COMPUTATIONAL MODELS OF THE MAMMALIAN CELL CYCLE

by

MICHAEL CHRISTIAN WEIS

Submitted in partial fulfillment of the requirements
For the degree of Doctor of Philosophy

Department of Electrical Engineering and Computer Science

CASE WESTERN RESERVE UNIVERSITY

January, 2012
CASE WESTERN RESERVE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

We hereby approve the dissertation of

__________________________________________________________________________

Michael Christian Weis

candidate for the Doctor of Philosophy degree *.

__________________________________________________________________________

(signed) Sree N. Sreenath (chair of the committee)

__________________________________________________________________________

James W. Jacobberger

__________________________________________________________________________

Kenneth A. Loparo

__________________________________________________________________________

Vira Chankong

__________________________________________________________________________

Mihajlo D. Mesarovic

__________________________________________________________________________

(date) November 28, 2011

*We also certify that written approval has been obtained for any proprietary material contained therein.
Contents

List of Tables viii

List of Figures ix

Acknowledgments xiii

Abstract xv

1 Introduction 1

1.1 Context ....................................................... 1

1.2 Contributions .................................................. 4

2 Background 6

2.1 The Cell Cycle .................................................. 6

2.2 The Cell Cycle Control System ................................... 9

2.2.1 Basic Control Scheme ........................................ 9

2.2.2 Detailed Mechanisms ....................................... 12
CONTENTS

2.3 The History of Cell Cycle Modeling ........................................... 17
   2.3.1 A Timeline of Published Models ....................................... 17
   2.3.2 Models and Data ......................................................... 24

3 Computational Modeling Methodology ......................................... 30
   3.1 Modeling Overview ......................................................... 30
   3.2 Deterministic Modeling ..................................................... 35
      3.2.1 Mass Action and Michaelis-Menten Formulations ................. 36
      3.2.2 Hill Kinetics and Goldbeter-Koshland Functions ............... 41
   3.3 A Note on ODE’s ............................................................ 45
   3.4 Calibration ................................................................. 46
      3.4.1 Math Formulation ....................................................... 48
      3.4.2 Selecting a Set of Parameters to Optimize ....................... 49
      3.4.3 Biophysically Reasonable Parameter Values ....................... 54
      3.4.4 Optimization Algorithms ............................................ 55
      3.4.5 Model Adjustment ..................................................... 59

4 Measuring Cell Cycle Regulatory Molecules .................................. 60
   4.1 Asynchronous Data Extraction ............................................ 61
      4.1.1 Data Pre-processing ................................................... 62
      4.1.2 Segmentation of Data Continuum ................................... 62
      4.1.3 Cumulative Frequency Plot ......................................... 65
CONTENTS

4.1.4 Frequency to Time Conversion ........................................ 67
4.1.5 Further Scale Conversions ............................................. 69
4.2 Resulting Expression Profiles ............................................ 71

5 An Improved Computational Model of the Mammalian Cell Cycle 74
  5.1 We could not fit previous computational models to our data ...... 75
    5.1.1 Published models compared to the K562 data ................. 76
    5.1.2 Adjusting numerical rate constants ............................. 82
    5.1.3 Defining Goodness of Fit .......................................... 82
      5.1.3.1 Parameter Sensitivities and Non-Identifiability ........ 84
      5.1.3.2 Optimization Algorithms .................................... 90
  5.2 Model Modifications Improve the Fit ................................. 91
    5.2.1 Cyclin A expression is undetectable in G1 .................... 91
    5.2.2 cyclin A/cdk drives the activation of cyclin B/cdk1 ........ 93
    5.2.3 Expression of cyclins A2 and B1 is highly elevated in G2 ... 94
    5.2.4 An Improved Fit to the General Shape of the K562 Data ... 95
    5.2.5 Cyclin A2 is degraded before cyclin B1 ....................... 95
    5.2.6 Calibration of the Modified Model ............................. 101
  5.3 Re-analysis of the K562 Data .......................................... 104
  5.4 Cyclin A Binds Both Cdk1 and Cdk2 ................................. 105
  5.5 Discussion ............................................................... 110
# CONTENTS

## 6 Variation

6.1 Single Cell Measurement Technologies .......................................................... 115

6.2 Measurement Noise ............................................................................................ 117

6.3 Biological Variation ........................................................................................... 123

6.3.1 Modeling Intrinsic Noise ............................................................................. 125

6.3.2 Distributions of Initial Protein Concentrations or Kinetic Rate Constants .......... 127

6.3.3 Calibration .................................................................................................... 131

6.4 Population Interactions ..................................................................................... 134

6.5 Concluding Thoughts ....................................................................................... 135

## 7 Conclusions and Future Work

7.1 Thesis Review .................................................................................................... 136

7.2 Opportunities for Future Research ................................................................. 138

7.2.1 Improving the Model .................................................................................. 138

7.2.2 Fundamental Research ............................................................................... 138

7.2.3 Application to Cancer Research ............................................................... 139

## A Computational Models Developed in this Dissertation

A.1 Section 5.2: Original Data Set ......................................................................... 142

A.1.1 Equations .................................................................................................... 142

A.1.2 Definitions .................................................................................................. 144
List of Tables

2.1 Natural frequency distribution of cells in an exponentially growing population of K562 cells. .......................................................... 7
2.2 Approximate length of cell cycle phases. Calculated by correcting for the exponential distribution of cell ages. ................................. 8
2.3 Functionally related, and evolutionarily conserved, cell cycle regulatory proteins in different species ............................................. 16
2.4 Scope of published mammalian cell cycle models. .......................... 22
2.5 Classification of published cell cycle models .................................. 25

3.1 Online resources for determining relevant system components and interactions [1]. ................................................................. 34
3.2 Comparison of published and re-estimated parameters from an example calibration of a cell cycle model. The most sensitive parameters of the 2004 Novak and Tyson model [2] were re-calibrated using ideal, “artificial data.” ................................................................. 59

5.1 Names of the most globally sensitive parameter identified in Figure 5.5. Refer to the model in Appendix I for the meanings of the parameter names ................................................................. 86
5.2 Names of the 30% most globally sensitive parameter identified in Figure 5.12. Refer to the model in Appendix I for the meanings of the parameter names ................................................................. 102
List of Figures

2.1 Cell cycle phases and their significant regulatory proteins. .......................... 10
2.2 Basic architecture of the cell cycle control system. Diagram inspired by [3]. ................................................................. 11

3.1 The systems biology process. (Image courtesy of Prof. Olaf Wolkenhauer.) ................................................................. 32
3.2 Example convergence trajectory of a hybrid simulated annealing optimization routine, in the calibration of an example cell cycle model [2]. 57
3.3 Example re-calibration of the top 9 most sensitive parameters of a cell cycle model [2] using a hybrid simulated annealing routine. ............. 58

4.1 Flow cytometry data pre-processing steps include (a) Fluorescence compensation (via application of a bias) (b) Removal of doublets (through gating out cells) (c) Minimizing the effects of non-specific binding (application of a bias) [4]. These steps were completed in WinList (from Verity Software). ................................................................. 63
4.2 Derivation of the cyclin A2 expression profile. A. Cells are first partitioned into interphase and mitotic fractions. B. Mitotic cells are then further subdivided into arbitrary partitions. C. A Gaussian distribution is fit to determine the standard deviation of cells in prophase. D. Partitioning along the cyclin A2 trajectory, G1 cells are defined by the absence of cyclin A, R5. G2 cells are defined by R6. S phase cells have intermediate levels of Cyclin A and are partitioned orthogonal to the direction of greatest variation. E. A series of Gaussian distributions is fit to describe the movement of cells through G2 phase. The Gaussians are assumed to have a standard deviation roughly equal to that of prophase. F. Median cyclin A2 expression as a function of cumulative cell frequency. See [5] [6] [7] for full details. ................................. 66
LIST OF FIGURES

4.3 Summary of cyclin A2 expression profile extraction. A-D The gated sections of the K562 data set which define movement through indicated sections of the expression profile in F-H. The profile was extracted using the methodology represented in Figure 4.2 and discussed in the text. ........................................ 67

4.4 Derivation of the cyclin B2 expression profile. A. G1 and G2 cells are separated from S and M. B. The trajectory of the G2 cells are not further resolvable. C. Gaussian subpopulations are fit to a one dimensional histogram of cyclin B1 levels in G2. D. G1 cells do not show significant variation which would allow two dimensional gating. E. Guassians are fit to approximate movement through G1. F. Median cyclin B1 expression as a function of cumulative cell frequency. See [5] [6] [7] for full details. ........................................ 68

4.5 (a) Normalized expression profiles of cyclin A2, cyclin B1, and PHH3. (b) Magnified view of the final 2% of the cell cycle. K562 cells were stained, measured, and analyzed as described in Materials and Methods. 73

5.1 Schematic of the Conradie et al. 2010 model [8]. Solid lines indicate chemical reactions and dashed lines regulatory effects. .............. 77

5.2 Comparison of Conradie et al. [8] model outputs (solid lines) with K562 data (markers). (a, b, c) Total simulated levels (bound and free, inactive and active) of cyclins A and B, and cyclin B/cdk1 activity are compared to the measured expression profiles of cyclins A2 and B1, and PHH3. Simulations were carried out using the published protocols and kinetic rate constants. ........................................ 78

5.3 Schematic of the Csikasz-Nagy et al. 2006 model [9]. Solid lines indicate chemical reactions and dashed lines regulatory effects. ............ 80

5.4 Comparison of Csikasz-Nagy et al. [9] model outputs (solid lines) with K562 data (markers). (a, b, c) Total simulated levels (bound and free, inactive and active) of cyclins A and B, and cyclin B/cdk1 activity are compared to the measured expression profiles of cyclins A2 and B1, and PHH3. Simulations were carried out using the published protocols and kinetic rate constants. ........................................ 81

5.5 Global parameter sensitivities (weighted average least squares method) of the Csikasz-Nagy 2006 model, evaluated at the available data time points. ........................................ 86

5.6 Correlation matrix between local, time-dependent, parameter sensitivities evaluated at published parameter values. The absolute values of the correlations are indicated by the darkness of the color index, provided along the right of the matrix plot. .............. 88
LIST OF FIGURES

5.7 Correlation matrix between local, time-dependent, parameter sensitivities of the top 20 most globally sensitive parameters, evaluated at published parameter values. The absolute values of the correlations are indicated by the darkness of the color index, provided along the right of the matrix plot. .................................................. 89

5.8 Comparison of the modified model outputs (solid lines) to K562 data (markers). (a, b, c) Total simulated levels (bound and free, inactive and active) of cyclins A and B, and cyclin B/cdk1 activity are compared to the measured expression profiles of cyclins A2 and B1, and PHH3. 96

5.9 Magnified view of the final 2% of the cell cycle. ......................... 97

5.10 Magnified view of the final 2% of the cell cycle, with added synthesis of cyclin B through mitosis. ......................................................... 99

5.11 Schematic of the modified model developed in this study. The black colored portion represents the original Csikasz-Nagy 2006 model and the grey portion is the incorporated section of the Conradie 2010 model. The red portion indicates modifications made to calibrate the model to the K562 data. Solid lines indicate chemical reactions and dashed lines regulatory effects. ...................................................... 100

5.12 Global parameter sensitivities (weighted average least squares method) of the modified model developed in this chapter, evaluated at available data time points of cyclins A and B and cyclin B/cdk1 activity. ..... 102

5.13 Current model (solid lines) calibrated to the re-analyzed data (markers). The data show pronounced concavity in the shape of cyclin B1; the model is unable to capture this trend. ................................. 106

5.14 Introducing E2F autoregulation and E2F regulated synthesis of cyclin B allows an improved fit (solid lines) to the re-analyzed K562 data (markers). ................................................................. 107

5.15 Magnified view of the final 3% of the cell cycle. Magnitudes were normalized to allow plotting on the same scaled axis. 108

5.16 Diagram of the model modifications introduced to better fit the re-analyzed K562 data. The original Csikasz-Nagy 2006 model is shown in black, the Conradie 2010 model portion is shown in grey, modifications introduced previously in this dissertation are in red, and the modifications required to fit the re-analyzed data are presented in blue. 109

6.1 Simulation compared to the original K562 data. The model output was offset and scaled to approximate the median trajectory. .......... 115
LIST OF FIGURES

6.2  (a) Model simulation with additive and multiplicative noise. An equal number of cells as the K562 data set (b) were simulated by evaluating the system at 100,624 equally spaced points (having rescaled the time domain to the frequency domain – the inverse of that described in Section 4.1.4) to generate a representative sample, and perturbing each point by the given noise equation. Here $\epsilon_A \sim N(300, 120), \eta_A \sim N(0, 0.18), \epsilon_B \sim N(500, 160), \eta_B \sim N(0, 0.15)$ to approximate the offset and distribution of the K562 data. . . . . . . . . . . . . . . . . . . 121

6.3  Estimate of the temporal dynamics of the measured coefficient of variation of cyclin A2. Values were estimated by calculating the coefficient of variation about the median value in each “gate” defined in Chapter 4.122
Acknowledgments

Among my many acknowledgements, I would like to begin by expressing my sincere gratitude to Prof. Sree N. Sreenath. As my freshman advisor, I met him my first week of college and over the following year he introduced me to the interdisciplinary opportunities of the Systems and Control Engineering program. Later, he introduced me to Systems Biology, and he’s been a wonderful advisor and mentor throughout my research career. I am also thankful for the opportunity I’ve had to study with Prof. Mihajlo D. Mesarovic. Ever since coining the phrase “systems biology” in 1968, he’s stayed true to its core philosophy. A profoundly brilliant systems scientist, what little of his enormous wisdom I’ve managed to absorb has had a tremendous impact. I am also deeply grateful for the chance to work with Prof. James W. Jacobberger. His insight, precision, and work ethic are truly inspiring. My appreciation to Prof. Vira Chankong and Prof. Kenneth A. Loparo for their time and advice. My thanks to my previous biological collaborators, Dr. Cheng-Kui Qu, Dr. Kevin Bunting, Dr. Sanjay Gupta, and Dr. Sanjeev Shukla. I would also like to thank two former members of the Sreenath Lab, Dr. Jayant Avva and Dr. Radina Soebiyanto, for all their help.
ACKNOWLEDGMENTS

and conversations throughout the years.

I owe a tremendous debt of gratitude to my parents, for their unwavering support and encouragement. I never would have made it through it all without them. Finally, to my wife Victoria, for all her love and support. An artist who moved from New York City to Cleveland - if that’s not true love, I’m not sure what is.
Computational Models of the Mammalian Cell Cycle

Abstract

by

MICHAEL CHRISTIAN WEIS

Systems biology has sometimes been defined as the application of systems science and engineering concepts to biological problems. This dissertation illustrates the usefulness of this approach in understanding the regulation of the mammalian cell cycle. Cell growth and division are fundamental properties of life, and the dysregulation of cell cycle control is central to the development of cancer. Understandably then, the cell cycle has historically been a popular subject for mathematical modeling efforts and we review 154 models developed over the past 80 years. Beyond mathematics however, understanding systems requires the evaluation of models against data. The work presented herein illustrates an approach for estimating the median dynamic expression profiles of cell cycle regulatory molecules from a flow cytometric snapshot of an asynchronous population, and applies this data to the modification and calibra-
ABSTRACT

tion of a computational model of mammalian cell cycle control. This contribution illustrates the value of the systems biology approach in integrating existing evidence, interpreting data, and driving new hypotheses regarding the organizing principles of biological systems. Having used single cell data to model the median trajectory of a population, we then investigate approaches to simulate cell-cell variation and reproduce the distribution of cells originally measured with flow cytometry. This comprehensive methodology also establishes an approach to studying proliferative diseases, such as hematopoietic cancers, which can be easily sampled and measured using flow cytometry. As only one static measurement is needed to define the underlying expression profile, this may provide an entry point to applying computational models and systems engineering methodologies to the treatment of individual patients.
Chapter 1

Introduction

1.1 Context

Cells grow in size and divide to reproduce. The sequence of cell mass and genome doubling, and the division of duplicated cell contents into two separate daughter cells is referred to as the cell cycle. A complex network of regulatory molecules orders and coordinates the entire process, while also responding to internal and external conditions to encourage or arrest the process when appropriate – entry into a new round of division commences in response to sufficient nutrients and external growth signals, while division can be arrested in response to DNA duplication errors or misalignment of chromosomes on the mitotic spindles (commonly referred to as cell cycle “checkpoints”). Understanding the cell cycle is central to understanding life. More practically, it is also central to understanding cancer, as the dysregulation of cell
CHAPTER 1. INTRODUCTION

division is a central factor in the development of human malignancies [10].

Proliferative disorders such as cancer are characterized by unregulated cellular growth and reproduction. Pro-growth signaling pathways are often up-regulated or constitutively active, and large numbers of specific mutations have been identified [11, 12]. The cell cycle control network lies downstream of these pathways and orchestrates the physical processes of growth and division. Many proteins from this network are also frequently mutated in cancers [13], particularly those controlling cell cycle checkpoints. This, combined with dysregulated upstream signaling, drives a cell to rapid and unchecked proliferation.

Owing to its central importance both to basic science and therapeutic research then, the cell cycle has historically been a popular subject for mathematical modeling. Models of cell reproduction have a long history, dating back at least to 1932 [14,15]. However, until the discovery of the first cell cycle regulatory molecules in the 1970’s, these models were largely statistical descriptions of population growth rates or state transition models of the progression of cells through distinct phases. The discovery of a “maturation promoting factor” [16] and, in the following decade, the logic of a basic molecular oscillatory system underlying the cell cycle, provided a biochemical basis for building new mechanistic models. A still-enlarging network was soon discovered of a complex regulatory system controlling the concentrations and activities of key regulatory molecules which order and coordinate the precise timing of cellular events. The following decades saw the development of at least 126 mathematical models of
CHAPTER 1. INTRODUCTION

the cell cycle control system.

Beyond mathematics however, understanding systems requires the evaluation of models against data. Models are formalized hypotheses and calibration therefore represents the feasibility testing of one’s presumptions. Once a model has been calibrated to data and shown to be feasible, it is also validated to strengthen one’s confidence in its fidelity. Validated models can then be used to experiment in silico or otherwise investigate a system’s structure or organizing principles. For instance, models of the cell cycle have shown the importance of positive feedback loops [17] and the existence of bifurcations and hysteresis [18, 19]. This process is often referred to as systems biology: the application of systems science and engineering concepts to biological problems.

However, most published models are unsupported by quantitative data, and instead are focused on qualitative dynamics and the relative timings of oscillations. The few exceptions which did incorporate quantitative data relied largely on imprecise synchronization, timed sampling, and immunoblotting of a limited number of molecules. New single cell measurement techniques provide an entry point for the development of improved cell cycle models, which are precisely calibrated to specific model systems, to understand the basic architecture of cell cycle control. Furthermore, single cell data provide estimates of both the median trajectory of molecules and the distribution of a population about that center. This work therefore addresses the goals of the two former periods of cell cycle studies, improving recent attempts
CHAPTER 1. INTRODUCTION

at biochemical modeling while also building on earlier efforts to discern the statistics of populations.

1.2 Contributions

The first contribution is an intensive review of the state of cell cycle modeling. Cell cycle modeling has a long history that dates before the actual discovery of cell cycle regulatory molecules in the 1970's. Recent review articles have tallied at least 81 published models [20] and we’ve found at least 154, but there currently lacks a comprehensive analysis of how these studies compare to each other in terms of depth and breadth of the included interactions, their assumptions, and their comparison to real data. We review the state of the field and assess how close have we come to the main goal: a comprehensive model of mammalian cell cycle regulation, which includes most (if not all) of the known biology and is calibrated to rigorously derived experimental time courses of these regulatory molecules.

Cell cycle molecules are inherently difficult to measure and therefore most published models are only loosely coupled to data. Using a new methodology, we have extracted the dynamic expression profiles of key regulatory molecules from a human cell line. Comparing the most recent and most comprehensive models to this new data showed that while the models capture basic, canonical cell cycle knowledge, they are unable to fit our data well. Extensive calibration attempts strongly suggest that ad-
CHAPTER 1. INTRODUCTION

justing numerical rate constants will not improve the model’s behavior. Including additional well-established and hypothesized regulatory interactions did improve the fit of an existing cell cycle model to real data. This new model represents the first complete model of the mammalian cell cycle, spanning all four phases, which is calibrated to dynamic profiles. This contribution provides a starting point for validation of these ideas through biology and demonstrates the usefulness of mathematical models for understanding the organization and behavior of complex biochemical systems.

The final contribution completes the loop of analysis: having extracted a median profile from a cloud of asynchronous population measurements, methods are explored to expand the ordinary differential equation model of the median trajectory to simulate biological variance and measurement noise.
Chapter 2

Background

2.1 The Cell Cycle

Growth and reproduction are defining features of life. On a cellular level, cells grow and then divide to form two daughter cells (except in the case of embryos where division occurs rapidly without significant growth). The cycle of cell reproduction is generally divided into four phases: Gap 1 (G1), Synthesis (S), Gap 2 (G2), and Mitosis (M). A fifth phase, Gap 0 (G0), is sometimes also added to represent quiescence, a resting state. G0, G1, S, and G2 are collectively known as interphase, the time between consecutive mitoses.

The K562 cell line used in this work averages a total cell cycle length of about 18 hours, although lengths in other cell lines vary. Much of the variation in cell cycle lengths is attributed to the G1 phase. G1 represents a period of cell growth, and
CHAPTER 2. BACKGROUND

evaluation of pro- and anti-growth signals which determine whether a cell enters S phase or G0. Once a cell commits to a new round of division, it crosses a so-called restriction point (RP) and enters into a period of DNA synthesis (S phase). G2 is the last period of interphase and entails additional cell growth, protein synthesis, and evaluation of DNA integrity. A G2/DNA damage checkpoint ensures chromosome duplication has been completed successfully. M phase encompasses both mitosis, the division of chromosomes between two nuclei, and cytokinesis, the division of a cell’s cytoplasm into two daughter cells.

Table 2.1 lists the relative numbers of cells in each phase, for an asynchronously growing population of K562 cells. These percentages give a rough approximation to the relative lengths of each phase, but, because a population contains an exponentially weighted distribution of cell ages (originating from the division of one cell into two), this relation is not exact. This is discussed in more detail in Section 4.1.4, but the true distribution of phase lengths is given approximately by Table 2.1.

<table>
<thead>
<tr>
<th>Phase</th>
<th>% Frequency of Cells</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>31.48</td>
<td>31.48</td>
</tr>
<tr>
<td>S</td>
<td>60.61</td>
<td>92.08</td>
</tr>
<tr>
<td>G2</td>
<td>7.00</td>
<td>99.09</td>
</tr>
<tr>
<td>M</td>
<td>0.91</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2.1: Natural frequency distribution of cells in an exponentially growing population of K562 cells.
CHAPTER 2. BACKGROUND

<table>
<thead>
<tr>
<th>Phase</th>
<th>% of Total Cell Cycle Time</th>
<th>Cumulative Cell Cycle Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>19.58</td>
<td>19.58</td>
</tr>
<tr>
<td>S</td>
<td>67.58</td>
<td>87.16</td>
</tr>
<tr>
<td>G2</td>
<td>11.29</td>
<td>98.45</td>
</tr>
<tr>
<td>M</td>
<td>1.55</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2.2: Approximate length of cell cycle phases. Calculated by correcting for the exponential distribution of cell ages.

Mitosis is often subdivided into five subphases, defined by chromosome condensation, alignment on mitotic spindles, and division between two daughter nuclei. Prophase can be considered the period from centrosome separation to nuclear envelope breakdown (NEB). Prometaphase begins with NEB and involves the attachment of the condensed chromosomes to the spindle and the movement of the chromosomes towards the equator. Metaphase is officially recognized as the period during which the chromosomes remain aligned along the metaphase plate. Anaphase involves the division of sister chromatids into two separate sets of chromosomes. Telophase is the final stage of mitosis during which the chromosomes de-condense and nuclear envelopes are re-established around the two sets of daughter nuclei. While the process of mitosis divides the duplicated chromosomes between two daughter nuclei, the physical division of the complete cell into two distinct daughter cells is referred to as cytokinesis. Cytokinesis is often conceptualized as following mitosis, however it may begin as early as late anaphase.
CHAPTER 2. BACKGROUND

Following division, a daughter cell may either immediately re-enter the cell cycle, or diverge into the resting G0 phase. Some cells remain in quiescence for their entire lives, or only divide under very specific circumstances. Others, such as epithelial cells divide continuously throughout the organism's life and may never enter a G0 phase. Cells re-enter the cell cycle, or avoid G0 entirely, in response to growth factor signaling such as the well-studied Ras-dependent mitogen-activated protein kinase (MAPK) pathway. Sufficient activation of these pathways trigger a cell cycle regulatory network of interacting molecules, which orders and enacts the program of cell growth, DNA duplication, and division.

2.2 The Cell Cycle Control System

2.2.1 Basic Control Scheme

While cell growth and proliferation has been studied for more than a century, the discovery of the underlying molecular regulatory system is fairly recent. A “maturation promoting factor” (MPF) was discovered in 1971 to control the entry into mitosis of Xenopus laevis oocytes [16]. MPF was later found to oscillate through the normal cell cycle with a period equal to that of the cell cycle length. During the 1980’s, MPF was found to be composed of two subunits: a regulatory cyclin and a cyclin dependent kinase (cdk). While the cdk subunit is present at relatively constant levels throughout the cycle, its activity is regulated through binding to oscillating levels
CHAPTER 2. BACKGROUND

of cyclin. The levels of cyclin are regulated by opposing synthesis and degradation rates: specific transcription factors and ubiquitin-dependent proteolysis systems (e.g. the anaphase-promoting complex/cyclosome (APC/C)) that are also regulated by additional controls. Coupled transcription/translation and degradation sequentially orders the periods of high concentration of specific cyclins. As part of MPF, oscillating B cyclin levels could trigger mitosis periodically, apparently requiring threshold levels [21]. Similarly, the other cell cycle phases are controlled by sequential activation of specific cyclin/cdk complexes (Figure 2.1).

![Cell cycle diagram](image)

Figure 2.1: Cell cycle phases and their significant regulatory proteins.

The basic logic of the cell cycle control system has been highly conserved through-
out evolution and is shown in Figure 2.2. Oscillations in the S phase and mitotic cyclins (cyclins A and B) are orchestrated through the interactions of promoters and antagonists [3, 13, 22]. The cycle is started by the bistable interpretation of growth factor signaling (via cyclin D/cdk4,6 and the E2F transcription factor) which causes the synthesis and activation of “starter kinases” such as cyclin E/cdk2, which inhibit antagonists of cyclins A and B. This causes a rise in S phase and mitotic cyclin/cdk activities, the destruction of the starter kinases, and progression through the cell cycle. However, the activation of these cyclin/cdk complexes also promote their own destruction through the activation of “exit proteins” which which degrade the cyclins and inactivate their corresponding cdk partners. This resets the system until it is re-started by growth factor signaling and the re-synthesis of the starter kinases.

Figure 2.2: Basic architecture of the cell cycle control system. Diagram inspired by [3].

Besides binding to cyclins, cdk’s can be regulated through phosphorylation. Op-
posing kinases and phosphatases modulate the activity of cyclin/cdk dimers in response to checkpoint and other controls. Active cyclin/cdk complexes can also be inactivated by binding to peptide cdk inhibitors (CKI’s).

2.2.2 Detailed Mechanisms

Cyclins activate cdk partners, which perform the catalytic functions of the dimers. It is believed cyclins also target cdk’s to appropriate cellular targets. The canonical pairings are cyclin D to cdk’s 4 and 6, cyclin E to cdk2, cyclin A to cdk’s 1 and 2, and cyclin B to cdk1. The finer details are enormously more complicated, encompassing cyclins A1, A2, B1, B2, B3, C, D1, D2, D3, E1, E2, F, G, H, K, L, T1, T2a, and T2b binding with different subsets of cdk’s 1-20. [23]. Most of these pairings have only minor, or not well understood roles in cell cycle regulation and many of the cyclin isoforms perform cell-type specific roles. Consider cyclins A and B. There are two forms of cyclin A in mammalian cells, A1 and A2. Cyclin A1 is expressed only in meiosis, early embryos, and some tumors. Cyclin A2 is the primary cyclin A isoform in mammalian somatic cells (and references to cyclin A generally refer to this form). Similarly, there are three forms of cyclin B, B1, B2, and B3. Cyclin B3 is restricted to developing germ cells. Cyclins B1 and B2 are present in most cells. Cyclin B2 is dispensable for mouse development, while cyclin B1 is not. Functionally, however, either cyclin B1 or cyclin B2 can be knocked down in somatic cell lines with only minor perturbations. Cyclin B1 co-localizes with microtubules, while cyclin B2
associates with the Golgi apparatus. In the K562 cells discussed later in Chapter 3, however, cyclin B1 is the isoform expressed at highest concentration and throughout this dissertation, cyclin B refers to this form.

Cell division in multicellular organisms is tightly controlled, and the decision to begin the cell cycle is made in response to pro-growth signaling, which tips the balance of the cell cycle control system beyond the “restriction point” (RP) and commits the cell to a new cycle. Pro-growth signaling pathways, such as those converging on MAPK, up-regulate cyclin D activity. Cyclin D/cdk complexes respond to upstream signaling pathways by activating the E2F transcription factor. In quiescent cells, E2F is held inactive by the retinoblastoma protein (Rb), which is frequently mutated in human cancers. Pro-division molecules such as cyclin D phosphorylate Rb and release E2F, allowing translocation to the nucleus and the transcription of several other cyclins. Among these, cyclin E can further contribute to Rb phosphorylation and this positive feedback loop contributes (along with other players) to the bistable nature of the Rb/E2F switch, which responds to sufficient levels of growth factors [24]. E2F also promotes the transcription of cyclins A and B, major regulators of S and M phases, respectively. Complicating this simple picture, there are seven known isoforms of E2F, although only E2F’s 1-3 bind to Rb and canonically regulate the G1/S transition as discussed. E2F’s also regulate the cell cycle in other complex, but less understood ways [25].

Oscillations in cyclin/cdk activities are the result of several sets of mutually an-
CHAPTER 2. BACKGROUND

tagonistic relations (Figure 2.2) [3]. Active cyclin/cdk’s inhibit APC/C complexes (bound to Cdh1 and Cdc20 specification factors), but cyclin B/cdk1 also activates APC/C at the end of the cycle. Additionally, the peptide inhibitors CKI’s, comprising two structurally distinct families, bind and regulate specific cyclin/cdk complexes [22]. The INK4 gene family includes p14, p15, p16, p18, and p19, which control G1 phase through inhibition of cyclin D/cdk4 and cyclin D/cdk6. The CIP/KIP gene family includes p21, p27, and p57 and regulate cyclin D/cdk4,6, cyclin E/cdk2, cyclin A/cdk2 and cyclin B/cdk1 complexes. For example, during G1, p27 binds to cyclins A/cdk and cyclin E/cdk dimers to form inactive trimers. On the other hand, p27 is degraded in response to phosphorylation on threone-187 by active cdk’s. The activation of cyclin D/cdk4,6 binds and titrates away p27 (remaining active despite p27 inhibitory effects on other cyclin/cdk’s). This tips the balance towards the synthesis of cyclins E and A, and the rapid degradation of p27.

The synthesis of DNA and the centrosome (microtubule organizing center) begins soon after cells cross the RP and enter S phase. Initiation and continuation of DNA synthesis is regulated by cyclin A/cdk2. Cyclin A/cdk1 also likely plays a role, although the distinct actions of cdk1 or cdk2 containing dimers are still not well understood.

The G2 phase is a gap between the end of DNA synthesis and the beginning of mitosis. During this gap, cells assess the state of chromosome replication and prepare to begin nuclear and then cytoplasmic division. The transition to M phase is regulated
by cyclin B/cdk1. Mitosis is a complex and critical event, and there are thus many
controls within a cell ensuring proper entry. Many of these regulations converge
on the activity of cyclin B/cdk1. While cyclin B is synthesized throughout the S
and G2 phases, it is kept inactive through phosphorylation of cdk1 on Threonine-14
and Tyrosine-15 by the kinases Wee1 and Myt1. Cyclin B/cdk1 is then activated
by dephosphorylation of the cdk1 sites by the phosphatase Cdc25. Cdc25 is also
phosphorylated and activated by cyclin B/cdk1, creating a bistable feedback loop
which causes cyclin B/cdk1 activity to rise abruptly at the G2/M transition once
Wee1/Myt1 is inactivated and a small fraction of either cyclin B/cdk1 or Cdc25
is activated \[?\?\].

During interphase, cyclin B/cdk1 is localized to the cytoplasm. However, following activation, cyclin B/cdk1 catalyzes NEB and migrates to the
nucleus.

Exit from mitosis requires the degradation of the mitotic cyclins. This is largely
accomplished by the APC/C complex which is activated via phosphorylation by PLK1
and/or cyclin B/cdk1 and binding to Cdc20. APC/C also causes the destruction of
securins and the subsequent separation of sister chromatids. Towards the end of mito-
sis, as the mitotic cyclins are degraded by APC/Cdc20, APC/Cdh1 also contributes
in removing cyclin activity and restoring the cell to a G1 state in preparation of
subsequent cycles.

The molecular control system regulating cell division has been highly conserved
throughout the evolution of eukaryotes, and a functionally analogous but evolution-
CHAPTER 2. BACKGROUND

<table>
<thead>
<tr>
<th>Budding Yeast</th>
<th>Fission Yeast</th>
<th>Xenopus</th>
<th>Mammal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cdc28/Clb1,2</td>
<td>Cdc2/Cdc13</td>
<td>Cdc2/CyclinB</td>
<td>Cdc2/CyclinB</td>
</tr>
<tr>
<td>Cdc28/Clb5,6</td>
<td>Cdc2/Cig2</td>
<td>Cdk1,2/CyclinA</td>
<td>Cdk1,2/CyclinA</td>
</tr>
<tr>
<td>Cdc28/Cln1,2</td>
<td>-</td>
<td>Cdk2/CyclinE</td>
<td>Cdk2/CyclinE</td>
</tr>
<tr>
<td>Cdc28/Cln3</td>
<td>Cdc2/Puc1</td>
<td>Cdk4,6/CyclinD</td>
<td>Cdk4,6/CyclinD</td>
</tr>
<tr>
<td>Sic1</td>
<td>Rum1</td>
<td>Xic1</td>
<td>p27Kip1</td>
</tr>
<tr>
<td>Cdh1</td>
<td>Ste9</td>
<td>Fzr</td>
<td>Cdh1</td>
</tr>
<tr>
<td>Swe1</td>
<td>Wee1</td>
<td>Xwee1</td>
<td>Wee1</td>
</tr>
<tr>
<td>Mih1</td>
<td>Cdc25</td>
<td>Xcdc25</td>
<td>Cdc25C</td>
</tr>
<tr>
<td>Cdc20</td>
<td>Slp1</td>
<td>Fizzy</td>
<td>p55Cdc, Cdc20</td>
</tr>
<tr>
<td>Swi4/Swi6, Mbp1/Swi6</td>
<td>Cdc10/Res1</td>
<td>XE2F</td>
<td>E2F</td>
</tr>
<tr>
<td>Whi5</td>
<td>-</td>
<td>Rb</td>
<td>Rb</td>
</tr>
</tbody>
</table>

Table 2.3: Functionally related, and evolutionarily conserved, cell cycle regulatory proteins in different species

A highly distinct network has also been identified in the prokaryotic Caulobacter crescentus [26,27]. As discussed later in this chapter, model organisms have been central to understanding cell cycle control. Fertilized Xenopus laevis eggs represent the simplest known and easiest manipulated system. They quickly cycle through repeated mitoses utilizing a highly reduced core of the control system, and this system has been recapitulated in cell free extracts. Use of this system led to important discoveries in the feedback loops between Wee1, Cdc25 and cyclin B/cdk1 [17,28]. Similarly, the large number of established mutant yeast strains has allowed the construction of complex models and the qualitative validation of these models against a large library of experimental evidence [29]. This dissertation generally uses the language and acronyms from the mammalian literature, but analogues exist for most model species. A conversion chart, inspired is provided in Table 2.2.2 showing the names of functionally related proteins in different species.
CHAPTER 2. BACKGROUND

The molecular regulatory network governing cell division is highly complex. This complexity arises not only from the sheer size of the network, but also from overlapping functions and nonlinear responses. Some of the special difficulties which complicate analysis are that: (1) enzymes are substrates of other enzymes, (2) many of the proteins are transcription factors that affect the synthesis rates of other pathway components, and (3) molecules are activated or inhibited by forming multimers with other molecules [30]. The system is largely self-organized and contains a number of closed loops which produce staggered oscillations of phase-specific regulators. Mathematical modeling provides a way to organize the many interactions discovered so far, allows reasoning with these facts to predict the system’s responses to perturbations (e.g. drugs), and aids in interpreting and directing new biological experiments.

2.3 The History of Cell Cycle Modeling

2.3.1 A Timeline of Published Models

The first models considered cell cycle control to be purely random [14,15,31,32]. Histograms of generation time were modeled with the assumption that the probability of division is a function of a cell’s present age. Eventually, these single state, probabilistic models were then expanded to include two states, one random, and one deterministic (representing random entry into a deterministic growth state) [33,34]. With increasing evidence that cell size is a significant factor, such models were then
further extended to include size control in the probability of division [35–46].

A molecular basis for cell cycle regulation was completely unknown until the 1970’s. MPF was discovered in Xenopus oocytes in 1971 [16] and its oscillatory role in cell cycle control was identified a bit later in 1978 [47]. With a biochemical basis for cell cycle transitions, models became deterministic and largely based on ordinary differential equations (ODE’s) of oscillatory systems. John Tyson published one of the first models of the hypothetical control of mitosis by an autonomous biochemical oscillator (a limit cycle generating biochemical system, generically termed a “Brusselator”) [48]. This first biochemical model was accompanied by models of a hypothesized unstable inhibitor [37] or activator [49–51] of replication. These models also represent the first use of nonlinear systems of ODE’s, which remain the dominant form of models.

The first models incorporating our modern understanding of cyclin/cdk regulation followed the discovery of this logic in the 1980’s [52–54], used small systems of ODE’s [55–62] and captured the basic oscillatory trend of cell cycle regulation. Among these, the 1991 Tyson model was more comprehensive and similar to the “systems biology” models most familiar today - incorporating cyclin synthesis and destruction, cyclin binding phosphorylated cdc2 to form “pre-MPF”, autocatalytic dephosphorylation and activation of pre-MPF to form active MPF, breakdown of MPF and degradation of cyclin, and finally the re-phosphorylation of cdc2. It appears to be the first Mass Action based ODE model of the mitotic control core. Tyson and Novak
CHAPTER 2. BACKGROUND

would next expand the model to include Wee1 and Cdc25 control of MPF activation, particularize the model for a specific cell type (Xenopus oocytes) and compare it to experiments in extracts and intact embryos [28]. The model was later tailored to fission yeast, and validated against the qualitative behaviors of 20 temperature sensitive mutants reported in the literature [63]. These two papers mark a transition in cell cycle modeling from small models of generic, qualitative dynamics, to much larger and more complex models which are more rigorously validated against (qualitative) experimental data.

Early cell cycle models focused on the control of M phase, and centered on MPF and the regulation of cdc2. Less was known about the control of earlier cell cycle phases, and these were often completely ignored or modeled in a “black box” fashion. Embryos replicate and divide autonomously and so this exclusion was often insignificant in modeling, for example, Xenopus oocytes. In most eukaryotes, however, there exists a G1 “waiting period” before replication of DNA commences in S phase (the random wait time before the deterministic synthesis phase in the probabilistic early state transition models discussed earlier). To account for this, the 1995 Novak and Tyson model [63] employed an “automaton model” – logical rules which simulate physiological observations without knowledge of the underlying biochemistry.

The next logical step for extending cell cycle modeling was then the inclusion of other biochemicals controlling G1, S, or G2 phases. Bertuzzi et al. [64] published a small model, inspired by the analytic forms of previous non-mechanistic
models [58, 59], of the activities of an S phase promoting cyclin and cdk, and a separate cyclin and cdk for mitotis (comprising four ODE’s). Novak and Tyson also published an extension of their fission yeast model [63], adding Cig2/cdc2 and Rum1 (equivalent to Cyclin A/Cdk1 and p27 in mammals) [65] and expanded the model further by including APC interactions to yield a model of nearly the complete cell cycle, incorporating G1/S, G2/M, and metaphase/anaphase checkpoint controls in the form of Cdc13/Cdc2 (equivalent to cyclin B/cdk1) interactions with Rum1, Wee1 and Cdc25, and APC [66]. This model was expanded further in 2000 to include the APC specification factors Cdc20 and Cdh1 [67].

These models by Novak and Tyson also mark the beginning of a continuing trend in cell cycle modeling. The model does not consider the separate dynamics of cyclins and cdks, instead only modeling the evolution of dimer concentrations. Cdk’s are known to be expressed at a much higher level than their cyclin partners and the binding of cyclin to cdk, and subsequent activation of dimers by, for example, cyclin H/Cdk7, is assumed to occur sufficiently fast as to be neglected. This modeling convention is still largely in use today, where the binding of cdk to cyclin is assumed to be in “quasi-steady state” and ignored, dramatically simplifying the number of modeled reactions.

Early cell cycle modeling focused on reduced systems such as the fertilized frog egg which oscillates autonomously between S phase and mitosis, with no real cell growth, and no significant G1 and G2 phases. While more regulated, single celled organisms
CHAPTER 2. BACKGROUND

such as yeast generally grow and divide as rapidly as nutritional conditions allow. On the other hand, the normal mammalian cell cycle is necessarily more constrained. Cell division and tissue growth is highly restricted and loss of this regulation is contributes to the development of many cancers. Mammalian cells have thus evolved sophisticated mechanisms for sensing and responding to growth and anti-growth signals. Growth signals are generally interpreted through the funneling of signaling pathways to the Rb/E2F “switch” which controls the transition through the restriction point, beyond which a cell commits to beginning the mitotic program. Obeyesekere and coworkers developed models of the G1 phase of mammalian cells based on the molecular interactions of cyclin D, cdk4, cyclin E, cdk2, E2F, and Rb [68, 69]. However, it should be noted that, like many other early cell cycle models [58, 59, 64], these G1/S models were not formulated according to strict Mass Action formalisms, instead relying on phenomenological (non-mechanistic) reaction rates to produce desired nonlinear effects. Other groups have also modeled the G1/S transition in more molecular detail. In fact, fully 16 of the 27 published mammalian models focus exclusively on the G1/S core regulatory system (Table 2.3.1). For example, Kohn et al. published an intensive investigation of the functional units of the system, beginning with a simple system of E2F and Rb, and then adding first cyclin E and cdk2, and then cyclin D and cdk4 as triggers [30]. Others have since published similarly detailed, mechanistic models of this transition [70–73], eventually incorporating more complex interactions involving inhibitors such as p27, or cyclin A /cdk2 [74–79]. These mech-
CHAPTER 2. BACKGROUND

anisms were eventually folded into earlier S/G2/M models (originally derived from Xenopus and yeast studies). This coupling of G1/S models with core mechanisms from other species represents the current state of the art [2, 8, 9]. A few models incorporate rudimentary growth factor signaling (generally approximated with one or two equations) [2, 8, 80, 81], but, surprisingly, none seem to incorporate realistic models of relevant signaling pathways, despite the existence of several relevant models of MAPK, for instance [82].

<table>
<thead>
<tr>
<th>FORMALISM</th>
<th>PHASES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G1/S</td>
</tr>
<tr>
<td>Ordinary Differential Equation</td>
<td>[30, 68, 69, 71–78, 80, 81, 83, 84]</td>
</tr>
<tr>
<td>Logical</td>
<td></td>
</tr>
<tr>
<td>Stochastic</td>
<td>[74]</td>
</tr>
<tr>
<td>Hybrid</td>
<td></td>
</tr>
<tr>
<td>Delay Differential Equation</td>
<td>[79]</td>
</tr>
</tbody>
</table>

Table 2.4: Scope of published mammalian cell cycle models.

All of the models discussed so far involve the conversion of biochemical reactions into systems of ODE’s which, because they are analytically intractable, are then numerically simulated on a computer. Simulation and study of ODE’s requires
CHAPTER 2. BACKGROUND

specification of numerical rate constants, which define the strengths of the interactions and the exact dynamics of the system (this process is discussed earlier in the “Mathematical Modeling” section). On a more basic level however, understanding molecular interactions first requires identifying the network they comprise. The first step in extending this analysis towards dynamics is what might be termed “qualitative modeling” [94]. Among these, logical models begin with a network diagram, and define Boolean rules to determine the effects of different combinations of interactions. The dynamic behavior of components then is often modeled in a digital manner - present or absent and the sequence of activations/inactivations can then be propagated through the network from an initial starting state. Boolean modeling of the cell cycle is a relatively recent approach, but several models have been proposed, ranging from a simplified model of the mammalian cell cycle [90] to more complex models of budding yeast [95,96], fission yeast [97,98], and mammals [89].

In the other direction, arguably towards more detail, many stochastic models of the cell cycle have also been published. Deterministic ODE models describe the behavior of an “average” cell. However, this neglects the differences among cells in culture. Additionally, as discussed in Chapter 6, stochastic models capture aspects of biochemistry that may not be reliably approximated with deterministic approaches such as the discrete nature of a small number of molecules binding and reacting within a large volume. Examples of published stochastic models are listed in Table 2.3.1. For a more complete discussion, see Chapter 6.
CHAPTER 2. BACKGROUND

Finally, a few hybrid models have been published, combining stochastic and deterministic elements, as well as two models based on delay differential equations (DDE), which couple typical ODE’s with an ad hoc “waiting time” to account for un-modeled dynamics.

2.3.2 Models and Data

The scientific method requires the evaluation of mental models (hypotheses) against data. Similarly, systems biology requires the coupling of mathematical models and data for calibration and validation. Obtaining cell cycle relevant data has always been a challenge and cell cycle modeling unsurprisingly has had a complicated relationship with data. The earliest, phenomenological models were generally tightly coupled to observations of cell cycle transit times (“alpha” and “beta” curves). However, until recently, models of the molecular interactions governing cell cycle control have only loosely been coupled to experimental observations. The vast majority of early models were mainly conjectural (generating biological hypotheses of basic control mechanisms) and unsupported by data. In tandem with common biological reasoning, most models have been calibrated or validated against “qualitative” observations or general information about the timings of events. For example, Hatzimanikatis et al. [71] explain their calibration process as “The parameters were chosen to obtain a reasonable distribution of cell cycle phases in time (i.e., S phase shorter than G1 phase) where the G1-phase is defined as the time for which the amount of hypophosphorylated
<table>
<thead>
<tr>
<th>Formalism</th>
<th>Prokaryotes or Slime Mold</th>
<th>Yeast</th>
<th>Xenopus or Drosophila</th>
<th>Mammals</th>
<th>Generic or Unspecified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenomenological</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[33–38, 40, 42–46, 49, 99–113]</td>
</tr>
<tr>
<td>Logical/Boolean</td>
<td></td>
<td>[95, 114–118]</td>
<td></td>
<td>[89,90]</td>
<td></td>
</tr>
<tr>
<td>Ordinary Differential Equation</td>
<td>[26, 27, 48, 51, 120,121]</td>
<td>[26,29,63,65–67, 116,122–137]</td>
<td>[17,28,55,57,61, 62,138–148]</td>
<td>[2,8,9,30,68–78, 80,81,83–88,149]</td>
<td>[9,50,51,56,58–60,64,149–159]</td>
</tr>
<tr>
<td>Stochastic</td>
<td></td>
<td>[114–117, 124–126,131,135,137, 160–167]</td>
<td></td>
<td>[74,91,92]</td>
<td></td>
</tr>
<tr>
<td>Hybrid</td>
<td></td>
<td>[131,168]</td>
<td></td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>Delay Differential Equation</td>
<td></td>
<td>[169]</td>
<td></td>
<td>79</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.5: Classification of published cell cycle models
pRb exceeds the amount of hyperphosphorylated pRb.” This is despite the precise mathematical formalism of the models. Correctness is usually judged on a superficial “present/not present” level of reasoning, in line with biological observations. Understanding the control and behavior of nonlinear control systems however requires much more detailed information on the exact trajectories of observable state variables. It wasn’t until the past decade that dynamic expression profiles were first compared to mathematical models. However, even these were only crude approximations based on synchronization studies.

Models of the cell cycle system have therefore been largely calibrated based on a qualitative understanding of the timings of events. Model parameters are loosely tuned to produce sequences of oscillations in molecules known to be active in corresponding cell cycle phases. Extending this approach is the use genetic perturbations for qualitative validation of the proposed system structure. For example, researchers have assembled a large catalogue of yeast with mutated cell cycle genes. Chen et al. [29] utilized over 131 genetically engineered strains to confirm the canonical “wiring diagram” of the yeast cell cycle and constrain the calibration of a computational model.

The use of quantitative time series in models of the cell cycle is rare. This is likely due to the inherent difficulties in measuring cell cycle regulatory proteins. Cells are naturally asynchronous, and so measurements require the physical or chemical synchronization of cells, which is imperfect and only provides rough estimates of
CHAPTER 2. BACKGROUND

the levels and activities of regulatory molecules (discussed in more detail in Chapter 4). Single cell measurements should allow better estimates of the true dynamics of cell cycle regulation. Methods such as flow cytometry or fluorescence microscopy allow measurements of individual cells and the precise separation of cells by cell cycle progression to better estimate corresponding levels of regulatory molecules.

Xenopus oocytes have been central to the understanding of the cell cycle. After fertilization, the early embryo begins a series of rapid, highly simplified cell cycles. These simplified cycles have been reconstructed in cell-free extracts and allow a high degree of flexibility in experimentation. Xenopus extracts have also provided the earliest examples of quantitative time series data of cell cycle regulation. In 1993, Bela Novak and John Tyson published a model [28] of the core control network (MPF interacting with Wee1 and Cdc25) and compared their model to a set of data from the literature. A similar analysis was also published on the same model 16 years later by Zwolak et al. [144], comparing the model behavior to a much larger collection of published experiments. Similarly, Pomerening et al. [17] utilized densely sampled extract measurements to calibrate a model of the same core system (MPF, Wee1, and Cdc25). The novelty of their publication however, lies in the clever manipulation of the system to illustrate the importance of positive feedback in coordinating entry to mitosis. Introducing a Wee1- and Cdc25-insensitive Cdc2 mutant shortcuts the traditional regulatory network and results in faster, damped oscillations in Cdc2 activity which interferes with DNA replication and M phase exit. This paper brilliantly
illustrates both the use of modeling in dissecting the role of regulatory interactions and the use of quantitative time series data in evaluating system behavior.

The yeast cell cycle has also served as a highly popular model system. However, despite the large number of published models, few of them have been calibrated to dynamic data, instead using large sets of mutants to evaluate the validity of modeled interactions (qualitative validation). The few exceptions include Barberis et al. [131] who utilized time courses of Sic1 and Clb5 and Charvin et al. [135] which used single cell fluorescent microscopy to measure the response of Whi5 and Cdc10 to various mutations and chemical cell cycle “blocks.” This last model however, was a highly reduced one equation model. Finally, in addition to the qualitative validation discussed in the previous section, Cross et al. [128] used a novel method of separating growing yeast according to cell size (a proxy for time), in order to measure the levels of key regulatory molecules. This allowed comparison of model simulations of several mutants to the corresponding time course gathered from various mutant strains.

Most relevant to this dissertation are models of the mammalian cell cycle. Of those reviewed in the previous section, only two appear to have been rigorously coupled to dynamic expression data. Both of these used synchronization, timed sampling, and immunoblotting of human cell lines to estimate expression of molecules relevant to the G1/S transition. Haberichter et al. [79] synchronized HCT116 cells by density arrest and, upon re-plating at lower density and subsequent re-entry to the cell cycle, measured the levels of phosphorylated Rb, p130, Cdk2, Cdk4/6, cyclins D1 and D2,
CHAPTER 2. BACKGROUND

cyclin A2, Emi, p27, and cyclin B. Similarly, Alfieri et al. [78] synchronized NIH3TC cells by serum starvation and, upon stimulation, measured cyclin D, Cdk4, p27, cyclin E and Cdk2. Limitations of these studies include the inherent inexactness of synchronization and the small scope of the models, only measuring and simulating G1 and S phases.
Chapter 3

Computational Modeling

Methodology

3.1 Modeling Overview

“The aim of science is not things in themselves, but the relations between things; outside these relations there is no reality knowable.”
- Henry Poincare

Models are ubiquitous in biology. Until recently, these models have largely been conceptual. In their most abstract sense, conceptual models separate levels of reality according to spatial and temporal domains. Hierarchical representations recognize organisms, organs, tissues, cells, functional modules of proteins such as signaling pathways, protein-protein interactions, and, at an even lower level, the atomic level interactions which determine the folding of amino acids into a functional protein. A
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

Key goal of modern biology is understanding both the exact interactions of elements on a given level, and also the evolutionary purpose of elements in a larger context—how the functioning of elements on one level contribute to the adaptive behavior of the system when viewed on a higher hierarchical plane. A protein may be defined as “pro-growth” for its role in promoting cell cycle entry, which results in the replication of a cell, or as “pro-death” if it contributes to apoptotic programs. Systems biology recognizes the importance of this context and seeks an integrated understanding of biological interactions [170].

Biology also employs conceptual models to organize accumulated understanding into narrative accounts of interactions and cause/effect relationships. Mathematical and computational models simply represent further levels of precision, and aid intuition as the number of variables and the complexity of a system increase. As explained by Olaf Wolkenhauer [171] “Mathematical models are the extended arm of common sense; the only means we have to deal with non-intuitive complexity – no more but also no less.” Mathematical modeling forces formalization of biological concepts into a precise language, and models allow analytic evaluation of the behavior of complex systems. The iterative process of mathematical model building and comparison to biological data clarifies the assumptions and understandings inherent in traditional narrative models of biology.

Beyond mathematics, computational models require specification of numerical rate constants, the initial system state, and an algorithm to numerically approximate the
behavior of the system. Computational models are calibrated to data by tuning the rate constants and the initial system state. In contrast, validation represents the testing of a computational model against an independent data set, ideally from perturbed system conditions. Once a computational model is built which faithfully represents known biology and is able to reproduce experimental data, formal analysis can be conducted to determine the organizing principles of the system. In silico experiments can also be conducted to better understand the effects of observed mutations or to predict the system’s behavior to perturbations (e.g. drug treatments). The iterative process of hypothesis generation, data gathering, mathematical and computational modeling, and simulation analysis is often referred to as the “systems biology process” (Figure 3.1).

Figure 3.1: The systems biology process. (Image courtesy of Prof. Olaf Wolkenhauer.)
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

Mathematical models can be characterized on a continuum between abstracted and more specified models [172]. At their most abstract, models can simply identify the connections between components through statistical mining or bayesian network analysis. As more details are specified, Boolean models can be created which describe qualitative relationships and timings of events. Very low level models can also be created describing the temporal evolution of molecular concentrations with ordinary differential equations, or even the spatial effects of protein diffusion and localization to solid substrates using partial differential equations. The level of detail required depends on the questions being asked and the data which is available to help answer them. Systems biology typically employs ODE models as they are relatively easy to construct and analyze, and are helpful in both predicting behavior and understanding functioning.

The basic modeling process can be identified as

1. Define the relevant parts of the system

2. Define the inputs (stimuli) and outputs (measurables)

3. Define known or hypothesized interactions (create a reaction diagram/“cartoon”/causality flow)

4. Convert the list of interactions to mathematical form

5. Calibration: estimate kinetic rate constants and initial system state from available data
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

6. Validation: test the model against new experiments

Defining the relevant parts of a system is the first step in constructing a computational model. This generally requires searching the existing literature for review publications or previously published models. Internet databases are also sometimes helpful. Example resources are listed in Table 3.1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biocarta</td>
<td>Signaling pathway diagrams</td>
</tr>
<tr>
<td>BioModels</td>
<td>Mathematical Models</td>
</tr>
<tr>
<td>BOND</td>
<td>Biomolecular interactions</td>
</tr>
<tr>
<td>Cell Cycle</td>
<td>Cell cycle specific genes, proteins and mathematical models</td>
</tr>
<tr>
<td>Database</td>
<td></td>
</tr>
<tr>
<td>CellML Model</td>
<td>Mathematical Models</td>
</tr>
<tr>
<td>Repository</td>
<td></td>
</tr>
<tr>
<td>DIP</td>
<td>Protein interactions and kinetic parameters</td>
</tr>
<tr>
<td>DOQCS</td>
<td>Protein interactions, kinetic parameters, mathematical models</td>
</tr>
<tr>
<td>JWS Online</td>
<td>Mathematical models and online simulator</td>
</tr>
<tr>
<td>KEGG</td>
<td>Signaling and metabolic pathway diagrams with links to relevant literature</td>
</tr>
<tr>
<td>PANTHER</td>
<td>Signaling pathway diagrams and interaction lists</td>
</tr>
<tr>
<td>PathGuide</td>
<td>List of other signaling pathway databases</td>
</tr>
<tr>
<td>SPAD</td>
<td>Signaling pathway diagrams</td>
</tr>
<tr>
<td>The Cancer Cell Map</td>
<td>Cancer related signaling pathway maps</td>
</tr>
</tbody>
</table>

Table 3.1: Online resources for determining relevant system components and interactions [1].

As a traditionally narrative subject, biological observations are usually described in text, or with diagrams. Such diagrams are sometimes referred to as “cartoons” to indicate the obvious lack of rigorous detail. Nevertheless, they have been fairly standardized and do an efficient job communicating the basic relations between system elements, at least at a qualitative level (e.g. a protein positively or negative regulates
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

another protein). However, these diagrams only provide a roadmap of the system, and “what we really seek to know are the traffic patterns, why such traffic patterns emerge, and how we can control them” [173]. Dynamic models fulfill this need.

Converting a diagram to mathematics first requires identification of an appropriate framework. Models can be deterministic or stochastic, discrete or continuous time, discrete or continuous state, or even a hybrid. Systems biology traditionally embraces a continuous deterministic approach based on ordinary differential equations (which are covered in much more detail in the following subsection). However many other approaches have been published and one’s choice depends on the questions asked and how faithfully such approaches capture important system details.

3.2 Deterministic Modeling

Systems biology seeks to understand biological causality, and “causation... is a relationship, not between components, but between changes of states of a system” [174]. Differential equations then are a natural formalism. The rate of change of a system state $x$ is defined by

$$\frac{dx}{dt} = f(x(t), k)$$

The rate of change $\frac{dx}{dt}$ (also written as $\dot{x}$) is defined by a function (or mapping) $f$ of the current state vector $x(t)$ of the system and a set of parameters $k$. 

35
3.2.1 Mass Action and Michaelis-Menten Formulations

The concentrations of biochemical species are often assumed to change according to the Law of Mass Action, which states that the rate of a chemical reaction is directly proportional to the product of the concentrations of the reactants. Consider, for example, the classic enzymatic reaction

\[ \text{S} + \text{E} \xrightleftharpoons[k_1]{k_2} \text{SE} \xrightarrow{k_3} \text{P} + \text{E} \]

in which a substrate S is catalyzed by an enzyme E, initially forming the complex SE before producing the product P and original enzyme E. Such a reaction scheme is very common in biochemical networks. When a protein kinase “activates” a downstream target, it often accomplishes this by transferring a phosphate group from a high-energy ATP molecule to a substrate. The substrate is thus converted to a product (the phosphorylated form of the substrate), while the enzyme remains unaffected.

To derive equations for this simple system, we first decompose the scheme into three elementary reaction steps

\[ \text{E} + \text{S} \xrightarrow{k_1} \text{ES} \]
\[ \text{ES} \xrightarrow{k_2} \text{E} + \text{S} \]
\[ \text{ES} \xrightarrow{k_3} \text{E} + \text{P} \]

The rates of each of these reactions can then be given by
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

\[ v_1 = k_1[E][S] \]
\[ v_2 = k_2[ES] \]
\[ v_3 = k_3[ES] \]

Where \([X]\) denotes the concentration of the chemical species \(X\). The state space model can then be defined by combining these rates according to whether a species is destroyed or created in the corresponding reaction. For this example then,

\[ \dot{[E]} = -v_1 + v_2 + v_3 = -k_1[E][S] + k_2[ES] + k_3[ES] \]
\[ \dot{[S]} = -v_1 + v_2 = -k_1[E][S] + k_2[ES] \]
\[ \dot{[ES]} = v_1 - v_2 - v_3 = k_1[E][S] - k_2[ES] - k_3[ES] \]
\[ \dot{[P]} = v_3 = k_3[ES] \]

While the Law of Mass Action serves as the initial assumption for deterministic modeling, the Michaelis-Menten approximation is frequently used to model enzymatic kinetics. This approximation begins with the Law of Mass Action, but utilizes a few additional assumptions in order to reduce the model’s dimension.

Firstly, a quasi-steady state approximation is used. The concentrations of the intermediate complexes ([ES] in the above example) are assumed to equillibrate much faster than those of the substrate and product. This implies a time derivative of zero

\[ \dot{[ES]} = k_1[E][S] - k_2[ES] - k_3[ES] = 0 \]
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

which allows us to write

\[ [ES] = \frac{k_1[E][S]}{k_2 + k_3} \]

Then, defining a Michaelis-Menten constant

\[ K_m = \frac{k_2 + k_3}{k_1} \]

We can write

\[ [ES] = \frac{[E][S]}{K_m} \]

One can then assume a conservation law, such that the total amount of enzyme remains constant. Then

\[ [E] = [E_{total}] - [ES] \]

And the concentration of the complex can be rewritten as

\[ [ES] = \frac{([E_{total}] - [ES])[S]}{K_m} \]

Which can be rearranged as

\[ [ES] = [E_{total}] \frac{1}{1 + \frac{K_m}{[S]}} \]
Finally, the rate of change of the product can then be written as

$$[\dot{P}] = k_3[ES] = k_2[E_{total}] \frac{[S]}{K_m + [S]} = V_{max} \frac{[S]}{K_m + [S]}$$

Where $V_{max}$ is the maximum reaction rate and is determined solely by the total amount of available enzyme. This underscores the third assumption of Michaelis-Menten kinetics: that the concentration of the substrate is much more than the available enzyme (that the enzyme is the limiting reactant). In theory, the parameters $K_m$ and $V_{max}$ can be measured directly from carefully controlled in vitro experiments, making Michaelis-Menten models historically popular and important.

Systems of ordinary differential equations based on Mass Action and/or Michaelis-Menten kinetics thus have very specific structures. If we separate the kinetics derived from the two sets of assumptions, we can define the system as

$$\Sigma : \begin{cases} \dot{x} = f_1(x, k_{MM}) + f_2(x, k_{ma}) + g(x, k_{ip})u \\ y = h(x) \end{cases}$$

Where $x \in X \subseteq \mathbb{R}^n$ is a vector of system states (taken to be concentrations of biochemicals), $u \in U \subseteq \mathbb{R}^l$ is the vector of input concentrations (taken to be ligands or upstream components), $y \in Y \subseteq \mathbb{R}^m$ is the vector of output measures (taken to be some function of measurable state variables). Here $h(.)$ is the output function. In the case of immunoblotting for example, $h(.)$ is a function which converts the concentrations (or combinations of concentrations) of measurable state variables into
intensity measures. $f_1(.)$ and $f_2(.)$ are functions of Michaelis-Menten and Mass Action reaction rates (respectively) and are parameterized by vectors of rate constants $k_{MM}$ and $k_{ma}$.

We can further specify this system more precisely by defining the stoichiometry matrix $S$, such that the stoichiometry of chemical species $i$ in reaction $j$, $s_{i,j}(i = 1, ..., n; j = 1, ..., p)$, takes an integer value equal to the number of molecules consumed (negative integer) or produced (positive integer) in that reaction. Then $S \subseteq \mathbb{Z}^{n \times p}$. We can then define vector functions for the reaction rates such that $v_{MM}$ corresponds to the rates derived from Michaelis-Menten approximations and $v_{ma}$ those from Mass Action. In this case,

$$f_1(x, k_{MM}) = S_{MM}v_{MM}(x, k_{MM})$$

$$f_2(x, k_{ma}) = S_{ma}v_{ma}(x, k_{ma})$$

The Michaelis-Menten reaction rates typically have the form mentioned earlier (or are modified slightly if cooperative/inhibitory binding is concerned, in which case Hill Kinetics are often modeled) and so each element of $v_{MM}$ will be of the form given earlier but cannot easily be put in a more compact notation. On the other hand, the special form of Mass Action kinetics allows us to specify the above expression a bit further. The vector of reaction rates can be written (as derived in [175])

$$v_{ma} = diag(k_{ma})exp(S_{in}^Tlog(x))$$

Where $S_{in} = \{s_{i,j} | s_{i,j} \in \mathbb{Z}^- \}$ with species $i$ participating as a reactant in reaction $j$. 
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

Similarly, we can define $S_{out} = \{s_{i,j} | s_{i,j} \in \mathbb{Z}^+\}$ where species $i$ is formed as a product in reaction $j$. Therefore $S = S_{in} + S_{out}$. Additionally, in the above expression, $\text{diag}(.)$ and $\exp(.)$ are element-wise operators. An individual reaction rate appears as an element in the vector $v_{ma}$ and can be formulated as

$$v_{ma,j} = k_j \prod x_{in,i,j}^{s_{i,j}} \forall j \in [1, ..., p]$$

Where $x_{in,i,j}$ is the chemical species $i$ participating as a reactant in reaction $j$.

As a final derivation, we can write the function

$$g(x, k_{ip}) = \text{diag}(k_{ip}) \exp(S_{ip}^T \log(x)))C$$

Where again, $\text{diag}(.)$ and $\exp(.)$ are element-wise operators and $C$ is the stoichiometry matrix for the input variables.

### 3.2.2 Hill Kinetics and Goldbeter-Koshland Functions

Beyond these two most common modeling conventions, other approximations are sometimes used. These require similar or additional assumptions to the common Mass Action framework.

Hill kinetics are used to represent cooperative binding of several molecules of
substrate to the same macromolecule. The rate equation is given by

\[ v = \frac{[S]^n}{K_h + [S]^n} \]

where \([S]\) is the concentration of substrate, \(n\) is the Hill coefficient describing the level of cooperativity, and \(k_h\) is the apparent dissociation constant derived from the appropriate Mass Action formulation.

The Goldbeter-Koshland function \[176\] represents the balance of a molecule between two forms (such as unphosphorylated and phosphorylated) which is controlled by opposing forces (such as kinases and phosphatases), both assumed to obey Michaelis-Menten kinetics. This function is used to model bistable switches between active and inactive forms of protein concentrations. The basic mechanism is represented diagrammatically by

\[
\begin{align*}
  A & \quad \xrightarrow{k_1} \quad X^* \\
  \downarrow & \\
X^* & \xrightleftharpoons[k_2]{k_1} X \\
  \uparrow & \\
  B &
\end{align*}
\]

Combining the rates of dephosphorylation \(r_1\) and phosphorylation \(r_2\) where

\[
  r_1 = \frac{k_1[A][X^*]}{K_{M1} + [X^*]}, \quad r_2 = \frac{k_2[B][X]}{K_{M2} + [X]}\]
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

We can arrive at the derivation

\[
\frac{dx}{dt} = \frac{r_1 - r_2}{J_1 + (1 - x)} = \frac{v_1(1 - x) - v_2x}{J_2 + x}
\]

where

\[
x = \frac{[X]}{[X]_0}, \ [X]_0 = [X] + [X]^*, \ v_1 = k_1[A], \ v_2 = k_2[B], \ J_1 = \frac{K_{M1}}{[X]_0}, \ J_2 = \frac{K_{M2}}{[X]_0}
\]

Goldbeter-Koshland kinetics also often assume that the dynamics of X are in quasi-steady state and so approximate

\[
\frac{dx}{dt} = \frac{r_1 - r_2}{J_1 + (1 - x)} = \frac{v_1(1 - x) - v_2x}{J_2 + x} = 0
\]

In this case, the steady state concentration is given by

\[
x = \frac{[X]}{[X]_0} = G(v_1, v_2, J_1, J_2) = \frac{2v_1J_2}{B + \sqrt{B^2 - 4(v_2 - v_1)v_1J_2}}
\]

where here

\[
B = v_2 - v_1 + J_1v_2 + J_2v_1
\]

This approximation is used in several models of the cell cycle, particularly those by John Tyson and Bela Novak \([2, 9]\). For example, the non-steady state dynamic equation is used to model Cdh1 activation by \([2]\). The resulting differential equation
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

is of the form

\[
\frac{dCdh1}{dt} = \frac{k_{ah1p}(1 - Cdh1)}{J_{ah1} + 1 - Cdh1}
\]

\[-(k_{ih11} + k_{ih12}CycA + k_{ih13}CycB + k_{ih13}CycE + k_{ih14}CycD)Cdh1\]

\[J_{ih1} + Cdh1\]

Additionally, the steady state approximation is used in [2] to represent E2F binding with RB

\[ [E2F : RB] = \frac{2[E2F_T][RB]}{[E2F_T] + [RB] + L - \sqrt{([E2F_T] + [RB] - L)^2 - 4[E2F_T][RB]} \]

\[ L = \frac{k_{26r}}{k_{26}} + \frac{k_{20}}{k_{26}}(\lambda_D[CycD] + \lambda_E[CycE] + \lambda_A[CycA] + \lambda_B[CycB]) \]

These assumptions blur the line between mechanistic modeling based on first principles and ad hoc, phenomenological models which approximate subsystem responses when few mechanistic details are known. In the examples given, the basic regulatory interactions and response is given, but the true mechanisms are likely more complicated. The underlying systems are likely bistable - for the case of E2F, this was shown quite convincingly by [24] - however the bistability may be created by more complex mechanisms than opposing rates of activation/inactivation as assumed by the Goldbeter-Koshland approximation. Such sources of bistability include multi-site phosphorylation of the substrate or positive-feedback loops, as was demonstrated through an exemplary combination of modeling and experimentation by James Ferrell.
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

in the case of Cdc25 and cyclin B/cdk1 interactions [177].

3.3 A Note on ODE’s

Chemical reactions were first modeled in vitro under controlled, well-mixed conditions (relatively dilute, pH buffered, homogeneous, and aqueous). On the contrary, the cytoplasm of a typical cell has a total macromolecular density of 50-400mg/ml, which is significantly higher than typical in vitro conditions of 1-10mg/ml, and leads to a much slower apparent translational diffusion speed (5-20 times slower than in saline solutions) [178]. With sufficient concentrations of reactants, in vitro reactions were observed to obey simple rate equations. This led to the Law of Mass Action and the Michaelis-Menten approximation just discussed. It has since been shown that, for an infinitely large system, Mass Action kinetics are themselves an approximation of the discrete, stochastic events described by the Chemical Master Equation (CME) [179].

The CME description conceptualizes the system as a Markov Jump Process, such that any change in the system state occurs discretely after a random waiting period, defined by the “propensities” for reactions to occur (themselves defined by the relative concentration of reactants in the simulated volume). Approximations to the CME such as the Two Moment Approximation (2MA) or Chemical Langevin Equation (CLE) are sometimes also used. Generally, however, the magnitude of stochastic fluctuations scales with $\frac{1}{\sqrt{N}}$ (where $N$ is the number of molecules in a given compart-
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

Thus, deterministic models are considered "good enough" for reactions having more than 100-1000 molecules (occurring in a typical cellular volume), and have been used quite successfully for models of eukaryotic signaling pathways. Single cell measurement techniques, however, have increased awareness and interest in stochastic modeling approaches. Applications of various approaches are discussed in more detail in Chapter 6.

On the whole, ODE approximations work quite well for most modeling purposes. They are intuitively understandable, easily constructed using the methods discussed in the previous sections, have robust numerical algorithms for calibration and simulation, are supported by a large body of mature analytical methods, and can be rigorously compared to experimental data. Unless spatial or stochastic effects are truly believed to be significant, ODE models are often sufficient.

3.4 Calibration

Having built a mathematical model, we are left with a large number of parameters defining the reaction kinetics. These parameters are often completely unknown, or only roughly estimated. The process of estimating parameters is known as calibration. Calibration is distinct from validation, which involves testing these estimated values on new observations. Calibration is generally cast as an optimization problem, comprising several steps:
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

1. Define the function to be minimized (objective/error function)

2. Define the set of parameters to be optimized

3. Define feasible estimates/ranges of the parameter values

4. Choose and run an appropriate optimization algorithm (often repeatedly) until an adequate solution is found

5. Reconsider the model formulation if unable to calibrate adequately

Computational models integrate a hierarchy of assumptions [29]: the first set are the most basic, the “wiring diagram” which defines the hypothesized interactions of the system; secondly, the mathematical equations which translate the diagram into precise dynamics (here the assumptions might include the choice to use Mass Action or Michaelis-Menten representations, for example); finally, exact values are assigned to represent the initial system state and the kinetic rate constants. The standard calibration routine outlined above seeks to adjust the last set of parameter assumptions to fit the data. If possible, this would confirm the plausibility of the hypothesized interactions. If calibration is not possible, one needs to move up the hierarchy and address the mathematical assumptions used to represent the hypothesized biology. If that is not helpful, one must reconsider the original definition of the system – perhaps a necessary system element or interaction is missing, or hypothesized interactions preclude fitting the data.
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

3.4.1 Math Formulation

The basic goal of calibration is to minimize an objective function by adjusting a set of parameters. The objective function is most commonly defined as minimizing the squared error between model predictions and biological observations.

\[ \text{obj}(\text{parameters}) = (\text{model} - \text{data})^2 \]

The parameters here are the rate constants and initial system state values of the computational model. More formally, we can use the notation derived in the Section 3.2, and consider the simulation output to be

\[ y_{\text{sim}} = y = h(x) \]

where \( h(x) \) is often a scaled combination of model state variables. For instance, an antibody may not discriminate between inactive/active forms, or bound/free forms of a protein and the proper comparison would thus be the measured data to the scaled sum of a subset of modeled protein states.

Given experimental data (sampled at \( N \) time points)

\[ y_{\text{exp}}(t_n) \quad 1 \leq t_n \leq N \]
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

At each data point \( t_n \), we evaluate the least-squares difference (\( \chi^2 \)) function:

\[
obj_{tn}(k) = \frac{1}{2\sigma^2} [y_{sim}(t_n) - y_{exp}(t_n)]^2
\]

The complete objective function is usually taken to be the sum of the \( \chi^2 \) errors at each sampled time point:

\[
obj(k) = \sum_{t_n=1}^{N} \frac{1}{2\sigma^2} [y_{sim}(t_n) - y_{exp}(t_n)]^2
\]

Or, for multiple output measurements:

\[
obj(k) = \sum_{m=1}^{M} \sum_{n=1}^{N} (y_{sim}^m(t_n) - y_{exp}^m(t_n))^T Q^m(t_n) (y_{sim}^m(t_n) - y_{exp}^m(t_n))
\]

Where \( Q \) is a matrix of weights for the reliability of each measurement. In the \( \chi^2 \) formulation, \( Q = \frac{1}{2\sigma^2} \) which weights better constrained data less than "noisier" measurements.

3.4.2 Selecting a Set of Parameters to Optimize

Ideally one would like to adjust all of the parameters of a model. Sometimes, however, this is not possible or necessary. Estimating too many parameters provides more degrees of freedom, and increases the likelihood an algorithm will converge to an unsatisfactory local optima. Generally, while a minimum number of parameters is
required to adequately fit a model to data, algorithm performance deteriorates significantly as the number of parameters is increased [181]. This choice of parameters is made with respect to two properties: parameter sensitivity and identifiability. Parameter sensitivity analysis is used to isolate the subset of parameters which most effect the behavior of the observable system states. Generally, only a small subset of the system parameters will be found to be highly significant, while a large number will have little to no effect. If at all possible, one should try to experimentally define the most significant parameters, or at least include them in the optimization attempt, as they are most crucial for fitting experimental data. Having defined a candidate subset of parameters to calibrate, a priori identifiability analysis can sometimes identify parameters which are not uniquely determinable from the available data. Functional relationships often exist between system parameters such that changes in one parameter produce effects that are either mirrored or exactly compensated by changes in another [182]. Such relationships complicate parameter estimation through the introduction of redundant degrees of freedom. One can thus reduce the dimension of the parameter subset to be optimized by only including one parameter of a functional subset and fixing the others to constant values.

Different parameters have different effects on different state variables. A given state variable is more strongly affected by changes in a given parameter than others. The simplest and most widespread sensitivity measure is gradient based, and defined by the normalized metric:
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

\[ s_{ij}(t) = \left. \frac{\partial y_i(t)/y_i(t)}{\partial k_j/k_j} \right|_{k^*} \]

where, for the generic ODE system given earlier, \( y_i \) represent the measurable outputs and \( k_j \) the parameters of the model. This can be calculated analytically for simple systems. For larger systems, the sensitivities can be estimated numerically through solving \( \frac{\partial}{\partial t} \frac{\partial y_i/y_i}{\partial k_j/k_j} \bigg|_{k^*} \) simultaneously with the original system [183]. For dynamic systems, this sensitivity index is time-dependent. Computation of these values provides information on the timing of a parameter’s effects on a given state variable. To compare or rank parameters in terms of their total effects then, it is necessary to compute a scalar value, often taken to be the integral of the time dependent sensitivity.

\[ s_{ij} = 1/T \int_0^T dt |s_{ij}(t)| \]

The above sensitivity metric is often called a “local sensitivity” because it is evaluated at a chosen set of nominal parameter values \( k^* \). Parameter sensitivities can be highly variable depending on the nominal parameter set \( k^* \). To capture a sense of a parameter’s overall importance on the output, one also seeks a “global sensitivity” measure. The most straightforward approach is to calculate the average of local sensitivities from multiple, randomly selected parameter sets. More complex, variance-based methods (e.g. Sobol’s [184]) have also been proposed wherein the variance of a model output is decomposed into partial variances determined by parameter
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

variances. A sensitivity metric is then computed as the ratio of the partial variance to the total output variance. Parameters with low sensitivity have little effect on measurable outputs. Practically, this means that such parameters will be difficult to constrain with the available data. Luckily, a low sensitivity also means that their chosen value should not have a significant effect on the behavior of the measurable outputs.

Having calculated time-dependent sensitivities, one can also compute a correlation matrix which summarizes the covariances between them. Sensitivities which are highly correlated indicate parameters which have similar impact on the output of interest. Parameters may have similar or compensatory effects on a given output, such that they cannot be uniquely estimated from the measured data. This provides a type of a priori identifiability analysis which may prevent needlessly increasing the degrees of freedom during calibration. Significantly, however, this is a local identifiability metric, as it depends on calculating the time-dependent sensitivities with respect to defined parameter values $k^*$. For a given observation function

$$y(t, k, u) = h(x(t, k, u))$$

Global, structural identifiability requires that distinct parameter sets generate distinct
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

observations:

\[ \forall k_1, k_2 \in \mathbb{R}^p, \ k_1 \neq k_2 \Rightarrow \exists t \ such \ that \]
\[ h(x(t, k_1, u)) \neq h(x(t, k_2, u)) \]

The correlation matrix of local parameter sensitivities approaches an estimate of the degree of local identifiability, defined in a neighborhood \( \epsilon \) of the “true” parameters \( k_0 \): 

\[ \forall k_1, k_2 \in \{ k \in \mathbb{R}^p \mid \| k - k_0 \| < \epsilon \}, \ k_1 \neq k_2 \Rightarrow \exists t \ such \ that \ h(x(t, k_1, u)) \neq h(x(t, k_2, u)) \]

Unfortunately, a priori global sensitivity analysis is still an open problem and analytic, a priori local methods are not broadly applicable to large, complex systems [185].

In practice however, if one is able to deduce functional relationships existing between pairs or sets of parameters, one can fix all but the one which is to be estimated. By ranking parameters by their sensitivity indices, with respect to measurable outputs, and then cross-referencing these against an estimate of the sensitivity covariances, it is hopefully possible to assemble a subset of the total parameters which are highly significant to the output of the model, and which can be uniquely determined from the available data. This provides a starting set of parameters to attempt model calibration.
3.4.3 Biophysically Reasonable Parameter Values

If estimates are available for biophysically reasonable parameter values, they can provide starting values for calibration efforts, or allow estimating feasible bounds on possible values. For the simple enzymatic equation given previously, a theoretical diffusion-limited rate constant for small molecules has been estimated as \( k_1 \sim 10^8 - 10^9 M^{-1} sec^{-1} \). However, for the case of enzymes binding substrates, the limited range of collision geometries which provide access to active sites significantly lowers this speed to perhaps \( k_f \sim 10^5 - 10^6 M^{-1} sec^{-1} \). If conformational changes in the enzyme are required, e.g. during the binding of imatinib (Gleevec) to Bcr-ABL kinase, this value can be even lower [180]. The unbinding rates are determined by dissociation enthalpies and entropies. A reasonable value has been suggested to be \( k_d \sim 10^{-1} sec^{-1} \) [180]. Catalytic rates, e.g. for the phosphorylation of peptide substrates by receptor kinases are estimated to be in the range \( k_{cat} \sim 10^{-1} - 10^{-2} sec^{-1} \) [79, 186]. These estimates provide first guesses for calibrating a quantitative model. Generally, one then defines a range of at least two orders of magnitude around these values as the feasible set to explore through numerical optimization. Unfortunately, however, these estimates are not always useful, as the concentration scales of involved proteins are often completely unknown and the corresponding model is in terms of “arbitrary units” of magnitude.
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

3.4.4 Optimization Algorithms

Biochemical models are widely understood to be non-identifiable, both practically (given reasonable biological data sets) and structurally (functional dependencies exist between parameters such that each cannot be uniquely determined [182]). Non-identifiability, and other structural issues, generally result in a highly multi-modal objective function landscape (consisting of many local minima surrounded by less optimal solutions). Calibration therefore requires broad and repeated searches through the parameter space [79,186,187]. State of the art algorithms generally couple global stochastic searches to cover space and leap local minima with deterministic local methods to refine the broad search once the neighborhood of the optimum has been located.

The simplest case is a multi-start local search, wherein one uses a deterministic local method (such as Gauss-Newton), but repeatedly runs the algorithm from different, randomly started, initial points. The hope is that one can randomly begin the simulation in the neighborhood of the optimum and then simply refine the solution. This method was used by [187], but admittedly at great computational expense. Estimating 159 of the model’s 233 parameters (chosen due to limited knowledge or high sensitivity), the authors report an average of 192 computational hours per solution set found. More sophisticated methods exist for stochastically searching the parameter space before commencing the local method.

Recent publications [79,186] have utilized more complex approaches such as Ge-
CHAPTER 3. COMPUTATIONAL MODELING METHODOLOGY

Genetic Algorithms and Simulated Annealing. Similarly, a comparative study of global optimization methods showed Stochastic Ranking Evolutionary Strategy (SRES) to have the best performance [188]. Genetic algorithms maintain a population of solutions, mixing the best fits (with slight added variation) after each iteration to improve the fitness of the population. A benefit of this approach is that it is highly parallelizable and thus considerable computational improvement can be gained by deploying the algorithm on a cluster of processors.

Simulated annealing is inspired by metal cooling processes. At each iteration, the current parameter guess is varied according to a probabilistic function which depends both on whether the move improves the solution, but also a time-varying “cooling temperature” parameter (generally exponentially or logarithmically decaying). This routine can also be time consuming. Chen et al. [186] estimated 75 of 229 parameters and report that finding one good fit required 100 simulated annealing runs and 24 hours on a 100-node cluster computer (estimated to be \(\sim 2000\) CPU hours).

To test the usefulness of these methods for calibrating cell cycle models, we can use a published model to create ideal, noiseless, “artificial” data and attempt to re-estimate the parameters of the model which produced it. This gives an indication of the ability of the algorithm to calibrate models of similar dimension and scope, assuming the model is able to fit available experimental data without modification to the mathematics or underlying system structure. An exercise using the 2004 mammalian cell cycle model published by Bela Novak and John Tyson [2] indicates that at least
one of these algorithms, a hybrid simulated annealing routine (here coupled to a final, adaptive mesh routine), is able to easily and routinely calibrate a cell cycle model. As an example, estimating the top 9 most sensitive parameters, when initialized from a random starting point, resulted in convergence to adequate solution after just a few optimization routines. An example convergence is shown in Figure 3.2, illustrating the stochastic nature of the search. The re-calibrated outputs are shown in Figure 3.3. The solution is not optimal, but an adequate approximation to the data. Interestingly, it converged on a solution set vastly different from the original parameter values (Table 3.4.4). This is likely because the landscape of the objective function is largely flat in multiple dimensions and lacks well-defined minima [186, 189]. The algorithm therefore converges to locally, but not necessarily globally optimal solutions. It is also likely that the system is non-identifiable and functional relationships exist between sets of parameters which yield identical outputs [182].

Figure 3.2: Example convergence trajectory of a hybrid simulated annealing optimization routine, in the calibration of an example cell cycle model [2].
Figure 3.3: Example re-calibration of the top 9 most sensitive parameters of a cell cycle model [2] using a hybrid simulated annealing routine.
### 3.4.5 Model Adjustment

Numerical rate constants represent one level of a hierarchy of modeling assumptions. If calibration is unsuccessful, it may be necessary to re-evaluate the underlying structure of the model (the assumed interactions between system elements) or the mathematical formalism (how the presumed interactions are encoded). Goldbeter-Koshland, Michaelis-Menten, or even Mass Action may not be appropriate approximations of the true biology. More fundamentally, the assumed interactions of system elements may be incorrect or previously un-modeled dynamics may actually be significant. This aspect of modeling is largely an art and cuts to the heart of the scientific method. As models are, fundamentally, organized hypotheses, the failure of a model to represent reality requires a re-assessment of one’s knowledge or the precision with which that knowledge is mathematically represented.

---

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Published Value</th>
<th>Re-estimated Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2Ft</td>
<td>5</td>
<td>6.214</td>
</tr>
<tr>
<td>J17</td>
<td>0.3</td>
<td>16.061</td>
</tr>
<tr>
<td>k18</td>
<td>10</td>
<td>623.207</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>1</td>
<td>0.989</td>
</tr>
<tr>
<td>k31</td>
<td>0.7</td>
<td>3.789</td>
</tr>
<tr>
<td>k32</td>
<td>1.8</td>
<td>28.598</td>
</tr>
<tr>
<td>k33</td>
<td>0.05</td>
<td>0.733</td>
</tr>
<tr>
<td>Rbt</td>
<td>10</td>
<td>6.462</td>
</tr>
<tr>
<td>k34</td>
<td>0.05</td>
<td>2.307</td>
</tr>
</tbody>
</table>

Table 3.2: Comparison of published and re-estimated parameters from an example calibration of a cell cycle model. The most sensitive parameters of the 2004 Novak and Tyson model [2] were re-calibrated using ideal, “artificial data.”
Chapter 4

Measuring Cell Cycle Regulatory Molecules

It is inherently difficult to obtain precise quantitative data of cell cycle regulatory proteins. Cultured cells are naturally asynchronous and a naive measurement approach would result in averaging the values over all the phases present in a given cell population. Measurement is therefore generally accomplished by physical or chemical synchronization, timed sampling, and immunoblotting or immunohistochemistry and microscopy. For mammalian cells, synchronization by physical or chemical methods is always imperfect and cells released from arrested states progress through the remainder of the cell cycle with a broad, and altered distribution of transit times. Imperfect synchronization averages both synchronous variability (natural variation in the expression of a molecule between cells at the same point of the cell cycle) and
asynchronous variability (variation between cells at different points of the cell cycle), resulting in inaccurate estimates of temporal expression profiles [190]. However, the information for the programmed expression of these same molecules is contained in a single, randomly sampled, asynchronous population.

4.1 Asynchronous Data Extraction

Asynchronous cells are distributed at each state within the data space of the programmed expression of said molecule. Therefore, because the cell cycle is a closed loop, single cell measurement data (e.g., flow cytometry) contain a sampled version of the dynamic expression profile of the measured molecules [6,7,191]. In a cycling population, the frequency of cells in each phase or state is proportional to the time spent in that state. In multiparameter cell cycle flow cytometry, if correlated parameters are chosen to unambiguously isolate a continuum along an expression profile, that expression profile can be unambiguously derived. Guided by heuristics assembled from canonical cell-cycle knowledge, we have derived a (computer-assisted) methodology to extract the embedded dynamic profiles of cell-cycle proteins from statically sampled, multiparameter cytometry data. This approach is illustrated in previous publications [5–7, 191] and summarized in Figures 4.2, 4.3, and 4.4. Specifics of the approach are discussed in the following subsections.

For the data set explored here, K562 cells (an immortalized myelogenous leukemia
cell line) were collected from an exponentially growing culture and fixed and stained with antibodies for cyclins A2 (cycA-PE) and B1 (cycB-A647), phospho-S10-histone H3 (PHH3) and DAPI.

4.1.1 Data Pre-processing

There is significant pre-processing required, prior to profile extraction. Firstly, spectral overlap was corrected between the secondary antibody Alex Fluor 488 and R-Phycoerythrin. Next, background fluorescence was subtracted using G1 phase as a guide, because neither cyclin A2 nor cyclin B1 are expressed in early G1 phase. This reduces the fluorescence of early G1 cyclin A2 or B1-negative cells to near zero. Lastly, outliers (e.g. S phase cells negative for cyclin A2) and endoreduplicated cells (which have more than 2 sets of chromosomes) were removed by creating a Boolean “gate” on a bivariate plot of interphase cyclin A2 versus DNA. This process is represented in Figure 4.1 and described in more detail elsewhere [4].

4.1.2 Segmentation of Data Continuum

The resulting data set consists of a sample of asynchronous cells passing through a closed trajectory of expression levels. The goal of the analysis is to segment these cells into non-overlapping contiguous subsets, in order to estimate the relative time spent by an average cell traversing each bounded region. Generically, the procedure
Figure 4.1: Flow cytometry data pre-processing steps include (a) Fluorescence compensation (via application of a bias) (b) Removal of doublets (through gating out cells) (c) Minimizing the effects of non-specific binding (application of a bias) [4]. These steps were completed in WinList (from Verity Software).

requires the isolation of an unambiguous trajectory of the measured molecules. This is fundamentally related to the monotonicity of dynamic expression profiles, and is especially difficult for oscillating proteins. For any region of the data space over which a given protein oscillates, another monotonically varying molecule must be measured. This allows separation of the rise and fall of the primary protein. Consider, for example, the measurement of only DNA content and PHH3 expression shown in Figure 4.1.A. DNA content is constant during G1 phase, increasing during S phase, and then constant again during G2 and M. PHH3 remains nearly constant throughout interphase, peaks abruptly at the start of mitosis, and then falls precipitously as cells exit mitosis. Therefore, DNA content remains constant during the period over which PHH3 rises and then falls. On a two dimensional scatter plot then, it is impossible to separate out cells entering mitosis from those leaving. Decomposing this movement
CHAPTER 4. MEASURING CELL CYCLE REGULATORY MOLECULES

requires measurement of an additional molecule. In this case, cyclin A2 levels provide a useful guide [191]: cyclin A2 rises as PHH3 does, falls as PHH3 remains constant, and then remains constantly absent as PHH3 falls. This allows isolation of a non-overlapping, unidirectional trajectory through the data space.

In practice, in order to segment the K562 data into contiguous regions, ordered unidirectionally from G1 to cytokinesis, we first use PHH3 expression to separate interphase from mitotic cells. Interphase cells range from low PHH3 intensity and one genome to approximately twice that in G2, and mitotic cells have higher PHH3 intensity and 2 genomes. Next, a plot of DNA content versus cyclin A2 is used to separate G1, S, and G2 phase cells. S phase manifests as a continuum of expression levels rising from 1 genome (DNA content) and absent cyclin A2 to 2 genomes (DNA content) and higher levels of cyclin A2. This phase was segmented by defining contiguous regions with defining borders set orthogonal to the slope of greatest variation (the direction of the underlying expression profile an “average” cell would traverse). G1 and G2 phases do not contain bivariate variation (at least for the markers measured) and instead require analysis of single parameter histograms for cyclins A2 and B1.

The analysis of G1 and G2 phase cells is completed differently for cyclins A2 and B1. G1 cells lack measurable cyclin A2 cells (expression is equal to background fluorescence) and DNA content is also constant (G1 occurring before the Synthesis phase). A2 expression is therefore defined to be zero for the length of G1.
CHAPTER 4. MEASURING CELL CYCLE REGULATORY MOLECULES

contrary, cyclin B1 expression increases in late G1 and thus the data contain useful information during G1 phase for this marker, even though DNA content does not provide an informative bivariate plot. To resolve the increasing expression of cyclin B1 through late G1, before the bivariate segmentation of S phase data, we use multi-Gaussian modeling to fit the sum of several subpopulations of increasing median expression to the G1 cyclin B1 distribution. Multi-Gaussian modeling is also used to resolve the expression of G2 cells. For cyclin A2, the distribution during prophase was used as estimates for the coefficients of variation of the Gaussian sub-components. Similar analysis is conducted to resolve G2 cyclin B1 expression.

Mitotic cells were segmented using similar, but more complex, logic as S phase. The simplest case uses a bivariate plot to segment the cells as they first degrade cyclin A2 and then cyclin B1. A more complex analysis, illustrated in [7] and used in Section 5.3, involves the segmentation of cells first when PHH3 rises as cells enter mitosis, then as cyclin A2 is degraded against constant PHH3 levels, followed by segmentation of cyclin B degradation when viewed against cyclin A2 and finally of falling PHH3 levels (viewed against constant cyclin B1) as cells re-enter G1.

4.1.3 Cumulative Frequency Plot

The result of the discussed segmentation is a collection of contiguous subsets of the data. Using the logic mentioned earlier, the time an “average” cell takes to traverse a given segment of the expression trajectory is proportional to the number of
Figure 4.2: Derivation of the cyclin A2 expression profile. 

A. Cells are first partitioned into interphase and mitotic fractions. 

B. Mitotic cells are then further subdivided into arbitrary partitions. 

C. A Gaussian distribution is fit to determine the standard deviation of cells in prophase. 

D. Partitioning along the cyclin A2 trajectory, G1 cells are defined by the absence of cyclin A, R5. G2 cells are defined by R6. S phase cells have intermediate levels of Cyclin A and are partitioned orthogonal to the direction of greatest variation. 

E. A series of Gaussian distributions is fit to describe the movement of cells through G2 phase. The Gaussians are assumed to have a standard deviation roughly equal to that of prophase. 

asynchronous cells sampled in that data space. An estimate of the median expression profiles can therefore be extracted by plotting the median expression values (the median data value each subset) against the cumulative frequency of cells (ranging from 0% to 100% of the sample population). A median expression value is plotted in the center of a frequency length defined by the relative number of cells in that subphase.

### 4.1.4 Frequency to Time Conversion

The result of this analysis is an estimate of expression levels as a function of cell frequency within the sampled population. Because one mitotic cell gives rise to
Figure 4.4: Derivation of the cyclin B2 expression profile. A. G1 and G2 cells are separated from S and M. B. The trajectory of the G2 cells are not further resolvable. C. Gaussian subpopulations are fit to a one dimensional histogram of cyclin B1 levels in G2. D. G1 cells do not show significant variation which would allow two dimensional gating. E. Guassians are fit to approximate movement through G1. F. Median cyclin B1 expression as a function of cumulative cell frequency. See [5] [6] [7] for full details.
two daughters, it is necessary to introduce an additional transformation to convert the expression profiles to the time domain. This transformation is well established [192, 193] and relies on the observation that an asynchronous, exponentially growing cell population has an exponential distribution of cell ages. To convert between cell frequency and the time domain (in this case a function of “total cell cycle time”) one simply divides the $N$ frequency values ($freq(i), i = 1, ..., N$) by an exponential value between 2 and 1, i.e.

$$T = \log_e(2)$$

$$weight(i) = 2e^{-T\cdot freq(i)}$$

$$time(i) = \frac{freq(i)}{weight(i)}$$

The resulting expression profiles are now functions of total cell cycle time (taken to be 0-1 or 0-100%). The expression profiles of cyclin A2, cyclin B1, and PHH3 are shown in Figure 4.5. Here the relative magnitudes of the expression profiles are normalized for illustration purposes.

### 4.1.5 Further Scale Conversions

As mentioned earlier, the K562 cell line studied here typically has a cell cycle time of around 18 hours, and so a truly quantitative time scale would range from 0 to 18 hours (or a more precise estimate obtained through additional experimentation). A second point regarding data scaling is that the basic approach mentioned so far produces relative expression data (in terms of arbitrary units of magnitude)
that provides correlated timing information, but does not allow comparison of magnitudes. While not used in this dissertation, we recently demonstrated the possibility to convert the arbitrary expression magnitudes to a relative scale. Then, if a purified protein is available, it is technically possible to convert one arbitrary magnitude to absolute molecular units [5,194]. Because each expression profile is scaled to relative magnitudes, it is then possible to scale all measured proteins to absolute molecular units.

To convert each marker on the same scale, one needs to conduct several fixation/staining experiments using identical samples of an exponential cell population. In [7], two separate samples were collected and stained indirectly for either cyclin A2 or B1 with the same unconjugated antibodies, the same secondary, and propidium (to quantify DNA content). The first sample was termed “multi-color” and the others “single color.” Then, to convert the multi-color data to the same scale, we used DNA content histograms to map the other markers in the same relative data space in each file. We then created seven gates, centered in S phase. Plotting the median values of the single color data versus multi-color showed a linear relationship between expression values and linear regression allowed estimation of the exact conversion equation to convert multi-color to single color scales. Because the single color experiments were conducted using the same secondary antibody, the multi-color data are now on the same relative scale.

As mentioned, if a purified protein is available for one of the markers, it is now
CHAPTER 4. MEASURING CELL CYCLE REGULATORY MOLECULES

possible to convert all of the multi-color sample to exact molecular units. This of
course assumes that the difference between recognition of the monoclonal marker
antibodies by the secondary polyclonal antibody is negligible, that the differences in
affinity for each epitope by the two monoclonal antibodies are negligible, that epitope
exposures are approximately equal (that they are not masked in a biased manner), and
that the dissociation rates of the monoclonal antibodies are the same. The application
of these approaches provides an extensible measurement system that should provide
time-correlated, truly quantitative, estimates of the dynamic expression profiles of as
many cell cycle regulatory molecules that are currently measurable using commercial
antibodies.

4.2 Resulting Expression Profiles

Figure 4.5 presents the expression profiles of cyclin A2, cyclin B1, and PHH3 in
the K562 cell line. Cyclin A2 synthesis begins at or near the start of S phase and
continues into M, wherein it abruptly decreases. The data show a two-phase increase
– approximately linear through S phase, followed by a rapid increase in expression at
the start of G2 phase. The pattern of cyclin B1 expression is similar. However, its
expression begins earlier than cyclin A2; the primary increase in expression is non-
linear and the secondary rate of increase is more rapid than cyclin A2. Cyclin B1
decays later in mitosis than cyclin A2. PHH3 is a marker of mitotic cells, increasing
very early and remaining elevated through to cytokinesis. This is evident in Figure 4.5.

Histone H3 is likely phosphorylated at serine 10 by Aurora kinase B during mitosis [195–197], and the link between Cdk1 activation and Aurora kinase B activation is presently unknown. Since cyclin B1/cdk1 is also activated at the beginning of mitosis, the onset of cyclin B1/cdk1 activity should be highly correlated with PHH3 expression during early mitosis. While we do not measure cyclin B1/cdk1 activity directly, we use PHH3 as a proxy for the timing of cyclin B1/cdk1 activity in mathematical models.

Having obtained these expression profiles, our goal was to determine how closely they matched published models and rate constants and, if the fits were not close, what changes were needed in existing models to make them fit the data – in essence, we were attempting to calibrate the published models.
Figure 4.5: (a) Normalized expression profiles of cyclin A2, cyclin B1, and PHH3. (b) Magnified view of the final 2% of the cell cycle. K562 cells were stained, measured, and analyzed as described in Materials and Methods.
Chapter 5

An Improved Computational Model of the Mammalian Cell Cycle

As reviewed in Chapter 2, cell cycle modeling has a long history. It is characterized by a gradual transition from purely statistical, phenomenological models of growing cell populations, to the incorporation of basic molecular modules of cell cycle regulation, as such information became available. Over several decades, more and more details have been added to these models to increase their complexity and explanatory/predictive power. Most models are unsupported by quantitative biological data. While these models have been useful in describing and validating hypothesized mechanisms based on biological research, especially on feedback mechanisms, hystere-
sition, and bifurcation [19,147,198] in cell cycle control, there have been few attempts to fit these models to quantitative measurements. Exceptions include the few discussed in Chapter 2.

In the work presented herein, our goals were to examine how well previously published models matched the expression profiles derived using the methodology previously discussed, calibrate a published model if possible, or identify necessary changes in existing models to better fit the data. We found that while published models capture broad trends in the expression profiles of cyclins A and B, modifications were required to refine their dynamics and illustrate the importance of modeling in testing hypotheses and explaining biological data.

5.1 We could not fit previous computational models to our data

We tested two existing models. Both models captured canonical cell cycle knowledge. They incorporated growth factor-induced activation of cyclin D, which phosphorylated Rb, activated E2F and resulted in the synthesis of cyclins A and B. Cyclin E initiated the deactivation of APC/Cdh1 and allowed cyclins A and B to accumulate. Cyclin B activated APC/Cdc20, which in turn degraded cyclins A and B, completing mitosis and restarting the cell cycle control system. Both models also incorporated the antagonistic relationship between p27 and cyclins E, A, and B. However, be-
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

Beyond these basics, there were key differences between these models and each captured different aspects of the K562 expression profiles, as shown in Figures 5.2 and 5.4.

5.1.1 Published models compared to the K562 data

The most recent comprehensive model [8] of the mammalian cell cycle appears to be an updated version of that used to investigate restriction point control [2] (comprising 23 ODE’s and 74 kinetic rate constants). It therefore incorporates detailed mechanisms and equations for cyclin D activation, E2F interactions with Rb, and the role of cyclin E. Notably absent from the model were Wee1 and Cdc25, or any mechanisms to account for at least the net rates of synthesis for cyclins A and B during interphase. Figure 5.1 presents a diagram of this model’s interactions and Figure 5.2 a comparison of the published model outputs with K562 derived expression profiles. The chosen antibodies do not discriminate, and so total protein amounts (bound and free, inactive and active) from the model are compared to K562 measurements. As discussed in Chapter 4, PHH3 is used a proxy for the timing of cyclin B/cdk1 activity. The model captured basic trends in the expression of cyclins A and B, but the model outputs did not display the two distinct biphasic rates of synthesis evident in the K562 data. Furthermore, because the model lacked Wee1/Myt1 and Cdc25, it did not capture the delayed activation of Cyclin B and abrupt onset of mitosis demonstrated by the PHH3 expression profile.
Figure 5.1: Schematic of the Conradie et al. 2010 model [8]. Solid lines indicate chemical reactions and dashed lines regulatory effects.
Figure 5.2: Comparison of Conradie et al. [8] model outputs (solid lines) with K562 data (markers). (a, b, c) Total simulated levels (bound and free, inactive and active) of cyclins A and B, and cyclin B/cdk1 activity are compared to the measured expression profiles of cyclins A2 and B1, and PHH3. Simulations were carried out using the published protocols and kinetic rate constants.
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

The second model we examined [9] is the mammalian implementation of a “generic cell cycle” model, comprising 13 ODEs with 66 kinetic rate constants. This model did not include any upstream signaling, and limited dynamics of cyclin D (modeled as an initial concentration which grows exponentially) and E2F activation (approximated with an ultrasensitive Goldbeter-Koshland function [176]). The model included more comprehensive dynamics of the core downstream modules including cyclins A and B, which is our focus, and incorporated Wee1 and Cdc25 regulation of cyclin B/cdk1. Additionally, a G2 transcription factor was modeled for cyclin B, to implement secondary, auto-catalytic transcriptional control. These interactions and a comparison of the published model outputs to our data are shown in Figures 5.3 and 5.4. Cyclin B dynamics were captured fairly well with the model. Cyclin A dynamics started immediately and were only transcribed at one characteristic rate throughout the cycle. The K562 data, however, demonstrated a delay in cyclin A expression for the first 23% of the total cell cycle time and showed a distinct change in the expression rate during G2.
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

Figure 5.3: Schematic of the Csikasz-Nagy et al. 2006 model [9]. Solid lines indicate chemical reactions and dashed lines regulatory effects.
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

Figure 5.4: Comparison of Csikasz-Nagy et al. [9] model outputs (solid lines) with K562 data (markers). (a, b, c) Total simulated levels (bound and free, inactive and active) of cyclins A and B, and cyclin B/cdk1 activity are compared to the measured expression profiles of cyclins A2 and B1, and PHH3. Simulations were carried out using the published protocols and kinetic rate constants.
Although the published model state variable outputs do not fit our expression profiles well, both models captured canonical cell cycle dynamics. The Csikasz-Nagy model comes closest, and incorporates Wee1 and Cdc25 regulation of cyclin B/cdk1 to delay activation until M phase - essential for fitting the PHH3 data. We therefore selected this model as a starting point for calibration efforts. To better fit this model to our data, we used the ideas of Chen et al. [29] and worked backwards through the hierarchy of model assumptions: numerical rate constants, mathematical approximations, and the underlying biological network. However, without structural changes, we were unable to satisfactorily calibrate this model to our data.

5.1.2 Adjusting numerical rate constants

Manual and automated routines failed to significantly improve the fit between the Csikasz-Nagy et al. model [9] and our data. To attempt automated calibration routines, the entire parameter set was evaluated, as well as subsets defined by the top 10% and 30% of the parameters, as defined by a global sensitivity analysis.

5.1.3 Defining Goodness of Fit

In order to compare model outputs to data, an error is calculated at each available time point (see Chapter 3 for more details on the construction of this “objective function”). As discussed earlier, the result of the data analysis process is an estimate
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

of expression as a function of relative “cell cycle time”. Time here is relative to the total length of the cell cycle. K562 cells have a cell cycle time of approximately 18 hours [James W. Jacobberger, personal communication]. For calibration purposes then, the relative time scale generated in the Chapter 4 is transformed from 0-1 (0%-100%) to 0-18 hours. The underlying kinetic parameters of the model to be calibrated will therefore have units of $\text{hours}^{-1}$.

A second scaling issue is one of magnitudes. As discussed, procedures exist, in theory, for converting the arbitrary time scales of the expression profiles to exact molecular units. Such procedures were not implemented for the data presented in this dissertation. Therefore, magnitude is presented in arbitrary units. For calibration, the arbitrary expression units were scaled to the magnitudes of the corresponding model outputs (when simulated using the published parameter set). This constrained the optimization problem to one of matching the shape of the outputs and not one of also scaling magnitudes.

Finally, the expression profiles derived from K562 cells are not state variables in the model. Instead, they are related to the model through output functions $y = h(x)$. For the two cyclins, antibodies do not discriminate between bound and free forms or phosphorylation states (e.g. “active” and “inactive” states). The expression profiles therefore are most appropriately compared to the sum total of cyclin A or B forms in the model. PHH3, on the other hand, is not a direct measurement of any state variables. Rather, as discussed, it is used as a proxy for the timing of cyclin B/cdk1
 CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

activity, as both occur at the onset of mitosis. For model calibration purposes, we can use this timing information to construct an “ideal” time curve for cyclin B/cdk1 activity. The timing of PHH3 onset was used to construct “synthetic” cyclin B/cdk1 data consisting of zeros at all PHH3 time points before the onset and values equal to total cyclin B expression at all time points afterwards. This constrains cyclin B/cdk1 activity to be undetectable throughout interphase and abruptly triggered at M, roughly coincident with PHH3 expression. The data set used for calibration therefore consists of the magnitude and time scaled cyclin A2 and cyclin B1 expression profiles, and the synthetic active cyclin B1/cdk1 expression profile. These were compared to the model outputs using the sum of squared errors function discussed in Section 3.1.2.1.

5.1.3.1 Parameter Sensitivities and Non-Identifiability

Ideally, one would calibrate a model by adjusting all of the kinetic rate constants. However, simultaneously calibrating too many parameters increases the degrees of freedom and the likelihood that the algorithm will converge to a non-optimal local solution. It also greatly increases the required computational time and generally degrades algorithm performance [181]. On the other hand, selecting too few parameters to estimate decreases the quality of the model fit. To select a subset of parameters to calibrate, parameter sensitivity can be used to rank the parameters according to their relative impact on the outputs measured.

As described in chapter 3, sensitivity indices are calculated based on $s_{ij}(t) =$
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

\[ \frac{\partial y_i(t)}{\partial k_j} \bigg|_{k^*} \]. Numerically, this is usually estimated by varying the parameters a small percentage - small enough to have a minor impact on the state variables, yet large enough to be distinguishable from errors introduced by the algorithms used to numerically approximate the system trajectory. For dynamical systems, this calculation is often numerically approximated by solving \( \frac{\partial}{\partial t} \frac{\partial y_i}{\partial k_j} \bigg|_{k^*} \) simultaneously with the original system [183], and yields a time-dependent measure of the local sensitivity. It is local because it depends on the starting parameterization of the system.

To determine the global sensitivity, local sensitivities are calculated at multiple random points within the feasible parameter space. The absolute values of the sensitivities for these parameter points are then integrated over the experimental time frame and a weighted average is computed to summarize the total sensitivity. The approach implemented here used a Boltzmann distribution weighting function \( e^{-E/k_bT} \) where \( E \) is the least squared error between the model simulation and experimental data and \( k_bT \) is a customizable scaling factor (taken here to be the maximum \( E \) value). This weighting function ensures that the randomly selected parameter points which are closest to the optimum contribute the most to the global parameter sensitivity estimate [199].

The goal is to fit this model to our data for cyclins A and B and the cyclin B/cdk1 activity. The global sensitivity was therefore calculated with respect to these outputs. As shown in Figure 5.5 a small minority of the parameters (listed in Table 5.1) have a significantly larger impact on the behavior of the measured outputs. There is then
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

Table 5.1: Names of the most globally sensitive parameter identified in Figure 5.5. Refer to the model in Appendix I for the meanings of the parameter names.

<table>
<thead>
<tr>
<th>Rank</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>u</td>
<td>ksbp</td>
<td>ksapp</td>
<td>kweep</td>
<td>k25p</td>
<td>kdbp</td>
<td>kafb</td>
<td>ksbpp</td>
<td>kifb</td>
<td>Jifb</td>
</tr>
<tr>
<td>Rank</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>Name</td>
<td>kdap</td>
<td>kweep</td>
<td>kih1pp</td>
<td>kah1p</td>
<td>kdbpp</td>
<td>Ji1h</td>
<td>ka25</td>
<td>ki25</td>
<td>k25pp</td>
<td>Ji25</td>
</tr>
</tbody>
</table>

a long tail of parameters which have comparatively little effect. These parameters which have nearly zero sensitivity indices will mostly likely be difficult to estimate – changing their values has little to no effect on the measured outputs – but this analysis also indicates refined estimates are probably not necessary.

Figure 5.5: Global parameter sensitivities (weighted average least squares method) of the Csikasz-Nagy 2006 model, evaluated at the available data time points.

Another approach to judge the ability to estimate parameters a priori is to examine the correlation of the time-dependent sensitivity functions. This is a local approach, as it measures the correlation of the sensitivity functions evaluated at the current
parameter set, but it does give an indication of possible functional relations between parameter sets (identified, e.g. in [182]) which introduce redundant degrees of freedom into the estimation procedure and makes calibration difficult. The correlation matrix of the local sensitivities of the Csikasz-Nagy 2006 model (evaluated using the published parameters) is illustrated in Figure 5.6. From this, it is strikingly obvious that the effects of the parameters on the measured outputs are very correlated and the model is probably “non-identifiable”. Indeed, large correlations exist even among the top most globally sensitive parameters (Figure 5.7).
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

Figure 5.6: Correlation matrix between local, time-dependent, parameter sensitivities evaluated at published parameter values. The absolute values of the correlations are indicated by the darkness of the color index, provided along the right of the matrix plot.
Figure 5.7: Correlation matrix between local, time-dependent, parameter sensitivities of the top 20 most globally sensitive parameters, evaluated at published parameter values. The absolute values of the correlations are indicated by the darkness of the color index, provided along the right of the matrix plot.

Due to the very large correlations between parameter sensitivities, it is difficult to draw any meaningful subset of identifiable, yet significantly sensitive, parameters. We therefore attempted calibration of the model using 1.) all 66 parameters 2.) the top 10% most sensitive (top 7) and 3.) the top 30% most sensitive (top 20).
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

5.1.3.2 Optimization Algorithms

Biochemical models generally present a difficult calibration problem: they are non-convex and the landscape of the objective function is extremely rugged. This, combined with non-identifiability, means that calibration generally requires broad and repeated searches through the parameter space [79, 186, 187]. State of the art algorithms often couple global stochastic searches (to cover space and leap local minima) with deterministic local methods (to refine the broad search). We tried three global algorithms: stochastic ranking evolutionary strategy (SRES), a standard genetic algorithm, and simulated annealing. SRES showed the best performance in a comparative study of global methods [188] and has been used successfully in calibrating a model of similar size to the ones presented here [181]. The standard genetic algorithm [79] and simulated annealing [186] have also been used to calibrate biochemical models.

We searched both the entire parameter space and subspaces defined by the most sensitive 10% and 30% of the parameters. Despite evaluating well over $10^5$ parameter sets with each of these methods, we were unable to obtain a satisfactory fit. This is most easily explained by the model structures, which define the space of possible system trajectories and preclude capturing certain data features (most notably the increase in cyclin A2 expression during G2). It bears noting however, that it is impossible to fully invalidate a model structure through calibration attempts alone. Algorithms are imperfect (in this case convergence is purely stochastic) and there are nearly infinite combinations of parameters. Nevertheless, our results strongly imply
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

that certain unmodeled dynamics are likely significant in capturing features of the data.

5.2 Model Modifications Improve the Fit

Numerical values of kinetic rate constants represent the first level of assumptions tested in the calibration process. The mathematical formulation of interactions and the hypothesized network structure are more basic assumptions. We obtained a better fit to data by modifying the Csikasz-Nagy model [9]. As shown in Figure 5.4, the simulated output of total cyclin B was similar to our data, but simulated cyclin B activation occurred much too soon, and cyclin A synthesis was generally unsatisfactory.

5.2.1 Cyclin A expression is undetectable in G1

In G1, K562 cells do not express cyclin A2 above our ability to detect it. As shown in Figure 5.4 however, cyclin A in the Csikasz-Nagy model was transcribed throughout G1. The measured delay in cyclin A2 is due to at least two mechanisms. Firstly, transcription is delayed by the requirement for cyclin D to inhibit Rb and free the E2F transcription factors. Secondly, cyclin A is actively degraded by the APC/Cdh1 complex [200]. We better incorporated these mechanisms to improve the simulation.
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

The Csikasz-Nagy model does not include detailed mechanisms of cyclin D, Rb, or E2F, but instead models the activation of E2F using a Goldbeter-Koshland approximation. E2F therefore switches on abruptly near the start of the simulation and remains at a constant level until switching off abruptly near the end. While lacking in its ability to capture measured cyclin A and cyclin B dynamics, an advantage of the Conradie model [8] is its detailed dynamics of E2F / RB and cyclin D. The model decomposed the Goldbeter-Koshland function used in an earlier incarnation of the model [2], into elementary Mass Action reaction rates and represented E2F activation in greater detail. Furthermore, cyclin D was modeled in more detail, incorporating transcription and proteolysis, as well as binding to a CKI such as p27/Kip1. To better model G1 phase and the restriction point, we therefore replaced the Csikasz-Nagy mechanisms with those from the Conradie model. As an added benefit, the model incorporated “highly stylized” dynamics for basic upstream signaling pathway activity (representing, e.g., MAPK), which provides an entry point for future, more complex models that include growth factor signaling, an essential element of cell cycle regulation.

Secondly, the Csikasz-Nagy model does not include degradation of cyclin A by APC/Cdh1. Additional mechanisms such as Skp2 likely contribute to the proteolysis of cyclin A during G1 [201], but including Cdh1 regulation provides an effective and easily incorporated modification. As shown in Figure 5.8, including this regulation allowed us to model the delay in detectable cyclin A2 levels until the start of S
phase. Biologically, this second mechanism is likely dominant, since cells with inactivating mutations in Rb (resulting in constitutively active E2F) still do not express appreciable cyclin A2 in G1.

5.2.2 cyclin A/cdk drives the activation of cyclin B/cdk1

The Csikasz-Nagy model [9] considered cyclin B/cdk1 as the key antagonist of Wee1 and activator of Cdc25, and created an ultrasensitive, bistable, feedback cycle where cyclin B/cdk1 essentially activated itself. Ultrasensitivity has been identified in several biochemical systems [202, 203] and it has been reported that competition between Wee1 and alternative cyclin B/cdk1 substrates generates ultrasensitivity in Xenopus oocytes [204]. In mammalian cells however, evidence has been accumulating that Wee1 inactivation may be triggered by cyclin A/cdk’s [136, 205, 206]. It has similarly been suggested that ultrasensitivity may arise from the competition between Wee1 and other cyclin A/cdk substrates in human somatic cells [136]. The original model [9] approximated Wee1 activity with a Goldbeter-Koshland function, and so ultrasensitivity was already inherent. However, we modified the trigger of this inactivation from cyclin B/cdk1 to also include cyclin A/cdk. This updates the model in accordance with recent literature. Once Wee1 is inactivated, it is known that cyclin B/cdk1 activates Cdc25 in an autocatalytic loop, which is responsible for the rapid
activation of cyclin B/cdk1. The inactivation of Wee1 by cyclin A/cdk is likely a trigger mechanism which allows a small amount of active cyclin B/cdk1 to accumulate and start an abrupt autocatalytic rise in activity.

5.2.3 Expression of cyclins A2 and B1 is highly elevated in G2

Cyclin B1 has constitutive and cell-cycle dependent transcriptional start sites [207]. As a key regulator of mitosis, cyclin B1 transcription is known to continue through G2. Key transcription factors include USF, NF-Y, B-Myb, and FoxM1 [208, 209]. In addition to constitutive synthesis, a mechanism for elevated G2 transcription of cyclin B was already incorporated in the Csikasz-Nagy model. This secondary transcription was modeled as being dependent on active cyclin B/cdk1 levels. While it is known that cyclin B/cdk1 can indirectly influence its own transcription, for example, through the activation of Bora, Plk1, and FoxM1, it is difficult to see how it could accomplish significant effect prior to appreciable cyclin B1/cdk1 activity. Further, FoxM1 has been shown to require the phosphorylation of an autoinhibitory domain by cyclin A/cdk’s before the transcription of cyclin B and other G2 phase targets [210]. Similarly, B-Myb also requires activation by cyclin A/cdk’s [211–213]. This evidence supports basic reasoning that cyclin A/cdk, which is already active, might drive the accumulation and activation of cyclin B. We therefore modeled the G2
cyclin B transcription factor here as being dependent on cyclin A/cdk both to keep
the model consistent with current literature, and to serve as a testable hypothesis for
future cell cycle studies.

Similar to cyclin B1, our data indicate a change in net synthesis rate of cyclin
A2 in late $S / G2$. We therefore also propose, and modeled, two phases for cyclin
A transcription. Without evidence, we assume cyclins B1 and A2 share a common
transcription factor - although this transcription factor has a lesser effect on cyclin
A2 compared to cyclin B1. This is not unreasonable - E2F, NF-Y and B-Myb, for
example, have been shown to bind both cyclin A and cyclin B promoters [212,213].

5.2.4 An Improved Fit to the General Shape of the
K562 Data

By including the mechanisms so far discussed and tuning the relevant parameters
(explained in the Calibration of the Modified Model section of this chapter), we are
able to improve the general shape of the simulation relative to the K562 expression
profiles. This correspondence is shown in Figure 5.8.

5.2.5 Cyclin A2 is degraded before cyclin B1

However, as shown in Figure 5.9., a magnified view of the final 2% of the cell
cycle illustrates an additional inconsistency. Currently, cyclins A and B are modeled
Figure 5.8: Comparison of the modified model outputs (solid lines) to K562 data (markers). (a, b, c) Total simulated levels (bound and free, inactive and active) of cyclins A and B, and cyclin B/cdk1 activity are compared to the measured expression profiles of cyclins A2 and B1, and PHH3.
as degraded simultaneously upon the activation of Cdc20. To our knowledge, this is the case for all published models. The K562 data, however, clearly demonstrate a delay in cyclin B degradation relative to cyclin A. There are several possibilities which might explain this.

It has recently been shown that cyclin B is subject to the spindle assembly checkpoint (SAC), while cyclin A is not. Cyclin A is degraded soon after APC/C phosphorylation by active cyclin B/cdk1, while cyclin B (and other crucial mitotic complexes such as securin) are protected by SAC-promoted sequestering of Cdc20 by Mad2 and the formation (with BubR1 and Bub3) of the inhibitory mitotic checkpoint complex (MCC) which binds to and inactivates the APC/C [214]. Cyclin A is likely degraded regardless of SAC activity by either or both of two proposed mechanisms: cyclin A may have a high enough affinity to compete with Mad2 for free Cdc20 before being targeted to the APC/C by Cks, or cyclin A may activate MCC-inhibited Cdc20 even after it has bound to the APC/C [215]. The different degradation pathways
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

of cyclins A and B enforced by the SAC therefore explains at least a portion of the staggered degradation of the K562 cyclins. Incorporating these mechanisms in a computational model is, at present, a difficult task. Progress has been made in dissecting the full molecular details of the SAC, and several simplified models have been constructed [216], but a more refined description requires additional quantitative data to clarify crucial gaps in pathway knowledge (e.g. how SAC inhibition is relieved and the MCC dissociates), and additional modeling approaches are required to couple the Mass Action style ODE models explored here with the biophysical modeling of microtubule forces and spatial distributions of molecules required to account for the fine structural mechanisms underlying the SAC [217].

A much simpler explanation might also be sufficient for explaining the staggered degradations of cyclins A and B. Contrary to other cyclins, synthesis of cyclin B has been found to continue through mitosis [218]. The G2 transcription factor, shared between cyclins A and B is currently modeled such that it degrades as cells enter mitosis. Introduction of an additional source of synthesis, here presumed to be dependent on active cyclin B/cdk1, can effectively prolong cyclin B expression through mitosis and delay its degradation by Cdc20/APC/C. This is illustrated in Figure 5.10. The mechanics underlying this continued synthesis are complicated, but including a synthesis rate proportional to the concentration of active cyclin B/cdk1 provides a convenient way to model continued expression of cyclin B through mitosis and serves as a placeholder for future studies.
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

Figure 5.10: Magnified view of the final 2% of the cell cycle, with added synthesis of cyclin B through mitosis.

This final modification results in an improved simulation of all four cell cycle phases, encompassing both interphase and mitosis. A diagram of the model's modifications is provided in Figure 5.11.
Figure 5.11: Schematic of the modified model developed in this study. The black colored portion represents the original Csikasz-Nagy 2006 model and the grey portion is the incorporated section of the Conradie 2010 model. The red portion indicates modifications made to calibrate the model to the K562 data. Solid lines indicate chemical reactions and dashed lines regulatory effects.
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

5.2.6 Calibration of the Modified Model

The Csikasz-Nagy model [9] was modified, combined with the initiating portion of the Conradie model [8], and expanded to include 26 ODEs and 121 kinetic rate constants. The equations and kinetic rate constant values are provided in the Appendix. Equations were solved in MATLAB using the stiff ode15s solver. The modified model was calibrated manually, using Figure 5.11 and the differential equations provided in the Appendix as guidance to adjust the strengths of relevant rate equations. Manual calibration was completed using the Systems Biology Toolbox 2 and SBPD packages for MATLAB [219], which allow live updating of model outputs relative to experimental data, in response to parameter adjustments.

Automated calibration was attempted using the SRES, simulated annealing, and standard genetic algorithms discussed earlier. These algorithms are implemented in the same MATLAB SBT2 and SBPD toolboxes used for manual calibration, which numerically simulates the system and attempts optimization using the C programming language. This allows simulation and calibration several orders of magnitude faster than the native MATLAB software. The entire parameter set, the 10% and the 30% of most globally sensitive parameters (illustrated in Figure 5.12 and listed in Table 5.2), supplemented with the unknown parameters introduced with the modifications, were tuned using each algorithm. Despite evaluating over $10^5$ parameter sets with each combination of parameter dimensionality and routine, no satisfactory fit was discovered.
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

Figure 5.12: Global parameter sensitivities (weighted average least squares method) of the modified model developed in this chapter, evaluated at available data time points of cyclins A and B and cyclin B/cdk1 activity.

Table 5.2: Names of the 30% most globally sensitive parameter identified in Figure 5.12. Refer to the model in Appendix I for the meanings of the parameter names.
This is not unprecedented. Chen et al. [186] also reported that despite being able to calibrate a biochemical model to “artificial data” (produced by simulating the model with a predefined parameter set) with a simulated annealing algorithm, they were unable to obtain a satisfactory fit to experimental data. This was despite using a large set of experimental data and a 100-node cluster computer. For the model developed in this chapter, there are several possible explanations. The first set of explanations arise from the structural difficulties of the calibration problem, for example the ruggedness of the objective function landscape, the “sloppy” nature of parameter sensitivities [189], and functional relations between parameters which contribute to non-identifiability. Additionally, this is likely the result of model imperfections (e.g. the approximated behavior of the hypothetical G2 transcription factor) or un-modeled dynamics which would improve the exact numerical trajectory of the simulation. Finally, while the manually tuned model outputs appear satisfactory, they are likely numerically non-optimal. Successful automated calibration may require reformulation of the objective function to better reflect the qualitative judgements which, when combined with the exact quantitative squared error calculation, yield a satisfactory calibration.
5.3 Re-analysis of the K562 Data

The primary data of Chapters 4 and 5 were extracted early in the learning phase of profile extraction by our research group. Later refinements on region setting (number of regions and boundary rules) and Gaussian fitting (numbers of Gaussians and variable constraints) resulted in more informative and sometimes smoother expression profiles. Therefore, in this section, we examine what we consider to be more accurate representations of expression. The new expression profiles are similar in shape, but with key differences. The shapes are, in general, smoother and more uniform, without the sharply defined transitions between S phase and G2. Additionally, the shape of the cyclin B1 expression profile is markedly more exponential. These new expression profiles are shown in Figures 5.13, 5.14, and 5.15.

Fitting this data required further modifications to the model. These modifications result from the more pronounced exponential character of cyclin B1 revealed in this re-analysis. The existing model includes three sources of cyclin B transition: constitutive transcription (which synthesizes cyclin B at some constant rate), that induced by the shared cyclin A/B transcription factor in G2, and active cyclin B/cdk1 driven synthesis during M phase. It is not possible to produce an exponential increase in cyclin B through parameter changes alone (Figure 5.13). Interestingly, studies have also identified E2F promoter elements on the cyclin B1 gene [213]. This creates a model with four distinct sources of cyclin B transcription: constitutive (cell cycle independent), E2F promoter driven, G2 phase specific (possibly via cyclin A activation of
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

e.g. NF-Y or B-Myb), and continued M phase synthesis. The additional E2F-driven synthesis, however, does not account for concavity inherent in the re-analysis of the K562 data, at least not given the existing model which assumes nearly constant levels of active E2F throughout S phase. E2F levels are not constant, however, and transcription has been found to be autoregulatory (E2F promoters contain E2F-binding sites) [220]. This establishes a system where E2F may increase exponentially throughout the cell cycle. No data is currently available to precisely determine the shape E2F expression, but assuming autocatalytic synthesis of E2F and E2F-regulated synthesis of cyclin B allows the model to be calibrated much more closely to the re-analyzed K562 data. One example calibration is shown in Figure 5.14 and 5.15. The additional model modifications are illustrated in the diagram of Figure 5.16.

5.4 Cyclin A Binds Both Cdk1 and Cdk2

An additional modification was made to the model, which does not affect the simulation output, but which improves the model’s fidelity with known biology, and allows the model to be easily extensible when more data becomes available. The Csikasz-Nagy model only considered one population of cyclin A. However, cyclin A is known to bind both cdk1 and cdk2. Cyclin A preferentially binds cdk2, which was shown to be enforced by differential activation by the Cdk Activating Kinase (CAK) cyclin H/cdk7 [221], reportedly until a significant fraction of available cdk2 is bound.
Figure 5.13: Current model (solid lines) calibrated to the re-analyzed data (markers). The data show pronounced concavity in the shape of cyclin B1; the model is unable to capture this trend.
Figure 5.14: Introducing E2F autoregulation and E2F regulated synthesis of cyclin B allows an improved fit (solid lines) to the re-analyzed K562 data (markers).
Figure 5.15: Magnified view of the final 3% of the cell cycle. Magnitudes were normalized to allow plotting on the same scaled axis.

in complex [221]. Cdk2 can be phosphorylated as a monomer and then stably binds cyclin A, whereas cdk1 can only be phosphorylated when unstably bound to cyclin A and this phosphorylation then stabilizes the dimer. Cyclin A then presumably binds cdk2 much more readily until cdk2 is more or less completely bound (between cyclins A and E). Then the newly synthesized cyclin A begins binding cdk1 in sufficient amounts to allow CAK to stabilize a significant fraction. It has been argued that only once the level of cyclin A/cdk2 reaches a plateau in late S does a significant fraction of cyclin A bind cdk1. However, cdk2 has been estimated to be present at levels at least 8-times the maximum levels of either cyclin E or cyclin A [222]. Whether the level of cyclin A/cdk2 truly reaches a plateau, or whether this is an artifact of the measurement process, remains to be determined. Most previous models, including Csikasz-Nagy, did not explicitly consider the binding of cyclins to cdk’s, as cdk levels are relatively constant and in excess of their cyclin partners [222] and binding is
Figure 5.16: Diagram of the model modifications introduced to better fit the re-analyzed K562 data. The original Csikasz-Nagy 2006 model is shown in black, the Conradie 2010 model portion is shown in grey, modifications introduced previously in this dissertation are in red, and the modifications required to fit the re-analyzed data are presented in blue.
presumed to occur sufficiently fast as to be ignored. Similarly, in our model, both cdk1 and cdk2 are activated by binding to cyclin A. We propose and modeled the synthesis of two distinct cyclin A fractions. The binding preference for cdk2 was enforced by appropriately proportioning the two synthesis rates. Presumably, the two fractions perform separate regulatory tasks. However, little evidence exists as to their distinct functions. We’ve modeled the two fractions as preparation for future efforts, which can further refine the model once more is known about the distinct functions of cyclin A/cdk1 and cyclin A/cdk2. For now, both are modeled as acting in unison. In keeping with the Csikasz-Nagy model, we did not consider inhibitory phosphorylation of cyclin A/cdk2, which was shown to be unimportant in unperturbed cells [223]. Similarly, we considered the inhibition cyclin A/cdk1 by Wee1 and activation by Cdc25 to be insignificant [198, 224, 225].

5.5 Discussion

Systems biology is an iterative process (Figure 3.1). This work represents one iteration. We’ve found a mismatch between expression data from an asynchronously growing population of a human cell line and the state variable output from previously published computational models. Extensive calibration attempts were unable to tune the model parameters and improve the fit. We therefore updated a model with biological mechanisms that improve the correspondence between output and data, and
that we believe to be significant in terms of biology. Including these mechanisms improved the model fit to data and provides a starting point for validation of these ideas through biology. Mathematical modeling formalizes biological knowledge and helps understand data that represent a large network of interacting variables. Comparing an existing model to data, we’ve discovered previously unmodeled dynamics to be significant factors determining the dynamic expression profiles of cyclins A2 and B1. The majority of these model additions are well-supported by biological experiments. However, the cyclin A-dependent activation of a G2 transcription factor, shared by both cyclins A and B, represents an hypothesis that needs to be tested. We’ve shown this hypothesis to be consistent with dynamic expression profiles in unperturbed cells, and previously published observations do not contradict it. Modeling therefore also generated a testable hypothesis.

The proposed model is undoubtably non-unique. It represents one possible solution and there are many other un-modeled interactions which might also improve the model output. Nevertheless, the modifications are either fully supported or at least not contradicted by known biology. Through an extensive literature search, many other biological mechanisms were identified which could potentially augment the model and improve the simulation output. Consultation with a biologist was a crucial aspect of model building, to judge the significance and feasibility of proposed model changes.

Mathematical approximations represent a level of the hierarchy of modeling as-
CHAPTER 5. AN IMPROVED COMPUTATIONAL MODEL OF THE MAMMALIAN CELL CYCLE

sumptions, between numerical rate constants and model structure. A simple cyclin A/cdk-activated transcription factor which abruptly “turns on” at the start of G2 (due to an ultrasensitive, Goldbeter-Koshland switching function) is an over-simplification of the real biochemistry. Such simplifications are limitations of the model. They allow us to reason about the organization of the cell cycle regulatory network and the relationship of our measured markers to that system, but do not generate the highly refined expression data we’ve observed. Future work requires models that are more complex, based on information that is available now, but such models invoke a need for significantly more data for model calibration. The system of data generation that is described in Chapter 4 is one approach that provides correlated data that covers the whole system. The aim would be a better understanding of transcriptional regulation of cyclins A and B, and the reflection of this understanding in a working, complex model.

There currently exists very limited quantitative data on the absolute concentrations of cell cycle regulatory proteins. While estimates are available for a small subset of proteins [5, 222], the vast majority are unknown. Frisa et al. [5] provide a technique to couple absolute concentration calculations with the multi-parametric flow cytometric methods discussed previously. Until such methods are fully implemented, the outputs of the discussed models are in terms of arbitrary units of magnitude. The rate constants of the models therefore only capture the time-scales of processes. They do not reflect the concentration scales of the related proteins, which, for now,
are in arbitrary units. When quantitative concentration data does become available, this information can be incorporated into the model by appropriately scaling the magnitudes of the kinetic rate constants for the reactions in which a given protein participates.

In this model, we assumed a single compartment, freely mixed, aqueous solution (as do all Mass Action -based models). The cell cycle regulatory network (like all events in a cell) however, is compartmentalized and encompasses both solution and solid state chemistry. The model does not consider details such as the shuttling of proteins between the nucleus and cytosol or the localization of proteins to solid substrates (such as cyclin B association with the centrosome and mitotic spindle). For realistic models, these events need to be accounted.
Chapter 6

Variation

Systems biology models have traditionally captured the “average” behavior of a “typical” cell. Cell-to-cell variability is a relatively new and very active research topic [226–230]. With the advances of imaging modalities such as flow cytometry and time-lapsed microscopy, researchers are gaining an ever-clearer picture of both the average dynamic behavior of cells and their heterogeneity. Dynamic proteomics utilizing microscopy can follow individual cells through their cell cycle [226,230], but flow cytometry offers the advantages of numbers: hundreds of thousands of cells can be sampled and much better statistics can be obtained regarding the distribution of these cells about the population averages. We utilized the advantage of flow cytometry’s scale in deriving the median expression profiles used for the construction and calibration of the ODE model in the previous chapter. The simulation trajectory through the original “cloud” of flow cytometric data is illustrated in Figure 6.1.
An interesting extension of this work is to explore possible origins of the population variation we’ve measured and modeling approaches which might capture it. In the following sections, we explore the sources of measured variance in biological data and discuss applicable modeling techniques.

Figure 6.1: Simulation compared to the original K562 data. The model output was offset and scaled to approximate the median trajectory.

6.1 Single Cell Measurement Technologies

Before molecular measures of protein variability, cell-cell variability was first examined through differences in cell generation times [33, 44] and sizes at [44] or after division [46]. Such measures gave rise to the phenomenological, statistical models discussed in Chapter 2. Recently, single cell measurements have provided new data
CHAPTER 6. VARIATION

on the variation of molecular measures of cell state.

Molecular biology has traditionally been measured using averages of populations of cells. Immunoblotting such as Western Blots, for instance, report the total amount of protein in a lysed sample of cells. If cells respond asynchronously to an input, or oscillatory cell cycle proteins are measured, the resulting total is an inaccurate estimate of individual cell behavior. For example, p53 dynamics were first described as damped oscillations in response to DNA damage [231]; single cell measurements later revealed that individual cells exhibit differing number of p53 pulses of equal amplitude and duration [232]. There are two broad classifications of single-cell techniques: those which track individual cells over the measurement period and others which provide a single, static sample of the population.

Live cell imaging is often accomplished with time-lapse microscopy. Molecules of interest are fused with fluorescent reporters and images are captured to estimate their levels and spatial localization. Live cell reporters are created using a wide range of strategies - transiently transfecting cells with vectors expressing the fluorescent reporters, establishing clonal cell lines, or integrating a fluorescent protein in the endogenous gene locus via bacterial artificial chromosomes, random viral insertion and screening, or targeted somatic recombination [233]. The latter method has the advantage that the reporter is under the same transcriptional regulation as the targeted gene. In addition to expression levels, protein activity can also be estimated using Forester Resonance Energy Transfer (FRET). However, while these methods allow
the tracking of an individual cells behavior or response to stimuli, it is sometimes
difficult to create a reliable reporter, and nearly impossible to guarantee the tagged
protein is behaving naturally and that the natural system is therefore unperturbed.

To reduce disruption of the natural system, single cells can be measured using
immunofluorescent microscopy or flow cytometry. In these cases, unperturbed cells
are fixed and then stained for the marker of interest. While these techniques cannot
follow the behavior of living, individual cells, they do have the advantage that large
numbers of cells can be sampled. We’ve previously shown in Chapter 4 that, using
these large samples, one can construct an estimate of the underlying expression pro-
files of the measured molecules. This then provides an estimate both of the median
expression profile and of the variation.

6.2 Measurement Noise

Variance in biological data can be classified into biological variance and measure-
ment noise. Measurement noise has an unavoidable, and often large, contribution
to observed variation. Measurement noise can be due to the biochemical basis of
epitope tagging, electromechanical imprecision of instruments, or operator error. For
flow cytometry, many contributions to measurement error have been identified. Many
of these errors are related to the flow cytometry machine (e.g. electric noise, laser
variation, stay light sources, etc.). However, while a typical flow cytometry measure
may show a 10- to 1000-fold range in levels of a constitutively expressed protein, machine error has been estimated (using standardized fluorescence beads) to account for less than a 10-fold range of the observed variation [234]. This estimate represents a minimum bound on the likely total measurement error. Epitope tagging and measurement of cells is much more complex than standardized fluorescence beads. Additional errors likely creep in due to the fixation process, molecular crowding, inaccessibility of binding sites due to the structural chemistry of cell biology, etc.

The first measurement step is to add a primary antibody to cells. This introduces the first source of error, as not all of this antibody binds any molecules, and some portion of the antibody binds the wrong epitope. This can be termed non-specific binding. Thus, the measured amount of antibody is proportional to the level of epitope, but not identical. A secondary antibody is then often used, which contains the fluorescent reporter and binds the primary antibody. This introduces an additional amount of uncertainty, as not all of this antibody binds the primary antibody. Quantification of the levels of the fluorescent chemical is also imperfect. Flow cytometers irradiate each cell with a laser in order to quantify the subsequent fluorescence, which is picked up by detectors tuned to specific wavelength ranges. Cells are dense, and so only a fraction of fluorescent molecules are irradiated and only a fraction of these are sensed by the flow cytometer. Errors are also introduced when multiple molecules are tagged with different fluorescent chemicals. Fluorescence is measured by quantifying the levels of specific wavelengths of light. The emission spectra of different fluorescent
CHAPTER 6. VARIATION

chemicals often overlap and this error is therefore referred to as “spectral overlap.” When spectral overlap occurs, the researcher must subtract out the light that each specially tuned detector was not meant to receive. This is done manually and the inherent imprecision also corrupts the data. Finally, detectors are extremely sensitive and detect autofluorescence of cells and other stray light sources.

Proteins therefore cannot be measured directly and a significant fraction of measured variation is likely due to the measurement process. A series of procedures fixes biochemical tags to the epitope of interest \( x \) and a related measure, e.g. fluorescence intensity \( y \), is measured. The most common theoretical treatments consider measurement noise \( \epsilon \) to be additive and normally distributed, yielding

\[
y = \alpha + \beta x + \epsilon, \; \epsilon \sim N(0, \sigma_{\epsilon})
\]

Here \( \alpha \) is a systemic offset and \( \beta \) is a scaling factor. Ideally experiments are conducted in the linear range between the fluorescence intensity and protein concentration. If purified proteins are available, this scaling factor can often be estimated experimentally [5].

While standard noise models usually assume additive Gaussian noise, non-negative data (e.g. fluorescence intensities) are often affected by multiplicative noise [235]. Such an error model was proposed [236] for immunoblotting

\[
\hat{y} = \beta_0 + \beta_1 x (1 + \eta), \; \eta \sim N(0, \sigma_{\eta})
\]
CHAPTER 6. VARIATION

This represents a specific case of

\[ \bar{y} = \beta_0 + \beta_1 x^{\beta_2} \eta, \; \eta \sim e^{N(0, \sigma_\eta)} \]

where \( \beta_2 = 1 \) [235].

For multiplicative noise, a log-transformation will result in additive Gaussian noise and the applicability of related methodologies. Specifically, setting \( y = \log(\bar{y} - \beta_0), \alpha = \log(\beta_1), \beta = \beta_2, \epsilon = \log(\eta), \sigma_\epsilon = \sigma_\eta \) transforms Equation 6.1 to the additive noise model for \( \log(x) \).

Flow cytometry data likely includes a combination of the two noise sources,

\[ \bar{y} = \beta_0 + \beta_1 x(1 + \eta) + \epsilon \]

Here \( \eta \sim N(0, \sigma_\eta) \) represents multiplicative noise sources (e.g. specific-binding variation) and \( \epsilon \sim N(0, \sigma_\epsilon) \) accounts for additive background and other sources. For illustrative purposes, we can use this formula to create artificial flow cytometry data. As an example, we can take the offset values of original flow cytometry data (which were subtracted during processing of the time curves) as the mean of the additive noise sources and choose coefficients of variation which match the general width of the data distribution. A comparison between simulated noise and the original data is shown in Figure 6.2. Here the simulation output was rescaled to the cumulative frequency domain (from the time domain, the inverse of that discussed in Section
CHAPTER 6. VARIATION

4.1.4) to generate an adequately weighted sample, and perturbed by the noise model.

![Simulation Diagram](image)

Figure 6.2: (a) Model simulation with additive and multiplicative noise. An equal number of cells as the K562 data set (b) were simulated by evaluating the system at 100,624 equally spaced points (having rescaled the time domain to the frequency domain – the inverse of that described in Section 4.1.4) to generate a representative sample, and perturbing each point by the given noise equation. Here $\epsilon_A \sim N(300, 120)$, $\eta_A \sim N(0, 0.18)$, $\epsilon_B \sim N(500, 160)$, $\eta_B \sim N(0, 0.15)$ to approximate the offset and distribution of the K562 data.
CHAPTER 6. VARIATION

The simple noise model we’ve constructed gives a decent, but inexact, approximation to the raw flow cytometry data. The temporal variation of the real expression profiles are likely more complicated than we’ve approximated. The estimate shown in Figure 6.3 was constructed by calculating the variation of cyclin A2 values assigned to each “gate” originally used for deriving the median profile (described in detail in Chapter 4). This is a very inexact estimate, owing to the combination of variance both parallel and orthogonal to the programmed expression profile which is captured in a given gate. However, for the time being, it provides a rough approximation. Here we see the coefficient of variation of the data varies considerably over time, and in a manner not expected by the simple noise model introduced above. This is likely the result of a more complicated measurement noise process as well as biological mechanisms producing real variation in the distribution of cyclin levels.

Figure 6.3: Estimate of the temporal dynamics of the measured coefficient of variation of cyclin A2. Values were estimated by calculating the coefficient of variation about the median value in each “gate” defined in Chapter 4.
CHAPTER 6. VARIATION

6.3 Biological Variation

While it is difficult at present to precisely measure and thus separate the contributions of measurement noise, biological variance likely contributes a non-trivial amount of variation to the flow cytometry data. As explained by Sui Huang [237], there are many sources of heterogeneity in cell populations, and it is important to be precise and clarify which aspect we are interested in. The most broad classification is between genetic (e.g. in tumors) and non-genetic (e.g. in a clonal population).

Among isogenic clones, with the same history and environment, one can distinguish between extrinsic and intrinsic noise. Experimentally, such noise classification is obtained by placing two gene reporters in the same cell and quantifying their gene expression. Intrinsic noise creates differences between the two reporters in the same cell, while extrinsic noise produces the same effect on both reporters within a given cell but creates differences between cells [238]. In other words, “Intrinsic noise has been defined as fluctuations arising from smallness of molecule numbers in reactions. Extrinsic noise accounts for the rest of the noise, originating from fluctuating physiological conditions” [116]. Intrinsic noise is due to the inherent, discrete, stochastic nature of biochemical collisions and reactions within a cell. Extrinsic noise is due to a large variety of sources: cell-to-cell interactions, distance from neighboring cells, oxygen and growth media gradients, etc.

Huang further identifies macro- and micro-heterogeneity. Macro-heterogeneity is manifested in multimodal distributions of a given measure within a presumed ho-
CHAPTER 6. VARIATION

Homogeneous population. Microheterogeneity is the distribution around a mean for a unimodal distribution of a given measure. These distributions can be created either by extrinsic or intrinsic noise (e.g. through the stochastic activation of an ultrasensitive system). A key distinction is that extrinsic noise can theoretically be isolated from the system, but intrinsic noise results from the discrete nature of finite and random molecular interactions.

Finally, Huang distinguishes between temporal noise, which is the change of measurement over time in a random, disordered manner, and population noise, due to distinct stationary trait differences. Most “gene expression noise” studies focus on temporal noise and attribute this variation to the random synthesis and degradation of mRNA and proteins [239], and stochasticity arising from a finite number of molecules diffusing, colliding, and reacting in a complex media. Temporal noise generally satisfies ergodicity conditions, and allows estimation of the average cell level from a snapshot of a population. Population noise is a bit more insidious and, in the simplest sense, represents a population distribution of a time-invariant measurement. Human cell lines appear to represent a combination of these two sources. While probably not as prominent as prokaryotes or yeast (owing to human cells having a larger number of molecules), temporal noise is still evident in single-cell tracking data and gene expression studies [230]. Additionally, human cell lines exhibit slow fluctuations of protein levels that persist through multiple cell generations. Cells have been observed to exhibit a mixing effect, where correlations between sister cells decay over
CHAPTER 6. VARIATION

time [226].

Restricting ourself to micro-heterogeneity within an isogenic cell line, how then can we model temporal and population heterogeneity?

6.3.1 Modeling Intrinsic Noise

Systems biology models are most often constructed within the deterministic framework, formulated with Mass Action assumptions or Michaelis-Menten approximations. The mass action assumption states that the rate of a reaction is proportional to the concentration of the reactants. This formulation can be shown to be an approximation of the stochastic chemical master equation (CME) description, in the limit of an infinitely large number of molecules [179]. In practice, it has been shown that the magnitude of stochastic fluctuations scales with $1/\sqrt{N}$, where $N$ is the number of molecules in a given compartment. Deterministic approximations are therefore considered “good enough” when the total number of molecules is at least $> 10^2\cdot10^3$, depending on the question of interest [180]. However, small numbers and infrequent reactions require discrete stochastic methods such as the Markov jump process formulation and the resulting CME system (simulated using the so-called Gillespie algorithm). This approach has been used to study stochastic effects on the cell cycle in yeast [135,137,162,165,166] and mammals [91].

While most mammalian signaling molecules are present in sufficiently large numbers (typically on the order of $10^3\cdot10^6$ copies/cell for mammals), stochasticity can
still arise from more infrequent events such as gene expression. In yeast, the numbers of molecules are much smaller (on the order of 100’s to 1000’s per cell [128,240]) which might result in more stochasticity. In mammals, Arooz et al. [222], have measured cdk levels to be on the order of $10^5$ molecules. However, cdk levels are known to be in excess and are often even excluded from models (as they are not rate-limiting). It is important that Arooz et al. also measured the concentrations of cyclins A2 and B1 to be about 30 fold less than cdk1, with cyclins A2 and E1 about 8-fold less during G2 and G1 phases respectively. These peak values are only on the order of 1000s, with lower levels, e.g. cyclins A2 and B1 during G1, being significantly less, and may therefore be impacted by stochastic effects. Exact measures of cyclin molecules vary significantly however. The above estimates can be contrasted with those by Frisa and Jacobberger who recently found a range of 150,000-2,500,000 (a 16-fold range with a 24% coefficient of variation) [5]. Other estimates of cyclin B1 levels (e.g. for Hela cells) range from 3,000 to 3,000,000 molecules/cell. Concentrations of other cell-cycle regulatory molecules may vary significantly from these estimates. The numbers of cell cycle regulatory molecules vary throughout the cell cycle, however, and it is possible that some species may approach significantly small concentrations - small enough to warrant CME modeling. Until more data is available on the exact quantification of cell cycle regulatory proteins, however, it is difficult to construct useful models.

An alternative approach to capture intrinsic noise is an approximation of the CME called the Chemical Langevin Equation (CLE), which consolidates stochasticity into a
CHAPTER 6. VARIATION

general noise term which is appended to each ODE of the corresponding Mass Action system, e.g.

\[
\frac{dC(t)}{dt} = vr(C) + v\sqrt{r(C)}\omega(t)
\]

where \(v\) is the stoichiometry matrix and \(r(C)\) is the rate law of the equation. Here \(\omega(t)\) is a Gaussian white-noise process with zero mean and unit variance. This is sometimes simplified by approximating

\[
\frac{dx_i}{dt} = f(x, k) + \omega_i(t)\sqrt{2D_i x_i}
\]

where \(\frac{dx_i}{dt} = f(x, k)\) is the original ODE formulation of a given rate equation, \(\omega(t)\) is the same zero mean, unit variance Gaussian white noise process, and \(D_i\) is a characteristic noise amplitude term. This formulation has been used for several preliminary attempts at studying intrinsic noise in the yeast \([124,160]\) and mammalian \([92]\) cell cycles.

6.3.2 Distributions of Initial Protein Concentrations or Kinetic Rate Constants

Long-lasting correlations in protein concentrations, across several generations, have been observed in human cell lines \([226]\). This variability is akin to Huang’s “population” noise, and is likely separate from the temporal, intrinsic noise discussed
above.

One possible cause of this variation is the unequal partitioning of cell contents at cytokinesis. Unequal partitioning has been postulated (going back at least to 1962) as an important influence on cell cycle transit times [35]. The simplest instance is the observation that cells born smaller at G1 take a comparatively longer time to complete the cell cycle [105]. Modeling this variation could be accomplished quite easily by creating a population of independent simulations initialized with a distribution of cell masses [126]. Extending this thinking further, it is likely that cells also partition the number/concentration of regulatory biochemicals/proteins unevenly. A cell population could then also be modeled using initial distributions of proteins (state variables) and Monte Carlo-style simulations. This method has been used quite successfully for modeling the distribution of individual cell responses to apoptotic signaling [228] by simulating a population of independently, log-normally, distributed initial protein concentrations prior to simulating a model of TRAIL signaling. On the other hand, Feinerman et al. demonstrated the existence of co-regulation of regulatory molecules and the need to include such covariances for biologically meaningful simulations [241].

Separate from initial protein distributions, one could presume a distribution of kinetic rate constants. Such rate constants would vary between cells, but would remain constant over the cell cycle simulation [74, 125, 125, 131, 165]. As explained by Chiorino and Lupi, “changes in reaction rates could be related to microenvironment differences (e.g. local temperature, pH, ionic strength) among cells, particularly in
CHAPTER 6. VARIATION

the nucleus, since the synthesis of all the products involved in cell growth, DNA replication and cell division is regulated by DNA and nuclear proteins” [74].

Unequal partitioning at cell division and distributions of kinetic rate constants could both be accommodated for by the traditional deterministic framework. In this case, the initial protein concentrations \( x(t = 0) \) or kinetic rate constants \( k \) can be conceptualized as constituents of the vector of random variables \( \theta \).

The effects of these random variables on the output of the model (i.e. the distribution of state variables or measurable outputs at later time points) can be directly determined by a transformation. Generically, given a random vector \( X = (X_1, \ldots, X_n) \) with probability density function (pdf) \( f_X \), we define a transformation \( Y = g(X) \), which is an injective map with continuous first partial derivatives. Denote the Jacobian matrix of \( g(X) \) by \( J_g(X) \). Then the pdf of \( Y \) is given by

\[
f_Y(Y) = \frac{f_X(X)}{|\det(J_g(X))|} \tag{6.1}
\]

Note that \(|\det(J_g^{-1}(Y))| = \frac{1}{|\det(J_g(X))|} \) and so an equivalent definition is

\[
f_Y(Y) = |\det(J_g^{-1}(Y))|f_X(X)
\]

Now, consider the initial state variables \( X(t = 0) \) to be random variables with pdf
CHAPTER 6. VARIATION

$f_0$, then later time points are given by the map

$$Y_t(X_0) = g_t(X_0)$$

Then, for $t > 0$

$$f_{Y_t}(Y_t) = |det(J_{g_t^{-1}}(Y))|f_0(g_t^{-1}(Y))$$

In either case, we require an analytic solution of the ODE system (either to calculate the Jacobian matrix of this solution $g(x,k)$ or its inverse). However, for nontrivial models of biochemical systems, analytic solutions are unknown. Thus it is difficult to predict analytically how the pdf of initial conditions (or similarly, kinetic rate constants) produces a distribution of state variables over time. Numerical techniques have been developed to estimate the necessary inverse functions and the propagation of variance, but they are often quite difficult to implement in practice [242–244].

An alternative approach to estimate the propagation of variance through a complex system is Monte Carlo simulation. Monte Carlo samples the distribution of the random variables to estimate the distribution of the output. When the initial distribution is adequately sampled, the distribution of the output can be accurately estimated. It should be noted that an inherent limitation of sampling is that infrequent events may not be adequately captured using this method. Nonetheless, it is often the most practical, or sometimes even the only possible, means to estimate the effects of a random variable on a complex system.
CHAPTER 6. VARIATION

In this case, we could use Monte Carlo-style simulation to create a population of individual cells. The dynamic behavior of each cell $i$ is given by the ODE’s of the original system, parameterized and initialized using values sampled from pre-defined distributions:

$$
\Sigma(\theta^i) : \begin{cases} 
\dot{x}^i = f(x^i, \theta^i) & x^i(0) = x_0(\theta^i) \\
y^i = h(x^i, \theta^i)
\end{cases}
$$

with state variables $x^i \subseteq \mathbb{R}^n_+$, outputs $y^i(t) \subseteq \mathbb{R}^m_+$, and parameters $\theta^i \subseteq \mathbb{R}^l_+$. Here the functions $f : \mathbb{R}^n_+ \times \mathbb{R}^l_+ \to \mathbb{R}^n$, $h : \mathbb{R}^n_+ \times \mathbb{R}^l \to \mathbb{R}^m_+$ and $x_0 : \mathbb{R}^l_+ \to \mathbb{R}^n_+$ are continuous. The parameters $\theta$ are distributed according to a probability density function $\Theta : \mathbb{R}^l_+ \to \mathbb{R}_+$. The ensemble cell population model is then the set

$$
\Sigma_{pop} = \{\Sigma(\theta^i)|i = \{1, ..., N\}, \theta^i \sim \Theta\}
$$

The resulting simulation could then be analyzed graphically or quantitatively to judge its similarity to experimental data.

6.3.3 Calibration

All of this analysis so far, however, has been focused on the “forwards problem”: given a model, parameters, an initial condition, and characteristics of the initial variance, predict the output. Calibration, on the other hand, (e.g. fitting a simulation
CHAPTER 6. VARIATION

model to our K562 data) represents the inverse problem: given data, determine the underlying values of initial conditions and parameters. The calibration of stochastic models is complicated by the additional parameters characterizing the stochasticity (e.g. the variance as well as the mean of the initial system states). There are many ways to approach the inverse problem.

For the problem of varying initial conditions or kinetic rate constants, the analytic derivations presented above demonstrate that there is no closed form solution for the biochemical models considered here. Calculating the distribution of the initial conditions (or kinetic rate constants, using an analogous derivation for the transformation of the vector of random parameter values $K$) requires knowledge of the explicit solution or its inverse. So we instead must turn to computational approaches.

If such data is available, the most direct method is to use individual output measurements of the “final conditions” of a biochemical system and backwards integration of the system of ODEs to obtain a collection of points which produce the output distribution. This generates an exact collection of initial values which produce the exact measurement points. This is theoretically possible for data sets which include a population sampled at incremental time points (e.g. in response to drug treatment). However, for the single snapshot flow cytometric data used in this thesis, we do not know the exact values of the measurements at a final time point, as the time domain is implicitly encoded in the frequency of the data. We therefore only have a rough estimate of the variance of measured cells about the median at approximated time
points (e.g. Figure 6.3), where time points are defined as distinct “gates” which en-circle groups of cells expressing similar levels of measured proteins (discussed in more detail in Chapter 2).

A second approach is Inverse Monte Carlo by Iteration (IMCI). IMCI is formulated as an optimization problem which is very similar to the model calibration approach outlined earlier in Chapter 3. The main difference is the characterization of the objective function. Where deterministic model calibration considers the objective function as simply minimizing the squared error between model output trajectories and data, the IMCI approach to fitting distributions may require more complex cost calculations involving the differences in time-dependent variances between simulation estimates and data, or using a completely different form of analysis such as Bayesian maximum likelihood estimation [245]. Nevertheless, the basic calibration procedure remains the same:

1. Choose an initial estimate of the parameters characterizing the distribution of initial conditions or parameters (e.g. mean and variance of the underlying pdf)

2. Use forward Monte Carlo simulations to estimate the characteristics of the simulated output

3. Compare the estimate to the measured characteristics

4. Choose a new estimate, using an appropriate optimization algorithm

5. Repeat until convergence, or model adjustment

133
CHAPTER 6. VARIATION

This process no doubt requires a very large computation burden, as it entails the Monte Carlo simulation of a large cell population. Luckily, however, it is highly parallelizable as each cell is presumed to be independent of the others. Similar to the deterministic model calibration, this problem generally requires a stochastic optimization routine. In this case, the value of the objective function is completely unknown a priori and has to be evaluated through Monte Carlo simulation. Evaluation of the objective function is thus very noisy, as output metrics can vary between different instances of the same simulation parameters.

6.4 Population Interactions

Not all variation is random. Recent evidence has highlighted the importance of cell-context (the microenvironment and interactions with neighboring cells) in determining the precise behavior of individual cells. This effect is deterministic and cell-cell variation therefore largely results from dynamics which are presently un-modeled. As explained by Snijder and Pelkmans [246]: “Variations are largely determined by the inherent properties of growing cell populations that create a large spectrum of microenvironmental differences to which cells adapt, combined with the non-genetic memory of phenotype states and protein levels. As soon as a single cell starts to divide, whether it is a bacterium, a yeast cell or a mammalian cell, differences in cell-cell contacts and the available space will arise among the single cells, even in experimen-
CHAPTER 6. VARIATION

tal settings where culture conditions are kept constant.” A true model of a growing cell population therefore needs to incorporate both the stochastics of gene regulation, protein interactions, and unequal cell division as well as the deterministic effects of cell-cell interactions and microenvironmental differences. This requires building on the core cell cycle module developed in this dissertation to include relevant upstream signaling pathways and mechanisms of cell population and microenvironmental actions on individual cells.

6.5 Concluding Thoughts

The origin of measured variation is multifactorial and poorly quantified. Many applicable methodologies exist for modeling these processes, and the Inverse Monte Carlo by Interaction method provides a practical, but inelegant method to calibrate population models. However, more information on the relative importance and contributions of each noise source on the data variation is required. Better understanding of variability will be greatly improved when we’re better able to quantify the balance between deterministic and stochastic sources of variability and the relative contributions of various stochastic processes.
Chapter 7

Conclusions and Future Work

7.1 Thesis Review

Cell cycle modeling has a deep history. However, an extensive review of the literature revealed a lack of significant research coupling dynamic expression profiles to computational models. Modeling the cell cycle is complicated by the intrinsic difficulty of reliably measuring oscillating proteins in asynchronous populations. A new methodology provided improved estimates of the dynamic expression profiles of cell cycle regulatory proteins. This new data was not well correlated to the outputs of previously published models and calibration of numerical rate constants did not improve this fit. Reconciling a model to data requires a re-analysis of modeling assumptions. Including additional well-established and hypothesized regulatory interactions was shown to be significant in improving the model's output. This new model represents
CHAPTER 7. CONCLUSIONS AND FUTURE WORK

the first complete model of the mammalian cell cycle, spanning all four phases, which is calibrated to quantitative dynamic profiles. Many of the required modifications are supported by experimental evidence, and this work therefore illustrates the useful of modeling in integrating previously discovered interactions and testing biological hypothesis against experimental data. We’ve shown existing hypotheses of the dominant relations underlying the cell cycle to be inadequate to explain the temporal dynamics of measured proteins. This process also led us to introduce a hypothetical regulatory feature: the cyclin A/cdk activation of a transcription factor which increases the rates of both cyclin A and cyclin B synthesis upon entry to G2 phase. This hypothesis represents one plausible mechanism to improve the model fit, which remains to be tested by biological experiments. This then illustrates the usefulness of modeling in generating new, testable hypotheses. Finally, having shown the usefulness of single-cell measurements for deriving the median expression profile of asynchronously oscillating molecules, we turned our attention to the measured variation of the population about the median trajectory. We reviewed possible origins of this variation and applicable modeling techniques. Practical application of these methods however will likely require better understanding of the fundamental mechanisms underlying biological stochasticity and more precise estimates of the effects of epitope tagging on measurement error.
CHAPTER 7. CONCLUSIONS AND FUTURE WORK

7.2 Opportunities for Future Research

7.2.1 Improving the Model

Advances in cytometry will soon allow the simultaneous measurement of dozens of markers. Using the methodology outlined in Chapter 4, the number of extracted dynamic expression profiles is limited only by the availability of antibodies. This then establishes an extensible system where large numbers of cell cycle regulatory molecules can be measured and compared to the model developed in this dissertation. Similar to the process illustrated in Chapter 5, this will likely involve updating and extending the model to better improve the integrated working hypothesis of cell cycle regulation.

Other opportunities exist for validation of model-driven hypotheses through biological experimentation. Such experiments include validating the hypothesized G2 transcription factor as well as more fundamental assumptions of the basic wiring diagram. Similar to the work done on yeast [29], possibilities exist for inducing cell cycle checkpoints in vitro and checking the correlation of the resulting behavior with that predicted in silico.

7.2.2 Fundamental Research

The cell cycle control system is but one subsystem of the larger network controlling cell phenotype. As discussed, mammalian cells only enter new rounds of division
in response to significant growth factor signaling and can be halted at distinct points in cell cycle progression in response to anti-growth signaling. Despite this known regulation, no published models have investigated the integration of upstream signaling with the cell cycle control core invested here. A natural extension of this work therefore is the integration of previously derived upstream signaling modules (such as the critical MAPK signaling pathway [82]) with the new cell cycle module developed here. Additional opportunities exist for integrating other checkpoint controls, such as the DNA damage or spindle assembly checkpoint.

Furthermore, intracellular signaling pathways represent one level of the biological hierarchy. As discussed in Chapter 6, cell-to-cell communication is crucial in determining the behavior of individual cells. A more accurate understanding of cell growth regulation therefore will require integrating the cell cycle control and upstream signaling pathway models with feedback mechanisms underlying population growth.

### 7.2.3 Application to Cancer Research

A final opportunity is in cancer research. Cancers exhibit dysregulated growth [10]. The cell cycle regulatory system is no doubt central. An interesting extension of the basic modeling work presented here would be the calibration of models to normal and diseased cell types. Such comparison would provide an opportunity for exploring hypothesized or known mutations in the cancer cell cycle and the in silico exploration of resulting phenotypes. Having constructed a cancer model, it would also be possible
CHAPTER 7. CONCLUSIONS AND FUTURE WORK

to conduct in silico explorations of the effects of known or hypothetical anti-cancer treatments.

In the long term, it may be possible to use modeling to interpret patient samples. Hematopoietic cancers provide easy sampling of cancer cells and flow cytometric measurement of cell cycle regulatory molecules. The methodology outlined in Chapter 4 also illustrates the ability to derive dynamic expression profiles of cell cycle molecules from one snapshot of a patient’s blood sample. This therefore provides a unique entry point for studying the molecular regulation of hematopoietic cancers and the response of individual patients to treatment.
Appendix A

Computational Models Developed in this Dissertation
APPENDIX A. COMPUTATIONAL MODELS DEVELOPED IN THIS DISSERTATION

A.1 Section 5.2: Original Data Set

A.1.1 Equations

\[
\frac{d(ppRB)}{dt} = v_{29} + v_{30} + v_{43} - v_{44}
\]
\[
\frac{d(E2F)}{dt} = v_{29} + v_{45} + v_{47} - v_{46} - v_{48}
\]
\[
\frac{d(pE2F)}{dt} = v_{30} + v_{49} + v_{46} - v_{47} - v_{50}
\]
\[
\frac{d(RB)}{dt} = v_{44} + v_{45} + v_{49} - v_{48} - v_{50} - v_{43}
\]
\[
\frac{d(E2FRB)}{dt} = v_{51} + v_{48} - v_{52} - v_{29} - v_{45}
\]
\[
\frac{d(pE2FRB)}{dt} = v_{52} + v_{50} - v_{51} - v_{30} - v_{49}
\]
\[
\frac{d(ERG)}{dt} = k_{15}/(1 + (DRG/J15)^2) - k_{16} \ast ERG
\]
\[
\frac{(d(DRG))}{dt} = k_{17}p \ast ERG + k_{17} \ast (DRG/J17)^2/(1 + (DRG/J17)^2) - k_{18} \ast DRG
\]
\[
\frac{d(actCycD)}{dt} = k_{9} \ast DRG + V_{di} \ast TriD + k_{24}r \ast TriD - k_{24} \ast actCycD \ast freeCI - k_{10} \ast actCycD
\]
\[
\frac{d(TriD)}{dt} = k_{24} \ast actCycD \ast freeCI - k_{24}r \ast TriD - V_{di} \ast TriD - k_{10} \ast TriD
\]
\[
\frac{d(actCycACdk1)}{dt} = a_{frac} \ast ((ksap + ksapp \ast E2F + ksappp \ast TFAB) \ast mass \ast 2 + (V_{di} + kdia) \ast TriA) - (V_{da} + kasa \ast freeCI) \ast actCycACdk1
\]
\[
\frac{d(actCycACdk2)}{dt} = (1 - a_{frac}) \ast ((ksap + ksapp \ast E2F + ksappp \ast TFAB) \ast mass \ast 2 + (V_{di} + kdia) \ast TriA) - (V_{da} + kasa \ast freeCI) \ast actCycACdk2
\]
\[
\frac{d(actCycB)}{dt} = V_{sb} \ast mass \ast 2 + V_{25} \ast (cycB - actCycB) + (V_{di}) \ast (cycB - preMF) - actCycB - (V_{db} + V_{wee}) \ast actCycB
\]
APPENDIX A. COMPUTATIONAL MODELS DEVELOPED IN THIS DISSERTATION

\[
\frac{d(\text{actCycE})}{dt} = (k\text{sep} + k\text{sepp} * E2F) * \text{mass} \times 2 + \\
(V\text{di} + k\text{die}) * \text{TriE} - (V\text{de} + k\text{ase} * f\text{reeCKI}) * \text{actCycE}
\]

\[
\frac{d(\text{cycA})}{dt} = (k\text{sap} + k\text{sapp} * E2F + k\text{sappp} * TFAB) * \text{mass} \times 2 \\
-V\text{da} * \text{cycA}
\]

\[
\frac{d(\text{cycB})}{dt} = V\text{sb} * \text{mass} \times 2 - V\text{db} * \text{cycB}
\]

\[
\frac{d(\text{cycE})}{dt} = (k\text{sep} + k\text{sepp} * E2F) * \text{mass} \times 2 - V\text{de} * \text{cycE}
\]

\[
\frac{d(\text{CKI})}{dt} = V\text{si} - V\text{di} * \text{CKI}
\]

\[
\frac{d(\text{Cdh1})}{dt} = \frac{(k\text{ah}1p + k\text{ah}1pp * \text{Cdc}20A) * (1 - C\text{dh}1)}{J\text{ah}1 + 1 - C\text{dh}1} \\
- \frac{(k\text{ih}1pp * (\text{actCycACdh1} + \text{actCycACdh2}) + k\text{ih}1pp * \text{actCycB} * \text{Cdh1}) * \text{Cdh1}}{J\text{ih}1 + C\text{dh}1}
\]

\[
\frac{d(\text{preMPF})}{dt} = V\text{wee} * (\text{cycB} - \text{preMPF}) - (V25 + V\text{db}) * \text{preMPF}
\]

\[
\frac{d(\text{TriA})}{dt} = k\text{sa} * (\text{cycA} - \text{TriA}) * \text{freeCKI} - (k\text{dia} + V\text{da} + V\text{di}) * \text{TriA}
\]

\[
\frac{d(\text{APCP})}{dt} = \frac{k\text{aAPC} * \text{actCycB} * (1 - \text{APCP})}{J\text{aAPC} + 1 - \text{APCP}} \\
- \frac{k\text{iAPC} * \text{APCP}}{J\text{aAPC} + \text{APCP}}
\]

\[
\frac{d(\text{Cdc}20A)}{dt} = \frac{k\text{a20} * \text{APCP} * (\text{Cdc}20T - \text{Cdc}20A)}{J\text{a20} + \text{Cdc}20T - \text{Cdc}20A} \\
- (k\text{i20} / (J\text{i20} + \text{Cdc}20A) + k\text{d20}) * \text{Cdc}20A
\]

\[
\frac{d(\text{Cdc}20T)}{dt} = (k\text{s20pp} * \text{actCycB}) / (J\text{20} + \text{actCycB}) - k\text{d20} * \text{Cdc}20T
\]

\[
\frac{d(\text{mass})}{dt} = u * \text{mass}
\]
APPENDIX A. COMPUTATIONAL MODELS DEVELOPED IN THIS DISSERTATION

A.1.2 Definitions

\[ v_{29} = E_{2}FRB \times (K_{20} \times ((actCycD + TriD) \times LD + LA \times (actCycACdk1 + actCycACdk2) + LB \times actCycB + LE \times actCycE)) \]

\[ v_{30} = pE_{2}FRB \times (K_{20} \times (actCycD \times LD + LA \times (actCycACdk1 + actCycACdk2) + LB \times actCycB + LE \times actCycE)) \]

\[ v_{43} = RB \times (K_{20} \times (actCycD \times LD + LA \times (actCycACdk1 + actCycACdk2) + LB \times actCycB + LE \times actCycE)) \]

\[ v_{44} = ppRB \times (K_{19}a \times (PP_{1}T - PP_{1}A) + K_{19} \times PP_{1}A) \]

\[ v_{45} = K_{26}R \times E_{2}FRB \]

\[ v_{46} = E_{2}F \times (K_{23a} + K_{23} \times (actCycACdk1 + actCycACdk2) + K_{23b} \times actCycB) \]

\[ v_{47} = K_{22} \times pE_{2}F \]

\[ v_{48} = K_{26} \times E_{2}F \times RB \]

\[ v_{49} = K_{26}R \times pE_{2}FRB \]

\[ v_{50} = K_{26} \times RB \times pE_{2}F \]

\[ v_{51} = K_{22} \times pE_{2}FRB \]

\[ v_{52} = E_{2}FRB \times (K_{23a} + K_{23} \times (actCycACdk1 + actCycACdk2) + K_{23b} \times actCycB) \]

\[ V_{atf} = katfpp \times (actCycACdk1 + actCycACdk2) + katfppp \times actCycE \]

\[ +katfpppp \times actCycD \]

\[ V_{de} = kdep + kdepp \times actCycE + kdeppp \times (actCycACdk1 + actCycACdk2) \]

\[ +kdepppp \times actCycB \]
APPENDIX A. COMPUTATIONAL MODELS DEVELOPED IN THIS DISSERTATION

\[ V_{da} = kdap + kacdh1 \cdot Cdh1 + kdapp \cdot Cdc20A \]

\[ TFAB = G(kafb \cdot (actCycACdk1 + actCycACdk2), kifb, Jafb, Jifb) \]

\[ V_{si} = ksip \]

\[ Cdk1CycBCKI = cycB - actCycB - preMPF \]

\[ Cdk1PCycB = cycB - actCycB \]

\[ V_{sb} = ksbp + ksbpp \cdot TFAB + ksbppp \cdot actCycB \]

\[ V_{db} = kdbp + kdbpp \cdot Cdh1 + kdbppp \cdot Cdc20A \]

\[ Wee1 = G(kawep, kiwee \cdot (actCycACdk1 + actCycACdk2) + kiweeb \cdot actCycB, Jaweew, Jiwee) \]

\[ V_{wee} = kweep + kweepp \cdot Wee1 \]

\[ Cdc25 = G(ka25 \cdot actCycB, ki25p, Ja25, Ji25) \]

\[ V25 = k25p + k25pp \cdot Cdc25 \]

\[ V_{di} = (kdip + kdipp \cdot (actCycACdk1 + actCycACdk2) + kdippp \cdot actCycB \]

\[ + kdipppp \cdot actCycE) \]

\[ TriE = cycE - actCycE \]

\[ freeCKI = CKI - TriA - TriE - TriD \]

\[ totalA = actCycACdk1 + actCycACdk2 + TriA \]

\[ totalB = actCycB + preMPF \]

\[ totalCKI = freeCKI + TriA + TriD + TriE \]

\[ totalD = actCycD + TriD \]

\[ PP1A = \frac{PP1T}{K21 \cdot (FE \cdot (actCycACdk1 + actCycACdk2 + actCycE) + FB \cdot actCycB + 1)} \]
APPENDIX A. COMPUTATIONAL MODELS DEVELOPED IN THIS DISSERTATION

Where \( G(...) \) is the Goldbeter-Koshland Function:

\[
B(A_1, A_2, A_3, A_4) = A_2 - A_1 + A_3 \ast A_2 - A_4 \ast A_1
\]

\[
G(A_1, A_2, A_3, A_4) = \frac{2 \ast A_4 \ast A_1}{B(A_1, A_2, A_3, A_4) + \sqrt{B(A_1, A_2, A_3, A_4) - 4 \ast (A_2 - A_1) \ast A_3 \ast A_1}}
\]
APPENDIX A. COMPUTATIONAL MODELS DEVELOPED IN THIS DISSERTATION

A.1.3 Kinetic Rate Constants

\[ K_{20} = 10 \]
\[ LA = 30 \]
\[ LB = 0.5 \]
\[ LE = 5.48239 \]
\[ LD = 3.3 \]
\[ K_{19} = 2 \]
\[ K_{19a} = 25 \]
\[ K_{26} = 1000 \]
\[ K_{26R} = 2 \]
\[ K_{23} = 0.01 \]
\[ K_{23a} = 0.0005 \]
\[ K_{22} = 1 \]
\[ PPIT = 1 \]
\[ FE = 25 \]
\[ FB = 2 \]
\[ K_{21} = 1 \]
\[ K_{23b} = 1 \]
\[ k_{15} = 0.172959 \]
\[ J_{15} = 0.1 \]
\[ k_{16} = 2.5 \]
\[ k_{17p} = 0.35 \]
\[ k_{17} = 100 \]
\[ J_{17} = 0.3 \]
\[ k_{18} = 10 \]
\[ k_{9} = 2.6 \]
\[ k_{24r} = 10 \]
\[ k_{24} = 1000 \]
\[ k_{10} = 5 \]
\[ J_{20} = 100 \]
\[ J_{a20} = 0.005 \]
\[ J_{aAPC} = 0.01 \]
\[ J_{afb} = 0.01 \]
\[ J_{a1} = 0.01 \]
\[ J_{af} = 0.01 \]
\[ J_{awee} = 0.05 \]
\[ J_{i20} = 0.005 \]
\[ J_{i25} = 0.0316228 \]
\[ J_{iAPC} = 0.001 \]
\[ J_{ifb} = 0.001 \]
\[ J_{ih1} = 0.01 \]
\[ J_{if} = 0.01 \]
\[ J_{iwee} = 0.05 \]
\[ k_{25p} = 3.4859 \]
\[ k_{25pp} = 1730.42 \]
\[ ka_{20} = 16.5959 \]
\[ ka_{25} = 0.622711 \]
\[ ka_{APC} = 0.132351 \]
\[ ka_{fb} = 11.1190 \]
\[ ka_{1pp} = 8.83872 \]
\[ ka_{1pp} = 100000 \]
\[ k_{asa} = 1119 \]
\[ k_{ase} = 1119 \]
\[ kat_{fpp} = 3.329 \]
APPENDIX A. COMPUTATIONAL MODELS DEVELOPED IN THIS DISSERTATION

\[ katfpp = 5.5462 \quad katfpppp = 4.40257 \quad kawee = 13.8188 \]
\[ kd20 = 1 \quad kdap = 0.0292653 \quad kdapp = 2000 \]
\[ kdbp = 0.04838 \quad kdbpp = 1.2 \quad kdbppp = 150 \]
\[ kdep = 0.1112 \quad kdepp = 1.1119 \quad kdeppp = 20 \]
\[ kdepppp = 500 \quad kdia = 11.1187 \quad kdie = 11.1187 \]
\[ kdp = 11.1187 \quad kdipp = 55.4618 \quad kdippp = 111.1901 \]
\[ kdippp = 55.4618 \quad k20 = 1.7 \quad ki25p = 1.82 \]
\[ kiAPC = 0.219011 \quad kfb = 0.55727 \quad kih1pp = 560 \]
\[ kih1pp = 10 \quad kfp = 2.7764 \quad kftp = 1.1119 \]
\[ kftp = 1.1119 \quad kiw = 0.1 \quad ks20pp = 6 \]
\[ ksapp = 0.045 \quad ksbp = 0.195 \quad ksbpp = 0.66896 \]
\[ ksep = 0.0886 \quad ksepp = 3.3290 \quad ksp = 22.1714 \]
\[ kweep = 13.3162 \quad kweepp = 334.487 \quad u = 0.039 \]
\[ ksap = 0.0300964 \quad kafbb = 0.1 \quad kifbb = 0.001 \]
\[ kafbb = 0.001 \quad ja = 0.01 \quad ksbpp = 0.1 \]
\[ kiweeb = 10 \quad Jaweeb = 0.05 \quad ksapp = 0.285 \]
\[ kafab = 0.08 \quad afrac = 0.081283 \quad kacdh1 = 10 \]

A.1.4 Initial Conditions

All state variables at time zero are equal to zero, except mass\( (t_0) = 1 \), Cdh1\( (t_0) = 1 \), ppRB\( (t_0) = 10 \), and pE2F\( (t_0) = 5 \).
A.2 Section 5.3 Reanalysis of the Data

A.2.1 Equations, Definitions, and Initial Conditions

All Equations and Definitions are identical to those in Appendix A.1, except for the following:

\[ V_{sb} = k_{sbp} + k_{sbpp} \ast TFAB + k_{sbppp} \ast actCycB + k_{sbpppp} \ast E2F \]

\[ \frac{d(E2F)}{dt} = v29 + v45 + v47 - v46 - v48 + ke2f \ast E2F \ast mass \]
APPENDIX A. COMPUTATIONAL MODELS DEVELOPED IN THIS DISSERTATION

A.2.2 Kinetic Rate Constants

\[ K^{20} = 4.47 \quad LA = 30 \quad LB = 0.5 \]
\[ LE = 5.48239 \quad LD = 3.3 \quad K^{19} = 0.894 \]
\[ K^{19a} = 11.175 \quad K^{26} = 447 \quad K^{26R} = 0.894 \]
\[ K^{23} = 0.00 \quad K^{23a} = 0.0002235 \quad K^{22} = 0.447 \]
\[ PP^{1T} = 1 \quad FE = 25 \quad FB = 2 \]
\[ K^{21} = 0.447 \quad K^{23b} = 0.0894 \quad k^{15} = 0.1341 \]
\[ J^{15} = 0.0447 \quad k^{16} = 1.1175 \quad k^{17p} = 0.0894 \]
\[ k^{17} = 67.05 \quad J^{17} = 0.3 \quad k^{18} = 4.47 \]
\[ k^{9} = 1.1622 \quad k^{24r} = 4.47 \quad k^{24} = 447 \]
\[ k^{10} = 2.235 \quad kaf^{bb} = 0.0447 \quad kif^{bb} = 44700 \]
\[ Jif^{bb} = 0.001 \quad Ja^{fbb} = 0.01 \quad ksbppp = 0.0447 \]
\[ kiweeb = 2.235 \quad Jaweeb = 0.05 \quad J^{20} = 100 \]
\[ Ja^{20} = 0.005 \quad Ja^{25} = 0.005 \quad Ja^{APC} = 0.01 \]
\[ Ja^{fb} = 0.01 \quad Jah1 = 0.01 \quad Jat^{f} = 0.01 \]
\[ Jawee = 0.05 \quad Ji^{20} = 0.005 \quad Ji^{25} = 0.0316228 \]
\[ Ji^{APC} = 0.001 \quad Ji^{fb} = 0.001 \quad Ji^{h1} = 0.01 \]
\[ Ji^{tf} = 0.01 \quad Ji^{wee} = 0.05 \quad k^{25p} = 1.5582 \]
APPENDIX A. COMPUTATIONAL MODELS DEVELOPED IN THIS DISSERTATION

\[
\begin{align*}
    k^{25pp} &= 773.4977 & ka^{20} &= 7.4184 & ka^{25} &= 0.2784 \\
    kaAPC &= 0.0592 & ka \_f b &= 4.9702 & kah1p &= 3.9509 \\
    kah1pp &= 44700 & kasa &= 500.193 & kase &= 500.193 \\
    katfpp &= 1.4881 & katfppp &= 2.4792 & katfpppp &= 1.9679 \\
    kaweep &= 6.177 & kd20 &= 0.447 & kdap &= 0.0131 \\
    kdapp &= 894 & kdbp &= 0.0216 & kdbpp &= 0.2235 \\
    kdbppp &= 84.93 & kdep &= 0.0497 & kdepp &= 0.497 \\
    kdeppp &= 8.94 & kdepppp &= 223.5 & kdia &= 4.9701 \\
    kdie &= 4.9701 & kdi p &= 4.9701 & kdi p p &= 24.7914 \\
    kdi p p p &= 49.702 & kdi p p p p &= 24.7914 & ki20 &= 0.7599 \\
    ki25p &= 0.894 & kiAPC &= 0.0979 & kifb &= 0.2491 \\
    kih1pp &= 187.74 & kih1ppp &= 134.1 & kitfp &= 1.2411 \\
    kitfpp &= 0.4970 & kitfppp &= 0.4970 & kiwee &= 0.00447 \\
    ks20pp &= 2.6820 & ksapp &= 0.0018 & ksbp &= 0.1565 \\
    ksbpp &= 0.2235 & ksep &= 0.0396 & ksepp &= 1.4881 \\
    ksip &= 9.9106 & kweep &= 5.9523 & kweepp &= 447 \\
    u &= 0.0174 & ksap &= 0.1565 & ksapp &= 0.2235 \\
    kafab &= 0.0375 & a1frac &= 0.0812831 & kacd1 &= 4.47 \\
    ksbpppp &= 0.0224 & ke2f &= 0.1788
\end{align*}
\]
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


[27] S. Li, P. Brazhnik, B. Sobral, and J. J. Tyson, “A quantitative study of the divi-
BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


[57] ——, “Modeling the mitotic oscillator driving the cell division cycle,” *Communications in Theoretical Biology*, vol. 3, pp. 75–107, 1993.
BIBLIOGRAPHY


[64] A. Bertuzzi, D. D. Meo, and A. Gandolfi, “An oscillator for the control of entry into s and m phases of cell cycle,” in *Complexity in the Living: A Modelistic*
BIBLIOGRAPHY


BIBLIOGRAPHY


[76] K. Iwamoto, Y. Tashima, H. Hamada, Y. Eguchi, and M. Okamoto, “Mathematical modeling and sensitivity analysis of g1/s phase in the cell cycle in-


BIBLIOGRAPHY


BIBLIOGRAPHY


[102] M. Kimmel and D. Axelrod, “Unequal cell division, growth regulation and
BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY

[148] T. Tsai, Y. S. Choi, W. Ma, J. R. Pomerening, C. Tang, and J. E. F. Jr.,
“Robust, tunable biological oscillations from interlinked positive and negative

[149] B. D. Aguda, “A qualitative analysis of the kinetics of the g2 dna damage
checkpoint system,” *Proceedings of the National Academy of Sciences USA*,

[150] C. D. Thron, “Bistable biochemical switching and control of the events of the

[151] Z. Qu, W. R. MacLellan, and J. N. Weiss, “Dynamics of the cell cycle: Check-


[153] L. Yang, W. R. MacLellan, Z. Han, J. N. Weiss, and Z. Qu, “Multisite phos-
phorylation and network dynamics of cyclin-dependent kinase signaling in the

[154] D. Battogtokh and J. J. Tyson, “Periodic forcing of a mathematical model of

BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


[175] M. Farina, R. Findeisen, E. Bullinger, S. Bittanti, F. Allgower, and P. Wellstead, “Results towards identifiability properties of biochemical reaction networks,” in
BIBLIOGRAPHY

Proceedings of the 45th IEEE Conference on Decision and Control, San Diego, CA, USA, December 2006.


BIBLIOGRAPHY


BIBLIOGRAPHY


[201] C. H. Yam, R. W. Ng, W. Y. Siu, A. W. Lau, and R. Y. Poon, “Regulation of


[208] E. Cheung, P. Mayr, F. Coda-Zabetta, P. G. Woodman, and D. S. Boam,


BIBLIOGRAPHY


BIBLIOGRAPHY


[233] A. Loewer and G. Lahav, “We are all individuals: Causes and consequences of


[239] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O’Shea, Y. Pilpel, and
BIBLIOGRAPHY


[245] A. Y. Weisse, R. H. Middleton, and W. Huisinga, “Quantifying uncertainty,