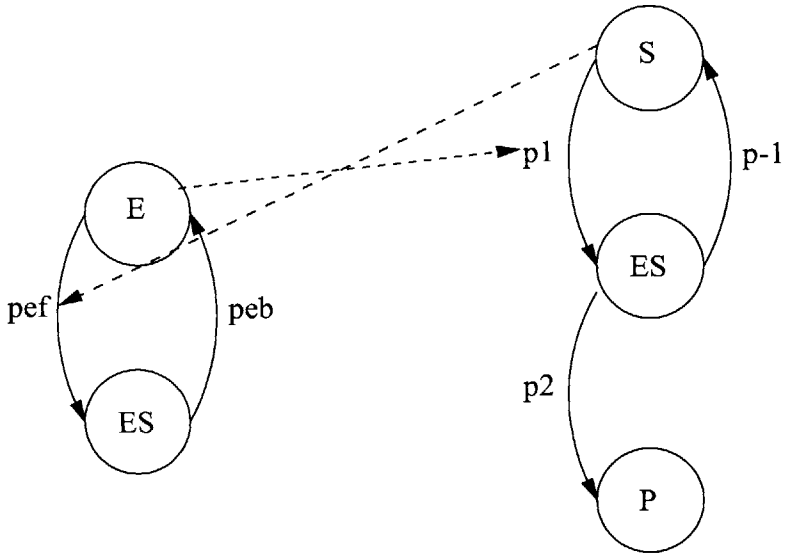


Figure 9-11. Interacting Markov chains illustration of enzyme kinetics.

the figure, $p_1[n]$ is the probability that a substrate collides and binds to a free enzyme and is therefore a function of the probability of an enzyme being in the free state at time n, $p_E[n]$. $p_{ef}[n]$ is the probability that a free enzyme collides and binds to a substrate and is therefore a function of the probability of the substrate-product chain being in the substrate

state at time n, $p_S[n]$. Specifically, we have:

$$\begin{aligned} p_1[n] &= \gamma_1 v_r N_{Et} p_E[n] \\ p_{ef}[n] &= \gamma_1 v_r N_{(S,P)_t} p_S[n] \end{aligned}$$

where N_{Et} and $N_{(S,P)_t}$ are the total numbers of enzyme and substrate-product species respectively. Since the two events (the enzyme transitioning from the free state to the bound state and the substrate transitioning from the free state to the bound state) correspond to the same reactions, the probability of a reaction happening given that the molecules have collided is the same, namely γ_1. In addition, the probability of an enzyme in the bound state transitioning to the unbound state, $p_{eb}[n]$, is equal to the probability of the substrate-enzyme bound complex transitioning out of that state (into the free substrate state or the product state), i.e. we have:

$$p_{eb}[n] = p_{-1}[n] + p_2[n] \tag{9.16}$$

where $p_{-1}[n]$ and $p_2[n]$ are constant. All transition probabilities can be derived from the kinetic rates constants as was demonstrated in the previous chapter. We therefore have:

$$\begin{aligned} p_1[n] &= k_1 \Delta t [E]_t p_E[n] \\ p_{-1}[n] &= k_{-1} \Delta t \\ p_2[n] &= k_2 \Delta t \\ p_{ef}[n] &= k_1 \Delta t [S,P]_t p_S[n] \\ p_{eb}[n] &= (k_{-1} + k_2) \Delta t \end{aligned}$$

where k_1, k_{-1}, and k_2 are the rate constants associated with the enzymatic reaction.

■ 9.5.2 Applying the Michaelis-Menten approximation

In the Michaelis-Menten formalism [92], it is assumed that the enzyme-substrate concentration ([ES]) quickly reaches a steady-state that persists until almost all of the substrate has been consumed. Using the steady-state assumption and the conservation of enzyme species, the Michaelis-Menten equation which gives the rate of change of the concentration of the product can be derived and is given by:

$$\frac{d[P]}{dt} = \frac{k_2}{k_M + [S]} [E]_t [S] \tag{9.17}$$

where k_M is the Michaelis constant and is defined as:

$$k_M = \frac{k_{-1} + k_2}{k_1} \tag{9.18}$$

In a similar manner, we can derive a Michaelis-Menten approximation to the interacting Markov chains model shown in Figure 9-11. Using the substrate-product chain, the probability of being in state ES at time n, $p_{ES}[n]$, can be written as:

$$p_{ES}[n] = p_1[n-1] p_S[n-1] + (1 - p_{-1} - p_2) p_{ES}[n-1] \tag{9.19}$$

where $p_1[n-1]$ is given by:

$$p_1[n-1] = \gamma_1 v_r N_{Et} p_E[n-1] \tag{9.20}$$

Substituting back the expression for $p_1[n-1]$ and using the steady-state assumption for $p_{ES}[n]$, we get:

$$p_{ES} = \gamma_1 v_r N_{Et} p_E[n-1] p_S[n-1] + (1 - p_{-1} - p_2) p_{ES} \tag{9.21}$$

Solving for p_{ES}, we get:

$$p_{ES} = \frac{\gamma_1 v_r N_{Et} p_E[n-1]}{p_{-1} + p_2} p_S[n-1] \tag{9.22}$$

Since the enzyme is assumed to be at steady-state, $p_E[n-1]$ is constant and can be obtained using the enzyme chain. Specifically, we have:

$$p_E[n] = (1 - p_{ef}[n-1]) p_E[n-1] + p_{eb}[n-1] p_{ES}[n-1] \tag{9.23}$$

Using the steady-state assumption and substituting the values for the transition probabilities, we get:

$$p_E[n] = (1 - \gamma_1 v_r N_{(S,P)_t} p_S[n-1]) p_E + (p_{-1} + p_2)(1 - p_E) \tag{9.24}$$

where we have used the fact that $p_E + p_{ES} = 1$. Solving for p_E, we get:

$$p_E = \frac{p_{-1} + p_2}{\gamma_1 v_r N_{(S,P)_t} p_S[n-1] + p_{-1} + p_2} \tag{9.25}$$

Substituting back the value of p_E into equation 9.22, we get:

$$p_{ES} = \frac{\gamma_1 v_r N_{Et}}{p_{-1} + p_2 + \gamma_1 v_r N_{(S,P)_t} p_S[n-1]} p_S[n-1] \tag{9.26}$$

$$= \frac{N_{Et}}{\frac{p_{-1}+p_2}{\gamma_1 v_r} + N_{(S,P)_t} p_S[n-1]} p_S[n-1] \tag{9.27}$$

The probability of being in state P at time n, $p_P[n]$, is given by:

$$p_P[n] = p_2 p_{ES}[n-1] + p_P[n-1] \tag{9.28}$$

substituting the expression for p_{ES}, we get:

$$p_P[n] = \frac{p_2 N_{Et}}{\frac{p_{-1}+p_2}{\gamma_1 v_r} + N_{(S,P)_t} p_S[n-1]} p_S[n-1] + p_P[n-1] \tag{9.29}$$

Defining $p_{sf}[n]$ as:

$$p_{sf}[n] \equiv \frac{p_2 N_{Et}}{\frac{p_{-1}+p_2}{\gamma_1 v_r} + N_{(S,P)_t} p_S[n]} \tag{9.30}$$

we can rewrite equation 9.29 as:

$$p_P[n] = p_{sf}[n-1] p_S[n-1] + p_P[n-1] \tag{9.31}$$

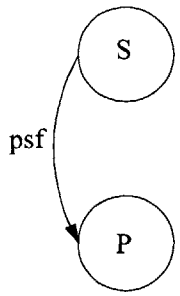


Figure 9-12. Simplified Markov chains model using the Michaelis-Menten approximation.

which can be represented by the simplified Markov chain shown in Figure 9-12. $p_{sf}[n]$ can also be rewritten in terms of the rate constants as follows:

$$p_{sf}[n] = \frac{k_2 \Delta t N_{Et}}{\frac{(k_{-1}+k_2)\Delta t}{(k_1 \Delta t)/(A_V V)} + N_{(S,P)_t} p_S[n-1]} \tag{9.32}$$

$$= \frac{k_2 \Delta t \frac{N_{Et}}{A_V V}}{\frac{k_{-1}+k_2}{k_1} + \frac{N_{(S,P)_t}}{A_V V} p_S[n-1]} \tag{9.33}$$

$$= \frac{k_2 \Delta t [E]_t}{k_M + [S,P]_t p_S[n-1]} \tag{9.34}$$

where k_M is the Michaelis constant of the reaction.

■ 9.6 Updated Two-State Receptor Model

■ 9.6.1 Stochastic enzyme kinetics implementation for methylation by CheR

The two-state receptor model was first updated to include the stochastic implementation of enzyme kinetics for the methylation of the receptor by CheR. The updated conditional receptor chains along with the CheR and CheA chains are shown in Figure 9-13. Tables 9.6 and 9.7 give the new interactions and parameters associated with the chains. All other chains, interactions, and parameters are left unchanged.

$p_{tf1}[n]$	=	$k_{tf1} \Delta t [R] p_R[n]$
$p_{tb1}[n]$	=	$k_{tb1} \Delta t$
$p_{tf2}[n]$	=	$k_{tf2} \Delta t$
$p_{tb2}[n]$	=	$k_{tb2} \Delta t [B] p_{bp}[n-1] A(Tm)$
$p_{tfl1}[n]$	=	$k_{tfl1} \Delta t [R] p_R[n]$
$p_{tbl1}[n]$	=	$k_{tbl1} \Delta t$
$p_{tfl2}[n]$	=	$k_{tfl2} \Delta t$
$p_{tbl2}[n]$	=	$k_{tbl2} \Delta t [B] p_{bp}[n-1] A(TmL)$
$p_{rf}[n]$	=	$k_{rf} \Delta t [T_{tot}] p_T[n]$
$p_{rb}[n]$	=	$k_{rb} \Delta t$

Table 9.6. Interactions of the updated two-state receptor model.

The state probabilities of the chains in the model in response to the input stimulus

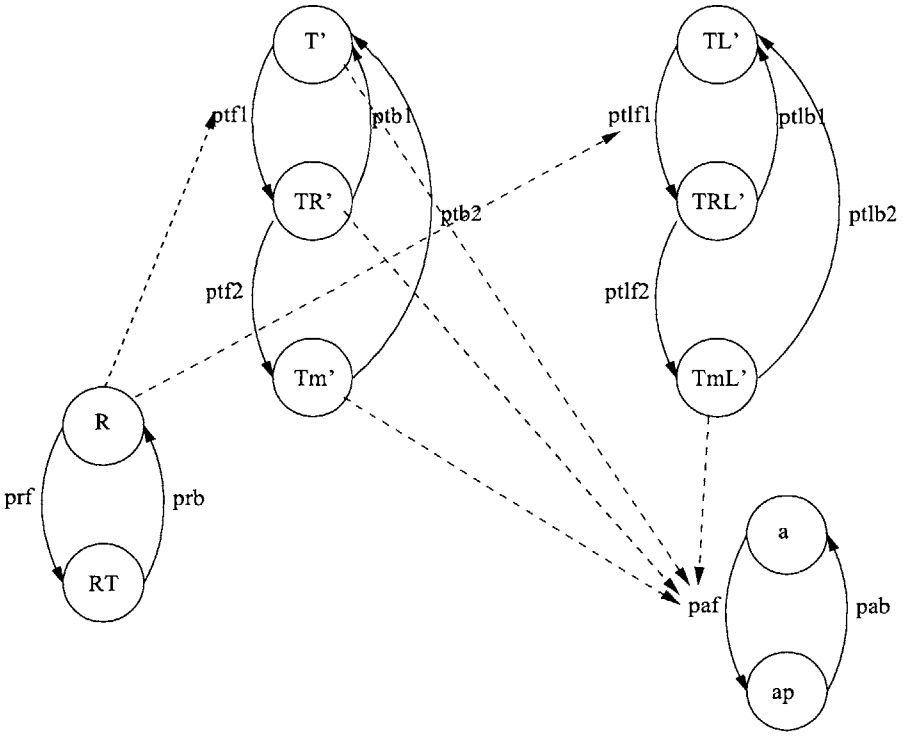


Figure 9-13. Receptor model updated to include the stochastic implementation of enzyme kinetics for the methylation of the receptor by CheR.

shown in Figure 9-6 were computed using the *a*3MC model implementation and are shown in Figure 9-14. The effect of the total number of molecules of CheB and CheR on the behavior of the state probabilities was also investigated. Figures 9-15 and 9-16 show the phosphorylated CheY state probabilities as well as the motor bias computed using the default parameters as well as half and ten times the concentration of CheB and CheR respectively. The results indicate that CheB and CheR have opposite effects on the state probability of CheYp and the bias. This should not be surprising since CheR and CheB have opposite functions (receptor methylation and receptor demethylation). The magnitude of the effect on the state probability of CheYp and the bias due to a change in CheB is comparable to the magnitude of the effect in response to a similar change in CheR. In addition, the enzyme kinetics model implementation for methylation by CheR leads to significantly less methylated receptors than the original model. For example, at saturating stimulus levels (10^{-3}M), the updated two state model predicts less than 20% of the receptors will be methylated (Figure 9-14) while the original model predicts 40% of the receptors will be methylated (Figure 9-7). This deviation between the two models is expected since at this input stimulus level, the probability of the receptor being in the unmethylated ligand bound case is much greater than 0.2. As a result, using the stochastic enzyme kinetics model for the methylation by CheR leads to better adaptation since the lack of adaptation in the original model was mainly due to the large increase in methylated receptors in response to an increase in ligand concentration (input stimulus). The magnitude of the response is also increased using the updated model.

k_{tf1}	=	$8 \times 10^7 M^{-1}s^{-1}$
k_{tb1}	=	$100s^{-1}$
k_{tf2}	=	$0.1s^{-1}$
k_{tb2}	=	$79992M^{-1}s^{-1}$
k_{tlf1}	=	k_{tf1}
k_{tlb1}	=	k_{tb1}
k_{tlf2}	=	k_{tf2}
k_{tlb2}	=	k_{tb2}
k_{rf}	=	k_{tf1}
k_{rb}	=	$k_{tb1} + k_{tf1}$

Table 9.7. Parameter values of the updated two-state receptor model.

■ 9.6.2 Using the Michaelis-Menten approximation

We applied the Michaelis-Menten approximation to the updated two-state model, we call this new updated model: the Michaelis-Menten two-state model. Specifically, we use the same model as the one shown in figure 9-4 however the probabilities of a non-ligand bound and a ligand bound receptor being methylated at time n, $p_{tf}[n]$ and $p_{tlf}[n]$ respectively are now given by:

$$p_{tf}[n] = (k_{tf}\Delta t N_R)/(k_{mtf}A_V V + N_{Ttot}p_T[n-1]) \quad (9.35)$$

$$p_{tlf}[n] = (k_{tlf}\Delta t N_R)/(k_{mtlf}A_V V + N_{Ttot}p_{TL}[n-1]) \quad (9.36)$$

where:

$$k_{tf} = 0.1s^{-1} \quad (9.37)$$

$$k_{tlf} = 0.1s^{-1} \quad (9.38)$$

$$k_{mtf} = 1.25 \times 10^{-6} M^{-1}s^{-1} \quad (9.39)$$

$$k_{mtlf} = 1.25 \times 10^{-6} M^{-1}s^{-1} \quad (9.40)$$

The rest of the interactions and parameters are left unchanged and are as in Tables 9.1 and 9.2. Figure 9-17 shows the state probabilities of the a3MC model in response to the multi-step input shown in figure 9-6. The state probability of CheY being in the phosphorylated state and the bias for varying concentrations of CheB and CheR were also computed and are shown in Figures 9-18 and 9-19 respectively. As was observed for the two-state model using the enzyme kinetics implementation, CheB and CheR have opposite effects on CheYp and bias but similar magnitude.

A 3MC stochastic simulation of the Michaelis Menten two-state model was also implemented. In this implementation, phosphorylated CheB can only demethylate active receptors, i.e. receptors that have their attached CheA phosphorylated. The conditional transition probabilities, which can be readily obtained using the description of the model in Table 9.1 and equation 9.36 are then realized through a Monte-Carlo simulation. As an example, in the implementation, the conditional probability of an unphosphorylated CheA

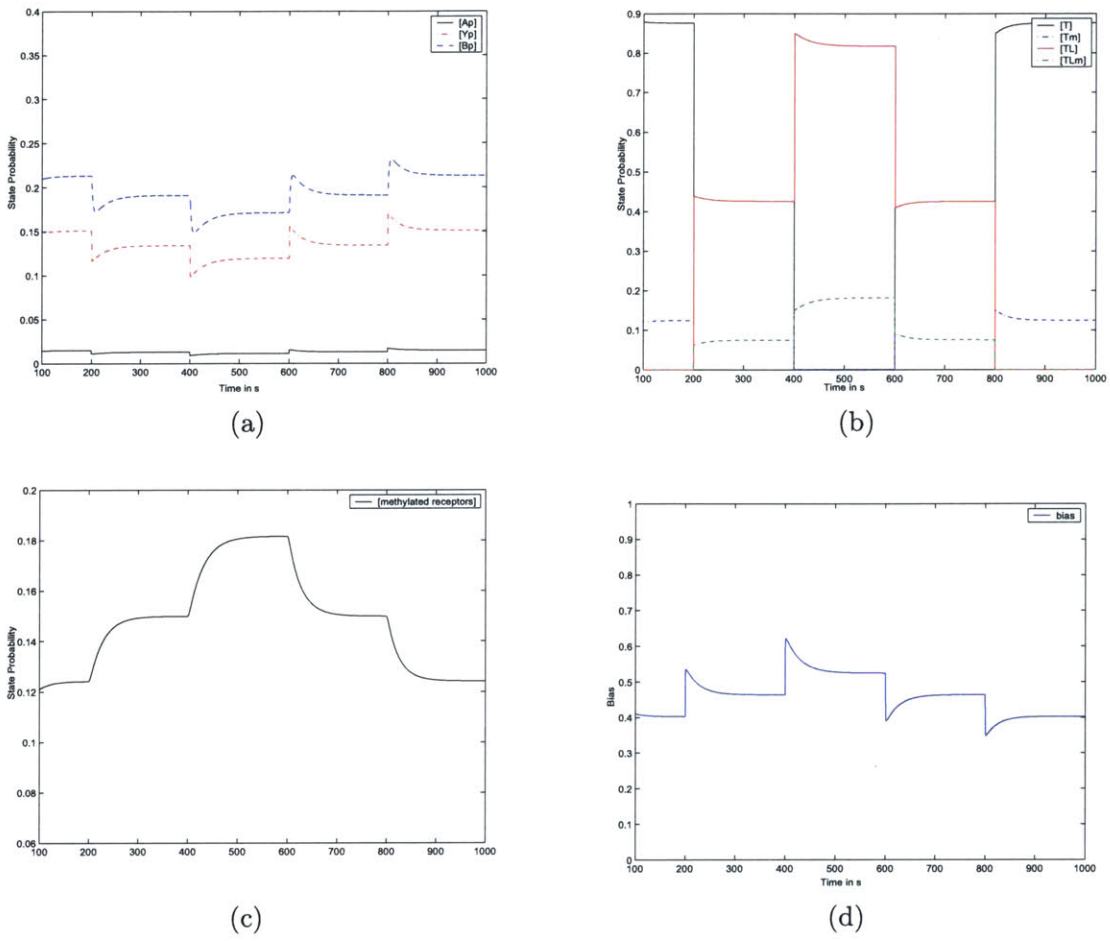
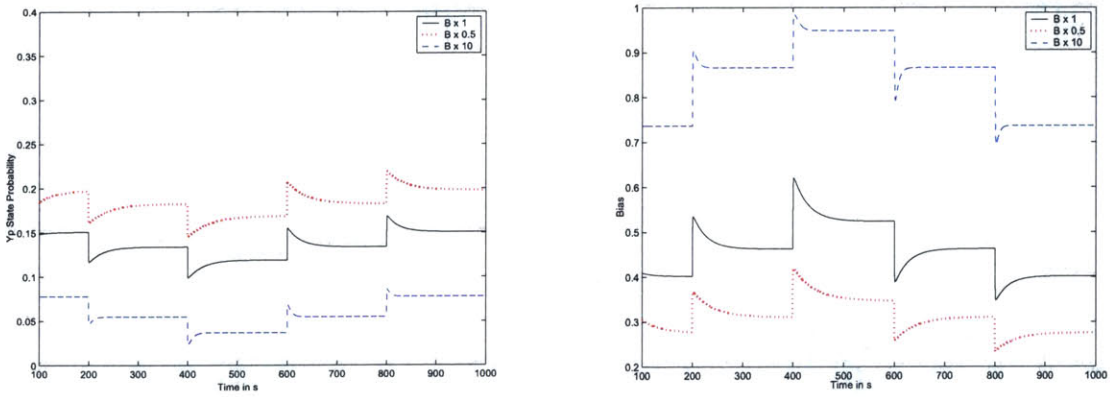


Figure 9-14. State probabilities for the two-state model with enzymatic kinetics implementation. (a) ap, yp and bp, (b) receptor states (c) methylated receptors, (d) motor bias.

getting phosphorylated at time n given the receptor state, $p_{af|t}[n]$, is given by:

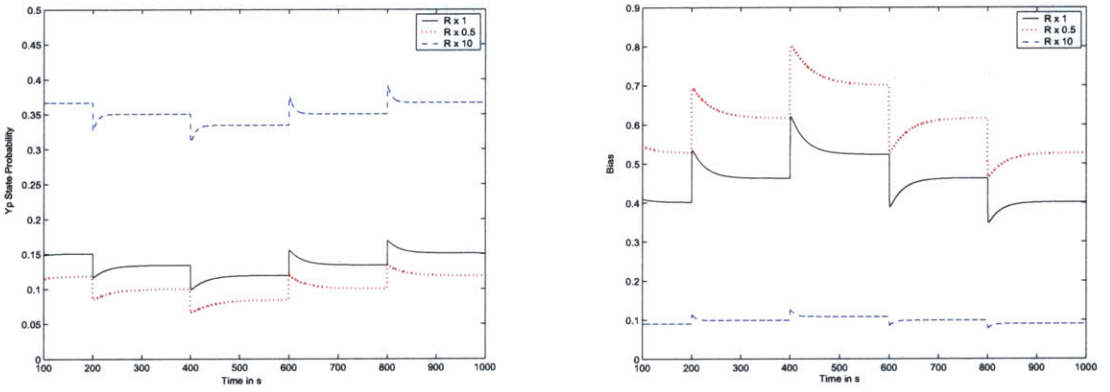
$$p_{af|t}[n] = \begin{cases} k_{af}\Delta tA(T) & : \quad t = T \\ k_{af}\Delta tA(Tm) & : \quad t = Tm \\ k_{af}\Delta tA(TL) & : \quad t = TL \\ k_{af}\Delta tA(TmL) & : \quad t = TmL \end{cases} \tag{9.41}$$

Similar expressions hold for the other states. In addition, it is assumed in these simulations that the CheA protein can fall off the receptor and re-attach to another receptor, i.e. at each time step, a CheA molecule (whether phosphorylated or not) is randomly (using a uniform distribution) paired with a receptor whose state determines the transition probability given in equation 9.41. We performed Monte Carlo simulations of the network response to a step change in input concentration at 400 seconds from 0 to 10^{-3}M. 100 molecules of the receptor and of each signaling protein (CheY, CheB, and CheA) as well as 1 motor were simulated and a total of 100 independent simulations were performed. The average phosphorylated CheY and average bias are shown in Figure 9-20. Note that the average of the 3MC model



(a) Yp (b) Bias

Figure 9-15. State probabilities for different concentrations of CheB of the two-state enzymatic kinetics model.



(a) Yp (b) Bias

Figure 9-16. State probabilities for different concentrations of CheR of the two-state enzymatic kinetics model.

at 0 and 10^{-3}M input concentrations match the state probability obtained from the a3MC model shown in Figure 9-17 which is expected since the number of molecules being simulated is large.

In the 3MC implementation presented so far, it was assumed that the CheA protein can fall off the receptor and pair up with a new one randomly at each time step, i.e. at every Δt. While there is biological evidence supporting the fact that CheA is not always attached to the receptor, the measured time scales at which CheA attaches and falls off are much longer than the time step used in our simulations and therefore a more realistic model would keep CheA attached to the receptor at all times. We therefore carried a second set of stochastic simulations using the 3MC model where now each CheA is attached to the same receptor and, at each time step, the state of the receptor to which a CheA is attached is checked and the transition probabilities of that CheA are set according to equation 9.41. Again, 100 stochastic simulations were carried out using 100 molecules of each signaling protein

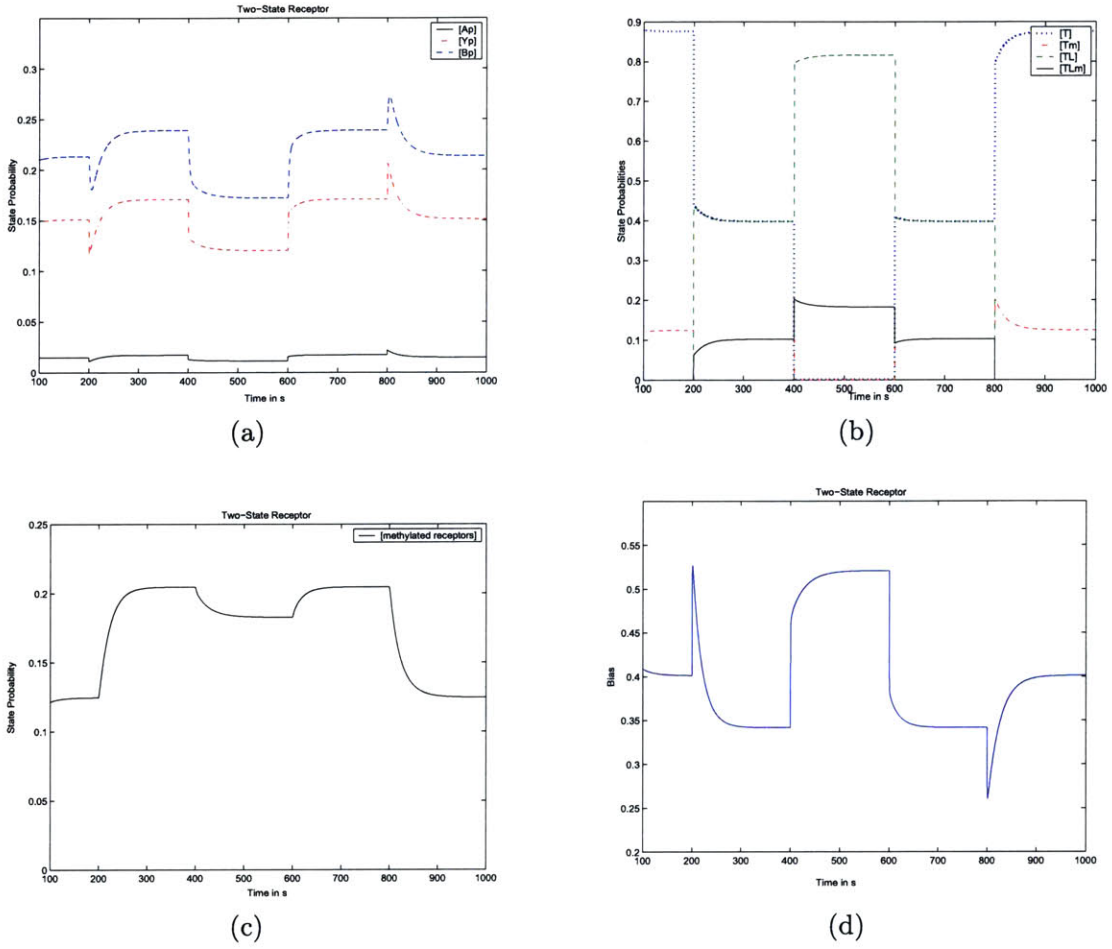
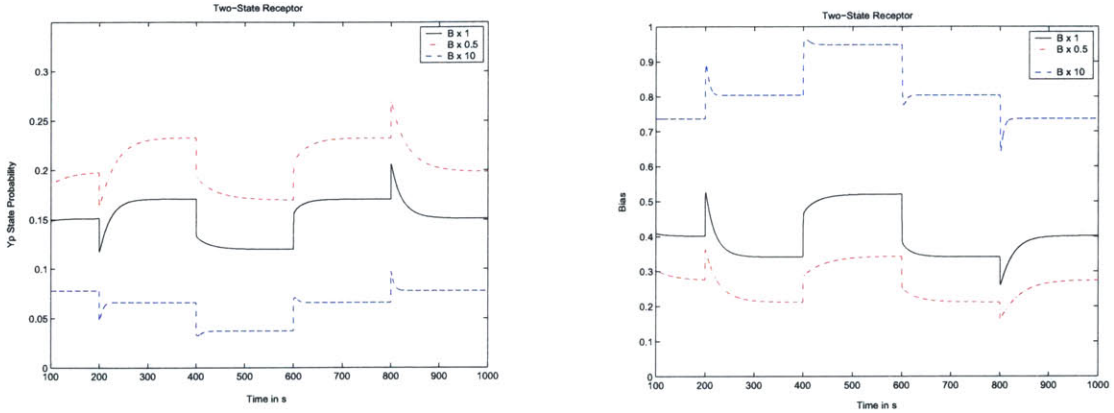


Figure 9-17. State probabilities for the Michaelis-Menten two-state model. (a) ap, yp and bp, (b) receptor states (c) methylated receptors, (d) motor bias.

and one motor. The average number of CheYp and the average bias in response to a step input stimulus at 400 seconds from 0 to 10^{-3}M are shown in Figure 9-21. As is clear from the figure, the average signal levels still match the a3MC model as well as the first version of the stochastic 3MC simulations. However, the variance using this version of the model is slightly smaller than the variance obtained using the first version of the model as can be seen from comparing Figures 9-20 and 9-21. This is due to the fact that keeping the CheA attached to the receptor eliminates one source of variability associated with CheA pairing up with a receptor and therefore the fluctuations in the response are reduced.

Finally, we performed a third set of stochastic simulations where now in addition to CheA not being allowed to fall off the receptor, phosphorylated CheB can only demethylate phosphorylated (as opposed to active) receptors, i.e. receptors that have their attached CheA phosphorylated. This assumption effectively leads to less receptors being demethylated and therefore a higher level of activity compared to the previous versions. 100 simulations using the 3MC model were carried out using 100 molecules of each signaling protein in response to a step input concentration at 400 seconds from 0 to 10^{-3}M. We also computed responses using different total CheR and CheB concentrations. Specifically, we used the default CheR



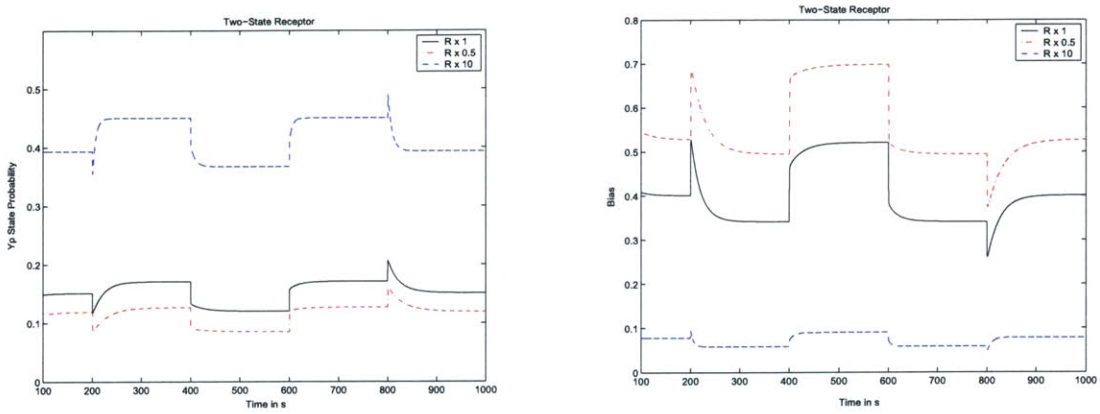
(a) Yp (b) Bias

Figure 9-18. State probabilities for different concentrations of CheB for the Michaelis-Menten two-state model.

and CheB concentrations as well as a ten fold increase in concentration and half the concentration of CheR and CheB. For each modification, 100 realizations were computed. The average phosphorylated CheY for each modification is shown in Figure 9-22. As expected, the level of activity (number of CheYp) is higher for this version of the model than for the previous versions. Furthermore, as was observed in the *a*3MC model implementation, changes in the total concentration of CheR and CheB have opposite effects on the probability of CheYp but similar in magnitude. The distributions of the phosphorylated CheY state for different CheR and CheB concentrations with zero input stimulus are shown in figures 9-23 and 9-24 respectively. The shape of the distribution changes for different CheR and CheB concentrations. Higher concentrations lead to narrower distributions. Furthermore changes in the concentration of CheR tend to have a larger effect on the shape of the distribution than changes in the concentration of CheB. These plots are intended to be illustrations of the types of results one can obtain using this model. To further investigate the implications of the effect of changes in CheR or CheB on the network response, a larger number of simulations needs to be run in order to get a better estimate of the distribution. In addition, while the magnitude of the effect of changes in CheR and CheB concentrations is comparable, the effect of these changes on adaptation precision seems to be different. Specifically, Figure 9-22 suggests that a ten fold increase in CheR concentration leads to a larger adaptation error than a ten fold increase in CheB concentration. Figure 9-25 plots the average percent adaptation error for different values of CheR and CheB as well as the standard deviation. While at wild type CheR and CheB values as well as at half that concentration, the adaptation precision is similar, at ten fold CheR concentration, the adaptation error is double the adaptation error obtained at ten fold the concentration of CheB. However, the standard deviation follows the opposite trend: it is higher for a ten fold increase in CheB than for a ten fold increase in CheR.

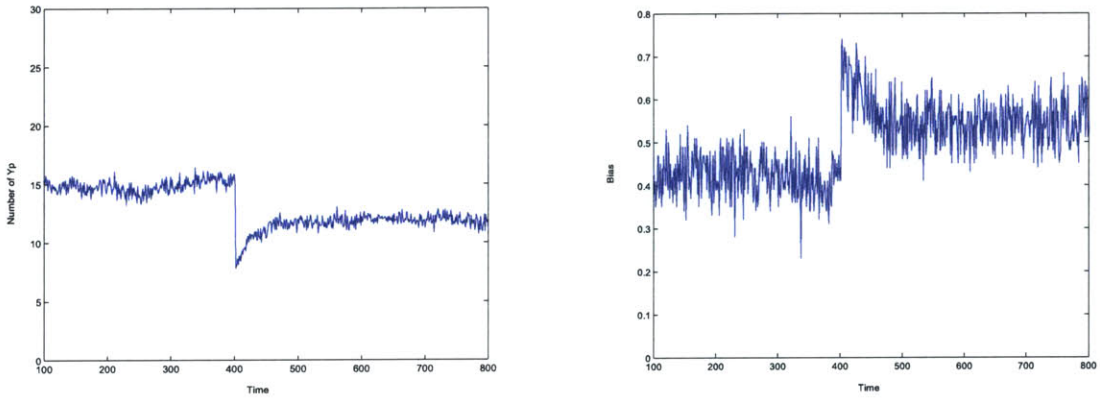
■ 9.7 Six-State Receptor

So far, we have only considered two states for the receptor (methylated and non-methylated), in this section we extend the model and incorporate three methylation states for the recep-



(a) Yp (b) Bias

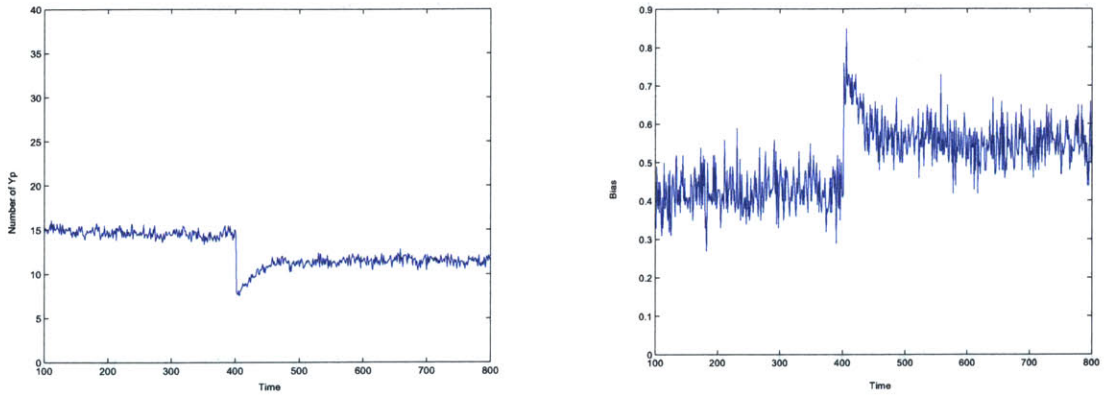
Figure 9-19. State probabilities for different concentrations of CheR for the Michaelis-Menten two-state model.



(a) Yp (b) Bias

Figure 9-20. 3MC implementation of the Michaelis-Menten two-state model.

tor. In fact, the Tar receptor is known to have three different methylation sites [25]. In this version of the model, the receptor is modeled as a six-state receptor where the singly methylated, dually methylated and triply methylated states are explicitly modeled and each one of these states can be in the active form (i.e. the attached CheA is phosphorylated) or in an inactive form. For simplicity, we fuse the unmethylated state into the singly methylated state since their behavior (in terms of activity) is very similar. Figure 9-26 shows the six-state receptor model. In addition, since ligand binding is much faster (more than three orders of magnitude) than methylation and phosphorylation, the effect of the ligand on receptor activation and methylation is modeled through its effect on the respective transition probabilities by using an effective rate constant that is the weighted average of the rates in the ligand bound and ligand unbound cases. Specifically, let $\alpha(L_n)$ be the probability that



(a) Yp (b) Bias

Figure 9-21. 3MC implementation of the Michaelis-Menten two-state model with CheA always attached to the receptor.

a receptor is ligand bound at time n, i.e.

$$\alpha(L_n) = \frac{[L]_n}{K_D + [L]_n} \tag{9.42}$$

and let k_{tf1}^L be the rate constant for receptor methylation if the receptor is singly methylated and ligand bound and k_{tf1} be the rate constant for receptor methylation if the receptor is singly methylated but not ligand bound. Then, the effective rate constant, k_{tf1}^{eff}, for a singly methylated receptor to be methylated at time n, i.e. to become dually methylated is defined as:

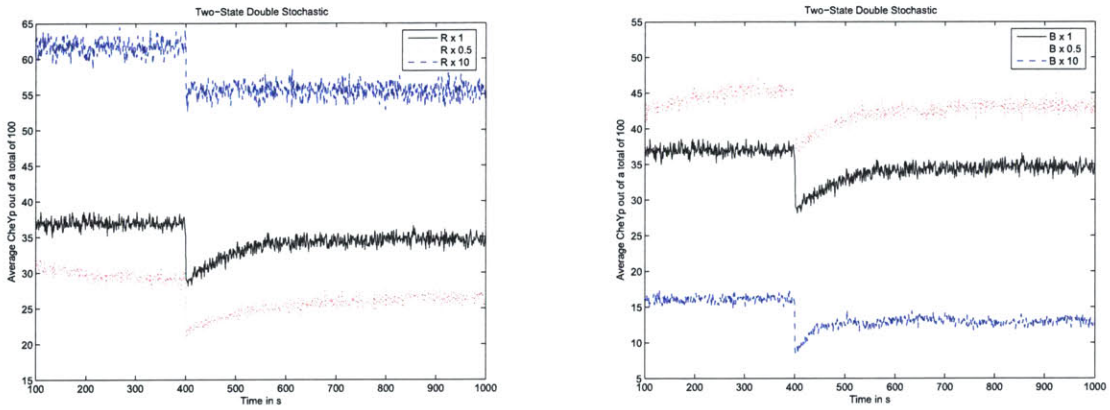
$$k_{tf_1}^{eff} \equiv \alpha(L_n)k_{tf1}^L + (1 - \alpha(L_n))k_{tf1} \tag{9.43}$$

Similar expressions hold for the rate constants associated with transitions to higher methylation states as well as for transitions from unphosphorylated (inactive) to phosphorylated (active) states. In addition, in this model, demethylation is assumed to be independent of ligand binding and the activity of the receptors.

Tables 9.8 and 9.9 give expressions for the transition probabilities as well as the values of rate constants used. The values of the rate constants are obtained from Spiro *et al.* [118] except for the rate constant for the methylation of the receptor in the ligand bound case which is assumed to be 18 times higher than the rate constant for the methylation of the receptor in the ligand unbound case. This value is in agreement with experimentally measured values [118].

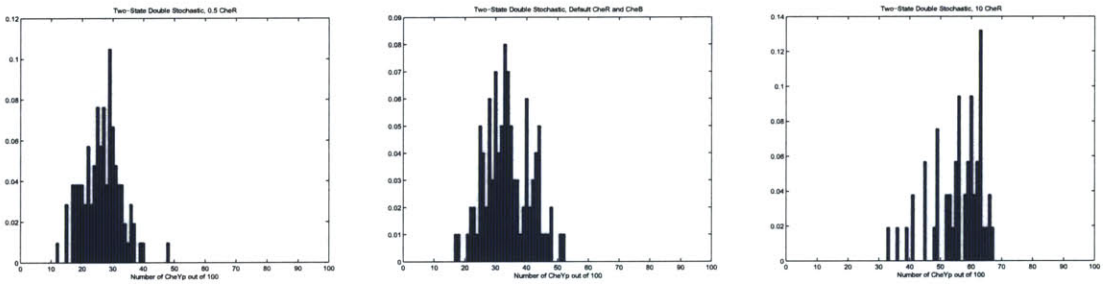
■ 9.7.1 Dynamics of the state probabilities of the a3MC model

Figure 9-27(a) shows the state probabilities for the phosphorylated signaling molecules of the six-state receptor a3MC model in response to the multistep input shown in Figure 9-6. The motor bias is shown in Figure 9-27(b). Interestingly, the six-state receptor leads to perfect adaptation and has a much larger response. In addition, as can be seen from Figure 9-27(b) adding aspartate (i.e. increasing the stimulus) leads to a stronger response than removing aspartate (i.e. decreasing the stimulus) which is in agreement with previous results [15].



(a) (b)

Figure 9-22. Average number of phosphorylated CheY for varying concentrations of CheR (a) and CheB (b) The average is based on simulations of 100 CheY molecules using the Michaelis-Menten two-state model where CheB only demethylates phosphorylated receptors.



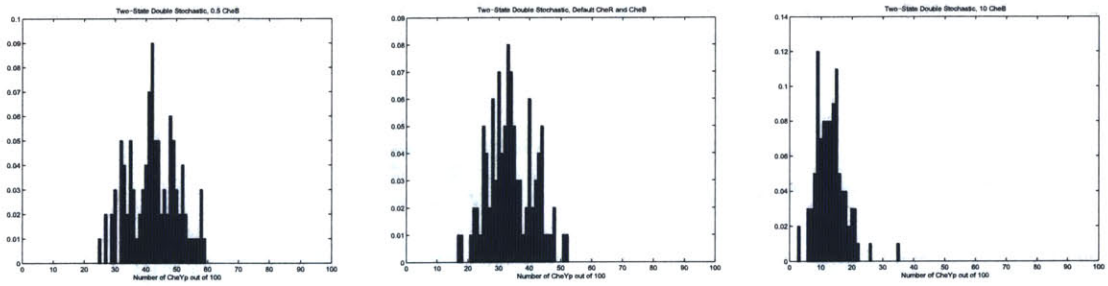
(a) 0.5 x CheR (b) Default (c) 10 x CheR

Figure 9-23. Distribution of the phosphorylated CheY state for different concentrations of CheR using the Michaelis-Menten two-state model where CheB only demethylates phosphorylated receptors.

The state probabilities in response to changes in the total concentration of CheB and CheR were also computed and are shown in Figures 9-28 and 9-29 respectively. Surprisingly, the model is robust to changes in the concentration of CheB but not CheR. Figure 9-28 shows that the response is insensitive to as much as a ten fold increase in CheB concentration however as little as a 0.5 change in the total concentration of CheR leads to a loss of perfect adaptation and a change in the magnitude of the response: higher CheR corresponds to a smaller response while lower CheR leads to a magnified response. Understanding the implication of these findings requires carrying out biological experiments to test the predictions and measure their impact in vivo which is suggested as further work.

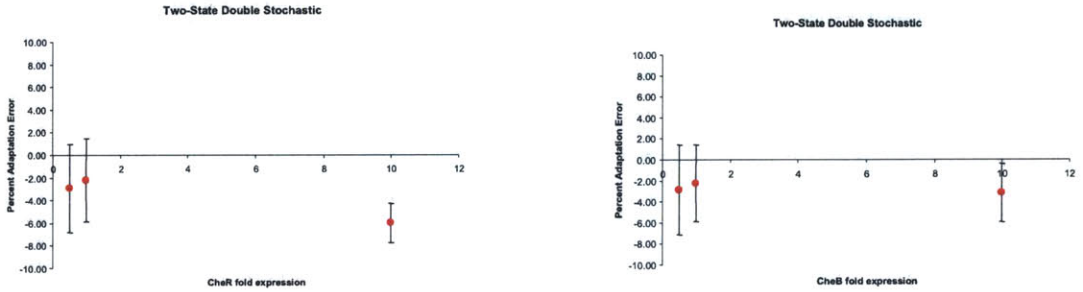
■ 9.7.2 Stochastic simulations using the a3MC model

In order to further investigate the effect of the total concentration of CheR on the network response, we performed stochastic simulations using the *a*3MC model. 1000 molecules of CheY were simulated using different concentrations of CheR and 100 simulations. Histograms of phosphorylated CheY in the absence of stimulus are shown in Figure 9-30. The histograms at different CheR concentrations have similar shape. Specifically, increasing



(a) 0.5 x CheB (b) Default (c) 10 x CheB

Figure 9-24. Distribution of the phosphorylated CheY state for different concentrations of CheB using the Michaelis-Menten two-state model where CheB only demethylates phosphorylated receptors.



(a) (b)

Figure 9-25. Adaptation error for different concentrations of CheR (a) and CheB (b) using the Michaelis-Menten two-state model where CheB only demethylates phosphorylated receptors. Error bars correspond to standard deviations around the average.

the CheR concentration does not make the distribution narrower as was observed for the two-state model (Figure 9-23).

The average normalized adaptation time as well as the average percent adaptation error for different CheR concentrations are shown in Figure 9-31 along with the corresponding standard deviation (error bars). When the error bars are not visible, they are smaller than the size of the circle representing the average. The results indicate that, as was observed for the other models, the adaptation time decreases with increasing CheR concentrations while the standard deviation decreases. The magnitude of the adaptation error, on the other hand, greatly increases with any deviation from the wild type CheR concentration and the standard deviation is extremely low.

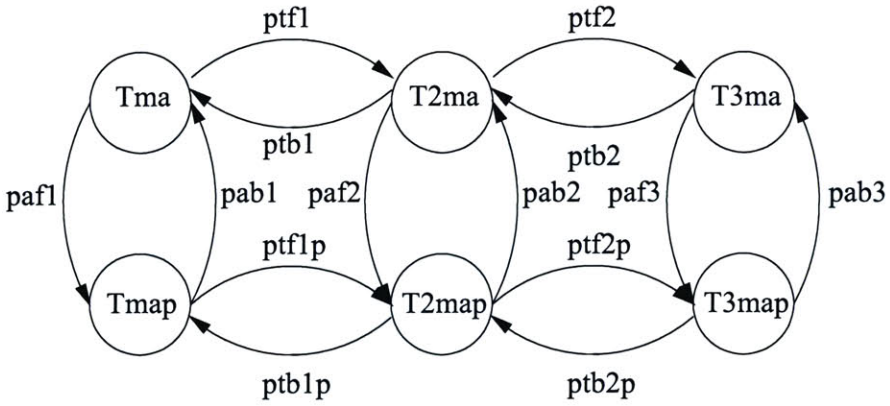
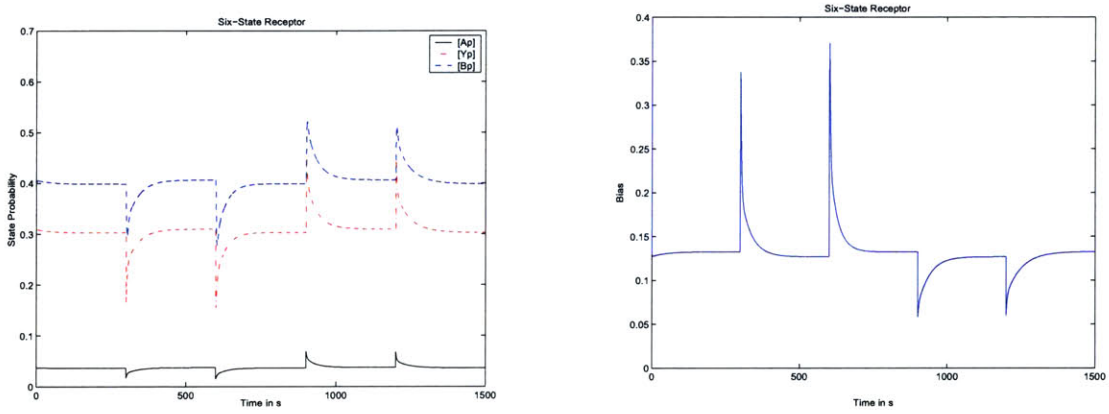
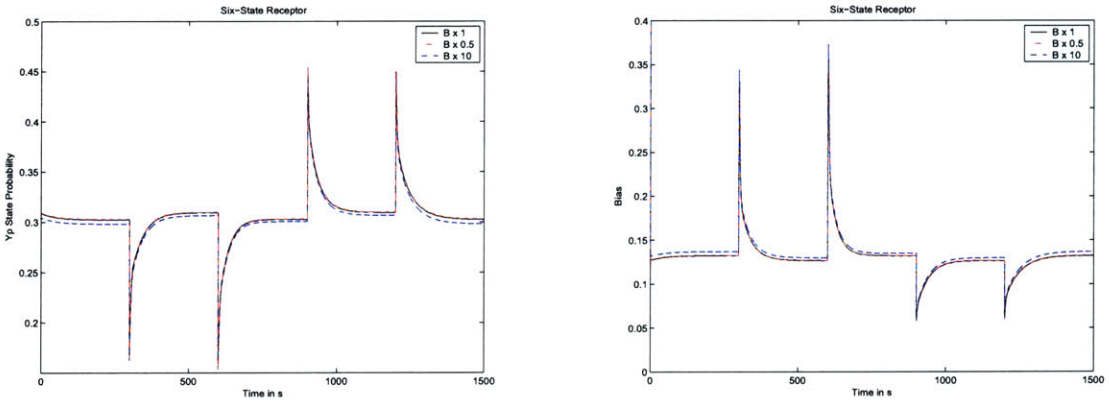


Figure 9-26. Interacting Markov chains model of the six-state receptor



(a) (b) Bias

Figure 9-27. State probabilities for the six state receptor.



(a) Yp (b) Bias

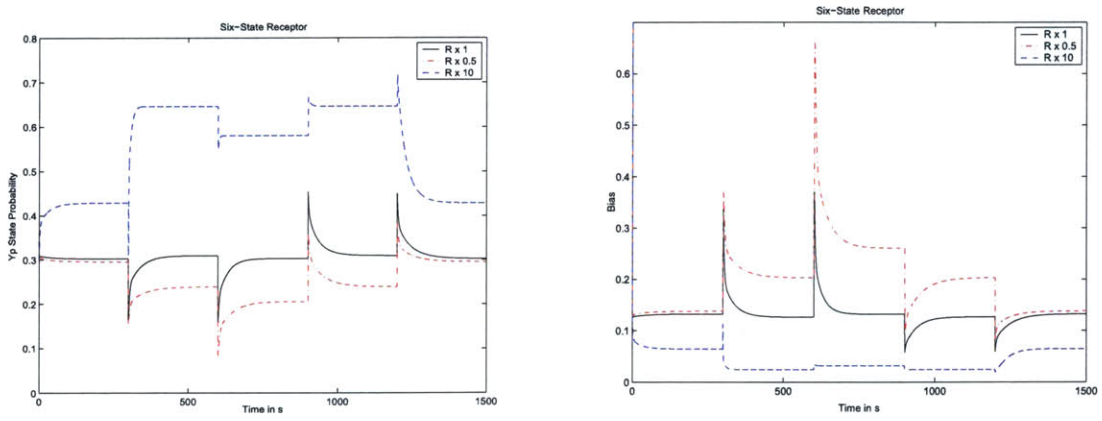
Figure 9-28. State probabilities for varying concentrations of CheB using the six-state receptor model.

$p_{tf1}[n]$	=	$(\alpha(L_n)k^L_{tf1} + (1-\alpha(L_n))k_{tf1}) \times (\Delta t N_R)/(k_{mtf}A_V V + N_{Ttot}p_{Tma}[n-1])$
$p_{tf2}[n]$	=	$(\alpha(L_n)k^L_{tf2} + (1-\alpha(L_n))k_{tf2}) \times (\Delta t N_R)/(k_{mtf}A_V V + N_{Ttot}p_{Tma}[n-1])$
$p_{tf1p}[n]$	=	p_{tf1}
$p_{tf2p}[n]$	=	p_{tf2}
$p_{tb1}[n]$	=	$\gamma_{tb}v_r N_B p_{bp}[n-1]$
$p_{tb2}[n]$	=	$\gamma_{tlb}v_r N_B p_{bp}[n-1]$
$p_{tb1p}[n]$	=	p_{tb1}
$p_{tb2p}[n]$	=	p_{tb2}
$p_{af1}[n]$	=	$(\alpha(L_n)k^L_{af1} + (1-\alpha(L_n)k_{af1})\Delta t$
$p_{af2}[n]$	=	$(\alpha(L_n)k^L_{af2} + (1-\alpha(L_n)k_{af2})\Delta t$
$p_{af3}[n]$	=	$(\alpha(L_n)k^L_{af3} + (1-\alpha(L_n)k_{af3})\Delta t$
$p_{ab1}[n]$	=	$\gamma_{ab_b}v_r N_B p_b[n-1] + \gamma_{ab_y}v_r N_Y p_y[n-1]$
$p_{ab2}[n]$	=	$p_{ab1}[n]$
$p_{ab3}[n]$	=	$p_{ab1}[n]$
$p_{bf}[n]$	=	$\gamma_{bf}v_r N_A(p_{Tmap}[n-1] + p_{T2map}[n-1] + p_{T3map}[n-1])$
$p_{bb}[n]$	=	$k_{bb}\Delta t$
$p_{yf}[n]$	=	$\gamma_{yf}v_r N_A(p_{Tmap}[n-1] + p_{T2map}[n-1] + p_{T3map}[n-1])$
$p_{yb}[n]$	=	$k_{yb}\Delta t$

Table 9.8. Transition probabilities for the six-state receptor model. For bimolecular reactions, the relevant γ is obtained from the relevant k (in $\mathrm{M^{-1}s^{-1}}$) using the expression $\gamma = \frac{k\Delta t}{v_r A_V V}$.

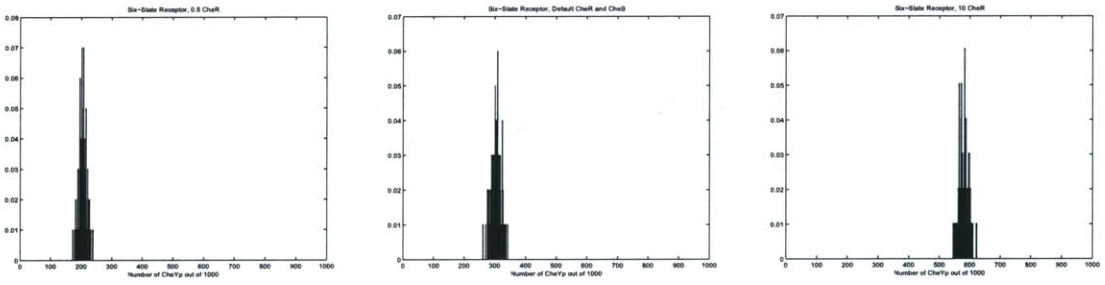
k_{tf1}	= 0.17	k_{af3}	= $3.2k_{af1}$
k_{tf2}	= $0.1k_{tf1}$	k^L_{af1}	= 0
k^L_{tf1}	= $18k_{tf1}$	k^L_{af2}	= $1.1k_{af1}$
k^L_{tf2}	= $18k_{tf2}$	k_{af2}	= $0.72k_{af3}$
k_{mtf}	= 1.7×10^{-6}	k_{ab_b}	= 8×10^5
k_{tb}	= 79992	k_{ab_y}	= 3×10^7
k_{tlb}	= 79992	k_{bf}	= k_{ab_b}
k_{af1}	= 15	k_{yf}	= k_{ab_y}
k_{af2}	= $3k_{af1}$	k_{yb}	= 20

Table 9.9. Six state receptor model rate constants.



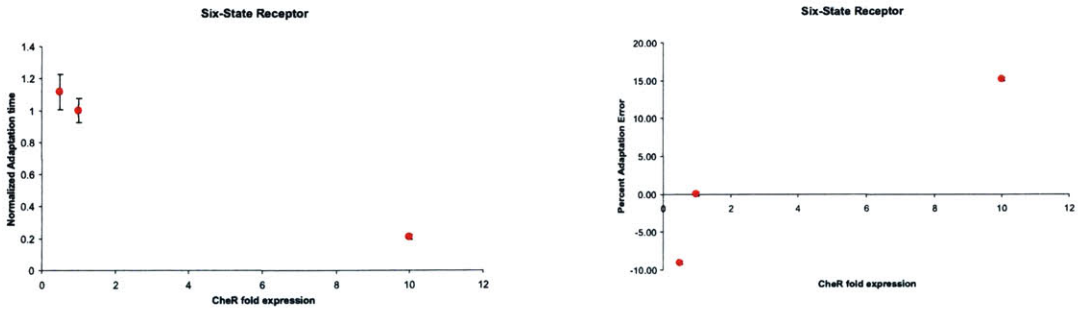
(a) Yp (b) Bias

Figure 9-29. State probabilities for varying concentrations of CheR using the six-state receptor model.



(a) 0.5 x CheR (b) Default (c) 10 x CheR

Figure 9-30. Distribution of the phosphorylated CheY state for different concentrations of CheR using the six-state receptor model.



(a) Adaptation Time

(b) Adaptation Error

Figure 9-31. Normalized adaptation time (a) and percent adaptation error (b) for different concentrations of CheR using the six-state receptor.

■ 9.8 Stochastic Analysis of Single Cell Behavior

The power of a stochastic modeling approach is that it provides a framework for studying the stochastic nature and properties of cellular behavior. In particular, for the case of bacterial chemotaxis, it allows investigation of the behavioral variability in single cells. In fact, Korobkhova *et al.* [78] have recently experimentally measured this variability and suggested that variability is a selected property of the bacterial chemotaxis adaptive system. In particular, they analyzed the power spectrum of the binary time series constructed by monitoring switching events of individual flagellar motors from cells in the absence of attractant (unstimulated). Specifically, the time series is constructed from the clockwise and counterclockwise rotations of a single motor. They found that correlations exist between time points separated by up to 15 minutes which would be unexpected if using the standard assumption that switching events are independent and governed by a Poisson process. In this section, we use our model to generate similar time series and analyze the effect of changes in CheR and CheB concentrations on these events.

■ 9.8.1 Correlations underlying the binary time series of switching events

Binary time series of switching events of unstimulated bacteria were generated using three models: the a3MC implementation of the six-state receptor model, the 3MC implementation of the updated two-state model with the enzyme kinetics implementation for CheR, and the 3MC implementation of the Michaelis-Menten two-state model with CheB acting only on phosphorylated receptors. For the six-state receptor model, each time series was 7200 seconds long while for the two-state receptor models, the time series were 2000 seconds long. Simulations were carried out using wild type concentrations of CheR and CheB as well as half and ten times the concentration of CheR and half and ten times the concentration of CheB. The estimate of the power spectral density of the time series was then computed using periodogram averaging with a rectangular window length of 10 seconds, an overlap of half the window size, and a 2^{18} point FFT. The power spectra of individual time series with varying CheR concentration for the six-state receptor are shown in Figure 9-32. Figures 9-33 and 9-34 show power spectra computed using the two-state receptor model with stochastic CheR enzyme kinetics as well as the Michaelis-Menten two-state receptor model with varying concentrations of CheR and CheB. The average power spectral density is shown in Figure 9-35 for the Michaelis-Menten two-state receptor model obtained by

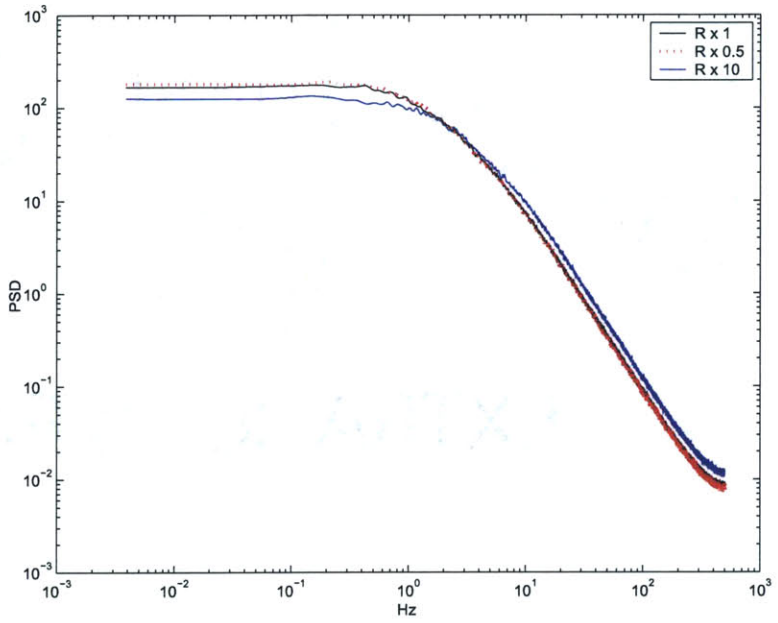


Figure 9-32. Power spectral density of the six-state receptor a3MC model.

averaging ten individual power spectra.

The results in Figures 9-32, 9-33, 9-34, and 9-35 indicate that some degree of correlation exist between different time points in the binary time series. Interestingly, while the correlation seems to disappear for frequencies below 1Hz in the a3MC six-state receptor indicating that events more than one second apart are uncorrelated, this does not seem to be the case for the 3MC two-state model implementation. In the 3MC two-state model implementation, correlations seem to exist between events more than 100 seconds apart for wild-type concentration of CheR and CheB which is in agreement with the results reported by Korobkhova *et al.* [78] even though the correlation observed for distant time points is much weaker than the one reported in [78]. However, interestingly increasing the CheB concentration leads to a less flat power spectrum (Figure 9-34b) i.e. it leads to correlations extended for longer times in the binary time series. These results are interesting because based on the changes in CheR concentration results one may speculate that temporal behavioral variability has been selected for through evolution because small changes in CheR lead to a suppression of this variability which seems to have been selected against. These results therefore suggest that some degree of *controlled* temporal behavioral variability seems to have been selected for. The advantages of this variability remain to be determined, However one can start analyzing its effects using the kind of models developed here.

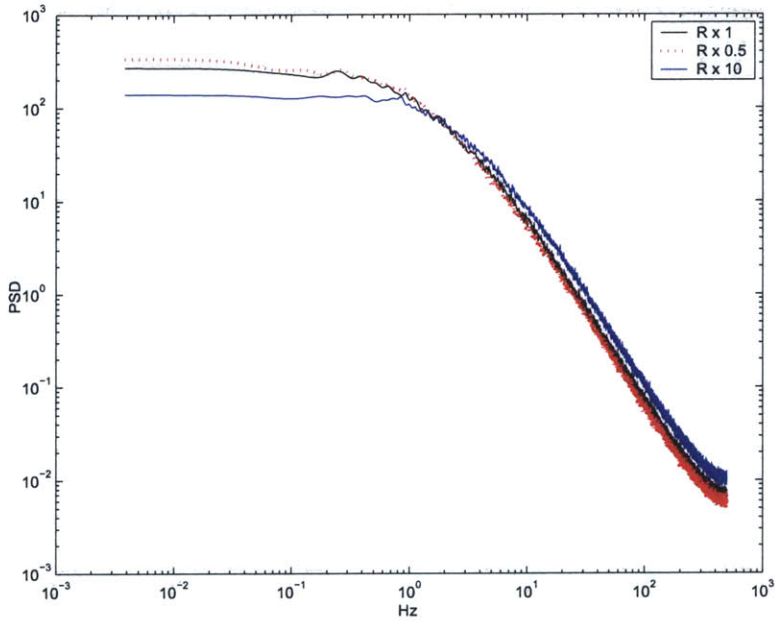
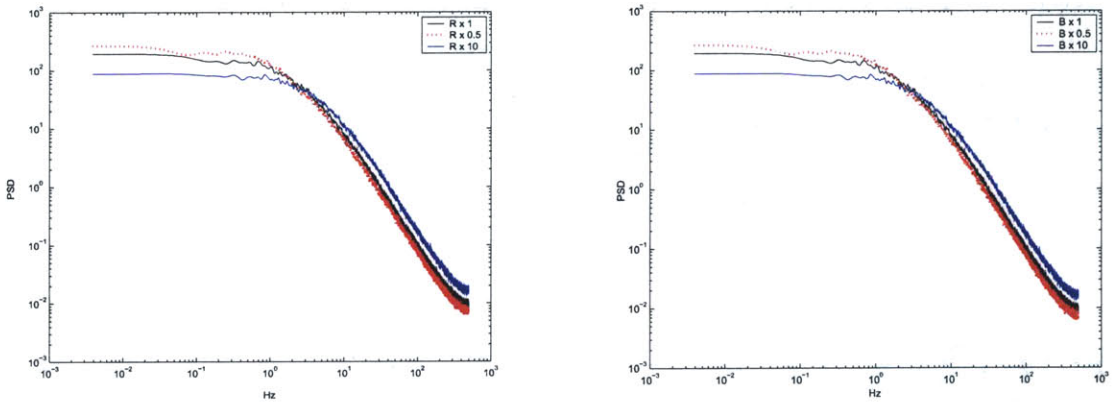


Figure 9-33. Power spectral density of the two-state model with stochastic enzyme kinetics 3MC implementation.



(a) (b)

Figure 9-34. Power spectral density of the Michaelis-Menten two-state 3MC implementation.

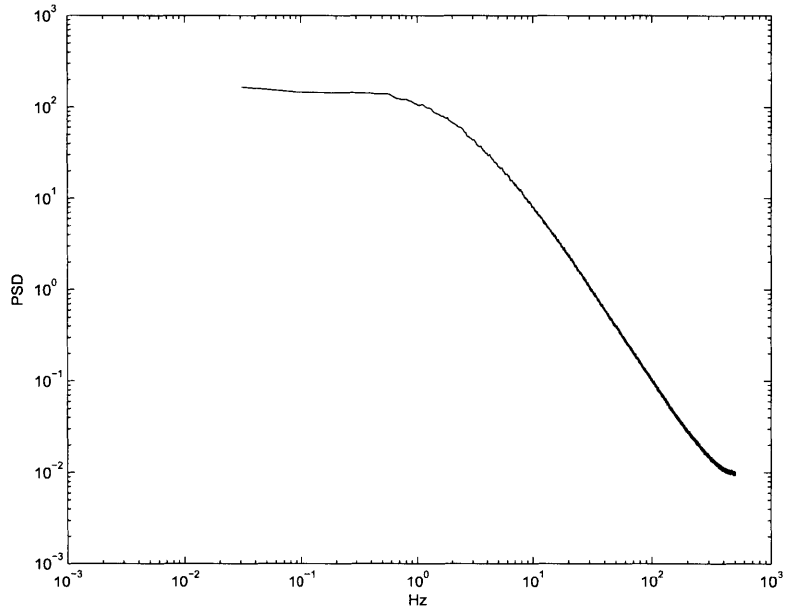
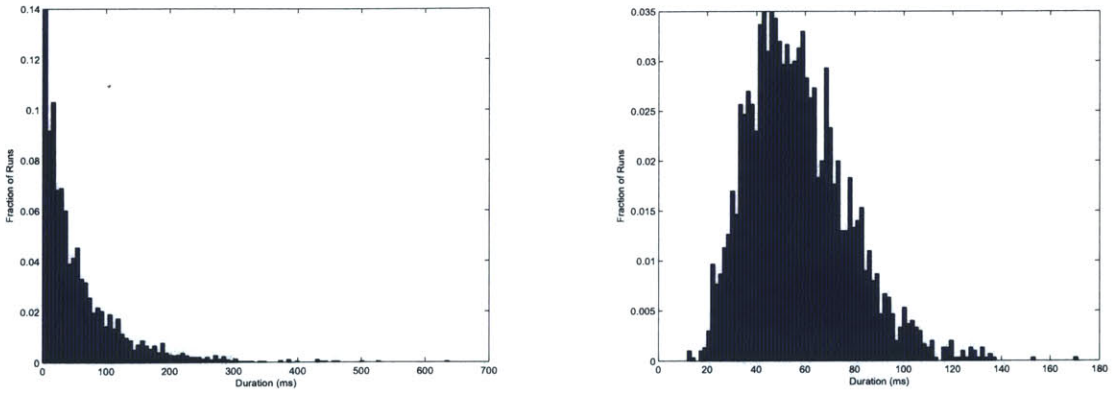


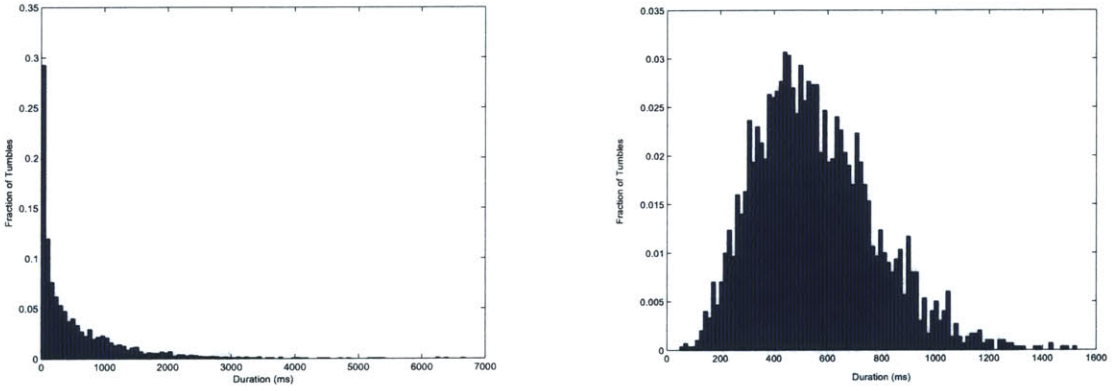
Figure 9-35. Average power spectral density of the two-state Michaelis-Menten 3MC model. This plot was generated using a 2^{15} point FFT.

■ 9.8.2 Distributions of run and tumble lengths



(a)Single cell (b) Average cell

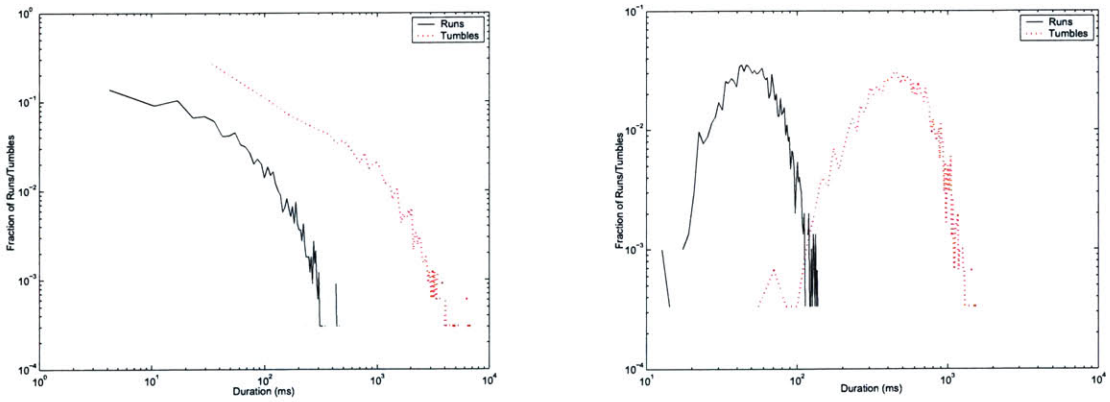
Figure 9-36. Distribution of run lengths computed using the Michaelis-Menten two-state model.



(a)Single cell (b) Average cell

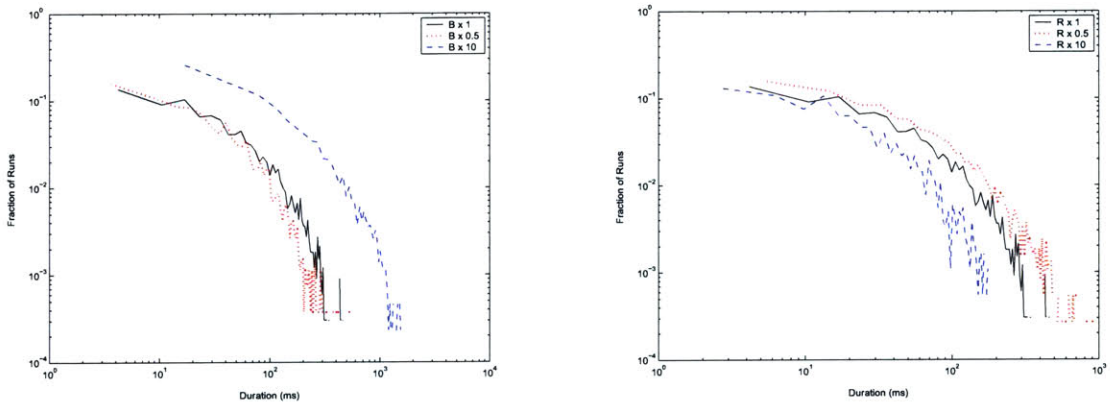
Figure 9-37. Distribution of tumble lengths computed using the Michaelis-Menten two-state model.

In addition to correlations in the binary time series, we have analyzed the nature of the runs and tumbles durations for the 3MC Michaelis-Menten two-state receptor model. Specifically, we have computed the distribution of run and tumble lengths for one simulation lasting 2000 seconds as well as for an average of ten simulations which correspond to taking the output from ten bacteria and averaging the lengths of each run and tumble. The individual and average cell distributions of run and tumble lengths are shown in Figures 9-36 and 9-37 respectively as well as in Figure 9-38 on a logarithmic scale. The single cell results are quantitatively and qualitatively very different from the average cell behavior stressing the importance of single cell measurements. The distributions of run and tumble lengths for a single cell with varying total concentration of CheR and CheB was also computed. The results are plotted on a logarithmic scale and are shown in Figures 9-39 and 9-40. Again, as was observed for the stochastic simulations of the signaling molecules, changes in the concentration of CheB and CheR have opposite effects on changes in the distributions of run and tumble lengths. Furthermore, the effect of one species (CheR or CheB) is opposite



(a) Single cell (b) Average cell

Figure 9-38. Distribution of run and tumble lengths computed using the Michaelis-Menten two-state model.

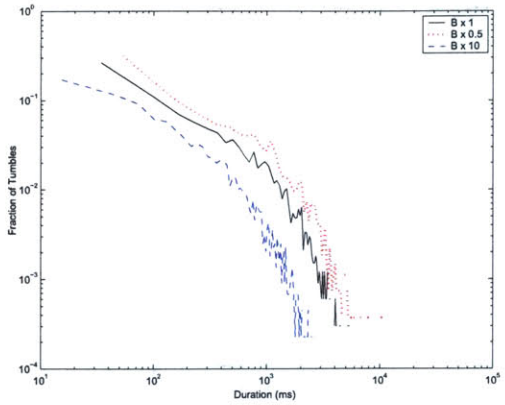


(a)Changing CheB (b) Changing CheR

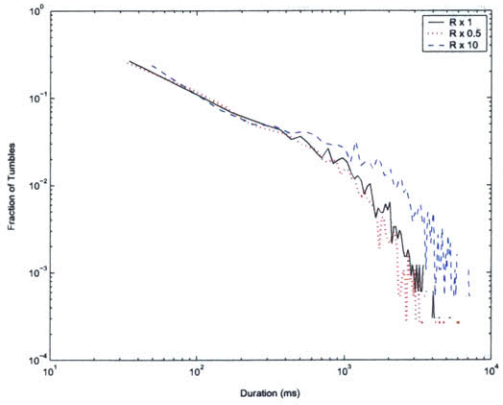
Figure 9-39. Distribution of run lengths for various CheB and CheR concentrations computed using the Michaelis-Menten two-state model.

on the run lengths distributions than on the tumble lengths distributions. Surprisingly, the nature of the effect on the distribution is very different whether the concentration of CheR or the concentration of CheB is modified. As can be seen from Figures 9-39(a) and 9-40(a), changing the CheB concentration simply shifts the distribution of runs and tumbles length, i.e. it increases or decreases the duration of all tumbles or runs by the same amount. However changing the CheR concentration has a very different effect. In fact, changing the concentration of CheR has essentially little to no effect on the relative number of short tumbles or runs however the relative number of longer tumbles and runs is changed dramatically. This observation is much more marked for tumbles than for runs.

These results reveal some insights as to the mechanism of CheR and CheB actions and the level of control that can be obtained by modifying their concentrations. Of course, experiments need to be carried out to further validate the results. However, this type of analysis stresses the importance of developing theoretical frameworks for studying the stochastic behavior of single cells.



(a)Changing CheB



(b) Changing CheR

Figure 9-40. Distribution of tumble lengths for various CheB and CheR concentrations computed using the Michaelis-Menten two-state model.

■ 9.9 Summary

In this chapter, we have demonstrated the use of the interacting Markov chains framework developed in Chapter 7 and have expanded it to include enzymatic kinetics as well as approximations based on Michaelis-Menten assumptions. The specific case of bacterial chemotaxis was chosen because of the wealth of both theoretical and experimental information available for this system as well as the fact that it is naturally suited for stochastic modeling since the behavior of the system is inherently random. We developed several versions of chemotactic models and explored both the a3MC and 3MC implementations which allowed us to draw conclusions about various aspects of the signaling pathway and the difference between individual and average cell behavior. The main purpose of the results presented in this chapter was to provide the reader with an in-depth example as to how our modeling framework could be used to investigate signaling pathways. It also provided validation of the models by comparing them to experimental data. While one can draw some hypotheses and preliminary conclusions using the results presented here, biological experiments need to be carried out to further validate the predictions of the models. It is the hypotheses and predictions as well as the proposal of new sets of relevant biological experiments to validate the model that is the most valuable contribution of this modeling technique.

Chapter 10

Using Biology as a Metaphor: A Surface Mapping Algorithm Based on Bacterial Chemotaxis

We have so far, in this thesis, focused on developing models for understanding the transmission and processing of signals in biological cells. Our work has led to the formulation of a new stochastic framework that was tested on a number of signaling networks including bacterial chemotaxis. While the understanding of the mechanisms underlying biosignaling is an important scientific endeavor, exploiting our understanding to formulate new algorithms for non-biological applications is an important engineering objective. This chapter presents a preliminary example of such an effort by formulating a new surface mapping algorithm inspired by the bacterial chemotaxis pathway investigated in the previous chapter. This algorithm falls under a general class of algorithms that use nature as a metaphor. Many of these algorithms have been developed in the context of optimization. We therefore start by presenting some examples of such algorithms in the context of optimization followed by the formulation of a bacterial chemotaxis a3MC based algorithm for surface mapping. Results based on different test functions are then presented and discussed.

■ 10.1 Nature as a Metaphor

The field of optimization has been the source of many problems addressed by algorithms inspired from Nature. It is defined in the context of an optimization problem which broadly refers to the problem of having a different number of possible solutions and a clear notion for assessing and comparing the quality of different solutions. Optimization is simply the collection of techniques leading to the best (in some well defined meaning) possible solution to the optimization problem. Some optimization problems can be fundamentally hard, i.e. they cannot be solved in a reasonable time [64]. In these cases, approximate methods provide efficient methods for solving hard problems by relaxing the constraint that the solution needs to be the best, instead a 'good enough' solution is generated. A number of approximate methods use various natural phenomena as a source of inspiration for formulating new optimization techniques. In the following subsections we briefly review a few examples of such methods. For a more extensive review, the reader is referred to [36] and [108].

■ 10.1.1 Simulated annealing

Simulated annealing falls under the broad class of local search methods where a special 'current' solution is maintained and its neighbors, i.e. new candidates that are only slightly different from it, are explored to find a better solution [36]. The algorithm is based on an analogy between the way a metal cools and freezes into a minimum energy crystalline structure (the annealing process) and the search for a minimum in a more general sense. It is based on the algorithm of Metropolis *et al.* [91] which provides a means of finding the equilibrium configuration of a collection of atoms at a given temperature. This algorithm was connected to mathematical minimization by Pincus [103] and was later proposed as the basis for an optimization technique for combinatorial problems by Kirkpatrick *et al.* [77]. The implementation of the simulated annealing algorithm is very straightforward and consists of four elements: a representation of possible solutions, a generator of random changes in solutions, a means of evaluating the problem functions, and an annealing schedule i.e. an initial temperature and rules for lowering it as the search progresses. Furthermore the algorithm is guaranteed to converge to the global optimum as the number of iterations approaches infinity if some conditions on the annealing schedule are met. For a detailed description of the algorithm, the reader is referred to [70] and [33]. The major advantage of the simulated annealing algorithm over other methods is that it avoids becoming trapped

at local minima by using a random search which not only accepts changes that decrease the objective function but also accepts, with a small probability, some changes that increase it [66]. In the course of the optimization process the process of accepting deteriorations decreases slowly towards zero.

■ 10.1.2 Genetic algorithms

In contrast to simulated annealing algorithms, genetic algorithms are population based search algorithms where the notion of a single current solution is replaced by a population of current solutions. The algorithms generate new solutions by first selecting members of the population as 'parents' and then making changes to the parents to produce 'children' [36]. In the case of genetic algorithms, the selection and reproduction process is inspired by Darwin's theory of evolution where problems are solved by an evolutionary process resulting in a best (fittest) solution (survivor), i.e. the solution is evolved. Specifically, the algorithm begins with a set of solutions which are represented by chromosomes and are called a population. Solutions from one population are selected and used to form a new population that is hopefully better than the first one. The selection process uses a fitness function which assesses the suitability of each solution: the more suitable a solution is the higher the chance it has to reproduce. This process is repeated until some condition, such as the number of populations or the quality of the best computed solution is satisfied. In addition, the reproduction process (creating new solutions using previously computed ones) uses two diversifying operations inspired from biology: genetic recombination and mutation. Genetic recombination is implemented through a crossover probability which is used to cross over the parents to form new offspring (children). If no crossover was performed, offsprings are exact copies of parents. Furthermore, a mutation probability (which is usually low) mutates new offspring at each locus by, for example, flipping a bit in the individual's bit string representation. For more details on this algorithm, the reader is referred to [69] and [132].

■ 10.1.3 Ant colony optimization

Ant algorithms in general are multi-agent systems in which the behavior of a single agent (artificial ant) is inspired by the behavior of real ants [36]. A particularly successful class of algorithms is the Ant Colony Optimization (ACO) algorithms which are inspired from experiments by Goss *et al.* [60] which showed that, after going through a brief transitory phase, most ants use the shortest paths if faced with different lengths paths leading to the same food source. ACO algorithms have been applied successfully to a large number of difficult combinatorial problems like the quadratic assignment and the traveling salesman problems, to routing in telecommunication networks, scheduling, and other problems [36]. The basic idea is that a large number of simple artificial agents are able to build good solutions to hard combinatorial optimization problems via low-level based communications such as the one used by real ants. A common memory which corresponds to the pheromone deposited by real ants is used in order for artificial ants to cooperate. The artificial pheromone is then accumulated at run-time through a learning mechanism. A detailed description of this class of algorithms can be found in [87].

■ 10.1.4 Optimization based on bacterial chemotaxis

In the same way as ants, genetics, and physics have been an inspiration for optimization algorithms, there have been a number of studies investigating the motion of bacteria as a

k_{tf}	=	$79992/1.2$
k_{tlf}	=	$79992/1.2 \times \sqrt{8.403}$
k_{tb}	=	79992×1.2
k_{tlb}	=	$79992 \times 1.2/\sqrt{8.403}$
k_{af}	=	45
k_{ab_b}	=	8×10^5
k_{ab_y}	=	3×10^7
k_{bf}	=	k_{ab_b}
k_{bb}	=	0.35
k_{yf}	=	k_{ab_y}
k_{yb}	=	20

Table 10.1. Model rate constants used for the surface mapping algorithm.

potential metaphor for optimization techniques specifically for the search for the maximum of a function [29], [30], [8], [16], [17], [98], and [99]. While the algorithms developed in these studies are inspired from bacterial chemotaxis, they are mainly behavioral algorithms i.e. are built using our understanding of the behavior of bacteria searching for food. In particular, they do not incorporate mechanistic knowledge as to how bacteria process environmental information internally leading to system behavior. For example, none of these algorithms incorporate the notion of adaptation which, as was investigated in the previous chapter, is a fundamental system property allowing the bacterium to have a wide dynamic range.

In the remainder of this chapter we present a preliminary formulation of an algorithm based on the molecular mechanism of bacterial chemotaxis. The potential applications of the algorithm is extended beyond the search for the maximum of a function to the problem of surface mapping and flattening.

■ 10.2 An Interactive Markov Chains Algorithm for Surface Mapping Based on Bacterial Chemotaxis

Bacterial chemotaxis is ultimately about searching a volume for nutrients (aspartate) where the bacterium moves towards higher nutrient concentrations and moves away from lower concentrations or repellents effectively mapping the nutrient concentration surface. As a result, bacteria can be thought of as surface mappers or optimizers whereby they scan a three-dimensional function (the nutrient concentrations) and try to spend most of their time on the function's maxima. This behavior can be directly translated into a signal processing algorithm that uses local information to evaluate and map a function. While most current optimization algorithms inspired from Nature focus on searching for the extremum of a function, the algorithm we formulate here provides an approximate solution to the value of a function at every point in the space and therefore constitutes a broader optimization algorithm. Later in this chapter, we will modify the algorithm to solve a newly formulated optimization problem consisting of flattening a surface. In this section, we formulate the surface mapping algorithm and we evaluate its performance in the following sections.

■ 10.2.1 BASM: Bacterial Algorithm for Surface Mapping

The Bacterial Algorithm for Surface Mapping (BASM) we propose is based on the interacting Markov chains model of bacterial chemotaxis presented in the previous chapter. The model schematic is the same as the one in Figures 9-3, 9-4, and 9-5 in the previous chapter where the interactions are given by Tables 9.1, 9.3, and 9.4. However the parameter values we use here are modified in order to amplify the response of the network and reach perfect adaptation. This is achieved by scaling the rate constants for the methylation and demethylation of the receptor as is shown in Table 10.1, increasing the rate constants associated with the backward probabilities in the motor chain as shown in Table 10.2, and changing the activities of the different receptor states as follows:

$$\begin{aligned} A(T) &= 0.15 \\ A(Tm) &= 0.95 \\ A(TmL) &= 0.4 \\ A(TL) &= 0 \end{aligned}$$

k_{mf1} =	7×10^6	k_{mb1} =	1.43×2.5
k_{mf2} =	6×10^6	k_{mb2} =	2.86×2.5
k_{mf3} =	5×10^6	k_{mb3} =	4.29×2.5
k_{mf4} =	4×10^6	k_{mb4} =	5.72×2.5
k_{mf5} =	3×10^6	k_{mb5} =	7.15×2.5
k_{mf6} =	2×10^6	k_{mb6} =	8.58×2.5
k_{mf7} =	1×10^6	k_{mb7} =	10.01×2.5

Table 10.2. Motor model rate constants used for the surface mapping algorithm.

The a3MC implementation of the model is used in order to make the algorithm more efficient. Specifically, the state probabilities are determined and only the relevant Markov chain, the motor chain, is stochatically simulated. Figures 10-1 and 10-2 show the dynamics of the state probabilities for a step increase in input from 0 to 10^{-3}M and from 10^{-3}M to 10^{-6}M respectively. The initial transient response corresponds to the chains equilibrating and should be ignored. When running the algorithm, the chains are initially trained so that they equilibrate before processing the input thereby eliminating the initial transient response. The figures show the amplified magnitude of the response and perfect adaptation over different input stimuli. Using this model as a basis, we formulated both two-dimensional and one-dimensional versions of the algorithm.

■ 10.2.2 Two-dimensional BASM

The motion and position of the bacterium in a two-dimensional search space can be characterized by three variables: (1) position, i.e. its current (x,y) coordinates, (2) angle or direction of motion, and (3) the state of the motor, i.e. running (rotating counterclockwise) or tumbling (rotating clockwise) between time indices n and $n+1$. During a run, the position changes but the direction does not while during a tumble the direction changes while the position remains the same.

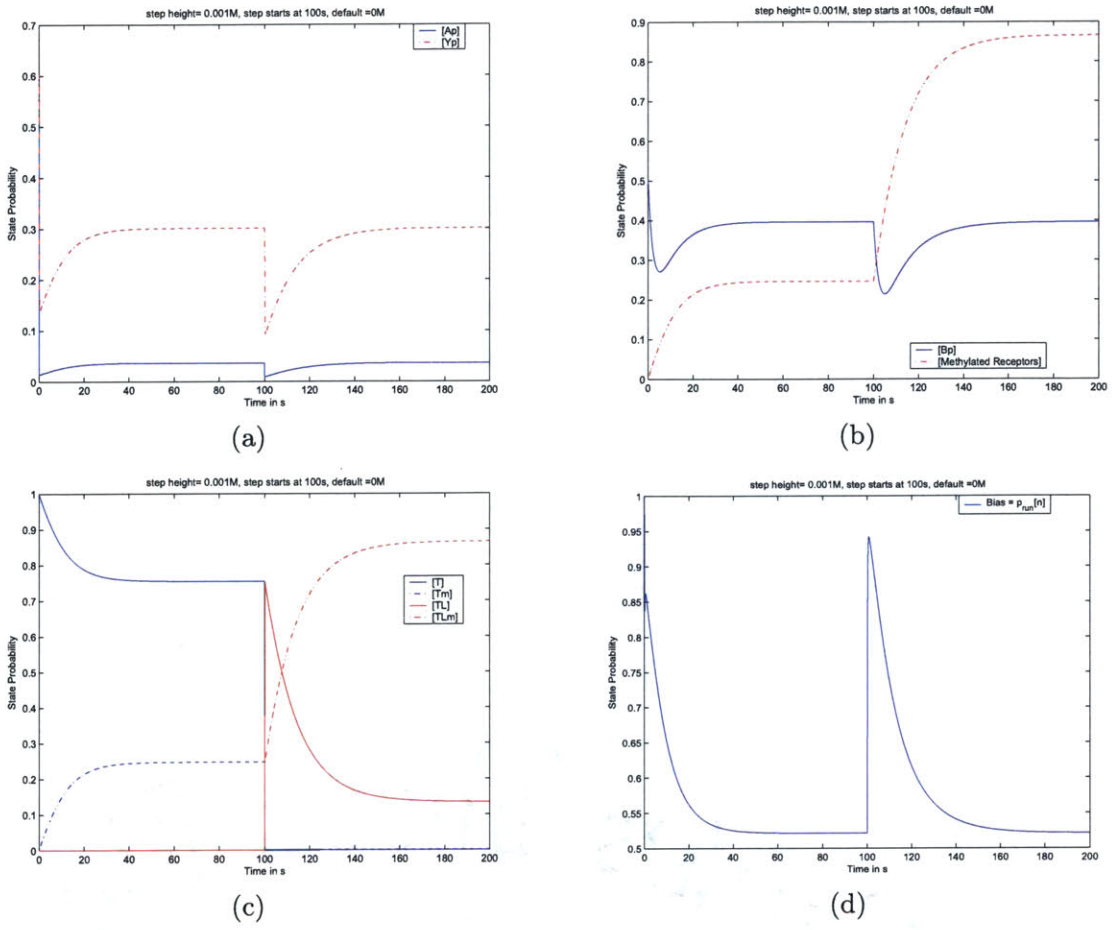


Figure 10-1. State probabilities time evolution for a step input stimulus at 100s from 0 to 10^{-3}M. (a) [Ap] and [Yp], (b) [Bp] and methylated receptors, (c) receptor states, (d) Motor bias.

Tumbles

A simple model is used to encode the effect of a tumble: for each tumble, the angle of the cell rotates counterclockwise or clockwise with equal probability. Note that the cell rotation is distinct from the flagellar rotation which is always in the clockwise direction for tumbles. The direction of rotation is constant throughout the tumble and a constant angular velocity (w_r) is used to determine the total change in direction due to the tumble. Specifically, the equation governing the angular direction in rad/s during a tumble is as follows:

$$\theta[n+1] = \theta[n] + w_r \times dt \tag{10.1}$$

where dt is the sampling period in seconds and $\theta[n]$ is the angular direction at time n.

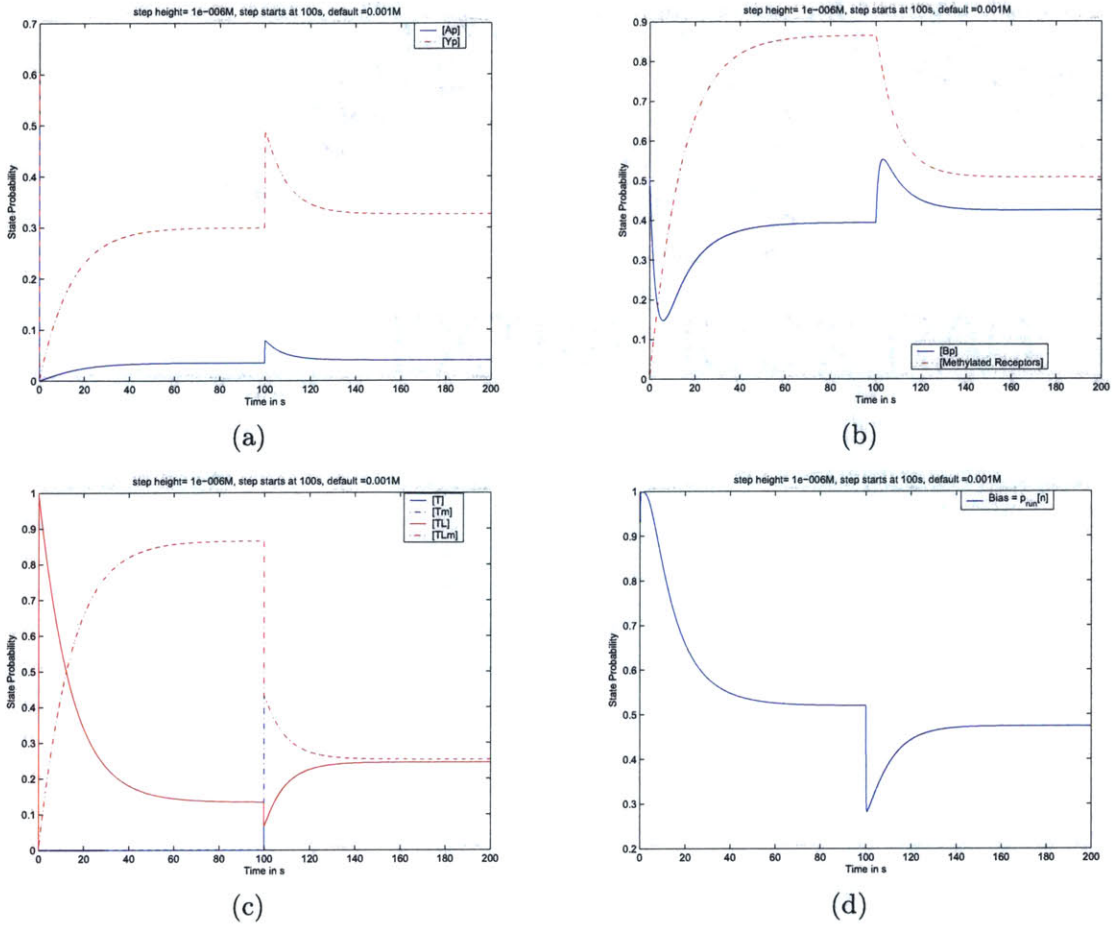


Figure 10-2. State probabilities time evolution for a step input stimulus at 100s from 10^{-3}M to 10^{-6}M. (a) [Ap] and [Yp], (b) [Bp] and methylated receptors, (c) receptor states, (d) Motor bias.

Runs

The run speed of a bacterium is constant and given by v. As a result, the updated position of the bacterium during a run is given by:

$$x[n+1] = x[n] + v \times dt \times cos(\theta[n]) \tag{10.2}$$

$$y[n+1] = y[n] + v \times dt \times sin(\theta[n]) \tag{10.3}$$

Motor behavior: single motor case

In one scheme the motor state is updated every dt, which means the cell could potentially switch between a run and a tumble every dt in the single flagellum per cell case.

An alternative scheme effectively simulates a two-state motor (run and tumble) whose transition probabilities are based on the 8 state model. That is once a motor is in run mode, the probability of exiting the run state is:

$$\frac{p_{56}P(M_n = 5)}{P(M_n = 5) + P(M_n = 4) + P(M_n = 3) + P(M_n = 2) + P(M_n = 1)} \tag{10.4}$$

where p_{56} represents the transition probability from state 5 to 6 at time index n, and M_n is the state of the motor at time n. This scheme amounts to finding the probability of being in the 'edge state', i.e. the state that is one transition away from changing the behavior of the motor, and multiplying that by the transition probability from that edge state to a tumble state. The expression is similar for the probability of transitioning from a tumble to a run.

Another scheme is to only check on the motor state every $dt' = L \times dt$ where L is an integer greater than 1. In this scenario, L should be large enough so that the effect due to a change in input ligand concentration propagates to the CheY or motor chains. One can then ignore all the states the motor is in between dt' checks or alternatively take an average of the motor states for the past dt'.

Multiple motors/flagella

There are about 6 to 10 flagella/motors per bacterium [94]. As a result, we can use multiple motors in our simulation such that a single motor cannot determine on its own whether the cell is running or tumbling. Motors are simulated independently and overall behavior is determined using a voting scheme. Specifically, if half the flagella or more are in the run state, the cell runs, otherwise it tumbles. We chose an odd number of tails so that there is always a clear winner when comparing the number of motors in the run and tumble states. An alternative scheme is to average the states of the motors, and compare that average to a threshold between the edge run and tumble states to decide whether the cell runs or tumbles. This scheme suggests that a motor can be 'tumbling more' if it has more CheYp bound than another motor that is also in a tumble state, and similarly for motors in a run state.

Algorithm flow

The algorithm is initialized with the cell position, direction, internal states, and motor state (run or tumble) at time 0. The flow of the algorithm is then as follows, for every dt:

1. Find current ligand concentration using current position.
2. Update the motor state using the current concentration and internal states.
3. Update the internal states and probabilities using current concentration and internal states.
4. Update position using current position and motor state.
5. go to 1 until reaching the simulation end time.

■ 10.2.3 One-dimensional BASM

The one dimensional implementation of the algorithm has exactly the same formulation as the two-dimensional case except that the cell moves in the positive x direction when the cosine of the angle is positive, and in the negative x direction when the cosine is negative. The speed of the run is constant and is independent of the angle.

■ 10.3 One-Dimensional Simulations

■ 10.3.1 Unimodal test function

We performed simulations of BASM on a one-dimensional function. Specifically, the test function was:

$$f(x) = 10^{-4} e^{-|\frac{3}{4}x|} \tag{10.5}$$

and is plotted in figure 10-3. Each simulation was run for 1000 seconds where the initial

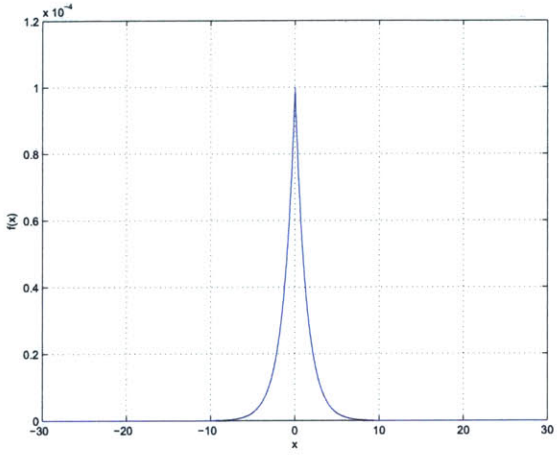
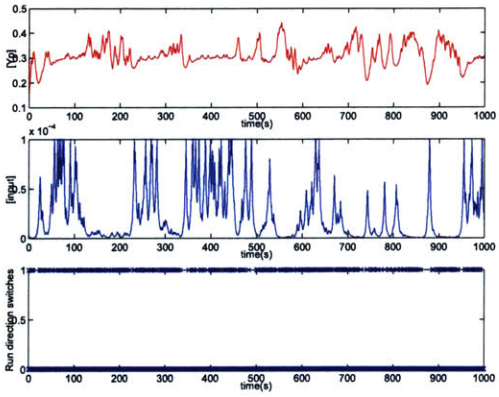


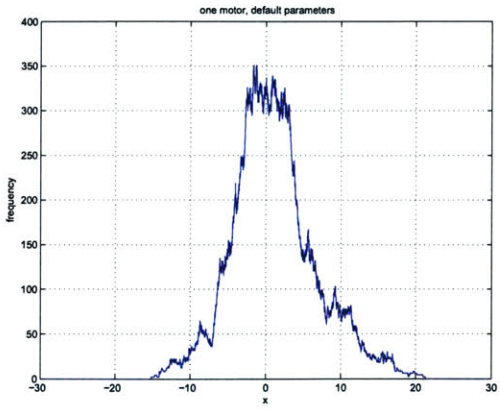
Figure 10-3. One-dimensional test function.

starting point for the cell was drawn from a uniform distribution between $x = -10$ and $x = 10$ and was equally likely to start in the $+x$ or $-x$ direction. We performed 10 to 40 such simulations for each condition described below. The default values for the run speed, v, and angular velocity, w_r, were 0.75 units/s and π rad/s respectively. Density functions displaying how often the bacterium visited certain parts of the x-axis (the search space) were computed as well as time plots for the internal variable $[Yp]$, the input function that the bacterium sees as it is traveling, and the points where the direction of running switches between the positive and negative x directions. This last measurement gives an indication of how long tumbles last.

Figure 10-4 shows the results for the simulation using a one motor configuration. Figure 10-5 shows the results for the simulation using nine motors as described earlier. Figure 10-6 shows the results for the simulation using one motor but where now the angle of rotation is only in the counterclockwise direction for a tumble. This change will effectively cause the run direction to change only after the cell tumbles for a long enough time. Figure 10-7 shows the results for the simulation using nine motors but with the angle of rotation only in the counterclockwise direction for a tumble (as in the previous figure), and with a factor of two increase in the backward probabilities of the motor Markov chain (i.e. the probability that lead to the unbinding of CheYp). This last change makes the backward transition probabilities more comparable to the transition probabilities in the forward direction, effectively leading to shorter tumbles and longer runs. Note the oscillations in the time plots of CheYp. The density plot which is flat near zero also suggests that changes in directions rarely occur near the maximum of the function. In order to check which set of conditions is driving these oscillations, we performed each combination of two out of the three conditions and checked the behavior. The closest behavior to the one seen in the previous figure is the

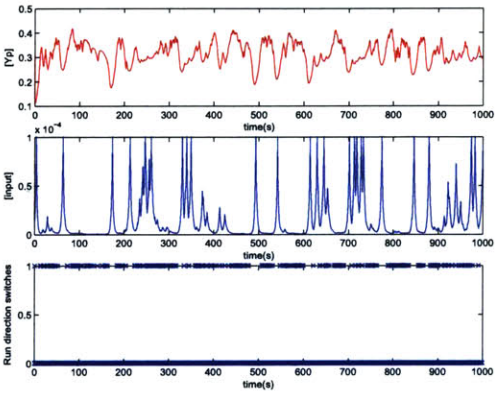


(a)

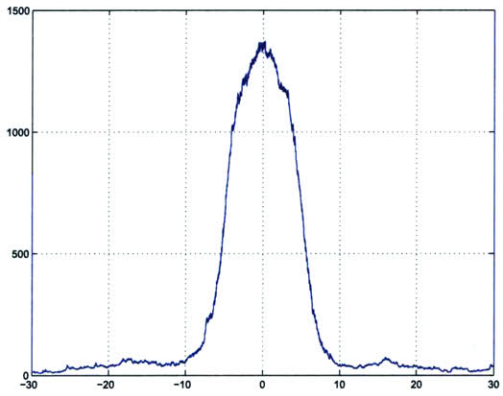


(b)

Figure 10-4. Simulations using the default parameters and a one motor configuration. (a) individual bacterium variables. (b) Average density plot using 10 simulations.



(a)



(b)

Figure 10-5. Simulations using the default parameters and a nine motor configuration. (a) individual bacterium variables. (b) Average density plot using 40 simulations.

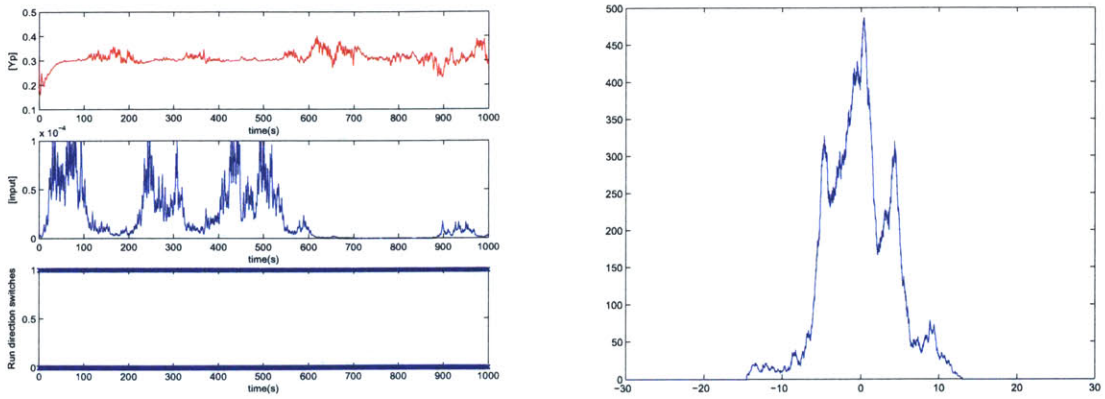
modification that makes the angle always rotate counterclockwise during a tumble together with the backward rate constants multiplied by 2. The results from this last modification are shown in Figure 10-8.

■ 10.3.2 Multimodal test function

We also performed one-dimensional BASM simulations on a multimodal function. Specifically, the test function was:

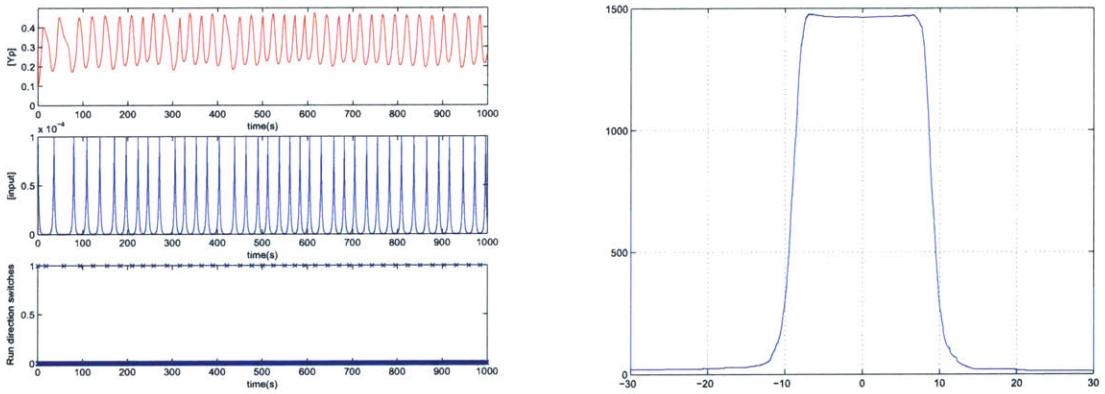
$$f(x) = 5 \times 10^{-5} \left| \frac{\sin(0.25x)}{0.25x} \right| \tag{10.6}$$

which is shown in Figure 10-9. 10 to 40 simulations were performed for different conditions. In each simulation, the bacterium was dropped uniformly between -10 and +10 and each run lasted 1,000 seconds. The results for the nine motor simulations are shown in Figure



(a) (b)

Figure 10-6. Simulations using the rotation in only one direction. (a) individual bacterium variables. (b) Average density plot using 10 simulations.

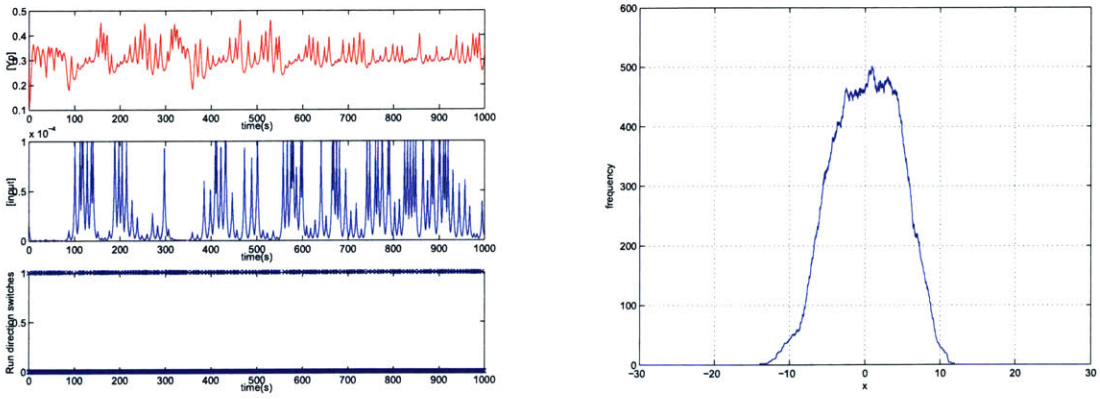


(a) (b)

Figure 10-7. Simulations using the rotation in only one direction, 9 motor chains, and an increase in the motor backward probabilities by a factor of two. (a) individual bacterium variables. (b) Average density plot using 40 simulations.

10-10 and correspond to a run speed of 0.75 units/s and an angular velocity of π rad/s. Figure 10-11 shows the average density of 40 simulations using a reduced run speed of 0.6 units/s while keeping the angular velocity at π rad/s. The angular velocity in Figure 10-12 is increased to 1.25π rad/s while the run speed is further decreased to 0.5 units/s.

The results show that while the bacterium spends most of its time on the main lobe of the test function, it also spends a fair amount of time on the side lobes thereby mapping the different modes of the test function.



(a) (b)

Figure 10-8. Simulations using the rotation in only one direction, and an increase in the motor backward probabilities by a factor of two. (a) individual bacterium variables. (b) Average density plot using 10 simulations.

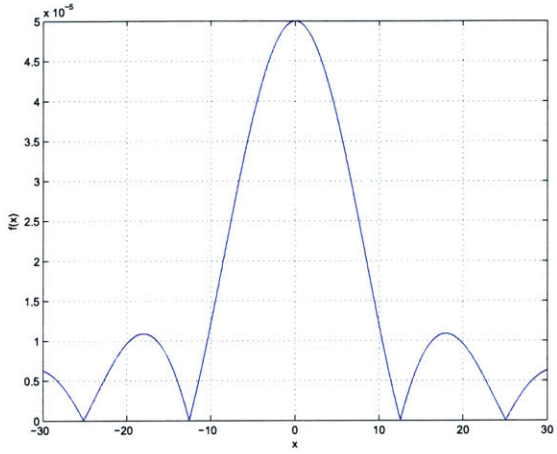


Figure 10-9. One-dimensional multimodal test function.

■ 10.4 Two-Dimensional Simulations

■ 10.4.1 Unimodal test function

We also simulated the two-dimensional version of BASM on different test functions. The test function used in the first set of simulations is given by:

$$f(x,y) = 10^{-4}e^{-0.5\sqrt{x^2+y^2}} \tag{10.7}$$

and is shown in Figure 10-13. In these simulations, the nine motor configuration is always used and the cell is allowed to reach steady state with the initial concentration corresponding to the concentration of the test function at its initial position before the search begins.

Figure 10-14 shows a trace of the bacterium running for 5000 seconds using double the backward probabilities for the motor, a running speed of 0.3 units/s and an angular velocity of 2π rad/s. Figure 10-15 shows a trace and density of a bacterium running for 5000 seconds

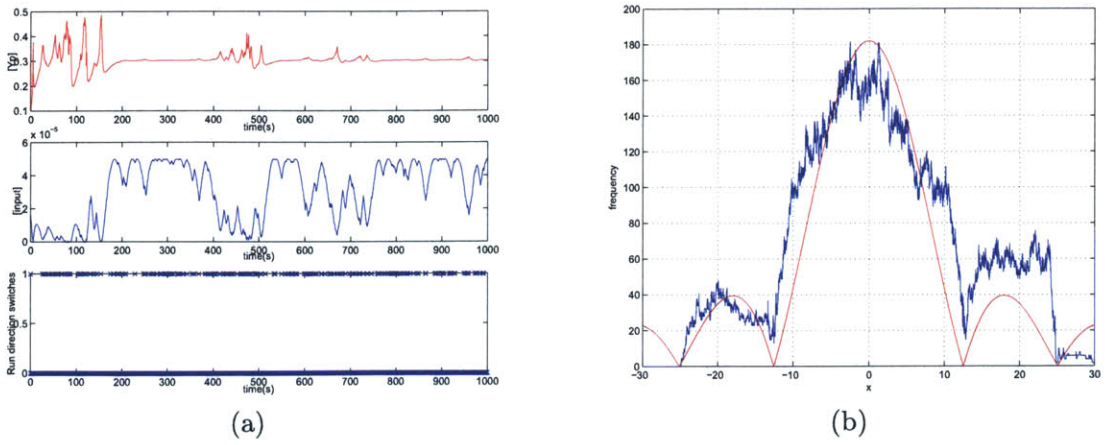


Figure 10-10. Multimodal simulations using the nine motor configuration. The test function (scaled) is shown in solid red. (a) individual bacterium variables. (b) Average density plot using 10 simulations.

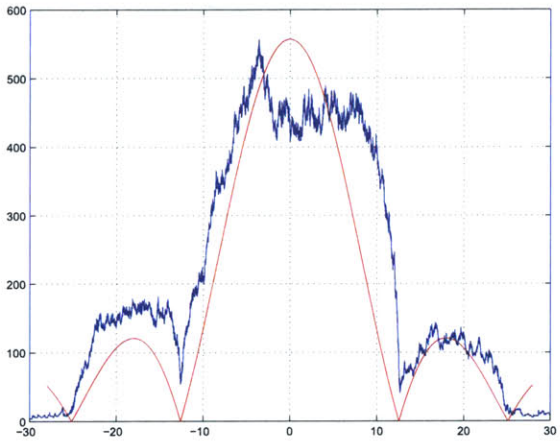


Figure 10-11. Average density plot using 40 multimodal simulations with nine motor configuration. $v = 0.6$ units/s and $w_r = \pi$ rad/s.

dropped uniformly on a radius of 7 around (0,0) and with a uniform angle using double the backward probabilities for the motor, a running speed of 0.3 units/s and an angular velocity of 2π rad/s. The average density plot of ten runs lasting 4000 seconds each and using a running speed of 0.3 units/s and an angular velocity of 10 π rad/s is given in Figure 10-16. The results show that the bacterium spends most of its time on the areas where the magnitude of the function is high.

■ 10.4.2 Multimodal test function

The multimodal test function we used in the second set of simulations is given by:

$$f(x,y) = 5 \times 10^{-5} \frac{\sin(x/4)\sin(y/4)}{(x/4)(y/4)} \tag{10.8}$$

and is shown in Figure 10-17. The average density plot of ten runs lasting 4000 seconds each and using a speed of 0.3 units/s and an angular velocity of 10 π rad/s is given in Figure 10-

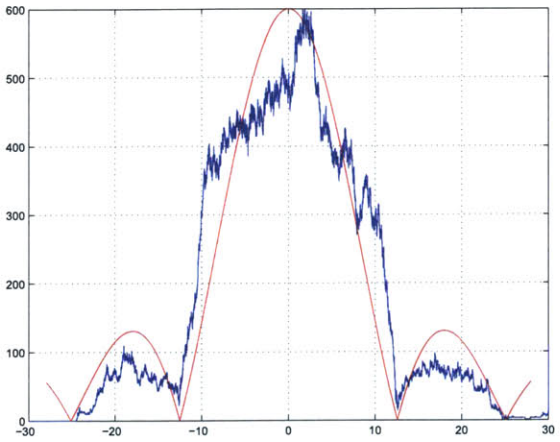


Figure 10-12. Average density plot using 40 multimodal simulations with nine motor configuration. $v = 0.5$ units/s and $w_r = 1.25\pi$ rad/s.

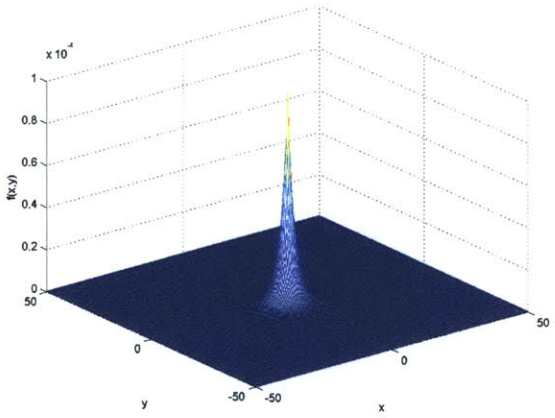


Figure 10-13. Two-dimensional test function.

18. In these simulations, the bacterium was uniformly dropped at the start of each run on a radius of 7 units around (0,0) and with a uniform angle. The double backward probabilities version of the parameters was used along with the nine motor configuration. While the results are not as dramatic as for the other test functions, the bacterium is still able to map the different modes of the function by spending more time on them then elsewhere in the function space.

■ 10.5 Bacterial Algorithm for Surface Flattening (BASF)

The chemotaxis pathway allows bacteria to search their environment for nutrients and move towards higher nutrient concentrations. However, as bacteria move, they eat and therefore change the nutrient concentration in their environment. As a result, these organisms are not only mapping the nutrient surface but effectively flattening it by reducing its value at the place they visit.

One can formulate by analogy a surface flattening algorithm where a surface rewriting step is added to the BASM algorithm presented in the previous sections. We call this

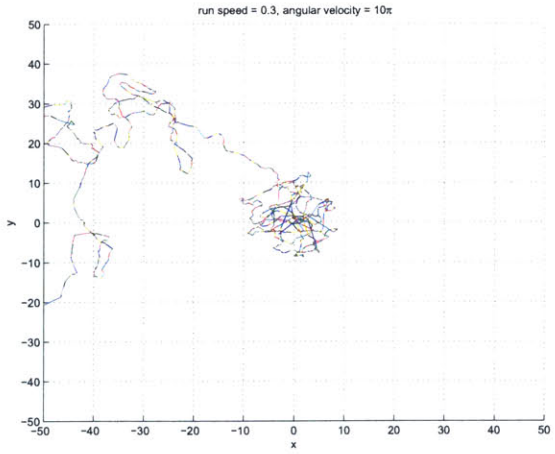
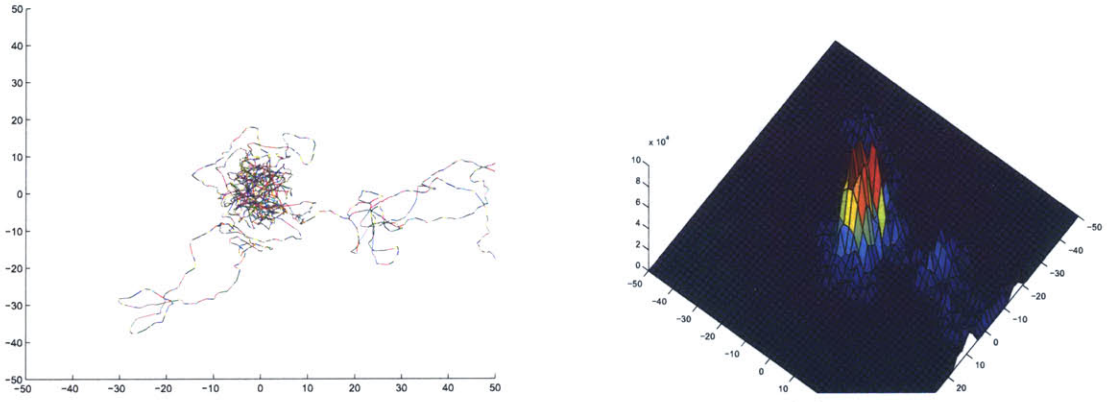


Figure 10-14. Two-dimensional bacterial run.



(a) trajectory (b) density

Figure 10-15. Two-dimensional run using a uniformly dropped bacterium.

algorithm: Bacterial Algorithm for Surface Flattening (BASF). Specifically, every time a point in the surface is read as an input to BASM, it is rewritten by reducing the value of the surface at that point by the flattening factor. In the one-dimensional version of BASF, the surface update is given by:

$$f[x, n+1] = \epsilon f[x, n] \tag{10.9}$$

where $f[x, n]$ is the value of the function at x at time n and $0 < \epsilon < 1$ is the flattening factor.

We implemented a one-dimensional version of BASF using the exponential test function shown in Figure 10-9. The default parameters and nine motor configuration were used for the $a3$MC model implementation. The run speed was 0.75 units/s and the angular velocity was π rad/s. The flattening factor was 0.8. The discretization step, which determines the region of the function being reduced in one step, was vdt. Furthermore, the search space was limited to the closed interval [-90 90], i.e. if the bacterium reached -90 or 90, it would stay at that position until it switched directions. We performed 20 simulations that lasted

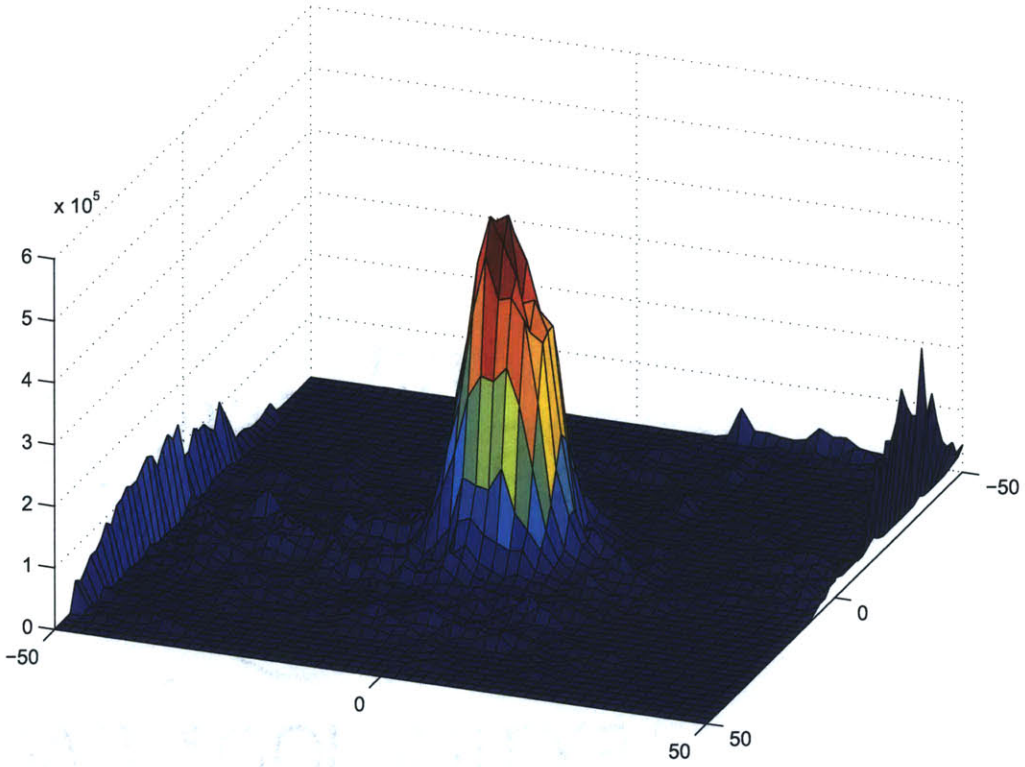


Figure 10-16. Average density plot of 10 uniformly dropped bacteria.

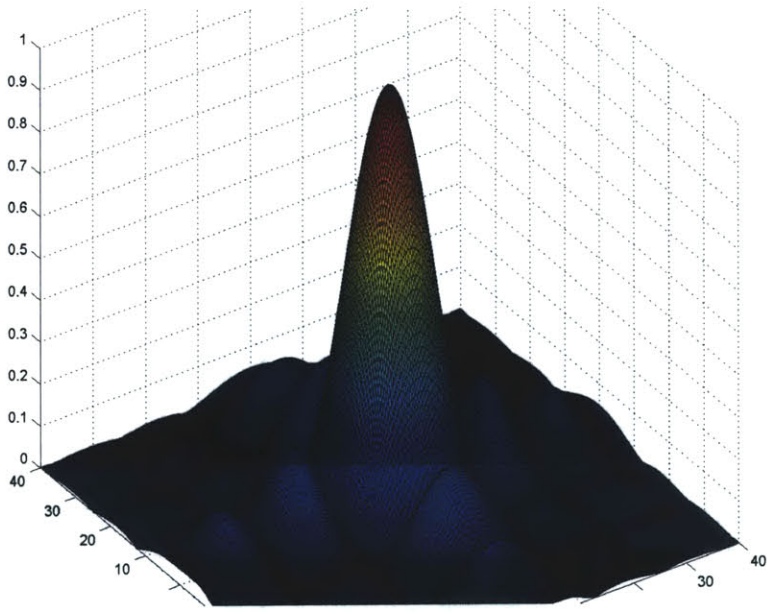


Figure 10-17. 2D multimodal test function.

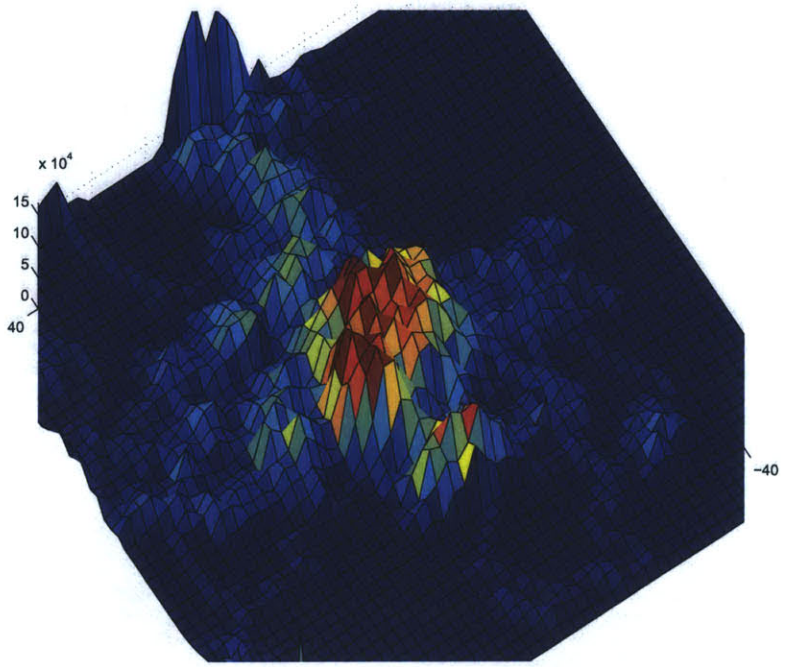


Figure 10-18. Average density of 10 runs using the multimodal test function.

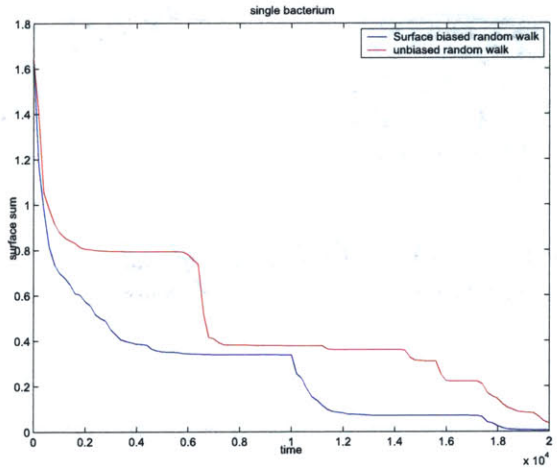


Figure 10-19. Surface flattening using one bacterium.

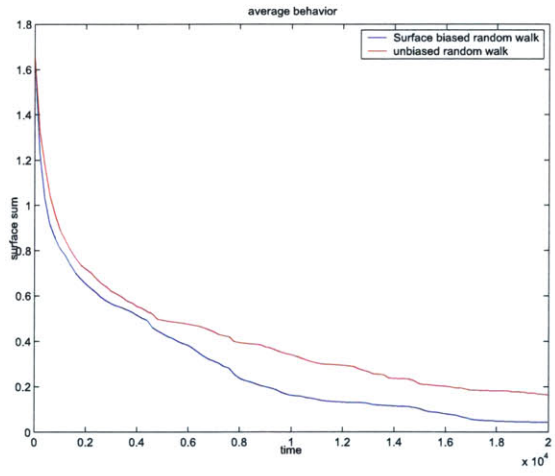


Figure 10-20. Average of 20 independent runs using one bacterium.

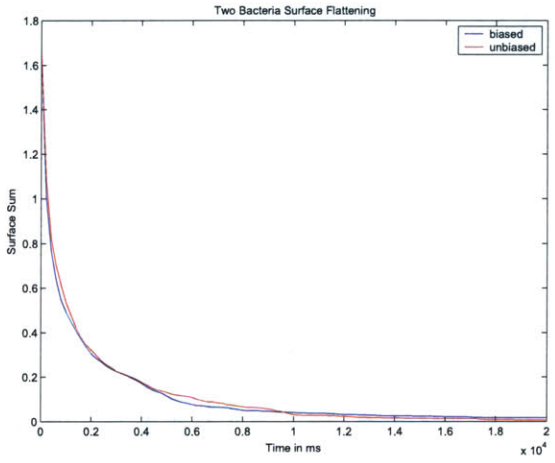


Figure 10-21. Average of two simultaneous bacteria.

20,000 seconds each. For each simulation, a bacterium was uniformally dropped in the interval [-10 10] and the total area under the surface was recorded as a function of time.

Figure 10-19 shows the total surface sum as a function of time for one bacterium as well as results obtained using an unbiased random walk, i.e. a bacterium that does not use surface information to guide its motion, that is the input to the $a3$MC model is always zero for that bacterium. The average of twenty runs is plotted in Figure 10-20. The results suggest that the BASF algorithm lead to faster surface flattening than the unbiased random walk.

We have also implemented a two-bacteria version of the algorithm where two bacteria are simultaneously dropped and evolve according to the BASF algorithm. The average of 16 runs using two simultaneous bacteria is shown in Figure 10-21 along with the average obtained from 10 runs of the unbiased random walk. As expected, using two bacteria leads to a faster flattening of the surface than using one bacterium. However Figure 10-21 indicates that the difference between the unbiased random walk and BASF is not as pronounced as for the one bacterium case. This may be due to the limited size of the flattening space used.

■ 10.6 Summary

In this chapter, we have started to explore the potential of using our understanding of cellular signaling to formulate new signal processing algorithms by developing algorithms for surface mapping and flattening based on bacterial chemotaxis. While the results presented here are still preliminary, there would seem to be potential advantages of such algorithms, namely simplicity and parallelism. These algorithms can be implemented on a collection of cheap, dispensable sensors that can be deployed in a field where a surface of interest (such as the chemical concentration of a given agent) needs to be mapped and potentially neutralized (flattened). Obviously, in order to determine the power of these algorithms, one needs to benchmark the performance of these algorithms against other common algorithms such as those presented in the first section of this chapter. This type of analysis is the subject of our future research.

Chapter 11

Conclusions and Contributions

In this thesis, we have developed new frameworks directed towards understanding the information processing in biological cells at different levels of abstraction. We have also presented a preliminary example as to how the results might be exploited to develop a new generation of algorithms for engineered distributed networks.

■ 11.1 Graph Theoretic Modeling

At the highest modeling level, the focus was on the network topology rather than on the dynamical properties of the components of the signaling network. In this regime, we focused on interconnectivity of nodes and introduced concepts from random graph theory to examine and analyze the distribution and properties of the network graph. Among the contributions of this thesis is the first graph theoretic model of the whole-genome relationship between cell genotype (genomic content) and phenotype (pathophysiological behavior) in response to toxic agents (chemicals and radiation) in the environment, for yeast as the currently most genomically-complex available experimental system. This model enabled exploration of how a biomolecular network processing input stimuli leading to output behavior in living cells operates in terms of network topology properties. Specifically, we have applied protein-protein interaction network analysis to the global genotype/phenotype data-set recently developed for the recovery of *S. cerevisiae* from exposure to DNA damaging agents. The data was analyzed in the context of the full yeast interactome and in newly defined network structures. On average, essential and damage-recovery proteins displayed greater direct interactions, smaller shortest-path-length characteristics, increased connectedness and higher local clustering than non-essential and no-phenotype proteins (i.e. proteins not required for recovery from DNA damage) respectively. We have also shown that other functional yeast networks do not necessarily share similar quantitative features. These results suggest that cells initiate highly coordinated responses to damage. With mapped genotype/phenotype information, we have further identified toxicologically-important protein complexes, pathways, and modules. In addition, we have presented a method for combining expression profiling data with network information to enhance the probability of predicting phenotypic behavior. And through partitioning analysis, we have identified specific interactions that could be tested for essentiality.

■ 11.2 Multiresolution Modeling

Another contribution of the thesis is the development of a new model referred to as Markov modulated Markov chains, for examining the dynamics of cellular signal processing based on interacting Markov chains. This model represents a unified framework for simultaneously studying the fluctuations of signaling pathways (stochastic behavior) and computing their average behavior therefore allowing modeling at multiple resolutions within the same framework. The use of this framework was demonstrated on two classical signaling networks: the Mitogen Activated Protein Kinase (MAPK) cascade and the signaling pathway underlying bacterial chemotaxis.

■ 11.3 Biologically Inspired Surface Mapping and Flattening Algorithm

In the thesis, we have started exploring the potential of using our understanding of cellular signaling to formulate novel signal processing algorithms by developing a surface mapping and flattening algorithm based on bacterial chemotaxis. This algorithm can be implemented

on a collection of cheap, dispensable sensors that can be deployed in a field where a surface of interest (such as the chemical concentration of a given agent) needs to be mapped and potentially neutralized (flattened).

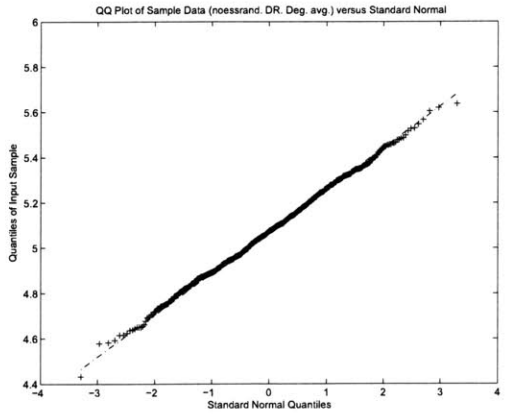
Appendix A

Randomizations

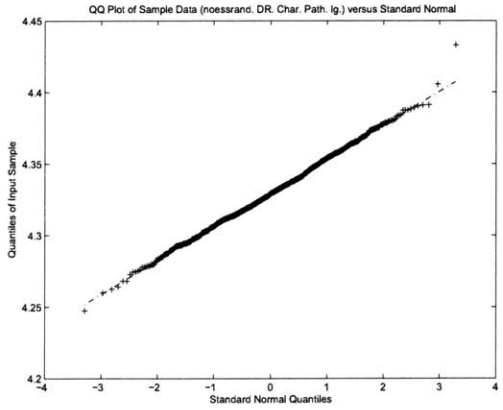
	Average degree	Char. Path length	LC size	C_{avg}	C> 0	Isolated Nodes
Essential	1	1	1	1	1	1
Metabolic	1	151	502	81	340	653
Non-Essential	1	1	21	1	1	4
No-Phenotype	1	1	46	5	1	24
Metabolic (N-E)	134	588	427	4	7	190
Damage-Recovery	1	1	4	1	1	7
MMS	1	1	1	1	1	1
4NQO	1	1	1	1	1	1
UV	3	5	7	1	1	1
t-BuOOH	39	247	102	1	1	1

Table A.1. Rank of each statistic in the tested network with respect to the values obtained in the 1,000 randomized sets.

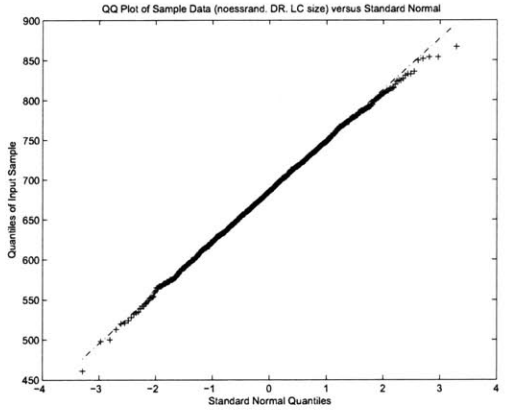
The quantiles of each randomized data set are plotted versus the quantiles of a standard Normal distribution in Figure A-1. The purpose of this kind of plot is to determine whether the sample is drawn from a Normal (i.e. Gaussian) distribution. If the sample comes from a Normal distribution, even if one distribution is shifted and re-scaled from the other, the plot will be linear.



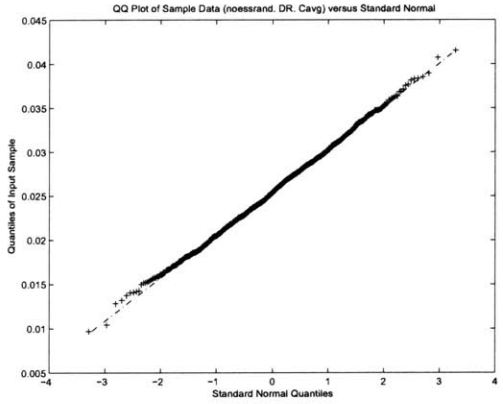
(a) Average degree



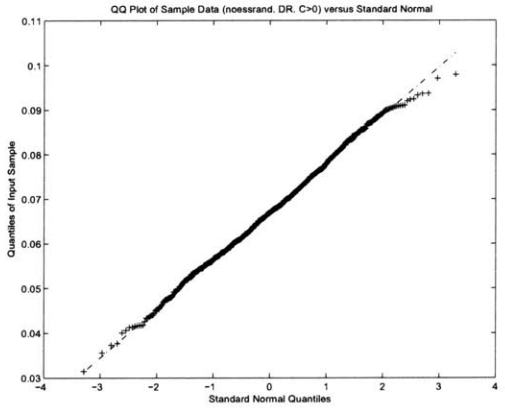
(b) Characteristic path length



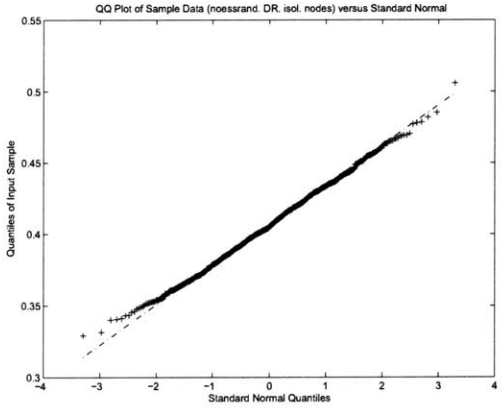
(c) LC size



(d) Average clustering coefficient



(e) $C > 0$



(f) Isolated nodes

Figure A-1. Representative Normal quantile-quantile (q-q) plots for different measures of the randomized networks corresponding to the damage-recovery category.

Bibliography

[1] A.V. Oppenheim, R.W. Schafer with John R. Buck. *Discrete-Time Signal Processing.* Prentice Hall, 1999.

[2] T. Achacoso and W. Yamamoto. *Neuroanaotomy of C. elegans for Computation* (Boca Raton, FL, CRC Press), 1992.

[3] H. Agrawal. Extreme Self-Organization in Networks Constructed from Gene Expression Data. *Physical Review Letters*, Vol. 89, No. 26, 2002.

[4] R. Albert, H. Jeong, and A. L. Barabasi. "Error and attack tolerance of complex networks." *Nature* 406, 378-382, 2000.

[5] U. Alon, M.G. Surette, N. Barkai, and S. Leibler, "Robustness in bacterial chemotaxis" *Nature* 397:168-171, January 1999.

[6] R. Alves, R.A.G. Chaleil, and M.J.E. Sternberg " Evolution of Enzymes in Metabolism: A Network Perspective" *J. Mol. Biol.*, 320:751-770, 2002.

[7] J.C. Ameisen " On the Origin, Evolution, and Nature of Programmed Cell Death: a Timeline of Four Billion Years" *Cell Death and Differentiation*, 9:367-393, 2002.

[8] R.W. Anderson. "Biased random-walk learning: a neurobiological correlate to trial-and-error" in *Neural networks and pattern recognition* O.M. Omidvar abd J. Dayhoff, Eds. New York: Academic, 1998, pp. 221-244. 1998.

[9] C. Asavathiratham, S. Roy, B. Lesieutre and G. Verghese, "The Influence Model." *IEEE Control Systems Magazine*, Vol.21, No. 6. pp. 52-64, 2001.

[10] A.R. Asthagiri and D.A. Lauffenburger, "A computational Study of Feedback Effects on Signal Dynamics in a Mitogen-Activated Protein Kinase (MAPK) Pathway model", *Biotechnol. Prog*, 17:227-239, 2001.

[11] G.D. Bader, D. Betel, C.W. Hogue, "BIND: the Biomolecular Interaction Network Database", *Nucleic Acids Res.*, 31(1):248-250, 2003.

[12] C.P. Bagowski and J.E. Ferrell Jr, "Bistability in the JNK cascade", *Current Biology*, 11:1176-1182, 2001.

[13] A. Bahn, D.J. Galas, and T. Dewey " A Duplication Growth Model of Gene Expression Networks" *Bioinformatics*, Vol. 18 no. 11:1486-1493, 2002.

[14] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286: 509–512, 1999.

[15] N. Barkai and S. Leibler. "Robustness in simple biochemical networks". *Nature* 387, 913-917, 1997.

[16] R.L. Barron. "Self-organizing and learning control systems", in *Cybernetic Problems in Bionics-Bionics Symposium, Dayton, May 1966.* New York:Gordon and Breach, pp. 147-203, 1968.

[17] R.L. Barron. "Neuromine nets as the basis for predictive component of robot brains," in *Cybernetics, Artificial Intelligence, and Ecology - Fourth Annual Symposium American Society of Cybernetics*, H.W. Robinson and D.E. Knight, Eds. Washington, DC:Spartan, pp.159-193, 1972.

[18] T.J. Begley, A.S. Rosenbach, T. Ideker, and L.D. Samson. "Damage Recovery Pathways in *Saccharomyces cerevisiae* Revealed by Genomic Phenotyping and Interactome Mapping", *Molecular Cancer Research* December, 1:103-112, 2002.

[19] T.J. Begley, A.S. Rosenbach, T. Ideker, and L.D. Samson. "Hot spots for modulating toxicity identified by genomic phenotyping and localization mapping", *Mol Cell.*, 1:117-25, October 2004.

[20] BIND: http://www.bind.ca/

[21] BIOCARTA: http://www.biocarta.com

[22] S. Blatt, "Distributed Sensor Fusion for Sensor Networks," *Proc. Fourth Annual Conf. Information Fusion*, pp. TuC3-8, Aug 2001.

[23] B. Bollobás *Modern Graph Theory.* Springer, New York, 1998.

[24] K.A. Borkovich, N. Kaplan, J.F. Hess, M.I. Simon. "Transmembrane signal transduction in bacterial chemotaxis involves ligand-dependent activation of phosphate group transfer" *Proc. Natl. Acad. Sci. USA*, Vol. 86, pp. 1208-1212, February 1989.

[25] A. Boyd, M.I. Simon. "Multiple electrophoretic forms of methyl-accepting chemotaxis proteins generated by stimulus-elicited methylation in Escherichia coli." *J. Bacteriol.* Vol. 143, pp.809-815. August 1980.

[26] D. Bray, R.B. Bourret, M.I. Simon. "Computer simulation of the phosphorylation cascade controlling bacterial chemotaxis" *Molecular Biology of the Cell*, Vol. 4, 469-482, May 1993.

[27] D. Bray, S. Lay. "Computer simulated evolution of a network of cell-signaling molecules" *Biophysical Journal*, Vol. 66, pp. 972-977, April 1994.

[28] D. Bray. "Signaling complexes: Biophysical Constraints on Intracellular Communication" *Annu. Rev. Biophys. Biomol. Struct.*', 27:59-75, 1998.

[29] H.J. Bremermann. "Chemotaxis and optimization" *J. Franklin Inst.*, vol. 297, pp.397-404, 1974.

[30] H.J. Bremermann and R.W. Anderson. "How the brain adjusts synapses-maybe" in *Automated Reasoning: Essays in Honor of Woody Bledsoe* R.S.Boyer, Ed. Norwell, MA:Kluwer, pp.119-147, 1991.

[31] T. Brody " The Interactive Fly. Cell Death Regulation in Drosophila: Conservation of Mechanisms and Unique Insights" *http://sdb.bio.purduc.cdu/fly/aignfam/apoptosis.htm*, 2002.

[32] J. Chen, K. Yao, and R.E. Hudson. "Source Localization and Beamforming," *IEEE Signal Processing Magazine*, V. 19 No. 2, pp. 30-39, March 2002.

[33] S. Chen and B.L. Luk "Adaptive simulated annealing for optimization in signal processing applications" *EURASIP Signal Processing Journal 79*(1):pp. 117-128. 1999.

[34] S.A. Chervitz and J.J. Falke. "Molecular mechanism of transmembrane signaling by the aspartate receptor: A model" *Proc. Natl. Acad. Sci. USA*, Vol. 93, pp. 2545-2550, March 1996.

[35] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein. *et al.* "The transcriptional program of sporulation in budding yeast." *Science* 282, 699-705, 1998.

[36] D. Corne, M. Dorigo, and F. Glover *New Ideas in Optimization*. McGraw-Hill, England. 1999.

[37] http://dip.doe-mbi.ucla.edu/

[38] C.M. Deane, L. Salwinski, I. Xenarios, D. Eisenberg. *Molecular and Cellular Proteomics* 1.5. 349-356, 2002.

[39] J.L. Devore. *Probability and Satatistics*, 6 edn (Belmont, CA, Brooks/Cole-Thomason Learning), 2004.

[40] K. Dolinski, R. Balakrishnan, K.R. Christie, M.C. Costanzo, S.S. Dwight, S.R. Engel, D.G. Fisk, J.E. Hirschman, E.L. Hong, L. Issel-Tarver, *et al.*. "Saccharomyces Genome Database". *http://www.yeastgenome.org/*, 2004.

[41] S.N. Dorogovtsev and J.F.F. Mendes *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, New York, 2003.

[42] A.H. Enyenihi and W.S. Saunders. "Large-Scale functional genomic analysis of sporulation and meiosis in *Saccharomyces cerevisiae*". *Genetics* 163, 47-54, 2003.

[43] P. Erdos and A. Rényi "On Random Graphs", *Publ. Math. Debrecen*, 6:290, 1959.

[44] P. Erdos and A. Rényi "On the Evolution of Random Graphs", *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17, 1960.

[45] I. Farkas, I. Derényi, A. L. Barabási, and T. Vicsek. Spectra of 'real-world' graphs: Beyond the semicircle law. *Physical Review E*, 64, 2001.

[46] I. Farkas, H. Jeong, T. Vicsek, A.-L. Barabási, and Z.N. Oltvai, "The Topology of the Transcription Regulatory Network of the Yeast, *Saccharomyces cerevisiae*", *Physica A*, 318:601-612, 2003.

[47] J.E. Ferrell Jr. "Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability," *Current Opinion in Cell Biology*, Volume:14 2, pp140-148, April 2002.

[48] J. Forster, I. Famili, P. Fu, B. Palsson, and J. Nielsen. "Genome-Scale Reconstruction of the *Saccharomyces cerevisiae* Metabolic Network." *Gen.Res.* 13, 244-253, 2003.

[49] L.C. Freeman " A set of Measures of Centrality Based on Betweenness" *Sociometry*, 40:35, 1977.

[50] M. Fromont-Racine, J.C. Rain, and P. Legrain. "Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens." *Nat Genet* 16, 277-282, 1997.

[51] C. Furusawa, K. Kaneko. Zipf's Law in Gene Expression. *Physical Review Letters*, Vol. 90, No. 8, 2003.

[52] R.G. Gallager, *Discrete Stochastic Processes*. Kluwer Academic Publishers 1996.

[53] A.P. Gash, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein, and P.O. Brown. "Genomic expression programs in the response of yeast cells to environmental changes". *Molecular Biology of the Cell.* 11, 4241-4257, 2000.

[54] A-C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A-M. Michon, C-M. Cruciat, *et al. Nature* 415. 141-147, 2002.

[55] D.T. Gillespie, "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions", *Journal of Computational Physics*, 22:403-434, 1976.

[56] D.T. Gillespie, "Exact Stochastic Simulation of Coupled Chemical Reactions", *Journal of Physical Chemistry*, Vol 81 No 25:2340-2361, 1977.

[57] D.T. Gillespie, "A Rigorous Derivation of the Chemical Master Equation", *Phsycia A*, 188:404-425, 1992.

[58] L. Giot, J.S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y.L. Hao, C.E. Ooi, B. Godwin, E. Vitols, *et al.* "A protein interaction map of *Drosophila melanogaster.*" *Science* 302, 1727-1736, 2003.

[59] K.-I. Goh, B. Kahng, and D. Kim. Spectra and eigenvectors of scale-free networks. *Physical Review E*, 64, 2001.

[60] S. Goss, S. Aron, J.L. Deneubourg, J.M. Pasteels. "Self-organized shortcuts in the Argentine ant" *Naturwissenschaften* 76, pp. 579-581. 1989.

[61] P.J.E. Goss and J. Peccoud. "Quantitative Modeling of Stochastic Systems in Molecular Biology by Using Stochastic Petri Nets", *Proc. Natl. Acad. Sci. USA*, 95:6750-6755, 1998.

[62] N. Guelzim, S. Bottani, P. Bourgine, and F. Képès. "Topological and Causal Structure of the Yeast Transcriptional Regulatory Network", *Nature Genetics*, 31:60-63, May 2002.

[63] J.D. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L.V. Zhang, D. Dupuy, A.J. Walhout, M.E. Cusick, F.P. Roth, M. Vidal. "Evidence for dynamically organized modularity in the yeast protein-protein interaction network." *Nature* 430(6995):88-93, 2004.

[64] D. Harel. *Algorithmics: The Spirit of Computing* Addison-Wesley. 1987.

[65] Y. Ho, A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, *et al.* "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry." *Nature* 415, 180-183, 2002.

[66] R. Horst and R.M. Pardalos. *Handbook of Global Optimization* Kluwer Academic, Dordrecht, Netherlands, 1995.

[67] C.F. Huang and J.E. Ferrell Jr, "Ultrasensitivity in the mitogen-activated protein kinase cascade." *Proc. Natl. Acad. Sci.* 93:10078-10083, 1996.

[68] T. Ideker and D. Lauffenburger "Building with a scaffold: emerging strategies for high- to low-level cellular modeling", *TRENDS in Biotechnology*, Vol.21 No.6:255-262, June 2003.

[69] L. Ingber and B. Rosen, "Genetic algorithms and very fast simulated re-annealing". *Mathematical Computer Modeling*, 16(11):87-100, 1992.

[70] L. Ingber. "Adaptive simulated annealing (ASA): Lessons learned" *J. Control and Cybernetics* Vol. 25 pp33-54, 1996.

[71] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. "A comprehensive two-hybrid analysis to explore the yeast protein interactome." *Proc Natl Acad Sci U S A* 98, 4569-4574, 2001.

[72] S.A. Jelinsky, P. Estep, G.M. Church, and L.D. Samson. "Regulatory networks revealed by transcriptional profiling of damaged saccharomyces cerevisiae cells: rpn4 links base excision repair with proteasomes." *Mol Cell Biol* 20, 8157-8167, 2000.

[73] S.A. Jelinsky and L.D. Samson. "Global response of *Saccharomyces cerevisiae* to an alkylating agent." *Proc Natl Acad Sci U S A* 96, 1486-1491, 1999.

[74] H. Jeong, S.P. Mason, A.-L. Barabási, and Z.N. Oltvai, "Lethality and Centrality in Protein Networks", *Nature*, 411:41-42, May 2001.

[75] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabasi. "The large-scale organization of metabolic networks." *Nature* 407, 651-654, 2000.

[76] G. Karypis and V. Kumar. "A Fast And High Quality Multilevel Scheme for Partitioning Irregular Graphs," *SIAM J. Sci. Comput.* V. 20, No. 1, pp. 359-392. 1998.

[77] S. Kirkpatrick, C.D.Jr. Gerlatt, and M.P. Vecchi. "Optimization by Simulated Annealing", *Science* 220, 671-680, 1983.

[78] E. Korobkova, T. Emonet, J.M.G. Vilar, T.S. Shimizu, and P. Cluzel, "From molecular noise to behavioural variability in a single bacterium," *Nature* Vol. 428, pp. 574-578. April 2004.

[79] S. Kumar, F. Zhao, and D. Shepherd. "Collaborative Signal and Information Processing in Microsensor Networks," *IEEE Signal Processing Magazine*, V. 19 No. 2, pp. 13-14, March 2002.

[80] T.G. Kurtz, *J. Chem. Phys.* 57, 2976, 1972.

[81] D.A. Lauffenburger. "Cell signaling pathways as control modules: complexity for simplicity?" *Proc. Nat. Aca. Sci.* 97. 5031-5033, 2000.

[82] E.J. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, D.B. Shmoys. *The traveling salesman problem: a guided tour of combinatorial optimization.* Wiley, Chichester. 1985.

[83] L.Lee and A.V.Oppenheim, "Distributed Signal Processing", *Proc. ICASSP'98.* Seattle, WA, May 1998.

[84] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, *et al.* "Transcriptional regulatory networks in *Saccharomyces cerevisiae.*" *Science* 298, 799-804, 2002.

[85] S. Li, C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.O. Vidalain, J.D. Han, A. Chesneau, T. Hao, *et al.* "A map of the interactome network of the metazoan C. elegans." *Science* 303, 540-543, 2004.

[86] D. Li, K. Wong, Y. Hu, and A. Sayeed. "Detection, Classification, and Tracking of Targets," *IEEE Signal Processing Magazine*, V. 19 No. 2, pp. 17 - 29, March 2002.

[87] V. Maniezzo, L.M. Gambardella, and F. De Luigi. "Ant Colony Optimizations" *New Optimization Techniques in Engineering* by Onwubolu, G.C., and Babu, B.V., Springer-Verlag Berlin Heidelberg, 101-117, 2004.

[88] S. Maslov, and K. Sneppen. "Protein interaction networks beyond artifacts." *FEBS Lett* 530, 255-256, 2002.

[89] S. Maslov and K. Sneppen, "Specificity and Stability in Topology of Protein Networks", *Science*, 296:910-913, May 2002.

[90] D.A. McQuarrie, *J. Appl. Probability*, 4, 413, 1967.

[91] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller "Equations of State Calculations by Fast Computing Machines" *J. Chem. Phys.* 21, 1087-1092, 1958.

[92] L. Michaelis and M.I. Menten "Die Kinetik der Invertinwirkung", *Biochem. Z.*, 336-369, 1913.

[93] E. Moreno, Y. Minhong, and K. Basler " Evolution of the TNF Signaling Mechanisms: JNK-Dependent Apoptosis Triggered by Eiger, The Drosophila Homolog of the TNF Superfamily" *Current Biology* Published online July8, 2002.

[94] C.J. Morton-Firth. "Stochastic Simulation of Cell Signalling Pathways" *http://www.zoo.cam.ac.uk/zoostaff/morton/index.htm*, September 1999.

[95] C.J. Morton-Firth, D. Bray, "Predicting temporal fluctuations in an intracellular signaling pathway". *J. theor. Biol.* 192:117-128, 1998.

[96] C.J. Morton-Firth, T.S. Shimizu, and D. Bray. "A Free-energy-based Stochastic Simulation of the Tar Receptor Complex" *J. Mol. Biol.*, 286:1059-1074, 1999.

[97] C.J. Morton-Firth. *Stochastic simulation of cell signaling pathways.* PhD Thesis. University of Cambridge. September 1998.

[98] A.N. Mucciardi. "Adaptive flight control systems", in *Principles and Practise of Bionics - NATO AGARD Bionics Symp.* Sept., pp.119-167, 1968.

[99] S.D. Muller, J. Marchetto, S. Airaghi, and P/ Koumoutsakos. "Optimization Based on Bacterial Chemotaxis" *IEEE Transactions on evolutionary computations*, Vol. 6, No. 1, February 2002.

[100] K.M. Ottemann, W. Xiao, Y.K.Shin, D.E.Jr Koshland. "A piston model for transmembrane signaling of the aspartate receptor" *Sience*, 285(5434):1751-4, Sep 10 1999.

[101] R. Overbeek *et al.* " WIT: integrated system of high-throughphut genome sequence analysis and metabolic reconstruction" *Nucleic Acids Res.*, 28:123-125, 2000.

[102] R. Pastor-Satorras, E. Smith, and R.V. Solé " Evolving Protein Interaction Networks through Gene Duplication" *J. Theor Biol.*, 222(2):199-210, May 2003.

[103] M. Pincus. "A Monte Carlo Method for the Approximate Solution of Certain Types of Constrained Optimization Problems" *Oper. Res.* 18, 1225-1228, 1970.

[104] B. Plateau and K. Atif, "Stochastic Automata Network For Modeling Parallel Systems". *IEEE Trans. on Software Engineering*, Vol.17, No.10, pp.1093-1108, 1991.

[105] V. Raghunathan, C. Schrugers, S. Park, and M. Srivastava. "Energy-Aware Wireless Microsensor Networks," *IEEE Signal Processing Magazine*, V. 19 No. 2, pp. 40-50, March 2002.

[106] J.C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, *et al.* "The protein-protein interaction map of Helicobacter pylori." *Nature* 409, 211-215, 2001.

[107] S. Redner "How popular is your paper? An empirical study of citation distribution", *Eur. Phys. J.*, B 23:267, 1998.

[108] C.R. Reeves. *Modern Heuristic Techniques for Combinatorial Problems.* McGraw-Hill, England, 1995.

[109] B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, *et al.* "Genome-wide location and function of DNA binding proteins." *Science* 290, 2306-2309, 2000.

[110] J. Rung, T. Schlitt, A. Brazma, K. Freivals, J. Vilo. Building and Analyzing Genome-Wide Gene Disruption Networks. *Bioinformatics*, Vol. 18 Suppl. 2, pp. s202-s210, 2002.

[111] M.R. Said, T.J. Begley, A.V. Oppenheim, D.A. Lauffenburger, L.D. Samson. "Global Network Analysis of Phenotypic Effects: Protein Networks and Toxicity Modulation in Saccharomyces cerevisiae". *Proc. Natl. Acad. Sci. USA.* 101(52):18006-11, December 2004.

[112] M.R. Said, A.V. Oppenheim and D.A. Lauffenburger, "A New Framework for Modeling Biochemical Signaling Networks across Evolutionary Boundaries", *Proc. Int. Conf. on Systems Biology (ICSB'02)*, Stockholm, Sweden, December 2002.

[113] M.R. Said, A.V. Oppenheim and D.A. Lauffenburger, "Modeling Cellular Signal Processing Using Interacting Markov Chains", *Proc. Int. Conf. on Accoustics, Speech, Signal Processing (ICASSP'03)*, Hong Kong, pp. VI-41 - VI-44, April 2003.

[114] C.K. Sestok, M.R. Said and A.V. Oppenheim, "Randomized Data Selection in Detection with Applications to Distributed Signal Processing", *Proceedings of the IEEE*, Vol. 91, Issue8, August 2003.

[115] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome Res.* 13, 2498-504, 2003.

[116] T.S. Shimizu, S.V. Aksenov, and D. Bray, "A spatially extended stochastic model of the bacterial chemotaxis signalling pathway" *J.Mol.Biol.* 329:291-309, 2003.

[117] V. Spirin, and L.A. Mirny. "Protein complexes and functional modules in molecular networks." *Proc Natl Acad Sci U S A* 100, 12123-12128, 2003.

[118] P.A. Spiro, J.S. Parkinson, and H.G. Othmer, "A model of excitation and adaptation in bacterial chemotaxis". *Proc. Natl. Acad. Sci. USA* 94, 7263-7268, 1997.

[119] N.C. Spitzer and T.J. Sejnowski Biological Information Processing: Bits of Progress *Science*, Vol. 277, pp 1060-1061, August 1997.

[120] A.H. Tong, M. Evangelista, A.B. Parsons, H. Xu, G.D. Bader, N. Page, M. Robinson, S. Raghibizadeh, C.W. Hogue, H. Bussey, *et al.* "Systematic genetic analysis with ordered arrays of yeast deletion mutants." *Science* 294, 2364-2368, 2001.

[121] A.H. Tong, G. Lesage, G.D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G.F. Berriz, R.L. Brost, M. Chang, *et al.* "Global mapping of the yeast genetic interaction network." *Science* 303, 808-813, 2004.

[122] TRANSFAC: http://transfac.gbf.de/TRANSFAC/

[123] P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, *et al.* "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." *Nature* 403, 623-627, 2000.

[124] A. Van Oudenaarden. "Modeling *Escherichia coli* chemotaxis," *Systems Biology Lecture Notes.* MIT, October 2004.

[125] P. Waage, C.M. Gulberg, and H.I. Abrash "Studies Concerning Affinity", *Journal of Chemical Education*, 63:1044-1047, 1986.

[126] A. Wagner and D.A. Fell " The small world inside large metabolic networks" *Proc. R. Soc. London*, B 268:1803, 2001.

[127] A. Wagner. "The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes." *Mol Biol Evol* 18, 1283-1292, 2001.

[128] A. Wagner " How the Global Structure of Protein Interaction Networks Evolves" *Proc. R. Soc. London*, B 270:457-466, 2003.

[129] A.J. Walhout, R. Sordella, X. Lu, J.L. Hartley, G.F. Temple, M.A. Brasch, N. Thierry-Mieg, and M. Vidal. "Protein interaction mapping in *C. elegans* using proteins involved in vulval development." *Science* 287, 116-122, 2000.

[130] D.J. Watts and S.H. Strogatz "Collective dynamics of small-world networks", *Nature*, 393:440, 1998.

[131] D.J. Watts *Small Worlds*, Princeton University Press, Princeton, NJ, 1999.

[132] M.S. While, and S.J. Flockton "Adaptive recursive filtering using evolutionary algorithms" in *Evolutionary Algorithms in Engineering Applications* (Editors: D. Dasgupta and Z. Michalewicz). Springer Verlag, 1997.

[133] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. PruB, I. Reuter, F. and Schacherer, "TRANSFAC: an integrated system for gene expression regulation Nucleic Acids Res. ", *Nucleic Acids Res.* 28:316-319, 2000.

[134] Wuchty, S. (2002). "Interaction and domain networks of yeast." *Proteomics* 2, 1715-1723.

[135] I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, and D. Eisenberg. "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions." *Nucleic Acids Res* 30, 303-305, 2002.

[136] YOUNG LAB: http://web.wi.mit.edu/young/

[137] F. Zhao, J. Shin, and J. Reich. "Information-Driven Dynamic Sensor Collaboration," *IEEE Signal Processing Magazine*, V. 19 No. 2, pp. 61-72, March 2002.